Sentiment EDA

Research objective is to underline possible relations between VADER extracted wheat related sentiment observations and International Grains Council Wheat Price Index movements. Different trend time windows will be compared to sentiment observations as in the previous analysis, but both positive and negative sentiment data have been added.

Data

IGC Price Index observations are composed by a series of 4181 daily prices, spanning from 03/01/2000 to 12/01/2016.

Sentiment observations are stored in the labelled sentiment matrix obtained by using the python file main.py.

In this version articles have been roughly labelled using as feature the presence of the words "wheat", "maize", "rice" and "soybeans" in each text. In each tokenized article the number of times that the words "wheat", "maize", "soybeans" and "rice" is counted and the article is labelled accordingly to the most used word. If two or more words are used an equal number of times and no other word is used more, then a "mixed" label will be applied; if no word is found, then a "general" label will be applied.
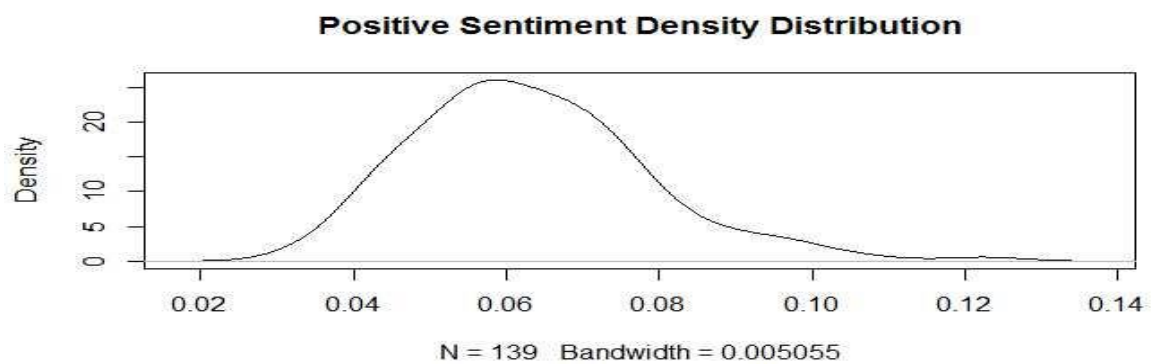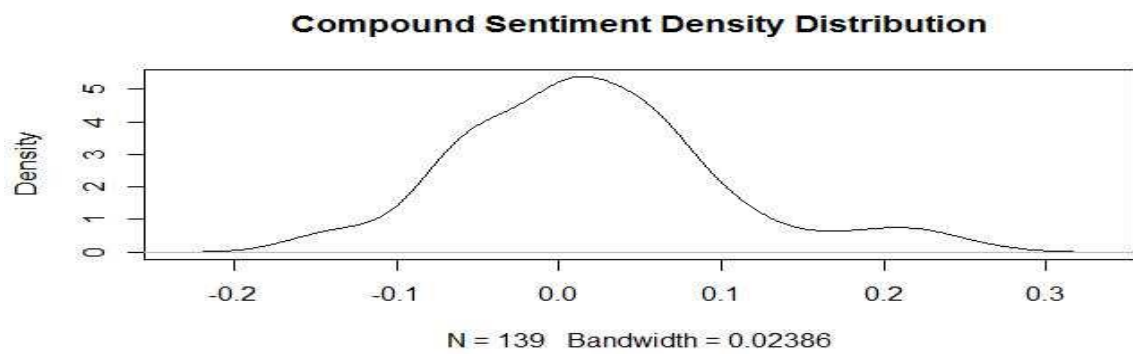
By analysing the results it appears that more than half of the articles are labelled as "general", therefore it has to be widened the set of words used and/or has to be selected a more robust labelling method. On the other hand it appears that a vast majority of non-general articles are labelled as "wheat". Moreover most of the articles labelled as "wheat" show a significantly high "wheat"word counter.
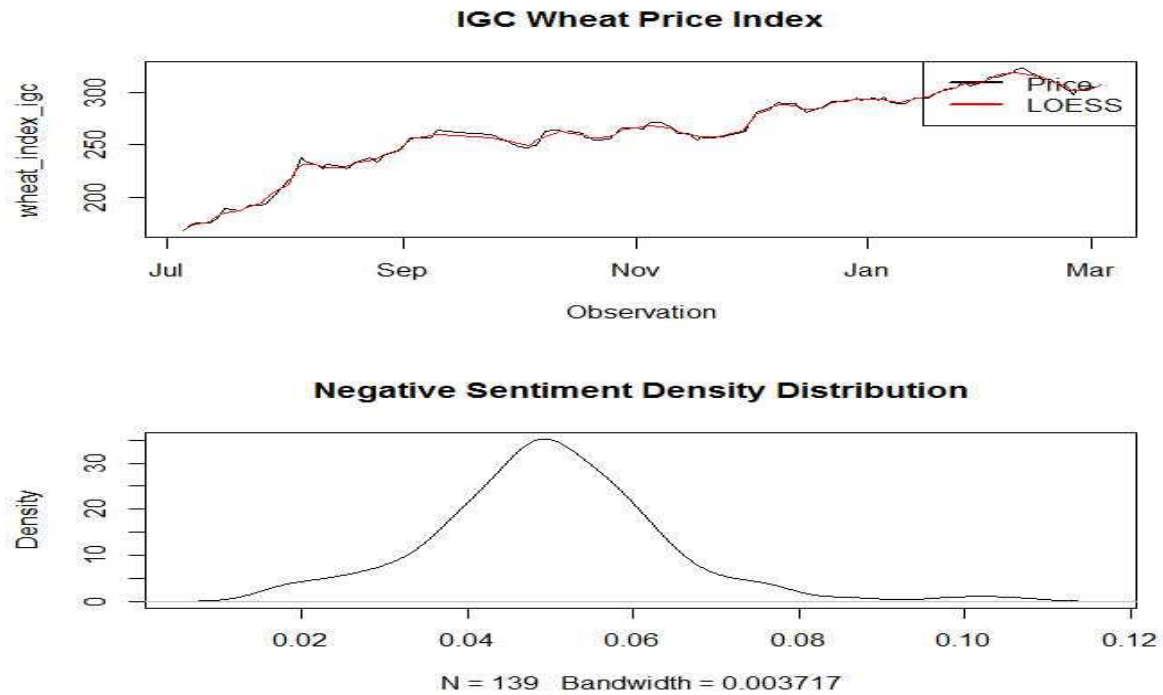
In particular, 14317 is the number of wheat labelled articles considered in this analysis.

