Trinity College Dublin

Masters Thesis

Innovations and Market Dynamics:

Weather and sentiment analysis in the commodities market

Author:
Gary White

Supervisor:
Prof. Khurshid Ahmad

A thesis submitted in fulfilment of the requirements
for the degree of Masters in Computer Engineering

in the

School of Engineering

May 2015

Declaration of Authorship

I, Gary White, declare that this thesis titled, 'Weather and sentiment analysis in the commodities market' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a degree at this University.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- The library may lend or copy the thesis or any part thereof on request.

Signed: ____

Date: ____

TRINITY COLLEGE DUBLIN

Summary

Faculty of Engineering, Mathematics and Science
School of Engineering

Masters in Computer Engineering

Weather and sentiment analysis in the commodities market

by Gary White

This project involves an investigation of the effectiveness of exogenous and qualitative analysis techniques such as weather and sentiment analysis on the spot price of commodities. The commodities market is a complex system and to estimate the effect of weather and sentiment articulated in news and comments requires a number of different analysis techniques. To calculate the impact of the weather analysis on the commodities market a large collection of historical data was aggregated. This was completed by writing a ruby script which could interact with an online weather API to get daily historical weather records for all the states of America over a 15 year period from the start of 2000 to the end of 2014. The weather data was then aggregated to a regional and national level using population data from the 2014 US census.

To make practical use of the large collection of historical data that had been collected a weather model was developed to be put a price on the impact that the weather would have. The model was formed using the yearly seasonality, mean temperature, cumulative heating degree days and a stochastic process. The model was then used to estimate the effect that the weather would have on energy commodity prices in the spot market. The effectiveness of the model was tested by comparing the price from the model with historical commodity records using vector autoregression which allows for the estimation of economic relationships.

The sentiment analysis system developed during the project provides a comparison to the effectiveness of the weather analysis. The investigation into the impact of sentiment analysis on the project was conducted by identifying formal and social news sources and then analysing the sentiment content of the news and comments. A large collection of historical articles over the 15 year period was needed to perform an effective comparison with the weather analysis.

In this project I use Lexis Nexis which provides access to 1782 formal newspapers and 37 social media/blog sites which allowed me to create a representative corpus of text with minimal bias of the investigator. Almost 9000 unique articles were collected for the various commodities over the 15 year period. The sentiment of the corpus of text was then analysed using the Rocksteady program which has been developed in Trinity, using a general language dictionary and domain specific dictionary which contained financial terms. The time series sentiment analysis is then analysed through the use of vector autoregression to allow comparison with the weather analysis.

The key findings from the research into the analysis techniques in the commodities market has shown that weather analysis is more effective at predicting changes in the commodity markets due to the lower P-values and higher F-tests returned by the vector autoregression. It was also observed that the analysis results of individual commodities varied to a large extent even for the same analysis techniques for example natural gas was heavily influenced by weather analysis while gasoline showed very little response to this analysis, this is somewhat expected as natural gas is used for heating while gasoline is used in cars. The sentiment analysis produced statistically significant results for WTI Crude Oil but was not as effective for the other commodities.

To test the real world application of the analysis techniques that were developed during the project an algorithmic trading program was developed. This allowed the analysis techniques to be used to make bids on commodities if they predicted a positive change in the market over a specific look-back window. Both analysis techniques produced positive average returns with the weather analysis slightly outperforming the sentiment analysis. As a conclusion to the project I developed an interactive method for viewing the results of the analysis using the D3 JavaScript library. I created a series of interactive choropleth maps which provided an intuitive way to view the findings that were observed over the course of the research. These visualizations can be published online which allows for other researchers to view and interact with the results that I have observed.

The project has led to a number of interesting results in the use of weather and sentiment analysis in the commodities market. The methodology developed during the project can be used for the inclusion of a number of other exogenous variables into the model such as shipping analysis which was briefly discussed during the project. With the increase in public information being released as part of the open government partnership there is great opportunity to add to and improve the analysis techniques developed in this project through the inclusion of further exogenous variables. This project provides clear methods on how to include and test exogenous variables as part of an analysis model.

“Imagine how much harder physics would be if electrons had feelings!”

– Richard Feynman

TRINITY COLLEGE DUBLIN

Abstract

Faculty of Engineering, Mathematics and Science
School of Engineering

Masters in Computer Engineering

Weather and sentiment analysis in the commodities market

by Gary White

This project involves an investigation of the impact of the effect of exogenous and qualitative analysis on the spot price of commodities. The commodities market is a complex system and to estimate the effect of weather and sentiment articulated in news and comments requires a number of different analysis techniques. To calculate the impact of the weather analysis on the commodities market a large collection of historical data was aggregated. A weather model was then formed using cumulative heating degree days, mean temperature and a stochastic process, the accuracy of this model is then tested using vector autoregression.

The investigation into the impact of sentiment analysis on the project was conducted by identifying formal and social news sources and then analysing the sentiment content of the news. The sentiment of the corpus of text is analysed using the Rocksteady program which has been developed in Trinity, using a general and domain specific dictionary. The time series sentiment analysis is then analysed through the use of vector autoregression.

The key findings from the research into the analysis techniques in the commodities market has shown that weather analysis is more effective at predicting changes in the commodity markets due to the lower P-values and higher F-tests returned by the vector autoregression. It was also observed that the analysis results of various commodities varied to a large extent even for the same analysis techniques for example Natural Gas was heavily influenced by weather analysis while gasoline showed very little response to this analysis. The sentiment analysis produced statistically significant results for WTI Crude Oil but was not as effective for the other commodities.

As a conclusion to the project I develop an algorithmic trading program which makes trades based on trends in the weather and sentiment analysis which shows how the research that was conducted has useful applications in industry. I also look at some visualization options which show how the research results can be best illustrated.

Acknowledgements

I would like to thank my project supervisor Professor Khurshid Ahmad for introducing me to this interesting project and for his constant support and guidance throughout the development of the project. I would also like to thank family and friends for their constant support and encouragement.

Contents

List of Figures

List of Tables

Abbreviations

EMH	**E**fficient **M**arket **H**ypothesis
VAR	**V**ector **A**uto **R**egression
FFA	**F**orward **F**reight **A**greement
NLP	**N**atural **L**anguage **P**rocessing
GP	**G**enetic **P**rogramming
HDD	**H**eating **D**egree **D**ays
NCDC	**N**ational **C**limate **D**ata **C**enter
CGE	**C**omputable **G**eneral **E**quilibrium
NIG	**N**ormal **I**nverse **G**aussian

Chapter 1

Introduction

1.1 Aims of the project

Information plays a large part in the valuation of commodities and stocks and the market will constantly adjust to the announcement of new information [1]. In recent years there has been great progress made in the development of knowledge and data engineering systems which has seen a large number of new techniques developed for information mining [2]. This in conjunction with the large amount of public data being released through the Open Government Partnership has led to a number of knowledge based systems being used for financial applications [3]. In this project I explore the possibility of using sentiment and weather analysis for financial applications in the commodities market. The initial phase of the project involves a literature review of the current techniques that are being used to make investment decisions in the commodity market and collecting a large dataset to test the final system on.

The project considers a number of techniques for weather and sentiment analysis and is tested on daily data from a number of different commodity exchanges. The corpus of text for sentiment analysis is built from two main categories, formal media and social network sites. The weather data is collected through querying an API which provides access to a large range of historical data. The final portion of the project deals with the application of this of this analysis by building an algorithmic trading system which makes trades based on trends in the weather and sentiment analysis.

1.2 A Reader's Guide to the Report

Chapter 2 - Literature Review: This chapter is concerned with a review of the current literature that is available in relation to the project. It is broken down into a number of sections which deal with the different types of analysis that have already been used in commodity markets and where my research will fit into the state of the art. This chapter takes the reader from some of the more traditional methods that are used to price commodities to the latest techniques that have been developed in weather and sentiment analysis. There is also a section dedicated to explaining how the variables will be aggregated and how the causality of the variables will be determined.

Chapter 3 - Methods & Data: This chapter is broken down into two major sections: Data and Exogenous & Qualitative Analysis. The data section deals with how the large collection of data that was required for the project was collected. It is broken down into sections about how the weather data, text corpus and commodity data were collected and aggregated. In these sections a number of different techniques are described and I outline the reason why the final methods were chosen. The other sections deal with how the data would be analysed, both in terms of the weather and sentiment analysis.

Chapter 4 - Results & Discussion: This chapter presents the results that have been observed by carrying out this project. It is broken down into sections about weather and sentiment analysis and also looks at how individual commodities responded to both types of analysis. The results of the algorithmic trading are also shown and discussed with reference to current trading techniques. The final conclusion section gives a brief overview of the results that were observed and the conclusions that can be drawn from the conduction of this project which is expanded upon in the final chapter.

Chapter 5 - Conclusions: The final chapter in the thesis deals with the conclusion that have been learned by carrying out this research and the implications that these conclusions have. I break down the final chapter to discuss the individual effectiveness of the weather and sentiment analysis. The final section deals with what I have learned over the course of the project, the project has been so broad in scope that it has allowed me to learn about a number of different areas of computer science and statistics.

Chapter 2

Literature Review

2.1 Introduction

This chapter is an introduction to the previous work that has been carried out in financial analysis and shows where my research fits in to the state of the art. There are a large amount of techniques which have been developed to conduct analysis into the financial markets, the review is broken into three sections which discuss quantitative analysis, qualitative analysis and the aggregation of variables and causality.

2.2 Quantitative Analysis

Quantitative analysis focuses on the analysis of the measurable figures related to a commodity, such as the current value or the projected demand. A large number of mathematical techniques have been developed to try and predict changes in the markets based on these measurable figures. These techniques can generally be divided into three categories of analysis: conventional analysis, econometric analysis and exogenous variables. Conventional analysis deals with two of the most widely used forms of analysis which are fundamental and technical analysis. Econometric analysis applies statistical methods and computer science techniques to economic data to give empirical content to economic relations. The section on exogenous variables shows how we can add independent variables such as weather analysis and shipping costs which can then be incorporated into our system to increase the reliability.

2.2.1 Conventional Analysis

Conventional analysis considers the two most widely used forms of analysis, fundamental and technical analysis. These sections discuss the latest developments in these forms of analysis as well and showing how they can be improved through the inclusion of exogenous variables such as weather or shipping or by the inclusion of qualitative analysis techniques such as sentiment analysis.

Fundamental Analysis

Fundamental analysis is a method of evaluating a commodity that entails attempting to measure its intrinsic value by examining related economic, financial and other quantitative factors. Fundamental analysts attempts to study everything that can affect the security's value, including macroeconomic factors like the overall economy and industry conditions [4].

There are two general approaches to fundamental analysis which are top-down and bottom-up. The top-down approach starts with the overall economy and then works down from commodity groups to a specific commodity and the bottom-up approach works in the opposite way. There has been some success in integrating bottom-up and top-down analysis in energy economic modelling [5], which shows how computable general equilibrium (CGE) models can be used to estimate how the economy might react to changes in policy or external factors.

To take a top down approach for example it has been shown that real interest rates and the price of the dollar can have a large impact on commodity prices [6]. This research shows that commodity prices increase significantly in response to reductions in real interest rates and also that there is evidence to suggest that a weaker dollar leads to higher commodity prices. Figure 2.1 shows the relationship between commodity prices and real interest rates over a 40 year period. Once a scenario for the overall economy has been developed, an investor can break down into the various commodity groups.

Commodities tend to move in groups for example energy commodities tend to have high correlations to each other [7]. As can be seen from historical evidence over 40 years in Figure 2.2, different commodity groups have large differences in performance. There can also be a number of variations within each commodity group for example in recent years

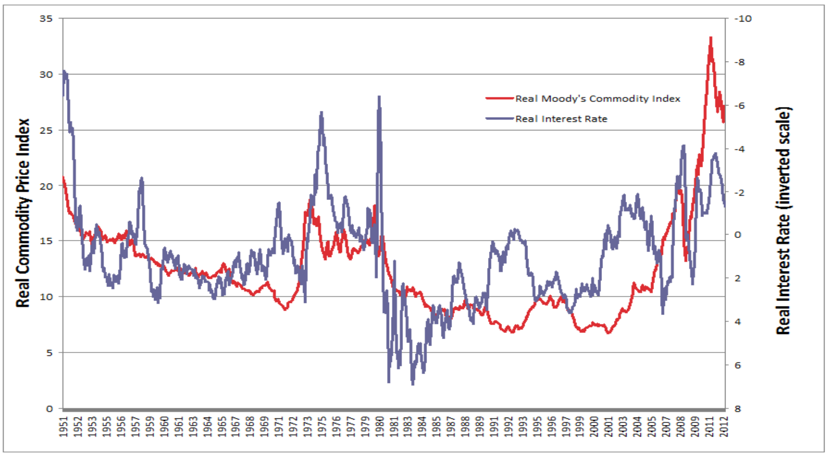
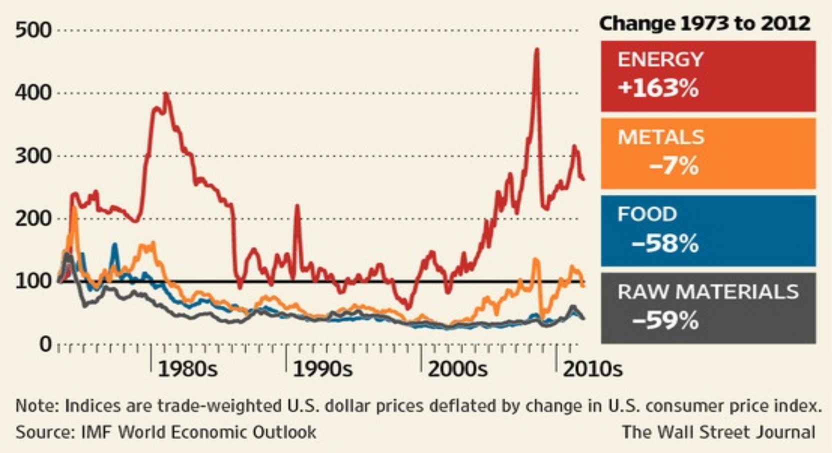


Figure 2.1: Commodity price spikes in the 1970s, 2008 & 2011 can be partially explained by real interest rates that are zero or even negative [6].



Note: Indices are trade-weighted U.S. dollar prices deflated by change in U.S. consumer price index.
Source: IMF World Economic Outlook
The Wall Street Journal

Figure 2.2: Adjusted for inflation, energy prices are now well above the 1973 levels. Metals prices are about the same and food and raw materials cost far less.

a glut in US crude sent WTI prices into a tailspin compared to Brent [8]. Figure 2.3 shows how dramatic these changes have been in recent years and also the importance of picking the individual commodity as well as the correct group.

As I have shown the commodity group has a large impact on the price of a commodity however there can be a number of variations within it group for example in recent years after a glut in US crude sent WTI prices into a tailspin compared to Brent [8] as can be seen in Figure 2.3. This shows the importance of researching a number of individual aspects which could affect the commodity such as location.

Fundamental Analysis is good for long-term investments based on very long-term trends. The ability to identify and predict long-term economic, demographic or consumer trends can benefit patient investors who pick the right commodity groups. Valuation techniques

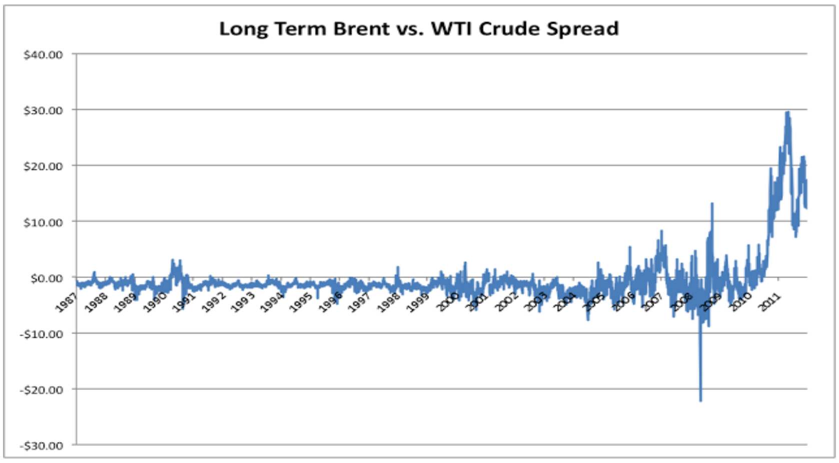


FIGURE 2.3: Long Term Brent vs. WTI Crude Spread

vary depending on the commodity group and specifics of each commodity, for this reason, a different technique and model is required for different groups. The end goal of performing fundamental analysis is to produce a value that an investor can compare with the commodity's current price, with the aim of figuring out what sort of position to take with that commodity e.g. if under-priced then buy and if overpriced then sell.

Technical Analysis

The other type of conventional analysis which is used in state of the art systems is technical analysis, which differs largely from fundamental analysis and believes that there is no reason to analyse a commodities fundamentals because these are all accounted for in the commodity price. Technical analysts believe that all the information they need about a commodity can be found in its charts and a large number of techniques have been developed to use this chart information. John Murphy [9] and Stephen Taylor [10] have written excellent books on Technical Analysis in financial markets which provided a great reference to a number of useful techniques.

Much of the criticism of technical analysis has its roots in academic theory, specifically the efficient market hypothesis (EMH). This theory says that the market's price is always the correct one and any past trading information is already reflected in the price of the stock and commodities and therefore any analysis to find undervalued commodities is useless. There are three forms of EMH: weak, semi-strong and strong. According to weak form efficiency, technical analysis can't predict future movements because all past information has already been accounted for and therefore analysing the stock's past

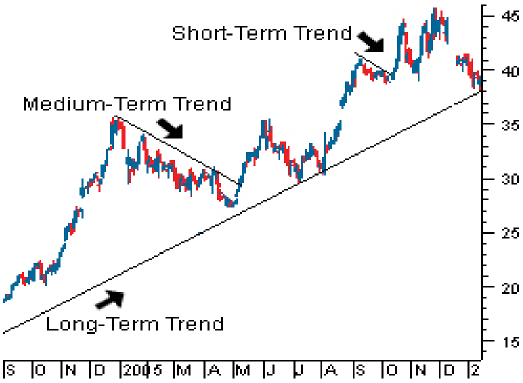


FIGURE 2.4: Trend lengths classifications

price movements will provide no insight into its future movements, however fundamental analysis may still be able to provide excess returns.

The strong and semi-strong form of the efficient market hypothesis that the market is even more efficient and that the price of commodities adjust to all publicly and privately available information, and that neither fundamental nor technical analysis will reliable be able to produce excess returns. There has been an extensive amount of research on the EMH [11, 12] and it is still widely argued by academics today, it is presented here to show how some researchers view the limitations of technical analysis.

One of the most important concepts in technical analysis is that of trend. There are three major types of trends up-trends, down-trends and horizontal trends. As the names imply when each successive peak and trough is higher it's referred to as an upward trend. If the peaks and troughs are getting lower it's a down-trend, when there is little movement up or down in the peaks and troughs, it's a sideways or horizontal trend.

Along with these three trend directions, there are three trend classifications. A trend of any direction can be classified as a long-term, intermediate or a short-term trend. In terms of the market, a major trend is generally categorized as one lasting longer than a year. An intermediate trend is considered to last between one and three months and a near-term trend is anything less than a month. A long-term trend is composed of several intermediate trends, which often move against the direction of the major trend. If the major trend is upward and there is a downward correction in price movement followed by a continuation of the up-trend, the correction is considered to be an intermediate trend. The short-term trends are components of both major and intermediate trends, examples of the different types of trends can be seen in Figure 2.4.

