5-2017

Optimized Forecasting of Dominant U.S. Stock Market Equities Using Univariate and Multivariate Time Series Analysis Methods

Michael Schwartz

Chapman University, schwa168@mail.chapman.edu

Recommended Citation

Optimized Forecasting of Dominant U.S. Stock Market Equities

Using Univariate and Multivariate Time Series Analysis Methods

A Dissertation by

Michael D. Schwartz

Chapman University

Orange, California

Schmid College of Science and Technology

Submitted in partial fulfilment of the requirements for the degree of

Doctor of Philosophy in Computational and Data Sciences

May 2017

Committee in charge:

Cyril Rakovski, Ph.D., Committee Chair
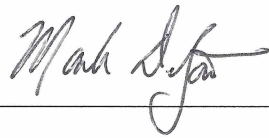
Mark DeSantis, Ph.D.
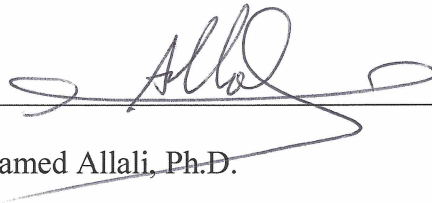
Mohamed Allali, Ph.D.

The dissertation of Michael D. Schwartz is approved.

_____

Cyril Rakovski, Ph.D., Committee Chair

_____

Mark DeSantis, Ph.D.

_____

Mohamed Allali, Ph.D.

May 2017

Optimized Forecasting of Dominant U.S. Stock Market Equities

Using Univariate and Multivariate Time Series Analysis Methods

Copyright © 2017

by Michael D. Schwartz

ACKNOWLEDGEMENTS

DEDICATION

To Elaine, Juliet, Mark, and Jason

"Intellectual growth should commence at birth and cease only at death"

– Attributed to Albert Einstein (1879-1955)

ABSTRACT

Optimized Forecasting of Dominant U.S. Stock Market Equities

Using Univariate and Multivariate Time Series Analysis Methods

by Michael D. Schwartz

This dissertation documents an investigation into forecasting U.S. stock market equites via two very different time series analysis techniques: 1) autoregressive integrated moving average (ARIMA), and 2) singular spectrum analysis (SSA). Approximately 40% of the S&P 500 stocks are analyzed. Forecasts are generated for one and five days ahead using daily closing prices. Univariate and multivariate structures are applied and results are compared. One objective is to explore the hypothesis that a multivariate model produces superior performance over a univariate configuration. Another objective is to compare the forecasting performance of ARIMA to SSA, as SSA is a relatively recent development and has shown much potential.

Stochastic characteristics of stock market data are analyzed and found to be definitely not Gaussian, but instead better fit to a generalized t-distribution. Probability distribution models are validated with goodness-of-fit tests. For analysis, stock data is segmented into non-overlapping time "windows" to support unconditional statistical evaluation. Univariate and multivariate ARIMA and SSA time series models are evaluated for independence. ARIMA models are found to be independent, but SSA models are not able to reach independence. Statistics for out-of-sample forecasts are computed for every stock

in every window, and multivariate-univariate confidence interval shrinkages are examined. Results are compared for univariate, bivariate, and trivariate combinations of highly-correlated stocks. Effects are found to be mixed.

Bivariate modeling and forecasting with three different covariates are investigated. Examination of results with covariates of trading volume, principal component analysis (PCA), and volatility reveal that PCA exhibits the best overall forecasting accuracy in the entire field of investigated elements, including univariate models. Bivariate-PCA structures are applied in a back-testing environment to evaluate economic significance and robustness of the methods. Initial results of back-testing yielded similar results to those from earlier independent testing. Inconsistent performance across test intervals inspired the development of a second technique that yields improved results and positive economic significance. Robustness is validated through back-testing across multiple market trends.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

"Essentially, all models are wrong, but some are useful."

– George E.P. Box (1919-2013), in the book in "Robustness in Statistics," 1979

Mankind has a strong desire to be able to reliably forecast future events. Throughout history there have been narratives with soothsayers predicting developments yet to happen. In the modern world of facts and science, there are practical matters to address for progress, improvements, efficiencies, and profits. Among possible outcomes, credible forecast results support (a) risk management, (b) economic, business, and production planning, (c) inventory and production control, (d) control and optimization of industrial processes, and (e) evaluation of different policies and models. Inclusion of the term "events" connotes a structure revolving around an ordered set of data called a time series. Forecasts of this nature consider problems with the following properties:

1. Observations of a single variable or multiple variables are collected over some period of time, T, and sampled, either randomly or at fixed instants. This analysis uses data sampled in uniform, discrete intervals $n = 1, 2, \ldots, N$, resulting in an (ordered) time series: $Z_n = z_1, z_2, \ldots, z_N$. Times series Z_n may be either univariate or multivariate.

2. A data point Z_{n+h} is a "future" value to be estimated, i.e., forecasted, located at a position h sample intervals beyond the last observed point N.

Forecasts are also specified as either in-sample or out-of-sample. In-sample forecasts rely on all data within the observation range and predict* new values within the same period, e.g., missing values. Out-of-sample forecasts estimate future values at time points $N + h$ using only observation data available at time point N.

This dissertation addresses only out-of-sample forecasts. Technical, mathematically-driven approaches are utilized to predict values, constrained by information available at or before the beginning of the forecast time point. Clearly, an underlying premise of these forecasts is that future values can be estimated based solely on past information. Another way to view this concept is that patterns in the subject time series may be identified, modeled, and subsequently evaluated at future time points.† Performance of a forecast is qualified using the training-validation method:

1. An input data set of $N + h$ values is separated into two sets of lengths N and h, with $N \gg h$.

2. Data subset N is used to train a model or form a basis from which to forecast.

3. A forecast at a time greater than N and less than or equal to h is computed based only on N values.

4. The quality of a forecast is "validated" by comparing predicted values to corresponding time values in subset h, using some computable measure of error.

---

* The terms forecast and predict, defined to have the same meaning, are used interchangeably.

† This premise is by definition in conflict with one commonly-held assumption about financial markets called the random-walk hypothesis. This hypothesis states that stock prices are not predictable because of the belief that prices are utterly random.

1.1 Input Data

The analyses herein target out-of-sample forecasting of stock prices of a select group of securities traded in the US stock market. A search was performed using a freely available stock screener [1], filtering for stocks with the following characteristics:

- Standard & Poor's (S&P) 500 Index

- Market Capitalization > $10B $\left(B = 10^9\right)$

- Average Daily Trade Volume > 1M shares $\left(M = 10^6\right)$

- Initial Public Offering (IPO) Date > 25 years ago

- Price > $7

These five attributes were chosen for the following reasons. The S&P 500 Index includes large companies listed on the New York Stock Exchange (NYSE) or the National Association of Securities Dealers Automated Quotations (NASDAQ) stock market, respectively, the first and second largest exchanges in the world. Equities with large capitalizations were selected to limit the number to be analyzed (as a computational convenience). Average daily trading volume greater than one million shares was selected as an indicator of liquidity, i.e., the ability to freely and easily trade with small bid/ask* spreads, such that these spreads may be reasonably treated as insignificant relative to stock price. Stocks with historical data longer than 25 years were selected for study primarily to provide multiple non-overlapping observation windows to support analytical statistical

---

* Simply, the ask price is defined as the minimum amount (per share) that a seller or sellers are willing to receive and the bid price is the maximum amount (per share) that a buyer or buyers are willing to pay for a security. (A transaction transpires when the buyer and seller agree on a price.)

independence. Additionally, with data covering a 25 year span, all possible market conditions were expected to be encountered and consequently produce a wide range of stock behaviors. Finally, stock prices greater than $7 were chosen as a threshold of study interest with the intent to avoid higher bid/ask spreads as reported in [2].

The search yielded 197 stocks (see Appendix 8.1) with varying inception dates before 1990. The stock with the latest start date was identified and the time range of all stocks was trimmed so that all securities spanned exactly the same time range, June 18, 1990 to March 21, 2016, yielding 6491 samples for each. Data for every stock on the list was downloaded from Yahoo! Finance [3]. Daily observation variables included date, open, high, low, and close prices, adjusted closing price,* and trading volume.

1.2 Stock Price Time Series

Examining long-term stock patterns spanning many years reveals that every stock has a unique time-domain behavior. Typical examples are given in Figure 1-1. Time-dependent patterns, also called trends, continuously vary, displaying characteristics (in a global sense) ranging from rising to falling to somewhat flat (also known as range-bound). Within each trend (up, down, or sideways) a stock also exhibits varying amounts of statistical variance, also termed volatility.

---

* The adjusted closing price is a stock's daily closing price adjusted to include any distributions and corporate actions that transpired at any time prior to market open on the following day.
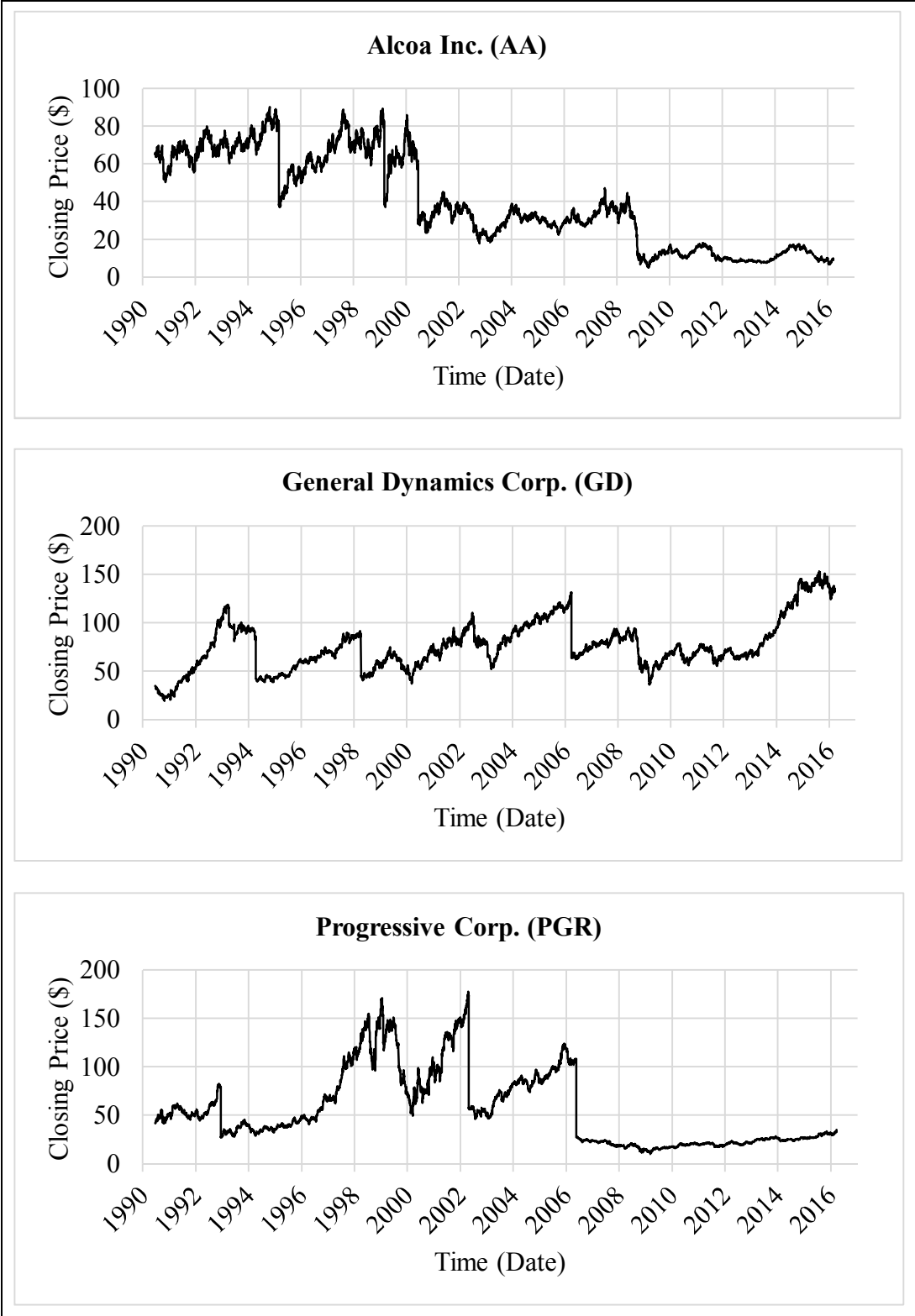
Figure 1-1. Typical Stock Time Series (Daily)

1.3 Stochastic Characteristics of Stock Prices

The origin of modern financial mathematical modeling is credited to French mathematician Louis Bachelier, as published in his Ph.D. thesis, "The Theory of Speculation," in 1900. Bachelier proposed that stock prices follow a Brownian motion[*]. In 1973, Fisher Black and Myron Scholes published their Black-Scholes derivative pricing model in a paper titled "The Pricing of Options and Corporate Liabilities." Subsequently, Robert Merton extended the Black-Scholes formula and in 1997 Scholes and Merton received the Nobel Prize in Economic Sciences.[†] In modern financial theory, the Black-Scholes model[‡] is used extensively and is reputed to be one of the foremost concepts for determining fair prices of stock options. A key assumption of the Black-Scholes model is that stock prices follow a lognormal distribution (since prices cannot be negative).

In contrast to the normal or lognormal models, results of a stock price study by Eugene Fama [4] in 1965 pointed to a stable Paretian distribution.[☼] Through analyses of the thirty stocks of the Dow Jones Industrial Average, he further states, "The presence, in general, of leptokurtosis in the empirical distributions seems indisputable." Stable Paretian density

---

[*] Brownian motion in physics is equivalent to a Wiener process in mathematics.

[†] Black was ineligible to receive the prize due to his death in 1995, but was mentioned as a contributor.

[‡] It is also known as the Black-Scholes-Merton model.

[☼] Benoit Mandelbrot, a Polish-born, French and American mathematician, identified non-Gaussian stable distributions as "stable Paretian" distributions [48].

functions, except for the special limiting case of a Gaussian, are heavy-tailedЖ and leptokurtic.

While the Wiener process proposed by Bachelier follows a Gaussian (or Normal) distribution, the Black-Scholes model obeys a lognormal distribution, and Fama's study yielded a stable Paretian distribution, there remains much ambiguity of actual stock price stochastic characteristics. Stable distributions, i.e., the Paretian, have seen limited use because in general they have no closed-form expression. Nontrivial numerical approximations exist, but are computationally onerous [5].

There have been a few studies of the distribution of stock returns in various world markets (see for example [6], [7], and [8], and included references), but these typically address the assumption of normality versus other distributions such as the Paretian, logistic, exponential power, scaled-t, or mixtures of Gaussian distributions. Data time-indexing of these stock return analyses range from monthly or weekly down to intraday high-frequency increments. Investigations primarily seek to quantify the lack of fit. Most, if not all, empirical studies of late reject the assumption of normality.

Ж Usage here connotes heavier (thicker) tails than the Gaussian distribution.

1.4 Stock Distribution Models

Herein, the investigation considers three fluctuation models for characterizing stock distributions: 1) simple daily return, 2) normalized model residuals, and 3) normalized detrended stock movement. These are defined as follows.

(1) Simple Daily Return – The one-period simple return, R_n, is defined as [9]

$$R_n = \frac{P_n - P_{n-1}}{P_{n-1}} = \frac{P_n}{P_{n-1}} - 1 \ , \tag{1-1}$$

where P_n is the price of an asset at time sample n. This (or a logarithmic version) is the most commonly reported criterion used for stock distribution analyses and produces normalized values centered about zero. The formula clearly reveals a normalized first-order difference equation, a simple high-pass filter.

(2) Normalized Model Residuals – The set of results produced in Chapter 2 yields a univariate autoregressive integrated moving average (ARIMA) model for each stock in the input data set. Model residuals are the difference between the observed value and the estimated value from the model. Residuals represent an "observable estimate of the unobservable statistical error." [10] "If the trend model is reasonably correct, then the residuals should behave roughly like the true stochastic component, and various assumptions about the stochastic component can be assessed by looking at the residuals." [11] Mathematically, the normalized model residuals, NMR_n, are defined as

$$NMR_n = \frac{MR_n}{D_n} \ , \tag{1-2}$$

where MR_n is the model residual and $D_n = P_n$ is the data at time sample n.

(3) Normalized Detrended Stock – As mentioned previously and seen in Figure 1-1, stock trends continually vary. In order to analyze the probability distribution of a time series, detrending, i.e., removal of the trend, is required. It is standard practice in the discipline of signal processing to reduce noise by filtering, exposing the desired signal for further processing and analysis. For extracting the trend from a stock time series, a two-sample moving-average filter[*] is applied and the smoothed filter output is subtracted from the original time series,[†] leaving a set of stochastic residuals. Mathematically, the normalized detrended stock, NDS_n, is defined as

$$NDS_n = \frac{D_n - F(D_n)}{F(D_n)} = \frac{D_n}{F(D_n)} - 1$$
$$= \frac{D_n - D_{n-1}}{D_n + D_{n-1}} \qquad (1\text{-}3)$$

where $F(D_n)$ is the filtered version of the observation data D_n. Conceptually, this approach includes elements of both models (1) and (2). Normalization is performed relative to the filtered signal which represents the estimated model.

Normalization is applied to the model residuals and the detrended residuals to allow direct comparison with the daily-returns model.

_____

[*] The two sample moving-average filter is defined as $F(D_n) = (D_n + D_{n-1})/2$.

[†] This is equivalent to high-pass filtering the data.

1.5 Stock Distribution Analysis

1.5.1 Candidate Distributions

The approach taken for analyzing the stock distribution models was to construct frequency distributions, i.e., histograms, for individual stocks. Preliminary visual evaluation of histograms of the distribution models of a few randomly selected stocks from the stock list implied several candidate distributions: 1) Cauchy, 2) Laplace, 3) generalized-t, and 4) Gaussian.[*] The more general skewed versions of these distributions were utilized in the analysis to account for any possible asymmetry in the observation data. As developed in [12] and [13], all of these distributions may be derived from the skewed generalized t-distribution (SGT). The SGT distribution exhibits a high degree of adjustability and many special cases. Five parameters describe the distribution: μ, σ, λ, p, and q. Figure 1-2 depicts the relationships between the many special cases and identifies the parameter values required to represent any particular distribution.[†]

---

[*]   While the Gaussian was visually a poor match for the data, it was included as a reference due to its widespread historical usage.

[†] The figure is adapted from [13] and [49].
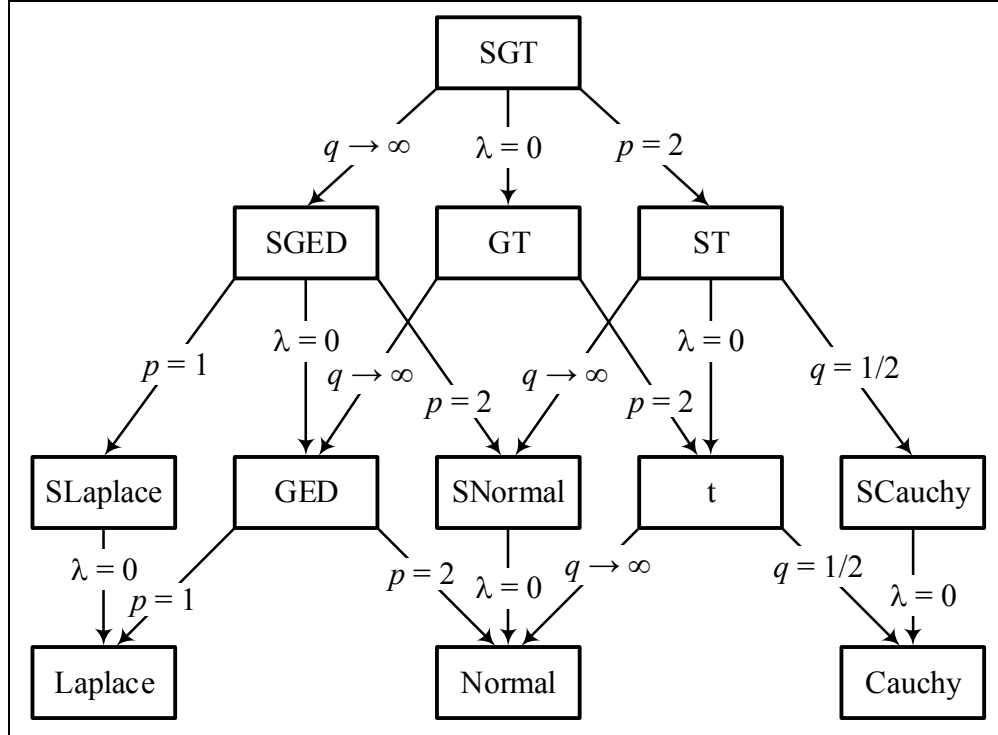
Figure 1-2. Skewed Generalized t-Distribution Interrelationships[*]

The probability density function (PDF) of the skewed generalized t-distribution is defined

as

$$f_{SGT}\left(y;\mu,\sigma,\lambda,p,q\right)$$

$$= \frac{p}{2\sigma q^{1/p}\,\mathrm{B}\left(\dfrac{1}{p},q\right)\left(1+\dfrac{\left|y-\mu\right|^{p}}{q\sigma^{p}\left(\lambda\cdot sign\left(y-\mu\right)+1\right)^{p}}\right)^{q+\frac{1}{p}}}, \qquad (1\text{-}4)$$

---

[*] Abbreviations used in the figure: S = Skewed, G = Generalized, ED = Error Distribution, T = t = t-distribution.

where $B(\cdot,\cdot)$ is the beta function, $p > 0$ and $q > 0$ govern the kurtosis (height and tails) of the density[*], μ is the mean (i.e., location parameter), $\sigma > 0$ controls the variance, and $-1 < \lambda < 1$ regulates the skewness. Table 1-1 summarizes the fixed parameters seen in Figure 1-2 that are required for the tested special-case distribution models. By definition, none of the SGT parameters in Figure 1-2 has a fixed value.[†]

Table 1-1. Summary of SGT Fixed Parameters

	Cauchy	Laplace	Normal
p	2	1	2
q	1/2	∞	∞

Thus, the PDF for the skewed Cauchy distribution simplifies to

$$f_{SCauchy}(y;\mu,\sigma,\lambda) = \frac{\sqrt{2}}{\sigma\pi\left(1 + \dfrac{2 \cdot |y-\mu|^2}{\sigma^2(\lambda \cdot sign(y-\mu)+1)^2}\right)}. \tag{1-5}$$

Additionally, the PDF for the skewed Laplace distribution reduces to

$$f_{SLaplace}(y;\mu,\sigma,\lambda) = \frac{1}{2\sigma}\exp\left(\frac{-|y-\mu|}{\sigma(\lambda \cdot sign(y-\mu)+1)}\right), \tag{1-6}$$

---

[*] Kurtosis increases as p and q decrease resulting in fatter tails, i.e., a leptokurtic distribution. Increasing p and q reduces kurtosis (possibly yielding negative excess kurtosis), resulting in thinner tails and a distribution that is more platykurtic.

[†] It may be noted that in the common t-distribution, also known as Student's t-distribution, parameter $p = 2$ and q is unfixed.

and the PDF for the skewed Normal distribution is

$$f_{SNormal}(y;\mu,\sigma,\lambda) = \frac{1}{\sigma\sqrt{\pi}}\exp\left[-\left(\frac{|y-\mu|}{\sigma(\lambda\cdot sign(y-\mu)+1)}\right)^2\right]. \qquad (1\text{-}7)$$

Computations were performed using the R-language version 3.2.5 and the R-Package 'sgt' version 2.0. Model optimization was executed with the function *sgtmle* to identify all five SGT parameters. Function *sgtmle* fits data to the skewed generalized t-distribution using maximum likelihood estimation and function *dsgt* was used to compute the model values for each density profile.

1.5.2   Goodness of Fit

After retrieving the input data and preparing the three model sets described in §1.4, a goodness-of-fit measure was computed for each model for each of the 197 stocks in the list. Goodness-of-fit models are used to compare fitted (or theoretical) models with observed data. Several measures for goodness of fit were considered as given in the subsequent five formulas. Note that in the following, y_i are the observation values and y_i^* are the fitted model values.

1. Mean Absolute Error (MAE):

$$MAE = \xi_1 = \frac{\sum_{i=1}^{n}\left|y_i - y_i^*\right|}{n} \qquad (1\text{-}8)$$

With MAE, outliers have less influence, but the value is easily understood. There is a disadvantage with noisier data since it places equal weighting on all deviations.

2. Root Mean Square Error (RMSE):

$$RMSE = \xi_2 = \sqrt{\frac{\sum\limits_{i=1}^{n}\left(y_i - y_i^*\right)^2}{n}} \qquad (1\text{-}9)$$

With RMSE, there is more emphasis or influence by outliers due to the squaring operation. This measure is used often and is an absolute measure.

3. Normalized Mean Absolute Error (NMAE):

$$NMAE = \delta_1 = \frac{\xi_1}{\frac{1}{n}\sum\limits_{i=1}^{n} y_i} = \frac{\sum\limits_{i=1}^{n}\left|y_i - y_i^*\right|}{\sum\limits_{i=1}^{n} y_i} \qquad (1\text{-}10)$$

This relative measure is used to compare data sets with different scales.

4. Normalized Root Mean Square Error (NRMSE):

$$NRMSE = \delta_2 = \frac{\xi_2}{\frac{1}{n}\sum\limits_{i=1}^{n} y_i} = \frac{\sqrt{\dfrac{\sum\limits_{i=1}^{n}\left(y_i - y_i^*\right)^2}{n}}}{\frac{1}{n}\sum\limits_{i=1}^{n} y_i} \qquad (1\text{-}11)$$

NRMSE, additionally known as the Coefficient of Variation, is also used to compare data sets with different scales.

5. RMSE Normalized to standard deviation (σ):

$$NRMSE_\sigma = \delta_3 = \frac{\xi_2}{\sqrt{\dfrac{1}{n}\displaystyle\sum_{i=1}^{n} y_i^2}} = \frac{\xi_2}{\sigma} = \sqrt{\frac{\displaystyle\sum_{i=1}^{n}\left(y_i - y_i^*\right)^2}{\displaystyle\sum_{i=1}^{n} y_i^2}} \qquad (1\text{-}12)$$

This approach is most appropriate and commonly used when the distribution is Normal, since each multiple of σ indicates a percentage of values lying within the σ band, i.e., 68.3–95.5–99.7%. In the case of non-Gaussian distribution, normalization to σ does not represent these same well-known percentages; however, for a unimodal distribution with a known mode, the Gauss Inequality (also known as the Camp-Meidell Inequality) may be utilized to find associated percentage bands 55.6–88.9–95.1%. (See [14] and Appendix 8.4 for a discussion.)

After preliminary evaluation of some densities, NRMSE (1-11) was selected as the best goodness-of-fit measure for use in the investigation since 1) non-Gaussian densities were evident, 2) there was uncertainty regarding the occurrence of outliers, 3) it was desired to maintain the significance of outliers, 4) normalization was required to compare the diverse price ranges of included stocks, and 5) resulting magnitudes were larger compared to $NRMSE_\sigma$ which was the next desirable choice.

For additional assessment, the Cramer-Von Mises (CVM) test was applied as a type of goodness-of-fit test based on the empirical distribution function. The basis of the method is to determine the quality of agreement for an observed sample cumulative distribution

function (CDF) to a hypothesized CDF. The Cramer-Von Mises test statistic for evaluating the null hypothesis $H_0 : X \sim F(x)$, that the distribution X follows a hypothetical distribution $F(x)$, is defined as [15] [16]

$$\begin{aligned} CVM &= \frac{1}{12n} + \sum_{k=1}^{n} \left(F(x_{k:n}) - \frac{2k-1}{2n} \right)^2 \\ &= \frac{1}{12n} + \sum_{k=1}^{n} \left(U_{k:n} - \frac{2k-1}{2n} \right)^2 \end{aligned} \quad , \tag{1-13}$$

where $x_{1:n}, \ldots, x_{n:n}$ refers to an ordered random sample of size n, and $U_{k:n}$ are ordered uniform variables.[*] The computed CVM statistic value is subsequently compared against a table of critical values, $CVM_{1-\alpha}$, to find its corresponding p-value. The null hypothesis, H_0, should be rejected if $CVM \geq CVM_{1-\alpha}$ for an α level of significance. It is important to note that the CVM test evaluates a Type II statistical error.[†] Computations for the Cramer-von Mises test were performed with the function *cvm.test* available in the R-Package 'goftest' version 1.0-3.

---

[*] The CVM test of an arbitrary distribution is equivalent to a test of uniformity as developed in [15] and [16].

[†] The Type II error is defined as incorrectly retaining a false H_0. In other words, there is a failure to have enough statistical evidence to reject H_0, which is in no way equivalent to having strong evidence to support H_0.

1.6 Distribution Analysis Results

Each of the four distributions previously identified (skewed versions of Cauchy, Laplace, Gaussian, and Generalized-t) were individually fit to all 197 stocks in the stock list. Moreover, every fit was performed for the daily returns, model residuals, and detrended residuals. For a sense of fits for the various distributions, Figure 1-3 shows an example of stock distributions falling near the average of the range of fit errors.[*] Clearly, all three sets of observation data have similar types of distributions. The horizontal scales of daily returns and model residuals are very similar, but the horizontal scale for the detrended residuals is much narrower. Inasmuch as all three fluctuation models are normalized, there appears to be less spread, i.e., smaller variance, in the filtered (detrended) model. This is explained by the additional D_{n-1} term in the denominator of equation (1-3) which reduces the magnitude of the first difference, i.e., the numerator, by approximately two. The vertical axes are set in logarithmic form because a linear scale provided extremely poor visibility for distinguishing between curves. With the logarithmic scale it clear that the Gaussian fit is worst in the tail regions. Corresponding numerical results for Figure 1-3 are given in Table 1-2.

---

[*] Note that "Observations" in Figure 1-3 represent conventional histogram bar heights. Bars were not included in an attempt to enhance visualization.
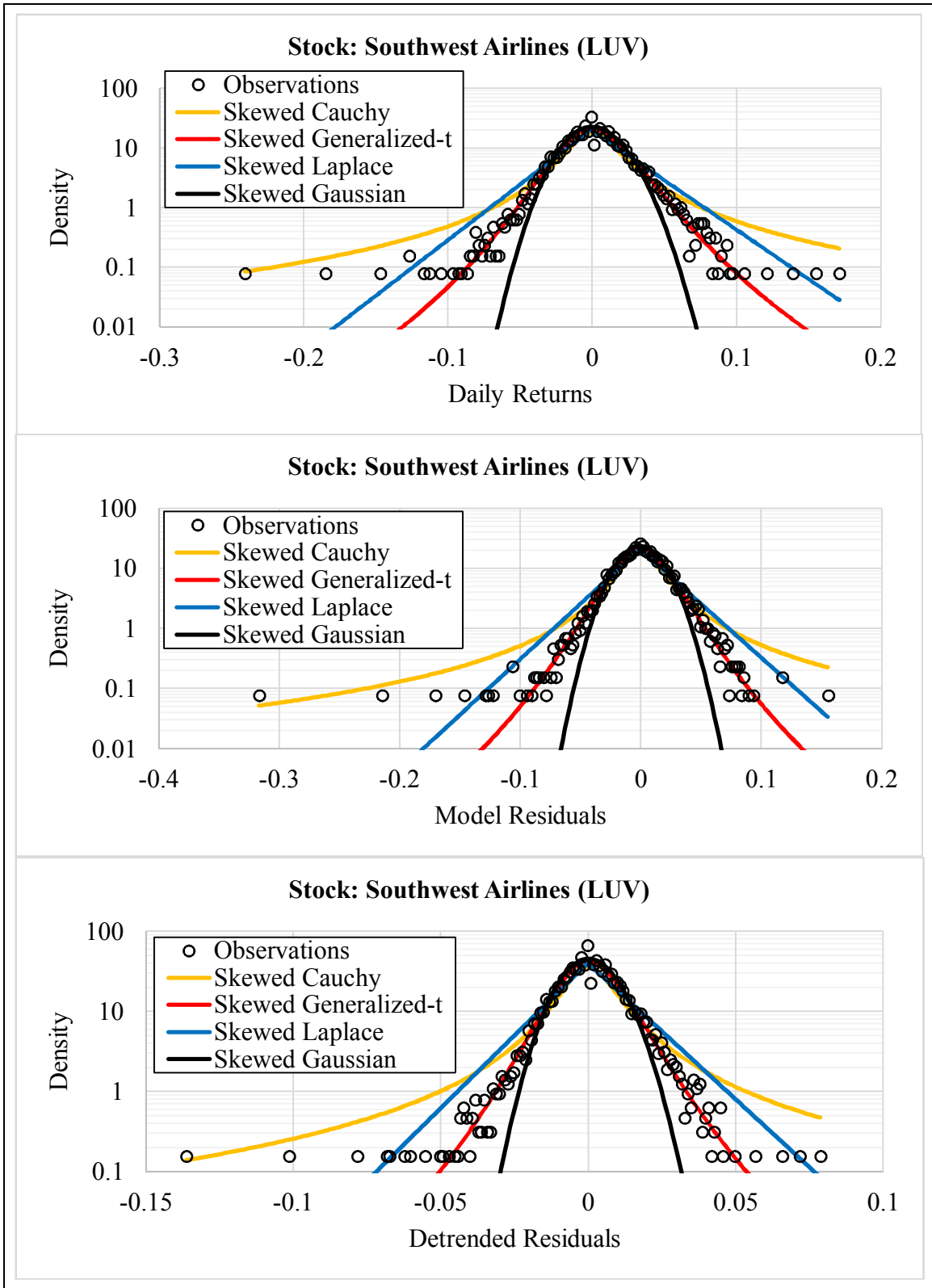
Figure 1-3. Typical Example of Distribution Fits

Table 1-2. NRMSE Summary for Southwest Airlines (LUV)

	Daily Returns	Model Residuals	Detrended Residuals
Skewed Cauchy	0.510	0.374	0.509
Skewed Laplace	0.487	0.344	0.493
Skewed Gaussian	0.411	0.265	0.409
Skewed Generalized-t	**0.361**	**0.150**	**0.359**
μ	-0.001	-0.001	0.000
σ	0.025	0.024	0.012
λ	0.059	0.014	0.034
p	1.786	1.693	1.790
q	3.163	3.815	3.146

The top section of Table 1-2 shows the fit errors for all distributions and fit models for example stock symbol LUV. Lowest errors are in bold. Although all three models were normalized in an attempt to allow comparisons, of principal importance were relative magnitudes within each column. A key result was that the SGT distributions yielded the best fit (lowest error) in all cases. That the SGT showed the best fit was apparent in the curves shown in Figure 1-3 as observations tended to fall on both sides of the SGT curves. It was indisputable that the Gaussian distribution possessed a visibly poor fit as it fell off in the tail regions, which tended to support rejection of normality as noted in other studies. Additionally, the goodness of fit for the Cauchy distribution was limited due to its characteristic fatter tails.

Parameter estimates from maximum likelihood estimation of the SGT distribution are shown in the bottom section of Table 1-2. Mean, μ, and skew parameter, λ, were negligibly small, indicating a high degree of symmetry about zero. The value of the variance-control parameter, σ, was virtually the same for daily returns and model residuals, but was half as large for the detrended model. This discrepancy was previously attributed to the effects of the moving average filter. Parameter values for p and q match closely across all models. With $p \approx 1.76$ and $q \approx 3.4$, the estimated curve was leaning toward a Student's t-distribution. A review of the graphs revealed that the optimum SGT fit for all three fluctuation models appeared nearly equidistant from both Laplace and Gaussian distributions.

Figure 1-4 shows the normalized RMS errors across all 197 stocks in the stock list for daily returns and detrended residuals. Although not shown, behavior of the model residuals was similar. As a general rule, the SGT distributions yielded the lowest errors. This is seen in the red series in the plots. Also noteworthy was that the Gaussian distributions typically have the highest errors, although there are several exceptions.

Table 1-3 provides the numerical statistics of the normalized RMS errors for the entire stock list for each of the three models. The SGT distribution yielded the lowest errors in every category, as indicated in bold.
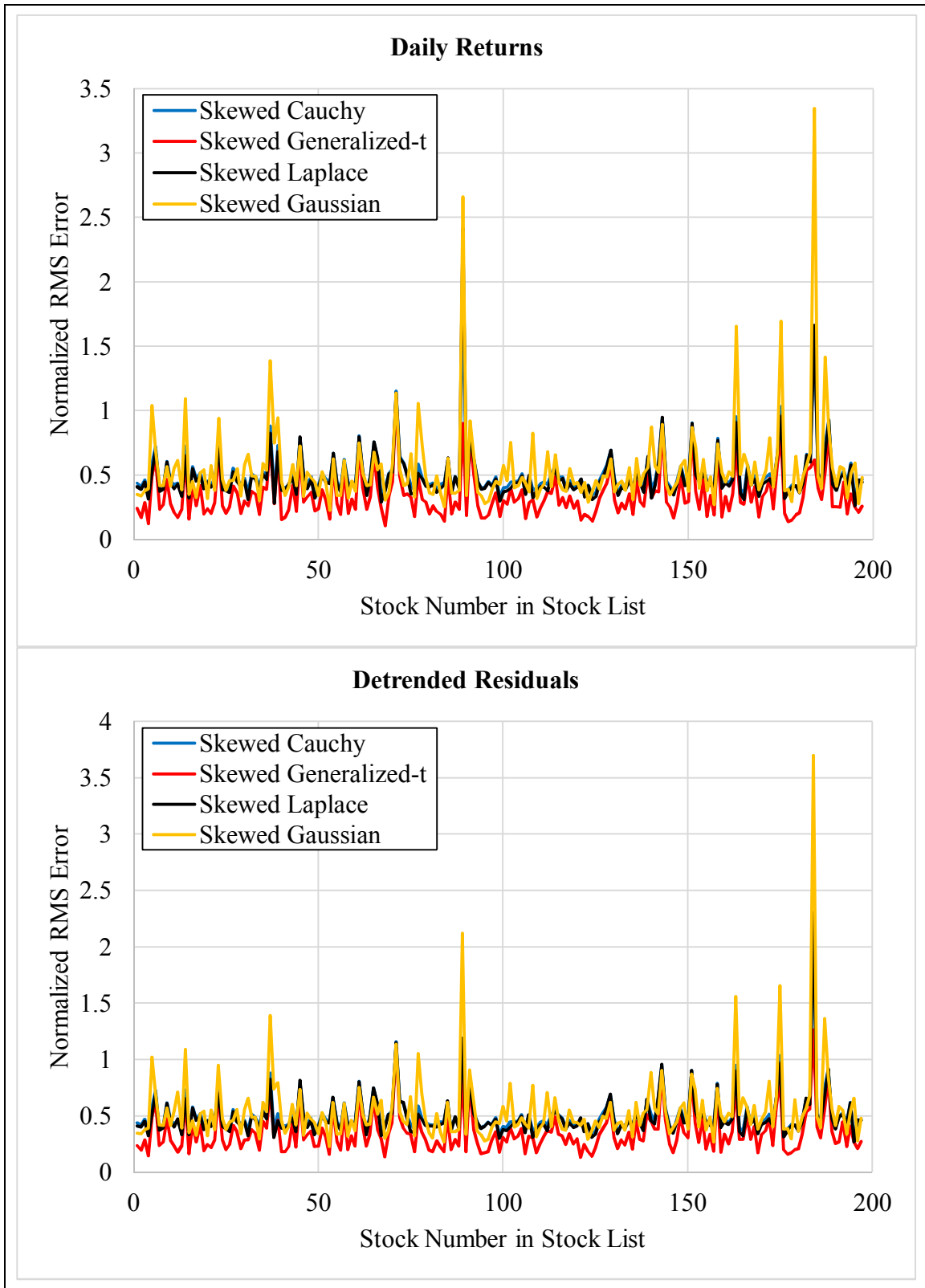
Figure 1-4. Normalized RMS Errors Across All Stocks

Table 1-3. Summary Statistics of NRMSE for All Distribution Estimates

	Skewed Cauchy	Skewed Laplace	Skewed Gaussian	Skewed Generalized-t
Daily Returns				
Median	0.454	0.439	0.460	**0.292**
Mean	0.508	0.488	0.548	**0.341**
Max	1.888	2.414	3.348	**1.099**
Min	0.317	0.265	0.230	**0.110**
Std Dev	0.185	0.212	0.335	**0.169**
Model Residuals				
Median	0.381	0.356	0.414	**0.174**
Mean	0.407	0.376	0.465	**0.204**
Max	2.029	2.018	3.616	**0.701**
Min	0.239	0.164	0.244	**0.054**
Std Dev	0.146	0.150	0.291	**0.108**
Detrended Residuals				
Median	0.458	0.446	0.466	**0.306**
Mean	0.509	0.491	0.547	**0.351**
Max	2.227	2.303	3.698	**1.264**
Min	0.316	0.269	0.229	**0.134**
Std Dev	0.187	0.194	0.332	**0.177**

A graph of results for the second goodness-of-fit evaluation using the Cramer-von Mises test is shown in Figure 1-5. The frequency of occurrence (count) for values less than the significance level (threshold = 0.05) are given in the legend for each model. These computed to approximately 12%, 11%, and 3% for daily returns, detrended residuals and model residuals, respectively. Overall, there was good agreement that most of the stock distributions ($\approx$ 90%) were accurately classified as generalized t-distributions.
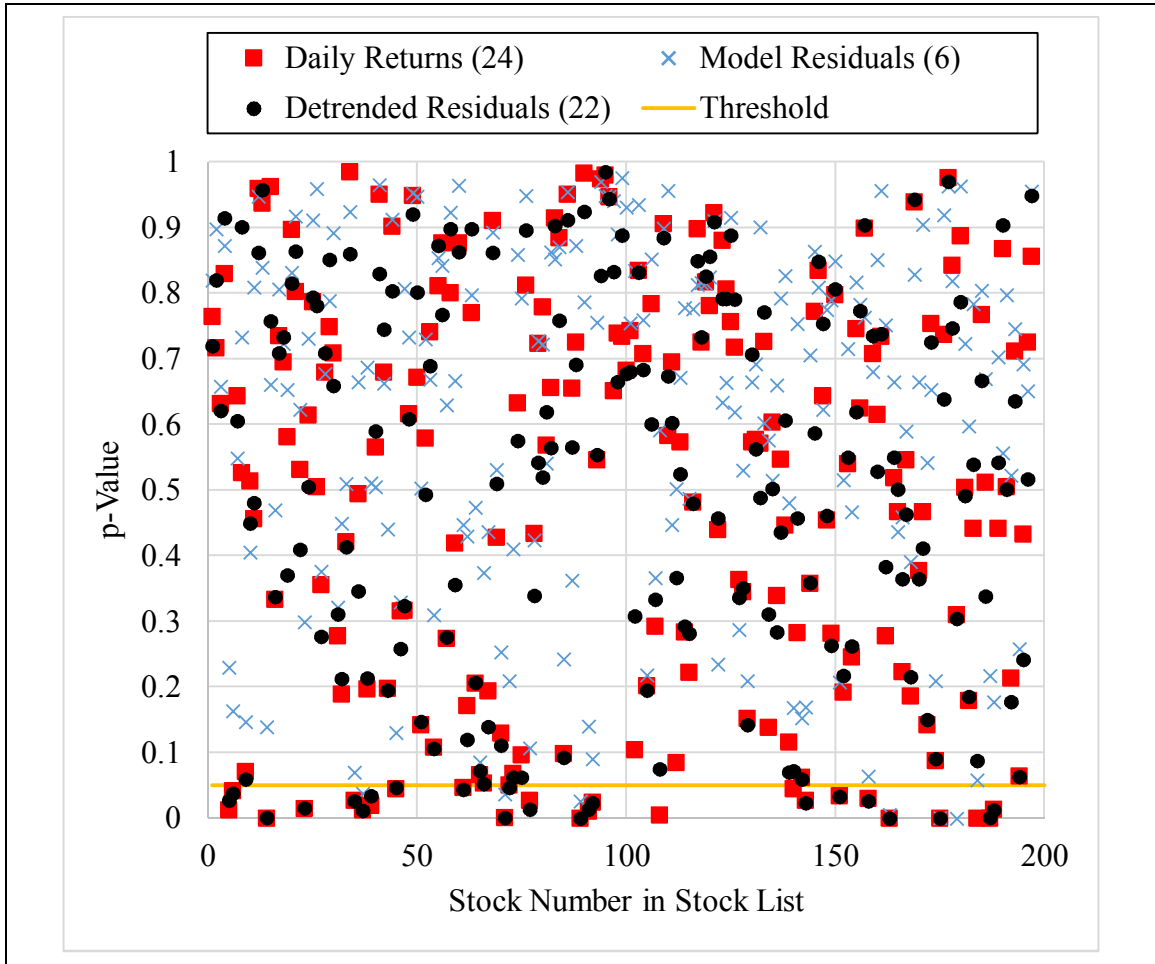
Figure 1-5. p-Values for Empirical CDF Goodness-of-Fit

A summary of the statistics of the estimated parameters is given in Table 1-4. Parameters μ, λ, and σ were consistent with characteristics mentioned for the example stock LUV. Parameter $p \approx 1.64$ for the daily returns and detrended residuals and was slightly lower at 1.58 for the model residuals, all similar to the LUV example. The median of parameter q ranged from 2.66 to 2.94, while the mean was quite variable due to some extreme values. Overall, there was a high degree of consistency for each model across the array of stocks.

