APPLICATION OF TIME SERIES ANALYSIS AND STATISTICAL PROCESS CONTROL TOOLS TO FINANCIAL DATA

Mohammed Siyam Ibrahim

Imran Mohammed

Mark Venneri

A thesis submitted in partial fulfilment

of the requirements for the degree of

BACHELOR OF APPLIED SCIENCE

Supervisor: V. Makis

Department of Mechanical and Industrial Engineering

ABSTRACT

The application of time series analysis to financial data continues to be an area of interest to day-traders, quantitative finance specialists and select investment professionals. By successfully combining classical time series analysis with Statistical Process Control (SPC) tools, the authors propose a highly capable technical analysis system which models and monitors market performance of a variety of financial instruments. In addition, the paper develops a 'data pooling' technique to generalize the standard Yule-Walker equations for autoregressive coefficients; applies the Box-Jenkins' methodology of model identification, estimation and validation to generate ARIMA models based on multiple non-sequential histories of stock data; and compares the accuracy of pooled-history models to that of conventional single-history models using graphical residual analysis techniques. The results indicate that while pooled-history models are generally as adequate as the single-history models, the former are particularly well-suited for situations involving multiple short run-length histories which are often observed in financial time series. Using CUSUM control charts to monitor model residuals provided a valuable proof-of-concept that validated the use of time series analysis in conjunction with SPC tools in modeling and monitoring the behaviour of financial instrument.

ACKNOWLEDGEMENT

This undergraduate thesis could not have been completed without Dr. Viliam Makis who not only supervised the thesis but also provided significant direction and input throughout the project. For his invaluable time and continued efforts, the authors are grateful. The authors would also like to acknowledge Michael Kim who provided guidance in the form of mathematical resources and constructive criticism.

March 26, 2009

1. INTRODUCTION 9

2. OBJECTIVES 10

3. STOCK AND DATA SELECTION 12

3.1. COMPANY BACKGROUND: YAMANA GOLD CORP. 12

3.2. THE DATA SELECTION PROCESS 13

3.3. DATASETS 14

4. OVERVIEW OF TIME SERIES ANALYSIS AND STATISTICAL PROCESS CONTROL 16

4.1. INTRODUCTION TO TIME SERIES ANALYSIS 16

4.2. DEVELOPING AN AUTOREGRESSIVE AR(P) MODEL 17

4.3. DEVELOPING A MIXED AUTOREGRESSIVE, MOVING AVERAGE MODEL 20

4.4. MODIFYING THE STANDARD *AR(P)* EQUATIONS TO ALLOW FOR POOLING 21

4.5. INTRODUCTION TO STATISTICAL PROCESS CONTROL (SPC) 21

4.6. SHEWHART CONTROL CHARTS 22

4.7. TABULAR (ALGORITHMIC) CUSUM 23

5. OVERVIEW OF THE BOX-JENKINS APPROACH 28

5.1. MODEL IDENTIFICATION 28

Run Sequence Plot 28

Autocorrelation Plot 29

Partial Autocorrelation Plot 31

5.2. MODEL ESTIMATION 32

Direct Inversion 33

Yule Walker Equations 35

Using Yule-Walker Equations To Compute PACF 37

5.3. MODEL VALIDATION 38

Run Sequence Plot 39

Lag Plot *40*

Histogram *40*

Normal Probability Plot *41*

6. APPLYING THE BOX-JENKINS METHODOLOGY 42

6.1. MODEL IDENTIFICATION 42

Run Sequence Plots of Pooled Data *42*

ACF and PACF plots of raw data *44*

Run Sequence Plots of Second Order Differenced Datasets *45*

ACF and PACF Plots of Second Order Differenced Datasets *46*

6.2. MODEL ESTIMATION 48

6.3. MODEL VALIDATION 48

Increasing Model Residual Analysis *49*

Decreasing Model Residual Analysis *50*

Analysis of the 4-plot for the residuals of the Fluctuating Model *50*

6.4. COMPARING POOLED AND SINGLE HISTORY ARIMA MODELS 51

Comparison of Increasing Models *52*

Comparison of Decreasing Models *53*

Comparison of Horizontal Models *54*

6.5. STATISTICAL PROCESS CONTROL (SPC) RESULTS 55

Shewhart's I-MR Charts *56*

CUSUM Analysis *58*

7. THESIS SUMMARY 62

7.1. REVIEW OF BOX-JENKINS METHODOLOGY AND SPC 62

7.2. INTERPRETATION OF CUSUM CONTROL CHART SIGNALS 63

7.3. MODEL LIMITATIONS 65

7.4. REAL WORLD APPLICATIONS 66

8. LITERATURE REVIEW 67

8.1. ESTIMATING STOCK MARKET VOLATILITY USING ASYMMETRIC GARCH MODELS 67

Executive Summary 67

Introduction 67

Data and Computation 69

Model Specification 69

Assessing Volatility Models 70

Forecasting 71

Working Example 71

Conclusion 77

8.2. PREDICTING CORPORATE FINANCIAL DISTRESS: A TIME-SERIES CUSUM METHODOLOGY 78

Executive Summary 78

Motivation for Time Series CUSUM Model 78

Statistical Process Control 79

Comparing Shewhart's and CUSUM control charts 80

CUSUM Example - B 89

CUSUM Example - C 90

Conclusions 91

8.3. A GARCH FORECASTING MODEL TO PREDICT DAY-AHEAD ELECTRICITY PRICES 93

Executive Summary 93

Motivations for using GARCH 93

Model Development 94

Model Validation and Application 96

A GARCH(1, 1) Example 98

Conclusions 102

9. HISTORY OF THE TSX **103**

9.1. THE PIONEERS OF TRADING 103

9.2. THE HISTORY OF THE DOW JONES 104

9.3. THE HISTORY AND EVOLUTION OF THE TORONTO STOCK EXCHANGE 106

9.4. HISTORY OF STANDARD & POOR'S 108

9.5. WHAT IS THE S&P/TSX 60 109

10. FINANCIAL MARKETS AND MARKET INDICATORS **111**

10.1. INDICATORS OF INFLATION AND INTEREST RATES 111

The price of gold *112*

Prime Rate *112*

Gross Domestic Product *112*

Employment Situation: Payroll Employment *112*

Employment Situation: Unemployment Rate *113*

Consumer Credit *113*

Housing Starts *113*

11. BIBLIOGRAPHY **114**

12. APPENDIX **115**

LIST OF TABLES

TABLE 3-1 KEY FINANCIAL, STOCK PERFORMANCE AND COMPANY RATIOS FOR YAMANA CORP. ADAPTED FROM GOOGLE FINANCE (MARCH 15, 2009). 13

TABLE 3-2 DATA SETS USED TO DEVELOP UNTREND, DOWNTREND AND FLUCTUATING MODELS 14

TABLE 4-1 STANDARD AND MODIFIED EQUATIONS FOR THE AR(P) MODEL 21

TABLE 5-1 BEHAVIOURS OF ACF PLOTS AND RECOMMENDED MODELS 31

TABLE 5-2 SUMMARY OF 4-PLOT COMPONENTS 39

TABLE 6-1 PARAMETER ESTIMATES FOR 'INCREASING', 'DECREASING' AND 'FLUCTUATING' DATASETS 48

TABLE 6-2 ARIMA EQUATIONS RECAST INTO ORIGINAL TIME SERIES 48

TABLE 6-3 SINGLE AND POOLED HISTORY MODEL PARAMETERS FOR 'INCREASING', 'DECREASING' AND 'FLUCTUATING' DATASETS 52

TABLE 6-4 ESTIMATES OF THE SAMPLE STANDARD DEVIATION AS CALCULATED BY MINITAB 57

TABLE 6-5 CUSUM DESIGN PARAMETERS AS OUTPUTTED BY MATLAB 58

TABLE 8-1 QUANTITATIVE FORECASTING PERFORMANCE MEASURES 69

TABLE 8-2 ARCH TEST MATLAB OUTPUT (PRE-ESTIMATION ANALYSIS) 73

TABLE 8-3 MATLAB OUTPUT FOR GARCH FITTING TO NASDAQ RETURNS 75

TABLE 8-4 ARCH TEST MATLAB (POST-ESTIMATION ANALYSIS) 77

LIST OF FIGURES

FIGURE 3-1 STOCK CHART OF YAMANA GOLD INC., (TSE: YRI) 13

FIGURE 3-2 DATA SETS SELECTED FOR POOLING, ILLUSTRATED BY THE STOCK CHART 14

FIGURE 5-1 RUN PLOT OF ALL HISTORIES CONTAINED WITHIN THE 'HORIZONTAL' DATASET 29

FIGURE 5-2 AUTOCORRELATION PLOT OF DECREASING DATASET 31

FIGURE 5-3 DOUBLE DIFFERENCED ACF PLOT OF DECREASING DATASET 31

FIGURE 5-4 PACF PLOT OF 'INCREASING' POOLED DATA, SECOND-ORDER DIFFERENCED 32

FIGURE 5-5 RUN SEQUENCE PLOT OF INCREASING DATASET 39

FIGURE 5-6 RUN SEQUENCE PLOT OF HISTORY 2 - DOUBLE DIFFERENCED 40

FIGURE 5-7 LAG PLOT OF THE DECREASING ARIMA(1,2,0) RESIDUALS 40

FIGURE 5-8 HISTOGRAM OF THE FLUCTUATE ARIMA(2,2,0) RESIDUALS 41

FIGURE 5-9 NORMAL PROBABILITY PLOT OF FLUCTUATE ARIMA(2,2,0) RESIDUALS 41

FIGURE 6-1 RUN SEQUENCE PLOTS FOR THE 'INCREASING', 'DECREASING' AND 'FLUCTUATING' DATASETS 43

FIGURE 6-2 ACF AND PACF PLOTS FOR UNDIFFERENCED DATASETS 44

FIGURE 6-3 RUN SEQUENCE PLOT OF SECOND ORDER DIFFERENCED DATASETS 45

FIGURE 6-4 ACF AND PACF PLOTS OF SECOND ORDER DIFFERENED DATASETS 47

FIGURE 6-5 4-PLOT OF THE RESIDUALS OF THE INCREASING ARIMA MODEL 49

FIGURE 6-6 4-PLOT OF THE RESIDUALS OF THE DECREASING ARIMA MODEL 50

FIGURE 6-7 4-PLOT OF THE RESIDUALS OF THE FLUCTUATING ARIMA MODEL 51

FIGURE 6-8 TIME SERIES USED FOR ONE-STEP AHEAD FORECASTS 52

FIGURE 6-9 NORMALITY PLOTS OF THE POOLED & SINGLE HISTORY MODELS (INCREASING) 53

FIGURE 6-10 NORMALITY PLOTS OF THE POOLED & SINGLE HISTORY MODELS (DECREASING) 54

FIGURE 6-11 NORMALITY PLOT OF THE SINGLE AND POOLED HISTORY MODELS (HORIZONTAL) 55

FIGURE 6-12 I-MR CHARTS OF THE RESIDUALS OF THE MODELS 56

FIGURE 6-13 MATLAB OUTPUT OF 'IN-CONTROL' CUSUM CHART (SLIGHTLY INCREASING) 59

FIGURE 6-14 MATLAB OUTPUT OF 'OUT-OF'CONTROL' CUSUM CHART (STEEPLY INCREASING) 60

FIGURE 6-15 MATLAB OUTPUT OF 'IN-CONTROL' CUSUM CHART (SLIGHLY DECREASING) 60

FIGURE 6-16 MATLAB OUTPUT OF 'OUT-OF-CONTROL' CUSUM CHART (STEEPLY DECREASING) 60

FIGURE 6-17 MATLAT OUTPUT OF AN 'OUT-OF-CONTROL' CUSUM CHART (HORIZONTAL) 61

FIGURE 7-1 GRAPHICAL OVERVIEW OF THE BOX-JENKINS METHODOLOGY 62

FIGURE 7-2 WORKSTATION CONFIGURATION OF MODELING AND MONITORING SYSTEM 66

FIGURE 8-1 FORECASTING ANALYSIS FOR THE TA25 INDEX: DENSITY COMPARISON (DIMA ET. AL) 71

FIGURE 8-2 RUN SEQUENCE PLOT OF NASDAQ COMPOSIT INDEX: CLOSING PRICE 72

FIGURE 8-3 RUN SEQUENCE PLOT OF NASDAQ COMPOSITE INDEX: PRICE-TO-RETURN 72

FIGURE 8-4 AUTOCORRELATION FUNCTION PLOT FOR NASDAQ RETURNS 72

FIGURE 8-5 PARTIAL AUTOCORRELATION PLOT FOR NASDAQ RETURNS 72

FIGURE 8-6 AUTOCORRELATION PLOT OF THE SQUARED RETURNS 73

FIGURE 8-7 INNOVATIONS, STANDARD DEVIATIONS AND OBSERVED RETURNS OF THE MODEL 76

FIGURE 8-8 PLOT OF STANDARDIZED RESIDUALS 76

FIGURE 8-9 AUTOCORRELATION PLOT OF SQUARED STANDARDIZED RESIDUALS 76

FIGURE 8-10 FORECASTING RESULTS FROM GARCH(1,3) MODEL (GARCIA ET. AL.) 97

FIGURE 8-11 REAL VS. FORECAST PRICE IN THE SPANISH ELECTRICITY MARKET (GARCIA ET. AL.) 97

FIGURE 8-12 RUN SEQUENCE PLOT OF TSX RETURNS 98

FIGURE 8-13 ACF PLOT OF SQUARED RETURNS 98

FIGURE 8-14 CONDITIONAL STANDARD DEVIATION OF THE RESIDUALS 100

FIGURE 8-15 RESIDUALS AND RETURNS: RUN SEQUENCE PLOT 101

FIGURE 8-16 ACF PLOT OF THE STANDARDIZED RESIDUALS 101

FIGURE 10-1 A GRAPHICAL REPRESENTATION OF THE FINANCIAL MARKETS 111

1. INTRODUCTION

The financial market is divided into four components: capital markets, commodities, foreign exchange and derivatives[1]. Capital markets are further divided into two components: debt and equity financing (of which, the stock market forms the public forum). Debt financing involves accepting loans from lenders and making interest payments, while equity financing raises funds by accepting payment from parties interested in fractional ownership of the company. Buyers and sellers come together at stock exchanges where publicly listed companies' stocks are traded. The behaviour of share prices on these public exchanges has traditionally been modeled using random walk techniques[2], although empirical analyses point to some serial correlation.

Analysis techniques that seek to explain and model future values of financial instruments fall into two categories: fundamental and technical analysis. Fundamental analysis looks at a firm's fundamental financial information which is typically disclosed in periodic financial statements. In contrast, technical analysis relies on historical stock data to generate models or detect patterns that help forecast the direction of stock price. This thesis employs techniques described in the latter, and adds to the toolkit of the practitioner.

---

1 We use the term derivates as an umbrella term for exotic instruments including but not limited to futures, options, contracts, etc.

2 For a non-technical approach to the random walk model, we recommend Burton Malkiel's very readable *Random Walk Down Wall Street.*

2. OBJECTIVES

The main objective of this undergraduate thesis is to implement univariate time-series methodologies and Statistical Process Control (SPC) tools in jointly developing a system to forecast equations and monitor charts that assist in stock trading decisions. Secondary objectives, which include performing literature reviews of published papers, modifying time series equations, and designing CUSUM control charts support and facilitate the primary objective stated above. The first phase of the thesis is concerned with performing time series analysis within the classical Box-Jenkins framework; the second phase of the project applies SPC tools to generate quality control charts which enable the user to detect changes in stock behaviour; the system, consisting of the mathematical time-series model combined with SPC control techniques, is intended to be a valuable addition to the technical stock trader toolbox.

The thesis project begins with individual case studies of published papers in the econometrics and financial statistics domain. The individual literature reviews were largely motivated by the authors' desire to understand and evaluate papers published in their respective fields of interest; needless to say, aspiring practitioners and academics alike must be able to comprehend concepts presented in papers to remain up-to-date with developments in their chosen fields. Furthermore, the literature reviews helped refine the scope of the thesis given time constraints.

The time-series analysis phase consists of three univariate models: one each for up-trends, horizontal trends and down trends. The SPC phase consists of quality control charts which will be designed to track real-time performance and produce signals indicating a shift in trend. Based on the graphical output of the charts, various investment strategies may be implemented with respect to the stock in question. The remainder of this section provides details into the two phases described above.

A stock listed on both the S&P 500 and the TSX 60 was chosen: this requirement was established since both exchanges have specific financial, operational and management requirements and guidelines that a firm must continue to meet in order to remain listed. Once selected, historical adjusted closing data was obtained and visually inspected for trends. The data was into three categories based on the prevalent trend between two desired dates; it is a key requirement of this thesis for each category to contain not only sufficient data points for each history, but for each category to contain multiple, visually analogous histories. This requirement stems from the need to pool datasets to generate one model for each category.

Using an approach similar to Box-Jenkins', standard formulae for the autocorrelation (ACF) and partial autocorrelation (PACF) functions were modified to accommodate the 'pooled' nature of the data sets. Based on the pooled ACF and PACF charts, the degree of differencing required was inferred. The paper then used the Yule-Walker set of linear equations to arrive at estimates of the autoregressive terms.

Based on the pooled, autoregressive models, one-step ahead prices were forecasted and the associated residuals captured. SPC tools (specifically CUSUM charts) were developed to assist in the process of monitoring residuals generated from the models, with the ultimate goal of detecting shifts in the mean or standard deviation of the overlying process. These signals—to be detected visually on the CUSUM charts, indicating changes in the underlying process—are intended to prompt a buy, sell or hold investment decision.

3. STOCK AND DATA SELECTION

For the purpose of this thesis, a stock that is listed on both the S&P 500 and the TSX 60 indices was desired. This requirement was established since both exchanges have specific financial, structural and management requirements and guidelines that a firm must continue to meet in order to be listed. Furthermore, a stock which exhibited varying performance and trends was desired to allow for the development of three distinct models. Based on visual inspection, three categories of adjusted close prices (uptrend, downtrend and horizontal data sets) were gathered for Toronto-based gold producer Yamana Gold Corp. Below, a brief overview of the firm, its condensed financial statements and stock charts are provided.

3.1. Company Background: Yamana Gold Corp.

Yamana Gold Inc. is a Toronto-based company engaged in the acquisition, exploration, development and operation of mineral properties; specifically, Yamana is a gold producer with significant gold production, gold development stage properties, exploration properties, and land positions in Brazil, Argentina, Chile, Mexico and Central America. Yamana is cross-listed on the Toronto (TSX: YRI), New York (NYSE:AUY) and London (LSE: YAU) stock exchanges.

Financials	
(In millions of USD)	2008
Income Statement	
Total Revenue	$1,054.61
Gross Profit	$557.84
Operating Income	$206.95
Net Income	$434.77
Balance Sheet	
Total Current Assets	$522.11
Total Assets	$9,337.35
Total Current Liabilities	$360.80

Stock Performance	
Open	$ 10.81
High	$ 10.88
Low	$ 10.29
Volume	6.11M
Market Capitalization	7.69B
52-Week High	$ 19.79
52-Week Low	$ 4.29
Shares Outstanding	733.49M
Price-to-Earnings (P/E)	13.33
Stock Beta	0.95
Earnings-per-share (EPS)	$ 0.79

Total Liabilities	$2,827.23
Total Equity	$6,510.12

Cash Flow	
Net Income/Starting Line	$434.77
Cash from Operating Activities	$328.68
Cash from Investing Activities	($559.72)
Cash from Financing Activities	$131.58
Net Change in Cash	($114.76)

Key Ratios

	2008
Net profit margin	41.20%
Operating margin	19.60%
EBITD margin	27.90%
Return on average assets	4.50%
Return on average equity	7.00%

Table 3-1 Key Financial, Stock Performance and Company Ratios for Yamana Corp. Adapted from Google Finance (March 15, 2009).

The company's stock chart was obtained from Google Finance (Figure 3-1), while the adjusted closing prices were downloaded to a spreadsheet using Yahoo! Finance.

Figure 3-1 Stock Chart of Yamana Gold Inc., (TSE: YRI)

3.2. THE DATA SELECTION PROCESS

The data selection process involved visually inspecting the stock chart and identifying data sets for up trends, downtrends and horizontal trends, each consisting of a minimum of fifty data points. This requirement will ensure statistical significance of the captured data. In Figure 3-2, colour-coded boxes enclose regions of interest on Yamana Gold's stock chart. The nomenclature referring to increasing, decreasing and horizontal histories consists of letters I_i, D_d and H_h, respectively.

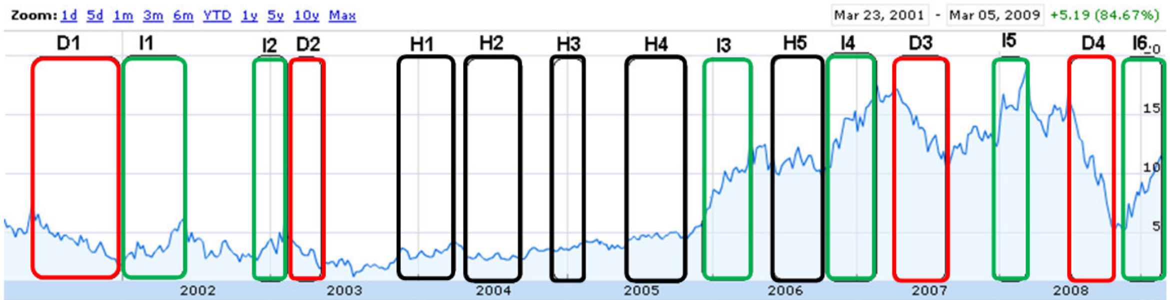


Figure 3-2 Data sets selected for pooling, illustrated by the stock chart

3.3. Datasets

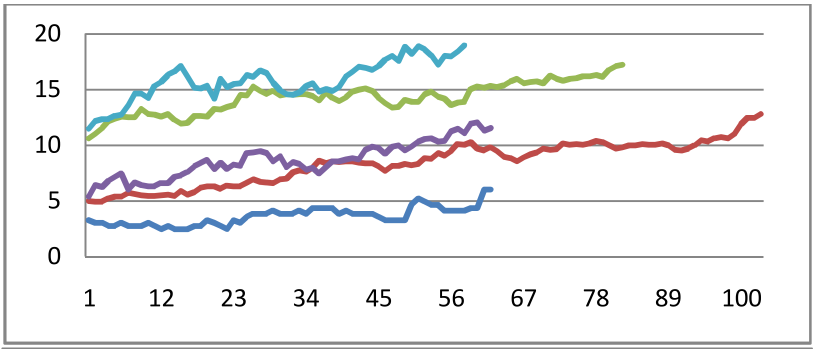
The data sets used in the Box-Jenkins framework are summarized in Table 3-2 along with summary statistical information; run plots of each category (and associated histories) are given under *Run Sequence Plot* on page 28.

	Start	End	N	$\bar{x}$	s	Kurtosis	Skew
INCREASING	8-Nov-02	7-Feb-03	62	3.640	0.833	0.526	0.697
	11-Nov-05	7-Apr-06	103	8.413	1.930	-0.797	-0.187
	27-Oct-06	23-Feb-07	82	14.279	1.434	-0.371	-0.269
	21-Nov-07	20-Feb-08	62	8.646	1.623	-0.629	0.174
	20-Dec-07	14-Mar-07	58	15.824	1.801	-0.216	-0.328
DECREASING	25-May-01	14-Aug-01	77	4.967	0.808	1.241	0.726
	7-Feb-03	6-May-03	61	3.636	0.802	0.785	0.352
	13-Apr-07	17-Aug-07	87	13.592	1.691	-0.705	0.084
	27-Jun-08	27-Oct-08	83	10.652	3.101	-0.652	0.008
FLUCTUATING	28-Nov-03	23-Apr-04	103	3.329	0.416	-0.339	0.472
	7-May-04	17-Sep-04	92	2.830	0.194	0.192	0.691
	8-Oct-04	11-Feb-05	87	3.562	0.185	0.302	0.799
	10-Jun-05	4-Nov-05	102	4.622	0.201	-0.390	0.434
	19-May-06	27-Oct-06	111	10.575	0.739	0.122	-0.035

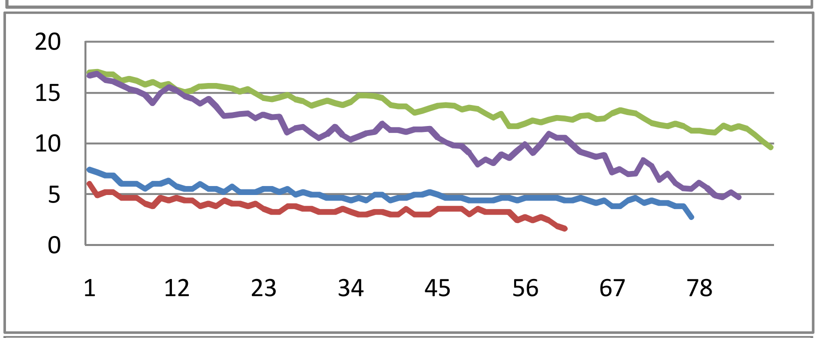
Table 3-2 Data Sets Used to Develop Untrend, Downtrend and Fluctuating Models

The kurtosis is a measure of the peak for a distribution. High kurtosis means more of the variance is due to infrequent, extreme deviation while low kurtosis means that the variance is due to frequent, modestly-sized deviations (Kurtosis). Skew is a measure of the asymmetry of the probability distribution: a negative skew implies the left tail is longer and the mass of the distribution is concentrated on the right of the figure while a positive skew implies the right tail is longer and the mass of the distribution is concentrated on the left of the figure (Skewness).

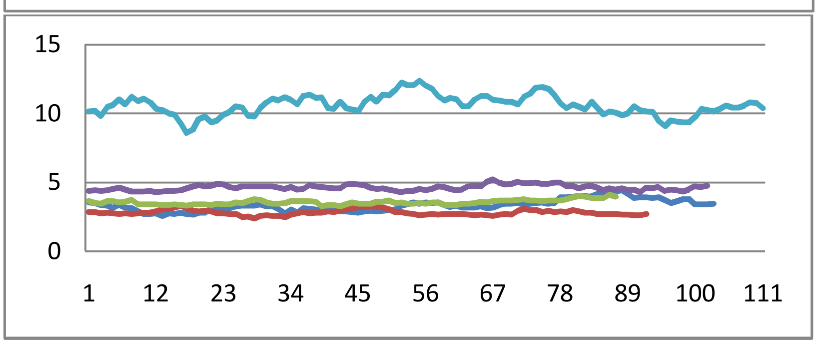
INCREASING



DECREASING



FLUCTUATING



4. OVERVIEW OF TIME SERIES ANALYSIS AND STATISTICAL PROCESS CONTROL

In this section, brief introductions to time series analysis and statistical process control are provided to prepare the undergraduate reader in appreciating the mathematical framework of this thesis. The authors assume a university level understanding of statistics; readers interested in the application may skip forward to Overview of the Box-Jenkins Approach on page 28 without loss of continuity.

4.1. INTRODUCTION TO TIME SERIES ANALYSIS[3]

At the most elementary level, observations taken sequentially in time constitute a time series; hourly yield of chemical processes, sequential measurements of volume on filling line, and daily closing prices of stocks are all examples of univariate (i.e., single response variable) time series. Time series analysis pertains to the construction, fitting and validation of models which are intended to allow for the forecasting and detection of changes in the underlying process.

Time series phenomena are largely driven by a stochastic engine in that future values are only partly determined by previously observed values. For a stochastic process which is stationary, the mean and the sample mean are given by (4.1) and(4.2), respectively.

$$\mu = E\left[z_t\right] = \int_{-\infty}^{\infty} zp(z)dz \tag{4.1}$$

[3] Adapted from (Box, Jenkins, & Reinsel, 1994)

$$\overline{z} = \frac{1}{N}\sum_{t=1}^{N} z_t \tag{4.2}$$

Where μ is process mean

$\bar{z}$ is the sample mean

N is the number of observations

zt is an observation at time t

p(z) is the probability distribution

E[z] is the expected value

The stochastic process variance and sample variance are yielded by (4.3) and(4.4), as follows:

$$\sigma_z^2 = E\left[\left(z_t - \mu\right)^2\right] = \int_{-\infty}^{\infty}\left(z - \mu\right)^2 p(z)dz \tag{4.3}$$

$$\hat{\sigma}_z^2 = \frac{1}{N}\sum_{t=1}^{N}(z_t - \overline{z})^2 \tag{4.4}$$

Where σ^2 is the process variance

$\hat{\sigma}$2 is the sample variance

μ is process mean

$\bar{z}$ is the sample mean

N is the number of observations

z_t is an observation at time t

p(z) is the probability distribution

E[z] is the expected value.

4.2. DEVELOPING AN AUTOREGRESSIVE AR(P) MODEL

In this section, an elementary autoregressive model is developed and the *AR(p)* notation is introduced; this will later be expanded to include differencing and moving average terms. A general autoregressive model consisting of *p* autoregressive terms is given in equation (4.5)

$$X_t = \xi + \sum_{i=1}^{p}\varphi_i X_{t-i} + \varepsilon_t \tag{4.5}$$

In equation(4.5), if *p* is limited to 1, the number of autoregressive terms is thusly restricted to one; this yields a first-order autoregressive model:

$$X_t = \xi + \varphi X_{t-1} + \varepsilon_t \tag{4.6}$$

Where ξ is a constant

φ (-1< φ <1) is the model parameter to be estimated

ε_t ~ IID N(0,σ^2)

The error term ε_t is white noise, mathematically understood to be Gaussian IID (identical and independently distributed) with a mean (or expected value) of zero and an unknown but fixed variance σ^2. Applying the expectation function ***E*[.]** to both sides of equation (4.6), yields the expected value (or mean) of the observation X_t as follows:

$$\begin{aligned} &\mathrm{E}[X_t] = \mathrm{E}[\xi] + \varphi \mathrm{E}[X_{t-1}] + \mathrm{E}[\varepsilon_t] \\ &\mu = \xi + \varphi\mu + 0 \end{aligned} \tag{4.7}$$

Rearranging for the mean results in the following equation for μ:

$$\mu = \frac{\xi}{1-\varphi} \tag{4.8}$$

The variance of X_t is derived by applying the ***VAR*[.]** operator thusly:

$$\mathrm{VAR}[X_t] = \mathrm{E}[X_t^{\ 2}] - \mu^2 = \frac{\sigma^2}{1-\varphi^2} \tag{4.9}$$

The autocovariance between observations separated by K time increments is given by

$$B_K = \mathrm{E}[X_{t+K} X_t] - \mu^2 = \frac{\sigma^2}{1-\varphi^2}\varphi^{|K|} \tag{4.10}$$

The correlation coefficient is given by φ^K; furthermore, we can see that that the autocovariance function decays exponentially[4].

