TRINITY COLLEGE DUBLIN
COLÁISTE NA TRÍONÓIDE, BAILE ÁTHA CLIATH

Multiresolution Analysis of the Relationship between Markets and Sentiment

Oisín BROGAN

Supervisor: Dr. Khurshid Ahmad

Submitted in partial fulfilment of the requirements for the
degree of M.A.I. (St) to the University of Dublin, Trinity College, April 2014

Master in Engineering (MAI) **09368400**
Electronic and Computer Engineering **April 2014**

Declaration of Authorship

I, Oisín Brogan, declare that the following dissertation, except where otherwise stated, is entirely my own work; that it has not previously been submitted as an exercise for a degree, either in Trinity College Dublin, or in any other University; and that the library may lend or copy it or any part thereof on request.

Signed: ______________________

Date: ______________________

Summary

Fundamentally, this thesis is about trying to explain the variance in market prices better. This variance present in the prices is not unexpected. Markets are part of a complex economic system, and respond and reflect a large amount of different information sources. Researchers can try to explain this variance using an information source from outside the market, instead of using time series analysis.

Sentiment analysis has reached the stage where it can be expressed as a time series, and as such incorporated into financial models. In particular, this work uses the sentiment analysis system Rocksteady, developed here in Trinity College, University of Dublin. (Ahmad et al., 2011) Rocksteady can work with online financial newspaper archives and news blogs to create a market specific sentiment times series. It is the output of Rocksteady that this thesis uses in it financial models.

The techniques used to model the possible relationship between sentiment and market data were Vector Autoregression (VAR) and Granger Causality tests. VAR models essentially create a linear equation for each of the variables in the model, using the lagged values of all variables in the model for the equation for each variable. The Granger causality test tests to see if a change in one variable occurs before a change in another variable, looking for changes as cause and response relationships. The Granger test fits well with VAR models, as in it based on many of the same assumptions (e.g. linear models, stationarity)

The final part of this work is wavelet analysis. The wavelet transform adds frequency information to the signal, but does not lose all time resolution, instead balancing the trade-off between time and frequency resolution. This allows the wavelet transform to show when in time events of certain frequencies occurred.

This behaviour allows the multiresolution analysis of a time signal. The time series can be broken down into different time series that reflect activity of different timescales. So market data can be split into daily and weekly and yearly trading, each time series orthogonal to the other. This allows sentiment and market returns to examined on different timescales, and how the relationship might change with timescale. This approach has been extensively used in economic research, uncovering relationships over timescales that remain hidden if one only considers the sum of the different activities i.e. the original time series.

That is a layout of the work this thesis attempts. The oil market was chosen as an initial test-bed. It was felt a commodity market would be preferable to a stock market,

as there it would be less likely to see the extreme volatility, and more likely to show a relationship.

Where this thesis seeks to make a contribution is a closer look at the relationship between sentiment and market. Combining the wavelet analysis discussed in section 2.1 we take a robust approach to seeing if the relationship actually exists, and if so what shape it takes and across what time-frames. This thesis sits well within the research done before, and attempts to add to the growing knowledge of how news effects markets.

The main findings of the thesis are that at smaller timeframes, the daily and weekly trading, sentiment does play a role in explaining returns. At timeframes longer then a forenight, we found no evidence for sentiment having an influence on the market. While we expected there to be a transition between short and long timeframes, that traders operating on a forenightly basis were ignoring sentiment was a little surprising.

The other main finding related to the delays, with a over a week of delay in sentiment appearing and the market responding. It seem that even daily traders take time to process sentiment before acting upon it. Finally, the need to decompose the time series was made evident that undecomposed, the signals showed that returns cause sentiment.

The method has proven to be able to uncover relationships between senitment and markets, and sugestions for future work include using different kinds of sentiment and exploring other types of markets.

Acknowledgements

I have to acknowledge the work of my supervisor, Dr. Khurshid Ahmad. He was an ideal supervisor - knowing when I was struggling and needed his direct help or if I just needed more time to figure a problem out on my own. He was also very understanding when I sometimes had to put other assignments before this project work.

I'd also like to thank my friends and family for both putting up with my stress over this year and reminding me I sometimes needed a break from the work. Special mention must go to my father, Alan, who proof read my thesis under quite some time pressure.

Contents

Chapter 1

Introduction

This thesis was completed for an M.A.I. for Electronic and Computer Engineering. Both areas of engineering covered in the degree are part of this thesis. Through one lens, the thesis is fundamentally about signal processing, filtering and comparing different signals. Through the other, this is about computer modelling and automating a task that used to only be done by humans - sentiment analysis of text. Both are valid interpretations of the work, and the hope is that this work will be relevant to both areas.

1.1 Motivation

Sentiment analysis has in recent years become a well developed field reporting considerable success. While there is certainly still work to be done, the results from the field have become increasingly reliable and accepted.

Coupled with this establishing of sentiment analysis as a respected field is an abundance of publicly available sentiment heavy texts on the internet. Of course, the internet is older than the last five years, but with the emergence of social media including Twitter, Facebook, blogs, message boards and forums, as well as the proliferation of the consumer review driven by Amazon.

Given the time the research area came of age, it is unsurprising that many of the applications of sentiment analysis beyond academia are focused around social media and the internet (Feldman, 2013). Classifying consumer reviews as positive or negative, or sorting tweets about politicians running for election were the subject of this work.

However, the applications are much broader than classifying tweets or Facebook statuses. As laid out in section 2.2, using sentiment analysis to help financial models is a growing body of research, with the more sophisticated methods now available seeing a greater success in this area.

The idea that market sentiment would have an effect on market movements seems intuitive, and also seems to align with our discussion in section 2.3 about irrational traders and the influences outside the market. The sentiment behind financial news seems like the perfect candidate for an influence on these irrational traders.

Where this thesis seeks to make a contribution is a closer look at that relationship. Combining the wavelet analysis discussed in section 2.1 we take a robust approach to seeing if the relationship actually exists, and if so what shape it takes and across what time-frames. This thesis sits well within the research done before, and attempts to add to the growing knowledge of how news effects markets.

1.2 Outline of Work

Fundamentally, this thesis is about trying to explain the variance in market prices better. Market prices show a high auto-correlation, with the price today largely depending on the price yesterday. However, even ignoring the shocks spikes and falls that are part of stock and commodity markets, prices still show large variance around the high auto-correlation.

This variance can be clearly seen when working with the returns, the difference in price data between one day and the next, rather then working directly with prices. This variance present in the prices is not unexpected. Markets are part of a complex economic system, and respond and reflect a large amount of different information sources.

Market price data is clearly a time series, and time series analysis has been used extensively in the study of this area. Random walks, Holt-Winters smoothing, autoregression models and various other techniques have to used to some success in modelling and forecasting financial data. None of these methods fully capture the behaviour of the data, there remains improvements to be made.

The problem with time series analysis is only the market prices themselves are used in the models. (Black, 1986) makes a good case as to why a financial data cannot be regressed purely on itself - there are too many other factors at play. Contrary to the Efficient Market Hypothesis, market prices do not reflect all relevant information instantaneously. (see 2.3 for a larger discussion on this.)

As mentioned in the motivation, sentiment analysis has reached the stage where it can be expressed as a time series, and as such incorporated into financial models. In particular, this work uses the sentiment analysis system Rocksteady, developed here in Trinity College, University of Dublin. (Ahmad et al., 2011) Rocksteady can work with online financial newspaper archives and news blogs to create a market specific sentiment times series. It is the output of Rocksteady that this thesis uses in it financial models.

The techniques used to model the possible relationship between sentiment and market data were Vector Autoregression (VAR) and Granger Causality tests. Both of these techniques are used modelling economic variables, as modeling the relationship between different variables plays a much larger role in economics then in financial modelling.

VAR models essentially create a linear equation for each of the variables in the model. The difference between autoregressive models is VAR used the lagged values of all variables in the model for the equation for each variable. The Granger causality test tests to see if a change in one variable occurs before a change in another variable, looking for changes as cause and response relationships. The Granger test fits well with VAR models, as in it based on many of the same assumptions (e.g. linear models, stationarity)

The final part of this work is wavelet analysis. The wavelet transform, like the Fourier transform, adds frequency information to the signal at the expense of resolution of the signal in time. Unlike with Fourier, the wavelet transform does not lose all time resolution, instead balancing the trade-off between time and frequency resolution. This allows the wavelet transform to show when in time events of certain frequencies occurred.

This behaviour allows the multiresolution analysis of a time signal. The time series can be broken down into what are known as details - different time series that reflect activity of different timescales. So market data can be split into daily and weekly and yearly trading, each time series orthogonal to the other. This allows sentiment and market returns to examined on different timescales, and how the relationship might change with timescale.

This approach has been extensively used in economic research, uncovering relationships over timescales that remain hidden if one only considers the sum of the different activities i.e. the original time series.

That is a layout of the work this thesis attempts. The oil market was chosen as an initial test-bed. It was felt a commodity market would be preferable to a stock market, as there it would be less likely to see the extreme volatility, and more likely to show the relationship. However, there is no reason this analysis cannot be applied to stock markets, and this topic is a suggestion for future work.

A final note on terminology. Given the mixed background of this thesis, sometimes different terms are used to mean the same thing. In particular, ‘time series’, ‘process’ and ‘signal’ are all used interchangeably. They mean slightly different things in different contexts, but within this work, they all refer to a data series where each datum has an associated time of occurrence, in our case a date.

1.3 Project Objectives

- *Multiresoltuion Analysis*

  The first objective is simply to breakdown the different time series, both sentiment and returns on oil, into their wavelet details. This will allow us to use the modelling techniques on compare the two variables on different timescales.

- *Test for relationships between sentiment and returns using Granger*

  Once the two processes have been decomposed, we want to test for a relationship between the two variables on corresponding levels. The initial test for a relationship (i.e. sentiment causing returns, or vise versa) will be the Granger Causality test.

- *Model any relationship found using VAR*

  If the Granger test shows a relationship, we want to model that relationship using a VAR model. The Granger test only shows a relationship exists - we use a VAR model to give a more detailed picture of the relationship, showing the delays and direction (positive or negative) in the relationship.

- *Learn the techniques used by Rocksteady*

  As mentioned, Rocksteady is a system of sentiment analysis developed here in Trinity. As part of this project I would like to become familiar with the techniques used by Rocksteady, and not leave it as a black box for me. If time allows, I may try to develop something for the system, though given the amount of other work, this may prove unlikely.

- *Learn to use various financial software packages*

As part of this thesis, I want to improve my own skills in financial modelling. Where possible, I intended to use different software packages and tools to implement different aspects of the work. Two that might prove useful are all the financial tools implemented for use in Matlab, and the Gretl Open-Source statistics package.

1.4 Layout of Thesis

In Chapter 2, Literature Review, the state of research of the different techniques employed in this work is discussed. Section 2.1 outlines the different properties that make the wavelet transform well suited to economic and financial study, before looking at research in finance that has already taken advantage of those properties.

Section 2.2 turns to Sentiment Analysis and the state of research in that area. It lays out the different techniques and methods developed so far, before again turning to research done in the area of finance and economics that has used sentiment analysis already.

The final section in Chapter 2, 2.3 looks at financial and economic theory, as well as behavioural science, to try find some possible mechanisms for out intuitive sense that sentiment should influence markets.

Chapter 3 explores the method used by the work, and the theory behind the different techniques and models. Section 3.1 discusses wavelets and the decomposition analysis. It lays out the theory behind the wavelet transform and why it can be used to decompose a signal into different timeframes.

The following section in Chapter 3, 3.2 looks at the sentiment analysis tools used by the Rocksteady system, as well exactly how the input data for the system was curated.

Section 3.3 lays out the theory and implementation of the VAR models, as well as exploring the underlying assumptions and limitations of the model. The variations of the VAR model, some of which were developed to deal with those limitations, are also discussed. The final section in the chapter 3.4 explains the theory and the implementation of the Granger Causality test.

Chapter 4 is the Results chapter, where all the results of the thesis are presented. The first section 4.1 describes exactly the data used by the project, before the following section 4.2 displays the results of the wavelet decompositions of the various time series.

The following section 4.3 lays out the different relationships discovered and the models fitted to those relationships. It goes through each of the different data-sets used in turn:

original sentiment and decomposed returns, both sentiment and returns decomposed and decomposed moving averages of sentiment and returns.

A brief discussion of follows each results of the Granger tests and VAR models, but the final section, 4.4 is a through discussion of the results and their implications.

The final chapter is the Conclusion, which first 5.1 summarizes the whole work, laying out what was done and how it was done. The next section 5.2 summarizes the main findings of the thesis. The final section in the thesis 5.3 gives ideas for possible future work based on this thesis, including items we would have liked to have done if time allowed.

Chapter 2

Literature Review

2.1 Wavelets and Financial Series

In an economic system, there is a large number of participants all with decentralized interactions and a wide range of different goals and ways to interact. These participants' actions effect the whole system but they often make decisions based only on local (often in time and space) information. And all the while the system as a whole is constantly changing, as participants enter and leave the system and it reacts to external stimuli.

As such, it is unsurprising that a number of different tools of analysis and inference have grown out of this field. One of the tools, wavelets and their associated filters certainly do not solve all of economists' problems. To quote directly from (Ramsey, 1999), page 2, talking about the ideal tool of economic analysis

> Key issues to be considered by the putative analyst are robustness of procedure to erroneous assumptions, flexibility of regression fit to deal with imprecise model formulations, the ability to handle complex relationships, efficiency of estimators to be able to make useful distinctions on a few data points and simplicity of implementation. This, the econometrican's Holy Grail, may be impossible to achieve, but is nevertheless a worth while overall objective

2.1.1 Wavelet Properties

However, wavelets to provide a number of useful and interesting properties, especially towards dealing with the third feature of Ramsey's Holy Grail. We can identify three key properties here:

- The ability to analyse non-stationary time series.
- Localisation in time
- Orthogonal time-scale decomposition of the data

Handling non-stationarity

The ability of wavelets to handle non-stationary processes comes from their fundamental properties. As their name suggests, wavelets are waves with a finite support, with values quickly returning to zero. (Gençay et al.) This means that as the descriptive statistics of a time series change with time, different wavelets will change with them without being influenced by the series' earlier behavior.

Early work in this area started in biology, (von Sachs and MacGibbon, 1997), which was an extension of an earlier paper (Johnstone and Silverman, 1997).

This ability is of obvious benefit to modelling and analyzing financial data, well known to be non-stationary processes. The wavelet transform allows the meaningful extraction of lower frequency events without undue influence of high frequency events that die away quickly, and don't effect the whole series. This is not a property of the Fourier transform, for example.

Localisation in time

The finite support of wavelets also give the wavelets ability to express frequency information localized in time. Non-stationary processes can feature events that appear for only a certain length of time before dying away. Wavelet's finite support allow that event to be captured in the transform, as discussed above. However, in addition, the wavelet retains some information as to when this event occurs.

No transform can give perfect resolution in time and frequency. (Hubbard) Wavelets adapt the trade-off between time and frequency resolution, with greater time resolution as frequency increases. As these temporary events are more likely to be of a higher frequency, wavelets manage this resolution trade-off well.

The ability to point to when events of certain frequency occur is a boon to economists. It allows the identification of shocks and structural breaks in the system.

Time-scale decomposition

An important property is that wavelets of different frequencies do not affect each other. Each time-frame or scale is isolated from the other. (Mallat, 1989) This allows a multi-resolution analysis that separates a time series into different series, each explaining the activity of the time series on different timescales. Wavelet transform allow these time series to differ by a factor of two - one series explaining every day activity, another for every 2nd day, every 4th day, every 8th day etc. These different timescales are referred to as levels and/or scales in wavelet literature.

The multilevel view wavelets provide is important to economics is it allows work to acknowledge what has long been known in the financial theory - that economic signals are the summation of many different effects and events that operate on many different levels.

The price of a stock is influenced both by the daily trading as well as so called "market fundamentals". Traders work on different levels as well, with day traders worried and acting upon much different information then someone trading futures in commodity markets. External stimuli to the economic system also operate on different time-scales: compare the after effect of the Eyjafjallajökull to 9/11, or a long term civil war.

2.1.2 Applications of Wavelets in Finance

These properties of wavelets make them well suited to the area of finance, and many papers have been published using and extending the techniques. Here we lay out three of the largest areas of use of wavelets in economics.

De-noising

Wavelets allow de-noising, as opposed to noise smoothing. When dealing with noisy signals, the oft simple formula used as a starting point is separating the observed signal into the actual signal plus some added noise:

$$y_t = f_t + \varepsilon_t \tag{2.1}$$

Often, the unstated assumption here is that the signal, f_t, is smooth, unlike the noise. So, in order to extract the signal for the observed data some kind of smoothing procedure is used. However, any smoothing will be counter-productive if the underlying signal itself is not smooth, or contains discontinuities. Much economic data is considered of this type.

What is required instead is de-noising. Here, the principal idea is to set some noise threshold - variations in the data below this threshold are considered noise, while variations greater are the signal. With a wavelet transform, we can analyze the wavelet coefficients, which show variation in data, to set this noise threshold, and shrink the coefficient accordingly, removing the noise.

We can shrink the wavelet coefficients using one of several formula. Soft shrinkage:

$$\delta_s(w) = \begin{cases} 0 & \text{if } |w| \leq c \\ sgn(w)(|w| - c) & \text{if } |w| > c \end{cases} \tag{2.2}$$

and hard shrinkage

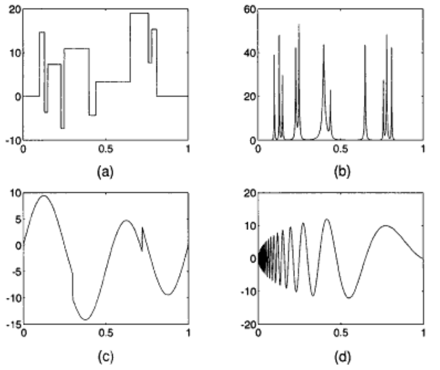
$$\delta_h(w) = \begin{cases} 0 & \text{if } |w| \leq c \\ w & \text{if } |w| > c \end{cases} \tag{2.3}$$

or Brieman's Garrote (Breiman, 1995)

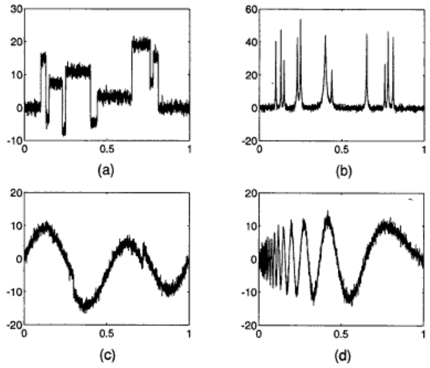
$$\delta_s(w) = \begin{cases} 0 & \text{if } |w| \leq c \\ w - \frac{c^2}{w} & \text{if } |w| > c \end{cases} \tag{2.4}$$

where c is the noise threshold, and w is the wavelet coefficient.

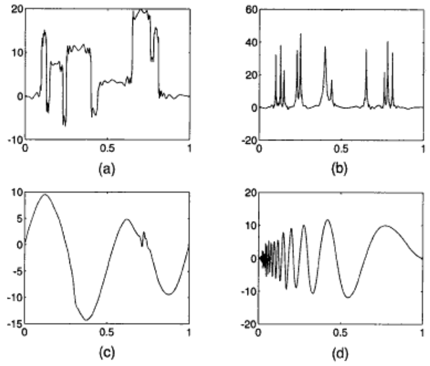
Much of the early work done here was done by these two papers (Donoho and Johnstone, 1998) and (Donoho and Johnstone, 1995) An additional benefit of wavelets is they allow this process to work on different scales, with different thresholds set for different scales. This is especially useful, as noise in markets is associated with high frequencies, so it seems obvious the wavelet coefficients of the lower scales would contain more of the noise. In fact, pragmatically, the first scale of wavelet coefficients are usually entirely associated with noise.



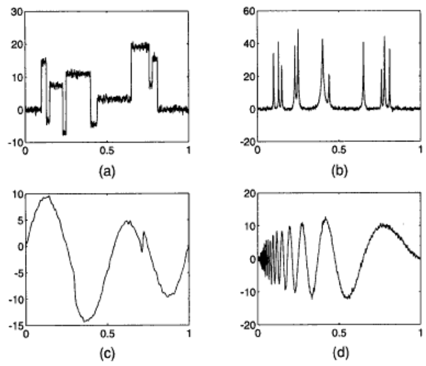
(a) Original signals, all of which display non-smooth behavior. (a) Blocks, (b) Bumps, (c) Heavsine and (d) Doppler



(b) Signals with added noise



(c) De-noised using wavelets



(d) De-noised using spline smoothing

The benefits of this approach can be seen in the diagrams, taken from (Donoho et al., 1995).

Seasonalities

As already mentioned, economics studies a system that operates on many timescales, with different processes making different decisions based on different information integrating with each other. The ability of wavelets to decompose the aggregate activity that is the fluctuations of prices and stocks into different time-frames is of great use in economics.

A common problem noted in the literature is how the presence of seasonalities can obstruct basic analysis of a time series. This can be illustrated with a simple example. We can simulate a very simple signal, $y_t = 0.95y_{t-1} + \varepsilon_t$. We expect this function to have very high auto-correlation at low lags, while dropping off exponentially as the lag increases. This is confirmed by the diagram below.

However, if we alter the equation to add in seasonalities, we lose this high auto - correlation.

$$y_t = 0.95y_{t-1} + \sum_{s=1}^{4} \left[3sin(\frac{2\pi t}{P_s})\right] + \varepsilon_t \tag{2.5}$$

with $P_1 = 3, P_2 = 4, P_3 = 5$ and $P_4 = 6$. Instead, the auto-correlation function shows peaks corresponding to multiples of 6.