Although fundamental and technical analysis are seen by many as polar opposites many market participants have experienced great success by combining the two. For example, some fundamental analysts use technical analysis techniques to figure out the best time to enter into an undervalued commodity. Alternatively, some technical traders might look at fundamentals to add strength to a technical signal. For example, if a sell signal is given through technical patterns and indicators, a technical trader might look to reaffirm his or her decision by looking at some key fundamental data. It is often the case that using both the fundamental and technical analysis can produce the best results.

2.2.2 Econometric Analysis

Econometric analysis applies statistical methods to the quantification and critical assessment of hypothetical relationships using data. It is with the aid of econometrics that we can discriminate between competing theories and are able to choose the most effective. There are a number of texts which outline the field of econometrics two of the texts which I found most useful were by William Greene [13] and Maddala [14]. Both of these texts deal with a large number of different techniques, in this section I will describe some of the techniques that are widely used in research to compare competing theories.

Simple Regression

Regression is a statistical and mathematical methods used to try and explain or predict a variable, known as the dependent variable, given one or more independent variables and is one of the most commonly used tools in econometric work. It is used to examine and quantify the effect of one variation upon another, for example the effect of higher personal income on television sales or the effect of a person's height on their weight. We will discuss the case of one explained variable which we will denote by Y and one explanatory variable which we will denote by X. The relationship between X and Y can be denoted:

$$Y_t = \alpha + \beta_1 X_t + u_t \tag{2.1}$$

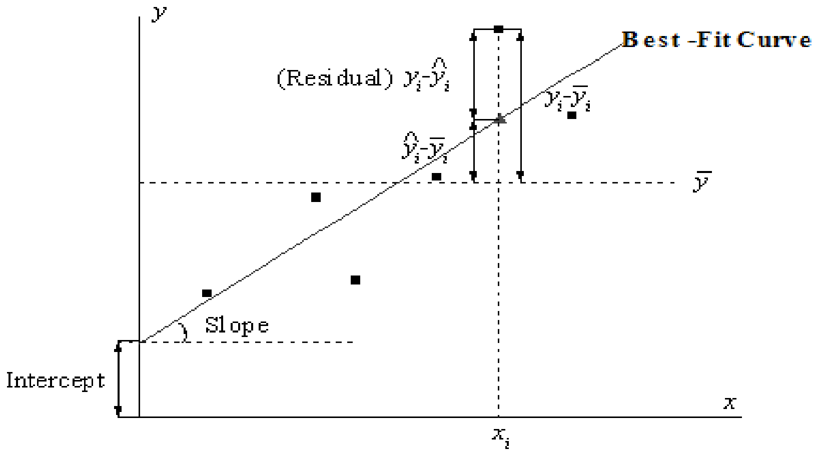


FIGURE 2.5: Method of least square error.

where α and β are the unknown parameters and u_t is the error. To determine α and β we use the method of least squares. The method of least squares requires that we choose $\hat{\alpha}$ and $\hat{\beta}$ as estimates of α and β so that:

$$Q = \sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta} x_i)^2 \tag{2.2}$$

is a minimum. Q also gives the sum of the squares of the prediction errors when we predict y_i given x_i and the estimated regression equation. The intuitive idea behind the least squares procedure can be seen in Figure 2.5.

Multiple Regression

The simple regression model can be extended to include a number of explanatory variables instead of only one. In this section I give a brief introduction to how one explained variable which we will denote by Y and a number of explanatory variables which we will denote by $X_1, X_2, \ldots, X_k$. The relationship between X and Y can be denoted:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + u \tag{2.3}$$

where α, β_1 and β_2 are the unknown parameters and u is the error. To determine α, β_1 and β_2 we use the method of least squares. The method of least squares requires that we choose $\hat{\alpha}$, $\hat{\beta}_1$ and $\hat{\beta}_2$ as estimates of α, β_1 and β_2 so that:

$$Q = \sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i})^2 \tag{2.4}$$

is a minimum, this allows for much more complicated data series to be analysed than would be possible with the simple regression model.

Vector Autoregression

A univariate autoregression is a single-equation, single-variable linear model in which the current value of a variable is explained by its own lagged values. A VAR is an n-equation, n-variable linear model in which each variable is in turn explained by its own lagged values, plus current and past values of the remaining n-1 variables.

This simple framework provides a systematic way to capture rich dynamics in multiple time series. Sims [15] was one of the first advocates of the of the VAR analysis and argued that VARs held out the promise of providing a coherent and credible approach to data description, forecasting, structural inference, and policy analysis. A VAR can be thought of as the reduced form of a dynamic economic system involving a vector of variables z_t. That is, starting from the so-called structural form:

$$Az_t = B_1 z_{t-1} + B_2 z_{t-2} + ... + B_p z_{t-p} + u_t \tag{2.5}$$

$$E_{uu'} = \Sigma_u = \begin{bmatrix} \sigma^2_{u1} & 0 & \dots & 0 \\ 0 & \sigma^2_{u2} & \dots & 0 \\ 0 & 0 & \dots & \sigma^2_{un} \end{bmatrix} \tag{2.6}$$

a VAR of lag length p (VAR (p)) can be written as:

$$z_t = A^{-1}B_1 Z_{t-1} + A^{-1}B_2 Z_{t-2} + \cdots + A^{-1}B_p Z_{t-p} + A^{-1}u_t$$

E(u_t) = 0, E($u_t u'_\tau$) = Σ_u for t = τ, and 0 otherwise. Thus, a vector autoregression is a system in which each variable is expressed as a function of own lags as well as lags of each of the other variables. This makes it useful in econometrics for forecasting [16, 17] and discriminate between competing theories. In this project we run VAR(5) regression to reflect the five day nature of the return prices.

2.2.3 Exogenous Variables

Some economic variables are determined by our conventional analysis, while others are usually assumed to be determined by factors outside of this analysis. We call the former endogenous variables and the latter exogenous variables. For econometric applications, the crucial difference between an endogenous and an exogenous variable is that we must assume that exogenous variables are not systematically affected by changes in the other variables of the model, especially by changes in the endogenous variables [18].

For example if we are modelling the individual supply of corn produced in a year by Farmer Jones, the endogenous variable would probably be the amount of corn sold. This is the output of our economic model of Farmer Jones's production decision. The inputs would be the explanatory variables that influence the amount he sells, which might include the market price and the amount of rain that falls during the summer.

These inputs to the model can be treated as exogenous variables as an increase in production would have no effect on the weather. Corn prices is a more difficult question to answer, as if Farmer Jones was to form a cooperative with all the other farmers in Iowa it is likely that an increase in production would lower the price of U.S. corn, thus market price would be an endogenous variable in this model.

Endogenous variables are important in econometrics and economic modelling because they show whether a variable causes a particular effect. Economists employ causal modelling to explain outcomes based on a variety of factors and to determine to what extent a result can be attributed to an endogenous or exogenous cause. In this section I look at two of the exogenous variables that were considered for use in this project which were shipping and weather.

Shipping

While commodities such as oil have been transported on ships since 1861 [19], its transition from a logistical exercise controlled by the major commodity companies to a spot market it relatively recent, with 70% of spot chartering in the 1990s versus only 20% in 1973. In the spot market, tankers are chartered for a single voyage through brokers with all costs included in the price. There are a number of reference routes for dirty and clean tankers, numbered TD1 to TD18 for dirty i.e. heavy oils such as heavy fuel oils or crude oil and TC1 to TC11 for clean i.e. refined petroleum products such as gasoline, kerosene, or chemicals [20]. At the end of each trading day the Baltic Exchange polls and publishes an assessment of the price levels for each of these routes, making up the Baltic Exchange dirty and clean tanker indices, this is the recognised spot price in the tanker market.

The tanker market has also seen the relatively recent development of a Forward Freight Agreement (FFA). While the Baltic Index had introduced a futures exchange in 1985 it lost popularity and has been replaced by the route specific FFAs. FFAs settle financially at the end of their contract month on the arithmetic average of daily values of the underlying Baltic Exchange index during that month.

As can be seen by my brief outline modelling shipping as an exogenous variable is quite a complex process as there are a number of variables to take into account. There are a large collection of excellent books which provide an introduction to risk management in the shipping market and how to model it as an exogenous variable such as German [21] and Geomelos [22]. Germen provides an excellent introduction of how to model risk from the shipping market in the commodities market which would prove very useful when trying to model shipping as an exogenous variable.

Weather

The idea that climate can influence economic performance is an old one, featuring predominantly in the writings if the ancient Greeks. Montesquieu argued in The Spirit of Laws [23] that an excess of heat made men slothful and dispirited, this would effect economic outcomes such as agricultural output and economic growth. In more recent times there has been a large growth in weather derivatives and a number of excellent texts have been published in this area by Jewson [24] and Kluwer et al. [25] to mitigate against these risks.

Jewson gives an excellent introduction to meteorological, statistical and financial tools that are needed for pricing and risk management of weather derivatives. A number of data cleansing methods are described such as gap filling and value checking in historical records which proved useful when collecting my own historical data.

In this project we use the concept of heating degree day which is a relative measure of how cold a day is according to some base temperature. In American historical weather the commonly accepted baseline for this measure is 65 °F or 18 °C. To calculate the number of HDD's on a specific day, subtract the average temperature from the baseline which can be seen in equation 2.7. For example if the outside temperature was 2 degrees below the base temperature for 2 days, then there would be a total of 4 heating degree days over that period. If a day is warmer than the baseline the HDD for that day is 0, the number of HDD's on a particular day can be expressed:

$$Z_i = max(T_o - T_i, 0) \tag{2.7}$$

where T_o is the baseline temperature and T_i is the average temperature of the day in question. A HDD index X over an N_d day period is defined as the sum of the HDD's over all days during that period:

$$X = \sum_{i=1}^{N_d} z_i \tag{2.8}$$

A record of monthly and especially daily HDD totals can be very useful for trading weather derivatives. If we can see patterns of temperatures for particular regions, we can make more accurate decisions as to the value of a weather derivative contract. The temperature will also effect a number of other commodities such as cereals and grains whose production is dependent on the weather. One area where the project could further analyse the impact that the weather has on the commodities market would be through the prediction of weather shocks, which are shock weather events such as hurricanes and floods, which do a large amount of damage. The inclusion of weather shocks in this analysis would give a better prediction of the impact that the weather data will have on commodities.

2.3 Qualitative Analysis

Qualitative analysis is a method of inquiry traditionally used in the social sciences, which uses non-quantifiable information. This analysis technique is different to quantitative analysis which focuses on measurable figures [26], however the two techniques are often used together and there is evidence that a combination of both approaches produces better results [27]. In this section I review some of the latest techniques which have found application in financial markets to increase insight into commodity prices.

2.3.1 Market Behaviour

One of the underlying theories in economics and finance is that actors in the market will make unbiased decisions to maximize their self-interest and that if they make suboptimal decisions they will be punished through poor outcomes. This process will lead to the stakeholder either making better decisions or leaving the marketplace [28]. This model proposes that agents are not fully rational, the field which has been established to try and explain these irrational decisions is called behavioural finance and is in contrast with the EMH described earlier. The field has two building blocks: limits to arbitrage, which argues that it can be difficult for rational traders to undo the dislocations caused by less rational traders, and psychology which catalogues the kinds of deviations from full rationality we might expect to see when traders are making decisions [29].

A classic objection to behavioural finance, namely that even if some agents in the economy are less than fully rational, rational agents will prevent them from influencing security prices for very long, through a process known as arbitrage that goes back to 1953 [30]. One of the biggest successes of behavioural finance is a series of theoretical papers showing that in an economy where rational and irrational traders interact, irrationality can have a substantial and long-lived impact on prices which is known as limits to arbitrage [31, 32].

The theory of limited arbitrage shows that if irrational traders cause deviations from fundamental value, rational traders will often be powerless to do anything about it. In order to say more about the structure of these deviations, behavioural models often assume a specific form of irrationality. For guidance on this, economists turn to the extensive experimental evidence compiled by cognitive psychologists on the systematic biases that arise when people form beliefs and on people's preferences such as the surveys of Rabin [33] and the edited volumes of Kahneman [34, 35].

These works outline a number of theories about how people appear to form beliefs in practice. There is a large range of techniques discussed in these texts so I shall outline the major discoveries and refer the reader to the references. One of the major discoveries and extensive evidence has shown that people are overconfident in their judgements. For example the confidence intervals people assign to their estimates of quantities are far to narrow for example their 98% confidence intervals for example include the true quantity only about 60% of the time [36].

Another important discovery observed by Kahneman and Tversky [37] is anchoring. This shows how when forming estimates people often start with some initial possibly arbitrary value and then adjust away from it. Experimental evidence shows that the adjustment is often insufficient and that people anchor too much on the initial value. In one experiment subjects were asked to estimate the percentage of United Nations countries that are African. More specifically, before giving a percentage, they were asked whether their guess was higher or lower than a randomly generated number between 0 and 100. Their subsequent estimates were significantly affected by the initial random number. Those who were asked to compare their estimate to 10, subsequently estimated 25%, while those who compared to 60, estimated 45%.

The final discovery that I will describe in this subsection is availability biases which when judging the probability of an event such the likelihood of getting mugged in Chicago, people often search their memories for relevant information. While this is a perfectly sensible procedure, it can produce biased estimates because not all memories are equally retrievable or available, in the language of [37]. More recent events and more salient events such as the mugging of a close friend will weigh more heavily and distort the estimate.

This section gives a brief description of behavioural finance and how it seeks to combine behavioural and cognitive psychological theory with conventional economics and finance to provide explanations for why people make irrational financial decisions. This is the background to the next subsections which focus on the use of text and sentiment analysis to estimate the markets belief about particular commodities which could anchor the decisions made by other traders.

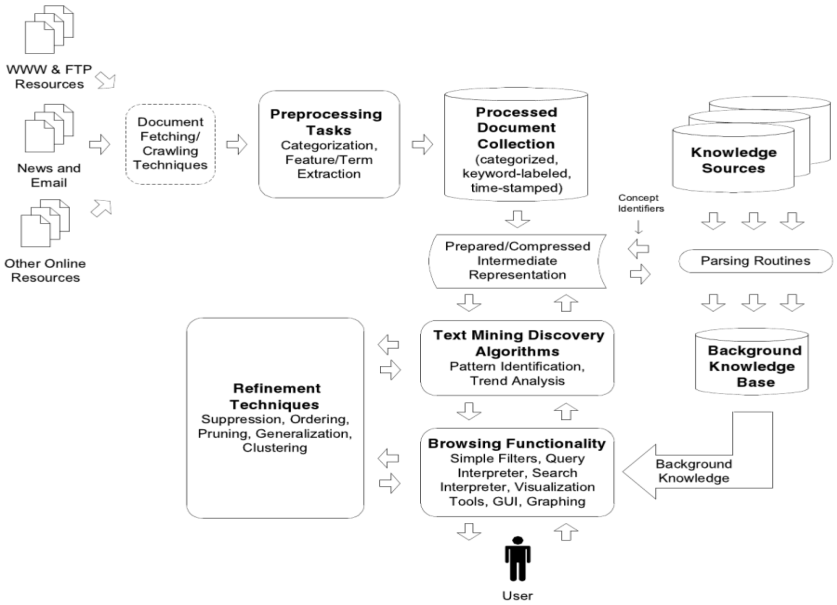

WWW & FTP Resources
News and Email
Other Online Resources
Document Fetching/ Crawling Techniques
Preprocessing Tasks
Categorization, Feature/Term Extraction
Processed Document Collection
(categorized, keyword-labeled, time-stamped)
Knowledge Sources
Concept Identifiers
Prepared/Compressed Intermediate Representation
Parsing Routines
Refinement Techniques
Suppression, Ordering, Pruning, Generalization, Clustering
Text Mining Discovery Algorithms
Pattern Identification, Trend Analysis
Background Knowledge Base
Browsing Functionality
Simple Filters, Query Interpreter, Search Interpreter, Visualization Tools, GUI, Graphing
Background Knowledge
User

FIGURE 2.6: System architecture for an advanced text mining system with background knowledge base

2.3.2 Text Mining

The information age has made it easy to store large amount of data, the proliferation of text and documents available on the web is overwhelming. Text mining refers generally to the process of extracting interesting information and knowledge from unstructured text. Text mining tries to solve the information overload problem by using techniques from data mining, machine learning, natural language processing (NLP) and knowledge retrieval. Text mining involves the preprocessing of document collections such as text categorization, information extraction, term extraction and the storage of intermediate representations such as distribution analysis, clustering, trend analysis and associate rules and visualization of results.

A large number of techniques have been developed to try and deal with the problems that text analysis faces in recent years and fortunately there are a number of articles which provide an excellent survey of the state of the art in text mining [38, 39]. Figure 2.6 provides a good description of the architecture for an advanced text mining system with a background knowledge base. In this project I focus on the identification of sentiment from different text documents which is described in more detail in the following section.

2.3.3 Sentiment Analysis

Sentiment analysis also known as opinion mining uses text analysis, NLP and computational linguistics to identify and extract subjective information in source materials. This is very valuable information as what other people think influences how individuals value commodities which has been described in section 2.3.1 as herd behaviour [40]. This has led to some interesting publications which have used a number of different corpora from the balanced corpus of Antweiller and Frank [41] to relying on commentary columns which appear in the back of financial newspapers such as the Wall Street Journal [42].

One of the early studies that combined text data with evolutionary methodologies was Thomas and Sycara [43] which examined whether the measure of message volume as distinct from message content on internet message boards could be used as an effective predictor of stock price movements. The study used a GP methodology to build trading rules based on the message volume data for a selection of the largest Russell 1,000 stocks. While initial studies looked at raw message count information, the next step is to consider the content or sentiment of these messages in order to assess whether investors are favourably disposed towards a stock [41, 44, 45].

The initial research has been quite positive and has shown that there is some useful information to be retrieved from message boards and all that talk is not just noise [41]. My methodology will be quite similar to these works except I will focus on sentiment analysis in the commodity markets. The design of the actual sentiment analysis system that was used is discussed in the next chapter.

2.4 Aggregation of Variables and Causality

The method of aggregation has crucial effects on the results of the analysis. The movement of commodity prices is highly sensitive to changes in fundamentals of the economy and to the change in expectations about future prospects. Expectations are influenced by the micro and macro fundamentals which may be formed either rationally or adaptively on economic fundamentals, as well as by many subjective factors which are unpredictable and also non quantifiable.

In the globally integrated economy, domestic economic variables are also subject to change due to the policies adopted and expected to be adopted by other countries or some global events. The common external factors influencing the commodity return would be stock prices in global economy, the interest rate and the exchange rate. When conducting analysis into a number of data series it is necessary to use some statistical techniques to ensure that the data series can be accurately compared against one and other. In this section I outline a number of the more sophisticated statistical techniques and formulas that were used in this project, I also list some of the more basic concepts in Appendix A for reference.

2.4.1 Inferential statistics

Inference statistics allow for the deduction of properties of an underlying distribution by the analysis of data. Inference statistics allow us to reach conclusions that extend beyond the immediate data alone. In this project I use a number of inference statistics, which allows for a variety of conclusions to be drawn from the data such as the dependence of two data series or one of the variables may have happened by chance in this study.