Table 1-4. Summary Statistics of Parameters for All Distribution Estimates

	μ	σ	λ	p	q
Daily Returns					
Median	0.000	0.017	0.028	1.639	2.662
Mean	0.000	0.018	0.027	1.622	386.9
Max	0.001	0.039	0.089	2.698	33408.4
Min	-0.002	0.0005	-0.045	0.345	0.860
Std Dev	0.001	0.005	0.021	0.314	2800.3
Model Residuals					
Median	0.000	0.017	-0.003	1.584	2.942
Mean	0.000	0.018	-0.005	1.576	338.6
Max	0.001	0.037	0.027	2.576	22890.0
Min	-0.001	0.006	-0.065	0.572	0.926
Std Dev	0.000	0.004	0.016	0.255	2695.9
Detrended Residuals					
Median	0.000	0.009	0.013	1.647	2.665
Mean	0.000	0.009	0.009	1.625	435.1
Max	0.001	0.019	0.071	2.701	25197.5
Min	-0.001	0.002	-0.602	0.675	0.614
Std Dev	0.000	0.002	0.048	0.310	2604.8

Figure 1-6 shows a graph of the *p-q* relationships of the parameter estimates for each model for each stock. Clearly, most of the estimates fell in the range of $1 \le p \le 2.7$ and $0.9 \le q \le 10$. It is rather deceiving that the estimates fall closest to the Cauchy distribution in Figure 1-6 in contrast to the example of Figure 1-3 which shows the Laplace distribution as a better visual match to $(p,q) \approx (1.76, 3.4)$. Of course, some data points fell on the standard t-distribution (vertical) line. Evidently, the SGT model has a substantial sensitivity to *p* and *q* values with respect to distribution tails.
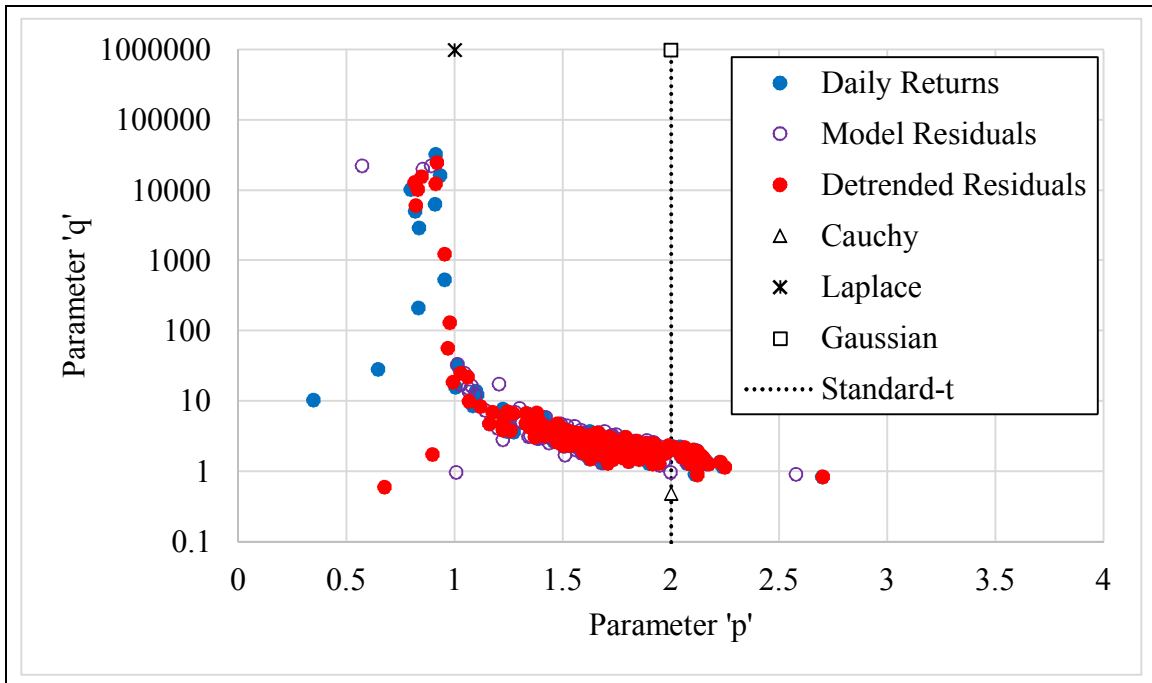
Figure 1-6. Parameter Estimates

1.7 Summary of Stock Characteristics

The evidence from analyzing approximately forty percent of the stocks in the S&P 500 resulted in a strong rejection of the assumption of Gaussian distribution of major U.S. equities. Studied stock data belonged to the category of large market capitalization (large-cap) and spanned over 25 years. Distributions of daily stock returns, model residuals, and detrended residuals all showed fat tails and high peaks as displayed in the example histogram of daily returns in Figure 1-7 (using a linear vertical scale)[*]. The alternative distributions of Cauchy, Laplace, and Gaussian all resulted in poorer fits than the Generalized-t model. Skewness and mean offsets were found to be negligible.

---

[*] This is the same observation data shown in Figure 1-3.
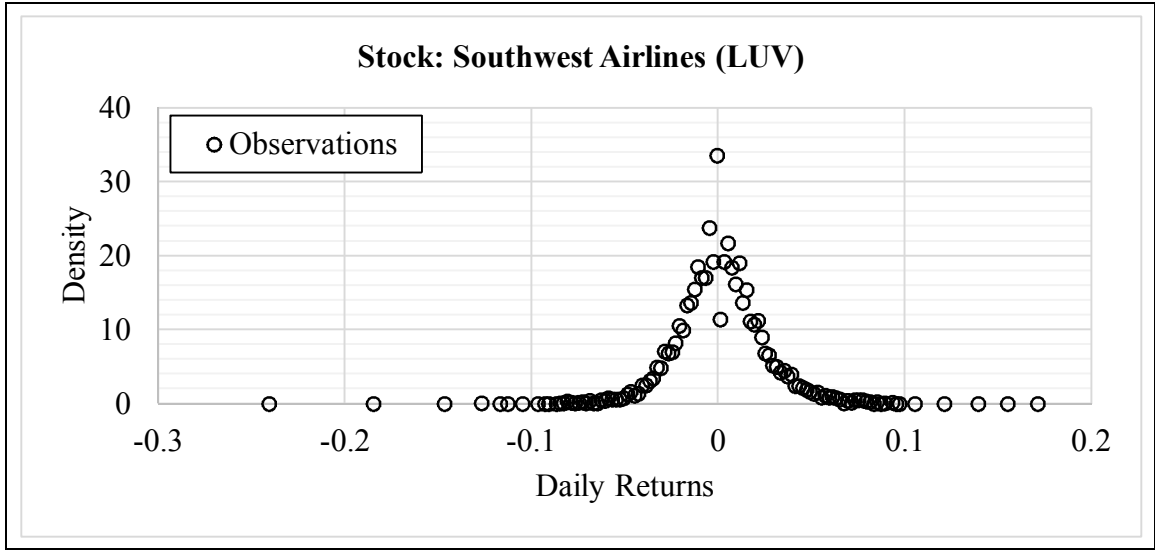
Figure 1-7. Typical Example of Non-Logarithmic Histogram Scale

2. STATISTICAL FORECASTING OF TIME SERIES

"Probability is expectation founded upon partial knowledge."

– George Boole (1815-1864), in the book "Collected Logical Works," vol. 2, 1940

This chapter addresses an investigation into the forecasting of stock time series based on statistical methods. All of the concepts and definitions introduced in Chapter 1 are applicable here as well. In addition to computing forecasts, their accuracies must be specified and this is expressed with probability limits, i.e., confidence intervals, on each side of a forecast value. While any convenient limits may be used, 95% confidence intervals (CI) are utilized in this investigation. In other words, there is a 95% probability that a realized value will occur within the CI limits.

The study compares the results of statistical forecasting of univariate time series to multivariate time series. Forecasting methods are applied to a time series, Z_n, assuming that the time series can be modeled as a stochastic process with an identifiable form. The paradigm used is the very general autoregressive integrated moving average (ARIMA) model. The hypothesis investigated is whether a multivariate time series of correlated stocks produces better forecast results than a single (univariate) series alone.

2.1 Analysis Approach – Windows and Splits

Each stock series examined spanned a range of 25 years of daily samples for a total of 6491 data points. To obtain independent statistics, a stock time series was divided into non-overlapping windows of 100 samples each*. This generated 64 completely independent sample windows for each stock. Analyzing 100-sample windows avoided any seasonal or repetitive cycles within or between windows that may have occurred over the data span of 25 years. One-hundred sample windows also provided a sufficient set of training data to establish an adequate model.

Stock time series have two unique aspects that separate them from naturally-occurring or scientifically-based time series: splits and ex-dividend days. A stock split causes no change in a shareholder's portfolio value nor in a firm's asset value. The occurrence of a split appears as a noticeable instantaneous shift in the time series pattern at some specified date.† For securities in the stock list, splits typically occurred 3-5 times over the 25 year data range. A split was managed as follows. For each stock, split dates and corresponding split ratios were obtained from Yahoo! Finance via the *getSplits* function in the R-Package 'quantmod' version 0.4-5. If a stock splits on day n with ratio r, e.g., a 2-for-1 split returns $r = 0.5$, then all price data in the window prior to day n are multiplied by r.‡ Implementing this adjustment produces a price time series without the large shift, providing continuity

---

* One hundred daily samples correspond to approximately 4.75 calendar months.

† Examples can be seen as vertical price changes in Figure 1-1.

‡ This approach works for forward or reverse splits.

for model building. On the other hand, the processing of any dividend payout on a stock was ignored and considered insignificant for this study.

2.2 Univariate Analysis

To analyze the univariate time series, computations were performed in the R language employing the *auto.arima* function from the R-Package[*] 'forecast' version 6.2. The *auto.arima* function returns the best-fit ARIMA model when evaluated against the Corrected Akaike Information Criterion (AICC), by searching over all possible models within specified model order constraints. The search range for model order was limited to a maximum of 5 for p and a maximum of 5 for q.

Univariate ARIMA Model

The ARIMA model combines the autoregressive (AR) and moving average (MA) models with differencing, also known as the integrated (I) component. Differencing parameter d specifies the number of first differences applied to the series to achieve stationarity. Some noteworthy models are identified in Table 2-1 [17].

---

[*] All R packages referenced herein may be downloaded from the R Archive Network [50].

Table 2-1. Special ARIMA Models

Name	Model
White Noise	ARIMA(0,0,0)
Random Walk	ARIMA(0,1,0) with no constant
Random Walk with Drift	ARIMA(0,1,0) with constant
AR(p)	ARIMA(p,0,0)
MA(q)	ARIMA(0,0,q)

A time series is required to be stationary in order to produce a proper model estimate. If the model is estimated in the presence of any non-stationary parameters, the estimated coefficients may be incorrect. Differencing makes the time series stationary. The number of required differences can be determined by checking for unit roots. In the *auto.arima* function, the Kwiatkowski–Phillips–Schmidt–Shin (KPSS) unit-root test was used.[*]

The non-seasonal ARIMA(p,d,q) model[†] is defined as

$$\varphi(B)(1-B)^d \, y_t = c + \theta(B) \, \varepsilon_t \, , \tag{2-1}$$

where $\varphi(z)$ is the AR polynomial of order p and $\theta(z)$ is the MA polynomial of order q, B is the backshift operator, $\{\varepsilon_t\}$ is a white noise process with zero mean and variance σ^2, and c is a constant. Equation (2-1) may be expanded as

---

[*] In the KPSS test, the null hypothesis is stationarity and the alternate hypothesis is a unit root, i.e., a non-stationary process.

[†] These parameters p and q are entirely different from those mentioned in Chapter 1.

$$\left(1 - \varphi_1 B - \cdots - \varphi_p B^p\right) y_t = c + \left(1 + \theta_1 B + \cdots + \theta_q B^q\right) \varepsilon_t , \qquad (2\text{-}2)$$

which may restructured into the predictive form as

$$y_t = c + \varphi_1 y_{t-1} + \cdots + \varphi_p y_{t-p} + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t . \qquad (2\text{-}3)$$

It is clearly seen in equation (2-3) that predictors include lagged values of y_t and lagged error terms of ε_t.

The *auto.arima* function cycles through various combinations of (p,q) model orders. After selecting a model order, estimation of parameters c, φ_1, ..., φ_p, θ_1, ..., θ_q is performed using maximum likelihood estimation (MLE). Maximization is accomplished via the AICC, defined as [18]

$$AICC = AIC + \frac{2(p+q+k+1)(p+q+k+2)}{T-p-q-k-2} , \qquad (2\text{-}4)$$

where the Akaike Information Criterion (AIC) is

$$AIC = -2\ln(L) + 2(p+q+k+1) , \qquad (2\text{-}5)$$

and L is the likelihood of the data[*], T is sample size, $k=1$ if $c \neq 0$, and $k=0$ if $c=0$. The (p,q) model, with estimated parameters, associated with the optimum AICC value is the final selected model.

---

[*] The likelihood is calculated with the *arima* function in the R-Package 'stats', which computes the exact likelihood using a state-space representation of the ARIMA process, and the innovations, i.e., errors, and their variance are found via a Kalman filter.

Forecasting characteristics of ARIMA models depend on the values taken by parameters c and d in the above formulas. Table 2-2 lists the expected performance traits [17]. Additionally, as parameter d increases, the size of forecast confidence intervals increase as well. On the low end, when $d = 0$, the forecast standard deviation tends to the standard deviation of the historical (training) data. Interpretation of the long-term forecasting behavior of ARIMA models (Table 2-2) infers convergence to the data sample mean. This implies that the utility of stationary models is principally for short-term predictions.

Table 2-2. ARIMA Model Forecasting Behavior

c	d	Long-Term Forecast
0	0	Tends to zero
0	1	Tends to a non-zero constant
0	2	Follows a straight line
$\neq 0$	0	Tends to the mean of the data
$\neq 0$	1	Follows a straight line
$\neq 0$	2	Follows a quadratic trend

2.3 Univariate Results

Results of applying the *auto.arima* function yielded a maximum model order of 4 for parameters p or q across all windows of all stocks. In contrast, the frequency of occurrence of random-walk models, $(p, d, q) = (0, 1, 0)$, seen in Figure 2-1, was prominent, ranging

from 43 to 77%, with a mean of 63% (indicated by the red line).[*] All random-walk models

occurred with a first difference in the analyzed time series. From the properties of a random

walk, the forecast for a model without drift is simply a horizontal projection attached to the

last observed sample, while a model with drift may project along a sloped path up or down

from the last observation. While actual stock price movements have reasons for moving

as they do,[†] the implication of the random-walk model is that forecast direction and

magnitude cannot be predicted. Hence, a random-walk model is ineffective for producing

useful forecasts. Fortunately, random-walk models, although frequent, did not occur in

every sample window.
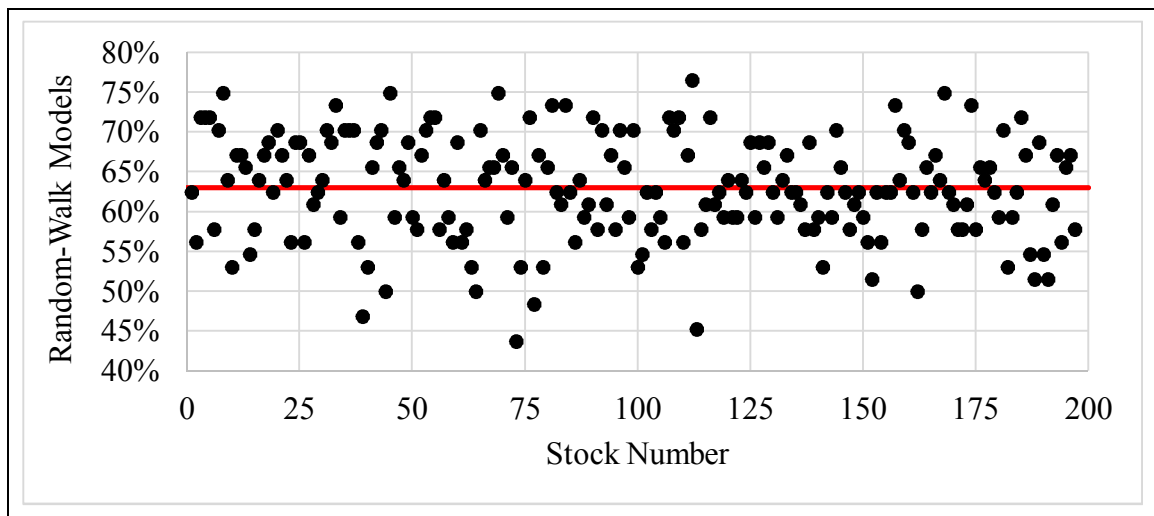
Figure 2-1. Frequency of Univariate Random-Walk Models

---

[*] Drift may or may not be present. Note that a random-walk model with drift indicates
   that the distribution of step sizes has a non-zero mean as opposed to no drift equating to
   a zero mean.

[†] Although beyond the scope of this work, there are many theories regarding stock price
   movement. To encapsulate, a stock's change in price is not random, without purpose,
   but may sometimes be explained subsequently.

It is generally believed that a well-fit ARIMA model exhibits a lack of serial correlation within the model's residuals [19]. While it is a common practice to use the Ljung-Box portmanteau test [20] for evaluating serial dependence, no theoretical basis has been developed for identifying the proper number of lags for the test [21]. For T data points, Hyndman [22] recommends using l lags where $l = \min(10, T/5)$. Since Ljung and Box published their test approach in 1978, alternative tests with more power have been developed [23] [24]. While the Ljung-Box test is time-based, these alternate tests are based in the spectral (density) domain and perform better on both Gaussian and non-Gaussian distributions, with only small loss of power. Normality of stock data is clearly absent as established in §1.6, and therefore any test must have the ability to verify serial correlation for non-Gaussian distributions.

To establish the reliability of the forecasting analyses, serial correlation tests were performed on the model residuals for all stocks in all windows using both the Ljung-Box portmanteau test and the test statistic, T_n, developed in [23]. The Ljung-Box test is implemented in the function *Box.test* in the R-package 'stats'[*] and the spectral density test for T_n is implemented by the function *UnivTest* in the R-package 'dCovTS' version 1.0. The Ljung-Box test was executed with a lag of 10 and degrees of freedom equal to ARMA

---

[*] The 'stats' package is included in the R core software.

model order $p+q$. The correlation T_n statistic was computed utilizing a Daniell

smoothing window,[*] bandwidth $= 10$,[†] and number of bootstrap replicates $= 499$.

Results of the serial correlation tests are shown in Figure 2-2. The top graph refers to the

Ljung-Box test results and the bottom plot refers to T_n test results. Both scatterplots show

for each stock the proportion of p-values (out of 64 windows) that are below the threshold

of 0.05 (shown as a red line). With a null hypothesis of independence, a small p-value

expresses strong implication against independence, rejecting the null hypothesis. Thus, if

a stock displays a small fraction of p-values below 0.05, then the hypothesis of

independence is accepted. The Ljung-Box results indicated 55 out of 197 stocks rejecting

the null hypothesis, whereas the spectral density statistic, T_n, rejected only two. These two

(PEP and VZ) were recomputed with the Daniell kernel and again found to have four

windows with p-values less than 0.05 as displayed in Figure 2-2. As seen in the figure, the

fractions appear in discrete jumps since they are increments of 1/64. Results for stocks

PEP and VZ produced a fourth window where the p-values were below 0.05, pushing them

over the 0.05 threshold. Any window with a low p-value may be interpreted as having a

model fit that is less than the best possible since some serial correlation of residuals

remains, but this does not invalidate its model.   Inasmuch as the T_n statistic has

---

[*] The computational time of the Daniell window was costly, but produced more accurate
     results (according to [23]). A Bartlett window was also run, yielding results showing
     no serial correlation for all stocks.

[†] Bandwidth was chosen to follow the Ljung-Box lag $= 10$.

demonstrated higher power [23], all models were accepted as independent and not serially correlated, validating the ARIMA model fits.
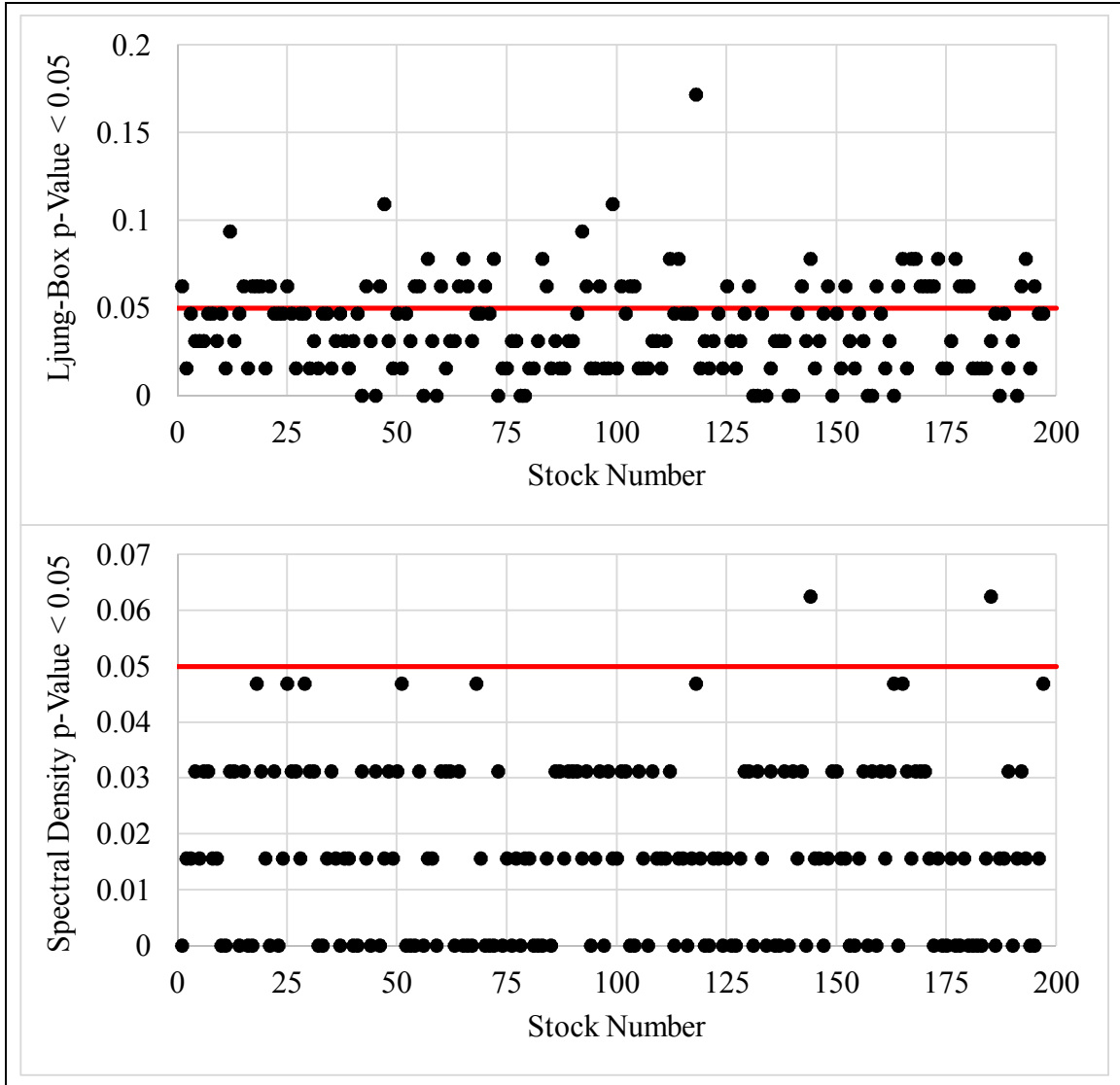
Figure 2-2. Serial Correlation p-Values for Univariate Model Residuals

Accuracy of forecasting direction of movement was found by comparing one-step ahead and five-step ahead predictions with out-of-sample data. Results of directional forecast accuracy for each stock are displayed in Figure 2-3 and summarized in Table 2-3. A 50%

threshold represents a type of result from a random occurrence, meaningless for effective forecasting. Results above (or below) the 50% threshold provide an improvement to a random guess, either forecasting the correct direction of movement, or forecasting incorrectly the direction of movement.* Table 2-3 also includes the number of occurrences for which the sign of forecast directional accuracy was the same for both one-step and five-steps ahead. In other words, if both the one-step and the five-step forecast were both greater than 50% or both less than 50%, that stock was counted in the total for being in the same direction. This may provide a stronger indication of expected directional movement.
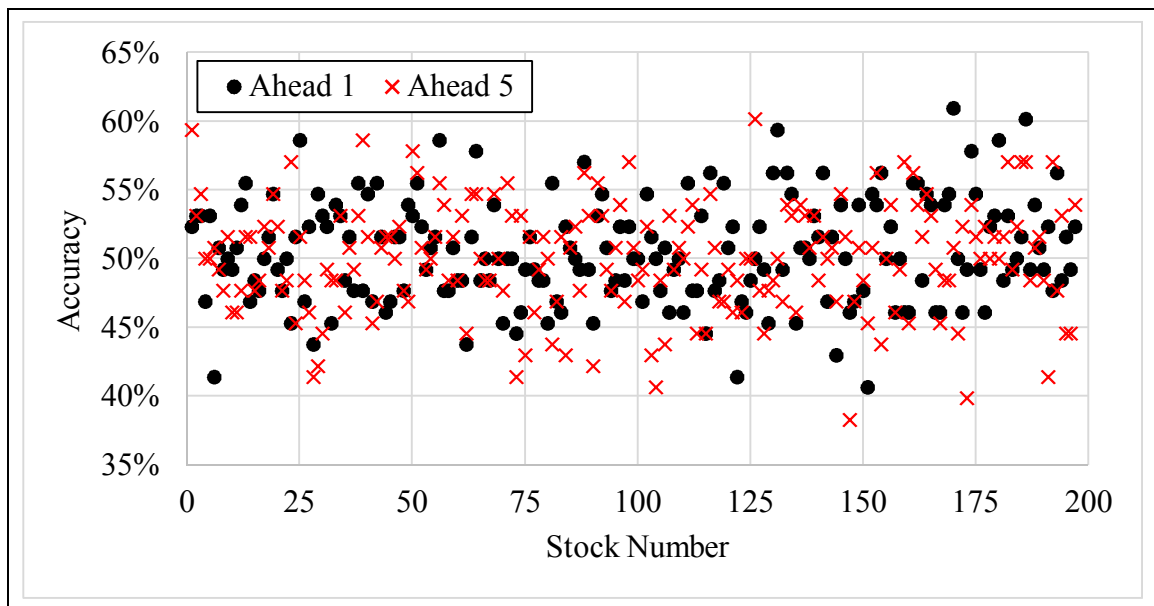
Figure 2-3. Univariate Directional Forecast Accuracy

---

* It may also be useful for trading to recognize (with some probability) that a forecast direction is incorrect.

Table 2-3. Univariate Directional Accuracy Summary

Forecast Interval	Accuracy	Frequency (Out of 197)
Ahead 1	> 60%	2
	> 55%	24
	> 50%	93
	< 45%	8
	< 40%	0
Ahead 5	> 60%	1
	> 55%	18
	> 50%	93
	< 45%	22
	< 40%	2
Ahead 1 & Ahead 5	Same Direction	111

Coverage refers to the percentage of forecasts that fall within a confidence interval. For each forecast window of each stock, the one-step- and five-steps-ahead forecasts falling within the 95% confidence interval were counted and the fraction of total windows was computed. Results are displayed in Figure 2-4. The results were greater than 82% for all stocks, with the average for one-step ahead at 92.5% and for five-steps ahead at 92.8%.
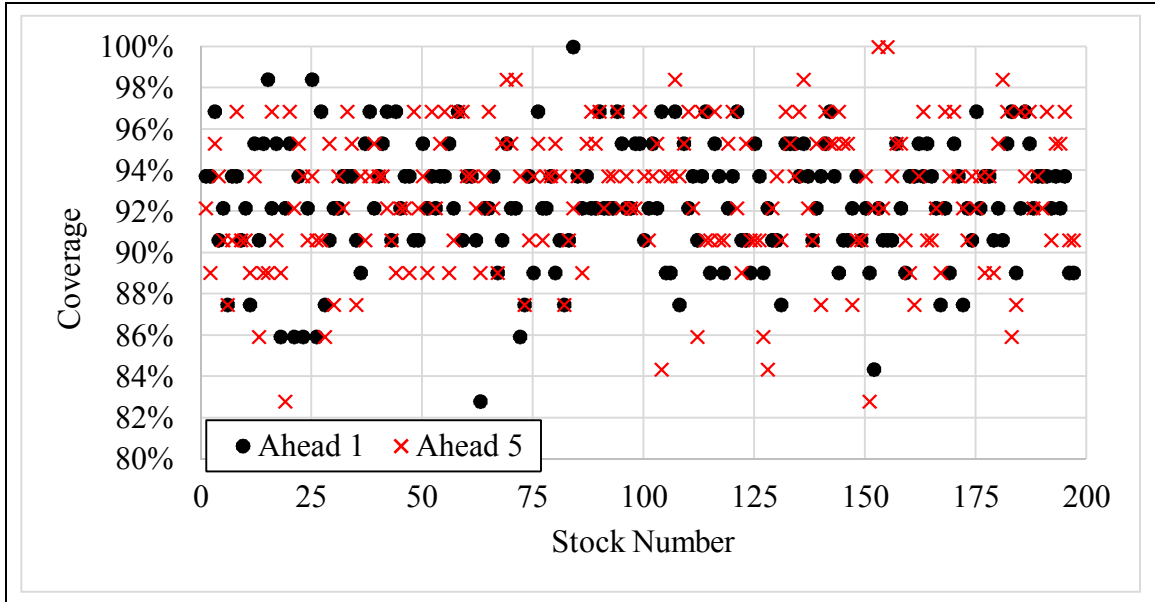
Figure 2-4. Univariate Forecast Coverage within 95% CI

2.4 Multivariate Analysis

Multivariate analyses were performed on pairs of correlated stocks. Vector time series

models were estimated utilizing the R language and a modified[*] version of Tsay's [25]

VARMA[†] function available from the R-Package 'MTS' version 0.33. Each time series

(100-sample window) was initially made stationary by differencing via the *ndiffs*[‡] function

available in the R-Package 'forecast' version 6.2. Search range for model order was limited

to $0 \le p \le 2$ and $0 \le q \le 2$, resulting in nine models computed for each multivariate series

---

[*] Any function described as a "modified" version connotes that the essential elements of
   the original function were retained, but that unused lines of code were eliminated.

[†] VARMA is the abbreviation for Vector Autoregressive Moving Average.

[‡] As in the univariate case, the KPSS test was applied as the unit-root test to determine
   the number of differences to reach stationarity. Computational results yielded a
   difference value of either zero, one, or two, for any window.

as given in Table 2-4. Forecasts were subsequently generated based on the "best" estimated model, i.e., the model with the lowest AIC, by employing a modified version of Tsay's *VARMApred* function found in the 'MTS' package. Inverse differencing was successively applied to recover the forecast in the original domain of the time series.

Table 2-4. VARMA Model Search Range

Model*	1	2	3	4	5	6	7	8	9
p	0	1	0	1	2	0	2	1	2
q	0	0	1	1	0	2	1	2	2

2.4.1 Pre-screening of Multivariate Pairs

As the computational cost of estimating VARMA models was found to be high during preliminary executions, a screening operation was performed to reduce the number of possible multivariate sets prior to estimating models. It was hypothesized that correlated stocks would improve forecast results. Multivariate correlation for all stocks across time-matched windows was performed with a modified version of Tsay's *ccm* function found in the 'MTS' package, implementing equations (2-8) to (2-11) as follows.

First, the lag-*l* cross-correlation matrix (CCM) for a stationary multivariate time series $\mathbf{z}_t$ is defined as [25]

---

* If identical AIC values were encountered, model-order preference was from 1 to 9.

$$\boldsymbol{\rho}_l = \mathbf{D}^{-1} \boldsymbol{\Gamma}_l \, \mathbf{D}^{-1} = \left[\rho_{l,ij} \right] , ^{*}$$
(2-6)

where $\mathbf{D} = diag\left\{ \sigma_1, \dots, \sigma_k \right\}$ is the diagonal matrix of standard deviations of the elements

of $\mathbf{z}_t$, and the lag-l cross-covariance matrix is

$$\boldsymbol{\Gamma}_l = \mathrm{cov}\left(\mathbf{z}_t, \mathbf{z}_{t-l} \right) = E\left[\left(\mathbf{z}_t - \boldsymbol{\mu} \right) \left(\mathbf{z}_{t-l} - \boldsymbol{\mu} \right)^{tr} \right] , ^{\dagger}$$
(2-7)

with $\boldsymbol{\mu} = E\left[\mathbf{z}_t \right]$ as a k-dimensional constant vector and the expected value of time series

$\mathbf{z}_t = \left(z_{1t}, z_{2t}, \dots, z_{kt} \right)^{tr}$. Clearly, k is the number of time-series vectors.

Correspondingly, the sample CCM components for the sample $\left\{ \mathbf{z}_t \right\}_{t=1}^{T}$ are estimated as

follows:

$$\hat{\boldsymbol{\mu}}_z = \frac{1}{T} \sum_{t=1}^{T} \mathbf{z}_t ,$$
(2-8)

$$\hat{\boldsymbol{\Gamma}}_0 = \frac{1}{T-1} \sum_{t=1}^{T} \left(\mathbf{z}_t - \hat{\boldsymbol{\mu}}_z \right) \left(\mathbf{z}_{t-l} - \hat{\boldsymbol{\mu}}_z \right)^{tr} ,$$
(2-9)

$$\hat{\boldsymbol{\Gamma}}_l = \frac{1}{T-1} \sum_{t=l+1}^{T} \left(\mathbf{z}_t - \hat{\boldsymbol{\mu}}_z \right) \left(\mathbf{z}_{t-l} - \hat{\boldsymbol{\mu}}_z \right)^{tr} ,$$
(2-10)

---

* Bold letters represent a matrix. $\left[\rho_{l,ij} \right]$ symbolizes a $k \times k$ matrix of correlation elements.

† $\mathbf{A}^{tr}$ is the transpose of $\mathbf{A}$ (not to be confused with T which is the sample size).

$$\hat{\boldsymbol{\rho}}_l = \hat{\mathbf{D}}^{-1}\,\hat{\boldsymbol{\Gamma}}_l\,\hat{\mathbf{D}}^{-1} = \left[\hat{\rho}_{l,ij}\right],\tag{2-11}$$

where $\hat{\mathbf{D}} = diag\left\{\sqrt{\hat{\gamma}_{0,11}},\ldots,\sqrt{\hat{\gamma}_{0,kk}}\right\}$ and $\hat{\gamma}_{0,ii}$ is the element in location (i,i) of $\hat{\boldsymbol{\Gamma}}_0$. To determine significance of correlation, $S_{l,ij}$, individual elements of the sample CCM $\hat{\boldsymbol{\rho}}_l$ were evaluated against the approximate 5% significance level with the test:

$$S_{l,ij} = \begin{cases} 1, & \text{if } \left|\hat{\rho}_{l,ij}\right| > 2/\sqrt{T} \quad \text{(significant)} \\ 0, & \text{if } \left|\hat{\rho}_{l,ij}\right| \le 2/\sqrt{T} \quad \text{(not significant)} \end{cases}^{*}.\tag{2-12}$$

The 19306 stock pair combinations were then sorted in order of frequency of significantly-correlated windows. Those with greater than 90% correlated windows were selected for forecasting analyses.

2.4.2   Multivariate ARIMA Model

The multivariate ARMA model, or vector ARMA (VARMA), is defined as a multi-dimensional form of the univariate ARMA model:

$$\mathbf{z}_t = \boldsymbol{\varphi}_0 + \sum_{i=1}^{p} \boldsymbol{\varphi}_i\,\mathbf{z}_{t-i} - \sum_{j=1}^{q} \boldsymbol{\theta}_j\,\boldsymbol{\varepsilon}_{t-j} + \boldsymbol{\varepsilon}_t\,.\tag{2-13}$$

---

* Under the white noise assumption, a more exact test would be against $1.96/\sqrt{T}$ or two standard errors. Precision was unnecessary as the significance test was simply seeking to reduce the number of time series combinations to evaluate.

For VARMA, p and q define model order as in the univariate case, $\boldsymbol{\varphi}_0$ is a $k \times 1$ constant vector, $\boldsymbol{\varphi}_i$ and $\boldsymbol{\theta}_j$ are $k \times k$ constant matrices, and $\{\boldsymbol{\varepsilon}_t\}$ is a zero-mean sequence of i.i.d.[*] random vectors with positive-definite covariance matrix $\boldsymbol{\Sigma}_\varepsilon$. As in the univariate case, the VARMA may be written in compact form using back-shift operator B as

$$\boldsymbol{\varphi}(B)\mathbf{z}_t = \boldsymbol{\varphi}_0 + \boldsymbol{\theta}(B)\,\boldsymbol{\varepsilon}_t \; . \tag{2-14}$$

In this form, $\boldsymbol{\varphi}(B) = \mathbf{I}_k - \boldsymbol{\varphi}_1 B - ... - \boldsymbol{\varphi}_p B^p$ and $\boldsymbol{\theta}(B) = \mathbf{I}_k - \boldsymbol{\theta}_1 B - ... - \boldsymbol{\theta}_q B^q$, both matrix polynomials in B, and $\mathbf{I}_k$ is the $k \times k$ identity matrix.

Thus, multivariate or vector autoregressive integrated moving average (VARIMA) was implemented by first making stationary each member of the (correlated) time-series pair, then applying the *VARMA* function to estimate nine models to find the best one. The best multivariate model was then forecast using the *VARMApred* function, followed by inverse differencing, i.e., discrete integration, to project the forecasts back into the original (un-differenced) time domain.

Confidence intervals for the multivariate time-domain forecast results were sought for comparison with the univariate cases. While the *VARMApred* function produced estimates of standard errors, model forecasts were estimated within the constraints of a "differenced" domain. With the assumption of a Gaussian distribution for the standard errors,[†] it was

---

[*] i.i.d. is defined as independent and identically distributed.

[†] Standard errors in *VARMApred* are based on the assumption of a Gaussian distribution.

ascertained that integration of the squared standard errors, i.e., the variances, produced a valid output reconstruction of the variances in the un-differenced domain [26].[*] Thus, to obtain the reconstituted forecast errors, the predicted errors were inverse-differenced either zero, once, or twice, corresponding to the number of differences applied to make the particular series stationary for VARMA modeling. For a one-to-one comparison with the univariate analysis, the 95% confidence intervals were computed as the square root of the integrated standard-error variances scaled by 1.96×2 (both sides of the interval).

2.5 Multivariate Results

Screening for significance reduced the number of multivariate pairs from 19306 to 45 pairs with correlations above 92% ("Tier 1") and 40 pairs with correlation at 90.6% ("Tier 2").[†] The majority of significant correlations for these 85 pairs occurred at a lag of zero as shown in Figure 2-5. Maximum correlation occurred with a frequency of 4733 at lag zero, with the next most frequent lag occurrence of 85 times at lag 1.

---

[*] An integrator is a linear system, and it is known that a Gaussian input process to a linear system yields a Gaussian output process.

