Results and Analysis

Kapil Thakkar and Reshma Kumari

June 11, 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 centers are considered over here, 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 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.

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 **143**. 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.

Now, we present detailed analysis for each of the different type of time-series. Fist type of such analysis is in table 1.3. This table shows distribution of news articles present year-wise for each method when retail price time series is compared with average retail price time series. Fist 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 is described in different tables for different analysis. Result for various analysis is described in tables 1.7, 1.8, 1.9 and 1.10.

Such results of different cities can also be calculated.

Methods	Slope Based (1)	Window 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	1245	353	100	177	1871	192	136
Retail Vs Arrival	420	120	353	100	167	810	267	173
Retail Vs Wholesale	658	1230	310	100	167	1819	229	132
Wholesale Vs Arrival	448	525	282	100	186	1165	286	217

Table 1.1: Anomalies Reported

Methods	Slope Based (1)	Window 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	1245	353	100	177	1871	192	136
Retail Vs Arrival	420	120	353	100	167	810	267	173
Retail Vs Wholesale	658	1230	310	100	167	1819	229	132
Wholesale Vs Arrival	448	525	282	100	186	1165	286	217

Table 1.2: Number of news articles matched with system

Distribution of All Articles	Articles Present	Slope Based (1)	Window Correlation (2)	Linear Regression (3)	Graph Based (4)	Multivariate (5)	1 U 2 U 3 (6)	4 U 5 (7)	6 ∩ 7
2010	7	0	5	6	0	0	6	0	0
2011	7	0	3	0	2	0	2	2	2
2012	2	1	1	0	0	0	1	0	0
2013	77	46	31	17	30	52	50	52	50
2014	37	6	18	14	0	1	20	1	1
2015	13	0	12	0	0	0	12	0	0
Total	143	53	70	37	32	53	91	55	53

Table 1.3: Retail Price VS Average Retail Price

Distribution of All Articles	Articles Present	Slope Based (1)	Window Correlation (2)	Linear Regression (3)	Graph Based (4)	Multivariate (5)	1 U 2 U 3 (6)	4 U 5 (7)	6 ∩ 7
2010	7	0	5	6	0	0	6	0	0
2011	7	0	3	0	2	0	2	2	2
2012	2	1	1	0	0	0	1	0	0
2013	77	46	31	17	30	52	50	52	50
2014	37	6	18	14	0	1	20	1	1
2015	13	0	12	0	0	0	12	0	0
Total	143	53	70	37	32	53	91	55	53

Table 1.4: Retail Price VS Arrival data of onion

Distribution of All Articles	Articles Present	Slope Based (1)	Window Correlation (2)	Linear Regression (3)	Graph Based (4)	Multivariate (5)	1 U 2 U 3 (6)	4 U 5 (7)	6 ∩ 7
2010	7	0	5	6	0	0	6	0	0
2011	7	0	3	0	2	0	2	2	2
2012	2	1	1	0	0	0	1	0	0
2013	77	46	31	17	30	52	50	52	50
2014	37	6	18	14	0	1	20	1	1
2015	13	0	12	0	0	0	12	0	0
Total	143	53	70	37	32	53	91	55	53

Table 1.5: Retail Price VS Wholesale Price

Distribution of All Articles	Articles Present	Slope Based (1)	Window Correlation (2)	Linear Regression (3)	Graph Based (4)	Multivariate (5)	1 U 2 U 3 (6)	4 U 5 (7)	6 ∩ 7
2010	7	0	5	6	0	0	6	0	0
2011	7	0	3	0	2	0	2	2	2
2012	2	1	1	0	0	0	1	0	0
2013	77	46	31	17	30	52	50	52	50
2014	37	6	18	14	0	1	20	1	1
2015	13	0	12	0	0	0	12	0	0
Total	143	53	70	37	32	53	91	55	53

Table 1.6: Wholesale Price VS Arrival data of onion

Distribution of All Articles	Slope Based (1)	Window Correlation (2)	Linear Regression (3)	Graph Based (4)	Multivariate (5)	1 U 2 U 3 (6)	4 U 5 (7)	6 ∩ 7
2006	35	30	0	78	0	58	78	7
2007	63	0	0	22	0	63	22	0
2008	28	0	0	0	0	28	0	0
2009	28	15	0	0	0	43	0	0
2010	48	15	46	0	0	82	0	0
2011	36	30	40	0	0	105	0	0
2012	77	0	0	0	0	77	0	0
2013	59	0	168	0	161	203	161	161
2014	46	30	99	0	6	151	6	5
2015	0	0	0	0	0	0	0	0
Total	420	120	353	100	167	810	267	173