Z-score

The z-score is the number of standard deviations a data point is from the mean of the series. It is used to standardise data series allowing them to be compared directly. For example we can compare variations in closing price and volume traded by first converting them to z-scores which brings the two series onto the same scale.

$$z = \frac{x_i - \bar{x}}{\sigma} \tag{2.9}$$

Examining a series after it has been converted to z-scores is also useful for identifying outliers in data. Outliers can either be mistakes in data or rare legitimate occurrences. Either way, they can drastically affect statistical analysis. If they are found consideration should be given to removing them to improve the accuracy of any models created with their data series. Figure 2.7 shows how z-scores are calculated and compares them to a number of other grading methods.

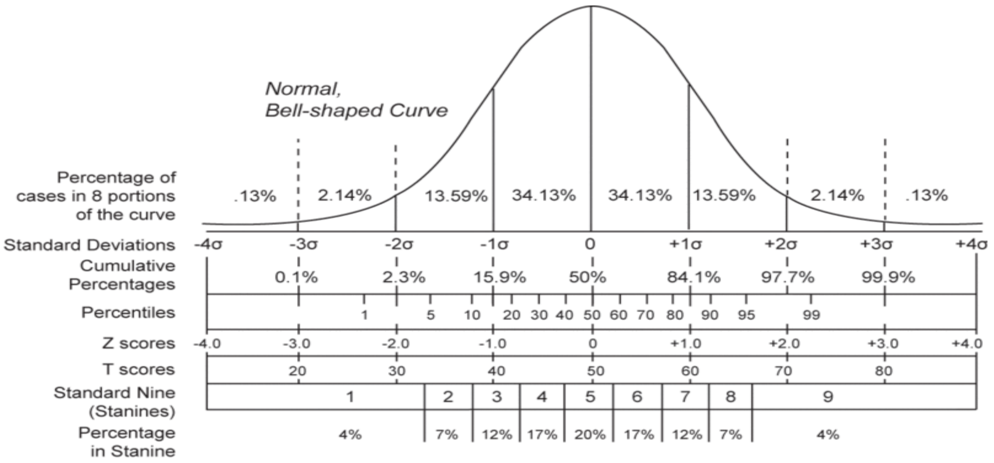


FIGURE 2.7: Comparison of the various grading methods in a normal distribution.

T-Test

T-test is a statistical hypothesis test in which the test statistic follows the t-distribution if the null hypothesis is supported. This value is used to test if the independent variable in question has an effect on the dependent variable. It does this by testing the null hypothesis that a zero coefficient would have the same effect on the dependent variable as the estimated coefficient. A large t-ratio rejects the null hypothesis showing that the coefficient calculated is significantly different from zero.

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} \tag{2.10}$$

where $\overline{x}$ is the sample mean, s is the sample standard deviation of the sample and n is the sample size.

P-Value

The p-value is the estimated probability of rejecting the null hypothesis of a study question when that hypothesis is true, it is the probability of coming up with the coefficient calculated if the true coefficient was actually zero. The p-value determines what significance level the null hypotheses of a zero coefficient can be rejected at. The critical values for this test are 0.1, 0.05 and 0.01 indicating significance levels of 10%, 5% and 1% respectively.

F-Test

The F-test is used for testing hypothesises involving multiple regression coefficients. To do this it takes the number of observations and the number of parameters in a regression and tests the null hypothesis that all the coefficients of the regression are zero. A low f value suggests the variation in the dependent variable can be attributed to chance rather than any meaningful relationship with the independent variables.

$$\sum_i n_i(\bar{Y}_{i\cdot} - \bar{Y})^2/(K-1) \tag{2.11}$$

where $\bar{Y}_{i\cdot}$ denotes the average in the i^{th} group, n_i is the number of observations in the i^{th} group, $\bar{Y}$ denotes the overall mean of the data, and K denotes the number of groups.

Akaike Criterion

The Akaike information criterion is a way of measuring the quality of a model by comparing the models complexity to its goodness of fit, it is an entropy maximization principle [46]. Suppose that we have a statistical model of some data. Let L be the maximized value of the likelihood function for the model, let k be the number of parameters in the model i.e. k is the number of degrees of freedom. Then the AIC value of the model is the following:

$$\mathrm{AIC} = 2k - 2\ln(L) \tag{2.12}$$

Given a set of candidate models for the data, the preferred model is the one with the minimum AIC value. Hence AIC rewards goodness of fit (as assessed by the likelihood function), but it also includes a penalty that is an increasing function of the number of estimated parameters. The penalty discourages over-fitting as increasing the number of parameters in the model almost always improves the goodness of the fit.

Schwarz Criterion

Also known as the Bayesian information criterion, the Schwarz criterion is an alternative to the Akaike criterion for comparing models. This method also uses the log likelihood to measure goodness of fit but the difference between this and the Akaike criterion is it applies a heavier penalty to over specification of the model. Over specification in the context of regression analysis means having more independent variables than necessary to model the dependent variable.

$$\mathrm{SC} = -2 \cdot \ln \hat{L} + k \cdot \ln(n) \tag{2.13}$$

where x is the observed data, θ is the parameters of the model, n is the number of data points in x, $\hat{L}$ is the maximum value of the likelihood function of the model.

Rho

Rho, or ρ is the correlation coefficient it is a measure of how well the relationship between the dependent variable and the independent variables is described using a linear model. It can range from -1 to 1 with -1 indicating a perfect negative linear relationship, 0 indicating no linear relationship and 1 indicating a perfect positive linear relationship.

Durbin-Watson

The Durbin-Watson statistic is a test statistic used to detect the presence of autocorrelation in the residuals from regression analysis.

If e_t is the residuals associated with the observation at time t, then the test statistic is:

$$d = \frac{\sum_{t=2}^{T} (e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2}, \tag{2.14}$$

where T is the number of observations. Since d is approximately equal to 2(1 - r), where r is the sample autocorrelation of the residuals, d = 2 indicates no autocorrelation. The value of d always lies between 0 and 4, if the Durbin–Watson statistic is substantially less than 2, there is evidence of positive serial correlation.

If the Durbin–Watson statistic is small such as less than 1.0, it indicates successive error terms are, on average, close in value to one another, or positively correlated. If d ¿ 2, successive error terms are, on average, much different in value from one another, i.e., negatively correlated. In regressions, this can imply an underestimation of the level of statistical significance.

2.4.2 Vector Autoregression

To analyse the results of this project which involves long time series and a number of dynamic variables there are a number of analysis techniques that can be used, for example multivariate generalized ARCH models within simultaneous equations systems [47]. Sims [15] advocated the use of VAR models which were able to describe the dynamic structure of multiple variables. They can also be used for economic analysis however, because they describe the joint generalised mechanism of the variables involved.

Traditionally VAR methods have been designed for stationary variables without time trends. Trending behaviour can be captured by including a number of deterministic polynomial terms. The importance of stochastic trends in economic variables and the development of the concept of co-integration by Granger [48] and other have shown that stochastic trends can be captured by VAR models. If there are trends in some of the variables it may be desirable to separate the long-run relations from the short-run dynamics of the generation process of a set of variables. Vector error correction models offer a convenient framework for separating long-run and short-run components of the data generation process (DGP).

One of the key papers that has been published in the area of sentiment analysis in recent years is by Tetlock [42]. In this paper Tetlock proposes the use of VAR estimations which include all lags up to 5 days prior to market activity. This allows for the modelling of endogenous and exogenous variables over a large time series. This simple framework provides a systematic way to capture rich dynamics in multiple time series. As Sims [15] and others argued in a series of influential early papers, VARs held out the promise of providing a coherent and credible approach to data description, forecasting, structural inference, and policy analysis. A more detailed description of VAR is given in the Econometric Analysis section 2.2.2

2.5 Conclusions

This chapter has covered a large body of work which makes the design of the system much easier as I can build on the work of the current state of the art systems. I have tried to make this a self-contained document and to be able to justify my design decisions from the initial literature review to the final conclusions. In this chapter I have looked at the conventional analysis which has been explored in the past and how this has been expanded on by developments in economic analysis and the modelling of exogenous variables. I have also looked at qualitative analysis and how market behaviour can be used to predict changes in the market, this led to the discussion of text mining and sentiment analysis to try to estimate the mood of the market.

This chapter also contains a brief investigation into how exogenous variables can be included in the model to improve performance, I look at the modelling of shipping and weather as exogenous variables that could be included in the model. Section 2.4 on the aggregation of variables and causality is important and is overlooked by many other projects, even though it can have a profound effect on the results. In this section I have described some of the possible techniques that can be used, after careful consideration of the current literature on the subject I have decided to use VAR analysis due to its success in a number of related projects such as Tetlock [42].

Chapter 3

Methods & Data

3.1 Introduction

This chapter outlines the methods and the data that were used throughout the project, because of the broad nature of this project it involved collecting, aggregating and cleaning data from a wide variety of sources. The chapter is broken into sections which describe the data that has been collected, the analysis that has been used and a final conclusion section which gives a summary of the chapter. The data section is broken down into a number of subsections each of which deal with the collection of a specific set of data for the project, and the analysis section deals with the different forms of analysis that were used throughout the project.

3.2 Data

This section describes the methods that were used for the collection of data for our hypothesis to be tested on. There are three main datasets which are used in this project historical weather data, historical text corpus and the historical commodity prices. Each of these data sets are broken down into subsections which describe how the data was collected and cleaned so that it could be used by the final system.

3.2.1 Weather Data

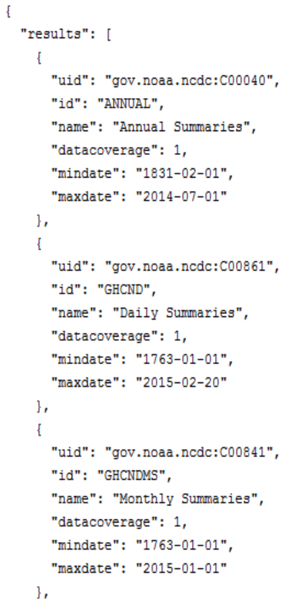
To analyse the impact that weather has on the commodity market it is important to have a large collection of historical weather data. Fortunately there are a large range of resources that have become available which give access to historical weather such as the United States National Climate Data Center (NCDC). Their website provides access to historical weather data for the United States through the use of an API `http://www.ncdc.noaa.gov/cdo-web/webservices/v2`. Another excellent resource that I found for access to historical weather is the weather underground site `http://www.wunderground.com/weather/api`. Both of these services provide excellent documentation and are free to use up to a certain number of requests.

NCDC

The United States National Climate Data Center (NCDC) has the worlds largest archive of weather data. The center has a historical record of more than 150 years of data and from 2004 to 2013 its digital archive has increased from 2 to 14 petabytes. This shows the vast amount of weather data that is available from this center and increasing sophistication of data collection equipment. New satellites and data holdings are expected to grow exponentially in the next 10 years.

The NCDC has a number of different datasets and a number of methods for interacting with these datasets. One of the easiest ways to query the large selection of data is through the use of the API. The website provides excellent documentation to the different datasets that can be accessed through the use of the API `http://www.ncdc.noaa.gov/cdo-web/webservices/v2`.

Figure 3.1 shows some of the datasets that are available through the use of the API, it also shows how large the collection is with daily summaries going back to 01-01-1763 for some stations. I found that the API was quite slow at responding to calls and I would occasionally get time-out errors when making calls. There also did not seem to be an active a community using the API which meant that there were few resources when problems where encountered with the API. I think that this resource it better for access to historical records that are much older than was needed for the project, so I decided to look at some other options to allows easier access to more recent weather information.



```
{
  "results": [
    {
      "uid": "gov.noaa.ncdc:C00040",
      "id": "ANNUAL",
      "name": "Annual Summaries",
      "datacoverage": 1,
      "mindate": "1831-02-01",
      "maxdate": "2014-07-01"
    },
    {
      "uid": "gov.noaa.ncdc:C00861",
      "id": "GHCND",
      "name": "Daily Summaries",
      "datacoverage": 1,
      "mindate": "1763-01-01",
      "maxdate": "2015-02-20"
    },
    {
      "uid": "gov.noaa.ncdc:C00841",
      "id": "GHCNDMS",
      "name": "Monthly Summaries",
      "datacoverage": 1,
      "mindate": "1763-01-01",
      "maxdate": "2015-01-01"
    },
```

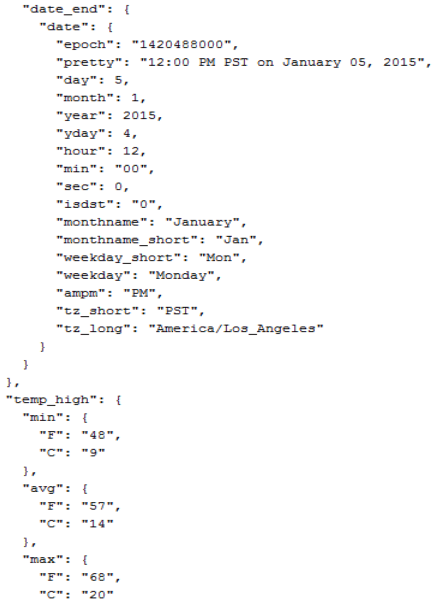
Figure 3.1: NCDC JSON Response

Weather Underground

Weather Underground is a commercial weather service that provides real-time weather information over the internet. It is based in the US and has over 100,000 weather stations across the country. They provide an API which can be used to get access to historical weather information. They provide a developer plan which allows free access to the API but restricts usage to 500 calls per day and 10 calls per minute. This allows for access to historical weather as well, however to get access to their next plan with historical weather cost $520 per month this allows access to 5000 calls per day and 100 calls per minute.

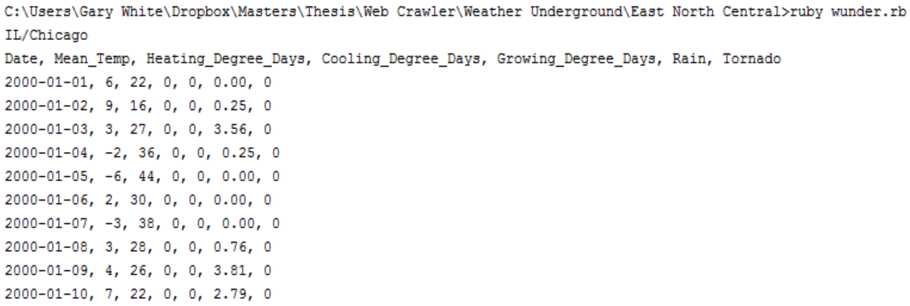
There is an active community involved with the API which has led to a number of ruby gems being written. One of the best is gems is Wunderground which allows for easy access to the API. All requests are made over HTTP and data features are returned in JSON or XML. Figure 3.2 shows how a typical request is formatted using JSON. The API provides access to lots of information such as minimum and maximum temperatures, but it also provides percentage estimates of shock weather events such as tornadoes and hurricanes. Figure 3.3 shows how these request were processed and how I was able to generate a .csv file containing all the information.

The developer plan provides 3 raindrops which are tokens which you can use to get unlimited access to the API for 24 hours. Registration of a key only requires an email



```
  "date_end": {
    "date": {
      "epoch": "1420488000",
      "pretty": "12:00 PM PST on January 05, 2015",
      "day": 5,
      "month": 1,
      "year": 2015,
      "yday": 4,
      "hour": 12,
      "min": "00",
      "sec": 0,
      "isdst": "0",
      "monthname": "January",
      "monthname_short": "Jan",
      "weekday_short": "Mon",
      "weekday": "Monday",
      "ampm": "PM",
      "tz_short": "PST",
      "tz_long": "America/Los_Angeles"
    }
  }
},
"temp_high": {
  "min": {
    "F": "48",
    "C": "9"
  },
  "avg": {
    "F": "57",
    "C": "14"
  },
  "max": {
    "F": "68",
    "C": "20"
```

Figure 3.2: Weather Underground JSON Response



```
C:\Users\Gary White\Dropbox\Masters\Thesis\Web Crawler\Weather Underground\East North Central>ruby wunder.rb
IL/Chicago
Date, Mean_Temp, Heating_Degree_Days, Cooling_Degree_Days, Growing_Degree_Days, Rain, Tornado
2000-01-01, 6, 22, 0, 0, 0.00, 0
2000-01-02, 9, 16, 0, 0, 0.25, 0
2000-01-03, 3, 27, 0, 0, 3.56, 0
2000-01-04, -2, 36, 0, 0, 0.25, 0
2000-01-05, -6, 44, 0, 0, 0.00, 0
2000-01-06, 2, 30, 0, 0, 0.00, 0
2000-01-07, -3, 38, 0, 0, 0.00, 0
2000-01-08, 3, 28, 0, 0, 0.76, 0
2000-01-09, 4, 26, 0, 0, 3.81, 0
2000-01-10, 7, 22, 0, 0, 2.79, 0
```

Figure 3.3: Weather Underground Processed Information

address so it easy to build up an array of keys using temporary email addresses. Using this method I was able to get enough access to the API to build up the collection of weather data that was needed.

Figure 3.5 shows the time series of this data and the variation between the different regions as the area that they take up is relative to the amount of heating degree days. We can see from this figure how the number of heating degree days follows the seasonal pattern as we would expect but also there are quite large variations in the weather series on a yearly basis. This gives a good opportunity to see how these variations in temperature caused variations in the price of commodities over this time period. To get the national aggregate figure I grouped each of the regions into national level region which can be seen on the map as either North, South, East or West region. Once they were aggregated to this level I could combine the regions based on population data from the US national census which can be seen in Table 3.1.

Region	Population	Percentage
Northeast	56,152,333	17.6%
Midwest	67,745,108	21.2%
West	75,187,681	23.6%
South	119,771,934	37.6%

Table 3.1: 2014 US Population by Region

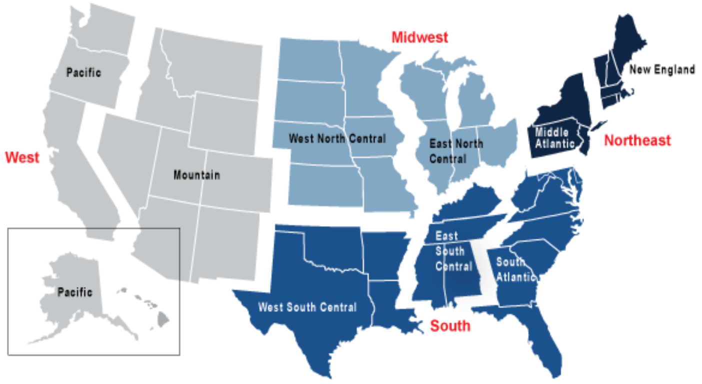


Figure 3.4: Divisions of the United States

This is an important step as it allowed regions with the most amount of people and therefore houses to have the greatest impact on the national model. The south has the biggest impact on the national figure as it contains 37.6% of the population.

3.2.2 Text corpus

To preform sentiment analysis a large collection of text is required this large collection of structured text is called a corpus. A corpus is a systematically motivated collection of texts to provide evidence about the use of language in general and the evidence tends to be more credible if the text collection comprises texts dealing with a specialist area of human endeavour. The work in corpus linguistics and in related areas of content analysis is a forerunner of modern sentiment analysis which is based on the observation that frequency of usage of a token relates directly to its acceptability in a community of language users Quirk et al [49] for work in general language and Ahmad [50, 51] for specialist languages.

