Computer-Based High-Frequency Analysis of Financial Time Series: An API Architectured Solution

Tomás Hanley
B.A.I. Engineering
Final Year Project April 2014
Supervisor: Dr. Khurshid Ahmad

School of Computer Science and Statistics
O’Reilly Institute, Trinity College, Dublin 2, Ireland
Declaration of Authorship

I hereby declare that this project is entirely my own work and that it has not been submitted as an exercise for a degree at this or any other university.

Signed: __________________________________________

Date: __________________________________________
TRINITY COLLEGE DUBLIN

Abstract

School of Computer Science and Statistics

BAI Computer Engineering

Computer-Based High-Frequency Analysis of Financial Time Series: An API Architectured Solution

by Tomás Hanley

In this project, I studied the steps involved in creating an econometric model, and built a system, COMHAIRLE, to perform many of these steps automatically. COMHAIRLE is an acronym for Computer-based High-frequency Analysis of Returns-Lagged, and is also the Irish word for advice. The system comprises access to financial data via API calls to online data providers, and access to an econometric and time series library (GRETl) via API calls. Typically, stock market analysis of the performance of a stock is carried out by purely looking at the past prices of the stock. However, the price of the stock of a company depends on a number of external factors, including its inputs and outputs. In this project I have focused on one of the key inputs to airline companies, the price of crude oil. I used Autoregression (AR) and Vector Autoregression (VAR) models to create an analysis of the company itself, and the inputs to the company. The system applies the Hannan-Quinn Information Criterion (HQIC) to the data to select the number of lags to include in the model. The system then provides the adjusted R-squared value and standard error of the model for the user to evaluate its correctness. The system then produces forecasts for the airline stock, which can be used to aid in financial decision making.
Acknowledgements

I would like to sincerely thank my supervisor for this project, Dr. Khurshid Ahmad, for his expert advice, help, and encouragement extended to me throughout the entire project.
Contents

Declaration of Authorship i

Abstract ii

Acknowledgements iii

Contents iv

List of Figures vi

List of Tables vii

Abbreviations viii

Symbols ix

1 Introduction & Motivation 1
  1.1 Introduction .......................... 1
  1.2 Motivation .......................... 3

2 Method & Data 5
  2.1 Method ................................... 5
    2.1.1 Returns in Econometric Modelling .......... 5
    2.1.2 Econometric Models ....................... 9
      2.1.2.1 Autoregression (AR) Models ............ 9
      2.1.2.2 Vector Autoregression (VAR) Models ..... 10
      2.1.2.3 How to Compare Models ................ 10
      2.1.2.4 Forecasting .......................... 11
    2.1.3 APIs ................................... 12
      2.1.3.1 Data-Collection API ................... 12
      2.1.3.2 GRETL API ........................... 13
  2.2 Data ........................................ 13
    2.2.1 Airline Company Data ..................... 14
    2.2.2 Crude Oil Data .......................... 16

3 Prototypes and Experiments 18
  3.1 COMHAIRLE’s User Interface ................... 18

iv
3.2 Comparing & Evaluating Models ........................................ 21
   3.2.1 Comparison of Returns ............................................. 22
   3.2.2 EasyJet ..................................................................... 22
   3.2.3 Ryanair ...................................................................... 25
   3.2.4 Aer Lingus ................................................................. 28
   3.2.5 Colinear Interdependence Between Oil Variables ............... 29

4 Conclusion ............................................................................ 33

5 Bibliography ............................................................................ 34
## List of Figures

1.1 Steps in creating an Econometric Model (Brooks: 9 [1]) .......... 1  
2.1 Time Series Of Ryanair Returns (2001-2012) .................. 6  
2.2 Summary Statistics, (Taylor: 52 [6]) .......................... 7  
2.3 Distribution Of Ryanair Returns (2001-2012) .................. 7  
2.4 COMHAIRe’s Architecture Diagram ............................... 12  
2.5 Unadjusted Closing Price of Ryanair ............................ 15  
2.6 Adjusted Closing Price of Ryanair ............................... 15  
3.1 COMHAIRe’s Opening Screen .................................... 18  
3.2 Histogram of Returns Tab ....................................... 19  
3.3 Summary Statistics Tab ........................................... 20  
3.4 AR/VAR Model Tab ................................................ 21  
3.5 Prediction of Returns Tab ........................................ 21  
3.6 EasyJet VAR Model ................................................ 23  
3.7 Past Prediction of EasyJet Returns ............................... 24  
3.8 5 Day Forecast of EasyJet Returns ............................... 24  
3.9 Response of EasyJet to shocks in 1.EasyJet, 2.WTI, 3.CL1, 4.Kerosene ... 25  
3.10 Ryanair VAR Model ............................................... 26  
3.11 Past Forecasts of Ryanair Returns ............................... 26  
3.12 5 Day Forecast of Ryanair Returns ............................. 27  
3.13 Response of Ryanair to shocks in 1.Ryanair, 2.WTI, 3.CL1, 4.Kerosene ... 27  
3.14 Aer Lingus VAR Model ............................................. 28  
3.15 Past Forecasts of Aer Lingus Returns ........................... 29  
3.16 5 Day Ahead Forecast of Aer Lingus Returns ................... 29  
3.17 Response of Aer Lingus to shocks in 1.Aer Lingus, 2.WTI, 3.CL1, 4.Kerosene ... 29  
3.18 Scatter Plot of CL1 Returns vs WTI Returns ................... 30  
3.19 Scatter Plot of Kerosene Returns vs WTI Returns ............... 31  
3.20 Scatter Plot of Kerosene Returns vs CL1 Returns ............... 31
List of Tables

