Results and Analysis

Kapil Thakkar and Reshma Kumari

June 14, 2016

Chapter 1

Results and Analysis

We executed our library functions on the onion data. This data consists of Wholesale Price, Retail Price and Arrival since 1st January 2006 to 6th July 2015. In this chapter, we will show results produced by our system and will analyse these results along with each method.

1.1 Results

We have performed 4 types of analysis and result for each of this method is as follows. Note that these are primary results. Data for 2 centres are considered , Mumbai and Delhi.

Here are some results related with Mumbai Center. Table 1.1 stats the result of anomalies reported by our system, with details about anomalies reported by each method. So here First 5 columns corresponds to each method. Column 6 is union of results of first 3 methods and column 7 is union of results of method 4 and 5, as described in table. Column 8 is intersection of results of column 6 and column 7, which is final result of our system.

Table 1.2 stats the result of number of articles matched with the dates reported by our system as anomaly, for each method. So here First 5 columns corresponds to each method. Column 6 is union of results of first 3 methods and column 7 is union of result of method 4 and 5, as described in table. Column 8 is intersection of results of column 6 and column 7, which is final result of our system.

Note that total number of articles present for center Mumbai is **99**. Note one thing that articles are present from 2010 onwards. Apart from Graph Based Anomaly method, all methods are producing results from 2006 onwards as input data is from that time. The following pie chart shows the analysis of article showing what news articles states as the reason for the price hikes of onion.

Now, we present detailed analysis for each of the different type of time-series. First type of such analysis is in table 1.3. This table shows distribution of news articles present (that matched with system results) year-wise for each method when retail price time series is compared with average retail price time series. Second type of such analysis is in table 1.4. This table shows distribution of news articles present year-wise for each method when retail price time series is compared with arrival data of onion time series. Third type of such analysis is in table 1.5. This table shows distribution of news articles present year-wise for each method when retail price time series is compared with wholesale price time series. Fourth type of such analysis is in table 1.6. This table shows distribution of news articles present year-wise for each method when wholesale price time series is compared with arrival data of onion time series.

Distribution of anomalies present year-wise, for each method is also shown in table. Result for various analysis is described in tables 1.7, 1.8, 1.9 and 1.10.

Such results for different cities can also be calculated.

Methods	Slope Based (1)	Correlation (2)	Linear Regression (3)	Graph Based (4)	Multivariate (5)	1 U 2 U 3 (6)	4 U 5 (7)	6 ∩ 7
Retail Vs Average	742	255	353	150	177	1206	228	84
Retail Vs Arrival	420	795	353	150	167	1381	317	256
Retail Vs Wholesale	658	420	310	150	167	1243	279	124
Wholesale Vs Arrival	448	705	282	150	186	1315	336	244

Table 1.1: Anomalies Reported

Methods	Slope Based (1)	Correlation (2)	Linear Regression (3)	Graph Based (4)	Multivariate (5)	1 U 2 U 3 (6)	4 U 5 (7)	6 ∩ 7
Retail Vs Average	67	47	50	95	122	142	144	61
Retail Vs Arrival	42	74	167	0	119	220	119	119
Retail Vs Wholesale	30	55	40	39	119	107	124	44
Wholesale Vs Arrival	64	64	174	0	139	219	139	139

Table 1.2: Number of news articles matched with system

Distribution of All Articles	Articles Present	Slope Based (1)	Correlation (2)	Linear Regression (3)	Graph Based (4)	Multivariate (5)	1 U 2 U 3 (6)	4 U 5 (7)	6 ∩ 7
2010	6	0	0	8	2	0	8	2	0
2011	3	0	0	0	12	0	0	12	0
2012	2	1	0	0	0	0	1	0	0
2013	54	42	14	18	78	119	61	124	61
2014	23	24	0	24	3	3	39	6	0
2015	11	0	33	0	0	0	33	0	0
Total	99	67	47	50	95	122	142	144	61

Table 1.3: Retail Price VS Average Retail Price

Distribution of All Articles	Articles Present	Slope Based (1)	Correlation (2)	Linear Regression (3)	Graph Based (4)	Multivariate (5)	1 U 2 U 3 (6)	4 U 5 (7)	6 ∩ 7
2010	6	17	0	17	0	0	17	0	0
2011	3	1	0	12	0	0	12	0	0
2012	2	0	0	0	0	0	0	0	0
2013	54	10	39	126	0	119	137	119	119
2014	23	14	35	12	0	0	54	0	0
2015	11	0	0	0	0	0	0	0	0
Total	99	42	74	167	0	119	220	119	119

Table 1.4: Retail Price VS Arrival data of onion

Distribution of All Articles	Articles Present	Slope Based (1)	Correlation (2)	Linear Regression (3)	Graph Based (4)	Multivariate (5)	1 U 2 U 3 (6)	4 U 5 (7)	6 ∩ 7
2010	6	6		2	0	0	6	0	0
2011	3	1	0	0	0	0	1	0	0
2012	2	2	7	0	0	0	7	0	0
2013	54	7	21	18	39	119	45	124	44
2014	23	14	27	20	0	0	48	0	0
2015	11	0	0	0	0	0	0	0	0
Total	99	30	55	40	39	119	107	124	44

Table 1.5: Retail Price VS Wholesale Price

Distribution of All Articles	Articles Present	Slope Based (1)	Correlation (2)	Linear Regression (3)	Graph Based (4)	Multivariate (5)	1 U 2 U 3 (6)	4 U 5 (7)	6 ∩ 7
2010	6	17	0	17	0	0	17	0	0
2011	3	1	0	12	0	0	12	0	0
2012	2	0	0	0	0	0	0	0	0
2013	54	25	15	125	0	123	129	123	123
2014	23	21	49	20	0	16	61	16	16
2015	11	0	0	0	0	0	0	0	0
Total	99	64	64	174	0	139	219	139	139

Table 1.6: Wholesale Price VS Arrival data of onion

Distribution of All Articles	Slope Based (1)	Correlation (2)	Linear Regression (3)	Graph Based (4)	Multivariate (5)	1 U 2 U 3 (6)	4 U 5 (7)	6 ∩ 7
2006	133	30	65	0	0	204	0	0
2007	14	15	30	0	0	53	0	0
2008	42	15	0	0	0	47	0	0
2009	82	15	0	1	0	96	1	0
2010	72	45	28	6	0	142	6	0
2011	56	60	115	27	0	182	27	0
2012	84	0	33	0	0	100	0	0
2013	77	15	56	108	161	125	170	80
2014	140	0	26	8	16	155	24	4
2015	42	60	0	0	0	102	0	0
Total	742	255	353	150	177	1206	228	84

Table 1.7: Distribution of Anomalies reported by system for Retail Price VS Average Retail Price

Distribution of All Articles	Slope Based (1)	Correlation (2)	Linear Regression (3)	Graph Based (4)	Multivariate (5)	1 U 2 U 3 (6)	4 U 5 (7)	6 ∩ 7
2006	35	185	0	78	0	199	78	48
2007	63	85	0	72	0	148	72	42
2008	28	105	0	0	0	126	0	0
2009	28	45	0	0	0	73	0	0
2010	48	60	46	0	0	107	0	0
2011	36	60	40	0	0	128	0	0
2012	77	45	0	0	0	118	0	0
2013	59	90	168	0	161	263	161	161
2014	46	105	99	0	6	204	6	5
2015	0	15	0	0	0	15	0	0
Total	420	795	353	150	167	1381	317	256

Table 1.8: Distribution of Anomalies reported by system for Retail Price VS Arrival data of onion

Distribution of All Articles	Slope Based (1)	Correlation (2)	Linear Regression (3)	Graph Based (4)	Multivariate (5)	1 U 2 U 3 (6)	4 U 5 (7)	6 ∩ 7
2006	126	0	0	63	0	126	63	28
2007	77	30	0	0	0	107	0	0
2008	70	15	0	24	0	85	24	0
2009	63	75	0	0	0	126	0	0
2010	62	30	2	0	0	92	0	0
2011	50	45	67	0	0	155	0	0
2012	72	82	64	20	0	160	20	9
2013	50	38	44	43	161	124	166	81
2014	60	75	99	0	6	182	6	6
2015	28	30	34	0	0	86	0	0
Total	658	420	310	150	167	1243	279	124

Table 1.9: Distribution of Anomalies reported by system for Retail Price VS Wholesale Price

Distribution of All Articles	Slope Based (1)	Correlation (2)	Linear Regression (3)	Graph Based (4)	Multivariate (5)	1 U 2 U 3 (6)	4 U 5 (7)	6 ∩ 7
2006	0	65	0	71	0	65	71	27
2007	42	100	0	79	0	142	79	31
2008	21	120	0	0	0	141	0	0
2009	56	60	0	0	0	109	0	0
2010	83	75	51	0	0	162	0	0
2011	64	60	34	0	0	157	0	0
2012	42	45	0	0	0	87	0	0
2013	80	45	157	0	152	246	152	152
2014	46	105	40	0	34	162	34	34
2015	14	30	0	0	0	44	0	0
Total	448	705	282	150	186	1315	336	244

Table 1.10: Distribution of Anomalies reported by system for Wholesale Price VS Arrival data of onion

1.2 Analysis of Each Method

In this section, we try to analyse each method, what is limitation of each method and where it is performing good. So, we will describe each method one by one and study them. Note that we have articles from 2010 onwards, so we will be focussing on anomalies reported after 2010 and comparing with them news articles which we have.

Note that all the graphs and results described in following sections are for centre Mumbai. As Mumbai is in Maharashtra state, which is largest producer of onion in India. Also in graphs, Yellow highlighted region corresponds to anomalies reported by system, red region corresponds to dates for which our system reported anomaly and news article was present for that and blue region corresponds to date for which news article was present but that date was not reported as anomaly by our system.

1.2.1 Slope Based Anomaly Detection

The main functionality of this method is to find change in one variable with respect to other. Given two time-series, here we try to find, between two points in time series, how much dependent variable changed corresponding to independent variable. If this change is huge, than it is reported as anomaly.

We have four types of analysis which are as follows:

1. **Retail Price vs Average of Retail Price**: Here, we first take average of retail price at all centres and than compare change in retail price with respect to change in average of retail price for different time window.
2. **Retail Price vs Arrival of Onion**: Here, we try to find change in retail price with respect to change in arrival of onion for different windows.
3. **Retail Price vs Wholesale Price**: Here, retail price is dependent on wholesale price and we try to find change in retail price with respect to change wholesale price for different windows in this method.
4. **Wholesale Price vs Arrival of Onion**: Here, we try to find change in wholesale price with respect to change in arrival of onion for different window size.

