A Study Of Onion Price Fluctuations in India

A thesis submitted in partial fulfillment
of the requirements for the degree of

BACHELOR OF TECHNOLOGY &
MASTER OF TECHNOLOGY

in

Computer Science & Engineering

by

Praneet Khandelwal (2013CS50292)

Under the guidance of

Prof. Aaditeshwar Seth
&
Prof. Parag Singla

Department of Computer Science and Engineering,
Indian Institute of Technology Delhi.
July 2018

Certificate

This is to certify that the thesis titled **A Study Of Onion Price Fluctuation in India** being submitted by **Praneet Khandelwal** for the award of **Bachelor of Technology & Masters of Technology** in **Computer Science & Engineering** is a record of bona fide work carried out by him under my guidance and supervision at the **Department of Computer Science & Engineering**. The work presented in this thesis has not been submitted elsewhere either in part or full, for the award of any other degree or diploma.

Prof. Aaditeshwar Seth & Prof. Parag Singla
Department of Computer Science and Engineering
Indian Institute of Technology, Delhi

Abstract

Fluctuations in the prices of food articles have always been a concern for both the consumers and the producers. In this project, we look at one such indispensable ingredient of Indian cooking, that is onion and aim to study the variations in its price over the past years. Onions have been a staple part of diet for many Indian families and therefore play an important role in our everyday life. Large fluctuations in its prices have also made it an area of focus in the political scenario of this nation. Our study mainly involves studying and understanding the entire onion trading system that is in place in India and try to explain some of the key events of the past based on the time series data available to us. Furthermore, we also analyse and evaluate a system to identify the anomalies in onion prices and classify the reasons behind those anomalies. In the end, we also present how advanced time series models can be used for the prediction of onion prices. These models altogether help us in understanding the entire onion trading system in a better way and help regulate the agricultural markets in a fair manner.

Acknowledgments

I would like to thank my supervisors, *Prof. Aaditeshwar Seth* and *Prof. Parag Singla* for providing me with the opportunity to work on this interesting project as my MTech. Project. Their unfailing support, guidance and help have been invaluable during the course of this project. I am grateful for all the help I received from them.

I would also like to thank Shivank Goel who has helped a great deal by providing me assistance whenever needed.

Praneet Khandelwal

Contents

Chapter 1

Introduction

Onion is one of the most indispensable raw commodity used in the traditional Indian cooking. It is used by all sections of the society, whether it be rich or poor. This makes studying about the price variations and fluctuations in the onion prices a very important area of research. A lot of study has been done to crack down the dynamics of the supply chain that this commodity goes through in the context of Indian markets. With the help of this project, we also intend to bring more insights to get a better understanding of the entire system.

There have been lot of ups and downs in the pricing of onions in the past decade. This makes studying this commodity a lot more important. There is slight increase and decrease in the price series of almost every commodity. It is, when these fluctuations grow significantly high, that it becomes a reason for concern. And onion is no exception to this. There have been quite a few price disturbances or anomalies that it has messed with the lives of common people. There have been instances in the past where it has also been able to dictate outcome of election in Indian politics. And this has been the case even after the fact that India is the second largest onion producing country in the world.

Generally these fluctuations are caused by natural factors like erratic rainfall, crop failure etc. But at times they are caused by man-made reasons such as strikes, poor management, government policies, hoarding. So in this study, we focus our attention on factors such as weather abnormalities and man-made factors like hoarding which turn out to be the top reasons for any anomaly.

Objective & Scope

The aim of this project is to go into the depths of the working of the onion supply chain in India. It also includes building an understanding of the onion trading in the local Indian markets. Apart from this, we also intend to learn how hoarding takes place within the system and the factors that lead to it. This helps in pointing out ways to reduce such instances of artificial price rise in the future which can be useful for the concerned authority. We can list down the broad objectives as follows:

- *Data Collection and Domain Understanding*- Collect and organise the required data in the form of time series. We would like to understand the working of the entire system to gain domain knowledge in order to make better interpretations.

- *Qualitative Analysis*- Study about the past occurrences of anomalous events and try to explain them using the data available. We also aim to explore the dependence of different factors on each other and how they influence the price.

- *Quantitative Analysis*- After going through the dynamics of how onion price anomalies occur, we would like to build a two-step system based on techniques from machine learning to learn what happens during anomalous events. The first step would be to see how well we can detect the anomalies and the second would be to try and classify these instances based on the reason for their cause.

- *Time Series Prediction using Advanced methods*- In the end, we would like to look into the traditional approach to time series analysis by using the standard seasonal ARIMA models and present an evaluation of how well these models perform for the particular case of onion price prediction.

Terminology

- **Mandi:** Mandi is basically a market or hub where farmers are provided a a system to bring their produce and auction it to sell to the traders.
- **Mandi Price:** In our context, this means the price at which onion is sold to the traders at the mandis. The seller in this case is the farmers who bring their produce to the mandi and sell it in wholesale to the traders. We will be using Mandi price and Wholesale price interchangeably throughout the report.
- **Centers:** These, in our context, refers to the main markets in a state where the produce is brought from various mandis associated to it and sold to the customers or the end-buyers.
- **Retail Price:** As the name suggests, retail price is an indicative measure of the price at which the customers purchase the raw commodity or in our case, onions, at the center or city markets.
- **Anomaly Event/Period:** Any continuous window or period of approximately weeks or 43 days which show inflation in onion prices will be referred to as anomaly event or period.

Chapter 2

Data Used & Pre-Processing

2.1 Dataset for Onion Prices & Arrivals

To get started with the analysis of the onion market scenario, we first obtained various data related to onions. We decided to obtain the data for *mandi prices, retail prices* and *mandi arrivals* for onions for various locations spread across India.

Mandi Price & Arrivals: The Government of India runs an *Agmarknet (Agricultural Marketing Information Network)* website which provides various information about different parameters like maximum, minimum and modal price of onion at a particular mandi. This data is publicly available on their website. We wrote scripts for web crawler to obtain this data for mandis spread across the entire nation. We obtained daily mandi price data for about 1514 mandis from different parts of India for the period starting from Jan 1, 2006 to Nov 30, 2017. The website also provided us daily data about volume of mandi arrivals being brought at each mandi on a daily basis.

Retail Price: We obtained the daily retail price of onions for about 30 centres from a portal run by the *National Horticulture Board.* Similar crawlers were written to download the retail price data for the period mentioned in the previous paragraph. After this the mandis were mapped to corresponding centers and this comes out to be a many to one matching. This means that there are a number of mandis associated with a single center.

Rainfall Data: We also obtained information about the rainfall for the above period for the region of western Maharashtra. Rainfall data is obtained as monthly averaged rainfall observed in the particular region contrary to the daily data obtained for the prices and arrivals.

Newspaper Labelled Anomalies: Additionally we manually read all the newspaper reports/articles related to 'onion prices' available on the Times of India archive. Each newspaper article was read in order to identify and label the reason, if any, behind the price fluctuations mentioned in the article.

These labellings formed the ground truth labels which would be later used for classification purposes. After identifying such newspaper articles in the entire period, we clubbed those newspaper articles together which were close to each other temporally and also depicted the same anomaly. By this clubbing technique, we were able to obtain a number of anomaly periods(explained in the previous chapter) each of which spanned over a duration of 43 days. So each anomaly period or event was a period of 43 days.

2.2 Selection Criteria

For proper analysis to be conducted, it is very important that we have sufficient amount of data. And this becomes all the more important for time series data because the values at any point may be dependent on previous values or there may be seasonality in the data. The reason for missing data points to the poor management, lack of responsibility, changing policies and weakly enforced rules.

One of the problems that we faced with the obtained data was that a lot of mandis had missing data. Another issue that came out was even if some mandis had enough data, they had some huge gaps in the data for long contiguous periods. This can introduce error in our analysis if not handled properly. Since this would hinder in quantitative analysis(discussed later),

Mandi	**Center**	**Mandi Price Data Available (%)**	**Max Missing Period(in days)**
Bangalore	Bengaluru	79.3	50
Azadpur	Delhi	89.4	36
Pune	Mumbai	72	27
Lasalgaon	Mumbai	65.5	58
Bahraich	Lucknow	82.1	22

Table 2.1: Selected Mandis

we decided to shortlist mandis based on the following two criteria:
Only those mandis would be selected which

1. Had atleast 65% of data(both mandi price and arrivals) available for the chosen duration.

2. Had no missing data(both mandi price and arrivals) for a contiguous period of more than 60 days.

Based on our selection criteria, we were able to get mandis listed in the table 2.1. We have also mentioned the corresponding centers that are associated with the selected mandis.

Mandi	**Mandi Arrival Data Available (%)**	**Max. Missing Period(in days)**
Bangalore	83.5	25
Azadpur	93.8	19
Pune	72.1	27
Lasalgaon	65.6	58
Bahraich	86.1	22

Table 2.2: More information on Selected Mandis

2.3 Interpolation

In the previous section, we observed that our selection criteria outputted the mandis from Maharashtra, Karnataka, Delhi and Uttar Pradesh. Even after passing the selection criteria, these mandis had few missing points. So we used polynomial interpolation to fill these missing points. It is a cubic spline method which takes into account the curve and tangent to interpolate between two points. After interpolation, figure 2.1 shows how mandi price series looks for the Azadpur(Delhi) mandi.

2.3 Interpolation

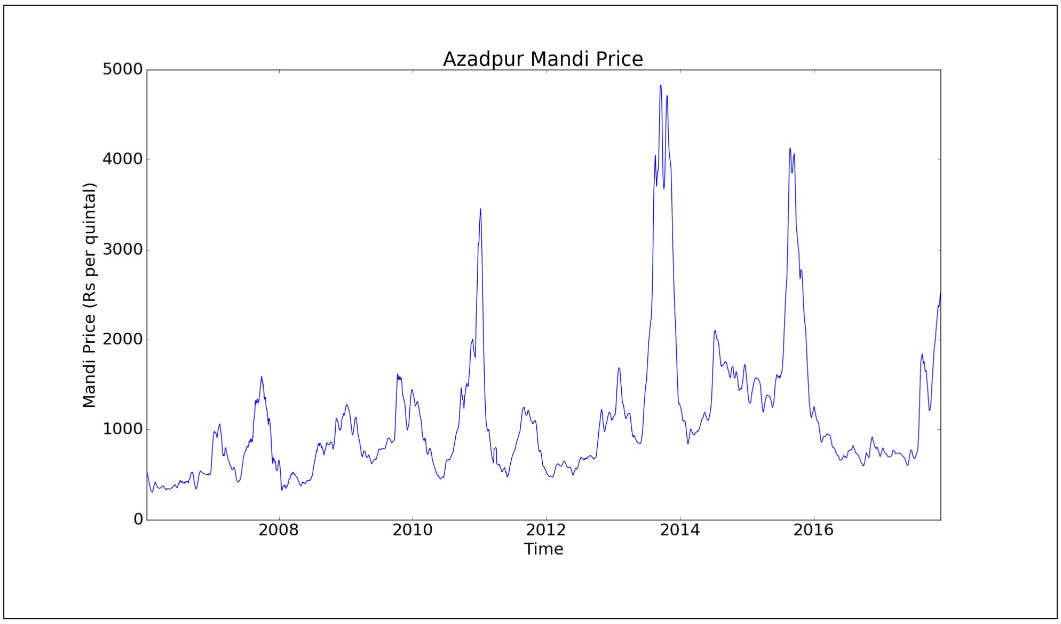


Figure 2.1: Interpolated Azadpur Mandi Price Series

Chapter 3

Context Of Onion Production in India

In this chapter we will be discussing about the onion cultivation and supply chain in the context of Indian markets. We will try to understand the normal working of the supply chain and how disturbances in it in the form of weather abnormalities can initiate fluctuations in the onion prices and give incentive to certain components of the supply chain to manipulate the prices the way they want in order to make extra money.