Suppose now that $\hat{\varphi}$ an estimate of φ obtained from analysis of sample data from the process, and $\hat{X}_t$ is the fitted value of X_t. Then, the residuals may be calculated as

$$\mathrm{e}_t = X_t - \hat{X}_t \tag{4.11}$$

and are found to be approximately Gaussian IID ~$(0,\sigma^2)$.

Generally available Statistical Process Control charts may be employed to analyze the sequence of residuals being generated in real-time, almost mirroring a manufacturing process. In particular, we apply CUSUM charts which generate a visual signal when the process parameters ξ or σ change, implying that the process has gone out of control; if we had been tracking a fluctuating time-series, a signal implies that the stock has changed behaviour, and based on whether the lower or upper control limit was breached, the price is most likely heading into sell or buy territory, respectively. We can easily extend the *AR(1)* model developed in the section above to a second-order model, as follows:

$$X_t = \xi + \varphi_1 X_{t-1} + \varphi_2 X_{t-2} + \varepsilon_t \tag{4.12}$$

---

4 Rewriting as $B_K = G\varphi^{|K|}$ with G independent of K. Note the similarity between $\varphi^{|K|} = e^{|K|\ln\varphi}$ and $\varphi^{|K|} = e^{-n/\tau}$

4.3. Developing a Mixed Autoregressive, Moving Average Model

It is sometimes useful to model the dependency observed between the error term ε_t. A simple way to do this is through first-order moving average model, *MA(1)*:

$$X_t = \mu + \varepsilon_t - \theta\varepsilon_{t-1} \tag{4.13}$$

In this model, the correlation between x_t and x_{t-1} is $\rho_1 = -\theta / 1 + \theta^2$ and zero at all other lags. Thus, the correlative structure in x_t only extends backwards one time period. Sometimes combinations of autoregressive (AR) and moving average (MA) terms are useful. A first-order mixed model is given by equation (4.14):

$$X_t = \xi + \varphi X_{t-1} + \varepsilon_t - \theta\varepsilon_{t-1} \tag{4.14}$$

Simply, the underlying process variable x_t is first-order autoregressive (φX_{t-1}) and first order moving average of the error term ($\varepsilon_t - \theta\varepsilon_{t-1}$).

The equations derived above belong to a class of time series models call the Autoregressive Integrated Moving Average (ARIMA) models. ARIMA models are parameterized by *p, d* and *q*—the order of the autoregressive term, the number of times the time series is to be differenced, and the order of the moving average term. To illustrate, an ARIMA(2,1,0) model differences the time series once, the resulting equation contains two autoregressive terms and no moving average terms. In order to preserve the readership of this thesis to that of an undergraduate audience, the models developed herein do not contain moving average terms, and while a MA term may be cast as an infinite series of autoregressive terms, this exercise is beyond the scope of this paper.

4.4. MODIFYING THE STANDARD $AR(p)$ EQUATIONS TO ALLOW FOR POOLING

The sample mean and sample autocovariance provided by Box-Jenkins needed to be modified in order to accommodate the ‘pooled’ nature of the datasets; below is a step by step development of the pooled equations from the standard equations.

	Standard formulation	**‘Pooled’ formulation**
Process mean	$\mu = E\left[z_t\right] = \int_{-\infty}^{\infty} zp(z)dz$	
Sample mean	$\bar{z} = \frac{1}{N}\sum_{t=1}^{N} z_t$	$\hat{\mu} = \frac{1}{\sum_{i=1}^{H} T_i}\sum_{i=1}^{H}\sum_{n=1}^{T_i} x_n^i$
Mean subtraction	-	$y_n^i = x_n^i - \hat{\mu}$
Process Autocovariance	$\gamma_k = \mathrm{cov}\left[z_t, z_{t+k}\right] = E\left[\left(z_t - \mu\right)\left(z_{t+k} - \mu\right)\right]$	
Sample autocovariance	$c_k = \frac{1}{N}\sum_{t=1}^{N-k}\left(z_t - \bar{z}\right)\left(z_{t+1} - \bar{z}\right)$	$c_k = \frac{1}{\sum_{i=1}^{H} T_i}\sum_{i=1}^{H}\sum_{n=1}^{T_i-k}\left(y_n^i \times y_{n+k}^i\right)$
Process autocorrelation	$\rho_k = \frac{E\left[\left(z_t - \mu\right)\left(z_{t+k} - \mu\right)\right]}{\sqrt{E\left[\left(z_t - \mu\right)^2\right]E\left[\left(z_{t+k} - \mu\right)^2\right]}} = \frac{E\left[\left(z_t - \mu\right)\left(z_{t+k} - \mu\right)\right]}{\sigma_z^2} = \frac{\gamma_k}{\gamma_0}$	$c_0 = \frac{1}{\sum_{i=1}^{H} T_i}\sum_{i=1}^{H}\sum_{n=1}^{T_i}\left(y_n^2\right)^2$
Sample autocorrelation	$r_k = \frac{c_k}{c_0} = \frac{\frac{1}{N}\sum_{t=1}^{N-k}\left(z_t - \bar{z}\right)\left(z_{t+1} - \bar{z}\right)}{\frac{1}{N}\left(z_0 - \bar{z}\right)}$	$r_k = \frac{c_k}{c_0} = \frac{\frac{1}{N}\sum_{t=1}^{N-k}\left(z_t - \bar{z}\right)\left(z_{t+1} - \bar{z}\right)}{\frac{1}{\sum_{i=1}^{H} T_i}\sum_{i=1}^{H}\sum_{n=1}^{T_i}\left(y_n^2\right)^2}$

Table 4-1 Standard and Modified Equations for the AR(p) Model

4.5. INTRODUCTION TO STATISTICAL PROCESS CONTROL (SPC)

When analyzing a set of data, a Shewhart control chart must be constructed to determine whether any large shifts (>1.5 σ) have taken place. If one may conclude from the Shewhart control chart that the process is in control, a CUSUM chart may then be constructed to detect small shifts (<1.5 σ).

4.6. Shewhart Control Charts[5]

To being implementing Shewhart control charts, the following statistics need to be calculated for the process:

Estimate of the population standard deviation

$$\hat{\sigma} = \frac{\overline{R}}{d_2(n)} = \frac{\overline{MR}}{d_2(n)} = \frac{\overline{S}}{C_4(n)} \tag{4.15}$$

Sample variance

$$S^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{(n-1)} \tag{4.16}$$

Estimate of the in-control process mean

$$\hat{\mu}_0 = \overline{\overline{x}} \tag{4.17}$$

Restricting sample size to one for simplicity, the sample control chart consisting of the individual and moving range, henceforth referred to as the *(I, MR)* chart may be generated with the following upper and lower control limits and center lines:

[5] Adapted from (Montgomery, 2005).

I-Chart	MR-Chart
$UCL = \bar{X} + z_{\alpha/2}\hat{\sigma}$	$UCL = \dfrac{\overline{MR}}{d_2(2)} W_{0.999}(2)$
$CL = \bar{X}$	$CL = \overline{MR}$
$LCL = \bar{X} - z_{\alpha/2}\hat{\sigma}$	$LCL = \dfrac{\overline{MR}}{d_2(2)} W_{0.001}(2)$

If, upon consulting the Shewhart control charts, the process is deemed to be in-control, we may proceed with the tabular cumulative sum (CUSUM) control charts.

4.7. TABULAR (ALGORITHMIC) CUSUM

The CUSUM chart plots two one-sided series: upper one-sided values (C_i^+) and lower one-sided value (C_i^-). The upper one-sided values aggregate deviations *above* the process target mean of μ_0, and are calculated as:

$$C_i^+ = \max[0, x_i - (\mu_0 + k) + C_{i-1}^+] \tag{4.18}$$

Similarly, the lower one-sided value aggregate deviations *below* μ_0, and are calculated as:

$$C_i^- = \max[0, (\mu_0 - k) - x_i + C_{i-1}^-] \tag{4.19}$$

Where x_i is the ith observation of the process
μ_0 be the mean or target value
k is the reference value

The above equations only pertain to the data series that are to be plotted on the CUSUM chart; the process of determining CUSUM control limits are discussed stepwise in the following:

STEP 1 – DETERMINE THE CRITICAL SHIFT DELTA

The critical size of the shift δ must be determined using the equation (4.20) which relates the shift from the in-control mean μ_0 to the out of control upper and lower means μ_1 and μ_2 respectively. Again, restricting the sample size to one for simplicity:

$$\begin{aligned} \mu_1 &= \mu_0 + \delta_1\sigma \\ \mu_2 &= \mu_0 + \delta_2\sigma \end{aligned} \tag{4.20}$$

where μ_0 be the mean or target value
μ_1 be the upper shift mean
μ_2 be the lower shift mean
δ_1 be the upper critical shift
δ_2 be the lower critical shift
σ be the standard deviation of the process

STEP 2- DETERMINE THE REFERENCE VALUE K

The reference value K is half the magnitude of the shift, thus a value half way between the target μ_0 and the upper CUSUM mean μ_1, or the lower CUSUM mean μ_2. The reference value is a line, which if crossed, provides an early warning of a shift in the process mean. The upper and lower reference values K_1 and K_2 are calculated as follows:

$$\begin{aligned} K_1 &= \frac{|\mu_1 - \mu_0|}{2} = \frac{\delta_1\sigma}{2} \\ K_2 &= \frac{|\mu_2 - \mu_0|}{2} = \frac{\delta_2\sigma}{2} \end{aligned} \tag{4.21}$$

STEP 3 – DETERMINE THE DECISION VALUE H

A decision variable H, which acts as a control limits, must be created to determine the state of the process. The decision variable H, if exceeded either from above (by the one-sided upper CUSUM) or below (by the one-sided lower CUSUM), determines a shift in the process mean. The upper and lower decision variables H_1 and H_2 are calculated as follows:

$$
\begin{aligned}
H_1 &= h_1\sigma \\
H_2 &= h_2\sigma
\end{aligned} \tag{4.22}
$$

However, to calculate *H*, the average run length (ARL) parameter *h* for both the one sided upper (h_1) and lower (h_2) must be determined. Keeping step 3 in mind, the ARL is therefore determined first.

STEP 4 – THE AVERAGE RUN LENGTH (ARL)

The two-sided average run length is related to the one-sided upper and lower CUSUM average run lengths by the following equation:

$$
\frac{1}{ARL_{II}(\delta_1, h_1, \delta_2, h_2, a)} = \frac{1}{ARL_I(\delta_1, h_1, a)} + \frac{1}{ARL_I(\delta_2, h_2, -a)} \tag{4.23}
$$

One-sided CUSUM ARL's may be calculated as:

$$
ARL_I(\delta, h, a) = \begin{cases} \dfrac{e^{-c} + c - 1}{2(a - \frac{\delta}{2})^2}, & if\ \ a \neq \dfrac{\delta}{2} \\[2ex] (h + 1.166)^2, & if\ \ a = \dfrac{\delta}{2} \end{cases} \tag{4.24}
$$

where $\mathrm{c} = 2(a - \frac{\delta}{2})(\mathrm{h} + 1.166)$

a = actual shift from μ_0 to $\mu_1 = \mu_0 + a\sigma$

h = find, control limit $H = h\sigma$

δ = critical size of shift

$H_1 = h_1 \sigma_{\bar{X}}$

STEP 5 – CALCULATE ARL IN CONTROL

When the ARL is in control, the actual shift is zero (i.e., $a = 0$). Therefore, equation (4.23) simplifies into

$$\frac{1}{ARL_{II}(\delta_1,h_1,\delta_2,h_2,0)}=\frac{1}{ARL_I(\delta_1,h_1,a)}+\frac{1}{ARL_I(\delta_2,h_2,-a)} \tag{4.25}$$

and the one-sided ARL in equation (4.24) may be rewritten as

$$ARL_{IN}(\delta_1,h_1,0) = \frac{e^{-c}+c-1}{2\left(\frac{\delta_1}{2}\right)^2} \tag{4.26}$$

$$where \begin{array}{l} \mathrm{c}=-\delta_1(\mathrm{h}_1+1.166) \\ ARL_{II}(\delta_2,h_2,0) \end{array}$$

If b is replaced with $-c$ (since $b = -c > 0$), ARL_I can be written as function g_1 in b_1:

$$g_1(b_1) = ARL_I(\delta_1,h_1,0)=\frac{e^b-b-1}{\frac{1}{2}\delta_1^2} \tag{4.27}$$

Re-arrange $g_1(b_1)$ into a non-linear equation:

$$\begin{aligned} g_1(b_1)&=e^b-b-1=ARL_I(\delta_1,h_1,0)\left(\frac{\delta_1^2}{2}\right) \\ &=e^b-b-1-ARL_I(\delta_1,h_1,0)\left(\frac{\delta_1^2}{2}\right)=0 \end{aligned} \tag{4.28}$$

In equation (4.28), b may be ignored since $e^b >> b$ which yields:

$$g_1(b)=e^b-1-ARL_I(\delta_1,h_1,0)\left(\tfrac{1}{2}\delta_1^2\right)=0 \tag{4.29}$$

Rearranging and take the natural logarithm of both sides of equation (4.29) gives:

$$b_0=\ln\left[1+ARL_I(\delta_1,h_1,0)\left(\tfrac{1}{2}\delta_1^2\right)\right] \tag{4.30}$$

Equation (4.30) is sufficiently in a form where Newton's method can be applied to approximate, through iteration, a value for b_{i+1}:

$$b_{i+1} = b_i - \frac{g_1(b_i)}{g_1'(b_i)} \tag{4.31}$$

where the stopping rule is defined as $|b_{i+1} - b_i| / |b_{i+1}| < 1\times10^{-6}$ *and* $g_1'(b_i) = e^b - 1$.

Using a suitable of b_i, h_1 is calculated by rearranging $b_i = \delta_1(h_1 + 1.166)$ into

$$h_1 = (b_i/\delta_1) - 1.166 \tag{4.32}$$

Once h_1 is obtained, decision variable H_1 is arrived at using equation(4.22). Recalculate the in-control decision variable for the lower one sided CUSUM (H_2) using the same method used to determine the in-control upper one sided CUSUM (H_1) using $ARL_{II}(\delta_2, h_2, 0)$. If the assumption is made that the control limits for the process are symmetric, the in-control decision variable H_2 will be mirrored over by negating the upper decision variable.

5. OVERVIEW OF THE BOX-JENKINS APPROACH

The Box-Jenkins methodology is an econometric framework, named after statisticians George Box (b. 1933 – d. 1982) and Gwilym Jenkins (b. 1919 –) which applies autoregressive moving average (ARMA) models or autoregressive integrated moving average (ARIMA) models to time-series data to facilitate forecasting. The original model uses an iterative three-stage modeling approach consisting of *model identification and selection*; *parameter estimation*; and *model validation*.

5.1. MODEL IDENTIFICATION

In the first stage of the Box-Jenkins approach, charting techniques consisting of *run sequence plots, autocorrelation plots* and *partial autocorrelation plots* are used to check stationarity, seasonality and autoregression, amongst other characteristics of the time series. The model identification stage helps decide whether the time series may be modeled by ARIMA, and if so, what order of autoregressive and moving average terms should be chosen for the validation stage.

RUN SEQUENCE PLOT

A run sequence plot of the data is generated to determine time series stationarity i.e., whether the time series exhibits constant mean and scale. A time series is non-stationary if it appears it appears to have no fixed level. While time series may display some periodic fluctuations, most of the tools and techniques discussed in this paper apply only to stationary time series. Should the data appear to be non-stationary, there exist techniques to transform it into a stationary time series.

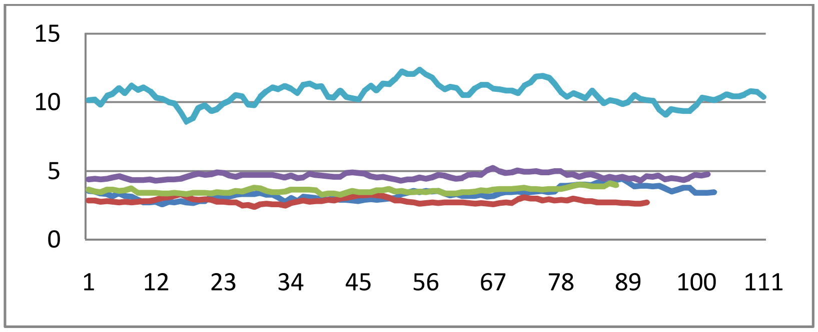


Figure 5-1 Run Plot of all histories contained within the 'Horizontal' dataset

There are two main types of stationary transformations. First, the data may be differenced if the stochastic process has an unstable mean. This type of transformation is used to remove polynomial trends that may be exhibited by the data. Second, logarithmic, square root and curve-fitting transformation are covered in literature, but the differencing approach is recommended by Box-Jenkins. These transformations are used if the series being examined has a non-constant mean and variance and it results in a straighter curve plot. Run plots also help identify the presence of any unusual points or outliers, and any seasonality that may be exhibited by the time series.

Autocorrelation Plot

The autocorrelation function (ACF) measures the degree of correlation between lagged values of the times series. For example, the autocorrelation coefficient at lag 1 (R_1) describes the over autocorrelation of the data series between successive observations *viz.,* between observations 1 and 2, 2 and 3, 3 and 4, and so on and so forth. Similarly, R_2 captures the correlation between observations 1 and 3, 2 and 4, 3 and 5 and so on and so forth.

The autocorrelation value is bounded by the interval [-1,1] where a value close to 1 indicates strong, positive correlation; a value close to -1 indicates strong negative correlation;

and a value close to 0 indicates weak or no correlation. The sample ACF at lag k is the autocovariance at lag k normalized by the covariance $C_0 = \frac{1}{N}\sum_{t=1}^{N}(z_t - \bar{z})^2$ of the time series:

$$r_k = c_k c_0^{-1} \tag{5.1}$$

The k-lagged process autocovariance γ_k and sample autocovariance c_k are:

$$\begin{aligned} \gamma_k &= \operatorname{cov}\left[z_t, z_{t+k}\right] = E\left[\left(z_t - \mu\right)\left(z_{t+k} - \mu\right)\right] \\ c_k &= \frac{1}{N}\sum_{t=1}^{N-k}\left(z_t - \bar{z}\right)\left(z_{t+1} - \bar{z}\right) \end{aligned} \tag{5.2}$$

where $cov[z_t,z_{t+k}]$ *is the covariance between* z_t *and* z_{t+k}
$E[z]$ is the expected value
μ is process mean
$\bar{z}$ is the sample mean
N is the number of observations
z is the observation

Meanwhile, the k-lagged process autocorrelation ρ_k and the sample autocorrelation r_k are:

$$\begin{aligned} \rho_k &= \frac{E\left[\left(z_t - \mu\right)\left(z_{t+k} - \mu\right)\right]}{\sqrt{E\left[\left(z_t - \mu\right)^2\right]E\left[\left(z_{t+k} - \mu\right)^2\right]}} = \frac{E\left[\left(z_t - \mu\right)\left(z_{t+k} - \mu\right)\right]}{\sigma_z^2} = \frac{\gamma_k}{\gamma_0} \\ r_k &= \frac{c_k}{c_0} = \frac{\frac{1}{N}\sum_{t=1}^{N-k}\left(z_t - \bar{z}\right)\left(z_{t+1} - \bar{z}\right)}{\frac{1}{N}\left(z_0 - \bar{z}\right)} \end{aligned} \tag{5.3}$$

where σ^2 is the process variance
c_k the sample autocovariance
c_0 is the variance function
γ_k: the estimate of autocovariance
γ_0: the process variance of a stationary process

A chart that plots the sample autocorrelation coefficient r_k against the associated lag k is known as an ACF plot. ACF plots exhibit a variety of behaviour that is characterized by rates of decay, waveform behaviour, outlier behaviour, or seasonal behaviour. Figure 5-2 shows an ACF plot of the pooled decreasing dataset which exhibits very weak decay, while

Figure 5-3 shows an ACF plot with rapid decay. Table 5-1 outlines various behaviours of ACF plots that may be encountered in industry and the associated model to be used.

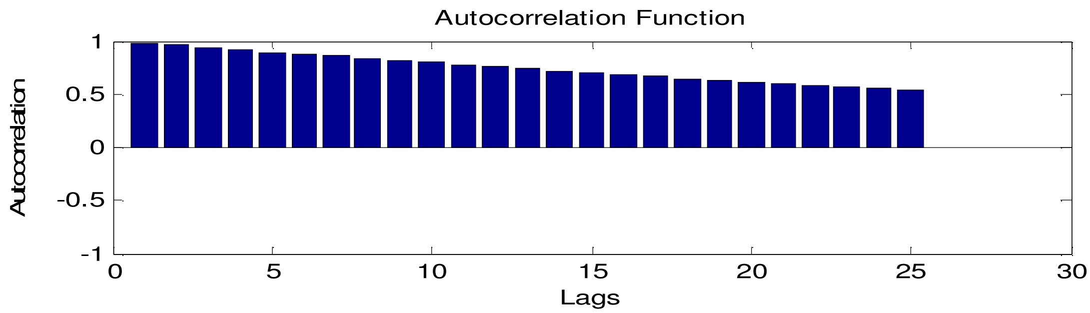


Figure 5-2 Autocorrelation Plot of Decreasing Dataset

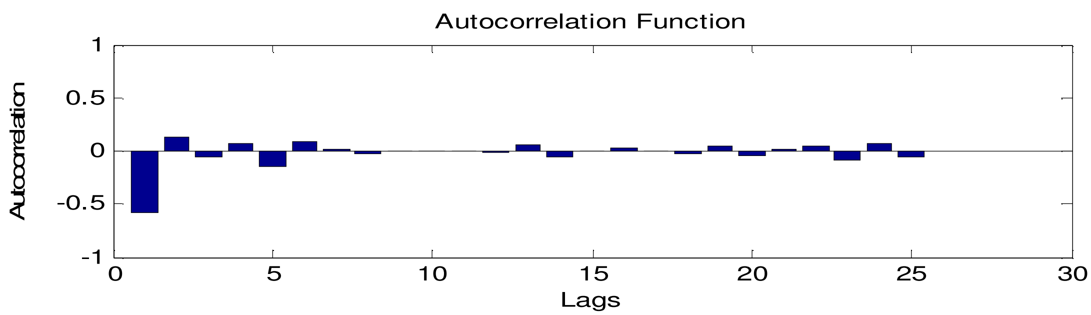


Figure 5-3 Double Differenced ACF Plot of Decreasing Dataset

Shape	Indicated Model
Exponential, decaying to zero	Autoregressive model - use PACF plot to identify the order.
Alternating positive and negative, decaying to zero	Autoregressive model - use PACF plot to identify the order.
One or more spikes, rest are essentially zero	Moving average model - order identified by where plot becomes zero.
Decay, starting after a few lags	Mixed autoregressive and moving average (ARMA) model.
All zero or close to zero	Data is essentially random.
High values at fixed intervals	Include seasonal autoregressive term.
No decay to zero	Series is not stationary.

Table 5-1 Behaviours of ACF Plots and Recommended Models

Partial Autocorrelation Plot

The partial autocorrelation coefficient at lag k is the autocorrelation between observations X_t and X_{t-k} that is *not* explained by lag $k = 1$ through to lag k-1. Similar to the ACF, the

PACF is bounded on the [-1,1] interval; the numerical interpretation of the PACF with respect to correlative behaviour and strength are also similar to the ACF. Solving a system of equations (called the Yule-Walker equations, discussed in the following section) helps arrive at the PACF. Like the ACF plot, a PACF plots the a *k*-lagged PAC coefficient against *k (*Figure 5-4*)*.

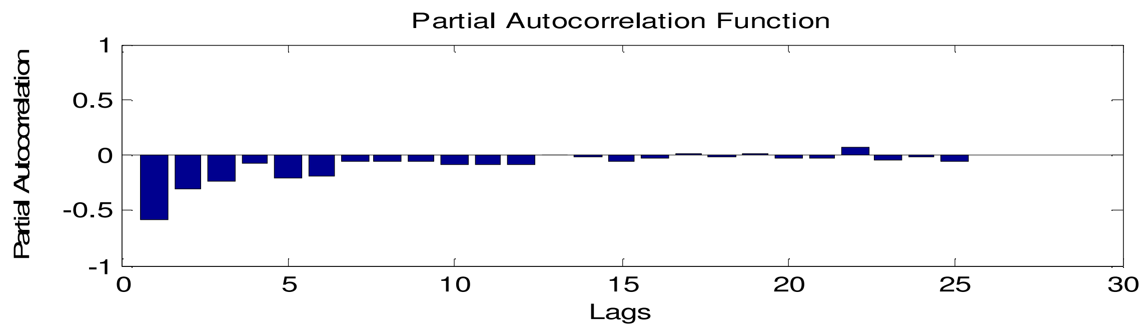


Figure 5-4 PACF Plot of 'Increasing' pooled data, second-order differenced

5.2. Model Estimation[6]

There are two methodologies to estimate the parameters of the *AR(p)* model: direct inversion and the Yule-Walker set of linear equations. Since the Yule-Walker equations are useful in determining the PACF, the paper will focus on the same, although a brief overview of direct inversion method follows.

---

[6] Adapted from (Eshel).

For a generic AR (1) process $x_t = \phi_1 x_{t-1} + \varepsilon_t$, the following over-determined system is formulated as:

$$\mathbf{b} = \mathbf{A}\phi$$

$$\begin{pmatrix} x_2 \\ x_3 \\ \vdots \\ x_N \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_{N-1} \end{pmatrix} (\phi_1) \tag{5.4}$$

Using a least squares estimator approach, estimates can be obtained thusly

$$\hat{\phi}_1 = \left(A^T A\right)^{-1} A^T b = \frac{\sum_{t=1}^{N-1} x_t x_{i-1}}{\sum_{t=1}^{N-1} x_t^2} = \frac{c_1}{c_0} = r_1 \tag{5.5}$$

where c_i *is the ith autocovariance coefficient*

r_i *is the ith autocorrelation coefficient*

Generalizing for an AR (2) process $x_t = \phi_1 x_{t-1} + \phi_2 x_{i-2} + \varepsilon_t$, a similar over-determined system can be formulated:

$$\mathbf{b} = \mathbf{A}\phi$$

$$\begin{pmatrix} x_3 \\ x_4 \\ \vdots \\ x_N \end{pmatrix} = \begin{pmatrix} x_2 & x_1 \\ x_3 & x_2 \\ \vdots & \vdots \\ x_{N-1} & x_{N-2} \end{pmatrix} \begin{pmatrix} \phi_1 \\ \phi_2 \end{pmatrix} \tag{5.6}$$

which is solved using the usual least squares estimator (now in matrix notation):

$$\Phi = \left(A^T A\right)^{-1} A^T b \tag{5.7}$$

The solution is not trivial and expanding the term $\mathbf{(A^T A)^{-1}}$ yields:

$$
\begin{aligned}
\left(A^T A\right)^{-1} &= \left[\begin{pmatrix} x_2 & x_3 & \cdots & x_{N-1} \\ x_1 & x_2 & \cdots & x_{N-2} \end{pmatrix}\begin{pmatrix} x_2 & x_1 \\ x_3 & x_2 \\ x_{N-1} & x_{N-2} \end{pmatrix}\right]^{-1} \\
&= \begin{pmatrix} \sum_{i=2}^{N-1} x_i^2 & \sum_{i=2}^{N-1} x_i x_{i-1} \\ \sum_{i=2}^{N-1} x_i x_{i-1} & \sum_{i=1}^{N-2} x_i^2 \end{pmatrix}^{-1} \\
&= \frac{1}{\left(\sum_{i=2}^{N-1} x_i^2 \sum_{i=1}^{N-2} x_i^2\right) - \left(\sum_{i=2}^{N-1} x_i x_{i-1} \sum_{i=2}^{N-1} x_i x_{i-1}\right)} \begin{pmatrix} \sum_{i=2}^{N-2} x_i^2 & -\sum_{i=2}^{N-1} x_i x_{i-1} \\ -\sum_{i=2}^{N-1} x_i x_{i-1} & \sum_{i=1}^{N-1} x_i^2 \end{pmatrix}
\end{aligned}
\tag{5.8}
$$

Since the time series is stationary, the autocovariance is a function of the lag only:

$$
\begin{aligned}
\left(A^T A\right)^{-1} &= \frac{1}{c_o^2 - c_1^2}\begin{pmatrix} c_o & -c_1 \\ -c_1 & c_o \end{pmatrix} \\
&= \frac{1}{c_o^2(1-r_1^2)}\begin{pmatrix} c_o & -c_1 \\ -c_1 & c_o \end{pmatrix} \\
&= \frac{1}{c_o(1-r_1^2)}\begin{pmatrix} r_o & -r_1 \\ -r_1 & r_o \end{pmatrix}
\end{aligned}
\tag{5.9}
$$

The second part of the equation then becomes

$$
A^T b = \begin{pmatrix} x_2 & x_3 & \cdots & x_{N-1} \\ x_1 & x_2 & \cdots & x_{N-2} \end{pmatrix}\begin{pmatrix} x_3 \\ x_4 \\ \vdots \\ x_N \end{pmatrix} = \begin{pmatrix} \sum_{i=3}^{N} x_i x_{i-1} \\ \sum_{i=3}^{N} x_i x_{i-2} \end{pmatrix}
\tag{5.10}
$$

And since the time series is stationary, rewrite as $A^T b = \begin{pmatrix} c_1 & c_2 \end{pmatrix}^T$

Combining equations (5.9) and (5.10) yields:

$$
\begin{aligned}
\left(A^T A\right)^{-1} A^T b &= \frac{1}{c_o\left(1-r_i^2\right)}\begin{pmatrix} r_o & -r_1 \\ -r_1 & r_o \end{pmatrix}\begin{pmatrix} c_1 \\ c_2 \end{pmatrix} \\
&= \frac{1}{1-r_1^2}\begin{pmatrix} 1 & -r_1 \\ -r_1 & 1 \end{pmatrix}\begin{pmatrix} r_1 \\ r_2 \end{pmatrix}
\end{aligned}
\tag{5.11}
$$

Individually, the estimates are given by:

$$\hat{\phi}_1 = \frac{r_1(1-r_2)}{1-r_1^2}$$
$$\hat{\phi}_2 = \frac{r_2 - r_1^2}{1-r_1^2} \tag{5.12}$$

This method may be extended for any p in an AR(p) model. The process becomes computationally tedious; instead, the Yule-Walker equations offer a better algorithm.

