Sentiment EDA

Alberto Munisso

Study objective is to detect possible links between daily compound sentiment values provided by VADER sentiment dictionary and the daily International Grains Council Wheat Price Index. The detection will focus on the comparison between two increasing trend periods and two decreasing trend periods, namely the decreasing trend intervals 28/05/2011 – 03/10/2011 and 03/01/2015 – 03/06/2015 and the increasing trend intervals 03/07/2010 - 03/03/2011 and 15/06/2012 - 03/10/2012.

The expected outcome is an average sentiment significant difference in each opposite trend window comparison. Moreover compound sentiment and price index returns linear relationship will be checked.

Data

Data is composed by International Grain Council daily series and a sentiment matrix. Sentiment matrix is composed by articles index, timestamp and negative, neutral, positive and compound registered sentiment.
In particular wheat index series is composed by 4181 daily prices, spanning from 03/01/2000 to 12/01/2016.
Sentiment is represented by 60934 negative, neutral, positive and compound observations spanning from 01/09/1991 to 28/07/2016 are available. Since most of the observations are clustered in the the last years, it must be taken into account that sentiment values are not equally distributed in time. It has to be underlined that VADER sentiment dictionary focuses more on "human sentiment" rather than "economic sentiment", therefore it gives a 0 negative and positive sentiment value to neutral economic statements as "A +/-3.5% price increase/decrease is expected". This fact explains why many near to zero sentiment values are registered, mirroring many near to 1 neutral scores. Sentiment matrix has been generated using the script main.py (available on Sentiment-Test-Alberto branch).

Dataframe composition

In order to compare International Grain Council Wheat Index and Sentiment observations, it's mandatory to build a data frame that is homogeneous in time, given many discrepancies in data availability. On one hand not all daily wheat index observations have at least one corresponding sentiment bearing article on the same date, while on the other hand may exist more than one sentiment bearing news article on the same day.

A time consistent data frame is built using the following methodology:

1) Sentiment Timestamps are converted back into date values.
2) On each date Sentiment values are collapsed into daily sentiment values using the arithmetic mean.
3) For each Wheat Index value is linked the corresponding day negative, neutral, positive and compound sentiment value by cross-checking International Grains Council Wheat date and Sentiment date (and viceversa).

Analysis

The analysis is composed by a list of dataframe items that have a compound sentiment higher (lower) than an arbitrary threshold, compound sentiment summary, compound sentiment variance, compound sentiment skewness, a wheat price index plot, a compound sentiment density plot and a wheat index returns and compound sentiment scatter plot.

The list of top compound sentiment values will be used to underline outliers.

Compound sentiment summary gives compound sentiment mean and median values: they will be compared across different periods.

Compound sentiment variance will be used to underline eventual pikes in sentiment variance, as skewness will be used to understand sentiment distribution.

Wheat price index plot is used to underline underlying trend through LOESS regression.

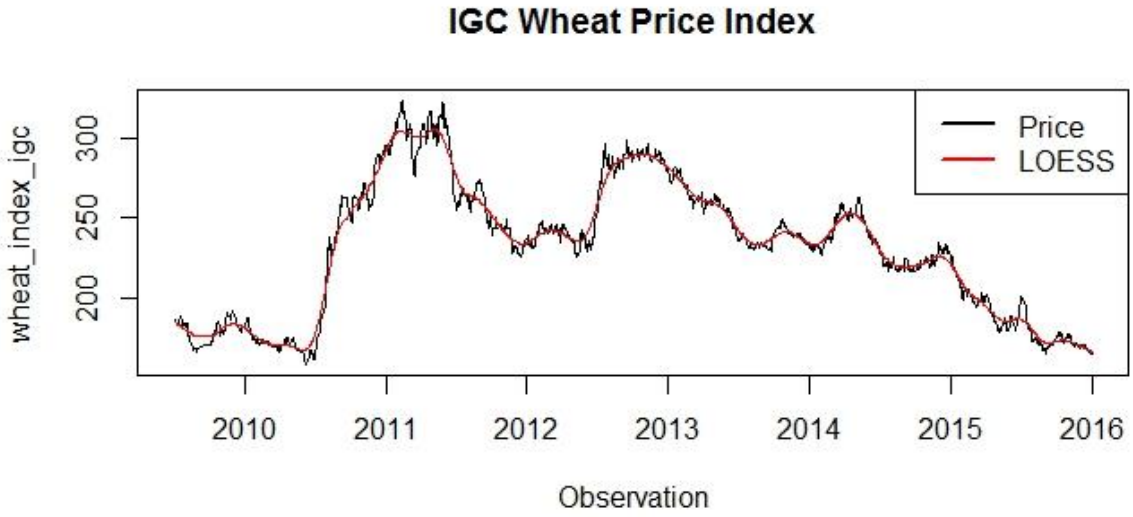
Wheat price index returns and compound sentiment scatterplot is used to detect an eventual trivial relationship through a basic linear model.