Since neutral sentiment has been dropped off, it's important to underline that in this analysis positive, negative and compound sentiment are used, therefore the relation that links negative and positive sentiment with compound sentiment through neutrality is ignored.
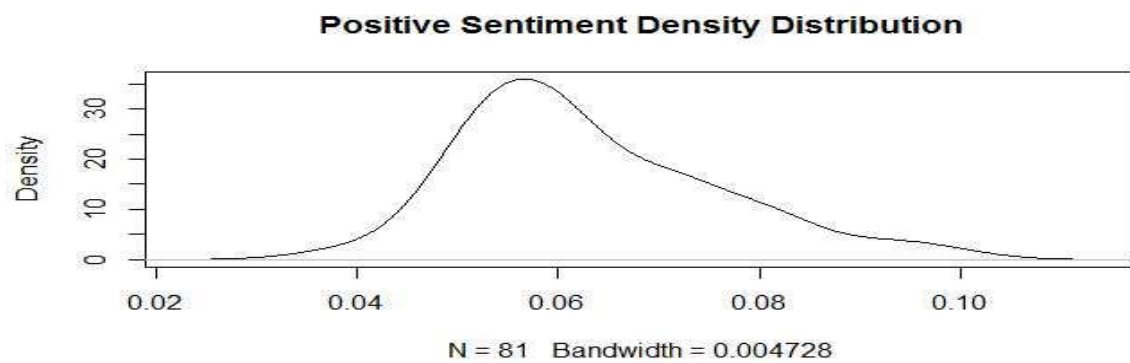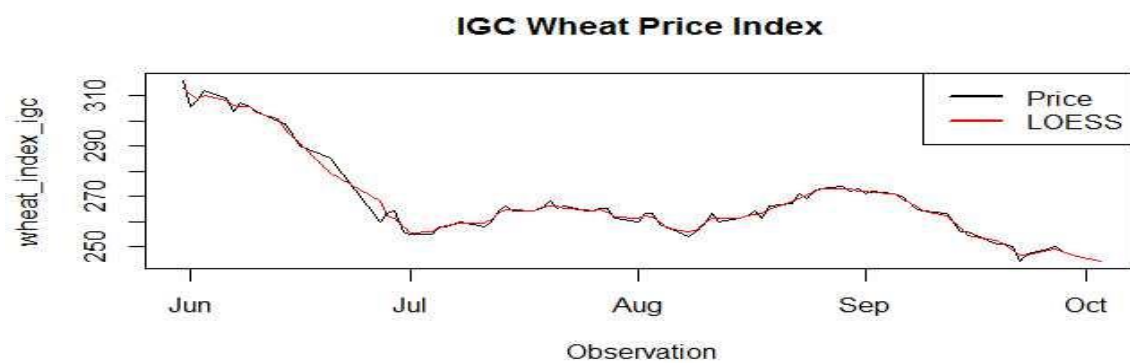
Results

