Financial Time Series Analysis and the Impact of Announcement

Stephen Kelly

April 6, 2011
Abstract

This paper studies a high frequency time series of oil futures trading contracts and the affect that publicly announced macroeconomic information in the form of an Oil Inventory announcement given by the Energy Information Association in the US has upon it.

Using the time series sampled at a five-minute frequency, it is the intention to detect any impacts to the price of the contracts after the announcement has been made and obtain an overall view of the affect on price around the time of the announcement.

It is from modelling this behaviour that a forecast is generated in an attempt to predict future values derived from the empirically built model of price distribution.
Acknowledgements

I would like to thank my family for their continued support and enthusiasm throughout the year and always.
My friends for their constant support and motivation.
And lastly my supervisor Professor Khurshid Ahmad for his support, dedication and continued patience throughout the project.
Contents

1 Introduction ......................................................... 4
  1.1 High frequency Financial Times Series and possible impact of an Innovation ........................................ 4
  1.2 Motivation and Literature Review ........................................ 4
  1.3 Dataset ................................................................... 6
  1.4 Event, Oil Inventory Announcement ........................................ 6
  1.5 Thesis Outline .......................................................... 8

2 Background, Time series analysis, Modelling and Forecasting 9
  2.1 High frequency time series ............................................ 9
  2.2 Stochastic Processes and Asset price .................................... 10
  2.3 Forecasting Models and the Monte Carlo Method ...................... 10
  2.4 Probability Models ...................................................... 11
    2.4.1 Statistical Analysis of the Distributions ............................ 14
    2.4.2 Forecasting future values ............................................ 19
  2.5 Summary ............................................................... 19

3 Analysis and Modelling .............................................. 21
  3.1 Modelling Noise ....................................................... 21
  3.2 Price adjustment, impact size and timing ............................... 22
  3.3 Distribution analysis ................................................... 24
  3.4 Process Summary ....................................................... 31

4 Implementation of the Analysis Program ............................ 32
  4.1 Program Design ....................................................... 32
  4.2 Java API’s ............................................................. 33
  4.3 Data Extraction ........................................................ 34
  4.4 Data Handling and Modelling ......................................... 36
  4.5 Simulation and Results ................................................ 36
  4.6 Concluding remarks .................................................. 37
Chapter 1

Introduction

1.1 High frequency Financial Times Series and possible impact of an Innovation

An asset’s price is of fundamental importance in financial economics. How to characterize the price of an asset is an ongoing topic of interest and under constant study. For the purpose of this study, the main focus is taken upon the fluctuations in crude oil prices. This consideration is distinct from the usual analysis, which is often conducted on bond, stock or foreign exchange markets[2]. The large scale dependence on the oil industry, its ubiquitous presence across many industries while making up a sizable portion of the energy sector highlight the significance of analyzing this particular commodity’s price and behavior.

1.2 Motivation and Literature Review

The conjecture made here is that announcements may have a discernible impact on the price of oil futures when examined in a high frequency time series. We build on the work of Almeida (1998) and of Anderson et al (2003) on exchange rates, and Balduzzi 2001 on bond prices. These studies have attempted to multiple and concurrent announcement impacts and news shocks. In my study, I have explored the impact of announcements on oil futures, another example of high frequency trading of assets or commodities. Some of the previous studies emphasised the analysis of exchange rate markets (Almeida 1998),(Anderson et al 2003), (Goodharte et al 1993) with relative announcements relating to each currency pair’s country of origin with various currency pairs being studied. Interest has also be shown in the bond market (Balduzzi 2001) where bond price and Bond trading in particular the volume traded, price volatility and fluctuations of bid-ask prices, and their response to public announcements of economic news was examined. A paper explored the future’s market but again with regard to foreign exchange (Ed-
Ederington and Lee (1993) looked into the relative impact on interest rates in this paper also and the price adjustment and volatility due to macroeconomic news announcements.

Similar in respect to Almeida (1998) and later investigations, this study is conducted using very high frequency data, earlier studies sampled at a frequency of a number of hours or more on exchange rate data. Performing the study on higher frequency data, in this case data was taken for a time period of five minutes. This allowed a more accurate method of demonstrating the reaction of the price to the event.

The approach taken here is the actual determination of price deviation as a result of the announcement impact. This is similar in respect to (Anderson et al. 2003) analysis of currency markets where the focus was on determining the actual exchange rate’s rather than the volatility of this financial instrument. This is of comparable interest, but with the analysis carried out here on the oil futures contract prices, which is of intrinsic interest. Many past studies investigated the volatility, the price variation of the financial instruments in question over a period of time but as stated in Anderson et al. 2003, high-frequency discrete-time volatility cannot be extracted accurately unless the conditional mean is modeled adequately first. In Balduzzi (2003), bond price and the resulting price change is regressed with the surprise as a result of an exogenous variable and with variables announced simultaneously, this emphasis on the resulting price change is more in accordance with the dataset used in this study.

As with Fleming and Remolona (1999) and Balduzzi et al. (2001) study the bond market, Mark J. Flannery and Aris Protopapadakis (2002) study the stock market, Anderson et al (2003) studies the foreign exchange market, this study is concerned specifically with oil futures and their movement in the market.

The effect of macroeconomic news announcements causing “jumps” in exchange rates has been suggested in studies and has been accounted for or identified as being somewhat related to particular announcements, primarily announcements that have some association with the financial instruments under consideration.