YULE WALKER EQUATIONS

Begin with a modified version of the generic AR(p) model as developed in equation (4.5):

$$X_{t+1} = \phi_1 x_i + \phi_1 x_{i-1} + \ldots + \phi_p x_{i-p-1} + \xi_{i+1} \tag{5.13}$$

For lag $k = 1$, multiply in x_i

$$x_i(x_{i+1}) = \sum_{j=1}^{p}\left(j_j x_i x_{i-j+1}\right) + x_i \varepsilon_{i+1} \tag{5.14}$$

where i is the time indices
j is the term indices

Apply the expectation function ***E[.]*** and realizing that $E\left[x_i \varepsilon_{i+1}\right] = 0$ if the error term is Gaussian IID ~N(0,σ):

$$E\left[x_i x_{i+1}\right] = \sum_{j=1}^{p}\left(\phi_j E\left[x_i x_{i-j+1}\right]\right) \tag{5.15}$$

Divide throughout by $N-1$ and knowing $c_{-l} = c_l$ due to the covariance's symmetry:

$$c_1 = \sum_{j=1}^{p} \phi_j c_{j-1} \tag{5.16}$$

Divide throughout by C_0 to arrive at the autocorrelation coefficient at lag 1 r_1:

$$r_1 = \sum_{j=1}^{p} \phi_1 r_{j-1} \tag{5.17}$$

If the process is replicated for lag 2, r_2 is given by:

$$r_2 = \sum_{j=1}^{p} \phi_1 r_{j-2} \tag{5.18}$$

Generalizing for any desired lag k, the autocorrelation coefficient r_k is given by

$$r_k = \sum_{j=1}^{p} \phi_j r_{j-k} \tag{5.19}$$

Lastly, the autocorrelation at the lag relating to the autoregressive order p is computed as:

$$r_p = \sum_{j=1}^{p} \phi_j r_{j-p} \tag{5.20}$$

Jointly, equations (5.17) through to (5.20) can be collected into the system of equations

$$\begin{aligned}
r_1 &= \phi_1 r_0 + \phi_2 r_1 + \phi_3 r_2 + \cdots + \phi_{p-1} r_{p-2} + \phi_p r_{p-1} \\
r_2 &= \phi_1 r_1 + \phi_2 r_0 + \phi_3 r_1 + \cdots + \phi_{p-1} r_{p-3} + \phi_p r_{p-2} \\
&\vdots \\
r_{p-1} &= \phi_1 r_{p-2} + \phi_2 r_{p-3} + \phi_3 r_{p-4} + \cdots + \phi_{p-1} r_0 + \phi_p r_1 \\
r_p &= \phi_1 r_{p-1} + \phi_2 r_{p-2} + \phi_3 r_{p-3} + \cdots + \phi_{p-1} r_1 + \phi_p r_0
\end{aligned} \tag{5.21}$$

and conveniently expressed into matrix notation as[7]

7 Recall that $r_0 = 1$

$$\begin{pmatrix} r_1 \\ r_2 \\ \vdots \\ r_{p-1} \\ r_p \end{pmatrix} = \begin{pmatrix} 1 & r_1 & r_2 & \cdots & r_{p-2} & r_{p-1} \\ r_1 & 1 & r_1 & \cdots & r_{p-3} & r_{p-2} \\ \vdots & & & & & \vdots \\ r_{p-2} & r_{p-3} & r_{p-4} & \cdots & 1 & r_1 \\ r_{p-1} & r_{p-2} & r_{p-3} & \cdots & r_1 & 1 \end{pmatrix} \begin{pmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_{p-1} \\ \phi_p \end{pmatrix} \tag{5.22}$$

In the classic form, equation (5.22) becomes $\mathbf{r} = \mathbf{R\Phi}$; this is a well-posed system with $\mathbf{R}$ full-rank and symmetric, which guarantees invertibility and yields $\hat{\mathbf{\Phi}} = \mathbf{R}^{-1}\mathbf{r}$.

Using Yule-Walker Equations To Compute PACF

Equation (5.22) provides an effective recursive method for computer the partial autocorrelation of the time series. The process is outlined below to allow the reader to put the authors' MATLAB code in computational context.

STEP 1 – COMPUTE THE ACF UP TO P = 0.25N

Begin by computing the autocorrelation function coefficients of a quarter of the data points.

STEP 2 – COMPUTE $R^{(i)}$ AND $r^{(i)}$

Let $R^{(i)}$ and $r^{(i)}$ denote the coefficient matrix and the right-hand side of equation (5.22) respectively, then while looping on $1 \le i \le p$, compute $R^{(i)}$ and $r^{(i)}$.

STEP 3 – INVERT FOR Φ

$$\hat{\phi}^{(i)} = \left(\mathbf{R}^{(i)}\right)^{-1} \mathbf{r}^{(i)} = \begin{pmatrix} \hat{\phi}_1 \\ \hat{\phi}_2 \\ \vdots \\ \hat{\phi}_i \end{pmatrix}$$

Compute the estimate

STEP 4 – RETAIN PACF

Eject all $\widehat{\phi}_j$ for $1 \leq j \leq i - 1$, but retain $\widehat{\phi}_i$ since PACF(i) = $\widehat{\phi}_i$.

STEP 5 – PLOT PACF

The last step calls for plotting the PACF coefficients against the lag values.

5.3. MODEL VALIDATION

Once the parameters have been estimated, the quality of the model must be established. If the estimation is determined to be inadequate, the Box-Jenkins methodology espouses returning to the *model identification* stage to re-examine the appropriateness of the ARIMA model to the data series. Validating the model involves examining the residuals using charts and test statistics, which separately form the qualitative and quantitative bases of analysis, respectively, and together constitute the graphical residual analysis toolkit.

In following the Box-Jenkins methodology, if a practitioner has developed a robust model, then the residuals should themselves behave like a univariate process. Moreover, the residuals from a robust model will exhibit random drawings, fixed distribution, fixed location and fixed variance. A powerful technique to perform residual analysis is to develop and examine 4-plots. 4-plots consist of a collection of exploratory graphical analysis (EDA) graphs which are especially suitable for the graphical residual analysis of univariate models. A 4-plot typically consists of a run sequence plot, a lag plot, a histogram, and a normal probability plot, and helps answer questions relating to the underlying assumptions of the univariate model. Table 5-2 summarizes the components of the 4-plot.

Graph	Axis Definitions	
	Y-axis	*X-axis*
Run Sequence Plot	Residuals	Index
Lag Plot	Residuals	Lagged Residuals
Histogram	Frequency	Residuals
Normal Probability Plot	Ordered Residuals	Normal order medians

Table 5-2 Summary of 4-plot Components

Below, each plot is illustrated with an example and sample analysis.

RUN SEQUENCE PLOT

Run sequence plots graphically summarize univariate data, and are effective in locating outliers and shifts in location and scale. An ideal run sequence plot of the residuals would be flat and non-drifting, allowing the fixed-location assumption to hold, and a vertical spread roughly the same over the entire plot, allowing the fixed-variation assumption to hold. Figure 5-5 shows the run sequences of the 'increasing' data set which contains five histories none of which are of fixed mean. To achieve stationarity, Box-Jenkins recommends differencing. Figure 5-6 shows a run sequence plot of history 2 following second-order differencing; it now exhibits fixed location, as desired.

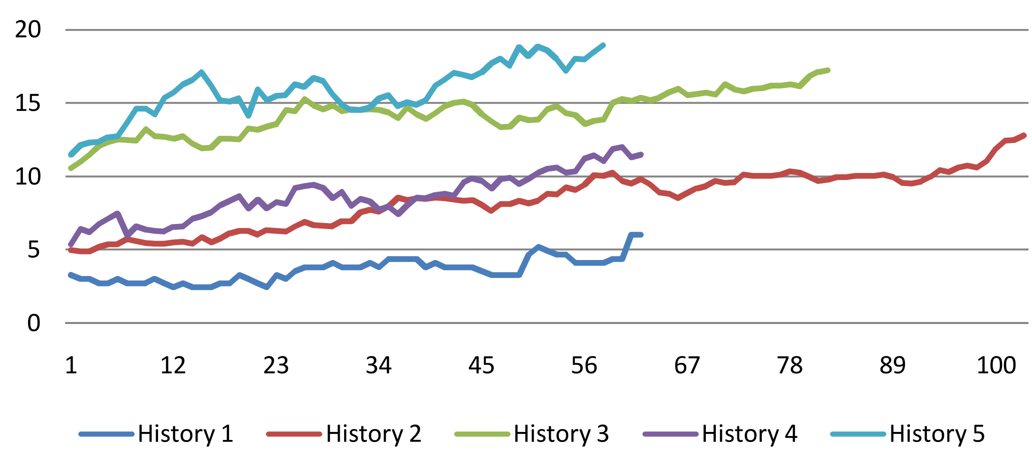


Figure 5-5 Run Sequence Plot of Increasing Dataset

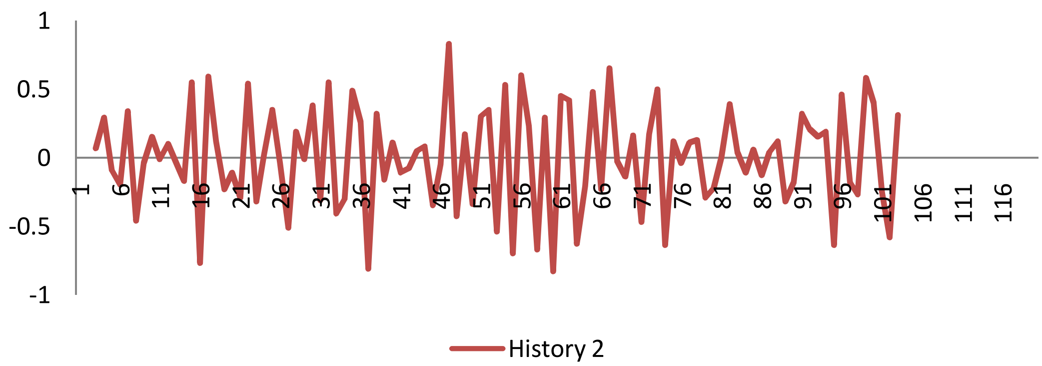


Figure 5-6 Run Sequence Plot of History 2 - Double Differenced

LAG PLOT

Lag plots effectively determine the randomness of data by plotting the residuals against a lagged version of themselves. The lag plot highlights any serial correlation that the model may not be capturing. Ideally, the lag plot would show white noise and no discernible trend, as Figure 5-7 shows.

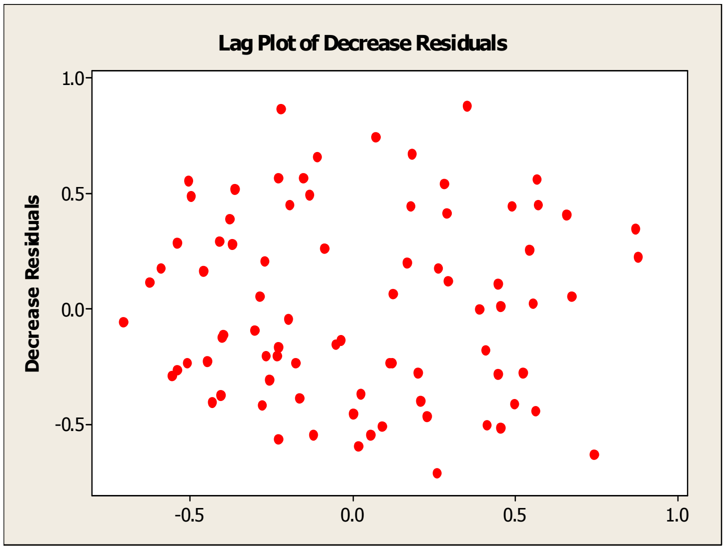


Figure 5-7 Lag Plot of the Decreasing ARIMA(1,2,0) Residuals

HISTOGRAM

Histograms depict the underlying characteristic of the distribution of the data. For white-noise residuals, the histogram would be Gaussian (or bell-shaped), signified by its unimodal, symmetrical nature. Figure 5-8 shows a fitted histogram

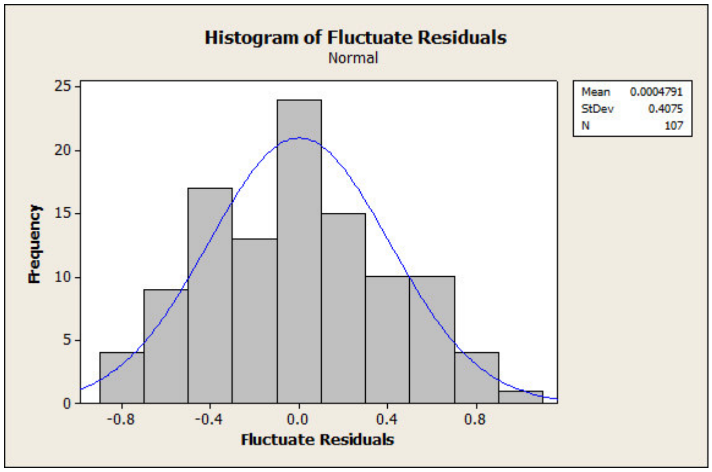


Figure 5-8 Histogram of the Fluctuate ARIMA(2,2,0) Residuals

NORMAL PROBABILITY PLOT

Probability plots (Figure 5-9) help assess whether a given dataset follows a specific distribution. To test the normality assumption of the residuals, a normal probability plot is generated which charts the ordered residuals against normal order statistic means, or more simply, the data's distribution against a theoretical normal distribution. If normal, the plot will show a straight line; points that plot far away from the normality line therefore indicate departures from normality. Lines that curl indicate kurtosis; lines that are biased to the left or to the right of x-axis indicate skew.

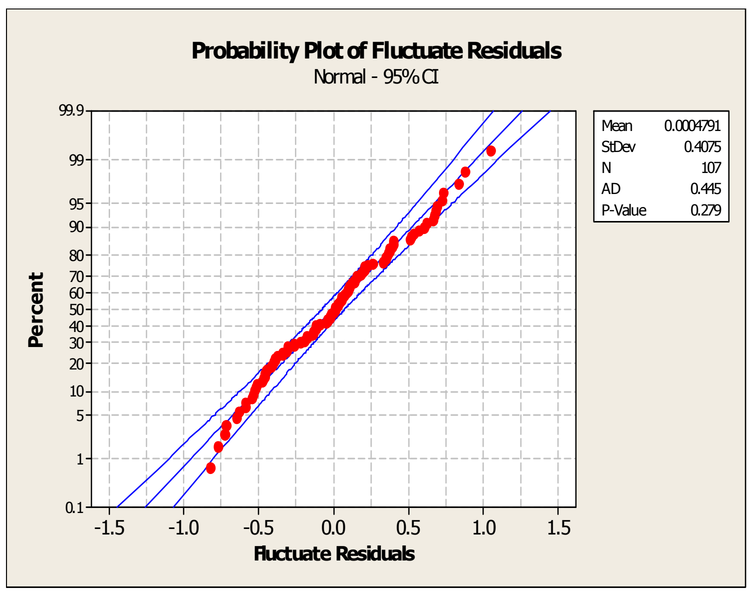


Figure 5-9 Normal Probability Plot of Fluctuate ARIMA(2,2,0) Residuals

6. APPLYING THE BOX-JENKINS METHODOLOGY

In this section, equations that were modified to accommodate pooling were used in the Box-Jenkins framework; model identification, estimation and validation was conducted using Minitab commands and custom MATLAB code. The model developed through use of pooling historical data was then used to forecast future data. In addition, the paper compares the relative forecasting strengths of the pooled models to the strength of single history models as generated by Minitab. The residuals calculated from the various models will be compared amongst each other to test the fit of the data and the implications of using multiple pooled histories as opposed to single histories.

6.1. MODEL IDENTIFICATION

The Box-Jenkins approach was initiated by generating simple run sequence plots which test for stationarity on the various histories.

RUN SEQUENCE PLOTS OF POOLED DATA

Based on the run plots of the following groups of histories (Figure 6-1) the following initial observations were made:

1. The increasing and decreasing datasets do not exhibit any outliers or seasonal trends, however each graph clearly indicates non-stationarity,
2. The horizontal dataset does not indicate any outlier or seasonal trends, either, yet the data fluctuates around a mean and may be initially regarded as a stationary.

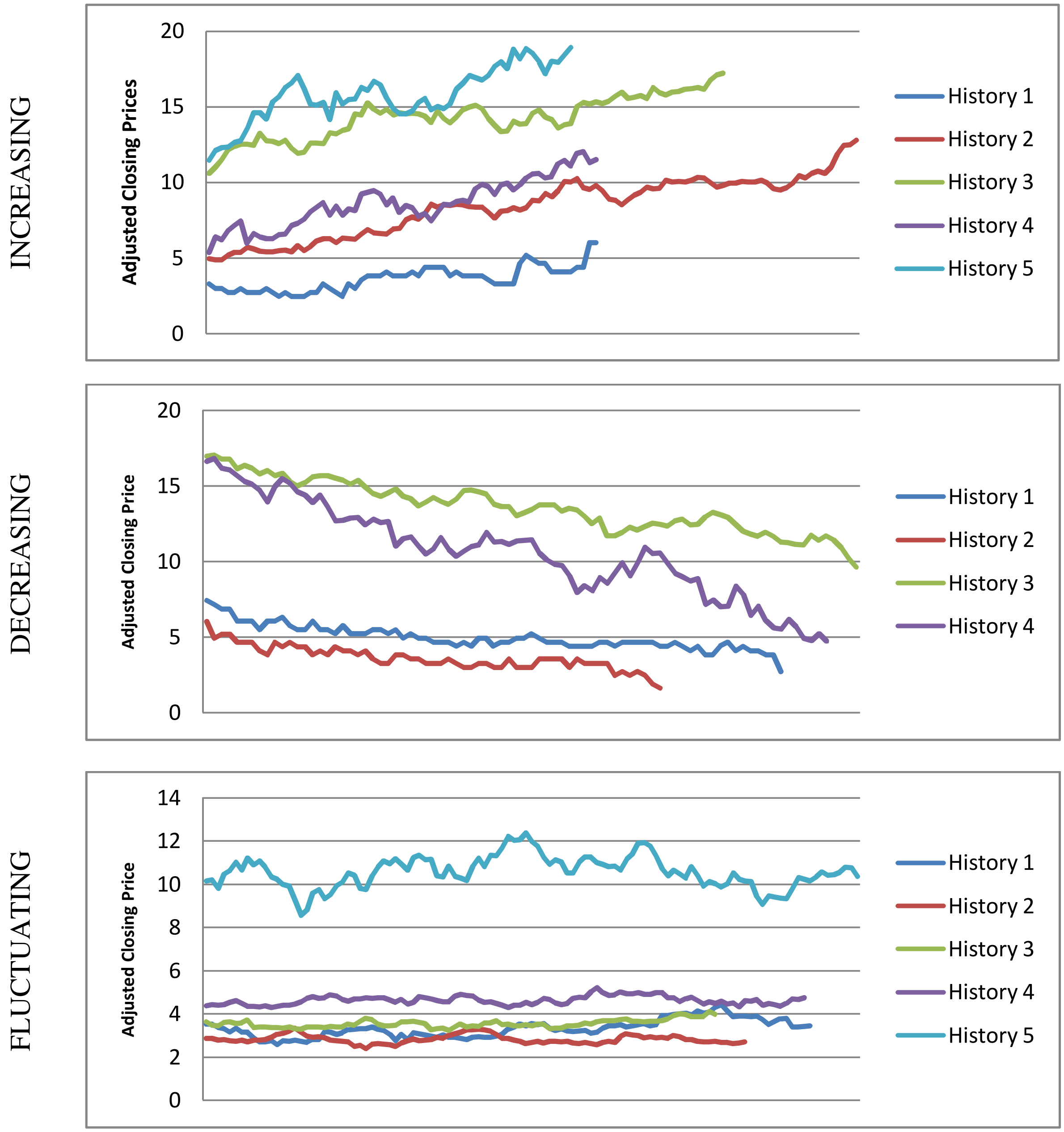


Figure 6-1 Run Sequence Plots for the 'Increasing', 'Decreasing' and 'Fluctuating' Datasets

Using the MATLAB program (Appendix A), pooled autocorrelation and partial autocorrelation plots for the initial raw data for each group were generated. This step was key in not only establishing stationarity but also in identifying the order of the autoregressive term and the order of the differencing.

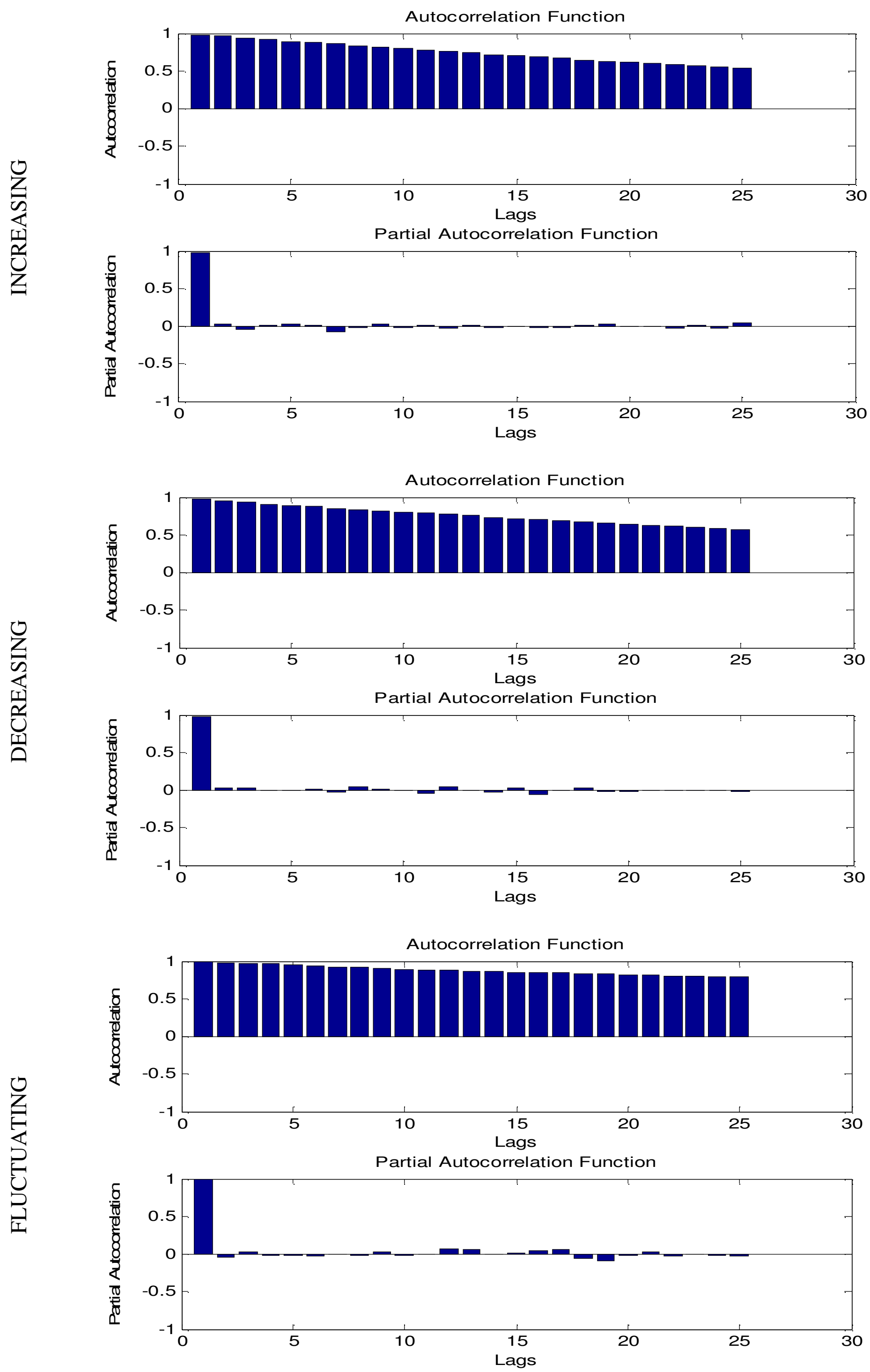


Figure 6-2 ACF and PACF Plots for Undifferenced Datasets

From the plots in Figure 6-2, the following general observations were made:

1. All groups display strong positive autocorrelation;
2. There is extremely weak decay especially with the horizontal data
3. The datasets are not stationary and need to be appropriately differenced

First order differencing was applied to the datasets, followed by second order differencing in order to impose stationarity. The following section covers the latter.

Run Sequence Plots of Second Order Differenced Datasets

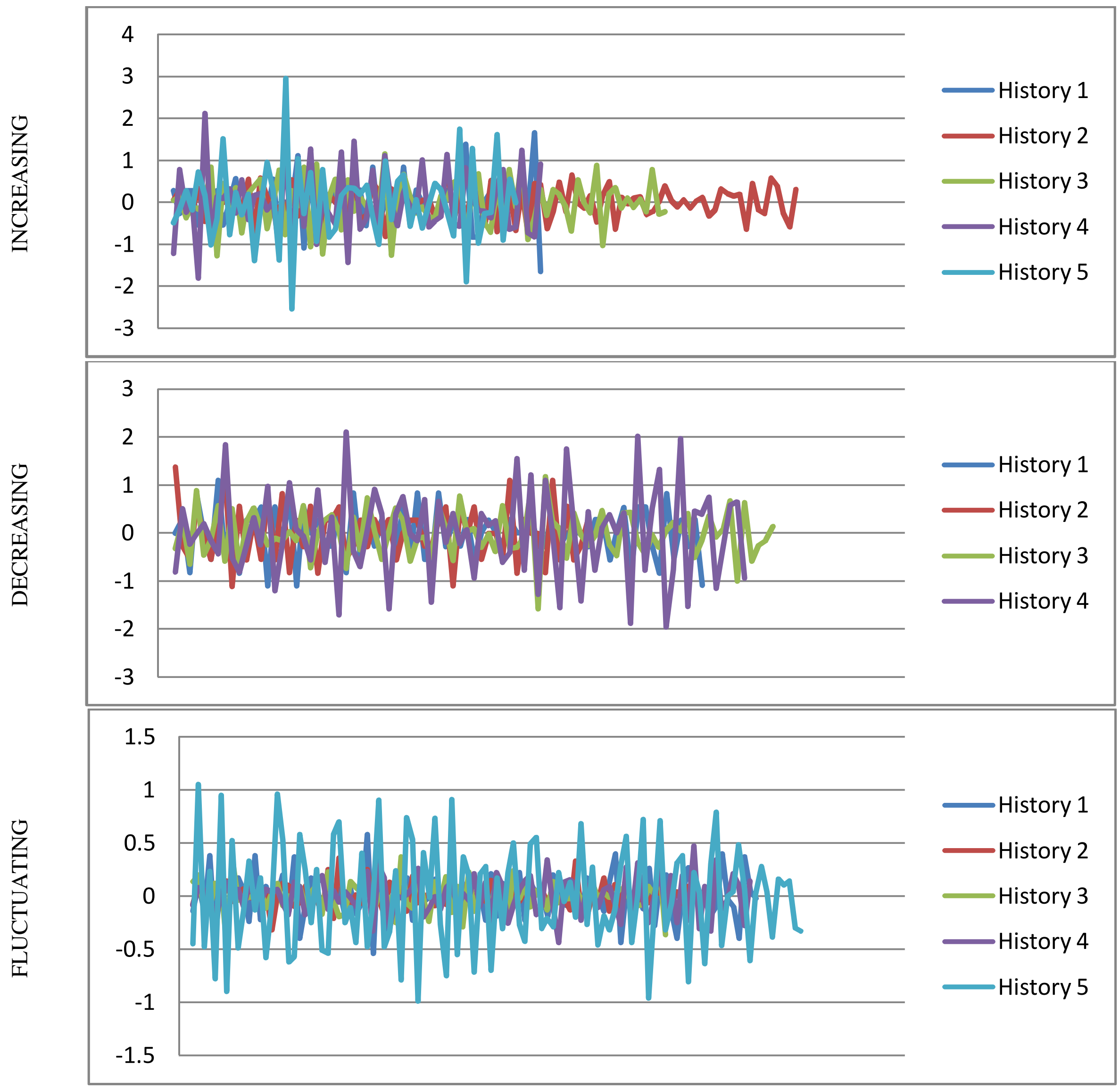


Figure 6-3 Run Sequence Plot of Second Order Differenced Datasets

The run plots of the differenced datasets (Figure 6-3) are mean-reverting and stationary. The plots indicate that the differenced data is less autocorrelated than the initial raw data.