Window 1: 2010/07/03 - 2011-03-03

IGC Wheat Price Index

Compound Sentiment Density Distribution

N = 139   Bandwidth = 0.02386

IGC Wheat Price Index

Positive Sentiment Density Distribution

N = 139   Bandwidth = 0.005055

IGC Wheat Price Index

Negative Sentiment Density Distribution
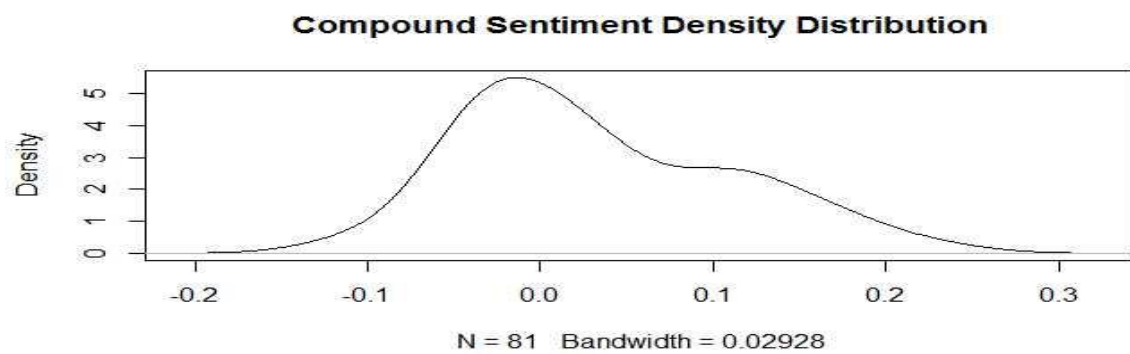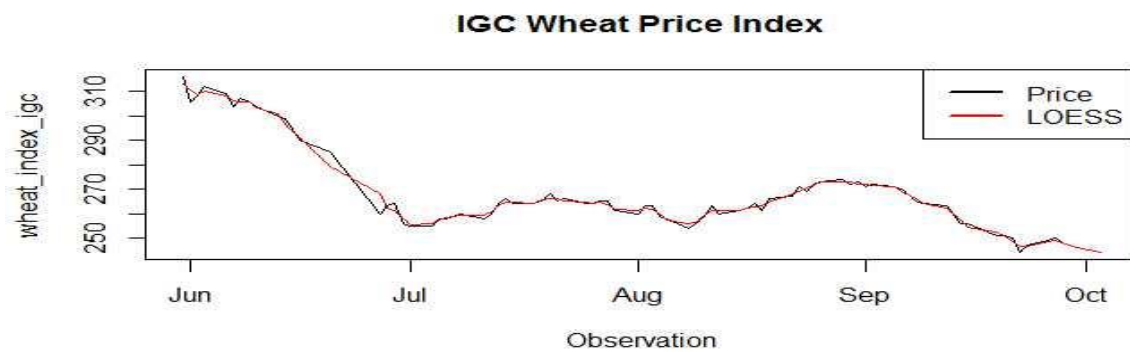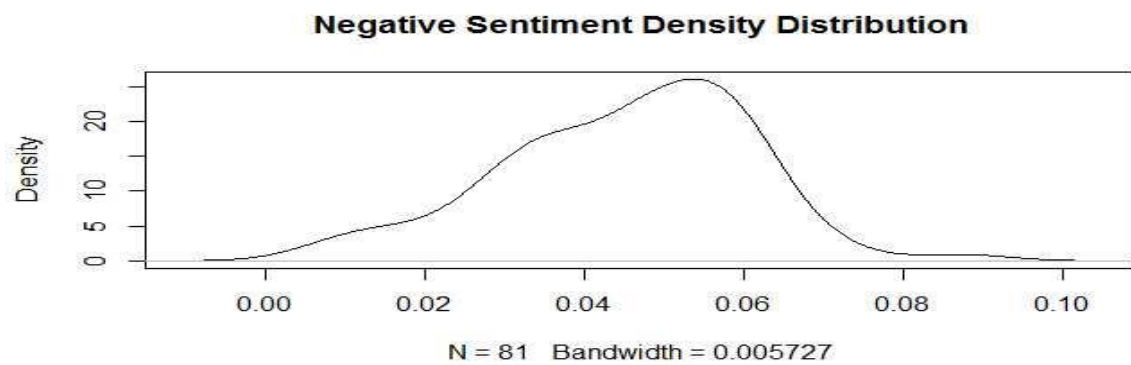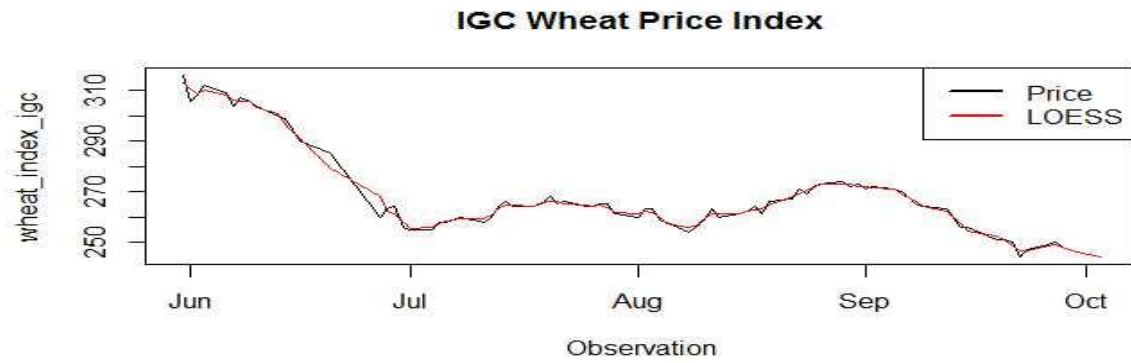