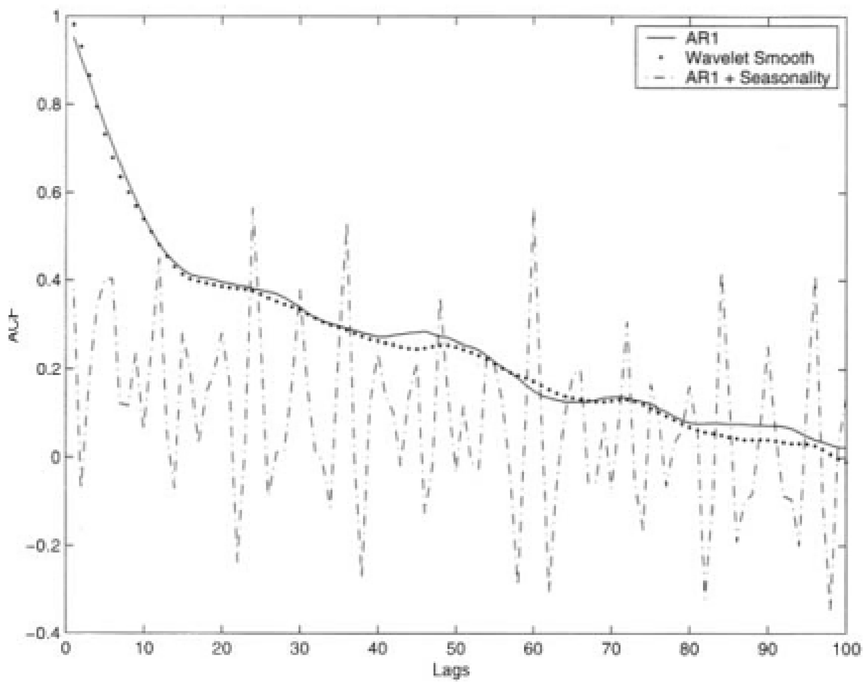


FIGURE 2.1: The high auto-correlation present in the signal is lost with added seasonalities. Wavelets can recover the signal to reveal the high auto-correlation once again

Using wavelets, we can create a wavelet smooth (see 3.1.4 for further explanation, but breifly a smooth is the time series constructed using only the lower frequencies i.e. higher scales of a wavelet decomposition) With this wavelet smooth, we can remove these higher frequency seasonalities, leaving just the fundamental signal again. With this approach, the auto-correlation of the signal reappears, as the diagram above shows.

Decomposition

Another application of timescale decomposition is the ability to examine economic relationships across different timescales. (Ramsey and Lampart, 1997) were the first to take this approach, and really set the example many following works, including this thesis, followed.

The author studied two classic economic relationships in two papers:

- Permanent Income; the hypothesis that temporary fluctuations in income have little to no effect on consumer spending behavior, but consumption is instead only a function of permanent income.
- Velocity, the relationship between income and money in an economy.

For ease, we focus on the paper dealing with income and consumption here. This relationship had been and still is to an extent the subject to major debate, with the exact nature of the relationship uncharacterized.

(Ramsey and Lampart, 1997) proposed the use of wavelets as a means of gaining insight to the relationship. The timescale decomposition would play a key role in the study of this relationships, as it was already posited that the relationship might change depending on the time frame involved in making the decision. Indeed, that was the original idea of Friedman when he proposed permanent income.

The paper used monthly data from the years 1960-1994, and decomposed the signals into six scales - five details and the resulting smooth. They then used Granger causality tests to examine the relationship between the economic variables. (see 3.3 for details on the Granger Causality test). Their final results are presented below.

	Results	Null Hypotheses	
		M1 $\not\Rightarrow$ NP1	NP1 $\not\Rightarrow$ M1
D6 (5 lags)	feedback	0.000	0.000
D5 (20 lags)	feedback	0.000	0.000
D4 (19 lags)	M1 $\Rightarrow$ NP1	0.000	0.695
D3 (17 lags)	M1 $\Rightarrow$ NP1	0.023	0.193
D2 (23 lags)	M1 $\Rightarrow$ NP1	0.000	0.293
D1 (14 lags)	NP1 $\Rightarrow$ M1	0.089	0.000
log diff. (12 lags)	inconclusive	0.892	0.186

Table 2.1: Table showing the results of Granger Causality tests on different wavelet levels. The values shown are probabilities of keeping the Null Hypotheses.

Note here how the relationship changes for different timescales. At lowest frequencies, there's feedback between the two variables. The monthly relationship shows that NP causes M, but for all other scales the relationship was reversed. Note that, without wavelet decomposition the interactions among the differing relationship are lost, leaving them unexaminable.

Ramsey also uncovered that the delay between money and income were functions of the state of the system, not constant as previously assumed. This was a major breakthrough in economics, showing that the timing of the purchase was as much an economic decision as the purchase itself. (Ramsey and Lampart, 1997)

But (Ramsey and Lampart, 1997) is not the only example of this approach. In (Gençay et al., 2001) and (Gençay et al.), the authors explore the relationship between money growth and inflation as well as the money-income relation across several countries. It confirmed the complexity of these relations which without multiresolution and timescale decomposition remainded hidden. The relationship between money growth and inflation differed in strength across timescales, and even flipped in causality in Japan for one timescale.

Forecasting

Wavelets contribute two new approaches to economic forecasting, both based on the technique of decomposing the signal to be predicted into separate timescales. The first, outlined in the paper (Ario, 1996) and developed in (Schlter and Deuschle, 2010) fits a regular ARIMA model to each time scale and forecasts each scale independently. The overall signal is obtained by a summation of the different predicted values. (Ario, 1996) used Spanish concrete production and car sales and obtained some promising results.

The second method also decomposes the signal, but instead uses neural networks on each component to produce an overall forecast after recombining the components. The approach was first outlined in (Aussem and Murtagh, 1997) and developed further in a Master's thesis in (Tan and Pedersen, 2009).

A criticism leveled towards wavelet based forecasting is there is a problem in gathering enough data, especially for the lower frequency components. However, this is an issue for any non-causal filter, and not limited to wavelets. While still a valid criticism, if enough data is available, wavelet based approaches are yielding promising results.

Identifying Structal Breaks

The final interesting application we look at here is the ability of wavelet transform to identify structural breaks and shocks in a time series. This application works better with a continuous wavelet transform, but can be used with a discrete transform as well.

This technique takes advantage of wavelets' ability to model non-stationary processes. A common event in non-stationary series, especially in financial data, is a shock to the system, where a abnormally large positive or negative change occurs. Wavelets can capture these data changes and identify them, sometimes with indications before the events occur.

Below are two diagrams taken from (Setz, 2011). The first is an artificial series, derivation of Brownian motion with artificial jumps and shocks added to the system, each shock bigger then the last. The second is the log returns of real financial data, the Commodities Research Bureau index, with historical events overlaid. The highlighted red areas are regions where the power of wavelet coefficients is significantly higher than what we would expect if we model the time series as a one-lag auto-regressive model.

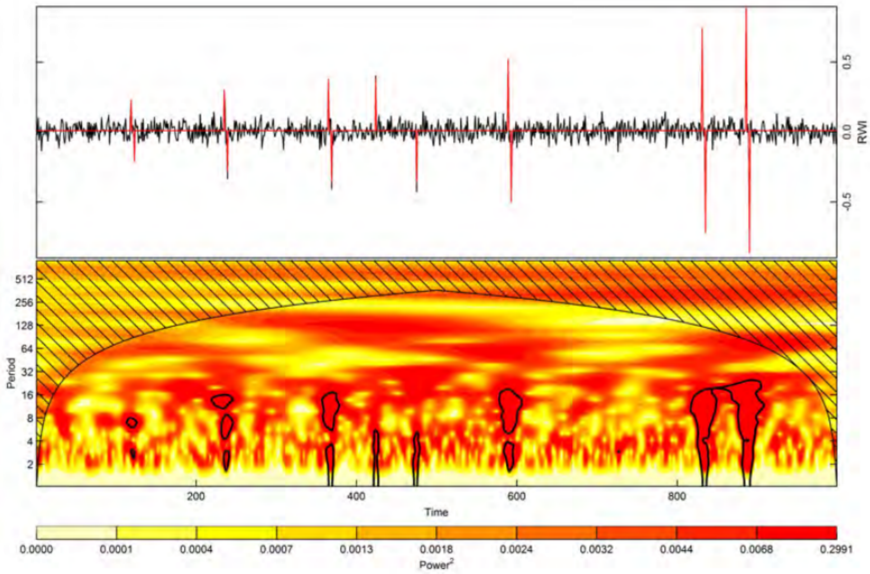


Figure 2.2: Derivation of Brownian motion with artificial jumps and shocks added to the system, with the CWT shown below. Areas of statistically high value are shown.

As can been seen in both diagrams, times around jumps and shocks to the time series have vertical bands of areas where the wavelet coefficients are higher then we would expect, spanning several scales. Sometimes, this rise in power actually starts before the jump, suggesting some sort of predictive power may be possible.

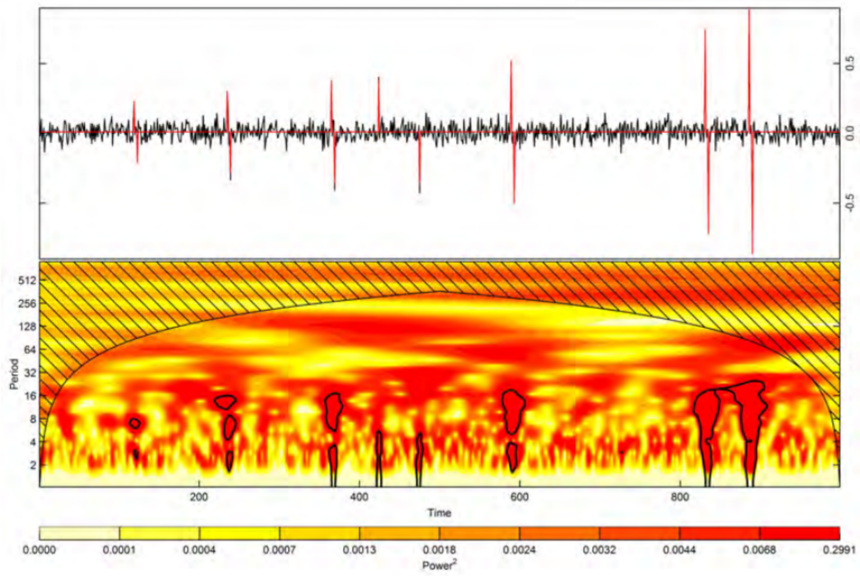


FIGURE 2.3: CRB Index from 1904-2010 with the CWT shown below. Areas of statistically high value are shown.

2.2 Sentiment Analysis

The area of sentiment analysis has a basic objective. Any text will contain more then the objective facts it is trying to convey. There will also be emotions, sentiment, affect, opinions. There is a reason another name for this field is opinion mining. While humans can pick up on if a movie review is positive or negative, it is a still a task that gives computers trouble. Sentiment analysis draws from several other fields to help solve its problems, including natural language processing techniques and machine learning classification tools.

Sentiment analysis has seen a boon in recent years, with the advance of the Internet age. It is easy to see why. More and sentiment laden texts are available quickly and easily to researchers. Social media, such as websites like Facebook or Twitter, provide short snippets of sentiment heavy text. Blogs on everything from politics to media abound. Consumer reviews for products and brands are becoming increasingly important.

Companies and politicians are also seeking access to information that only sentiment analysis can provide. Companies would like to know if it is possible to extract more complicated opinions about their products then the simple star rating systems. Politicians would like to distill social media concerning them and their campaign into real-time sentiment information, to allow them to react quickly to their public. In the age of big data as well, sentiment analysis could provide more metadata to allow quicker searching through reviews or blogs or other kinds of data.

The field of sentiment analysis has responded and risen to these challenges, as discussed below. However, sentiment analysis is far from a solved question, and there is still plenty of work to be done in the area.

This section gives an overview of sentiment analysis and some of the techniques used, with a focus on finding affect in text, before moving on to look at its numerous applications in the financial world.

2.2.1 Overview

The desire to automatically extract opinion or emotion have a text is not new. Early work that could be considered sentiment analysis tried to interpret a wide a range of topics including belief, metaphor, narrative, point of view and affect. (Hearst, 1992) (Carbonell, 1979) (Huettner and Subasic, 2000) Since the year 2001 there has been an accelerated output from the field, with literally hundreds of papers published since then. (Pang and Lee, 2008)

The rise in the field has been put down to, as already mentioned, the emergence of the internet and the new datasets it brings and the need of corporations and other entities to deal intelligently with these new datasets, as well as the improvement of machine learning and natural language processing methods.

Sentiment analysis is a mix of natural language processing techniques and machine learning classification. The main paradigm of opinion mining is trying to learn to classify texts based how they are written and what they contain. Natural language processing provides the features of the text, from as simple as word frequency to identifying gradation of adjectives. (Wiebe et al., 2005) Machine learning techniques can then learn which of of those features are useful to classify the text as containing certain emotions or sentiment. (Melville et al., 2009)

While there are many aspects of sentiment analysis, we focus here in this review on affect, mainly because that is the area used in this thesis. Affect has generally been broken down into three categories: positive, negative and neutral, especially in earlier work. (Turney, 2002) This is referred to as "sentiment polarity". Sometimes this is even simplified further, into simply detecting if a text had strong affect or was simply neutral. (Wiebe et al., 2005)

There are still challenges within the field. Much of the early work that has seen success was focused on reviews of products or movies, because of the given score attached to these types of text (the star ratings). However, this means sentiment analysis sometimes finds it difficult to deal with more complex texts, which may express multiple sentiments, even within one sentence.

Another problem is when a text has more then one topic it is expressing an opinion on. An example sentence that would prove a challenge to earlier work in the field is "Jurassic Park is great as a book, lousy as a movie." Work is being done to improve these issues, borrowing from linguistics to gain more insight into the structure of texts, as well as the words used. (Argamon et al., 2007)

2.2.2 Outline of Techniques

As mentioned previously, sentiment analysis is often broken into two areas - natural language processing methods to extract features from texts, before feeding those features into machine learning algorithms to learn patterns of classification or regression. While some have questioned the basic assumptions behind this breakdown (Argamon et al., 2007) it has yielded much success. For this discussion we use that breakdown, first looking at the natural language processing before moving on to the machine learning techniques.

Natural Language Processing

Probably the simplest model used to present a text is the Bag of Words model, which a simply a count of each of the individual terms used in the text. This representation couples well with the General Inquirer method (Stone et al., 1966) where the bag of words in compared with a pre-defined sentiment lexicon, counting terms associated with positive or negative sentiment. This approach has seen success, for example in (Tetlock et al., 2008)

A similar approach is a vector based representation, with each text represented as a vector $(w_1, w_2, ...w_N)$ where w_i is a Boolean variable indicating if the associated term is contained in the text. The set of terms W_1 to W_N could come from a pre-defined lexicon, or could be a list of all words in the corpus. For example, "Alice said hello", "Bob said hi", "Carol yelled hello". There are 7 words in this corpus: Alice, said, hello, Bob, hi, Carol and yelled. If this is the order of the words, then the vectors that correspond to these three documents are (1, 1, 1, 0, 0, 0, 0), (0, 1, 0, 1, 1, 0, 0) and (0, 0, 1, 0, 0, 1, 1).

Complementary methods to ensure results are not confused are important to these techniques that use simply the terms used in the text. First, a distinction needs to be made as to the grammatical use of a word. The terms cannot simply be treated as strings - the noun "good" has much less sentiment value then the corresponding adjective. Sentiment analysis borrows lexicographical methods to identify words and their grammatical contexts.

The second technique is known as stemming. This reduces words back to their stems, removing much ambiguity around term matching. For example, both "positive" and "positively" would be reduced to "positv". This reduces the labour required in defining the lexicons, removing the onus to enter every possible derivative word, and catches many sentiment laden words that might otherwise be missed. This complements the grammatical knowledge from above, sorting words from the the same stem but different uses into different terms.

Finally, usually stopping words which rarely have sentiment attached, such as 'the' or 'a' are removed from the text, to avoid the machine learning algorithms misappropriating significance. This can happen simply because such words are so prevalent the they make up a significant proportion of any document.

As already said, these approaches have seen success, but they are obviously simple approaches. They remove all information about the structure of the texts, or the relation between words. There have been proposed improvements to these methods, trying to retain some of this information.

An easy improvement is to the vector model. Instead of simple Boolean variables indicating a word's presence or not in the text, this can be changed to total number of terms, or text frequency (the number of times the term is used divided by the total terms in the text).

The most complicated version of the vector representation used in the literature is the term frequency-inverse document frequency (TFIDF). (Chakrabarti, 2002) This is a product of two terms: the term frequency as described before, and the inverse document frequency, the inverse of the number of documents the terms appears in. The idea behind this metric is words that appear often in different texts are less important to classification, and should have reduced importance. However, (Pang et al., 2002) questions whether any real benefit is gained with this approach.

Approaches that try to include some information on the structure include Parts of Speech, (Turney, 2002) which includes phrases, semantic orientation of phrases using Pointwise Mutual Information between words and phrases (Turney, 2002) and a cohesion-based approach (Fellbaum, 1998). This latter representation builds a graph of the text, with terms as nodes and edges based on relationships borrowed from linguistics and lexical cohesion (Halliday and Hasan, 1976).

Machine Learning

Once the features are created from the text, these features are used as input to machine learning algorithms. The first step of pre-processing on the features can help improve the results.

Natural language processing techniques, by their nature, output a lot of features; sometimes the numbers of different terms used in the corpus. Given enough features, this high dimensionality problem not only makes machine learning process slower, but it is more likely for the output to be random rather than meaningful patterns. So some sort of feature selection is required.

One common criterion used for feature selection is information gain. Borrowing from Shannon's information theory, we can define the entropy of a system where we are trying to classify texts as positive or negative as

$$H(D) = -p_D log(p_D - (1 - p_D)log(1 - p_D) \tag{2.6}$$

where D is the set of documents, p_D is the probability of finding a random text of D has positive sentiment. The entropy variesx between 0 and 1, as the probability of a random text being classed positive increases.

We want to select features that gain a lot of information, reducing the entropy of our set. Remaining with the information framework, we can define information gain as

$$IG(D, x) = H(D) - \frac{|D_0|}{|D|}H(D_0) - \frac{|D_1|}{|D|}H(D_1) \tag{2.7}$$

where x is a feature, D_0 and D_1 are two sets the feature splits the larger set D into. An example of how this feature might split D is with the Boolean vector representation is simply if x is 1 or 0. We can then choose features that maximize this measure of IG.

With the features chosen, an array of machine learning algorithms can classify texts. Support Vector Machines, K-Nearest Neighbours and Naive Bayes have all been used, and are well suited to the binary classification problem. Supervised machine learning algorithms have two stages - training, where they build the different models based on subset of the data that have been humanly labeled; and testing, where the models are evaluated with unlabeled (to the algorithm) data.

Naive Bayes classifies documents by estimating the joint probability of the words in a text and the different categories. The naivety comes from the assumption that all words are independent of each other. This assumption makes the calculation process much quicker and more efficient. This model can be extended to beyond two categories of positive and negative, and has been used for a broader range of emotion. Naive Bayes has been used extensively for sentiment analysis. (Xia et al., 2011) (Melville et al., 2009)

Support Vector Machines (SVM) in their simple form can only handle the binary classification problem, but is one of the best performers in the field. The SVM creates a decision plane separating the data into two classes. The algorithm creates a plane between the "support vectors", individual texts selected from the corpus. The support vectors are chosen in such a way that the fewest data are is placed on the wrong side of the plane i.e. misclassified and then the distance between the support vectors is maximised. Examples of the use of SVM in the field include (Xia et al., 2011) (Prabowo and Thelwall, 2009). SVMs can be expanded to non-linear and multi-classification problems using kernels, which has been tried in sentiment analysis. (Xu et al., 2011)

K-Nearest Neighbours (KNN) is an example of a system that does not define a explicit model for each category. Instead, in simply applies a label to an unclassified text which is the majority of the labels of a user specified K nearest of the texts from the training set. "Nearest" is a flexible definition, but often the simple Euclidean distance is used, combined with the vector representation from the NLP techniques. See (Tan and Zhang, 2008) for use of KNN in sentiment analysis.

There are of course a number of other machine learning methods, including unsupervised learning which do not need human provided labels to build models of the data. As this is not a review paper, we will not cover them all here, but this provides an insight to the work of machine learning. Studies have been done comparing the different machine learning methods, and these generally find SVM models come out on top (Tan and Zhang, 2008) (Ye et al., 2009).

Sentence Level Challenges

All the discussion so far has been focused on document based sentiment analysis, where an entire body of text is given one classification, whether in the binary positive/negative or a more complicated classification structure. Inherent in this approach is the assumption that the text only expresses one kind of sentiment or opinion about one entity. While sometimes this assumption can be reasonable, even in something like a product review it is often the case that multiple sentiments are being expressed about the one entity. "X is great because of this, this and this. However, that and that could be better."

This gives rise to the need to break a text into smaller pieces, each of which is given a separate classification. The pieces the text is broken into are usually sentences or clauses of sentences. This new lens comes with its own challenges, the main one being that not all sentences have a sentiment. While this is also true of documents, the objective sentence (one without sentiment e.g. Ireland has a land mass of 84,421 km^2) is much more common and more difficult to identify.

Generally, the approach had been to use supervised machine learning to classify sentences into objective and subjective, before further processing the subjective sentences into positive/negative or more complicated classifications. This comes with a large amount of man hours, as any decent sized corpus will have thousands of sentences that would all need to be classified by a human.

Work has been done to try save that large about of labour, with (Riloff and Wiebe, 2003) suggesting a bootstrapping approach to cut down the labeling required. (Pang and Lee, 2004) suggests using a minimum cut approach, with their underlying assumption being that neighbouring sentences should tend to have the same labeling.

Once all the sentences are classified as objective or subjective, the focus then becomes assigning sentiment labels. The approach here is much like the document based level, with NLP providing features for machine learning algorithms to classify. Recent research has suggested an interesting modification to this approach, where instead off applying the same strategy to each sentence, instead the algorithm adapts to different sentence types. (Narayanan et al., 2009)

Sentence based sentiment analysis can also easily provide a document level classification, by simply assigning the document label the majority of the sentence labels (or average if the classification is a regression).

New Research

This is a broad field with many different aspects to the study. Two interesting pieces of research are worth mentioning here to show the array of challenges being tackled. (Jindal and Liu, 2006) is making ground on the study of comparative sentiment. A very common sentiment expressed in reviews is comparing the subject of review to one or more products of a similar type. This paper found using a list of comparative words following a simple rule set, they were able to extract 98% of comparative opinions. The list included only comparative adverbs (more, less, words ending in 'er'), superlative adjectives and adverbs (most, least, words ending with 'est' and nine other phrases, including 'exceed' and 'prefer').

Another topic of research is mining for opinions on aspects of a main topic of sentiment. Reviewers often express opinions on different aspects of a product, not just the product itself. The main difficulty in Aspect-based sentiment analysis, as it is known, is identifying all the different aspects referred to by the review. One approach is to identify all noun phrases (borrowed from linguistics) and keep all that appear in the text above a certain defined frequency are kept as aspects (Popescu and Etzioni, 2005). An alternative approach is to keep noun phrases used in conjunction with certain known sentimental phrases. These phrases would have to be identified with a phrase dependency parser (Wu et al., 2009).