Recent work in financial sentiment analysis has led to the creation of diachronically well designed corpora, covering both periods of recession and expansion as described by Garcia [52]. Some corpora are synchronically well designed in that the researcher

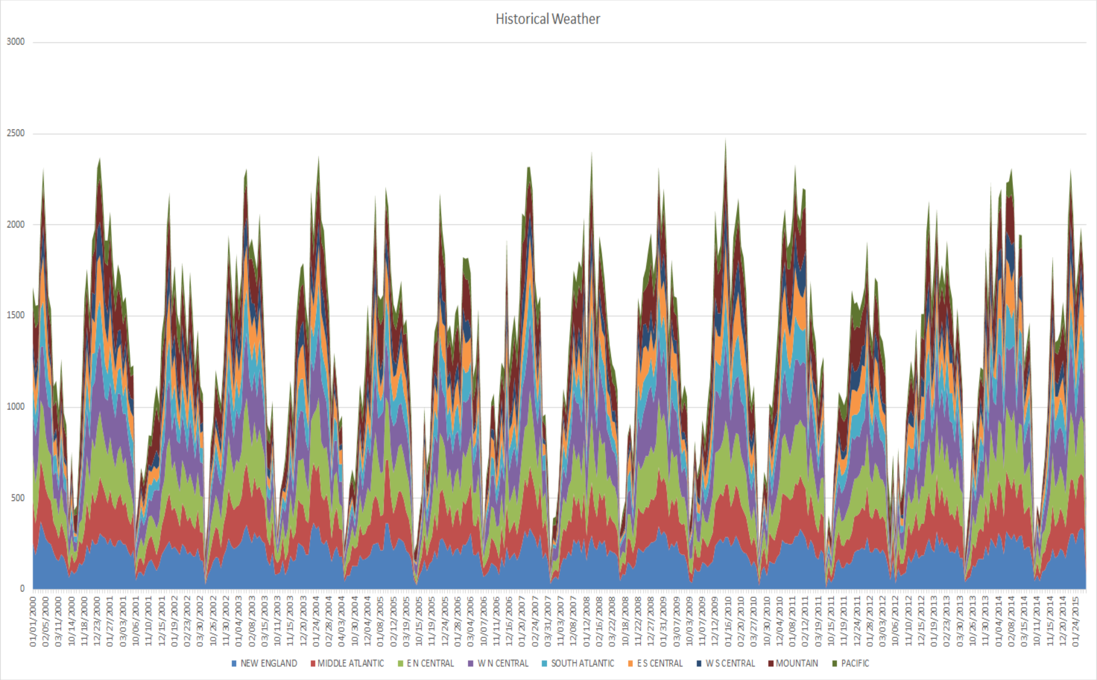


Figure 3.5: Stacked heating degree day series by region

has included texts from social media such as financial blogs as well as more formal print/electronic media sources such as financial newspapers as in the case of Antweiller and Frank [41]. Other researchers are content with only a single text type and usually rely on commentary columns appearing in the back of newspapers as in the case of Tetlock where he uses the Abreast of the Market column in the Wall Street Journal for a 10 year period containing the 1987 crash. Garcia [53] has a centenary collection of such similar columns in the New York Times, running from 1907-2008. I explored two methods of building a collection of sentiment data which are described in the following sections.

Web Crawler

My initial approach to building my own corpus was to use a web crawler to build up a database of text from the internet. There are a number of excellent resources online with advice on how to build a webcrawler, as I already had experience with ruby I decided to look for some web crawler gems that would allow for a much simpler implementation than to having to create the crawler from scratch. I designed my system using the mechanize gem, which allowed me to start collecting texts relatively quickly. I was easily able to set up a system capable of following links and collecting the text from web pages.

After achieving some initial success using the web crawler I discovered that a lot of the historical articles that I was searching for where behind paywalls for most of the major newspapers. This meant that I has to change from my initial idea of building a web crawler to using an academic resource called Lexis Nexis which would provide free access to the historical articles that I needed.

Lexis Nexis

I have used a public source which is available to most academic colleagues for building my English language text corpus where a majority of articles deal with issues in the commodities market and were published in the USA and the UK over a 15 year period. Lexis Nexis sources its news from various news vendors including those specialising in news sources in geographical location, and also has the ability to translate from the language spoken in the location to English. Lexis Nexis News and Business comprises articles from newspapers, newswires, press releases, trade magazines and web-based publications. The Lexis-Nexis service provides access to 1782 formal news sources, ranging from papers with international circulation and small regional newspapers that appear in print and electronically, 37 social media/blog sites and 115 magazines dedicated to business and management.

The use of a broad range of information sources is motivated by a desire to construct a representative corpus of texts such that evidence derived from the corpus has a minimal bias of the investigator. In order to conduct time-series analysis of sentiment it is essential that there is a large pool of news stories in order to minimise the number of days with no articles about a particular commodity.

As commodities are often written about in groups it is quite possible that other commodities may be mentioned in the articles collected. I therefore have only selected those news items whose primary topic is a particular commodity. Techniques developed for retrieving texts using keywords have led to statistical methods and metrics that can be used as a proxy for overall relevance of a text to the keywords, or the concepts based on the frequency distribution of one or more keywords [54]. These metrics include the

Commodity	No. of articles
Brent	2054
WTI	2342
Natural Gas	1900
Gasoline	2652

TABLE 3.2: Number of articles collected for sentiment analysis

pragmatic layout of the texts as well, if a term appears in the title of a document then the document will be more relevant to the term compared to a document where the term appears as part of a paragraph. The frequency of a term in a document will also be used as a key factor in determining the relevance of a document to a particular keyword.

In compiling my corpus I selected articles that had the commodity in the headline as well as at least three mentions in the text. Lexis Nexis also offers a number of relevance options to ensure that the retrieved articles were written specifically about each particular commodity so I choose to only include strong reference. I was also able to remove any duplicate articles that may have appeared in different newspapers or been posted online. This is essential because articles are often duplicated as a result of national and international syndication and also because websites often pick up and repeat newspapers and other articles and commentaries.

The results of the corpus collection can be seen in table 3.2, which shows that a large amount of texts were able to be collected for each of the commodities after all the duplicates had been removed. We can see that Gasoline had the most amount of articles over the period which could have been expected as gasoline prices are frequently mentioned in news articles. It is clear that it would have been almost impossible to build a web crawler in the time frame of this project that would have been able to build up a collection of data as large as the one that is readily accessible in the Lexis Nexis database. It would also have been much more difficult to include relevance indicators which are readily accessible using Lexis Nexis.

3.2.3 Commodity Data

There is a large amount of resources online for collecting historical commodity prices, some of them provide analysis as well as price and some require payment to access the data. The best resources that I found were *Quandl* and the *World Bank*. The world bank provides access to a large amount of different data sets including commodity data. The data contains indices from 1960 to present, updated on the third working day of each month. I felt that monthly data however would possibly miss some events that that were happening in the news or weather and wouldn't be able to provide as accurate results, therefore I decided to access the more frequently recorded data on Quandl.

Quandl is a data platform that hosts data from hundreds of publishers on a single website. This makes it much easier to get access to the correct data source as they already cover 10 million datasets from 500 sources, and also allow the data to be downloaded in a number of different formats. Quandl also has an API which makes it easy to query the large amount of datasets that are available on the site. Table 3.3 shows the set of commodities that were collected for this project and also the frequency and date. The table also includes the source where Quandl has aggregated the data from.

Commodity	Frequency	Start Date	Source
WTI Crude Oil	Daily	01/01/1986	US Dept. of Energy
Brent Crude Oil	Daily	20/05/1987	US Dept. of Energy
New York Harbour Gasoline	Daily	02/06/1986	Federal Reserve Economic Data
Henry Hub Natural Gas	Daily	01/05/1991	US Dept. of Energy

TABLE 3.3: Available historical data

3.3 Exogenous & Qualitative Analysis

This section describes the methods that were used in the analysis of the exogenous and qualitative variables. They are quite different systems and require different analysis techniques which is why they have been broken down into two different sections. This section describes the methods that were used in the analysis of the two different systems, the results are published and discussed in the next chapter.

3.3.1 Weather Analysis

In this section I aim to put a price on the large collection of historical temperature data that had been collected using the methods described in section 3.2.1. The history of modelling weather is quite old and one of the first works on the probabilistic modelling of precipitation was by Quetelet who in 1852 reported that runs of constitutive rainy and dry days in Brussels exhibited persistence [55]. Since then there has been a large

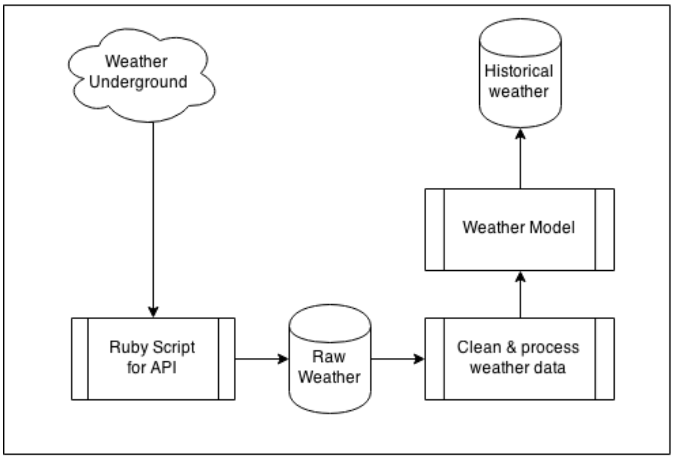


FIGURE 3.6: Architecture of the weather analysis system

number of developments in the collection of data and the analysis techniques that are used.

The basic outline of the system can be seen in figure 3.6, in section 3.2.1 I discuss the initial portion of the project where I collect the data from the Weather Underground public API. In this section I describe the analysis that was conducted on this collection of data. I use the research outlined by Benth and Benth [56] as the basis for my weather model which estimates the daily price as a time-discrete stochastic process P_t with days as time units. I use an arithmetic model with fundamental deterministic components describing the yearly seasonality and a function of temperature. I include a stochastic process to describe the short term fluctuations and long-term uncertainty of the model:

$$P_t = S_t + f(\Theta) + \Lambda_t + Y_t \tag{3.1}$$

where S_t is the yearly seasonality, $f(\Theta)$ is a function of the mean temperature, (Θ) itself is a stochastic process, Λ_t is the normalized cumulated heating degree days and Y_t is a long-term stochastic process.

Yearly Seasonality

As the temperature changes throughout the seasons it is natural that the consumption of energy would be seasonal. In this section I outline how the yearly seasonality S_t is modelled by using a trigonometric function with a linear trend.

$$S_t = a \sin\left(\frac{2\pi t}{365.25}\right) + b \sin\left(\frac{2\pi t}{365.25}\right) + ct + d \tag{3.2}$$

where the factors a and b describe the height of the amplitudes and the phase shift of the yearly seasonality, c describes the deterministic drift and d describes the intercept. The regression problem is solved using ordinary least-squares regression.

Cumulative Heating Degree Days

It is reasonable to assume that temperature plays an important role in the commodity spot price market behaviour. However, with more precise analysis it can be seen that using temperature directly does not improve the quality of the model significantly compared with a purely seasonal model. It can be seen that in long and cold winters prices are higher and in warm winters prices are below the mean level. Therefore, we use an aggregate of temperature during winter periods. Cumulating heating degree days over a winter leads to a number indicating how cold the winter was in this context I refer to the cumulative heating degree days over each year:

$$CHDD = \sum_{k=1}^{d} HDD_{k,w} \tag{3.3}$$

The impact of cumulated heating degree days on the price depends on the comparison to a normal winter. This information is included in normalized cumulated heating degree days:

$$\Lambda_{d,w} = CHDD_{d,w} - \frac{1}{w-1}\sum_{k=1}^{w-1} CHDD_{k,w} \tag{3.4}$$

I use Λ_t instead of $\Lambda_{d,w}$ w for simplicity, if t is a day in a winter. The definition of Λ_t for a summer day is described by a linear return to zero during summer. Although there might be cold days between April and September this time of the year is usually used to refill gas storages. The impact of cold days on the price decreases due to increasing filling levels. We take account for this situation by the linear part of normalized cumulated heating degree days as seen in Figure 3.7. Positive values of $\Lambda_{d,w}$ describe winters colder than the average. Λ_t is included into the gas price model by a regression approach.

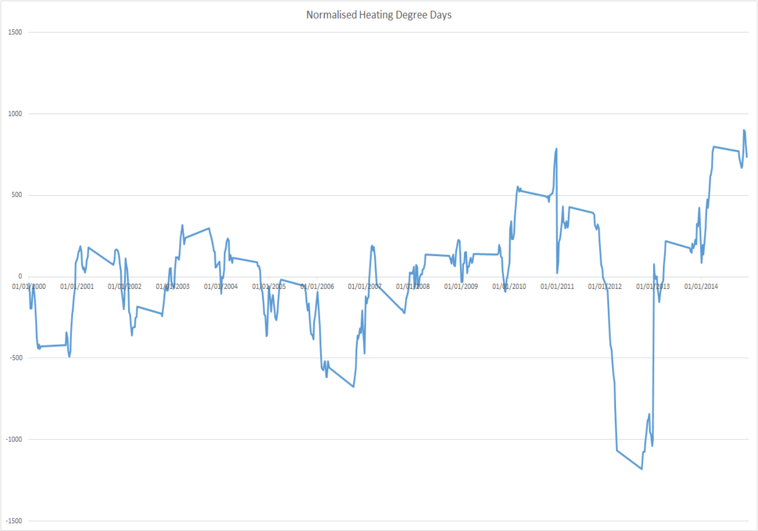


FIGURE 3.7: Normalized cumulated heating degree days for 15 year period.

The temperature process

For temperature I use an additive model with yearly seasonality and linear trend W_t and a stochastic component X_t:

$$\Theta = W_t + X_t \tag{3.5}$$

where W_t represents the yearly seasonality and the trend:

$$W_t = a \sin\left(\frac{2\pi t}{365.25}\right) + b \sin\left(\frac{2\pi t}{365.25}\right) + ct + d \tag{3.6}$$

where the factors a and b describe the height of the amplitudes and the phase shift of the yearly seasonality, c describes the deterministic drift and d describes the intercept. The stochastic component of the temperature process is modelled as a normal-inverse gaussian (NIG) which is a special case of the generalized inverse Gaussian distribution.

The long term process

The stochastic mean-reverting temperature and the additional short-term processes model very short-term fluctuations. For times t far in the future, volatilities for long-term forward contracts tend to zero [57]. In addition, the forward dynamics do not follow

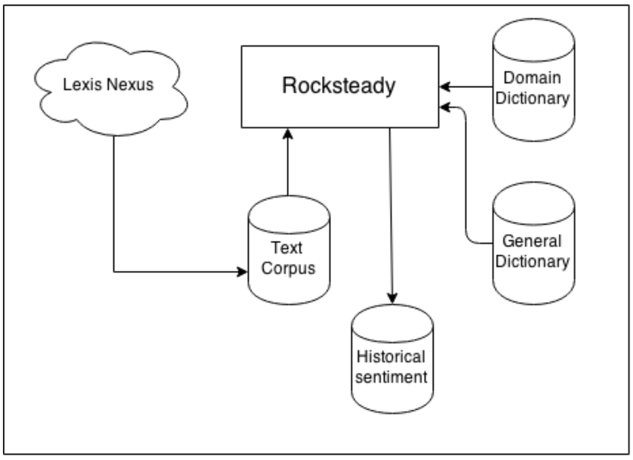


Figure 3.8: Architecture of the sentiment analysis system

a mean-reverting process. Therefore we have to add an additional long-term process Yt which is modelled by a geometric Brownian motion given by:

$$Y_t = Y_0 \; exp((\mu - \tfrac{1}{2}\sigma_Y^2)t + \sigma_Y \epsilon_t^Y) \tag{3.7}$$

where the ϵ_t^Y are serially uncorrelated normally distributed random variables. Once each of the individual parts of the model have been completed it only remains to aggregate them using equation 3.1. The results of this aggregation can be seen in the results and discussion chapter section 4.3.

3.3.2 Sentiment Analysis

The outline of the sentiment analysis system can be seen in figure 3.8, I have explained my reasons for using the Lexis Nexis system in section 3.2.2 and will in this section discuss the analysis of the corpus of text that has been built up. The are a number of different approaches and applications that can be used to conduct sentiment analysis on a corpus of text. In this project I use the Rocksteady software package that has been developed in Trinity College Dublin.

The Rocksteady package allowed for the input, analysing and display of the sentiment from the corpus that had been collected as described in the previous section. It allows for the corpus to include numerical time series data of prices and volumes of one or more continuously traded entities. The system conducts affect analysis and computes the impact of affect on the prices of the entity and vice versa. The focus of the system is on finding the distribution of affect words in text, the search for affect words can be integrated simultaneously with search for pre-specified keywords.

The corpus is analysed using an ontologically organised dictionary that comprises specific information about the entity, and information about affect words indicating sentiment expressed, strength associated with the sentiment and intended actions. The ontological information comprises the details of the relationship of the entity with other entities. This allows the information about the entity and affect to be maintained separately and merged by the system at run time. After researching some other sentiment analysis systems I found that they did not have this aggregation of domain specific information with affect information sets which is one of the reasons that I choose Rocksteady.

The system can compute the distribution of pre-specified keywords, using term frequency and inverse-document frequency, at different scales. The distribution of keywords, which may include names of organisation, places or people, and can also be used. The results of affect analysis can be made sensitive to different time and spatial scales by computing not only the raw sentiment score, or the score normalised to the length of the text comprising the affect words, but by computing the deviation in the frequency of affect words in each individual texts when compared the texts collected and analysed to date.

One of the most important decisions when conducting sentiment analysis is to choose the correct dictionary or collection of dictionaries to be used. One of the advantages of the Rocksteady system is that is allows for the use of multiple dictionaries. I use a General Language dictionary as my base dictionary and an integrated oil & financial economics dictionary for the specialist knowledge dictionary. This is important as a number of financial terms can have negative meaning in a general language dictionary but the financial meaning would be different such as crude oil, or share. In these cases crude in a regular dictionary would have a negative sentiment and share would have a positive sentiment but there actual financial meaning is totally different. The specialist dictionary over-rules the general dictionary in all cases of conflict, which ensures that we are always getting the best estimation of sentiment for each word.

3.4 Algorithmic Trading

In this section I make a practical application of the research that has been conducted in the previous sections on weather and sentiment analysis, I use this analysis in a simple algorithmic trading strategy. In this section I give a brief introduction to algorithmic trading and the strategies that were used in my system. There are a large number of statistical techniques and methods which can be used in algorithmic trading which has evolved from the original 'pairs trading' which was first used by Morgan Stanley in 1985. This strategy finds a pair of stocks with similar historical price behaviour and when the prices of the stock diverge bet on a subsequent convergence.

One of the books which I used while researching this topic is by Andrew Pole [58], who provides a great introduction to a number of techniques which can be used to trade algorithmically. This book introduces a number of techniques like mean-reversion, noise and structured models. One of the other articles that I found most useful to show the limitations of algorithmic trading and which outlines a number of real world cases where algorithmic trading has been used and has failed catastrophically is by Kirilenko [59]. This paper highlights a number of vulnerabilities created by algorithmic trading such as the flash crash in 2010 and Knights Capital in 2012 which lost more than $457.6 million after 'a technology issue at the open of trading at the NYSE related to a software installation that resulted in Knights sending erroneous orders into the market'.