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

Distribution of All Articles	Slope Based (1)	Window Correlation (2)	Linear Regression (3)	Graph Based (4)	Multivariate (5)	1 U 2 U 3 (6)	4 U 5 (7)	6 ∩ 7
2006	35	30	0	78	0	58	78	7
2007	63	0	0	22	0	63	22	0
2008	28	0	0	0	0	28	0	0
2009	28	15	0	0	0	43	0	0
2010	48	15	46	0	0	82	0	0
2011	36	30	40	0	0	105	0	0
2012	77	0	0	0	0	77	0	0
2013	59	0	168	0	161	203	161	161
2014	46	30	99	0	6	151	6	5
2015	0	0	0	0	0	0	0	0
Total	420	120	353	100	167	810	267	173

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

Distribution of All Articles	Slope Based (1)	Window Correlation (2)	Linear Regression (3)	Graph Based (4)	Multivariate (5)	1 U 2 U 3 (6)	4 U 5 (7)	6 ∩ 7
2006	35	30	0	78	0	58	78	7
2007	63	0	0	22	0	63	22	0
2008	28	0	0	0	0	28	0	0
2009	28	15	0	0	0	43	0	0
2010	48	15	46	0	0	82	0	0
2011	36	30	40	0	0	105	0	0
2012	77	0	0	0	0	77	0	0
2013	59	0	168	0	161	203	161	161
2014	46	30	99	0	6	151	6	5
2015	0	0	0	0	0	0	0	0
Total	420	120	353	100	167	810	267	173

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

Distribution of All Articles	Slope Based (1)	Window Correlation (2)	Linear Regression (3)	Graph Based (4)	Multivariate (5)	1 U 2 U 3 (6)	4 U 5 (7)	6 ∩ 7
2006	35	30	0	78	0	58	78	7
2007	63	0	0	22	0	63	22	0
2008	28	0	0	0	0	28	0	0
2009	28	15	0	0	0	43	0	0
2010	48	15	46	0	0	82	0	0
2011	36	30	40	0	0	105	0	0
2012	77	0	0	0	0	77	0	0
2013	59	0	168	0	161	203	161	161
2014	46	30	99	0	6	151	6	5
2015	0	0	0	0	0	0	0	0
Total	420	120	353	100	167	810	267	173

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.

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 type 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 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 here 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)
- *Analysis 3*: Apr Jul Oct Dec 2010, Jan 2011, May Jun 2014

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

- This observation is limitation of this method. There exist few cases where, drop in retail price for one center is quite huge as compared to drop in price for average retail price (in case of analysis 1) or wholesale price (in case of analysis 3). This is good thing for centres, and should bot be 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, Jun Aug 2012, Jan 2013, Feb Oct 2014, Feb 2015
  - *Analysis 3*: Feb May 2010
- 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 was not reported by our system. Such cases are reported for the following tenure by this method:
  - *Analysis 1*: Dec 2010, Jan Feb 2012, Jun 2013, May June 2015
  - *Analysis 3*: Feb 2013, Jul 2014
- 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.
- In some cases, tenure reported as anomaly is quite large, because situations were abnormal for long time. But may be for such large tenure, news articles should be present, is not necessary. And anomaly reported was justifiable. For *Analysis 1*, such tenure was reported for March end to May start 2014.
- 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 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.
- There exist some cases in *Analysis 3* where retail price were decreasing but wholesale price was keep on increasing, this creatd 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.

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
  - *Analysis 4*: Sept Oct Nov 2010, Jun 2013, Jun 2015
- 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

- *Analysis 4*: Oct Dec 2011, Jan 2012, Jan 2014, Mar 2011

- 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 so, 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
  - *Analysis 4*: Jan Feb 2013, July 2014, June 2015

- There exist some cases, where first due to low arrival prices went too high and when arrival started entering into market slowly and slowly prices were going down. This period of slowly 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 and in *Analysis 4* for Nov 2013. 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.

- 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 our system. Such cases in *Analysis 2* exist for Aug Sept 2013

- 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

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 type 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, 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:

- 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 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
  - *Analysis 3*: Feb Mar 2011, Jan 2012, June 2012, Jan Feb 2014, Apr 2015
- One limitation of this method is both the series have high values for some time period and difference between them is not so huge then that will not be reported. 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
  - *Analysis 3*: Feb 2013, Aug Sept Oct Nov 2013, June July 2015
- 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.

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. Such cases are reported for the following tenure by this method:
  - *Analysis 2*: Dec Jan Feb 2011, Aug Sept Oct Nov 2013, Oct Dec 2014
  - *Analysis 4*: Dec 2010, July Aug Sept Oct Nov Dec 2013, July 2014
- 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
  - *Analysis 4*: Jan Feb 2013, June July 2013, June 2015

1.2.3 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** and **Retail Price vs Wholesale Price**, 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.

For **Retail Price vs Arrival of Onion** and **Wholesale Price vs Arrival of Onion**, this method is not producing good results. Many points are reported as anomaly which are close to each other. And due to that for limited number of points, which we are passing as 150, number of anomalies matching with news articles are quite less.