It is argued in many papers that the most significant changes occur within a short time frame around the time of large macroeconomic announcements. In Dominguez it is debated that the largest change in exchange rate occurs ten seconds of macroeconomic news announcement and close timing of central bank interventions to news announcements increases their impact. The correlation between the announcements and the impact on relevant time series within a certain time frame is of particular significance. The “speed of the impact” (Balduzzi 2001) or price adjustment due to an announcement is a constituent of each study. With this the determination of the size of the response to the impact caused is another constituent.

By referencing previous studies that have examined how scheduled an-
 announcements affect exchange rates and their fluctuations, I wish to explore other relevant announcements under the same conditions and their subsequent impact on relevant financial time series data. It is my intention to extend this concept further with regard to the price of oil futures contracts and the impact that the scheduled oil inventory announcement may have upon it.

1.3 Dataset

The original dataset used is for the Oil future contracts (CL1), which are traded ‘continuously’ on the Chicago Board of Exchange. The time series used is derived from the last quoted price of the crude oil futures contracts in a five-minute period. The series is examined using this high-frequency intraday data ranging from c2008 to 2011. The prices at the time of the announcement, thirty minutes before and also thirty minutes after were noted. Taking $P(0)$ as the price at the time of the announcement, $P(30)$ is price thirty minutes after the announcement and $P(-30)$ before the announcement. The Price change between these times is calculated as well as the price change across the entire hour around the announcement, from 10.00 to 11.00, these figures are then used for analysis and are also used as the working dataset that is modelled and used to train the system. The following demonstrates the price change which is being examined:

\[
\text{Price deviation 30 minutes before} = P(0) - P(-30)
\]
\[
\text{Price deviation 30 minutes after} = P(30) - P(0)
\]

Using these price deviations as the dataset the empirical distribution functions are generated giving the solution space. Modelling the behavior in this way is useful in determining the event and can be easily extended to examine other innovations and their impact.

1.4 Event, Oil Inventory Announcement

The particular event that is looked at is the Oil inventory report released by the Energy Information Administration in the US. The report contains general information regarding crude oil supply, consumer consumption and production and refinery utilization. Released every week on Wednesday at 10.30am it serves as a variable of interest. The main element of interest is the oil inventory portion of the announcement, which gives an account of the change in the number of barrels of crude oil held in inventory by commercial firms during the past week. This announcement is of particular significance as it has a direct affect on oil prices and oil products and extends to influence industries that are reliant on oil as a result.

The preceding figure1.1 shows a representation of previous oil inventory announcements with the results of the forecasted and actual oil inventories
Figure 1.1: Forex Factory forecast and record of physical oil inventories from weekly Oil Inventory Announcement (EIA) [1]

held for each week of that announcement. It is recorded as the number of barrels of crude oil per million held in inventory. The forecasted values are predicted a week before the announcement and can be seen here as the highlighted yellow lines, with the actual number of barrels of oil represented as the separate bar. The number of barrels is counted in the millions while each instance on the graph is one announcement, each week, over a six-month period.

To evaluate the affect of the oil inventory announcement it is necessary to know the date and time of its statement as was the case with Anderson et al 2003 and previous studies which studied the timing of the innovation. This was determined from the forexfactory calendar [1] that gave a historical record of previous announcements and also the next subsequent announcement’s time and date.

Then the identification of significant jump components in oil futures prices is undertaken and an analysis made of the relative contribution of jumps to the realized price change in the futures contracts price. The investigation leads on to see if the significant jumps are due to the Energy Information Administration’s (EIA) inventory news announcement related by certain factor such as the announcement’s timing.
1.5 Thesis Outline

I will first discuss the background knowledge base for the project and the related areas of interest and discuss particular topics relevant to this project. I will then move on to discuss how exactly the process and methodology for this project was undertaken and demonstrate some of the results obtained and analysis done. Building on this knowledge I will then describe how these techniques and the system was implemented to create an independent java application, which automated the analysis process. Finally I will discuss conclusions drawn and some considerations for extending the project further and the future work involved.
Chapter 2

Background, Time series analysis, Modelling and Forecasting

2.1 High frequency time series

A time series is a collection of observations of defined data items obtained from repeated measurements made sequentially over a period of time. Financial time series analysis is the specific application to financial data and concerned with the theory and practice of asset valuation over a period of time[18].

As outlined, this thesis is concerned with monitoring price behavior, in particular, the fluctuations that occur due to a specific announcement of oil inventories. By monitoring the price behavior frequently and periodically and attempt to understand it thoroughly, it may be possible to estimate the likely development of prices in the future. Observing a series based on financial prices and data, such as stock, currency or commodity prices, an attempt can be made to describe the events and by employing standard techniques maybe develop a system to forecast the behavior.

To understand how prices behave analysis must be conducted on the actual prices of interest using statistical methods. Forecasting events such as price change, volatility, has a degree of uncertainty but the values and events can be described by a probability distribution and using past data and observations it may be possible to estimate the average value of future distributions.