So, in each of the case, we try to find change with respect to another, and if this change is huge, crossing threshold than it is reported as anomaly. Now, not that in analysis 1 and 3 stated above, both the time series are directly proportional to each other and in the analysis 2 and 4 both the time series are inversely proportional to each other. So, limitations faced by this method for analysis 1 and 3 will be similar and for analysis 2 and 4 will be similar. While describing this method, each analysis will be referenced by its corresponding number.

First we will start with analysis 1 and 3. Here, we have few observations as follows:

- Dates are reported as anomalies, if retail price at centre is increasing more as compared to average retail price for analysis 1 or if retail price at centre is increasing more as compared to wholesale price for analysis 3.
  Such cases are reported for the following tenure by this method:
  - *Analysis 1*: June 2010, August 2010, May 2011, June 2011, May Jun Nov 2012, Apr May 2013 (Prices went too high as compared to average) (See Figure 1.1)
  - *Analysis 3*: Apr July Oct Dec 2010, Jan 2011, May June 2014 (See Figure 1.2)

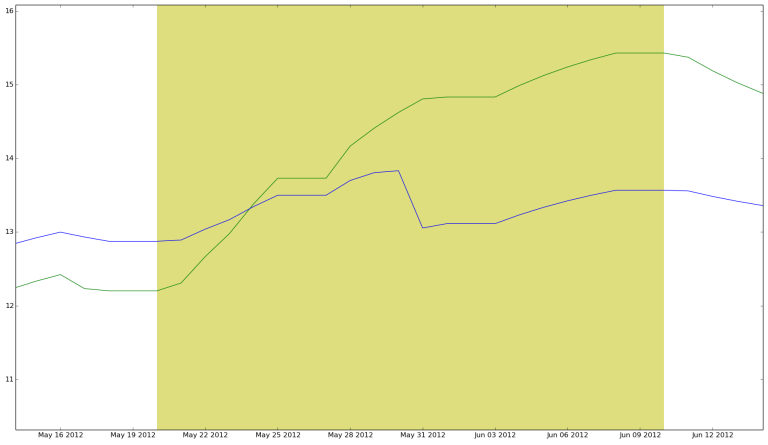

Figure 1.1: Slope Based Anomaly Detection (Green line - Centre Retail Price, Blue Line - Average Retail Price)

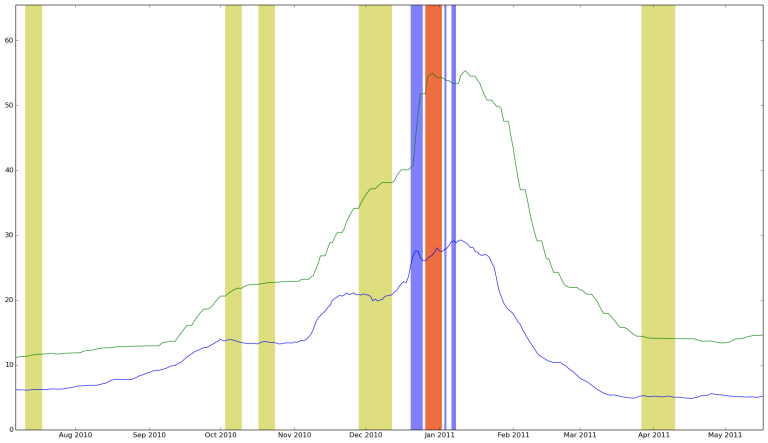

Figure 1.2: Slope Based Anomaly Detection (Green line - Retail Price, Blue Line - Wholesale Price)

So, even though we do not have articles for these anomalies, but method is behaving as it should be.

- This method also has limitations. There exist few cases where, drop in retail price for one center is quite huge as compared to drop in average retail price (in case of *analysis 1*) or wholesale price (in case of *analysis 3*). This is good thing for centres, and should not be treated as anomaly. But in this case, slope value goes high and that's why our method reports that tenure as anomaly as well.
  Such cases are reported for the following tenure by this method:
  - *Analysis 1*: January 2010, June Aug 2012, Jan 2013, Feb Oct 2014, Feb 2015 (See Figure 1.3)
  - *Analysis 3*: Feb May 2010 (See Figure 1.4)

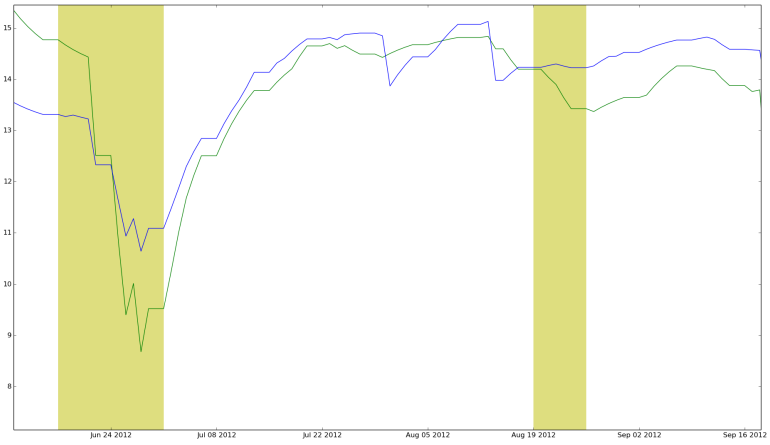

Figure 1.3: Slope Based Anomaly Detection (Green line - Centre Retail Price, Blue Line - Average Retail Price)

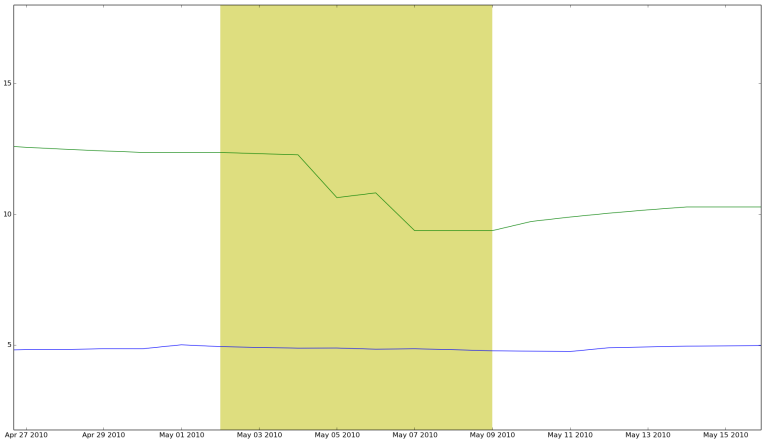

Figure 1.4: Slope Based Anomaly Detection (Green line - Retail Price, Blue Line - Wholesale Price)

- Other observation is related to why few anomalies were reported in news but not by our system. Reason for that is we are comparing relative change in two time series. Now for some dates, where news article is present but our system did not report, value of both time series increased together. Although, prices went too high, but still relative change, i.e. slope value remained relatively low as compared to others, and so that was not reported by our system.

  Such cases are reported for the following tenure by this method:

  - *Analysis 1*: Dec 2010, Jan Feb 2012, June 2013, May June 2015 (See Figure 1.5)
  - *Analysis 3*: Feb 2013, July 2014 (See Figure 1.6)

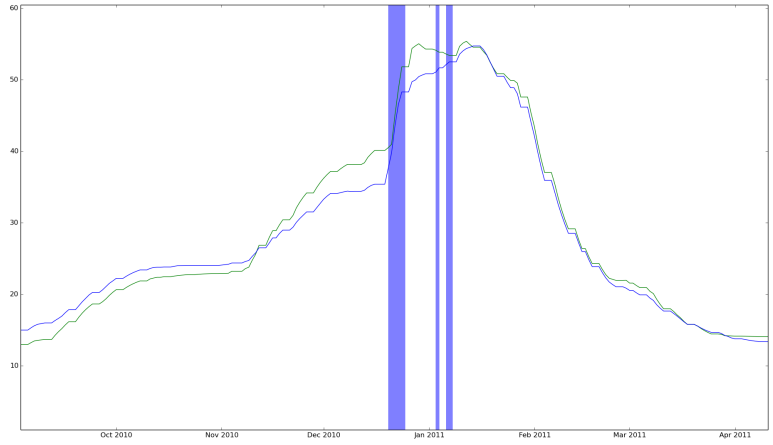

Figure 1.5: Slope Based Anomaly Detection (Green line - Centre Retail Price, Blue Line - Average Retail Price)

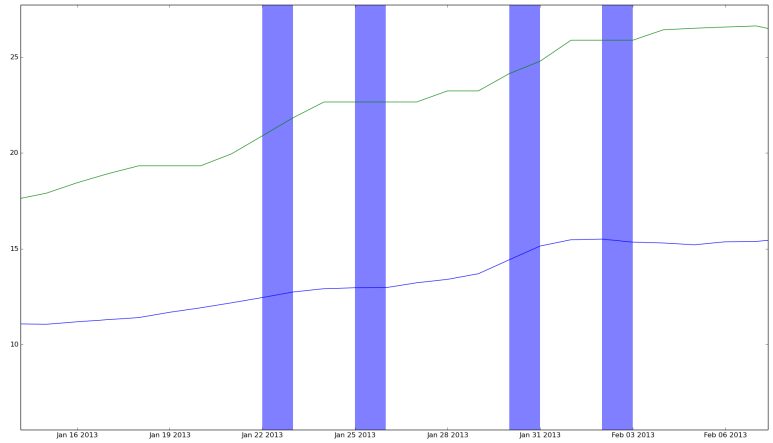

Figure 1.6: Slope Based Anomaly Detection (Green line - Retail Price, Blue Line - Wholesale Price)

- In some cases, original retail price was running less than average retail price for some time and then suddenly prices in the centre increased drastically. So such cases were reported as anomaly in this method, which is quite normal. Such cases were found in *Analysis 1* for Nov 2011, Feb Mar 2012 and Dec 2014. (See Figure 1.7)