2.2.3 Sentiment in Finance

Attempting to incorporate the information provided by financial news into models of those same financial systems that news is reporting on has a long history. (Niederhoffer, 1971) analysed twenty years of New York Times headlines, classified by him into nineteen different categories to try see how markets reacted to good or bad news. His main finding was that markets do react to news, but bad news seemed to have a more pronounced and long term effect than good.

The advance of computers saw more work in this area. Without sentiment analysis at a stage where it could provide meaningful evaluations of sentiment in financial news, studies used proxies instead. The proxies had to be easily quantifiable aspects of the news, and included news arrival, type, provenance and volume (Cutler et al., 1989) (Mitchell and Mulherin, 1994).

Some even turned to market based proxies to emulate news: (Engle et al., 1993) used abnormal stock returns (as defined using a Fama-French model) instead of news, and studied the impact of these abnormal events. Indeed, this proxy influenced this industry, and even newer research focusing on this aspect of the market e.g. (Tetlock et al., 2008) following the number of abnormal returns following news stories, or (Azar, 2009) reversing the direction of information, using abnormal returns to classify financial news rather then assigning human labels.

With the advance of sentimental analysis, this work continued, seamlessly incorporating actual news sentiment now that it could be quantified. There were two, slightly different approaches using the new methods. Some worked with whole classified texts, labeling news stories as negative or positive and looking at how markets reacted to a number of news articles. The alternative is to work with summation of total negative or positive features used in the news articles.

This area of research has grown in tandem with the advance of sentiment analysis. (Tetlock et al., 2008) is an oft cited paper in this area. It worked with both classified news articles and the number of sentiment terms used in the news. It had only two sources of news - the Dow Jones News Service and the Wall Street Journal over a period from 1984 to 2004. It used the General Inquirer method to count the number of negative and positive terms, as well as a self defined lexicon of words associated with a firm's fundamentals. The paper used both the term frequency and the z-score of sentiment terms as input to their models.

The paper studied the reaction of two market variables: abnormal returns and standardized analyst forecast error. It used simple regression models to examine how a firm's stock reacted to the values extracted by sentiment analysis. They also modeled the firm's behavior just prior to and just after the release of negatively or positively labelled news stories. They focused on the cumulative abnormal returns as proxies for a firm's behavior. See the diagram below.

The paper's main finding was that a firm's stock did react badly to frequency of negative terms used in firm-specific news stories, even more so when those news stories the stories had a high frequency of firm fundamental terms.

Another paper (Oh and Sheng, 2011) turned to the mircoblogging site StockTwits to see if firm performance could be modeled using tweets with the firm's stock ticker. This paper not only studied returns on a firm's stock, but also market adjusted returns, to compensate for whole market movements. This avoids the misjudgment where a particular firm's stock overall sees negative returns after positive mentions on Twitter,

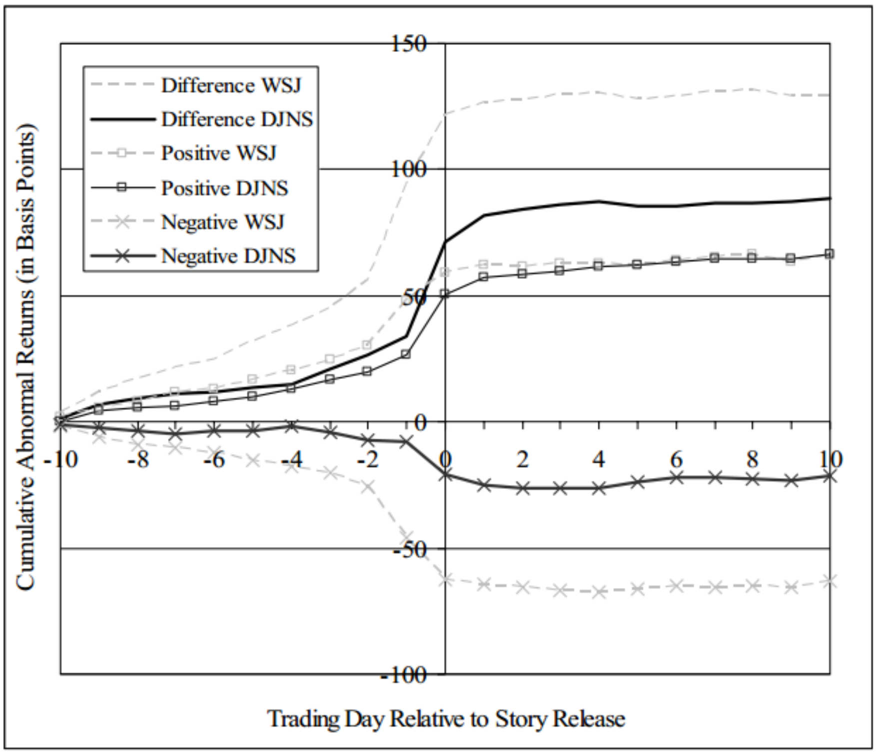


FIGURE 2.4: A graph showing a firms abnormal event returns from 10 trading days preceding a news storys release to 10 trading days following its release. All news stories focus on S&P 500 firms and come from either Dow Jones News Service (DJNS) or The Wall Street Journal(WSJ). Also shown is the difference between the reaction to positive and negative news stories for each source. From (Tetlock et al., 2008)

but this return was due to a overall negative movement of the market. The adjusted returns simply compare the behaviour of a firm's stock to the market as a whole, rather than in isolation.

The reason for choosing mircoblogs is succinctness, allowing for easy sentiment analysis; high volume, giving confidence in statistics like average negative term frequency; and the real time aspect of the information. The paper fitted two models - one based simply on the volume of tweets, and another with sentiment information added. The sentiment was slightly different from positive/negative, instead classifying tweets as bullish or bearish.

The paper found that volume of tweets alone was good enough to give predictive power of markets adjusted returns, but including sentiment data increased the accuracy, and could include unadjusted returns as well. Furthermore, they found that tweets with bearish sentiment had more predictive accuracy that bullish tweets, reflecting somewhat the result from (Tetlock et al., 2008) that negative sentiment played a larger role in the models.

A interesting alternative approach is studied in the thesis by (Azar, 2009) Perhaps because the main goal was to train models to classify other types of texts, this thesis reversed the direction of information gain. Instead of using sentiment of financial news to help model financial data, they use abnormal returns to provide labels to news articles. Articles published on days of abnormal negative returns are labelled as negative, and similarly for positive.

The paper then uses the previously discussed NLP and machine learning techniques to learn classifiers for texts. While the paper failed in its goal of using these models to classify movie reviews, they also implemented a hypothetical stock buying strategy. Training on only articles from the previous year, the trading strategy was if a news article came out about firm i, it would buy or short that firm's stock if the news was positive or negative respectively. 2.5 shows the results of this strategy, showing several of the classification mehtods making positive returns.

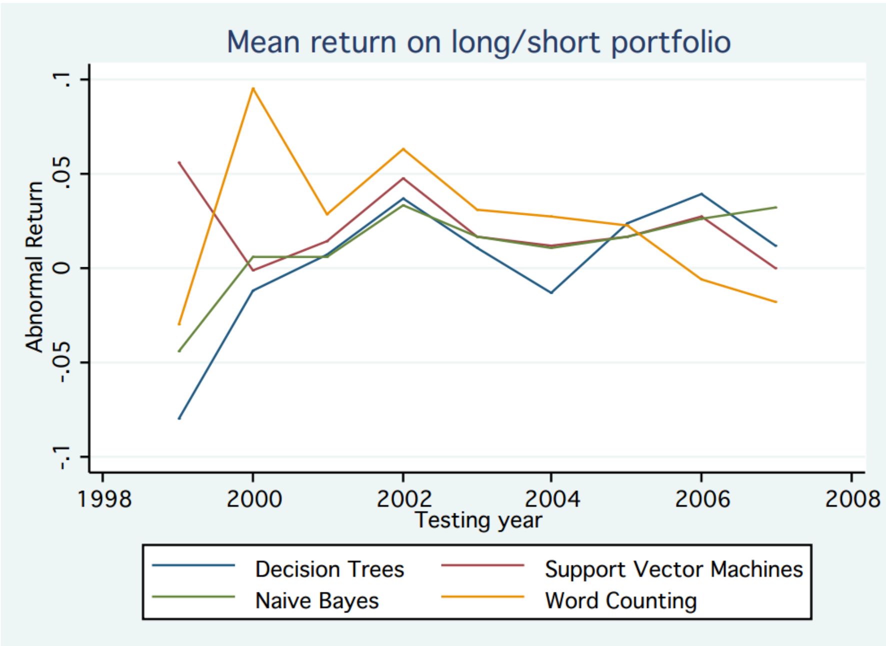


Figure 2.5: Mean returns for different classifiers using the strategy outlined. From (Azar, 2009)

2.3 Possible Mechanisms

2.3.1 Efficient Market Hypothesis

If a trader reads in the Financial Times a news article favourable towards a certain firm, this will raise his evaluation of the firm's stock. They will quickly try to buy its stock, in expectation of a rise in the current price to his new greater evaluation of the firm. However, sellers of the stock presumably also read the Financial Times, and have also seen this favourable article. As they and the trader are rational actors, their evaluations will agree, and the trader will be forced to buy at this higher price.

There is a body of economic theory that delineates the above intuitive grasp of how financial news interacts and influences markets. The Efficient Market Hypothesis (EMH) states that the price of any stock or commodity immediately reflects all available information. In a world of rational buyers and sellers with perfect information, there is no way to "beat the market" and make a greater return then the market overall.

There are three versions of the EMH: weak, semi-strong and strong.

The weak form states that no profit can be made by looking at transaction information - previous prices, volume levels etc. If true, it negates the possibility of finding patterns in past prices and asking bids. It makes no allowances for information from outside the market.

The semi-strong version takes a harder line, stating that any public information, past or current, is already summarized by the current price. It might still be possible to make money with private information, but this kind of information is either expensive and/or illegal to obtain, due to insider trading laws.

The strong version states that it is impossible to make a profit greater then the market as a whole no matter what information one holds. This is because, despite laws to the contrary, insider trading takes place anyway, and the market price changes to reflect this trading.

2.3.2 Fischer Noise

Of course, the real world is not as simple as EMF supposes. There are all sorts of complications to the underlying assumptions to EMF. An asymmetry of information, even if it is public, irrational traders and imperfect communication all cut against EMF.

In addition, there is also basic noise in the market. (Black, 1986) is probably still the best work on noise in financial markets. Fischer identified several sources of noise - future tastes and technology down to traders following their gut instead of rational.

Noise makes it possible to make profit in trading on the market, but also makes it difficult to take advantage of inefficiencies and make that profit at all. When working in the real world, traders still react to financial news, but their evaluations will not always agree or be totally rational. Sentiment in financial news is fundamentally about uninformed people trying to make sense of the noise in the market.

Traders can over or under react to news, or have a delayed reaction once confirmation from another source comes in. See the different reactions to good and bad news (Niederhoffer, 1971). The sentiment is not always rational either - journalists are no better then traders when it comes to rationality. The interaction of sentiment, which may or may not be rational, with traders, of equally questionable rationality, can have profound effect.

A common approach is to split traders into informed traders and noise traders. Informed traders (De Long et al., 1990) know the true value of a good or firm, and they make profit from irrational mispricing of stock or goods by the noise traders (Koski et al., 2004). Of course, noise prevents one from knowing which trader is which with certainty. Noise traders do not necessarily lose on their trades in the short-term, nor do informed traders necessarily win. (Black, 1986)

Sometimes noise traders can have such an influence on the market that is overpowers the checking nature of the informed traders, and the irrational strategies can make short term profit. This over or undervaluing of the market cannot be sustained however, and the bubble eventually pops. One of the ways this state of affairs can come about is discussed in the next section.

2.3.3 Herd Behaviour

The term 'Herding' used in the context of in financial markets means following the crowd with your investment decisions. A trader has been influenced by the herd if they make an investment they otherwise would not have made upon hearing a number of other traders have made that investment, or visa versa. Herd behaviour is inherent bias that has been studied in behavioural psychology. (Asch, 1956) It is often of major benefit to us outside the financial world. If everyone you encounter does not do a certain activity, it is probably not a good idea to engage in such. Following the crowd also has a social aspect, as resisting group opinion can create tension and lead to isolation of the rebelling individual.

Where sentiment starts to play a role is providing the initial push to get a herd going. In market equilibrium, noise traders' over-optimism/pessimism is kept in check by the informed traders. However, if a particularly damning or laudatory article, or series of articles, is published that does not reflect reality, it can change behaviour. If this convinces enough noise traders to move in one direction, informed trader will start to doubt themselves, wanting to follow the crowd. If enough people follow, it can create a bubble, and short term gains soon turn into a long term crash.

2.3.4 Feedback

There is the possibility that the relationship discovered will not be one way between sentiment and market prices, but instead see some feedback. As much as financial news contains information, objective and subjective, to form opinions and trading strategies, financial articles also contain sentiment trying to make sense of the market noise. This can be as simple as a news report on the bad showing of a firm's stock in the past week/quarter. Journalists are also subject to herding mentality, where a rational journalist changes the sentiment in their article to match how the crowd of their peers are writing.

Chapter 3

Method

3.1 Decomposition Analysis

Wavelets and the wavelet transform were first developed for geophysical applications (Goupillaud et al., 1984) but have since found a wide array of applications in various fields, including finance, geology, atmospheric science and the study of turbulence (Gençay et al.).

This section will give an overview of wavelet theory and the mathematics behind the various transformations and techniques, in addition to providing some of the most prominent applications. It borrows from (Gençay et al.), while (Burrus et al.) was also useful.

3.1.1 Time and Frequency Resolution

When processing signals, typically the raw signal is some time series - voltage across a component, stock prices or a speech signal are typical examples. It is well known that extra useful information about that signal relating to its frequency components are available through it Fourier transform. The Fourier transform gives the frequency representation of a time signal, exploiting the fact that any signal can be represented as a sum of different sine and cosine waves.

The choice of basis functions for the Fourier transform (sine and cosine) give a complete resolution in frequency as they have a value across all times. This makes the Fourier transform ideal for representing stationary time series. The unchanging repetitive signals can be easily represented with a small number of frequency components.

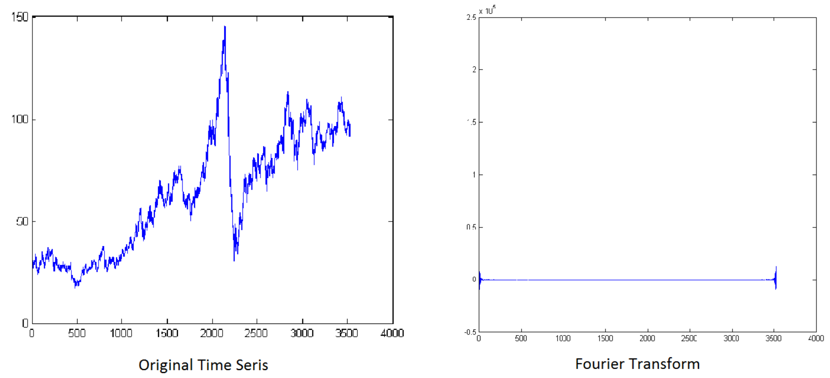


FIGURE 3.1: The original time series of price data on the right, with the Fourier transform on the left. The Fourier transform captures the large drop in price in 2001 in the high frequency component, but gives no indication when it occurs.

However, many real world signals are not stationary, including two of the examples given above, speech recordings or stock prices. These real world signals contain events and parts that are localized in time - they appear, influence the signal and die away again. Or even more fundamentally, there is no “underlying” signal, but the signal is made up of a series of smaller (in time) signals e.g. speech.

With the Fourier transform, these extremely time-localized signals have an influence on the entire Fourier transform. Indeed, these events are difficult to capture with a Fourier transform, requiring a combination of a large amount of sines and cosines to be expressed. It can be the case that these signals drown out the underlying signal (if present) in the Fourier plot, obscuring analysis.

Another drawback of the Fourier transform is thet for prefect frequency resolution it must give up time resolution entirely. A high-frequency time-localized part of the signal will be captured by the transform, but no indication can be given was to when in time that high-frequency component occurred.

These drawbacks were well known in academia, and various attempts were made to overcome them, including the Short-Time Fourier Transform (STFT) (Gabor, 1946) This technique involves taking a sliding window across a time series, and only applying the Fourier transform to the part of the signal that falls within the window. An immediately obvious drawback is this will fail to capture events that appear within the width of the window. However, the technique does still provide extra information then usin just the Fourier transform, and continues to be used today, for example in speech processing. (McAulay and Quatieri, 1986)

Wavelet transform goes some way to solving these problems. The wavelet transform adapts the trade-off between time and frequency resolution depending on the frequency. Lower frequency events can be resolved with less time resolution, and vice versa for high frequency events. This is what makes wavelets well suited to studying economic signals - the high frequency events are highly localized in time, and are captured by the wavelet transform with their times intact, thanks to the high resolution in time. The diagram below gives a good summary of how different transforms deal with time and frequency resolution trade-off.

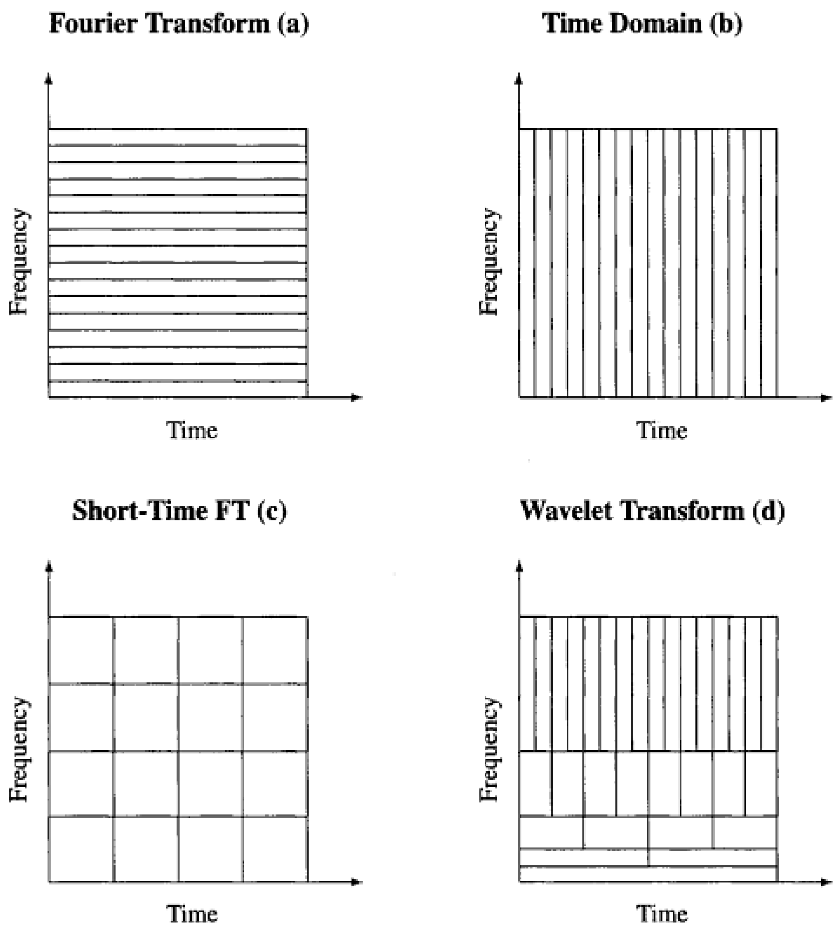


FIGURE 3.2: Diagram taken from (Gençay et al.) showing the difference resolutions in time and frequency for different transforms.

3.1.2 Wavelets

The underpinnings of the wavelet transform and all associated techniques are the wavelets themselves. There are many different kinds of wavelets, but they all obey a basic rule, called the admissibility rule:

$$C_\psi = \int_0^\infty \frac{|\Psi(f)|}{f} df < \infty \tag{3.1}$$

where $\Psi(f)$, a function of frequency f, is the Fourier transform of the wavelet, $\psi(t)$.

Fulfilling this condition ensures that the integral of the wavelet across all time must equal zero and the wavelet has unit energy. In effect, the wavelet must have non-zero values but positive values are cancelled out by negative values. Wavelets do not have full support across time, but are instead localized.

$\psi(t)$ is often called the *mother wavelet* and all wavelets used in the transform are scaled and/or time-translated versions of this mother wavelet, as explained in the following section. Below is a diagram outlining some commonly used mother wavelets.

The different wavelets have different properties and typical uses but, as discussed in (Ramsey and Lampart, 1997) and (Michis, 2011), the choice of wavelet has little impact on the type of application used in this thesis, so long as the wavelet is relatively smooth. As such, we will not go into that topic here, but (Gençay et al.) has an interesting discussion on the subject. For this work we used Coiflet (5), Daubechies (4) and the Mexican Hat, with the choice having no influence on the results. The results presented in Chapter 4 are from the work using Coiflet (5).

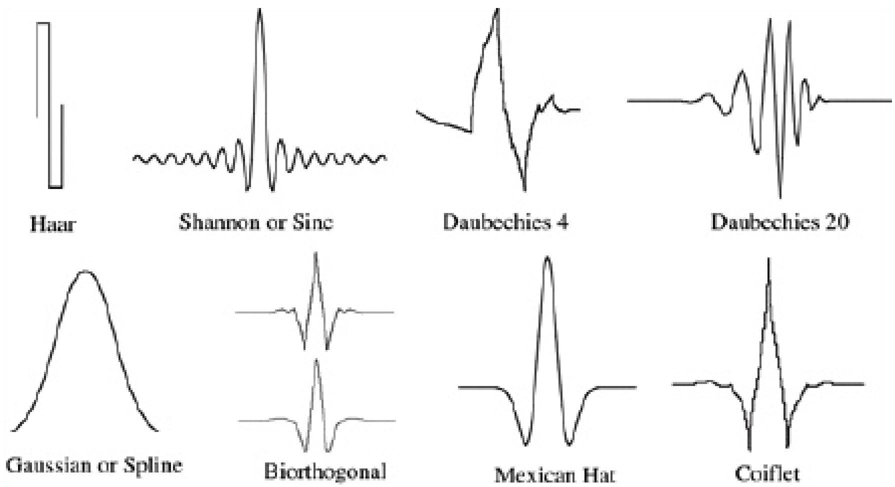


FIGURE 3.3: Different types of wavelets

3.1.3 Transforms

Continuous Wavelet Transform

The actual wavelet transform itself is a projection of the signal $x(t)$ through the wavelet ψ. The equation

$$W(u,s) = \int_{-\infty}^{\infty} x(t)\psi_{u,s}(t)dt \tag{3.2}$$

gives that projection. This is the Continuous Wavelet Transform (CWT). $\psi_{u,s}(t)$ is the transformed version of the mother wavelet, given by

$$\psi_{u,s}(t) = \frac{1}{\sqrt{s}}\psi\Big(\frac{t-u}{s}\Big) \tag{3.3}$$

Note how the transform is a function of two variables, u and s. u is obviously related to time - it translates the wavelet by u along the time scale. s is referred to as the scale of the wavelet, and it is the part of the transform related to frequency. Scale is inversely related to frequency - at larger scales, the wavelet it dilated by a factor of s, expanding its range. This wider range allows the capture of lower frequency events. The wavelet adaptation with scale is what gives the transform its varying time and frequency resolution.