High frequency data is generated from observations taken at five time intervals. In the case of financial data it is often daily or even finer intervals such as hourly or per minute. The importance of high frequency data lies in its ability to reveal the market microstructure. From macroeconomic realizations and events “jumps” can be observed (Anderson et al 2003) and in
In this case it is intended to relate this to the oil inventory announcement. From high frequency time series, the result of the innovation or any information event can be observed and understood. It is from this understanding then that an attempt is made to model the behavior and ultimately predict it for future, similar occurrences.

It is necessary to know the exact time of the innovation to properly predict it’s particular behavior on price change. In this regard there is no ambiguity as the time and date of the selected announcement is predetermined weekly. The impact of the event can therefore be assessed consistently and accurately.

2.2 Stochastic Processes and Asset price

Much statistical theory is concerned with random samples of independent observations. However, successive observations are not often independent and the analysis must take into account the time order of the observations. With dependency existing between successive observations, it is possible to predict and forecast future values from past observations. If in some cases the values can be explicitly predicted then the system is said to be deterministic. This, however, is rarely the case in real world scenarios where most time series are stochastic and future values can only be partly determined by past values and occurrences[2]. Explicit predictions for values are consequently impossible and instead are exchanged for a probability distribution which is built from a knowledge of past values.

The price of a financial asset varies over time and can be identified as a stochastic process[16]. This describes the progression of a random variable over a specific period of time. The prices that are observed in this case are a typical example of this. There are two types of stochastic process for modeling in time series; these are discrete-time and continuous-time processes. Continuous processes are defined when observations are made continuously while discrete are taken only at specific times in most cases equally spaced. Although many financial instruments are traded continuously for the purpose of analysis the data is sample at discrete time intervals, in this case 5mins intervals, giving a discrete time series, which is used here.

2.3 Forecasting Models and the Monte Carlo Method

By developing a forecasting model an estimate is made of an expected value in the future. An exact value cannot be determined but based on previous occurrences and by making observations on the dataset of past values an estimate can be made. This estimate, although it may be accurate in some instances, still has a degree of uncertainty. It can, however, give an indication to the most likely or most probable next value in the future and this is the
benefit of the forecasting model. One such model that will be looked at and examined out is the Monte Carlo method.

Monte Carlo method is a term used to describe the broad range of problem solving techniques that use randomly generated values and the statistics of probability. The key components of a Monte Carlo algorithm are:

- **Random Number Generation**
  - Reliable number generation is essential for the success of a Monte Carlo simulation. If the randomly generated values show non-random behavior in a short interval of time then the simulation is compromised. Simulations are carried out with multiple iterations and often with a pseudo random number generator as is the case here.

- **Sampling**
  - The algorithm works by using the randomly generated values to sample the solution space of the problem. For a data set some probability can be applied to each point. This is can be determined in several different ways. For the approach taken here, the probability of a point is determined by its frequency in a larger dataset. The accuracy of the solution from sampling increases with the amount of points sampled. It is also noted that not all points in the solution space contribute equally to the solution, this can have a decided affect on the overall outcome as will be seen.

- **Probability Distribution Functions**
  - Sampling is done according to a probability distribution function. The Probability distribution can give the likely probability of a random variable’s occurrence at a given point or the probability of the value falling between particular intervals.

- **Error estimation**
  - Estimation of the error of the algorithm used is necessary in analyzing it’s performance and the iterative process used.

### 2.4 Probability Models

A probability distribution identifies the probability of each value of a discrete random variable or the probability of a continuous value falling within a particular interval. A histogram can give an illustration of how data is
Figure 2.1: Histogram of dataset built from many Price Changes across a period of 1 Hour around the time of an event

distributed and an estimation of the probability distribution of the variable. An example of a histogram printed out from program after analysis is as follows:

The histogram consists of frequencies defined over discrete intervals or bins. The height of each rectangle of the histogram is the frequency density, which is the frequency divided by the interval width. The frequency or height of each rectangle represented, as seen above, gives an indication to the number of data points or values in that particular bin. The bin interval is calculated simply by:

$$I = \frac{\text{Max}(x) - \text{Min}(x)}{k}$$

Where $I$ is the interval, $k$ the number of bins, $\text{Max}(x)$ and $\text{Min}(x)$ are the extreme points in the range of the dataset. A more formal definition of this function is if $m_i$ is a count of the number observations that fall into each of the disjoint categories (bins). Then letting $n$ equal the total number of observations and $k$ the total number of bins:

$$\sum_{i=1}^{k} m_i$$
In the illustrated histogram, the dataset is built from price changes ranging from several announcement days. The approximation of the frequency distribution of the price change is made then once the bin intervals have been created. The cumulative distribution is then built from the histogram by keeping a count of the cumulative number of observations in all the bins up to the present bin. The cumulative distribution function $M_j$ of histogram $m_i$ is thus:

$$\sum_{j=1}^{k} M_j$$

Using the Cumulative distribution it can be seen that the probability of a random variable $X$ will be found at a value equal to or less than $x$ [14].

Figure 2.2 is an example cumulative distribution printed out from the program which was written for this project to automate the analysis procedure. Where the bins (x value), are represented as the range of price changes that have occurred over many consecutive announcements. The x values correspond with the histogram x values as the cumulative distribution represents the normalized accumulation of the frequencies from the histogram, providing an alternative representation which makes further data interpretation possible.
2.4.1 Statistical Analysis of the Distributions

Using traditional statistical analysis on the dataset more information can be inferred and a greater understanding of the events occurring can be attained. The mean can give an idea of the central tendency of the distribution and characterize the dataset. From this the variability of the distribution can be examined by looking at its variance and standard deviation. The equation to calculate the variation being:

$$\sigma^2 = \frac{\sum(X-\mu)^2}{N}$$

Where $\mu$ is the mean and $N$ the total number of elements in the set, $X$ each element of the set.