In this section I describe the simple trend following strategy that was used to analyse the effectiveness of weather and sentiment analysis when applied to an algorithmic trading strategy. This builds upon the research that was carried out in the previous sections to collect and analyse the data and shows how this research can have commercial applications such as in algorithmic trading systems.

Trend Following

One of the most common trading strategies within the commodities market is trend following. Trend following is an absolute momentum strategy in that it assumes that a particular commodity is trending if its latest closing price is above the highest high or below the lowest low over the predefined look back window, which is briefly described in section 2.2.1. It is also possible to use the weather and sentiment analysis which has been calculated in the previous section to try and predict changes in the market. I use Matlab to implement my trading system and base the design off *this* webinar.

Although trend following is quite a simplistic strategy is has been known to produce good results especially in the commodities market [60]. In my algorithmic trading system I go long on a particular commodity at the next period's opening price if the previous period's closing price is higher than the highest high in a specified look-back window. I compare this traditional trend following strategy which the results that are observed using trends in the weather and sentiment analysis.

Each of these trading techniques are then looked at for a number of different look back windows and holding periods. The number of different look back windows shows how variable even the same strategy can be in the market an why it is so difficult to build an effective algorithmic technique. The comparison of different look back windows shows how strategies can have different levels of volatility and risk.

3.5 Conclusions

This chapter has outlined a number of the challenges and decisions that were faced when collecting the data and exploring the different analysis methods that could be used in this project. In the data section I explored a number of sources that provide access to historical data. For the historical weather collections I explored the use of the Weather Underground API as well as the NCDC website. I decided to use the Weather Underground API as it allowed for more broad data queries and has a more active community. For the historical corpus collection I decided that it was better to use the Lexis Nexis dataset to get a more complete coverage of the media than could be obtained by the use of a web crawler. Finally I decided to use Quandl to access the historical commodity prices as it had a public API and a large dataset of daily commodity prices.

In this chapter I also decided on the analysis techniques that would be used over the course of the project. In section 3.3.1 I outline the mathematical model that will be used to analyse the large collection of historical weather data that has been collected. The model that is outlined is developed from the research that was carried out in the literature review section. The next section deals with the sentiment analysis system that I would be using for the project. After much deliberation I decided to use the Rocksteady program which has been developed in Trinity College for sentiment analysis, this was due to its ability to deal with a large corpus and use multiple dictionaries.

In the final section I also describe a simple algorithmic trading strategy which uses the analysis which has been conducted in the previous sections. This shows that the research that has been conducted has real world applications. This section also describes some of the limitations of algorithmic trading strategies and how a number of safety measures must be incorporated into the system to avoid catastrophic errors.

Chapter 4

Results & Discussion

4.1 Introduction

This chapter presents the results that were observed over the course of the project and a brief discussion of what can be learned from these results which is expanded upon in the next chapter. The analysis is broken down into a number of different sections which concentrated on individual portions of the project. These are all then discussed in the final conclusion section of the chapter.

4.2 Sentiment Analysis

This section presents the results that were observed for the sentiment analysis section of this report. There are a number of different statistical methods that can be used to measure the influence of the sentiment analysis which are described in Section 2.4. The first subsection uses OLS regression and shows how this method of analysis is not suitable for analysing a large financial time series, I then use VAR analysis and show how this overcomes the faults of OLS regression.

4.2.1 OLS Regression

When analysing the impact of sentiment on the value of WTI I converted the price to a zscore so that the two time series could easily be graphed against each other, I also did the same for the return values which can be seen in figure 4.1 and 4.2. These figures provide a good initial insight into the data that we are observing and the patterns that are being followed. We can see from figure 4.1 how the sentiment appears to lag the actual market price, and from figure 4.2 how the daily market returns are much more volatile than the sentiment analysis.

Figure 4.2 shows how violent the market can change and the amount of outliers that are present in the series, we can see that we get zreturns of over 10 which is a large deviation from the mean. The probability of being 10 standard deviations from the mean is very small (1.524×10^{-23}) so we can see how difficult it is to predict next day returns. After getting some initial insights into what to look for in the data it necessary to use some the statistical techniques that have been outlined already to provide a comparison to other research.

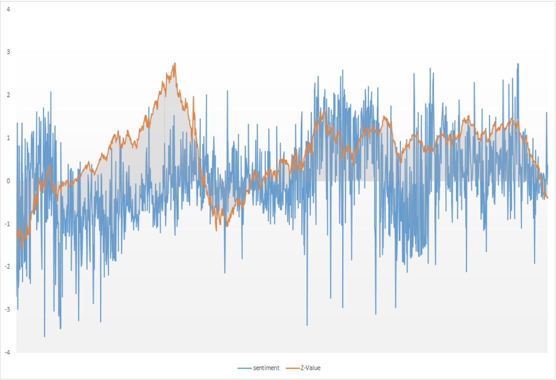


FIGURE 4.1: Sentiment v ZValue WTI

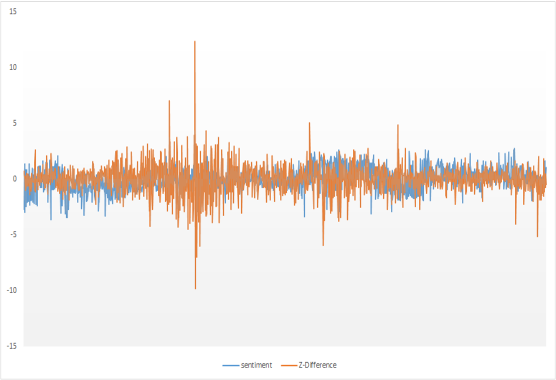


FIGURE 4.2: Sentiment v ZReturn

The first technique that was used to find the relationship between sentiment and the return value of commodities was ordinary least squares. I model the return value with a lag in sentiment value to analyse the impact of sentiment in the series. As I conducted more analysis into the fundamentals of the OLS method it became clear that it wasn't the best method of regression available for the data series being analysed. One failure of OLS for the needs of this project is that one of the assumptions it works on is that the independent variables cannot be correlated with each other. To get around the problem of the data violating this assumption it was decided to replicate Tetlocks regression analysis to establish a link between sentiments and return using vector auto regression.

In this section I present the OLS results for WTI crude oil to give an overview of how this analysis was conducted. When concentrating on the other commodities that were analysed as part of this series I only present the VAR results, this allows for direct

	Coefficient	Std. Error	T-Ratio	P-Value
sentiment_1	0.0199820	0.1349473	1.404	0.4163
sentiment_2	0.0222803	0.1212519	0.660	0.2140
sentiment_3	0.0545211	0.0197218	2.765	0.0058
sentiment_4	0.0456026	0.0197695	2.307	0.0212
sentiment_5	0.0463629	0.2253422	1.829	0.5675

R-squared	0.040021	Adjusted R-squared	0.036641
F(6, 1393)	1.725364	P-value(F)	0.111458
ρ	0.994476	Durbin-Watson	0.010555

Table 4.1: WTI Return OLS Regression

comparison between the analysis results from the different commodities. The major results of the OLS regression analysis can be seen in Table 4.1, the explanation of these statistical estimates is given in Section 2.4. From this table we can observe some interesting results, we can see that the P-value reaches its minimum at sentiment_3 and from that point gradually reduces in significance. The large T-Ratio combined with a low P-Value rejects the null hypothesis and shows that sentiment analysis has a statistical significance on the return price of commodities.

We can also observe how the length of the sentiment lag also has a large impact on its ability to predict the return value. We can see in this case that sentiment_3 which is a lag of 3 days returns the results with the least amount of standard error. This observation was not true for all of the commodities and it seemed to be an individual attribute to each commodity, as to how fast the news would impact the price. The R^2 values indicate that although there is a large amount of statistical significance to the sentiment values against the return price it can be unpredictable and have a large amount of outliers.

From this analysis we are able to observe a number of key results about the data, such as whether there is a linear relationship between the variables in the regression. We can look at figure 4.3 which plots actual values of the dependent variable against the values predicted by the model. In this case we can see that due to the large amount of values that have been observed over such a long period of time that there are a large number of outliers in the system. I also test for Heteroscedasticity using White's Test, the results of which are:

$$Test\ statistic:\ LM = 42.7515$$

$$with\ p{-}value = P(Chi - square(27) > 42.7515) = 0.0277103$$

The null hypothesis of this test is that heteroskedasticity is not present. Since there is a high p-value the null hypotheses can be accepted, therefore the conclusion is that the errors are homoscedastic. We can also look at figure 4.4 to check the assumption of normally distributed errors. In this case we can clearly see that the errors are distributed normally as the points on the plot fall close to the line.

The reason that OLS regression is not very effective in this data set is that there is an assumption of independent errors and regressors. To test for this assumption we use the Durblin-Watson test, the result of which is outlined in table 4.1. The Durbin-Watson statistic ranges in value from 0 to 4, a value near 2 indicates non-autocorrelation, a value toward 0 indicates positive autocorrelation and a value toward 4 indicates negative autocorrelation. We can see from the table that our results have a value close to zero which indicates a positive autocorrelation. This violates one of the assumptions of OLS, to solve this I used VAR analysis, the results of which are described in the next section.

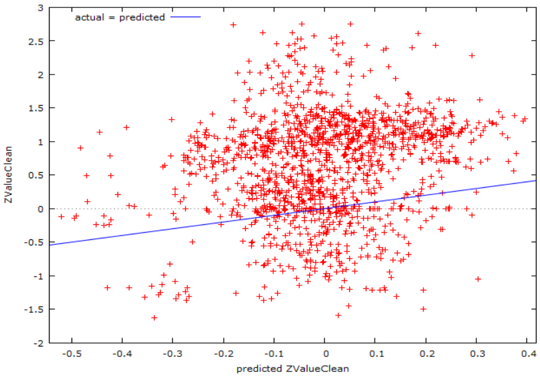


FIGURE 4.3: Actual v Predicted WTI

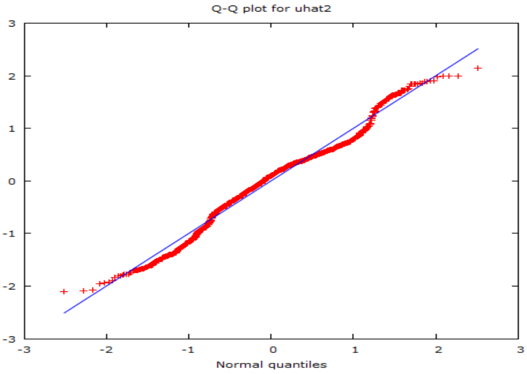


FIGURE 4.4: QQ Plot for WTI

4.2.2 VAR Analysis

From the OLS regression section we had proved that the assumptions of linearity, homoscedasity and normality of errors were demonstrated to hold true for the data series. The Durbin-Watson value indicated that there was autocorrelation in the data, we can see in tables 4.2 - 4.8 that the VAR model has compensated for the regressors not being independent, as the Durbin-Watson value is now very close to 2 in all of the tables indicating no autocorrelation. There is a large collection of results described in this section, so I break down the analysis by commodity and then complete a summary of the overall results in the final subsection. On each commodity we conduct VAR analysis on the z-value of the series and also the z-return.

	coefficient	std. error	t-ratio	p-value
sentiment_1	0.192301	0.0315414	6.097	1.40e-09
sentiment_2	0.148124	0.0314993	4.702	2.83e-06
sentiment_3	0.116018	0.0319198	3.635	0.0003
sentiment_4	0.137664	0.0309880	4.442	9.60e-06
sentiment_5	0.100547	0.0312141	3.221	0.0013

R-squared	0.261775	Adjusted R-squared	0.258596
F(6, 1393)	65.20124	P-value(F)	1.60e-71
ρ	-0.006973	Durbin-Watson	2.011123

Table 4.2: WTI VAR Value, lag order 5

	coefficient	std. error	t-ratio	p-value
sentiment_1	0.00617084	0.0348062	0.1773	0.85930
sentiment_2	0.0154037	0.0350302	0.4397	0.66020
sentiment_3	-0.0416651	0.0365122	-1.1411	0.25401
sentiment_4	-0.082267	0.0382882	-2.1486	0.03183
sentiment_5	0.0116087	0.033851	0.3429	0.73170

R-squared	0.026033	Adjusted R-squared	0.019150
F(10, 1415)	2.086338	P-value(F)	0.022759
ρ	0.000668	Durbin-Watson	1.997834

Table 4.3: WTI VAR Return, lag order 5

WTI Crude Oil

The results of the WTI analysis can be seen in table 4.2 and 4.3, from these tables we are able to observe some interesting results about sentiment analysis. The high T-Ratio and low P-Values reject the null hypothesis and show that sentiment has a large influence on the value of the commodity for predicting the values. In the return series however there some lags which have a low T-Ratio and a high P-Value indicating that some lags agrees with the null hypothesis and have no correlation with the returns. This allows us to observe the statistically significant lags in the series which would be sentiment_4 for the returns.

Both of the Durbin-Watson values are close to 2 which shows that autocorrelation has been removed through the use of VAR analysis. The R^2 values for the value and returns are quite different, we can see that there is much more variance in the return series than in the value series. This is what we would expect as a return series is much harder to predict than a value series as it can have much larger changes in value from one time period to the next. We can also see from the P-Values and T-Ratios that sentiment has a different effectiveness at predicting the values and returns at different time periods.

	coefficient	std. error	t-ratio	p-value
sentiment_1	0.0787289	0.0340266	2.3137	0.02084
sentiment_2	0.139059	0.0337725	4.1175	0.00004
sentiment_3	0.14348	0.0345078	4.1579	0.00003
sentiment_4	0.106676	0.0338063	3.1555	0.00164
sentiment_5	0.0972273	0.0355247	2.7369	0.00629

R-squared	0.138981	Adjusted R-squared	0.134854
F(6, 1252)	22.25716	P-value(F)	5.26e-25
ρ	-0.006973	Durbin-Watson	2.017817

Table 4.4: Brent VAR Value, lag order 5

	coefficient	std. error	t-ratio	p-value
sentiment_1	0.0136908	0.0309282	0.4427	0.6581
sentiment_2	0.0025810	0.0105614	1.721	0.0856
sentiment_3	-0.0170048	0.0291357	-0.5836	0.5596
sentiment_4	-0.0385384	0.0323876	-1.190	0.2343
sentiment_5	-0.0513446	0.0303413	-1.692	0.0908

R-squared	0.027721	Adjusted R-squared	0.019961
F(10, 1253)	1.730399	P-value(F)	0.069212
ρ	0.001337	Durbin-Watson	1.997286

Table 4.5: Brent VAR Return, lag order 5

Brent Crude Oil

The results of the Brent analysis can be seen in table 4.4 and 4.5, from these tables we are able to observe some interesting results about sentiment analysis and are also able to compare it to the other crude oil commodity. The high T-Ratio and low P-Values reject the null hypothesis and show that sentiment has an influence on the value of the commodity for predicting the values. In the return series however there some lags which have a low T-Ratio and a high P-Value indicating that some lags agrees with the null hypothesis and have no correlation with the returns. This allows us to observe the statistically significant lags in the series which would be sentiment_2 for the returns.

Both of the Durbin-Watson values are close to 2 which shows that autocorrelation has been removed through the use of VAR analysis. Compared to the WTI value series we can see that there is much more variance in this time series which suggests that sentiment is not as good at predicting changes in Brent Crude Oil. It can also be observed that the P-Values and T-Ratios that sentiment has a different effectiveness at predicting the values and returns at different lag periods, which are different to the WTI lag periods.

	coefficient	std. error	t-ratio	p-value
sentiment_1	0.0453458	0.0248936	1.822	0.0687
sentiment_2	0.0821370	0.0257432	3.191	0.0014
sentiment_3	0.0791094	0.0260085	3.042	0.0024
sentiment_4	0.0874566	0.0250444	3.492	0.0005
sentiment_5	0.0786908	0.0260232	3.024	0.0025

R-squared	0.049008	Adjusted R-squared	0.045545
F(6, 1648)	14.98022	P-value(F)	9.44e-17
ρ	-0.003002	Durbin-Watson	2.005974

Table 4.6: New York Gasoline Value, lag order 5

	coefficient	std. error	t-ratio	p-value
sentiment_1	0.0155644	0.0250525	0.6213	0.5345
sentiment_2	0.00979851	0.0238126	0.4115	0.6808
sentiment_3	0.0173504	0.0250739	0.6920	0.4891
sentiment_4	0.0696467	0.0264806	2.630	0.0086
sentiment_5	-0.00902715	0.0228115	-0.3957	0.6924

R-squared	0.018483	Adjusted R-squared	0.012513
F(10, 1644)	1.502144	P-value(F)	0.132443
ρ	0.000542	Durbin-Watson	1.998908

Table 4.7: New York Gasoline Return, lag order 5

New York Gasoline

The results of the Gasoline analysis can be seen in table 4.6 and 4.7 we can see that again the null hypothesis has been rejected due to the high T-Ratio and low P-Values for each of the lags in the value series. For the return series the only lag that rejects the null hypothesis is sentiment_4 from the value series we can also see that the most significant lag in the series was sentiment_4 which suggests that there is a delay in the sentiment having an effect on gasoline prices.

We can also see that the R^2 values are quite low for both the value and return series which suggests a large amount of variance in the series. In this case I think that the corpus of text that was collected could have had less formal media sources included due to the terms ‘New York Gasoline’ which could pick up a number of articles which would not be related to the commodity which would effect the sentiment analysis.

	coefficient	std. error	t-ratio	p-value
sentiment_1	0.0122329	0.0303897	0.4025	0.6874
sentiment_2	0.0272862	0.0318850	0.8558	0.3923
sentiment_3	0.0443003	0.0312408	1.418	0.1565
sentiment_4	0.108424	0.0320700	3.381	0.0007
sentiment_5	0.0155830	0.0311315	0.5006	0.6168

R-squared	0.028434	Adjusted R-squared	0.022736
F(6, 1023)	6.292048	P-value(F)	1.65e-06
ρ	-0.001799	Durbin-Watson	2.001682

Table 4.8: Natural Gas Value, lag order 5

	coefficient	std. error	t-ratio	p-value
sentiment_1	-0.0108807	0.0295756	-0.3679	0.7130
sentiment_2	0.0378761	0.0339024	1.117	0.2642
sentiment_3	0.0775013	0.0290973	2.664	0.0079
sentiment_4	0.0279358	0.0234942	1.189	0.2347
sentiment_5	0.00230065	0.0264224	0.08707	0.9306

R-squared	0.061115	Adjusted R-squared	0.051901
F(10, 1019)	1.088915	P-value(F)	0.367433
rho	0.000425	Durbin-Watson	1.999108

Table 4.9: Natural Gas Return, lag order 5

Natural Gas

The results of the Gasoline analysis can be seen in table 4.8 and 4.9, we can see from these table that the natural gas analysis proved to be the least reliability and the T-Ratio is much lower and the P-Value is much higher than the other results. We can see that many of the sentiment lags agree with the NULL hypothesis which shows that sentiment analysis was not providing any information about the series.

There are some lags which provide an insight about gas prices, for instance the sentiment_4 gives a good indication of the value of the series while the sentiment_3 gives a good indication of the return prices. The R^2 and adjusted R^2 values are very low and indicates that there is a lot of variability in the analysis and that sentiment analysis has not been effective at predicting changes in the market. The overall F-test and P-values also indicate that sentiment analysis has not been effective for this commodity and that the results are too unreliable.