3.1 Background Study

India is the second largest onion producing nation in the world after China. In spite of the above fact, there is very little stability to the volatility of onion prices in the nation. But before moving to that, we would like to study about the onion cultivation cycle in India.
Table 3.1 shows the list of top onion producing states in India. We can see that Maharashtra ranks first in the overall production with a total share of 28.32% and Karnataka ranks second in the list. This also increases the validity of our study because 3 out of 5 selected mandis are from these two regions.

State	Production(in 1000 tonnes)
Maharashtra	5362
Karnataka	2985.8
Madhya Pradesh	2967.4
Bihar	1247.3
Gujarat	1126.5
Rajasthan	800.1
Haryana	667.1
Andhra Pradesh	575.6
Telangana	419.1
Uttar Pradesh	413.4

Table 3.1: List of Top onion producing states in India

3.1.1 Onion Cultivation Cycle

Onion is a seasonal crop and is mainly grown in two broad crop seasons in India, Rabi and Kharif. Table 3.2 shows the different steps in the onion harvesting cycle in the two seasons. Kharif onions are sown during the months of June to September which also happen to the monsoon months in India and are harvested during the period November-February. On the contrary, Rabi onions are sown after the monsoon months in October and November and harvested during the summer months of March to May. Out of these two seasons, onions grown as Rabi crops constitute 60% of the entire onion production in India. The rabi crops are used for consumption starting May and continue till October-November when the rabi stocks start depleting and kharif crops start coming in.

This cultivation cycle is also verified by the plot of average arrivals(fig 3.1) during a year using the data extracted from the internet. We can clearly see a peak in the arrivals during January-March which corresponds to the Kharif

Season	Sowing	Transplanting	Harvesting
Kharif	July-Sep	Sep-Oct	Dec-Mar
Rabi	Oct-Nov	Dec-Jan	Apr-May

Table 3.2: Onion Cultivation Cycle in India

arrivals. Following this we can see the next peak in arrivals during April-June when the rabi harvests start pouring in. After this the rabi harvests are stored and used till the festive season in Oct-November. The arrivals graph also shows a decreasing trend during this period. Moving ahead we begin to see an upward trend which corresponds to the kharif arrivals as discussed earlier.

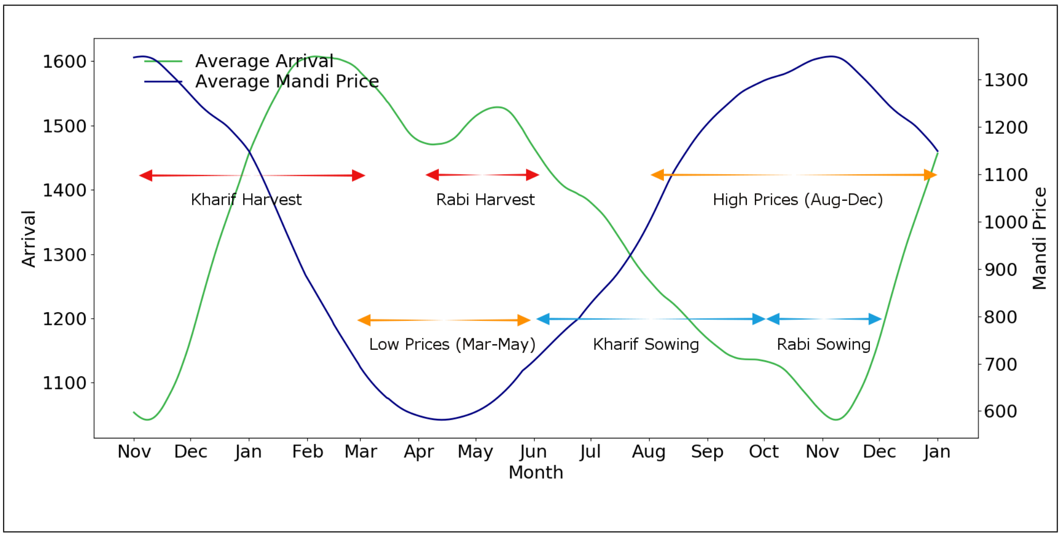


Figure 3.1: Average Mandi Price and Arrivals

3.1.2 Pricing Movements

In the previous section, we got to know about how onion is grown in India. In this section we look at how onion prices vary during a year and how it is related to the norms of the normal demand-supply pricing theory. We all know that price for any commodity increases in the period of scarcity of that product and the price takes a diving plunge when the good is in ample amount in the market.

In fig 3.1, we see that price starts decreasing when the kharif arrivals start coming in during December-February. After this they continue decreasing because of the rabi arrivals in the summer months where the prices hit the bottom most point. Following this, as the rabi harvests start getting used up and thereis increasing demand for onions during the festive season, the price begin to show a continuous upward movement before reaching their peak in the months prior to the kharif arrivals. This implies that onions also cater to the anticipated dynamics of the market throughout a year. Now we will learn how the onions reach to the consumers from the fields they are grown in and the intermediate steps involved.

3.1.3 Supply Chain

There are mainly three entities involved in the supply chain namely farmers, traders and consumers. After harvesting season, farmers bring their produce to the mandis where it is sold to the traders at a very nominal mandi price. The produce then is stored by the traders and jump through a number of intermediate traders to finally reach the retail shops where it is sold to the consumers at the retail price. Because there are a number of traders present in the chain, the price keeps on adding up and the final price is paid by the customers.

Under the APMC act, farmers are required to bring their produce to the established mandis where their crop is sold to the traders by the agents working there. Now the problem arises because of a combination of several factors. Any disturbances in rainfall affect the amount of produce being brought by the farmers in the mandis. Since the farmers do not have sufficient resources

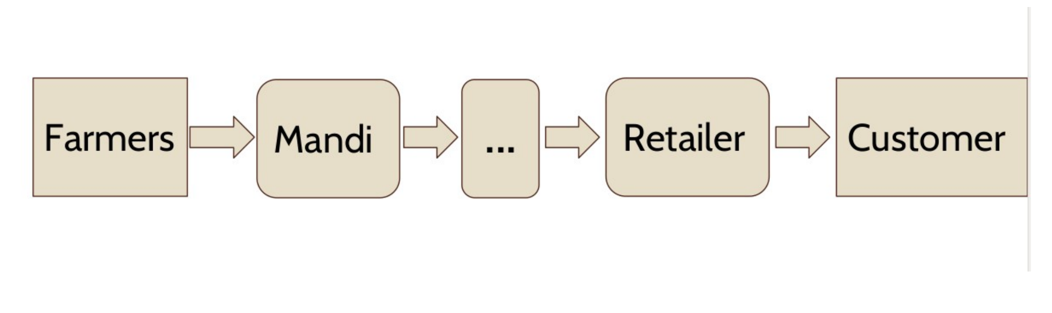


Figure 3.2: Onion Supply Chain

and are eager to sell their produce, the traders try to use this situation in order to extort the best prices out of the poor farmers. This happens because onions are still not supported by MSPs. The plight of the farmers is worsened if there is delayed rainfall which means destruction of harvested crops because of unavailability of proper storage structures with the farmers. So the farmers are forced to sell the crops at an even lower price. This ends up with farmers earning very low profits for their produce.

Now after the traders have purchased the crops at nominal prices, they store the crops in their storage facilities and also at times create artificial deficit which leads to increase in the retail prices. After the prices are soaring high, the traders now start releasing their produce slowly in the market to earn huge amounts of profits. This is how hoarding initiates with natural factors and then is exploited by the traders to make huge profits. These hoarding activities were confirmed when we studied the newspaper articles for the given period. After the government realises the artificial increase in the prices, they try to curb the prices by allowing the officials to raid the traders. A study conducted in 2012 by the Competition Commission Of India also found significant cartelization among onion traders to manipulate the onion prices and prevent the entry of new players into the network.

Chapter 4

Qualitative Analysis

In this chapter we would like to to get a deeper understanding of how the different factors are related to each other and how they behave during an anomalous event. We intend to do this by broadly studying about the following topics:

- We have 5 mandis and 4 centers under our consideration in the analysis. We would look at how we can classify the mandis as being source mandi or a terminal mandi. This will help us to understand the flow of crops in a broader context.
- The main time series that we have at our disposal is that of mandi and retail price and mandi arrivals. We would therefore look at how these factors are related with each other by using delayed correlation.
- In this chapter, we would take a deeper look at the past events that took place in the past decade and get a better overview of the price and arrival movements during the anomalous events and try to differentiate between the different anomalous events in a qualitative manner with the help of examples from the past.

4.1 Source v/s Terminal Mandis

Source Mandis are the mandis which are the major producers of the crops(here onions) and the produce is mainly supplied by the farmers directly by bringing into the mandis. So the demand in these mandis is fulfilled by

the crops grown and brought in the areas nearby.
Whereas **terminal mandis** are those where the demand for the prodcut is mainly fulfilled by the crops brought in from the source mandis. Therefore the supply of the produce depend on the demand of the population in the region. Now we look into a couple of statistics which help us differentiate between the two types of mandis in a better way.

4.1.1 Volatility in Arrivals

One of the key differences between source and terminal mandis is the volatility in the arrivals. We measure volatility in the arrivals by calculating *Z score* of daily arrivals. Z score is simply calculated by dividing the standard deviation of daily arrivals by the daily mean. It gives us a measure of how the arrival volume fluctuate for a given mandi. Table 4.1 shows the mandis with highest z scores and table shows the mandis with lowest z scores amongst all the mandis that we observed.

Center	**Mandi**	**Mean Arrival**	**Z Score**
Bengaluru	Bangalore	2761	0.51
Mumbai	Pune	1167	0.42
Mumbai	Lasalgaon	1339	0.25

Table 4.1: Mandis with Highest Z Scores

Center	**Mandi**	**Mean Arrival**	**Z Score**
Lucknow	Bahraich	110	0.032
Delhi	Azadpur	907	0.122
Hyderabad	Karimnagar	762	0.126

Table 4.2: Mandis with lowest Z Scores

4.1 Source v/s Terminal Mandis

We can clearly see that mandis with high z scores are the source mandis because the arrivals in the source mandis do not depend on the demand. It is dependent on the produce brought by the farmers which can show large fluctuations from day to day because of sudden weather changes leading to destruction of crops, poor storage, insufficient rainfall etc. On the other hand mandis with low z scores are the terminal mandis where the arrival volumes in the mandis depend on the demand in the city as the crop is not grown there. Also we can see that in general the source mandis have a higher mean arrival than the terminal mandis.

It can also be visualised in a better way by plotting the mandi arrival graphs(fig 4.1) for an average year. Here we take Lasalgaon as a source mandi and Azadpur to be a terminal mandi. We can clearly that there are large variations in the arrivals of Lasalgaon mandi(source mandi) whereas the Azadpur arrivals(terminal mandi) do not show such large variations. But this does not have any impact on the average mandi prices at the two places as can be clearly seen in fig 4.2.