The standard deviation is simply calculated as the square of the variance. It is used as a measure of the spread of the distribution and from knowing both the mean and standard deviation the percentile rank of a value can be found. The percentile rank is beneficial as it can indicate the number of standard deviations away from the mean certain proportions of the distribution are. To calculate the number of standard deviations a sample is from the mean the following equation is used:

$$z = \frac{X-\mu}{\sigma}$$

Where $z$ is the number of standard deviations, $X$ the sample value, $\mu$ mean and $\sigma$ the standard deviation of the dataset.

The following histogram, Figure 2.3, was constructed from price changes taken around an hour of an announcement. This particular example being used for illustration purposes has several events taken into consideration.

Table 1.1 shows the data values used to generate the histogram in Figure 2.3.

By graphing the data, information can be quickly interpreted, but the benefit of modeling the data in this way is in the manipulation of it and corresponding results. To demonstrate the information shown on the histogram it is easy to note that the probability of a zero or negligible price change is 0.0%(Table 1.1). We can also demonstrate that the probability of a price change of 0.4 is 13.3%, this is calculated from the ten independent occurrences of prices changes which gave a rounded value of 0.4 to the total number of values in the dataset.

Looking across at the cumulative percentage for a price change of 0 it is seen that there is a 50.7% chance of obtaining this value or less. This gives an overall representation that there is in fact a 50.7% probability that there will be no price change or a price decrease. This implies that there is a 49.3% probability of a price increase as a result. This distribution can completely describe the probability for a realization of a variable to be equal to or less than the corresponding $x$ value.
Figure 2.3: Bimodal distribution created from Price deviation from Price close before and after an Oil Inventory Announcement

Table 2.1: Tabulated Dataset of values used to generate Figure 2.3 histogram and Figure 2.4 Cumulative distribution
The resulting cumulative distribution figure 2.4 once graphed can show a representation of this data for faster interpretation.

The standard deviation is often used as a simple means of measuring volatility. If a stock’s return for example, varies greatly from the stock’s average return then the stock is more volatile. If a sample from the dataset is taken 1.42, its deviation from the mean can be calculated from $z$ above as:

$$\frac{1.42 - 0.01}{0.62937} = 2.2$$

It can be interpreted that this value is two standard deviations above the mean. Similarly we calculate for -0.62 that it is 1 standard deviation below the mean. On calculating the deviation for the dataset it is found that 5% are two standard deviations, 24% are one standard deviation from the mean and 71% are within one standard deviation from the mean. A graphical representation of the spread is seen here:

What can also be noted with respect to the above results is the consideration of the “68-95-99.7 rule”[15] or empirical rule, which states that for a normal distribution, nearly all values lie within 3 standard deviations of the mean[15]. Comparing graphical to a typical normal distribution:
Figure 2.5: Regraphing of the Histogram in Figure 2.3 showing the percentage deviation of values from the mean, blue lines indicate values are within one standard deviation, red lines indicate within two standard deviations and last period values are within three standard deviations of the mean.
Figure 2.6: A plot of a normal distribution (or bell curve). Each colored band has a width of one standard deviation.[21]
2.4.2 Forecasting future values

Forecasting the future values of a time series is the main concern once the distributions have been properly modelled. From an observed time series \( x_1, \ldots, x_n \) the basic premise is to estimate the value \( x_n + h \), where \( h \) is the lead time or forecasting horizon[1].

With respect to the dataset used in this instance, the series of price changes around the time of the announcement are modelled, from this model it may be possible to predict a price change for around the time of the next announcement. This predicted value is then added to the time \( P(t) \) to give an estimation of \( P(t + 1) \), the forecasted price of the oil futures contract after the announcement. The initial value of \( P(t) \) is determined by the what price change values are used to train the system, \( P(t) \) will be defined as the price at the start of the time frame used.

After the realization of the empirical distribution functions by modeling the price change, the model is then run to look at the distribution of synthetically generated price changes. This is done by generating random numbers between 0 and 1, the number of numbers generated corresponding to the number of simulations desired. Then the random numbers are interpolated with the cumulative distribution by creating a cubic spline interpolation of the dataset.

For the spline interpolation of the cumulative distribution multiple piece-wise continuous functions, in this case cubic splines, are fit to the data set for each point and data interpolation with the spline is now possible. The random numbers are interpolated giving a synthetic value for each. A mean of these synthetic values is obtained over multiple iterations. The synthetic mean represents the out of sample prediction used to find the forecasted \( P(t + 1) \).