ACF and PACF Plots of Second Order Differenced Datasets

The ACF and PACF plots of the datasets along with run sequence plots of the differenced data suggest that the differenced data are now stationary. The autocorrelation plots of each group suggests an AR(1) model, yet to examine other potential models, the partial autocorrelation plot needs to be consulted (Figure 6-4). The PACF plot for the increasing and decreasing histories concur with the results from the autocorrelation plots, both suggesting an AR(1) model. The PACF for the horizontal histories, however, strongly suggests that an AR(2) model may be better suited. Based on this visual inspection and reasoning, the following summarizes the orders of the autoregressive, differencing and moving average terms:

1. The increasing dataset will be modeled by ARIMA(1,2,0)
2. The decreasing dataset will be modeled by ARIMA(1,2,0)
3. The fluctuate dataset will be modeled by ARIMA(2,2,0)

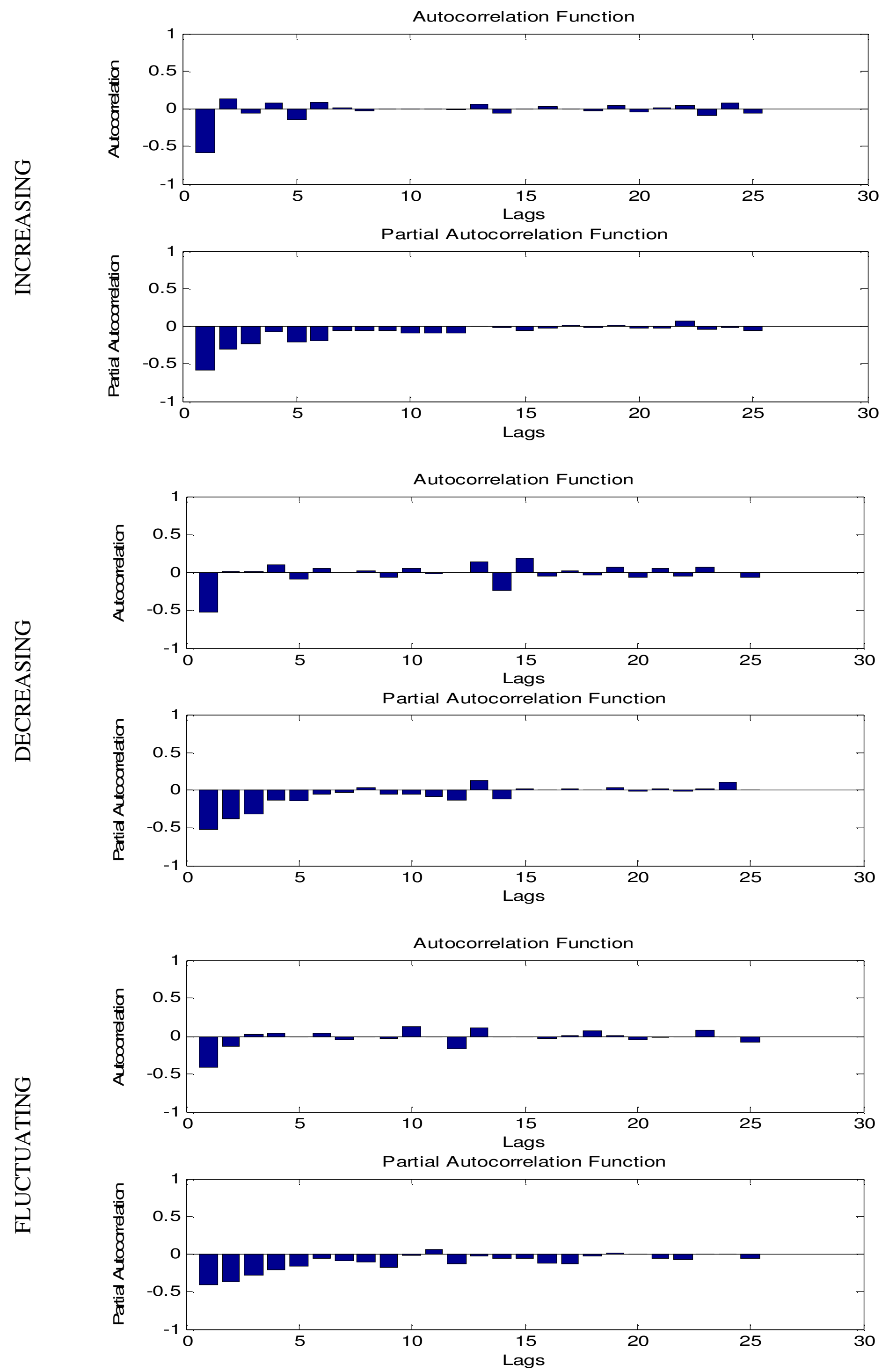


Figure 6-4 ACF and PACF Plots of Second Order Differened Datasets

6.2. Model Estimation

In order to estimate the parameters of the three different models, a custom MATLAB script was developed (Appendix B) to estimate the parameters using the Yule-Walker equations (page 35) using the multiple history (page 21). Executing the script with the models specified on page 46 yields the following estimates for the parameters:

Dataset	Model Specification	Model Parameters
Increasing	ARIMA(1,2,0)	AR (1) -0.5823
Decreasing	ARIMA(1,2,0)	AR (1) -0.5320
Fluctuating	ARIMA(2,2,0)	AR (1) -0.5601 AR (2) -0.3644

Table 6-1 Parameter Estimates for 'Increasing', 'Decreasing' and 'Fluctuating' Datasets

The parameter estimates obtained from the program are second order differenced. Therefore the equations were recast in terms of the original time series as follows:

Increasing

$$x_t - 2x_{t-1} + x_{t-2} = -0.5823\left(x_{t-1} - 2x_{t-2} + x_{t-3}\right)$$
$$x_t = 1.4177x_{t-1} + 0.1646x_{t-2} - 0.5823x_{t-3}$$

Decreasing

$$x_t - 2x_{t-1} + x_{t-2} = -0.5320\left(x_{t-1} - 2x_{t-2} + x_{t-3}\right)$$
$$x_t = 1.468x_{t-1} + 0.064x_{t-2} - 0.5320x_{t-3}$$

Fluctuating

$$x_t - 2x_{t-1} + x_{t-2} = -0.5601\left(x_{t-1} - 2x_{t-2} + x_{t-3}\right) - 0.3644\left(x_{t-2} - 2x_{t-3} + x_{t-4}\right)$$
$$x_t = 1.4399x_{t-1} - 0.2442x_{t-2} + 0.1687x_{t-3} - 0.3644x_{t-4}$$

Table 6-2 ARIMA Equations Recast into Original Time Series

6.3. Model Validation

In order to validate the parameter estimates for each of the three models, one-step ahead forecasts were generated using one of the existing histories from each model and the resulting residuals analyzed. This stage is key in determining the models' goodness-of-fit.

The run sequence plot contained within the increasing model's residual 4-plot (Figure 6-5) indicates that the residuals do not violate the constant location and scale assumption. It also shows that most of the residuals are within the (-1, 1) range. The lag plot indicates the residuals are not serially autocorrelated, and while the histogram does not give clear indication of a normal distribution, the poorness of fit may be attributed to the low number of residuals generated. The probability plot does show that the residuals are normally distributed with a *P*-value of 0.92 which is well above the critical value of 0.005.

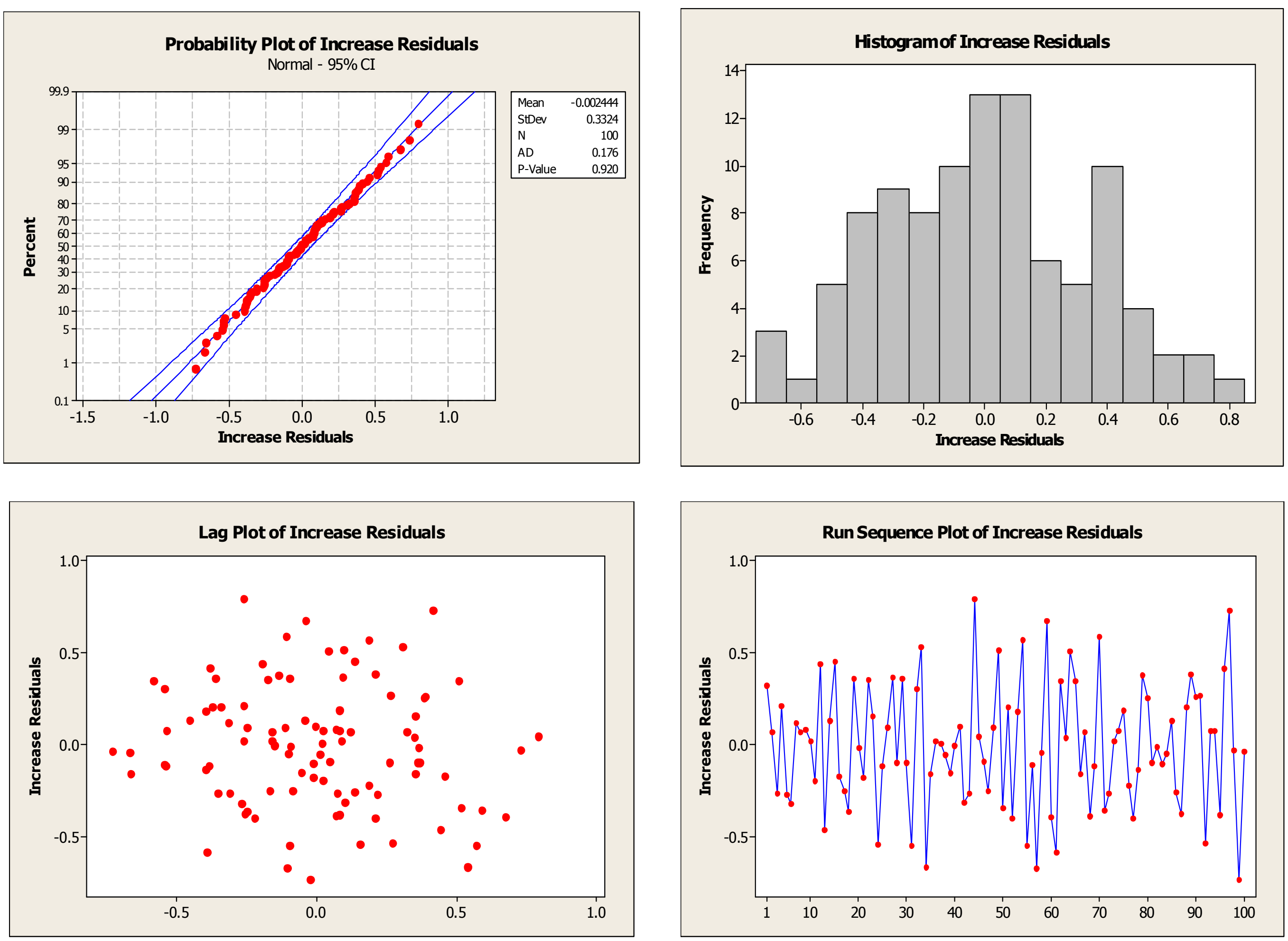


Figure 6-5 4-Plot of the Residuals of the Increasing ARIMA Model

Decreasing Model Residual Analysis

As with the residuals of the increasing model, the 4-plot of the residuals of the decreasing model (Figure 6-6) indicates a good fit. The lag plot indicates that there is no serial correlation while the run sequence plot is indicative of constant location and scale. The probability plot indicates a *P*-value of 0.103, well above the critical value.

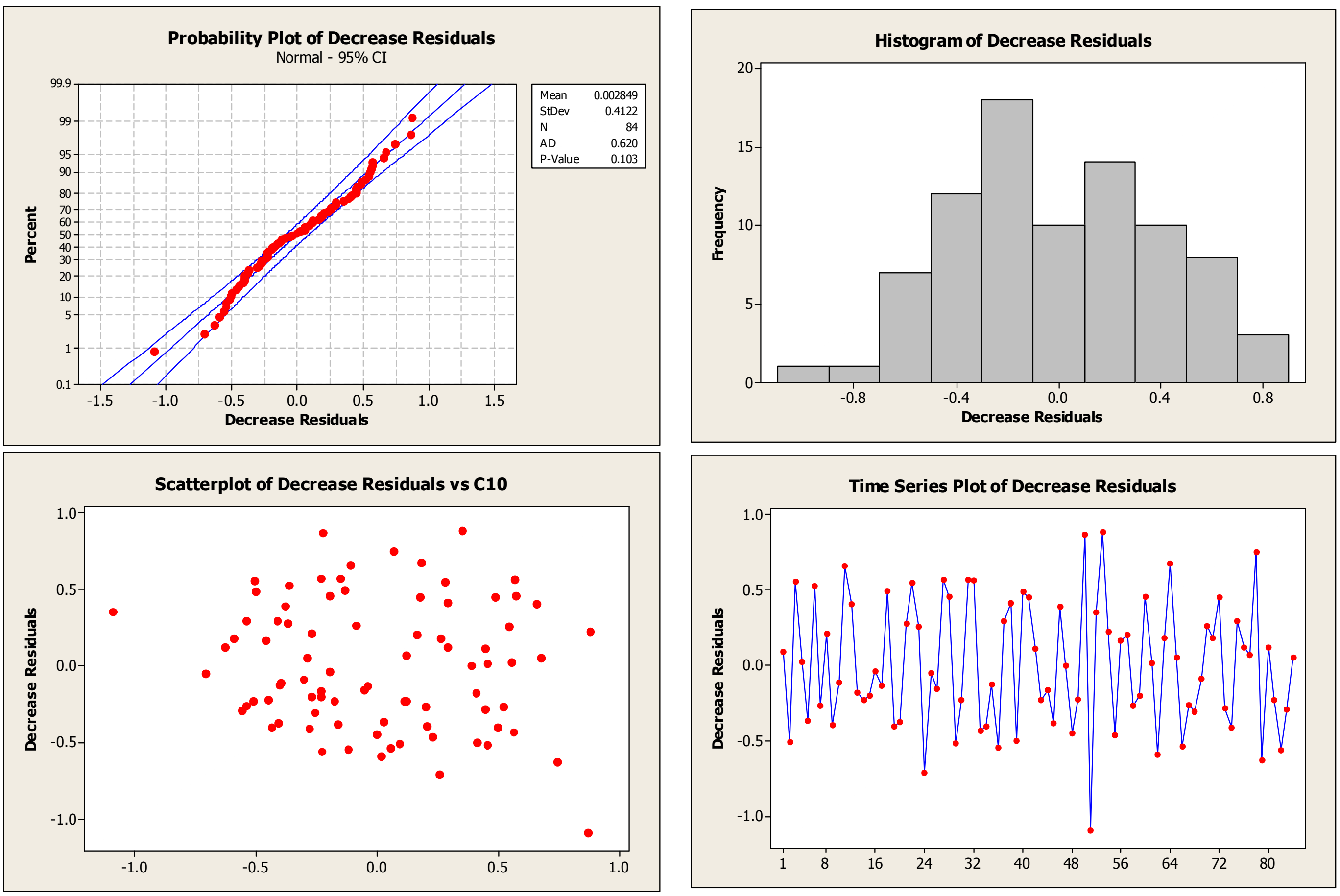


Figure 6-6 4-Plot of the Residuals of the Decreasing ARIMA Model

Analysis of the 4-Plot for the Residuals of the Fluctuating Model

The run plot for the horizontal model indicates that the residuals have constant location and scale. The lag plot indicates that there is no autocorrelation between lag 1 and the probability plot shows that the residuals follow a normal distribution with a p-value of 0.279.

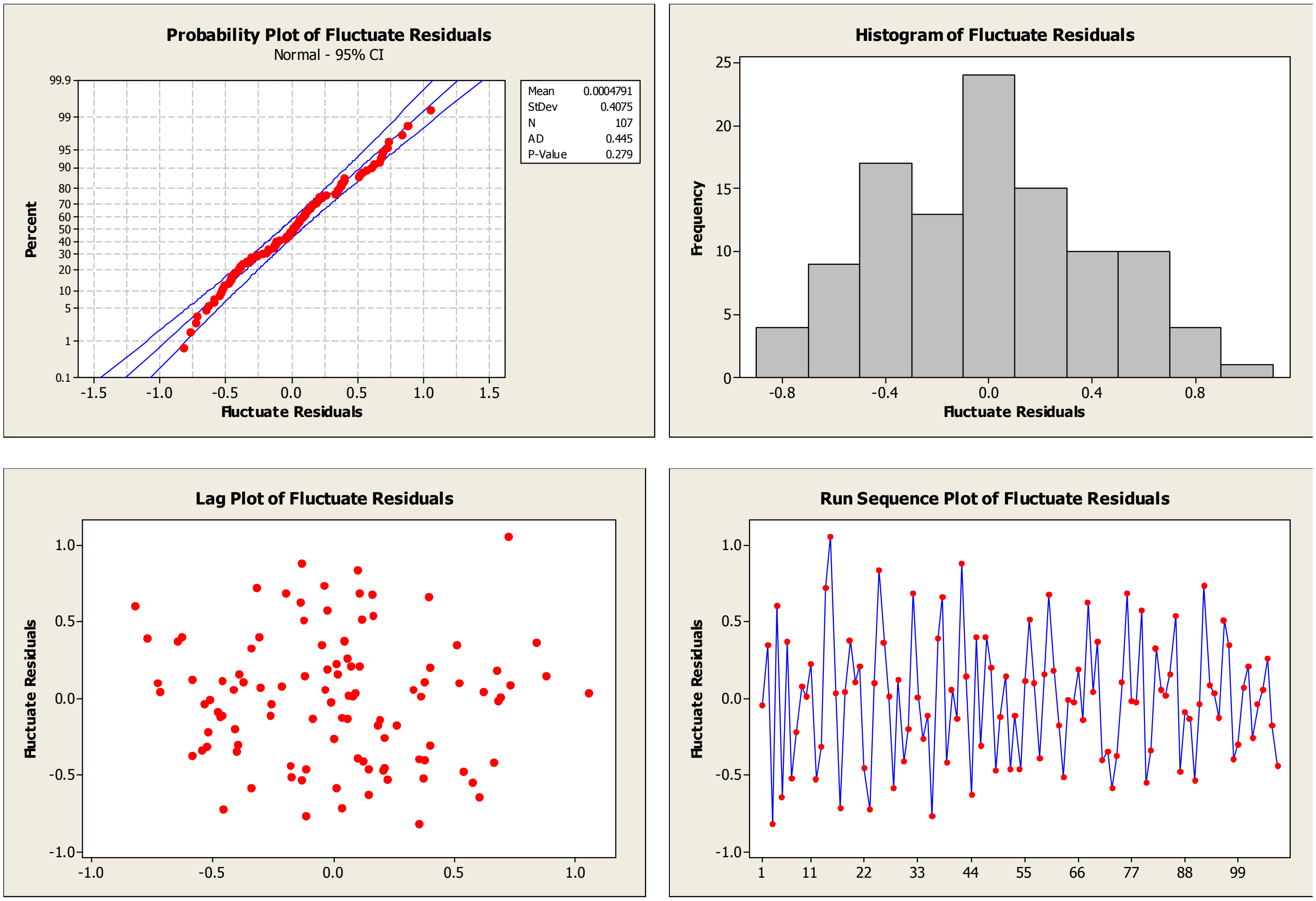


Figure 6-7 4-Plot of the Residuals of the Fluctuating ARIMA Model

While certain models fared better than others in the residual analysis, when comparing the normality plots, it is concluded that all models are satisfactory fits to the pooled histories.

6.4. Comparing Pooled and Single History ARIMA models

Using the same Box-Jenkins methodology presented in the prior section, single-history ARIMA models were built in Minitab using the longest history of each group. This was done in order to compare and contrast between the pooled and single-history models. Figure 6-8 provides the histories used to generate the one-step ahead forecasts; Table 6-3 provides the single-history and pooled history ARIMA model parameters side by side.

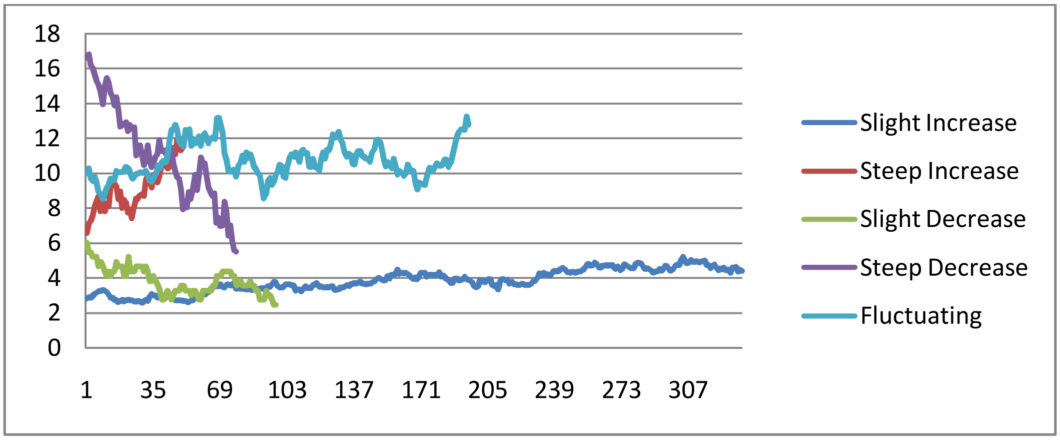


Figure 6-8 Time series Used for One-Step Ahead Forecasts

Dataset	Type	Model Specification	Model Parameters
Increasing	Pooled	ARIMA(1,2,0)	AR (1) = -0.5823
	Single	ARIMA(1,2,0)	AR (1) = -0.4628
Decreasing	Pooled	ARIMA(1,2,0)	AR (1) = -0.5320
	Single	ARIMA(1,2,0)	AR (1) = -0.4926
Fluctuating	Pooled	ARIMA(2,2,0)	AR (1) = -0.5601
			AR (2) = -0.3644
	Single	ARIMA(2,2,0)	AR (1) = -0.5275
			AR (2) = -0.4139

Table 6-3 Single and Pooled History Model Parameters for 'Increasing', 'Decreasing' and 'Fluctuating' Datasets

The new datasets were forecasted using both single and pooled history models. Residuals were then calculated for the various model forecasts, and probability plots were drawn to comparatively analyze the models' goodness-of-fit.

Comparison of Increasing Models

The normality plots show that the single history model was able to produce a better fit for the slightly increasing time series whereas the pooled history model was more adequate for the steeply increasing time series (Figure 6-9). It should be noted, however, that both models are technically acceptable at the 5% significance level.

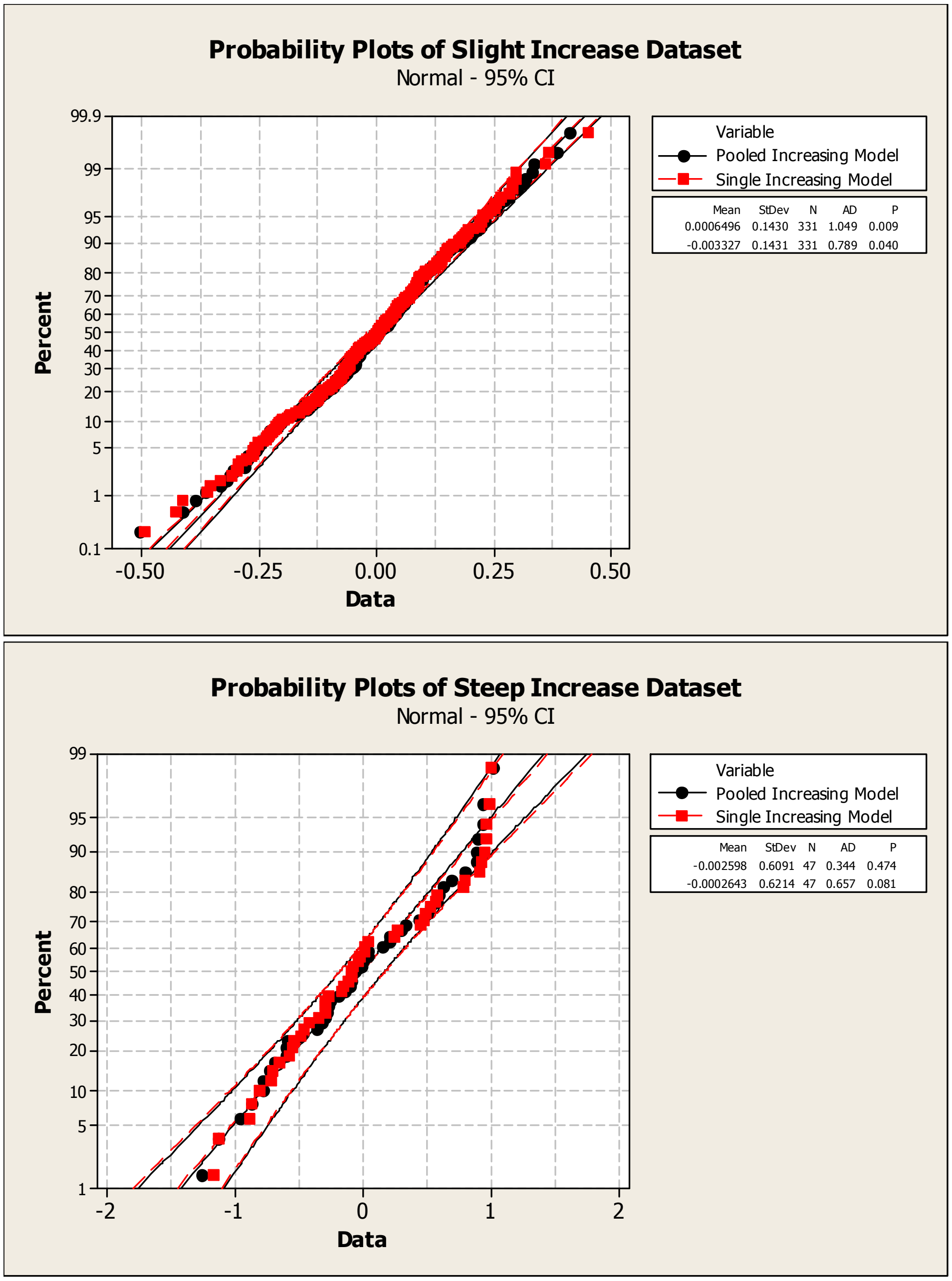


Figure 6-9 Normality Plots of the Pooled & Single History Models (Increasing)

COMPARISON OF DECREASING MODELS

The residual normal probability plots show that while both models perform reasonably well, the decreasing model using pooled histories is a slightly better fit as evidenced by higher *P*-values.

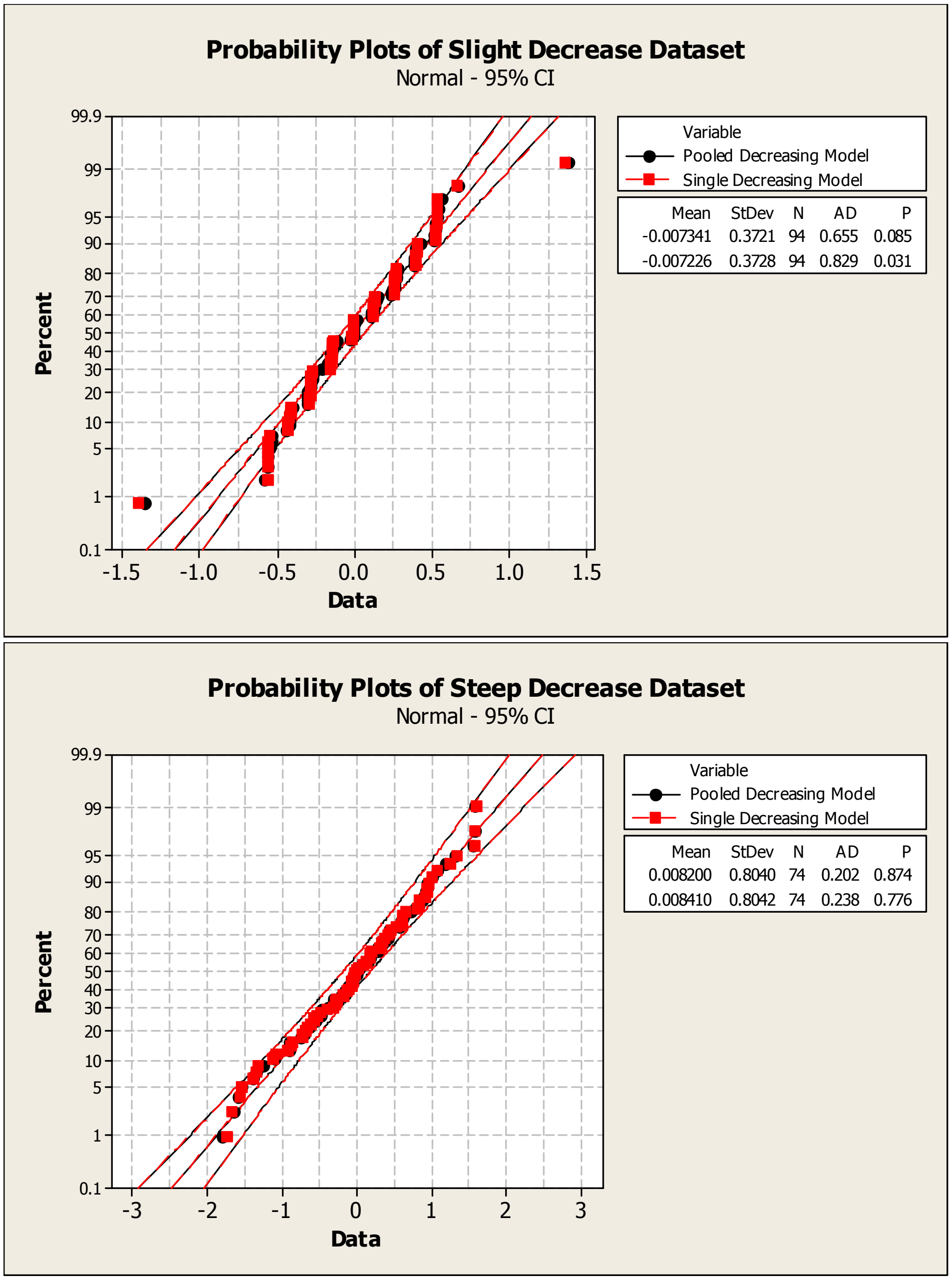


Figure 6-10 Normality Plots of the Pooled & Single History Models (Decreasing)

Comparison of Horizontal Models

The results of the probability plot generated from the residuals show again that both the pooled and single history models perform relatively well , yet in this specific analysis the horizontal model derived from the single history is a slightly better fit according to the *P*-value ($P_{single} = 0.544 > P_{pooled} = 0.540$).