Below is a visualization of how the wavelet is scaled and shifted when used for the transform.

Discrete Wavelet Transform

In theory, the CWT has a value for every combination of u and s, though of course any computer based implementation is going to be discrete at some level of granularity. However, CWT contains a lot of redundant information for signal representation. The signal can be fully reconstructed using only a sample of the information contained in the CWT. Much like the Discrete Fourier Transform, the Discrete Wavelet Transform (DWT) can be thought of as removing much of the redundant information by a sampling of the CWT.

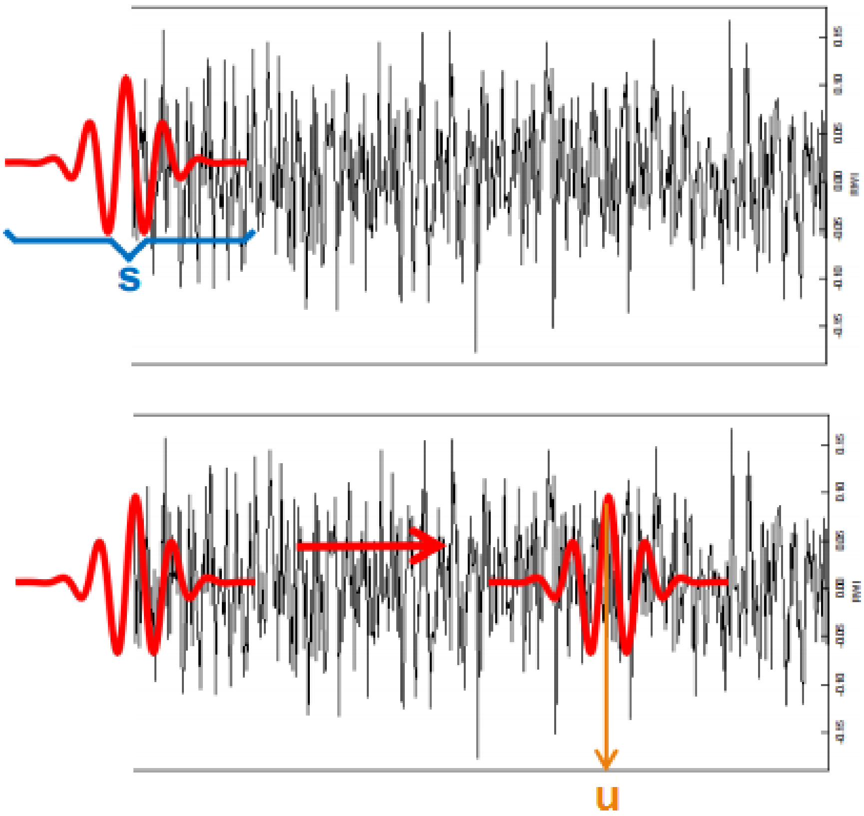


FIGURE 3.4: A wavelet dilated and shifted with respect to the time series to be transformed

A critical sampling (one that still allows a prefect reconstruction of the signal with a minimum number of coefficients) is obtained by sampling at $s = 2^j$ and $u = k2^{-j}$ where j and k and integers representing the set of discrete dilations and translations. [1]

Of course, the actual implementation of DWT does not involve sampling the CWT - that would negate the computational savings made. The DWT is very similar to the implementation of the CWT however, so it is explained only briefly here.

Again our basis is the wavelets themselves. The DWT uses discrete versions of wavelets, which must follow similar discrete versions of the wavelets used in CWT. Namely the wavelet must sum to zero

$$\sum_{n=0}^{L-1} h[n] = 0 \tag{3.4}$$

[1]It should be pointed out here that this is not the only choice of critical sampling, but the one most often used.

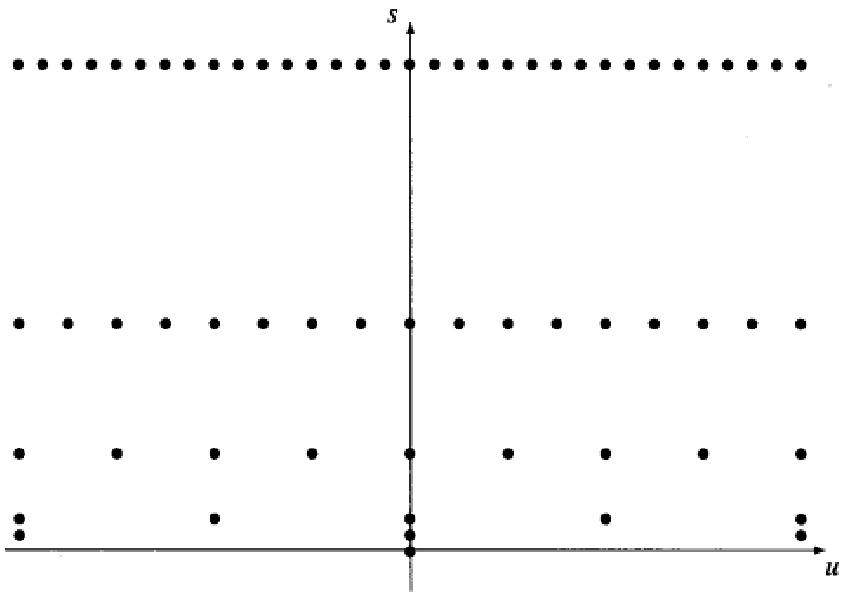


FIGURE 3.5: Diagram taken from (Gençay et al.) showing the critical sampling of the time frequency plane, taking the minimum number of points to allow the signal to be perfectly reconstructed

and have unit energy

$$\sum_{n=0}^{L-1} h[n]^2 = 1 \tag{3.5}$$

In addition to these conditions, discrete wavelets must be orthogonal to its even shifts i.e.

$$\sum_{l=0}^{L-1} h[n]h[n+2l] = 0 \text{ for all nonzero integers l} \tag{3.6}$$

These conditions ensure that the wavelet coefficients obtained from the DWT are orthogonal to each other. The DWT is then, like the CWT, obtained by projecting the signal through different dilations and transposes of the mother wavelet. However, in the discrete case, this projection is much more easily expressed as convolutions.

Think of the wavelet as a filter. When convolved with the signal for different ks, we filter out all parts of the signal with frequency higher then our scale 2^{-j}.

$$y_{high,j}[k] = \sum_n x[n]h[2k-n] \tag{3.7}$$

Our wavelet is, in effect, a high-pass filter. This wavelet filter gives the wavelet coefficients for the level j. A natural complement to a high pass filter is a low-pass filter, called in this context the scaling filter, $g[n]$. The signal convolved with the scaling filter

$$y_{low,j}[k] = \sum_n x[n]g[2k-n] \tag{3.8}$$

gives the signal with those higher frequency components removed from the signal. The process can now be repeated on the resulting $y_{low,j}[k]$ to extract $y_{high,j+1}[k]$ and $y_{low,j+1}[k]$ i.e. the wavelet coefficients for level $j+1$. Before subsequent filtering, the signal is sub-sampled. This is the process that changes the scale - the wavelet and scaling filters themselves are unchanged. This sub-sampling retains the desired orthogonality. Note how this is direct contrast to the MODWT, discussed below. $y_{low,j}[k]$ This process is repeated for as many levels as required. The figure below helps visualize this process.

As the level increases, the coefficients are being calculated on lower frequency data i.e. longer in time. This means, for a given signal $x[n]$, the higher levels will have less data to sample from, simply because the wavelet is becoming larger with the scale. This directly leads to less coefficients on each level. If we follow the critical sampling suggested above, the number of coefficients halve with each increase in level, as our scale is doubling.

Note the result of our process is J+1 vectors of coefficients, where J is the total number of levels specified by the user. These vectors can be gathered into one length N vector $w = [w_1, w_2...w_J, v_J]$ where w_j is the vector of coefficients at level j, and v_J is the result of the final scaling filter, and is equal to the averages of the signal at a scale 2^J.

Maximum Overlap DWT

The DWT gives up a number of desirable features to retain orthogonality. DWT cannot handle time series of arbitrary length N, but instead must be some multiple of 2^J. DWT coefficients also have an odd relationship in time with the original time series - events in the original time series may not exactly align with events in the multiresolution details (see 3.1.5) and the DWT is not invariant to time shifts or circular-shifting of the original time series. (Gençay et al.)

To get around these problems, the Maximum Overlap Discrete Wavelet Transform (MODWT) gives up orthogonality by dilating the wavelet and scaling filtering instead of sub-sampling the signal at each step. In MODWT, $h_j[n] = h[n]/2^j$ and $g_j[n] = g[n]/2^j$ where $h[n]$ and $g[n]$ are analogous to the mother wavelets used in CWT.

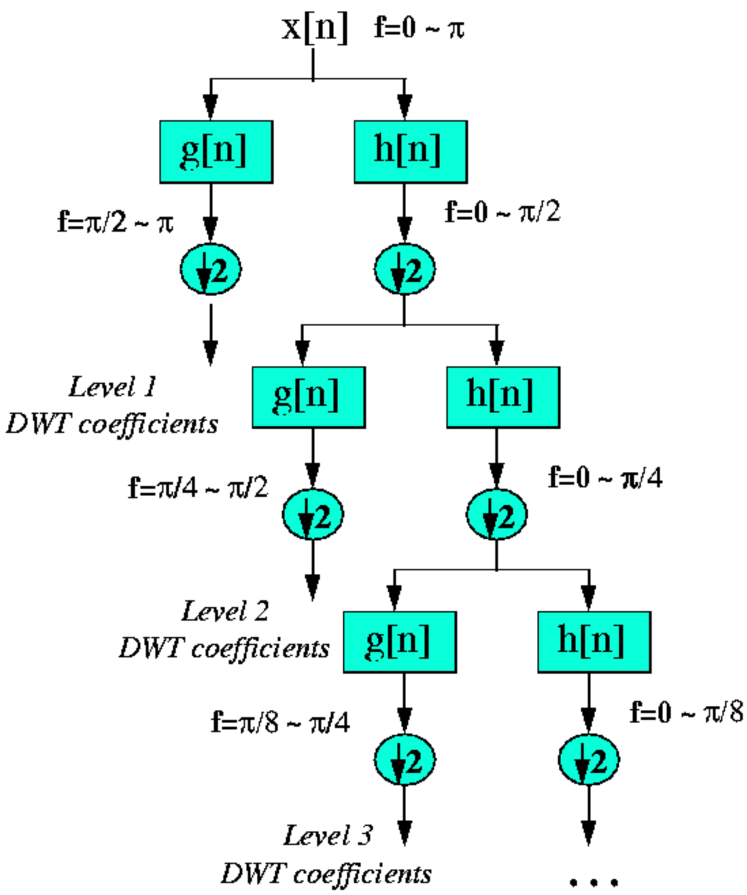


FIGURE 3.6: Diagram taken from (Polikar, 1999) showing filter implementation of the DWT

The flexibility of MODWT means it is the most widely used of the wavelet transforms, and is the transform used in this thesis. Care must be taken when making inferences based on the transform, due to the lack of orthogonality. See 3.1.5 for details.

3.1.4 Wavelet Details

An important extension to the DWT and the MODWT is an additive decomposition of the original time series. Using the wavelet coefficients, J different signals can be constructed, each the same length of the original times series.

These vectors are called wavelets details, d_j and are constructed via the following formula:

$$d_j[n] = \sum_k w_j[n]h_j[2k - n] \tag{3.9}$$

where d_j is the wavelet detail associated with level j and corresponding scale, s. As can be seen by the formula, the details are simply linear combinations of the scaled and translated wavelet. The final detail, d_{J+1} uses v_J, the result of the final scaling filter.

These details can be linearly combined to reconstruct the original signal. We can define the signal as

$$x[n] = \sum_{j=1}^{J+1} d_j[n] \tag{3.10}$$

These details are the isolated activity of the original signal on the particular scale, 2^j. Analogously, we can define a wavelet smooth, s_j which is the cumulative sum for the wavelet details from j+1 to J+1. As j increases, the smooth loses more of the higher frequency activity, and becomes smoother as a result. s_J is the representation of the lowest frequency activity in the original signal.

3.1.5 Multiresolution Analysis

Multiresolution Analysis is use of these details to view the signal decomposed over different scales. Since the different scales correspond to different timescales, the details allow analysis to be focusing only on events of a particular frequency, without the influence of other events.

This isolation of particular frequencies is of great practical use in many fields, particularly economics and finance, but also areas like image and audio processing. After the first paper on the subject (Mallat, 1989), multiresolution analysis saw an immediate adoption by the financial industry. See 2.1 for more details.

The reason the multilevel view wavelets provide is important to economics is it allows work to acknowledge what has long been known in financial theory - that economics signals are the summation of many different effects and events that operate on many different levels. The price of a stock is influenced both by the daily trading as well as so called “market fundamentals”.

Multiresolution analysis allows the analysis of one without having to consider the other.

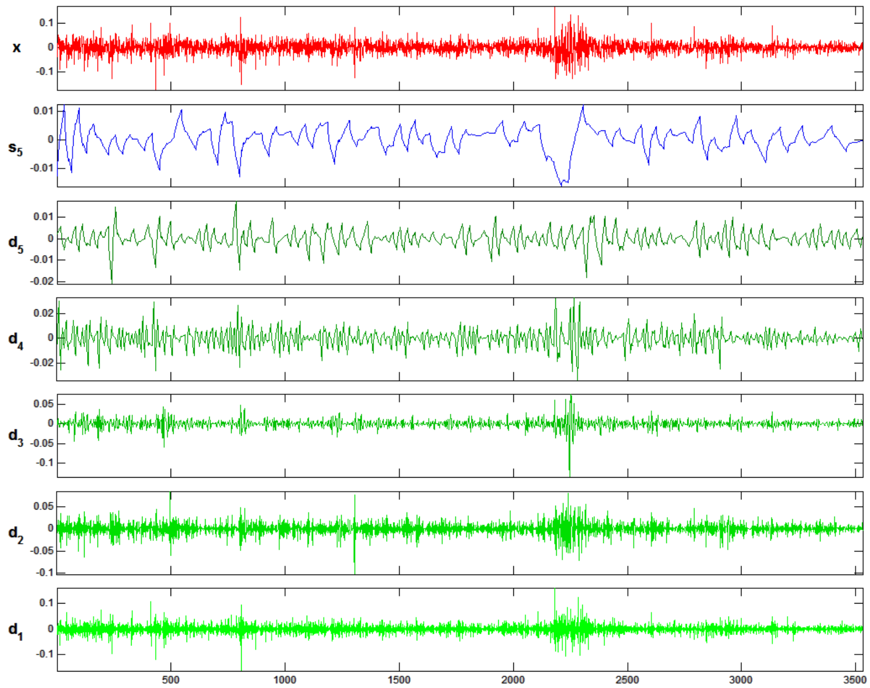


FIGURE 3.7: Example of details and smooth obtained from a 5 level DWT of returns of Oil sport prices

3.1.6 Analysis of Variance

With any transform, care must be taken with the assumptions made and their foreseeable implications. As discussed before, the main difference between DWT and MODWT is that DWT retains orthogonality at the expense of flexibility in dealing with signals of different sizes and invariance to time shifts. This difference has implications when considering the variance of the details obtained using the two methods.

In DWT, orthogonality is retained, as already stated. This means that energy is preserved by the transformation i.e. $\|x\|^2 = \|w\|^2$. This is easily proven below.

$$\|x\|^2 = x^T x = (Ww)^T Ww = w^T W^T Ww = w^T w = \|w\|^2 \tag{3.11}$$

where W is DWT represented in matrix form. (see (Burrus et al.) for details on this matrix representation). This is precisely the orthogonality of the transform that allows the substitution of the identity matrix for $W^T W$.

This relationship carries over to the details, allowing us to decompose the energy of the signal across the different levels like so:

$$\|x\|^2 = \sum_{j=1}^{J} \|d_j\|^2 + \|s_J\|^2 \tag{3.12}$$

Since variance is proportional to energy, this result indicates variance of each detail accurately represents the proportion of the overall variance attributable to the level of events making up the signal.

Sadly, this relationship does not carry over to the MODWT. To distinguish from the DWT, allow the vector of wavelet coefficients obtained by the MODWT to be represented as $\tilde{w}$, with details and smooths using similar notation, $\tilde{d}$ and $\tilde{s}$.

(Percival and Mofjeld, 1997) showed that in fact the first relationship still holds - energy is retained across wavelet coefficients i.e.

$$\|x\|^2 = \sum_{j=1}^{J} \|\tilde{w_j}\|^2 + \|\tilde{v_J}\|^2 \tag{3.13}$$

but the second relationship, the decomposition of energy across wavelets does not. This is because a relationship exploited in DWT, $\|d_j\|^2 = \|W^T w\|^2 = w^T W W^T w$ no longer holds. In fact, (Percival and Walden) shows that $\|d_j\|^2 \leq \|w_j\|^2$. As such, when reporting variances of different levels, care must be taken to report the variances of MODWT coefficients, not the details. All reported variances in this thesis naturally come from the coefficients.

3.1.7 Thesis Implementation

All work with wavelets done with this thesis was implemented in Matlab 2013a using the Wavelet package. No code was written - instead the work was done using the package's implemented GUI. As such there is no source code in the Appendix.

3.2 Sentiment Analysis

3.2.1 General Inquirer Method

The natural language processing (NLP) technique used by this thesis is the General Inquirer method, developed by (Stone et al., 1966). This method translates the text into a Bag of Words format, where the number of each word is simply counted, all relational information removed. The sentence "Billy saw the nicest green field in the world" will look like the following in Bag of Words format: Billy:1, field:1, green:1, in:1, nicest:1, saw:1, the:2, world:1.

While Bag of Words does remove the relation information from the sentence, it can be made more complicated to retain certain information of the grammatical context. The sentence "Billy saw a saw" would count the verb 'saw' as different from the noun.

General Inquirer method takes Bag of Words format and compares it to human defined dictionaries of words that have been associated with certain categories. (Stone et al., 1966) defines two lists of words for the most common results of affect in sentiment analysis: 1915 words for Positive and 2291 words for Negative. There are many more than two defined dictionaries, and as we will see later in section 3.2.3, local user defined dictionaries are in use.

The General Inquirer method simply gives the number of times a term in the dictionary appears in the text. Modifications used by Rocksteady, the NLP system used by this thesis, improve the approach, but this is the bedrock of our sentiment analysis.

3.2.2 Curating the Data

We need a source of data to perform sentiment on. This thesis used the news aggregater LexisNexis. LexisNexis has access to a large number of online newspaper archives and news blogs from around the world. It allows users to define search terms to appear in different parts of the text, define which news sources to consider and across which timelines.

We included only the terms "Crude Oil" within the headline or opening three paragraphs. This was our only criteria to ensure the news articles had some relevance to crude oil spot prices (see 4.1) We wanted this criterion to remain broad however, as we theorized the kind of news that would impact on oil prices would be a broader then directly related news with 'Oil' in the title.

We also limited the timeline of news from Jan 2000 to Feb 2014, during which we could remain confident of online news' reliability and coverage.

We limited our search to three main news sources: The New York Times, The Wall Street Journal and the Financial Times. These are three extremely well respected news institutions, known to be followed by Wall Street and beyond, with a worldwide coverage.

The reason we limited to just these three sources was because of data constraints. LexisNexis only grants ordinary users the use of a search that yields less than 3000 results, and then only download 500 at a time. Even with only three sources, with our timeline ranging from 2000 to 2014 our search yielded 35397 articles, and took two days to download in 500 article size blocks.

A key part of dealing with news data is removing reprints. Articles and pieces can be reprinted for a number of reasons: the paper's need to correct or update details, the story original comes from another source or one story developing live through the day may re-publish extended older articles as "new".

Thankfully, LexisNexis provides an option to check for similarity between different articles, and automatically removed duplicates. The setting chosen when obtaining this data was to remove articles of "High Similarity", to avoid the algorithm being too general. The similarity scores between two articles is based on the number of word changes to make two articles exactly the same, so this algorithm should not impact on our sentiment analysis.

3.2.3 Rocksteady

Rocksteady is a affect analysis system developed here at Trinity College, Dublin, Ireland. (Ahmad et al., 2011) Rocksteady uses a General Inquirer Method, with some modifications to add robustness. The system can interact directly with a LexisNexis corpus, parsing the database for meta-data and the actual text.

The first thing LexisNexis can do is to drop articles based on certain criteria. Articles below a given length can be removed, as well as articles with fewer than a certain number of dictionary terms in the text i.e. the text is overall neutral. This work dropped articles with fewer then twenty total terms or five Positive, Negative or Crude Oil (see below) terms total, as suggested by (Tetlock et al., 2008)

The second pre-processing step Rocksteady takes after this basic sanitation is aggregate all the articles into a chosen time-frame, ranging from hourly to yearly. This thesis used daily data, to align with the daily data obtained for oil prices.

Finally, Rocksteady generalizes the dictionaries to included derived words. For example, if "bad" is in the dictionary, Rocksteady will add "badly" if the word is not already there. The grammatical knowledge the dictionary provides makes this task relatively easy. This process ensures Rocksteady captures many of the negative or positive terms that it might otherwise have missed.

Rocksteady then calculates the number of total terms appearing in any one day (in our case) as defined by our dictionaries. We used three dictionaries: (Stone et al., 1966) Positive and Negative dictionaries, as well as a locally defined Crude Oil dictionary. This latter dictionary was used simply to confirm out articles were indeed relevant to our chosen market.

Rocksteady then calculates the percentage of total terms of the dictionary-defined terms for each day, and finally the z-score for each dictionary i.e. $p = \frac{P-\mu_P}{\sigma_P}$ where P is the total number of positive terms used in a day, μ_P is the mean of P and σ_P is the standard deviation of P.

(Tetlock et al., 2008) uses z-scores in their analysis, but this thesis uses percentages. While it may by a topic of future work to use the z-scores, we felt there was a danger in trying to model sentiment as a Gaussian function. Sentiment reacts to real world shocks and unexpected events, and trying to model such as Gaussian have proved futile in the past. (Taleb) We expect sentiment to fluctuate widely, rendering the z-score a misleading statistic.