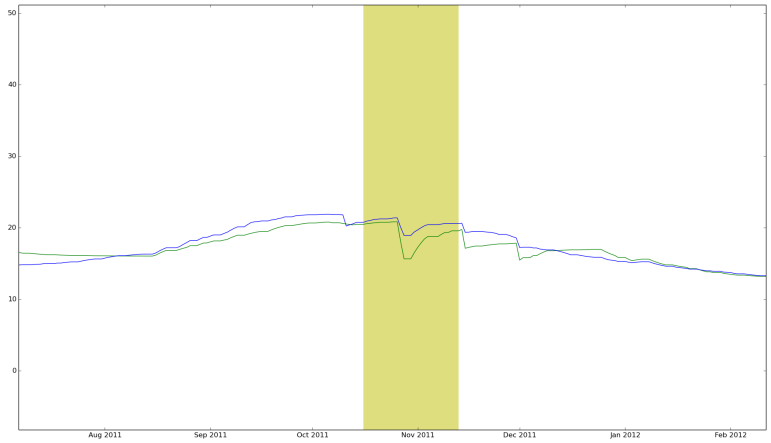

Figure 1.7: Slope Based Anomaly Detection (Green line - Centre Retail Price, Blue Line - Average Retail Price)

- In some cases, tenure reported as anomaly is quite large, because situations were abnormal for long time. But is not necessary that news articles should be present for such a large tenure. And anomaly reported was justifiable. For *Analysis 1*, such tenure was reported for March end to May start 2014. (See Figure 1.8)

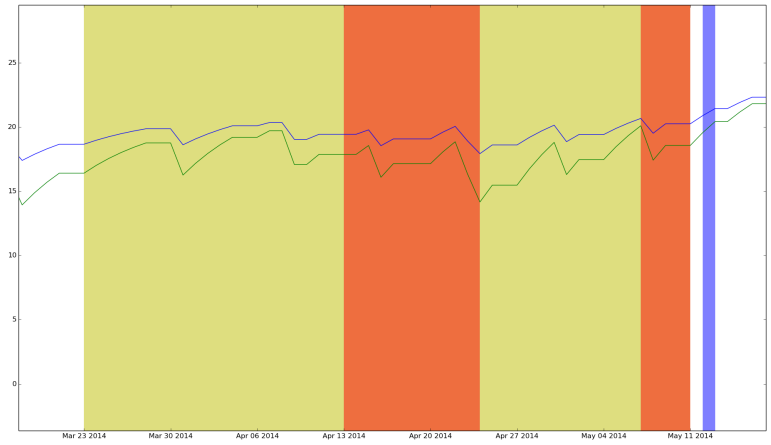

Figure 1.8: Slope Based Anomaly Detection (Green line - Centre Retail Price, Blue Line - Average Retail Price)

- In *Analysis 1*, For June 2014, method has reported tenure upto mid June when prices started increasing, but it remained high and due to that some news articles are present for June 21 around, we have missed, because at that time relative slope value became normal, but since prices were high, it was covered by news articles. (See Figure 1.9)

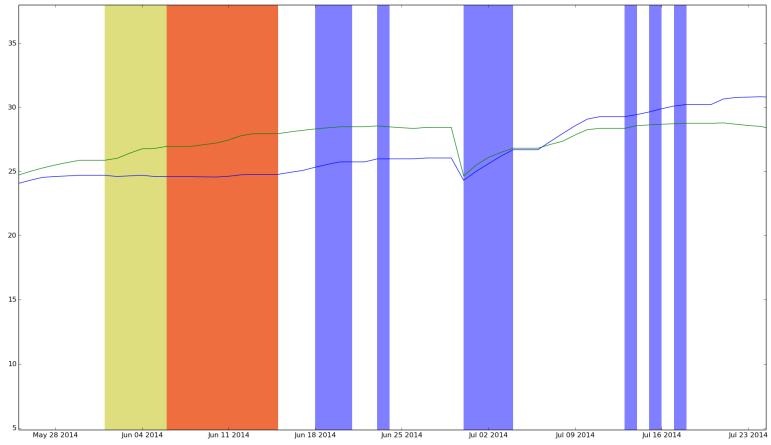

Figure 1.9: Slope Based Anomaly Detection (Green line - Centre Retail Price, Blue Line - Average Retail Price)

- There exist some cases in *Analysis 3* where retail price were decreasing but wholesale price kept on increasing, this created negative slope value, where as in this scenario, we were looking for only positive slopes and that's why this method missed it. Such periods were in July Aug Sept 2013, Nov Dec 2013. (See Figure 1.10)

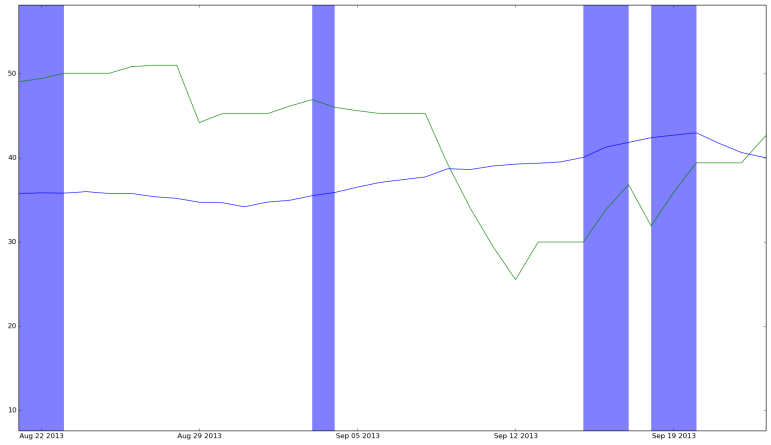

Figure 1.10: Slope Based Anomaly Detection (Green line - Retail Price, Blue Line - Wholesale Price)

Now we present few observation for Analysis 2 and 4.

- If change in retail price or change in wholesale price is more as compared to change in arrival (prices went too high, even for small drop in arrival), then it is reported as anomaly by this method.

  Such cases are reported for the following tenure by this method:

  - *Analysis 2*: Oct Dec 2010, May Jun 2013 (See Figure 1.11)
  - *Analysis 4*: Sept Oct Nov 2010, Jun 2013, Jun 2015 (See Figure 1.12)

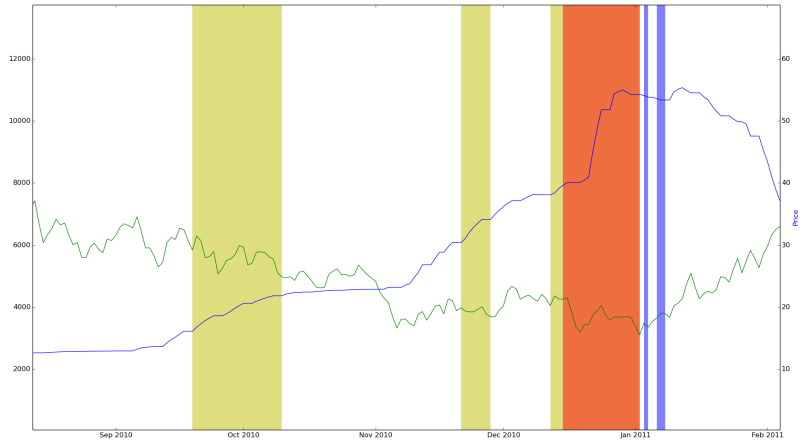

Figure 1.11: Slope Based Anomaly Detection (Green line - Arrival Data of Onion, Blue Line - Retail Price)

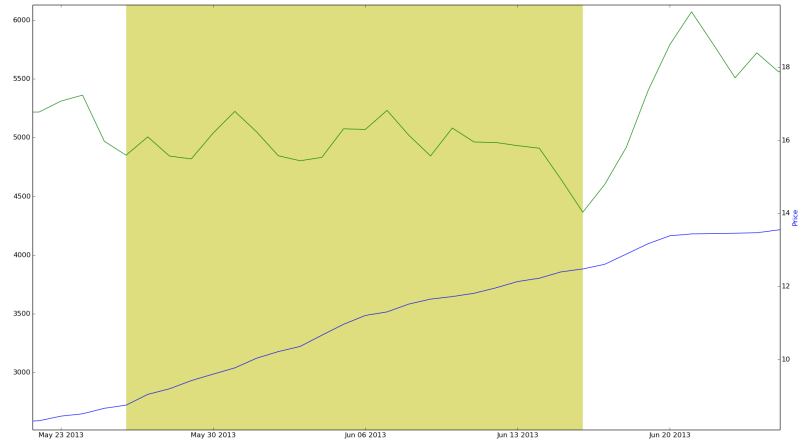

Figure 1.12: Slope Based Anomaly Detection (Green line - Arrival Data of Onion, Blue Line - Wholesale Price)

- But in above described scenario, when change in price is high, but prices are decreasing and when arrival is increasing, and if drop in price is too high, then also it will be reported as anomaly. So this is limitation of this method and reports false positives in this case.
  Such cases are reported for the following tenure by this method:
  - *Analysis 2*: Feb Mar 2011, Jan Feb 2014 (See Figure 1.13)
  - *Analysis 4*: Oct Dec 2011, Jan 2014, Mar 2011 (See Figure 1.14)

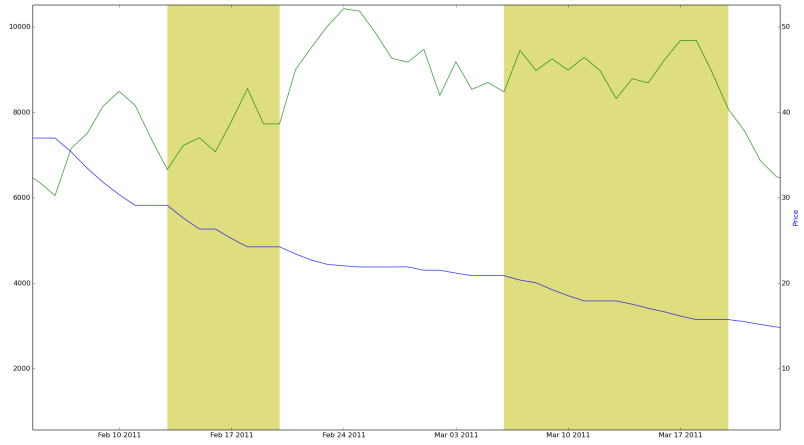

Figure 1.13: Slope Based Anomaly Detection (Green line - Arrival Data of Onion, Blue Line - Retail Price)

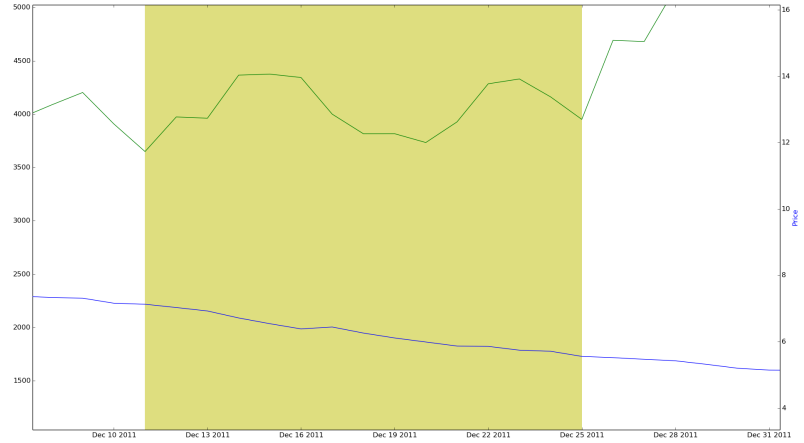

Figure 1.14: Slope Based Anomaly Detection (Green line - Arrival Data of Onion, Blue Line - Wholesale Price)

- One more limitation of this method is when arrival is increasing but along with that retail or wholesale price is also increasing. Since will come out as positive slope and this method, in this scenario is only looking for negative slope and that's why, this will not be reported as anomaly and news articles corresponding to this tenure will not be matched by results of this method.