The typical method for assessing a forecast is by calculating its mean square error. This is the error estimation element of the process and particularly good for evaluating the generated series. It is implemented by getting the mean value of the square of the difference between the forecasted(\( F \)) values and the actual(\( A \)) observed values:

\[
RMSerror = \sqrt{\frac{\sum_{t=1}^{n} A - F}{n}}
\]

2.5 Summary

The financial time series derived from the continuously traded oil futures trading contracts provides a meaningful and easy way to interpret the data as a sampled discrete time series. It is possible to quickly build a series of analytical techniques upon this basis to identify and better understand the data and recognize if any underlying patterns exist. Using statistical analysis further observations can be made and investigations carried out on the data.
It is the hope then that from interpreting past behaviors it may be possible to encounter similar events in the future and from the knowledge of data already obtained predict the likelihood of such events.
Chapter 3

Analysis and Modelling

3.1 Modelling Noise

The main aim of this project is to attempt to model an event’s influence on a time series by creating probability distributions through empirical analysis. Essentially we are attempting to model the noise in the series. We can examine this more closely through the following equation:

$$P(t) = \sum \alpha_i P(t - 1) + \beta_t + \epsilon_t$$

Where $P(t)$ is the actual value of the series, $\sum \alpha_i P(t - 1)$ is the expected value with regression coefficients:

$$\alpha_1 = \sum P(t).P(t - 1)$$
$$\alpha_2 = \sum P(t).P(t - 2)$$
$$\alpha_n = \sum P(t).P(t - n)$$

With $\beta_t$ as the variable which determines the time of the announcement, it is given the value of one for the time when there is an announcement and zero for any other time not specified.

The element that we wish to model is the noise given as $\epsilon_t$, essentially we re-arrange the equation:

$$\epsilon_t = P(t) - \sum \alpha_i P(t - 1) - \beta$$

The noise then essentially represents the effect of the announcement.

In the following analysis, the announcement’s impact is first examined and how it is represented and examined with the time series for the price changes across multiple announcement days. The resulting distributions from the datasets are then generated from the program and sample results are examined to see if there are any significant deviations. A method to forecast the future mean price change is suggested and carried out with suggestion to the results of the experimentation.
3.2 Price adjustment, impact size and timing

To monitor the price adjustment due to the announcements effect the resulting price change post-announcement is looked at with the dependent variable being the 30-minute window chosen. Previous studies (Ederington and Lee 1993, Payne 1996) have suggested a 15-minute window. Ederington and Lee suggest the major bulk of price adjustment occurs within one minute after the announcement with prices being considerably more volatile for a further 15 minutes and slightly more volatile for several hours after in comparison to a trading day with no significant innovations, again price volatility in the foreign exchange market is what is being observed but implies the same effect for the underlying prices.

In this study the fluctuation in price change in a 30-minute window prior to the announcement and again after the announcement is observed. This ensures the most significant portion of the movement is accounted for. The time frame is still close enough to the announcement to observe it’s main impact. The total price change across the hour is then looked at to get a general overview of the price movement. From looking at the two time frames and accounting for the price adjustment due to the respective event then it may be possible to forecast a value for price by calculating

\[ \text{Price at } 10.00 + \Delta_{60\text{mins}} = \text{price at } 11.00 \]

\[ \text{Price at } 10.30 + \Delta_{30\text{mins}} = \text{price at } 11.00 \]

If the news contained in the macroeconomic announcement is a fundamental determinant of price fluctuation then there should be a significant impact in the post release price. The impact, however, is only significant up to two hours after the release as macroeconomic information is often drowned out in the subsequent random fluctuation (Almeida 1998).

From the histograms generated in Figure3.1 of the price change 30-minutes before and after the announcement, it is possible to see a change in the distributions. The values are clearly distributed further away from the mean after the announcement resulting in a greater standard deviation and spread of the distribution this is also evident in the figures computed for each distribution of data from the figures in table2.

The result that values are distributed further away from the mean suggests that some exogenous variable maybe having an influence on price movement. The resultant histograms above and difference in standard deviations certainly suggest some evidence of this exogeneity.

For announcements of economic data, in particular announcements that have a regular pre-arranged time fixed to the minute, it is likely that some form of anticipation and planning such as consultation with technical analysts and economist will exist before hand. Due to this previous knowledge of the event’s scheduling, faster assimilation of information can re-
Table 3.1: Mean of price change and standard deviation occurring between 10.00 and 11.00 before announcement and between 10.30 and 11.00, which is price change after announcement

<table>
<thead>
<tr>
<th></th>
<th>Mean (-30)</th>
<th>Mean (30)</th>
<th>Standard Deviation (-30)</th>
<th>Standard deviation (30)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-0.0357971</td>
<td>-0.00477612</td>
<td>0.26309474</td>
<td>0.55633366</td>
</tr>
</tbody>
</table>

Figure 3.1: Histograms of price change occurring 30 minutes before and after the 10.30 oil inventory announcement for multiple announcements, takes forty days of oil inventory announcements
Figure 3.2: The resulting cumulative distribution generated from histogram Figure 3.1

sult as compared to announcements that have no pre-scheduling and longer lags result (Almeida 1998). Consequently reaction to the information post-announcement is concentrated into a short period of time.

3.3 Distribution analysis

The resulting histogram from the price adjustment in the thirty-minute time frame after the announcement is seen in Figure 3.1. The cumulative distribution that is then derived from this histogram is seen in Fig8:

Table 3 displays the data from the distributions created from the price change dataset after the announcement:

Due to larger spread of values from the mean and greater standard deviation for the dataset from the post announcement time it is interesting to examine these results in greater detail.