[†] Greater than 92% translates to 59 or more (out of 64) significantly-correlated windows, while 90.6% equals 58 (of 64) significantly-correlated windows.
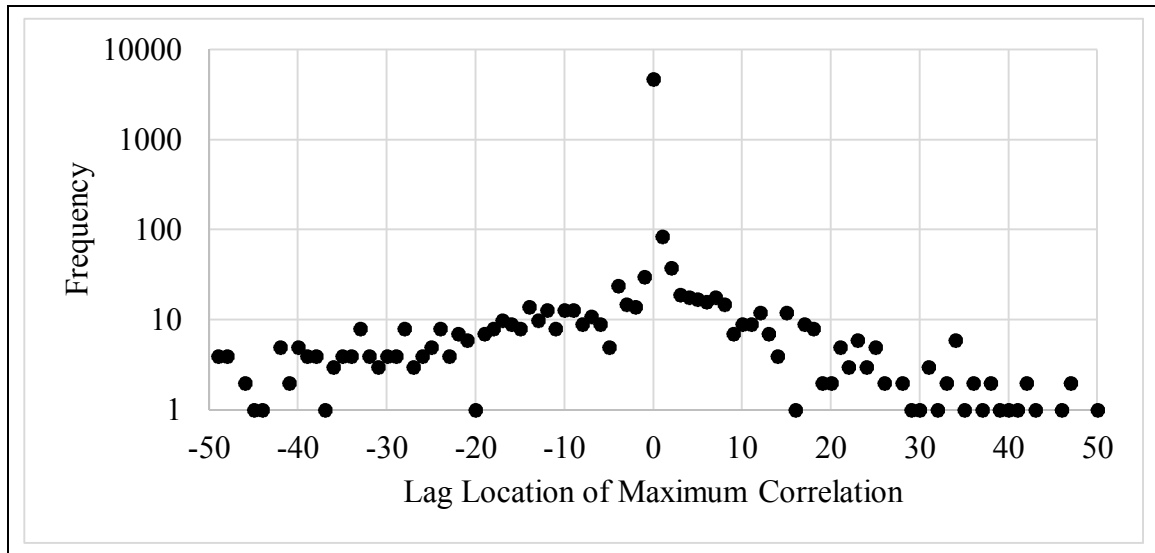
Figure 2-5. Lag Correlation

While the maximum order of the multivariate ARIMA model was restricted to fall in the range of zero to two for p and for q, the bivariate model inherently doubles the number of coefficients due to the matrix configuration of φ_i and θ_j. Thus, a second order model yields possibly four coefficients for either p and/or q. In opposition, the random-walk model with no coefficients for φ_i or θ_j, except possibly the constants φ_0, occurred with a mean frequency of 34.7% with the distribution as shown in Figure 2-6.[*] The implications of the multivariate random-walk models are identical to the univariate cases.

---

[*] All random-walk models were found to be first order: $(p,d,q) = (0,1,0)$.
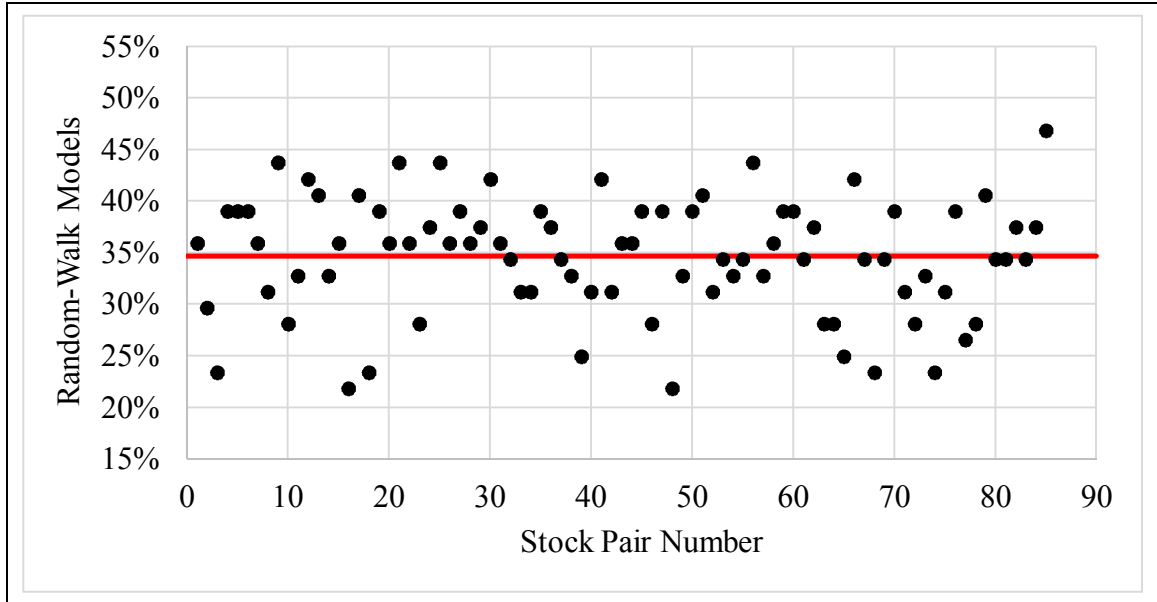
Figure 2-6. Frequency of Bivariate Random-Walk Models

As in the univariate case, a lack of serial correlation for VARIMA model residuals was also sought for representing a well-fit representation. To aid in validation, a multivariate version of the Ljung-Box portmanteau test was evaluated with a modified version of Tsay's *mq* function found in the 'MTS' R-package [25]. The multivariate version of the Ljung-Box statistic is defined as

$$Q_k(m) = T^2 \sum_{t=1}^{m} \frac{1}{T-l} Trace\left(\hat{\mathbf{\Gamma}}_l^{tr} \hat{\mathbf{\Gamma}}_0^{-1} \hat{\mathbf{\Gamma}}_l \hat{\mathbf{\Gamma}}_0^{-1} \right), {}^* \qquad (2\text{-}15)$$

where again T is the sample size, l is the lag, and $\hat{\mathbf{\Gamma}}_0$ and $\hat{\mathbf{\Gamma}}_l$ are defined previously in equations (2-9) and (2-10), respectively. Lag was set at 10 to match the univariate tests.

---

* m is some positive integer, and is set equal to l in the *mq* function.

46

Inasmuch as the Ljung-Box test assumes a Gaussian distribution and performs with lower power for non-Gaussian data, a multivariate version of the spectral-density-based correlation test was applied to support validation of serial independence of model residuals. The statistic proposed in [27] was computed using function *mADCFtest* in the 'dCovTS' R-package. Parameters were set for a Daniell smoothing kernel, bandwidth = 10, and number of bootstrap replicates = 499, as in the univariate analysis. The spectral density test statistic is defined as [27]

$$\bar{T}_n = \sum_{j=1}^{n-1} (n-j) \, k^2 \left(\frac{j}{bw}\right) Trace\left\{ \hat{\mathbf{V}}^*(j) \, \hat{\mathbf{D}}^{-1} \hat{\mathbf{V}}(j) \hat{\mathbf{D}}^{-1}\right\} ,^{\dagger} \qquad (2\text{-}16)$$

where $\hat{\mathbf{D}}^{-1} = diag\left\{ \hat{V}_{11}(0), ..., \hat{V}_{kk}(0)\right\}$, diagonal elements $\left(V_{rr}(j)\right)_{r=1}^{k}$ are the auto-distance covariance function of $\left\{ \mathbf{z}_{t,r}\right\}$, $\left(V_{rm}(0)\right)_{r=1}^{k}$ are off-diagonal auto-distance-correlation-matrix elements that measure the concurrent dependence between $\left\{ \mathbf{z}_{t,r}\right\}$ and $\left\{ \mathbf{z}_{t,m}\right\}$, $k(\cdot)$ is the smoothing kernel function, and *bw* is the bandwidth.

Serial correlation test results are plotted in Figure 2-7 for the bivariate Ljung-Box and spectral density $\bar{T}_n$ statistics. The plots show the fraction of p-values (out of 64 windows) that are below the 0.05 threshold (displayed as a red line). Here again, with the null hypothesis of independence, a small p-value rejects the null hypothesis and thus represents

---

$\dagger$ $\hat{\mathbf{V}}^*(\cdot)$ denotes the complex conjugate of matrix $\hat{\mathbf{V}}(\cdot)$

a strong indication against independence. Hence, as a test based in statistical probabilities, if test results show a small fraction of p-values below 0.05, then independence is affirmed.
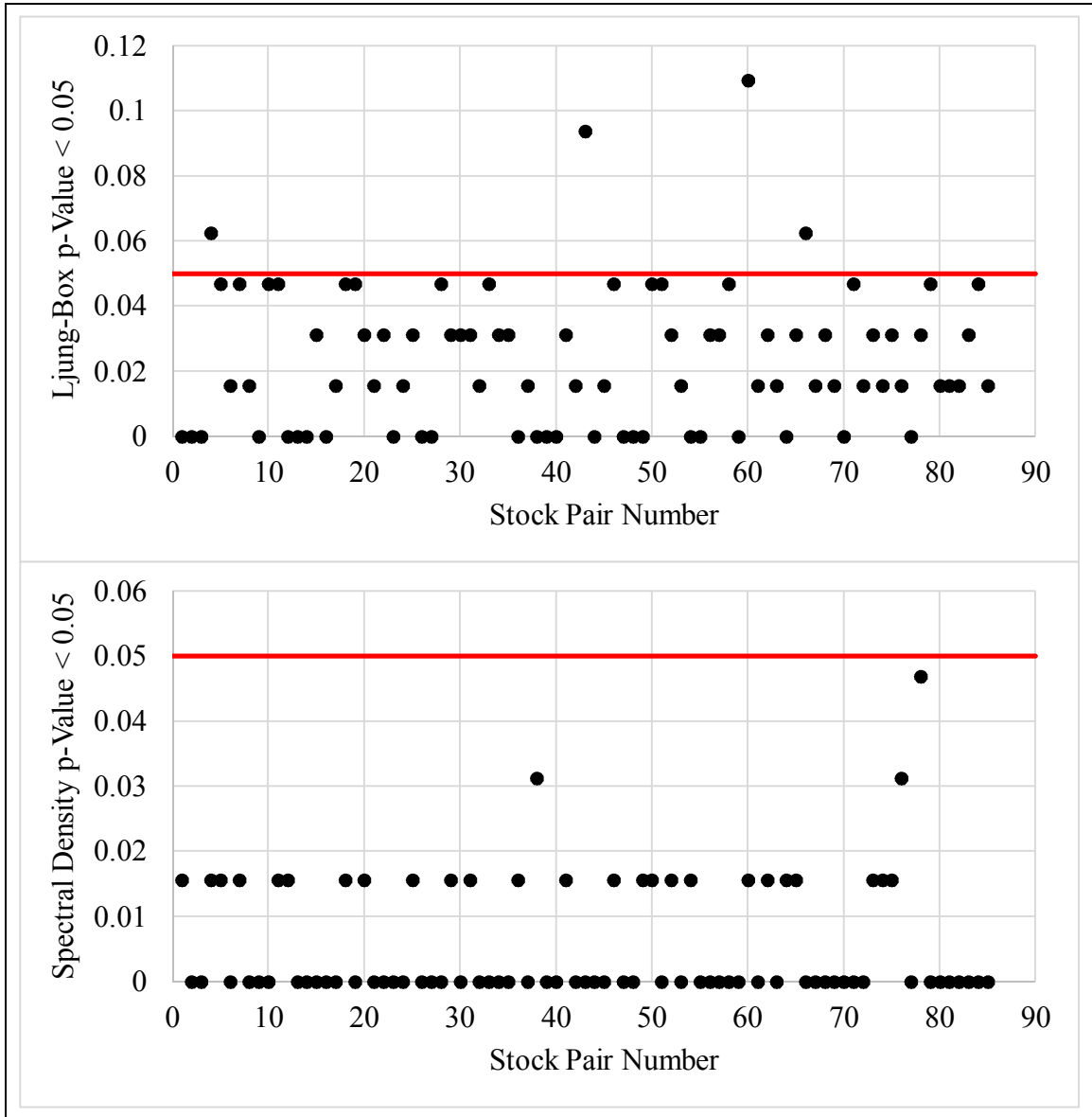
Figure 2-7. Serial Correlation p-Values for Bivariate Model Residuals

The Ljung-Box results rejected four stock pairs while the spectral test did not reject any. Fractions below 0.05 allowed the occurrence of zero to three windows with possible serial

dependence. Values appearing at 0.0625 or above indicate additional windows with possible serial dependence. Any serial dependence suggests that the model fit for that particular window may not be optimal. With no scores above 0.05, the spectral test appeared to show higher power than Ljung-Box. All models were accepted as independent, validating the VARIMA model fits.

Forecast directional accuracy was again investigated with one-step-ahead and five-steps-ahead predictions on out-of-sample data. Figure 2-8 shows the results for each pair of stocks.[*] Implications surrounding the 50% threshold are the same as stated for the univariate analysis. A summary of directional accuracy statistics is given in Table 2-5, identifying the number of occurrences for each accuracy condition. The last line gives the number of incidents where both stocks in the pair forecast the same direction (all correct or all incorrect).

---

[*] S1 refers to Stock 1 and S2 refers to Stock 2 in the stock pairs list provided in the Appendix.
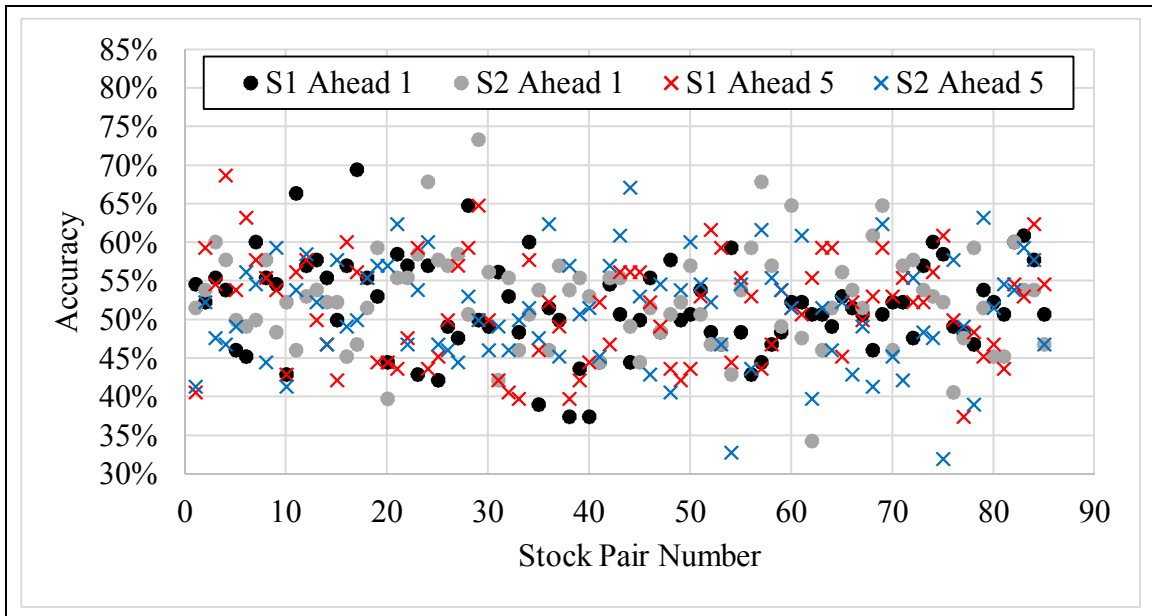
Figure 2-8. Bivariate Directional Forecast Accuracy

Table 2-5. Bivariate Directional Accuracy Summary

Forecast Interval	Accuracy	Stock 1 Frequency (Out of 85)	Stock 2 Frequency (Out of 85)
Ahead 1	> 60%	8	8
	> 55%	25	30
	> 50%	50	57
	< 45%	12	7
	< 40%	3	2
Ahead 5	> 60%	7	10
	> 55%	29	24
	> 50%	48	46
	< 45%	20	14
	< 40%	3	4
Ahead 1 & Ahead 5	Same Direction	45	39
Ahead 1 & Ahead 5	Same Direction	45	

Coverage analysis was also performed for the multivariate stock pairs. Results displayed in Figure 2-9 revealed greater than 79% coverage for all stocks in the pairs list, with the great majority above 85%. Coverage averages for the points displayed in Figure 2-9 are given in Table 2-6.
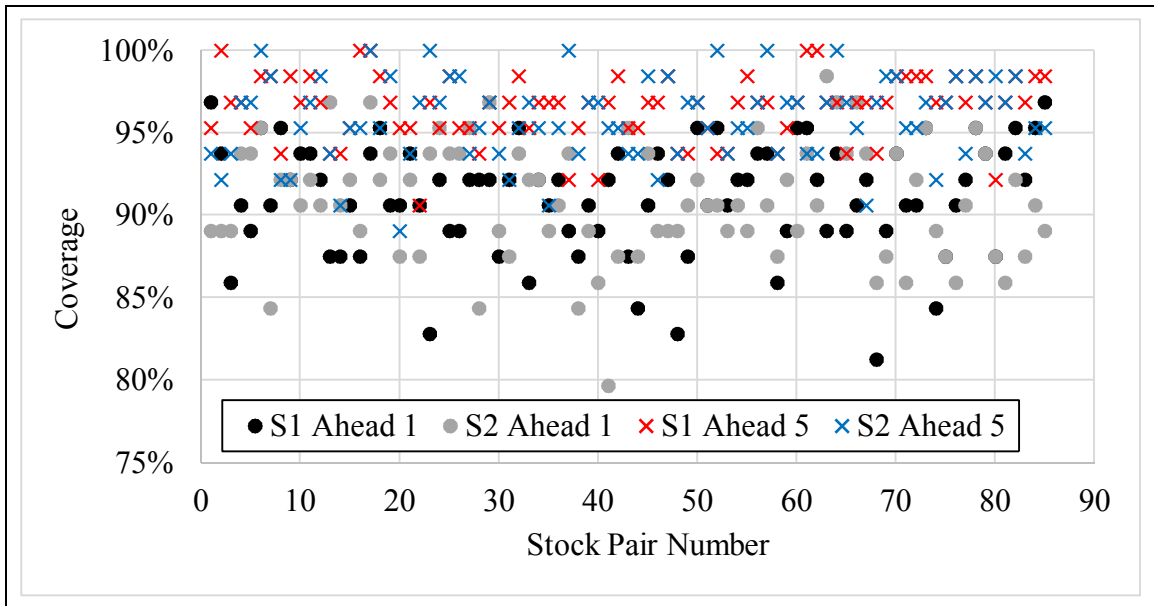
Figure 2-9. Bivariate Forecast Coverage within 95% CI

Table 2-6. Bivariate Mean Coverage

	Ahead 1 (%)	Ahead 5 (%)
Stock 1	91.1	96.5
Stock 2	90.9	95.8

2.6 Summary of Univariate–Bivariate Results

This study investigated the hypothesis that a multivariate analysis would outperform a univariate forecast. A summary of univariate and bivariate forecasting results provided in Table 2-7 supports this hypothesis.[*] The less-frequent occurrence of the random walk in the bivariate models reduced the obviously undesirable incidence of unpredictability. Moreover, directional forecast accuracy was superior for the bivariate model as demonstrated by the greater numbers in the regions farther from the 50% uncertainty midline. In the > 60% region, the bivariate models produced greater than 9% of their one-step forecasts compared to 1% for the univariate models. For the five-step forecasts in the > 60% region, the bivariate models yielded approximately 10% compared to the univariate results of 0.5%. Assuming that an individual who trades stocks seeks higher-probability trades,[†] the trader will presumably be more inclined toward particular stocks near the edges of the directional accuracy range as a technique to reduce risk. As another indicator of forecast quality, the bivariate confidence interval coverage was approximately 1.5% worse for the one-step forecasts but 3.4% better for the five-steps-ahead forecasts.

Additionally, the 95%-confidence-interval widths were compared for univariate and bivariate results. Each member of the 85 stock pairs was evaluated against its matching univariate competitor.[‡] The one-step- and five-steps-ahead bivariate confidence intervals

---

[*] The serial-correlation tests provide only validating support for the models and do not contribute to the quality of the forecasts, and therefore are not included in this table.

[†] A higher-probability trade is one that is more likely to be profitable.

[‡] Within the 85 pairs, there were 107 unique stocks, i.e., some pairs included one member stock that was included in one or more other pairs.

of these 170 stocks (some duplicates) were compared with their corresponding univariate confidence intervals. The objective was to determine if the additional computations required for the multivariate models produced smaller 95%-confidence intervals and thereby enhanced forecasts. Results are displayed in Figure 2-10. The top graph shows the mean frequency of occurrence for shrinkage of the bivariate confidence intervals. For the one-step forecast, all bivariate confidence intervals were smaller at least 78% of the time, with an average frequency of occurrence for shrinkage greater than 87%. For five-step forecast confidence intervals, the frequency of occurrence of smaller bivariate intervals was lower, averaging approximately 74%. It is noteworthy that every pair showed an average reduction of the confidence interval width compared to its univariate counterpart. This is demonstrated by all shrinkage frequency values appearing above the 50% threshold (blue line). The bottom graph in Figure 2-10 shows the amount of shrinkage produced by the bivariate models. Mean Percent Shrinkage (MPS) was defined as

$$MPS_x = \left(\frac{1}{M} \sum_{m=1}^{M} \left(\frac{UCIW_{x,m} - BCIW_{x,m}}{UCIW_{x,m}} \right) \right) \times 100\% \ , \qquad (2\text{-}17)$$

where $UCIW_{x,m}$ is the univariate confidence interval width for stock x in window m, $BCIW_{x,m}$ is the CI width for the corresponding stock in the bivariate pair, and M is the total number of windows ($= 64$). Values appearing in the region above the (blue) threshold line at zero indicate a shrinkage of the multivariate confidence interval, i.e., an improvement over the corresponding univariate model. In contrast, those with values below zero indicate an increase in the multivariate confidence interval. The one-step shrinkage was fairly consistent and averaged 4.2%, but was always positive, i.e., yielding shrinkage. The five-step CI shrinkage was more varied and yielded an average negative

53

shrinkage of 1.7%, resulting in an average slight increase of the multivariate confidence intervals. Table 2-8 summarizes the results. While the one-step forecast showed an improvement, narrowing the CI width in all cases, the five-step CI was more often wider, yielding only 25 and 30 CI reductions for Stock 1 and Stock 2, respectively, out of the 85 pairs. Thus, for short-term (one-step) forecasts, use of the bivariate model appeared advantageous. The bivariate model also showed advantage to the longer term five-step prediction, as well, but more focusing of asset selection was required. Such a narrowing is a natural outcome for identifying a stock to trade, and hence is neither a difficult nor unreasonable extra step.

Table 2-7. Univariate-Bivariate Comparison

Parameter		Univariate (%)	Bivariate (%)	
Random Walk		63.0	34.7	
Ljung-Box p-Value < 0.05		27.9	4.7	
Spectral Density p-Value < 0.05		1.0	0.0	
Directional Forecast Accuracy			**Stock 1**	**Stock 2**
Ahead 1	**> 60%**	1.0	9.4	9.4
	> 55%	12.2	29.4	35.3
	> 50%	47.2	58.8	67.1
	< 45%	4.1	14.1	8.2
	< 40%	0.0	3.5	2.4
Ahead 5	**> 60%**	0.5	8.2	11.8
	> 55%	9.1	34.1	28.2
	> 50%	47.2	56.5	54.1
	< 45%	11.2	23.5	16.5
	< 40%	1.0	3.5	4.7
Confidence Interval Coverage				
Ahead 1		92.5	91.1	90.9
Ahead 5		92.8	96.5	95.8
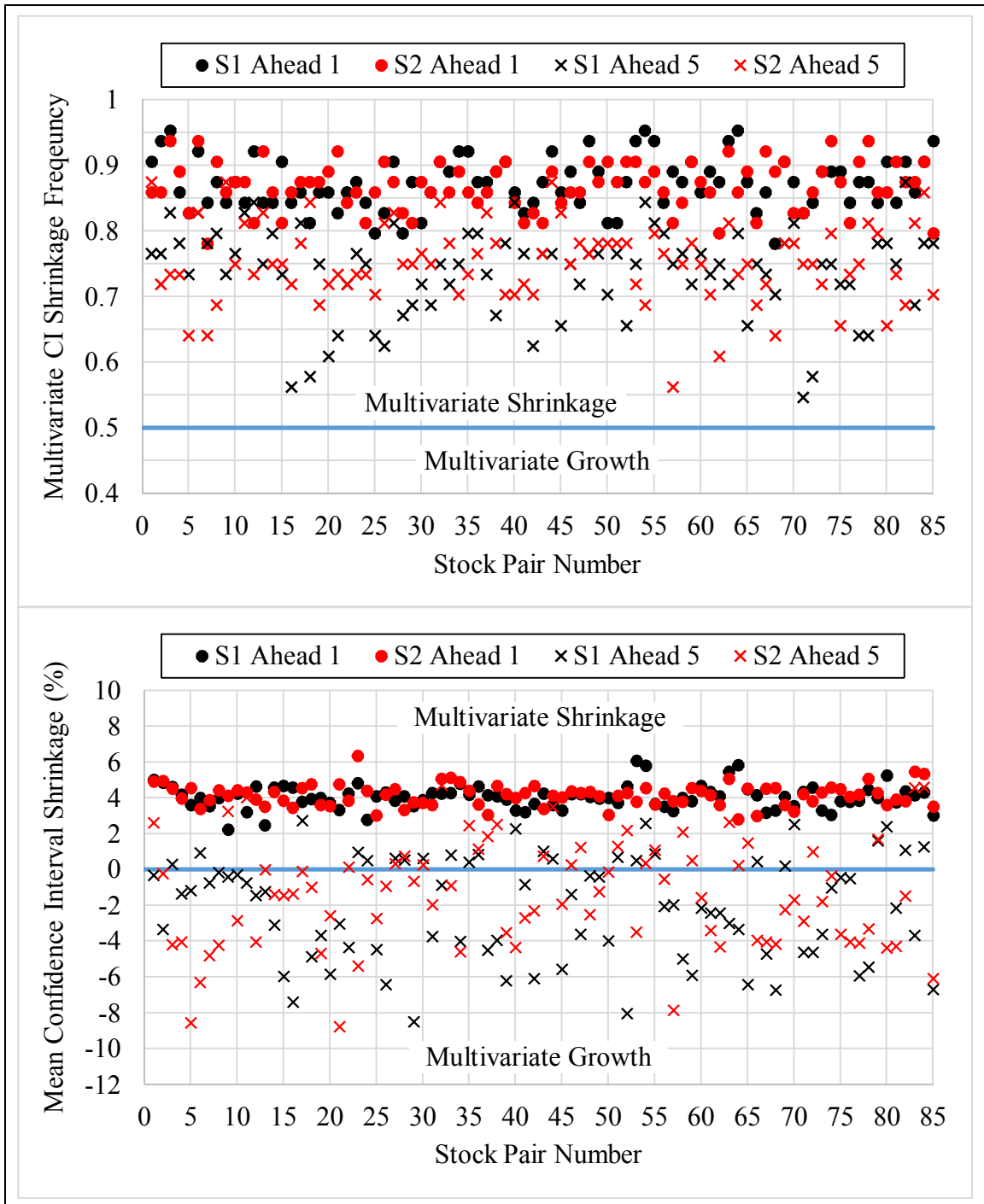
Figure 2-10. Univariate-Bivariate Confidence Intervals

56

Table 2-8. Univariate-Bivariate CI Statistics

Bivariate Shrinkage	Ahead 1 (%)	Ahead 5 (%)
Frequency (Stock 1)	87.4	73.8
Frequency (Stock 2)	86.9	75.1
Mean Shrinkage (Stock 1)	4.1	-2.1
Mean Shrinkage (Stock 2)	4.2	-1.3

Considering all these elements together, it is plausible to conclude that multivariate, i.e., bivariate, forecasting is superior to univariate prediction, within some constraints.

2.7 Extension to Trivariate Sets

With noticeable improvements resulting from application of the bivariate form, a supplementary investigation was conducted to determine whether additional performance enhancement was attainable with trivariate models. A set of stock triplets was built from the list of highly-correlated stock pairs by combining "overlapping" pairs. For example, if one stock pair constituted $\left(S_1, S_2\right)$ and another stock pair comprised $\left(S_1, S_3\right)$, then a triplet was formed as $\left(S_1, S_2, S_3\right)$. Furthermore, if three pairs existed as $\left(S_1, S_2\right)$, $\left(S_1, S_3\right)$, $\left(S_2, S_4\right)$, then all possible "overlapping" triplets were formed: $\left(S_1, S_2, S_3\right)$, $\left(S_1, S_2, S_4\right)$. Results of screening the stock pairs list (Appendix 8.2) yielded a set of 85 stock triplets as given in Appendix 8.3. Analogous tests to those performed on the pairs were adapted to execute on the triplets. Although the models were restricted to

$(0,0) \le (p,q) \le (2,2)$ as previously outlined, the trivariate model substantially increased the number of coefficients for the matrices φ_i and θ_j, yielding a 3×3 matrix for each (up to nine coefficients for each). In contrast, the random-walk model produces zero coefficients with the possibility of a constant φ_0 (3×1 matrix) and resulted in a mean frequency of occurrence of 44.8% with the distribution shown in Figure 2-11.
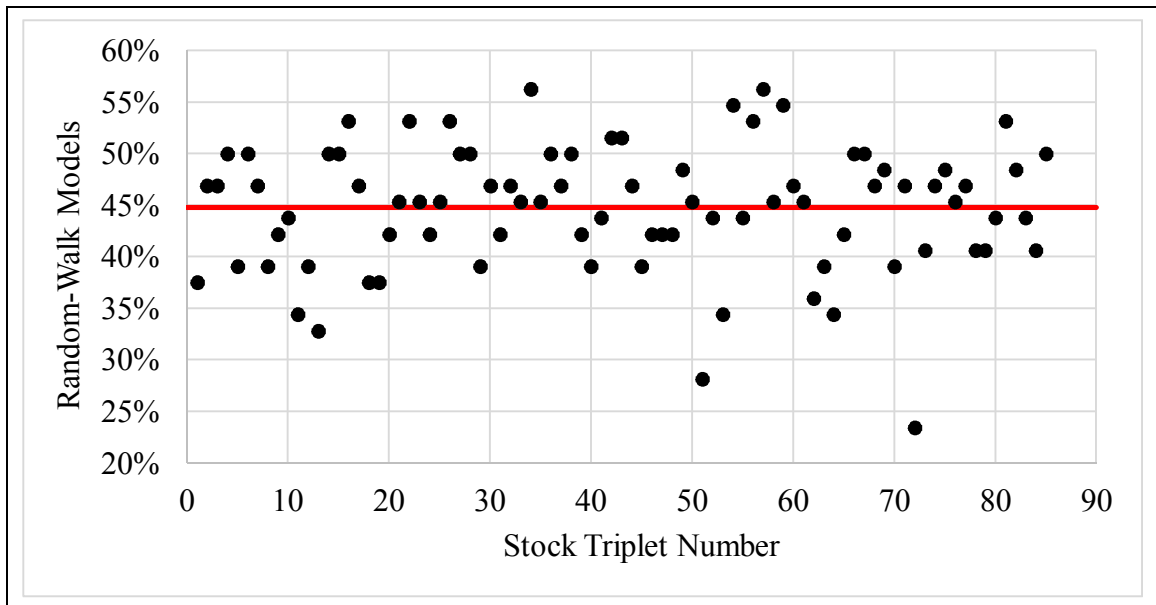
Figure 2-11. Frequency of Trivariate Random-Walk Models

Test results for serial correlation of the trivariate models are shown in Figure 2-12. As in the bivariate case, the graphs show the fraction of p-values (out of 64 windows) that fell below the 0.05 threshold (red line). As previously stated, a small p-value rejects the null hypothesis of independence, so a small fraction of resulting p-values below threshold indicates support for independence. The Ljung-Box test rejected 16 triplet models while the spectral test identified only a single triplet with four p-values above the threshold, rather

than the three allowed to stay below 0.05. Again, the spectral test produced superior results to the Ljung-Box technique, demonstrating its higher power. The trivariate VARIMA models were accepted as independent by this validation process.
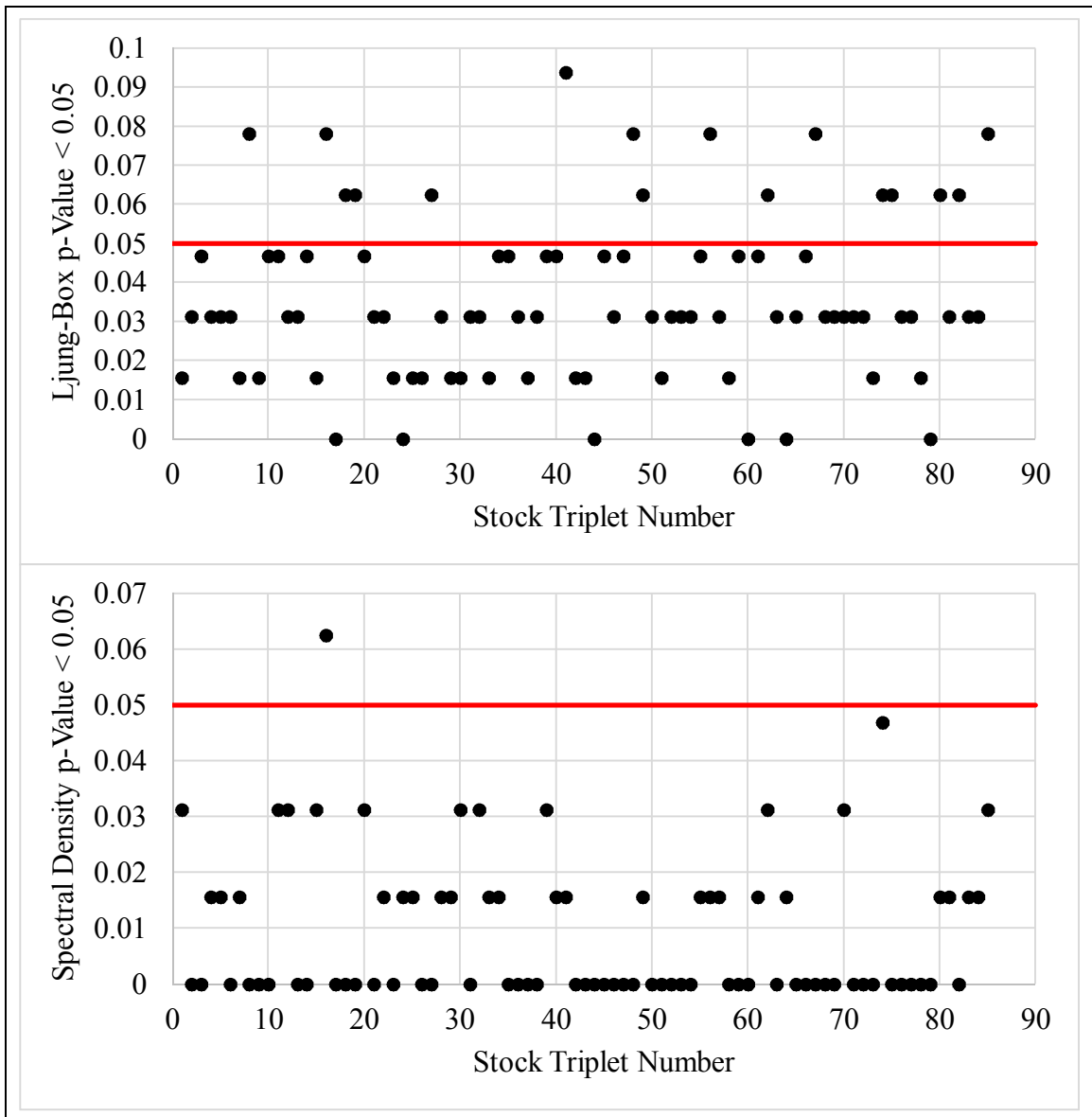
Figure 2-12. Serial Correlation p-Values for Trivariate Model Residuals

Forecast directional accuracy for the trivariate models was investigated for the one-step and five-step predictions for out-of-sample data. Results displayed in Figure 2-13 separate the one-step and five-step accuracies for clarity.[*] As stated previously, the 50% level signifies unpredictability. Triplets that fall farther from 50% indicate more consistency in their statistical forecasting accuracy. Greater than 50% represents correct directional forecasting, whereas less than 50% reflects incorrect prediction. A summary of directional accuracy statistics is provided in Table 2-9. The last line in the table shows the number of occurrences where all three stocks in the triplet forecast the same direction (either all correct or all incorrect).

Coverage analysis results for the trivariate models are shown in Figure 2-14. One-step-ahead coverage showed better than 82% with a mean of 91.5% and the five-steps-ahead coverage was greater than 89% with a mean of 96.5%. Table 2-10 provides a breakdown of the averages.

---

[*] S1, S2, and S3 refer to Stock 1, Stock 2, and Stock 3, respectively, in the stock triplet list provided in the Appendix.
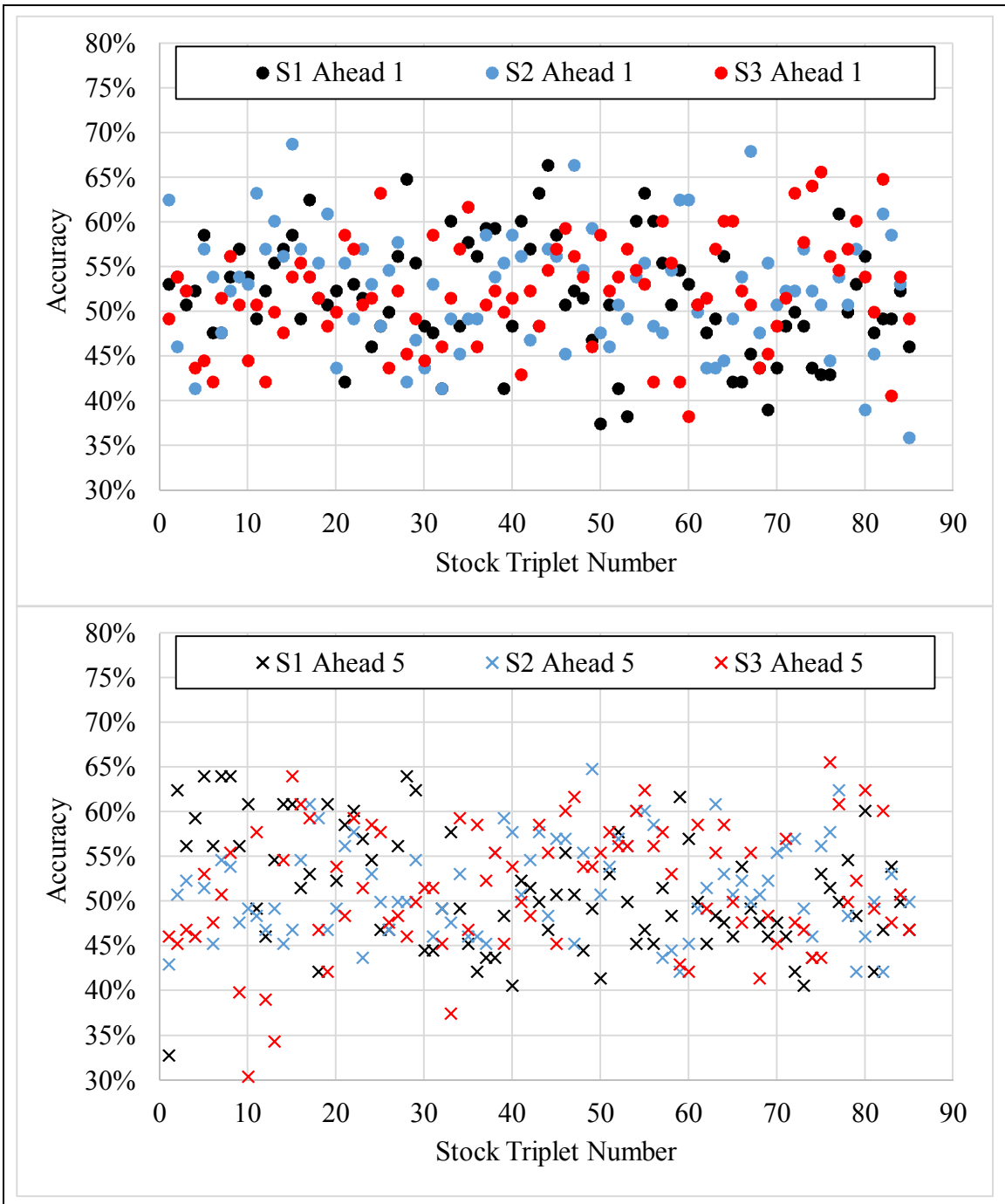
Figure 2-13. Trivariate Directional Forecast Accuracy

Table 2-9. Trivariate Directional Accuracy Summary

Forecast Interval	Accuracy	Stock 1 Frequency (Out of 85)	Stock 2 Frequency (Out of 85)	Stock 3 Frequency (Out of 85)
Ahead 1	> 60%	10	10	10
	> 55%	26	30	26
	> 50%	48	53	56
	< 45%	14	11	13
	< 40%	3	2	1
Ahead 5	> 60%	13	6	10
	> 55%	24	24	32
	> 50%	40	46	46
	< 45%	14	7	11
	< 40%	1	0	5
Ahead 1 & Ahead 5	Same Direction	52	35	54
Ahead 1 & Ahead 5	Same Direction	13		
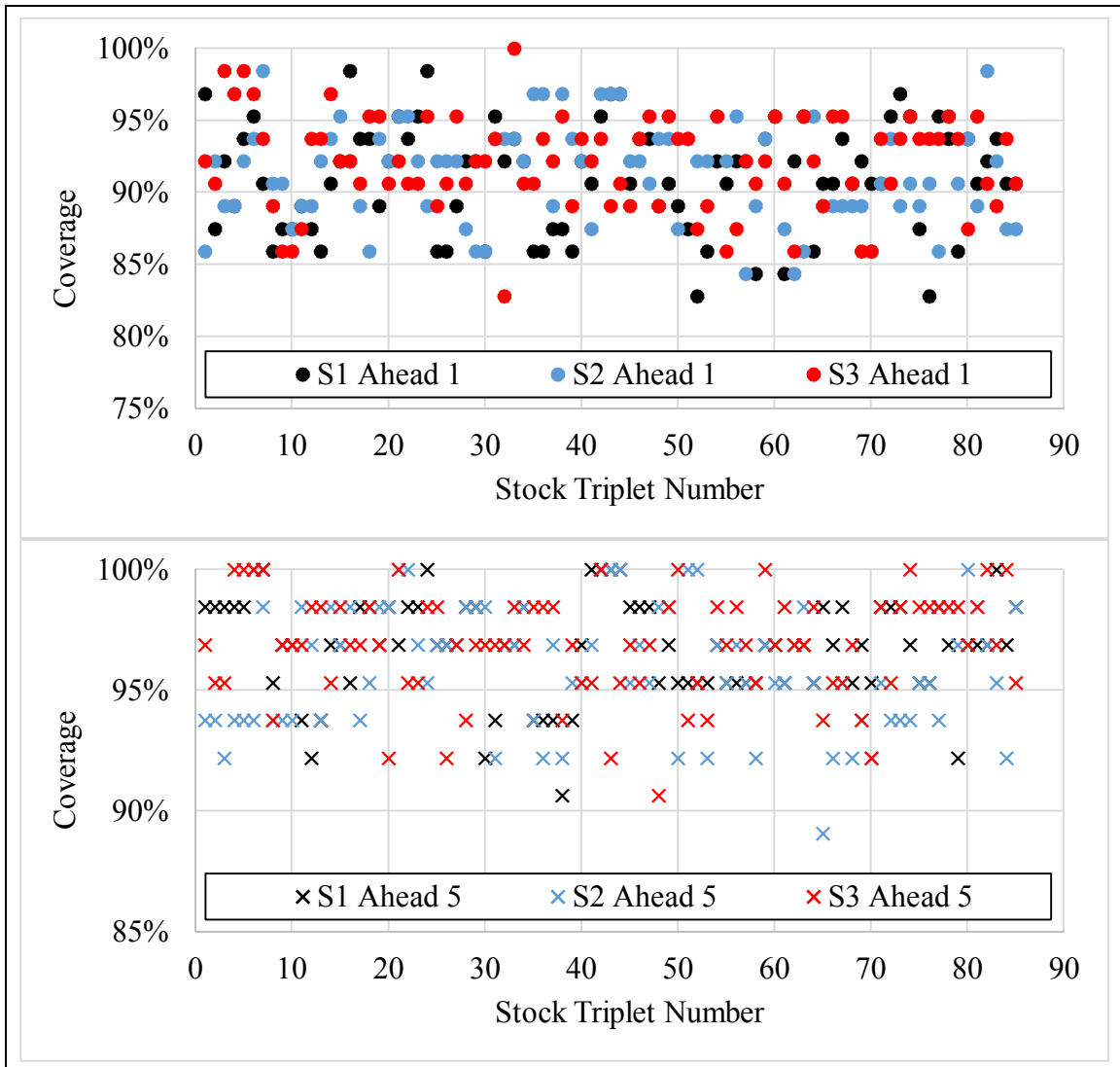
Figure 2-14. Trivariate Forecast Coverage within 95% CI

Table 2-10. Trivariate Mean Coverage

	Ahead 1 (%)	Ahead 5 (%)
Stock 1	91.0	96.8
Stock 2	91.4	95.9
Stock 3	92.1	96.9

2.8 Summary of Overall Multivariate Results

Investigating the hypothesis of improved time-series forecasting performance through multivariate modeling yielded a mixture of results. Table 2-11 provides a summary of results for univariate, bivariate, and trivariate forecast outcomes covering approximately twenty-five years of daily stock market data. Through application of traditional time series analysis techniques, it was evident that multivariate analysis outperformed a univariate approach. Furthermore, the bivariate form appeared to yield the best overall performance, even superior to the more complex trivariate model. A possible explanation is a lower correlation coefficient attributed to the third stock in the triplet. Inasmuch as triplets were built from highly-correlated stock pairs such that two of three pair combinations within the triplet had high correlation, it was most probable that the third doublet had a correlation coefficient less than 90%. A diminished interrelationship of the third doublet would inject a confounding factor into the triplet model, reducing the model's effectiveness. It was noteworthy that the trivariate models yielded a significant increase in the frequency of occurrence of random-walk models.