2.1 Distribution of Ryanair Returns Compared with the Normal Distribution 8
3.1 Airline Returns Summary Statistics ........................................ 22
3.2 Comparison of EasyJet Models ........................................... 22
3.3 Comparison of Ryanair Models ........................................... 25
3.4 Comparison of Aer Lingus Models ....................................... 27
3.5 Correlation Between Oil Variables ....................................... 32
## Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>API</td>
<td>Application Programming Interface</td>
</tr>
<tr>
<td>AR</td>
<td>Auto Regression</td>
</tr>
<tr>
<td>CL1</td>
<td>Continuous Front Month Crude Oil Futures Contract (see section 2.2.2)</td>
</tr>
<tr>
<td>COMHAIRLE</td>
<td>COMputer-based High frequency Analysis of Returns Lagged</td>
</tr>
<tr>
<td>GRETL</td>
<td>Gnu Regression Econometrics Time Series Library</td>
</tr>
<tr>
<td>HQIC</td>
<td>Hannan Quinn Information Criterion</td>
</tr>
<tr>
<td>SSR</td>
<td>Sum of the Squared Residuals</td>
</tr>
<tr>
<td>TSS</td>
<td>Total Sum of Squares of the dependent variable</td>
</tr>
<tr>
<td>VAR</td>
<td>Vector Auto Regression</td>
</tr>
<tr>
<td>WTI</td>
<td>West Texas Intermediary (Crude Oil benchmark, see section 2.2.2)</td>
</tr>
</tbody>
</table>
Symbols

\[ p_t \quad \text{Asset price at time } t \]
\[ r_t \quad \text{the continuously compounded return at time } t \]
\[ d_t \quad \text{dividends per share during time } t \]
\[ Z_t \quad \text{vector of continuously compounded returns at time } t \]
\[ \varepsilon_t \quad \text{residual, or error term at time } t \]
\[ k \quad \text{number of free variables} \]
\[ n \quad \text{number of observations in the sample period} \]
Chapter 1

Introduction & Motivation

1.1 Introduction

Econometrics is the application of statistical techniques to problems in finance. Financial econometrics can be used to determine the behaviour of asset prices or returns, and test hypotheses concerning the relationship between variables. There are many different approaches to formulating an econometric model, but a valid approach is described by Brooks [1] and shown in figure 1.1.

![Figure 1.1: Steps in creating an Econometric Model (Brooks: 9 [1])]
Steps 1a and 1b are a general statement of the problem. In this case, the hypothesis that the price of crude oil will impact the price of airline stocks. Step 2 is the collection of data relevant to the model. Step 3 is the choice of model to estimate. Step 4 asks if the model produced adequately describes the data. If the answer is yes, proceed to step 5. If the answer is no, you must return to step 1, 2 or 3 and reformulate the model. In step 5 you evaluate the model to see if it agrees with the hypothesis from step 1. If not, you return to step 1. If it does, and you are finally satisfied with the model, you can move on to step 6. The model can then be used for formulating forecasts or suggested courses of action.

I have developed a financial advisory system called COMHAIRLE, which effectively performs steps 2, 3, & 6 in real time on behalf of the user. The system consists of three main stages; the input stage, the processing stage, and the output stage. Firstly, the system input stage. It collects the raw price data of the selected airline, as well as the crude oil prices, through the use of API calls to online data providers. Next is the processing stage. Here the price data is transformed and used to create an econometric model. Key facilities are used to create the model such as the Hannan-Quinn Information Criterion to decide how many lags to include. The system creates Autoregressive models, and Vector Autoregressive models. It calculates the key statistics for the model such as the adjusted R-squared value and the standard error, for the user to evaluate it. It then produces forecasts based on the model which can be used to aid in decision making. COMHAIRLE also provides analysis on the distribution of the returns of the airline. The third stage in the system is the visualisation of the output. COMHAIRLE displays the relevant results in a clear and easy to read format for the user. COMHAIRLE is a java based system. The calculations of returns and graphs were implemented with the help of a java statistics package from http://commons.apache.org, and a chart library from http://www.jfree.org/jfreechart.

I created models for EasyJet, Ryanair, and Aer Lingus, including crude oil prices in the models. I analysed 2758 days of prices for Ryanair, 3474 days for EasyJet, and 1936 days for Aer Lingus. For the crude oil prices I analysed 3569 days of WTI prices, 3565 days of CL1 prices, and 3568 days of Kerosene prices. I found that the stock price of all three airlines was impacted by crude oil prices. Ryanair was the most heavily impacted, followed by EasyJet and then Aer Lingus. It seems that the more ‘low-cost’ the airline, the more susceptible they are to changes in the price of oil.
Chapter 1. Introduction & Motivation

In this report, I explain how the raw price data was transformed for analysis (section 2.1.1). I then describe the models used and how they are evaluated and compared, and look at the forecasts that can be made from them (section 2.1.2). I explain the API’s I created for the system and describe the system’s software architecture (section 2.1.3). I then provide detail on the raw data used (section 2.2). Finally, I discuss the current COMHAIRLE prototype (section 3.1), and the experiments I carried out using it (section 3.2).

1.2 Motivation

In most econometric textbooks, it was accepted that to understand the behaviour of any asset or price series, all you had to do was analyse the past prices of the asset itself. If you wanted to understand the value of a currency, you looked at its past prices, if you wanted to understand GDP, you looked at GDP in the past. The theory was that any other figures impacting the price would be encompassed in the past price. By analysing how historically, past prices have impacted prices, we can create a model to show how current prices will impact future prices. This econometric model is called an Autoregressive (AR) model. This model stipulates that the price of a stock depends linearly on its own previous values. (Taylor 2005 [6])

However, Sims in 1980 [5] argued that past prices alone do not contain all of the effects on the price. He proposed that the price of a stock is not a scalar value, but rather a vector value, composed of multiple scalar components. Analogous to how velocity is resolved into two scalars, speed and direction, price can be resolved into its various components. A stock price $p_t$, is a vector which contains the price itself, and any other variable which might affect it. If you only create an AR model, you may be missing some important components of the price. So in order to analyse a stock, we have to look at all other stocks that impact the stock, and put all of this data together to create a model. To do this, Sims advocated a model called a Vector Autoregression (VAR) model. This allowed you to add in as many variables to the model as you wish. This model provided superior forecasts and analysis of the variance in prices, than the previous AR models.