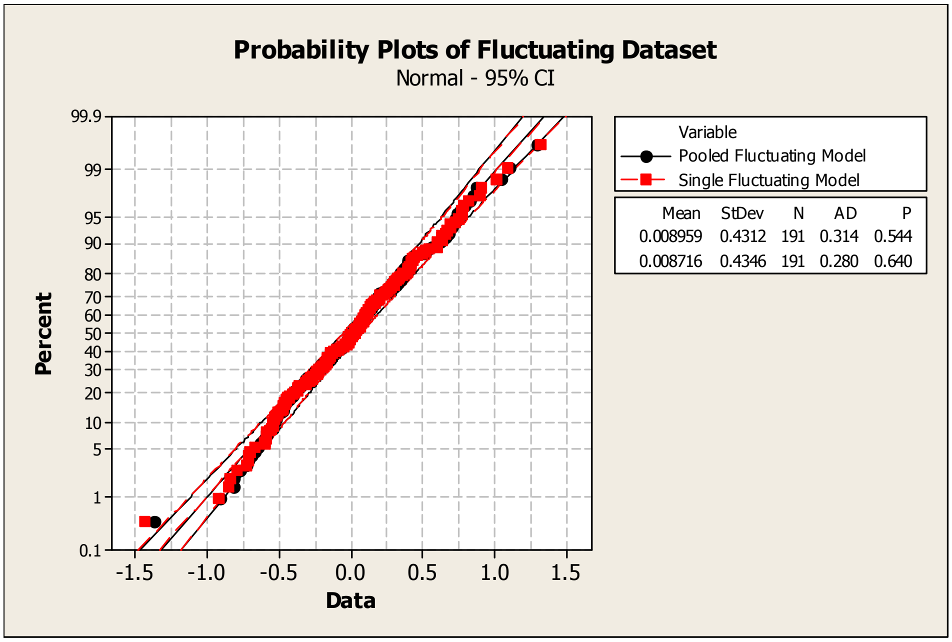


Figure 6-11 Normality Plot of the Single and Pooled History Models (Horizontal)

While both models are a reasonable fit to the data sets, what can be elucidated from the comparison is that using pooled data takes into account a much larger overall dataset and may be considered a better estimator. This can be true since modeling based on a single history will result in having to rely on fewer data points. In contrast, being able to pool segments that indicate similar trends helps expands the dataset upon which the model is based, thereby increasing the model's adequacy as measured through normality plots and *P*-values.

6.5. Statistical Process Control (SPC) Results

The benefit of applying SPC techniques to the stock data was twofold—firstly, the charts indicated whether a model being used to generate day-ahead forecasts is still relevant, and secondly, whether the stock is signally some type of investment action (buy, sell or hold). The design of CUSUM control charts for each type of trend (increasing, decreasing or horizontal) to monitor shifts in mean and variance helped make inferences into the stock movement. The assumption that the pooled historical data and resulting models were sufficiently robust and an in-control was tested by generating Individuals and Moving Range (I-MR) charts of the residuals of each model.

SHEWHART'S I-MR CHARTS

In order to design the CUSUM control chart, estimates for the process standard deviation need to be computed; these may be obtained from the I-MR charts (Figure 6-12).

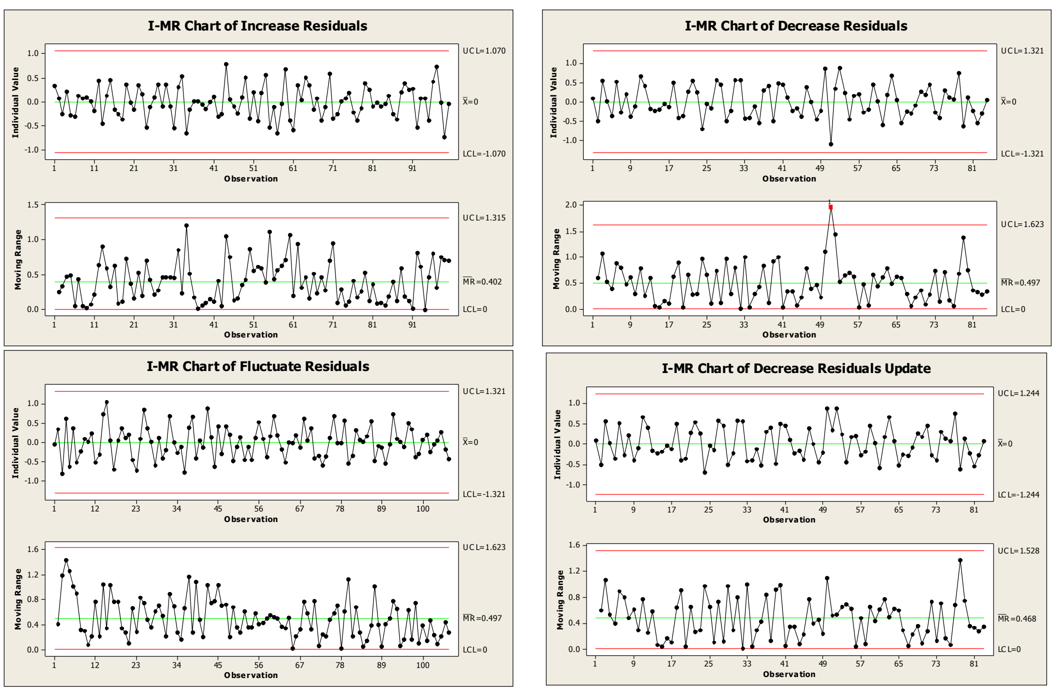


Figure 6-12 I-MR Charts of the Residuals of the Models

The I-MR charts show that the model residuals are in control expect for the 'decreasing' model where a single point is out of control. To remedy this situation, that single data point was removed[8], and the I-MR chart for the 'decreasing' model regenerated (bottom right quadrant of Figure 6-12). The updated I-MR chart shows the 'decreasing' model to now be 'in-control'; estimates of the sample standard deviation are provided in Table 6-4.

Standard Deviation Estimates	
Increasing Model	0.356691
Decreasing Model	0.414625
Horizontal Model	0.440377

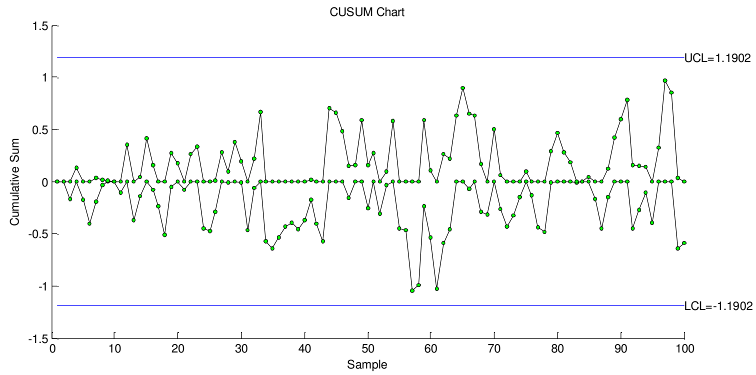
Table 6-4 Estimates of the Sample Standard Deviation as Calculated by Minitab

To design the CUSUM control charts, a MATLAB script executed (Appendix C)to test a range of values for the Average Run Length (ARL) and δ which is the shift size to be detected. Table 6-5 tabulates the results which were obtained from the MATLAB script; also, the resulting 'in-control' CUSUM charts are provided.

[8] See (Montgomery, 2005) for a detailed treatment of single out-of-control points.

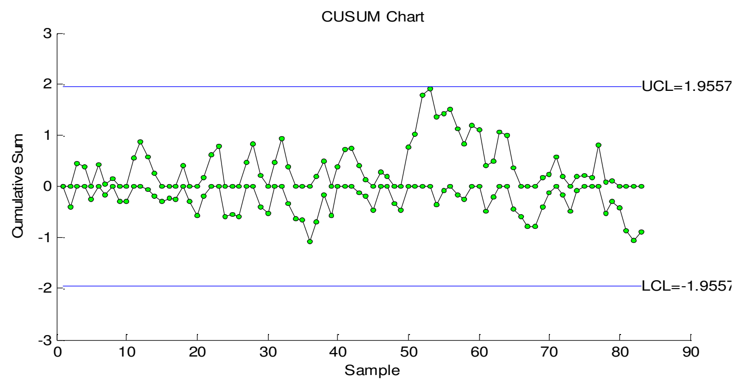
Increasing Model Design

```
H = 1.1902
h= 3.3369
k= 0.0892
ARL= 25
Delta= 0.5000
Sigma= 0.3567
Mean= 0
```



Decreasing Model Design

```
H= 1.9557
h= 4.7167
k= 0.1037
ARL= 60
Delta= 0.5000
Sigma= 0.4146
Mean= 0
```



Fluctuating Model Design

```
H= 1.5894
h= 3.6093
k= 0.1101
ARL= 30
Delta= 0.5000
Sigma= 0.4404
Mean= 0
```

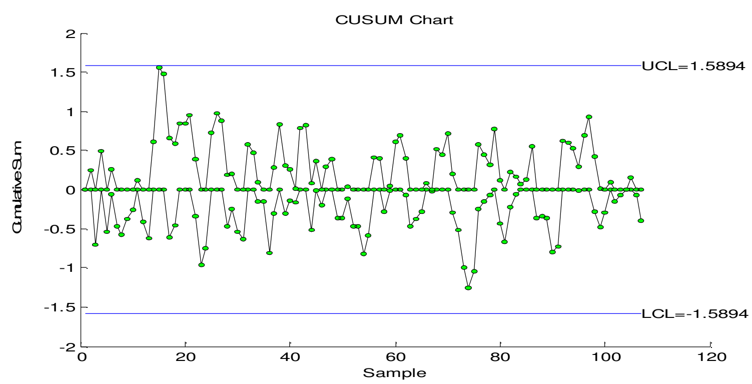


Table 6-5 CUSUM Design Parameters as Outputted by MATLAB

CUSUM ANALYSIS

Using the same data provided in the future historical datasets (Figure 6-8 on page 52), CUSUM analysis techniques were applied to the residuals of the model forecasts to investigate the CUSUM charts' ability to detect changes in the stock's behaviour. What follows is a discussion of the CUSUM chart and MATLAB output following each run.

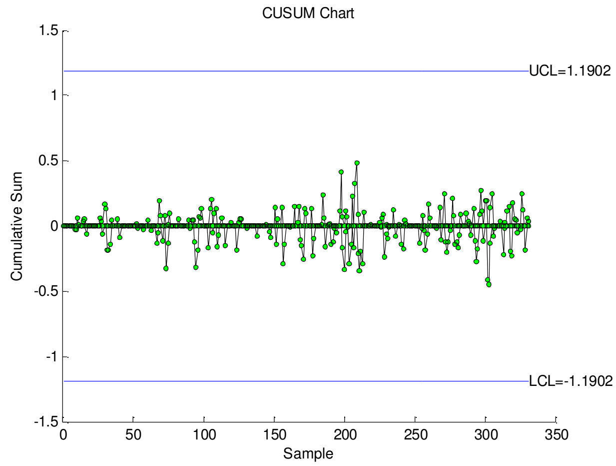
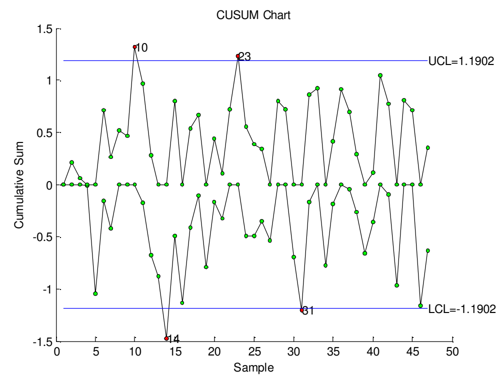


Figure 6-13 MATLAB Output of 'In-Control' CUSUM Chart (Slightly Increasing)

The ‘slightly increasing’ model residuals for the slight increase data show that the process is ‘in-control’, and an adequate estimator for datasets that have a slight gradual increase. The stock could be viewed at this point to be healthy (Figure 6-13). The steeply increasing time series CUSUM chart (Figure 6-14) however, indicates an ‘out-of-control’ process. This may indicate that the stock is behaving erratically and is currently in a volatile situation. It can also be noted as the best time to either buy or sell, depending on the strategy preferred by the investor.

The model residuals for the slight decrease data show that the process is ‘in-control’ (Figure 6-15), and is an adequate estimator for datasets that have a gradual decrease.




```
Test failed at points:
10 14 23 31
```

Figure 6-14 MATLAB Output of 'Out-of'Control' CUSUM Chart (Steeply Increasing)

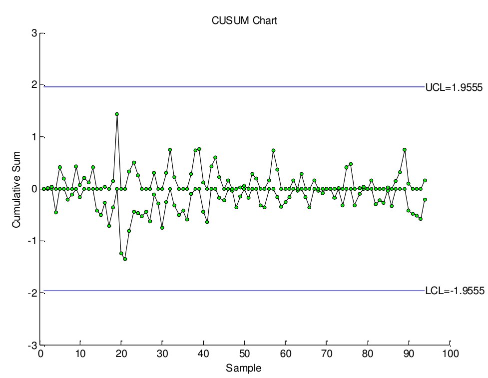
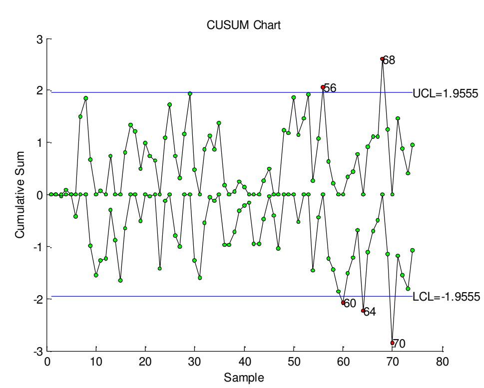


Figure 6-15 MATLAB Output of 'In-Control' CUSUM Chart (Slighly Decreasing)




```
Test failed at points:
56 60 64 68 70
```

Figure 6-16 MATLAB Output of 'Out-of-Control' CUSUM Chart (Steeply Decreasing)

When fed a steeply decreasing time series, the CUSUM indicates an ‘out-of-control’ process for the residuals (Figure 6-16). This can indicate that the stock is behaving erratically and is currently in a volatile situation. The major deviations indicate that the model may need to be updated, or that the stock is producing uncharacteristic behaviour. When presented with horizontal data, the CUSUM chart indicates a process that is ‘in-control’ Figure 6-17.

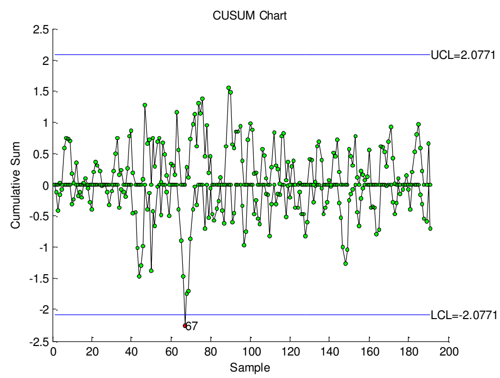


Figure 6-17 MATLAT Output of an 'Out-of-Control' CUSUM Chart (Horizontal)

7. THESIS SUMMARY

In this section of the thesis, a high level overview of the activities performed to arrive at the control charts is presented. Also, financial interpretations of the CUSUM chart signals are discussed in the context of the investor's personal risk profile. The limitations of the modeling framework are briefly touched upon, and real world applications are presented.

7.1. REVIEW OF BOX-JENKINS METHODOLOGY AND SPC

Broadly speaking, the thesis was tackled in two phases; first, ARIMA models were generated for each of three 'pools' of financial data depicting either increasing, decreasing or horizontal behaviour. Second, the model residuals were monitored on CUSUM control charts to detect signals that would be indicative of a change in the underlying stock trends.

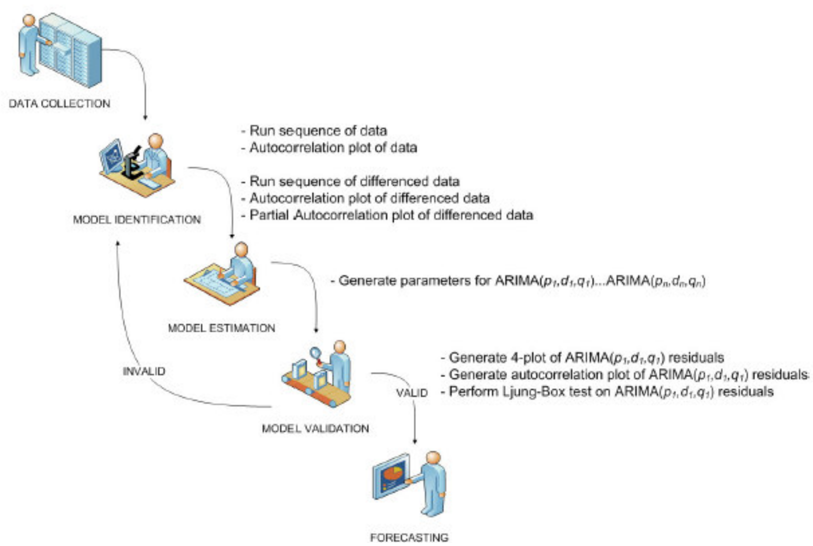


Figure 7-1 Graphical Overview of the Box-Jenkins Methodology

The model generation phase was conducted in the standard Box-Jenkins methodology (Figure 7-1) which was used to identify, estimate and validate ARIMA models. Run sequence, ACF and PACF plots were used to visually inspect the data, and second-order differencing applied to impose stationarity. Furthermore, the orders of the autoregressive terms were identified. To accommodate the pooled data, the sample mean and sample autocovariance formulas were modified before being in a MATLAB script that exploited the Yule-Walker system of equations to estimate the model parameters. The models were tested against single histories to validate the goodness-of-fit; this was primary achieved by examining the P-values of the normality plots of the model residuals. The models were accepted as being satisfactory based on residual analysis contained in 4-plots.

The control chart design and testing phase of the thesis started with the generation of the model residuals' I-MR charts to test whether the process was in-control. Once this requirement addressed, CUSUM control charts were developed based on two input parameters, namely Average Run Length (ARL) and the critical shift δ. The CUSUM control charts again confirmed that the residual process was in control. To simulate a change in stock trend, new histories were used to back-forecast, and the CUSUM charts monitored for signals that would indicate a change in the underlying process.

7.2. Interpretation of CUSUM Control Chart Signals

Although the value of control charts in manufacturing and production system cannot be underestimated, it is important to be recognize that control chart signals do not tell the operator what corrective action to take—rather, control charts only bring to the operator's attention a *possible* shift in the process. Therefore, control chart signals provide the operator a reliable method to investigate changes in process, and require the operator to put the signal in the context of prevailing conditions.

Within the financial markets context of this thesis, the signals are even more dependent on accurate diagnosis and interpretation by the operator (typically, a day trader or institutional investor) because of the multitude of trading strategies one could subscribe to. To illustrate this point, a few signalling scenarios, possible causes along with suitable investment decisions are presented in the following.

If a control chart exhibits multiple out-of-control points either above the upper control limit or below the lower control limit, it may be inferred that the stock has changed behaviour i.e., if the residuals are currently being generated based on a 'fluctuating' model, and the control chart plots multiple points above the control limit, it may be inferred that the stock is now exhibiting an upward trend. Depending on the risk-return profile of the investor, a 'buy' decision may be made (this strategy is comically called 'herd mentality') to capitalize on the market rally. On the other hand, if the investor is following a contrarian approach, he or she will sell the stock (or take up a 'short' position) believing that the rally will be short lived—and that prices will fall—thereby ensuring capital gains. This scenario succinctly illustrates the dual-nature of a simple control chart signal.

Investment decisions depend on the investor's risk-return profile, and a conservative trader may interpret an increasing trend signal as a 'sell' signal assuming the stock becoming overvalued, a horizontal trend as a 'hold' signal, and a decreasing trend as a 'buy' signal, assuming the stock will be undervalued. The conservative strategy is polar to that of an aggressive trader. An aggressive trader will interpret an increasing trend signal as a buy signal, assuming the stock is undervalued, a horizontal trend as a hold signal, and a decreasing trend signal as a sell signal, assuming the stock is overvalued.

7.3. MODEL LIMITATIONS

A risk inherent to all modeling techniques is the simplification of real-world phenomena into less complex model behaviour; instead, it is more prudent to address the shortcoming of the model in light of its computational or conceptual ease. There are two model limitations that were encountered which the authors would like to address in greater detail.

While conducting a review of the literature, the authors came across more sophisticated mathematical models which were even more suitable for financial application—for example, the GARCH family of models, with its heteroskedasticity modeling capability, would make for a more accurate model since homoskedasticity behaviour is rarely seen in the financial markets. On the contrary, time-variant variance, which is perhaps a defining hallmark of any stock index, is poorly modeled by standard ARIMA models due to the homoskedasticity assumption. However, the mathematical complexities of the GARCH model put it out of the reach of this undergraduate thesis.

A second constraint was the limited data set which was available for back-forecasting. It must be noted that using histories which were used to build the model to test its accuracy is almost a self-fulfilling fallacy: the predictive power of a model would ideally be tested against completely new histories that played no part in model building and estimation. Because the stock history for Yamana Gold Corp. was comparatively short-lived, it forced the use of histories that had earlier been used (either in part or in full) to estimate the parameters of the model. The authors point out that using a longer-lived index such as the New York Stock Exchange or the Dow Jones Industrial Average would help address this limitation, and would be a central consideration for any future work.

7.4. Real World Applications

This data modeling and financial control chart system is intend to be practical from a practitioner's point-of-view, who has been defined as an active day-trader or an institutional investor with a technical trading mandate. In the section below, one possible configuration of this system is presented, and how it adds value to the investment decision process.

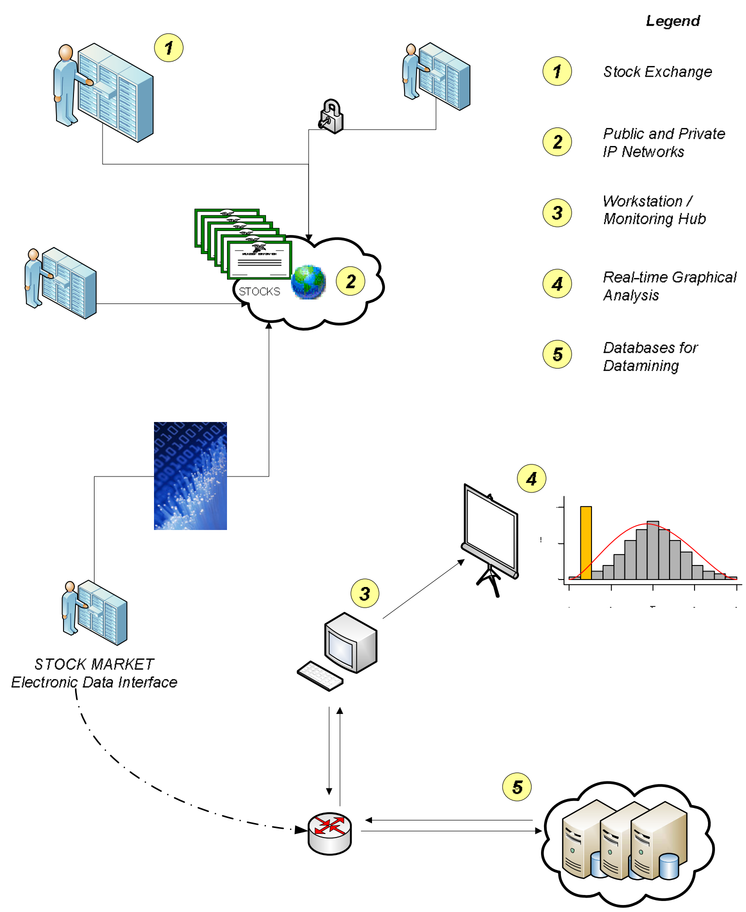


Figure 7-2 Workstation Configuration of Modeling and Monitoring System

In the figure above, the modeling and monitoring software would run on node 3; here the operator performs modeling activities and may either continuously monitor the control charts or setup an algorithmic to enable automated trades based predefined conditions.

8. LITERATURE REVIEW

In addition to studying Box-Jenkins' *Time Series Analysis: Forecasting and Control*, three published papers were reviewed, one each by the authors of this paper. Here, reviews of the three published papers are presented in case-study format along with examples of the mathematical models described within each.

8.1. Estimating Stock Market Volatility Using Asymmetric GARCH Models

Dima Alberg, Haim Shalita and Rami Yosef
Reviewed by
Imran Mohammed

Executive Summary

This paper takes an empirical analysis approach to forecasting the mean and variance of the Tel Aviv Stock Exchange (TASE) using various GARCH models. GARCH models allow users to model time series' serial volatility dependence, in contrast to ARIMA models which assume homoskedasticity i.e., constant variance. The paper is relevant to the thesis in that it develops and applies univariate time series models to financial data and provides a published, acceptable methodology on how to compare two forecasting models. Their findings suggest that the EGARCH model is the most accurate ARCH model, as applied to two Tel Aviv Stock Exchange indices.

Introduction

Volatility clustering, the leverage effect and excess kurtosis are empirically observed in financial market data, and accurately capturing market volatility with these complications is pertinent to many in the wealth, risk and investment management roles. Following Engle's

(1982) breakthrough work on the Auto-Regressive Conditional Heteroskedasticity (ARCH) model which addressed volatility clustering, the generalized ARCH (GARCH) model proposed by Bolleslev (1986) superseded the Exponentially Weighted Moving Average (EWMA) model that was being used at the time. Nelson's (1991) nonlinear, exponential extension of the GARCH model (EGARCH), Glosten et. al's GJR-GARCH (1993) and Ding et. al's Asymmetric Power ARCH (APARCH) loosen the symmetric restrictions on the ARCH/GARCH models by capturing negative biases in stock price movement due to volatility movements i.e., the leverage effect. Although heteroskedasticity explains some of the fat-tail behaviour, GARCH models often fail to fully capture the fat tails observed in asset return series, and therefore, to compensate for this limitation, fat-tailed distributions such as Student's t-distribution have been applied to GARCH modeling. Engle's original Gaussian implementation has now been extended to use the Student's *t*-distribution which better captures the leptokurtosis seen in financial data.

The paper investigates forecasting performance of GARCH, EGARCH, GJR-GARCH and APARCH models, each with three different underlying distributions: Gaussian, the Student's t-distribution and the non-central Student's t-distribution, using data from the TA100 and TA25 indices. Goodness-of-fit of the models were measured against test statistics while forecasting performance was measured against MSE, MedSE, MAE, AMAPE and TIC

Mean Square Error	$\frac{1}{h+1}\sum_{t=S}^{S+h}\left[\widehat{\sigma}_t^{\,2}-\sigma_t^{\,2}\right]^2$
Median Squared Error	$\mathrm{Inv}(f_{Med}(e_t))$ where $e_t = \left[\widehat{\sigma}_t^{\,2}-\sigma_t^{\,2}\right]^2$ and $t \in [S, S+H]$
Mean Absolute Error	$\frac{1}{h+1}\sum_{t=S}^{S+h}\left|\widehat{\sigma}_t^{\,2}-\sigma_t^{\,2}\right|$

Adjusted Mean Absolute Percentage Error	$\frac{1}{h+1}\sum_{t=S}^{S+h}\left\lvert\frac{\widehat{\sigma_t}^2-\sigma_t^2}{\sigma_t^2}\right\rvert$
Theil's Inequality Coefficient	$\frac{\sqrt{\frac{1}{h+1}\sum_{t=S}^{S+h}[\widehat{Y_t}-Y_t]^2}}{\left(\sqrt{\frac{1}{h+1}\sum_{t=S}^{S+h}[\widehat{Y_t}]^2}-\sqrt{\frac{1}{h+1}\sum_{t=S}^{S+h}[Y_t]^2}\right)}$

Table 8-1 Quantitative Forecasting Performance Measures

Data and Computation

3058 observations from the TA25 index and 1911 observations from the TA100 index were obtained; both are daily time series data. A price-to-return transformation was applied to the data to produce a stationary time series. Anticipating higher trading volumes on the first (Sunday) and last (Thursday) day of trading week, two regressions were used to isolate said 'calendar effects'. As suspected, the regressed sample descriptive statistics indicate skewness, excess kurtosis and Jarque-Bera statistics that are consistent with non-normally distributed returns. The estimation and forecasting were performed using G@RCH 2.0: an open-source toolkit based on the Ox mathematical programming language. Parameter estimation was accomplished computationally by relying on maximum likelihood estimators.

Model Specification

Conditional, time-variant volatility is captured in Engle's basic ARCH model by means of a quadratic function of the lagged values of the 'shocks' or 'surprises'. The core model is: $\sigma_t{}^2 = \alpha_0 + \sum_{i=1}^{q}\alpha_i\varepsilon^2{}_{t-i}$ when $\varepsilon_t{}^2 = \sigma_t{}^2 z_t{}^2$ such that z_t is χ^2 distributed. If an autoregressive moving average (ARMA) model can be applied to the error variance, then we've arrived at Bolleslev's (1986) broader GARCH model which parameterizes the above into two variables *p* (order of the variance term) and *q* (order of the error term) as follows:

$$\sigma_t{}^2 = \omega + \sum_{i=1}^{q}\alpha_i\varepsilon^2{}_{t-i} + \sum_{j=1}^{p}\beta_j\sigma^2{}_{t-j}$$

where α_i, β_j and ω are parameters to be estimated.

