

Data Analytics for Crop Recommendation System

Submitted by

Riddhi Patel

1421013

In the partial fulfillment of the requirements for the degree of

Master of Technology

in

Computer Science and Engineering (CSE)

to



AHMEDABAD UNIVERSITY

School of Engineering and Applied Science, Ahmedabad University

Ahmedabad, India

June 2016

Acknowledgements

It is obvious that a good work need support and encouragement of many people. I would like to express my sincere gratitude towards my thesis supervisor, Prof. Sanjay Chaudhary for all his patient guidance, encouragement and help throughout the period of research to giving a proper direction to my efforts. His constant inspiration and encouragement along with his valuable guidance has been instrumental in the successful completion of this research.

I am thankful to Mrs. Purnima Shah and Mr. Deepak Hiremath for their continuous support, good advice, and kind help to me at various stages. Also, I would like to thank to the faculty and staff of the Department of CSE for creating the beautiful academic environment.

I am thankful to all my friends of Mtech 2014 Batch for helped me get through some tough times. I would also like to thank to my parents and my family for their constant support, encouragement.

Riddhi D. Patel

Dedication

I would like to dedicate my thesis to my sweet loving Parents, whose affection, love, encouragement make me able to get such success.

Declaration

I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, contains no material previously published or written by another person nor material which has been accepted for the award of any other degree or diploma of the university or other institute of higher learning, except where due acknowledgment has been made in the text.



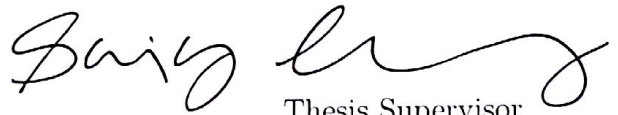
Signature of Student

Riddhi Patel

1421013

Certificate

*This is to certify that the thesis entitled **Data Analytics for Crop Recommendation System** submitted by Ms. Riddhi Patel (1421013) to the School of Engineering and Applied Science (SEAS), Ahmedabad University towards partial fulfilment of the requirements for the award of the Degree of Master of Technology in Computer Science Engineering, is a bonafide record of the work carried out by her under my supervision and guidance.*



Thesis Supervisor

Prof. Sanjay Chaudhary

Date 30th June, 2016

Place Ahmedabad

Contents

Acknowledgements	ii
Dedication	iii
Declaration	iv
Certificate	v
Abstract	ix
List of Symbols	x
List of Abbreviations	xi
List of Figures	xii
List of Tables	xiii
1 Introduction	1
1.1 Background	3
1.1.1 Agriculture in Gujarat	3
1.2 Profile of Study Area: Ahmedabad District	4
1.3 Motivation	6
1.4 Problem Statement	7
1.5 Organisation of the Thesis	8
2 Literature Review	9
2.1 Related Work	9

3	Research Methodology	13
3.1	Theoretical Framework	13
3.2	Measurement Framework	13
3.2.1	Concepts of Statistics	13
3.2.2	Linear Regression Analysis	14
3.2.3	Time series Analysis	17
3.3	Proposed Approach	19
3.3.1	Data Collection	20
3.3.2	RESTful Web Services	21
3.3.3	Data Analytics and Data Visualization	21
3.3.4	User Interface	22
3.4	Summary	23
4	Experimental Data	24
4.1	General	24
4.2	Data Organization and Use	25
4.3	Summary	25
5	Results	26
5.1	Experiment Environment	26
5.1.1	Hardware usage	26
5.1.2	Software usage	26
5.2	Results	27
5.2.1	Past Agro-Production Trend	29
5.2.2	Weather indices based MLR model	29
5.2.3	Seasonal Analysis	33
5.2.4	Time Series Analysis	35
6	Conclusion and Future work	37
6.1	Conclusion	38
6.2	Future Work	38
6.3	Limitations	39

Appendix A Experiment Datasets	42
A.1 Monthly Rainfall and Temperature Data for Ahmedabad District . . .	42
A.2 Cotton Crop Production Data for Ahmedabad District	43
A.3 Cotton Crop wholesale Monthly Price Data for Ahmedabad District .	43
A.4 Cotton Crop Seasonal Climate Parameters Data for Ahmedabad Dis- trict	44

Abstract

Agriculture is becoming increasingly information and knowledge centric today. Due to the large rural population, agriculture plays a vital role in Indian economy. In the current scenario, a large number of data is generated from various sources like weather, climate, geo-spatial, crop production, consumed by stakeholders, location specific crop disease in farm practice. But it is not used effectively and optimally by the experts due to lack of information flow. Thus, to bridge the gap between users and information, data analytics can be one of the solution.

Crop recommendation system model integrating with data analytics has been proposed. The system consist of components; web services, data analytics, and web application development. The RESTful weather and agriculture web services were built to interaction with various data sources. The web services are developed using JAX-RS in NetBeans IDE. Regression Analysis and Time Series Analysis are used to analyses the trends and pattern of agriculture Growth and Production. Crop Recommendation System is carried out for cotton crop in Ahmedabad District, Gujarat. The proto type is developed using MySQL, Java, NetBeans IDE, and RStudio.

List of Symbols

Y_i	Dependent Variable
X_i	Independent Variable
α_i	Regression Coefficients
β_i	Autoregressive Coefficients
θ_i	Moving Average Coefficients
ϵ_i	Error
R_i	Correlation Coefficient
p	Order of Autoregressive
d	Order of Integration
q	Order of Moving Average

List of Abbreviations

ICT	Information Communication Technology
GDP	Gross Domestic Product
MLR	Multiple Linear Regression
AR	Autoregressive
MA	Moving Average
ARIMA	Autoregressive Integrated Moving Average
ACF	Auto Correlation Function
PACF	Partial Auto Correlation Function
AIC	Akaike's Information Criterion
SBC	Bayesian Information Criterion
REST	REpresentational State Transfer
HTTP	Hypertext Transfer Protocol
URI	Uniform Resource Identifier
JSON	JavaScript Object Notation
IDE	Integrated Development Environment
NPK	Nitrogen, Phosphorus and Potassium fertilizer

List of Figures

1.1	Cotton crop trends in area, production and yield, Ahmedabad, 1995-2011	5
1.2	Variation in Rainfall and Temperature, Ahmedabad, 1995-2011	6
3.1	Proposed Architecture of Data Analytics for Crop Recommendation System	20
5.1	RESTful Web service for Cotton Crop	27
5.2	Homepage of Crop Recommendation System	28
5.3	Cotton crop diseases information	28
5.4	Area & Production of Cotton crop, Ahmedabad from year 1995-2011	29
5.5	Cotton crop yield prediction based on temperature	30
5.6	Cotton crop yield prediction based on temperature and rainfall . . .	32
5.7	Cotton crop yield prediction based on seasonal Analysis	34
5.8	Cotton crop wholesale monthly market price forecast	36

List of Tables

1.1	Main Agricultural crops of Gujarat	4
4.1	Sources of Dataset	24
5.1	Temperature based MLR Analysis	30
5.2	Prediction of temperature based MLR analysis for cotton yield	31
5.3	Temperature and Rainfall based MLR Analysis	31
5.4	Prediction of temperature and rainfall based MLR analysis for cotton yield	32
5.5	Result of Seasonal Analysis	33
5.6	Prediction of seasonal analysis for cotton yield	34
5.7	Forecast result based on the fitted ARIMA model	35
6.1	Weather indices based MLR experimental result	37
A.1	Monthly Rainfall and Temperature, Ahmedabad, 1995-2011	42
A.2	Cotton Area, Production and Yield, Ahmedabad, 1995-2011	43
A.3	Cotton wholesale monthly price, Ahmedabad, 2010-2015	43
A.4	Cotton crop seasonal climate parameters, Ahmedabad, 1995-2011 . .	44

Chapter 1

Introduction

In the world, data can be available from web logs, sensor network, Internet texts and Documents, internet search indexing, mobile devices, social networking. Every day 2.5 quintillion bytes of data are created according to the estimation done by IBM and it is very large amount so the 90% of data in the world has been created in last 2 years ¹ .

Data Science is the extraction of knowledge from data. Hal Varian, Google's Chief Economist, NYT, 2009 define Data Science is "The ability to take data, to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it - that's going to be a hugely important skill". Jeffrey Staton, Syracuse University School of Information Studies define data science is " Data Science refers to an emerging area of work concerned with the collection, preparation, analysis, visualization, management and preservation of large collection of information".

Data analytics is the process of transforming raw data into usable information, often presented in the form of a published analytical article, in order to add value to the statistical output. Big data analytics is the process of examining big data to discover hidden patterns, unknown correlations and other useful information that can be used to make better decisions. With big data analytics, data scientists and others can analyze huge volumes of data. Analyzing big data allows analysts, researchers, and business users to make better and faster decisions using data that was previously inaccessible or unusable. Using advanced analytics techniques such

¹Improving Decision Making in the World of Big Data: <http://www.forbes.com/sites/christopherfrank/2012/03/25/improvingdecision-making-in-the-world-of-big-data/>

as text analytics, machine learning, predictive analytics, data mining, statistics, and natural language processing, businesses can analyze previously untapped data sources independent or together with their existing enterprise data to gain new insights resulting in significantly better and faster decisions ². There are mainly three types of data analytics their specification is discussed below.

A. Predictive Analytics

Predictive analytics use data to identify historical patterns to predict the future. Predictive analytics provide estimates about the likelihood of a future outcome. For example, some companies are using predictive analytics for sales lead scoring. Some companies have gone one step further use predictive analytics for the entire sales process, analyzing lead source, number of communications, types of communications, social media, documents, CRM data, etc. Properly tuned predictive analytics can be used to support sales, marketing, or for other types of complex forecasts.

B. Descriptive Analytics

Descriptive analytics or statistics does exactly what the name implies them “Describe”, or summarize raw data and make it something that is interpretable by humans. Descriptive statistics are useful to show things like, total stock in inventory, average dollars spent per customer and Year over year change in sales. Common examples of descriptive analytics are reports that provide historical insights regarding the company’s production, financials, operations, sales, finance, inventory and customers.