The Efficient Market Hypotheses is an investment theory that states it is impossible to “beat the market” because stock market efficiency causes existing share prices to always
incorporate and reflect all relevant information. According to this theory, stocks will always trade at their fair value, making it impossible for investors to either purchase undervalued stocks, or sell stocks for inflated prices. (Investopedia [7])

However, Pesaran [4] shows that at times of market euphoria or gloom, widespread beliefs and individual irrationality can cause significant departures from market efficiency, and predictability tends to rise during these crisis periods. He showed how the correlation of daily returns of the S&P500, with the previous two day’s returns, increased during the 2008 global financial crisis. During normal times, serial correlations between returns are small and only marginally statistically significant, but become relatively large and attain a high level of statistical significance during crisis periods.
Chapter 2

Method & Data

2.1 Method

2.1.1 Returns in Econometric Modelling

We will focus on returns to investors rather than prices when analysing the price data. This is more appropriate for several reasons. The returns can tell you at a glance if the investors made a profit or a loss on that day. If the return is negative, they made a loss, and if it is positive they made a profit. It also allows you to normalise the price series and therefore compare and analyse the relationships between different series, despite originating from different absolute values and even different currencies. Returns have the benefit of being unit free. Returns also remove the effects of inflation. As the daily return is calculated using that day’s price, and the day before’s price, inflation does not impact the prices in one day and so does not affect the return. Also, unlike prices, returns are only weakly correlated through time, making them easier to analyse. For these reasons, we convert our time series of raw prices into a series of returns. Academic finance literature generally uses the continuously compounded returns formula, with the return over a period being the difference in the natural logarithm of the starting and end price, as shown in equation 2.1

\[ r_t = \ln \left( \frac{p_t}{p_{t-1}} \right) \] (2.1)
The advantage of using the continuously compounded return, is that the returns are time additive. This means that returns for different days can be summed. For example, to obtain the weekly return, you can simply sum the 5 daily returns for the week. This means that the frequency of the returns does not matter and also that it encompasses all of the price movements for that period.

Shown below is how the continuously compounded return can be expanded into an infinite series to encompass the sum of the returns for \( n \) days.

\[
\ln \left( \frac{p_t}{p_{t-1}} \right) + \ln \left( \frac{p_{t-1}}{p_{t-2}} \right) + \ln \left( \frac{p_{t-2}}{p_{t-3}} \right) + \ldots + \ln \left( \frac{p_{t-(n-1)}}{p_{t-n}} \right)
\]

Using the logarithmic identity \( \log_a x + \log_a y = \log_a (xy) \):

\[
\ln \left( \frac{p_t}{p_{t-1}} \frac{p_{t-1}}{p_{t-2}} \frac{p_{t-2}}{p_{t-3}} \ldots \frac{p_{t-(n-1)}}{p_{t-n}} \right)
\]

Which simplifies to:

\[
\ln \left( \frac{p_t}{p_{t-n}} \right)
\]

**Figure 2.1: Time Series Of Ryanair Returns (2001-2012)**

Shown in figure 2.1 is a time series of the returns of Ryanair from December 2001 until October 2012, a period of 2778 trading days. It can be seen from the graph, that the returns vary substantially around their average level, which is close to zero. Another
important feature of returns shown in the graph is that there are periods of high variance in returns, and periods of low variance. These are known as periods of high volatility and low volatility and can be used to forecast the risk involved in investing in a stock at a particular time. You can see the period of high volatility during the 2008/2009 global financial crisis. This feature of returns means that you can produce very different results for your model depending on what period of returns you take. For this reason it is recommended that whenever possible, you analyse at least 8 years of returns (more than 2000 observations) in order to have in your model, periods of both high and low volatility. (Taylor: 11 [6])

There are certain statistical properties that are usually present with a set of returns taken over a number of years, known as stylised facts. The first important stylised fact is that the distribution of returns is not normal. Instead, it has a high peak, fat tails, and is approximately symmetric. This departure from the normal distribution can be shown using summary statistics such as the mean ($\bar{r}$), standard deviation ($s^2$), skewness ($b$) and kurtosis ($k$). These statistics are calculated as shown in figure 2.2.

$$\bar{r} = \frac{1}{n} \sum_{t=1}^{n} r_t, \quad s^2 = \frac{1}{n-1} \sum_{t=1}^{n} (r_t - \bar{r})^2,$$

$$b = \frac{1}{n-1} \sum_{t=1}^{n} \frac{(r_t - \bar{r})^3}{s^3}, \quad k = \frac{1}{n-1} \sum_{t=1}^{n} \frac{(r_t - \bar{r})^4}{s^4}.$$

**Figure 2.2: Summary Statistics, (Taylor: 52 [6])**

![Histogram of Ryanair Returns](image1.png)

**Figure 2.3: Distribution Of Ryanair Returns (2001-2012)**
Table 2.1: Distribution of Ryanair Returns Compared with the Normal Distribution

<table>
<thead>
<tr>
<th>Range</th>
<th>Observed Percentage in Range</th>
<th>Normal Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu \pm \sigma$</td>
<td>77.17%</td>
<td>68.27%</td>
</tr>
<tr>
<td>$\mu \pm 2\sigma$</td>
<td>95.25%</td>
<td>95.45%</td>
</tr>
<tr>
<td>$\mu \pm 3\sigma$</td>
<td>98.56%</td>
<td>99.73%</td>
</tr>
<tr>
<td>$\mu \pm 4\sigma$</td>
<td>99.68%</td>
<td>99.99%</td>
</tr>
<tr>
<td>$\mu \pm 5\sigma$</td>
<td>99.82%</td>
<td>$\approx$100%</td>
</tr>
<tr>
<td>$\mu \pm 9\sigma$</td>
<td>99.93%</td>
<td>$\approx$100%</td>
</tr>
<tr>
<td>$\mu \pm 13\sigma$</td>
<td>99.96%</td>
<td>$\approx$100%</td>
</tr>
<tr>
<td>$\mu \pm 14\sigma$</td>
<td>100%</td>
<td>$\approx$100%</td>
</tr>
</tbody>
</table>

Shown in figure 2.3 is a histogram showing the observed distribution of Ryanair’s returns. Shown also are blue vertical lines representing multiples of the standard deviation. The returns were found to have a mean of $1.06 \times 10^{-4}$ with a standard deviation of 0.026. The table below shows the percentage of Ryanair returns within multiples of the standard deviation, compared with the normal distribution.