N = 139   Bandwidth = 0.003717

This period is an increasing trend one.

Compound sentiment median and mean are respectively 0.01219 and 0.01957. Positive sentiment median and mean are respectively 0.06197 and 0.06312. Negative sentiment median and mean are respectively 0.04916 and 0.04975.

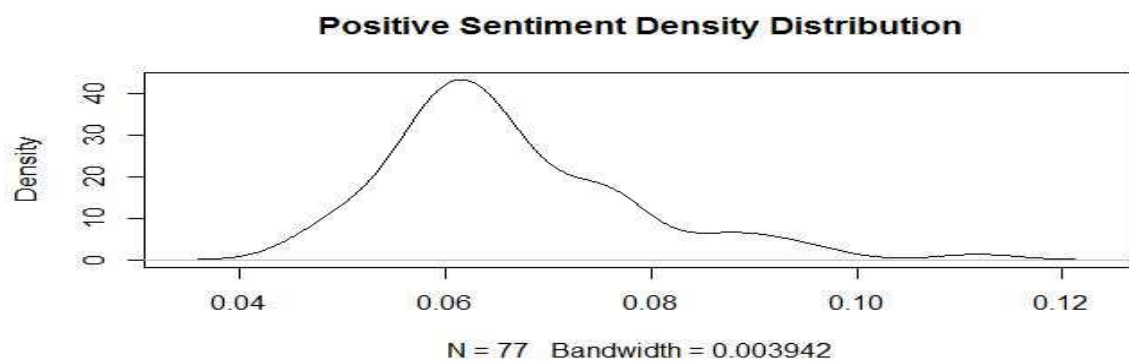## Window 2: 2011/05/28 – 2011/10/03

IGC Wheat Price Index

Compound Sentiment Density Distribution

N = 81   Bandwidth = 0.02928

IGC Wheat Price Index

Positive Sentiment Density Distribution

N = 81   Bandwidth = 0.004728

IGC Wheat Price Index

Negative Sentiment Density Distribution
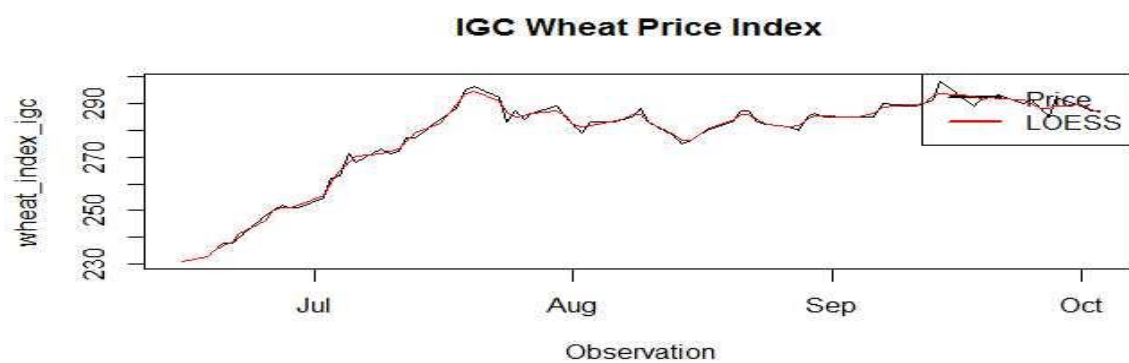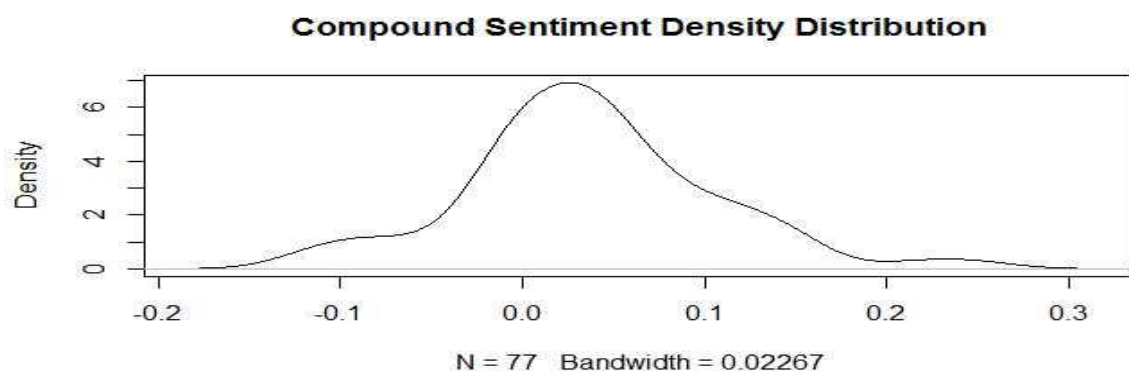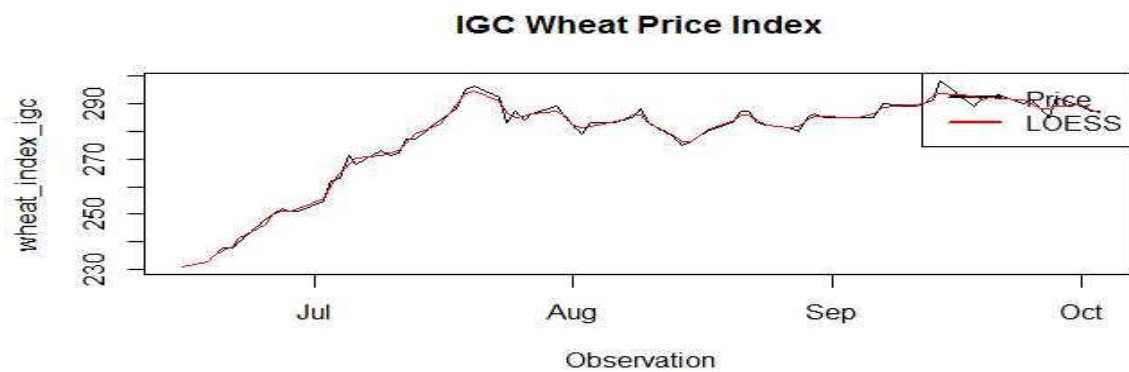

N = 81 Bandwidth = 0.005727

This period is a decreasing trend one.

Compound sentiment median and mean are respectively 0.01367 and 0.03234. Positive sentiment median and mean are respectively 0.05956 and 0.06306. Negative sentiment median and mean are respectively 0.047950 and 0.045290.