C. Prescriptive Analytics

Prescriptive analytics allows users to “prescribe” a number of different possible actions to and guide them towards a solution. Prescriptive analytics automatically synthesizes big data, mathematical sciences, business rules, and machine learning to make predictions and then suggests decision options to take advantage of the predictions. For example, in the health care industry, you can better manage the patient population by using prescriptive analytics to measure the number of patients who are clinically obese, then add filters for factors like di-

²http://www.sas.com/en_us/insights/analytics/big-data-analytics.html

abetes and LDL cholesterol levels to determine where to focus treatment. The same prescriptive model can be applied to almost any industry target group or problem.

1.1 Background

Due to the large rural population, agriculture plays a vital role in Indian economy. Agriculture development is essential part for the overall development of the economy. Agriculture provides rewarding employment and livelihood for majority of the population in India. At the time of Independence about 72% of working population was engaged in agriculture activity in India and the share of Agriculture and allied activities was 51.45% to total GDP of India. Even today, 49% of population is working in agriculture and allied sector (As per Economic survey 2013-14). According to 2011 census 68.8% of working population still engaged in agriculture activity and the share of agriculture and allied activity was only 14% to total GDP ³.

Present research is to analyses the trends and pattern of agriculture growth and Production in Gujarat. In the agricultural domain, the bother is to present the latest information and research to the agriculture experts and the farmers so that they can strengthen the power of ICT to improve their agricultural. In the current scenario, a large amount of data relating to the agricultural domain is collected from different sources like Government of India Reports, Directors of economics and statistics Gujarat reports, Ministry of Agriculture reports, books, articles, and Economic Survey of India. But it is not used effectively and optimally by the experts due to lack of information flow.

1.1.1 Agriculture in Gujarat

A huge variety of food grain like Wheat, Paddy, Bajara, Maize etc. and major pulses like Pigeon pea, Gram, Green gram and major oilseeds crops like Cotton, Castor, Ground nut, Mustard are sown in Gujarat. Major agriculture crops of Gujarat are in below Table 1.1.

³Website of planning commission: http://planningcommission.gov.in/hackathon/index.php?sector=Agriculture_and_Rural_Development

Crop Seasons	Crops Name
Main Kharif crops	Cotton, Groundnut, Castor, Paddy, Bajara, Maize, Green Gram
Main Ravi(Summer) Crops	Wheat, Rice, Mustard, Gram, Groundnut, Bajara, Sugarcane
Main Vegetable	Onion, Potato, Brinjal, Tomato, Cabbage, Cauliflower
Main Spices	Cumin, Fennel Garlic, Chilly, Coriander, Ginger, Turmeric, Fenugreek, Ajawan and Suva

Table 1.1: Main Agricultural crops of Gujarat

Cotton is important oilseed crop of Gujarat. Gujarat produces cotton variety like SHANKAR-6, B.T. COTTON, and V - 797 which provides lint. This lint provides high quality fiber to textile industry, a high protein meal to livestock feed, oil for human consumption, byproducts used as fertilizer, produce paper, cardboard, etc. In Gujarat cotton cropping season starts from late June to early July and ends towards the end of January to early February.

1.2 Profile of Study Area: Ahmedabad District

Ahmedabad District is a central part of Gujarat state. Ahmedabad is an industrial hub for textiles industries. The major crops of Ahmedabad are Cotton, Rice, Wheat, Jowar, Bajra, Maize, Gram, Groundnut etc. Cotton is a kharif crop which requires 6-7 months to cultivate. There are many varieties of cotton grown like Assam Comilla, Shanker 6 (B) 30mm FIne, V-797 22mm FIne in Ahmedabad District. Average annual rainfall 600-800 mm except for the years 2002 as Gujarat was affected by drought⁴. Minimum, Maximum, and Average temperatures in Ahmedabad District from year 1995 to 2011 are 18 °C, 31 °C and 26 °C respectively. Average relative humidity is 60 percent. Average cotton yield from 1995 to 2011 is 1.5 Bales/Hectare, whereas the average area of cultivation is 179206 in Hectare and average production is 251388 in Bales as shown in below figure 1.1.

⁴Web site of Gujarat Online: <http://www.gujaratonline.com/newsroom/drought/>

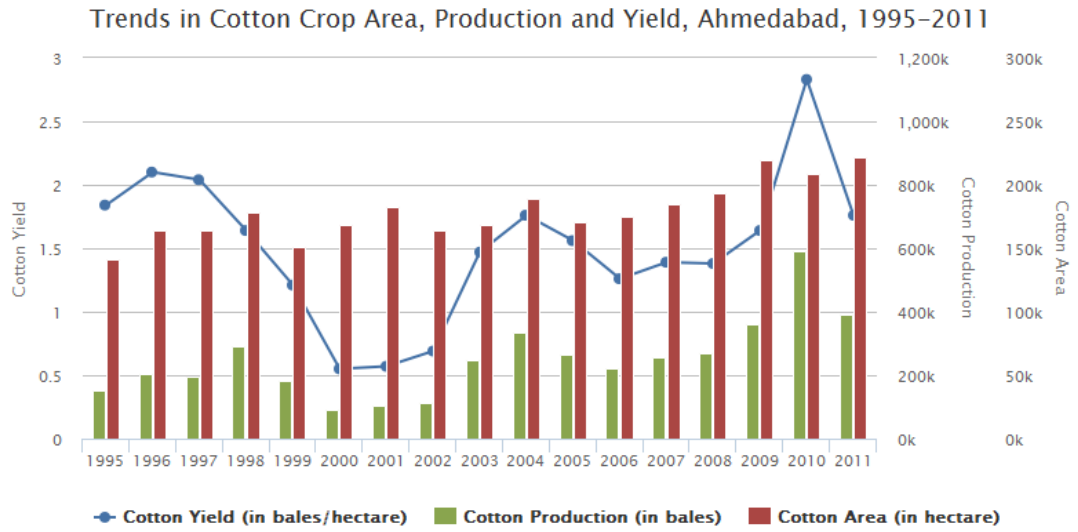


Figure 1.1: Cotton crop trends in area, production and yield, Ahmedabad, 1995-2011

Figure 1.2 shows the high variability in annual rainfall and temperature from year 1995 to 2011 in Ahmedabad District. High fluctuations in weather parameter like rainfall and temperature affects the crop production. As a result of high fluctuations in weather condition farmers are facing problems like low crop production or loss of crop. Less seasonal rainfall will not much affected crop production as Ahmedabad is being irrigated district. But more rainfall than the optimal seasonal rainfall will affect the crop production. Also uncertainty in monsoon rains and in temperature affects cotton production. These will create a problems to farmers. One of the solutions is to analyses the cotton productivity with respect the weather parameters like temperature, rainfall, relative humidity and forecast cotton productivity before sowing of cotton.

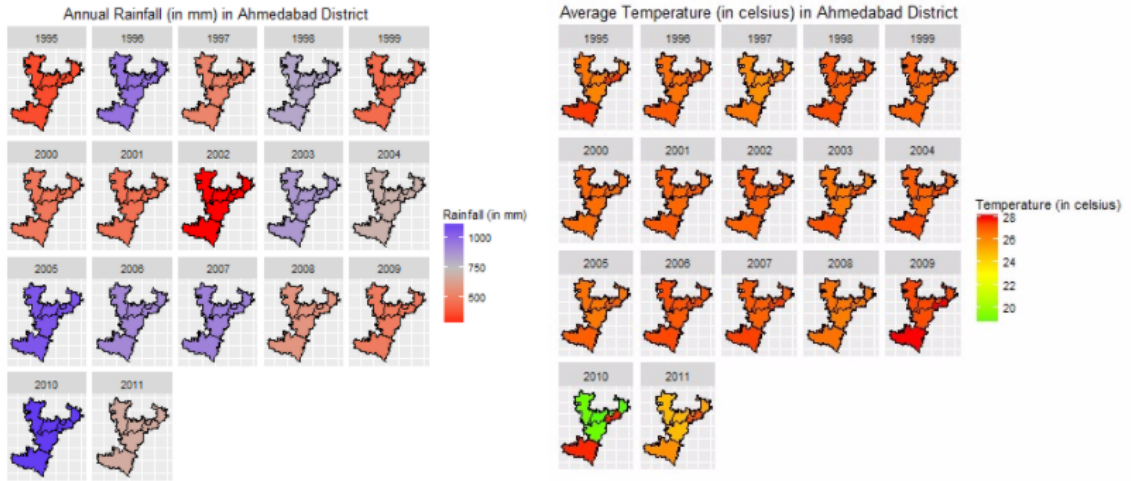


Figure 1.2: Variation in Rainfall and Temperature, Ahmedabad, 1995-2011

1.3 Motivation

Agriculture has become highly input and cost intensive. Crop production will become more difficult with weather change and environmental degradation. In various weather scenario forecasting of agriculture is essential. Reliable and timely delivery of forecast of agriculture is necessary for farmers. By using this forecast farmers will be able to do planning especially under uncertainties of weather, production, prices, etc.

Before each cropping period begins farmers are in a bother which crop they will grow because farmers do not have information for the outcome of their crops before the cropping season. So, the expectations for yield and price would help the farmers to estimate the expected return for each crop. The crop yield and crop market price forecasting are required for government organizations, farmers. They can make appropriate decisions pertaining to crop sowing, fertilizer uses and pesticide uses, decisions related to storage, price support, etc. on such forecasts.

Hence, forecast modeling is beneficial for farmers as it may increase crop production and more returns for a particular crop. An increase in agricultural productivity leads to higher incomes from low cost. Based on such observations, we are motivated to develop the data analytics for a crop recommendation system that provides recommendation to farmers to improve farm productivity.

1.4 Problem Statement

Indian Agriculture sector requires innumerable types of data analytics in various sectors such as crop productivity prediction models, economic models, pest and crop disease prediction models, crop price forecasting models, etc. The frequent changes in climate conditions are affecting more in cotton production. Most of the forecasts are seasonal and are available around 1-2 months before the crop harvesting. Farmers are benefited if recommendation and forecast of particular crop are available before sowing of crop.

Contribution of this research is to improve the agricultural productivity and provide the crop recommendation to farmers in North Gujarat region.