3.2.4 Final Pre-processing

The final pre-processing required was to take the output of Rocksteady and remove sentiment data for the dates for which no oil prices exist because the market was closed. While editing the data somewhat, this was felt to be a better solution to inserting values into the oil prices when the market was closed. While Rocksteady does add data from after the market closes to the data for the following day, it cannot take into account Bank Holidays or other miscellaneous market closures.

3.3 VAR Model

With these two types of data, the returns on oil prices and the negative terms from the associated sentiment analysis, there is now a need for a model to test the basic question of the thesis - what is the relationship between the two time series. The model chosen here is the vector autoregression (VAR) model, a commonly used model in econometrics, as it conforms well to the theory of economic variables.

An assumption often made in economic analysis is past values of an economic variable contain information for modelling the current (and future) values of that variable. This assumption comes both from experience and theoretically based on variable's autocorrelation with trends and seasonalities.

This view can be expressed as

$$\hat{y_t} = f(y_{t-1}, y_{t-2}, ...y_{t-\rho}) \tag{3.14}$$

where y_t is the value of some variable at time t, and $\hat{y_t}$ is our estimate of the variable at time t. A univariate time series analysis could propose an autoregression model, which uses a linear function for $f(.)$

$$y_t = c + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \ldots + \alpha_\rho y_{t-\rho} + e_t \tag{3.15}$$

where e_t is the model error, $\hat{y_t} - y_t$. We can write this more succinctly as

$$y_t = c + \sum_{j=1}^{\rho} \alpha_j y_{t-j} + e_t \tag{3.16}$$

(Sims, 1980) expanded this model into a multivariate system. With K economic variables, y_k, each could be explained as linear combinations of the past values (or lagged versions, as they are often called in terms of VAR models) of itself and all the other variables under consideration. That is, any variable y_k could be modeled as

$$y_{k,t} = c_k + \sum_{j=1}^{\rho} \alpha_{k1,j} y_{1,t-j} + \sum_{j=1}^{\rho} \alpha_{k2,j} y_{2,t-j} + \ldots + \sum_{j=1}^{\rho} \alpha_{kK,j} y_{K,t-j} + e_{k,t} \tag{3.17}$$

which, again, can be expressed more succinctly as

$$y_{k,t} = c_k + \sum_{i=1}^{K} \sum_{j=1}^{\rho} \alpha_{ki,j} y_{i,t-j} + e_{k,t} \tag{3.18}$$

This representation lends itself to vector and matrix notation. Since it will also help explain hope to solve for the regression coefficients later, we will rewrite the last equation in matrix form. We gather the values of all K economic variables at time t into a vector, y_t, and similarly for all constants and error terms: $c = (c_1, \ldots, c_K)$ and $e_t = (e_{1,t}, \ldots, e_{K,t})$. Then, if we collect all the regression coefficients $\alpha k, j$ into a matrix

$$A_j = \begin{pmatrix} \alpha_{11,j} & \cdots & \alpha_{1K,j} \\ \vdots & \ddots & \vdots \\ \alpha_{K1,j} & \cdots & \alpha_{KK,j} \end{pmatrix} \tag{3.19}$$

we can rewrite the VAR model as

$$y_t = c + \sum_{j=1}^{\rho} A_j y_{t-j} + e_t \tag{3.20}$$

This paper's motivation was dealing with what it termed "incredible" identification problems with autoregression and structural models. It seems obvious that in the economic system, different variables interact and contain information about each other. However, autoregressive models ignore other variables, instead using only past values of itself to model any variable. VAR models are an attempt to incorporate those interactions between variables within models.

VAR models also provide an additional benefit of describing those inter-variable relationships. With the discovered regression coefficients, economists can describe whether the correlation between variables is positive or negative, and a sense of the multiplicative factor. And, providing our data fits our assumptions, we can perform standard statistical tests to see the significance of these relationships.

3.3.1 Assumptions

As with any model, there are certain underlying assumptions behind VAR models. The basic assumptions are that the data is stationary and normal. These assumptions ensure our models are well built and that our tests for significance are sane.

The error terms give us an easy way to check if the variables conform to these assumptions. They should have the following properties:

- $E(e_t) = 0$ - all error terms should have zero mean
- $E(e_t, e_t - k) = 0$ for any non-zero k - There is no serial correlation for any of the error terms

- $E(e_{k,t}e_{i,t}) = \Omega$ - the contemporaneous covariance matrix is positive semi-definite.

This last property can be changed to equal zero with an adjustment to the VAR model - see Recursive VAR models below.

This is why we work with the log returns of the oil prices instead of the prices themselves. Like most economic variables, the straight prices violate the stationary assumption - there is a clear trend in the data. The log part of the log returns removes trends and gives some sense of stationarity.

Of course, the log returns of many economic variables still violate the stationary condition - their second order dynamics, such as volatility, vary with time. In this case, we need to consider the interaction between the variables. If variables are cointegrated (see 3.4.1 for explanation), we need to change to a Vector Error Correction model. (see 3.3.5)

If the variables are not cointegrated, we simply need to difference the variables d times, where d+1 is the max order of non-stationary. For economic variables, d typically is 1, hence why we work with log returns.

3.3.2 Estimating the regression coefficients

As with autogression, the regression coefficients are estimated using ordinary least squares criteria. This procedure is described below.

To help with ease of notation, let us first write out the VAR model in what is called General Matrix Notation. If we have in total T observations of each of K variables and we are modelling with ρ lags. If in addition, we have ρ pre-sample observations, $y_{-\rho+1}$ to y_0 we can define

$$Y = [y_1 \dots yT] = \begin{pmatrix} y_{1,1} & \cdots & y_{1,T} \\ \vdots & \ddots & \vdots \\ y_{K,1} & \cdots & y_{K,T} \end{pmatrix}$$

$$B = [c, A_1 \dots A\rho] = \begin{pmatrix} c_1 & a11,1 & a12,1 & \cdots & a_{1K,1} & \cdots & a_{11,\rho} & a_{12,\rho} & \cdots & a_{1K,\rho} \\ c_2 & a21,1 & a22,1 & \cdots & a_{2K,1} & \cdots & a_{21,\rho} & a_{22,\rho} & \cdots & a_{2K,\rho} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ c_K & aK1,1 & aK2,1 & \cdots & a_{KK,1} & \cdots & a_{K1,\rho} & a_{K2,\rho} & \cdots & a_{KK,\rho} \end{pmatrix}$$

$$Z = [Z_0 \dots Z_t \dots ZT-1] = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ y_0 & y_1 & \cdots & y_{T-1} \\ \vdots & \ddots & \vdots & \\ y_{-\rho+1} & y_{-\rho+2} & \cdots & y_{T-\rho} \end{pmatrix}$$

$$= \begin{pmatrix} 1 & 1 & \cdots & 1 \\ y_{1,0} & y_{1,1} & \cdots & y_{1,T-1} \\ y_{2,0} & y_{2,1} & \cdots & y_{2,T-1} \\ \vdots & \vdots & \ddots & \vdots \\ y_{K,0} & y_{K,1} & \cdots & y_{K,T-1} \\ y_{1,-1} & y_{1,0} & \cdots & y_{2,T-1} \\ y_{2,-1} & y_{2,0} & \cdots & y_{K,T-1} \\ \vdots & \vdots & \ddots & \vdots \\ y_{K,-1} & y_{K,0} & \cdots & y_{K,T-1} \\ \vdots & \vdots & \ddots & \vdots \\ y_{1,-\rho+1} & y_{1,-\rho+2} & \cdots & y_{2,T-1} \\ y_{2,-\rho+1} & y_{2,-\rho+2} & \cdots & y_{K,T-1} \\ \vdots & \vdots & \ddots & \vdots \\ y_{K,-\rho+1} & y_{K,-\rho+2} & \cdots & y_{K,T-\rho} \end{pmatrix}$$

$$U = [u_1 \dots uT] = \begin{pmatrix} u_{1,1} & \cdots & u_{1,T} \\ \vdots & \ddots & \vdots \\ u_{K,1} & \cdots & u_{K,T} \end{pmatrix}$$

With these matrices, we can rewrite our VAR model as

$$Y = BZ + U \tag{3.21}$$

or as

$$\begin{aligned} vec(Y) &= vec(BZ) + vec(U) \\ &= (Z' \otimes I_K) vec(B) + vec(U) \end{aligned} \tag{3.22}$$

where $\otimes$ is the Kronecker product, vec(A) is the vectorization of any matrix A and I_K is the identity matrix of size K. In which case, the ordinary least squares estimate for vec(B) minimizes

$$\begin{aligned}
S(vec(B)) &= vec(U)'(I_T \otimes \Sigma_u)^{-1} vec(U) \\
&= vec(U)'(I_T \otimes \Sigma_u^{-1}) vec(U) \\
&= [vec(Y) - (Z' \otimes I_K)vec(B)]'(I_T \otimes \Sigma_u)^{-1}[vec(Y) - (Z' \otimes I_K)vec(B)] \\
&= vec(Y - BZ)'(I_T \otimes \Sigma_u)^{-1} vec(Y - BZ) \\
&= tr[(Y - BZ)'\Sigma_u)^{-1}(Y - BZ)] \\
&= vec(Y)'(I_T \otimes \Sigma_u^{-1})vec(Y) + vec(B)'(ZZ' \otimes \Sigma_u^{-1})vec(B) \\
&- 2vec(B)'(Z \otimes \Sigma_u^{-1})vec(Y)
\end{aligned} \tag{3.23}$$

Differentiating with respect to vec(B) and equating to zero gives us the estimator for B

$$\begin{aligned}
\widehat{vec(B)} &= ((ZZ^{-1} \otimes \Sigma_u)(Z \otimes \Sigma_u^{-1})vec(Y) \\
&= ((ZZ^{-1}Z \otimes I_K)vec(Y) \\
&= YZ'(ZZ')^{-1}
\end{aligned} \tag{3.24}$$

This method is specific to the basic VAR model laid out previously, but is easily adapted to the models described below. This derivation is laid out in full here as this basic model was the one used for this thesis.

3.3.3 Selecting ρ

Up to now, we have been ignoring ρ, the maximum number of lags to be considered in the model, taking it as a given. Called the order of the VAR model, it is a user defined parameter (within the limits of the data, of course), and its choice must be addressed.

Of course, if ρ is a correct summary of the variables, $\rho + 1$ will also be a valid model, as the last variable can always be ignored. We therefore generally look for the minimum ρ that explains the data well.

Choosing ρ is essentially a model comparison choice. Fit different VAR models to the variables of order 1 to some chosen N, and use a model comparison statistic to chose the best order. N is typically chosen from eight to ten, and rarely greater than twenty. There are a number of model comparison metrics that fit well with the assumptions underlying the VAR models. Common choices are the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC) and the Log Likelihood Ratio Test.

AIC can be expressed as $2k - 2\ln(L)$ where k in the number of parameters in the model and L is the likelihood of the model. The criterion is based on the trade-off between likelihood of model against complexity and overfitting. In our context, AIC is given by

$$AIC = \ln|\hat{\Sigma_u}| + \frac{2\rho K^2}{T} \tag{3.25}$$

where $\hat{\Sigma_u}$ is an estimate of the error terms covariance matrix. The best model will minimize its AIC.

3.3.4 Types of VAR Models

There are three main kinds of VAR models, with some additional variations covered in 3.3.5 The kind discussed so far is called the reduced form. The reduced form is the simplest version, and easiest to implement. One of its drawbacks is that if the variables being modeled are correlated, the error terms at any time t will also be correlated. This correlation in the error terms prevents certain kind of analysis, such as the study of a response of other variables to a shock in one variable.

The recursive form removes this correlation by including contemporaneous terms of some of the other variables into the model of each of the variables. (Clearly, it makes no sense to include the contemporaneous term of a variable itself in its own model.) However, if we include all other variable's contemporaneous terms, we will no longer be able to estimate the regression coefficients using OLS, as the error terms will now be correlated with the regressors.

To see this, for simplicity say we have two variables, income and interest rates, and we choose $\rho = 1$. The VAR model with contemporaneous values would be

$$y_t = -a_{yr}r_t + \alpha_{yr,1}r_{t-1} + \alpha_{yy,1}y_{t-1} + e'_{y,t}$$
$$r_t = -a_{ry}y_t + \alpha_{ry,1}y_{t-1} + \alpha_{rr,1}r_{t-1} + e'_{r,t} \tag{3.26}$$

The correlation between Y and the error term is given by

$$\begin{aligned}
cov(y_t, e'_{r,t}) &= cov(-a_{yr}r_t + \alpha_{yr,1}r_{t-1} + \alpha_{yy,1}y_{t-1} + e_{y,t}, e'_{r,t}) \\
&= cov(-a_{yr}(-a_{ry}y_t + \alpha_{ry,1}y_{t-1} + \alpha_{rr,1}r_{t-1} + e'_{r,t}) \\
&+ \alpha_{yr,1}r_{t-1} + \alpha_{yy,1}y_{t-1} + e_{y,t}, e'_{r,t}) \\
&= a_{yr}a_{ry}cov(y_t, u_{rt}) - a_{yr}\sigma^2_{ur} \\
&= \frac{-a_{yr}}{1 - a_{yr}a_{ry}}\sigma^2_{ur}
\end{aligned} \tag{3.27}$$

a non-zero value, unless $-a_{yr}$ is equal zero.

The recursive VAR model gets around this problem by recursively setting those values to be zero. It is easiest explained with an example:

If we have a three-variable VAR, with inflation, unemployment rate and interest rates. In the first equation of the recursive VAR for inflation, only lagged values of the variables are used, no contemporaneous values. In the second equation for unemployment rate lagged values of all three variables are also used, plus the current value of the inflation rate. The final equation for interest rate includes lagged values and the current values for both inflation and the unemployment rate. Estimation of each equation by OLS produces residuals that are uncorrelated across equations.

Yes, the recursive VAR does depend on the order of the variables, and in general there are K! different models for a given chosen ρ.

The third type is the structural VAR. This allows the inclusion of some previously know economic knowledge or theory to set parameters of the model. One example, taken from (Stock and Watson, 2001) could be the "Taylor Rule", where the Federal Reserve sets the interest rate based on output gap and inflation. This would change our model for interest rate to

$$R_t = r^* + 1.5(\bar{\pi}_t - \pi^*)1.25(\bar{u}_t - u^*) + lagged values of R, \pi, u + e'_t \tag{3.28}$$

where r^*, π^* and u^* are the desired values of interest, inflation and output gap respectively and $\bar{\pi}_t$ and $\bar{u}_t$ are the average values of inflation and the output gap of the past four quarters.

3.3.5 Variations

As mentioned previously, there are several variations on the VAR model that incorporate additional features or handle different kinds of data.

VAR Moving Average (VARMA) incorporates a moving average aspect to the model, using lagged versions of the error terms to help explain the data. The VARMA equation becomes

$$y_t = c + \sum_{j=1}^{\rho} A_j y_{t-j} + \sum_{j=1}^{q} B_j e_{t-j} + e_t \tag{3.29}$$

This model gets rid of the requirement that error terms be sequentially uncorrelated, and instead actually takes advantage of this correlation to help explain the data. A stable and invertable VARMA can always be represented as a pure VAR or MA model.

We can also introduce exogenous variables. So far we have only discussed models that deal with endogenous variables - ones included in the model. These exogenous variables are not included in the model, but are still theorized to have some effect. In economic terms, exogenous variables are generally seen as “outside” the market in some way - unexpected weather or other natural occurrences, for example. A VARMA model with exogenous variables is shown below.

$$y_t = c + X_t b + \sum_{j=1}^{\rho} A_j y_{t-j} + \sum_{j=1}^{q} B_j e_{t-j} + e_t \tag{3.30}$$

where X_t is a k by r matrix, where r is the number of exogenous variables at time t. b is a vector of regression coefficients to be learned.

Mentioned before is that VAR models can only deal with stationary processes. If a process is not stationary, the are ways of processing the data to give it stationarity. These techniques leave the relationships we are studying intact, unless the variables are cointegrated. While a full discussion of what cointegrated means can be found in section

3.4.1, here we will lay out how to adapt a VAR model to a Vector Error Correction (VEC) model, which can handle this kind of data.

The VEC assumes all variables have an equilibrium relationship to each other, where they would remain if there were no changes. The key idea in a VEC model is modelling the speed at which different variables return to this equilibrium after a change in another variable. As the number of variables grow, it becomes harder to isolate changes in one variable, and thus estimates of these speeds becomes more difficult.

It is easiest again to explain with an example: Our two variables are two prices of the same good in different markets, y_1 and y_2. Assume at equilibrium, $y_{1t} = \beta y_{2t}$. If we assume for simplicity we include only one lag in our model,

$$\Delta y_{1t} = \alpha_1(y_{1,t-1} - \beta_1 y_{2,t-1}) + e_{1t}. \tag{3.31}$$

and similarly for y_{1t}. If we expand this to include changes of both variables, we have a VEC.

$$\Delta y_{1t} = \alpha_1(y_{1,t-1} - \beta_1 y_{2,t-1}) + \gamma_{11,1}\Delta y_{1,t-1} + \gamma_{12,1}\Delta y_{2,t1} + e_{1t}. \tag{3.32}$$

For a more in depth discussion of all the points covered here about VAR models, see this excellent book on the topic (Lütkepohl, 1991).

3.4 Granger Causality Tests

The Granger Causilty test is a test to try to move from correlation towards causation. It was first proposed in (Granger, 1969) and is a test for a simple idea. If a variable y_1 causes another variable y_2, any change in y_1 should be seen before y_2. Putting this in similar terms we used to describe VAR models, lagged values of y_1 should help model the current value of y_2. We can write very similar formal equations as we did for the VAR model

$$y_{1,t} = c_1 + \sum_{j=1}^{\rho} a_{11,j} y_{1,t-j} + \sum_{j=1}^{\rho} a_{12,j} y_{2,t-j} + e_{1t}$$
$$y_{2,t} = c_2 + \sum_{j=1}^{\rho} a_{21,j} y_{1,t-j} + \sum_{j=1}^{\rho} a_{22,j} y_{2,t-j} + e_{2t} \tag{3.33}$$

where the error terms have similar properties that they did for VAR models. Note how both sums consider the same amount of lagged values.

This gives us two null hypotheses. The first is

$$H_0 : a_{12,1} = a_{12,2} = \ldots = a_{12,\rho} \tag{3.34}$$

which states that y_1 does not Granger cause y_2. The second

$$H_0 : a_{21,1} = a_{21,2} = \ldots = a_{21,\rho} \tag{3.35}$$

states that y_2 does not Granger cause y_1. In order to say that y_1 Granger causes y_2, we must not only reject the first null hypothesis, but not reject the second, and vice versa. If both null hypotheses are rejected, we say there is feedback between the variables. If neither is rejected, the results are inconclusive.

It should be noted that Granger hedged his bets with his test, not calling his test a test for complete causality. Two variables that were both caused by a third variable could pass the Granger causality test if one responded quicker than the other. Instead, he and further researchers would refer to Granger causality, and state one variable Granger caused another. Only knowledge about the variable and the environment can establish true causality.

As should be obvious, the Granger test is very similar to the VAR model. The Granger Causality test can be performed using a series of F-tests on the VAR model, testing each pair of variables. This is why the Gretl software performs Granger causality tests automatically when asked to fit VAR models.

3.4.1 Cointegration

One important topic to both VAR models and the Granger causality test is the idea of cointegration. This idea is simple - if there exists some linear combination of a set of variables that is stationary, they are said to be cointegrated. This does not exclude the idea that the variables are individually integrated i.e. a linear combination of lagged versions of itself is stationary.

A classic example of cointegration variables is the price of one good in two separate markets. If two variables are cointegrated, they share a common stochastic drift, and move together in some way. The two individual variables might exhibit random walk behavior, but still be cointegrated, see the diagrams below for examples of this.

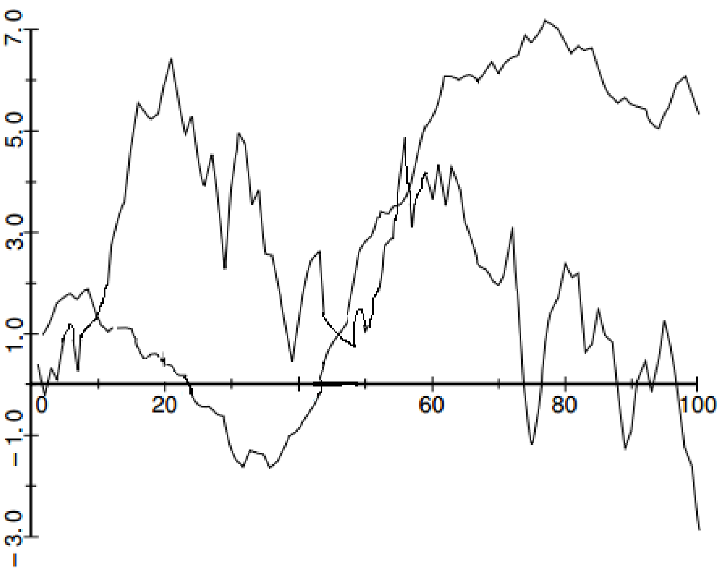


FIGURE 3.8: Two artificially generated random walks overlain. Diagram taken from (Lütkepohl, 1991)

The main test for cointegration is the two step Engle-Granger test (Engle and Granger, 1987). If two variables are cointegrated, $y_{1,t} - \beta y_{2,t} = u_t$ and u_t would be stationary, by definition. However, we do not know β, so the first step is estimating it by ordinary least squares. We can then run one of a number of stationary tests on u_t, for example the Dickey-Fuller test. (Dickey and Fuller, 1979) The test is calculated as

$$\Delta y_t = c + bt + (\gamma - 1)y_{t-1} + \sum_{i=1}^{\rho} \Delta y_{t-i} + e_t \tag{3.36}$$

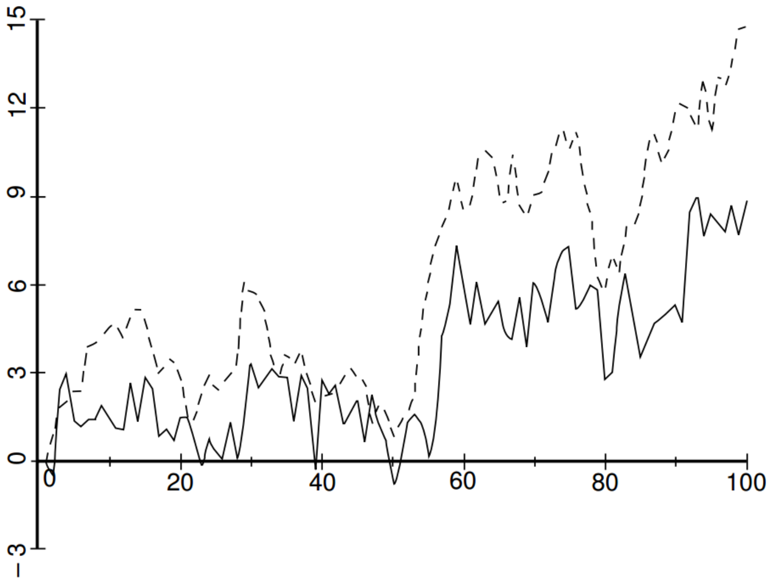


FIGURE 3.9: A bivariate cointegrated time series. Diagram taken from (Lütkepohl, 1991)

where $\Delta y_t = y_t - y_{t-1}$. If the null hypothesis of $\gamma = 1$ is not rejected, the variable is stationary.