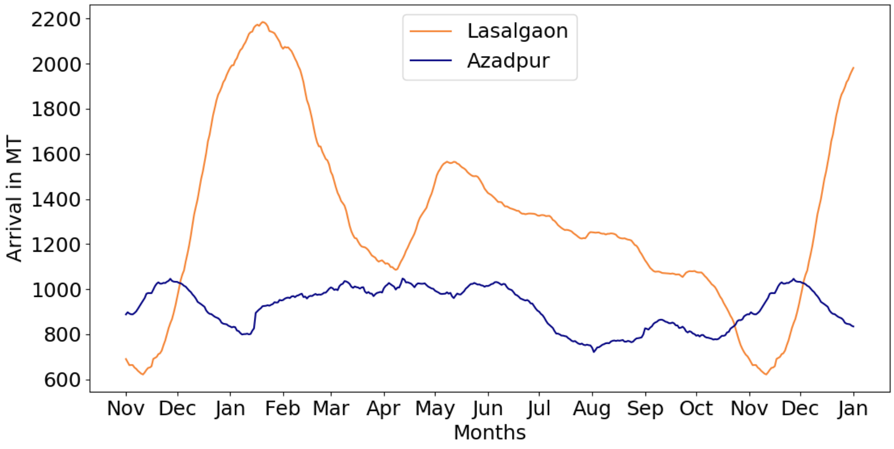


Figure 4.1: Difference in volatility in Arrivals between Source and Terminal mandi

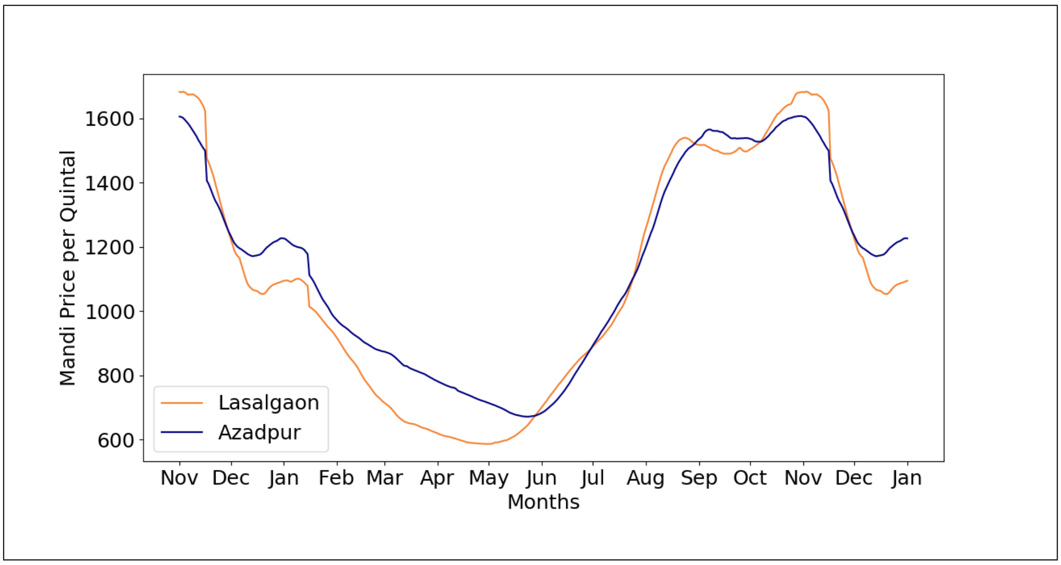


Figure 4.2: Similarity in the Mandi Prices of Source and Terminal Mandi

4.1.2 Delayed Correlations for various parameters

Now we see how the retail and mandi prices vary between the source and terminal mandis.

In fig 4.3, we plotted shifted or delayed correlations between average source retail prices and average terminal retail prices. We observe that a peak occurs at X = -1 which simply means that retail prices in terminal mandis(centers) follow the source retail prices by just 1 day. This is intuitive because any changes in the retail prices would initially begin at the source mandis and then will be quickly communicated to the terminal mandis within a day or two. This gives us some knowledge of how strong the network of the traders is in this nation.

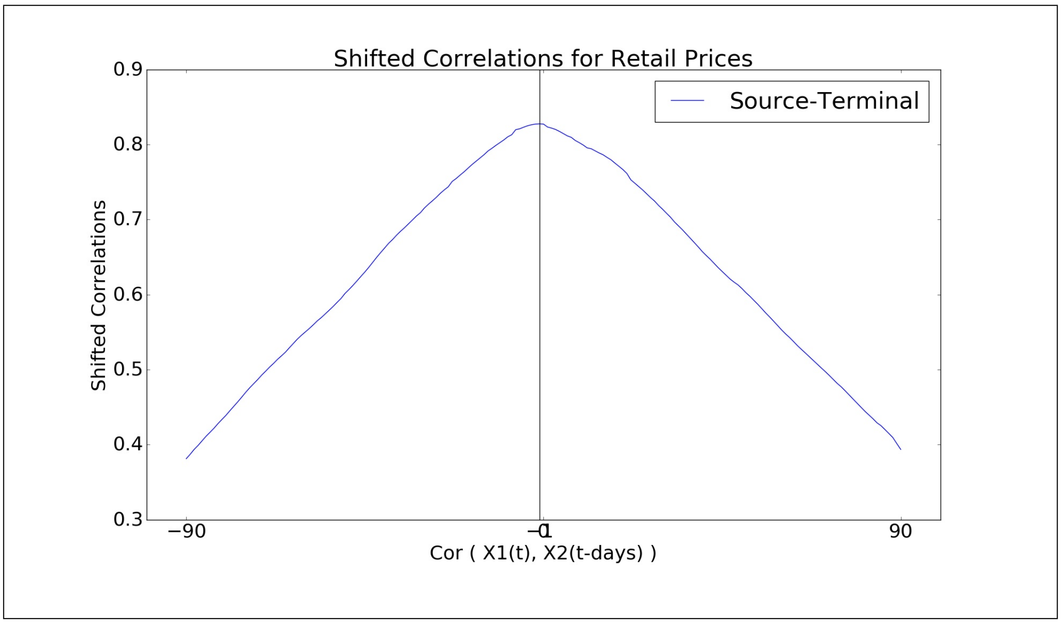


Figure 4.3: Shifted Correlation between Source and Terminal Retail Prices

Similarly, we have the shifted correlations between average source mandi prices and average terminal mandi prices in fig 4.4. This time the peak occurs at X = -5. This means that the mandi prices at terminal mandis follow the trend at source mandi prices after a duration of about 5 days. This leads us to the conclude that retail price information travel faster than the mandi price information because the network of the farmers is not as strong as that of the traders in the market.

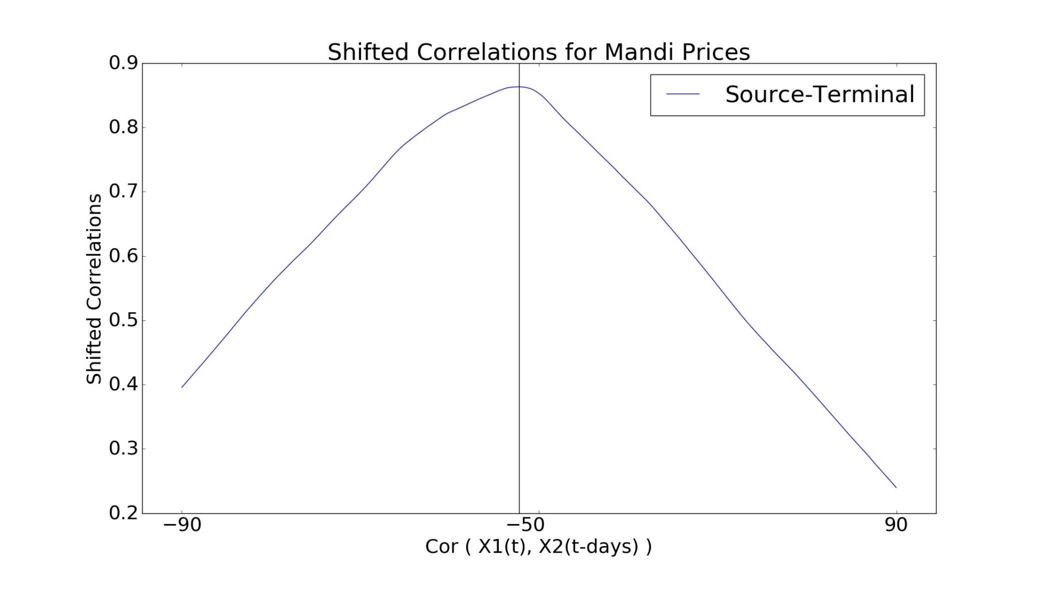


Figure 4.4: Shifted Correlation between Source and Terminal Mandi Prices

4.2 Analysis of Different factors

In the previous section, we learnt how the terminal mandis and source mandis are different from each other and also the correlations between the different parameters. In this section we will briefly look at the relations between the three factors(mandi price, retail price, mandi arrivals) in general.

In fig 4.5, the graph between mandi price and arrivals is a downward parabola with a trough at X = -5. This is in line with our intuition that when the arrivals increase, it takes about 5 days for the mandi price to decrease and vice-versa.

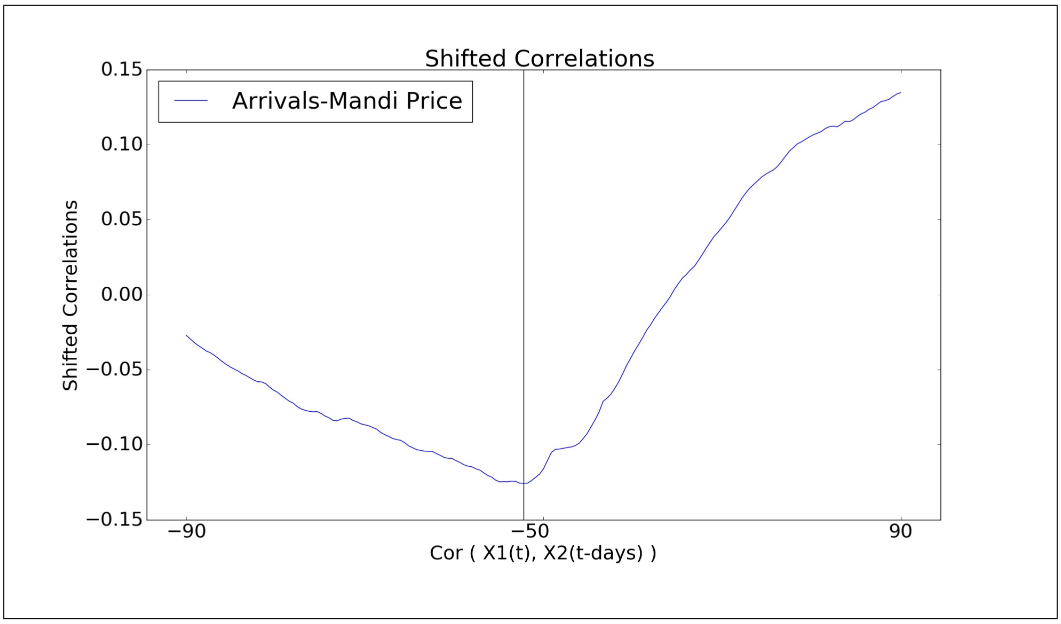


Figure 4.5: Shifted Correlation between Arrivals and Mandi Prices

Also we expect that once the mandi price increase or decrease, this information takes some time to be transmitted to the centers and so the change in retail price is observed with a delay of about 3 days. This means within a region, the changes in the retail price due to the changes in the mandi price start occurring in about 3 days. This is confirmed by the plot in fig 4.6.