While more flexible than ARCH, GARCH is limited by its symmetry in that it cannot fully capture the leverage effect. To remedy the notion of larger negative returns associated with strong volatility movements than positive returns, asymmetric ARCH models introduce additional parameters that address the financial leverage. The APARCH model is expressed:

$$\sigma_t{}^{\delta} = \omega + \sum_{i=1}^{q} \alpha_i (|\varepsilon_{t-i}| - \gamma_i \varepsilon_{t-i})^{\delta} + \sum_{j=1}^{p} \beta_j \sigma^{\delta}{}_{t-j}$$

where $\delta > 0$ and $-1 < \gamma_i < 1$. If we fix $\gamma_1 = 0$, then a positive surprise ($\varepsilon_1 > 0$) has just as large an impact on $\sigma_1{}^2$ as would a negative surprise $\varepsilon_1 < 0$). However, if the data supports a non-zero value for γ_i (specifically, $\gamma_i < 0$), we are lead to believe in the presence of leverage.

Assessing Volatility Models

The output of the G@RCH 2.0 software suite points to the adoption of an autoregressive AR(1) model for the time series mean and to GARCH, EGARCH, GJR-GARCH and APARCH models – all of specification (1,1). Twenty-lag Box-Pierce test statistics, which test whether any of a group of autocorrelations of a time series are different from zero, were computed for each model for each index. In addition, the Pearson goodness-of-fit for 50 cells, the Log-Likelihood value and the Akaike Information Criteria (AIC) were calculated in order to determine forecasting validity. These values suggest that the EGARCH, APARCH and GJR models better estimate the data than the GARCH model. Furthermore, the models with non-normal underlying distributions bested their Gaussian cousins. Evidence of this may be seen in higher Log-Likelihood and AIC values for the non-normal (Student *t*-

distribution) varieties of the models. This is not unexpected since earlier sample statistics pointed at data non-normality that could only be captured with asymmetric models.

	GARCH		EGARCH	GJR		APARCH
TA25	Student-t	Skewed-t	Skewed-t	Student-t	Skewed-t	Skewed-t
MSE(1)	0.187	0.187	0.188	0.187	0.187	0.188
MSE(2)	0.344	0.343	**0.269**	0.407	0.415	0.347
MedSE(1)	0.029	0.029	0.033	0.033	0.033	0.033
MedSE(2)	0.223	0.227	**0.128**	0.309	0.316	0.224
MAE(1)	0.274	0.274	0.272	0.273	0.272	0.272
MAE(2)	0.500	0.499	**0.397**	0.568	0.575	0.501
RMSE(1)	0.432	0.432	0.433	0.433	0.433	0.434
RMSE(2)	0.587	0.586	**0.519**	0.638	0.644	0.589
AMAPE(2)	0.782	0.782	**0.758**	0.795	0.796	0.781
TIC(1)	0.938	0.944	0.983	0.968	0.976	0.981
TIC(2)	0.559	0.559	0.565	0.565	0.566	0.560

Notes: (1) – mean equation, (2) – variance equation. The bold figures show the best result for the forecasting measures.

Figure 8-1 Forecasting Analysis for the TA25 Index: Density Comparison (Dima et. al)

FORECASTING

Again, using the G@RCH suite, 30 one-step-forward forecasts were made for each index using the totality of the data set as the rolling window for both the mean and the variance. Using MSE, MedSE, MAE, AMAPE and TIC performance measures, it was determined that the EGARCH model with underlying Student *t*-distribution consistently outperformed the GARCH, GJR-GARCH and APARCH models.

WORKING EXAMPLE

In this section of the case study, the reviewer presents a full example of the GARCH methodology using the MATLAB GARCH Toolbox. The sample period is from January 2, 1990, to December 31, 2001, yielding a total of 3028 daily equity index observations from the NASDAQ Composite index (Figure 8-2 presents the closing price run sequence; Figure 8-3 presents the same data now stationary after a price-to-return transformation).

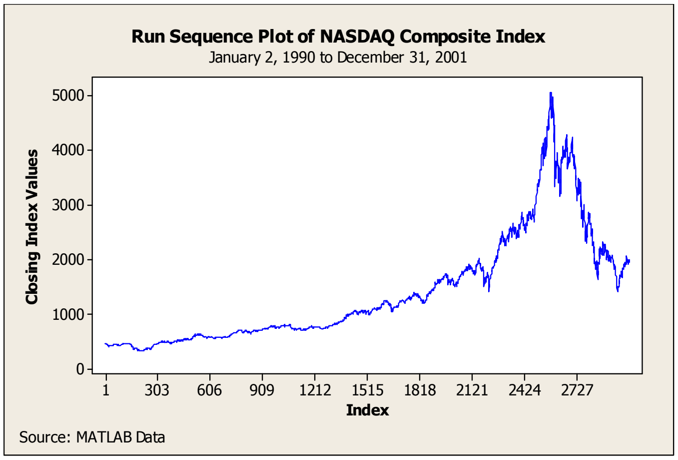


Figure 8-2 Run Sequence Plot of NASDAQ Composit Index: Closing Price

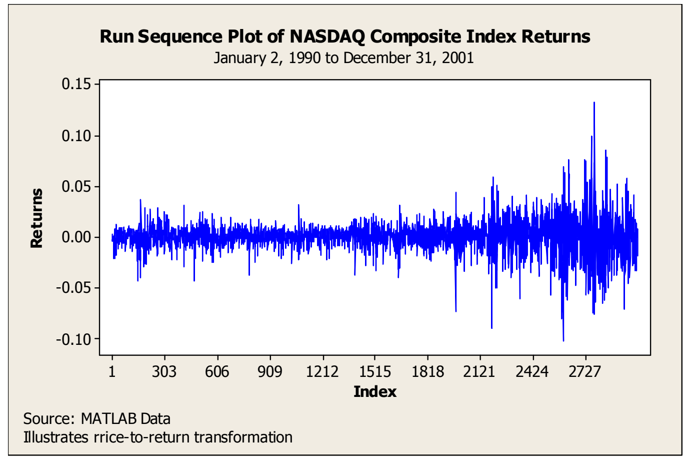


Figure 8-3 Run Sequence Plot of NASDAQ Composite Index: Price-to-Return

In the *pre-estimation analysis*, the autocorrelation and partial autocorrelation functions of the index returns are examined (Figure 8-4 and Figure 8-5) since they provide guidance on the correlation characteristics of the returns. Based on the plots generated in MATLAB, there does not seem to be a need for a correlation structure in the conditional mean.

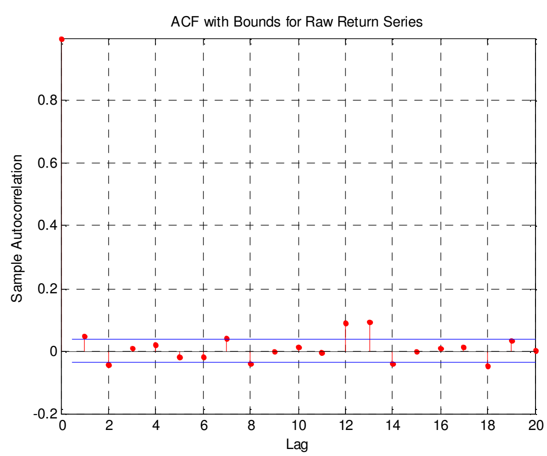


Figure 8-4 Autocorrelation Function Plot for NASDAQ Returns

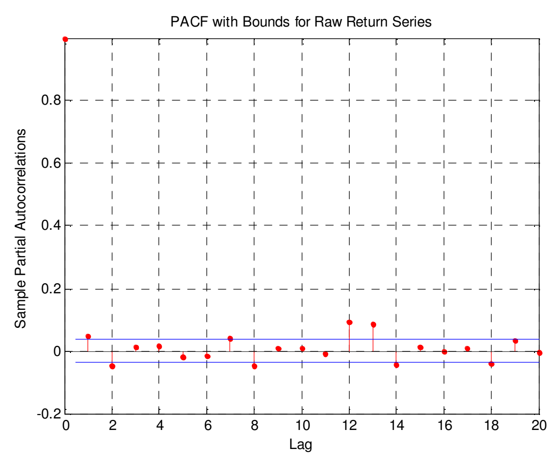


Figure 8-5 Partial Autocorrelation Plot for NASDAQ Returns

Similarly, any correlation between the squared returns is investigated using an ACF plot of the squared returns to determine the characteristics of the variance process (Figure 8-6).

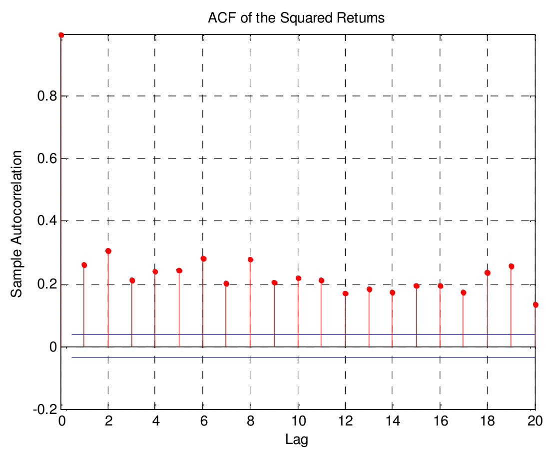


Figure 8-6 Autocorrelation Plot of the Squared Returns

H	pValue	Stat	CriticalValue
1.0000	0	605.6572	18.3070
1.0000	0	616.8817	24.9958
1.0000	0	682.4249	31.4104

Table 8-2 ARCH test MATLAB Output (Pre-Estimation Analysis)

Although the returns themselves are not largely correlated, the variance does appear to exhibit some correlation; furthermore, the ACF plot does not decay rapidly, raising the possibility that the variance process is not stationary. In the next step, Engle's own ARCH test is performed using built-in MATLAB commands to investigate homoskedasticity by testing the null hypothesis that a time series is a random sequence of Gaussian disturbances (i.e., no ARCH effects exist). The test (Table 8-2) indicates significant heteroskedastic effects.

In the *parameter estimation phase*, the `garchfit` command is used to estimate model parameters. Assuming the default GARCH model outlined below

$$y_t = C + \varepsilon_t$$
$$\sigma_t^2 = \kappa + G_1\sigma_{t-1}^2 + A_1\varepsilon_{t-1}^2$$

MATLAB provides an output estimating the model parameters when the returns array is applied as an argument to the `garchfit` command (Table 8-3).

```
       %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
   Diagnostic Information

Number of variables: 4

Functions
 Objective:                          garchllfn
 Gradient:                           finite-differencing
 Hessian:                            finite-differencing (or Quasi-Newton)
 Nonlinear constraints:              armanlc
 Gradient of nonlinear constraints: finite-differencing

Constraints
 Number of nonlinear inequality constraints: 0
 Number of nonlinear equality constraints:   0

 Number of linear inequality constraints:    1
 Number of linear equality constraints:      0
 Number of lower bound constraints:          4
 Number of upper bound constraints:          4

Algorithm selected
   medium-scale


       %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
 End diagnostic information
```

Max	Line	search	Directional	First-order			
Iter	F-count	f(x)	constraint	steplength	derivative	optimality	Procedure
0	5	-8776.15	-2.41E-05				
1	25	-8777.34	-2.41E-05	3.05E-05	-1.45E+04	5.90E+04	
2	31	-8925.44	-1.21E-05	0.5	1.57E+03	3.45E+07	
3	38	-8943.77	-3.86E-05	0.25	336	1.00E+06	
4	44	-9042.69	-2.12E-05	0.5	53.5	9.04E+06	
5	52	-9044.65	-2.00E-05	0.125	170	1.14E+07	
6	61	-9080.3	-1.87E-05	0.0625	90.6	2.41E+06	
7	67	-9130.62	-9.35E-06	0.5	-52.3	1.69E+06	
8	74	-9138.98	-7.01E-06	0.25	39.9	7.30E+06	
9	82	-9144.52	-7.41E-06	0.125	3.83	1.30E+06	
10	87	-9145.77	1.11E-16	1	53.7	2.01E+07	

11	93	-9163.22	0	0.5	90.9	7.42E+06	
12	102	-9163.44	-3.42E-06	0.0625	58.1	7.58E+06	
13	109	-9169.96	-2.78E-06	0.25	59.7	2.81E+06	
14	117	-9169.97	-3.06E-06	0.125	11.1	2.26E+06	
15	126	-9170.22	-3.40E-06	0.0625	2.26	9.57E+03	
16	132	-9173.3	-2.12E-06	0.5	-0.819	1.06E+06	
17	137	-9173.39	-1.97E-06	1	0.154	5.82E+05	
18	142	-9173.43	-2.02E-06	1	0.002	4.53E+04	
19	147	-9173.43	-2.02E-06	1	-1.31E-05	4.73E+03	
20	166	-9173.43	-2.02E-06	-6.10E-05	-3.34E-05	4.72E+03	Hessian modified twice
21	185	-9173.43	-2.02E-06	-6.10E-05	-9.40E-06	4.73E+03	Hessian modified twice

```
 Optimization terminated: magnitude of search direction less than 2*options.TolX
  and maximum constraint violation is less than options.TolCon.
 No active inequalities.

   Mean: ARMAX(0,0,0); Variance: GARCH(1,1)

   Conditional Probability Distribution: Gaussian
   Number of Model Parameters Estimated: 4
```

Parameter	Value	Standard Error	T-Statistic
C	0.000858	0.000183	4.6757
K	2.22E-06	3.34E-07	6.6526
GARCH(1)	0.87598	0.008925	98.1544
ARCH(1)	0.1157	0.0084721	13.6567

```
Log Likelihood Value: 9173.43
```

Table 8-3 MATLAB Output for GARCH Fitting to NASDAQ Returns

In the final stage, namely, the *post-estimation phase*, the relationship between the residuals derived from the fitted model, the corresponding conditional standard deviations, and the observed returns is graphically analyzed (Figure 8-7).

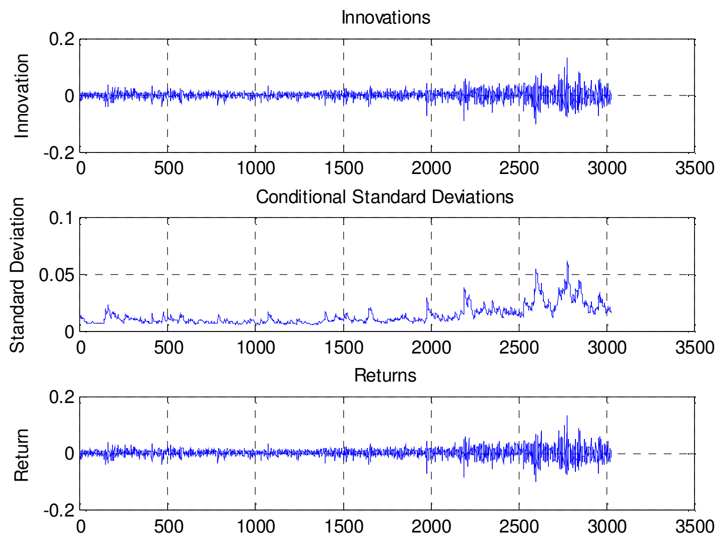


Figure 8-7 Innovations, Standard Deviations and Observed Returns of the Model

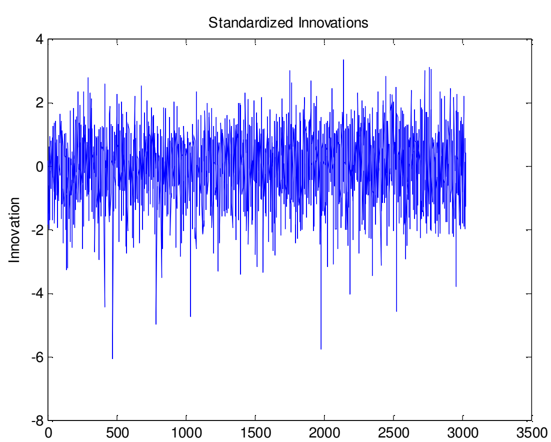


Figure 8-8 Plot of Standardized Residuals

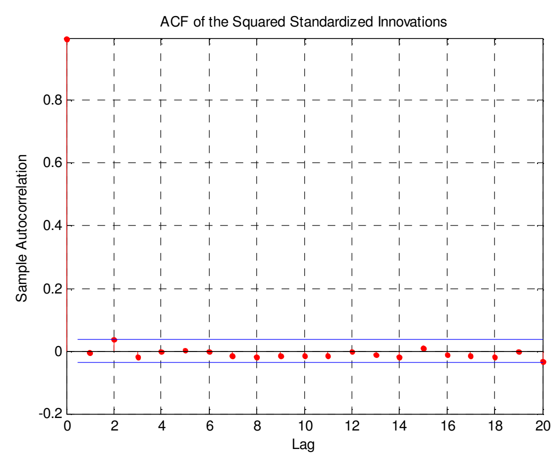


Figure 8-9 Autocorrelation Plot of Squared Standardized Residuals

The innovations show heteroskedasticity, although plotting the standardized innovations shows homoskedasticity.

Figure 8-8); furthermore, the ACF plot of the squared standardized residuals indicates no significant correlation (Figure 8-9). It appears that the default model adequately explains the heteroskedasticity in the returns. Again, the ARCH test statistic is applied using the MATLAB command `archtest`, but now to the standardized residuals of the fitted model at the 10th, 15th and 20th lags. The resulting MATLAB output indicates acceptance of the null hypotheses (Table 8-4).

H	pValue	Stat	CriticalValue
0	0.5672	8.6340	18.3070
0	0.7480	11.0642	24.9958
0	0.6436	17.1447	31.4104

Table 8-4 ARCH Test MATLAB (Post-Estimation Analysis)

CONCLUSION

Given the skewed, leptokurtotic, heteroskedastic and leveraged characteristics of the data set, the researchers found the EGARCH skewed Student *t*-distribution a natural candidate for forecasting the mean and variances of the TASE stock indices. The results not only establish the superiority of the standard GARCH models, but also give prominence to the flexibility and power of the asymmetric GARCH varieties of in forecasting the leverage effect.

8.2. Predicting corporate financial distress: A time-series CUSUM methodology

Executive Summary

This paper develops a stationary financial distress model for American Stock Exchange's (AMEX) and the New York Stock Exchange's (NYSE) manufacturing and retailing firms based on a Cumulative Sum (CUSUM) time-series methodology. This review examines the motivation behind a time-series CUSUM methodology and provides a general example to demonstrate the benefits of a general CUSUM model.

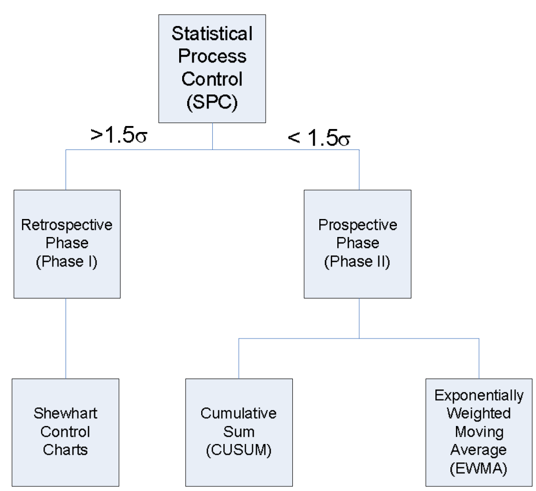
Motivation for Time Series CUSUM Model

A CUSUM time-series model allows the practitioner to distinguish between changes in the financial variables of a firm stemming from serial correlation and those stemming from a permanent shift in the mean structure of the variables due to financial distress. The paper develops a model based on the statistical methodology of CUSUM time series for a representative sample of manufacturing and retailing firms. A screening process was used to select firms 117 healthy and 72 failed firms from the AMEX and NYSE. Screening factors include the nature of financial distress (i.e., removing firms that filed for Chapter 11 protection for nonfinancial reason) and whether firms exhibited anomalous data. The second step called for verification of the economic relationships (i.e., selection of explanatory variables such as change in total assets, change in inventory over sales, etc) and data stationarity. The Vector Auto-Regression (VAR) model was used to capture the vectorized evolution and interdependencies between multiple time series. Each variable in the autoregression model is treated symmetrically by including, for each variable, an equation explaining its evolution based on its own lags and the lags of all other variables in the model.

The CUSUM model is then incorporated into a time series methodology to strengthen predictions.

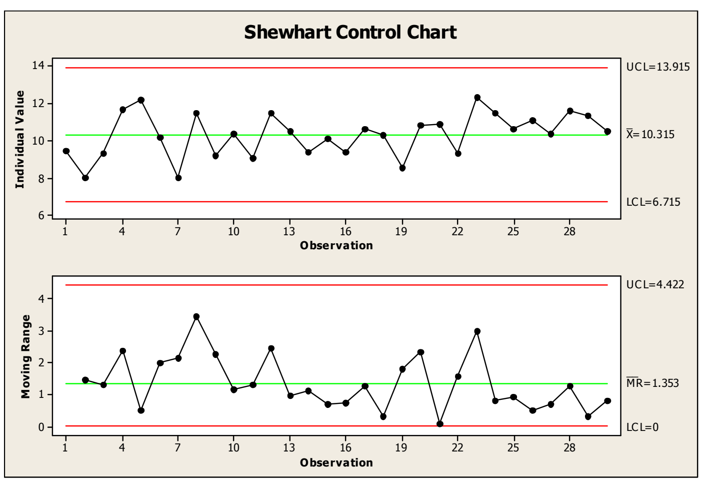
Statistical Process Control

Statistical process control (SPC) is the process of monitoring the shifts in the mean to determine whether the process is in control or out of control. The process is out of control if a shift in the process mean occurs, else it is in control. The implementation of statistical process control (SPC) is divided into two phases, the retrospective stage (phase I) and the prospective stage (phase II). The retrospective stage (phase I) is where the process is likely to be experiencing large shifts (>1.5σ) in the parameters. The Shewhart control charts are useful in the diagnostic aspect of bringing a process into statistical control. The prospective stage (phase II) is where the process tends to experience small shifts (<1.5 σ) in the parameters. The Shewhart control charts are relatively insensitive to small process shifts, which makes them less useful in prospective stage (Phase II). Two alternatives to the Shewhart control chart for the prospective phase (Phase II) are the Cumulative Sum (CUSUM) control charts and the Exponentially Weighted Moving Average (EWMA) control charts. The CUSUM monitoring process will be further discussed throughout the course of this paper.



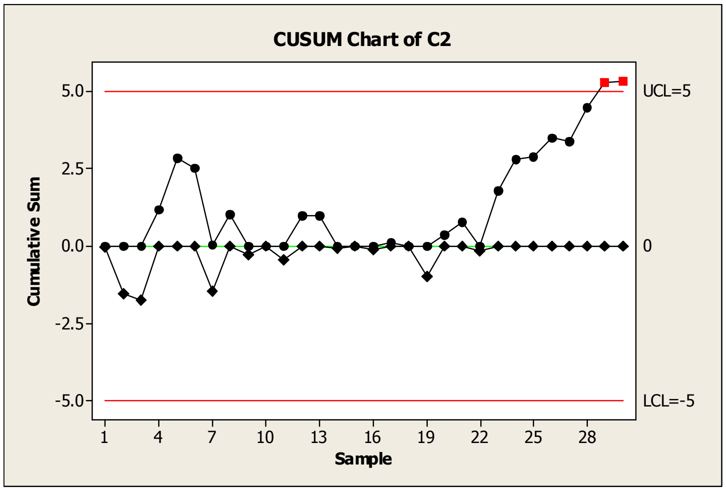
The data below demonstrates a small shift in the process mean of 1σ. The first 20 observations are N~ (10, 1) where as the final 10 observations are N~ (11, 1). Both a Shewhart control chart and CUSUM chart are analyzed below.

Sample	x_i	x_i - 10	C_i	Sample	x_i	x_i - 10	C_i
1	9.45	-0.55	-0.55	16	9.37	-0.63	-0.37
2	7.99	-2.01	-2.56	17	10.62	0.62	0.25
3	9.29	-0.71	-3.27	18	10.31	0.31	0.56
4	11.66	1.66	-1.61	19	8.52	-1.48	-0.92
5	12.16	2.16	0.55	20	10.84	0.84	-0.08
6	10.18	0.18	0.73	21	10.9	0.9	0.82
7	8.04	-1.96	-1.23	22	9.33	-0.67	0.15
8	11.46	1.46	0.23	23	12.29	2.29	2.44
9	9.2	-0.8	-0.57	24	11.5	1.5	3.94
10	10.34	0.34	-0.23	25	10.6	0.6	4.54
11	9.03	-0.97	-1.2	26	11.08	1.08	5.62
12	11.47	1.47	0.27	27	10.38	0.38	6
13	10.51	0.51	0.78	28	11.62	1.62	7.62
14	9.4	-0.6	0.18	29	11.31	1.31	8.93
15	10.08	0.08	0.26	30	10.52	0.52	9.45



The Shewhart control charts, which are based on the traditional process signals provided by the Western Electric Handbook (Appendix A), fail to detect the small shift in the process mean. In the

Shewhart chart none of the data points are plotted outside the control limits, thus we lack any strong evidence that the process mean has drifted upwards. However, there is an indication of a shift in the process mean for the last 10 points, because all but one of the points plots above the center line.



In the CUSUM chart, the one sided upper CUSUM is plotted in an increasing trend, surpassing the monitoring limits, thus providing strong evidence that the process mean has shifted upwards.

When analyzing a set of data, a Shewhart control chart must be constructed to determine whether any large shifts (>1.5 σ) are present. If the Shewhart control chart conclude that the process is in control, a CUSUM chart must be constructed to determine whether the Shewhart chart has overseen any small shifts (<1.5 σ).

SHEWHART CONTROL CHARTS

Sample size is restricted to one (n=1) for illustration and simplicity's sake in the following equations.

$$\hat{\sigma} = \frac{\overline{R}}{d_2(n)} = \frac{\overline{MR}}{d_2(n)} = \frac{\overline{S}}{C_4(n)}$$

$$S^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$$

$$\hat{\mu}_o = \overline{\overline{x}}$$

The Sample Control Chart (I, MR)

I-Chart

$$UCL = \overline{X} + z_{\alpha/2}\hat{\sigma}$$

$$CL = \overline{X}$$

$$LCL = \overline{X} - z_{\alpha/2}\hat{\sigma}$$

MR-Chart

$$UCL = \frac{\overline{MR}}{d_2(2)} W_{0.999}(2)$$

$$CL = \overline{MR}$$

$$LCL = \frac{\overline{MR}}{d_2(2)} W_{0.001}(2)$$

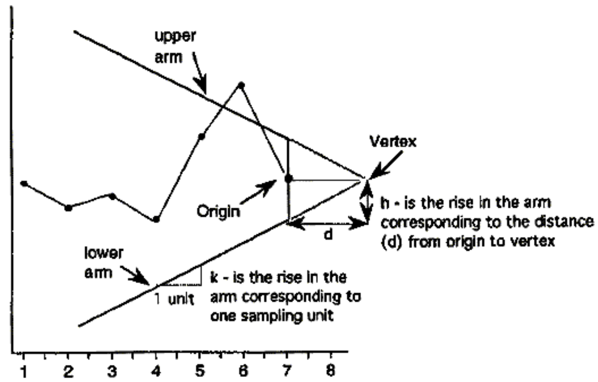
CUSUM CONTROL CHARTS

There are two ways to monitor the CUSUM process mean: The V-Mask CUSUM and the Tabular CUSUM. The Tabular CUSUM can be graphically represented two ways; the two sided CUSUM chart and the one sided CUSUM chart. The two sided CUSUM chart represents the cumulative sum through a line chart with one horizontal line. The horizontal line is free to traverse over and under the mean value. The one sided CUSUM chart represents the cumulative sum through a line chart with two horizontal lines. Each line is bounded by the mean value. The One sided upper CUSUM iteratively calculate all deviations above the mean, where as the one sided lower CUSUM iteratively calculates all deviations below the mean. A CUSUM chart has three potential trend developments:

1. **Horizontal (Neutral Drift):** The process mean remains at the target value μ_0.
2. **Upwards (Positive Drift):** The process mean shifts upwards to some value $\mu_0 > \mu_1$
3. **Downwards (Negative Drift):** The process mean shift downwards to some $\mu_1 > \mu_0$.

THE V-MASK CUSUM

The V-mask is applied to successive values of the CUSUM statistic as soon as it is plotted with both arms extending backwards to the origin.



Place the V-mask on the cumulative sum control chart with the point O on the last value of C_i and the line OP parallel to the horizontal axis. If all the previous cumulative sums lie with the two arms of the V-mask, the process is in control. However, if any of the cumulative sums lie outside the arms of the mask, the process is considered to be out of control

THE TABULAR (ALGORITHMIC) CUSUM – EXAMPLE A

Presented below is an application of the tabular CUSUM technique; for a full treatment of the tabular algorithmic CUSUM, please see Tabular (Algorithmic) CUSUM on page 23. The data below demonstrates a shift in the process mean from μ_0 to $\mu_1 = 56.25$ and from μ_0 to $\mu_2 = 55.62$, where μ_0 is the in-control estimate of the process mean. When the process is in control we want to have the ARL=500.

Cylinder	Diameter	Cylinder	Diameter
1	55.69	11	56.74
2	56.08	12	56.23
3	56.19	13	55.86
4	55.5	14	56.23
5	56.36	15	55.64
6	56.9	16	57.3
7	55.69	17	56.38

8	56.6	18	55.15
9	55.66	19	55.89
10	55.79	20	55.71

$\overline{X} = 56.08$ $\hat{\sigma} = 0.6194$

The process mean is determined $\hat{\mu} = \dfrac{\overline{\mathrm{MR}}}{\mathrm{d}_2(2)} = 56.08$

Shewhart Chart

The data provided is of sample size one (n=1) thus a Shewhart (I, MR) process control chart must be generate to indicate if the process is out of control, a large shift (>1.5σ) is present in the process mean and standard deviation.