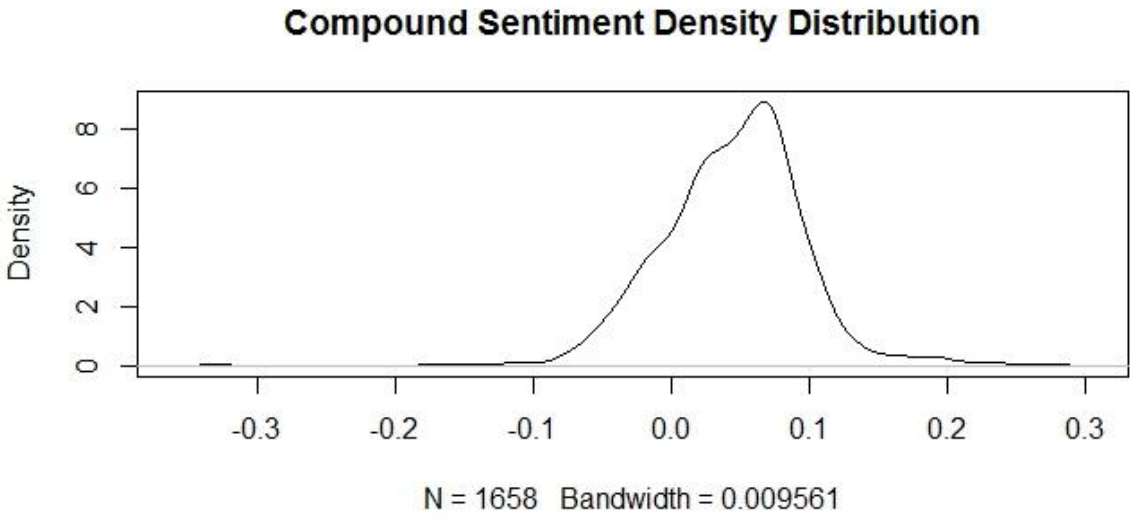
Results

2009-2016

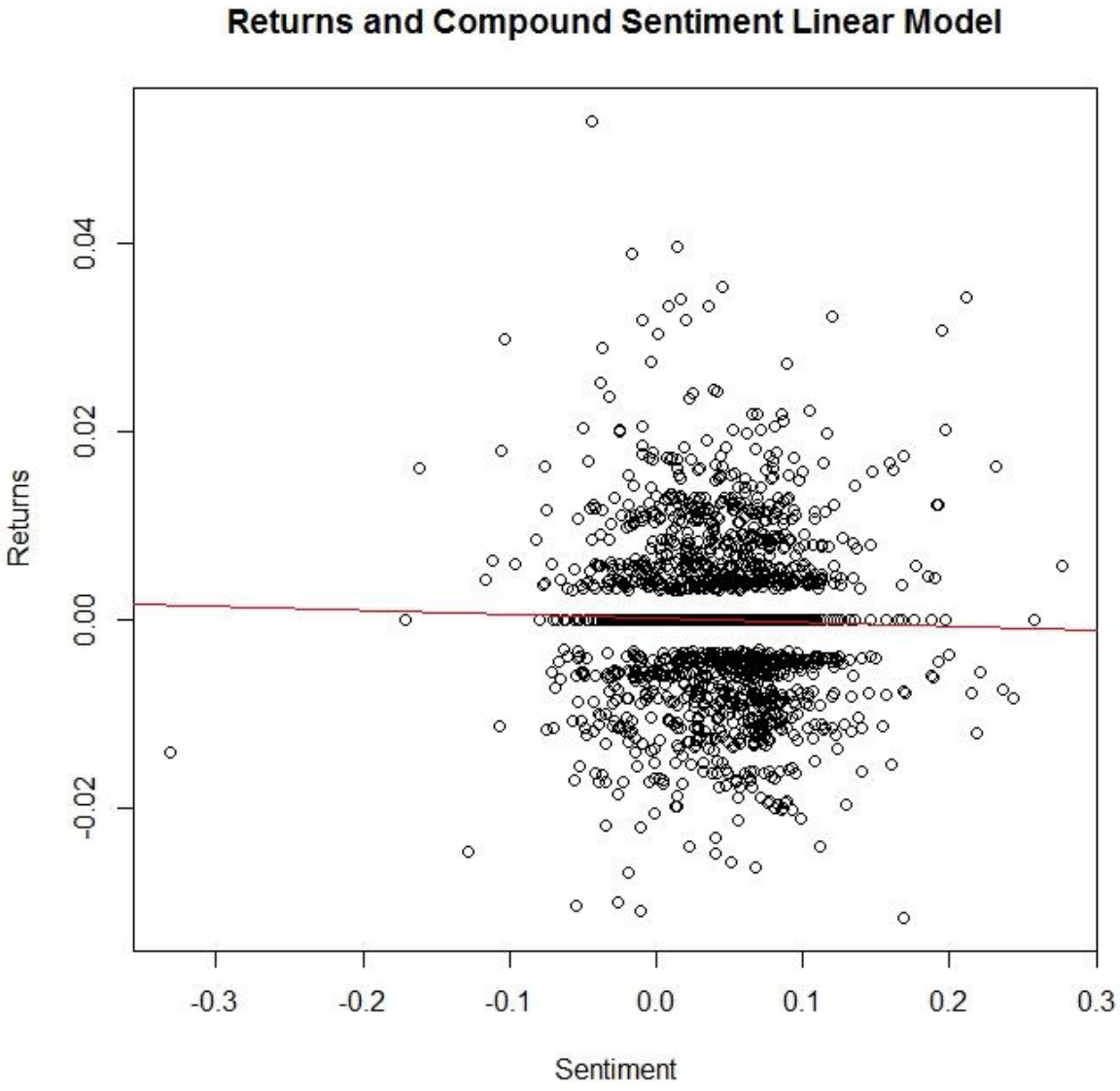
By analyzing 03/07/2009 and 03/01/2016 time interval two increasing trend windows appear, namely 03/07/2010 - 03/03/2011 and 15/06/2012 - 03/10/2012, while two decreasing trend windows are 28/05/2011 - 03/10/2011 and 03/01/2015 - 03/06/2015.

Compound sentiment is distributed (almost) normally around a mean equal to 0.045, a median equal to 0.048 and variance 0.003. Moreover a slightly negative skewness parameter is registered (-0.16) and a sentiment outlier equal to -0.33 (05/09/2012) exists.





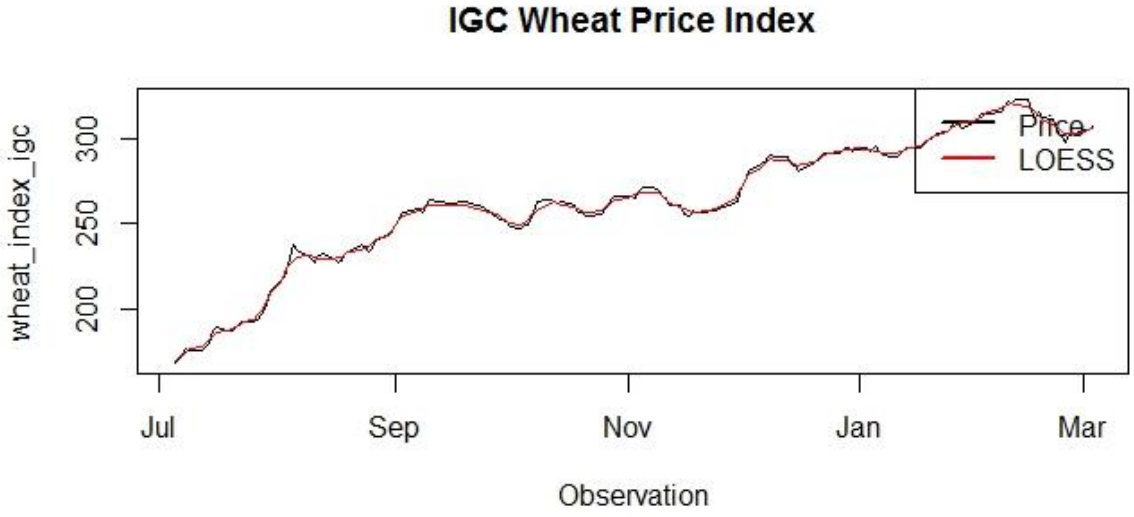
Overall it's not possible to assume any linear relationship between sentiment and price returns.