The adjustment needed to change the VAR model to VEC in the case of conintegration has already been explained. The Granger Causality test gets a similar adjustment, with an added term to each equation.

$$\begin{aligned} y_{1,t} &= c_1 + \delta_1(y_{1,t-1} - y_{2,t-1}) + \sum_{j=1}^{\rho} a_{11,j} y_{1,t-j} + \sum_{j=1}^{\rho} a_{12,j} y_{2,t-j} + e_{1t} \\ y_{2,t} &= c_2 + \delta_2(y_{1,t-1} - y_{2,t-1}) + \sum_{j=1}^{\rho} a_{21,j} y_{1,t-j} + \sum_{j=1}^{\rho} a_{22,j} y_{2,t-j} + e_{2t} \end{aligned} \tag{3.37}$$

where δ_1 and δ_2 represent the speed of adjustment of both variables. The null hypotheses both change slightly as well - they now both must check their respective $\delta = 0$ as well to hold.

(Engle and Granger, 1987) is the work that brought to attention that this required adjustment, and the hazards of trying to fit linear models to cointegrated variables.

3.4.2 Thesis Implementation

The VAR models and the Granger causality tests were implemeneted using the Open Source statistics software package Gretl. Like with the wavelet transformations and Matlab, there was no code written. Once the output of the wavelet transformations were input to Gretl, the models and tests were built using the software's various GUIs.

Three main functions were used within Gretl. First, Gretl provides a cointegration test, implementing the two step Engle-Granger test. The second is the VAR lag select tool, used to select the best ρ for the VAR model. Gretl builds a different model for the values of ρ between one and N (in our case ten) and calculates the best fit model. Gretl provides three separate model criteria for comparison: Akaike Information Criterion, Schwarz Bayesian Criterion and Hannan-Quinn criterion. The model with the best AIC was chosen for our model.

The third function actually fit the VAR models, as well as automatically computing the Granger causality tests for each pair of variables.

Chapter 4

Results and Analyis

In this chapter the results of various stages of the process are presented. In section 4.2 the decomposition of various time series into different timeframes using the methods described in 3.1.5 are shown. Section 4.3 deals with the results of the different vector autoregressive (VAR) models fit to the wavelet decompositions, and finally covers the Granger Causality Tests done using the VAR models.

4.1 Data Used

The oil prices used were the Cushing, OK WTI Spot Price FOB (Dollars per Barrel) obtained from the United States Energy Information Adminsitration (Government, 2014). The infromation was from January 2nd 1986 until January 27th 2014, though only data from January 4th 2000 was used in this work. As discussed before, the work was not done on the prices directly, but instead used the log returns as a basis.

This is because reliable online news data is only available from 2000. As previously mentioned, the news aggregater site LexisNexis was used to get raw news articles for this project. The search was limited to three news sources: The Finanical Times, Wall Street Jounral and The New York Times. The timeline matched the one used for the oil prices, January 4th 2000 until January 27th 2014. The only search terms used were “Crude Oil”, which could appear anywhere in the headline or body of the article.

4.2 Wavelet Decompositions

Using wavelets to decompose our time signals into seperate processes is crucial to this thesis. As already discussed in 2.1 there is good reason to think that economic relationships in particular will not remain constant across different timeframes, and these dynamics will be lost when looking at the overall time series.

The decompositions here use maximum overlap discrete wavelet transform (MODWT) to obtain the wavelet coefficents, before using those values to create the wavelet details, as laid out in 3.1.5.

The figure 4.1 shows the decompostion of the returns on the oil spot prices down to ten levels. We used ten levels to show the full range of activity in the market, from daily trading to the yearly business cycle (level ten shows activity of the order $2^9 = 512days \approx$ 1.5 years)

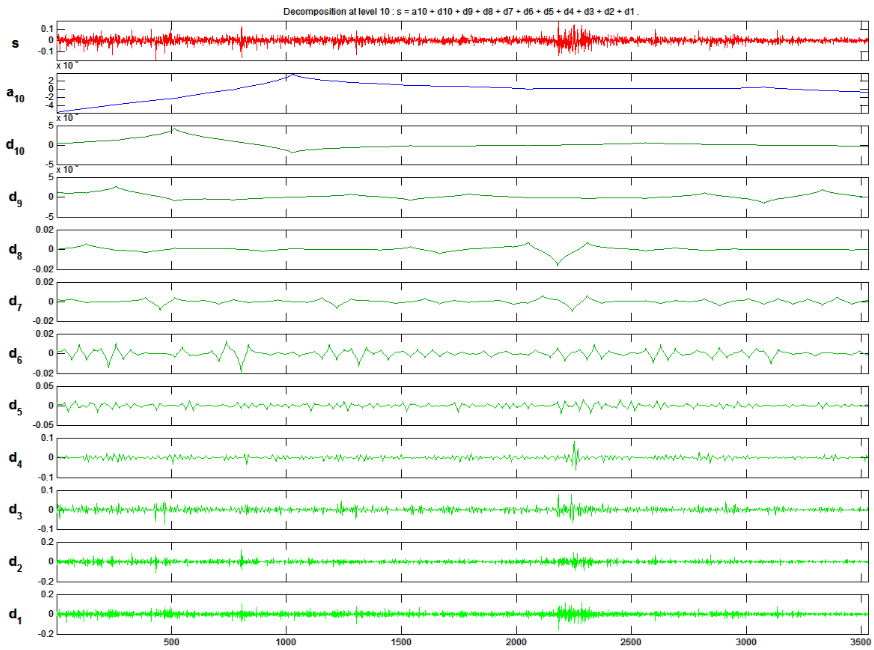


FIGURE 4.1: Details obtained using the DWT of the returns of oil prices. The top graph, s, is the original signal, a10 is the level 10 smooth and the rest of the graphs show the details in decending level.

The results presented here and below used the Coiflet (5), though work was done using the Daubchies (4) and Mexican Hat wavelets. The figure 4.2 shows the same results but using the Daubchies (4) wavelet. As can be seen, the produced details do vary, espeically at higher levels, but as mentioned previously, the use of different smooth wavelets do not effect the overall outcome of the VAR and Granger Causality tests.

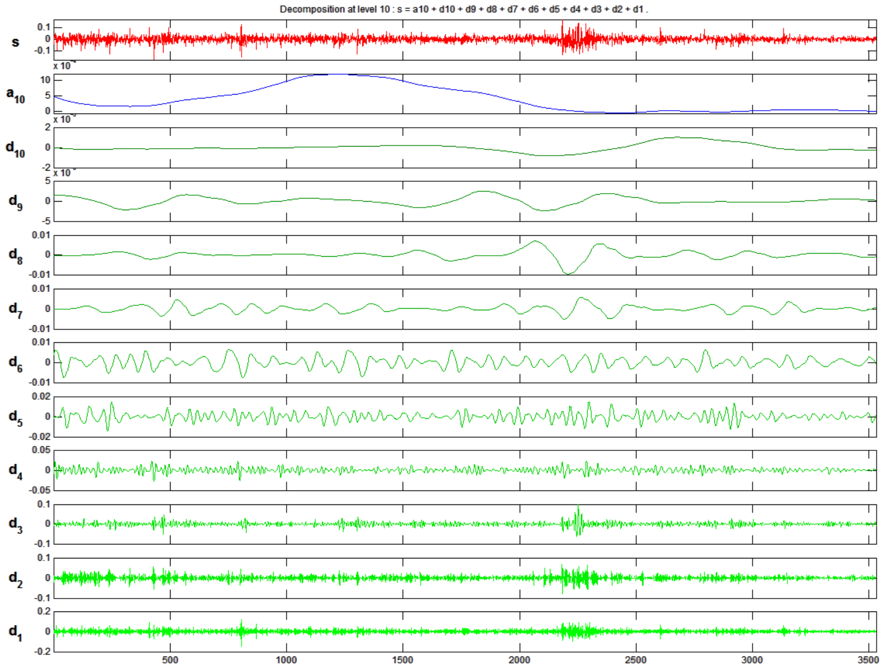


FIGURE 4.2: Details obtained using the DWT of the returns of oil prices using the Daubchies (4) wavelet.Graphs are laid out as above

Next we present the decompostion of the negative sentiment to the same amount of levels as the oil returns.

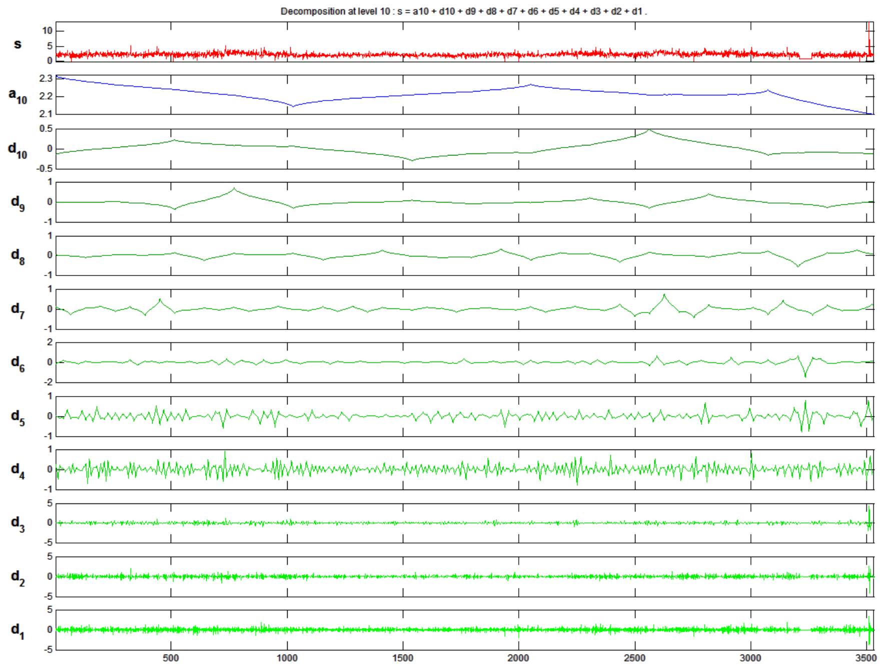


FIGURE 4.3: Details obtained using the DWT of the negative sentiment

The details of the two different time series look as we might expect them to. The lower levels are more noisy with their high frequency data, whereas the higher scales and smooth show much smoother behaviour.

Note how the period of high volatility in the oil returns between day 2000 and 2500 appears in details 1 to 4, but not in the higher levels. This shows the power of timescale decompostion - that event can be clearly seen to be caused by short scale activity with this decomposition.

Finally, we present the wavelet decomposition of the moving average of both the oil returns and negative sentiment, used to try to remove some of the outliers in the time series, such as the spike of negative sentiment towards the end of the signal.

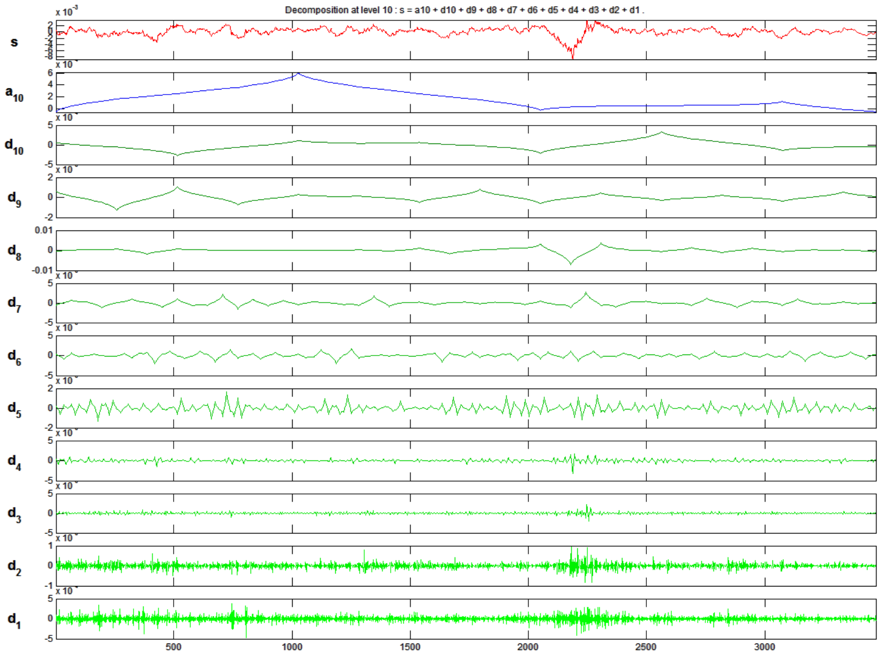


FIGURE 4.4: Details obtained using the DWT of the moving average of the returns of oil prices.

For tables describing all individual values of the wavelet details, please see the Appendix.

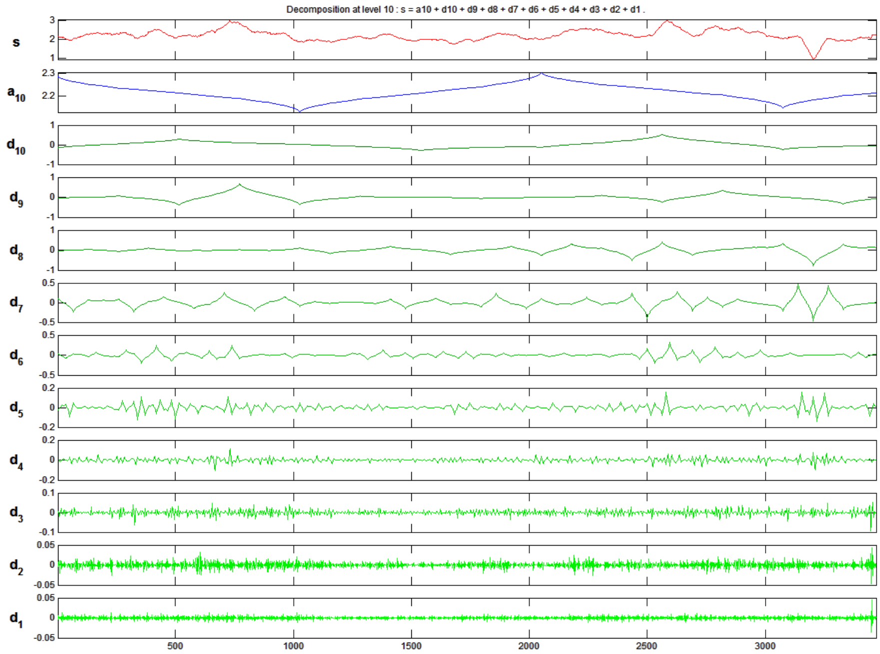


FIGURE 4.5: Details obtained using the DWT of the moving averager of the negative sentiment

Descriptive Statistics

We can take a closer look at the oil returns decomposition using descriptive statistics. As discussed previously in section 3.1.6, any statistics of higher order than the mean use the wavelet coefficents rather the details.

Summary Statistics for oil returns 2000/01/04–2014/01/24

Variable	Mean	Median	Minimum	Maximum
D1	2.67e–6	0.000140	−0.157	0.122
D2	−5.66e–6	−0.000101	−0.110	0.121
D3	4.28e–6	6.90e–05	−0.0764	0.0807
D4	1.41e–5	3.82e–05	−0.0659	0.0845
D5	1.91e–5	0.000148	−0.0222	0.0163
D6	3.04e–5	0.000132	−0.0190	0.0126
D7	3.83e–5	0.000189	−0.0107	0.00657
D8	7.85e–5	0.000105	−0.0163	0.00771
D9	0.000146	2.16e–5	−0.00163	0.00276
D10	0.000261	−1.15e–5	−0.00211	0.00430
Returns	0.000374	0.00123	−0.171	0.164

Summary Statistics for oil returns, 2000/01/04–2014/01/24 continued

Variable	Std. Dev.	Mean/Std. Dev.	Skewness	Ex. kurtosis
D1	0.0178	1.50e–4	−0.122	6.17
D2	0.0124	−4.57e–4	0.0147	11.3
D3	0.00843	5.08e–4	0.00567	12.2
D4	0.00621	2.269e–3	0.533	23.2
D5	0.00378	5.046e–3	−0.687	3.46317
D6	0.00275	0.0110	−0.835	4.89
D7	0.00177	0.0216	−0.997	4.20
D8	0.00222	0.0354	−2.09	11.9
D9	0.000661	0.222	0.952	1.10
D10	0.00100	0.260	1.41	2.47
Returns	0.0247	0.0151	−0.246	4.84

If we look at the ratio of mean to variance, we see that as you go up the levels the details become more mean dominated. The activity at lower frequencies is more predictable with less variance. We can interpert this as details of D6 and higher become less susceptible to outside influences, and will mostly continue to their patterns.

This makes intuitive sense - the detail at level 6 corresponds to activity of 32 (2^5) days, or roughly monthly data. The data is moving away from the higher frequency and variablity of daily traders, and toward more market fundementals. While we might expect sentiment to have an impact on lower level details, we would be surprised if levels higher then this showed influence, mean dominated processes that they are.

4.3 VAR Models and Granger Tests

In this section, the results of fitting the VAR models and the Granger Causality tests are presented. Three types of data had VAR models fitted: the wavelet decomposition of the returns compared with the raw negative sentiment series, the decomposition of both the returns and the negative sentiment and finally the decomposition of a sixty day moving average calculation of both returns and negative sentiment.

Each subsection below deals with these data types in turn. Of course, not all results are shown here. Each subsection fits ten separate VAR models, and showing each here would be excessive. Instead, a choice of the VAR models are presented here for illustration. However, the full set of the VAR models are available in the Appendix.

The subsections all follow a similar structure. First, the test for the choice of best number of lags is shown. Secondly, the results of the Granger Causality tests are presented, indicating where a relationship between sentiment and the market exists. Thirdly, the results of the VAR models themselves, including which of the regression coefficients are statistically significant, are presented. Finally, a brief discussion on what those results indicate completes each subsection. The justification of why these sets of data were used is given in this discussion.

Note that this is not the order of the method, where VAR models are fit before the Granger Causality tests are done, as once the VAR models are fitted the Granger Causality tests are simply F-tests on the regression coefficients. The results are presented this way as not all ten sets of VAR coefficients are presented here, as this would take up too much space. Instead, the VAR models are shown based on the Granger tests, as they show a closer look of the nature of the relationship the Granger tests have discovered. This was felt to be the most intuitive way to present the results.

All variables were tested for cointegration using the Engle-Granger test, outlined in 3.4.1, and no pairing showed signs of cointegration. As such, VAR models were used instead of Vector Error Correction, and Granger Causality tests could be used unaltered. Since all tests for cointegration were negative, this is not mentioned again below.

4.3.1 Just Returns Decomposed

In this subsection, we look at the relationship between the prices and sentiment by decomposing the data on the oil returns into ten levels, but leaving the sentiment as the raw untransformed data.

Choice of Lags

First, the choice of lags results from Gretl. As discussed in 3.4.2, Gretel provides the Akaike criterion, Schwarz Bayesian criterion and Hannan-Quinn criterion as criteria to choose the best number of lags. Up to a maximum number of ten lags were checked, to avoid overfitting. Besides one exception, all levels suggested the maximum ten lags as the best model. Typical examples were level 1 and level 9. The stars indicate the best choice, based on the different criteria.

Best Lag Selection for VAR System, Maximum order 10. Variables: Negative Sentiment and Detail 1 of Oil Returns

Lags	AIC	BIC	HQC	Choice
1	−1.75	−3.74	−3.75	
2	−4.06	−4.04	−4.05	
3	-4.243	-4.22	-4.23	
4	-4.36	-4.33	-4.35	
5	-4.46	-4.42	-4.44	
6	-4.53	-4.48	-4.51	
7	-4.57	-4.52	-4.55	
8	-4.60	-4.54	-4.58	
9	-4.64	-4.57	-4.62	
10	-4.67	-4.59	-4.76	***

Best Lag Selection for VAR System, Maximum order 10. Variables: Negative Sentiment and Detail 9 of Oil Returns

Lags	AIC	BIC	HQC	Choice
1	-18.38	-18.37	-18.38	
2	-19.52	-19.49	-19.51	
3	-19.48	-19.46	-19.47	
4	-19.56	-19.53	-19.55	
5	-19.57	-19.54	-19.56	
6	-19.62	-19.58	-19.61	
7	-19.63	-19.58	-19.61	
8	-19.63	-19.57	-19.61	
9	-19.64	-19.57	-19.62	
10	-19.68	-19.61	-19.66	***

The exception was the VAR model using the level six, for which the different criteria choose different lags.

Best Lag Selection for VAR System, Maximum order 10. Variables: Negative Sentiment and Detail 6 of Oil Returns

Lags	AIC	BIC	HQC	Choice
1	-11.34	-11.33	-11.33	
2	-12.07	-12.05	-12.06	
3	-12.10	-12.08	-12.09	
4	-12.14	-12.11	-12.13	
5	-12.16	-12.21	-12.15	*(BIC)
6	-12.28	-12.12	-12.21	*(AIC) *(HQC)
7	-12.21	-12.16	-12.19	
8	-12.22	-12.15	-12.19	
9	-12.22	-12.16	-12.20	
10	-12.22	-12.16	-12.19	

The decision made was to use the AIC to determine which lag order to use i.e. six.

Granger Causality Tests

Presented in the table below are the Granger Causality tests for each level of the decomposed returns with the negative sentiment values. The probability of retaining the null hypothesis of no causality indicates how statistically significant the relationship is. S in the table refers to sentiment, R returns, DX is the detail of level X.

	Results	Null Hypotheses	
		$S \not\Rightarrow R$	$R \not\Rightarrow S$
D10 (10 lags)	inconclusive	0.429	0.191
D9 (10 lags)	inconclusive	0.131	0.623
D8 (10 lags)	inconclusive	0.879	0.906
D7 (10 lags)	inconclusive	0.710	0.248
D6 (6 lags)	inconclusive	0.396	0.487
D5 (10 lags)	inconclusive	0.729	0.906
D4 (10 lags)	inconclusive	0.901	0.288
D3 (10 lags)	inconclusive	0.974	0.586
D2 (10 lags)	inconclusive	0.130	0.382
D1 (10 lags)	inconclusive	0.566	0.758

TABLE 4.1: Table showing the results of Granger Causality tests on different wavelet levels. The values shown are probabilities of keeping the Null Hypotheses.