The objectives of this research are:

1. Weather indices based Regression Analysis – To analyses relationship of crop yield ,monthly average temperature and monthly average rainfall in Ahmedabad District, Gujarat.
2. Seasonal Analysis – To analyses crop yield with respect to seasonal weather parameters in Ahmedabad District, Gujarat.
3. Time Series Analysis – To analyses and forecast the wholesale monthly market price of cotton crop in Ahmedabad District, Gujarat.

1.5 Organisation of the Thesis

The thesis is organized as follows.

Chapter 1 Basic introduction, background of research problem, which motivated us to do this research work is highlighted.

Chapter 2 Literature Survey, Research work done related to the stated problem.

Chapter 3 Theoretical Framework, Implementation Strategy, Proposed approach to address the problem statement.

Chapter 4 Data sources, organization of data used in the research work.

Chapter 5 Discusses the results obtained from the regression analysis and time series analysis.

Chapter 6 Conclusion and Future Work, describes the overall conclude of the research work and further Future Implementation.

Chapter 2

Literature Review

2.1 Related Work

In Agriculture domain especially in India, farmers are worried about making decision during the cropping season. So, that they get high yield productivity of crop in cropping season is aim of farmers. To achieve this, it should require recommendation from the experts so depending on that recommendation farmers can make decision related to crop cultivation. Number of research carried out in this section by agriculture institutes and agriculture research center. But in most of case recommendation do not reach to farmers due to lack of communication channel between researchers and farmers. To solve this problem some recommendation or advisory system are making exists to farmers. The Recommendation system have set of input dataset which is taken out from previous years actual data, after that they begin analysis. They process the data, step by step as directed by algorithm to get the result and make a conclusion. In this algorithm plays important role. So, knowledge is represented by the results of algorithm.

Previous research on crop recommendation system involved many methods or techniques over the time periods. An important findings' summary are discussed below.

IIIT Hyderabad and Media labs Asia, developed IT-based agro advisory system which is known as eSagu. It is helpful for the farmer to improve crop productivity by giving advice or recommendation to farmer. Agriculture experts put up their advices to each farmer in regular time period by knowing crop growth status which

is sent by the farmer in the form of text or photographs [1].

Agrisnet is a web based agriculture portal. Agrisnet provides the various kinds of information like seed, plant protection, fertilizer uses, soil health card, weather etc. to the registered farmers. This information will helpful for farmers to improve farm productivity [2].

The Karshaka Information Systems, Services and Networking system is developed during the Kissan Kerela project. This system has different modules like crop information, market information, query handling, weather information, administration, general statistics and soil data. The main objective of the project is to increased farm productivity, obtain better returns and improvement on diseases. Agrouusers can access information via online video channels, local TV networks and internet web portal [3].

mKrishi is developed by TCS group. It is a mobile based Agro Advisory system. It connects the farmer with ecosystem to make them accurate decision in their farming. It provides video-audio facilities on a mobile phone to express queries to experts. Farmers can send their queries by sending the pictures of crops or audio clip. This system provides the climate information of location's weather stations. After analyzing all information, the expert will advise to the farmer's via mobile phone [4].

Krishimantra is an agro advisory system for Gujarat state. The system gives the information regarding farms, farmers, weather information etc. And generate recommendations for pest and disease prevention based on inputs given by the user [5, 6].

In Agriculture crop production is significantly affected by change in climate. Crop management like planting, fertilizer, irrigation, etc. can be used to offset the loss in crop yield due to effects of climate. In Agriculture models like crop yield prediction, crop diseases, crop price, etc. can contain a large number of variables, such as many types of weather parameters, crop production variables (amount of fertilizer applied, planting date, total production, productivity etc.) and geographic variables. Crop yield predictions are useful in formulation of policies regarding crop stock, distribution and supply of agricultural products to different area in country. The relationships between various climate factors like meteorological information

and crop yield, has been the most common approach to predicting crop yields in past years. A result of crop yield forecasting is an important for optimizing crop yield. Various statistical approaches have been used for such agricultural recommendation system.

“Preharvest Crop Production Forecast Methodologies: IASRI Studies”, was study for the forecasting methodologies of crop yield. In this research they have define different models for forecasting yield of various crops at various locations, developed at IASRI (Indian Agricultural Statistics Research Institute, New Delhi) using various types of data or approaches. The developed methodology has been demonstrated at district, agro-climatic zone as well as state level. The methodology has been successfully used by various research workers [7].

“A Study of Crop Yield pattern with Climate Change based on Physical Parameters: Temperature and Rainfall in Western Uttar Pradesh to make future predictions for better Crop Management and Yield”, study was conducted to predict the wheat yield for ten district of Uttar-Pradesh. In this research weather indices based multiple linear regression (MLR) analysis is used to predict wheat yield of the district Ghaziabad. They have implemented MLR model for small area and found 90% of accuracy in result thus making it a very good model for crop prediction [8, 9].

“Applicability of ARIMA Models in Wholesale Vegetable Market: An Investigation”, research was study for investigated the ability of ARIMA model in wholesale vegetable market of Ahmedabad district. In this research models were built for sales data of one storage vegetable for Ahmadabad wholesales market in India. So, Ultimate Goal of this research was studying the application of ARIMA models on vegetable wholesale data and to forecast the future demand with accuracy [10].

“Alternative Forecasting Techniques for Vegetable Prices in Senegal”, study was conducted to investigate the performance for forecasting vegetable prices and to make recommendation to users. In this research work two forecasting (a parametric models and a non-parametric model) approaches used by the author. The parametric models consist of the naive model, exponential smoothing models and auto regressive integrated moving average model (ARIMA). The non parametric model uses the spectral analysis. The author collected monthly average price of tomato, potato and onion and apply the both the models and generate the forecast price for potato,

tomato and onion [11].

“Price Behaviour of Potato in Agra Market - A Statistical Analysis”, study was conducted to found market price pattern of potato crop in Agra’s local market. Also, predict the market price of potato using the time series auto regressive integrated moving average (ARIMA) model [12].

Chapter 3

Research Methodology

3.1 Theoretical Framework

There are many weather parameters that affect crop productivity during various stages of crop cultivation. So, it is necessary to know the distribution pattern of weather in the crop season. Cotton production is based on weather parameters like temperature, rainfall, soil moisture, etc. in irrigated areas. Crop market price is dependent on the total crop production. Hence, pattern of market price is important to farmers to get more returns.

3.2 Measurement Framework

Weather indices based multiple regression was used to analyze the effect of weather parameters on cotton yield in Ahmedabad district. Weather variables for cotton were selected through the literature survey on cotton research work done. Regression analysis was used to analyze the seasonal climate parameters on cotton yield. Time series ARIMA model was used to analyze the market trend of cotton production for Ahmedabad local market.

3.2.1 Concepts of Statistics

“A statistic is any summary number that represents the piece of information”. Generally, a statistic is used to estimate the value of a population parameter. Value of

statistics can be determined easily because the samples are manageable in size and then we use known statistics to learn about unknown parameters.

3.2.2 Linear Regression Analysis

Forecasting is a technique to estimate future trends using statistical methods. There are several methods of prediction, like quantitative and qualitative methods, naive approach, economic forecasting methods etc. Each method is fitting for certain type of situation only. Thus select the right method is most important step in forecasting.

Simple Linear Regression

Regression analysis is one of the most widely used methods in yield forecasting. This technique predicts the response variable, i.e. yield, in terms of explanatory variables such as weather, soil properties, etc. It is a statistical method that allows us to analyses and summarize relationships between the predictor, explanatory, or independent variable and the response, outcome, or dependent variable. Equation for simple linear regression is:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \quad (3.1)$$

Assumptions of Simple Linear Regression Model:

- Y_i dependent variable, is a linear function of the X_i and have equal variances for all X_i values.
- X_i is independent variable.
- The errors ϵ_i are independent.
- β_0 and β_1 is a regression coefficient.

Multiple Linear Regression

Multiple linear regression (MLR) is a method which allows us to define relationship between more than two variables, out of which one variable is dependent variable and all other variables are independent variables. Relationship between dependent

variable and each independent variable is linear. Weather indices based Multiple Linear Regression (MLR) model is used for prediction of yield. Given below is the MLR model which is used to predict yield using single weather parameter.

$$y = \alpha_0 + \sum_{i=1}^m \alpha_i z_i + \sum_{i=1}^m \alpha_{ij} z_{ij} + \epsilon \quad (3.2)$$

Where ϵ denotes random error, y denotes yield, $\alpha_0, \alpha_i, \alpha_{ij}$ are the regression coefficients, and z_i, z_{ij} are the independent variables which are function of basic weather variables like temperature and rainfall. The index 'm' represents the particular month of the season.

- Based on the individual effect of weather parameters the generated variables are given by

$$z_{ij} = \sum_{m=1}^n x_{im} R_{im}^j / \sum_{m=1}^n R_{im}^j \text{ where, } j = 0, 1 \quad (3.3)$$

Here, the index 'j' is degree of the equation. Where, x_{im} is the value of weather variable of particular month. R_{im} is the simple correlation coefficient between weather variable at m^{th} month and crop yield of particular year.

For $j = 0$,

$$z_{i0} = \sum_{m=1}^n x_{im} / n$$

and weighted generated variables

$$z_{i1} = \sum_{m=1}^n x_{im} R_{im} / \sum_{m=1}^n R_{im}^1$$

And the MLR model becomes,

$$y = \alpha_0 + \alpha_1 z_{i0} + \alpha_2 z_{i1} + \epsilon \quad (3.4)$$

- Based on the joint effect of weather parameters the generated variable is as

follows,

$$Q_{ii',j} = \sum_{m=1}^n R_{ii',m}^j x_{im} x_{i'm} / \sum_{m=1}^n R_{ii',m}^j \text{ where, } j = 0,1,2 \quad (3.5)$$

Where R_{im}^j is the correlation coefficient between crop yield y and product of weather variables x_{im} and $x_{i'm}$.