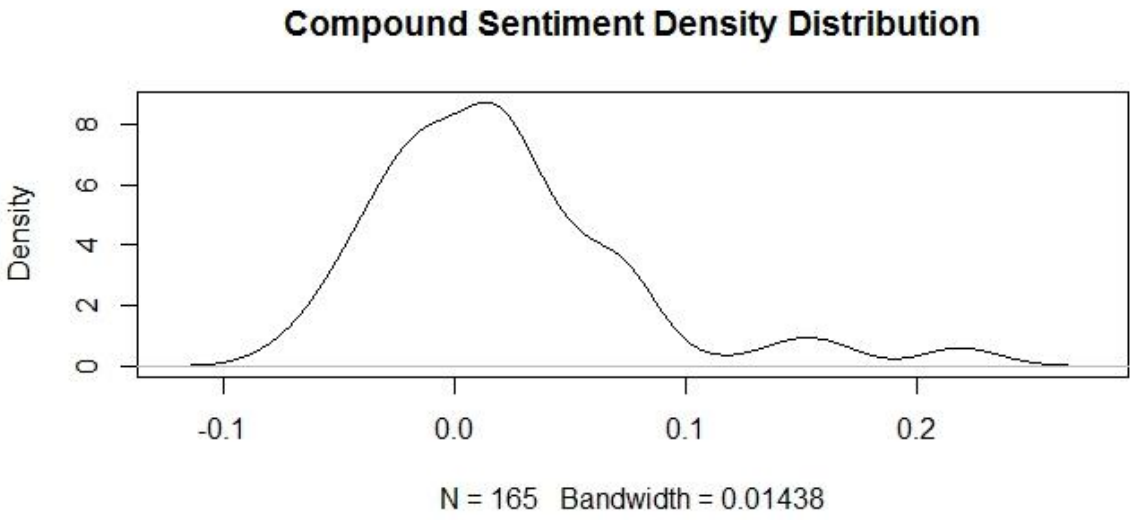


03/07/2010 - 03/03/2011

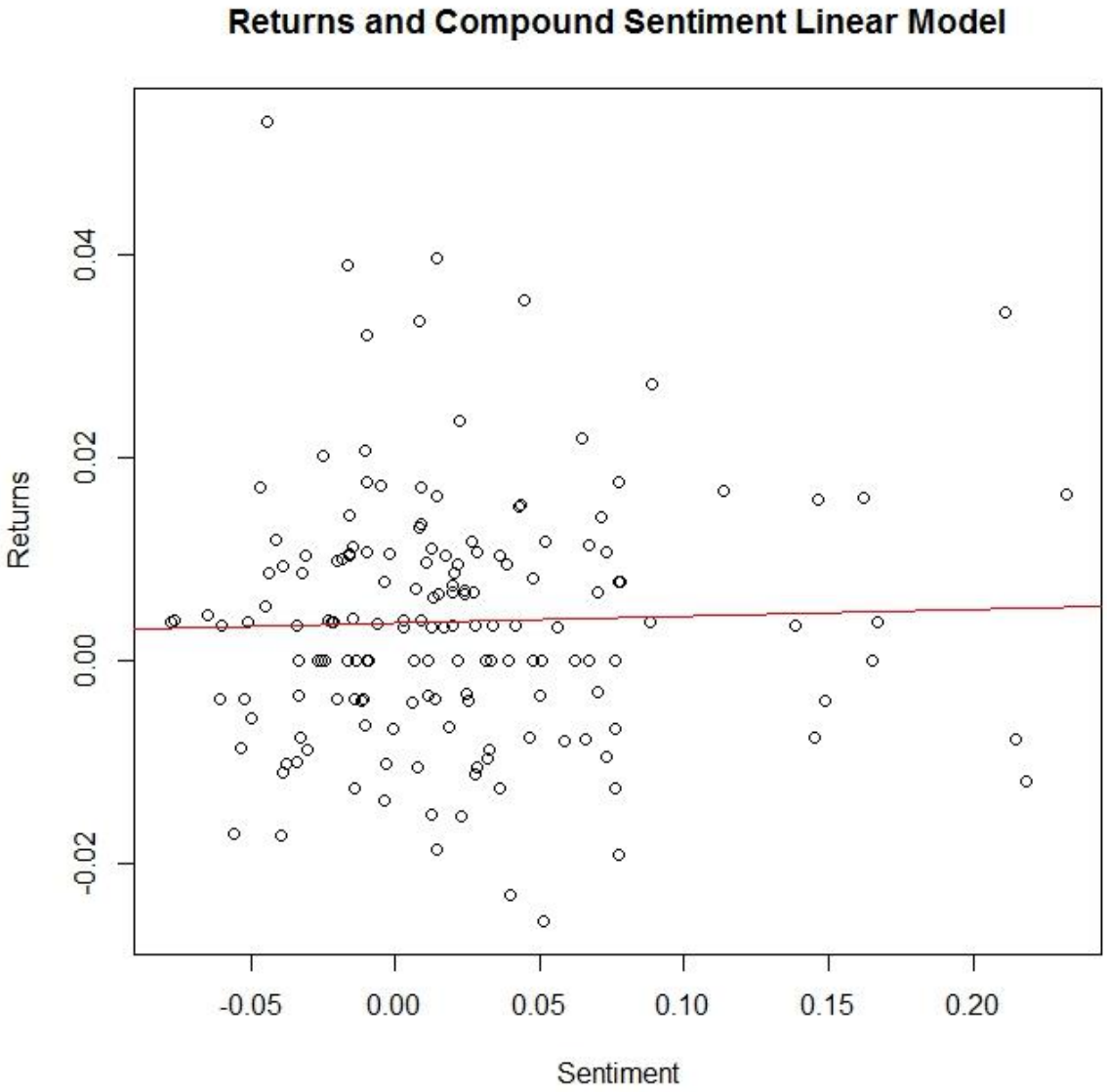
03/07/2010 - 03/03/2011 time interval is characterized by increasing price.

Sentiment mean and median are respectively equal to 0.021 and 0.013, therefore somehow lower than the 2009-2016 interval value. Variance is 0.003, which is equal to the 2009-2016 interval value. Moreover a positive skewness value is registered (+1.36), which stands for a big number of near to zero compound sentiment values.