  Such cases are reported for the following tenure by this method:

  - *Analysis 2*: Jan Feb July Nov 2013, June July 2014 (See Figure 1.15)
  - *Analysis 4*: Jan Feb 2013, July 2014, June 2015 (See Figure 1.16)

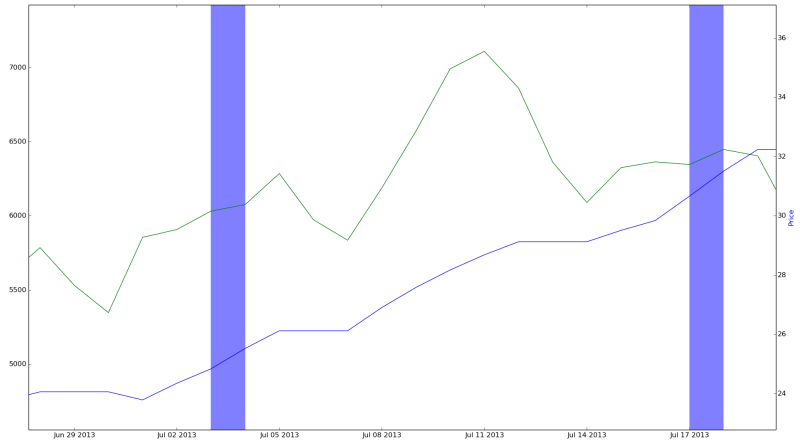

Figure 1.15: Slope Based Anomaly Detection (Green line - Arrival Data of Onion, Blue Line - Retail Price)

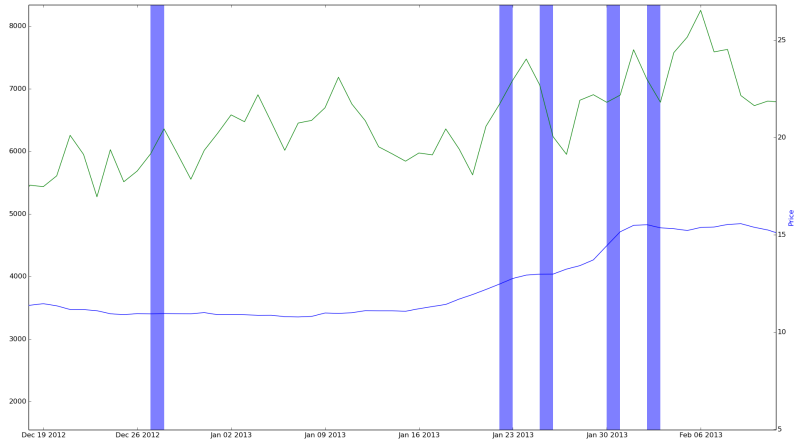

Figure 1.16: Slope Based Anomaly Detection (Green line - Arrival Data of Onion, Blue Line - Wholesale Price)

- There exist some cases, where due to low arrival, prices went too high and when slowly arrival started entering into market, prices were going down slowly. This period of slow decrement of prices is not reported by this method, but since prices were still high, this system could not report dates for news articles corresponding to this tenure. Such cases occurred in *Analysis 2* for Dec 2010 and Jan 2011 (See Figure 1.17)and in *Analysis 4* for Nov 2013 (See Figure 1.18). Also, note that these articles were mainly on Pakistan banned exports and article on inflation stating that inflation rate is high and onion prices are playing an important role in this.

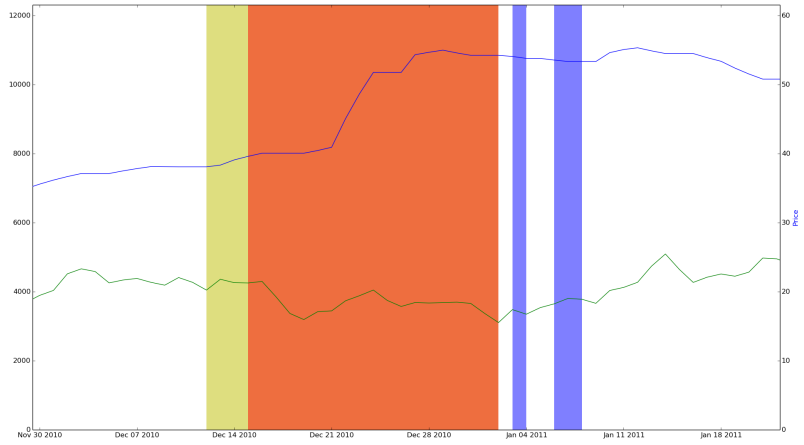

Figure 1.17: Slope Based Anomaly Detection (Green line - Arrival Data of Onion, Blue Line - Retail Price)

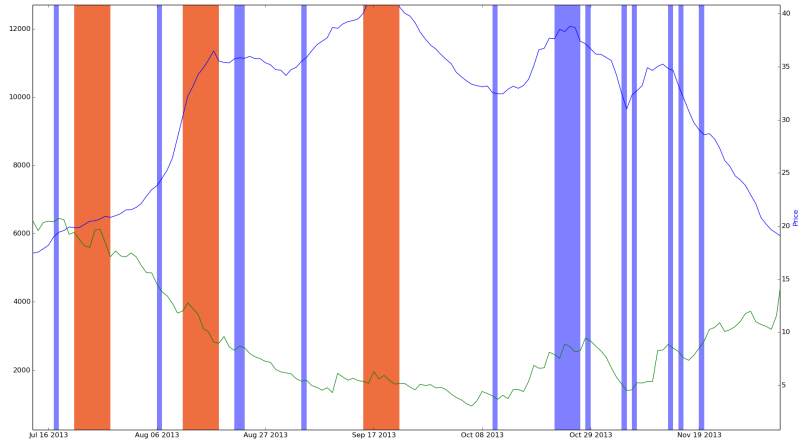

Figure 1.18: Slope Based Anomaly Detection (Green line - Arrival Data of Onion, Blue Line - Wholesale Price)

- In some of the cases, where arrival fell too much drastically, and due to that retail price went high drastically as well. And since retail prices went high too much it got reported in news articles, but this was expected, as arrival was less. But here both the changes were high, so ultimately slope value was not so high and was not reported by system. Such cases in *Analysis 2* exist for Aug Sept 2013 (See Figure 1.19).

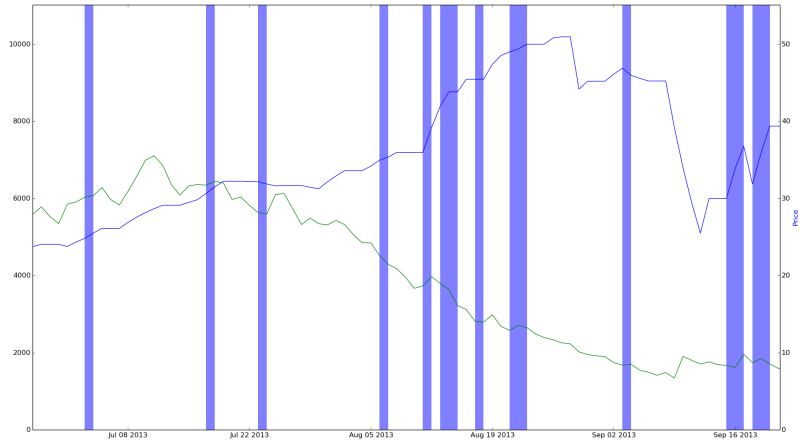

Figure 1.19: Slope Based Anomaly Detection (Green line - Arrival Data of Onion, Blue Line - Retail Price)

- Another limitation of this method is when retail price remained constant and there was change in arrival. As retail price was constant, slope value became zero and method did not report them and due to that few news articles could not be matched by dates reported by this method for example in *Analysis 2*, this thing occurred for June July 2015 (See Figure 1.20).

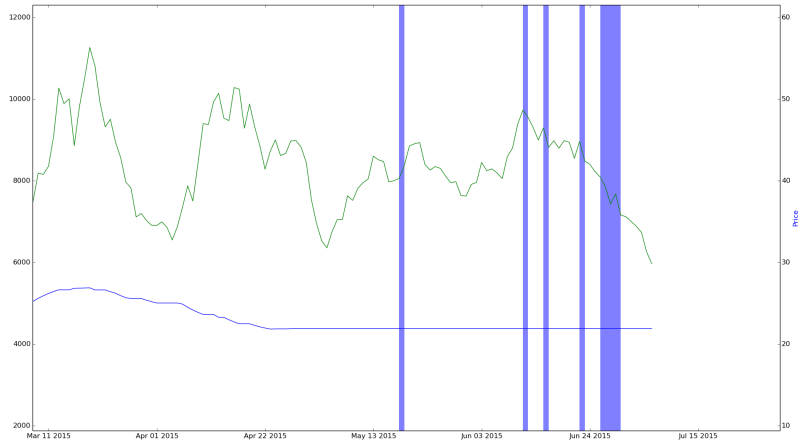

Figure 1.20: Slope Based Anomaly Detection (Green line - Arrival Data of Onion, Blue Line - Retail Price)

Also, note one thing that, this method reports anomaly as whole window of few days (here 7 days). So, because of that too, method tends to report more anomaly dates.