Kurtosis is essentially a measure of the “peakedness” of the distribution compared to a normal distribution. For a perfectly normal distribution this is 3, but for this series of returns it was found to be 17.78. This shows that many of the returns are close to the mean, giving a high peak on the distribution. This can be seen in the histogram, and also in the table, where 77.71% of the returns lie within one standard deviation of the mean. Kurtosis is however, very sensitive to extreme values as it uses the fourth power of the value. The skewness statistic assesses the symmetry of the distribution. For this series it was found to be -0.92. This would suggest that the distribution is leaning towards the left. However, as skewness uses the third order of the observations, it is also sensitive to extreme observations. We can see from the histogram and the table that there is one extreme negative return that is 9 standard deviations from the mean, and another that is 13 standard deviations away. Theory would tell us that in a normal distribution, we should only see results that far from the mean once every $1.2 \times 10^{16}$ and $2.4 \times 10^{35}$ years respectively. As we observe results like this in a relatively short period of 11 years, it is clear that a normal distribution would not provide a good model for the magnitude or frequency of such extreme values. These two outliers also have a large impact on both the skewness and kurtosis values.
2.1.2 Econometric Models

The two econometric models used in this project were Autoregression (AR) models, and Vector Autoregression (VAR) models.

2.1.2.1 Autoregression (AR) Models

An AR model is used to describe a time series of returns based on its own previous values. It stipulates that the current value of the return depends linearly on its own previous values. An AR model of order $p$ is defined as follows:

$$ r_t = c + \alpha_1 r_{t-1} + \alpha_2 r_{t-2} + \ldots + \alpha_p r_{t-p} + \varepsilon_t $$  \hspace{1cm} (2.2)

where $\alpha_1 \ldots \alpha_p$ are the coefficients of the model, $c$ is a constant, and $\varepsilon_t$ is the error term. $\varepsilon_t$ is also known as the residual of the equation and is assumed to be normally distributed white noise. It represents the value of $r_t$ that is not accounted for by the equation. The model is sensitive to the model order chosen, or the number of lags included. The method for choosing the model order is done by calculating the Hannan-Quinn Information Criterion (HQIC) over a range of model orders, and choosing the order that minimises this criterion. In brief, this criterion is the sum of two terms, one that characterises prediction error of the model, and a second term that characterises the number of freely estimated parameters in the model (which increases with increasing model order). The HQIC is calculated as follows:

$$ HQIC = n \ln \left( \frac{SSR}{n} \right) + 2k \ln (\ln (n)) $$  \hspace{1cm} (2.3)

where $n$ is the number of observations in the sample period, $k$ is the number of variables in the model, and $SSR$ is the Sum of the Squared Residuals. Each time the estimated equation is fitted to a set of returns over the sample period, there is an error term, or residual. $SSR$ is the sum of each of these residuals squared. For all autoregressive models, $k$ will be one. This will change when we start to use VAR models, as described in the next section, section 2.1.2.2.
2.1.2.2 Vector Autoregression (VAR) Models

A VAR model produces an equation, where the current value of the dependent variable depends on past movements in that variable, and in the past movements of all other variables in the system. It is a generalisation of the AR model to allow multiple variables to impact the dependent variable. In this case, the dependent variable will be the returns of the chosen airline stock, and the other variables will be the returns of the crude oil series. A VAR model of order \( p \) can be expressed in the following form:

\[
r_t = A_1 Z_{t-1} + A_2 Z_{t-2} + \ldots + A_p Z_{t-p} + \varepsilon_t
\]

(2.4)

where \( Z_t \) is a vector of variables at time \( t \), \( A_1 \ldots A_p \) are vectors of coefficients, \( p \) is the number of lags in the model or the model order, and \( \varepsilon_t \) is a vector of error terms or residuals.

Estimating a VAR model involves choosing which variables to include and how many lags to include. The resulting model can be sensitive to both of these choices. The variables included are generally decided by economic theory, in our case, crude oil prices impacting airline stock prices. The method for choosing the number of lags to include in the model is the same as with AR models.

2.1.2.3 How to Compare Models

After creating a number of models, you have a large selection of statistical criteria with which to compare them. One of the key measures of the correctness of the model is the \( R^2 \) value. This is essentially a goodness-of-fit measure and tells us what portion of the variance of the dependent variable is explained by the model. It is calculated as follows:

\[
R^2 = 1 - \frac{SSR}{TSS}
\]

(2.5)

Where \( SSR \) is the Sum of the Squared Residuals, and \( TSS \) is the total sum sum of squares for the dependent variable. However this can yield inaccurate results the more variables you add to the model. By adding more variables to the model you will often increase the \( R^2 \) value without it actually being a better model. The adjusted \( R^2 \) value, denoted as \( \bar{R}^2 \), takes the number of variables into account and thus gives you more
Chapter 2. Method & Data

accurate results.

\[ R^2 = 1 - \frac{SSR (n - k)}{TSS (n - 1)} \]  \hspace{1cm} (2.6)

where \( n \) is the number of observations in the sample period, and \( k \) is the number of variables in the model.