For $j=0$,

$$Q_{ii',j} = x_{i'm} x_{im} / n$$

And the weighted term will be,

$$Q_{ii',j} = \sum_{m=1}^n R_{ii',m}^1 x_{i'm} x_{im} / \sum_{m=1}^n R_{ii',m}^1$$

By, including these two interaction terms in the model,

$$y = \alpha_0 + \sum_{i=1}^2 \sum_{j=0}^2 \alpha_{ij} z_{ij} + \sum_{j=0}^1 \alpha_{ii',j} Q_{ii',j} + \epsilon \quad (3.6)$$

Where ϵ denotes random error, y denotes crop yield, $\alpha_0, \alpha_{ij}, \alpha_{ii',j}$ are the regression coefficients, z_{ij} is generated variable for single parameter, and $Q_{ii',j}$ is generated variable for joint effect of both parameters.

Some terminologies of linear regression analysis are as follows [13]:

- Prediction error: The difference between actual value and prediction value of the dataset is defined as prediction error of the model.
- Standard error: It is used for the accuracy measure of predictions. It is also known as sum of squares error.
- t-test: A t-test is one of the statistical hypothesis test. The test statistics analyze a statistical t-statistic, t-distribution and degrees of freedom to determine a P-value if the null hypothesis is supported. It is used for testing the mean of one population against a standard or comparing the means of two populations.

- P-test: The P-value is the probability of obtaining a sample statistic as extreme as test statistic. The test statistics is t statistics that was actually observed by assuming null hypothesis is true. If the P-value is less than the significance level than reject the null hypothesis.
- R-squared: R-squared is coefficient of determination. It is a number between 0 and 1. If $R^2 = 1$, it shows that all data points perfectly fitted on the regression line. If $R^2 = 0$, then the fitted regression line is perfectly horizontal due to some random errors.
- Correlation coefficient: Correlation coefficient is denoted by r . It is a number between -1 and 1, where both the boundaries are inclusive. Plus and minus sign represents positive and negative slop of the fitted regression line respectively. If $r = -1$, it shows perfect negative linear relationship. If $r = 1$ it shows perfect positive linear relationship. If $r = 0$, it shows that there is no linear relationship between the two variables.

3.2.3 Time series Analysis

Time series techniques are the models to predict future values based on previous values. The variety of time series techniques are Autoregressive (AR) model, Moving Average (MA) model, Autoregressive moving average (ARMA) model, Autoregressive Integrated Moving Average (ARIMA) model, trend analysis, double moving average model, double exponential smoothing etc.

The statistical properties like mean, variance, autocorrelation, etc. of the time series are constant over time period is known as stationary series. The models used to predict the stationary series are simple moving average model, Autoregressive model. On the other hand, the statistical properties like mean, variance, autocorrelation, etc. of the time series are not constant over time period is called nonstationary series. Some of the models that can handle this type of series include trend analysis, double moving average, and double exponential smoothing, autoregressive moving average model or autoregressive integrated moving average model [14]. ARIMA model is used to predict the market price of the crop.

The AR(p) model (autoregressive model) of y_t series, can be expressed as:

$$y_t = \beta_0 + \beta_1 y_{t-1} + \dots + \beta_p y_{t-p} + \epsilon_t \quad (3.7)$$

Where, y_t is the actual values or response variables at time t ,

$y_{t-1}, y_{t-2}, \dots, y_{t-p}$ is the respective variables at different time with lags;

$\beta_0, \beta_1, \dots, \beta_p$ are the model parameters;

and ϵ_t is random error over the time period.

Similarly, the MA(q) model (moving average model) of y_t series, can be expressed as:

$$y_t = \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \dots - \theta_q \epsilon_{t-q} \quad (3.8)$$

Where, y_t is the actual values or response variables at time t ,

$\epsilon_{t-1}, \epsilon_{t-2}, \dots, \epsilon_{t-q}$ is the respective error over the time period;

$\theta_0, \theta_1, \dots, \theta_q$ are the model parameters;

And ϵ_t is random error over the time period.

In general, an ARIMA model is characterized by the notation ARIMA (p, d, q), where p, d and q indicate orders of Autoregressive, Integration (differencing) and Moving Average, respectively.

An ARMA (p, q) model is defined by Equation:

$$y_t = \beta_0 + \beta_1 y_{t-1} + \dots + \beta_p y_{t-p} + \epsilon_t - \theta_1 \epsilon_t - \theta_2 \epsilon_{t-1} - \dots - \theta_q \epsilon_{t-q} \quad (3.9)$$

where, y_t and ϵ_t are the actual value and random error over the time period t , respectively β_i ($i=1, 2, \dots, p$) and θ_j ($j=1, 2, \dots, q$) are the model parameters. Identification of order of p and q the Auto-Correlation Function (ACF) and Partial Auto-Correlation Function (PACF) is applied. Then, the estimation and forecasting process are applied on data.

Some terminologies of linear regression analysis are as follows :

- Auto-Correlation Function (ACF): Auto-Correlation Function is also known as lagged correlation or serial correlation. It is the correlation between observations of a series as a function of the time lag between them.
- Partial Auto-Correlation Function (PACF): Partial Auto-Correlation Function

is a conditional correlation. While calculating PACF, the ACF with all the observations of a series within the lag are partialled out. All partial auto correlation is got from a different linear equations.

- Akaike information criterion (AIC): The Akaike information criterion (AIC) is a calculation of the relative quality of time series model for a given data sets. It is used for the compare the models for the same datasets and find the best fitting model. The model with smallest AIC value is preferred.
- Bayesian information criterion (BIC): The Bayesian information criterion is also known as Schwarz criterion (SBC). It is used for the selection of model from finite set of models, the model which have lowest SBC is preferred.

3.3 Proposed Approach

This system is developed to help the farmers of Gujarat region to improving their farm productivity. The proposed architecture of the system is shown in below figure 3.1. It is a web based system accessible via web browsers. The agro user can get information which can be location based, check information about crop diseases, information about crop nutrient, information about crop storage, information about crop market price and get recommendations regarding crop production. The farmers/users will be able to get current weather information of their locations through the system. Apart from that get past weather data to figure out patterns and weather conditions.

Basically, the whole proposed system is divided in four modules. These modules and their functioning along with standards and specifications are discussed below.

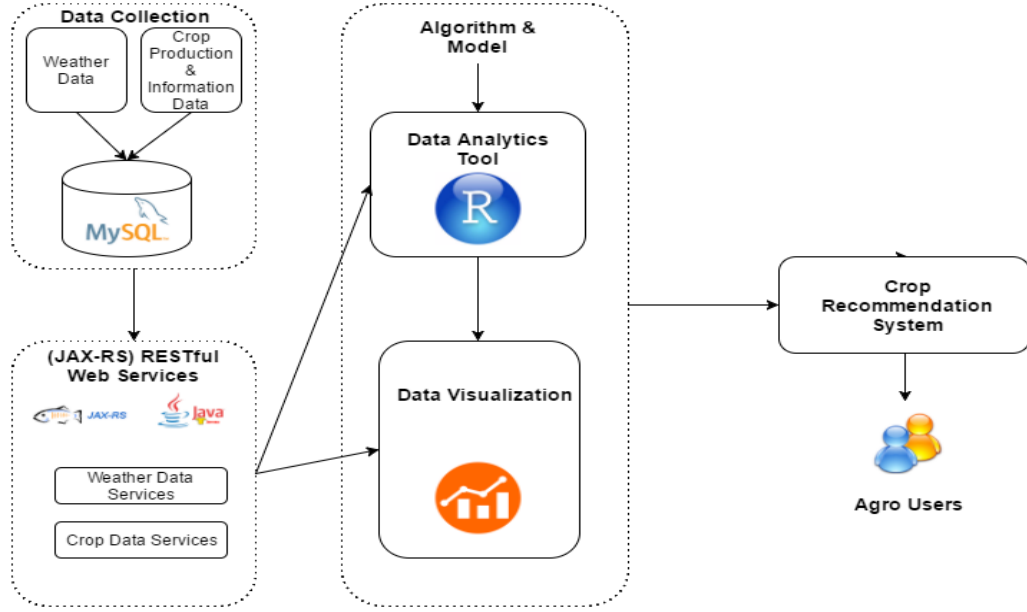


Figure 3.1: Proposed Architecture of Data Analytics for Crop Recommendation System

3.3.1 Data Collection

Data collection is the act or process of gathering data from data sources for further data processing or data storage to answer relevant questions and evaluate outcomes. The SQL database holds the general static kind of information that does not change frequently like crop production information, soil information, crop market price information etc. Apart from the static data, weather related data is also be stored in the SQL database. The data is fetched using SQL queries and it is be stored on a MySQL Workbench to make it accessible over the internet using RESTful services.

Massive quantities of disparate data like daily weather data, climate data cultivation area, production, soil conditions, fertilizer/ pesticide use, commodity market conditions, etc. are available at weather stations, meteorological departments, department of agriculture: state government and research organizations like SAC/ISRO. We have collected seventeen years of time series data ranging from 1995-2011 from various sources. The weather database consist of daily weather data parameters like minimum temperature, maximum temperature, rainfall, relative humidity and soil moisture. The crop database consist of yearly crop data parameters like crop area, crop production, crop productivity, crop diseases information, and monthly crop market price for Ahmedabad District, Gujarat. The data from these sources are

gathered and store in the MYSQL storage. In addition, to make the data sharable and reusable.

Data Sets	Parameters	Unit
Weather Database	Minimum Temperature, Maximum Temperature, Temperature	Celsius
	Rainfall	Millimeter (mm)
	Soil Moisture	mm/m
	Relative Humidity	Percentage (%)
Crop Database	Crop Yield	Tonnes/Hectare
	Crop Production	Hectare
	Crop Cultivation Area	Tonnes
	Crop Market Price	Rs/Quintal

3.3.2 RESTful Web Services

REST is a REpresentational State Transfer, web standards based architecture and uses HTTP Protocol. A web service is a collection of open protocols and standards used for exchanging data between applications or systems. Web services based on REST Architecture are known as RESTful web services. These web services use HTTP methods to implement the concept of REST architecture. A RESTful web service usually defines a URI (Uniform Resource Identifier) service, provides resource representation in JSON or XML and set of HTTP Methods.