Firstly for the maximum and minimum value of the dataset the following is observed:

Once the program has been properly trained with the dataset the main aim of its application is to give a prediction for the next price adjustment due to the announcement or event. Using the modelled data represented by
Table 3.2: Tabulated data for Figure 3.1 and Figure 3.2, represents the % frequency of data in the histogram and also cumulative percentage of values.

<table>
<thead>
<tr>
<th>bins</th>
<th>frequency</th>
<th>%freq</th>
<th>%cdf</th>
</tr>
</thead>
<tbody>
<tr>
<td>-2</td>
<td>0</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>-1.8</td>
<td>0</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>-1.6</td>
<td>0</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>-1.4</td>
<td>0</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>-1.2</td>
<td>0</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>-1</td>
<td>0</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>-0.8</td>
<td>1</td>
<td>2.5%</td>
<td>2.5%</td>
</tr>
<tr>
<td>-0.6</td>
<td>2</td>
<td>5.0%</td>
<td>7.5%</td>
</tr>
<tr>
<td>-0.4</td>
<td>0</td>
<td>0.0%</td>
<td>7.5%</td>
</tr>
<tr>
<td>-0.2</td>
<td>9</td>
<td>22.5%</td>
<td>30.0%</td>
</tr>
<tr>
<td>0</td>
<td>9</td>
<td>22.5%</td>
<td>52.5%</td>
</tr>
<tr>
<td>0.2</td>
<td>3</td>
<td>7.5%</td>
<td>60.0%</td>
</tr>
<tr>
<td>0.4</td>
<td>7</td>
<td>17.5%</td>
<td>77.5%</td>
</tr>
<tr>
<td>0.6</td>
<td>3</td>
<td>7.5%</td>
<td>85.0%</td>
</tr>
<tr>
<td>0.8</td>
<td>3</td>
<td>7.5%</td>
<td>92.5%</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>2.5%</td>
<td>95.0%</td>
</tr>
<tr>
<td>1.2</td>
<td>2</td>
<td>5.0%</td>
<td>100.0%</td>
</tr>
<tr>
<td>1.4</td>
<td>0</td>
<td>0.0%</td>
<td>100.0%</td>
</tr>
<tr>
<td>1.6</td>
<td>0</td>
<td>0.0%</td>
<td>100.0%</td>
</tr>
<tr>
<td>1.8</td>
<td>0</td>
<td>0.0%</td>
<td>100.0%</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0.0%</td>
<td>100.0%</td>
</tr>
</tbody>
</table>
The training set that was used to create the initial model was based on thirty-seven data points with a prediction made for the three next subsequent data points, which were then compared to the actual values. The cumulative distribution that is created by the program is of particular importance as it provides the data to which the analysis and forecasting is created from. An interpretation of the cumulative distribution of price changes by the program is seen here in Figure 3.4.

The results from the first analysis conducted on the data for the price deviation thirty minutes after the announcement are seen in Table 3.3. The simulation was run for a thousand iterations and the mean of all the simulations was obtained. Although the exact price change on a particular
Figure 3.4: Cumulative distribution for price change on post-announcement 30-minute time frame, graph generated from the application developed
day is difficult to interpret, a relation may be drawn between the actual mean price change calculated for all the announcement days. The result of the forecasted values relates closer to the distributions generated from the previous observations. As seen in the histogram generated

The distribution as seen suggests some form of bimodality. A strong peak exists in the area just left of no price change or zero, an indication that there is a greater frequency of minus values just left of the zero as compared to the strongest positive peak just right of the zero margin, which reveals the second highest frequency of price changes. The forecasted values are closer in line to the mean price change of the actual value than they are for the actual price change that resulted for any one particular day.

Important information that was returned by the program’s analysis is the probability of the particular price changes occurring due to the event. As seen in table the values were derived from the cumulative distribution. The probability of obtaining a price change result of -0.02 or less is 60.4%, which reflects the distributions obtained where a stronger possibility of a slight price decrease is suggested due to the higher frequency.

Similarly an actual price change of -0.72 occurred, which in some sense could be seen as an outlier in the distribution as its probability of occurrence
is low at 5.25% although still within two standard deviations of the mean.

The same analysis was repeated for the hour time frame. It was intended that looking at the price change across the entire hour around the announcement that some overall impact would be noted but still short enough that an impression from the announcement would still be detected.

The mean result of a thousand simulations was again found and compared to three actual price changes from three subsequent announcements and against the actual mean price changes up to each successive announcement day as before.

The distributions built from the price change across the hour assign a different weight to the probability of price changes due to the different time frame. As seen in figure 2.5 and in the histogram printed from the program:

The distribution suggests some bimodality with two distinct peaks, one positive and another negative. This can result in a more staggered cumulative distribution as opposed to a more pronounced sigmoidal representation.
Table 3.4: Mean forecasted price change for one hour around the time of the announcement, result from simulation run on trained system or previous price changes around the announcement over forty announcement days, compared to the actual price change for three days and actual mean price change on those days.