The I-Chart is provided to monitor the process mean

$$\mathrm{LCL} = \overline{X} - Z_{\alpha/2} * \hat{\sigma} = 56.08 - 3.09(0.6194) = 54.166$$
$$\mathrm{CL} = \overline{X} = 56.08$$
$$\mathrm{UCL} = \overline{X} + Z_{\alpha/2} * \hat{\sigma} = 56.08 + 3.09(0.6194) = 57.994$$

The MR-Chart is provided to monitor the process variance

$$\mathrm{LCL} = \frac{\overline{\mathrm{MR}}}{\mathrm{d}_2(2)} * W_{0.001}(2) = 56.08 * 0.000019595 = 0.0010989$$
$$\mathrm{CL} = \overline{\mathrm{MR}} = 0.7$$
$$\mathrm{UCL} = \frac{\overline{\mathrm{MR}}}{\mathrm{d}_2(2)} * W_{0.999}(2) = 56.08 * 0.051397646 = 2.88238$$

Observing the data provided, each data point is between the LCL and UCL, thus indicating the process is absent of large shift (>1.5σ) in the process mean and standard deviation. However, since the Shewhart chart is insensitive to small shift (<1.5σ) in the process mean and standard deviation a CUSUM chart must be generated.

CUSUM CHART

Step 1 – Determine the critical shift δ

The critical size of the shift δ can be determined by defining a shift of 1σ above and below the in control process mean μ_0 to the out of control upper mean μ_1 and lower mean μ_2.

$$\mu_1 = 56.25 = \mu_0 + \delta_1\sigma = 56.08 + \delta_1(0.6194)$$

$$\delta_1 = 0.274459$$

And

$$\mu_2 = 55.62 = \mu_0 + \delta_2\sigma = 56.08 + \delta_2(0.6194)$$

$$\delta_2 = 0.74265$$

Step 2- Determine the Reference Value K

The reference value K is a value that is half the magnitude of the shift, thus a value half way between the target μ_0 and the upper CUSUM mean μ_1 or the lower CUSUM mean μ_2. The reference value is a line, which if surpassed, determines early warnings of a shift in the process mean.

The Upper Reference Value K_1

$$K_1 = \frac{|\mu_1 - \mu_0|}{2} = \frac{\delta_1\sigma}{2} = \frac{(0.274459)(0.6194)}{2} = 0.08499$$

The Lower Reference Value K_2

$$K_2 = \frac{|\mu_2 - \mu_0|}{2} = \frac{\delta_2\sigma}{2} = \frac{(0.74265)(0.6194)}{2} = 0.22999$$

Step 3 – The Average Run Length (ARL)

The two sided Average Run length is denoted by its relation to the one sided upper and lower CUSUM average run length.

$$\frac{1}{ARL_{II}} = \frac{1}{ARL_I(\delta_1, h_1, 0)} + \frac{1}{ARL_I(\delta_2, h_2, 0)}$$

The one-sided ARL must be twice the value of a two-sided ARL.

$$ARL_I(\delta_1, h_1, 0) = 2(ARL_{II})\text{=}1000$$
$$ARL_I(\delta_2, h_2, 0) = 2(ARL_{II})\text{=}1000$$

Thus obeying the relational equation

$$\frac{1}{500} = \frac{1}{1000} + \frac{1}{1000}$$

Step 4 – Calculate ARL in control

Step 4.A: When the ARL is in control, the actual shift is zero (a=0)

$$\frac{1}{ARL_{II}(\delta_1, h_1, \delta_2, h_2, 0)} = \frac{1}{ARL_I(\delta_1, h_1, 0)} + \frac{1}{ARL_I(\delta_2, h_2, 0)}$$
$$\frac{1}{ARL_{II}(0.274459, h_1, 0.74265, h_2, 0)} = \frac{1}{ARL_I(0.274459, h_1, 0)} + \frac{1}{ARL_I(0.74265, h_2, 0)}$$

$$\text{ARL}_{IN}(\delta_1,\text{h}_1,0) = \frac{\text{e}^{-\text{c}}+\text{c}-1}{2\left(\frac{\delta_1}{2}\right)^2}$$

Where:

$$\text{c}=-\delta_1(\text{h}_1+1.166)$$

$$\text{ARL}_{IN}(0.274459,\text{h}_1,0) = \frac{\text{e}^{-\text{c}}+\text{c}-1}{2\left(\frac{0.274459}{2}\right)^2}$$

$$\text{c}=2\left(0-\frac{0.274459}{2}\right)(\text{h}_1+1.166) = -0.274459(\text{h}_1+1.166)$$

Step 4.B: Replace b with –c (b= -c> 0)

$$\mathrm{g}_1(\mathrm{b}_1) = \mathrm{ARL}_I(\delta_1,\mathrm{h}_1,0) = \frac{\mathrm{e}^{\mathrm{b}}\text{-b-1}}{2\left(\frac{\delta_1}{2}\right)^2} = \frac{\mathrm{e}^{\mathrm{b}}\text{-b-1}}{\frac{\delta_1^2}{2}}$$

$$\mathrm{g}_1(\mathrm{b}_1) = \mathrm{ARL}_I(0.274459,\mathrm{h}_1,0) = \frac{\mathrm{e}^{\mathrm{b}}\text{-b-1}}{2\left(\frac{0.274459}{2}\right)^2} = \frac{\mathrm{e}^{\mathrm{b}}\text{-b-1}}{\frac{(0.274459)^2}{2}}$$

Step 4.C: Re-arrange the equation into a non-linear equation

$$g_1(b) = e^b - b - 1 = ARL_I(\delta_1,\mathrm{h}_1,0)\left(\frac{\delta_1^2}{2}\right)$$

$$g_1(b) = e^b - b - 1 - ARL_I(\delta_1,\mathrm{h}_1,0)\left(\frac{\delta_1^2}{2}\right) = 0$$

$$g_1(b) = e^b - b - 1 - 1000\left(\frac{0.274459^2}{2}\right) = 0$$

Step 4.D: Since e^b>>b, we can ignore b in the non-linear equation

$$g_1(b) = e^b - 1 - ARL_I(\delta_1,\mathrm{h}_1,0)\left(\frac{\delta_1^2}{2}\right) = 0$$

$$g_1(b) = e^b - 1 - 1000\left(\frac{(0.274459)^2}{2}\right) = 0$$

Step 4.E: Solve for b_0 in the non linear equation

$$g_1(b) = e^{b_0} = 1 + ARL_I(\delta_1,\mathrm{h}_1,0)\left(\frac{\delta_1^2}{2}\right)$$

$$e^{b_0} = 1 + ARL_I(\delta_1,\mathrm{h}_1,0)\left(\frac{\delta_1^2}{2}\right)$$

$$b_0 = \ln\left[1 + ARL_I(\delta_1,\mathrm{h}_1,0)\left(\frac{\delta_1^2}{2}\right)\right]$$

$$b_0 = \ln\left[1 + 1000\left(\frac{(0.274459)^2}{2}\right)\right] = 3.65$$

Step 4.F: Apply Newton's Method to approximate a b_i

$$b_{i+1} = b_i - \frac{g_1(b_i)}{g_1'(b_i)}$$

$$b_1 = b_0 - \frac{g_1(b_0)}{g_1'(b_0)} = 3.65 - 3.6537.6641 = 3.7469$$

Step 4.G: Once b_i is obtained, solve for h_1

$$b_i = \delta_1 \left(h_1 + 1.166\right)$$
$$h_1 = \frac{b_i}{\delta_1} - 1.166 = \frac{3.7469}{0.274459} - 1.166 = 12.4859 \approx 12.5$$

Step 4.H: Once h_1 is obtained, solve for Decision Variable H_1

$$H_1 = h_1 \sigma_{\bar{X}} = 12.5(0.6194) = 7.7425$$

Step 4.I

Recalculate the in-control lower Decision Variable (H_2) using the same method denoted above. If the assumption is made that the control limits for the process is symmetric, the in control lower decision variable H_2, will be the negative of the upper decision variables (H_2 = -7.7425)

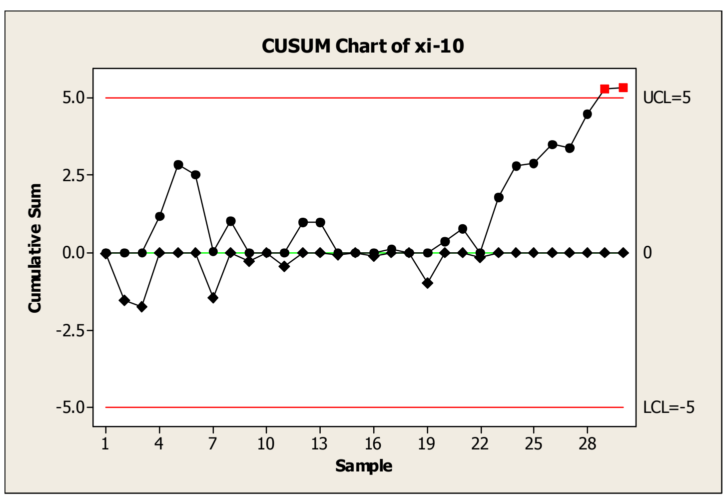
The Counter Variable *N*

The counter variable N is a value which determines the number of periods passed since the process mean was last in control. The counter variable determines the exact location by counting the previous consecutive period of data points plotted above or below the target value in between the decision variable. This allows for greater analysis of the process by locating the assignable causes for the shift in the process mean.

1. N^+: Record the consecutive one-sided upper CUSUM data point plotted in between the target value and the upper decision variable *i.e.,* $\mu_0 < N^+ < H$
2. N^- : Record the consecutive one-sided lower CUSUM data point plotted in between the target value and the lower decision variable *i.e.,* $\mu_0 < N^- < H$

The data below demonstrates a small shift in the process mean of 1σ. The Tabular CUSUM monitoring process is provided with $K = \frac{1}{2}$ and H=5. The first 20 observations are N~ (10, 1) where as the final 10 observation are N~ (11, 1).

Sample	Data	xi - 10.5	Ci+	N+	9.5 - xi	Ci-	N-	Sample	Data	xi - 10.5	Ci+	N+	9.5 - xi	Ci-	N-
1	9.45	-1.05	0	0	0.05	0.05	1	16	9.37	-1.13	0	0	0.13	0.13	1
2	7.99	-2.51	0	0	1.51	1.56	2	17	10.62	0.12	0.12	1	-1.12	0	0
3	9.29	-1.21	0	0	0.21	1.77	3	18	10.31	-0.19	0	0	-0.81	0	0
4	11.66	1.16	1.16	1	-2.16	0	0	19	8.52	-1.98	0	0	0.98	0.98	1
5	12.16	1.66	2.82	2	-2.66	0	0	20	10.84	0.34	0.34	1	-1.34	0	0
6	10.18	-0.32	2.5	3	-0.68	0	0	21	10.9	0.4	0.74	2	-1.4	0	0
7	8.04	-2.46	0.04	4	1.46	1.46	1	22	9.33	-1.17	0	0	0.17	0.17	1
8	11.46	0.96	1	5	-1.96	0	0	23	12.29	1.79	1.79	1	-2.79	0	0
9	9.2	-1.3	0	0	0.3	0.3	1	24	11.5	1	2.79	2	-2	0	0
10	10.34	-0.16	0	0	-0.84	0	0	25	10.6	0.1	2.89	3	-1.1	0	0
11	9.03	-1.47	0	0	0.47	0.47	1	26	11.08	0.58	3.47	4	-1.58	0	0
12	11.47	0.97	0.97	1	-1.97	0	0	27	10.38	-0.12	3.35	5	-0.88	0	0
13	10.51	0.01	0.98	2	-1.01	0	0	28	11.62	1.12	4.47	6	-2.12	0	0
14	9.4	-1.1	0	0	0.1	0.1	1	29	11.31	0.81	5.28	7	-1.81	0	0
15	10.08	-0.42	0	0	-0.58	0	0	30	10.52	0.02	5.3	8	-1.02	0	0



The chart concludes:

1. The process mean has shifted upwards at period 29 since the one-sided upper CUSUM exceeds the decision variable H i.e., $C_{29}^{+} = 5.28 > H$

2. The process was last in control at period 22, since 7 consecutive one-sided upper CUSUM data point were plotted in between the target value and the upper decision variable. A shift occurred between period 22 and 23 i.e.,

   29 (period) - 7 (N^+) = 22 (period)

<u>Reinitialize the CUSUM chart</u>

Once the monitoring process concludes that a shift in the process mean has occurred, the operator is required to take corrective action on the assignable causes and reinitialise the monitoring process. The CUSUM chart can be reinitialized by:

1. Setting the cumulative sums to zero i.e., $C_i^+ = C_i^- = 0$
2. Setting the cumulative sums to an estimate of the new process mean.
   a. If $C_i^+ > H$, then $\hat{\mu} = \mu_0 + k + \frac{C_i^+}{N^+}$
   b. If $C_i^- > H$, then $\hat{\mu} = \mu_0 - k - \frac{C_i^-}{N^-}$
3. The Fast Initial Response (FIR) or Headstart: setting the cumulative sums half way between the process mean and the decision variable *H* i.e.

   $C_0^+ = C_0^- = \frac{H}{2}$ (50% Headstart) is equal to the reference value *K*.

CUSUM EXAMPLE - C

The data below demonstrates the benefits of First Initial Response (FIR) or head start. The 10 samples are in control with N~ (100, σ) and K=3, H=12

Period i	x_i	x_i-103	C^+_i	N^+	97 - x_i	C^-_i	N^-
1	102	-1	5	1	-5	1	1
2	97	-6	0	0	0	1	2
3	104	1	1	1	-7	0	0
4	93	-6	0	0	4	4	1
5	100	-3	0	0	-3	1	2

6	105	2	2	1	-8	0	0
7	96	-7	0	0	1	1	1
8	98	-5	0	0	-1	0	0
9	105	2	2	1	-8	0	0
10	99	-4	0	0	-2	0	0

The calculations for the first cumulative sum period is provided

$$C_0^+ = C_0^- = \frac{H}{2} = 6$$

$$C_1^+ = \max\,[0,\, x_1 - (x_1 + 1) + C_0^+] = \max\,[0,\, 102\text{-}103\text{+}6\,] = 5$$

$$C_1^- = \max\,[0,\, 97 - x_1 + C_0^-] = \max\,[0,\, 97\text{-}102\text{+}6] = 1$$

Throughout each period, the process mean remains in control. Now, suppose the process has been out of control at process start-up, with mean 105

Period i	x_i	x_i-103	C^+_i	N^+	97 - x_i	C^-_i	N^-
1	107	4	10	1	-10	0	0
2	102	-1	9	2	-5	0	0
3	109	6	15	3	-12	0	0
4	98	-5	10	4	-1	0	0
5	105	2	12	5	-8	0	0
6	110	7	19	6	-13	0	0
7	101	-2	17	7	-4	0	0
8	103	0	17	8	-6	0	0
9	110	7	24	9	-13	0	0
10	104	1	25	10	-7	0	0

- With FIR (Headstart)

$$\left.\begin{array}{l} C_0^+ = C_0^- = 6 \\ C_3^+ \text{ exceed the limit H=12.} \end{array}\right\} \text{The process mean is out of control at period 3}$$

- Without FIR (Headstart)

$$\left.\begin{array}{l} C_0^+ = C_0^- = 0 \\ C_6^+ \text{ exceed the limit H=12.} \end{array}\right\} \text{The process mean is out of control at period 6.}$$

CONCLUSIONS

The paper's application of the Statistical Process Control (SPC) indicates the limitations of the Shewhart's and Cumulative Sum's (CUSUM) control charts. SPC incorporate these limitations into its framework; monitoring firstly for large shifts then secondly for smaller shifts in the process. The CUSUM charts initially do not have control limits and thus have to be constructed to strictly monitor the strength and health of the corporations listed on the NYSE and AMEX. Shewhart and CUSUM control charts provide clear signals of trend reversals in the process, which require the operator to investigate for assignable causes.

8.3. A GARCH Forecasting Model to Predict Day-Ahead Electricity Prices

Reinaldo C. Garcia, Javier Contreras, Marco van Akkeren, and João Batista C.Garcia.
Reviewed by
Mohammed Siyam Ibrahim

Executive Summary

This paper develops a model to predict next-day electricity prices based on the GARCH methodology. These models provide strong foundation for individuals seeking to broaden their understanding of applied econometrics. An in-depth analysis of this paper is relevant in understanding the GARCH methodology since it provides a detailed explanation of the method itself and uses empirical results from the electricity markets of Spain and California.

Motivations for using GARCH

Many models have been employed to forecast time-series data, prominently ARIMA (Autoregressive Integrated Moving Average) yet ARIMA assumptions limit the mean of the error term ε_t to zero and a constant variance σ^2. Formally, these limitations are $E(\varepsilon_t)=0$ and $E(\varepsilon_t^2)= \sigma^2$. The basic version of the least squares model assumes that the expected value of all error terms, when squared, is identical at any given point. This notion, called homoskedasticity (or homogeneity) of variance, is relaxed by ARCH/GARCH models. Data in which the variance is time-variant i.e., in which the error terms may reasonably be expected to be larger for some points or ranges of the data than for others, are said to exhibit heteroskedasticity.

In the presence of heteroskedasticity, the regression coefficients for an ordinary least squares regression may still be unbiased, but the standard errors and confidence intervals estimated by conventional procedures will be too narrow, giving a false sense of precision. The ARCH family of models treats heteroskedasticity as a characteristic of the data that is to be modeled. The error term is now assumed to be serially correlated and may be modeled by an autoregressive (AR) process. Thus, a GARCH process can measure the implied volatility of a time series due to price spikes. As a result, not only are the deficiencies of least squares address, but a prediction is computed for the variance of each error term. This prediction turns out often to be of interest, particularly in financial applications.

The modeling approach taken in the paper subscribes to the Box-Jenkins' idea on parsimony which leads to the fewest model parameters being selected (as supported by the data) to estimate an ARMA process with GARCH error components.

MODEL DEVELOPMENT

The ARMA model consists of an autoregressive (AR) and a moving average (MA) component. The model is therefore parameterized into ARMA(*p,q*), where *p* and *q* represent the orders of the autoregressive and moving average components, respectively.

$$X_t = c + \varepsilon_t + \sum_{i=1}^{p} \varphi_i X_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}.$$

Where Xt = the dependent (response) variable
c = a constant
εt = represents the error term
p = represents the autoregressive parameter or lag
q = represents the moving average term

Using the Box-Jenkins' parsimony approach, lower orders of *p* and *q* are preferred so that the data may be modeled using the fewest parameters. The GARCH(*p,q*) model is used specifically to model the error variance once data is identified as heteroskedastic. Formally, the GARCH(*p,q*) model is:

$$\sigma_t^2 = \alpha_0 + \alpha_1 \epsilon_{t-1}^2 + \cdots + \alpha_q \epsilon_{t-q}^2 + \beta_1 \sigma_{t-1}^2 + \cdots + \beta_p \sigma_{t-p}^2$$

or, more concisely:

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{q} \alpha_i \epsilon_{t-i}^2 + \sum_{i=1}^{p} \beta_i \sigma_{t-i}^2$$

In the equation presented above, α_0, α, β are to be estimated using Maximum Likelihood Estimation (MLE), ε_t^2 is the logarithmic squared error of the returns and σ_{t-i}^2 is the variance from the previous forecast (initialized with the long run variance). In the equations given in the paper, the error term for the GARCH model is re-cast as:

$$\begin{aligned} &\varepsilon_t^2 = \nu_t^2 h_t \\ &h_t = c + \sum_{i=1}^{1} \alpha_i h_{t-i} + \sum_{i=1}^{3} \beta_i \varepsilon_{t-i}^2. \end{aligned}$$

where ε_t represents the error term that will be used in conjunction with the specified ARMA model. Model selection based on significant lag values was deduced using EViews software and autocorrelation plots iteratively, with GARCH(1,3) being selected as the initial model. The historical data that was used in the model was also made stationary using logarithmic transformation. Log transformation is useful for data where the residuals get larger given larger values of the dependent variable. The effects of logarithmic transformation include variance stabilization, increasing slopes in x in relation to another variable becoming linearized and positively skewed distributions of x becoming normalized. Therefore, this technique was applied to ensure that the data can be fitted to a normal distribution as required by autoregressive models.

To validate the model, several statistical tests were performed on the residuals. Autocorrelation and partial autocorrelation plots were used to measure the quality of the fitted data. The Ljung-Box portmanteau test, which checks the overall randomness of the data, was used to validate the model at a 10% significance level. The model was then applied to forecast the volatility and prices of the Californian and Spanish electricity markets.

When analyzing the following results produced in the paper, the average prediction error was computed for each day of the last week of the respective month. Then, the Mean Week Error, that averages the seven daily errors of the week, was calculated as follows:

$$a.p.e. = \frac{1}{24}\sum_{t=1}^{24}\frac{\left|p_t - \widehat{p}_t\right|}{p_t}100$$

Where $\hat{p}$ is the actual hourly price
$\widehat{p_t}$ is the forecasted hourly price.

One can see that the Mean Week Error produced by a GARCH(1,3) model is far less compared to that of an ARIMA model which assumes the variance of the error to be constant. Only when there are periods of low volatility such as February and May does the ARIMA model slightly outperform the GARCH model. The results of the actual and forecasted electricity prices using the GARCH methodology are provided in Figure 8-10.

TABLE I
SPANISH MARKET MEAN WEEK ERROR MWE (IN PERCENT)

	GARCH	GARCH with demand	ARIMA
January	9.25	8.62	10.29
February	7.24	6.64	7.02
March	9.94	9.75	11.16
April	12.00	11.91	12.12
May	5.19	4.62	4.82
June	8.92	8.67	10.10
July	8.49	8.23	10.77
August	7.28	7.20	8.95
September	9.46	9.08	10.46
October	8.99	8.83	10.73
November	10.92	10.24	13.13
December	16.96	15.41	19.93
Average	9.55	9.10	10.79

TABLE II
CALIFORNIAN MARKET MEAN WEEK ERROR MWE (IN PERCENT)

	GARCH	GARCH with demand	ARIMA
January	5.76	5.76	8.65
February	5.07	4.19	5.02
March	5.76	5.33	5.71
April	9.14	8.54	13.23
May	11.55	10.58	13.24
June	15.91	15.40	22.26
July	12.27	12.26	15.09
August	14.03	12.36	17.57
September	10.97	9.72	11.78
October	8.10	7.83	9.51
November	5.69	5.43	5.68
December	13.63	13.22	14.38
Average	9.82	9.21	11.88

Figure 8-10 Forecasting Results from GARCH(1,3) Model (Garcia et. al.)

As noted in the graph, although the model could not accurately forecast the price spikes, the forecast following the spikes are appropriately adjusted by the GARCH model and are reasonable fits to the actual data.

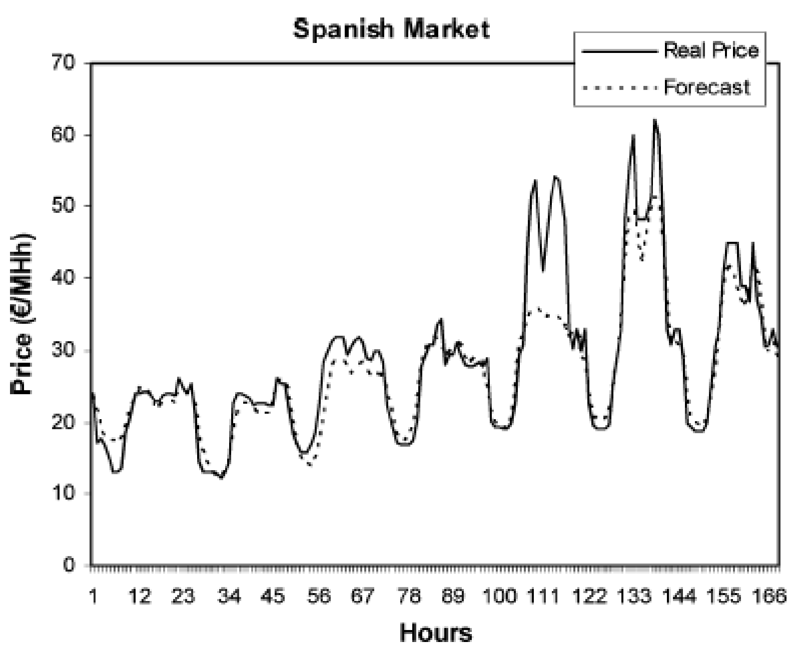


Figure 8-11 Real vs. Forecast Price in the Spanish Electricity Market (Garcia et. al.)

A GARCH(1, 1) EXAMPLE

This example uses historical data from the TSX 60 starting from Jan 3, 2000 to March 26, 2009, and applies standard model identification, estimation and validation stages to fit a GARCH(1,1) model.

MODEL IDENTIFICATION

First, a plot of the return of the adjusted closing prices is generated, and the presence of volatility clustering noted (Figure 8-12). An autocorrelation plot of the squared returns of the data is generated to determine if there is any correlation between lagged values (Figure 8-13).

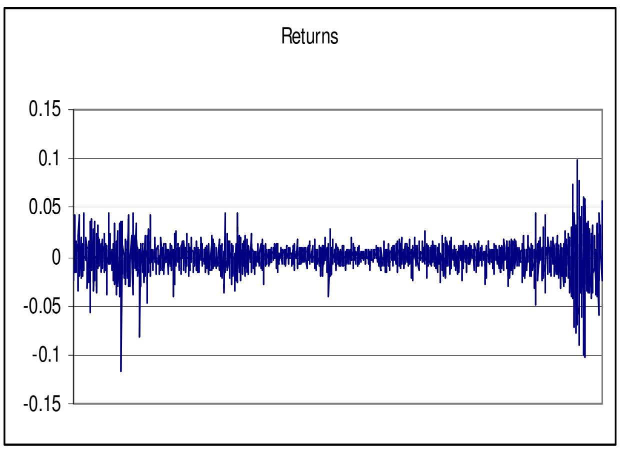


Figure 8-12 Run Sequence Plot of TSX Returns

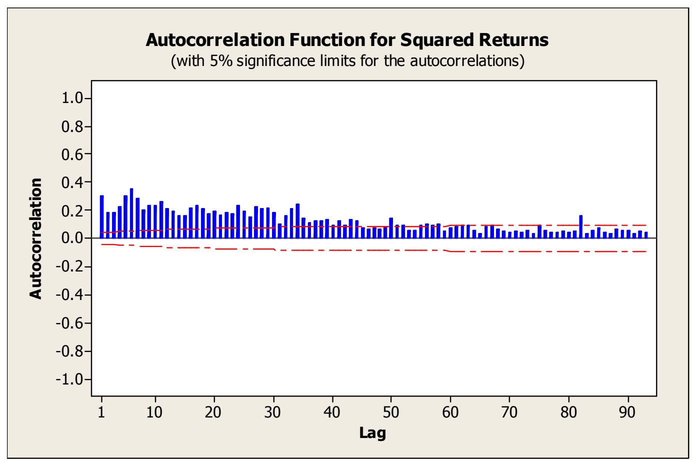


Figure 8-13 ACF Plot of Squared Returns

. As the graph indicates there seems to be some correlation between values, and the slow decay indicates that the data is close to being non-stationary. It is also possible to apply qualitative tests to determine for correlation between the returns using the Ljung-Box-Pierce Q-test and the ARCH test proposed by Engle. For the purposes of this example, a GARCH(1,1) is developed which identifies an autoregressive order of one for simplicity. Most stocks and index volatility models fall within this category.

MODEL ESTIMATION

Using the return data and Excel's Solver add-in a GARCH (1, 1) model of the data was fitting using Maximum Likelihood Estimation (MLE) in order to determine the required parameters. The model for forecasting the conditional variance is specified below:

$$\sigma_n^2 = \omega + \alpha u_{n-1}^2 + \beta \sigma_{n-1}^2$$

Where ω, α, β are parameters to be estimated using MLE for the long run variance, lagged return and the lagged variance of the return.

u_{n-1}^2 *The logarithmic squared return of the lagged value*

σ_{n-1}^2 *The conditional variance of the previous forecast*

Since Excel's solver add-is being used to estimate the parameters, certain manual calculations are required:

Log of Returns

$$y_t = ln \frac{u_t}{u_{t-1}}$$

Squared Returns

$$y_t^2 = \left(ln \frac{u_t}{u_{t-1}} \right)^2$$

GARCH

$$\sigma_t^2 = \omega + \alpha u_{t-1}^2 + \beta \sigma_{t-1}^2$$

M.L.E

$$ln \frac{1}{\sqrt{2\pi\sigma^2}} e^{-0.5\left(\frac{y_t^2}{\sigma_t^2}\right)}$$

Using Microsoft Excel's Solver Add-in, optimal solution to the parameters of the data was generated by maximizing the total M.L.E while adjusting the parameters used in the GARCH equation. The initial estimate for the variance is computed by calculating the unconditional variance of the log of the returns. Excel arrives at the following estimates for the parameters:

α	***β***	***ω***
0.117323249	0.876223774	0

These are reasonable estimates based on the approximation algorithm used in Excel to maximize the total M.L.E

MODEL VALIDATION

In order to determine the sufficiency of the model, residual analysis was performed to check whether the reisuals are are Independently and Identically Distributed (I.I.D).