Sentiment Analysis Discussion

Taking a broad view of the overall results of the sentiment analysis system as a whole it is clear from the results observed in the previous section i.e the high T-Ratio and the low P-Values that sentiment analysis has been successful rejected the null hypothesis and that the analysis is statistically significant. We can also look at other statistical figures such as the F-test and P-values in the table which also confirm the statistical significance of sentiment analysis.

The analysis of multiple commodities has provided interesting results on the varying impact of sentiment analysis on different commodities. The results shown above show how sentiment can lag commodities at different rates and provide better or worse estimates. The results from this section shows that sentiment analysis has a statistically significant impact on the return price of commodities, but can effect individual commodities to different extents.

From the analysis into the return and value of the commodity series we can see that is it much easier to predict values than returns. This is because returns tend to change much more quickly than the overall value of the commodity as can be seen in figure 4.1 and 4.2. Sentiment analysis is still and area being actively researched and there is still improvements needed in the tools for caring out automated sentiment analysis. From the results produced in this report we can conclude that the results of sentiment analysis are statistically significant but can also have a large variance.

4.3 Weather Analysis

In this section I describe the results that were observed through the conduction of the weather analysis on the historical data that is described in section 3.3.1. Having been able to aggregate the cumulative heating degree days to a national level it made it much easier to create the weather model. I use z-score to allow direct comparison between the model value and the values of the different commodity series. The first step in the model is to calculate the yearly seasonality of the temperature, figure 4.5 shows this first estimation of the model which gives a relatively good prediction of the Henry Hub Value z-score using the yearly seasonality over 15 years.

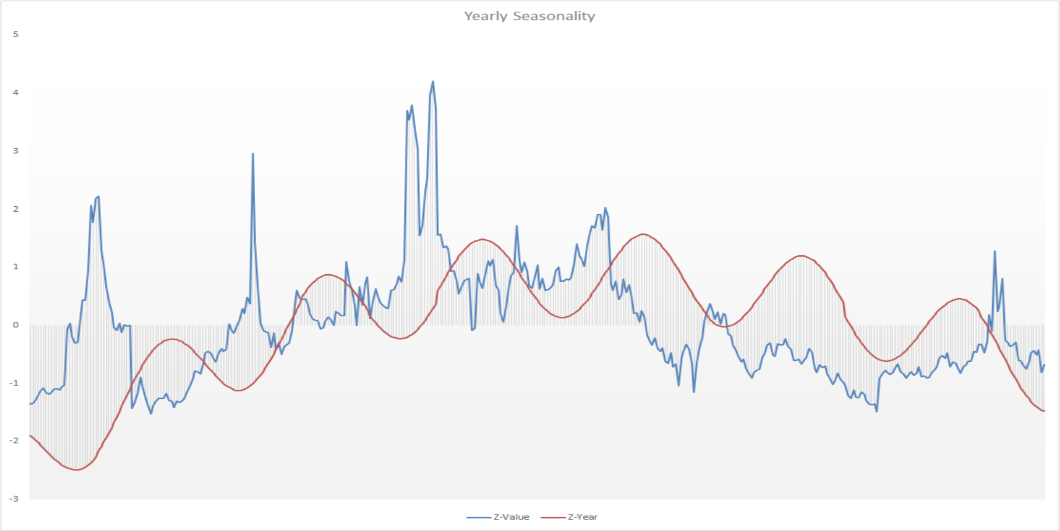


FIGURE 4.5: Henry Hub Natural Gas and yearly seasonality over 15 years

Commodity	R^2	*Adjusted* R^2	F-test	P-Value	Durbin-Watson
Brent	0.006406	-0.006737	0.098184	0.996536	0.011936
WTI	0.002580	-0.010613	0.050983	0.999461	0.022559
Natural Gas	0.492625	0.486137	12.60797	5.17e-13	0.227261
Gasoline	0.004474	-0.008695	0.128627	0.992726	0.019170

TABLE 4.10: OLS regression, lag order 5

After this encouraging result I completed the full model and investigated the impact of the model against the z-score of the values of the commodities. Figure 4.6 shows the weather model graphed against the commodities, it can be seen that the weather model closely simulates the natural gas model with the exception of a few outliers. After completing this initial comparison between the commodities I then used OLS regression to get a mathematical output of the similarities of the two series which can be seen in table 4.10. The low Durbin-Watson values indicated that autocorrelation was present in the series, so I used VAR analysis to remove this autocorrelation.

The results of the VAR analysis on the time series can be seen in table 4.11. This reinforces the OLS regression analysis that suggested that the weather would have the largest impact on the natural gas time series. It suggests however that there is a greater relationship between the weather and the other energy series than OLS regression had calculated. I think that this is more reflective than the OLS analysis and shows that the weather has an impact on a number of energy commodity series but impacts the natural gas series the most.

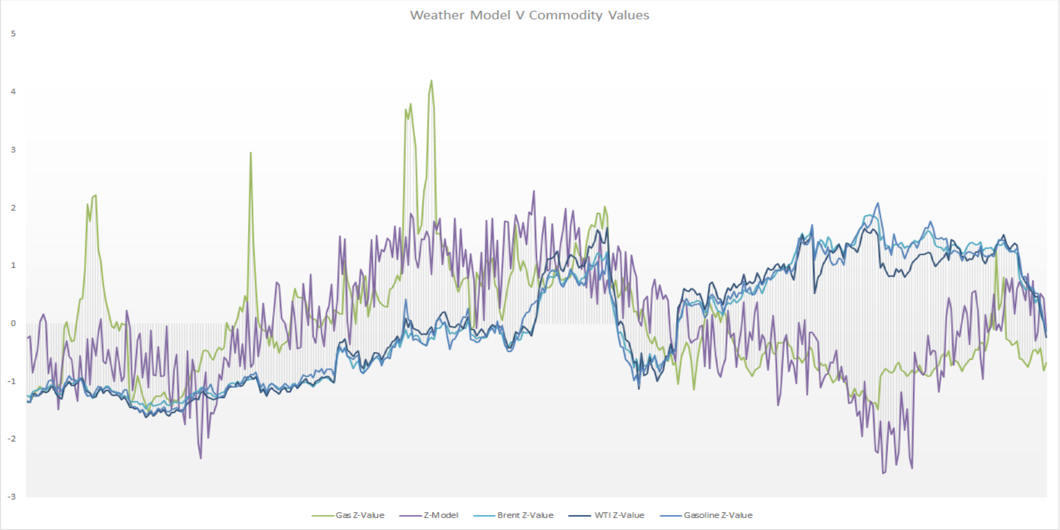


FIGURE 4.6: Weather analysis plotted against all the commodities

Commodity	R^2	*Adjusted* R^2	F-test	P-Value	Durbin-Watson
Brent	0.777739	0.772391	130.8708	1.4e-115	2.030686
WTI	0.787741	0.782396	133.5208	1.2e-116	2.030764
Natural Gas	0.903567	0.902292	590.3066	1.5e-188	2.027096
Gasoline	0.758741	0.752391	130.8708	1.1e-115	2.030738

TABLE 4.11: VAR System Value, lag order 5

Commodity	R^2	*Adjusted* R^2	F-test	P-Value	Durbin-Watson
Brent	0.061928	0.05974	1.263921	0.029180	1.997440
WTI	0.055071	0.051335	1.553688	0.051171	1.991982
Natural Gas	0.080912	0.071993	2.384634	0.024369	2.005997
Gasoline	0.042415	0.016742	0.715888	0.75373	1.991623

TABLE 4.12: VAR System Returns, lag order 5

Weather Analysis Discussion

The weather analysis that has been conducted in this project has produced some interesting results. From table 4.10 we can see how accurate the weather model was at predicting the value of different commodities. The Durbin-Watson values however are very low (almost 0), which indicates positive autocorrelation. While autocorrelation does not bias the OLS coefficient estimates, the standard errors tend to be underestimated and the t-scores overestimated when the autocorrelations of the errors at low lags are positive. We can see from the results of this table that the F-Test and P-Value for the analysis were quite poor except for Natural Gas, which is what would be expected after looking at the results in figure 4.6.

The VAR analysis which can be seen in table 4.11 shows the analysis for the commodity values, while the analysis in table 4.12 shows the analysis for the return values in the series. We can see from the Durbin-Watson values which are close to 2 that autocorrelation has successfully been removed which means that the analysis will provide more reliable results. The figures in Table 4.11 show how accurate the model was at predicting the values in the commodity series. We can see by the small P-Values and large F-Test results that all commodities have rejected the null hypothesis. We can also see by the high R^2 values that the weather analysis has been good at estimating the commodity values.

When we compare the results in table 4.11 and 4.12 we can see that is a large difference in predicting the daily returns of the series and the commodity values. This is due to the variability in return price which changes daily and can be many standard deviations away from the mean. The weather model does not perform well when trying to predict the next day returns and the F-Test values are quite low and the P-Values are quite high, with some of them agreeing with the null hypothesis, the natural gas series however did produce some statistically significant results. This shows that the model is good for estimating longer term changes in price rather than the next day return.

4.4 Algorithmic Trading

In this section I present the results of the algorithmic trading system design that was outlined in the previous chapter in section 3.4 using the results of the weather and sentiment analysis. This section is designed to show the practical applications of this research and how it can be used in live systems. There are a number of different parameters which are used when trading algorithmically, such as the look back window and holding period that can be used. These parameters can affect the results of the analysis dramatically, so I test a number of different holding and look back periods in my trading system.

Short Term

The short term trend in this case is modelled with a holding period of four days as this is the period that was found to have the greatest correlation with value from the VAR

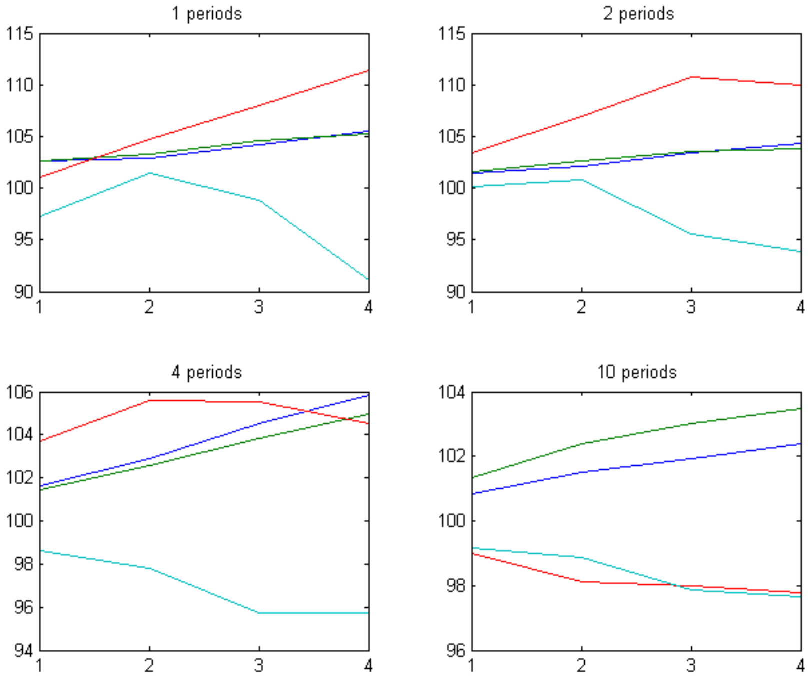


FIGURE 4.7: Average return for multiple look-back periods 1,2,4 and 10 of weather analysis with a holding period of 4

analysis conducted in the previous sections. I use four different look-back periods to analyse the impact of this factor on the algorithm so we get periods which look 1,2,4 and 10 days into the past.

From the short term analysis in Figures 4.7 - 4.9 we can observe some interesting results. We can see that over the numerous look-back periods that weather has generally outperformed the sentiment and opening trend analysis. I would have expected the weather analysis to outperform the sentiment analysis but to have them both outperform the traditional trend strategy is quite surprising. This shows that they analysis techniques that have been developed during the project have real world applications in the commodities market.

Long Term

The long term trend in this case is modelled with a holding period of 100 days, I use four different look-back periods to analyse the impact of this factor on the algorithm so we get periods which look at 10,20,40 and 100 days into the past. I increase the look-back

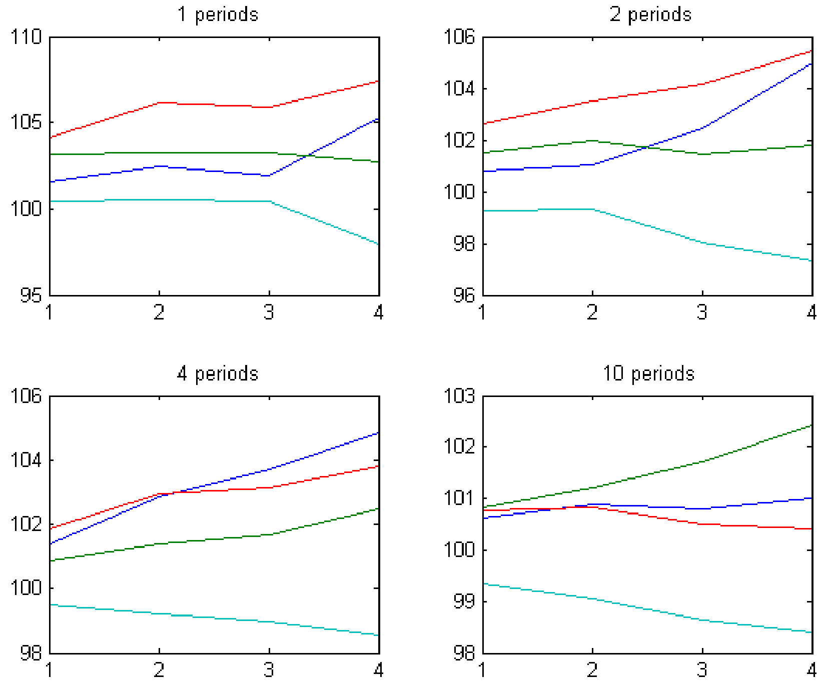


FIGURE 4.8: Average return for multiple look-back periods 1,2,4 and 10 of sentiment analysis with a holding period of 4

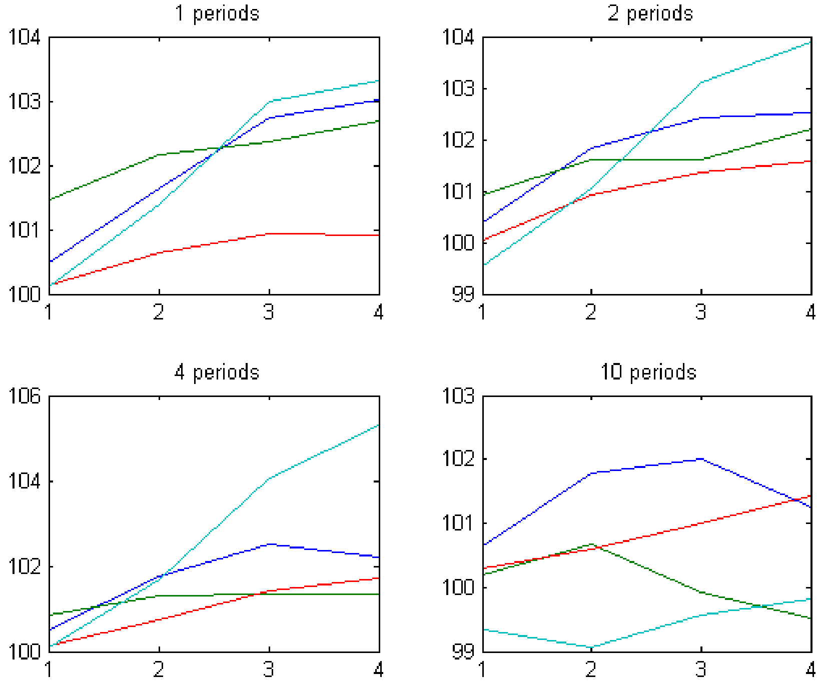


FIGURE 4.9: Average return for multiple look-back periods 1,2,4 and 10 of close price analysis with a holding period of 4

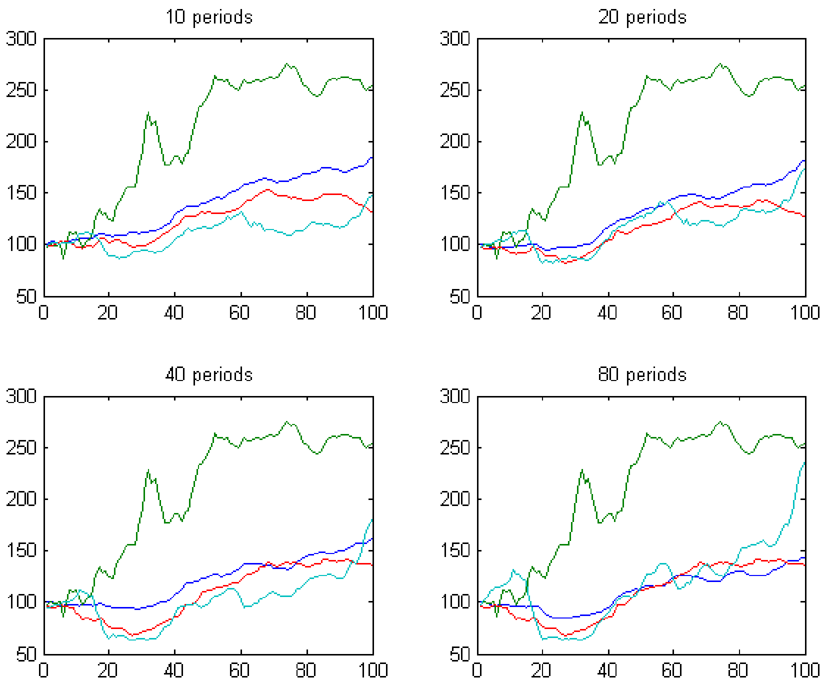


FIGURE 4.10: Average return for large look-back periods 10,20,40 and 100 of weather analysis with holding period of 100

period for each of the commodities in this longer term analysis as this gives a better prediction of outliers in the series which leads to better results.

From the long term analysis in Figures 4.10 - 4.12 we can observe some interesting results. We can see that as was the case with the short term analysis the weather analysis has proven the most reliable to build a long term trading strategy on. We can see that the weather analysis has given excellent results for different commodities but in-particular WTI crude oil. In this longer term strategy sentiment analysis has been found to be less reliable and preforms the worst out of the three systems.