<table>
<thead>
<tr>
<th>Forecasted Price Change(1hr)</th>
<th>Actual (day)</th>
<th>Actual (Mean)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.07</td>
<td>62.7%</td>
<td>75.6%</td>
</tr>
<tr>
<td>0.05</td>
<td>-0.64</td>
<td>12.5%</td>
</tr>
<tr>
<td>-0.02</td>
<td>-0.38</td>
<td>17.5%</td>
</tr>
</tbody>
</table>

Figure 3.7: Histogram of price changes around time of the announcement, between 10.00 and 11.00, generated from forty data points of from just announcement days.
3.4 Process Summary

Analysis is carried out only at the time on an announcement and around the time of the announcement. This way the announcement’s effect can be studied in isolation from the rest of the daily price data and doesn’t decay or disappear as rapidly as if were to examine low frequency data. The model is made from the distributions, which are generated by using the training data set of past observations. Once they have successfully generated the system a random value acts as one simulated price change. The simulation is repeated more than a 1000 times and a mean value of the simulations is created, this provides a forecast with a corresponding probability attributed to it which can be recognized by identifying it with respect to the cumulative distribution, which gives the probability of obtaining that value or one less within than that interval of values.
Chapter 4

Implementation of the Analysis Program

As the central part of my work I developed a program to independently handle the data to be used and then carry out the desired analysis, developing the model and distributions for prediction. The design features are outlined first in the following program design. The java implementation and in particular apis used are then discussed followed by how the GUI is implemented and displays with the last part of the chapter devoted to the program’s evaluations and concluding remarks.

4.1 Program Design

The application that was developed to conduct the calculations and analysis on the dataset was implemented in the java language. The choice of developing in the java language provided many benefits to this project. Primarily the existent Java API’s that provided classes for many of the fundamental operations necessary in modeling the data and interfacing with external files containing necessary data.

The first aspect of the program’s implementation is data extraction from excel. Preprocessing of the data is often necessary, in this case to determine the price change from quoted close prices of the oil futures. After the desired dataset is built properly from the actual price quotations the data is then used to create the histogram based of the frequency of certain values and the dataset’s range. Using the arrangement of data in the histogram a cumulative distribution is then built by passing the histogram dataset to it. Spline interpolation is then done on the cumulative distribution data that’s been created with random numbers generated between zero and one to give corresponding x values. These x values correlate with the dataset that has been modelled. A mean of the interpolated x values is then calculated over multiple iterations. The most important data that’s printed our from the
program are the two graphs of the resulting probability distributions for the histogram and the cumulative distribution and the forecast generated from the simulation. Other important information that is generated within the program is the histogram and cumulative distribution data and probabilities computed for the datasets.

### 4.2 Java API’s

The use of java api’s provided many benefits in removing coding overheads and improving the efficiency of the program. Some of the basic methods necessary for the data manipulation were coded and calculations done and compared with the results from the classes used from the external libraries and the result was found to be the same with little discrepancies. This provided adequate proof that the external classes were appropriate for continued development and progress in the project program.

The Api’s used were:

- Apache POI[17] - Java API for Microsoft Documents - In particular POI-XSSF for accessing Microsoft excel files (xlsx format)
• JFreeChart [19] – java library for displaying and representing data graphically and printing charts

• Apache Commons Math[20] - a small library with mathematics and statistics components and functions.

Although the api’s were restrictive in some sense and limited in some of their applications, their use meant it was possible to avoid having to code complex mathematical algorithms and interfacing with excel files would have been impossible.

4.3 Data Extraction

The first part in the process of data extraction was to create a java object from the excel file. Essentially a “HSSFWorkbook” object is created using the excel poi. This is the root object that is accessed which references the excel file being used. A sheet object “HSSFSheet” is created using poi excel which holds the sheet containing the relevant data. This information about the excel file is passed to the application by the user through the gui and where variables such as the root directory and the sheet number, arranged in numerical order as opposed to the actual sheet name, are set.

The data extraction can be performed in two ways depending on the information known and supplied by the user. The two methods differ on the way in which the data extraction is done while the actual analysis of the data remains the same in both implementations.

Excel Define region works reads in data directly from an excel sheet without doing any processing of the data. The user specifies where in the excel sheet the data is located by giving the row and column information and the application then stores the relevant excel information. The application runs a series of methods that were written that searches the region defined, checks that the values are numerical values and stores them accordingly in an array.

The excel search and store implementation is based on the user’s knowledge of the events in question and will preform the necessary preprocessing of data as opposed to the excel defined region which just reads in data directly. A user first supplies price and information data for a given financial instrument to analyze. Then specifies the dates of when the event occurs and the time frame around the event that is to be analyzed. The user also specifies the column number from the spreadsheet locating the financial data to be processed. The implementation shown for Excel Search and Store in the GUI above will find the price difference or change in price between the two times specified and store the values in an array. This portion of the application simply calls methods written to search through the excel sheet object and finds a match between the dates and times supplied and returns
Figure 4.2: GUI design, front end seen by user with two data extraction implementations, method of data analysis remains the same
the values from the specified column from that row. These values are then processed accordingly to give a price change for example.

In both cases some knowledge of the excel file and how the data is stored is required. It is also important to note that the dates and times must be properly recorded in the xlsx file or some errors or false results will be returned from the search and store function.

4.4 Data Handling and Modelling

Once the data has been properly extracted and processed it is then passed to the HistogramChart class. A function within this class then builds the dataset for the histogram and cumulative distribution. The final design settled on for the application utilizes the jfreechart library’s HistogramDataset class as it provided robust methods for handling multiple dataset types, greater exception handling and representing the information graphically was much easier as it already had specific built in functions for graphing the histogram data.