The distribution of confidence interval shrinkage is shown in Figure 2-15. The 50% threshold seen as a yellow line in the figure separates multivariate confidence-interval shrinkage from growth. Clearly, as seen in the top plot, the trivariate confidence interval widths showed a mean shrinkage for all stocks in the triplets as compared with their univariate match. The bottom graph in Figure 2-15 shows the amount of shrinkage. These results were similar to those seen in the bivariate case described earlier (§2.6). Numerical

results are summarized in Table 2-12. Again, statistics of the bivariate model showed a superior performance to the trivariate form in all categories.

Table 2-11. Univariate-Multivariate Comparison

Parameter		Univariate (%)	Bivariate (%)		Trivariate (%)		
Random Walk		63.0	34.7		44.8		
Ljung-Box p-Value < 0.05		27.9	4.7		18.8		
Spectral Density p-Value < 0.05		1.0	0.0		1.2		
Directional Forecast Accuracy			**Stock 1**	**Stock 2**	**Stock 1**	**Stock 2**	**Stock 3**
Ahead 1	**> 60%**	1.0	9.4	9.4	11.8	11.8	11.8
	> 55%	12.2	29.4	35.3	30.6	35.3	30.6
	> 50%	47.2	58.8	67.1	56.5	62.4	65.9
	< 45%	4.1	14.1	8.2	16.5	12.9	15.3
	< 40%	0.0	3.5	2.4	3.5	2.4	1.2
Ahead 5	**> 60%**	0.5	8.2	11.8	15.3	7.1	11.8
	> 55%	9.1	34.1	28.2	28.2	28.2	37.6
	> 50%	47.2	56.5	54.1	47.1	54.1	54.1
	< 45%	11.2	23.5	16.5	16.5	8.2	12.9
	< 40%	1.0	3.5	4.7	1.2	0.0	5.9
Confidence Interval Coverage							
Ahead 1		92.5	91.1	90.9	91.0	91.4	92.1
Ahead 5		92.8	96.5	95.8	96.8	95.9	96.9
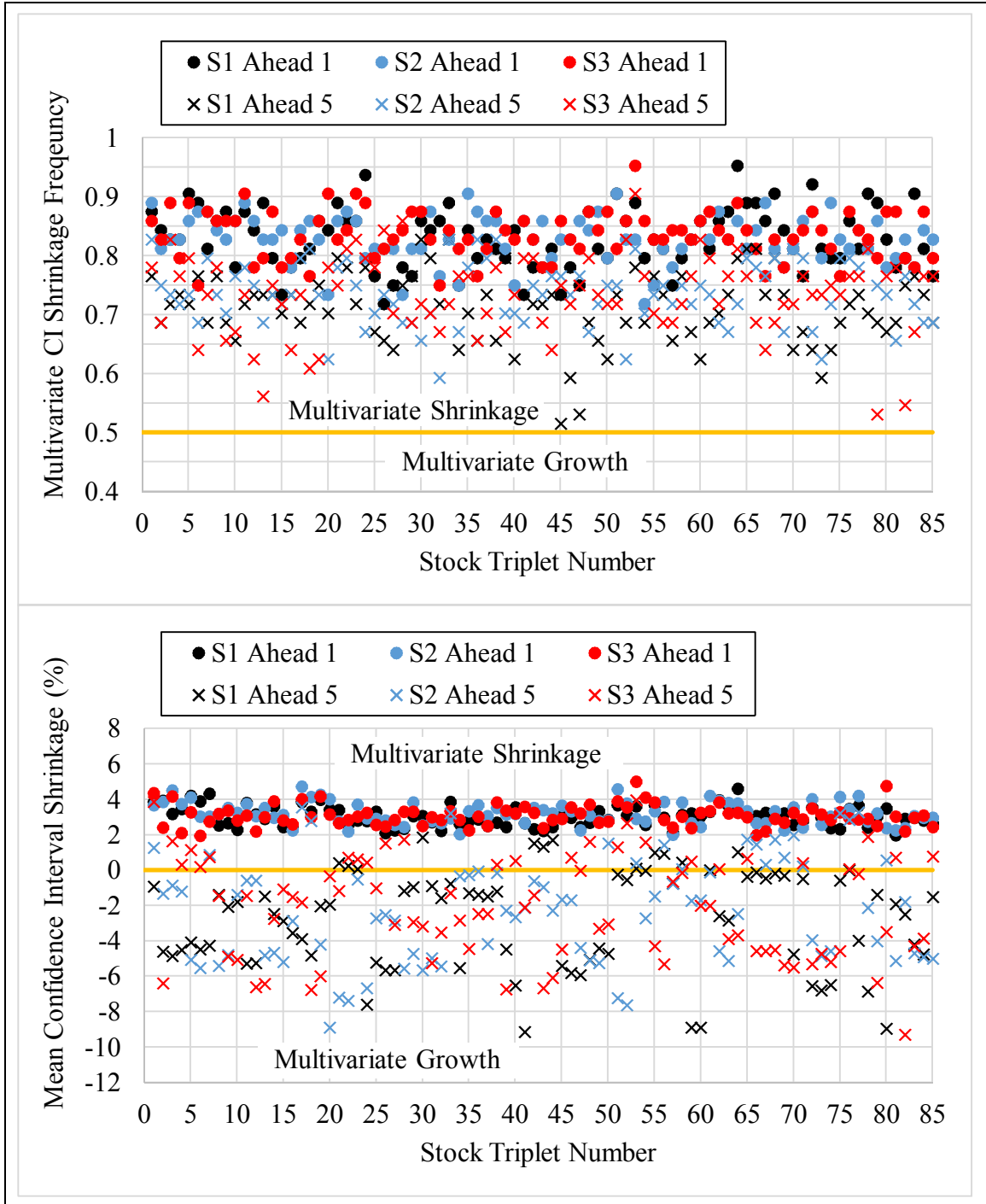
Figure 2-15. Univariate-Trivariate Confidence Intervals

Table 2-12. Bivariate-Trivariate Confidence-Interval Statistics

Multivariate Shrinkage	Ahead 1 (%)		Ahead 5 (%)	
	Bivariate	Trivariate	Bivariate	Trivariate
Frequency (Stock 1)	87.4	83.0	73.8	70.8
Frequency (Stock 2)	86.9	83.0	75.1	73.6
Frequency (Stock 3)		83.6		73.3
Mean Shrinkage (Stock 1)	4.1	3.1	-2.1	-2.8
Mean Shrinkage (Stock 2)	4.2	3.2	-1.3	-2.2
Mean Shrinkage (Stock 3)		3.1		-2.1

3. FORECASTING WITH SINGULAR SPECTRUM ANALYSIS

"It is far better to foresee even without certainty than not to foresee at all."

– Henri Poincare (1854-1912), in the book "The Foundations of Science," 1913

This chapter expounds on a method known as singular spectrum analysis (SSA) to investigate time-series modeling and forecasting. The recent introduction of the SSA approach into modern analytical methods is commonly identified with the 1986 articles of Broomhead and King [28] [29]. Subsequently, significant additional supporting theory and operational details have been developed and documented by Golyandina, et al., in two books [30] [31].* Inasmuch as SSA is a relatively new development, it is not widely known, particularly among statisticians and economists.† Although SSA is completely unrelated to ARIMA, GARCH,‡ and wavelet methods, it is associated with principal component analysis (PCA) and empirical mode decomposition (EMD).

SSA is a nonparametric technique that blends aspects of signal processing, linear algebra, nonlinear dynamical systems, finite-difference equations, and multivariate geometry. The usual objective of SSA is to decompose a time series into a collection of interpretable components consisting of a (possibly slowly-varying) trend, oscillatory elements, and

---

* See also their included references.

† However, it "has become a standard tool in meteorology and climatology; it is also a well-known technique in nonlinear physics and signal processing." [30]

‡ Generalized Autoregressive Conditional Heteroskedasticity

(unstructured) noise. In application to forecasting, the goal is simplified, requiring only identification and extraction of the composite signal model from a time series assumed to consist of signal plus noise. Exposing the underlying composition of the signal structure is unnecessary as it provides no additional benefit to automated forecasting.

In this chapter, parallel to analyses reported in Chapter 2, stock data described in Chapter 1 are forecast and analyzed to determine accuracies and 95% confidence intervals, within the framework of SSA. Performance of univariate and multivariate time series are compared. The study herein evaluates whether a multivariate time series of correlated stocks surpasses a univariate-based forecast and assesses performance results of SSA to ARIMA modeling. Similar to the classical ARIMA investigation, there is an underlying assumption that the time series can be modeled with an identifiable structure, although for SSA it is a nonparametric form.

3.1 Univariate SSA Methodology

In the SSA modeling process, a time series is subjected to a decomposition stage followed by a reconstruction stage. The decomposition stage (A) consists of two steps: 1) embedding, and 2) singular value decomposition (SVD). The reconstruction stage (B) also comprises two steps: 1) grouping, and 2) diagonal averaging. These are explained as follows.

A.1 Decomposition – Embedding

The process of embedding converts a one-dimensional time series $X_N = (x_1, x_2, \ldots, x_N)$ into a multidimensional series. With a series length $N > 2$, model parameters include integers L called the window length (with $1 < L < N$), and $K = N - L + 1$. Embedding is performed by mapping the sequence X_N into a set of K lagged vectors, each of length L as $X_i = (x_i, x_{i+1}, \ldots, x_{i+L-1})^{\mathrm{T}}$ with $1 \le i \le K$.* The trajectory matrix of the original series X_N is the multidimensional series:

$$\mathbf{X} = [X_1 : X_2 : \ldots : X_K] = (x_{ij})_{i,j=1}^{L,K} = \begin{pmatrix} x_1 & x_2 & x_3 & \cdots & x_K \\ x_2 & x_3 & x_4 & \cdots & \\ x_3 & x_4 & x_5 & \cdots & \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_L & x_{L+1} & x_{L+2} & \cdots & x_N \end{pmatrix}. \qquad (3\text{-}1)$$

Clearly, the X_i vectors are columns of $\mathbf{X}$, and both columns and rows of $\mathbf{X}$ are subseries of X_N. Note also that $\mathbf{X}$ is a Hankel matrix and as such has equal elements along the antidiagonals $i + j = \text{constant}$.

A.2 Decomposition – Singular Value Decomposition (SVD)

SVD is applied to the trajectory matrix $\mathbf{X}$. Defining $\mathbf{S} = \mathbf{X}\mathbf{X}^{\mathrm{T}}$, eigenvalues of $\mathbf{S}$ denoted as $\lambda_1, \lambda_2, \ldots, \lambda_L$, with eigenvalues in decreasing order of magnitude $(\lambda_1 \ge \lambda_2 \ge \ldots \ge \lambda_L)$, and $U_1, U_2, \ldots, U_L$ as the orthonormal system of corresponding eigenvectors, setting

---

* Here, A^{T} is the transpose of vector A.

$d = \operatorname{rank} X = \max(i, \text{ such that } \lambda_i > 0)$ and $V_i = \mathbf{X}^{\mathrm{T}} U_i / \sqrt{\lambda_i}$, the SVD of the trajectory

matrix may be written as

$$\mathbf{X} = \mathbf{Y}_1 + \mathbf{Y}_2 + \cdots + \mathbf{Y}_d \; , \tag{3-2}$$

where $\mathbf{Y}_i = \sqrt{\lambda_i} \, U_i \, V_i^{\mathrm{T}}$. The $\mathbf{Y}_i$ matrices have rank one and are called elementary matrices,

U_i and V_i are the left and right eigenvectors of the trajectory matrix and in SSA literature

are designated empirical orthogonal functions (EOFs) and principal components,

respectively. The ensemble $\left(\sqrt{\lambda_i}, U_i, V_i \right)$ is labeled the i-th eigentriple of $\mathbf{X}$, $\sqrt{\lambda_i}$

$(i = 1, 2, \ldots, d)$ are the singular values of $\mathbf{X}$, and the set $\left\{ \sqrt{\lambda_i} \right\}$ is called the spectrum of

matrix $\mathbf{X}$.

B.1 Reconstruction – Grouping

Grouping refers to collecting eigentriples into sets. The procedure corresponds to

partitioning the index set $\{1, 2, \ldots, d\}$ into m disjoint subsets $I_1, I_2, \ldots, I_m$. With a group

of indices $I = \left\{ i_1, i_2, \ldots, i_p \right\}$, the corresponding matrix $\mathbf{X}_I = \mathbf{X}_{i_1} + \mathbf{X}_{i_2} + \ldots + \mathbf{X}_{i_p}$ defines

the group partition. Combining the SVD expansion (3-2) and resultant matrices for groups

$I_1, I_2, \ldots, I_m$ yields the total decomposition

$$\mathbf{X} = \mathbf{X}_{I_1} + \mathbf{X}_{I_2} + \ldots + \mathbf{X}_{I_m} \; , \tag{3-3}$$

where $\mathbf{X}_{I_j} = \sum_{\upsilon \in I_j} \mathbf{Y}_\upsilon, \; j = 1, 2, \ldots, m$.

In standard SSA, the objective of this grouping step is to separate the original time series into independent sets, preferably orthogonal, so that the underlying trend, (possibly multiple) oscillation formations, and noise can be separated and identified. In forecasting applications, the objective is to distinguish the composite signal structure (trend and oscillations, collectively) from the noise-ladened signal.

B.2 Reconstruction – Diagonal Averaging

Diagonal averaging converts each $.\mathbf{X}_{I_j}.$ matrix from (3-3) into a vector , i.e., a time series, of length N. For an $L \times K$ matrix $\mathbf{Z}$ with elements z_{ij}, [*] the k-th term of the transformed vector $z = (z_1, z_2, \ldots, z_N)$, is produced by averaging z_{ij} over all i, j such that $i + j = k + 1$, i.e., along the antidiagonals. To illustrate, for $k = 1$, $z_1 = z_{1,1}$, and with $k = 2$, $z_1 = (1/2)(z_{1,2} + z_{2,1})$, and so forth. Diagonal averaging matrix $\mathbf{X}_{I_j}$ yields a reconstructed series

$$\mathbf{X} = \tilde{\mathbf{X}}_{I_1} + \tilde{\mathbf{X}}_{I_2} + \ldots + \tilde{\mathbf{X}}_{I_m} \ , \tag{3-4}$$

which is equivalent to the original series $X_N = (x_1, x_2, \ldots, x_N)$ decomposed into a sum of m series

$$x_n = \sum_{j=1}^{m} \tilde{x}_n^{(j)} \qquad (n = 1, 2, \ldots, N) \ , \tag{3-5}$$

---

[*] This corresponds to $1 \le i \le L$, $1 \le j \le K$, and $N = L + K - 1$.

72

where $\tilde{X}_N^{(j)} = \left(\tilde{x}_1^{(j)}, \tilde{x}_2^{(j)}, \ldots, \tilde{x}_N^{(j)} \right)$ corresponds to $\tilde{\mathbf{X}}_{I_j}$. Conceptually, the procedure of diagonal averaging in SSA may be regarded as an inverse of the embedding process, but operating on a modified data set that approximates the original input.

3.1.1 Univariate SSA Forecasting

Obtaining a plausible forecast with singular spectrum analysis assumes the following:[*]

1. The series possesses a structure.

2. The structure can be identified.

3. The structure leads to a method of time series continuation.

4. The identified structure holds through the (future) forecast interval.

SSA forecasting can be divided into two types: 1) recurrent, and 2) vector. Computations performed in this study focused exclusively on recurrent forecasting since results of preliminary investigations showed more stable, i.e., reliable, outcomes. Consequently, only recurrent forecasting is described below.

Recurrent Forecasting (R-SSA) [32]

Recurrent forecasting within the SSA framework is considered the primary algorithm [33]. With $j = 1, 2, \ldots, L$, defining $u_{i,j}$ as the i-th entry in eigenvector U_j of the trajectory matrix, and π_j as the last element of U_j, the forecast values $\left(\hat{x}_{N+1}, \hat{x}_{N+2}, \ldots, \hat{x}_{N+M} \right)$ are computed as

---

[*] These assumptions are not unlike assumptions for ARIMA-based forecasting.

$$\hat{x}_i = \begin{cases} \tilde{x}_i\ , & i = 1, 2, \ldots, N \\ \displaystyle\sum_{k=1}^{L-1} a_k\ \hat{x}_{i-k}\ , & i = N+1, N+2, \ldots, N+M \end{cases}, \tag{3-6}$$

where $\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_N$ are the reconstructed series,

$$a_i = \sum_{j=1}^{r} \left(\frac{\pi_j\, u_{L-1,j}}{1 - \left(\pi_1{}^2 + \pi_2{}^2 + \ldots + \pi_r{}^2\right)} \right), \qquad i = 1, 2, \ldots, L-1\ . \tag{3-7}$$

Also, r is the dimension of the linear space $\mathcal{L}_r$ spanned by the columns of the trajectory

matrix, and $r < L$.

3.1.2   SSA Implementation

Time series analyses via the SSA method are performed in the R language utilizing several

routines from the R-package 'Rssa' version 0.13-1. As in the ARIMA modeling, input

data for each stock is identically preprocessed, dividing it into the same 100-sample

windows[*] and split-adjusting as described in §2.1. Preliminary studies of the stock data

set revealed that sequential SSA is useful and necessary to produce any reasonable output

results. Sequential SSA attempts to overcome the complication of weakly-separable series

components, i.e., those with closely-spaced singular values, by first performing the SSA

---

[*] Hereupon, the 100-sample-input-data windows will be called data frames where
     necessary as a distinction from the SSA length-L window.

process on the input data with window length L_1, followed by a second application of SSA with window length L_2 on the residuals[*] from the first pass.[†]

The *ssa* function (from 'Rssa') was used to generate an SSA model for an input time series. Since window length, L, is a required input parameter and profoundly impacts the returned model, an automated approach was developed to select a suitable window length for each data frame. The essence of the approach assumed that the largest spectral peak in the waveform represented the most significant signal on which to focus. The spectral density of the data frame was first estimated with the periodogram function *spec.pgram*, provided in the base-R 'stats' package. Subsequently, all spectral peaks were found with the *fpeaks* function from the R-package 'seewave' version 2.0.5 and the peak with the largest magnitude was identified. The inverse of this peak's frequency, i.e., its period, after rounding, was selected as the window length L. A constraint on minimum window length was set to five because experimental observations showed that smaller limits produced poor forecast results.[‡]

The reconstruction process was carried out with the 'Rssa' *reconstruct* function, which requires a group (of eigenvectors) to be specified as an input parameter. Hence, a process

---

[*] Residuals are the remains after extracting the resulting signal model, as in ARIMA modeling.

[†] There is nothing preventing additional passes, but little improvement was found with more than two.

[‡] Moreover, with window lengths less than five, occasional results yielded unstable, i.e., unbounded, forecasts.

for automatic group selection was also developed. A measure of deviation of two series $F^{(1)}$ and $F^{(2)}$ is the weighted correlation (w-cor) defined as [30]

$$\rho_{12}^{(w)}\left(F^{(1)},F^{(2)}\right)=\frac{\left(F^{(1)},F^{(2)}\right)}{\left\|F^{(1)}\right\|_w\left\|F^{(2)}\right\|_w}\,, \tag{3-8}$$

where $\left\|F^{(i)}\right\|_w=\sqrt{\left(F^{(i)},F^{(i)}\right)_w}$, for $i=1,2$. A small absolute value of correlation infers that the two series are close to orthogonal and thus separable; conversely, with a large absolute value, separability is poor, if not impossible. To identify desirable groupings, the 'Rssa' function *wcor* was applied to the SSA model to compute a correlation matrix of all eigenvectors in the model.[*] Matrix elements are sorted by size in each column[†] and a threshold of 0.2 was applied to find eigenvectors with low correlation for addition into a reconstruction group. Furthermore, a test threshold of greater than 0.9 was imposed on adjacent eigenvectors to find possible harmonic pairs for inclusion into the group.[‡] Under nonharmonic conditions an eigenvector with any high correlation value is rejected from inclusion into the group.

Recurrent forecasting was performed using the 'Rssa' function *rforecast*. This function determined the mean forecast values only. Confidence intervals surrounding these mean values were computed with a modified version of the 'Rssa' function *bforecast* which

---

[*] The number of available eigenvectors in a model equals the window length L.

[†] The correlation matrix is symmetric, so sorting rows or columns is equivalent.

[‡] Harmonic pairs display a periodic characteristic when plotted against each other.

estimates confidence intervals using a bootstrap☼ technique. In *bforecast*, confidence intervals were estimated by taking multiple random permutations of model residuals (assumed to be uncorrelated, noise-like) finding the specified sample quantiles. After discovering that the residuals exhibited substantial autocorrelation, the *bforecast* function was modified to accept block resampling. Block sampling is a method that was introduced in 1989 to retain the correlation formation in bootstrapped time series by resampling blocks of consecutive data units, rather than single data points [34]. In support of this concept, block size was computed equal to the so-called "variance inflation factor," *VIF*, because this value is sometimes cited as the "time between effectively independent samples," or the "decorrelation time" [34] [35]:

$$\text{Block Size} = VIF = 1 + 2 \sum_{n=1}^{N-1} \left(1 - \frac{n}{N}\right) r_n \ , \qquad (3\text{-}9)$$

where N is the sample size, and r_n is the autocorrelation estimate at lag n.

With the methodology described above, for each of two sequential SSA passes, means and confidence intervals were computed for each forecast step from one to $h_{\max}$, the maximum forecast step. Composite results were obtained by summing mean outputs and root-sum-squaring (RSS) upper and lower CI values independently, according to

$$\sigma_x(h) = \sqrt{\sigma_{x_1}^2(h) + \sigma_{x_2}^2(h)} \ , \qquad (3\text{-}10)$$

☼ Bootstrapping is a computationally-intensive technique used to estimate statistics of a distribution by repeatedly resampling, with replacement, applicable data to produce numerous synthetic, equivalently-sized sample sets [34].

where $\sigma_x(h)$ represents either the upper or lower confidence interval offset[*] from the mean at forecast step h, and x_1 and x_2 refer to the first and second pass results, respectively. Interim results appeared implausible as typical returns from equation (3-10) yielded fairly consistent confidence intervals with increasing forecast horizon.[†] To realize a sensible CI offset with increasing forecast horizon h, a cumulative sum of squares was applied as

$$CI(h) = \sigma_{x_{Cum}}(h) = \sqrt{\sum_{i=1}^{h} \sigma_x^2(i)} \ . \tag{3-11}$$

Finally, initial experiments showed that 'Rssa'-based forecast results appeared to continue from the end point of the reconstruction approximation, rather than from the original time series.[‡] This outcome can contribute significant errors to the time series forecast when disparate endpoints exist between the two series. To alleviate this predicament, the offset of the forecast with respect to the reconstructed series was determined and an adjustment was applied to shift the values to present the same offset relative to the original series.

---

[*] Offset is in absolute value. Lower CI values are subtracted from the mean.

[†] The statistically-based supposition is that due to an increase in uncertainty, CI width increases as forecast horizon increases.

[‡] Furthermore, observations revealed that close alignment of series endpoints between the reconstruction and original series was window- and model-dependent and random.

3.2 Univariate SSA Results

Applying singular spectrum analysis to the stock data typically yielded multi-group models. At times, the group selection process (described in §3.1.2) yielded only one group for each sequential-SSA pass.   Frequency of occurrence for these "simple" one-group models was collected and plotted for comparison with ARIMA random-walk models. As seen in Figure 3-1, the frequency of occurrence of simple models was quite low, averaging 1.9% (indicated by a red line), as compared to the ARIMA random-walk mean of 63% (Figure 2-1). A one-group model only indicates a lack of separability in the eigenvector space. While the random-walk model forms unpredictable forecasts, a one-group SSA model, although simple, nevertheless produces viable forecasts.
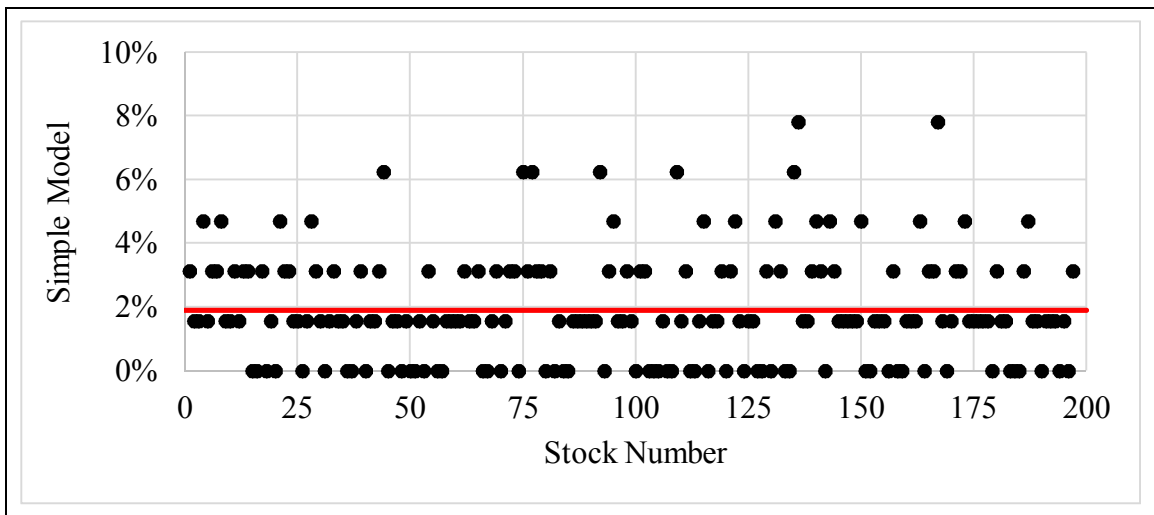
Figure 3-1. Frequency of Univariate SSA Simple Models

Unlike the ARIMA models, there was noticeable correlation remaining in the second-pass-SSA residuals as determined by the Ljung-Box portmanteau test.[*] The mean rejection rate was 97.3%.[†] Working under the assumption that correlation represented more information to be gleaned from the residuals, two additional experiments were conducted to try to realize uncorrelated residuals. First, for each window of each stock, the second-pass-SSA residuals were subjected to fitting by a standard ARIMA model. This yielded a substantially-reduced Ljung-Box rejection rate of 10.5%, but included a reduction in forecast accuracy. As a second experiment, a third pass of sequential SSA was performed. Results showed a small decrease of Ljung-Box rejection to 88.7% and a small loss in forecast accuracy. With no improvement in forecast accuracy, neither of these third processing steps was considered advantageous.

As in Chapter 2, accuracy of forecasting directional movement was evaluated by one-step and five-step predictions of out-of-sample data for each window of each stock. Figure 3-2 shows a plot of the results for each of 197 stocks.[‡] Here again, any results above (or below) 50% (blue line) indicates an improvement over a random guess. Examination of the numerical summary given in Table 3-1 reveals that univariate SSA analyses showed a marked improvement over ARIMA models for positive directional forecasts. In contrast,

---

[*] Although, visually the residuals did appear randomly distributed.

[†] Consequently, it did not make sense to execute the time-consuming spectral-density-based test for correlation due to the extremely high Ljung-Box rejection rate.

[‡] See Appendix 8.1 for a listing.
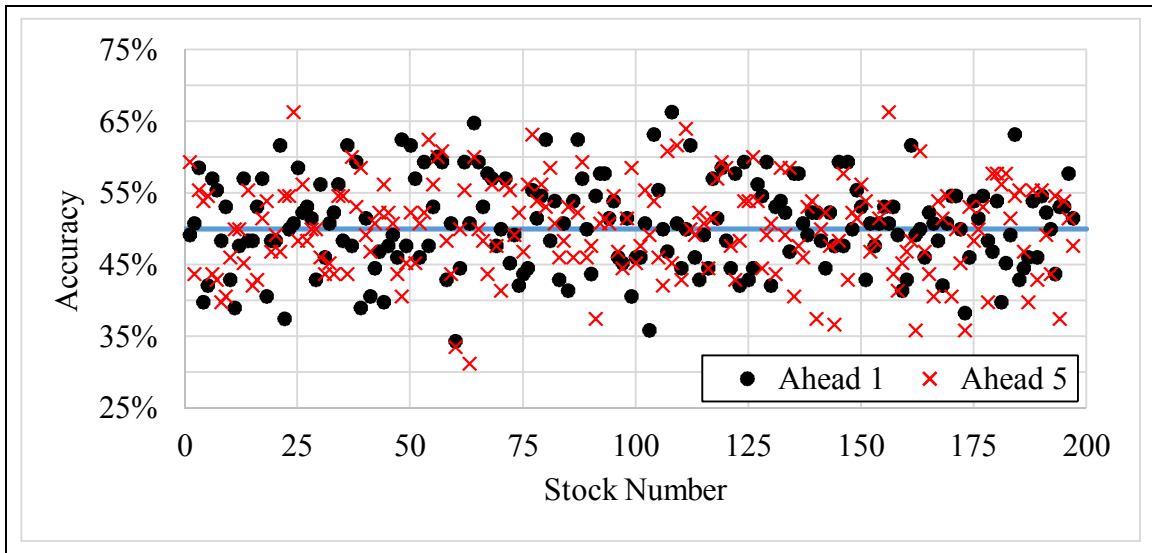
Figure 3-2. Univariate SSA Directional Forecast Accuracy

Table 3-1. Univariate ARIMA-SSA Directional Accuracy

Forecast Interval	Accuracy	ARIMA Frequency (Out of 197)	SSA Frequency (Out of 197)
Ahead 1	> 60%	2	13
	> 55%	24	47
	> 50%	93	102
	< 45%	8	40
	< 40%	0	9
Ahead 5	> 60%	1	13
	> 55%	18	44
	> 50%	93	97
	< 45%	22	41
	< 40%	2	11
Ahead 1 & Ahead 5	Same Direction	111	88

the number of incorrect forecasts also increased noticeably as seen in the counts below

45%. While the SSA method appeared promising, more extensive testing would be advised

to establish robustness of forecasts.

As previously defined with ARIMA, for each stock, the percentage of forecasts falling

within the 95% confidence interval is called the coverage. One-step and five-step

confidence interval coverage results for SSA are displayed in Figure 3-3. Overall results

were greater than 81% coverage for all stocks. The mean for one-step ahead was 90.8%

and for five steps ahead was 92.1%.
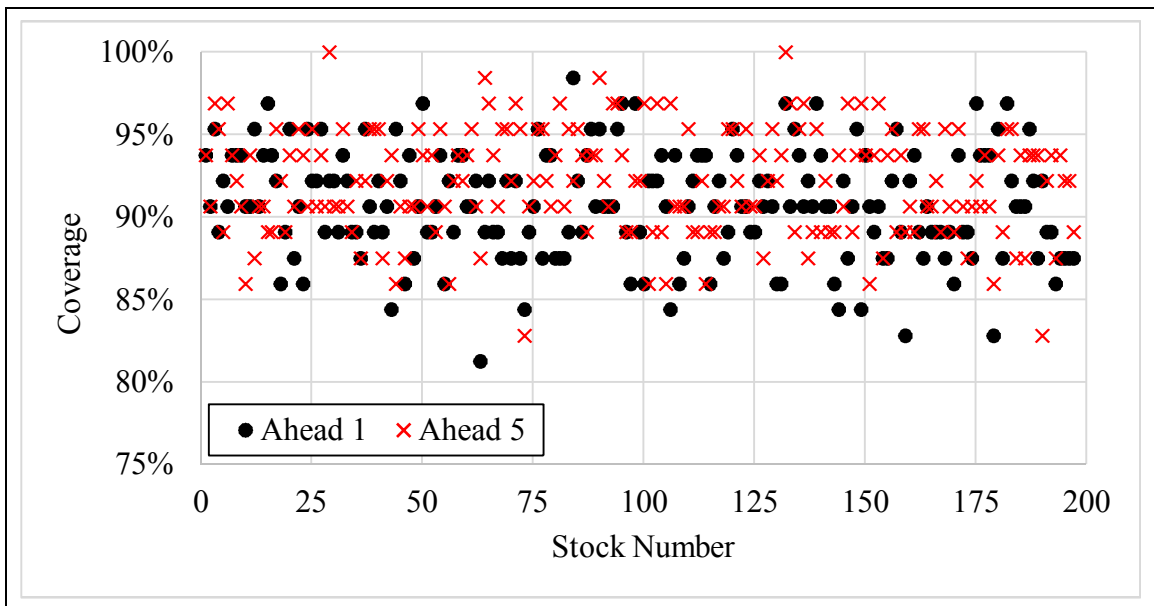
Figure 3-3. Univariate SSA Forecast Coverage within 95% CI

Confidence interval widths were compared between univariate ARIMA and univariate

SSA results. Each of the 197 stocks was evaluated for one-step and five-step forecasts.

Computations were similar to those for comparing univariate and bivariate ARIMA

models: 1) all windows of a given stock were evaluated for SSA CI shrinkage or growth with respect to the equivalent ARIMA CI and a frequency (proportion) was computed; and 2) the mean percentage of shrinkage (across all windows) was calculated for each stock. As seen in the top plot of results provided in Figure 3-4, mean frequency of shrinkage (as opposed to growth) for SSA only occurred in seven cases for the one-step forecasts and in only a single instance for the five-step forecasts. The bottom plot in Figure 3-4 shows that SSA CI growth always resulted (in the mean). Although mean SSA confidence intervals were wider than those for ARIMA, impact to accuracy results appeared unrelated since results for SSA accuracy were better than ARIMA accuracy. Finally, the statistical means across all 197 stocks are provided in Table 3-2.

Table 3-2. ARIMA-SSA Univariate CI Statistics

ARIMA-SSA Shrinkage	Ahead 1 (%)	Ahead 5 (%)
Frequency	39.0	31.3
Mean Shrinkage	-20.2	-32.5
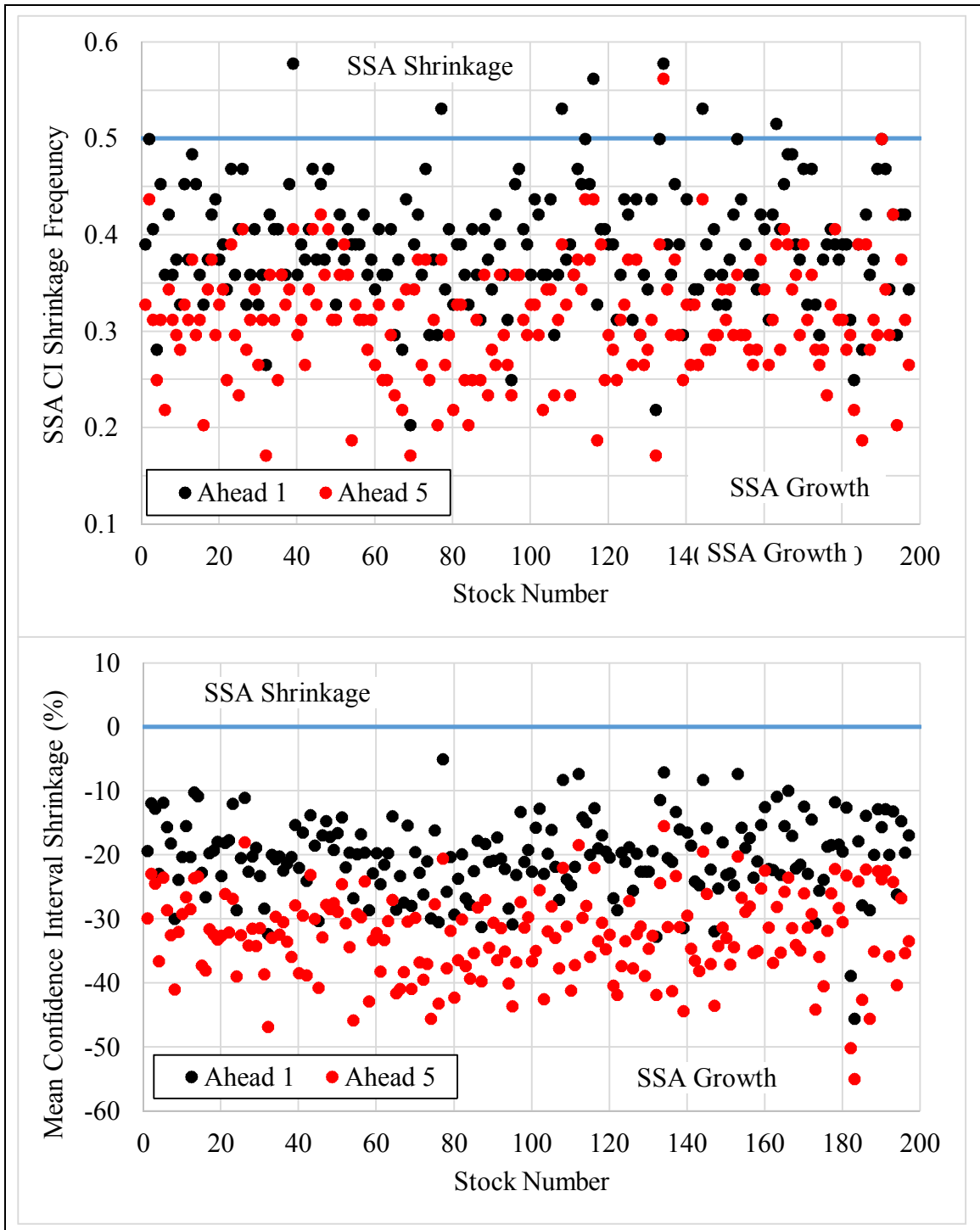
Figure 3-4. ARIMA-SSA Confidence Intervals

3.3 Multivariate SSA

Extending univariate SSA to a multivariate framework yields four possible forecasting algorithms: multivariate series may be stacked vertically or horizontally, then forecast by recurrent or vector methods [36]. For consistency with prior analyses, the multivariate SSA (MSSA) method applied herein uses horizontally-stacked, recurrent forecasting [37].* The modeling stages for MSSA are equivalent to those for univariate SSA: decomposition followed by reconstruction, each separated into two parts. The difference is in the structure of the trajectory matrix.

Multivariate Horizontal Model

A multivariate system is composed of multiple univariate time series $X^{(m)} = \left(x_j^{(m)} \right)_{j=1}^{N}$, $m = 1, ..., M$, of M series of length N. The embedding stage consists of constructing a trajectory matrix for each univariate series $X^{(m)}$, then horizontally stacking the individual matrices to form a composite multivariate trajectory matrix as

$$\mathbf{X} = \left[X_1^{(1)} : \dots : X_K^{(1)} : \dots : X_1^{(M)} : \dots : X_K^{(M)} \right] = \left[\mathbf{X}^{(1)} : \dots : \mathbf{X}^{(M)} \right], \qquad (3\text{-}12)$$

where each $X_k^{(m)}$ has length L and the $\mathbf{X}^{(m)}$ matrices are size $L \times K$. Subsequent processing steps for SVD, grouping, and diagonal averaging, are identical to the univariate case.

---

* Horizontal stacking is the only format implemented in the 'Rssa' package at this time.