Another key metric used to compare the models is the root mean squared (RMS) error, or standard error. It is calculated as follows:

\[ RMS\ Error = \sqrt{\frac{SSR}{n}} \]  \hspace{1cm} (2.7)

When comparing models that have the same dependent variable and the same sample period, the RMS error goes down as adjusted R-squared goes up. Hence, the model with the highest adjusted R-squared will have the lowest root mean squared error, and you can use either of these statistics to determine the best model. However, when comparing models with different estimation periods, it is more reliable to look at the RMS error as a guide to quality.

In examining the model we look at not just the size of the coefficients, but the statistical significance of each coefficient. This will tell us what the most statistically significant impacts on today's return are.

2.1.2.4 Forecasting

Forecasting is one of the main objectives of financial econometric models. Forecasting from an AR model and a VAR model is quite similar and straightforward once the model has been estimated. Use \( t \) to refer to the first period for which data is not yet available and substitute the known prior values into the equation. The error term \( \varepsilon_t \) is set to zero. As it is assumed to be normally distributed white noise, this is its expected value. The result of the equation is the one-step-ahead forecast. This result can then be used to produce the two-step-ahead forecast and so forth for an arbitrary number of days.

There are four sources of uncertainty surrounding these forecasts. Firstly, uncertainty as to whether the model chosen is the correct model. Secondly, uncertainty about the accuracy of the model coefficients. Thirdly, uncertainty about the value of the error term \( \varepsilon_t \) for the period being forecast. Finally, for forecasts more than one step ahead,
uncertainty about the accuracy of the forecasted values that are used as lagged values in the right side of the equation. The last three sources can be quantified and combined to give a confidence interval for the forecasts. The confidence interval will become wider as the number of steps ahead increases, because of the use of an increasing number of estimated values for the right-side variables.

2.1.3 APIs

As part of the financial advisory system, COMHAIRLE, I designed and wrote two APIs. One for collecting the relevant data from online sources, and the other for interacting with GRETL’s library. The software architecture diagram is shown in figure 2.4.

![Figure 2.4: COMHAIRLE’s Architecture Diagram](image)

2.1.3.1 Data-Collection API

This was used to collect the price data from different online sources. The system downloads the historical price data for the relevant assets for the selected time period using COMHAIRLE’s simple graphical user interface. The data is downloaded to a comma separated value text file (.csv file), which COMHAIRLE then extracts the price series
from. Quandl.com has built in API calls to download the selected time series to a choice of formats. Yahoo finance only provides csv files however so this is the format I used.

2.1.3.2 GRETL API

GRETL is an open source Gnu, Regression, Econometrics and Time series Library. Its easy to use graphical interface is its main attraction but it also provides a command line interface. In order to access this program from COMHAIRLE, I wrote GRETL scripts to run different actions in the program automatically, which could be called from COMHAIRLE. This was used to create all of the AR/VAR models. There were a number of steps involved in this process. First of all, the relevant time series must be loaded into GRETL’s workspace. As GRETL scripts only allow for one scalar argument to be passed into the script from the command line, this meant that I couldn’t simply pass in the relevant .csv file names. To get around this I used text files as the method of passing data between the two applications. Having downloaded the price series to csv files, COMHAIRLE then prints the names and locations of these files to a text file. COMHAIRLE then calls a GRETL script that reads from this text file, and imports the csv files listed here. Now that the relevant time series are in GRETL’s workspace, it calculates the HQIC for lags 1 to 10 inclusive and prints the results to a text file. COMHAIRLE then parses this file and chooses the lag with the minimum HQIC as the optimal lag selection. COMHAIRLE then calls another GRETL script to create the AR/VAR model, with the optimal lag number passed to the script as the scalar argument. GRETL then prints the results of the model to a text file which COMHAIRLE then reads to get the relevant information and display it in a readable fashion for the user.

2.2 Data

Any empirical investigation of the behaviour of asset prices through time requires price data. The historical daily price data for almost any publicly traded company or commodity can be obtained from many websites and organisations. The company price data was obtained from Yahoo Finance (http://finance.yahoo.com) as this provides the
adjusted daily closing price. The price is adjusted to incorporate the effect of cash dividends, stock dividends and stock splits. If we were using unadjusted prices we would have to adjust the returns calculation formula for dividends and splits. By providing the adjusted prices, Yahoo Finance makes that job easy for us. The oil prices were obtained through www.quandl.com, which obtains its data from various sources such as the US Department of the Environment, and the US Energy Information Administration.

However, the price series are not ideal discrete time series, as the intervals between observations is not uniform. Stock markets and commodity markets are closed on weekends, on public holidays and under certain exceptional circumstances. There is also the issue that free data providers such as Yahoo and Quandl tend to not be as reliable as paid data providers. They do make mistakes and often revise the series to correct these mistakes. They also don’t always update the series immediately when new data is available. However, when using these data providers, COMHAIRLE always uses the most up-to-date version of the series.

2.2.1 Airline Company Data

Ryanair is the largest Irish airline and has its primary stock listing is on the Irish Stock Exchange. The best available free historical data from this listing is on Yahoo Finance. However this data series only covers the period of the 12/12/2001 to 15/10/2012, meaning that the returns for the last 16 months are missing. This is unfortunate as current data is necessary to make useful forecasts. This is unfortunate as current data is necessary to make useful forecasts. However it is the best free dataset available. This sample period comprises 2831 weekdays. There are 2758 observations in the dataset, meaning there are 73 missing daily observations. Shown in figure 2.5 is the time series of the unadjusted daily closing prices for this period. This is the price of the stock at the end of trading each day. It does not take into account stock dividends, cash dividends or stock splits. This makes the data unsuitable for analysis. For example on the 27th of February 2007, the price of the stock appears to plummet by 50%. However what happened on that day was a 2:1 stock split. At the time, Michael O’Leary decided that traders had an aversion to buying shares that were above €10 in price, so in an attempt to improve the liquidity of the stock, he ordered a 2:1 stock split. In effect, each €12.35 share became two €6.175 shares. By doubling the number of shares held by each investor and halving the price, this has no effect on the value of the company.
In order to accurately represent the stock price we must take these changes into account. Fortunately, Yahoo finance provides the adjusted daily closing prices, which takes these changes into account. A graph of the adjusted closing prices is shown in figure 2.6 for the same series.