VAR Models

The tables below show the VAR models for some of the levels. Each VAR model generates two linear equations; one for each variable. Shown in the table is the regression coefficient for each of the lagged values used in the equation. The number of lags for the value is indicated in the brackets. This is followed by the standard error, a t-test and its statistical significance, represented as the probability (p-value) of the null hypothesis that the coefficient is used in the equation is. Statistically significant values are shown for ease of reading with stars - One star for significance of confidence 10%, two for confidence of 5%, and three for confidence of 1%. All tables of VAR models will follow similar structures. To save space, sometimes only the equation for one variable is shown, if this is enough to illustrate the point. But of course each VAR model creates two equations, one for each variable

As can be seen, the Granger Causality tests show no significant causal relationships at any level. When there is no Granger causality between two variables, the VAR model generally like table 4.2 Only lagged values of the variable itself have statistically significant regression coefficients, with lagged values of the of variable ignored in the model.

Equation 1: Returns Level 8

Lagged Value	Coefficient	Std. Error	T-ratio	P-value	Significance
Constant	2.70e-6	4.45e-6	0.608	0.5434	
Return (1)	1.73	0.0165	104.6	0.00	***
Return (2)	0.812	0.0328	24.76	0.00	***
Return (3)	0.301	0.0352	8.53	2.03e-17	***
Return (4)	0.471	0.0355	13.28	2.71e-39	***
Return (5)	0.492	0.0361	13.62	3.58e-41	***
Return (6)	0.256	0.0361	7.07	1.86e-12	***
Return (7)	0.0747	0.0355	2.10	0.0355	**
Return (8)	0.297	0.0351	8.45	4.15e-17	***
Return (9)	0.438	0.0327	13.38	7.79e-40	***
Return (10)	0.202	0.0165	12.24	9.35e-34	***
Neg Senti (1)	1.09e-6	1.22e-6	0.893	0.372	
Neg Senti (2)	6.20e-7	1.22e-6	0.504	0.613	
Neg Senti (3)	3.55e-7	1.23e-6	0.288	0.773	
Neg Senti (4)	7.43e-7	1.23e-6	0.600	0.548	
Neg Senti (5)	9.40e-7	1.23e-6	0.760	0.447	
Neg Senti (6)	6.43e-7	1.23e-6	0.519	0.603	
Neg Senti (7)	1.25e-6	1.23e-6	1.01	0.311	
Neg Senti (8)	9.27e-7	1.23e-6	0.751	0.452	
Neg Senti (9)	1.11e-6	1.22e-6	0.904	0.366	
Neg Senti (10)	7.32e-7	1.22e-6	0.599	0.549	

Table 4.2: Table showing the equation detail 8 of oil returns from the fitted VAR model.

However, this is not always the case. Some VAR models showed some lagged values of the other variable were significant in the created equation, but usually only one and at a low significance level. This shows that there are cases where it cannot be said with confidence that one variable causes another, but that does not rule out some values playing a role in explaining the variable. But those roles are not statistically significant to say one variable causes the other. An example of this is the VAR for level five of returns and sentiment, where the nine lagged and ten lagged values of detail 5 of the oil returns are significant values in the equation for negative sentiment 4.3.

Discussion

The data used here goes against the procedure in (Ramsey and Lampart, 1997) and (Gençay et al., 2001), which decompose both variables being studied. This is covered in the next section. The thinking here was that the raw sentiment data, which is the

Equation 1: Negative Sentiment

Lagged Value	Coefficient	Std. Error	T-ratio	P-value	Significance
Constant	0.614	0.0612	10.02	2.51e-23	***
Returns (1)	17.4	16.0	1.09	0.274	
Returns (2)	39.0	29.7	1.31	0.189	
Returns (3)	24.8	31.7	0.783	0.433	
Returns (4)	18.0	32.0	0.560	0.575	
Returns (5)	16.7	32.6	0.511	0.609	
Returns (6)	6.09	32.6	0.186	0.852	
Returns (7)	9.15	32.0	0.285	0.775	
Returns (8)	34.2	31.6	1.07	0.280	
Returns (9)	53.2	29.6	1.79	0.073	*
Returns (10)	26.8	16.0	1.67	0.094	*
Neg Senti (1)	0.113	0.0168	6.73	1.92e-11	***
Neg Senti (2)	0.108	0.0169	6.39	1.78e-10	***
Neg Senti (3)	0.101	0.0170	5.96	2.64e-9	***
Neg Senti (4)	0.0436	0.0171	2.547	0.011	**
Neg Senti (5)	0.0926	0.0170	5.418	6.44e-8	***
Neg Senti (6)	0.0712	0.0170	4.169	3.13e-5	***
Neg Senti (7)	0.0575	0.0171	3.364	0.001	***
Neg Senti (8)	0.0418	0.0170	2.455	0.014	**
Neg Senti (9)	0.0408	0.0169	2.408	0.016	**
Neg Senti (10)	0.0514	0.0169	3.049	0.0023	***

Table 4.3: Table showing the equation for negative sentiment from the fitted VAR model of itself and detail 5 of oil returns.

sum of the activity at different levels, might have a direct effect on the market. After all, this is what the trader would interact with. The market data was decomposed to see how that effect might differ with timescales.

However, as the results show, the effect of the negative sentiment does not come through. The different timeframes of the sentiment interact with the different timeframes of the market differently, as we shall see below, and without decomposition these relationships are lost.

4.3.2 Both Returns and Sentiment Decomposed

The data used in this section is decomposed versions of both the oil returns and negative sentiment. Ten VAR models were built comparing corresponding levels, plus one more comparing the original raw time series.

Choice of Lags

Again, up to a maximum ten lags were considered for all VAR models, and again all tests showed the maximum ten lags was the best choice with one exception, which was the original untransformed time series. The results from this test are shown in the table below.

Best Lag Selection for VAR System, Maximum order 10. Variables: Negative Sentiment and Oil Returns

Lags	AIC	BIC	HQC	Choice
1	-2.75	-2.74	-2.75	
2	-2.80	-2.78	-2.79	
3	-2.83	-2.80	-2.82	
4	-2.83	-2.80	-2.82	
5	-2.85	-2.81	-2.84	
6	-2.86	-2.814	-2.84	*(BIC)
7	-2.86	-2.81	-2.85	
8	-2.86	-2.81	-2.847	*(HQC)
9	-2.86	-2.80	-2.85	
10	-2.870	-2.79	-2.84	*(AIC)

These results do not show a clear choice for the number of lags to use in the VAR model. However, having fit the VAR model with all three of the suggested number of lags, the result of Granger causality test is the same for all three. With this in mind, the results of the model using ten lags is shown below, though the other results are in the appendix.

Granger Causality Tests

Displayed in table 4.4 are the results of the Granger causality tests. The table is laid out as before.

The interesting results are naturally the levels that show a result other than inconclusive. Below a closer look at the VAR models of these levels is presented.

	Results	Null Hypotheses	
		S $\not\Rightarrow$ R	R $\not\Rightarrow$ S
D10 (10 lags)	inconclusive	0.999	0.997
D9 (10 lags)	inconclusive	1.000	1.000
D8 (10 lags)	inconclusive	0.998	0.999
D7 (10 lags)	inconclusive	1.00	0.999
D6 (10 lags)	inconclusive	1.00	1.00
D5 (10 lags)	inconclusive	0.974	0.753
D4 (10 lags)	Feedback	0.000	0.082
D3 (10 lags)	inconclusive	0.999	0.984
D2 (10 lags)	inconclusive	0.359	0.797
D1 (10 lags)	S $\Rightarrow$ R	0.048	0.329
Overall (10 lags)	R $\Rightarrow$ S	0.342	0.035

TABLE 4.4: Table showing the results of Granger Causality tests on different wavelet levels. The values shown are probabilities of keeping the Null Hypotheses.

VAR Models

Here the VAR models for all the relationships that show Granger Causality. The tables follow the same format as before.

Lagged Value	Coefficient	Std. Error	T-ratio	P-value	Significance
Constant	8.22e-7	0.000168	0.00488	0.996	
Neg Senit D1 (1)	0.000454	0.000807	0.562	0.573	
Neg Senit D1 (2)	0.00113	0.00137	0.825	0.409	
Neg Senit D1 (3)	0.00174	0.00186	0.937	0.348	
Neg Senit D1 (4)	0.00392	0.00220	1.78	0.074	*
Neg Senit D1 (5)	0.00491	0.00238	2.06	0.039	**
Neg Senit D1 (6)	0.00667	0.00238	2.80	0.005	***
Neg Senit D1 (7)	0.00718	0.00220	3.26	0.001	***
Neg Senit D1 (8)	0.00585	0.00186	3.13	0.001	***
Neg Senit D1 (9)	0.00451	0.00137	3.27	0.001	***
Neg Senit D1 (10)	0.00211	0.000807	2.61	0.009	***
Returns D1 (1)	1.41	0.0166	85.16	0.000	***
Returns D1 (2)	1.75	0.0283	62.01	0.000	***
Returns D1 (3)	1.85	0.0391	47.21	0.000	***
Returns D1 (4)	1.78	0.0470	37.98	0.000	***
Returns D1 (5)	1.59	0.0511	31.24	0.000	***
Returns D1 (6)	1.33	0.0511	26.09	0.000	***
Returns D1 (7)	1.01	0.0470	21.56	6.40e-97	***
Returns D1 (8)	0.708	0.0391	18.09	5.75e-70	***
Returns D1 (9)	0.412	0.0283	14.59	7.26e-47	***
Returns D1 (10)	0.161	0.0160	9.64	9.26e-22	***

TABLE 4.5: Table showing the equation for detail 1 of oil returns from the fitted VAR model of itself and corresponding detail of negative sentiment.

Rather then waste room showing the entire data for each VAR equation for level four and the overall series, which are available in the Appendix, for the other Granger significant models we will simply list which lagged values have statisical significance in each equation. This is a adequate summary of the VAR model, as the coefficents themselves are not of great concern of this thesis, but instead which lagged values are significant in the equations.

Equation for detail 4 of oil returns.

- Negative Senitment D4 lagged values 8-10. (All of 1% confidence level)
- Returns D4 lagged values 1-6. (All of 1% confidence level)
- Returns D4 lagged values 8-10. (All of 1% confidence level)

Equation for detail 4 of negative sentiment.

- Negative Senitment D4 lagged values 1-6. (All of 1% confidence level)
- Negative Sentiment D4 lagged values 8-10. (All of 1% confidence level)
- Returns D4 lagged value 8. (At 5% confidence level)
- Returns D4 lagged value 9. (At 1% confidence level)

Equation for original negative sentiment series.

- Constant (At 1% confidence level)
- Negative Senitment lagged values 1-3. (All of 1% confidence level)
- Negative Sentiment lagged value 4. (At 5% confidence level)
- Negative Sentiment lagged values 5-8. (All of 1% confidence level)
- Negative Sentiment lagged value 9. (At 5% confidence level)
- Negative Sentiment lagged value 10. (At 1% confidence level)
- Returns lagged value 2. (At 1% confidence level)
- Returns lagged values 5. (At 10% confidence level)
- Returns lagged values 8. (At 10% confidence level)

Discussion

With the wavelet decomposition, the relationships that were hidden in the previous subsection using the raw negative sentiment series now appear. What is interesting is that this relationship changes across different levels, as other relationships studied using this method have also done (Ramsey and Lampart, 1997).

At level one, the daily activity shows that sentiment causes returns, but this relationship falls away at higher levels two and three. This seems to suggest that daily traders do put signifance in the sentiment of financial news of their market, but here they are heeding the quick-fire daily reporting. The slightly longer recap reports of two or four days are not significant enough to hold sway.

It is not an immediate reaction to the news articles either - only the lagged values between four and ten lags of sentiment are significant in the VAR created linear equation. Traders wait between four and ten days before moving on the financial news. This does not conflict with our observation that twice daily activity (level 2) is unaffected by sentiment. The distinction between longer timescale activity, and daily activity from the past, muist be kept clear. This is the distinction drawn above between quickfire news and longer recap reports.

The feedback at level four is interesting. This scale is the activity of the timrscale $2^3 = 8$ days, or approximatly a week. This seems like the intutative scale where such feeback would happen. At this timescale, news paper articles are being printed about the week in review, detailing how the market has performed this week. There may also be a kind of trader who would repsond to those longer, more thoughtful articles, instead of the daily quickfire reports. The reporting on how the market repsonds to the reporting also influences the market.

It is also interesting that both directions of causailty, returns causing sentiment and sentiment causing returns, are delayed by about a little over a week, 8-10 days. This is a longer delay than the daily activity, where four days was the shortest wait. This would seem to fit what we might expect for the longer timescale activity at level 4 - traders and journalists that consider a week's worth of information would need a week to consider before acting on sentiment or returns.

Longer term activity, i.e. from levels five and up, all show no signs of Granger Causality. This confirms our hypothesis that sentiment mainly effects the variance in the market, which is mainly captured at the higher frequency lower levels. Data from longer time-frames are unaffected by sentiment, instead moving with market fundementals. The hypothesis would be that by the time a news article would report on a change in a market fundemental, the change would already be seen in the price.

The fact that the original time series of returns Granger causes sentiment shows how the sum of the activities of different timescales, which is what the undecomposed time series is, it does not nessacrily display the same relationships as those at the decomposed levels. This has already being discussed in relation to (Ramsey and Lampart, 1997) and (Michis, 2011). Their result showed no relationship for the overall time series, even though relationships exisited at different timescales.

Here, we can see the relationship discovered between the overall time series does not hold at the different timescales. This does not mean we ignore this result, or the results of the wavelet details. They are both valid, and show that the relationship between economic variables is complicated.

4.3.3 Moving Average Decomposed

In this subsection, before doing any wavelet transformation, a simple sixty day moving average was taken of both negative sentiment and oil returns. This was done using the formula:

$$\hat{y}_t = \sum_{i=1}^{60} \frac{y_{t-i}}{60} \tag{4.1}$$

As the formula suggests, the moving average shortens the data length by 60, as the formula is only valid at t=61. This was an attempt to remove any outliers in the data, such as the 9/11 data or the spike of negative sentiment at day 3520. The diagrams below show the results of the taking the moving average of the data.

The moving averages were then decomposed to ten levels using the wavelet transform, and corresponding details were then compared, as well as the undecomposed moving averages. For a discussion of the merits and pitfalls of using moving averages, see 4.4

Choice of Lags

As before, up to a maximum ten lags were consider for all VAR models. This time the maximum number of lags ten proved to be the best model, without exception.

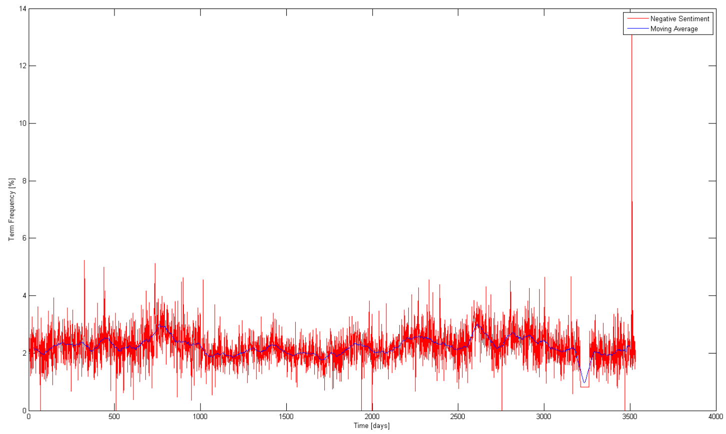


FIGURE 4.6: The moving average of negative sentiment in blue, with the original series in red.

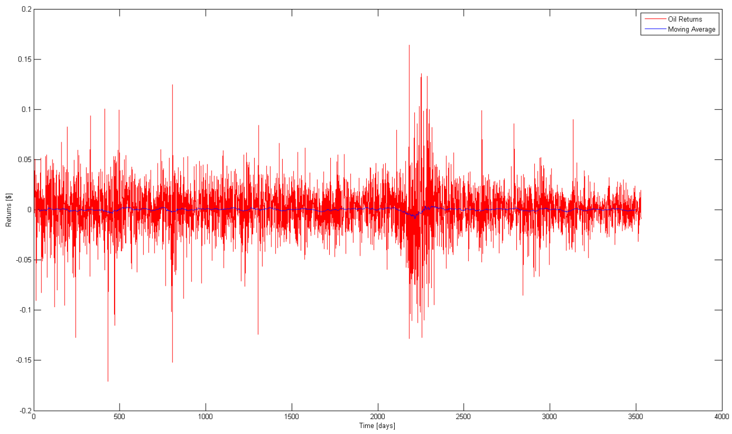


FIGURE 4.7: The moving average of oil returns in blue, with the original series in red.

Granger Causality Tests

Displayed in table 4.6 are the results of the Granger causality tests. The table is laid out as previously. Again, the next subsection looks at the VAR models for relationships that showed Granger causality.

VAR Models

Here we take a closer look at the the relationships discovered by the Granger causality tests. While we again save space as before by just listing which coefficents were significant, it is worthwhile to look at only complete discription of a VAR equation.

	Results	Null Hypotheses	
		S $\not\Rightarrow$ R	R $\not\Rightarrow$ S
D10 (10 lags)	inconclusive	0.996	0.999
D9 (10 lags)	inconclusive	0.999	0.999
D8 (10 lags)	inconclusive	0.999	0.999
D7 (10 lags)	inconclusive	0.999	1.00
D6 (10 lags)	inconclusive	1.00	1.00
D5 (10 lags)	inconclusive	0.962	0.480
D4 (10 lags)	S $\Rightarrow$ R	0.012	0.994
D3 (10 lags)	S $\Rightarrow$ R	0.004	0.577
D2 (10 lags)	S $\Rightarrow$ R	0.050	0.195
D1 (10 lags)	inconclusive	0.862	0.107
Overall (10 lags)	R $\Rightarrow$ S	0.558	0.012

Table 4.6: Table showing the results of Granger Causality tests on different wavelet levels. The values shown are probabilities of keeping the Null Hypotheses.

Lagged Value	Coefficient	Std. Error	T-ratio	P-value	Significance
Constant	2.26e-8	9.047e-7	0.0250	0.980	
Neg Senit D3 (1)	4.51e-5	0.000284	0.158	0.873	
Neg Senit D3 (2)	3.50e-5	0.000419	0.08357	0.933	
Neg Senit D3 (3)	4.75e-5	0.000400	0.118	0.905	
Neg Senit D3 (4)	0.000816	0.000402	2.029	0.042	**
Neg Senit D3 (5)	0.00101	0.000490	2.067	0.038	**
Neg Senit D3 (6)	0.000459	0.000491	0.934	0.349	
Neg Senit D3 (7)	0.000184	0.000403	0.458	0.646	
Neg Senit D3 (8)	0.000170	0.000401	0.424	0.671	
Neg Senit D3 (9)	0.000360	0.000420	0.857	0.391	
Neg Senit D3 (10)	0.000151	0.000284	0.532	0.594	
Returns D3 (1)	1.327	0.0152	87.28	0.000	***
Returns D3 (2)	0.608	0.0223	27.19	0.000	***
Returns D3 (3)	0.150	0.0213	7.050	2.15e-12	***
Returns D3 (4)	1.049	0.0214	48.93	0.000	***
Returns D3 (5)	1.313	0.0258	50.86	0.000	***
Returns D3 (6)	0.626	0.0257	24.27	0.000	***
Returns D3 (7)	0.0955	0.0213	4.465	8.28e-6	***
Returns D3 (8)	0.721	0.0213	33.86	0.000	***
Returns D3 (9)	0.911	0.0223	40.84	0.000	***
Returns D3 (10)	0.451	0.0151	29.81	0.000	***

Table 4.7: Table showing the equation for detail 3 of oil returns from the fitted VAR model of itself and corresponding detail of negative sentiment.

For the other Granger causal relationships, here are lists of the significant lagged values.

Equation for detail 2 of oil returns.

- Negative Senitment D2 lagged value 9. (At 5% confidence level)
- Returns D2 lagged values 1-10. (All of 1% confidence level)

Equation for detail 4 of oil returns.

- Negative Senitment D4 lagged value 8. (At 5% confidence level)
- Negative Senitment D4 lagged value 9. (At 1% confidence level)
- Negative Senitment D4 lagged value 10. (At 10% confidence level)
- Returns D4 lagged values 1-6. (At 1% confidence level)
- Returns D4 lagged values 8-10. (At 1% confidence level)

Equation for original negative sentiment series.

- Constant (At 1% confidence level)
- Negative Senitment lagged values 1-3. (All of 1% confidence level)
- Negative Sentiment lagged value 4. (At 5% confidence level)
- Negative Sentiment lagged values 5-8. (All of 1% confidence level)
- Negative Sentiment lagged value 9. (At 5% confidence level)
- Negative Sentiment lagged value 10. (At 1% confidence level)
- Returns lagged value 2. (At 1% confidence level)
- Returns lagged values 5. (At 10% confidence level)
- Returns lagged values 8. (At 10% confidence level)

Discussion

The brief discussion here will focus on the actual results of the moving average decomposition in isolation. A comparision between these results and those obtained without using the moving average, see section 4.4 below.

These results seem to answer the thesis question. On three seperate timeframes of activity, sentiment causes returns with varying levels of significance. The reason the full

VAR model for level three was shown above is because it shows that a large number of the lagged values need not be significant for Granger causality. The equation for oil returns has two lagged values of negative sentiment detail, but these have a large enough influence on their own for sentiment to Granger cause returns on level three.

It is interesting to see which lagged values are significant at different levels. At level two and four, there is a delay of at least nine days before the reaction of the markets. At level three however, the response is much quicker, with a delay of only four to five days.

Again, the result for the overall undecomposed series does not reflect the direction of causality of different levels. This is still explained by noting again that the sum of activities on different levels does not nessacrily reflect the individual levels. The benefit of breaking the time series into different timeframes is not to explain the overall relationship, but to see how that realtionship changes with those timeframes, information lost in the sum of those activities.

4.4 Discussion

4.4.1 Need to decompose

From all these various results, one thing is clear - to discover the full nature of the relationship between two variables, there is a need to decompose the signals by timescale. Even decomposing one of the variables is not enough, as the first section of results show. Like must be compared with like. The overall time series of sentiment cannot be compared with activities of only one time scale of the returns.