Multivariate Recurrent Forecasting [37]

After identification of the r leading eigentriples and the selected groups, a system of reconstructed signals is defined as $\tilde{X}^{(m)} = \left(\tilde{x}_j^{(m)}\right)_{j=1}^{N}$, $m = 1, \ldots, M$. With the multivariate recurrent algorithm, forecasting may be performed using the left eigenvectors $\left\{U_j\right\}_{j=1}^{r}$ or the right eigenvectors $\left\{V_j\right\}_{j=1}^{r}$. For consistency with the previous univariate analysis, left eigenvectors were used as described in the following.

In a concept parallel to the univariate scenario, the last $L-1$ values of the reconstructed signals are collected in a matrix

$$\tilde{\mathbf{X}} = \begin{pmatrix} \tilde{x}_{N-L+1}^{(1)}, \ldots, \tilde{x}_N^{(1)} \\ \tilde{x}_{N-L+1}^{(2)}, \ldots, \tilde{x}_N^{(2)} \\ \vdots \\ \tilde{x}_{N-L+1}^{(M)}, \ldots, \tilde{x}_N^{(M)} \end{pmatrix}, \tag{3-13}$$

U_j^{∇} denotes the vectors of the first $L-1$ coordinates of the U_j eigenvectors, and π_j defines the last coordinates of the eigenvectors. With $\upsilon = \sum_{j=1}^{r} \pi_j^2$, if $\upsilon < 1$, then the forecast, $R_N = \left(\hat{x}_N^{(1)}, \hat{x}_N^{(2)}, \ldots, \hat{x}_N^{(M)}\right)^{\mathrm{T}}$, exists and may be computed with

$$R_N = \tilde{\mathbf{X}}\, G_L, \quad \text{where} \quad G_L = \frac{1}{1-\upsilon^2} \sum_{j=1}^{r} \pi_j U_j^{\nabla}. \tag{3-14}$$

Equation (3-14) expresses that forecasts of each series are generated from the same linear recurrent formula produced by considering all series in the multivariate system.

Multivariate Implementation

The same basic approach taken for SSA-based univariate analysis was utilized for multivariate SSA investigation. Refer to §3.1.2. Routines from the R-package 'Rssa' are multivariate-compatible by specifying the input parameters accordingly. Sequential SSA was again used to separate complex component commingling.

The window-length input parameter for the 'Rssa' function *ssa* used the previously-described method for finding the period of the largest spectral peak, but selected the largest peak of multiple series as the guide for determining the value. The working assumption was that the highest peak corresponded to the strongest separable signal. Five was again used for the minimum window length.

The process for automatic group selection was identical to the univariate case because only a single weighted-correlation matrix is generated by the process. Thresholds were set to the same levels.

Mean values of recurrent forecasting were also found via the *rforecast* function, but with the direction set for left eigenvectors, as mentioned previously. Confidence intervals were computed with a modified *bforecast* function designed to accept joint block sampling to maintain intra-series- and system-correlation structure. Whereas the univariate analysis used block sampling, multivariate block sampling was executed by finding the block size [see equation (3-9)] for each series and selecting the largest block size to perform parallel

block sampling. In this manner, time-aligned blocks of residuals from each series were resampled to compute the bootstrapped confidence intervals.

Upper and lower confidence interval offsets for each series were computed separately, parallel to the univariate model. Lastly, as observed in the univariate outputs, 'Rssa' multivariate forecasts appeared relative to the reconstructed series, rather than the original series. Consequently, to remedy this manifestation, an adjustment was applied individually to multivariate series forecasts so that offsets relative to the reconstructed series were shifted to exhibit the same offset relative to the original series.

3.4 Bivariate SSA Results

Bivariate singular spectrum analysis was applied to the same stock-pair list obtained from correlation screening as described in §2.4.1 and §2.5 for ARIMA analyses. Inasmuch as SSA is a nonparametric method, differences in processing and forecasting do not contribute supplementary coefficients, but simply generate additional outputs due to the augmented trajectory matrix. The properties of model groups are identical to univariate SSA (see §3.2), but naturally, the selected groups are different. Frequency of simple models was tracked and is plotted in Figure 3-5. The mean frequency of occurrence was 3.1% (shown as a red line), slightly higher than for univariate SSA, but still significantly below the 34.7% found for bivariate ARIMA. Here again, the one-group model only expresses a lack of separability in the eigenvector space, but nevertheless yields valid forecasts.

Similar to the univariate-SSA outcome, noticeable correlation remained in the bivariate second-pass residuals as found with the Ljung-Box portmanteau test (not shown). The mean rejection rate was 99.7%. Based on the results from the univariate analysis, no attempt was made to extract additional information by either ARIMA modeling of the residuals or a third pass of SSA.
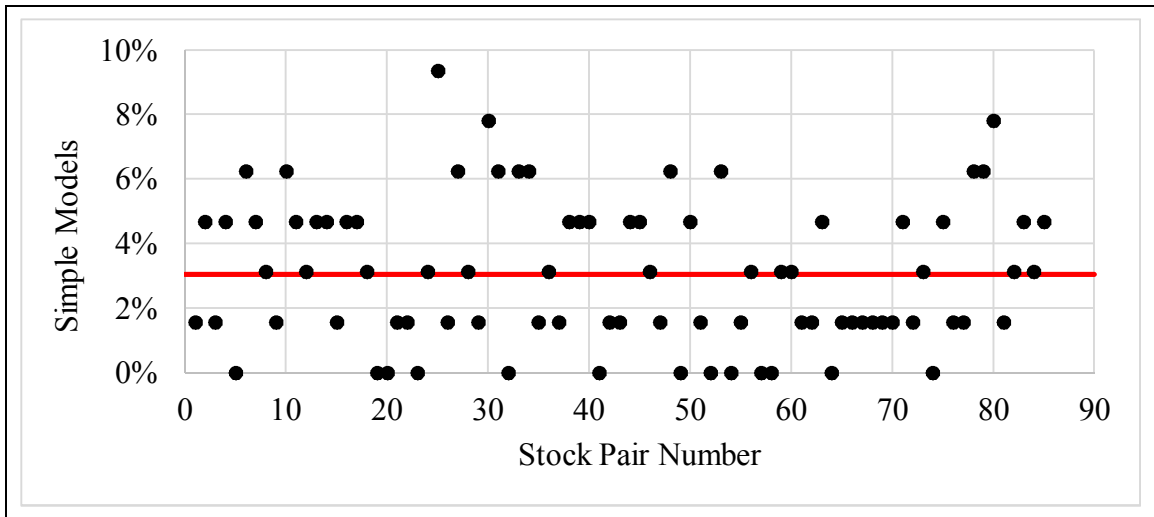
Figure 3-5. Frequency of Bivariate SSA Simple Models
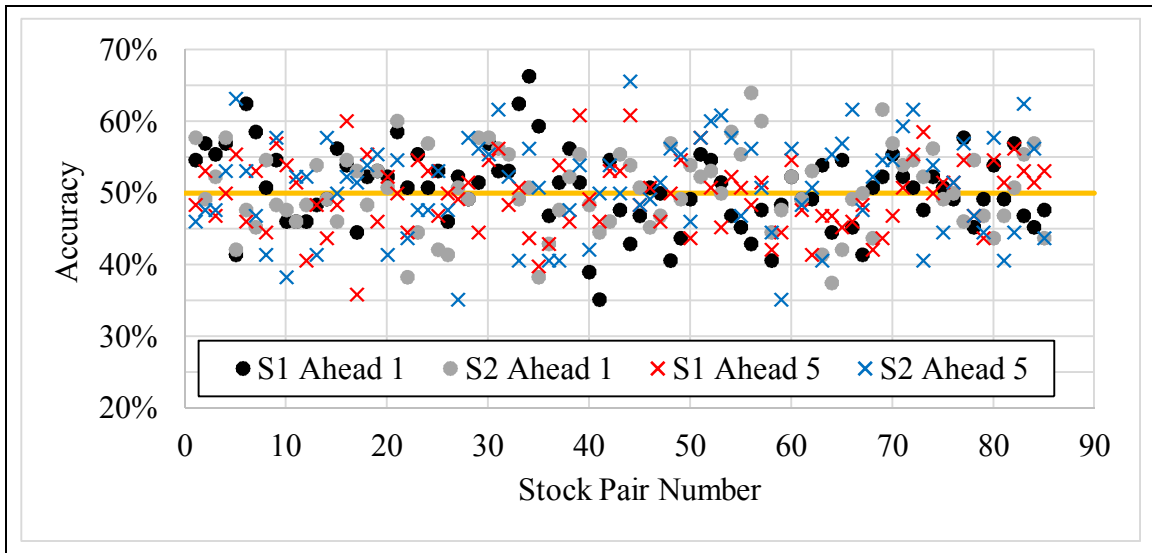
Figure 3-6. Bivariate SSA Directional Forecast Accuracy

Results for one-step and five-step directional forecast accuracy for out-of-sample data are plotted in Figure 3-6 for the 85 stock pairs. Results above (or below) 50% (yellow line) represent probability improvement over a random guess. Examination of Table 3-3, which summarizes and compares results of bivariate ARIMA and SSA, revealed similar performance, but the counts for ARIMA forecasts appeared somewhat better.

Table 3-3. Bivariate ARIMA-SSA Directional Accuracy Summary

Forecast Interval	Accuracy	ARIMA		SSA	
		Stock 1 Frequency (Out of 85)	Stock 2 Frequency (Out of 85)	Stock 1 Frequency (Out of 85)	Stock 2 Frequency (Out of 85)
Ahead 1	> 60%	8	8	3	4
	> 55%	25	30	17	20
	> 50%	50	57	47	42
	< 45%	12	7	11	15
	< 40%	3	2	2	3
Ahead 5	> 60%	7	10	3	8
	> 55%	29	24	11	27
	> 50%	48	46	40	48
	< 45%	20	14	16	19
	< 40%	3	4	2	3
Ahead 1 & Ahead 5	Same Direction	45	39	43	46
Ahead 1 & Ahead 5	Same Direction	45		46	

Computational results for confidence interval coverage for bivariate SSA forecasts are given in Figure 3-7. One-step and five-step CI coverages are plotted for both stocks in the pair, designated as S1 and S2.* Overall results were greater than 68% coverage for all stocks, much lower than the bivariate-ARIMA minimum coverage of 79%. A summary-comparison for ARIMA-to-SSA is given in Table 3-4. Mean coverage was clearly lower for bivariate SSA.

---

* S1 and S2 correspond to Stock 1 and Stock 2, respectively, in the list in Appendix 8.2.
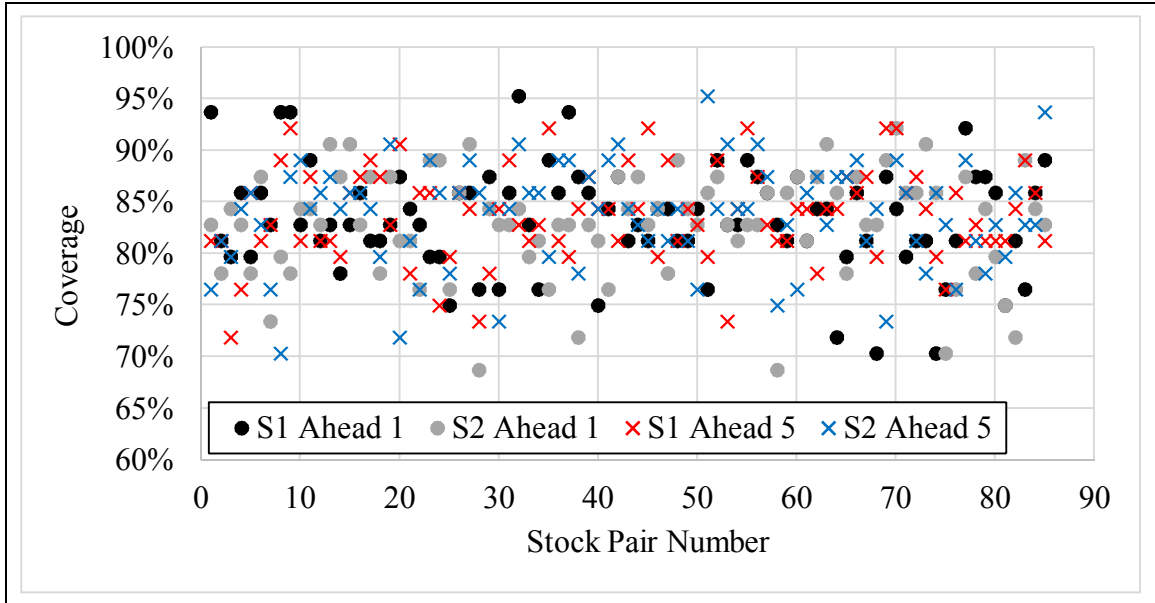
Figure 3-7. Bivariate SSA Forecast Coverage within 95% CI

Table 3-4. Bivariate Mean Coverage Comparison

	ARIMA		SSA	
	Ahead 1 (%)	Ahead 5 (%)	Ahead 1 (%)	Ahead 5 (%)
Stock 1	91.1	96.5	83.3	83.8
Stock 2	90.9	95.8	82.9	83.8

3.5 Summary of Univariate-Bivariate SSA Results

The objective of this investigation was to evaluate the hypothesis that an SSA-based bivariate forecast would outperform its univariate version, as was demonstrated for the ARIMA process. Results were mixed, as seen in the summary provided in Table 3-5. The occurrence of the simple model was slightly higher in the bivariate analysis, indicating

more difficult separability of the time series. The more complex bivariate trajectory matrix may explain the slight increase. Directional accuracy results were mixed. The univariate model displayed better one-step forecast accuracy, as it produced higher percentages for > 55% and > 60% regions; however, for five-step forecasts in these regions, the univariate results landed between the Stock 1 and Stock 2 predictions. Thus, for five-step forecasts, it is inconclusive whether bivariate or univariate accuracy is better. A possible explanation for the wide difference in Stock 1 and Stock 2 forecast results may be that discrepancies within series pairs return model compromises during the group selection process.* Lastly, recall that CI coverage provides the statistics for the number of forecasts falling inside confidence intervals. Examination of the CI coverage clearly revealed that bivariate coverage averaged 9.4% and 9% less for one-step and five-step forecasts, respectively. Although such degradation in coverage was substantial, degradation in forecast accuracy was mixed, producing an ambiguous indicator with respect to CI coverage and a nebulous response since the CI widths are computed based on residuals only.

---

* In an effort to identify other possible explanations, analyses were repeated with Stock 1 and Stock 2 categories interchanged in the pair definitions, producing swapped, identical results.

Table 3-5. SSA Univariate-Bivariate Comparison

Parameter		Univariate (%)	Bivariate (%)	
Simple Model		1.9	3.1	
Directional Forecast Accuracy			**Stock 1**	**Stock 2**
Ahead 1	**> 60%**	6.6	3.5	4.7
	> 55%	23.9	20.0	23.5
	> 50%	51.8	55.3	49.4
	< 45%	20.3	12.9	17.6
	< 40%	4.6	2.4	3.5
Ahead 5	**> 60%**	6.6	3.5	9.4
	> 55%	22.3	12.9	31.8
	> 50%	49.2	47.1	56.5
	< 45%	20.8	18.8	22.4
	< 40%	5.6	2.4	3.5
Confidence Interval Coverage				
Ahead 1		90.8	83.3	82.9
Ahead 5		92.1	83.8	83.8

Bivariate confidence interval widths were compared against univariate CI widths, like that done for ARIMA modeling (see §2.6). Results, given in Figure 3-8, show a high frequency of occurrence (> 81%) and significant CI shrinkage (> 22%) for bivariate SSA. Mean values are summarized in Table 3-6 and mean percent shrinkage was computed as previously defined in equation (2-17).   Substantial bivariate CI shrinkage while maintaining similar forecast accuracy pointed toward advantage of the bivariate model.
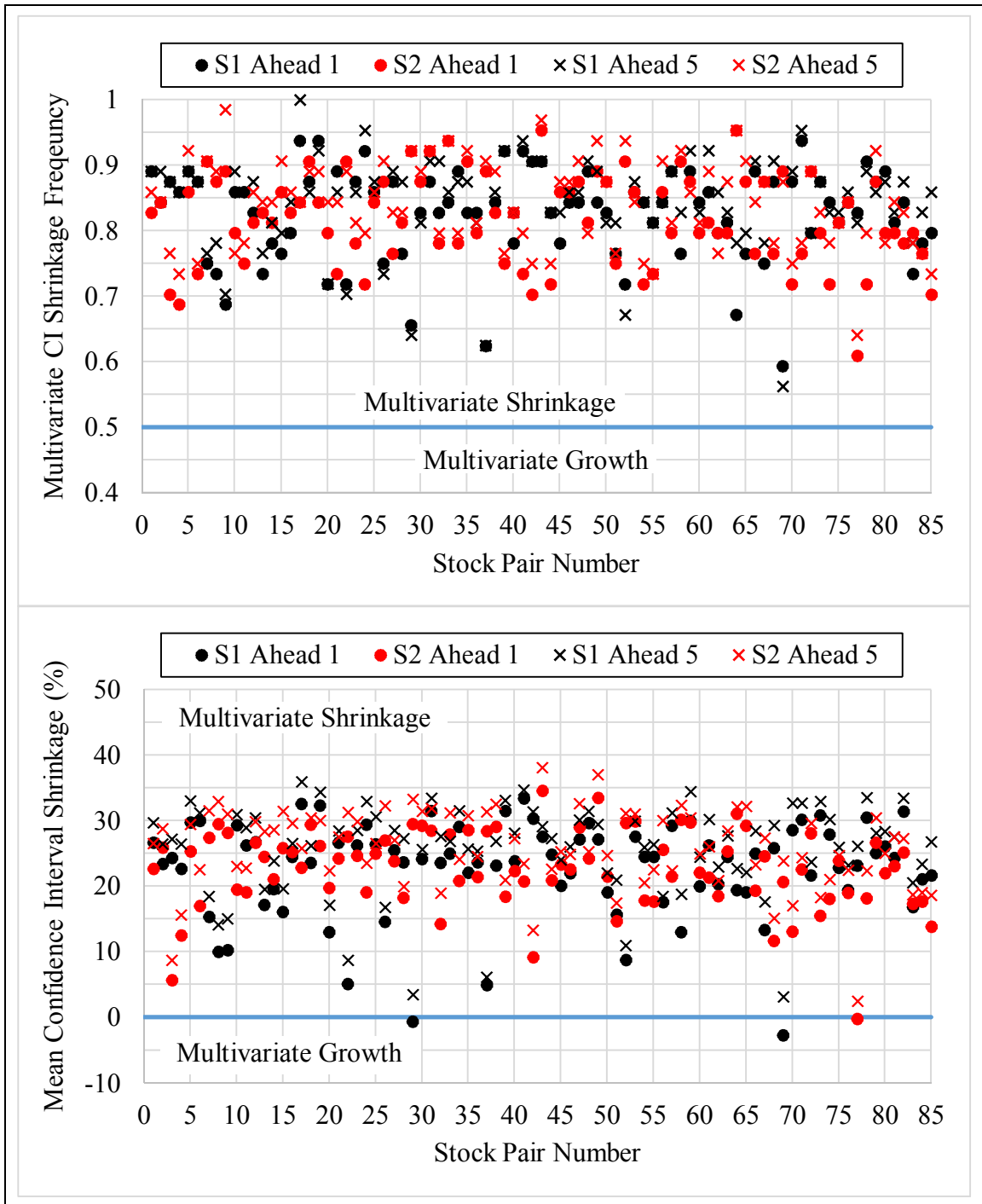
Figure 3-8. Univariate-Bivariate SSA Confidence Intervals

Table 3-6. Univariate-Bivariate SSA CI Statistics

Bivariate Shrinkage	Ahead 1 (%)	Ahead 5 (%)
Frequency (Stock 1)	82.6	84.2
Frequency (Stock 2)	81.7	84.2
Mean Shrinkage (Stock 1)	22.9	25.8
Mean Shrinkage (Stock 2)	22.7	25.9

3.6 Trivariate SSA

For complete comparison, trivariate SSA analysis was applied to the set of stock triplets (see Appendix 8.3) used for ARIMA modeling. Trivariate SSA generates an extension of the trajectory matrix like that described for bivariate SSA. Frequency of simple models was found, as shown in Figure 3-9, to have a mean of 4.1% (displayed as a red line), slightly higher than bivariate SSA results.
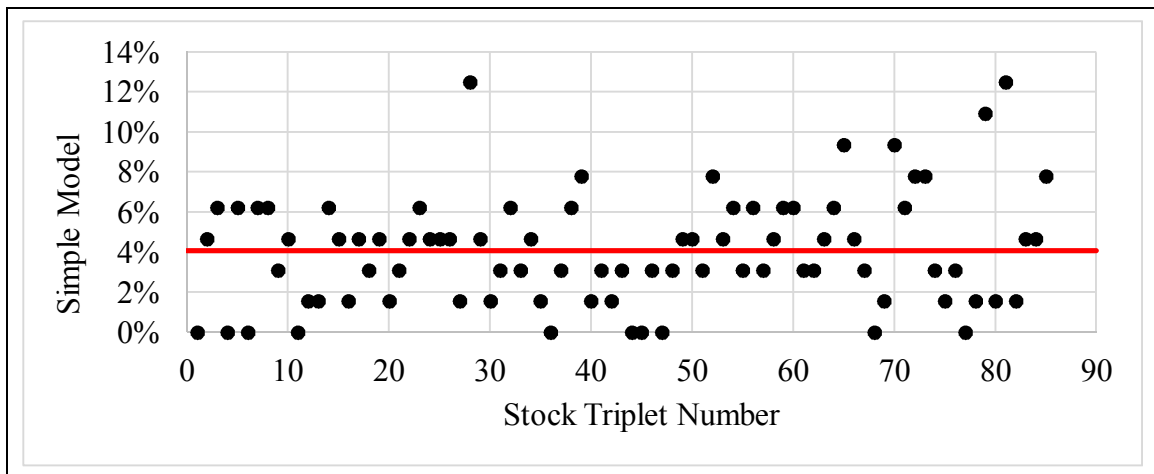
Figure 3-9. Frequency of Trivariate SSA Simple Models

Results of the Ljung-Box portmanteau test showed obvious residual correlation of the sequential-SSA second-pass (not shown) with a mean rejection rate of 99.7%. As previously stated for the bivariate results, no effort was made to extract additional information due to the poor results obtained from the univariate-based experiments.

For the 85 triplets, results of one-step and five-step directional forecast accuracies for out-of-sample data are displayed in Figure 3-10. Results farther away from the 50%-threshold (yellow line) random guess represent improvements.[*] Table 3-7 summarizes the trivariate results for both SSA and ARIMA models. ARIMA yielded superior one-step directional accuracy forecasts over SSA. Five-step directional accuracy was very similar between SSA and ARIMA, with a slight advantage to SSA for better consistency encompassing the three-stock groups. Trivariate SSA appeared to surpass ARIMA with regard to identifying triplets all moving in the same direction. Generally, SSA demonstrated greater concordance in forecasting across three stocks comprising the triplets, giving SSA a slim advantage over ARIMA.

---

[*] Even consistently incorrect forecasts may be more utilitarian than random guesses.
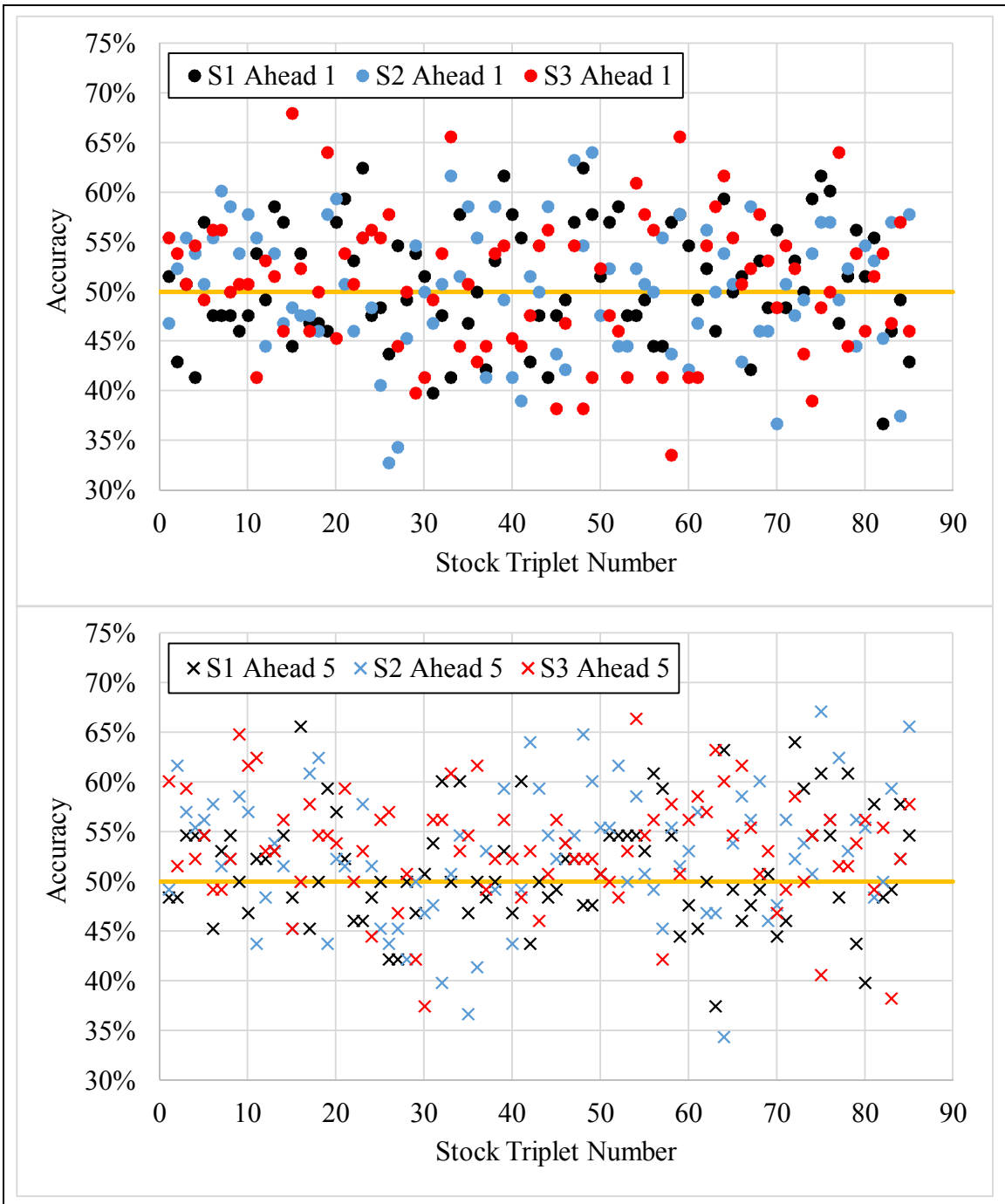
Figure 3-10. Trivariate SSA Directional Forecast Accuracy

Table 3-7. Trivariate SSA Directional Accuracy Summary

Forecast Interval	Accuracy	Frequency (Out of 85)					
		ARIMA			SSA		
		Stock 1	Stock 2	Stock 3	Stock 1	Stock 2	Stock 3
Ahead 1	> 60%	10	10	10	5	4	7
	> 55%	26	30	26	24	24	21
	> 50%	48	53	56	41	45	47
	< 45%	14	11	13	14	17	19
	< 40%	3	2	1	2	5	5
Ahead 5	> 60%	13	6	10	9	11	10
	> 55%	24	24	32	15	32	31
	> 50%	40	46	46	41	55	64
	< 45%	14	7	11	8	9	6
	< 40%	1	0	5	2	3	2
Ahead 1 & Ahead 5	Same Direction	52	35	54	43	46	48
Ahead 1 & Ahead 5	Same Direction	13			20		

Confidence interval coverage for trivariate SSA forecasts are shown in Figure 3-11, with mean statistics provided in Table 3-8. The top plot in the figure shows the one-step CI coverage and the bottom plot shows the five-step coverage. Examination of the plots revealed Stock 3 to have the most variation in coverage, extending both above and below Stock 1 and Stock 2, but exhibited a mean approximately 6% lower. SSA coverage was

clearly lower than ARIMA coverage, but SSA also demonstrated more consistency between one- and five-step coverage.
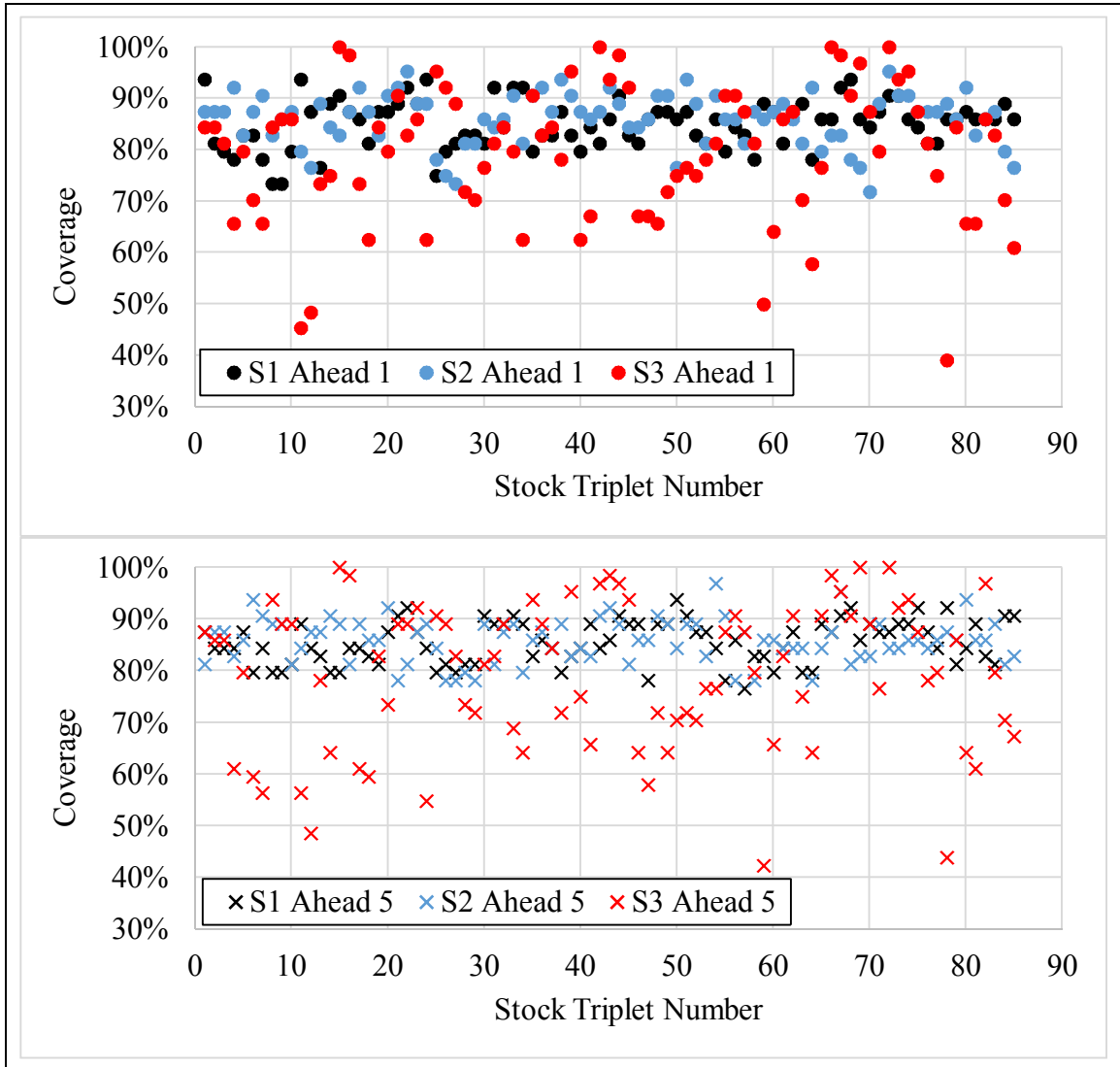
Figure 3-11. Trivariate SSA Forecast Coverage within 95% CI

Table 3-8. Trivariate Mean Coverage Comparison

	ARIMA		SSA	
	Ahead 1 (%)	Ahead 5 (%)	Ahead 1 (%)	Ahead 5 (%)
Stock 1	91.0	96.8	85.0	85.5
Stock 2	91.4	95.9	86.0	85.8
Stock 3	92.1	96.9	79.3	79.0

3.7 Summary of Trivariate SSA Results

A statistical summary of SSA results is given in Table 3-9.   Overall, trivariate SSA appeared to produce superior forecasting results; however, univariate forecast statistics were better in the one-step, > 60%-accuracy category for all but trivariate Stock 3 which was best. For five-step forecasting, trivariate SSA was clearly the best. It was interesting to note that for confidence interval coverage, trivariate SSA landed between the other two, with the exception of the Stock 3 category which showed the worst coverage. A review of the shrinkage plots in Figure 3-12 revealed good consistency for Stock 1 and Stock 2 categories, whereas there was a wide variation in shrinkage results for Stock 3, which yielded more extreme values than with Stock 1 or Stock 2.* As seen in the CI shrinkage statistics given in Table 3-10, bivariate SSA produced more CI shrinkage than trivariate SSA, but both showed substantial shrinkage over univariate SSA, except for the trivariate Stock 3 category. If mean statistics are ignored and more focus is applied to the Stock 3

---

* For clarity, not shown on the lower plot for S3 Ahead 1 are 8 points below -100% (which are > -322%) and for S3 Ahead 5 are 7 points below -100% (which are > -305%).

category, then selecting a specific more extreme stock triplet might yield better returns. Moreover, it is reasonable to assume that a typical trader/investor, operating with a portfolio consisting of a limited amount of operating capital, would therefore focus on higher-probability trades.

Table 3-9. SSA Multivariate Comparison

Parameter		Univariate (%)	Bivariate (%)		Trivariate (%)		
Simple Model		1.9	3.1		4.1		
Directional Forecast Accuracy			**Stock 1**	**Stock 2**	**Stock 1**	**Stock 2**	**Stock 3**
Ahead 1	**> 60%**	6.6	3.5	4.7	5.9	4.7	8.2
	> 55%	23.9	20.2	23.5	28.2	28.2	24.7
	> 50%	51.8	55.3	49.4	48.2	52.9	55.3
	< 45%	20.3	12.9	17.6	16.5	20.0	22.4
	< 40%	4.6	2.4	3.5	2.4	5.9	5.9
Ahead 5	**> 60%**	6.6	3.5	9.4	10.6	13.9	11.8
	> 55%	22.3	12.9	31.8	17.6	37.6	36.5
	> 50%	49.2	47.1	56.5	48.2	64.7	75.3
	< 45%	20.8	18.8	22.4	9.4	10.6	7.1
	< 40%	5.6	2.4	3.5	2.4	3.5	2.4
Confidence Interval Coverage							
Ahead 1		90.8	83.3	82.9	85.0	86.0	79.3
Ahead 5		92.1	83.8	83.8	85.5	85.8	79.0
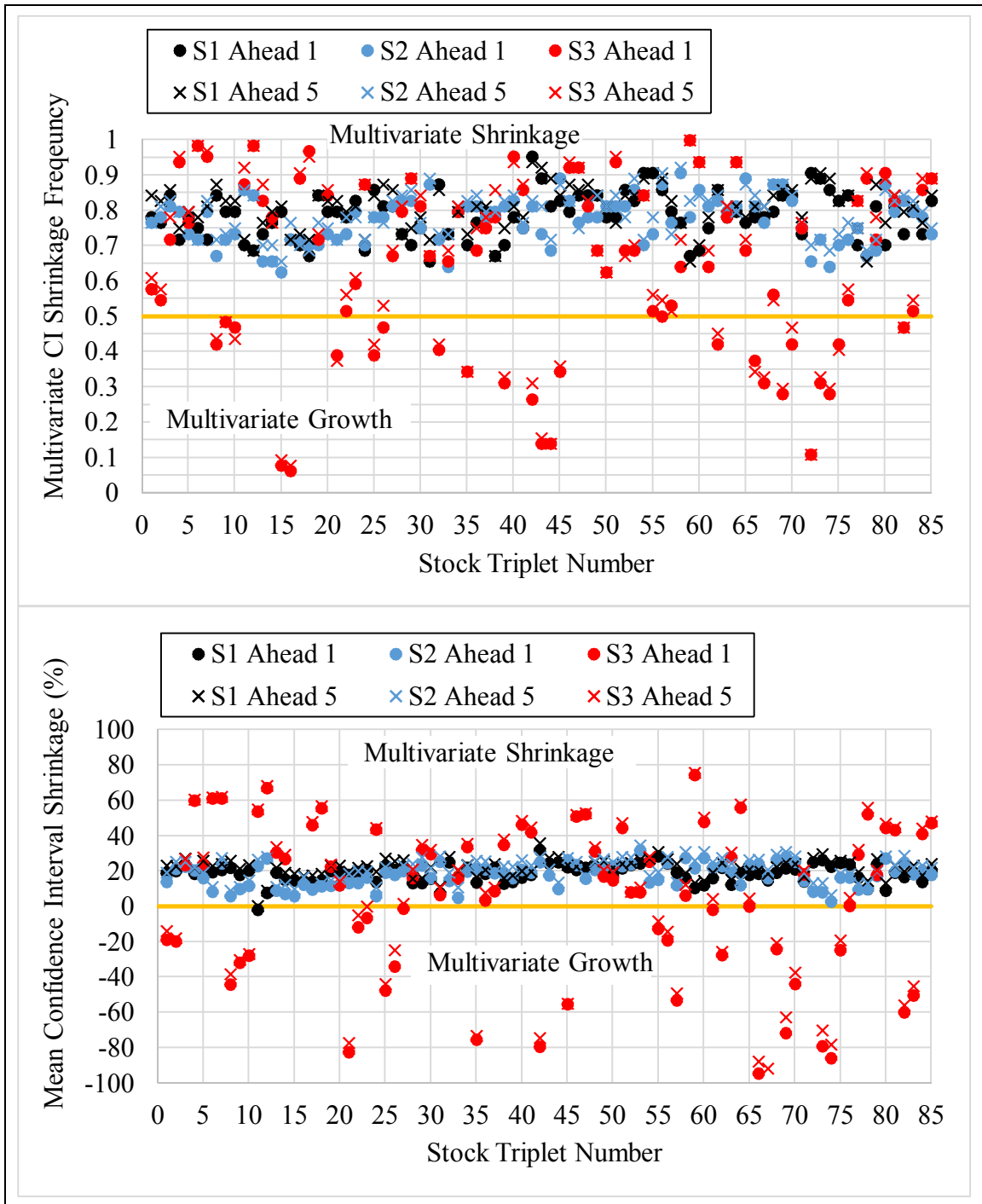
Figure 3-12. Univariate-Trivariate SSA Confidence Intervals

Table 3-10. Bivariate-Trivariate SSA Confidence-Interval Statistics

Multivariate Shrinkage	Ahead 1 (%)		Ahead 5 (%)	
	Bivariate	Trivariate	Bivariate	Trivariate
Frequency (Stock 1)	82.6	78.5	84.2	80.9
Frequency (Stock 2)	81.7	77.3	84.2	79.2
Frequency (Stock 3)		64.0		65.7
Mean Shrinkage (Stock 1)	22.9	18.9	25.8	21.9
Mean Shrinkage (Stock 2)	22.7	17.8	25.9	21.4
Mean Shrinkage (Stock 3)		-13.8		-9.8

3.8 Overall Summary of ARIMA-SSA Results

A side-by-side summary of ARIMA and SSA results is provided in Table 3-11. A review indicated that forecast accuracy rates were generally better for bivariate and trivariate ARIMA for both one and five steps. Univariate SSA showed substantially higher forecast accuracies, but also higher rates of incorrect forecasts. SSA yielded generally higher rates of incorrect forecasts across all models and lower confidence interval coverages. A comparison of these overall results confirmed the hypothesis that the multivariate model produces improved forecasts over the univariate model.

Table 3-11. Composite Summary of ARIMA-SSA Results

Parameter		Univariate (%)		Bivariate (%)				Trivariate (%)					
		ARIMA	SSA	ARIMA		SSA		ARIMA			SSA		
Random-Walk | Simple Model		63.0	1.9	34.7		3.1		44.8			4.1		
Directional Forecast Accuracy				S1	S2	S1	S2	S1	S2	S3	S1	S2	S3
Ahead 1	**> 60%**	1.0	6.6	9.4	9.4	3.5	4.7	11.8	11.8	11.8	5.9	4.7	8.2
	> 55%	12.2	23.9	29.4	35.3	20.2	23.5	30.6	35.3	30.6	28.2	28.2	24.7
	> 50%	47.2	51.8	58.8	67.1	55.3	49.4	56.5	62.4	65.9	48.2	52.9	55.3
	< 45%	4.1	20.3	14.1	8.2	12.9	17.6	16.5	12.9	15.3	16.5	20.0	22.4
	< 40%	0.0	4.6	3.5	2.4	2.4	3.5	3.5	2.4	1.2	2.4	5.9	5.9
Ahead 5	**> 60%**	0.5	6.6	8.2	11.8	3.5	9.4	15.3	7.1	11.8	10.6	13.9	11.8
	> 55%	9.1	22.3	34.1	28.2	12.9	31.8	28.2	28.2	37.6	17.6	37.6	36.5
	> 50%	47.2	49.2	56.5	54.1	47.1	56.5	47.1	54.1	54.1	48.2	64.7	75.3
	< 45%	11.2	20.8	23.5	16.5	18.8	22.4	16.5	8.2	12.9	9.4	10.6	7.1
	< 40%	1.0	5.6	3.5	4.7	2.4	3.5	1.2	0.0	5.9	2.4	3.5	2.4
Confidence Interval Coverage													
Ahead 1		92.5	90.8	91.1	90.9	83.3	82.9	91.0	91.4	92.1	85.0	86.0	79.3
Ahead 5		92.8	92.1	96.5	95.8	83.8	83.8	96.8	95.9	96.9	85.5	85.8	79.0

Note: S1, S2, and S3 refer to Stock 1, Stock 2, and Stock 3, respectively.

4. BIVARIATE FORECASTING WITH ALTERNATE COMPONENTS

"The most important function of statistics is to produce evidence showing the

relation of one group of phenomena to another"

– Arthur Lyon Bowley (1869-1957), in the book "Elements of Statistics," 1937

This chapter reports on an investigation of time series forecasting based on a bivariate structure incorporating alternate second components. While bivariate analyses in previous chapters included a primary stock and a highly-correlated secondary stock as the auxiliary component in the series, this bivariate study evaluates forecasting with three distinctive secondary components: 1) trading volume, 2) first principal component (from PCA), and 3) volatility. These elements are detailed in the following sections.