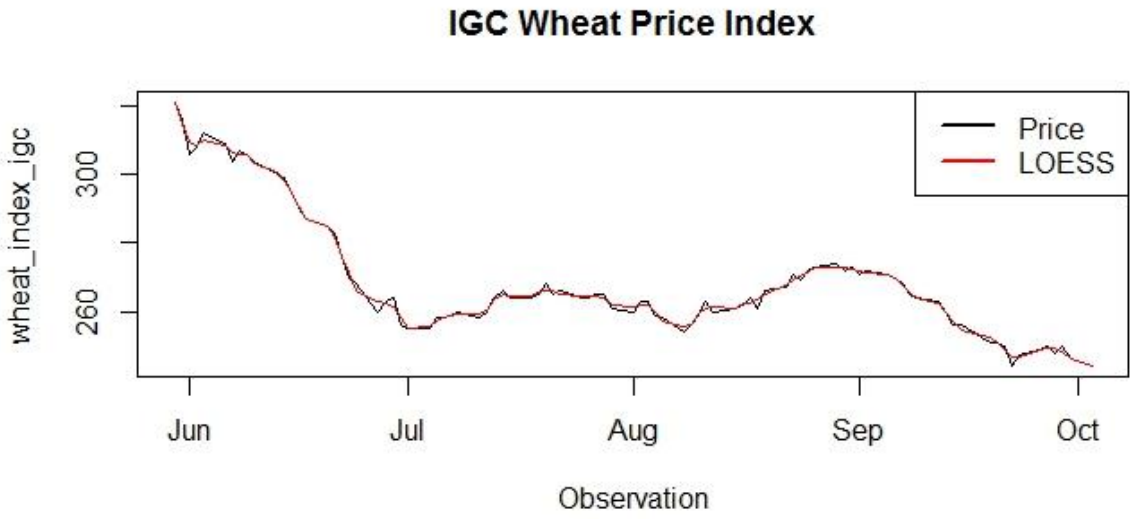
Also in this case no linear relationship between sentiment and returns can be assumed.

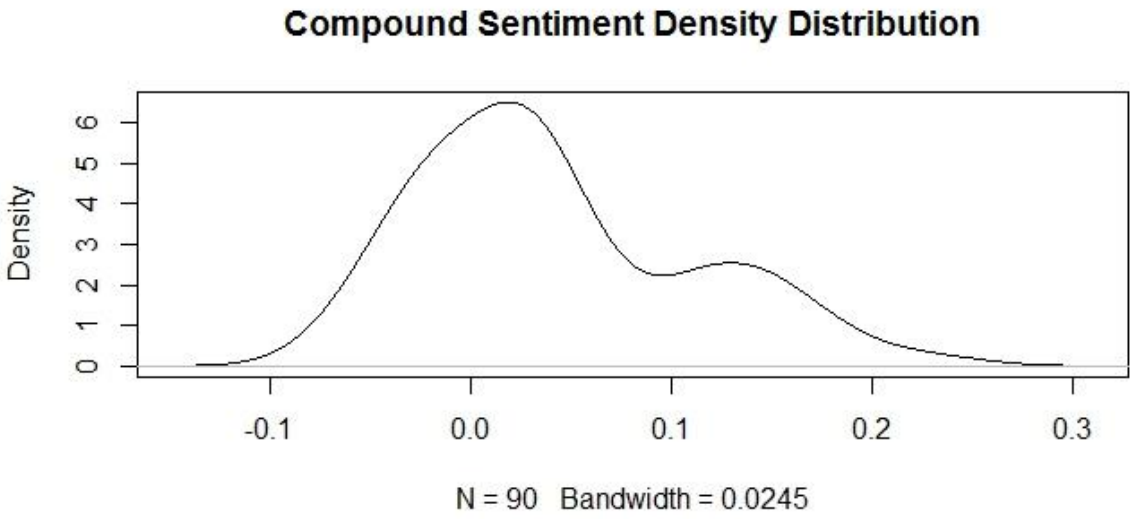


28/05/2011 - 03/10/2011

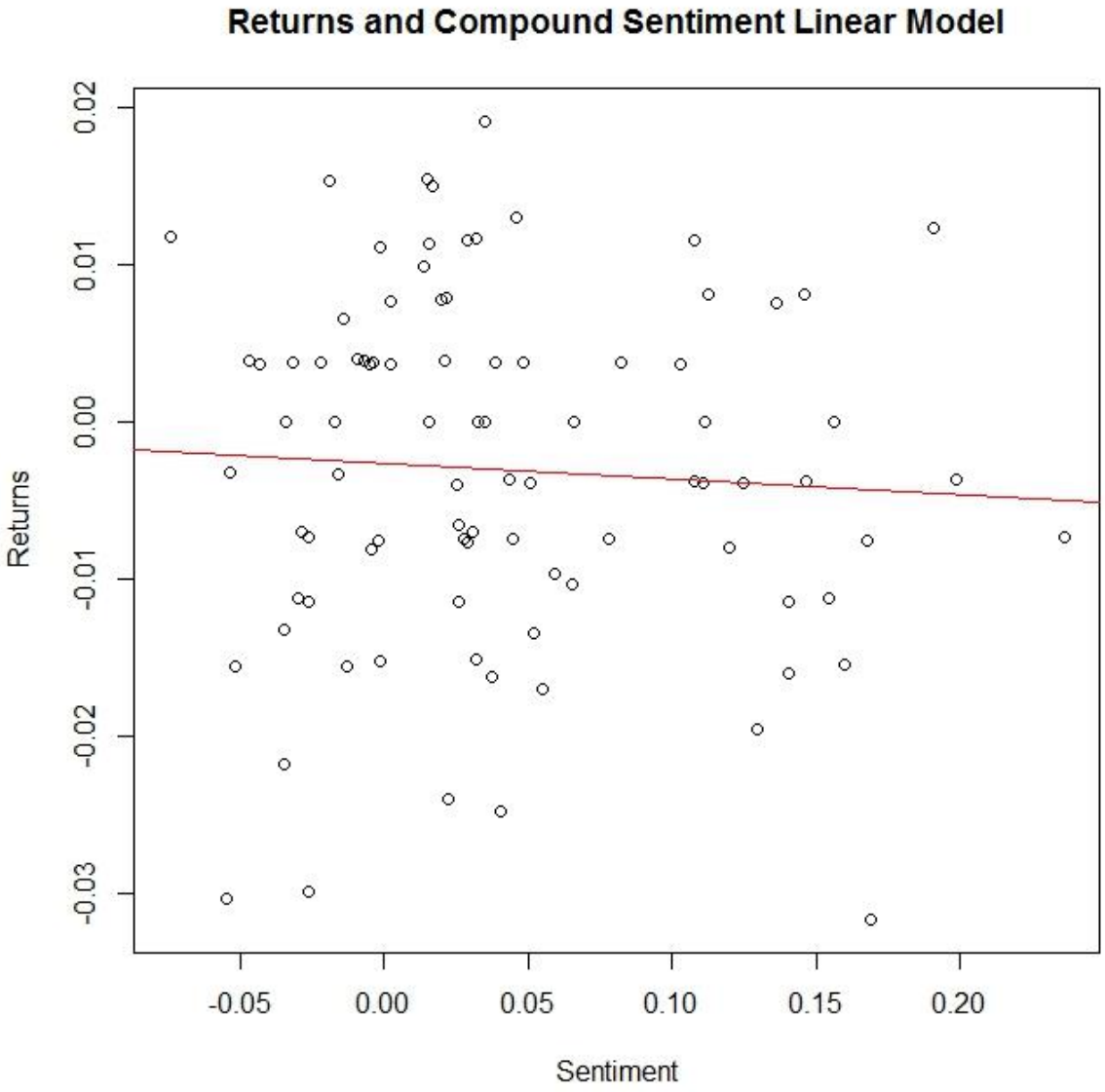
28/05/2011 - 03/10/2011 is a decreasing price window.