Easyjet is the largest UK airline and has its primary stock listing on the London Stock Exchange. The historical prices are available from Yahoo Finance for the period of 16/11/2000 to present. This dataset has 3475 daily observations with 9 days missing. This dataset is much better than the Ryanair dataset as it is up-to-date and has much fewer missing days.
Aer Lingus is an Irish airline with its primary stock listing on the Irish Stock Exchange. The Irish state had an 85% share in the company until it was floated on the Irish and London stock exchanges in 2006. Currently, the Irish state has a 25% share and Ryanair has a 29% share in the company. Because the company is partly state owned, it is more shielded from turbulence in the markets than other airlines. The historical prices are available from Yahoo Finance from when it was first floated on 28/09/2006, to present. This dataset contains 1936 daily observations, with 12 observations missing.

2.2.2 Crude Oil Data

The oil prices are obtained through the data provider Quandl.com, which sources its data sets from various online sources. Unlike the company data, these prices are unadjusted. This is fine however as these are commodity prices and therefore no dividends or stock splits need to be accounted for. There are three different oil price series included in COMHAIRLE at the moment. These are the crude oil spot price, the crude oil futures price, and the spot price of kerosene.

The crude oil spot price used is the WTI price per barrel. WTI stands for West Texas Intermediary. It is a grade of crude oil used as a benchmark in pricing. The historical prices of this are provided by the US Department of Energy going back to 1986. The sample period that was used here was from 05/01/2000 to its most recent observation at the time of writing (17/03/2014). This dataset contains 3569 daily observations with 140 missing daily observations.

The crude oil futures price is a continuous one month contract, called a CL1 contract. It is calculated by the Open Financial Data Project (www.ofpd.org) and obtained through Quandl.com. A one month oil futures contract is where you agree now on the price you will pay for the oil, which you will receive in one months time. It essentially involves betting on what the price will be in one months time, and is therefore a projection of the spot price based on investor sentiment. The historical prices are provided as far back as 1983 but for our calculations we used a sample period from 05/01/2001 to present (26/03/2014). This contains 3565 daily observations with 146 missing days.

Kerosene is refined from crude oil for use as jet fuel. The kerosene price data is the spot price of U.S. Gulf Coast Kerosene-Type Jet Fuel, provided by the US Department of
Energy. It is obtained through the data provider Quandl.com. The sample period used for our calculations was from 05/01/2000 to 24/03/2014. This period contains 3568 daily observations with 141 days missing from the series.
Chapter 3

Prototypes and Experiments

In this chapter I will detail the current COMHAILRE prototype, and the experiments I performed using it.

3.1 COMHAILRE’s User Interface

![COMHAILRE’s Opening Screen](image)

Figure 3.1: COMHAILRE’s Opening Screen

The opening screen of the program is shown in figure 3.1. The user selects the company to analyse from a list containing Ryanair, Aerlingus, and EasyJet. The user then selects the observation period over which to analyse the company by entering the start and end
date. If the ‘Predict Returns’ checkbox is ticked, the series is analysed up to today’s date, and forecasts are enabled for an arbitrary number of days ahead. The user then selects from the different oil price series to include in the model. There is an option to standardise the returns to their z-scores, however this option is not available when the ‘Predict Returns’ checkbox is ticked. This is because the forecasted return is not useful when it is in z-score format. When the screen has been filled out correctly and the calculate button is clicked, COMHAIRLE runs with the selected parameters and displays its results. If the screen is not filled out correctly, i.e. incorrect date format or no company selected, an error message is shown when calculate is clicked, telling the user what is wrong so they can fix it.

The results screen of the program then pops up containing a number of different tabs. The first tab, ‘Distribution of Returns’ details the distribution of the returns. It has two tabs within it, the first of which is the ‘Histogram of Returns’ tab. This is shown in figure 3.2. Here we can see the relative frequency of the distribution. Note also the vertical blue lines showing multiples of the standard deviation.

![Histogram of Returns Tab](image)

**Figure 3.2: Histogram of Returns Tab**

The other tab within the ‘Distribution of Returns’ tab is the ‘Summary Statistics’ tab, as shown in figure 3.3. The top table on this screen tabulates the summary statistics of the returns. This includes the mean, standard deviation, skewness, kurtosis, max return
and date it occurred on, and the min return and the date it occurred on. The lower table details the observed distribution and compares it to the normal distribution.

![Summary Statistics Tab](image)

**Figure 3.3: Summary Statistics Tab**

The next tab is the ‘Vector Autoregression Model’ tab, as shown in figure 3.4. This shows us the VAR model produced by COMHAIRLE. In this case we have not added in any of the oil price series to the model, so it is an Autoregressive (AR) model. The table here shows us each of the lags in our model, their coefficients, and their statistical significance. There is a check box above the table, ‘Rank Lags by Significance’. This orders the lags in the table based on their statistical significance, allowing the user to see the dominant impacts on the return. Above the table we also see the key statistics to evaluate the model, the adjusted R-squared and the RMS error. We also see the HQIC used to select the number of lags in the model.

The third and final tab, ‘Prediction of Returns’ is only visible when ‘Predict Returns’ has been selected on the opening screen. This tab is shown in figure 3.5. Above the table the user can input the number of days ahead to forecast. The table then shows the selected number of day’s forecasts. This includes the predicted return, the standard error of the prediction, and the 95% confidence interval. As you can see the error becomes larger the more days ahead you forecast.
3.2 Comparing & Evaluating Models

For each of our three airlines, Aer Lingus, EasyJet, and Ryanair, we will first examine the distribution of their returns. We will then create and evaluate multiple different
models for each airline. Firstly we will look at how well an AR model describes the behaviour of the returns. We then look at how this can be improved by including oil prices in the model and creating a VAR model. Then we choose our best model and look at the forecasts that can be made using it.