In contrast to the 85 pairs studied in the prior two chapters, all 197 members of the stock data list (see Appendix 8.1) outlined in Chapter 1 are examined to determine forecast accuracies and 95% confidence intervals.[*] Forecasting methods described in the previous chapters on ARIMA-based statistical estimation and singular spectrum analysis (SSA) are applied with each of these secondary components to investigate the hypothesis of improved out-of-sample forecasting with secondary alternatives.

---

[*] Limiting the number of bivariate pairs is unnecessary due to the disassociated nature of the specified secondary components.

4.1 Trading Volume

In this study, volume refers to the quantity of shares that are traded between buyers and sellers during a sampling interval, i.e., one day. Many traders/investors use volume as a measure of participation and an aid to identify market strength or weakness.* John Bollinger, creator of the Bollinger bands technical indicator, argues that "volume is everything," particularly when combined with another trading indicator. Moreover, he points to volume as "the crux of analysis." [38] Daily trading volume was included in the input data download as mentioned in §1.1. Bivariate analyses were conducted for each of the 197 stocks to assess possible forecast improvements by applying each stock's associated daily volume as a second variate in the ARIMA and SSA models.

4.1.1   ARIMA-Volume Model Results

As described previously in §2.4, bivariate ARIMA modeling was implemented with an ARMA model preceded by differencing for stationarizng† the data. Volume data was preconditioned by normalizing to 1 million to bring it into the range of stock prices, mean centered (by subtracting the [windowed] mean), and stationarized by differencing using the *ndiffs* function from the R-Package 'forecast' version 6.2 (verified with a KPSS test).

Unlike the results for univariate ARIMA modeling, the frequency of occurrence of random-walk models in the bivariate-volume analyses showed an order of magnitude lower average

---

* Beyond the scope of this report, there are numerous strategic approaches that active traders have taken to utilize volume to enhance trading outcomes.

† In the field of time series analysis, the word stationarize means "to make stationary."

of 5.6% (compared to the univariate 63%) and six times lower than the bivariate stock average of 34.7%. The distribution of random-walk models is shown in Figure 4-1. As previously stated, a lower occurrence of random-walk models indicates a higher probability of predictability.
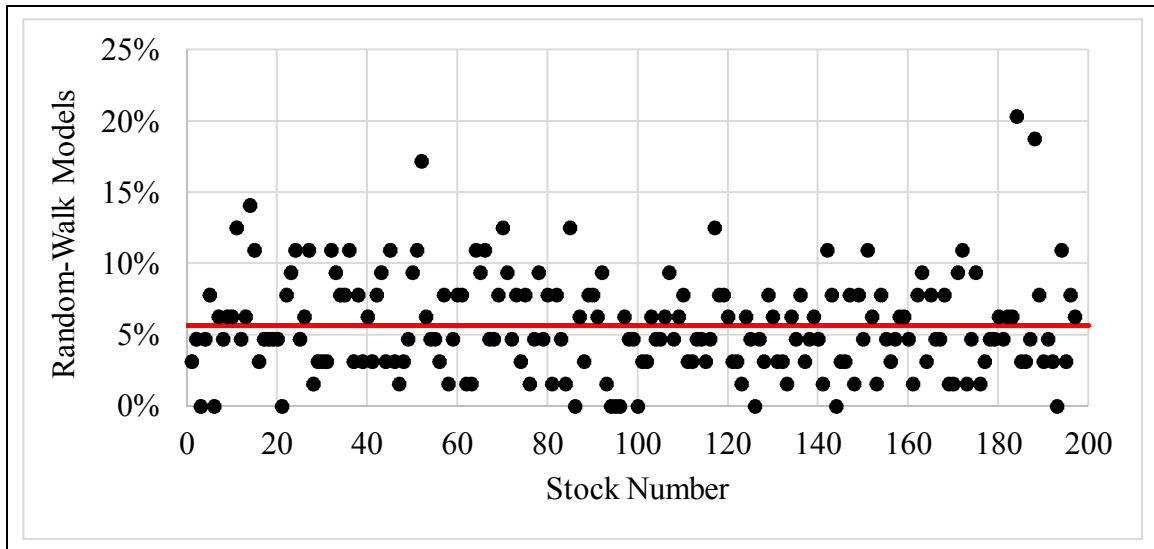
Figure 4-1. Frequency of Bivariate (Volume) Random-Walk Models

Results of serial correlation tests seen in Figure 4-2 show the percentage of p-values falling below 0.05 for the Ljung-Box and spectral density tests of bivariate-volume model residuals. While the Ljung-Box results showed only five stocks with p-value percentages above 5%, the spectral density outputs yielded 35 stocks exceeding 5% frequency, with the maximum at 12.5%. The spectral density test was performed using identical parameters as specified for the bivariate-stock models in §2.5, except that a Bartlett smoothing kernel was used for increased computational speed.[*] Assuming that the spectral density test is a

---

[*] Test evaluations showed output p-values very close to those obtained with the Daniell kernel. At least an order of magnitude in speed was gained by using the Bartlett kernel.

more powerful measurement as demonstrated in [23] and [24], greater numbers of stocks exceeding 0.05 indicated more information retained in the model residuals. Nevertheless, the bivariate-volume results yielded significant improvements over the univariate models in the directional forecast accuracy evaluations.
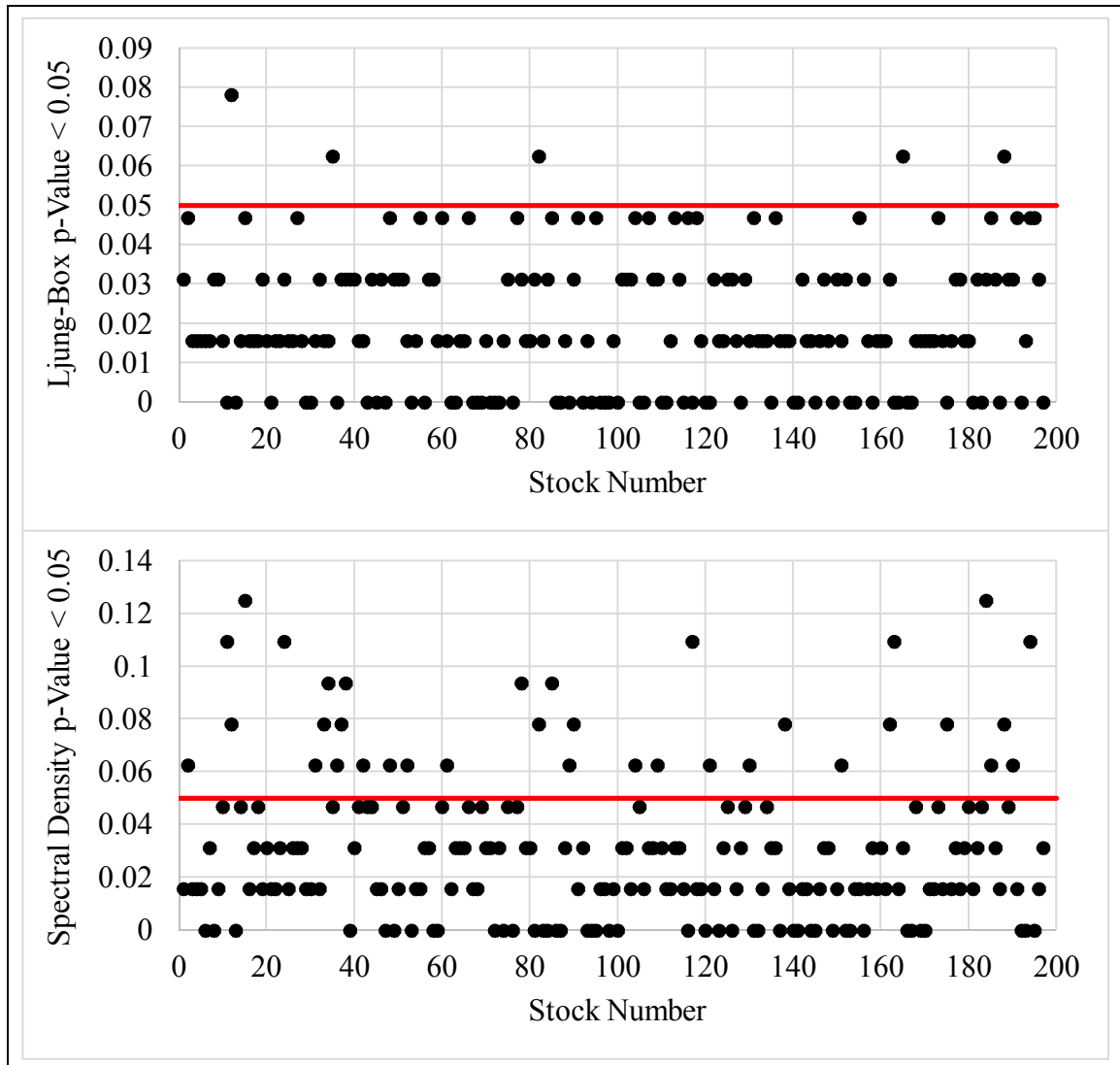
Figure 4-2. Serial Correlation for Bivariate (Volume) Model Residuals

Results of volume-augmented directional forecast accuracy are shown in Figure 4-3 and summarized in Table 4-1. Overall, frequency of bivariate-volume accurate directional one- and five-step forecasts were substantially higher than for the univariate ARIMA models. As seen in Figure 4-3, the highest one-step directional forecast accuracy was 67.2% (one occurrence) and the highest five-step accuracy was 65.6% (four occurrences). Clearly, incorporation of a stock's trading volume generated a significant enhancement to forecasting directional movement.
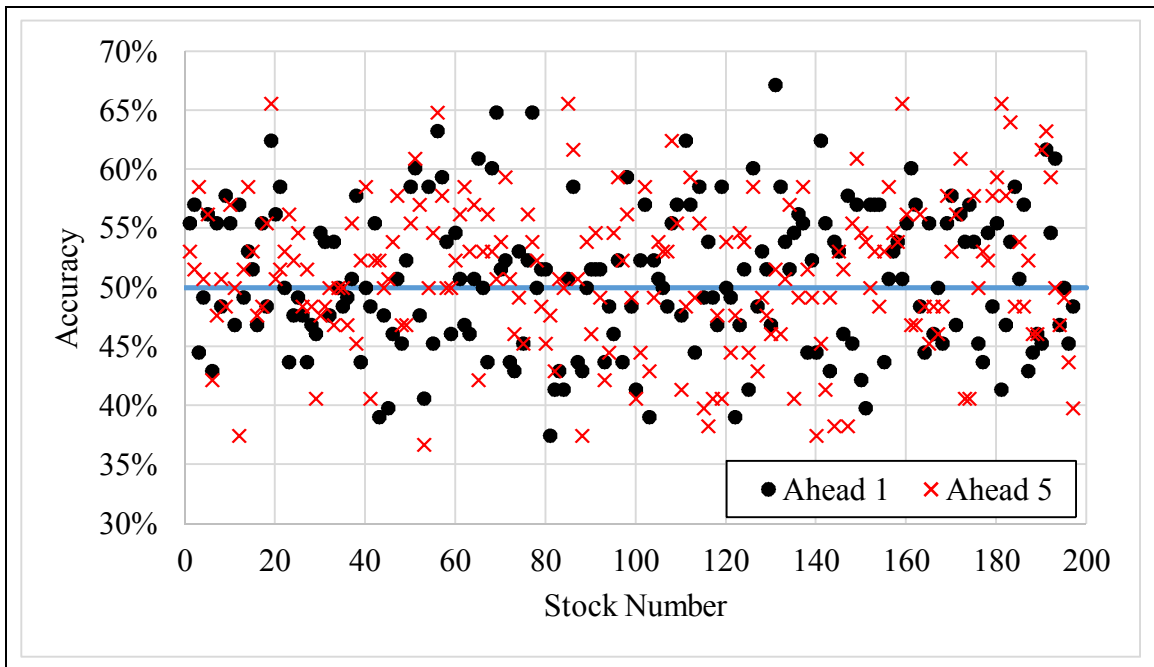
Figure 4-3. ARIMA Bivariate (Volume) Directional Forecast Accuracy

Table 4-1. ARIMA Univariate-Bivariate (Volume) Directional Accuracy

Forecast Interval	Accuracy	Frequency (Out of 197)	
		Univariate	Bivariate
Ahead 1	> 60%	2	14
	> 55%	24	56
	> 50%	93	102
	< 45%	8	35
	< 40%	0	6
Ahead 5	> 60%	1	13
	> 55%	18	51
	> 50%	93	106
	< 45%	22	30
	< 40%	2	9
Ahead 1 & Ahead 5	Same Direction	111	99

Results of coverage analyses are given in Figure 4-4. It is evident that five-step forecast coverage displays a higher mean than the one-step coverage. This may be partially explained by wider confidence intervals at the farther forecast range; however, this was also revealed in the bivariate-stock ARIMA coverage results of §2.5. A comparison of univariate-bivariate (volume) coverage is provided in Table 4-2.
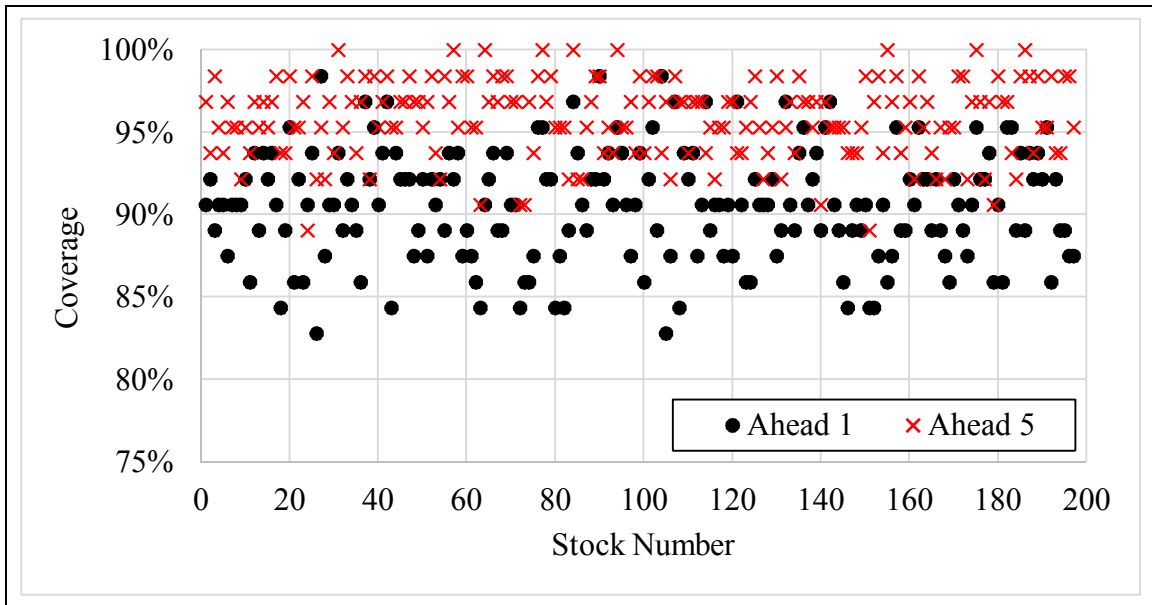
Figure 4-4. ARIMA Bivariate (Volume) Forecast Coverage within 95% CI

Table 4-2. ARIMA Univariate-Bivariate (Volume) Mean Coverage

	Univariate (%)	Bivariate (%)
Ahead 1	92.5	90.6
Ahead 5	92.8	95.8

Confidence interval shrinkage is graphed in Figure 4-5 and summarized in Table 4-3. While the frequency of shrinkage for the bivariate-volume model was high for both one-step and five-step forecasts (averaging greater than 75% of windows), the five-step CI mean percent change actually achieved a growth at -1.7%. The figure clearly shows a consistent percent shrinkage for the one-step forecast, but the five-step forecast had more variation.

112
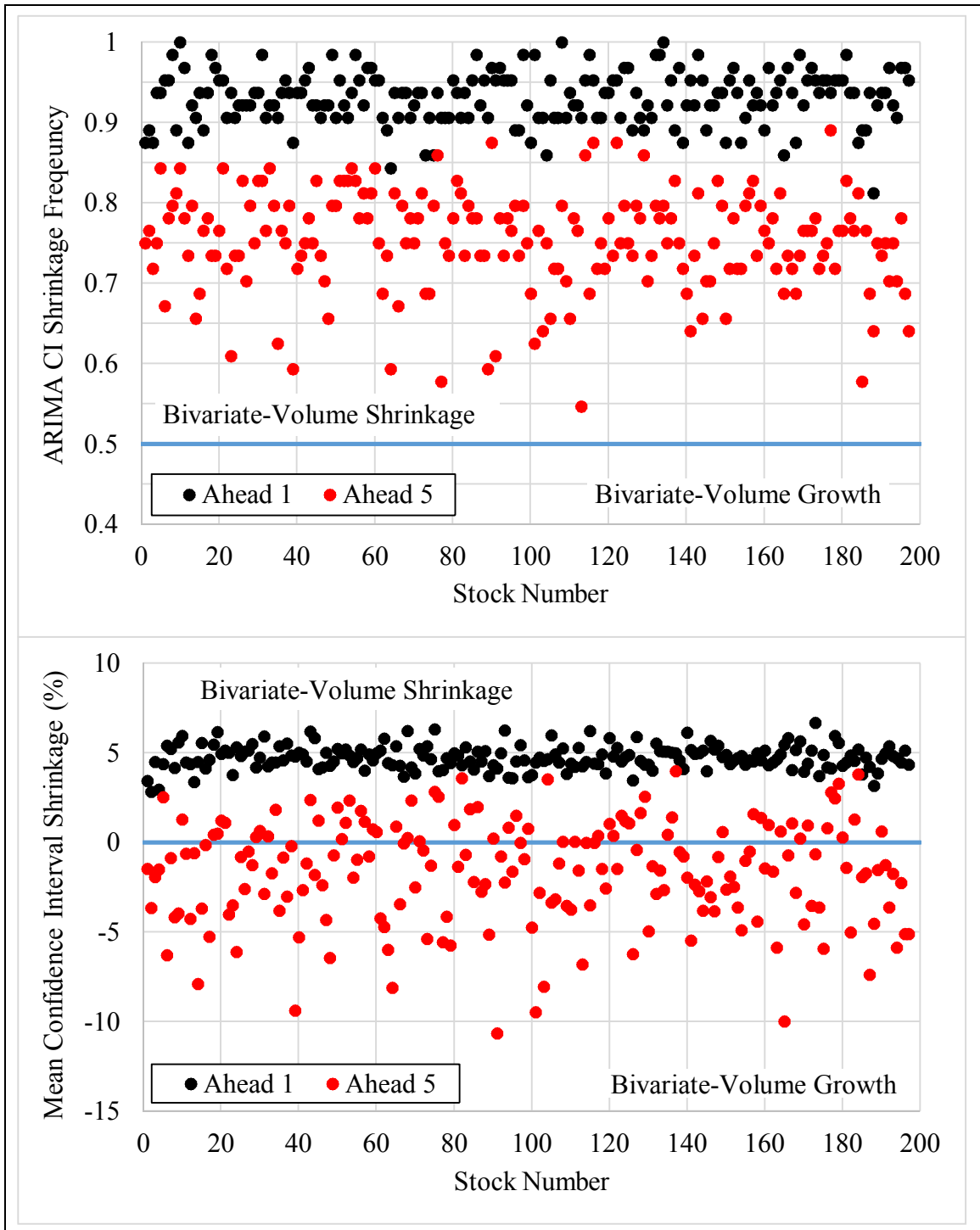
Figure 4-5. ARIMA Univariate-Bivariate (Volume) Confidence Intervals

Table 4-3. ARIMA Univariate-Bivariate (Volume) CI Statistics

Bivariate Shrinkage	Ahead 1 (%)	Ahead 5 (%)
Frequency	93.1	75.2
Mean Shrinkage	4.7	-1.7

4.1.2   SSA-Volume Model Results

Bivariate modeling with singular spectrum analysis (SSA) was performed by substituting the associated stock trading volume for the second variate. Unlike the ARIMA process, no differencing, for either stock or volume, was required for SSA. As in the previous bivariate ARIMA modeling, volume data was preprocessed by normalizing to 1 million to bring it into the range of stock prices[*] and mean centered (by subtracting the [windowed] mean).

Frequency of occurrence for the "simple"[†] SSA model was incrementally higher for the bivariate-volume scenario with a mean value of 4.5%, compared to the bivariate-stock models at 3.1% and the univariate structures at 1.9%. This was likely an indication of increased difficulty separating the structural components (in eigenvector space). The distribution for frequency of occurrence of simple models is shown in Figure 4-6.

---

[*]   Normalizing into a similar numerical range avoids any potential matrix singularities due to widely disparate values.

[†]   Recall that the "simple" SSA model is merely a classification defined as a one-group composition (for each sequential pass), devised as a comparison for the random-walk model.
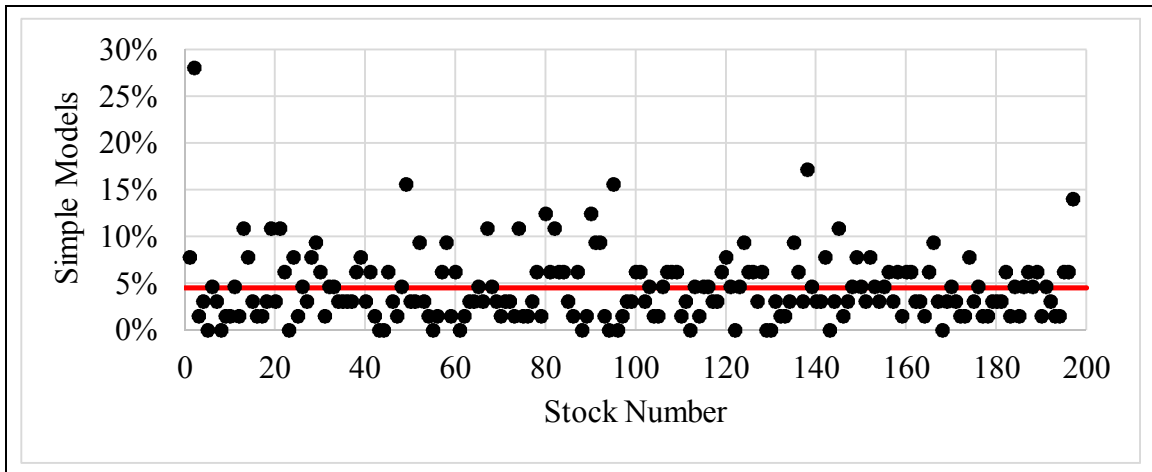
Figure 4-6. Frequency of Bivariate (Volume) SSA Simple Models

Results of forecasting directional accuracy for SSA bivariate-volume are shown in the plot of Figure 4-7. Maximum accuracy was reached at 70.3% for a one-step forecast and 68% for a five-step forecast. Table 4-4 summarizes the SSA bivariate-volume prediction accuracies. Examination of the table revealed results very similar between the bivariate-volume and univariate systems.
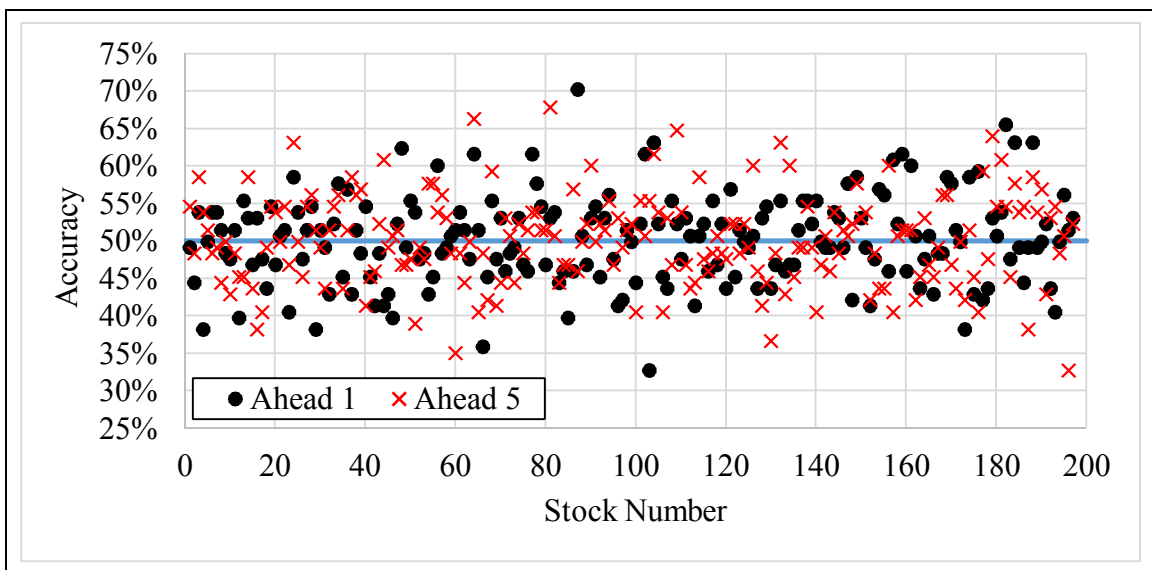
Figure 4-7. SSA Bivariate (Volume) Directional Forecast Accuracy

Table 4-4. SSA Univariate-Bivariate (Volume) Directional Accuracy

Forecast Interval	Accuracy	Frequency (Out of 197)	
		Univariate	Bivariate
Ahead 1	> 60%	13	13
	> 55%	47	37
	> 50%	102	98
	< 45%	40	36
	< 40%	9	8
Ahead 5	> 60%	13	13
	> 55%	44	36
	> 50%	97	98
	< 45%	41	37
	< 40%	11	6
Ahead 1 & Ahead 5	Same Direction	88	109

Coverage of the 95% confidence intervals had mean values of 76.3% and 76.5% for one-step and five-step forecasts, respectively. A distributional plot is provided in Figure 4-8. As seen in the comparison with univariate SSA coverage given in Table 4-5, the SSA bivariate-volume coverage was significantly lower, but as mentioned earlier, Table 4-4 shows similar accuracies.
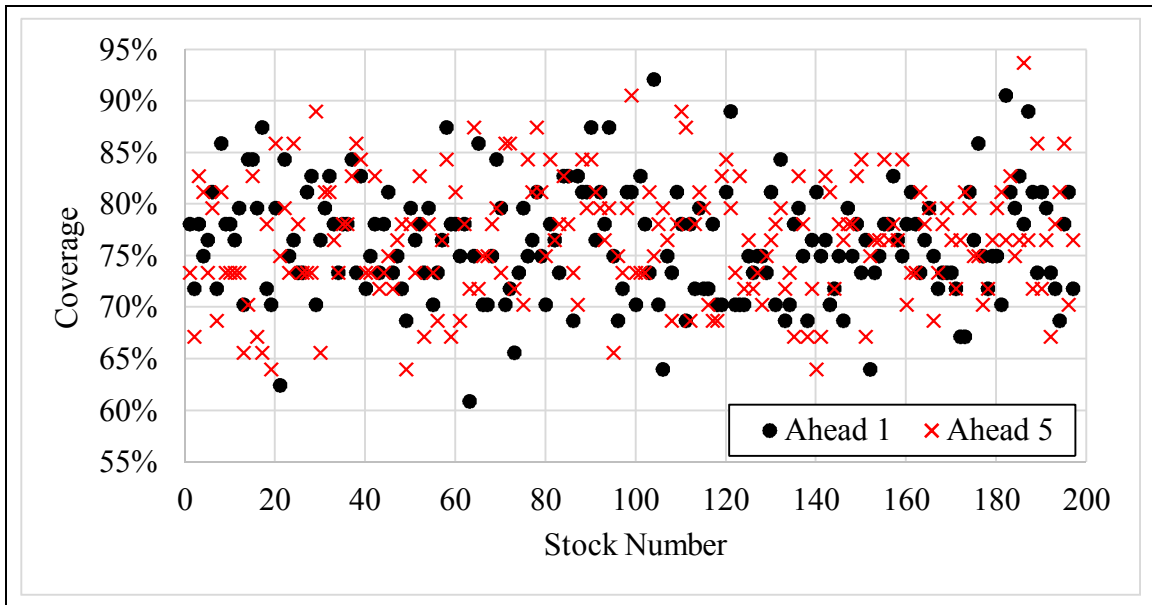
Figure 4-8. SSA Bivariate (Volume) Forecast Coverage within 95% CI

Table 4-5. SSA Univariate-Bivariate (Volume) Mean Coverage

	Univariate (%)	Bivariate (%)
Ahead 1	90.8	76.3
Ahead 5	92.1	76.5

As seen in Figure 4-9 and Table 4-6, the SSA bivariate-volume models produced a significant amount of confidence interval shrinkage in both frequency and percent reduction, with a few exceptions. With a consistent forecast accuracy compared to the univariate models, it was clear that inclusion of volume as a second variate was remarkable.
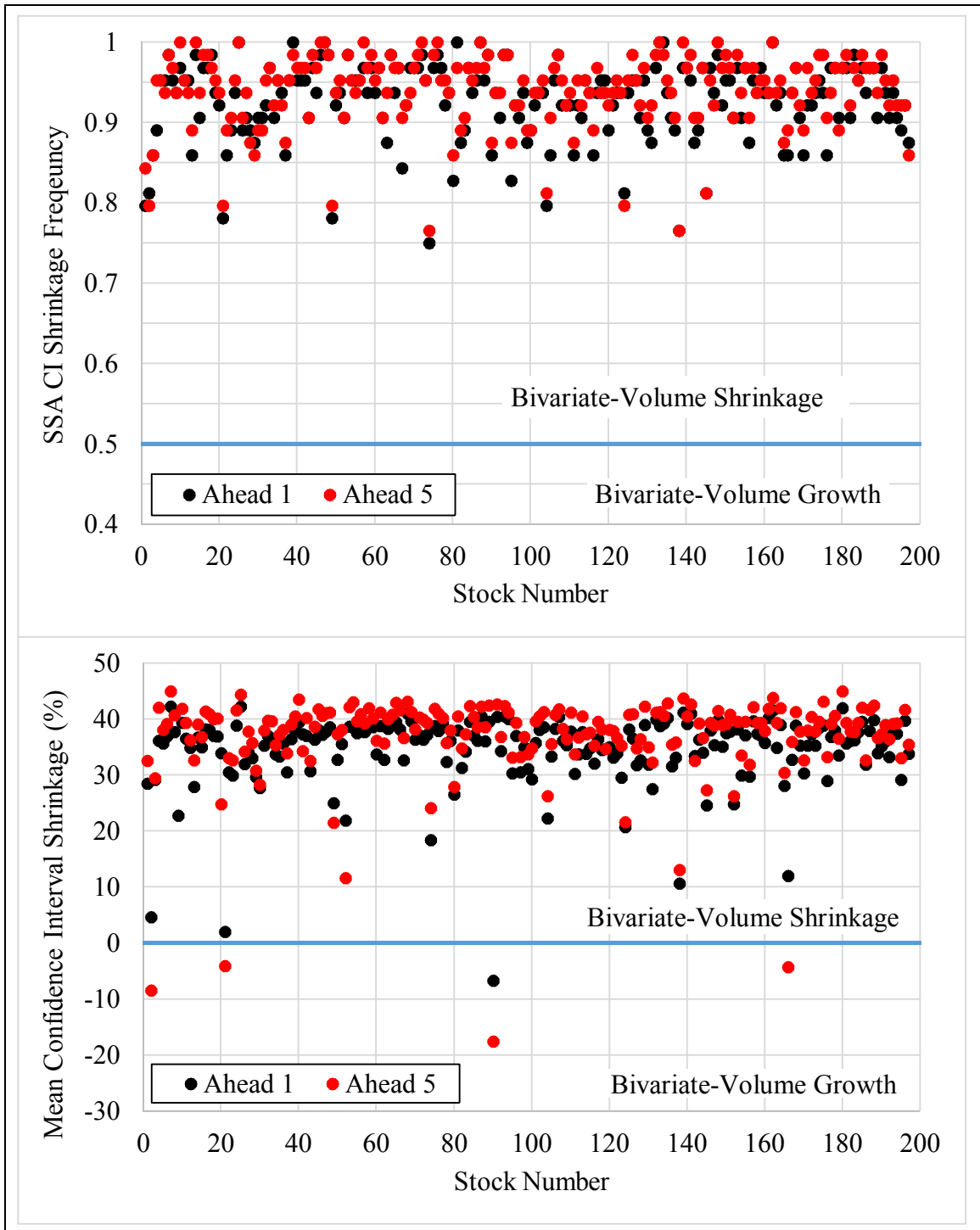
Figure 4-9. SSA Univariate-Bivariate (Volume) Confidence Intervals

Table 4-6. SSA Univariate-Bivariate (Volume) CI Statistics

Bivariate Shrinkage	Ahead 1 (%)	Ahead 5 (%)
Frequency	92.8	93.9
Mean Shrinkage	34.5	36.4

4.2 Principal Component Analysis (PCA)

The objective of principal component analysis (PCA) on a set of variables is to elucidate the variance-covariance structure through linear combinations of the variables. The general goals of PCA are 1) data reduction, and 2) interpretation. If p components (of length n) are required to account for total variability of a system, frequently a smaller number of principal components, k, can efficiently describe much of the variability. These k principal components may then supplant the original $p \times n$ data with a reduced set of size $k \times n$. Principal component analysis frequently exposes previously hidden relationships, allowing possibly unexpected interpretations, but usually serves as an intermediate step, rather than an end of analysis. In this study, application of PCA is focused only on data reduction to produce a single variate most representative of the entire set of 197 stocks in the stock list. The largest principal component was used as the second variate in a bivariate forecast analysis of each stock.

The essence of principal component analysis is to find another basis that is a linear combination of the original basis, to best describe the data set [39]. If $\mathbf{X}$ is the original data matrix of size $m \times n$, where m is the number of variables (e.g., stocks) and each column is

a time sample of all m members, then another matrix $\mathbf{Y}$ of size $m \times n$ can re-express $\mathbf{X}$ by a linear transformation $\mathbf{P}$, effecting a change of basis:

$$\mathbf{PX} = \mathbf{Y} \qquad\qquad (4\text{-}1)$$

Defining $\mathbf{p}_i$ as the rows of $\mathbf{P}$, $\mathbf{x}_i$ as the columns of $\mathbf{X}$, and $\mathbf{y}_i$ as the columns of $\mathbf{Y}$, yields

$$\mathbf{PX} = \begin{bmatrix} \mathbf{p}_1 \\ \vdots \\ \mathbf{p}_m \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 & \cdots & \mathbf{x}_n \end{bmatrix} = \begin{bmatrix} \mathbf{p}_1 \bullet \mathbf{x}_1 & \cdots & \mathbf{p}_1 \bullet \mathbf{x}_n \\ \vdots & \ddots & \vdots \\ \mathbf{p}_m \bullet \mathbf{x}_1 & \cdots & \mathbf{p}_m \bullet \mathbf{x}_n \end{bmatrix} = \mathbf{Y} \; , \qquad (4\text{-}2)$$

where the columns of $\mathbf{Y}$ are

$$\mathbf{y}_i = \begin{bmatrix} \mathbf{p}_1 \bullet \mathbf{x}_i \\ \vdots \\ \mathbf{p}_m \bullet \mathbf{x}_i \end{bmatrix} . \qquad\qquad (4\text{-}3)$$

Since the coefficients of $\mathbf{y}_i$ are dot products of $\mathbf{x}_i$ with the rows of $\mathbf{P}$, the j-th coefficient of $\mathbf{y}_i$ is a projection onto the j-th row of $\mathbf{P}$, thus forming a projection onto the $\{\mathbf{p}_1, \ldots, \mathbf{p}_m\}$ basis. These $\{\mathbf{p}_1, \ldots, \mathbf{p}_m\}$ row vectors are the principal components of $\mathbf{X}$.

Two important assumptions of principal component analysis are: 1) the basis vectors $\{\mathbf{p}_1, \ldots, \mathbf{p}_m\}$ are orthonormal, and 2) the magnitude of signal variance relates importance such that larger variance indicates more importance. Moreover, "principal components depend solely on the covariance matrix Σ " of $\mathbf{X}$, and PCA does not require multivariate normality [40].

The PCA method consists of: 1) finding the matrix **P** such that the covariance of **PX** is diagonal, and 2) putting the diagonal elements (i.e., eigenvalues) in descending order. Using results of linear algebra, is can be shown (see [40]) that the eigenvectors of the covariance matrix* $\mathbf{\Sigma} = \mathbf{X}\mathbf{X}^{\mathrm{T}}$ comprise the columns of matrix **P**, and corresponding eigenvalues constitute the associated variances. For this study, the first principal component was selected as the second variate for bivariate forecast analyses and PCA was performed using the *prcomp* function in the R-package 'stats' in the R core software.

4.2.1 ARIMA-PCA Model Results

For bivariate ARIMA models, the largest component was found by applying PCA to stationarized stock-price data. The proportion of total variance due to this principal component is shown in Figure 4-10, where the calculation is based on the contribution of all 197 stocks in each 100-sample window. The mean proportion of variance was 27.7%.
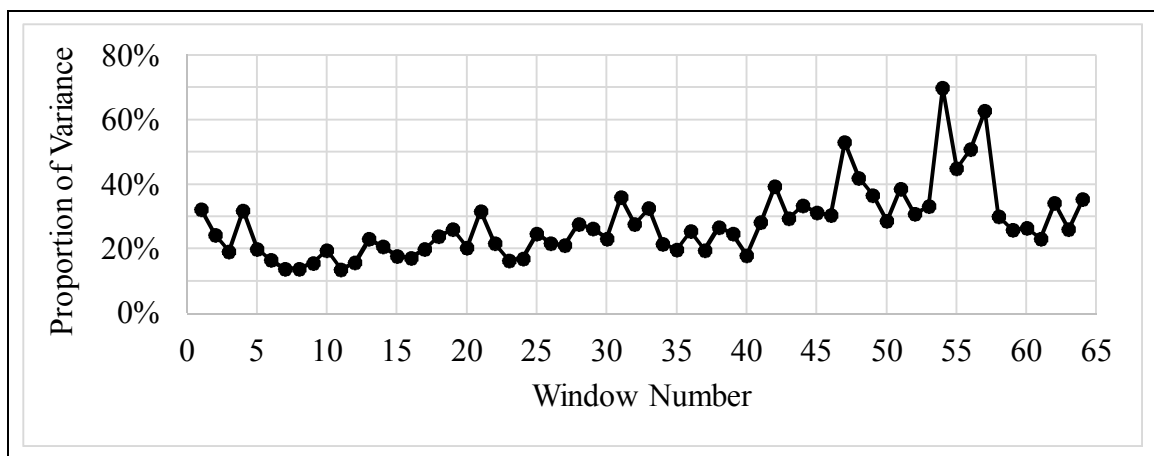
Figure 4-10. Proportion of Variance due to First Principal Component (ARIMA)

---

* Notation: $\mathbf{A}^{\mathrm{T}}$ is the transpose of matrix **A**.