4.4.2 Longer timeframes unaffected

Something all sets of results agree upon is that there is no relationship between the two variables at higher timescales i.e. lower frequencies. For level five up, no set of data revealed any kind of relationship. Level five corresponds to $2^4 = 16$ days, roughly a fortnight. While we would not have been surprised if this timescale would have seen a relationship, we would have if level eight (roughly half a year) or higher had seen a realtionship. At activities at this timescale, senitment from news articles would not have much effect, as traders are now thinking long into the future, and basing thier activities on more concrete data, such as the objective information in finanical news, or other market variables.

In fact, it is a noteworthy result from this thesis that by the time traders act on timeframe greater than a week, they start ignoring sentiment that used to play a role in weekly down to daily data. This conforms to our hypothesis that sentiment has more to do with the variance in the returns signal than the mean. The lower levels of timescale contain much more the the variance of the signals. These results seem to confirm the hypothesis that sentiment is largely an output of people trying to understand the variance in the market.

4.4.3 Delays in reaction

Another common element across datasets is that when there was a relationshio between the two variables at any level, there was a delay in the reaction. While the VAR models do not contain contemporous values, it is interesting to see that these delays were often substantive, up to ten days between the original change in one variable before the response in the other.

In fact, only two relationships, detail three of the moving average dataset and detail one without the moving average, had a delay of less then 8 days. It makes intuitive sense that the daily data would have faster responses, as these traders work on smaller timescales. It would be interesting to see, if the data were available, how sentiment impacted minute by minute. Perhaps the delays there would be much smaller?

4.4.4 Undecomposed relationship

The final common element between the different sets of data in the undecomposed signals, moving average or not, ts that returns Granger cause senitment. This result seems to indicate that overall, more sentiment filled articles are printed about the movements of markets then the markets move to those articles, and it is necessary to break down the time series into different time frames to see when the causality moves the other way.

4.4.5 Moving average

The big difference in results that must be addressed is between taking the decomposition of the moving average verus using the original time series. The relationship between the two variables at the daily time scale disappears using the moving average, along with the feedback at level four becoming Granger causality in one direction only. In addition, relationships at levels two and three can be seen where the results were inconclusive before. The lack of realtionship at the higher levels and the returns causing sentiment on the undecomposed signals were the only common features of both data sets.

Taking the moving average of any dataset removes the outliers in the dataset, but also removes much of the variance. While we performed this operation to acheive the former, it can clearly be seen in figures 4.7 and 4.6 that the latter effect also happened. Given much of the work is around explaining the variance in the market, it might seen counter-intutitive to make this operation.

However, there were clear outliers in the data we were working with, and ignoring the effects their presence could have is also a road to failure. We could have tried to replace the outliers with other data, but such manual selection of what consitutes an outlier is problematic at best. Taking the average at the expense of losing some of the variance was seen as the best comprimise.

Our interpertation of the two sets of results is as follows: Level one contains most of the variance of the two signals, and as such would be most affected by the moving average procedure. We feel at this level the benefit of removing outliers will be outweighed by the removal of much of the variance. This was also felt to be the case at level two. However, some of the variance remains in the signal after taking the moving average, and it was felt that enough was left to be caught by the higher levels, three to four. On these levels the benefit of removing outliers would be seen, so for these levels the results after taking the moving average were valid.

We acknowledge that this interpertation is open to debate, and it is why both sets of results are presented here. One of the suggestions for future work is an robust examination of the effect of taking a moving average on the data, which unfortunately there was no time for in this work.

4.4.6 Summary of Results

In summary, sentiment does Granger cause returns in the oil market; at least when the activity on a daily, every fourth day and weekly timeframe (levels one, three and four) are considered. At longer timeframes, the sentiment is largerly ignored by the market. Showing the complications of relationship of economic variables, the original undecomposed signals, returns Granger cause sentiment. These results agree with the analysis done in section 4.2, where we only expected sentiment to have an impact at the variance dominated lower levels.

There was an interesting indication of feedback between sentiment and returns at the weekly level, but the moving average results seem to suggest that this result was unreliable, and based on outliers in the data.

So, the answer to the thesis' main question is yes: sentiment causes returns, but only on short timeframes. The sum of the different timescales, the realtionship is reversed, and returns cause sentiment

Chapter 5

Conclusion

5.1 Summary

This work fits alongside a growing body of work of using data provided by sentiment analysis to try better financial models. The goal of any work trying to incorporate sentiment in this area is trying to better explain the variance in the market prices by incorporating information from a source outside the market. The past values of the price itself does not provide enough information to adequately model the data.

Sentiment seems like a good outside source for our models. It makes intuitive sense that markets should be influenced by the news reporting on them, and there is a body of theory as to how that influence might be felt. There is also a chance sentiment is influenced by the markets, perhaps creating a feedback loop. This would not have been a surprising result.

Sentiment comes from uninformed people trying to makes sense of a noisy market, and could be used in turn by people that influence the market. This has been recognized before sentiment analysis techniques became sophisticated enough to provide the kind of information required by financial modelers. Now sentiment analysis can distill a document of text to a value of positive or negative sentiment using natural language processing and machine learning, quantifying sentiment in a way in can be used in models.

The models this thesis used to incorporate sentiment were VAR models, testing for relationships between the two variables using Granger Causality tests. VAR models represent the data a linear equations of lagged values of the variables, and provide a flexible approach to relating two or more variables. These models were used to give a better picture of relationships pointed to by the Granger tests, which simply state if a relationship exists between two time series.

Fitting models to economic or financial variables can often lead to inconclusive results, or show no evidence of a relationship. However, variables of this kind are the cumulative effect of numerous activities, all operating on different timeframes. Some traders operate on a daily basis, whereas others plan six months or a year in advance.

It is possible to decompose these time series into different constituents, each relating only to activity within a certain timeframe using the wavelet transform. Doing this allows the two processes, sentiment and returns, to be compared on these different timetables. This multiresolution analysis gives a better view of the relationship between two complicated variables, and shows how that relationship changes with timeframes. (Ramsey and Lampart, 1997)

5.2 Main Findings

Here we summarize the main findings of the thesis.

- *In shorter timeframes, sentiment does cause returns*

  At the lower levels of 1, 3 and 4, the relationships found using the Granger Causality test showed that sentiment does indeed Granger cause returns. The results for levels 1 and 2 are from not using the moving average, but higher levels are from the moving average results.

- *The longer timeframes showed no relationship*

  At levels 5 and higher, the Granger Causality tests showed no evidence of a relationship between the two variables. This and the previous finding agree with our analysis of the descriptive statistics of the each detail of the decomposed return signal. If sentiment had an influence, we expected it to be at the lower levels, where the variance dominated over the mean.

- *Undecomposed, returns cause sentiment* The cumulative sum of the activities of different timeframes show a relationship in the opposite direction of lower levels, where instead returns cause sentiment. Why returns might cause sentiment has been discussed. One reason is that articles are written that comment on the movements of the market.

- *Moving Average reveals relationships*

  Taking the moving average of the two signals before decomposing them using wavelets revealed relationships on levels 2 and 3, while it changed the relationships found without using the moving average at levels 1 and 4. The relationship found at levels 2, 3 and 4 were that sentiment causes returns, with no relationship found at level 1.

  The moving average was taken to try remove outliers, but also removed variance from the signals. This removal of variance would have a greater affect at lower levels of the decomposition, as this is where the variance of the signals is concentrated. For this reason the results of the moving average were ignored for levels 1 and 2. It has been acknowledge this distinction is arbitrary, and a suggestion for future work is defining exactly when the moving average results are reliable.

- *There is a need to decompose the signal*

  Without decomposing the two signals, the relationships between these two signals would have remained hidden. Decomposing just returns is not enough as well - like must be compared with like. The is no point of comparing the overall sentiment activity with daily return timeframe data. Multiresolution analysis is important in this study.

- *Delays in reaction*

  The response in returns to sentiment were surprisingly long, with the delay at level 1 and 4 over a week long. Considering level one show daily trading activity, a delay this long shows unexpected consideration from day traders.

5.3 Future Work

The first recommendation for future work would be a robust analysis of the exact nature of the effect of taking moving average of the two time series. The interpretation taken in this thesis does seem reasonable, but there is a need for a statistical analysis of the proposed effects. The limit of ignoring the results of the moving average at levels one and two is somewhat arbitrary. A method for identifying the tipping point between the benefits of removing outliers against removing the variance we are trying to study needs to be formulated.

Other future work could take the method developed in this thesis and apply it to other contexts. An obvious context would be to look at positive instead of negative sentiment.

It would be interesting to see if the result from (Niederhoffer, 1971) of negative sentiment having a larger affect agreed with this analysis. It could also use more complicated dictionaries. Perhaps the locally defined Rocksteady dictionary for Crude Oil has some correlation with the oil market, though this would seem to be a return to just news volume.

Another extension could be looking at other markets. We choose a commodity market for its comparative lack of volatility, but there is no reason stock markets could not be explored with this analysis.

A final suggestion is looking at other types of sentiment sources. Using mircoblogs such as Twitter or StockTwits has seen some success in the literature. One area of research would be to see if bringing wavelet analysis to the table would gain much benefit for such a quick and frequent data source.

Appendix A

Data Used

Please see provided CD for all wavelet details and full Vector Autoregression tables.

Bibliography

K. Ahmad, N. Daly, and V. Liston. What is new? news media, general elections, sentiment, and named entities, 2011.

S. Argamon, K. Bloom, A. Esuli, and F. Sebastiani. Automatically determining attitude type and force for sentiment analysis. In *In Proceedings of the 3rd Language and Technology Conference (LTC07*, 2007.

M. A. Ario. Forecasting time series via discrete wavelet transform. Technical report, 1996.

S. E. Asch. Studies of independence and conformity: A minority of one against a unanimous majority. *Psychological Monographs: General and Applied*, 70(9):1–70, 1956. doi: 10.1037/h0093718.

A. Aussem and F. Murtagh. Combining neural network forecasts on wavelet-transformed time series, 1997.

P. D. Azar. Sentiment analysis in financial news, 2009. Thesis presented to Applied Mathematics at Havard College, Massachusetts.

F. Black. Noise. *The Journal of Finance*, 41(3):529–543, 1986. ISSN 1540-6261. doi: 10.1111/j.1540-6261.1986.tb04513.x. URL http://dx.doi.org/10.1111/j.1540-6261.1986.tb04513.x.

L. Breiman. Better subset regression using the nonnegative garrote. *Technometrics*, 37(4):373–384, 1995. doi: 10.1080/00401706.1995.10484371. URL http://www.tandfonline.com/doi/abs/10.1080/00401706.1995.10484371.

C. S. Burrus, R. A. Gopinath, and H. Guo. *Intoduction to Wavelets and Wavelet Transforms: A Primer.* Prentice Hall.

J. G. Carbonell. Subjective understanding: Computer models of belief systems, 1979.

S. Chakrabarti. *Mining the Web: Discovering Knowledge from Hypertext Data*. Morgan Kaufmann, 2002.

D. M. Cutler, J. M. Poterba, and L. H. Summers. What moves stock prices? Working Paper 2538, National Bureau of Economic Research, July 1989. URL http://www.nber.org/papers/w2538.

J. B. De Long, A. Shleifer, L. H. Summers, and R. J. Waldmann. Noise Trader Risk in Financial Markets. *Journal of Political Economy*, 98(4):703–38, August 1990. URL http://ideas.repec.org/a/ucp/jpolec/v98y1990i4p703-38.html.

D. A. Dickey and W. A. Fuller. Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, 74(366a):427–431, 1979. doi: 10.1080/01621459.1979.10482531. URL http://www.tandfonline.com/doi/abs/10.1080/01621459.1979.10482531.

D. L. Donoho and I. M. Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, 90(432):pp. 1200–1224, 1995. ISSN 01621459. URL http://www.jstor.org/stable/2291512.

D. L. Donoho and I. M. Johnstone. Minimax estimation via wavelet shrinkage. *The Annals of Statistics*, 26(3):pp. 879–921, 1998. ISSN 00905364. URL http://www.jstor.org/stable/120061.

D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard. Wavelet shrinkage: Asymptopia? *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(2):pp. 301–369, 1995. ISSN 00359246. URL http://www.jstor.org/stable/2345967.

R. F. Engle and C. W. J. Granger. Co-integration and Error Correction: Representation, Estimation, and Testing. *Econometrica*, 55(2):251–76, March 1987. URL http://ideas.repec.org/a/ecm/emetrp/v55y1987i2p251-76.html.

R. F. Engle, V. K. Ng, R. F. Engle, V. K. Ng, R. F. Engle, and P. O. Economics. Measuring and testing the impact of news on volatility. *Journal of Finance*, pages 1749–1778, 1993.

R. Feldman. Techniques and applications for sentiment analysis. *Communications of the ACM*, 56(4):82–89, 2013. doi: 10.1145/2436256.2436274.

C. Fellbaum. Wordnet: An electronic lexical database, 1998.

D. Gabor. Theory of communication. part 1: The analysis of information. *Electrical Engineers - Part III: Radio and Communication Engineering, Journal of the Institution of*, 93(26):429–441, November 1946. doi: 10.1049/ji-3-2.1946.0074.

R. Gençay, F. Selçuk, and B. Whitcher. *An Introduction to Wavelets and Other Filtering Methods in Finance and Economics*. Academic Press.

R. Gençay, F. Seluk, and B. Whitcher. Scaling properties of foreign exchange volatility. *Physica A: Statistical Mechanics and its Applications*, 289(12):249 – 266, 2001. ISSN 0378-4371. doi: http://dx.doi.org/10.1016/S0378-4371(00)00456-8. URL http://www.sciencedirect.com/science/article/pii/S0378437100004568.

P. Goupillaud, A. Grossmann, and J. Morlet. Cycle-octave and related transforms in seismic signal analysis. *Geoexploration*, 23(1):85 – 102, 1984. ISSN 0016-7142. doi: http://dx.doi.org/10.1016/0016-7142(84)90025-5. URL http://www.sciencedirect.com/science/article/pii/0016714284900255. Seismic Signal Analysis and Discrimination {III}.

U. Government. U.s. enery information administration, Jan. 2014. URL http://www.eia.gov/.

C. W. J. Granger. Investigating Causal Relations by Econometric Models and Cross-Spectral Methods. *Econometrica*, 37(3):424–38, July 1969. URL http://ideas.repec.org/a/ecm/emetrp/v37y1969i3p424-38.html.

M. Halliday and R. Hasan. *Cohesion in English*. Routledge, 1976.

M. A. Hearst. Direction-based text interpretation as an information access refinement, 1992.

B. B. Hubbard. *The World According to Wavelets: The Story of a Mathematical Technique in the Making*. A K Peters.

A. Huettner and P. Subasic. Fuzzy typing for document management. In *ACL 2000 Companion Volume: Tutorial Abstracts and Demonstration Notes*, pages 26–27, 2000.

N. Jindal and B. Liu. Identifying comparative sentences in text documents. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, pages 244–251, New York, NY, USA, 2006. ACM. ISBN 1-59593-369-7. doi: 10.1145/1148170.1148215. URL http://doi.acm.org/10.1145/1148170.1148215.

I. M. Johnstone and B. W. Silverman. Wavelet threshold estimators for data with correlated noise. *Journal of the Royal Statistical Society. Series B (Methodological)*, 59(2):pp. 319–351, 1997. ISSN 00359246. URL http://www.jstor.org/stable/2346049.

J. L. Koski, E. M. Rice, and A. Tarhouni. Noise trading and volatility: Evidence from day trading and message boards. *Available at SSRN 533943*, pages 529–543, 2004. URL http://ssrn.com/abstract=533943.

H. Lütkepohl. *Introduction to Multiple Time Series Analysis.* Springer, 1991.

S. Mallat. A theory for multiresolution signal decomposition: the wavelet representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 11(7):674–693, Jul 1989. ISSN 0162-8828. doi: 10.1109/34.192463.

R. McAulay and T. Quatieri. Speech analysis/synthesis based on a sinusoidal representation. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 34(4):744–754, Aug 1986. ISSN 0096-3518. doi: 10.1109/TASSP.1986.1164910.

P. Melville, W. Gryc, and R. D. Lawrence. Sentiment analysis of blogs by combining lexical knowledge with text classification. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pages 1275–1284, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-495-9. doi: 10.1145/1557019.1557156. URL http://doi.acm.org/10.1145/1557019.1557156.

A. Michis. Multiscale analysis of the liquidity effect. Working Papers 2011-5, Central Bank of Cyprus, 2011. URL http://EconPapers.repec.org/RePEc:cyb:wpaper:2011-5.

M. L. Mitchell and J. H. Mulherin. The impact of public information on the stock market. *The Journal of Finance*, 49(3):923–950, 1994. ISSN 1540-6261. doi: 10.1111/j.1540-6261.1994.tb00083.x. URL http://dx.doi.org/10.1111/j.1540-6261.1994.tb00083.x.

R. Narayanan, B. Liu, and A. Choudhary. Sentiment analysis of conditional sentences. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 180–189, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-59-6. URL http://dl.acm.org/citation.cfm?id=1699510.1699534.

V. Niederhoffer. The analysis of world events and stock prices. *The Journal of Business*, 44(2):pp. 193–219, 1971. ISSN 00219398. URL http://www.jstor.org/stable/2351663.

C. Oh and O. Sheng. Investigating predictive power of stock micro blog sentiment in forecasting future stock price directional movement. In D. F. Galletta and T.-P. Liang, editors, *ICIS*. Association for Information Systems, 2011. ISBN 978-0-615-55907-0. URL http://dblp.uni-trier.de/db/conf/icis/icis2011.html#OhS11.

B. Pang and L. Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics*, ACL '04, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics. doi: 10.3115/1218955.1218990. URL http://dx.doi.org/10.3115/1218955.1218990.

B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, 2008. doi: 10.1561/1500000001.

B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of EMNLP*, pages 79–86, 2002.

D. B. Percival and H. O. Mofjeld. Analysis of subtidal coastal sea level fluctuations using wavelets. *Journal of the American Statistical Association*, 92(439):868–880, 1997. doi: 10.1080/01621459.1997.10474042. URL http://www.tandfonline.com/doi/abs/10.1080/01621459.1997.10474042.

D. B. Percival and A. T. Walden. *Wavelet Methods for Time Series Analysis*. Cambridge University Press.

R. Polikar. Multiresolution analysis: The discrete wavelet transform, March 1999. URL http://users.rowan.edu/~polikar/WAVELETS/WTpart4.html.

A.-M. Popescu and O. Etzioni. Extracting product features and opinions from reviews. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 339–346, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics. doi: 10.3115/1220575.1220618. URL http://dx.doi.org/10.3115/1220575.1220618.

R. Prabowo and M. Thelwall. Sentiment analysis: A combined approach. *Journal of Informetrics*, 3(2):143 – 157, 2009. ISSN 1751-1577. doi: http://dx.doi.org/10.1016/j.joi.2009.01.003. URL http://www.sciencedirect.com/science/article/pii/S1751157709000108.

J. Ramsey and C. Lampart. The Decomposition of Economic Relationships by Time Scale Using Wavelets. Technical report, 1997.

J. B. Ramsey. The contribution of wavelets to the analysis of economic and financial data. *Philosophical Transactions of the Royal Society*, 357(1760):2593–2606, 1999. doi: 10.1098/1999.0450.

E. Riloff and J. Wiebe. Learning extraction patterns for subjective expressions. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, EMNLP '03, pages 105–112, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics. doi: 10.3115/1119355.1119369. URL http://dx.doi.org/10.3115/1119355.1119369.

S. Schlter and C. Deuschle. Using wavelets for time series forecasting: Does it pay off? Technical report, 2010.

T. Setz. Wavelet analysis on financial time series, 2011.

C. A. Sims. Macroeconomics and Reality. *Econometrica*, 48(1):1–48, January 1980. URL http://ideas.repec.org/a/ecm/emetrp/v48y1980i1p1-48.html.

J. H. Stock and M. W. Watson. Vector Autoregressions. *Journal of Economic Perspectives*, 15(4):101–115, Fall 2001. URL http://ideas.repec.org/a/aea/jecper/v15y2001i4p101-115.html.

P. J. Stone, D. C. Dunphy, M. S. Smith, and D. M. Ogilvie. *General Inquirer: A Computer Approach to Content Analysis*. The MIT Press, 1966.

N. N. Taleb. *The Black Swan: The Impact of the Highly Improbable*. Random House.

C. Tan and C. Pedersen. Financial time series forecasting using improved wavelet neural network, 2009.

S. Tan and J. Zhang. An empirical study of sentiment analysis for chinese documents. *Expert Systems with Applications*, 34(4):2622 – 2629, 2008. ISSN 0957-4174. doi: http://dx.doi.org/10.1016/j.eswa.2007.05.028. URL http://www.sciencedirect.com/science/article/pii/S0957417407001534.

P. C. Tetlock, M. Saar-Tsechansky, and S. Macskassy. More Than Words: Quantifying Language to Measure Firms' Fundamentals. *Journal of Finance*, 63(3):1437–1467, 06 2008. URL http://ideas.repec.org/a/bla/jfinan/v63y2008i3p1437-1467.html.

P. D. Turney. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. *CoRR*, cs.LG/0212032, 2002.

R. von Sachs and B. MacGibbon. Nonparametric curve estimation by wavelet thresholding with locally stationary errors. 1997.

J. Wiebe, T. Wilson, and C. Cardie. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2-3):165–210, 2005. ISSN 1574-020X. doi: 10.1007/s10579-005-7880-9. URL http://dx.doi.org/10.1007/s10579-005-7880-9.

Y. Wu, Q. Zhang, X. Huang, and L. Wu. Phrase dependency parsing for opinion mining. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3 - Volume 3*, EMNLP '09, pages 1533–1541, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-63-3. URL http://dl.acm.org/citation.cfm?id=1699648.1699700.

R. Xia, C. Zong, and S. Li. Ensemble of feature sets and classification algorithms for sentiment classification. *Information Sciences*, 181(6):1138 – 1152, 2011. ISSN 0020-0255. doi: http://dx.doi.org/10.1016/j.ins.2010.11.023. URL http://www.sciencedirect.com/science/article/pii/S0020025510005682.

K. Xu, S. S. Liao, J. Li, and Y. Song. Mining comparative opinions from customer reviews for competitive intelligence. *Decision Support Systems*, 50(4):743 – 754, 2011. ISSN 0167-9236. doi: http://dx.doi.org/10.1016/j.dss.2010.08.021. URL http://www.sciencedirect.com/science/article/pii/S0167923610001454. Enterprise Risk and Security Management: Data, Text and Web Mining.

Q. Ye, Z. Zhang, and R. Law. Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. *Expert Systems with Applications*, 36(3, Part 2):6527 – 6535, 2009. ISSN 0957-4174. doi: http://dx.doi.org/10.1016/j.eswa.2008.07.035. URL http://www.sciencedirect.com/science/article/pii/S0957417408005022.