1.2.2 Linear Regression

The main functionality of this method is to find what should be ideal value of the dependent variable given value of independent variable. This method first finds out linear relationship between 2 variables, whose time series is given as input and one of them is dependent on another. After finding out this equation, we see for a given value of independent variable what should be ideal value of dependent variable and note down the relative difference. If this difference is large, then it is reported as anomaly.

We have four types of analysis which are as follows:

1. **Retail Price vs Average of Retail Price**: Here, we first take average of retail price at all centres as independent variable and retail price as dependent variable.
2. **Retail Price vs Arrival of Onion**: Here, we take retail price as dependent variable and arrival of onion as independent variable.
3. **Retail Price vs Wholesale Price**: Here, we take retail price as dependent variable and Wholesale Price as independent variable.
4. **Wholesale Price vs Arrival of Onion**: Here, we take Wholesale price as dependent variable and arrival of onion as independent variable.

So, in each of the case, we try to find relative difference between ideal value and its real value, and if it is huge, crossing threshold than it is reported as anomaly. Now, note that in analysis 1 and 3 stated above, both the time series are directly proportional to each other and in the analysis 2 and 4 both the time series are inversely proportional to each other. So, limitations faced by this method for analysis 1 and 3 will be similar and for analysis 2 and 4 will be similar. While describing this method, each analysis will be referenced by its corresponding number.

First we will start with analysis 1 and 3. Here, we have few observations as follows:

- This method will report any tenure as anomaly when there is large gap i.e. more than expected between retail price of a center and average retail price (for *Analysis 1*) or wholesale price (for *Analysis 3*).

  Such cases are reported for the following tenure by this method:

  – *Analysis 1*: Dec 2010, Near to Jan 2011, May June July 2011, Jan May June 2012, June 2013 (See Figure 1.21)

  – *Analysis 3*: Feb Mar 2011, Jan 2012, June 2012, Jan Feb 2014, Apr 2015 (See Figure 1.22)

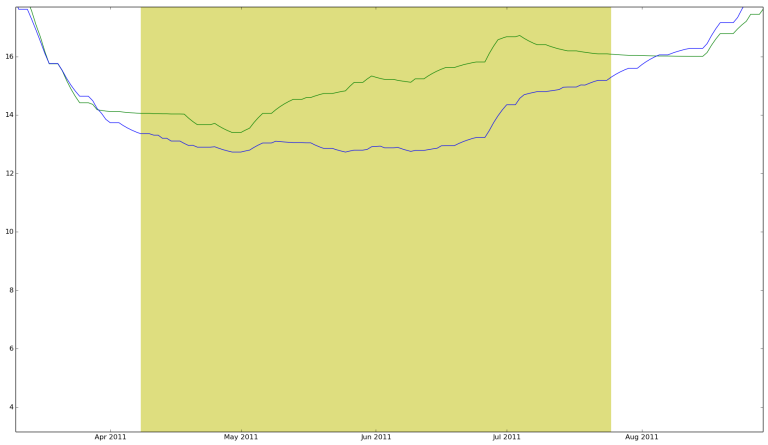

Figure 1.21: Linear Regression (Green line - Centre Retail Price, Blue Line - Average Retail Price)

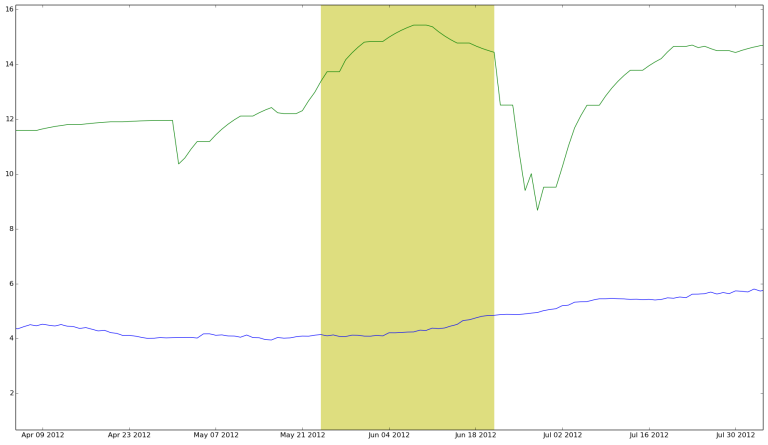

Figure 1.22: Linear Regression (Green line - Retail Price, Blue Line - Wholesale Price)

- One limitation of this method is that, if both the series have high values for some time period and difference between them is not so huge then that will not be reported as anomaly.

  Such cases are reported for the following tenure by this method:

  – *Analysis 1*: Jan 2011, Jan Feb Aug Sept Oct Nov 2013, July 2014, June July 2015 (See Figure 1.23)
  – *Analysis 3*: Feb 2013, Aug Sept Oct Nov 2013, June July 2015 (See Figure 1.24)

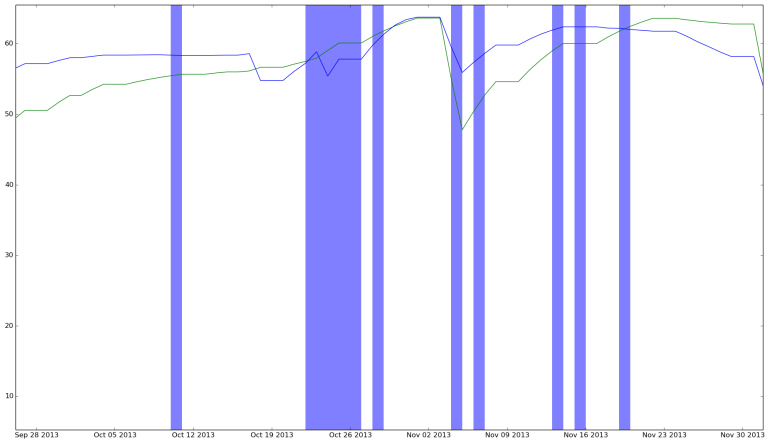

Figure 1.23: Linear Regression (Green line - Centre Retail Price, Blue Line - Average Retail Price)

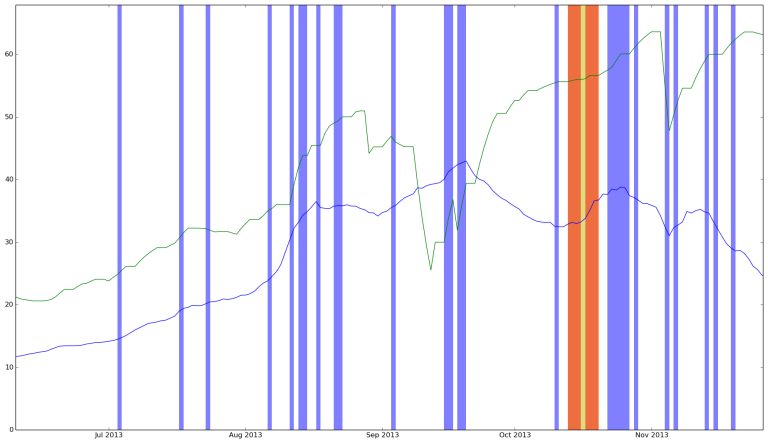

Figure 1.24: Linear Regression (Green line - Retail Price, Blue Line - Wholesale Price)

- Note that in the tenure of Oct Nov 2013 (for *Analysis 1*) prices are usually high and as the prices are high, tolerance level also increases little bit. So even, if for some difference it is reported as anomaly at lower price, it is not necessary that for the same difference, it will be reported as anomaly at higher prices. (See Figure 1.23)

Now we present few observation for Analysis 2 and 4.

- Here, in this method, it tries to predict what should be retail price or wholesale price based on the arrival of the product. So if the price is too high for the given arrival than it will be reported as anomaly. Such cases are reported for the following tenure by this method:
  - *Analysis 2*: Dec 2010, Jan Feb 2011, Aug Sept Oct Nov 2013, Oct Dec 2014 (See Figure 1.25)
  - *Analysis 4*: Dec 2010, July Aug Sept Oct Nov Dec 2013, July 2014 (See Figure 1.26)

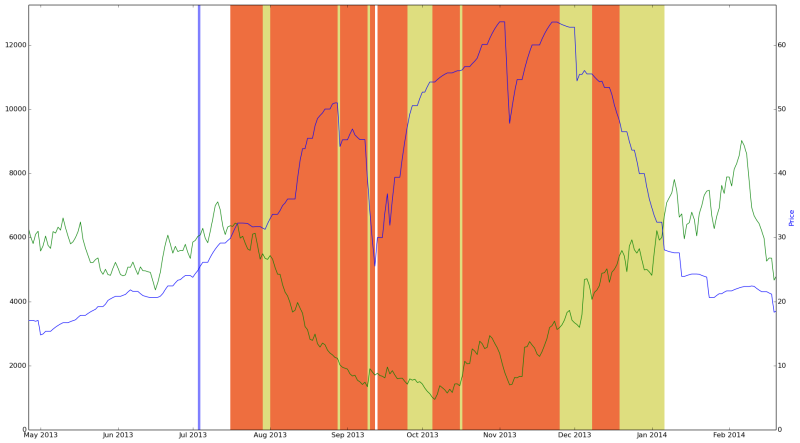

Figure 1.25: Linear Regression (Green line - Arrival Data of Onion, Blue Line - Retail Price)

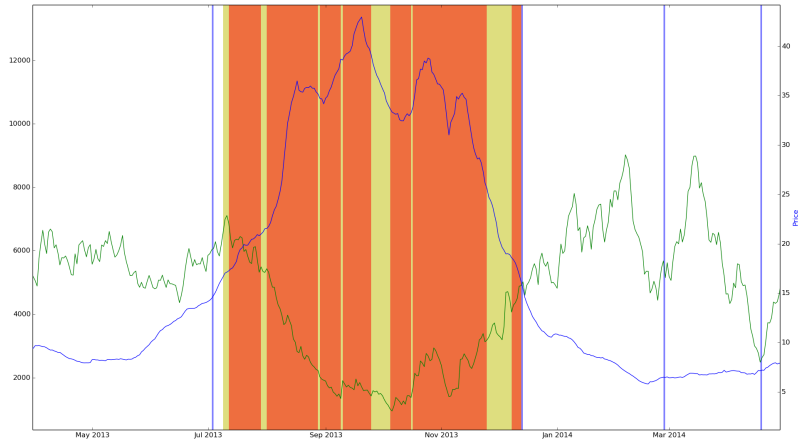

Figure 1.26: Linear Regression (Green line - Arrival Data of Onion, Blue Line - Wholesale Price)

- Now, this method has also missed few of the articles for this analysis as well. Now, looking at the graphs we could not interpret what may be exact reason why they were missed. But method may have found prices to be moderate and that's they might have been missed.