The frequency of occurrence for random-walk models resulting from bivariate-PCA ARIMA modeling is shown in Figure 4-11. The mean was 32.1%, very close to the 34.7% mean of the bivariate-stock models, but still indicating a substantial proportion of unpredictable models.
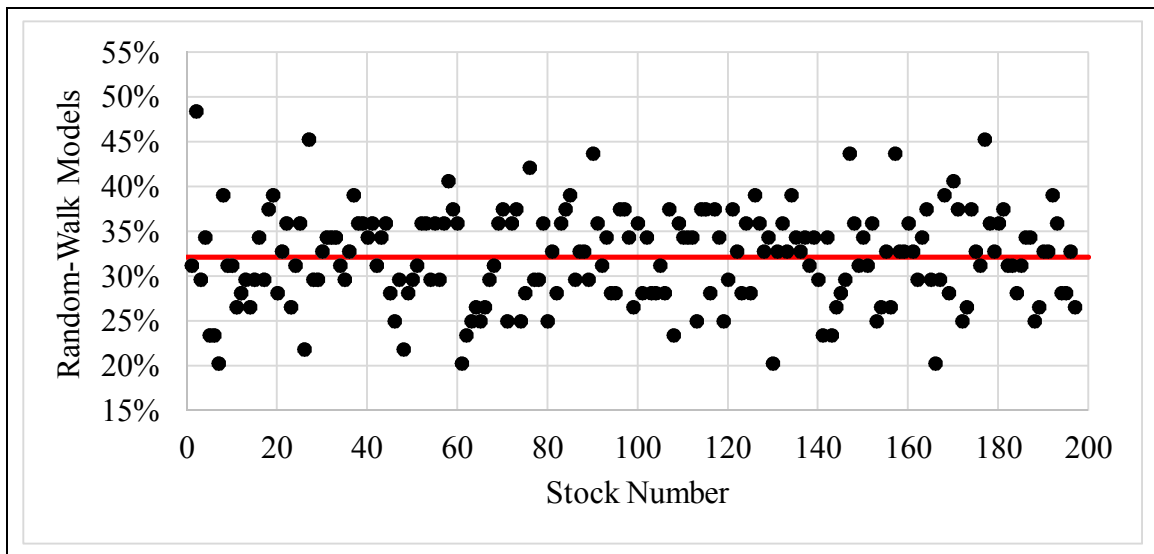
Figure 4-11. Frequency of Bivariate (PCA) Random-Walk Models

Results of serial correlation tests again advanced the perception that the spectral density test is more powerful than the Ljung-Box test as shown in Figure 4-12. While there is doubt promoted by the Ljung-Box test regarding data independence, the spectral density test clearly showed that model fits were very good, based on residual independence. The Bartlett kernel was used in the spectral density test, with all other parameters identical to previous configurations (see §2.5).
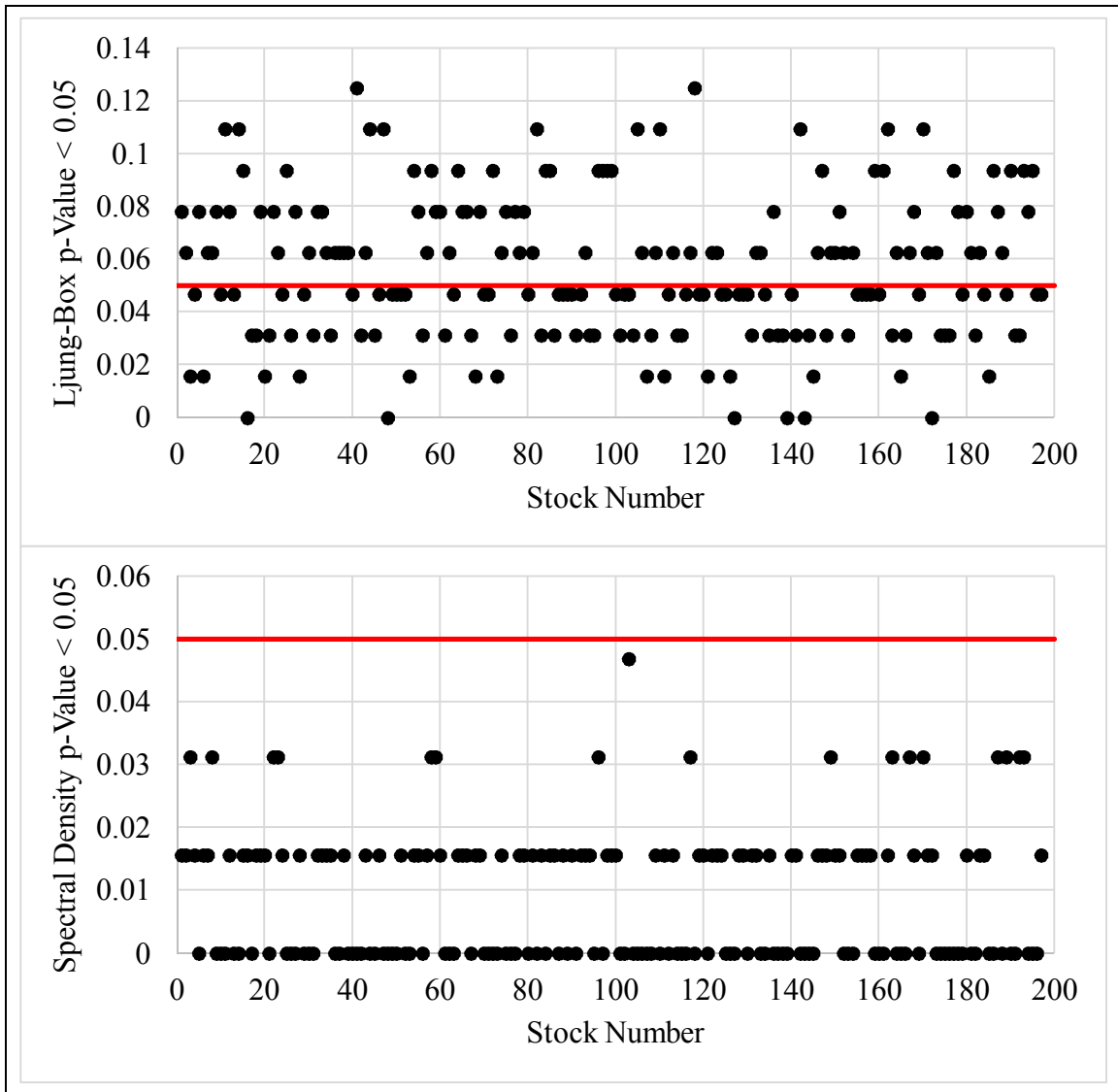
Figure 4-12. Serial Correlation for Bivariate (PCA) Model Residuals

Directional accuracy results for the bivariate-PCA models, given in Figure 4-13 and summarized in Table 4-7, showed significant improvements over the univariate ARIMA models. The best one-step forecast accuracies were 73.4% and 69.5% with 20 stocks performing better than 60%. For the five-step forecasts, there were 23 stocks with accuracy rates from 60–65%, but none greater than 64.1%. Augmenting the univariate with PCA yielded significant results to directional movement forecast accuracy.
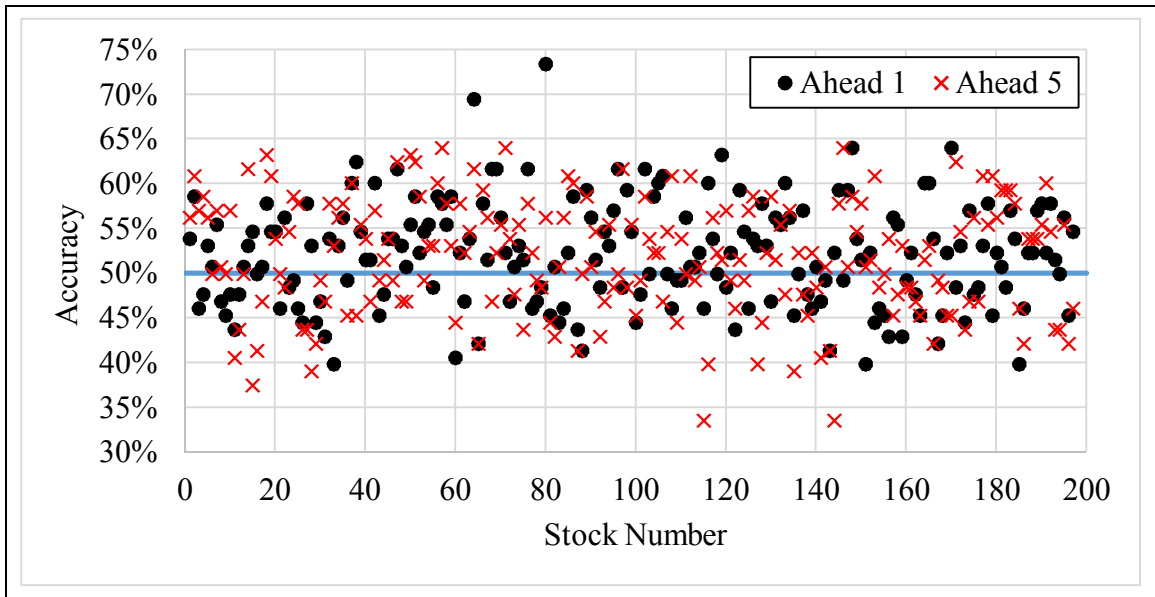
Figure 4-13. ARIMA Bivariate (PCA) Directional Forecast Accuracy

Table 4-7. ARIMA Univariate-Bivariate (PCA) Directional Accuracy

Forecast Interval	Accuracy	Frequency (Out of 197)	
		Univariate	Bivariate
Ahead 1	> 60%	2	20
	> 55%	24	59
	> 50%	93	120
	< 45%	8	20
	< 40%	0	3
Ahead 5	> 60%	1	23
	> 55%	18	69
	> 50%	93	114
	< 45%	22	30
	< 40%	2	7
Ahead 1 & Ahead 5	Same Direction	111	113

Bivariate-PCA confidence interval coverage provided 1% lower coverage compared with the univariate one-step forecasts, but showed an increase of 3.3% for the five-step intervals. Refer to Figure 4-14 and Table 4-8. Visually, there appeared to be a generally greater variation of the one-step coverage. These results are similar to the bivariate-volume outcomes.
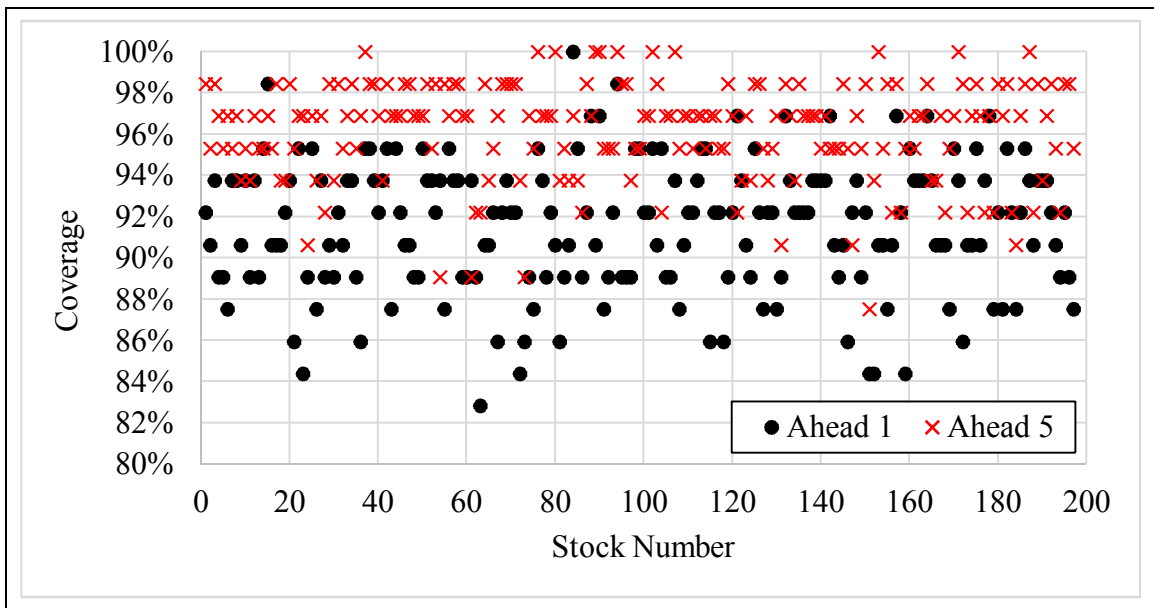
Figure 4-14. ARIMA Bivariate (PCA) Forecast Coverage within 95% CI

Table 4-8. ARIMA Univariate-Bivariate (PCA) Mean Coverage

	Univariate (%)	Bivariate (%)
Ahead 1	92.5	91.5
Ahead 5	92.8	96.1

Bivariate-PCA confidence interval shrinkage is shown in Figure 4-15, with Table 4-9 providing the mean values. Results are very similar to ARIMA bivariate-volume analyses.

125
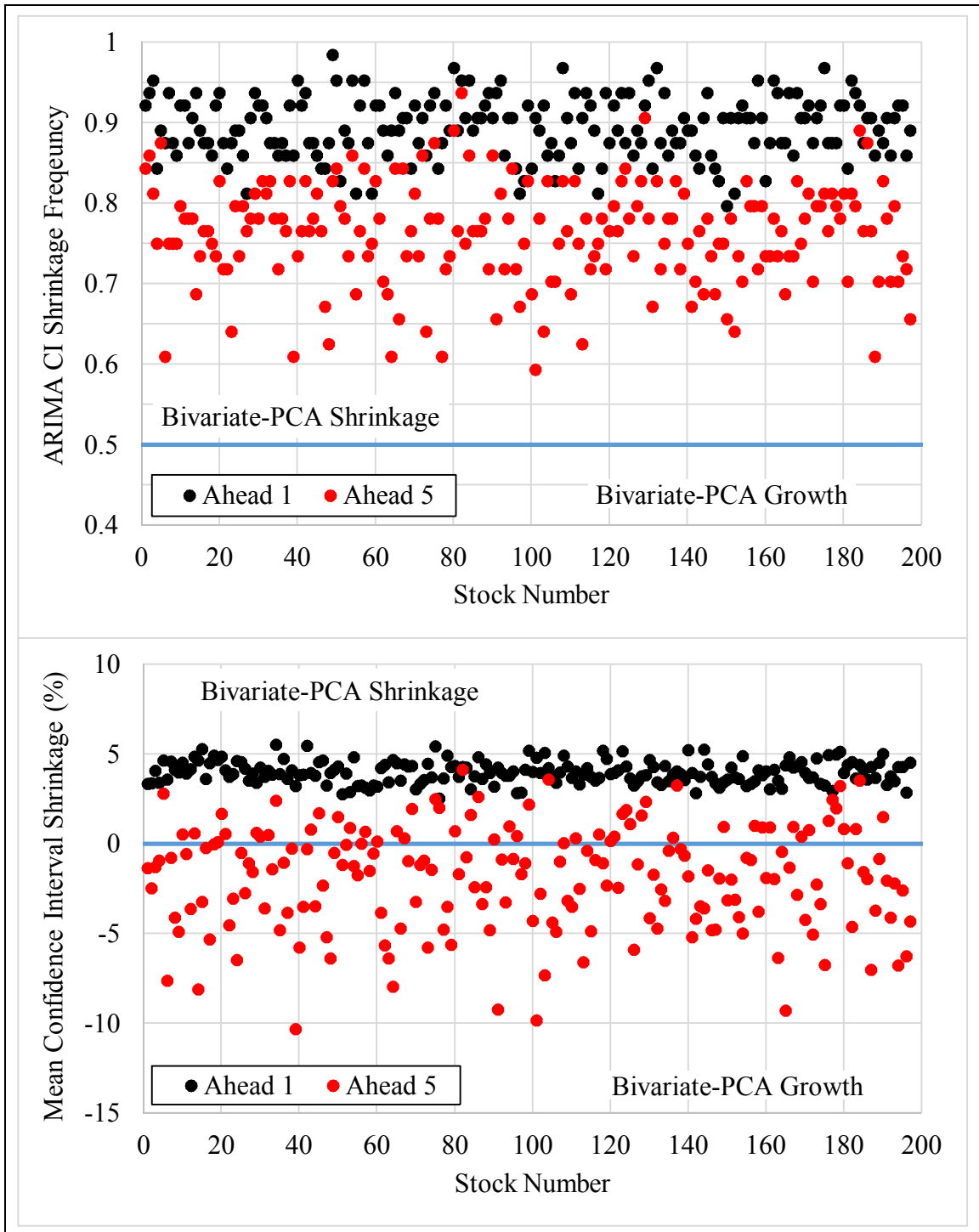
Figure 4-15. ARIMA Univariate-Bivariate (PCA) Confidence Intervals

Table 4-9. ARIMA Univariate-Bivariate (PCA) CI Statistics

Bivariate Shrinkage	Ahead 1 (%)	Ahead 5 (%)
Frequency	89.5	76.2
Mean Shrinkage	4.0	-1.9

As a final comment on the bivariate ARIMA-PCA investigation, all analyses (except for shrinkage) were repeated with data stationarizing of the greatest principal component after performing PCA on the non-stationary stock data. This was done to determine if any improvement was possible. (This sequence was in direct contrast to the preprocessing described heretofore in the chapter where PCA was performed on previously-stationarized data.) Associated results returned similar outcomes for all statistics, except forecasting accuracies, which were worse than those for PCA on the stationarized series.

4.2.2   SSA-PCA Model Results

One advantage of singular spectrum analysis is that there is no requirement for processed data to be stationary. Consequently, principal component analysis was applied to the stock-price data directly and the first principal component was found. Figure 4-16 shows the proportion of variance in each of the 64 independent windows. The mean proportion of variance was 65.4%, significantly greater than the comparable ARIMA statistic at 27.7%.
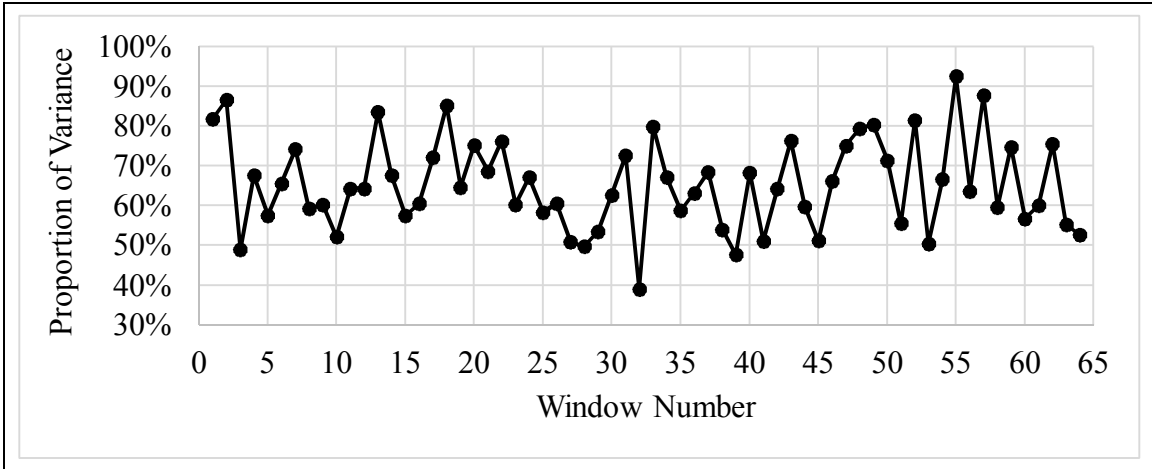
Figure 4-16. Proportion of Variance due to First Principal Component (SSA)

Applying PCA yielded a frequency of occurrence for the SSA simple bivariate model as depicted in Figure 4-17. The mean was 4.6%, compared to the SSA bivariate-stock models at 3.1% and SSA univariate models at 1.9%.
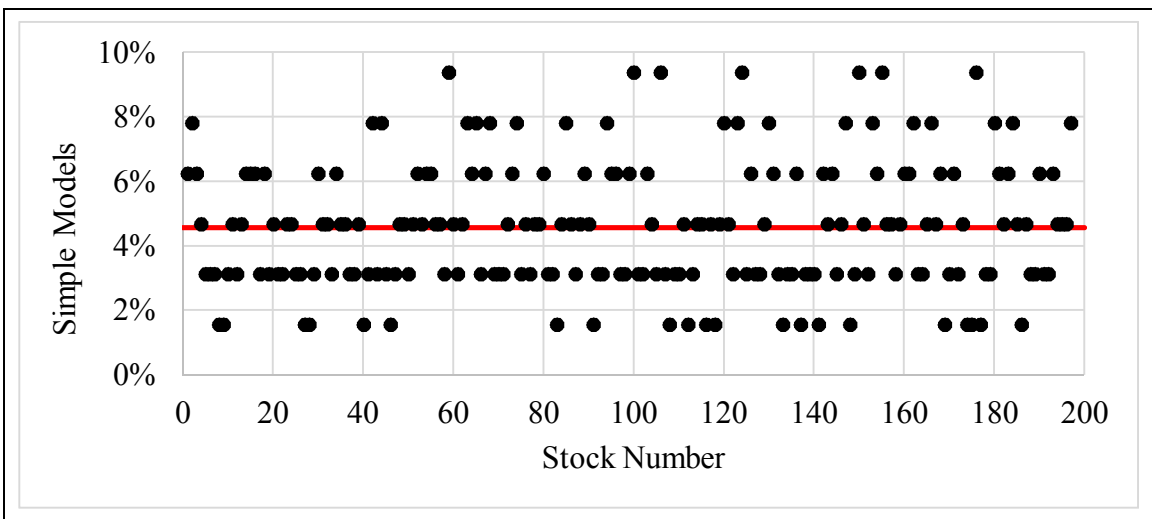
Figure 4-17. Frequency of Bivariate (PCA) SSA Simple Models

Directional forecast accuracy results, given in Figure 4-18 and summarized in Table 4-10, showed mixed results. One-day forecasts were better for the univariate models, while five-day forecasts were better for the bivariate-PCA models. Bivariate-PCA one-step predictions were as high as 64.1%, and five-day forecasts topped out at 68.8%.
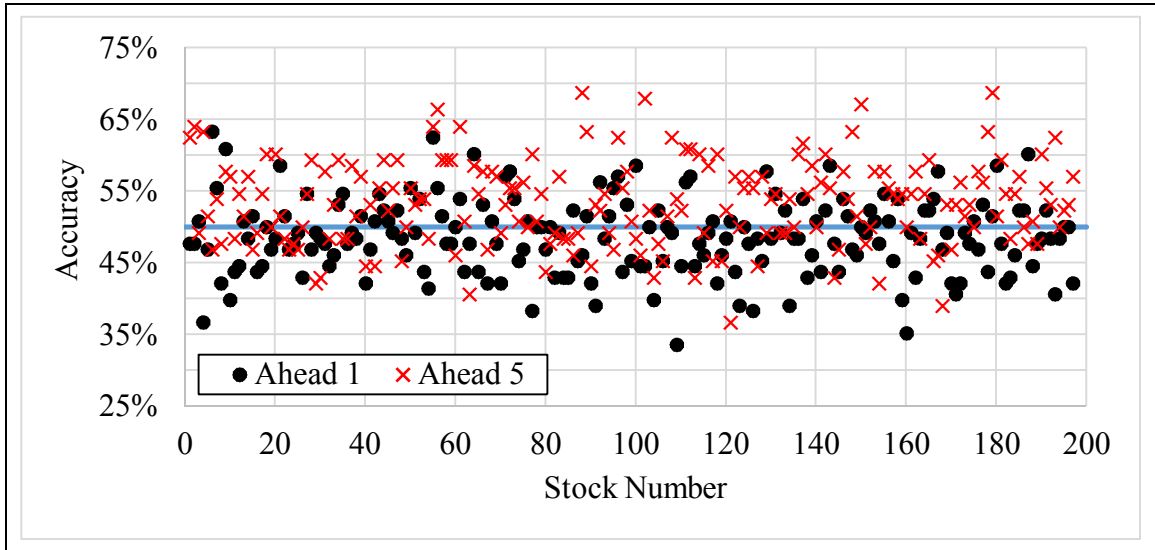
Figure 4-18. SSA Bivariate (PCA) Directional Forecast Accuracy

Table 4-10. SSA Univariate-Bivariate (PCA) Directional Accuracy

Forecast Interval	Accuracy	Frequency (Out of 197)	
		Univariate	Bivariate
Ahead 1	> 60%	13	5
	> 55%	47	21
	> 50%	102	69
	< 45%	40	49
	< 40%	9	11
Ahead 5	> 60%	13	27
	> 55%	44	75
	> 50%	97	126
	< 45%	41	14
	< 40%	11	2
Ahead 1 & Ahead 5	Same Direction	88	85

Figure 4-19 and Table 4-11 provide the 95% confidence interval coverage results for SSA bivariate-PCA. The bivariate-PCA results were 3.4% and 4.3% lower for one- and five-step confidence intervals, respectively. Coverage results appeared to have no correlation with forecast accuracy.
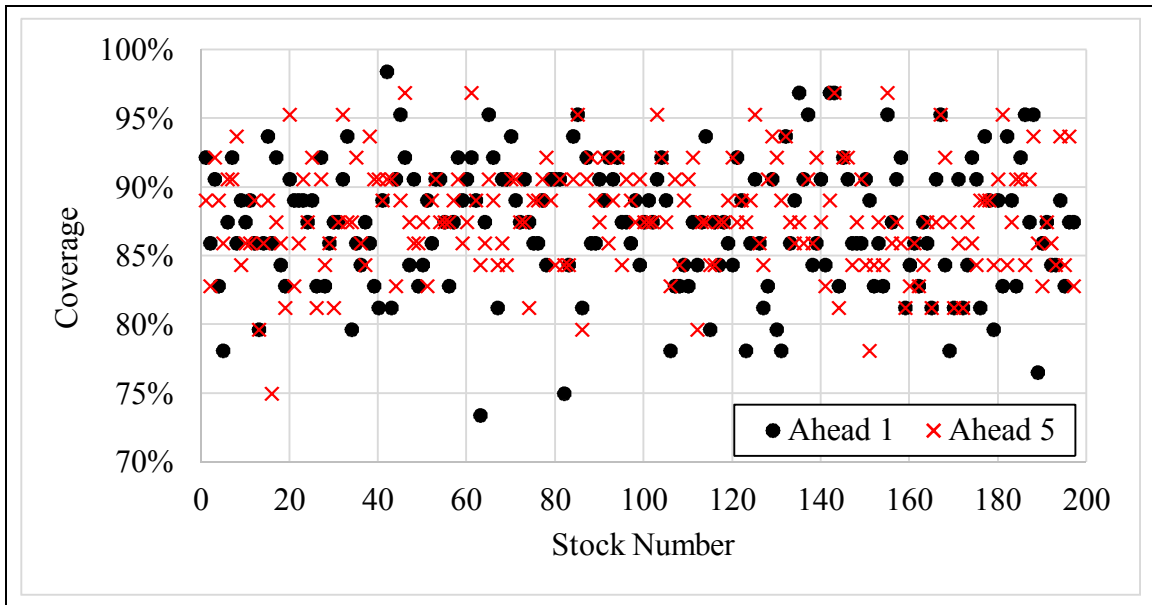
Figure 4-19. SSA Bivariate (PCA) Forecast Coverage within 95% CI

Table 4-11. SSA Univariate-Bivariate (PCA) Mean Coverage

	Univariate (%)	Bivariate (%)
Ahead 1	90.8	87.4
Ahead 5	92.1	87.8

Confidence interval shrinkage outcomes are shown in Figure 4-20 and Table 4-12. SSA bivariate-PCA revealed a mixed set of CI shrinkage. The frequency of shrinkage was just above 50% in frequency and between -20% and -25% in amount. These results showed that only a small percentage of the stocks produced a CI shrinkage, counting 18 and 25, for one- and five-step forecasts, respectively. For these stocks displaying CI reductions, the mean percent shrinkage was 6.3% and 7.9%, respectively.
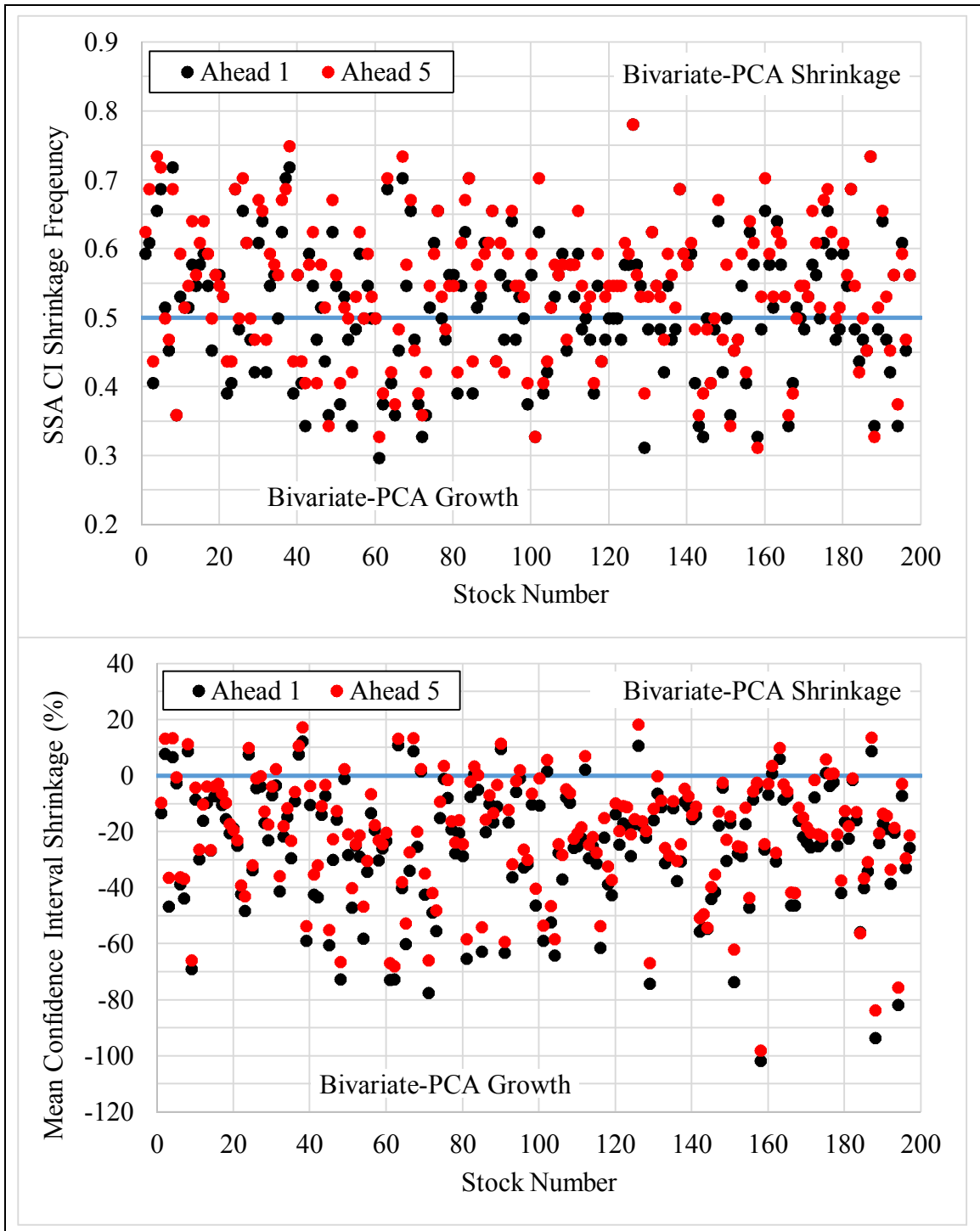
Figure 4-20. SSA Univariate-Bivariate (PCA) Confidence Intervals

Table 4-12. SSA Univariate-Bivariate (PCA) CI Statistics

Bivariate Shrinkage	Ahead 1 (%)	Ahead 5 (%)
Frequency	51.7	54.1
Mean Shrinkage	-25.3	-20.7

4.3 Volatility

In economic and finance literature, volatility is most often defined as a statistical measure of variation of logarithmic returns for a security (or index). Volatility is determined mathematically as either the standard deviation (σ) or the variance (σ^2), creating a source of confusion. In this study, volatility is defined as the standard deviation, since this is the predominant representation. A higher volatility signals a higher risk regarding a security's value, as the price can change substantially (up or down) over a short period of time, whereas a lower volatility indicates smaller short-term fluctuations without restricting longer term movement in either direction.

Volatility is not directly observable, but exhibits some notable characteristics. Values may appear in clusters, but mostly evolve continuously over time (jumps are unusual). Volatility fluctuates within some positive range, never diverging to infinity. Volatility displays a "leverage" effect, meaning that it reacts to a large price increase differently than to a large price reduction [41]. Economic and finance literature almost exclusively cites and applies volatility to assess risk. In stark contrast, this investigation simply employed volatility as a second variate to evaluate its effect on forecasting performance.

Volatility estimation may be established with either parametric or non-parametric models.

Models supporting generalized autoregressive conditional heteroscedasticity (GARCH) were introduced by T. Bollerslev in 1986 [42] as an enhancement of the autoregressive conditional heteroscedasticity (ARCH) model advanced by R. Engle in 1982 [43]. Both are parametric models that address the issue of non-constant volatility. In this study, initial attempts to compute volatility using a GARCH model yielded unusable results for 100-sample windows. Outputs produced negative eigenvalues[*] and minimum AIC[†] identified questionable volatility profiles. Subsequent experiments with windows of length as large as 1500 samples also showed similarly unusable results.

As an alternate to GARCH, a non-parametric approach for estimating volatility was conducted using one of the so-called "extreme-value" variance estimators [44]. To derive daily volatility, these estimators utilize an asset's open, high, low, and close (O,H,L,C) price values. The effectiveness of the extreme-value volatility estimators was investigated and validated in [44] by comparing extreme-value variance estimates with high-frequency intraday data. A high-level summary of characteristics for extreme-value estimators is provided in Table 4-13 (in chronological order of development).

---

[*] Valid eigenvalues are non-negative.

[†] Akaike Information Criterion (AIC) was used to find the best model fit from a search over 300 possible models for each window. Models consisted of both a GARCH(α,β) order for volatility and an ARMA(p,q) order for the mean model. Refer to R-Package 'rugarch' version 1.3-6 and [51] for details.

Table 4-13. Extreme-Value Variance Estimators

Estimator Name	Drift Independent	Considers Opening Jumps	Special Features
Close-to-Close (σ_{CC}^2)	No	No	Traditional
Mean-Adjusted Close-to-Close	No	No	Adjustment for mean
Parkinson (σ_P^2)	No	No	≈5X improved efficiency[*]
Garman-Klass (σ_{GK}^2)	No	No	Minimum variance; ≤ 7.4X more efficient
Rogers-Satchell (σ_{RS}^2)	Yes	No	Drift independence; efficiency comparable to σ_{GK}^2 [45]
Yang-Zhang (σ_{YZ}^2)	Yes	Yes	Minimum variance; considers overnight jumps; efficiency ≤ 14X [46]

The Yang-Zhang estimator was selected and applied to estimate volatility because it included utilities for handling drift and close-to-open jumps, and demonstrated high efficiency. The Yang-Zhang estimator is defined as

$$\sigma_{YZ}^2 = \frac{1}{N-1}\sum_{t=1}^{N}\left(\ln(\frac{O_t}{C_{t-1}})-\overline{o}\right)^2 + k\frac{1}{N-1}\sum_{t=1}^{N}\left(\ln(\frac{C_t}{O_t})-\overline{c}\right)^2 + (1-k)\sigma_{RS}^2 \ , \ n>1, \quad (4\text{-}4)$$

---

[*] Efficiency is defined for estimation variance relative to the traditional Close-to-Close estimator σ_{CC}^2.

where the Rogers-Satchell variance is given as

$$\sigma_{RS}^2 = \frac{1}{N}\sum_{t=1}^{N}\left(\ln\left(\frac{H_t}{C_t}\right)\ln\left(\frac{H_t}{O_t}\right) + \ln\left(\frac{L_t}{C_t}\right)\ln\left(\frac{L_t}{O_t}\right)\right), \ n \geq 1,$$ (4-5)

with $\bar{o} = \frac{1}{N}\sum_{t=1}^{N}\ln\left(\frac{O_t}{C_{t-1}}\right)$, $\bar{c} = \frac{1}{N}\sum_{t=1}^{N}\ln\left(\frac{C_t}{O_t}\right)$, $k = 0.34\Big/\left(1.34 + \frac{N+1}{N-1}\right)$, and N is the

number of samples over which to average.[*] Volatility estimation was computed with

$N = 2$, wherein highest efficiency is reached [46].

4.3.1 ARIMA-Volatility Model Results

Volatility estimates were computed for each window of each stock using the function

volatility in the R-Package 'TTR' version 0.23-1. Subsequent to computing values,

differencing was applied with the *ndiffs* function (verified with a KPSS test) to stationarize

the resultant volatility vector for compatibility with ARIMA processing.[†] Bivariate

ARIMA computations were performed using the same method described in §2.4.

---

[*] Note also that X_t refers to the value of X at time sample t.

[†] Additionally, a distinctive aspect of the *volatility* function in 'TTR' is that it returns NaN (not a number) values for the first N samples in the vector. To accommodate as many as two orders of differencing, the second value in the volatility vector was computed with the Rogers-Satchell two-sample estimate and the first value was computed with the Rogers-Satchell one-sample estimate. Moreover, under limited conditions, e.g., when a stock's price is constant for one or more samples, a small negative variance can ensue, resulting in a NaN output value. To manage this infrequent but problematic occurrence, a NaN volatility value was replaced with a small but non-zero value of 10^{-8}.

The frequency of random-walk models seen in the graph of Figure 4-21 showed another order of magnitude reduction from the bivariate volume results (two orders of magnitude below the univariate models). The mean was 0.5%, with a maximum of three random-walk models detected in any stock-volatility pair.
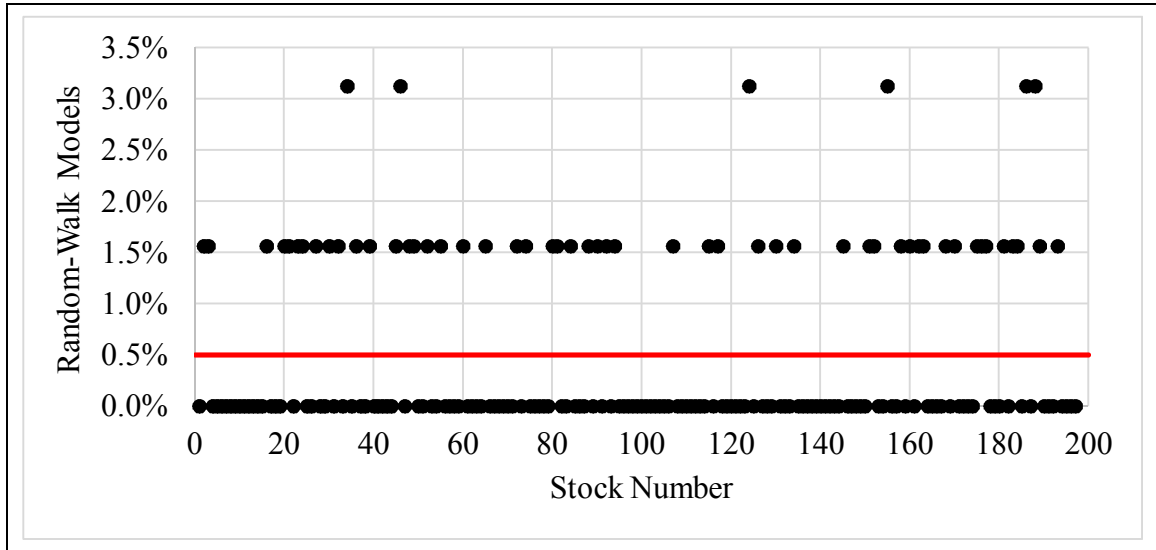
Figure 4-21. Frequency of Bivariate (Volatility) Random-Walk Models

Serial correlation results (Figure 4-22) for the Ljung-Box test showed only seven stocks with p-values above 0.05. Here, similar to the bivariate-volume outcome, the more powerful spectral density test produced a larger number above 5%, with a count of 35 and a maximum of 10.9%. The spectral density test was performed with the same parameters as specified in §2.5, with a Bartlett smoothing kernel as mentioned previously. While these results indicated the possibility of additional useful information contained in the model residuals, the directional accuracy results showed significant increases compared to the univariate evaluations.
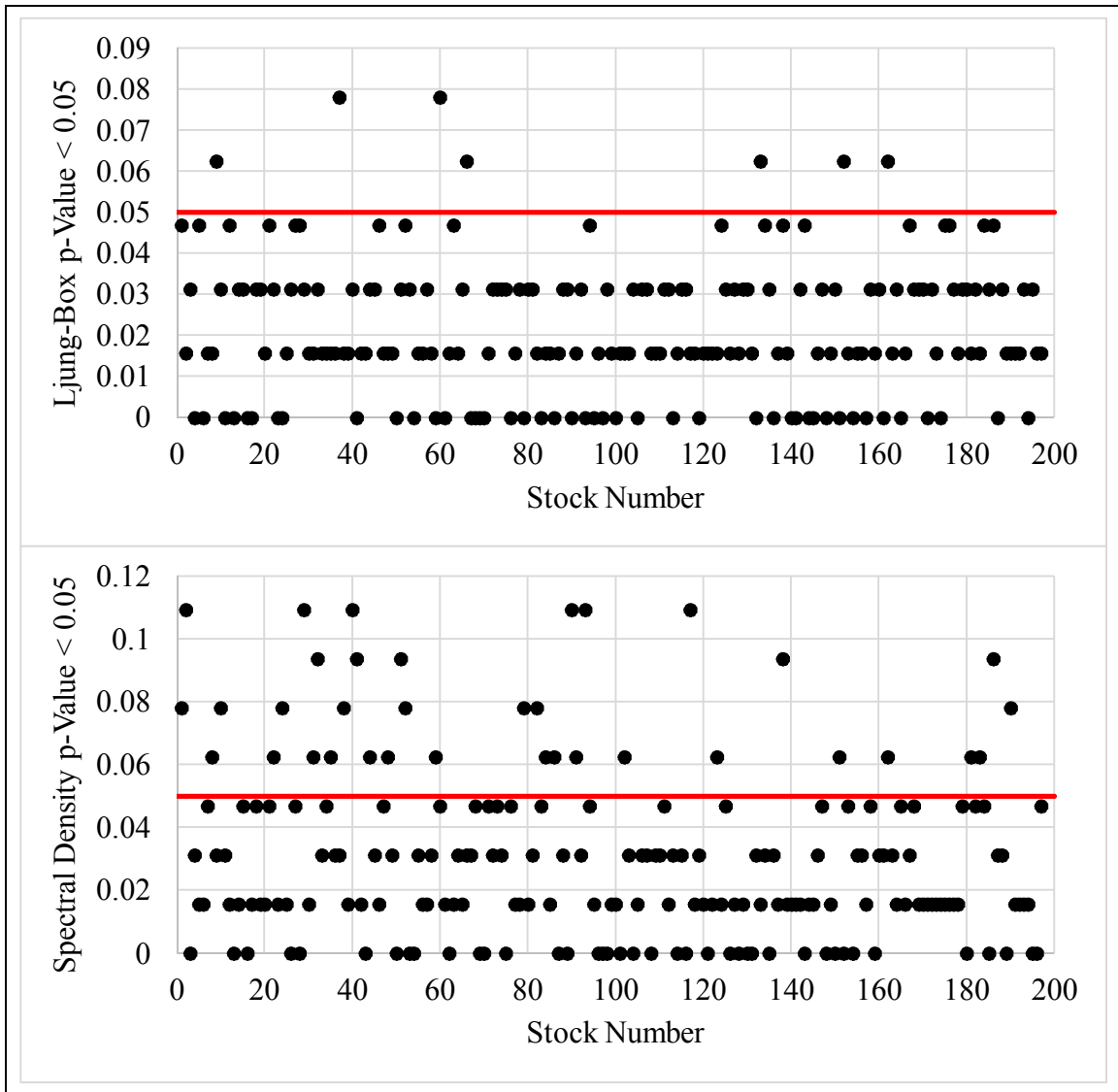
Figure 4-22. Serial Correlation p-Values for Bivariate (Volatility) Residuals

Figure 4-23 and Table 4-14 show directional accuracy results for the bivariate-volatility investigation. Four stocks showed five-step forecast accuracy of 68% or better, with two stocks performing above 70%. Overall effects were significantly better than the univariate products as seen in the summary table. These results demonstrated that working only with volatility estimates yielded a strong enhancement to directional forecasting.
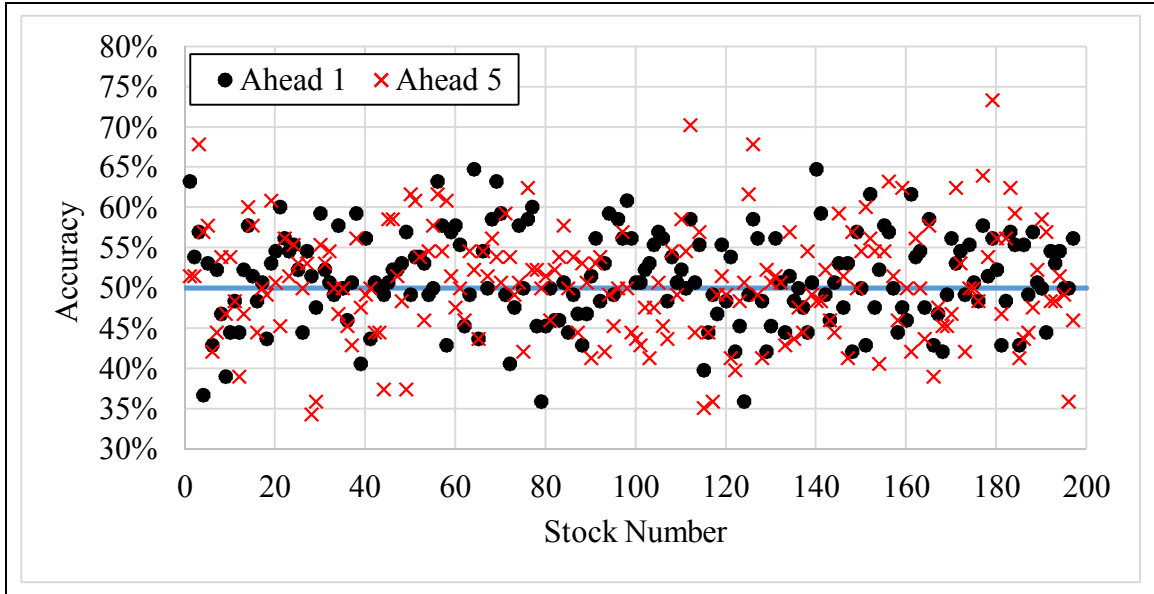
Figure 4-23. ARIMA Bivariate (Volatility) Directional Forecast Accuracy