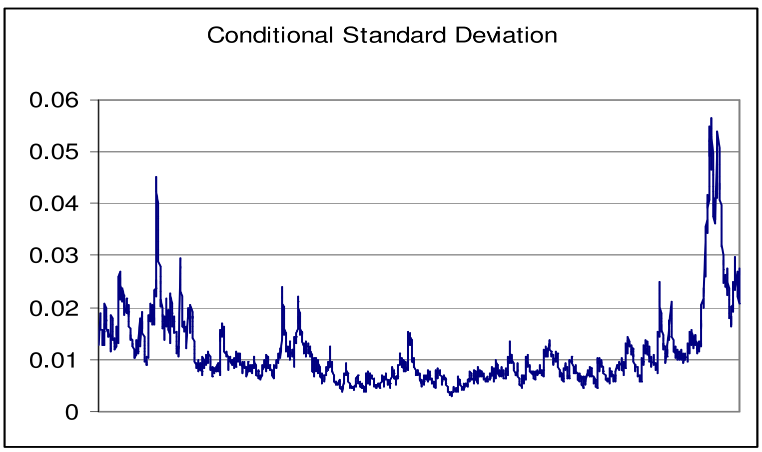


Figure 8-14 Conditional Standard Deviation of the Residuals

The residuals were calculated by the log of the returns divided by the conditional standard deviation (sigma). Formally, this is $Z_t = \left(\ln\frac{u_t}{u_{t-1}}\right)^2/\hat{\sigma}_t$. The residuals and the returns were plotted (Figure 8-15), and it was noted that there does not appear to be much volatility clustering for the residuals. Therefore, the correlation between the standardized residuals calculated as $Z_t = \left(\ln\frac{u_t}{u_{t-1}}\right)^2/\hat{\sigma}_t{}^2$ may be analyzed using the ACF plot (Figure 8-16).

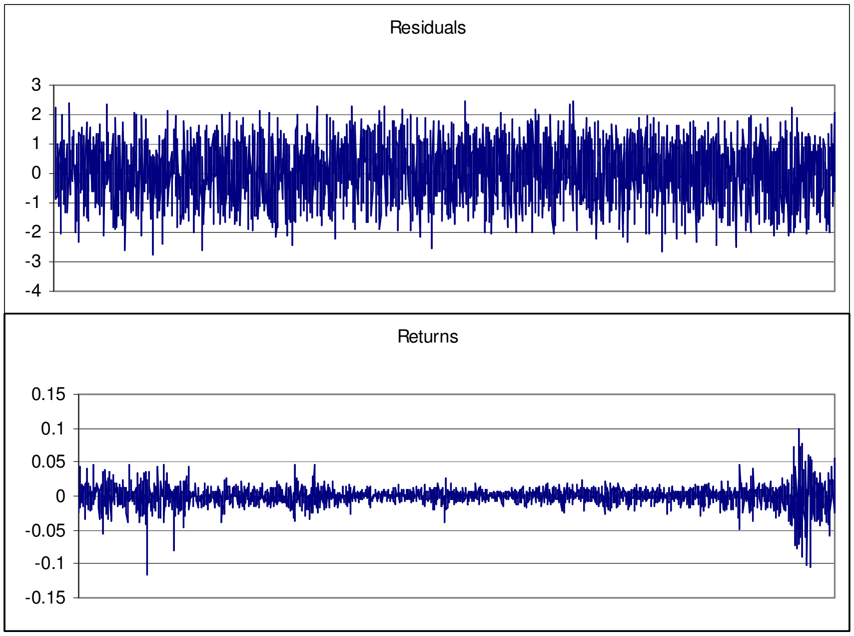


Figure 8-15 Residuals and Returns: Run Sequence Plot

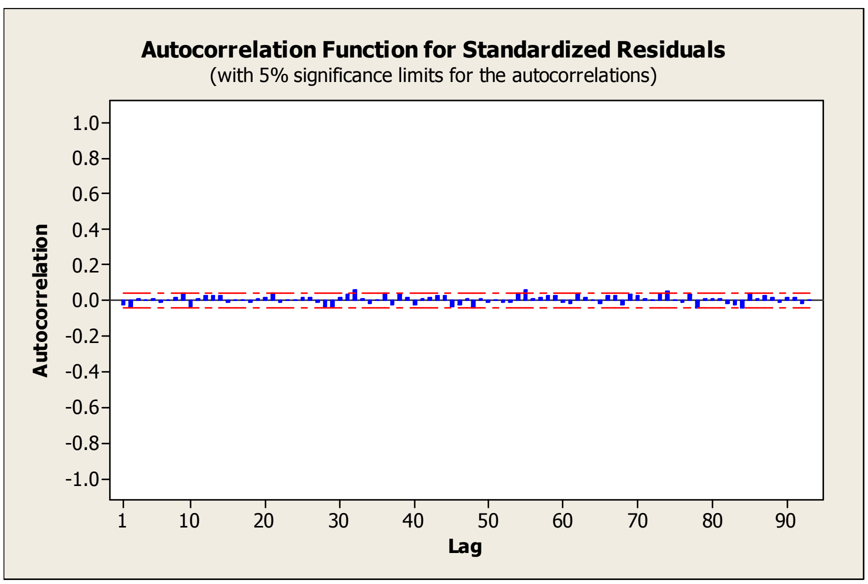


Figure 8-16 ACF Plot of the Standardized Residuals

Analysis of the ACF plot of the standardized residuals and the ACF plot of the square returns highlight some dissimilarity. It appears that the GARCH (1, 1) model sufficiently explains the heteroskedasticity in the actual returns.

The paper's application of the ARIMA and GARCH models showed that when markets were volatile, the GARCH model produced more accurate forecasts due to its ability to model conditional variance which was a salient feature of the electricity prices. Thus a GARCH model can be used to forecast prices with lower average prediction errors when modeling data subject to high volatility (heteroskedasticity) as opposed to other autoregressive models that assume constant variance (homoskedasticity).

9. HISTORY OF THE TSX

9.1. THE PIONEERS OF TRADING[9]

Historian Fernand Braudel suggests that in Cairo in the 11th century, Muslim and Jewish merchants had already set up every form of trade association and had knowledge of many methods of credit and payment, disproving the belief that these were originally invented later by Italians. In 12th century France the *courratiers de change* were concerned with managing and regulating the debts of agricultural communities on behalf of the banks. Because these men also traded with debts, they could be called the first brokers. A common misbelief is that in late 13th century Bruges commodity traders gathered inside the house of a man called "Van der Beurze", and in 1309 they became the "Brugse Beurse", institutionalizing what had been, until then, an informal meeting, but actually, the family Van der Beurze had a building in Antwerp where those gatherings occurred; the Van der Beurze had Antwerp, as most of the merchants of that period, as their primary place for trading. The idea quickly spread around Flanders and neighbouring counties and "Beurzen" soon opened in Ghent and Amsterdam.

In the middle of the 13th century, Venetian bankers began to trade in government securities. In 1351 the Venetian government outlawed spreading rumours intended to lower the price of government funds. Bankers in Pisa, Verona, Genoa and Florence also began

[9] Excerpted from (Stock Market).

trading in government securities during the 14th century. This was only possible because these were independent city states not ruled by a duke but a council of influential citizens. The Dutch later started joint stock companies, which let shareholders invest in business ventures and get a share of their profits or losses. In 1602, the Dutch East India Company issued the first shares on the Amsterdam Stock Exchange. It was the first company to issue stocks and bonds.

The Amsterdam Stock Exchange (or Amsterdam Beurs) is also said to have been the first stock exchange to introduce continuous trade in the early 17th century. The Dutch pioneered short selling, option trading, debt-equity swaps, merchant banking, unit trusts and other speculative instruments, much as we know them.

There are now stock markets in virtually every developed and most developing economies, with the world's biggest markets being in the United States, Canada, China (Hong Kong), India, UK, Germany, France and Japan.

9.2. THE HISTORY OF THE DOW JONES[10]

Dow Jones & Co. was founded in 1882 by Charles Dow, Edward Jones and Charles Bergstresser. Despite popular belief, the first averages were not published in the Wall Street Journal but in its precursor called the Customer's Afternoon Letter. The first averages didn't even include any industrial stocks. The focus was on the top twelve growth stocks of the

[10] Excerpted from (How Now, Dow? What Moves The DJIA?)

time, mainly transportation companies. This means that the first Dow Jones Index included nine railroad stocks, a steamship line and a communications company. This average eventually evolved into the Transportation Average. Dow felt that these twelve stocks provided a good indication of the economic health of the country. It wasn't until May 26, 1896, that Dow determined that two separate indices would better represent that health of the country. Dow created a 12 stock industrial index (Dow Jones Utilities Average (DJUA)), and a 20 stock rail index (The Dow Jones Transportation Average (DJTA)). By 1928 the industrial index includes 30 stocks, the number at which it stands today. In 1929 a utility index was added. In 1984, the year that marked the one hundredth anniversary of Dow's first publication, the Market technicians association presented a Gorham-silver bowl to Dow Jones & Co. According to the MTA, the award recognized "the lasting contribution that Charles Dow made to the field of investment analysis.

Over the years, companies in the index have been changed to ensure the index stays current in its measure of the U.S. economy. In fact, of the initial companies included, only General Electric remains as part of the modern-day average. Today, the DJIA is a benchmark that tracks American stocks that are considered to be the leaders of the economy and are on the NASDAQ and NYSE.

9.3. THE HISTORY AND EVOLUTION OF THE TORONTO STOCK EXCHANGE[11]

The Toronto Stock Exchange was born on October 25th, 1861 when a group of twenty-four men gathered at Masonic Hall to pass a resolution. At the time memberships were five dollars, granting access to the 18 securities listed at the time. Trading hours were limited to half hour periods daily consequently amounting two to three transactions per day. In 1878 the TSX became formally incorporated by an Act of the Ontario Legislature; and resided headquarters at 24 King Street East in Toronto. In 1901, the TSX moved to 20 King Street East introducing continuous auction trading. In 1913, the TSX built and moved into its own building on Bay Street introducing new technologies such as the first print-out-ticker which carried a series of trading prices and bid-ask offering quotations. On July 28th, 1914, the TSX along with the New York Stock Exchange (NYSE) ceased operations for three months due to financial panics relating to the declaration of World War I. The Armistice of 1918 marked the beginning of the first inflationary period in the economy, however, quickly followed by the first economic recession period in 1919 characterized by sharp declines in commodity prices. The TSX returned to economic equilibrium until 1933 when a worldwide depression inflicted financial hardship on Canadians. In 1934, the TSX merged with key competitors, the Standard Stock & Mining Exchange. In 1936 the TSX became North America's third largest exchange. In 1937, the TSX moved to a new facility at 234 Bay Street. In 1958, the TSX board of governors required listed companies to file statements

[11] Credit (TSX Group).

disclosing any changes in the company's affairs which might affect the price of its shares. In 1960, Lieutenant General of Ontario, Howard D. Graham became the first appointed President of the TSX. In 1977, the TSX was the first to launch Computer Assisted Trading Systems (CATS) as well as the for composite index the TSE300 Composite Index. In 1980, the TSX accounted 80% of all equity trading in Canada. In 1983, the TSX moved to its current location in the Exchange Towers at 130 King Street West. In 1987, the second Composite Index was introduced, the Toronto 35 Index. On October 19th, 1987 stock markets around the world suffered a major correction. The TSE 300 Composite Index dropped more than 400 points in very active trading. In 1996 the TSX became the first exchange in North America to introduce decimal trading and in 1997 became the largest exchange in North America to choose an afterhours electronic environment. In 1999, the TSX became Canada's sole exchange for trading of senior equities. The natural exchange assumed responsibility for the trading of derivatives and the Vancouver and Alberta stock exchange merged to form Canadian Venture Exchange (CDNX) handling trading in junior equities. The Canadian Dealing Network, Winnipeg Stock Exchange and equities potion of the Montreal Exchange later merged with CDNX. The TSX received Royal Assent from Lieutenant Governor of Ontario to convert the exchange into a for-profit organization. In 2000, the TSX opened an office in Montreal. In 2001, the TSX completes the acquisition of Canadian Ventures Exchange (CDNX) renamed TSX Venture Exchange migrating its list to the TSX trading platform. S&P/CDNX Index was launched on December 10, 2001. The index was renamed S&P/TSX Venture Composite Index in May 2002. Later in 2002, Standard & Poor's agreed to take over management of the TSE 300 Composite Index renaming it the S&P/TSX Composite Index. On April 8, 2002, The Toronto Stock Exchange

Inc, re-branded its organization and adopted the acronym TSX, TSX Venture Exchange and TSX Market. In 2003, the TSX Group Inc declared its first quarterly dividend as a public company, rising slowly ever since. In August 2003 the TSX Market launches Specialty Price Crosses, enabling participating organizations to have great flexibility in reporting trades. In 2004 the TSX Group acquires NGX Canada Inc., an electronic exchange that trades and clears natural gas and electricity contracts as well as the U.S. dollar order book for trading in selected stocks, also launching the TSX Market On Close, a facility to stabilize orders and pricing at the end of the trading day, also Multiple Give-Up a feature whereby TSX and TSX Ventures Exchange investors have greater choice in clearing & settlement decisions. In 2005, the TSX sent a delegation to China to attract international listings. S&P announced Canada's first independent, multi-dealer priced fixed income index, the S&P/TSX Canadian Bond Index. The TSX Venture announced the TSX Venture 50, the first ever ranking of Canada's top emerging public companies to increase awareness of the top performing companies on TSX Ventures Income trust added to the S&P/TSX Composite. In 2006, TSX Group signed an agreement with Standard & Poor's to secure exclusive use of S&P/TSX equity indices in connection with options and futures.

9.4. History of Standard & Poor's[12]

The first S&P index to be introduced in 1923 was the S&P 90 index which was published daily as well as the S&P423 which was published weekly. In 1957, S&P introduced the

[12] Adapted from (S&P 500).

S&P500 Index the real-time calculation of which was made possible due to advancements in technology. The S&P 500 is used widely as an indicator of the broader market including both growth and value stocks from both the NASDAQ and NYSE. The S&P 500 is a stock market index containing the stocks of 500 large American corporations. However, the index does include a handful of foreign, yet formerly American companies that are now incorporated outside of the United States, which were allowed to remain in the S&P 500. A requirement to be on the S&P 500 the company must have liquidity. The Fortune 500 attempts to list the 500 largest public companies in the United States by gross revenue, regardless of where their stocks trade or their liquidity. The S&P index forms part of the broader S&P 1500 and S&P Global 1200 stock market indices. It is often quoted using the symbol SPX or INX and may be prefixed with a caret or a dollar sign.

9.5. What is the S&P/TSX 60[13]

The S&P/TSX 60 is designed to represent leading companies in leading sectors; the S&P/TSX 60 covers approximately 73% of Canada's equity market capitalization. The S&P/TSX 60 is maintained by the S&P Canadian Index Committee, which comprises a team of seven, including four members from Standard and Poor's and three members from the Toronto Stock Exchange (TSX). The committee follows a set of published guidelines for maintaining the index.

[13] Adapted from (Standard & Poor's).

Entry criteria include the following considerations:

4. Eligibility: A company must be a constituent of the S&P/TSX Composite
5. Listing: Only stocks listed on the Toronto Stock Exchange are considered
6. Domicile: Only Canadian Incorporated companies are eligible.
7. Market Capitalization: Only the 60 largest companies, as measured by market cap.
8. Liquidity: Only stock that are actively and regularly traded are considered
   Fundamentals: The company must have an track record of both revenues and earnings
9. Sector Representation: The index Committee strives to maintain a balance with all ten sectors (Energy, Materials, Industrials, Consumer Discretionary, Consumer Staples, Health Care, Financials, Information Technology, Telecommunications Services and Utilities)

Grounds for delisting include:

10. Violations: A violation of one of the requirements for continued inclusion-
    Mergers/Acquisitions: Companies involved in M&A's or significant restructuring

10. FINANCIAL MARKETS AND MARKET INDICATORS

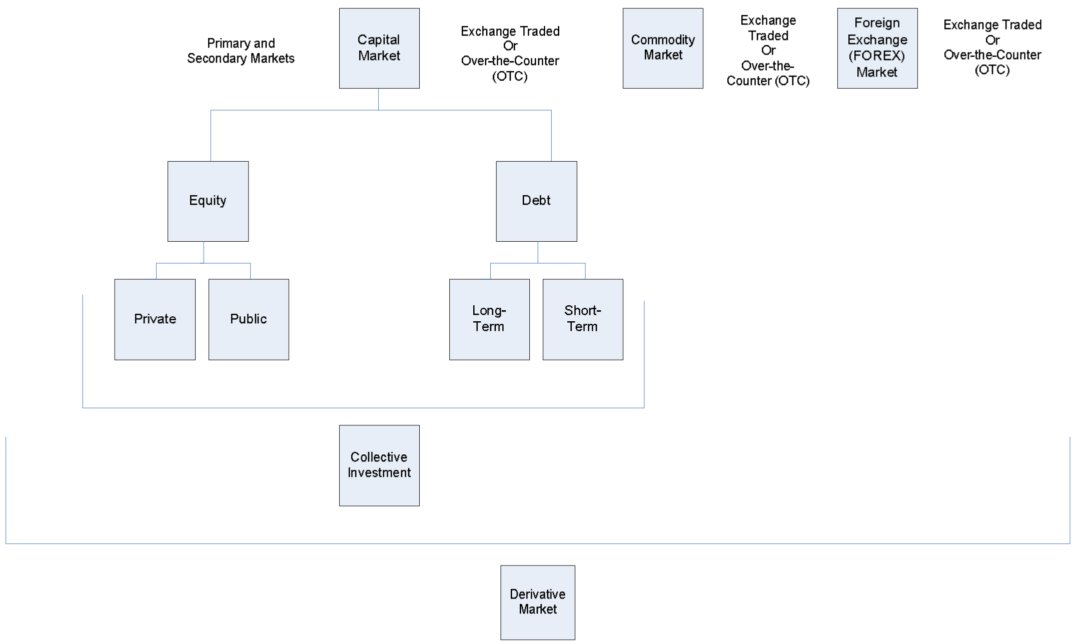


Figure 10-1 A Graphical Representation of the Financial Markets

10.1. Indicators of Inflation and Interest Rates[14]

The market is driven by two main forces; the inflation rate and interest rate. These two rates are inversely related. An increase in the interest rates inversely affects inflation rates. Interest rates provide insight on borrowing money (debt financing). A rise in interest rate will cause debt financing to become more costly, a decrease interest rates generally makes debt financing less expensive. A rise in inflation rate involves the erosion of purchasing

[14] Adapted from (Interest Rate Forecasting: Economic Indicators) and (Garner).

power, meaning the currency is worth less; this is not encouraging for the equity investor, however if inflation rates decrease, equity appears to be more affordable.

THE PRICE OF GOLD

The price of gold is a leading economic indicator. Rising inflation will cause investors to transfer investments out of financial assets and into gold. As a result, an increase in the price of gold might precede an increase in inflation rates and a decrease in interest rates. However, the price of gold will fluctuate in response to supply and demand and also foreign economic and political factors.

PRIME RATE

The prime interest rate is a leading indicator, which represents the interest rate that banking institutions require their most liquid, credit-worthy corporations to pay. Rising prime rates indicate a decrease in inflation rates.

GROSS DOMESTIC PRODUCT

The gross domestic product (GDP) is the most important economic indicator. The GDP represents the total output of goods and services produced by labour and property located in the country. A larger-than-expected quarterly increase is considered inflationary, causing concern the Fed might need to intervene and raise interest rates in order to slow growth.

EMPLOYMENT SITUATION: PAYROLL EMPLOYMENT

The payroll employment is the most significant indicator of current economic trends each month, along with the unemployment rate. A higher-than-expected monthly increase is considered inflationary, and can cause interest rates to rise.

Employment Situation: Unemployment Rate

The unemployment rate is a lagging indicator. A lower-than-expected unemployment rate or declining trend is considered inflationary, and can cause interest rates to rise.

Consumer Credit

Consumer credit data tracks debt levels for auto financing and commercial banking credit. Consumer credit report is generally considered to have little impact on interest rates, conversely little impact on inflation rates.

Housing Starts

Housing starts are a leading economic indicator. A higher-than-expected increase in housing starts triggers economic growth and is considered inflationary, causing interest rates to rise.

11. BIBLIOGRAPHY

Box, G. E., Jenkins, G. M., & Reinsel, G. C. (1994). *Time Series Analysis: Forecasting and Control.* New Jersey: Englewood Cliffs.

Eshel, G. (n.d.). *The Yule Walker Equations for the AR Coefficients.* Retrieved February 2009, from http://hope.simons-rock.edu/~geshel/geos31415/YW.pdf

Garner, C. A. (n.d.). *How Useful Are Leading Indicators of Inflation?* Retrieved from http://www.kc.frb.org/publicat/econrev/pdf/2q95garn.pdf

Interest Rate Forecasting: Economic Indicators. (n.d.). Retrieved 10 1, 2009, from http://mortgage-x.com/general/indicators.asp

Kurtosis. (n.d.). Retrieved March 19, 2009, from Wikipedia: http://en.wikipedia.org/wiki/Kurtosis

Mladjenovic, P. *Stock Investing For Dummies.* For Dummies.

Montgomery, D. C. (2005). *Introduction to Statistical Quality Control.* Wiley.

NIST/SEMATECH e-Handbook of Statistical Methods. (2006). Retrieved from http://www.itl.nist.gov/div898/handbook/index.htm

Skewness. (n.d.). Retrieved March 19, 2009, from Wikipedia: http://en.wikipedia.org/wiki/Skewness

Stock Market. (n.d.). Retrieved from Wikipedia: http://en.wikipedia.org/w/index.php?title=Stock_market&oldid=279780198

TSX Group. (n.d.). *TMX Group History at a Glance.* Retrieved 10 1, 2008, from http://www.tsx.com/en/pdf/TMXHistory.pdf

12. APPENDIX

Appendix A – ACF and PACF Plots Script

```
function arplots(f,lag, difference)

% Read excel data and precalculations
data = xlsread(f);
c0=0;
k=lag;
[maxrow,history]=size(data);

% Transform data for calculations
for i = 1:history
    for n= 1:maxrow
        if (isnan(data(n,i)))
           eval(strcat('y',num2str(i),'=','(h',num2str(i),');'));
           break
        elseif (n==maxrow)

eval((strcat('h',num2str(i),'(',num2str(n),',','1',')','=',num2str(data(n,i)),';'))
);
            eval(strcat('y',num2str(i),'=','(h',num2str(i),');'));
        else

eval((strcat('h',num2str(i),'(',num2str(n),',','1',')','=',num2str(data(n,i)),';'))
);
        end
    end
end

% Difference data
for i= 1:history
    for j= 1:difference
        eval(strcat('y',num2str(i),'=','diff(y',num2str(i),');'));
    end
end

% Calculate mean
mn_string='';
for i= 1:history
    mn_string = strtrim(strcat(mn_string,'y',num2str(i), ';'));
end
mcum = eval(strcat('[',mn_string,']'));
themean = mean(mcum);

% Subtract mean
for i= 1:history
    eval(strcat('y',num2str(i), '=', 'y',num2str(i), '-themean;'));
end
T=0;
for i= 1:history
    [yrow,ycol] = eval(strcat('size(y',num2str(i), ')'));
    T = T + yrow;
end

% Calculate ACF
csum = 0;
for i = 1:history
    [yrow,ycol] = eval(strcat('size(y',num2str(i), ')'));
```

```
    isum = 0;
    for n = 1:yrow
        isum = isum + 
(eval(strcat('y',num2str(i),'(',num2str(n),',',num2str(1),')')))^2;
    end
    csum = csum + isum;
end
c0=csum/T;
for l = 1:k
    j=1;
    accm = 0;
    for i = 1:history
        [thisrow,thiscol]=size(eval(strcat('y',num2str(i))));
        for n= 1:thisrow-l
            accm(j,1)= 
eval(strcat('y',num2str(i),'(',num2str(n),',',num2str(1),')*y',num2str(i),'(',num2s
tr(n+l),',',num2str(1),')'));
            j=j+1;
        end
    end
    r(l,1)=(sum(accm)/T)/c0;
end

% Calculate PACF
pacf=0;
for m=1:k
    p=eye(m);
    for i= 1:m
        for j= 1:m-i
            p(i,j+i)=r(j,1);
        end
        for j= 1:m-i
            p(j+i,i)=r(j,1);
        end
    end

    rpacf(m,1)=r(m,1);
    phipacf=inv(p)*rpacf;
    [m,n] = size(phipacf);
    pacf(m,1)=phipacf(m,1);
end

% Display results and plots
r
pacf
subplot(2,1,1); bar(r)
title 'Autocorrelation Function'
xlabel 'Lags'
ylabel 'Autocorrelation'
ylim([-1 1])
subplot(2,1,2); bar(pacf)
title 'Partial Autocorrelation Function'
xlabel 'Lags'
ylabel 'Partial Autocorrelation'
ylim([-1 1])
```

```
function armodel(f,param, difference)

% Read excel data and precalculations
data = xlsread(f);
c0=0;
k=param;
[maxrow,history]=size(data);

% Transform data for calculations
for i = 1:history
    for n= 1:maxrow
        if (isnan(data(n,i)))
           eval(strcat('y',num2str(i),'=','(h',num2str(i),');'));
           break
        elseif (n==maxrow)

eval((strcat('h',num2str(i),'(',num2str(n),',','1',')','=',num2str(data(n,i)),';'))
);
            eval(strcat('y',num2str(i),'=','(h',num2str(i),');'));
        else

eval((strcat('h',num2str(i),'(',num2str(n),',','1',')','=',num2str(data(n,i)),';'))
);
        end
    end
end

% Difference data
for i= 1:history
    for j= 1:difference
        eval(strcat('y',num2str(i),'=','diff(y',num2str(i),');'));
    end
end

% Calculate mean
mn_string='';
for i= 1:history
    mn_string = strtrim(strcat(mn_string,'y',num2str(i), ';'));
end
mcum = eval(strcat('[',mn_string,']'));
themean = mean(mcum);

% Subtract mean
for i= 1:history
    eval(strcat('y',num2str(i), '=', 'y',num2str(i), '-themean;'));
end

% Calculate total sample size
T=0;
for i= 1:history
    [yrow,ycol] = eval(strcat('size(y',num2str(i), ')'));
    T = T + yrow;
end

% Calculate Autocorrelation
csum = 0;
for i = 1:history
    [yrow,ycol] = eval(strcat('size(y',num2str(i), ')'));
    isum = 0;
    for n = 1:yrow
```

```
        isum = isum + 
(eval(strcat('y',num2str(i),'(',num2str(n),',',num2str(1),')')))^2;
    end
    csum = csum + isum;
end
c0=csum/T;
for l = 1:k
    j=1;
    accm = 0;
    for i = 1:history
        [thisrow,thiscol]=size(eval(strcat('y',num2str(i))));
        for n= 1:thisrow-l
            accm(j,1)= 
eval(strcat('y',num2str(i),'(',num2str(n),',',num2str(1),')*y',num2str(i),'(',num2s
tr(n+l),',',num2str(1),')'));
            j=j+1;
        end
    end
    r(l,1)=(sum(accm)/T)/c0;
end

% Yule-Walker method for estimating parameters
p=eye(k);
for i= 1:k
    for j= 1:k-i
        p(i,j+i)=r(j,1);
    end
    for j= 1:k-i
        p(j+i,i)=r(j,1);
    end
end
[rrow,rcol] = size(r);
phi = inv(p)*r;
phi
```

```
function cusum(f, arl, delta, mean, sigma)

% Read excel data and precalculations
residuals = xlsread(f);
k=delta*sigma/2;
hi=residuals-mean-k;
lo=mean-k-residuals;
shi(1,1)=0;
slo(1,1)=0;
[m,n] = size(residuals);

% Calculate CUSUM
for i=2:m
    shi(i,1)=max(0,shi(i-1,1)+hi(i,1));
    slo(i,1)=max(0,slo(i-1,1)+lo(i,1));
end
slo=-1*slo;

% Newton's Method for calculating control limits
b0=log(1+(2*arl*(delta^2)/2));
for i=1:100
    g=exp(b0)-b0-1-((delta^2)/2)*(2*arl);
    gprime=exp(b0)-1;
    b1=b0-(g/gprime);
    if (abs(b1-b0)/b1 < 10^-6)
        break;
    elseif i==100
        error( 'B did not converge' );
    end
    b0=b1;
end
b1;
h=(b1/delta)-1.166;
H=h*sigma;
ubound(1:m,1)=H;
lbound(1:m,1)=-1*H;

% Plot CUSUM chart
hold on
plot(slo,'-ko','LineWidth',1,'MarkerSize',4,
'MarkerEdgeColor','k','MarkerFaceColor','g')
plot(shi,'-ko','LineWidth',1,'MarkerSize',4,
'MarkerEdgeColor','k','MarkerFaceColor','g')
failed='';

% Identify points past critical bounds
for i=1:m
    if shi(i,1) >= H
        plot(i,shi(i,1),'-ko','LineWidth',1,'MarkerSize',4,
'MarkerEdgeColor','k','MarkerFaceColor','r')
        failed = [failed ,' ', num2str(i)];
        text(i,shi(i,1),num2str(i))
    end
    if slo(i,1) <= -1*H
        plot(i,slo(i,1),'-ko','LineWidth',1,'MarkerSize',4,
'MarkerEdgeColor','k','MarkerFaceColor','r')
        failed = [failed ,' ', num2str(i)];
        text(i,slo(i,1),num2str(i))
    end
end
```

```
% Display final results
failed = strtrim(failed);
plot(ubound)
plot(lbound)
text(m,H,strcat('UCL=',num2str(H)))
text(m,-1*H,strcat('LCL=',num2str(-1*H)))
title 'CUSUM Chart'
xlabel 'Sample'
ylabel 'Cumulative Sum'
hold off
plotedit on
results{1,1}='H';
results{1,2}=H;
results{2,1}='h';
results{2,2}=h;
results{3,1}='k';
results{3,2}=k;
results{4,1}='ARL';
results{4,2}=arl;
results{5,1}='Delta';
results{5,2}=delta;
results{6,1}='Sigma';
results{6,2}=sigma;
results{7,1}='Mean';
results{7,2}=mean;
disp(results)
if (failed)
    disp('Test failed at points:')
    disp(failed)
else
    disp('Process is in control')
end
```