Window 3: *2012/06/15 – 2012/10/03*

IGC Wheat Price Index

Compound Sentiment Density Distribution

N = 77   Bandwidth = 0.02267

IGC Wheat Price Index

Positive Sentiment Density Distribution

N = 77   Bandwidth = 0.003942

IGC Wheat Price Index

Negative Sentiment Density Distribution
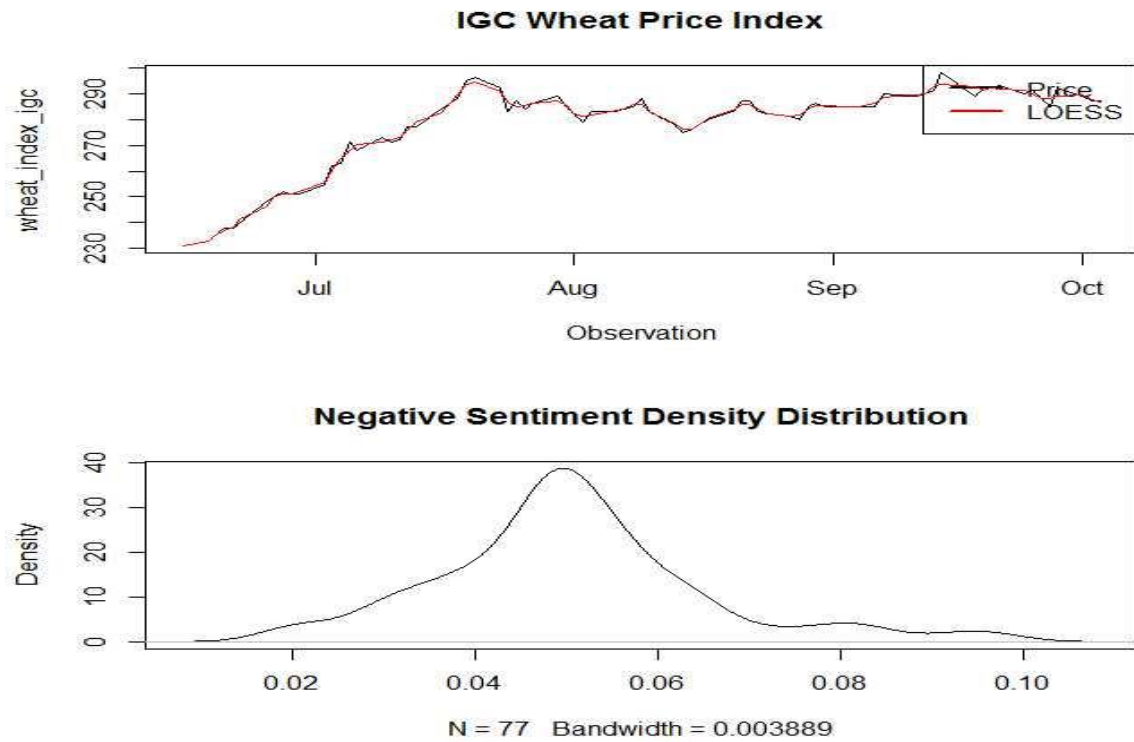

N = 77   Bandwidth = 0.003889

This period is an increasing trend one.

Compound sentiment median and mean are respectively 0.030380 and 0.036920.

Positive sentiment median and mean are respectively 0.06334 and 0.06630.
Negative sentiment median and mean are respectively 0.04973 and 0.05047.