Sentiment mean and median are still slightly lower than 2009-2016 ones: they are respectively equal to 0.041 and 0.028. Sentiment variance is equal to 0.005, which is higher than 2009-2016 variance. A positive skewness value is registered (+0.71), which is lower than the previous period one. In this window sentiment is slowly moving towards higher values.





Also in this case no sentiment and returns linear relation can be drawn.

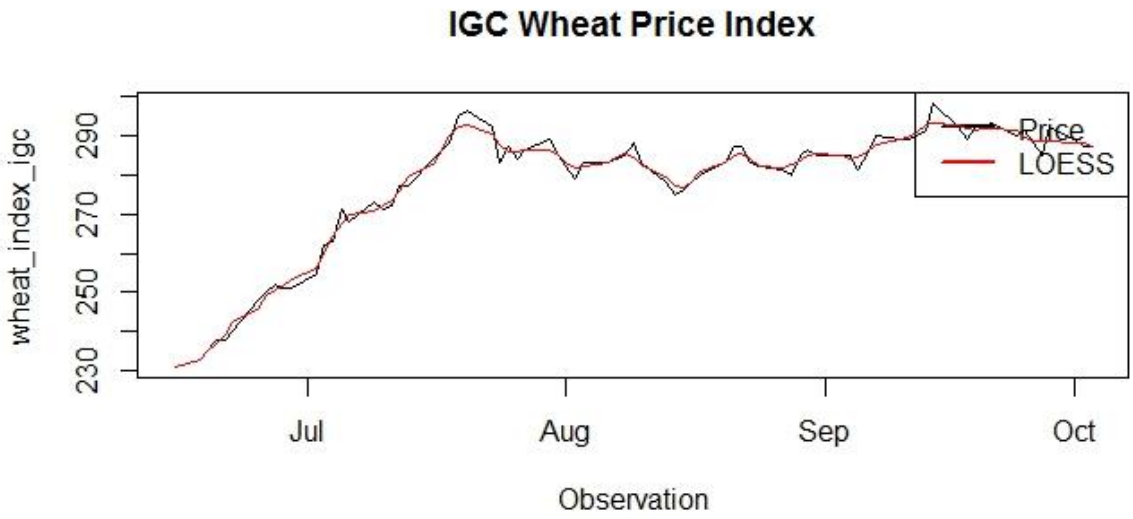


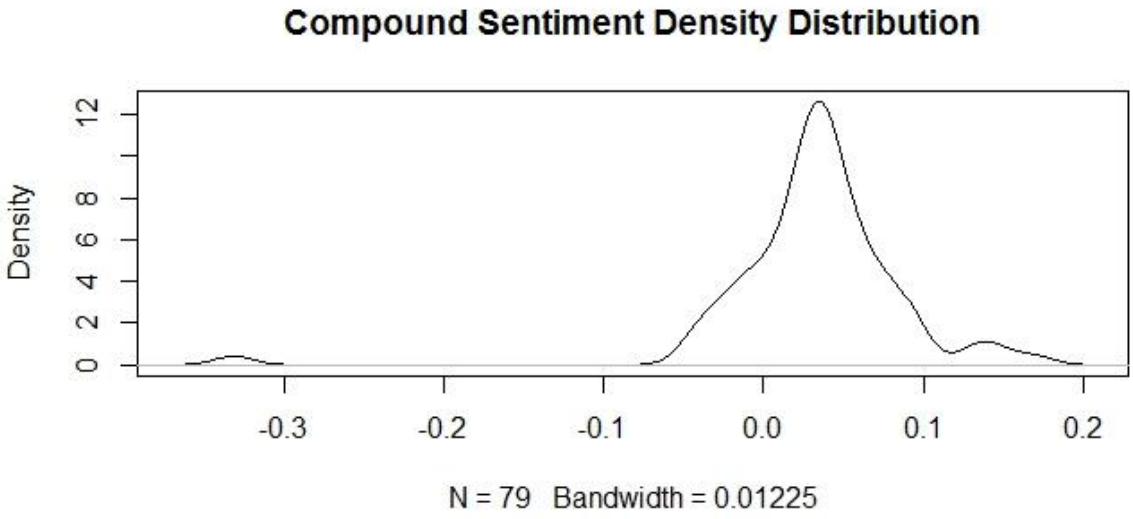
15/06/2012 - 03/10/2012

The interval is an increasing price period window.