4.5 Conclusions

This chapter has shown the results that were observed over the course of the project both in sentiment and weather analysis. This section provides a summary of the results that have been outlined in this chapter, through the different sections. We can compare the two sets of results that are given by VAR analysis on the two different analysis

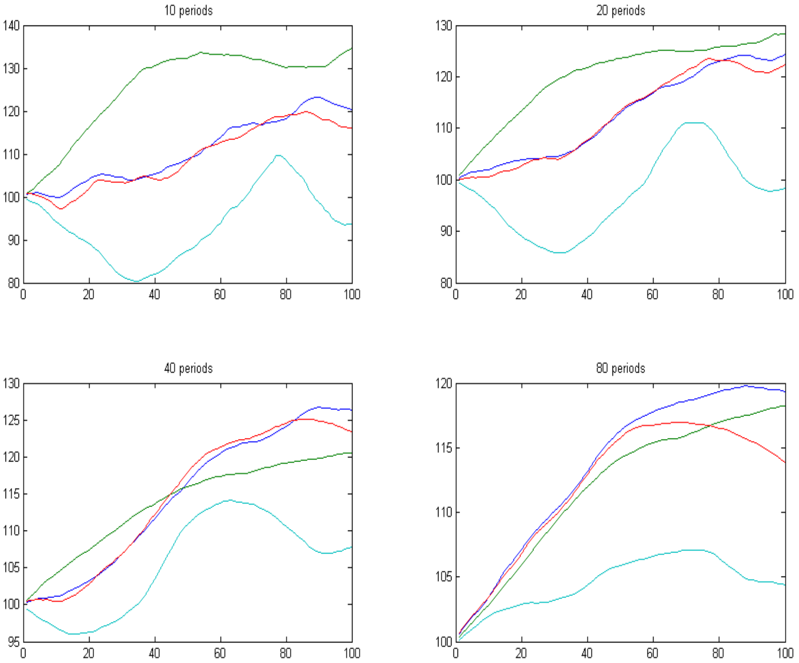


FIGURE 4.11: Average return for multiple look-back periods 10,20,40 and 100 of sentiment analysis with a holding period of 100

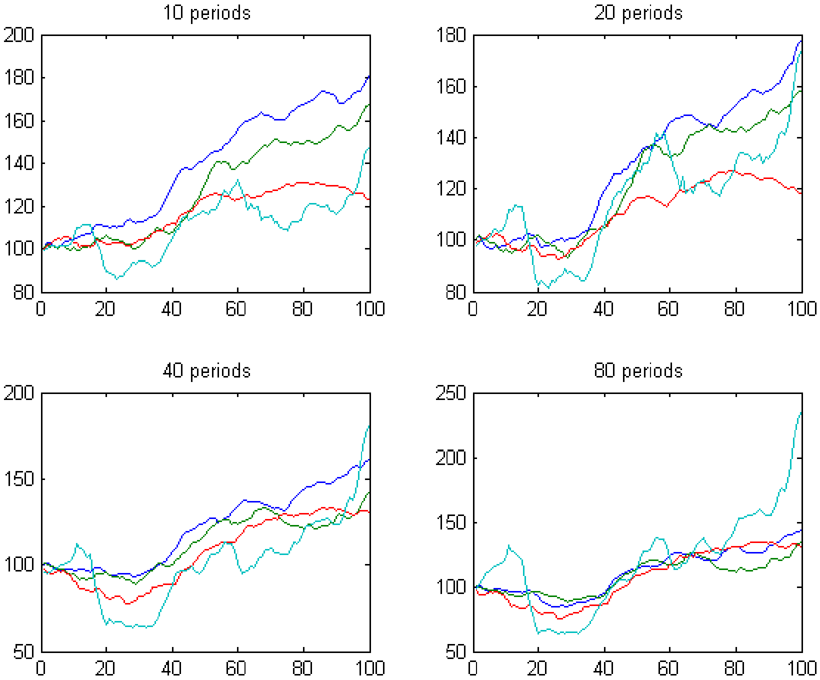


FIGURE 4.12: Average return for multiple look-back periods 10,20,40 and 100 of close price analysis with holding period of 100

methods in tables 4.2 to 4.9 for the sentiment analysis and table 4.11 and 4.12 for the weather analysis.

We can see that the weather analysis has proved to be more reliable than the sentiment analysis and that the F-Test and P-Values that are obtained for the weather analysis indicate a stronger rejection of the null hypothesis for both the value and return series. It has also been observed how the results of the different analysis techniques can vary with each commodity, for example the weather analysis had quite poor results for the gasoline series compared to the other commodities, which suggests that individual analysis still needs to be completed on the commodities that are used.

In the final section I apply the results that were obtained in the weather and sentiment analysis in an algorithmic trading program which is back-tested on the historical commodity data. The trading system showed that both weather and sentiment analysis could be used as part of either a long or short term trading strategy, with the weather analysis having slightly better results.

Chapter 5

Conclusions

5.1 Final Result

There are a number of results that have been observed through the conduction of this project as the scope of the project was quite broad. The conduction of both weather and sentiment analysis has led to some interesting comparisons between the two types of analysis. From the results observed in Chapter 4 it can be seen that weather analysis has proven more effective than sentiment analysis at predicting returns in energy commodities.

This can be due to a range of factors, one of the reasons I would suggest is that sentiment analysis systems have a much more difficult task due to the complexity and ambiguity of human language. There has been a large amount of tools developed for sentiment analysis however some of the most recent research has found that the latest sentiment tools only have an average accuracy of about 60% [61]. This obviously limits the analysis that can be conducted using sentiment, when the automatic sentiment analysis tools are giving such poor results.

The results for weather analysis are quantitative which makes it much easier to detect outliers and offsets especially as we know what temperature to expect in different regions within a certain limit. For example if we were receiving results of 20°C for New York in the middle of winter we could assume that the station was giving back erroneous results. For sentiment analysis however it is much more difficult to check the results

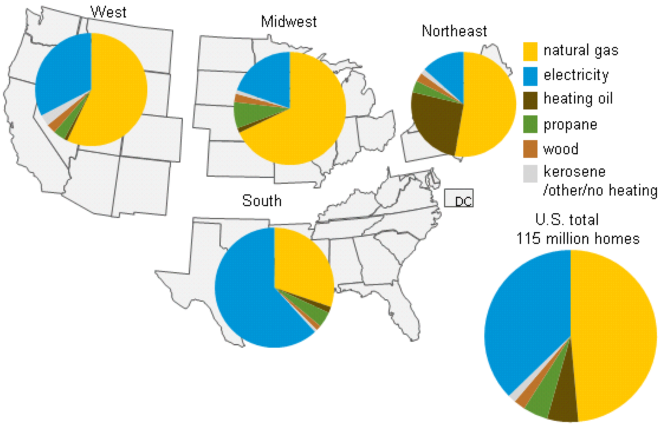


FIGURE 5.1: Number of homes by primary space heating fuel and region

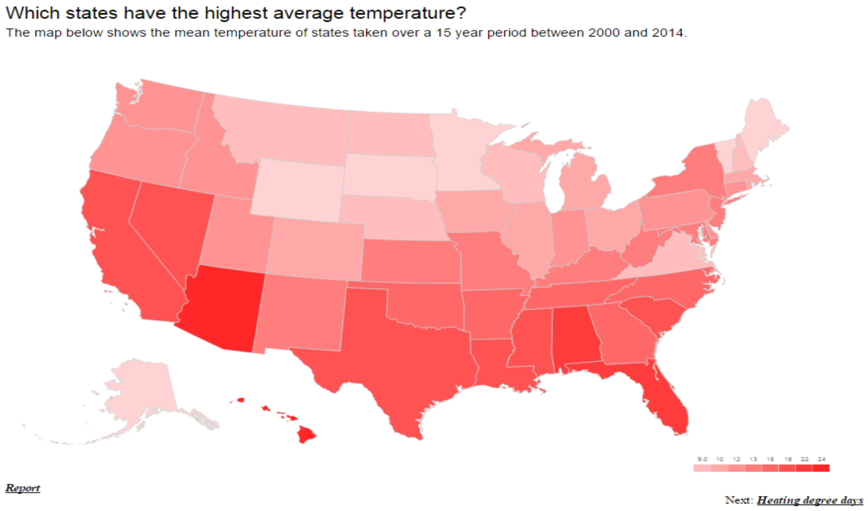
of the sentiment analysis system to make sure that they are within a certain degree of error.

The development of the trading program and the results that were observed in Chapter 4 has shown the real world application of the analysis techniques that have been developed in this project. Both of the systems had positive average returns and preformed better than conventional analysis techniques. The methodology that has been developed in this project means that it is also possible to easily include further exogenous variables such as shipping analysis for further improve the accuracy and reliability of the system.

After finding out the results of this analysis I looked for reasons as to why the weather model would be so good at predicting the natural gas spot prices. One of the reasons that is most evident can be seen in figure 5.1, which shows that natural gas is the most popular heating fuel used by almost half of all U.S households. Electricity which accounts for 38% of households also uses 27% Natural Gas to make the electricity so it becomes clear that there is a large number of reason why the gas spot price would be linked to weather analysis.

5.2 Visualisation Options

Having completed my final year project on ‘Particle Systems for Weather Simulation’ in computer graphics I felt that I was in a good position to explore some visualisation options which would make the results of the project easier to understand. One of the best resources that I found to develop these visualisation options was the *D3* JavaScript



Which states have the highest average temperature?

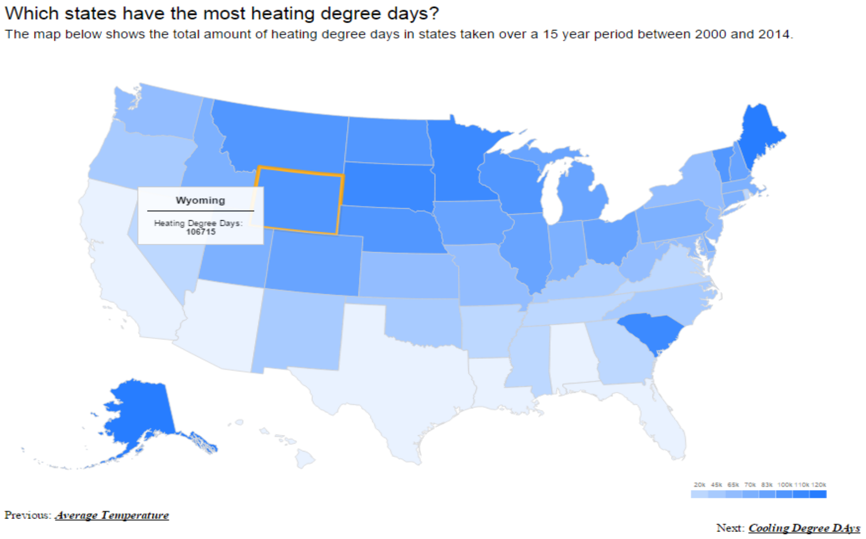
The map below shows the mean temperature of states taken over a 15 year period between 2000 and 2014.

FIGURE 5.2: Display of the mean temperature of different states

visualization library which is open source. The web is the best place to publish the visualizations and D3 gives the full capabilities of modern browsers and a data driven approach to DOM manipulation.

Using D3 I was able to load the weather information that was collected over a 15 year period and create a series of interactive choropleth images. These images were interactive and if you scrolled over them it would provide further information about the weather series. As I was able to collect a wide variety of information about each of the different states I could create a number of images which presented the heating degree days, cooling degree days and other useful information. Figure 5.2 and 5.3 show two of the images that were produced using the D3 library. Figure 5.3 shows the tooltip and the further information that appears if you hover over a state such as Wyoming in this case.

A choropleth image is intuitive to understand and also provides useful information to the researcher about outliers that are located in the dataset. The visualization is also good for public science outreach and people can understand the latest projects that researchers are working on. In the test site included on the CD I also include a link to this report on the index page, which makes it easy for people to learn more about how the research that was carried out.



Which states have the most heating degree days?

The map below shows the total amount of heating degree days in states taken over a 15 year period between 2000 and 2014.

Wyoming
Heating Degree Days: 106715
20k 45k 65k 70k 83k 100k 110k 120k
Previous: *Average Temperature*
Next: *Cooling Degree DAys*

FIGURE 5.3: Display of the total heating degree days of different states

5.3 Future Work

There is a large number of possibilities for future work in this area as there is a lot of information which is to be released to the public which could have a large impact on commodity prices. The Open Government Partnership which launched in 2011 has had its membership grow from 8 to 65 countries in these few years, we can see in figure 5.4 the countries that are part of the OGP and the current action plans that they are implementing. We can see from figure 5.5 that in this short period there has been a large number of datasets made available to the public. From the US open data site we can see that over 124,057 datasets have been made available to the public and in Europe 8128 datasets have been made publicly available.

There are a number of possibilities for the development of this project due to the large number of data sets that are being released, for example the ocean data set that has been released by the US open data site could be used to predict periods and areas where the ocean is dangerous which may lead to lost cargo or ships. This could then be factored into the shipping analysis which would be used as an exogenous variable in the commodities market analysis. The methodology that has been developed during this project is a resource that can be used in further projects in this area for example the analysis techniques that were used such as vector autoregression and the trading program that was developed.

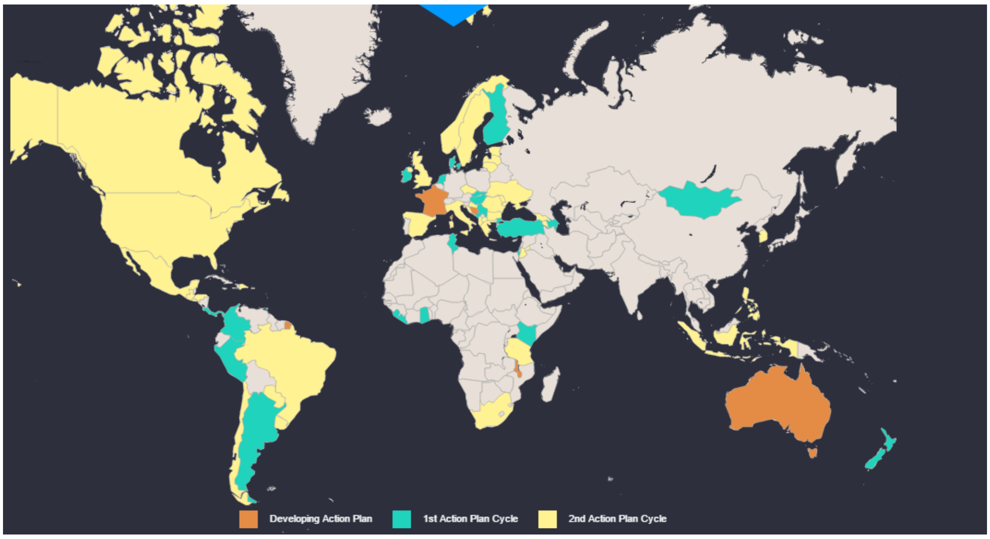


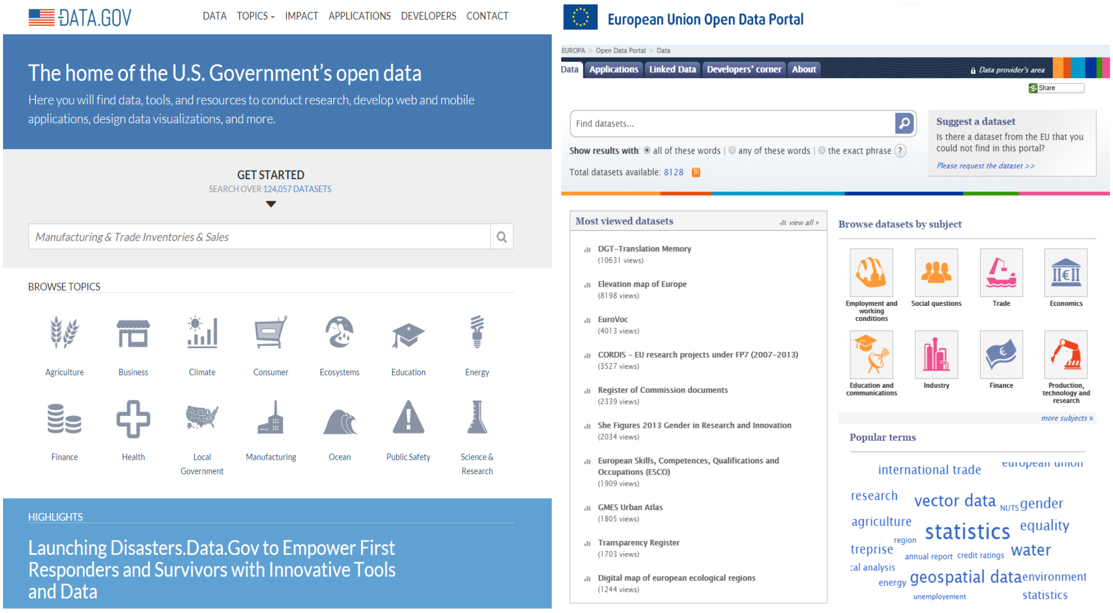
FIGURE 5.4: World map of countries action plans in the OGP

The results of this research project have also focused on energy commodities, there is obviously a large number of other areas which the research could expand into. As I have already conducted the weather analysis in America for the last 15 years, further research could be conducted into how effective weather analysis could be at predicting changes in other commodities which would be effected by the weather. There a number of other commodities which traders could use weather analysis to make better decisions, which would include grains and cereals such as barley, corn and oats.

The visualisation of results has a large potential for future growth due to some of the recent technologies that have been developed such as D3 and WebGL. Figure 5.3 shows a simple intuitive display of the amount of heating degree days which would be much more difficult to interpret from a table. With HTML5 allowed access to the GPU through WebGL there are a lot of options for advanced visual displays in browsers which would not have been possible a few years ago. Some academic experiments have been conducted in this area but there is still a large opportunity for developing a library which could be used to assist in the publishing of academic results which would improve accessibility.

5.4 Learning Outcomes

Working on a project this large has allows me to expand my knowledge in a number of different areas. I have learned a great deal about weather and sentiment analysis, about how to design and build these analysis systems and the limitations that they have. I have



The home of the U.S. Government's open data

Here you will find data, tools, and resources to conduct research, develop web and mobile applications, design data visualizations, and more.

HIGHLIGHTS

Launching Disasters.Data.Gov to Empower First Responders and Survivors with Innovative Tools and Data

European Union Open Data Portal

Show results with: ● all of these words | ○ any of these words | ○ the exact phrase

Total datasets available: 8128

Suggest a dataset

Is there a dataset from the EU that you could not find in this portal?

Please request the dataset >>

Most viewed datasets

- DGT-Translation Memory (10631 views)
- Elevation map of Europe (8198 views)
- EuroVoc (4013 views)
- CORDIS – EU research projects under FP7 (2007–2013) (3527 views)
- Register of Commission documents (2339 views)
- She Figures 2013 Gender in Research and Innovation (2034 views)
- European Skills, Competences, Qualifications and Occupations (ESCO) (1909 views)
- GMES Urban Atlas (1805 views)
- Transparency Register (1703 views)
- Digital map of european ecological regions (1244 views)

FIGURE 5.5: US and European Open Data Site

also learned a large amount about economic and statistical techniques such as vector autoregression which I never had any experience with before but will prove invaluable in any other research projects that I carry out in this area.

I have been able to design and implement a number of programs to aggregate and analyse information from different parts of the project which involved working with a number of different technologies. I was able to interact with a number of API's using Ruby scripts to aggregate a large collection of historical information which allowed the analysis techniques to be tested. I implement a number of excel macro which allowed the historical data to be cleaned and collected, I also implemented a Matlab script which allowed the analysis techniques to be tested as part of a trading system.

I have also been able to use my web programming skills and the D3 JavaScript visualisation library to produce visual interactive results from the analysis that was conducted. As can be seen by the list of programs and languages used, the project has improved my knowledge in a wide variety of areas of computer science from heavy processing in Matlab to web programming with Ruby and JavaScript.

I have also learned a lot about individual research and the many challenges that come with undertaking such a large project with many different resources. It emphasises the importance of good software engineering techniques such as code versioning. I used git to keep track of the various different programs that I was working on which allowed me to always know where I had left off when I moved to a different program. I learned the importance of having useful commit messages and to use dev branches when designing new features to ensure that the master branch always has fully working code.