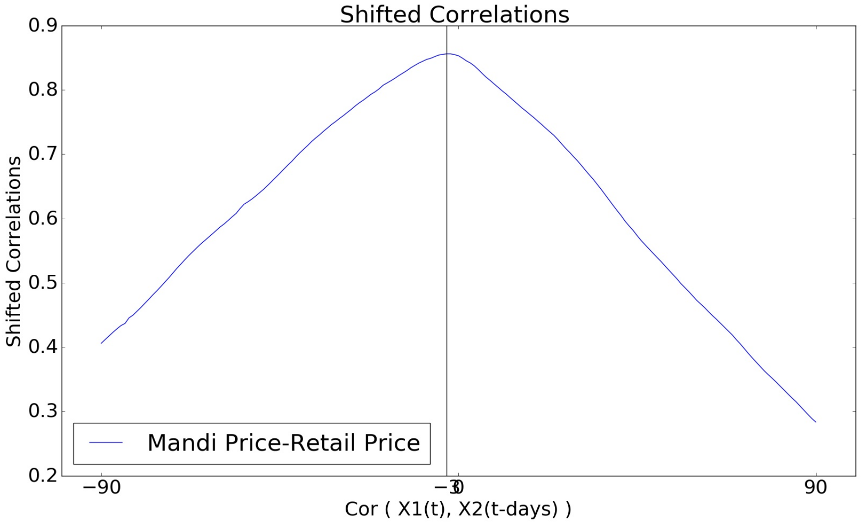


Figure 4.6: Shifted Correlation between Mandi and Retail Prices

4.3 Different Anomalous Events

In this section, we look at three major anomalies of 2007, 2010 and 2013 and describe the reasons of their occurrence based on the data available to us.

4.3.1 2013 Scenario

As we can see in the rainfall graph for the year 2013, the rainfall started a bit early and above average and was erratic in the later monsoon months. This led to destruction of the kharif crops that year which also confirmed with the newspaper reports stating the same. So we can see a huge dip in arrivals in the Sep-Oct period. This led to a gradual increase in both the mandi and retail price of onions.
But in the months from Jan-March of the next year, we observe that the arrivals are quite above the average arrivals during that time of the year.

This points to the possibility of hoarding of the crop by the traders and then releasing it in the later months to make huge profits because of the increased prices. This claim is also supported by the newspaper reports that stated that the hoarders had stocked the produce of the previous season and tried to hide this under the news of disturbances in rainfall in order for the dip in arrival to seem legitimate. So we see that this was a case of hoarding anomaly which initiated with weather disturbances but was later opportunistically leveraged by the traders to make huge profits.

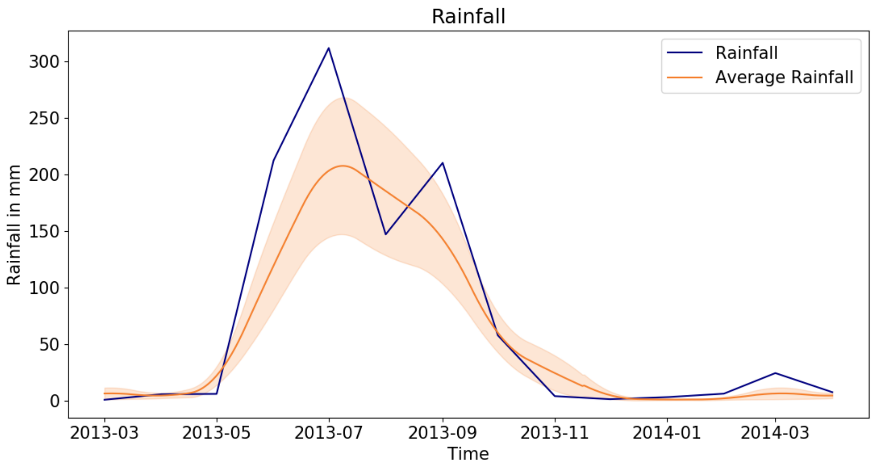


Figure 4.7: Rainfall During 2013

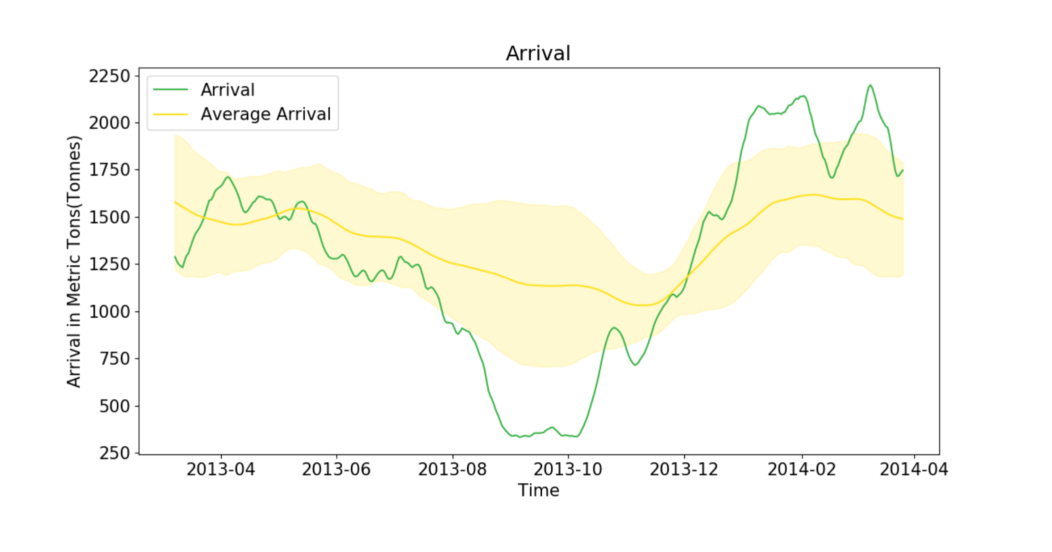


Figure 4.8: 2013 Mandi Arrivals

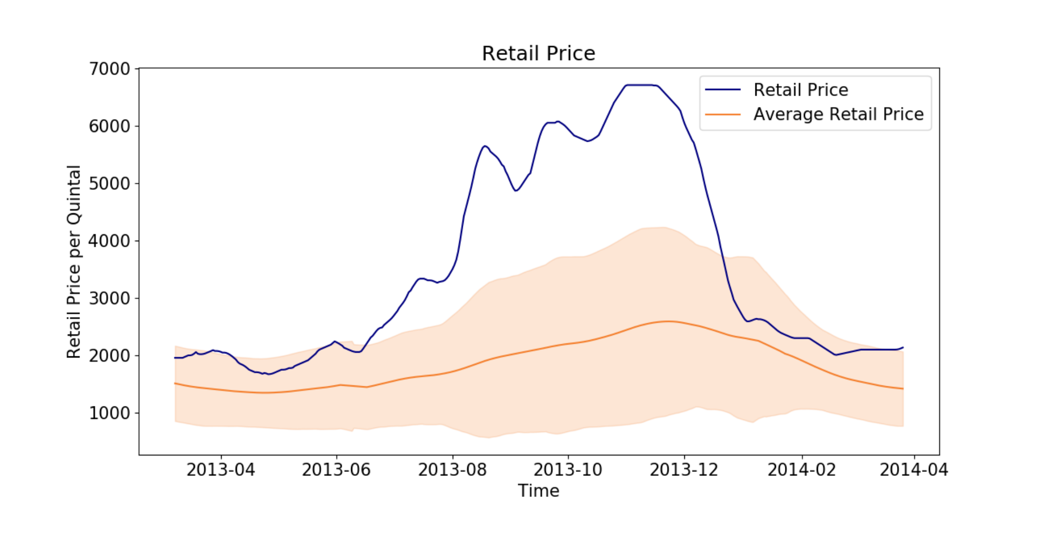


Figure 4.9: 2013 Retail Price

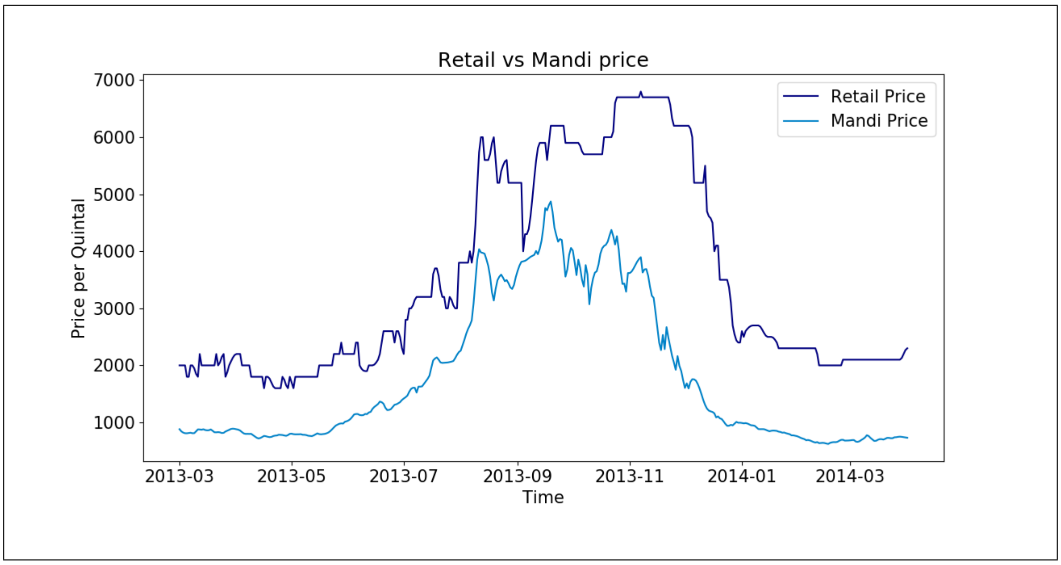


Figure 4.10: 2013 Retail Price and Mandi Price

4.3.2 2010 Scenario

2010 case was similar to the 2013 case. If we see the rainfall during 2010, we observe that there was unseasonal rainfall during November 2010. This excess rainfall was reported to have destroyed large quantities of kharif harvests in parts of Maharashtra and Karnataka. This led to poor arrivals in the following Decemeber-January which in turn led to sharp rise in prices of onions.

Soon there were steps taken by the government to curb the prices by raiding the traders in order to stop hoarding. In this case also we observe that soon after the rise in prices, there was increase in the arrival volumes during March 2011. This again points to the hoarding activities carried out by the traders in order to make the most out of a situation which started because of natural causes.