3.2.1 Comparison of Returns

The returns of the three airlines are summarised in table 3.1. We can see that EasyJet has the highest average daily return at 0.000453. It has a larger standard deviation and smaller kurtosis than Ryanair, suggesting that there is a larger spread in it’s returns.

3.2.2 EasyJet

Firstly we examine our AR model for EasyJet. The AR equation for this model is as follows:

$$ r_t = 0.028r_{t-1} - 0.047r_{t-2} + \epsilon_t $$

The significance of the first and second lags is 90% and 95% respectively. This tells us that the second lag of returns is the dominant impact on returns. The adjusted R squared value for this model is 0.0024 telling us that 0.24%, or 24 basis points, of the variance in $r_t$ is explained by this model.
We then look at how we can improve on this by adding in the oil prices to the model. Shown in table 3.2 are the evaluation statistics for each combination of the VAR model. We can see that the AR model, with no external impacts assumed, accounts for by far the least amount of variance in the returns of EasyJet at 24 basis points. The model that accounts for the most variance includes the impact of all three crude oil price series. This model has an adjusted R squared value of 0.0088 meaning that 88 basis points of the variance in $r_t$ is explained by this model. The optimal model order chosen for this model was five lags. Its equation is displayed in $COMHAIRLE$ as shown in figure 3.6. The table is ordered to show the most statistically significant lags at the top. We can see here that there are three lags with a 90% significance; the first and second lags of EasyJet itself, and the 5th lag of WTI returns. Choosing this as our best model, we will look at the forecasts that can be made from it. To evaluate how well this model will forecast returns, we can apply it to past movements and compare the forecast to the actual return. Shown in figure 3.7 is the time series of the entire sample period, showing the forecasted return (blue), the 95% confidence interval of the forecast (green), and the actual return (red). We can see from this that during periods of high variance, particularly in 2008/2009, the forecast also showed accordingly, a higher variance. $COMHAIRLE$’s 5-day forecast is shown graphically in
Chapter 3. Prototypes and Experiments

Figure 3.7: Past Prediction of EasyJet Returns

Figure 3.8: 5 Day Forecast of EasyJet Returns

The model can also be used to predict the reaction of EasyJet’s return to a major shock in any of the variables in the model. In figure 3.9 we graph the 15 day response of EasyJet’s return to a shock in the return of 1. EasyJet itself, 2. WTI, 3. CL1, and 4. Kerosene. The grey area is the 95% confidence interval.
Chapter 3. Prototypes and Experiments

Figure 3.9: Response of EasyJet to shocks in 1.EasyJet, 2.WTI, 3.CL1, 4.Kerosene

<table>
<thead>
<tr>
<th>External Impacts</th>
<th>Basis Points Variance Explained</th>
<th>Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>21</td>
<td>0.0258</td>
</tr>
<tr>
<td>WTI</td>
<td>99</td>
<td>0.0259</td>
</tr>
<tr>
<td>CL1</td>
<td>68</td>
<td>0.0260</td>
</tr>
<tr>
<td>Kerosene</td>
<td>32</td>
<td>0.0260</td>
</tr>
<tr>
<td>WTI &amp; CL1</td>
<td>123</td>
<td>0.0259</td>
</tr>
<tr>
<td>WTI &amp; Kerosene</td>
<td>97</td>
<td>0.0258</td>
</tr>
<tr>
<td>CL1 &amp; Kerosene</td>
<td>65</td>
<td>0.0260</td>
</tr>
<tr>
<td>WTI, CL1 &amp; Kerosene</td>
<td>138</td>
<td>0.0259</td>
</tr>
</tbody>
</table>

Table 3.3: Comparison of Ryanair Models

3.2.3 Ryanair

The AR model produced for Ryanair was as follows:

\[ r_t = -0.022r_{t-1} - 0.049r_{t-2} + \epsilon_t \]  \hspace{1cm} (3.1)

We can see here that the model order chosen was two. The significance of the first and second lags are 75% and 98% respectively. In the same way as the EasyJet model, the second lag of the returns is the dominant impact on the return. The $\bar{R}^2$ value of the model tells us that it accounts for 21 basis points in the variance of Ryanair’s returns.

Shown in table 3.3 are the evaluation statistics for each combination of the the VAR model. We can see that the AR model, accounts for the least amount of variance in the return at 21 basis points. The model that accounts for the most variance includes the returns of WTI, CL1, and Kerosene. This model explains 138 basis points of the variance. The model order chosen for this model was four lags, and the equation is displayed by $COMHAIRLE$ as shown in figure 3.10:

The first WTI lag has the biggest impact and has a significance value of 99.99%. There are four other lags above 95% significance; the 2nd, 3rd and 4th lags of Ryanair itself, and the 2nd lag of Kerosene. Taking this as our best model, we look at the forecasts
that can be made from it. We apply our model to the past movements in price and compare the forecast to the actual return. This is shown in figure 3.11.

![Parameter Estimates Table]

**Figure 3.10:** Ryanair VAR Model

Shown in figure 3.12 is a graph of COMHAIRLE’s forecast for 5 days, with the green stars being the out of sample predictions, and their 95% confidence interval error bars. The previous 100 days are also shown, with the forecasted value in blue, and the actual return in red.

![Forecast Graph]

**Figure 3.11:** Past Forecasts of Ryanair Returns
Next we look at how the model can forecast the reaction of Ryanair’s return to a shock in any of the variables in the model. In figure 3.13 we graph the 15 day response of Ryanair returns to a shock in the return of 1. Ryanair itself, 2. WTI, 3. CL1, and 4. Kerosene. The grey area is the 95% confidence interval.