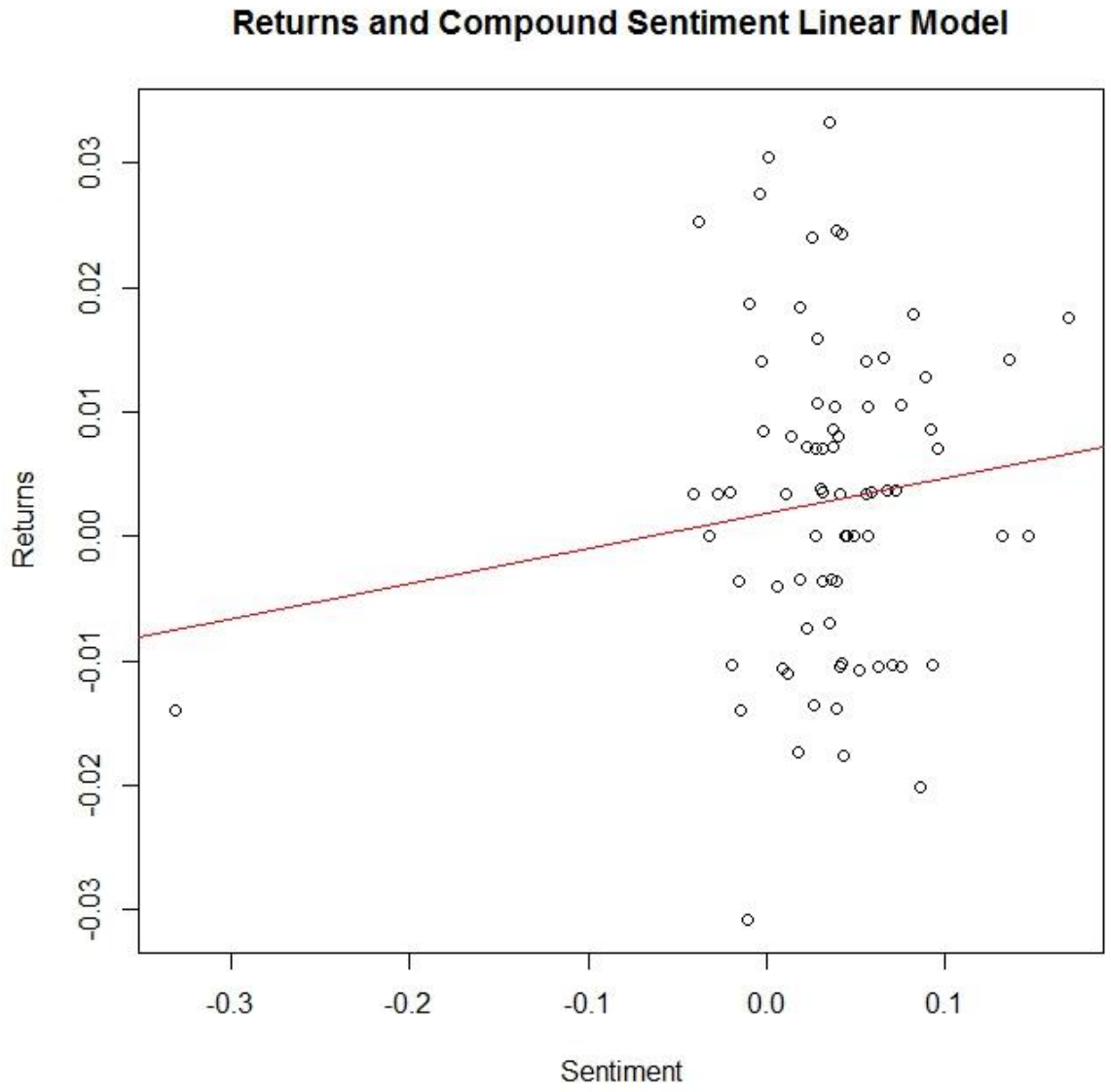
Compound Sentiment mean is equal to 0.032 and median is 0.036, which are both lower than the overall values. Sentiment variance is equal to 0.003, which is in line with the overall variance value. Compound sentiment skewness is highly negative (-2.81).

However all these values are highly affected by a compound sentiment outlier. By dropping the outlier skewness becomes positive (+0.62), therefore lower than the previous value. By considering mean and median values higher than 03/07/2010 - 03/03/2011 window, sentiment is still quite positive and in line of previous window one. This is also confirmed by a mean value equal to 0.037 and a median value of 0.036. Variance also suddenly drops to 0.001.





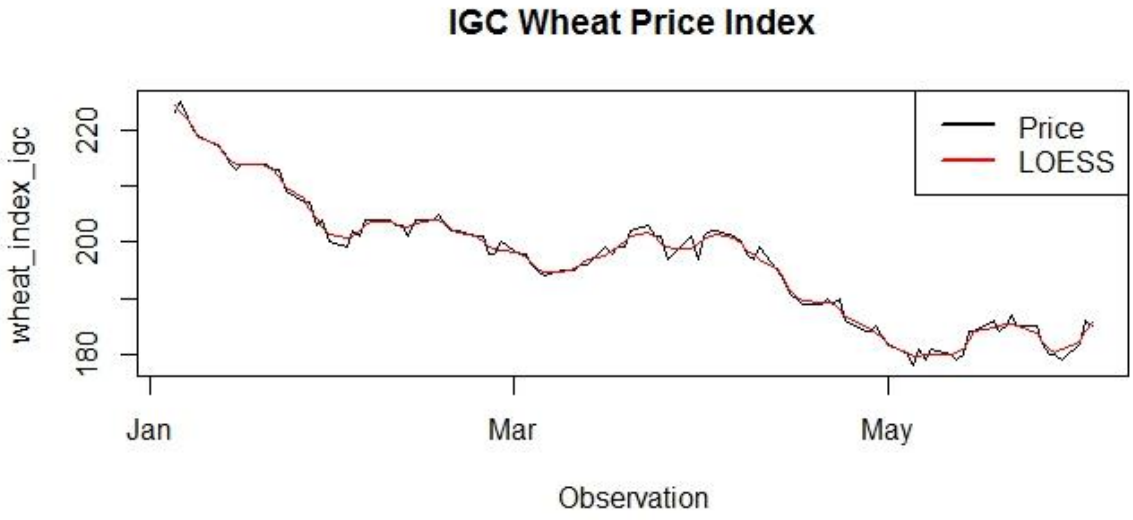
Keeping in the outlier value it seems that there exists some correlation, but dropping it will surely result in no linear relationship amongst the variables