The services are responsible for the communication that takes place between the components of the system. It connects the SQL database and the web browser end-user over the internet to fetch and deliver information and recommendations. These services are developed in Java v1.8 using JAX-RS (Java API for RESTful Web Services) implementation for REST and NetBeans IDE (Integrated Development Environment). The web services have been implemented using REST and output is in JSON or XML format. The output can be routed to a browser and can be used by performing standard JSON or XML parsing.

3.3.3 Data Analytics and Data Visualization

Data Analytics is a process of inspecting, cleaning, transforming, and modeling data with the goal of discovering valuable information, suggesting conclusions, and deci-

sion support systems. Main goal of data analytics is to analyze data (land data, crop, weather, geo-spatial data, etc.) coming from various sources using various analytic tools and techniques, to generate valuable information. The information generated can be used by the decision support system.

Data visualization is a technique that communicates information clearly and efficiently to end-users via the information graphics, such as tables, graphs and charts. Data visualization makes complex data more accessible, understandable and usable. The visualizations can be useful to easily identify crop patterns, crop future trends, market trends, etc.

The analytic results are stored in the data store and can be recycled to perform incremental analytics and optimize the analytics process. The valuable information will be integrated into the decision support system and knowledge management system to deliver useful recommendations to the agro user through web based interfaces. In this research work RStudio data analytical tool and d3 visualization library is used for developing the crop recommendation system. RStudio is an open-source IDE for R programming language which is used for statistical computing. D3 is a JavaScript library for producing interactive data visualizations.

3.3.4 User Interface

User interface of Crop Recommendation system is implemented using RShiny web framework which provides facility to turn your analyses result into interactive web applications. Hence, agro user can access this recommendation system via web browser. We have designed it to be very simple without manual inputs by keeping options for selection wherever possible.

3.4 Summary

In this chapter we discussed theoretical framework to accomplish the research work. It discusses the proposed approach and its solution. Also, it discusses statistical analysis of (i) monthly rainfall and temperature on overall cotton yield (ii) seasonal weather parameters on overall cotton yield (iii) crop market trend. For this we are using time series weather data like rainfall, minimum temperature, maximum temperature, average temperature, relative humidity, soil moisture, etc. and agriculture data like crop cultivation area, production, yield, price, etc. of 17 years (1995-2011) and then applied the data analytics techniques in order to get analysis result.

Chapter 4

Experimental Data

4.1 General

Agricultural and weather datasets are available in different formats like structured (excel, text, csv, etc.), unstructured (image files, audio, video, etc.) and semi-structured (pdf, emails, web logs, etc.), which need to be extracted for further processing. The dataset available at different sources like, government bodies, Meteorological department, research institutes, etc. are processed and extracted using RESTful web services. We have collected seventeen years of time series data ranging from 1995-2011 from various sources listed in Table 4.1.

#	Data Type	Year	Source
1.	Weather data (Temperature, Rainfall, Relative Humidity, Soil moisture)	1995-2011	www.Indiastat.com
2.	Agriculture data (Crop cultivation area, Crop production, Crop Productivity)	1995-2011	www.Indiastat.com, http://apy.dacnet.nic.in
3.	Crop Market Price	2013-2015	http://agmarknet.dac.gov.in

Table 4.1: Sources of Dataset

The weather data and crop data are collected on daily basis, yearly basis respectively. The crop data and weather data are extracted from SQL using RESTful web services and integrate into analytical engine to find valuable information. The daily weather data were converted into average monthly estimates so as to be fitted into the proposed yield prediction model. For rainfall and average temperature seven

months of data were considered in the yield prediction model as the cropping season of cotton starts from early July and ends towards the end of January in Ahmedabad District, Gujarat. Table A.1 data shows monthly total rainfall and temperature from year 1995-2010. We used monthly rainfall and average temperature to identify relationship with yields in those years. Table A.2 data shows cotton cultivated area in hectare, production in bales and productivity in bales/hectare from year 1995-2011. Cotton wholesale monthly market price in rs./quintal is present in Table A.3 from year 2010-2015.

4.2 Data Organization and Use

The weather parameters used for seasonal analysis for seventeen years is shown in below table. For analyzing the effect of change in different weather parameters on the cotton crop productivity we experiment the seasonal analysis. For the seasonal analysis used the weather parameters are total rainfall, average temperature, minimum temperature, maximum temperature, relative humidity, soil moisture, total rainy days, air temperature in surface over land area in particular season. The weather parameters are initially collected on daily basis and then it is aggregated into monthly and yearly data. Table A.4 shows cotton crop data with seasonal climate parameters.

4.3 Summary

In this chapter data type and data sources for weather and agriculture data sets are discussed. Also, organization and use of data according to different analysis scenarios is discussed.

Chapter 5

Results

5.1 Experiment Environment

Experimental setup for implementation of Crop recommendation system is:

5.1.1 Hardware usage

Processor : Intel Core i5 Processor
RAM : 4 GB
Hard Disk : 500 GB

5.1.2 Software usage

Operating System : Windows 10
MySQL Workbench 6.3
NetBeans IDE 8.1
JDK 1.8
RStudio 0.99.484
D3 JavaScript library

5.2 Results

As we have created RESTful web services for crop and weather data to explored data in JSON or XML format.

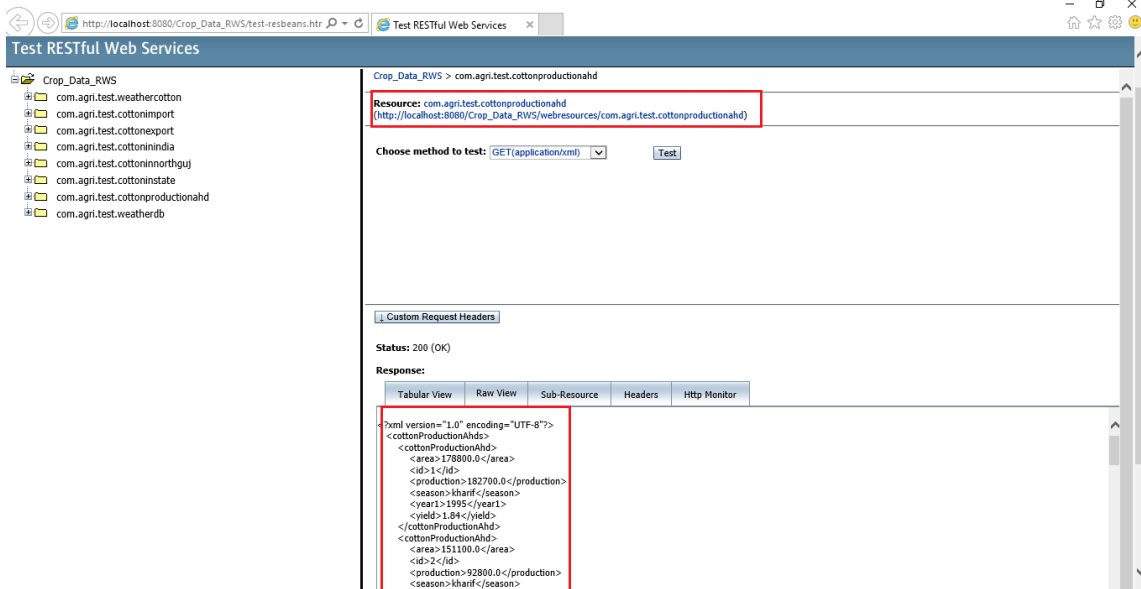


Figure 5.1: RESTful Web service for Cotton Crop

The weather and crop data are extracted in XML format using RESTful web-services and parsed into SQL format. The daily weather data were aggregated to average monthly estimates, so as to be fitted into the proposed prediction model. The weather data set and crop data set is integrated and then analytics are performed on them.

The agro users will be able to access a variety of data like basic information of crop, crop diseases & management information, crop yield prediction analysis, market price analysis and other relevant data on the browser. To access all this information firstly user have to select the location in map as shown in figure 5.2. After selection of location system will display the current weather condition. Now for accessing the crop information users have to select the specific crop from the sidebar.

Here, We have select the cotton crop so, system will provides information like basic information, crop diseases and management information, cultivation area vs

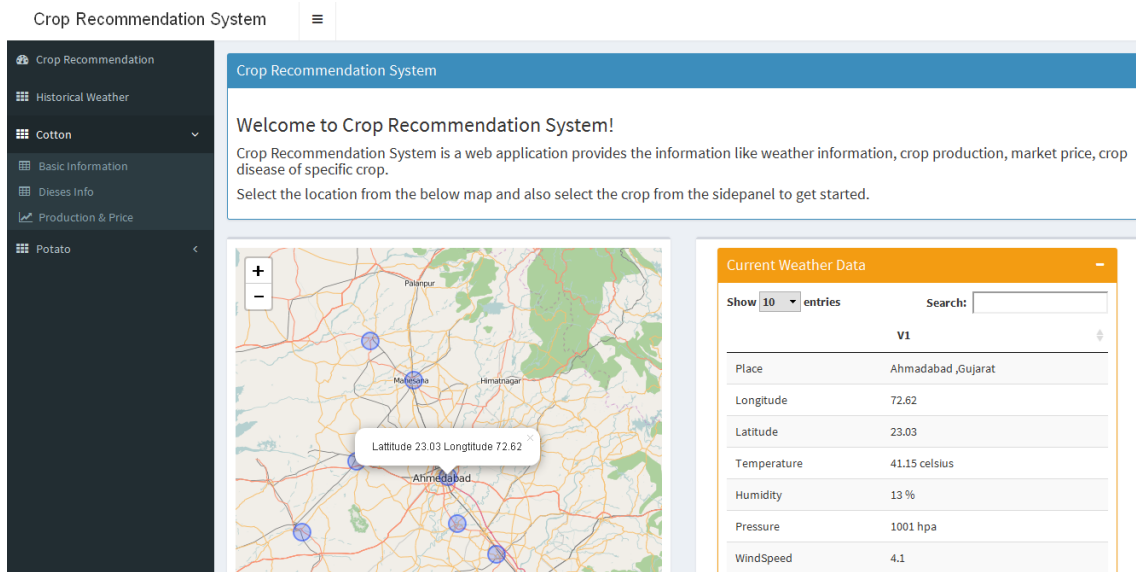


Figure 5.2: Homepage of Crop Recommendation System

production information, crop yield prediction analysis, seasonal analysis, market price forecast result. Cotton crop diseases information is displayed in below figure 5.3.