<table>
<thead>
<tr>
<th>External Impacts</th>
<th>Basis Points Variance Explained</th>
<th>Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>16</td>
<td>0.0286</td>
</tr>
<tr>
<td>WTI</td>
<td>22</td>
<td>0.0289</td>
</tr>
<tr>
<td>CL1</td>
<td>42</td>
<td>0.0288</td>
</tr>
<tr>
<td>Kerosene</td>
<td>7</td>
<td>0.0288</td>
</tr>
<tr>
<td>WTI &amp; CL1</td>
<td>62</td>
<td>0.0288</td>
</tr>
<tr>
<td>WTI &amp; Kerosene</td>
<td>91</td>
<td>0.0287</td>
</tr>
<tr>
<td>CL1 &amp; Kerosene</td>
<td>95</td>
<td>0.0286</td>
</tr>
<tr>
<td>WTI, CL1 &amp; Kerosene</td>
<td>87</td>
<td>0.0287</td>
</tr>
</tbody>
</table>

Table 3.4: Comparison of Aer Lingus Models
3.2.4 Aer Lingus

The AR model for Aer Lingus was estimated as follows:

$$ r_t = -0.046r_{t-1} + \varepsilon_t $$  \hspace{1cm} (3.2)

The optimal model order chosen was one. This model explains 16 basis points of the variance in Aer Lingus returns. We can see in table 3.4 how this can be improved by adding in the oil variables to the equation. The model that explains the most variance includes CL1 and Kerosene, but not WTI. This model accounts for 95 basis points of the variance in Aer Lingus returns. The model order chosen for this model was 4, and its details are shown in figure 3.14.

![Image](image.png)

**Figure 3.14: Aer Lingus VAR Model**

There are five lags with a 90% plus significance level. These are the 1st and 4th lags of Aerlingus itself, and the 1st, 2nd and 3rd lags of CL1. Choosing this model as our best model, we then examine the forecasts we can make from it. Shown in figure 3.15 is the forecasts made using the past values. The 5 day ahead forecast is shown in figure 3.16.

In figure 3.17 we graph the 15 day response of Aer Lingus returns to a shock in the return of 1. Aer Lingus itself, 2. WTI, 3. CL1, and 4. Kerosene. The grey area is the 95% confidence interval.
3.2.5 Colinear Interdependence Between Oil Variables

In this section, we discuss the suitability of including more than one oil variable in our model due to the collinear interdependence between these three variables. The spot price (WTI) and futures price (CL1) of crude oil are very heavily correlated. The futures prices is a projection of the spot price, one month into the future and is based on investor sentiment. We examine the relationship between the these two variables from
01/01/2000 to 01/03/2014. The persons correlation coefficient between the two variables is 0.87 showing a strong correlation between them. Figure 3.18 shows a scatterplot of two sets of returns. We can see a noticeable linear relationship between the two returns. By performing a simple linear regression between the two time series we obtain the following equation showing the relationship between the two variables.

\[ CL1(t) = -4.6 \times 10^{-5} + 0.88WTI(t) \] (3.3)

This equation has an \( R^2 \) value of 0.75, meaning that it explains 75% of the variance in CL1 returns.

Kerosene is a product of crude oil and therefore its price is heavily dependent of the price of crude oil. We can see the linear relationship between Kerosene returns and WTI returns in figure 3.19. The Pearson's correlation coefficient for these two variables is 0.62 showing a less strong linear relationship than between CL1 and WTI, but it is significant none the less.

\[ Kerosene(t) = 1.2 \times 10^{-4} + 0.66WTI(t) \] (3.4)

This equation has an \( R^2 \) value of 0.39.

We also examine the relationship between Kerosene and the futures price of crude oil,
A scatterplot of these two variables is shown in figure 3.20. The Pearson correlation coefficient of this relationship is 0.62, the same as Kerosene and WTI. The equation resulting from performing a simple linear regression of the two variables is shown in equation 3.5

\[ Kerosene(t) = 2.1 \times 10^{-4} + 0.65CL1(t) \]  

This equation has an \( R^2 \) value of 0.38.

We can see from these three relationships that there is a considerable amount of interdependence between the three oil variables. By including more than one of the in the
Chapter 3. Prototypes and Experiments

### Table 3.5: Correlation Between Oil Variables

<table>
<thead>
<tr>
<th></th>
<th>Pearsons Correlation Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>CL1 vs WTI</td>
<td>0.87</td>
</tr>
<tr>
<td>Kerosene vs WTI</td>
<td>0.62</td>
</tr>
<tr>
<td>Kerosene vs CL1</td>
<td>0.62</td>
</tr>
</tbody>
</table>

VAR models we are including some of the same information twice. However there is different information present in each series as they are not completely correlated. The respective Pearson correlations between the variables are shown in table 3.5.
Chapter 4

Conclusion

This was a study on how to use APIs to curate financial data, and to perform simulation and modelling. I have created a java based system to curate data from external sources. It uses API calls to leading world data authorities, and also uses well known public domain econometric models. Typically, when financial time series analysis is conducted, end users are expected to collect the data, transform it to a format suitable for analysis, ensure it is in the correct format, and then that have to find a model and run the model. Doing all of this manually can produce errors in the data, as well as the inconvenience involved. What I have done is to control access, so the user just specifies the company and oil series, and the dates over which to analyse them, without worrying about data formats, copy and paste errors, or learning how to run the model.

I used to the system to create models for three airline stocks, taking into account the impact of crude oil prices. Although the models explain a small portion of the variance in the airline returns, this can serve as useful advice, particularly in response to market shocks such as a spike in oil prices. The general wisdom is that historical price data cannot be relied upon to predict the future with great certainty, though may be useful as a appraisal of an investment decision. Asset returns are only weakly correlated through time, and thus, are difficult to forecast. However these correlations, however weak, can be useful as a basic estimate.
Bibliography