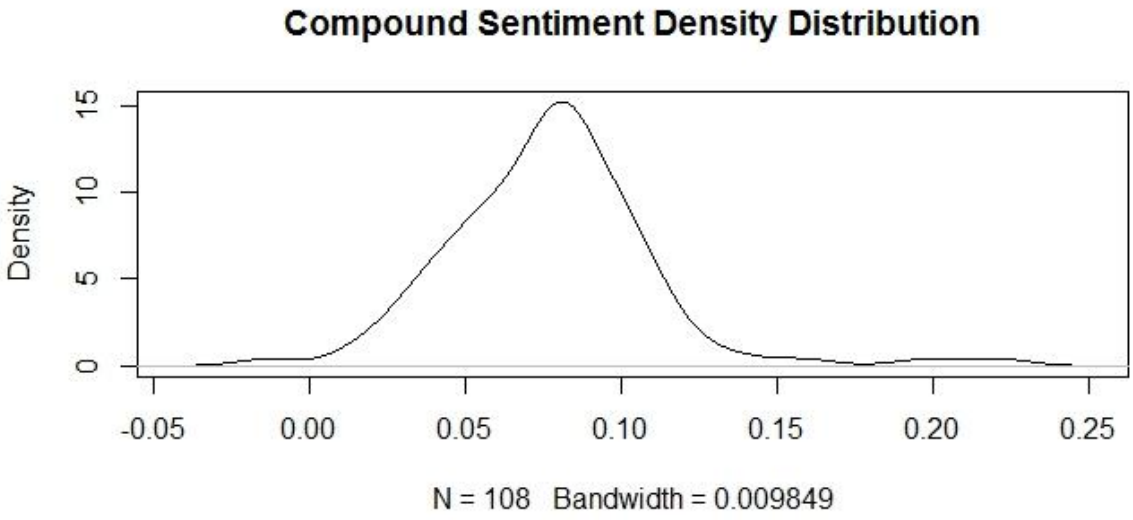


03/01/2015 - 03/06/2015

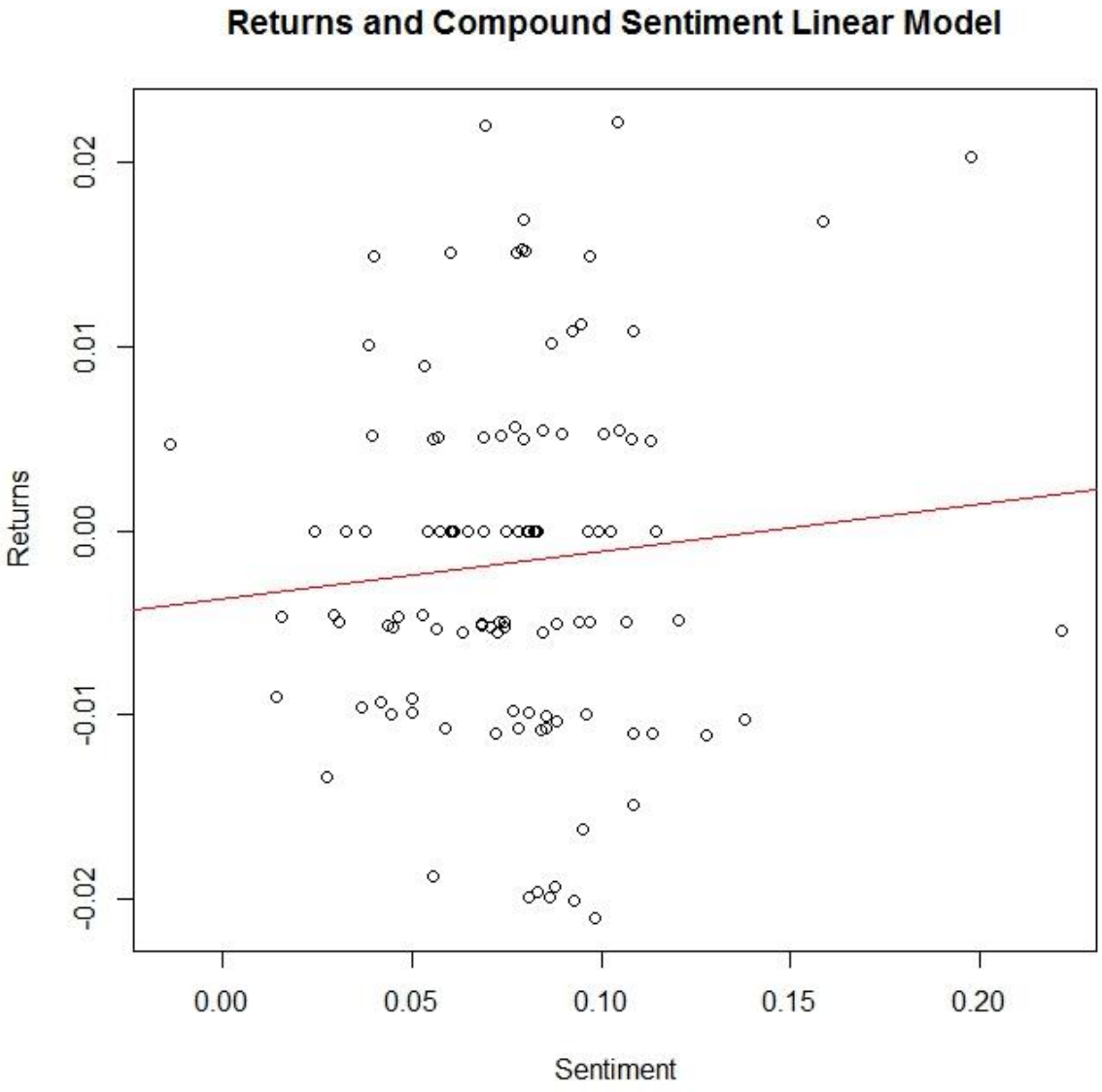
03/01/2015 - 03/06/2015 is a decreasing trend window.

Sentiment mean and median are higher than 2009-2016 values: this window value are respectively equal to 0.077 and 0.079. Sentiment variance is lower than the overall one and it's equal to 0.001. Compound sentiment skewness is still positive (+0.98).





Also in this case no linear relationship amongst the variables exists.



Conclusion

A considerable average compound sentiment difference is observed by comparing 2010 and 2015 opposite trend windows. Such difference is equal to 0.055, which is equal to 17% if compared to maximum absolute compound sentiment observed value.

Since these sentiment values are registered contemporaneously to price index observations, the evidence of different average compound sentiment values in opposite trend periods proves that sentiment is affected by price evolution.

In this context is observed that an increasing price trend is characterized by near to zero average compound sentiment, while positive compound sentiment seems to be associated to a decreasing price trend. This negative correlation is somehow the opposite that may be expected in equity markets, where usually positive sentiment is associated to rising asset prices. As possible explanation can be said that rising food commodities prices may lead to food crises, therefore negative sentiment is associated to increasing price trends (and viceversa).

However event sentiment polarity and magnitude really depends on dictionary used, therefore this result may change by changing dictionary.