  Such cases are reported for the following tenure by this method:

  - *Analysis 2*: Jan Feb 2013, July 2014, June July 2015 (See Figure 1.27)
  - *Analysis 4*: Jan Feb 2013, June July 2013, June 2015 (See Figure 1.28)

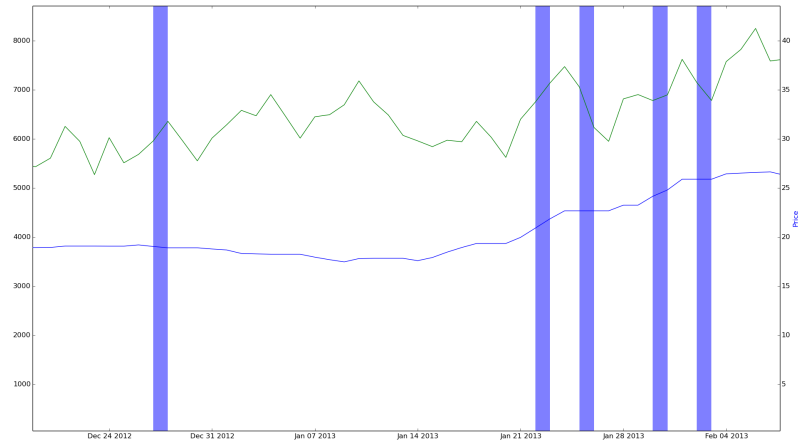

Figure 1.27: Linear Regression (Green line - Arrival Data of Onion, Blue Line - Retail Price)

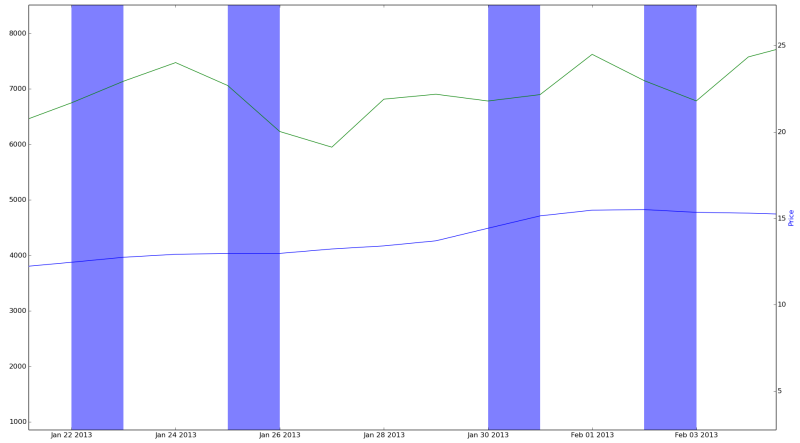

Figure 1.28: Linear Regression (Green line - Arrival Data of Onion, Blue Line - Wholesale Price)

1.2.3 Window Based Correlation

This method checks if the provided two timeseries moves in tandem or not. If not, then what are the time periods when they are not following the desired behavior. In order to find such time periods in the timeseries, the whole timeseries is divided into smaller windows and correlation value computed for that window is used to determine if that period is anomalous or not. The P-value of 0.01 is used in order to check the significance of correlation value.

The function is tested with different set of timeseries. Following are the four tests which were performed:

1. **Retail Price vs Average of Retail Price**: This test tries to find if any centre deviates from other centres abruptly. Average of timeseries of all the centres is taken as a representative timeseries for the behavior of all the centres which is compared with every centre in order to find time periods where these two did not move in tandem

2. **Retail Price vs Arrival of Onion**: Retail timeseries is expected not to move in tandem with Arrival timeseries. So, the time periods with positive correlation values are spotted in this.

3. **Retail Price vs Wholesale Price**: Retail timeseries is expected to move in tandem with wholesale timeseries. So, the time periods with negative correlation values are spotted in this.

4. **Wholesale Price vs Arrival of Onion**: Wholesale timeseries is expected to not move in tandem with Arrival timeseries. So, the time periods with positive correlation values are spotted in this.

The two timeseries are first aligned with each other at the maximum lag value. Then the entire timeseries is divided into small time periods of 15 days in order to locate anomalous situations through their respective correlation values.

Observations when retail price timeseries is compared with average of retail price timeseries:

- The timeseries showed the initial shift/lag of zero days which means both the timeseries are best aligned without any lag.

  Some of the tenures for which the method reported anomalies are:

  - 2008-03-06 to 2008-03-20 (See Figure 1.29)
  - 2009-09-12 to 2009-09-26 (See Figure 1.30)
  - 2011-07-04 to 2011-08-02 (See Figure 1.31)

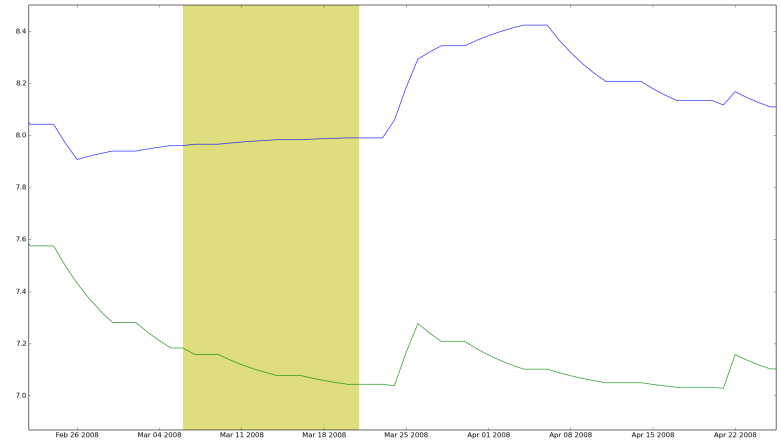

Figure 1.29: Window based correlation (Green line - Centre Retail Price, Blue Line - Average Retail Price)

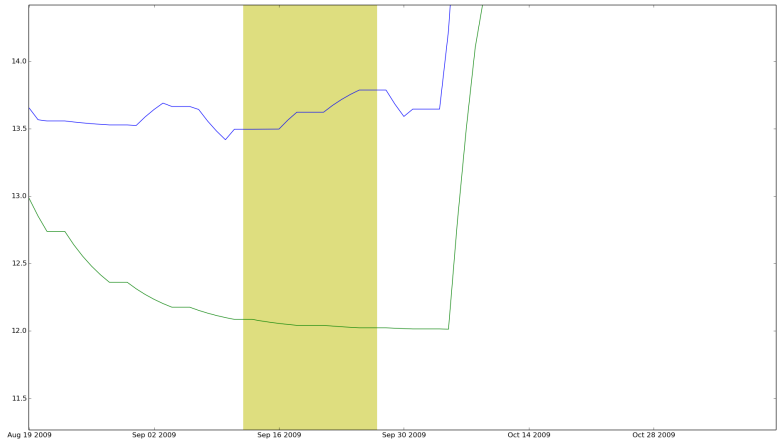

Figure 1.30: Window based correlation (Green line - Centre Retail Price, Blue Line - Average Retail Price)

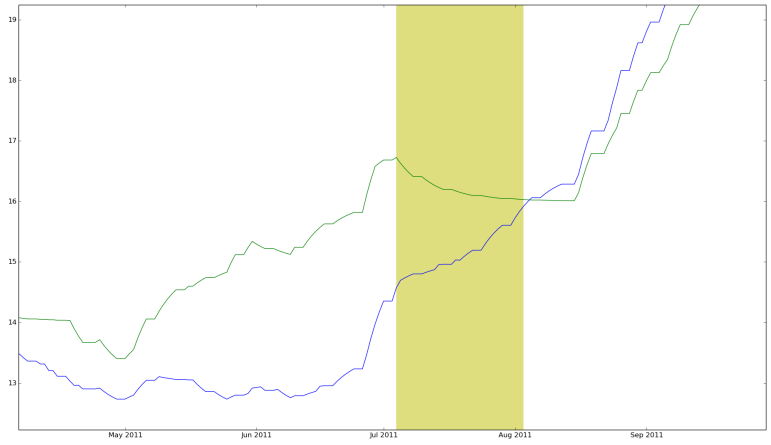

Figure 1.31: Window based correlation (Green line - Centre Retail Price, Blue Line - Average Retail Price)

But looking closely in the data, it is found that these tenures are reported because the two series were not in tandem which was mostly because prices of delhi faced some fluctuations whereas Mumbai prices did not show evident fluctuations. One of the reason for fluctuation in prices of Delhi over Mumbai could be because Delhi is dependent on other states for Onion whereas Mumbai does not have such issue.