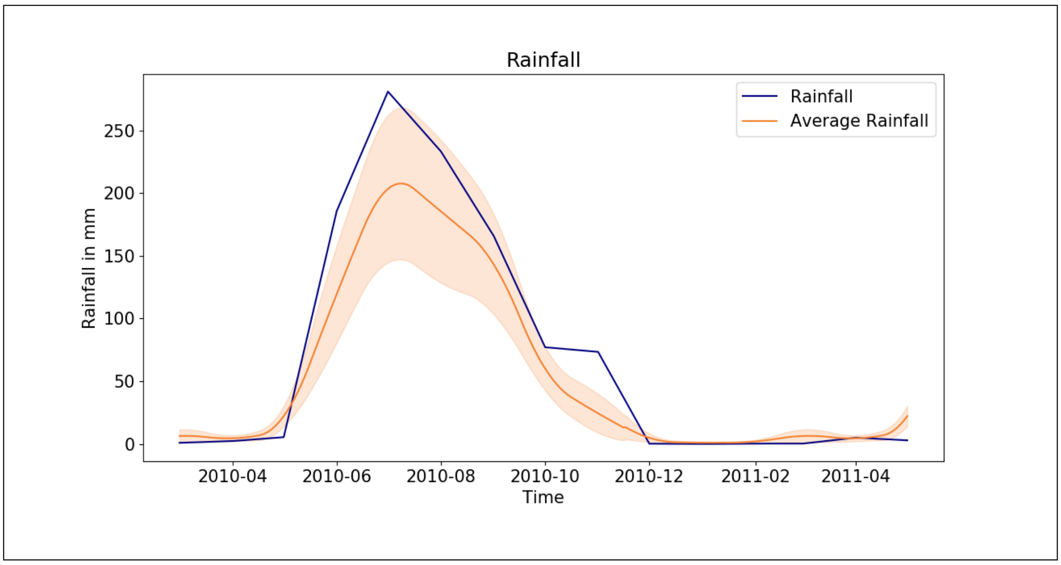


Figure 4.11: Rainfall During 2010

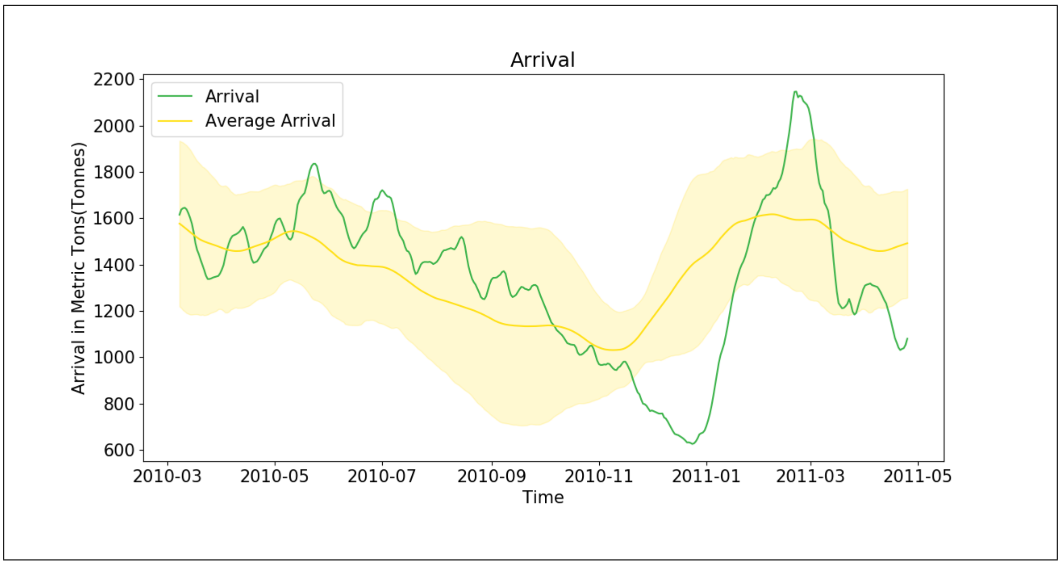


Figure 4.12: 2010 Mandi Arrivals

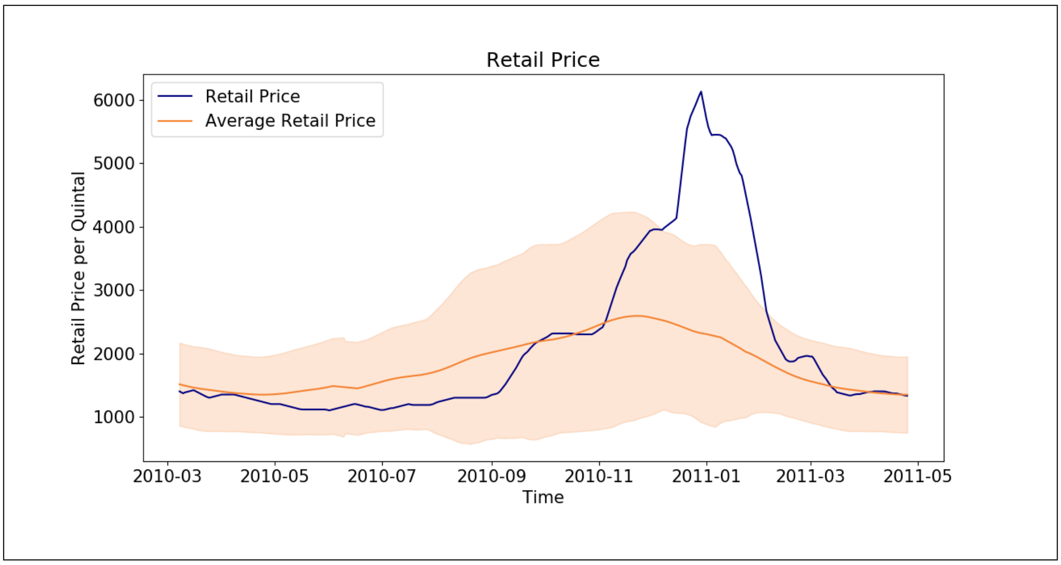


Figure 4.13: 2010 Retail Price

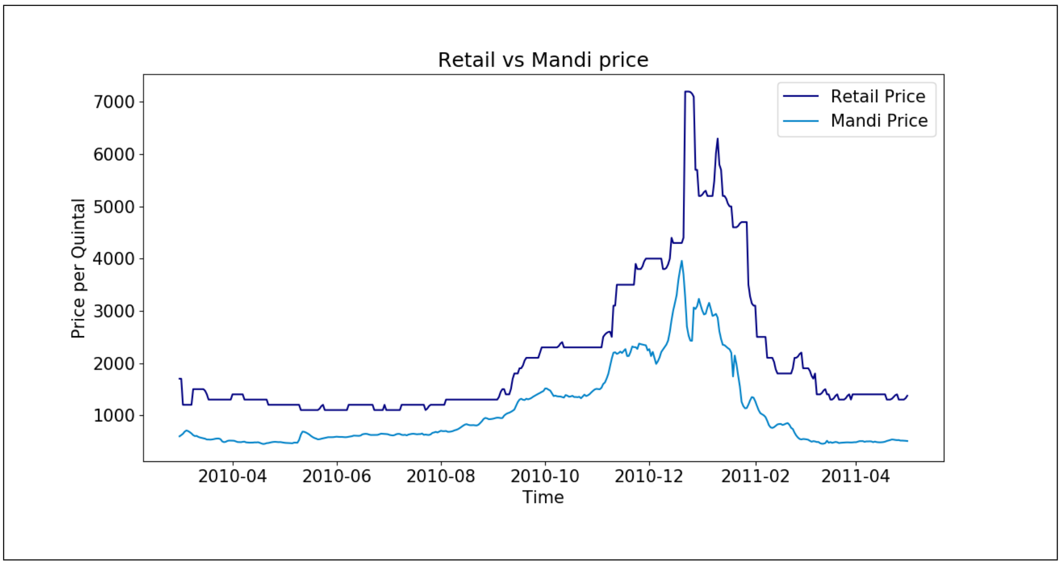


Figure 4.14: 2010 Retail Price and Mandi Price

4.3.3 2007 Scenario

This case is different from the previous two because the year 2007 witnessed a weather anomaly but was not followed by hoarding. To start with, this year experienced more than normal rainfall during the monsoon months which reportedly destroyed the crops. This led to increase in the prices but contrary to the previous cases, the increase was not significantly large. Because of the unexpected rainfall, there was dip in the arrivals during Sep-Nov.
But after this we observe that things got back to normal in the following months. We can see the arrivals also remained within limits contrary to the 2010 and 2013 cases. This was maybe because the traders did not exploit the situation to turn the tide in their favours. This was also confirmed with the newspaper which blamed the price rise to be solely because of weather disturbances and no hoarding activities were reported during this period. So this was an anomalous case which was only due to the weather and natural circumstances.