Diseases Information		
Diseases Name	SYMPTOMS	MANAGEMENT
Alternaria leaf spot	Small, circular brown lesions on cotyledons and seedling leaves which expand and develop a concentric pattern; necrotic areas coalesce and often have a purple margin; centers of lesions may dry out and drop from the plant creating a "shot-hole" appearance on the leaves	Plow crop residue into the soil to reduce inoculum levels; provide plants with adequate irrigation and nutrients, particularly potassium; applications of appropriate foliar fungicides may be required on susceptible cultivars
Aphids	Small soft bodied insects on underside of leaves and/or stems of plant; usually green or yellow in color, but may be pink, brown, red or black depending on species and host plant; if aphid infestation is heavy it may cause leaves to yellow and/or distorted; necrotic spots on leaves and/or stunted shoots; aphids secrete a sticky, sugary substance called honeydew which encourages the growth of sooty mold on the plants	If aphid population is limited to just a few leaves or shoots then the infestation can be pruned out to provide control; check transplants for aphids before planting; use tolerant varieties if available; reflective mulches such as silver colored plastic can deter aphids from feeding on plants; sturdy plants can be sprayed with a strong jet of water to knock aphids from leaves; insecticides are generally only required to treat aphids if the infestation is very high - plants generally tolerate low and medium level infestation; insecticidal soaps or oils such as neem or canola oil are usually the best method of control; always check the labels of the products for specific usage guidelines prior to use

Figure 5.3: Cotton crop diseases information

5.2.1 Past Agro-Production Trend

An analysis of the data like cotton cultivation area in hectare and cotton production in bales from 1995 to 2011, Ahmedabad District of Gujarat is presented in below graph. There has been steady increase both in terms of area and production except for the year 1999-2000 as Gujarat was affected by drought. There has been increase in production and cultivation area taken place because of good rainfall, improved groundwater recharge, increased the storage in surface reservoirs throughout the state, and improved soil moisture conditions. For the year 2006 declined in production was noticed as changes in climate conditions.

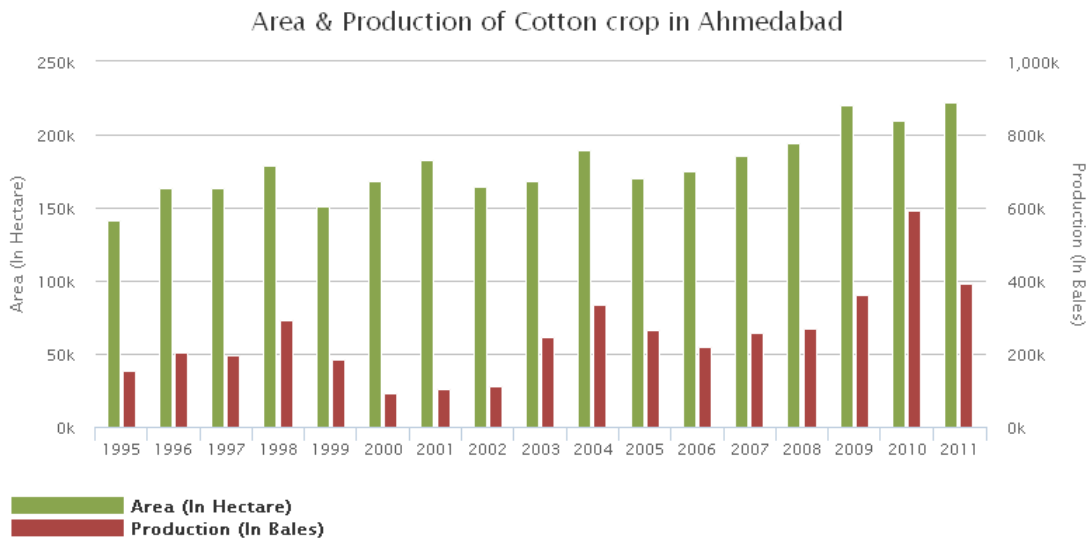


Figure 5.4: Area & Production of Cotton crop, Ahmedabad from year 1995-2011

5.2.2 Weather indices based MLR model

Weather indices based MLR model is used for the cotton yield prediction. The data consist of daily weather data parameters like minimum temperature (in degree celsius), maximum temperature (in degree celsius) and rainfall (in mm). Seven months data were considered in the yield prediction model as the cropping period of cotton starts from early July and ends towards the end of January in Ahmedabad District, Gujarat. The results of various cotton crop yield prediction models based on the weather parameters are presented below.

1. Temperature based MLR model

From the equation (3.2) MLR model based on temperature for cotton crop yield prediction is as follow,

$$yield = \alpha_0 + \alpha_1 z_{t0} + \alpha_2 z_{t1} + \alpha_3 z_{t2} + \epsilon \quad (5.1)$$

#	Variable	Coeff	Std. Error	t-Statistics	p-Value	R- Square
1	Intercept	7.87	4.34	1.81	0.09	0.55
2	average temperature	0.03	0.35	0.08	0.94	
		-0.56	0.19	-2.96	0.01	
		0.35	0.47	0.76	0.46	
		0.35	0.47	0.76	0.46	

Table 5.1: Temperature based MLR Analysis

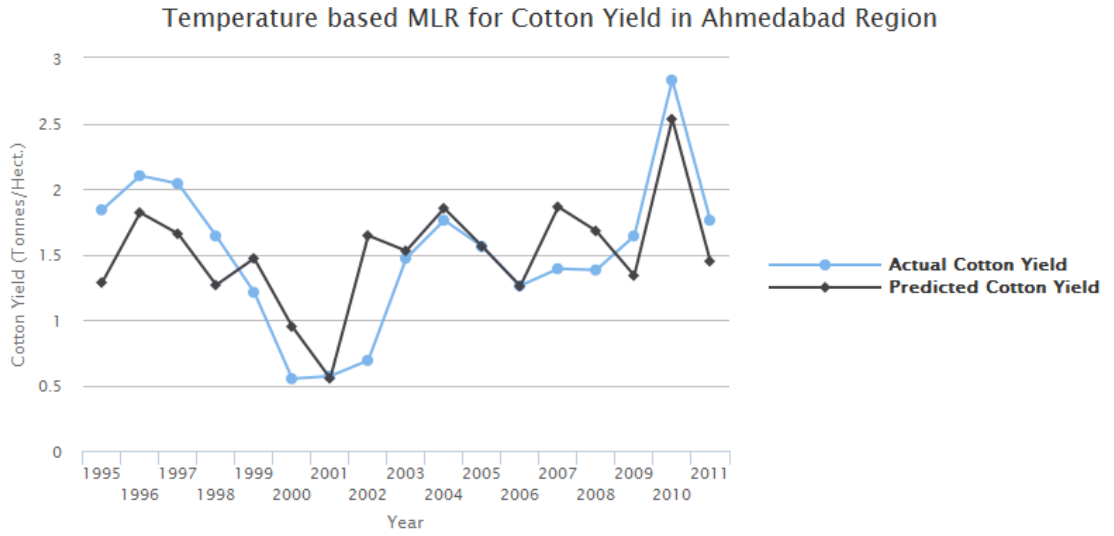


Figure 5.5: Cotton crop yield prediction based on temperature

Year	Actual Yield (tonnes/hectare)	Predicted Yield (tonnes/hectare)	Difference
1995	1.84	1.28	0.56
1996	2.1	1.82	0.28
1997	2.04	1.66	0.38
1998	1.64	1.27	0.37
1999	1.21	1.47	0.26
2000	0.55	0.95	0.40
2001	0.57	0.55	0.02
2002	0.69	1.64	0.95
2003	1.47	1.53	0.06
2004	1.76	1.85	0.09
2005	1.56	1.56	0.00
2006	1.26	1.26	0.00
2007	1.39	1.86	0.47
2008	1.38	1.68	0.30
2009	1.64	1.34	0.30
2010	2.83	2.53	0.30
2011	1.76	1.45	0.31

Table 5.2: Prediction of temperature based MLR analysis for cotton yield

2. Temperature and Rainfall based MLR model

By simplifying the equation (3.6) MLR model based on joint effect of temperature and rainfall for cotton crop yield prediction is as follow,

$$yield = \alpha_0 + \alpha_1 z_{0r} + \alpha_2 z_{1r} + \alpha_3 z_{0t} + \alpha_4 z_{1t} + \alpha_5 Q_0 + \alpha_6 Q_1 + \alpha_7 Q_2 + \epsilon \quad (5.2)$$

#	Variable	Coeff	Std Error	t-Statistics	p-Value	R- Square
1	Intercept	6.12	7.81	0.78	0.46	0.65
2	rainfall	-0.23	0.24	-0.97	0.35	
		0.10	0.17	0.63	0.54	
3	average temperature	0.22	0.29	0.75	0.47	
		-0.36	0.15	-2.36	0.04	
4	joint effect of both	0.01	0.01	1.01	0.34	
		-0.01	0.01	-0.79	0.45	
		0.00	0.00	0.95	0.37	