IGC Wheat Price Index

Compound Sentiment Density Distribution

N = 104   Bandwidth = 0.01333

IGC Wheat Price Index

Positive Sentiment Density Distribution

N = 104   Bandwidth = 0.003054

IGC Wheat Price Index

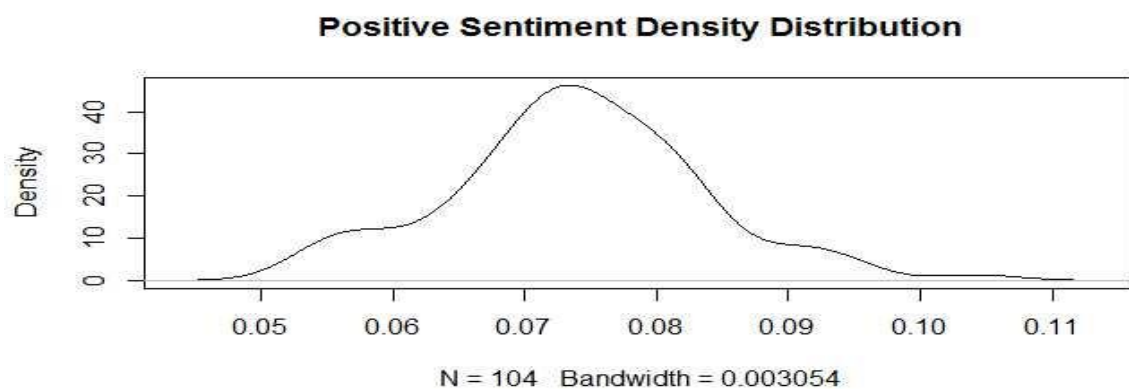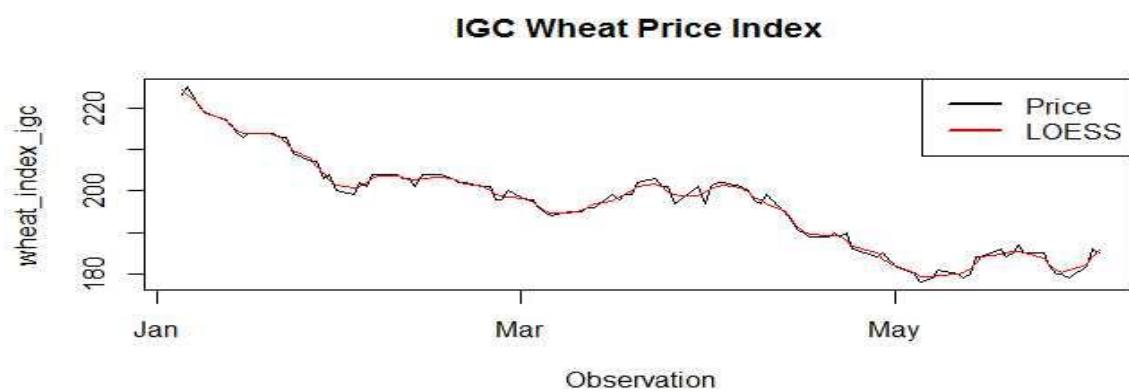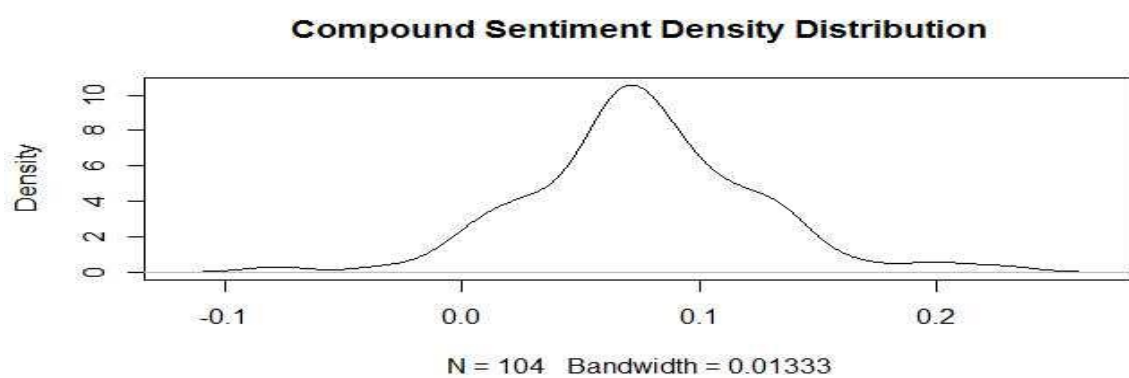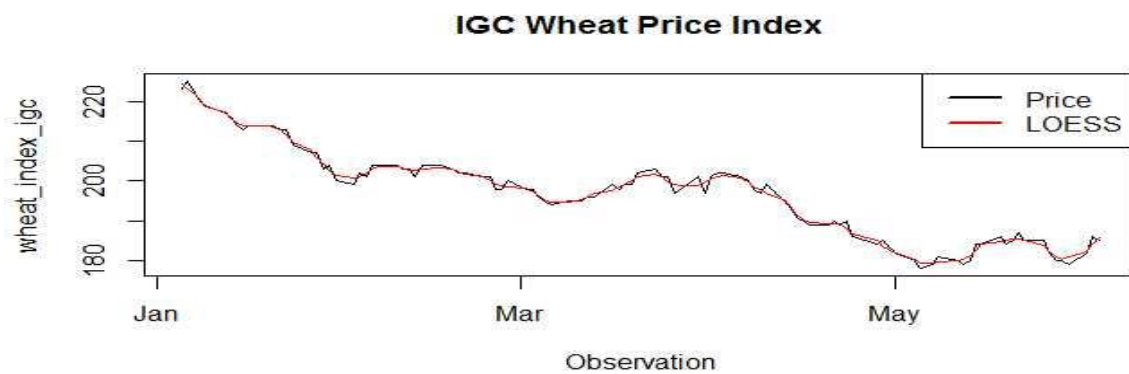## Negative Sentiment Density Distribution

N = 104 Bandwidth = 0.003132

This period is a decreasing trend one.