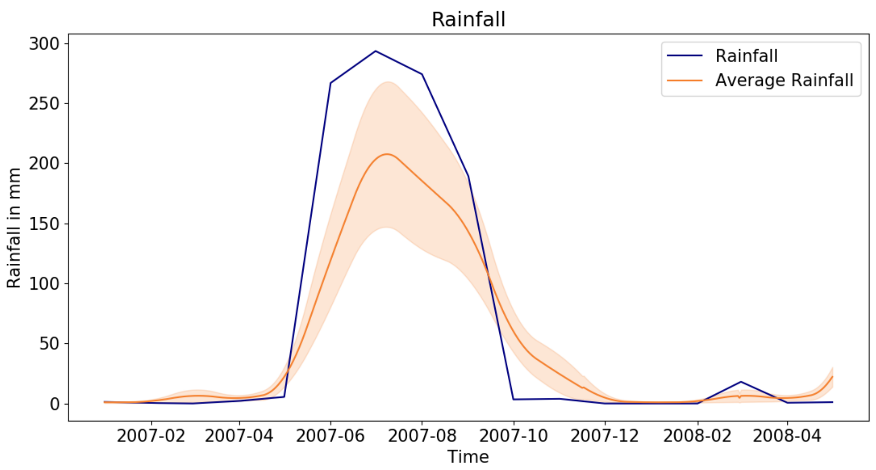


Figure 4.15: Rainfall During 2007

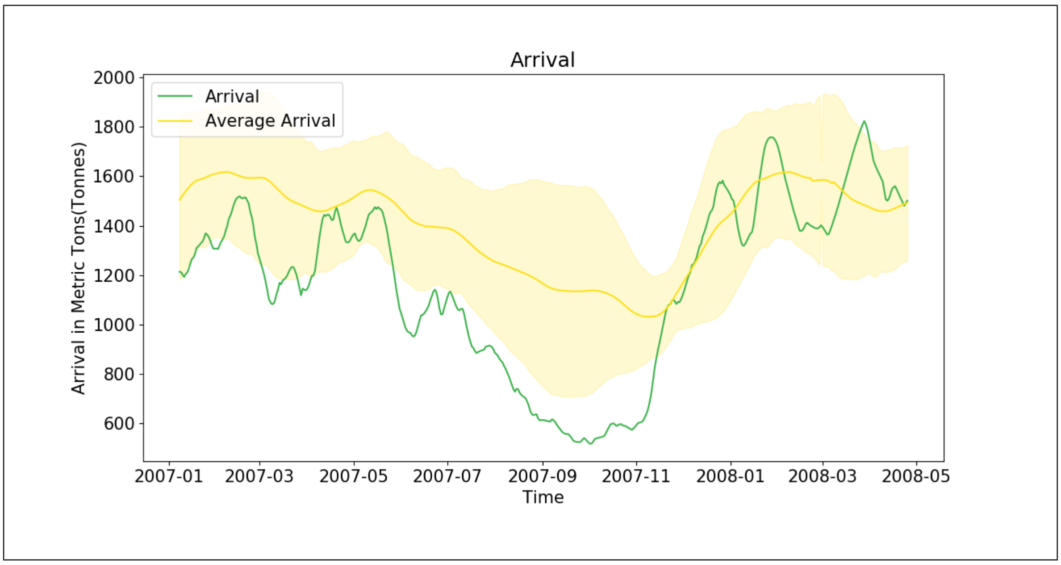


Figure 4.16: 2007 Mandi Arrivals

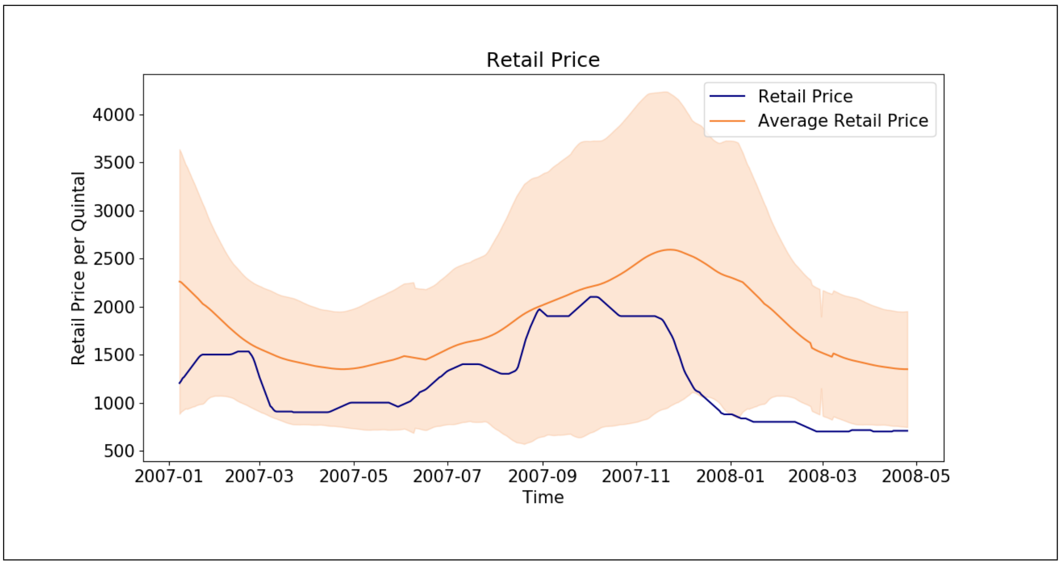


Figure 4.17: 2007 Retail Price

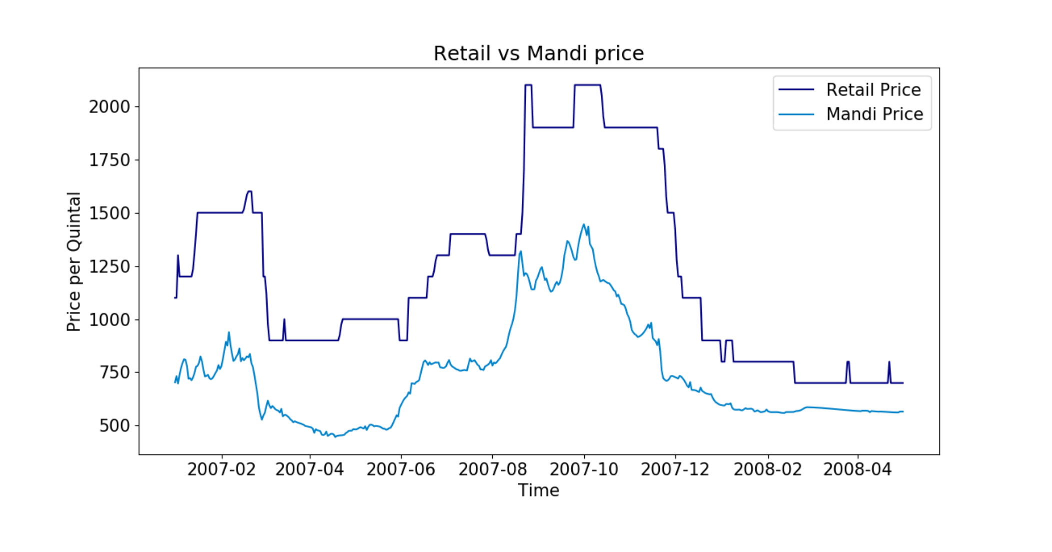


Figure 4.18: 2007 Retail Price and Mandi Price

In this section, we analysed the three anomalous cases and identified the major types of anomalies, weather anomalies and hoarding anomalies preceded by weather disturbances. We also formed basic rules to identify these types of anomalies that took place over large duration of time. In the next section, we now build a system to detect anomalies in the first place and then to classify them based on the reason of their occurrence.

Chapter 5

Quantitative Analysis

In this chapter, we would look at proposing a system which can help us in identifying anomalies and also classifying them based on different reasons. We would also look at how we can improve upon the labelling of the examples and the impact of this on the overall two step system. The two step system for anomaly detection and classification is represented in the fig 5.1.

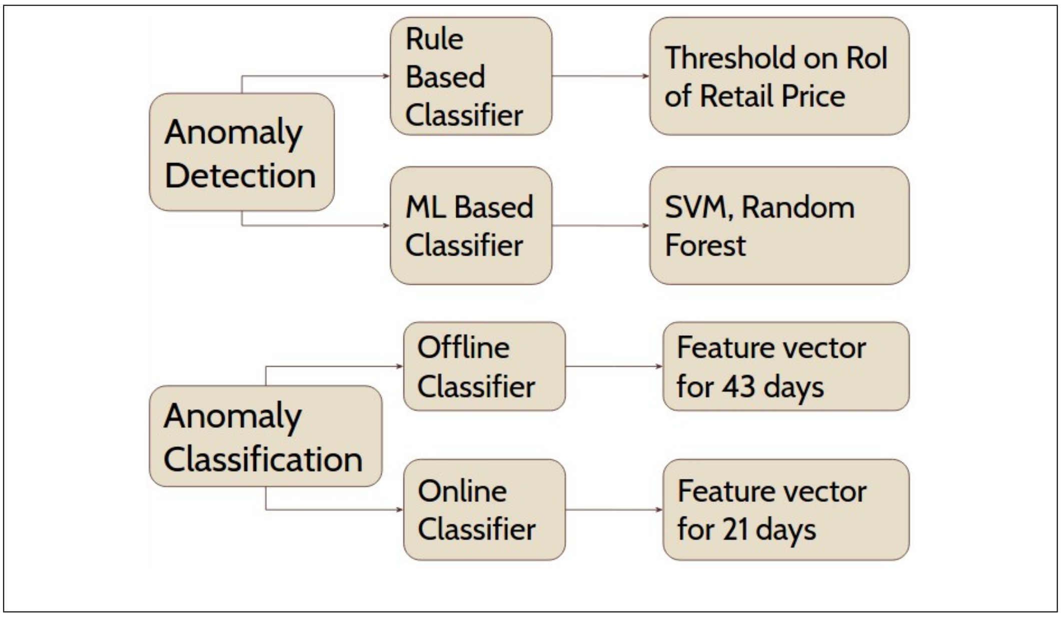


Figure 5.1: Two Step Classifier

- *Anomaly Detection:* In this first step, we basically just identify anomalies amongst all the events available. The windows for all events were created by reading the newspaper articles and identifying if they talk about any anomalous thing happening. After identifying these, we get a collection of all the events during the entire period. For a binary

classifier, our next step was to get non anomalous or normal periods during the same period.
We define normal period as a continuous duration of 43 days which does not overlap with the already present anomalies and the difference between the maximum and minimum retail price during that period does not cross rupees 300 per quintal(This value was approximately one fourth of the average difference between the max and min retail price for all the events). Using this process, we have 128 anomalous and 144 non anomalous periods.

- *Anomaly Classification:* In the second step, we use only the anomalous periods and classify them into hoarding or weather anomaly. According to our labelling, we have 58 hoarding and 70 weather anomalies. This step involves using the random forest classifier to perform the binary classification.

5.1 Anomaly Detection

As discussed earlier, this step involves using two methodologies to identify anomalies. One is the rule based method and the other is building a binary classifier.

5.1.1 Rule Based Method

The data available to us is the mandi and retail price series and mandi arrivals. In this method, we evaluated different statistics on the time series to differentiate between the anomalous and non anomalous events. Some of the statistics are rate of increase of retail price during the interval, or average difference between the actual and expected value of time series during that interval. The results in the table 5.1 are shown corresponding to the best statistic observed. The threshold was kept on the rate of increase of retail price during the first 21 days and if the slope of the regressed line was greater than 2, it was classified as an anomalous period. The accuracy in this case comes out to be 66.9%.

Actual-Predicted	**Anomalous**	**Non-Anomalous**
Anomalous	91	37
Non-Anomalous	53	91
Precision	0.63	0.71
Recall	0.71	0.63

Table 5.1: Result of Rule Based Classifier. Acc: 66.9%. Anomalous: 128, Non Anomalous: 144

5.1.2 ML Based Classifier

In this we used a random forest binary classifier for testing. We used various combinations of the base data and used them as feature vectors to train the classifier. In the table 5.2, we show the accuracies using the various combinations.

Feature Vector Set	**Accuracy**
RP	66.9
MP	70.6
RP, MP	71.3
RP-MP, MA	69.8
RP-MP	70.6
RP, MA	65.8
RP, MP, MA	70.22
RP/MP	50

Table 5.2: Accuracy with Different Sets of Feature Vector. RP: Retail Price. MP: Mandi Price. MA: Mandi Arrivals

Further in table 5.3, we show the confusion matrix corresponding to the best feature vectors. In our analysis, we have chosen the model which gives the best validation set accuracy. The accuracies shown in the table are the cross validation accuracies. The validation sets are formed by dividing the entire duration into 6 month periods and each set has a number of anomalies falling in it. The accuracies are then obtained in a cross validation format by leaving out single set for testing in each iteration. We observe that the accuracy of the ML based classifier is slightly better than the rule based classifier.

Actual-Predicted	**Anomalous**	**Non-Anomalous**
Anomalous	71	57
Non-Anomalous	21	123
Precision	0.77	0.68
Recall	0.55	0.86

Table 5.3: Result of ML Based Classifier. Acc: 71.3%. Anomalous: 128, Non Anomalous: 144

5.2 Anomaly Classification

In the step 2, we distinguish between the two major kinds of anomalies, namely hoarding and weather. In this, we again use a similar binary classifier as the previous case. Here in the table 5.4 for the offline case, we use the entire 43 day data for training whereas in the online case, we use incremental number of days for the training. The validation accuracies are shown for the feature sets with the maximum accuracy to avoid confusion. In the online case, we observe that accuracy increases till the middle and then plateaus without increasing much. This means that most of the information about the type of anomaly resides in the first half of the data rather than the later half.