The number of bins desired and the array of values are used to create the Histogram dataset. It is from the histogram and it’s data that the dataset for the cumulative distribution is built from. Very few existing libraries could easily build a cumulative distribution or interact with the HistogramDataset object easily so much of the calculation and preparation was done by creating a new defaultKeyedValues object from the jfreechart library, this provided a simple solution for creating the cumulative distribution dataset by indexing the frequency values (y values from the histogram) with the correspond bin values (x values of histogram). The cumulative data, specifically the frequency or y values, is then normalized to obtain values between 0 and 1. The resulting dataset is then graphed displaying the corresponding cumulative distribution.

4.5 Simulation and Results

The first part to developing the simulation was creating a random number generator method. A simple function was written, which fills an array with pseudo-randomly generated values between 0 and 1 with the number of values generated representing the number of random simulations to run. For the interpolation of the random numbers with the cumulative distribution a “SplineInterpolator” object was created by using the apache commons math api. From this a natural cubic spline interpolation for the data set is calculated. Using the supplied interpolate method a PolynomialSplineFunction (from the commons math library) is returned and the random numbers are interpolated with this spline function. These interpolated values are then stored in an array as the synthetic values that have been generated from
Figure 4.3: The GUI output after running the excel defined region implementation of a given dataset contained in excel

the simulation. The mean of the synthetic values is then calculated and returned.

The output of the calculations are sent to a textpane for the user to interpret them as shown:

The following graphs are an example of the graphical representation of the data that is done by the program. The files are outputted as png files to the directory containing the application.

4.6 Concluding remarks

Both methods of data extraction performed by the program require the user to be informed about how the data is stored and formatted within the excel sheet and if the formatting is in fact correct. This can prove to be prob-
Figure 4.4: Histogram and cumulative distribution graphs output from the application running on given dataset
lematic primarily for very large excel sheets containing thousands of data entries. Large datasheets such as these can often cause processing errors with the poi excel api and it is advised to parse the excel sheet into smaller sets to overcome this problem. This can help in evaluating the data entries formatting also.

The excel defined region approach can provide a robust, simplistic way of extracting the necessary data and in many cases is the best method to use with this application.

The program implemented with the GUI gives a lot of freedom and independence that other programs such as matlab do not have. Such programs require the user to have a firm grasp not just on the knowledge of the method but also in writing the actual script. The application that I have developed for this project requires minimal knowledge from the user except for what dataset is to be analyzed and with respect to what event. The GUI has a simple input and output interface to facilitate ease of use by the user but the underlying logic and methods written can provide the user with a more in depth view of the distributions, calculations on the data and new datasets and functions which have been calculated by the program.
Chapter 5

Conclusion

5.1 Review

In this paper I have studied the impact of macroeconomic news on a financial time series. This has been demonstrated by examining the reaction of the price of Oil futures trading contracts to a weekly announcement reporting on the Oil inventories held by the Energy Information Association in the US.

To analyze the price effect I used intraday high-frequency data consisting of price quotations sampled at the end of 5-minute periods. This facilitated the measurement of the announcement’s impact on price in a smaller time interval getting a more accurate estimation of the impact’s intensity. This followed on from work set out by Almeida(1998) and carried on in studies by Goodharte et al(1998) Anderson et al(2003), Balduzi(2003), and Chabaud at al(2008).

I draw some conclusions from the study that suggests a quick, sudden impact from the announcement on the price of oil futures similar to the propositions made in Almeida (1998) Balduzi (2003), where a significant effect or jump was experienced in the data analyzed. This observation was made due to the change in distributions that occurred after the event. It is also observed, similar to the previous references, that the impact is quickly incorporated into the price to the degree that on low frequency data the impact very quickly decays as implied by Balduzi(2003).

The forecast generated from random simulation on the modelled data provided some insight into the mean price change with a short forecasting horizon. An accurate determination of the next subsequent price change given a system trained with, as up-to-date data as possible can still prove somewhat difficult.
5.2 Comments and Future Work

Although a link between “jumps” and fundamentals can be postulated, with reference to Anderson (2007), it could be assumed or examined in later work that more than one event or fundamental can have a joint effect on determining the price even around this particular announcement time. Research papers have conducted joint effect analysis of announcements such as Anderson (2003) which could also be a point of interest to examine with the financial data I have used in this paper.

The program and system it implements could also be extended to increase its forecasting horizon by increasing the dataset available to it and also developing the method to which probabilities are weighted to past observations as looked at with Ederington (2010). Increasing the dataset used in this study to roughly a hundred data points from forty may also help in it’s ability to derive the distributions. More outliers may also be evident if the dataset was increased as would be expected from a statistical perspective. Extending the program to easily adapt to constantly new data keeping the dataset at a constant hundred data points without having to directly specify would help in maintaining the distributions and keeping model generated relevant. Implementing an automated moving average feature would help this consideration.

The scope of analysis in this area of financial modeling can increase to encompass many elements and have a broad spectrum. A direct consideration for future work would be to analyze the news content of the oil inventory announcement as stated from the forex factory[1] calendar and determine if a response to price exists to the forecast and actual inventories quoted.

In summary I have given an indication that a financial time series can be influenced by an exogenous variable. The oil inventory announcement can in fact impact the price of oil futures trading contracts with notable results and given past observations a system of prediction can be put in place to give the most likely occurrence and to characterize the behavior of the event.
Bibliography