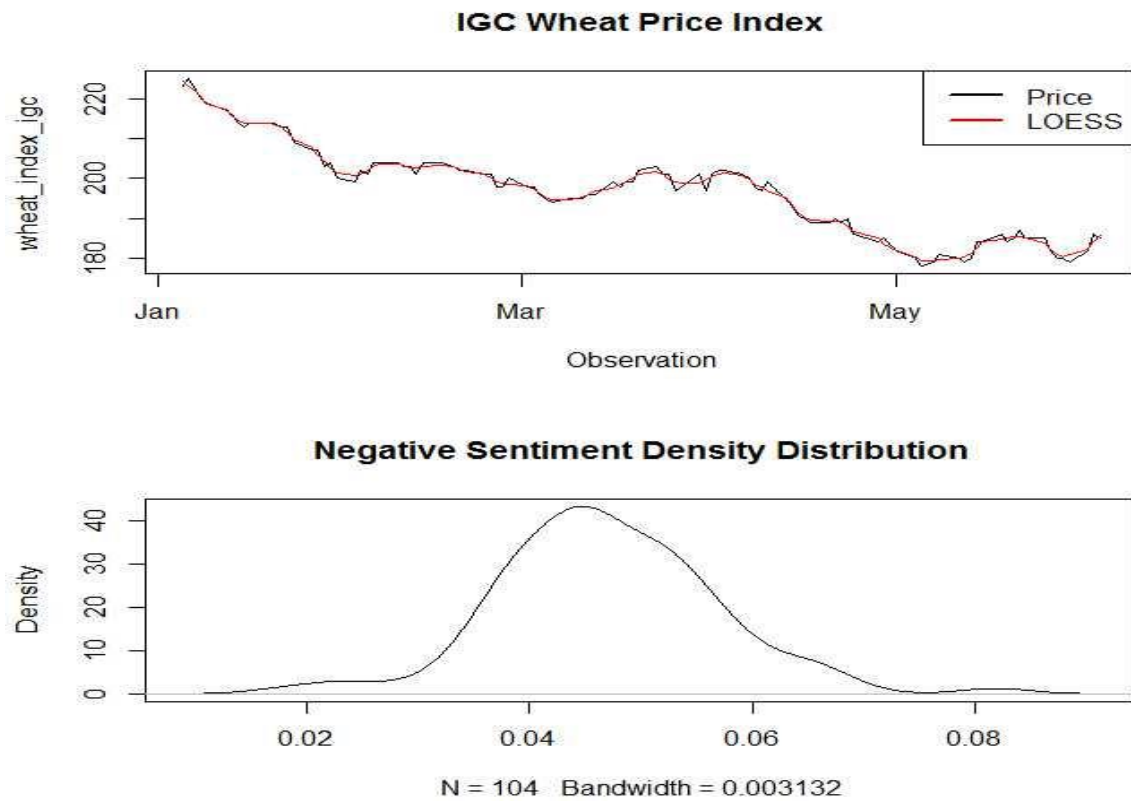Compound sentiment median and mean are respectively 0.07268 and 0.07614. Positive sentiment median and mean are respectively 0.07396 and 0.07396. Negative sentiment median and mean are respectively 0.04629 and 0.04699.

Conclusion

	$\bar{S}_{me}$	$\bar{S}$	$\bar{S}_{me}^{+}$	$\bar{S}^{+}$	$\bar{S}_{me}^{-}$	$\bar{S}^{-}$	Price Trend
Window 1	0.01219	0.01957	0.06197	0.06312	0.04916	0.04975	Increasing
Window 2	0.01367	0.03234	0.05956	0.06306	0.047950	0.045290	Decreasing
Window 3	0.030380	0.036920	0.06334	0.06630	0.04973	0.05047	Increasing
Window 4	0.07268	0.07614	0.07396	0.07396	0.04629	0.04699	Decreasing

$\bar{S}_{me}$: *compound sentiment median* ; $\bar{S}$: *compound sentiment mean*
$\bar{S}_{me}^{+}$: *positive sentiment median* ; $\bar{S}^{+}$: *positive sentiment mean*
$\bar{S}_{me}^{-}$: *negative sentiment median* ; $\bar{S}^{-}$: *negative sentiment mean*

From the table above it's possible to underline an increasing trend in compound sentiment in both median and mean terms, driven by a positive trend in positive sentiment (mean terms): unfortunately these movements don't seem to have connections with price trends, since they show different signs.

However it has to be noticed that negative sentiment decreases when price trend is decreasing and increases when price trend increases. This finding is somehow consistent with previous research assumption of inverse relation between sentiment and food commodities prices, namely high prices are positively correlated with negative sentiment given their linkage with food crises. Further research on this assumption has to be conducted.

Clearly it's underlined the necessity of a better classification method for the articles, since using this basic classification method many articles are left without a label. Moreover it can be of interest to have an idea of the context from which the sentiment is extracted (trade related or equity related articles for example) as to extract from the articles just the sentiment bearing sentences in order to avoid as much as possible data pollution.

Moreover it is repeated the necessity to work on the dictionary since VADER analysis tool has been used just for test purposes and it has to be refined to better capture sentiment from real commodities related articles. In fact in the previous research it has been underlined that this tool it's designed to capture twitter tweets sentiment, therefore it doesn't fit well when dealing with neutral economic statements. Given that sentiment negativity seems to follow the price trands, probably there is a bias on the positivity magnitude measure, namely VADER analysis tool tends to over-estimate articles positivity. However also in this case further research is needed.