Some of the tenures which went unnoticed by method:

- 2010-12-20 to 2010-12-25 (See Figure 1.32)
- 2011-01-06 to 2011-01-08 (See Figure 1.33)
- 2012-12-27 to 2013-01-05 (See Figure 1.34)

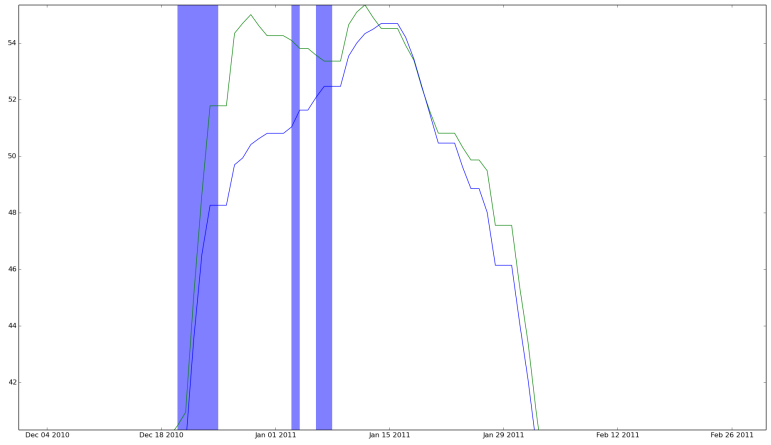

Figure 1.32: Window based correlation (Green line - Centre Retail Price, Blue Line - Average Retail Price)

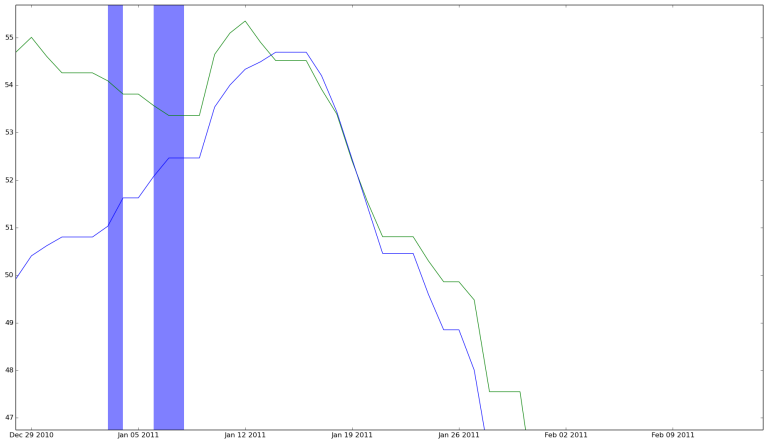

Figure 1.33: Window based correlation (Green line - Centre Retail Price, Blue Line - Average Retail Price)

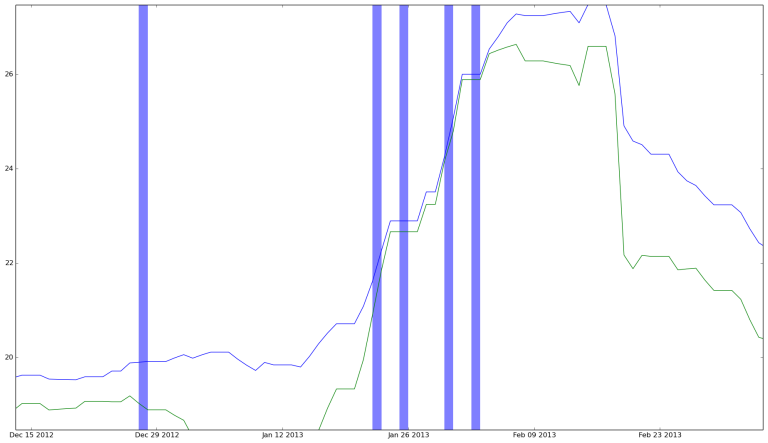

Figure 1.34: Window based correlation (Green line - Centre Retail Price, Blue Line - Average Retail Price)

Some of the tenures which were reported by news reports as well as method:

- 2013-10-06 to 2013-10-15 (See Figure 1.35)
- 2013-10-17 to 2013-10-21 (See Figure 1.35)

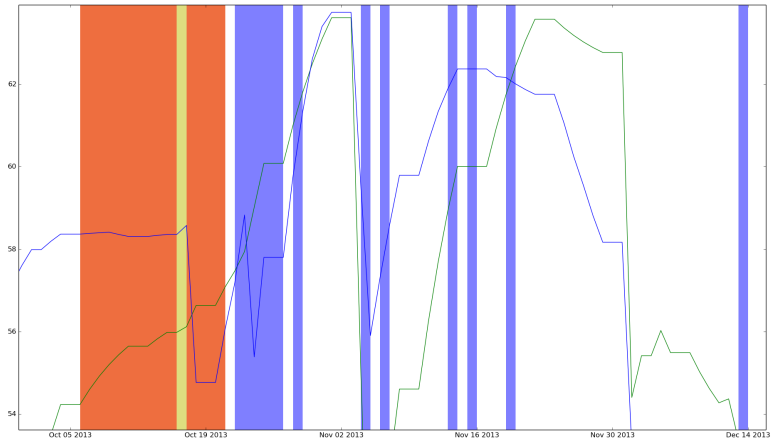

Figure 1.35: Window based correlation (Green line - Centre Retail Price, Blue Line - Average Retail Price)

Other interesting observations that were found while conducting above mentioned tests are as following:

- The lag chosen while comparing retail and arrival for Mumbai came as 15 days which means it takes 15 days of time for arrival to impact retail.
- The lag chosen by method while comparing retail and arrival for Delhi came out to be -9 days which gives hints that retail prices are impacting arrival of Onion which is not ideal in real time scenario.
- The lag chosen while comparing retail and wholesale for Mumbai comes out as -15 which means retail prices are impacted in approx. 15 days after change in the wholesale prices.
- The lag chosen while comparing retail and wholesale for Delhi comes out as -4 which is clearly very small than Mumbai. It might be because of smaller supply chain in Delhi compared to Mumbai.
- The lag chosen while comparing wholesale and arrival for Mumbai came as 11 days which means it takes 11 days of time for arrival to impact wholesale.
- The lag chosen by method while comparing wholesale and arrival for Delhi came out to be -14 days which gives hints that wholesale prices are impacting arrival of Onion which is not ideal in real time scenario.

Limitations of method:

- The method reports all the tenures where series are not in sync or are in sync(as needed). Like in case of 2008-03-06 to 2008-03-20 for retail vs average, though the correlation went really high because of opposite directions but the fluctuation in prices were not very prominent.
- If all the timeseries follows same anomalous behavior then it is not possible to find anomaly in any.
- If the anomaly is for very small tenure compared to the selected window size then it is likely to be missed or go unnoticed. In case of 2011-01-06 - 2011-01-08

1.2.4 Multivariate Time Series- Vector Autoregressive

The method uses vector autoregressive framework for multivaraiate time-series analysis in order to forecast values by using all the related variables/timeseries that can impact the timeseries. The error in the predicted value with the original value helps in determining anomalous points from the timeseries. MAD test is applied in order to fix the threshold value for the error. 60% (This can be configured as per need) of the data is used for calibration and rest of the data is checked for anomalous points. So, the start

date for the anomalous points begins after 2011-09-16
The function is tested with different set of timeseries. Following are the four tests which were performed:

1. **Retail Price vs Average of Retail Price**: All the centres should ideally move in tandem. So test is performed over every centre with other centres helping in predicting values.
2. **Retail Price vs Arrival of Onion**: Arrival is one of the major deciding factor for retail. So the predictions are made for retail based on arrival.
3. **Retail Price vs Wholesale Price**: Retail prices are predicted based on the wholesale prices.
4. **Wholesale Price vs Arrival of Onion**: Wholesale timeseries is predicted on the basis of arrival timeseries.

Observations when retail price timeseries is compared with average of retail price timeseries:

- The threshold selected by this method for error values are -19.0358099572 and 107.697818954 which means all the data points with error values less than -19.0358099572 and greater than 107.697818954 are reported by the method.

  Some of the tenures for which the method reported anomalies are:

  – 2013-07-19 to 2013-09-11 (See Figure 1.36)
  – 2013-09-17 to 2014-01-06 (See Figure 1.37)
  – 2014-12-03 to 2014-12-15 (See Figure 1.38)

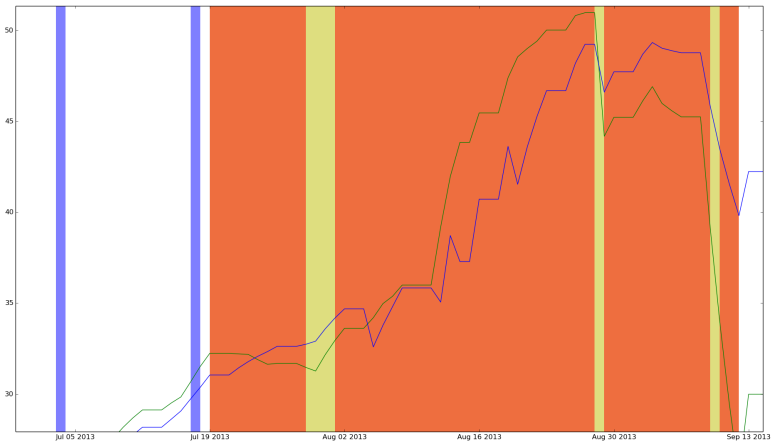

Figure 1.36: Vector Autoregressive (Green line - Centre Retail Price, Blue Line - Average Retail Price)

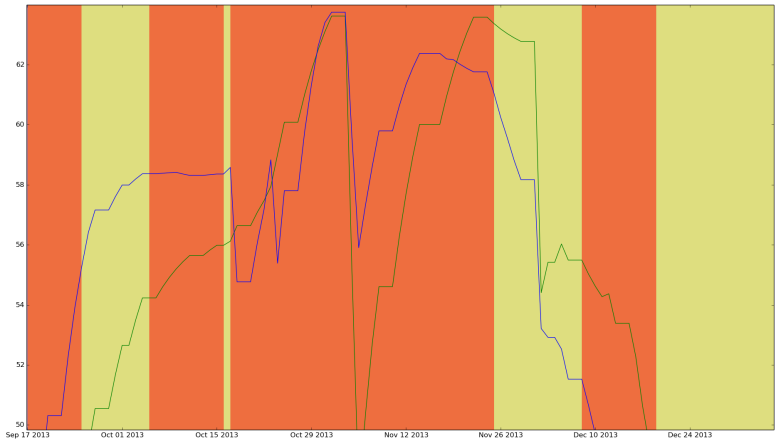

Figure 1.37: Vector Autoregressive (Green line - Centre Retail Price, Blue Line - Average Retail Price)

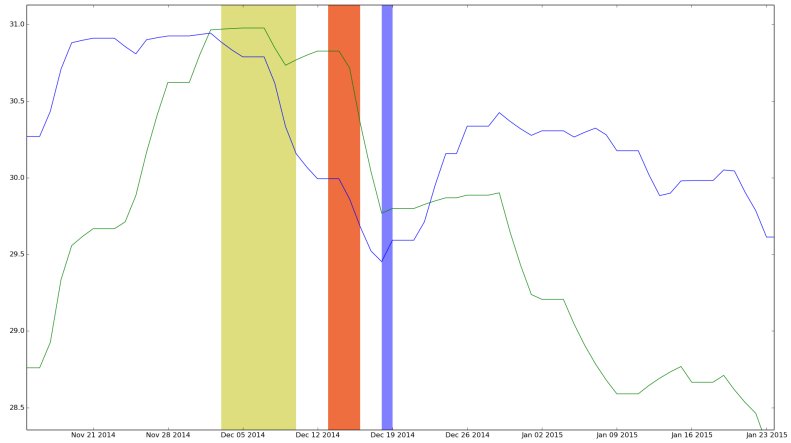

Figure 1.38: Vector Autoregressive (Green line - Centre Retail Price, Blue Line - Average Retail Price)

No data points were reported against news articles present in May, June, July 2014 despite of large error value. The threshold selected by the MAD test was higher than the error values. This could be corrected by manually setting threshold values which can be account for these data points.