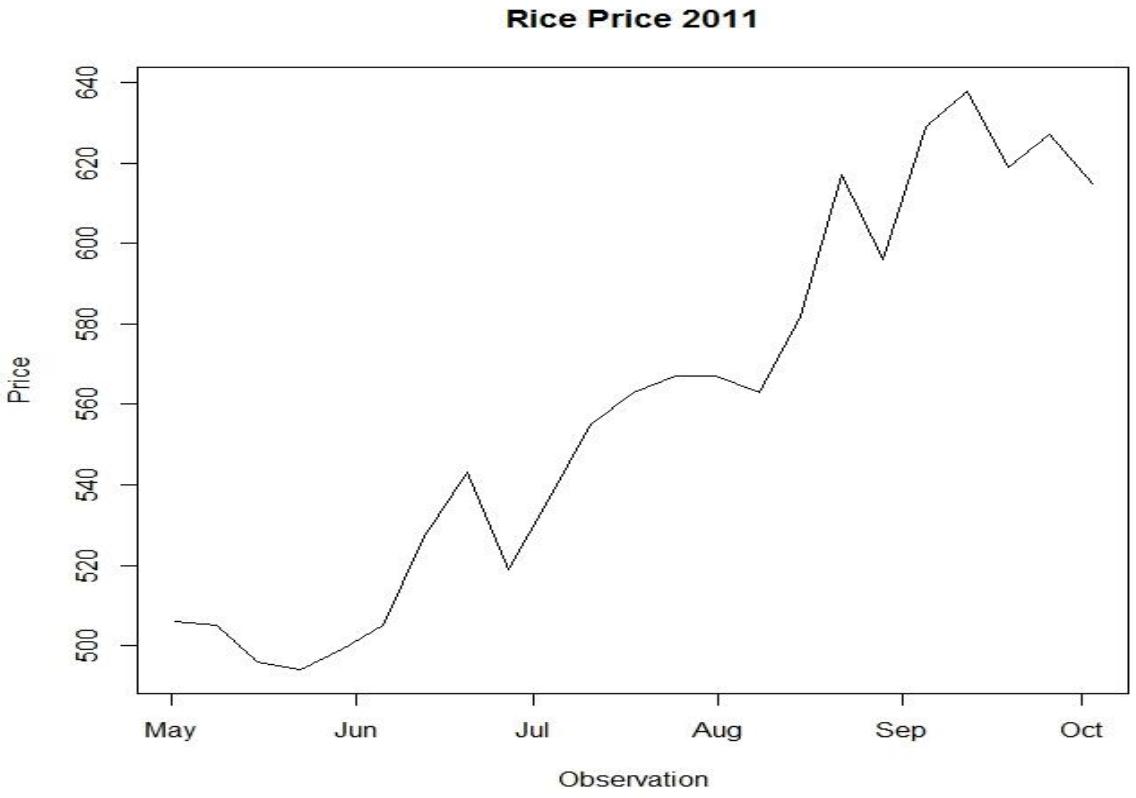
By the way such sentiment difference isn't observed in 2011 and 2012 opposite trend windows. It's legit to think that, since scraped compound sentiment is also relative to other commodities than wheat, then it's advisable to relate each commodity just to the corresponding sentiment. In particular it's observed that while wheat was decreasing in 2011 and 2012 windows, the same cannot be said for rice.

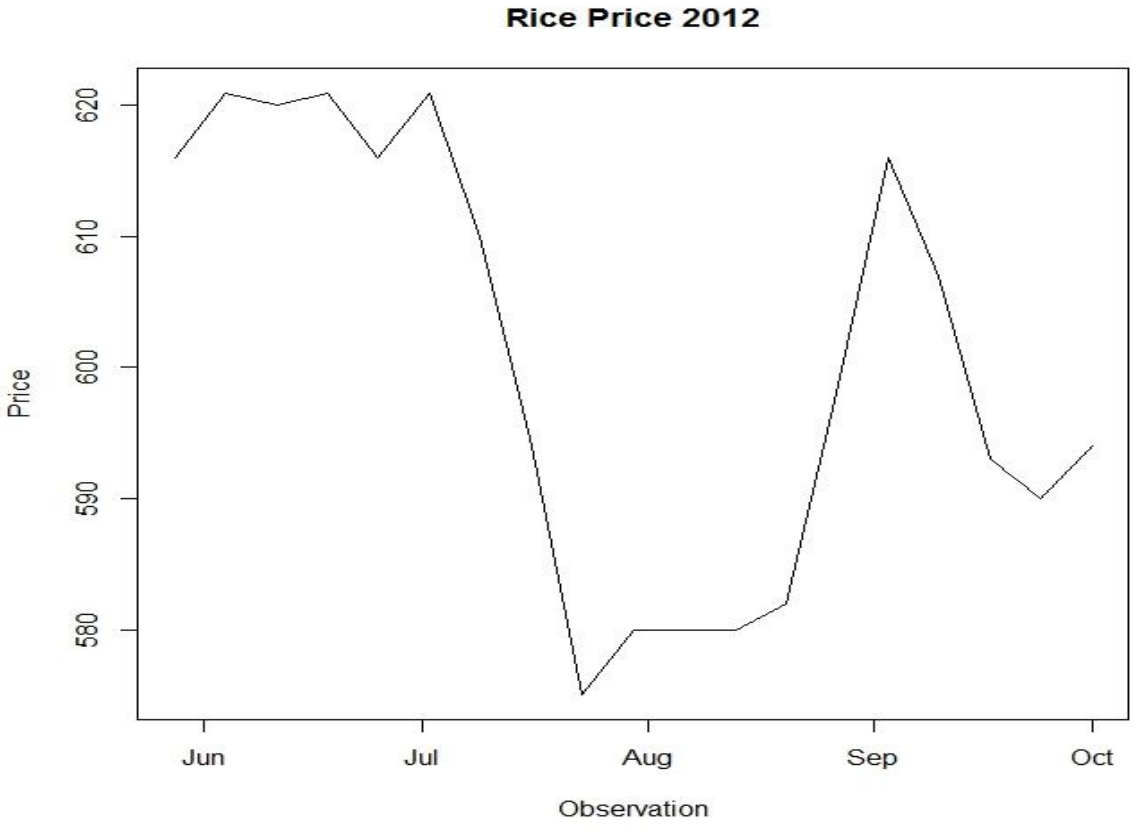
Having said that, it appears legit to suspect of compound sentiment data pollution caused by rice. Given 2010 and 2015 compound sentiment difference, by contrast this anomaly in sentiment can be seen as an ulterior proof of trend relevance in compound sentiment, namely can be said that if rice related sentiment is able to pollute data, therefore rice related sentiment is descriptive of rice price trend too.

This last intuition is however still an assumption that has to be tested, therefore it seems appropriate to cluster the articles commodity classes.

Once shown the linkage between compound sentiment and trend, it appears of interest to measure sentiment effect on prices and to study its predictive capabilities. For example, it could be tested if some predictive information can be found by looking at compound sentiment evolution in time, namely if a positive sentiment trend is related to a price decrease. If that's proven to be true, then price decreasing trend could be anticipated by a compound sentiment rise. However, since sentiment appears to be (almost) normally distributed, a sentiment trend can be described as a shifting of the whole distribution in the span of different time windows.

Unfortunately no evidence of linear relationship between compound sentiment and commodities returns has been found.


Rice Price 2011
Price
500
520
540
560
580
600
620
640
May
Jun
Jul
Aug
Sep
Oct
Observation


Rice Price 2012
Price
580
590
600
610
620
Jun
Jul
Aug
Sep
Oct
Observation