Actual-Predicted	**Hoarding**	**Weather**
Hoarding	30	28
Weather	20	50
Precision	0.6	0.64
Recall	0.52	0.71

Table 5.4: Result of ML Based Classifier. Acc: 62.5%. Hoarding: 58, Weather: 70

Days	**Accuracy**
14	60.9
21	63.2
28	60.15
35	61.7
43	62.5

Table 5.5: Accuracy for Online Classifier for Step 2

5.3 Improvements

In the previous section, we saw the performance of the binary classifier in both the steps. Since the performance was not quite good enough, we decided to look at ways to improve it. Some of the ways are:

- The first thing we did was to look back at the labels for the anomalous periods which are used as ground truths for the classifiers. Earlier we had labelled a window just based on the newspaper articles present in that window. This time we also looked at the various time series data for a particular event to see whether it really represents an anomaly or not. While labelling, it was observed that many of the anomalous events were not anomalous seeing the price variations during that interval. And some of them were not clear as to whether they were anomalous or not. So we created another class and put all such events under unlabelled events. So we now have 71 anomalous events, 90 non anomalous and 63 unlabelled events. This time, we use semi supervised algorithm for the step 1 along with the previous algorithms and compare them on the same validation set accuracies.
- Uptill now, we have only been using simple arithmetic combinations of the time series data we had for our feature vectors. Now we included the first derivative to the set of the feature vector.
- We also varied the window size (between 7 to 43 days) to look at the effects but this did not have any positive impact on the results.
- Our study only includes 5 mandis. So to make our data more robust, we relaxed the criteria for calculating average mandi arrivals. We selected all those mandis for a center which had more than 50% of the data. Now we looked at those periods where all of these mandis had the data available and calculated the contribution of each mandi to the average

mandi arrival. We used this contribution or weights of each mandi to adjust the average arrival in the time periods where a particular mandi had missing data. This helped make our data more robust.

Now let us again look at the results of the step 1 & 2 after making the above changes. For the step 1, the difference between the supervised and semi supervised algorithms is that supervised algorithm takes only the known labelled events for training but the semi supervised algorithm also takes the unlabelled events for training. The results for step 1 are tabulated in tables 5.6 & 5.7. For the step 2, we have listed the revised results in table 5.8.

Models	**Accuracy**
Rule Based	74.53
Semi Supervised	75.15
Supervised	86.95

Table 5.6: Comparision of step 1 accuracies for different models.

We observe that in table 5.5, the supervised learning algorithm performs the best for the case of anomaly detection. Also we observe that accuracies of both the rule based and the supervised algorithms have improved significantly than the earlier scenario which is mainly due to the relabelling of the events and inclusion of derivatives in the feature vector.

Actual-Predicted	**Anomalous**	**Non-Anomalous**
Anomalous	57	14
Non-Anomalous	7	83
Precision	0.85	0.86
Recall	0.8	0.92

Table 5.7: Revised Result of Supervised Learning Classifier. Acc: 86.95%. Anomalous: 71, Non Anomalous: 90. Feature Vector: RP and it Derivative.

For the step 2, we did not observe any improvements in the result. This can be due to the fact that the training examples have significantly reduced in number and that we need more useful features to differentiate between hoarding and weather. It happens because both the types of anomalies might have quite similar characteristics which complicates the learning.

Actual-Predicted	**Hoarding**	**Weather**
Hoarding	15	16
Weather	11	29
Precision	0.57	0.64
Recall	0.48	0.72

Table 5.8: Revised Result of ML Based Classifier. Acc: 61.97%. Hoarding: 31, Weather: 40. Feature Vector: RP-MP

Days	**Accuracy**
14	64.7
21	63.3
28	63.3
35	60.56
43	61.97

Table 5.9: Revised Accuracy for Online Classifier for Step 2

	Newspaper Covered		*Not Newspaper Covered*	
Stage-2 Label	*Predicted*	*True Labels*	*Predicted*	*True Labels*
Hoarding	73	*18 Hoarding* *55 Weather*	61	*39 Abnormal* *22 Normal*
Weather	62	*10 Hoarding* *52 Weather*	89	*54 Abnormal* *35 Normal*
Total	135		150	

Table 5.10: Output over the entire period

5.4 Evaluation of the Entire System

Now that we have looked at the individual steps, we present the evaluation of the entire system when run in a pipeline manner. In this we scan the entire duration and use our rule based decision to identify anomalous period. Once a period passes the step 1, it is named as an anomaly and passed to step 2 where prediction is done on whether the anomalous period corresponds to a hoarding or to a weather anomaly. The true label for any period is decided based on the overlapping of that period with the already known anomalies. The final results for this process are listed in table 5.10. We can see that the overall system performs quite poorly in terms of detecting hoarding because of no proper distinction between the hoarding and weather anomalies in step 2. There is a lot of scope of improvement in this area. Nevertheless, all of this analysis helps us in aligning things into perspective and build a proper understanding of the entire hoarding situation and forms a basis for further improvement and analysis.

Chapter 6

Time Series Analysis for Prediction

Uptill now we have been framing the anomaly problem in the form of detection and classification of anomalies. However, in this section we take a more common approach towards time series analysis and use time series models for prediction of time series during the test period.

6.1 Theory

Autoregressive Model(AR)

An Autoregressive model AR(p) model is a regression model where lag values of the variable in consideration are used as regressors. The number of lag variables used in the model is denoted by the parameter p. AR(p) model is defined as:

$$y_t - \sum_{i=1}^{p} \phi_i y_{t-i} = \epsilon_t$$

where ϵ_t is prediction error, and $\phi_1, \phi_2, \ldots, \phi_n$ are the unknown coefficients of the AR model.

Moving Average Model(MA)

A Moving average MA(q) model is a regression model where lag values of the error term are used as regressors. In this model the parameter q specifies the maximum lag of error terms being used in the model. A MA(q) model is defined as:

$$y_t - \sum_{i=1}^{q} \phi_i \epsilon_{t-i} = \epsilon_t$$

where ϵ_t is prediction error, and $\phi_1, \phi_2, \ldots, \phi_n$ are the unknown coefficients of the MA model.

Autoregressive Integrated Moving Average Model(ARIMA)

The ARIMA model is a combination of both the AR and MA model. It takes into account the lag values of the variable in question and also the error terms. These models can be applied only if the time series under consideration is stationary(A series is called stationary if the mean and variance of the series does not change with time and remains constant). If this assumption does not hold, then we need to first transform the series in order to make it stationary. One of the ways of doing this is by performing the differencing operation. So in this case, a new parameter d is introduced to denote the number of differencing operations required to make the time series stationary. The parameters p and q hold the same meaning as before. Hence ARIMA(p, d, q) model can be written as:

$$\phi(L)\nabla^d y_t = \theta(L)\epsilon_t$$

where $\nabla y_t \equiv (1-L)y_t$ is the lag difference operator, L is the lag operator, i.e. $L^h y_t \equiv y_{t-h}$, $\phi(L)$ is a shorthand notation for

$$\phi(L) = 1 - \phi_1 L - \ldots - \phi_p L^p$$

and $\theta(L)$ is a shorthand notation for

$$\theta(L) = 1 + \theta_1 L + \ldots + \theta_q L^q$$

6.1 Theory

Seasonal Autoregressive Integrated Moving Average Model(SARIMA)

A seasonal ARIMA model takes care of the seasonality present in the time series. The parameter S defines the number of time periods until the pattern repeats itself in the time series. In a seasonal ARIMA model, seasonal AR and MA terms predict the variable in consideration using data values and errors at times with lags that are multiples of S. SARIMA models take care of both the seasonal and the non seasonal components. A general seasonal ARIMA (p,d,q) X (P,D,Q)S model is defined as:

$$\Phi(L^S)\phi(L)(y_t - \mu) = \Theta(L^S)\theta(L)\epsilon_t$$

where the non seasonal components are:
AR:

$$\phi(L) = 1 - \phi_1 L - \ldots - \phi_p L^p$$

MA:

$$\theta(L) = 1 + \theta_1 L + \ldots + \theta_q L^q$$

and the seasonal components are:

Seasonal AR:

$$\Phi(L^S) = 1 - \Phi_1 L^S - \ldots - \Phi_P L^{PS}$$

Seasonal MA:

$$\Theta(L^S) = 1 + \Theta_1 L^S + \ldots + \Theta_Q L^{QS}$$

and P is the seasonal AR order, D is the seasonal differencing and Q is the seasonal MA model.

6.2 Approach

Before moving on to prediction, we first check the stationarity of the time series since the models apply only to stationary time series. We have used *ADF* hypothesis test for checking stationarity. For our analysis, we will be using the mandi price of Mumbai. The entire process can be similarly applied for all the other series. Table 6.1 shows the results of running the ADF test. We can clearly see that the magnitude of test statistic is quite low and the p-value is not below 0.02, so the series is not stationary. We therefore work with the log transformation of the given series.

Test Statistic	**p-value**
-3.29	0.072

Table 6.1: ADF test for Mandi Price Mumbai

Now for fitting and evaluation of time series, we take the data from 2006 to 2016 as the training data and the data for 2017 as the testing data. We now wrote R scripts to decide the best parameters for the ARIMA and seasonal ARIMA models. The model is selected based on the AIC values(Akaike Information Criteria). The parameters are selected for the model which result in the minimum AIC value for the training data. The best models for both ARIMA and SARIMA are then used to predict the mandi price for the test period. The results are shown in the fig 6.1.

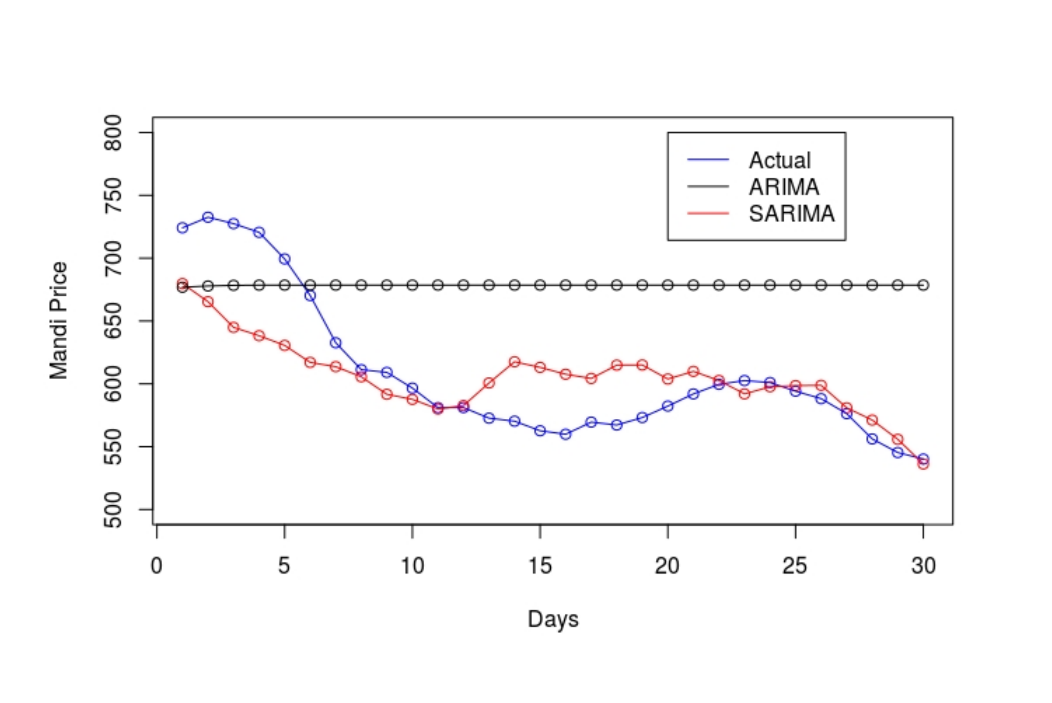


Figure 6.1: Prediction of ARIMA and seasonal ARIMA models for the test period Jan 2017