Table 5.3: Temperature and Rainfall based MLR Analysis

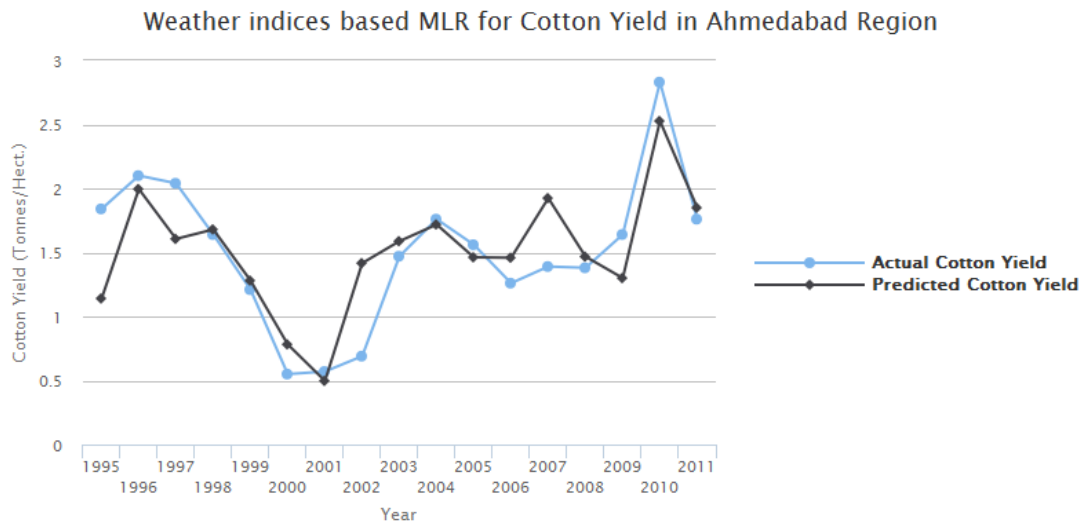


Figure 5.6: Cotton crop yield prediction based on temperature and rainfall

Year	Actual Yield (tonnes/hectare)	Predicted Yield (tonnes/hectare)	Difference
1995	1.84	1.14	0.70
1996	2.1	2.00	0.10
1997	2.04	1.61	0.43
1998	1.64	1.68	0.04
1999	1.21	1.28	0.07
2000	0.55	0.78	0.23
2001	0.57	0.50	0.07
2002	0.69	1.42	0.73
2003	1.47	1.59	0.12
2004	1.76	1.72	0.04
2005	1.56	1.46	0.10
2006	1.26	1.46	0.20
2007	1.39	1.92	0.53
2008	1.38	1.47	0.09
2009	1.64	1.30	0.34
2010	2.83	2.53	0.30
2011	1.76	1.85	0.09

Table 5.4: Prediction of temperature and rainfall based MLR analysis for cotton yield

5.2.3 Seasonal Analysis

The Experiment of seasonal analysis is done for analyzing the cotton crop productivity in terms of weather parameters and soil moisture for cropping period of cotton, which is starts from late June to early July and ends towards the end of January to early February in Ahmedabad District, Gujarat. Results of experiment is presented in Table 5.3.

$$\begin{aligned}
 Yield = & 8.62 + (-0.44) * Temp + (-0.24) * MaxTemp + (0.70) * MinTemp \\
 & + (0.03) * Rhumidity + (0.00) * Rainfall + (-0.01) * Rainyday \quad (5.3) \\
 & + (-0.01) * Soilmoisture + (-0.09) * Airtempland
 \end{aligned}$$

#	Variable	Coeff	Std Error	t-Statistics	p-Value	R- Square
1	Intercept	8.62	11.55	0.75	0.48	0.875
2	Temp	-0.44	0.58	-0.77	0.46	
3	Max Temp	-0.24	0.50	-0.49	0.64	
4	Min Temp	0.70	0.35	2.00	0.08	
5	Relative Humidity	0.03	0.03	0.78	0.46	
6	Rainfall	0.00	0.00	1.42	0.19	
7	Rainy Day	-0.01	0.01	-1.38	0.20	
8	Soil moisture	-0.01	0.00	-1.14	0.29	
9	Air Temp near surface over land area	-0.09	0.04	-2.06	0.07	

Table 5.5: Result of Seasonal Analysis

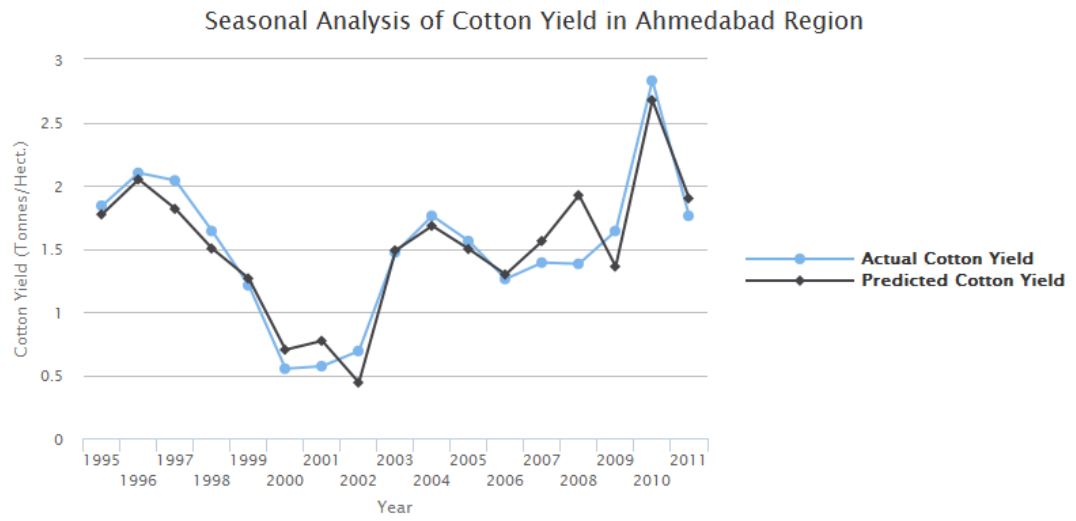


Figure 5.7: Cotton crop yield prediction based on seasonal Analysis

Year	Actual Yield (tonnes/hectare)	Predicted Yield (tonnes/hectare)	Difference
1995	1.84	1.77	0.07
1996	2.1	2.05	0.05
1997	2.04	1.82	0.22
1998	1.64	1.50	0.14
1999	1.21	1.26	0.05
2000	0.55	0.70	0.15
2001	0.57	0.77	0.20
2002	0.69	0.44	0.25
2003	1.47	1.49	0.02
2004	1.76	1.68	0.08
2005	1.56	1.50	0.06
2006	1.26	1.30	0.04
2007	1.39	1.56	0.17
2008	1.38	1.92	0.54
2009	1.64	1.36	0.28
2010	2.83	2.68	0.15
2011	1.76	1.90	0.14

Table 5.6: Prediction of seasonal analysis for cotton yield

5.2.4 Time Series Analysis

The time series data of wholesale monthly price of cotton in Ahmedabad market from January 2013 to December 2015 has been collected from agmarket of Indian agriculture ¹, and analyzed with ARIMA model. Based on minimum Akaike's Information Criterion (AIC) and Bayesian Information Criterion (SBC) values of the wholesale monthly price series, ARIMA(1,0,0) with non-zero mean model is selected. The results of monthly price forecasting is presented below.

Months	Actual Price (Rs./q)	Forecast Price (Rs./q)	Difference
Jan-15	4052.45	4328.118	275.7
Feb-15	3914.61	4275.392	360.8
Mar-15	3794.26	4227.024	432.8
Apr-15	4254.42	4184.794	69.6
May-15	4468.92	4346.263	122.7
Jun-15	4591	4421.531	169.5
Jul-15	4597.52	4464.368	133.2
Aug-15	4590.26	4466.656	123.6
Sep-15	4142.47	4464.109	321.6
Oct-15	4291.96	4306.98	15.0
Nov-15	4212.59	4359.436	146.8
Dec-15	4481.35	4331.585	149.8
Jan-16		4425.892	0.0
Feb-16		4406.432	0.0
Mar-16		4399.604	0.0
Apr-16		4397.208	0.0
May-16		4396.367	0.0

Table 5.7: Forecast result based on the fitted ARIMA model

¹<http://agmarknet.dac.gov.in/>

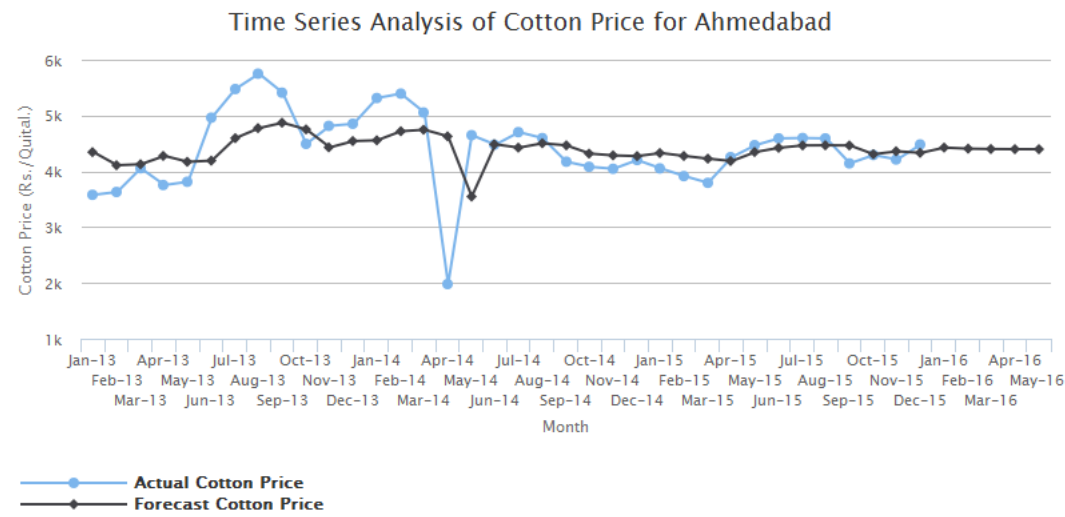


Figure 5.8: Cotton crop wholesale monthly market price forecast

Chapter 6

Conclusion and Future work

Result of various models of weather indices based MLR for cotton crop yield prediction is presented in Table 6.1. In this table model 1 is based on only one weather parameter temperature and model 2 is based on the joint effect of rainfall and temperature weather indices. From the analytical results it is clear evidence that both rainfall and temperature (model 2) have impact on cotton yield. Rainfall requirement in early stages of cotton growth is less. Cotton yield losses if temperature and rainfall is more than the optimal season value.

MLR model	R-Squared	Actual Cotton Yield in 2011	Predicted Cotton Yield in 2011	Difference
Model 1	0.55	1.76	1.45	0.31
Model 2	0.65	1.76	1.85	0.09

Table 6.1: Weather indices based MLR experimental result

Result of Seasonal Analysis shows in Table 5.5. Analysis is done for six climate variable for Ahmedabad District, Gujarat from year 1995-2011. Coefficient for minimum temperature shows that it affect more on cotton yield during the cotton cropping season. Negative coefficient indicate that there is negative relation between these two variables. Here, temperature and relative humidity contributes more than rainfall because Ahmedabad District is most irrigated region. Hence, if rainfall is low than optimal seasonal rainfall than water supply to cultivation area through Narmada Canal.