Working on a project this large has also allowed me to develop some of my softer skills as well such as being able to make an effective timetable which is key for a project this long and conducting regular meetings with my project supervisor. I was able to develop a Gantt chart early in the project which provided a clear structure to the research that I was undertaking and gave me enough time to plan, research, design, implement and report about the project.

Overall I think that the project has been a great success and I feel that I have learned a great deal about a number of different areas in computer science and statistics. I have greatly improved my research skills and now feel confident to propose, design, investigate, implement and test a research topic as I have been able to do in this project.

Appendix A

Inference Statistics

A.1 Moments of Data

A moment is a specific quantitative measure which is used in statistics to analyse a set of points. The four moments which are described in this section enables the succinct description of data series and allows us to communicate our findings to others in an unambiguous manner.

1^{st} Moment: Mean

The first moment of data is the mean which is the sum of a sample $x_1, x_2, \ldots, x_n$ denoted by $\bar{x}$ is the sum of the sampled values divided by the number of items in the sample:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} \tag{A.1}$$

2^{nd} Moment: Variance

The second moment of data is used to measure the variance of data point from the mean value. This indicates how spread out the points are and is important for standardizing series through using z scores. Its value is denoted by:

$$\sigma = \sqrt{\frac{1}{N} \sum (x_i - \bar{x})^2} \tag{A.2}$$

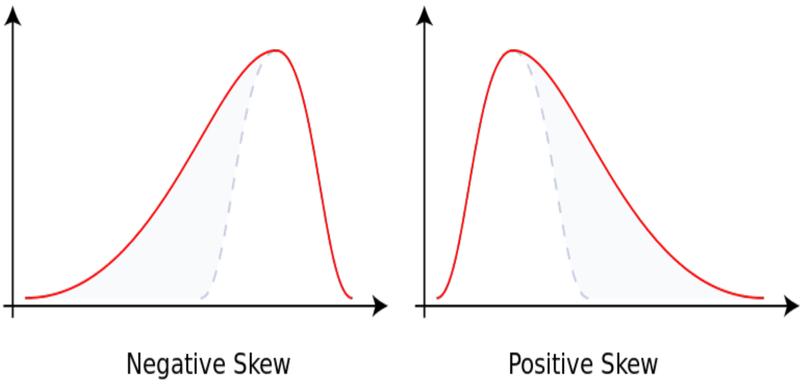


Figure A.1: Examples of a skewed data series

3^{rd} Moment: Skewness

The third moment is skewness which is a measure of the asymmetry of a distribution. Positive or negative skew indicates that the tail on one side of the distribution is longer or fatter than the tail on the other side and a skewness of 0 indicates a normal distribution. Examples of different levels of skewness can be seen in Figure A.1 and its value is denoted by:

$$\gamma_1 = E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \frac{\mu_3}{\sigma^3} \tag{A.3}$$

where μ_3 is the third central moment, μ is the mean, σ is the standard deviation.

4^{th} Moment: Kurtosis

The fourth moment is Kurtosis which is used to describe how peaked a distribution is. Larger kurtosis suggests the presence of extreme outliers while low kurtosis suggests data points are close to the mean. For a normal distribution the kurtosis is three, in equation A.4 I describe the excess kurtosis which is the kurtosis minus three. Therefore using this equation a normal distribution would have a kurtosis of zero. Any series with excess kurtosis of zero is known as mesokurtic while series with positive excess kurtosis are known as leptokurtic and negative excess are platykurtic which can be seen in Figure A.2.

$$\gamma_2 = \frac{\mu_4}{\sigma^4} - 3 \tag{A.4}$$

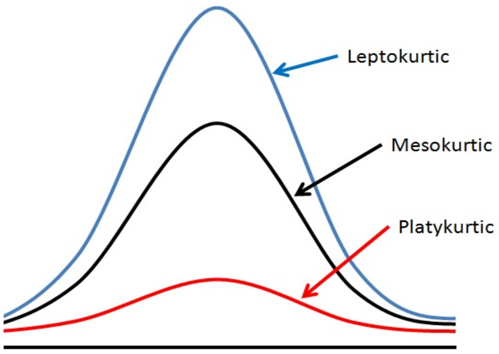


FIGURE A.2: General forms of kurtosis

where μ_4 is the fourth central moment and σ is the standard deviation.

Appendix B

Included Code

The full source code is available on the CD, the code is broken into a number of different folders which deal with the weather analysis, sentiment analysis, trend following program in Matlab and finally the visualisations that were produced by using the D3 library. I have made an effort to make the specific parts of the project as accessible as possible by including a detailed ReadMe markdown file in the top level directory. I have also tried to comment and layout the code and data in the most intuitive way possible.

Bibliography

[1] Eugene F Fama, Lawrence Fisher, Michael C Jensen, and Richard Roll. The adjustment of stock prices to new information. *International economic review*, 10(1):1–21, 1969. URL http://papers.ssrn.com/sol3/papers.cfm?abstract_id=321524.

[2] Oded Z Maimon and Lior Rokach. *Data mining and knowledge discovery handbook*, volume 1. Springer, 2005. URL http://www.springer.com/computer/database+management+%26+information+retrieval/book/978-0-387-09822-7.

[3] David Leinweber. Knowledge-based systems for financial applications. *IEEE Intelligent Systems*, 3(3):18–31, 1988. ISSN 0885-9000. URL http://www.it.iitb.ac.in/~palwencha/ES/J_Papers/ES3.pdf.

[4] Mark J Flannery and Aris A Protopapadakis. Macroeconomic factors do influence aggregate stock returns. *Review of Financial Studies*, 15(3):751–782, 2002. URL http://rfs.oxfordjournals.org/content/15/3/751.short.

[5] Christoph Böhringer. The synthesis of bottom-up and top-down in energy policy modeling. *Energy economics*, 20(3):233–248, 1998. URL http://www.sciencedirect.com/science/article/pii/S0140988397000157.

[6] Jeffrey A Frankel. Effects of speculation and interest rates in a “carry trade” model of commodity prices. *Journal of International Money and Finance*, 42: 88–112, 2014. URL http://www.sciencedirect.com/science/article/pii/S0261560613001083.

[7] Lukas Vacha and Jozef Barunik. Co-movement of energy commodities revisited: Evidence from wavelet coherence analysis. *Energy Economics*, 34(1): 241–247, 2012. URL http://www.sciencedirect.com/science/article/pii/S0140988311002350.

[8] Bahattin Buyuksahin, Thomas K Lee, James T Moser, and Michel A Robe. Physical markets, paper markets and the wti-brent spread. *Energy Journal*, 34(3), 2013. URL http://www.eia.gov/finance/markets/reports_presentations/2012paperbrentwti.pdf.

[9] John J Murphy. *Technical analysis of the financial markets: A comprehensive guide to trading methods and applications.* Penguin, 1999. URL https://212c2bce6facfd6fe7b214d47322d8d875e93dc5.googledrive.com/host/0B0rcwmcuCEWpQTY0djNCRjlqQVE/Murphy_I.pdf.

[10] Stephen J. Taylor. *Asset Price Dynamics, Volatility, and Prediction.* Princeton University Press, 2011. URL http://www.jstor.org/stable/j.ctt7t66m.

[11] Burton G Malkiel. The efficient market hypothesis and its critics. *Journal of economic perspectives*, pages 59–82, 2003. URL http://www.jstor.org/stable/3216840?seq=1#page_scan_tab_contents.

[12] Meredith Beechey, David WR Gruen, and James Vickery. *The efficient market hypothesis: a survey.* Reserve Bank of Australia, Economic Research Department, 2000. URL http://www-ho.rba.gov.au/publications/rdp/2000/pdf/rdp2000-01.pdf.

[13] William H Greene. *Econometric analysis.* Granite Hill Publishers, 2008. URL http://pages.stern.nyu.edu/~wgreene/Text/econometricanalysis.htm.

[14] Gangadharrao S Maddala and Kajal Lahiri. *Introduction to econometrics*, volume 2. Macmillan New York, 1992. URL http://eu.wiley.com/WileyCDA/WileyTitle/productCd-EHEP000871.html.

[15] Christopher A Sims. Macroeconomics and reality. *Econometrica: Journal of the Econometric Society*, pages 1–48, 1980. URL http://www.jstor.org/stable/1912017?seq=1#page_scan_tab_contents.

[16] Eric Zivot and Jiahui Wang. Vector autoregressive models for multivariate time series. *Modeling Financial Time Series with S-PLUS®*, pages 385–429, 2006. URL http://link.springer.com/chapter/10.1007/978-0-387-32348-0_11.

[17] Thomas Doan, Robert Litterman, and Christopher Sims. Forecasting and conditional projection using realistic prior distributions. *Econometric reviews*, 3(1): 1–100, 1984. URL http://www.nber.org/papers/w1202.

[18] Gangadharrao S Maddala. *Limited-dependent and qualitative variables in econometrics.* Number 3. Cambridge university press, 1986. URL https://books.google.ie/books?id=-Ji1ZaUg7gcC&dq=exogenous+variables+maddala&source=gbs_navlinks_s.

[19] Daniel Yergin. *The prize: The epic quest for oil, money & power.* Simon and Schuster, 2011. URL http://www.amazon.com/The-Prize-Quest-Money-Power/dp/1439110123.

[20] Per Einar Ellefsen. *Commodity market modeling and physical trading strategies.* PhD thesis, Massachusetts Institute of Technology, 2010. URL http://dspace.mit.edu/handle/1721.1/61602.

[21] Helyette Geman. *Commodities and commodity derivatives: modeling and pricing for agriculturals, metals and energy.* John Wiley & Sons, 2009. URL http://eu.wiley.com/WileyCDA/WileyTitle/productCd-0470012188.html.

[22] ND Geomelos and E Xideas. Forecasting spot prices in bulk shipping using multivariate and univariate models. *Cogent Economics & Finance*, 2(1), 2014. URL http://www.tandfonline.com/doi/full/10.1080/23322039.2014.932701#.VON3JfmsXOM.

[23] Charles-Louis De Secondat et al. *The Spirit of Laws.* Hayes Barton Press, 1748. URL http://oll.libertyfund.org/titles/837.

[24] Stephen Jewson and Anders Brix. *Weather derivative valuation: the meteorological, statistical, financial and mathematical foundations.* Cambridge University Press, 2005. URL http://ebooks.cambridge.org/ebook.jsf?bid=CBO9780511493348.

[25] Eckhard Platen and Jason West. A fair pricing approach to weather derivatives. *Asia-Pacific Financial Markets*, 11(1):23–53, 2004. ISSN 1387-2834. doi: 10.1007/s10690-005-4252-9. URL http://dx.doi.org/10.1007/s10690-005-4252-9.

[26] James Mahoney and Gary Goertz. A tale of two cultures: Contrasting quantitative and qualitative research. *Political Analysis*, 14(3):227–249, 2006. doi: 10.1093/pan/mpj017. URL `http://pan.oxfordjournals.org/content/14/3/227.abstract`.

[27] Charles C Ragin. *The comparative method: Moving beyond qualitative and quantitative strategies*. Univ of California Press, 2014. URL `http://www.ucpress.edu/book.php?isbn=9780520280038`.

[28] H Kent Baker and John R Nofsinger. *Behavioral finance: investors, corporations, and markets*, volume 6. John Wiley & Sons, 2010. URL `http://eu.wiley.com/WileyCDA/WileyTitle/productCd-0470499117.html`.

[29] Nicholas Barberis and Richard Thaler. Chapter 18 a survey of behavioral finance. In M. Harris G.M. Constantinides and R.M. Stulz, editors, *Financial Markets and Asset Pricing*, volume 1, Part B of *Handbook of the Economics of Finance*, pages 1053 – 1128. Elsevier, 2003. URL `http://dx.doi.org/10.1016/S1574-0102(03)01027-6`.

[30] Milton Friedman. *Essays in positive economics*, volume 231. University of Chicago Press, 1953. URL `http://www.econ.umn.edu/~schwe227/teaching.s13/files/articles-sug/friedman-1953.pdf`.

[31] Andrei Shleifer and Robert W Vishny. The limits of arbitrage. *The Journal of Finance*, 52(1):35–55, 1997. URL `http://onlinelibrary.wiley.com/doi/10.1111/j.1540-6261.1997.tb03807.x/full`.

[32] Malcolm Baker, Brendan Bradley, and Jeffrey Wurgler. Benchmarks as limits to arbitrage: Understanding the low-volatility anomaly. *Financial Analysts Journal*, 67(1):40–54, 2011. URL `http://www.cfapubs.org/doi/abs/10.2469/faj.v67.n1.4`.

[33] Matthew Rabin. Psychology and economics. *Journal of economic literature*, pages 11–46, 1998. URL `http://www.jstor.org/stable/2564950?seq=1#page_scan_tab_contents`.

[34] Daniel Kahneman and Amos Tversky. On the study of statistical intuitions. *Cognition*, 11(2):123–141, 1982. URL `http://www.sciencedirect.com/science/article/pii/0010027782900221`.

[35] Daniel Kahneman and Amos Tversky. *Choices, values, and frames.* Cambridge University Press, 2000. URL http://www.cambridge.org/ie/academic/subjects/psychology/cognition/choices-values-and-frames.

[36] Richard A Block and David R Harper. Overconfidence in estimation: Testing the anchoring-and-adjustment hypothesis. *Organizational Behavior and Human Decision Processes*, 49(2):188–207, 1991. URL http://www.sciencedirect.com/science/article/pii/074959789190048X.

[37] Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases. *science*, 185(4157):1124–1131, 1974. URL http://www.sciencemag.org/content/185/4157/1124.short.

[38] Michael W Berry and Malu Castellanos. Survey of text mining. *Computing Reviews*, 45(9):548, 2004. URL http://link.springer.com/book/10.1007/978-1-4757-4305-0.

[39] Andreas Hotho, Andreas Nürnberger, and Gerhard Paaß. A brief survey of text mining. In *Ldv Forum*, volume 20, pages 19–62, 2005. URL http://www.kde.cs.uni-kassel.de/hotho/pub/2005/hotho05TextMining.pdf.

[40] Zhiyong Dong, Qingyang Gu, and Xu Han. Ambiguity aversion and rational herd behaviour. *Applied Financial Economics*, 20(4):331–343, 2010. URL http://www.tandfonline.com/doi/abs/10.1080/09603100903299675#.VOSUQfmsXOM.

[41] Werner Antweiler and Murray Z. Frank. Is all that talk just noise? the information content of internet stock message boards. *The Journal of Finance*, 59(3):pp. 1259–1294, 2004. ISSN 00221082. URL http://www.jstor.org/stable/3694736.

[42] Paul C Tetlock. Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, 62(3):1139–1168, 2007. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1540-6261.2007.01232.x/full.

[43] James D Thomas and Katia Sycara. Gp and the predictive power of internet message traffic. In *Genetic Algorithms and Genetic Programming in Computational Finance*, pages 81–102. Springer, 2002. URL http://link.springer.com/chapter/10.1007/978-1-4615-0835-9_4.

[44] Fiacc Larkin and Conor Ryan. Good news: using news feeds with genetic programming to predict stock prices. In *Genetic Programming*, pages 49–60. Springer, 2008. URL `http://www.cse.ust.hk/~leichen/courses/comp630p/collection/reference-5-1.pdf`.

[45] Gabriel Pui Cheong Fung, Jeffrey Xu Yu, and Wai Lam. Stock prediction: Integrating text mining approach using real-time news. In *Computational Intelligence for Financial Engineering, 2003. Proceedings. 2003 IEEE International Conference on*, pages 395–402. IEEE, 2003. URL `http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1196287&tag=1`.

[46] Hirotugu Akaike. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6):716–723, 1974. URL `http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1100705&tag=1`.

[47] Robert F. Engle and Kenneth F. Kroner. Multivariate simultaneous generalized arch. *Econometric Theory*, 11:122–150, 2 1995. ISSN 1469-4360. URL `http://journals.cambridge.org/article_S0266466600009063`.

[48] Robert F Engle and Clive WJ Granger. Co-integration and error correction: representation, estimation, and testing. *Econometrica: journal of the Econometric Society*, pages 251–276, 1987. URL `http://www.jstor.org/stable/1913236?seq=1#page_scan_tab_contents`.

[49] Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. *A Comprehensive Grammar of the English Language.* Longman, London, 1985. URL `http://journals.cambridge.org/action/displayAbstract?fromPage=online&aid=2545148`.

[50] Khurshid Ahmad. Pragmatics of specialist terms: The acquisition and representation of terminology. In Petra Steffens, editor, *Machine Translation and the Lexicon*, volume 898 of *Lecture Notes in Computer Science*, pages 51–76. Springer Berlin Heidelberg, 1995. ISBN 978-3-540-59040-8. URL `http://dx.doi.org/10.1007/3-540-59040-4_20`.

[51] Khurshid Ahmad. Sentiment analysis & emotions and metaphors: A multidisciplinary perspective, 2008. URL `http://www.researchgate.net/`

publication/235960120_Sentiment_Analysis__Emotions_and_Metaphors_A_Multi-disciplinary_Perspective.

[52] Diego Garcia. Sentiment during recessions. *The Journal of Finance*, 68(3):1267–1300, 2013. URL http://onlinelibrary.wiley.com/doi/10.1111/jofi.12027/full.

[53] Casey Dougal, Joseph Engelberg, Diego Garcia, and Christopher A Parsons. Journalists and the stock market. *Review of Financial Studies*, page hhr133, 2012. URL http://rfs.oxfordjournals.org/content/early/2012/01/21/rfs.hhr133.full.pdf.

[54] Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to information retrieval*, volume 1. Cambridge university press Cambridge, 2008. URL http://www-nlp.stanford.edu/IR-book/.

[55] Allan H Murphy and Richard W Katz. *Probability, statistics, and decision making in the atmospheric sciences.* Westview Press, 1985. URL http://nldr.library.ucar.edu/repository/collections/NAB-000-000-000-116.

[56] Fred Espen Benth, JŪRATĖ ŠALTYTĖ BENTH, and Steen Koekebakker. Putting a price on temperature*. *Scandinavian Journal of Statistics*, 34(4):746–767, 2007. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1467-9469.2007.00564.x/full.

[57] Markus Burger, Bernhard Graeber, and Gero Schindlmayr. *Managing energy risk: An integrated view on power and other energy markets*, volume 426. John Wiley & Sons, 2008. URL http://eu.wiley.com/WileyCDA/WileyTitle/productCd-0470029625.html.

[58] Andrew Pole. *Statistical arbitrage: algorithmic trading insights and techniques*, volume 411. John Wiley & Sons, 2011. URL http://eu.wiley.com/WileyCDA/WileyTitle/productCd-0470138440.html.

[59] Andrei A. Kirilenko and Andrew W. Lo. Moore's law versus murphy's law: Algorithmic trading and its discontents. *The Journal of Economic Perspectives*, 27(2): pp. 51–72, 2013. ISSN 08953309. URL http://www.jstor.org/stable/23391690.

[60] Andrew C Szakmary, Qian Shen, and Subhash C Sharma. Trend-following trading strategies in commodity futures: A re-examination. *Journal of Banking & Finance*, 34(2):409–426, 2010. URL `http://www.sciencedirect.com/science/article/pii/S037842660900199X`.

[61] Mark Cieliebak, Oliver Dürr, and Fatih Uzdilli. Potential and limitations of commercial sentiment detection tools. In *ESSEM AI*, pages 47–58. Citeseer, 2013. URL `http://ceur-ws.org/Vol-1096/paper4.pdf`.