6.3 Results

MAPE: Mean Absolute Percentage Error. For analysis involving univariate time series models and prediction of time series values, MAPE is a common statistic to get a measure of the error. And 100 - MAPE is often used as a measure of the accuracy. MAPE is defined as:

$$MAPE = 100 * \sum_{i=1}^{no.\,of\,points} \frac{|Actual\,Value - Predicted\,Value|}{Actual\,Value}$$

Days	**MAPE**
7	8.40
14	5.55
21	5.92
28	4.73
35	5.229

Table 6.2: MAPE Values for SARIMA model for increasing number of days

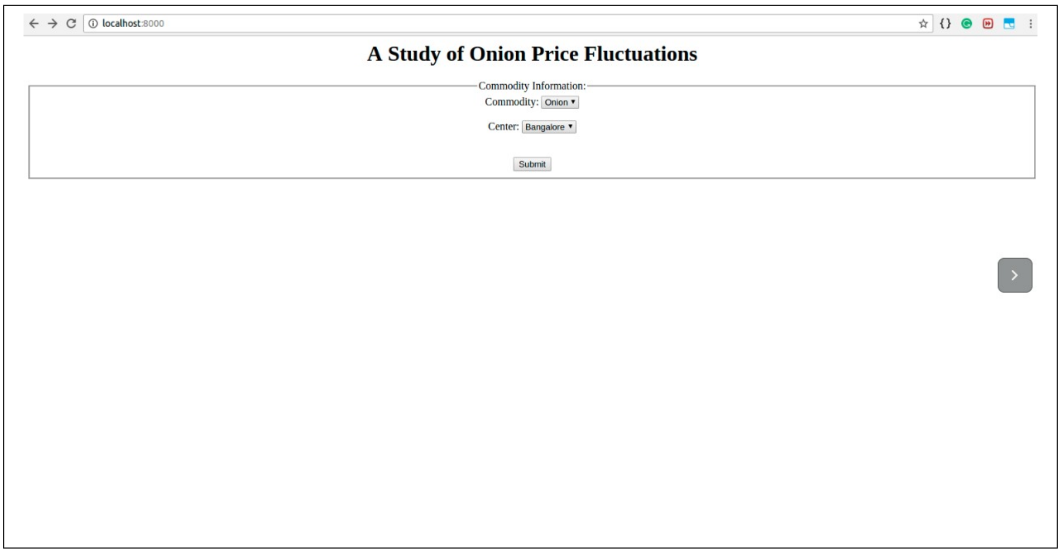
In the table 6.2 we have listed the MAPE values obtained when the learnt SARIMA model was used for prediction for varying number of days. In fig 6.1, we can clearly see that SARIMA model works better than the ARIMA model. The ARIMA predictions are just a simple straight line whereas the SARIMA model is able to capture the variations to some extent. This happens because there is inherent seasonality in the data and this seasonality is captured by the seasonal model. So we can say SARIMA models perform better than vanilla ARIMA models for prediction of onion prices.

Chapter 7

Web Portal

In this project, we have carried out both qualitative and quantitative analysis to get a deeper understanding about how the external dynamics affect the price of onions and how well we are able to explain them. To sum it all up, we also prepared a small and simple portal where we can visualise the price variations of different commodities(work is in progress for other commodities) and also analyse all the anomalous events with a closer detail. The portal allows to understand the price and arrival dynamics in any individual anomaly and also provides the links to the newspaper articles related to the anomaly. We now provide a step-by-step guide for a better understanding of the portal.

1. Fig 7.1 shows the home page of the portal. Here you can choose the commodity and the centre you want to look at. Then click the submit button.



A Study of Onion Price Fluctuations

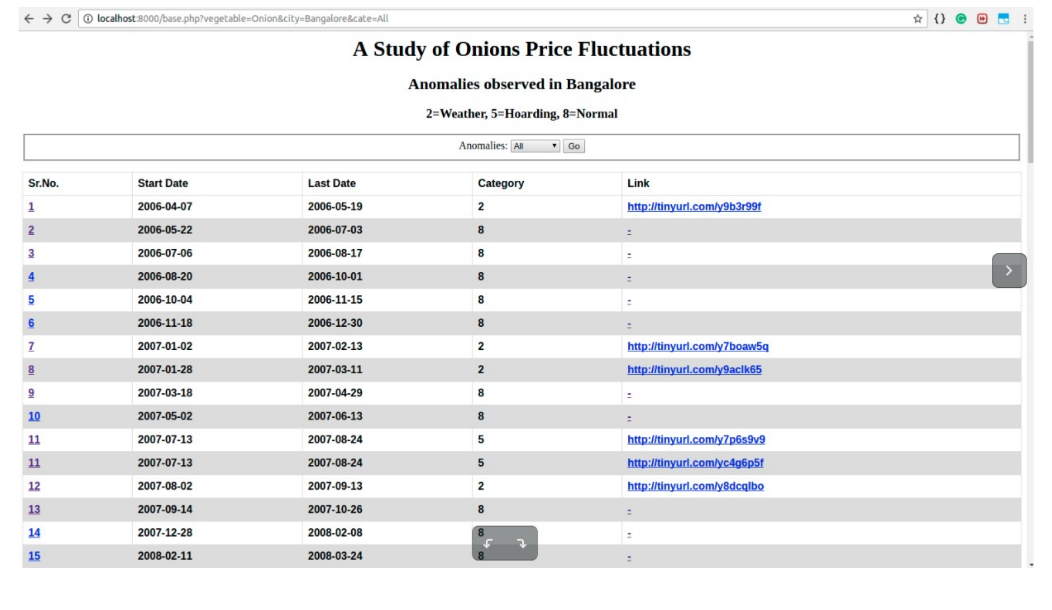
Commodity Information:

Commodity: Onion ▾

Center: Bangalore ▾

Figure 7.1: Home Page of the Web portal

2. The next page(fig 7.2) shows all the anomalies observed during the 12 year period. It shows the start and end date for each event and the label of each event. In the last column, it provides the link to the newspaper articles obtained during that particular anomaly.



localhost:8000/base.php?vegetable=Onion&city=Bangalore&cate=All

A Study of Onions Price Fluctuations

Anomalies observed in Bangalore

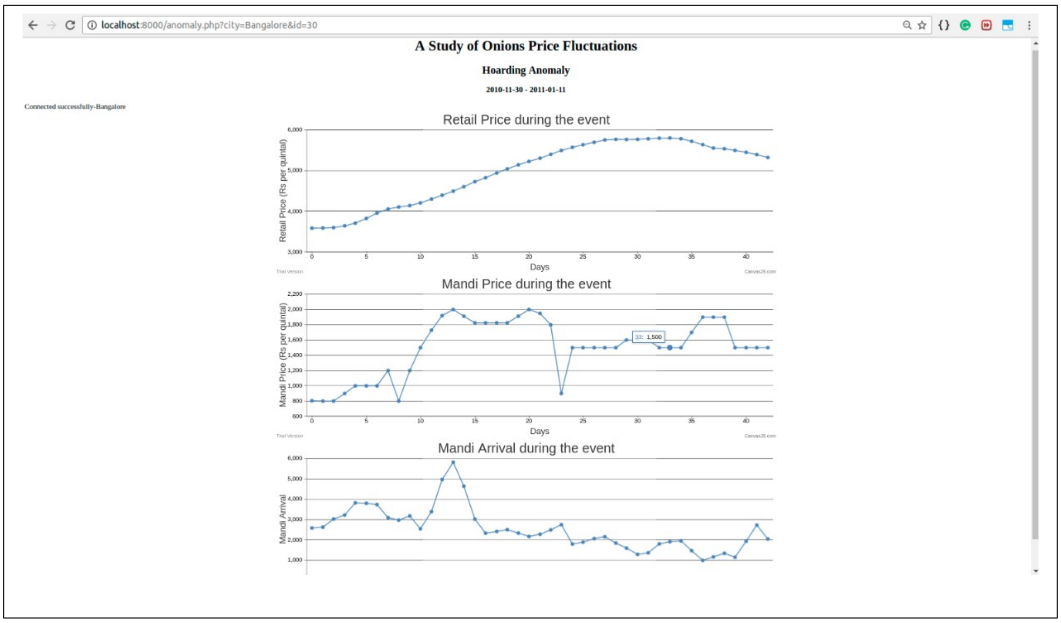
2=Weather, 5=Hoarding, 8=Normal

Anomalies: All Go

Sr.No.	Start Date	Last Date	Category	Link
1	2006-04-07	2006-05-19	2	http://tinyurl.com/y9b3r99f
2	2006-05-22	2006-07-03	8	-
3	2006-07-06	2006-08-17	8	-
4	2006-08-20	2006-10-01	8	-
5	2006-10-04	2006-11-15	8	-
6	2006-11-18	2006-12-30	8	-
7	2007-01-02	2007-02-13	2	http://tinyurl.com/y7boaw5q
8	2007-01-28	2007-03-11	2	http://tinyurl.com/y9aclk65
9	2007-03-18	2007-04-29	8	-
10	2007-05-02	2007-06-13	8	-
11	2007-07-13	2007-08-24	5	http://tinyurl.com/y7p6s9v9
11	2007-07-13	2007-08-24	5	http://tinyurl.com/yc4g6p5f
12	2007-08-02	2007-09-13	2	http://tinyurl.com/y8dcqlbo
13	2007-09-14	2007-10-26	8	-
14	2007-12-28	2008-02-08	8	-
15	2008-02-11	2008-03-24	8	-

Figure 7.2: List of Anomalies at a center

3. Clicking on hyperlink in the first column takes you to show detailed information for an individual anomaly. Here, in fig 7.3, you can see the mandi and retail price and the mandi arrival during the 43 day window of the event.



localhost:8000/anomaly.php?city=Bangalore&id=30

A Study of Onions Price Fluctuations

Hoarding Anomaly

2010-11-30 - 2011-01-11

Connected successfully-Bangalore

Figure 7.3: Detailed Information of Individual Anomaly

4. As of now, we have included only onions in our analysis. But work is being done on other commodities like potatoes, pulses and will be integrated in this portal.

Bibliography

[1] Wikipedia contributors. 2010 indian onion crisis-wikipedia, the free encyclopedia. 2017. Retrieved from `https://goo.gl/AiAhY6`.

[2] Agmarknet and national horticultural board: Obtaining the prices and arrivals of onions.

[3] Times of india archive search.

[4] Devesh Kapur and Mekhala Krishnamurthy. Understanding mandis: Market towns and the dynamics of india's rural and urban transformations, 2014. Center For The Advanced Study of India, University of Pennsylvania.

[5] PennState Eberly College Of Science. Applied time series analysis: Seasonal arima models. Retrieced from `https://onlinecourses.science.psu.edu/stat510/node/67/`.

[6] Tsay R.S. Analysis of financial time series. 2001.

[7] Purushottam Sharma, K C Gummagolmath, and R C Sharma. Prices of onions: An analysis. *Economic and Political Weekly*, 2011.

[8] Shoumitro Chatterjee and Devesh Kapur. Understanding price variation in agricultural commodities in india: Msp, government procurement, and agriculture markets. 2016. National Council of Applied Economic Research.

[9] Sunandan Chakraborty, Ashwin Venkataraman, Srikanth Jagabathula, and Lakshminarayanan Subramanian. Predicting socio-economic indicators using news events. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1455–1464. ACM, 2016.

[10] Ritesh Baldva. *Anomaly Detection & Classification in Commodity Prices*. 2017.

[11] Pradeep Rawat. *Time Series Analysis of Commodity Prices*. 2017.