6.1 Conclusion

While doing the research works the crop recommendation system has been proposed using data analytics architecture. The agriculture and weather data were collected for various crops of the Ahmedabad District, Gujarat and results were generated for crop yield prediction, crop production trends, and crop market price prediction.

As a result of weather indices based MLR model we can conclude that rainfall and temperature are most important parameter affecting the cotton production. With 800-1000 mm annual rainfall and temperature varying between 21 °C and 30 °C in cropping season of cotton is necessary for good quality and production of the cotton. Results of seasonal analysis describe contribution of different climate parameters on cotton yield. As per seasonal analysis relative humidity, soil moisture, number of rainy days, air Temperature near surface over land area are the important factor affecting the cotton yield in cropping season.

Result of crop yield prediction model shows that, if we get weather forecast before irrigation of crop than we can limit amount of irrigation so than we can reduce loss of crop production or will help to farmers to make decision on whether to grow cotton crop or go for the alternative crop. Crop market price forecast model will help to farmers to sell the crop, if crops are storable farmer may keep than sale at higher prices after harvest time.

6.2 Future Work

Improve the results of analysis model (yield prediction model, price forecasting model), using large number of crop datasets and more weather parameters. Building a strong yield prediction and price forecasting model for all the crops on this analysis. Generating the crop recommendation using natural or local language to make it user friendly.

6.3 Limitations

The use of data analytics in India, especially in agriculture domain is quite difficult because of there are few limitations in these kinds of research works. The main issue is in while collecting the different kind of data, like climate data, agricultural data, like amount of fertilizer and pesticides used, different varieties of crop grown, different soil parameters, etc.

There are many factors affects the crop production like soil quality, soil fertility rate, amount of NPK in soil, water holding capacity of land, quality of seed, time difference of sowing crop, etc. in the same area. Like all these factors affecting the crop production. It is extreme difficult to model all these factors mathematically and due to lack of data.

REFERENCES

- [1] P Krishna Reddy, GV Ramaraju, and G Syamasundar Reddy. esagu: a data warehouse enabled personalized agricultural advisory system. In *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*, pages 910–914. ACM, 2007.
- [2] Agrisnet: A mission mode project to promote agricultural informatics and communications. 2014.
- [3] Kissan kerela: An integrated multi-modal agricultural information system for kerela. 2014.
- [4] Sanjay Kimbahune Pankaj Doke Ajay Mittal Dineshkumar Singh Arun Pande, Bhushan G. Jagyasi and Ramesh Jain. Mobile phone based agro-advisory system for agricultural challenges in rural india. *IEEE Conference on Technology for Humanitarian Challenges*, 2009.
- [5] Sanjay Chaudhary and Bhise Minal. Restful services for agricultural recommendation system. *Proceedings of NSDI-2013, IITB, Mumbai*, pages 46–52, 2013.
- [6] Sanjay Chaudhary, Minal Bhise, Asim Banerjee, Aakash Goyal, and Chetan Moradiya. Agro advisory system for cotton crop. In *Communication Systems and Networks (COMSNETS), 2015 7th International Conference on*, pages 1–6. IEEE, 2015.
- [7] Ranjana Agrawal, RC Jain, and MP Jha. Models for studying rice crop-weather relationship. *Mausam*, 37(1):67–70, 1986.

- [8] Aditya Saxena, Surendra Pratap Singh, Manju Rani, Manoj Kumar, Vishal Parmar, Siddhant Shekhar, and Abhishek Gogna. A study of crop yield pattern with climate change based on physical parameters: Temperature and rainfall in western uttar pradesh to make future predictions for better crop management and yield.
- [9] Amender Kumar and V Ramasubramanian. Crop forecasting based on meteorological data using sas. *Reference Manual, IASRI, New Delhi* ([http://www.iasri.res.in/sscnars/socialsci/7-Weather% 20based% 20Forecasting% 20Models. pdf](http://www.iasri.res.in/sscnars/socialsci/7-Weather%20based%20Forecasting%20Models.pdf)), 2012.
- [10] Manish Shukla and Sanjay Jharkharia. Applicability of arima models in whole-sale vegetable market: an investigation. *International Journal of Information Systems and Supply Chain Management (IJISSCM)*, 6(3):105–119, 2013.
- [11] Alionue Dieng. Alternative forecasting techniques for vegetable prices in senegal. *Revue senegalais de recherches agricoles et agroalimentalress*, 1(3):5–10, 2008.
- [12] DS Dhakre and D Bhattacharya. Price behaviour of potato in agra market-a statistical analysis.
- [13] Norman R Draper and Harry Smith. *Applied regression analysis*. John Wiley & Sons, 2014.
- [14] William Wu-Shyong Wei. *Time series analysis*. Addison-Wesley publ Reading, 1994.

Appendix A

Experiment Datasets

A.1 Monthly Rainfall and Temperature Data for Ahmedabad District

Year	Average Temperature (celsius)							Rainfall (mm)						
	jul	aug	sep	oct	nov	dec	jan	jul	aug	sep	oct	nov	dec	jan
1995	30	29	30	30	24	22	20	210	71	67	0	3	6	0
1996	29	28	28	27	23	20	19	313	530	82	43	0	0	9
1997	29	28	28	27	25	20	20	166	268	86	1	3	0	0
1998	29	29	29	28	23	19	19	409	186	195	30	0	0	0
1999	29	28	29	27	24	20	20	303	38	13	48	0	0	40
2000	29	29	30	29	25	21	19	361	115	5	0	0	1	0
2001	28	28	30	29	25	21	20	208	226	0	28	0	0	0
2002	30	29	29	29	25	22	21	81	106	70	0	0	1	3
2003	28	28	28	28	25	20	20	296	534	27	6	0	0	0
2004	30	27	30	27	25	21	18	212	432	27	16	14	0	0
2005	29	27	28	27	23	18	20	416	216	400	0	0	31	0
2006	29	27	30	29	25	22	21	291	461	162	0	0	0	0
2007	29	28	29	27	24	21	19	495	359	59	13	11	0	0
2008	28	28	29	28	24	23	21	200	283	96	0	0	6	0
2009	30	30	30	28	24	22	20	305	159	23	6	0	0	1
2010	29	28	28	29	25	25	21	428	525	114	9	29	29	0
2011	30	29	29	30	28	23	20	290	310	73	0	0	0	0

Table A.1: Monthly Rainfall and Temperature, Ahmedabad, 1995-2011

A.2 Cotton Crop Production Data for Ahmedabad District

Year	Area (In Hectare)	Production (In Bales)	Yield (Bales/Hect)
1995	141100	152300	1.84
1996	163900	202500	2.1
1997	163900	196200	2.04
1998	178800	293100	1.64
1999	151100	182700	1.21
2000	168000	92800	0.55
2001	182600	104500	0.57
2002	164300	113400	0.69
2003	168000	246300	1.47
2004	189500	332700	1.76
2005	170300	266100	1.56
2006	174800	221100	1.26
2007	185200	256900	1.39
2008	193900	268200	1.38
2009	220000	361200	1.64
2010	209200	592000	2.83
2011	221900	391600	1.76

Table A.2: Cotton Area, Production and Yield, Ahmedabad, 1995-2011

A.3 Cotton Crop wholesale Monthly Price Data for Ahmedabad District

Year	Cotton Crop Wholesale Monthly Price (Rs./q)											
	jan	feb	mar	apr	may	jun	jul	aug	sep	oct	nov	dec
2010	2880	2863	2877	2870	2753	3210	3210	3210	2984	4452	4453	4074
2011	3967	5419	5629	5076	3308	2791	2250	3357	3975	4675	4428	4551
2012	4049	3830	3688	3637	3352	3711	4584	4888	3967	3967	4346	4024
2013	3579	3630	4055	3755	3813	4962	5475	5751	5409	4492	4815	4853
2014	5314	5393	5054	1980	4648	4478	4702	4596	4173	4082	4046	4203
2015	4053	3915	3794	4254	4469	4591	4598	4590	4143	4292	4213	4481

Table A.3: Cotton wholesale monthly price, Ahmedabad, 2010-2015

A.4 Cotton Crop Seasonal Climate Parameters

Data for Ahmedabad District

Year	Temp (°C)	Max Temp (°C)	Min Temp (°C)	Humidity (%)	Rainfall (mm)	Rainy Days	Soil Moisture (mm/m)	Temp near surface over land (°C)
1995	26.37	32.78	20.67	61.80	356.88	37	116.58	26.32
1996	24.86	31.47	19.49	64.99	977.34	61	125.68	26.66
1997	25.39	31.24	20.39	68.97	524.50	80	186.69	25.94
1998	25.11	31.89	19.39	68.49	820.44	65	156.69	27.04
1999	25.37	32.38	19.41	60.60	441.94	50	115.25	26.83
2000	26.17	33.23	19.45	52.80	482.10	38	114.43	26.82
2001	25.77	32.76	19.46	54.34	462.03	66	100.79	26.87
2002	26.45	33.73	19.76	49.51	260.61	55	77.04	27.00
2003	25.27	32.05	19.68	56.78	862.87	63	140.74	26.43
2004	25.20	32.03	19.91	63.43	701.79	61	150.31	26.90
2005	24.59	31.68	18.82	64.04	1062.73	54	175.23	26.49
2006	26.16	32.21	20.29	62.79	914.13	68	183.63	27.24
2007	25.20	31.80	19.72	64.95	936.75	67	177.42	27.00
2008	25.99	32.40	20.76	65.06	585.73	62	133.79	26.46
2009	26.34	33.11	20.44	57.19	494.30	42	116.49	27.39
2010	26.31	32.54	21.26	65.20	1134.09	87	163.49	18.71
2011	26.09	32.72	20.37	60.22	673.37	69	163.49	18.71

Table A.4: Cotton crop seasonal climate parameters, Ahmedabad, 1995-2011