Other interesting observations that were found while conducting above mentioned tests are as following:

- Some of the data points like July 2014 data were not captured despite of error value close to MAD threshold. These can be captured by lowering the threshold.
- For news articles in Dec 2012,Jan 2013, Feb 2013 the values does not show high error values despite of news articles reporting the crisis. The possible reason could be because similar trend have been seen for this tenure in the data.

Limitations of method:

- The method depends on the MAD Test in order to set threshold. So, even though the error value is close to threshold but less than it won't be reported. Other methods could be also used in order to decide threshold.

1.2.5 Graph Based Anomaly Detection

This method, treats each day as a node of a graph, and connects with other nodes if nodes are similar. This connecting edge is given similarity value and random walk is performed to get connectivity of each node. Node with the least connectivity values are reported as anomaly. Note that for the previous methods, we had threshold values either defined by user or calculated by using MAD test. But here we do not have that and we just ask method to report "n" number of nodes with least connectivity values.

The working of this method is quite complex and can not be generalised. For detailed information go through the paper. So we will just represent, how method has performed on the different analysis.

For **Retail Price vs Average of Retail Price** (See Figure 1.39) and **Retail Price vs Wholesale Price** (See Figure 1.40), this method has performed well. For **Retail Price vs Average of Retail Price**, every tenure of anomaly has been matched with some news articles. The anomalies which were not matched with news articles were part of large tenure which had some matching with news articles and usually, this tenure is large and for every date news articles are not present. Few articles are missed that might be due to limited number of points chosen. If number of points are increased, than it might be covered as well. For **Retail Price vs Wholesale Price**, apart from Jan 2013, July 2014, June 2015, all anomalies are matching with some news articles.

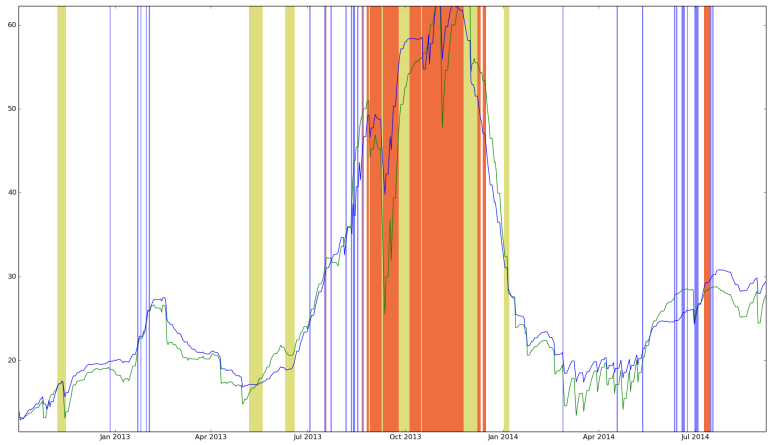

Figure 1.39: Graph Based Anomaly Detection (Green line - Centre Retail Price, Blue Line - Average Retail Price)

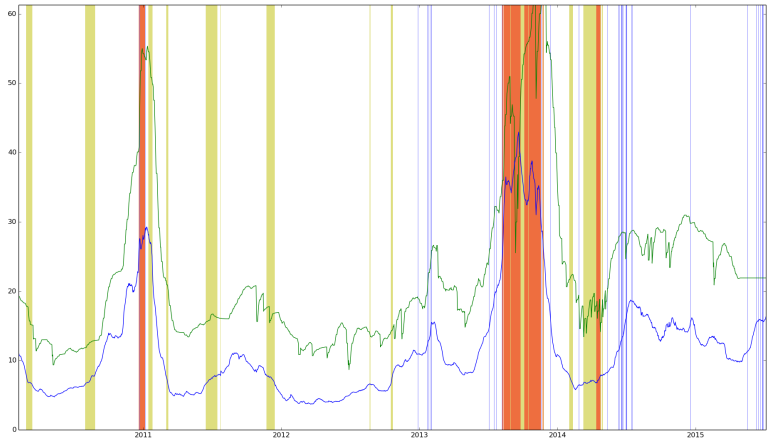

Figure 1.40: Graph Based Anomaly Detection (Green line - Retail Price, Blue Line - Wholesale Price)

For **Retail Price vs Arrival of Onion** (See Figure 1.41) and **Wholesale Price vs Arrival of Onion** (See Figure 1.42), this method is not producing good results. Many points are reported as anomaly which are close to each other. And due to limited number of points, number of anomalies matching with news articles are quite less. Figures 1.43 and 1.44 describe results of these both analysis for Delhi centre.

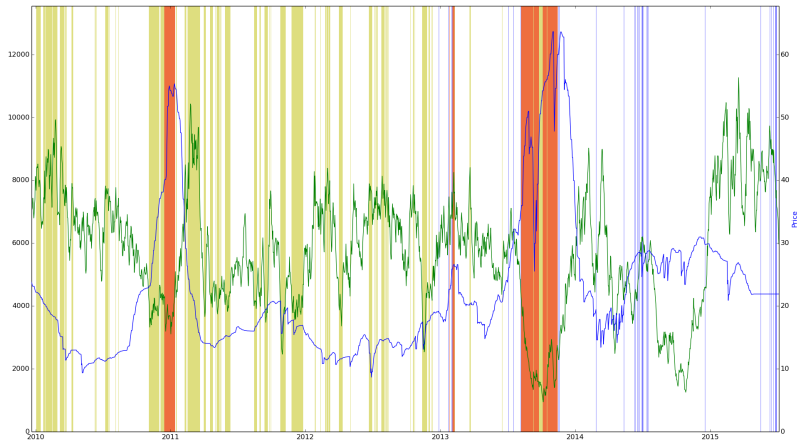

Figure 1.41: Graph Based Anomaly Detection (Green line - Arrival Data of Onion, Blue Line - Retail Price)

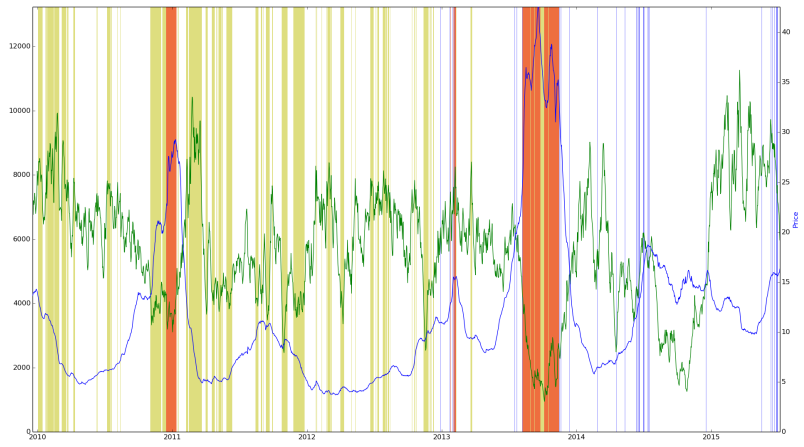

Figure 1.42: Graph Based Anomaly Detection (Green line - Arrival Data of Onion, Blue Line - Wholesale Price)

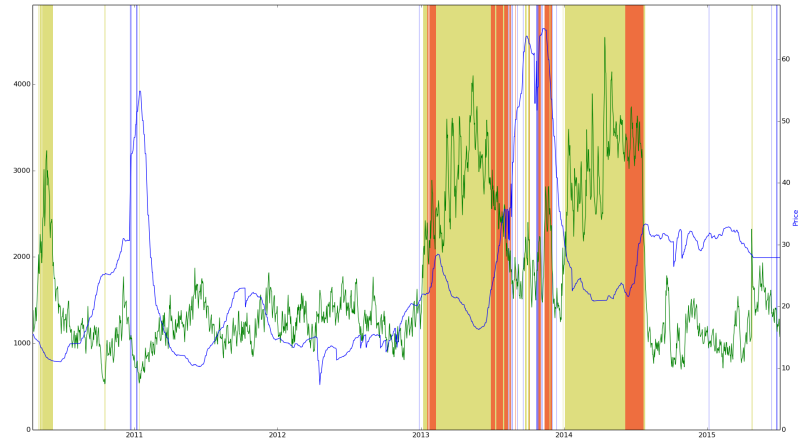

Figure 1.43: Graph Based Anomaly Detection (Green line - Arrival Data of Onion, Blue Line - Retail Price)

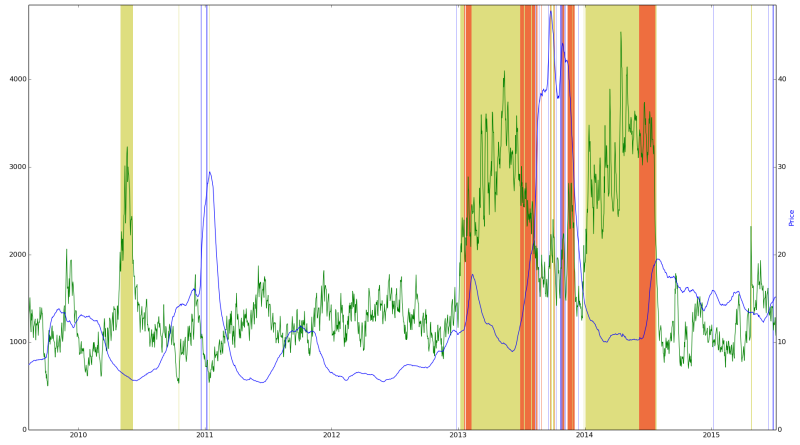

Figure 1.44: Graph Based Anomaly Detection (Green line - Arrival Data of Onion, Blue Line - Wholesale Price)