Table 4-14. ARIMA Univariate-Bivariate (Volatility) Directional Accuracy

Forecast Interval	Accuracy	Frequency (Out of 197)	
		Univariate	Bivariate
Ahead 1	> 60%	2	10
	> 55%	24	55
	> 50%	93	109
	< 45%	8	30
	< 40%	0	5
Ahead 5	> 60%	1	18
	> 55%	18	45
	> 50%	93	98
	< 45%	22	42
	< 40%	2	10
Ahead 1 & Ahead 5	Same Direction	111	101

Coverage analysis results, shown in Figure 4-24 and Table 4-15, revealed that five-step forecast coverage was better than one-step coverage; however, both were above 90%. Values were comparable to univariate coverage as the table revealed. The distributional plot was similar to bivariate-volume and bivariate-PCA coverage plots (respectively, Figure 4-4 and Figure 4-14).
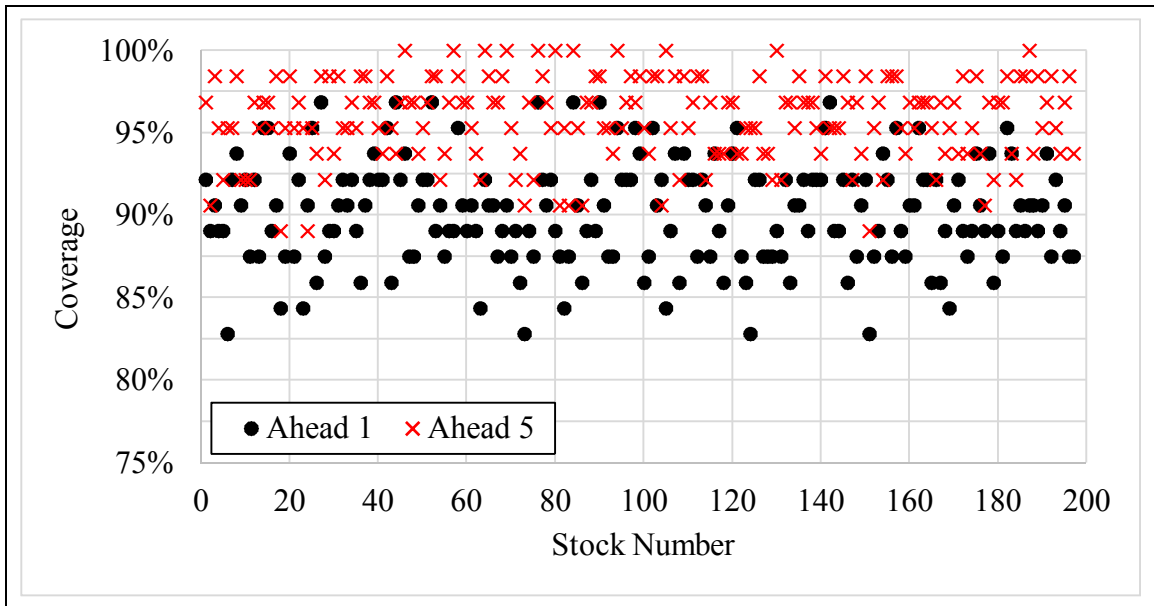
Figure 4-24. ARIMA Bivariate (Volatility) Forecast Coverage within 95% CI

Table 4-15. ARIMA Univariate-Bivariate (Volatility) Mean Coverage

	Univariate (%)	Bivariate (%)
Ahead 1	92.5	90.2
Ahead 5	92.8	95.8

Results of confidence interval shrinkage, seen in Figure 4-25 and Table 4-16 yielded consistent reduction of one-step forecast intervals in both frequency and mean value.
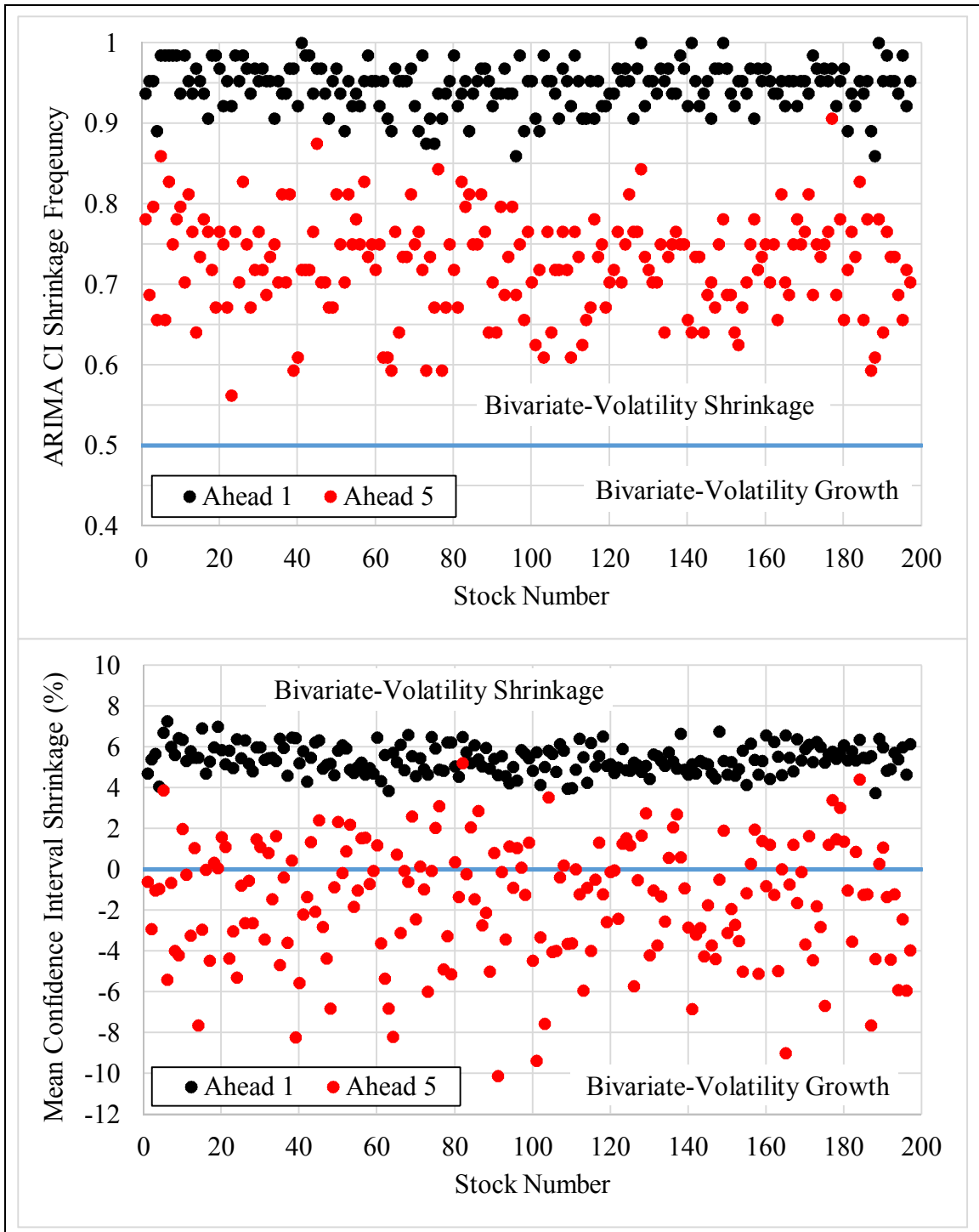
Figure 4-25. ARIMA Univariate-Bivariate (Volatility) Confidence Intervals

Five-step confidence intervals were also dominated by shrinkage as seen in the figure, but the mean value of reduction was -1.5% which equates to growth, not shrinkage. There were, however, 64 stocks with positive mean shrinkage, yielding a mean of +1.5%.

Table 4-16. ARIMA Univariate-Bivariate (Volatility) CI Statistics

Bivariate Shrinkage	Ahead 1 (%)	Ahead 5 (%)
Frequency	94.6	72.5
Mean Shrinkage	5.4	-1.5

4.3.2   SSA-Volatility Model Results

For singular spectrum analysis, volatility was computed using the Yang-Zhang variance estimator defined in §4.3. According to the SSA process, no differencing was needed for either stock price or volatility values. Frequency of occurrence for simple SSA models seen in Figure 4-26 yielded a mean of 1.8%, slightly better than the 1.9% for univariate models. Using volatility as a second variate appeared to avoid any confounding factors regarding separating structures in SSA eigenvector space.
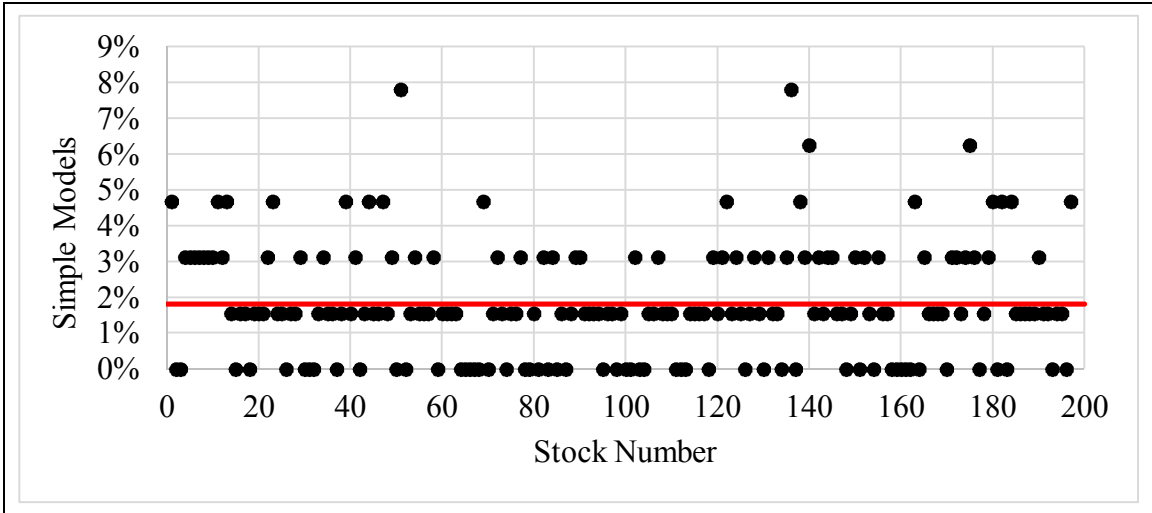
Figure 4-26. Frequency of Bivariate (Volatility) SSA Simple Models

Directional accuracy results for SSA bivariate-volatility are provided in Figure 4-27. Maximum accuracy was attained for a five-step forecast at 69.5%, with three others above 65%. The maximum for one-step forecasts reached 64.8%. The summary displayed in Table 4-17 shows similar results between bivariate-volatility and univariate models.
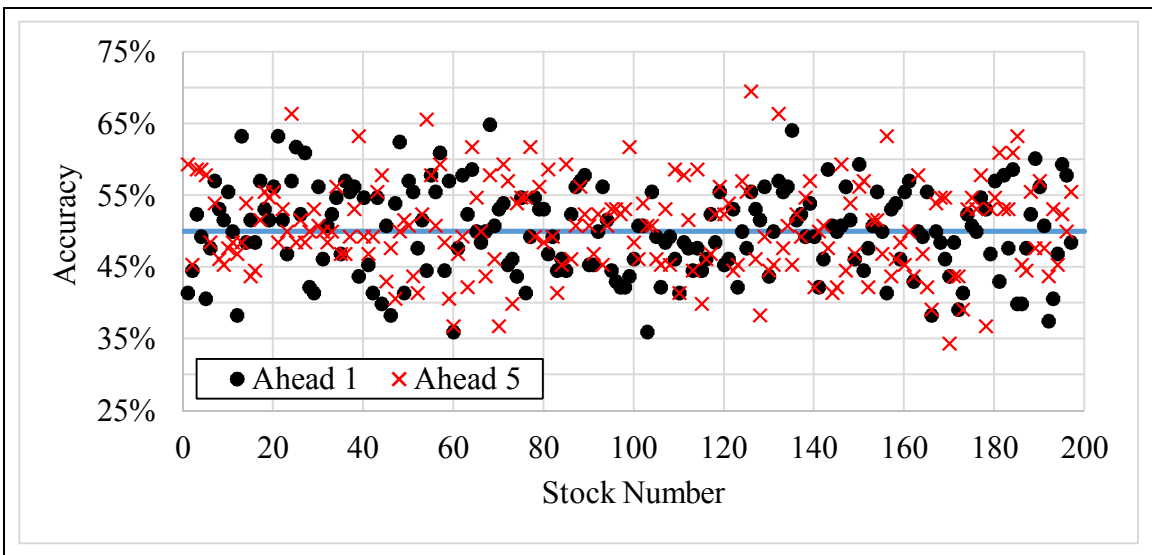
Figure 4-27. SSA Bivariate (Volatility) Directional Forecast Accuracy

143

Table 4-17. SSA Univariate-Bivariate (Volatility) Directional Accuracy

Forecast Interval	Accuracy	Frequency (Out of 197)	
		Univariate	Bivariate
Ahead 1	> 60%	13	9
	> 55%	47	49
	> 50%	102	96
	< 45%	40	43
	< 40%	9	10
Ahead 5	> 60%	13	12
	> 55%	44	45
	> 50%	97	98
	< 45%	41	35
	< 40%	11	9
Ahead 1 & Ahead 5	Same Direction	88	91

Confidence interval coverage for SSA bivariate-volatility showed mean values proximate to 76% for both one- and five-step forecasts. A plot is given in Figure 4-28. While the coverage was substantially lower compared to the univariate models as summarized in Table 4-18, accuracies were similar as seen in Table 4-17.
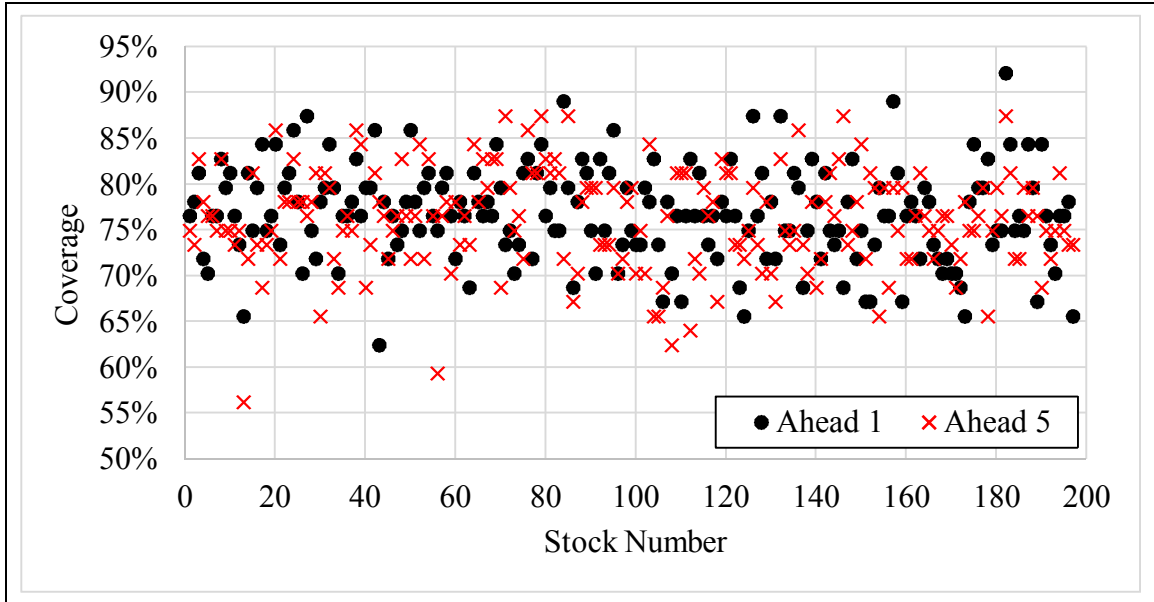
Figure 4-28. SSA Bivariate (Volatility) Forecast Coverage within 95% CI

Table 4-18. SSA Univariate-Bivariate (Volatility) Mean Coverage

	Univariate (%)	Bivariate (%)
Ahead 1	90.8	76.8
Ahead 5	92.1	76.1

Confidence interval shrinkage for SSA bivariate-volatility was consistent and significant across all stocks, without exception. Frequency of shrinkage averaged near 98% and amount of shrinkage averaged approximately 38%. Coverage was better than that for SSA bivariate-volume. SSA bivariate-volatility displayed exceptionally strong characteristics.

145
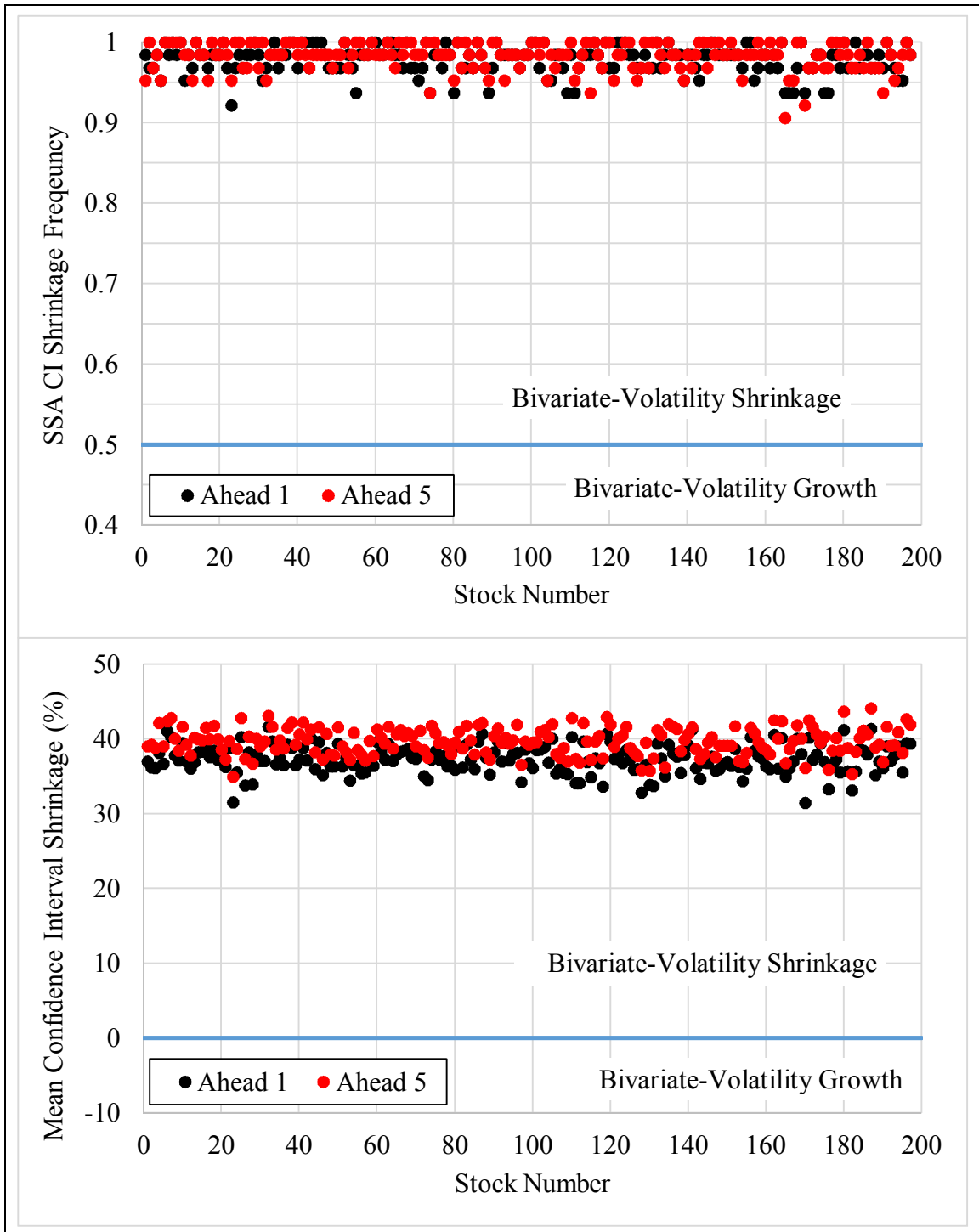
Figure 4-29. SSA Univariate-Bivariate (Volatility) Confidence Intervals

Table 4-19. SSA Univariate-Bivariate (Volatility) CI Statistics

Bivariate Shrinkage	Ahead 1 (%)	Ahead 5 (%)
Frequency	97.9	98.2
Mean Shrinkage	37.3	39.7

4.4 Summary of Alternate Bivariate Components

This chapter comprised a study of hypothesized time-series forecasting performance improvements utilizing bivariate models with covariates other than stock. Volume, greatest principal component from PCA, and volatility were processed as the second variate. Overall results pointed toward the method of principal component analysis to produce the highest forecasting accuracies. Table 4-20 provides a compendium of study outcomes.

Bivariate ARIMA models consistently outperformed their univariate counterpart, but such regularity was not found for bivariate SSA. Univariate SSA-based one-step forecasts generally surpassed any of the bivariate-SSA alternatives, but for five-step projections, SSA-based bivariate-PCA performed best in the entire field of both SSA and ARIMA trials. The highest performers are bolded in Table 4-20; these included bivariate ARIMA-PCA one-step- and SSA-PCA five-step forecast accuracies for $> 55\%$ and $> 60\%$ categories.

147

Table 4-20. Bivariate Performance Comparison

Parameter			ARIMA*					SSA*				
			Univ	S1 | S2	Vlm	PCA	Vty	Univ	S1 | S2	Vlm	PCA	Vty
Random-Walk | Simple Model			63.0	34.7	5.6	32.1	0.5	1.9	3.1	4.5	4.6	1.8
Directional Accuracy	Ahead 1	> 65%	0	**2.4 | 3.5**	0.5	1.0	0	0.5	1.2 | 0	1.0	0	0
		> 60%	1.0	9.4 | 9.4	7.1	**10.2**	5.1	6.6	3.5 | 4.7	6.6	2.5	4.6
		> 55%	12.2	29.4 | 35.3	28.4	**29.9**	27.9	23.9	20.2 | 23.5	18.8	10.7	24.9
		> 50%	47.2	58.8 | 67.1	51.8	60.9	55.3	51.8	55.3 | 49.4	49.7	35.0	48.7
		< 45%	4.1	14.1 | 8.2	17.8	10.2	15.2	20.3	12.9 | 17.6	18.3	24.9	21.8
	Ahead 5	> 65%	0	1.2 | 1.2	2.0	0	2.0	1.0	0 | 1.2	1.0	**2.5**	2.0
		> 60%	0.5	8.2 | 11.8	6.6	11.7	9.1	6.6	3.5 | 9.4	6.6	**13.7**	6.1
		> 55%	9.1	34.1 | 28.2	25.9	35.0	22.8	22.3	12.9 | 31.8	18.3	**38.1**	22.8
		> 50%	47.2	56.5 | 54.1	53.8	57.9	49.7	49.2	47.1 | 56.5	49.7	64.0	49.7
		< 45%	11.2	23.5 | 16.5	15.2	15.2	21.3	20.8	18.8 | 22.4	18.8	7.1	17.8
CI Coverage – Ahead 1			92.5	91.1 | 90.9	90.6	91.5	90.2	90.8	83.3 | 82.9	76.3	87.4	76.8
CI Coverage – Ahead 5			92.8	96.5 | 95.8	95.8	96.1	95.8	92.1	83.8 | 83.8	76.5	87.8	76.1
CI Mean Shrinkage – Ahead 1			–	4.1 | 4.2	4.7	4.0	5.4	–	22.9 | 22.7	34.5	-25.3	37.3
CI Mean Shrinkage – Ahead 5			–	-2.1 | -1.3	-1.7	-1.9	-1.5	–	25.8 | 25.9	36.4	-20.7	39.7

* All values in percent

Notation: Univ = Univariate; Bivariate: S1 | S2 = Stock 1 & Stock 2, Vlm = Volume, Vty = Volatility

An extra accuracy category of > 65% was incorporated in an attempt to isolate the best performers. While this helped solidify the five-step forecasts for SSA-PCA, it revealed the bivariate-stock model with the highest accuracy percentage for the one-step predictions. It should be recalled, however, that there were 85 stock pairs considered versus 197 PCA-related stocks; each model category showed only two forecasts > 65%.

Table 4-20 summarizes the most significant findings of the forecasting investigations contained in this dissertation. In conclusion, it was evident that accuracy results for one- and five-steps-ahead imply that no single analysis method dominated forecasting in both time horizons.[*]

---

5. ECONOMIC SIGNIFICANCE & ROBUSTNESS – BACKTESTING

"The only relevant test of the validity of a hypothesis is comparison of prediction with experience."

– Milton Friedman (1912-2006), in the book "Essays in Positive Economics," 1953

This penultimate chapter reports on an investigation of economic significance and robustness within a limited configuration for the two ostensibly-best techniques identified in the previous chapter, i.e., bivariate ARIMA-PCA and bivariate SSA-PCA. These two methods demonstrated the best performance potential based on results from all previous studies. The meaning of economic significance implies real-world consequences, connoting a change in value of an asset (after some passing of time), and considers both magnitude and sign. Robustness refers to the ability of a method to perform well under various market conditions such as up-trends, down-trends, and sideways (also called range-bound) patterns.*

As in previous chapters, all tests were conducted using out-of-sample forecasts. Economic significance was evaluated by "back-testing" the above-mentioned methods using historical data to determine gain and loss in simulated trading accounts. To establish a meaningful back-testing scenario, all initial account sizes were set at $10,000 to allow a realistic number of shares to be traded and create a feasible starting setup for most traders.

---

* Robustness could also concern modifications of model parameters, but these were not included in this investigation, hence a "limited configuration."

Robustness was assessed by applying the forecasting techniques to five recent historical market conditions highlighted in Figure 5-1. Each of these five time periods spanned 254 daily samples (approximately one calendar year) and is specified in Table 5-1. As seen in Figure 5-1, various trend patterns were explored.
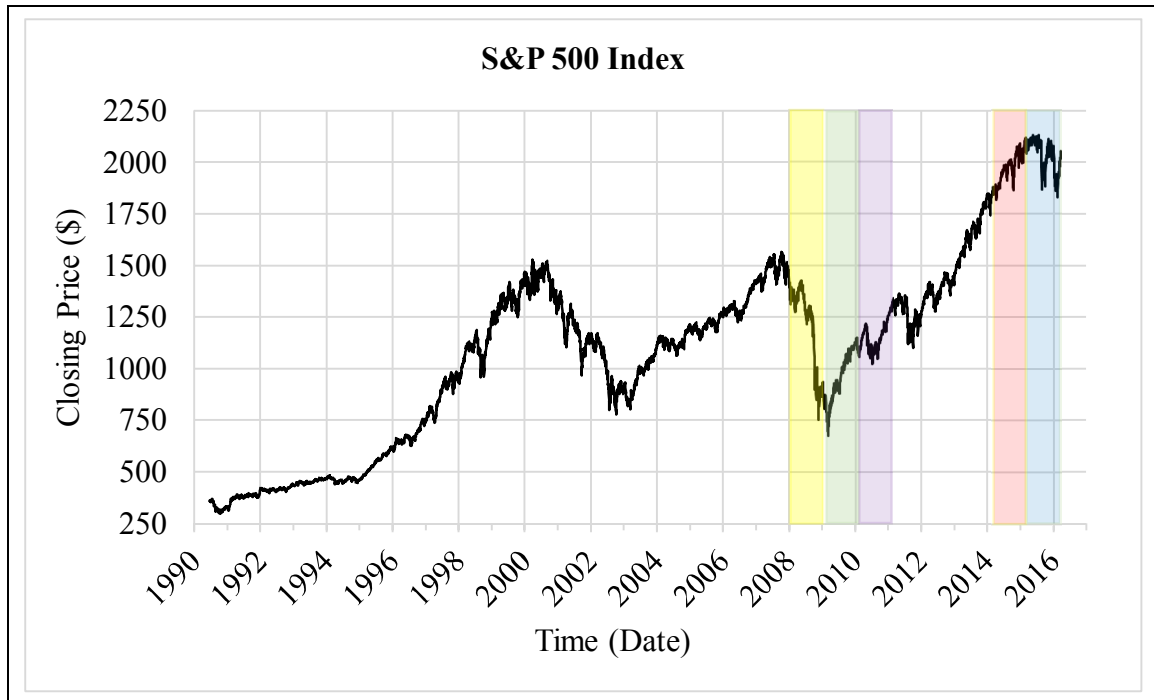
Figure 5-1. Study Intervals

Table 5-1. Back-Testing Intervals

Color	Interval Name	Start Date	End Date
Blue	2015	3-10-2015	3-10-2016
Red	2014	3-6-2014	3-9-2015
Violet	2010	1-6-2010	1-6-2011
Green	2009	1-2-2009	1-5-2010
Yellow	2008	12-31-2007	12-31-2008

151

5.1 Test Specifications

Back-testing evaluations were constructed as follows. For each of the specified test intervals defined in Table 5-1, one- and five-step out-of-sample-forecast models were computed for each day in the interval for each stock in the stock list (see Appendix 8.1). Forecast models were used to compute forecast directions and associated magnitudes. Based on the available account balance, a maximum-integer number of shares for each transaction was determined. A net change was computed as the number of shares multiplied by the change in value between the closing price on the forecast reference day t and the actual closing price on day $t + 1$, or in the case of a five-step forecast, the closing price on day $t + 5$. Account balance was updated by the net change multiplied by the sign of the forecast direction. Trade entries for positive forecasts were long (buy) while trade entries for negative forecasts were short (sell). Of course, trade exits were opposite to entries. Account balances were maintained for one- and five step forecast intervals, with and without transaction fees. Transaction fees, i.e., commissions, were instituted at \$4.95 for trade entry and the same for exit, totaling \$9.90 for each complete transaction.[*] For the five-step tests, the subsequent trade was delayed one day in an attempt to avoid any day-of-the-week bias. In a second configuration of tests, the five-step forecasts were repeated daily for five steps ahead, producing a set of shifted-overlapping intervals every five samples, to determine if there were any noteworthy findings through periodicity.[†] An additional test arrangement was included in which a threshold was invoked whereby a trade

---

[*] These amounts were found to be reasonable and readily available from multiple brokers.

[†] Naturally, day-of-the-week repetition only held until a non-trading day (i.e., a market holiday) occurred during a normal five-day week. No attempt was made to track or correct for such atypical occurrences.

entry would only proceed if the threshold was exceeded. This was evaluated to determine if a higher rate of return was attainable.

5.2 Individual Stock Tracking Results

Results of testing all stocks independently with bivariate models for ARIMA-PCA and SSA-PCA yielded distributions of accuracies similar to those found in previous chapters. As seen in the examples of Figure 5-2, no consistent, repeated, performance was observable in sequential years. Similarly inconsistent results were obtained from the other three year-long test intervals. These results contrasted sharply with the methodology that assumed that accurate predictability was of paramount importance to achieving positive economic significance, i.e., profit. A brief summary of outcomes follows.

Results of tracking account balances for interval 2015 are provided in Figure 5-3. The top plot shows final balances without including transaction fees and the middle chart shows terminal balances including fees. The bottom graph shows the results of including thresholding in deciding whether to enter a trade. Clearly, including fees substantially impacted any profits, especially for the one-step forecasts because there were five times the number of fees incurred for the five-step profiles. Other years showed similarly random results, some with higher ending balances for a few stocks, but no consistency. Deciding to enter a trade based on whether the forecast value exceeded a threshold (several were tested) yielded degraded performance at all trials compared to the no-threshold scenario. In an effort to understand these results, a correlation was compiled for account balance vs.

accuracy. As seen in Figure 5-4, there was somewhat of a correlation exposed, but there remained sufficient ambiguity to be of little use in clarifying apparently random results.
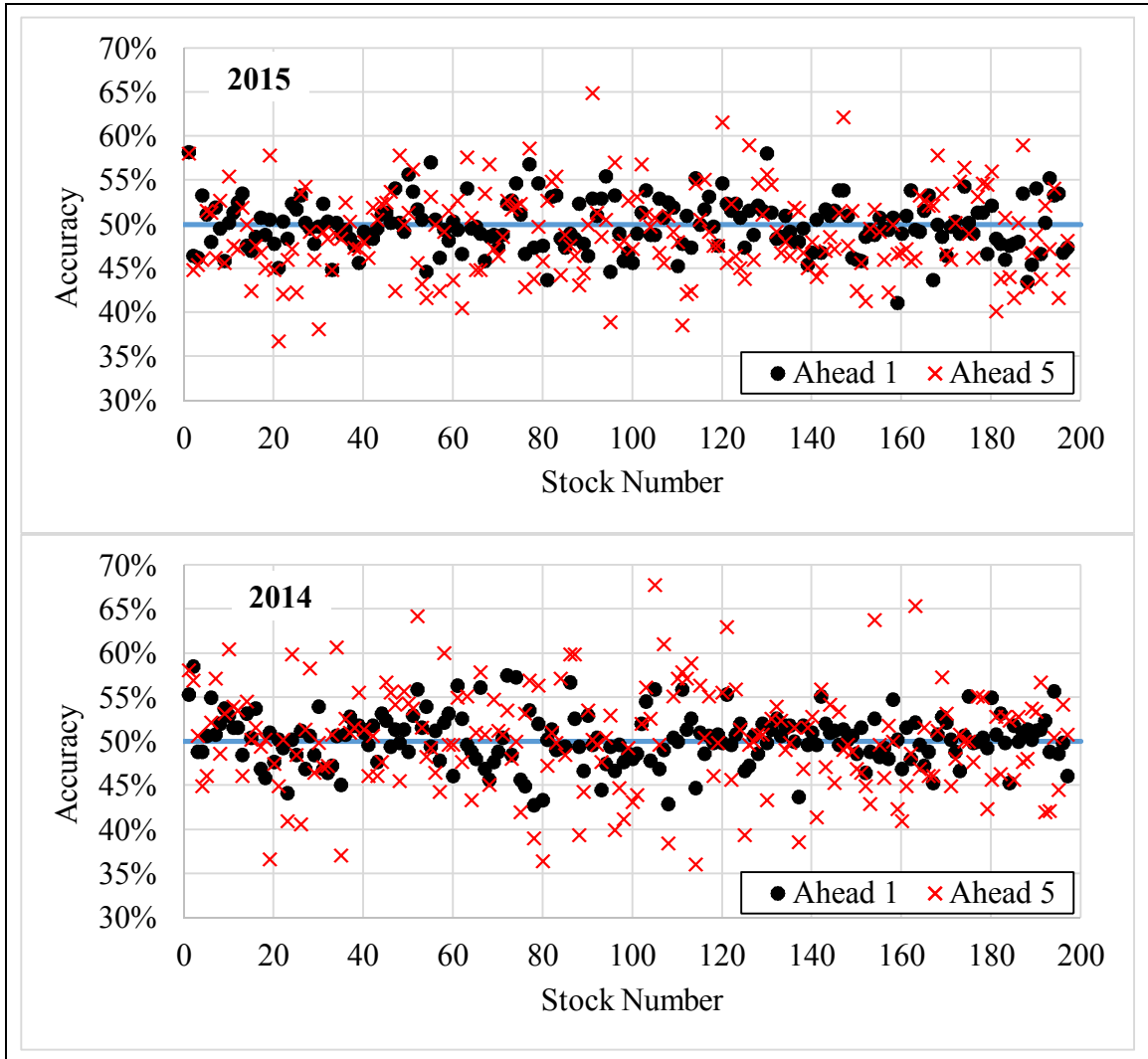
Figure 5-2. ARIMA-PCA Directional Forecast Accuracy
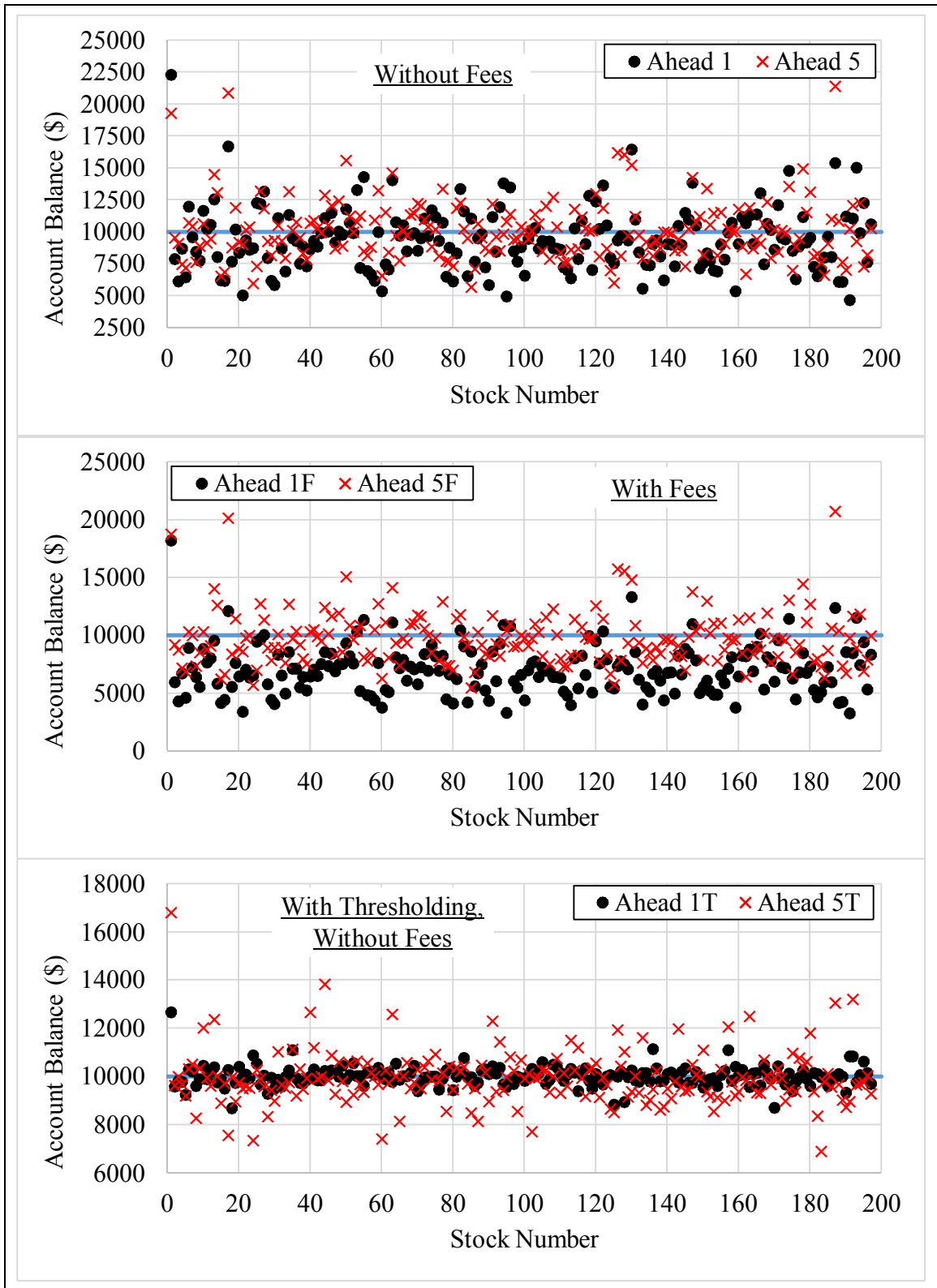
Figure 5-3. ARIMA-PCA Account Balance Totals (2015)

155
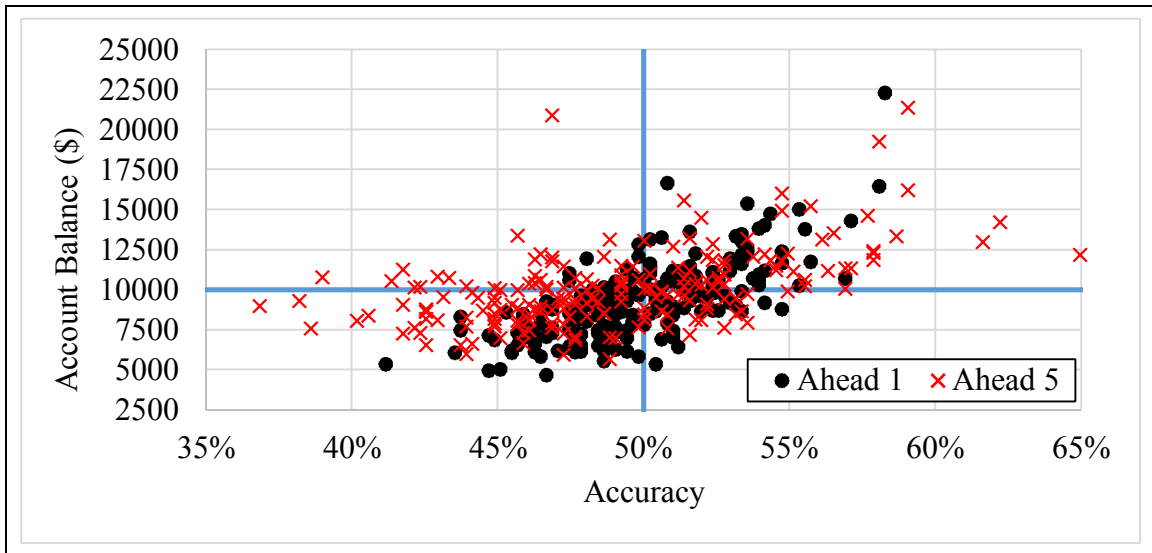
Figure 5-4. ARIMA-PCA Account Balance-Accuracy Correlation (2015)

Although many additional plots were compiled to evaluate results of independent stock tracking, they were not included in this document since they provided no additional insight into the investigational outcomes. The plots displayed in this section were representative of the excluded results. Moreover, no noteworthy differences were observed for bivariate SSA-PCA results compared to those for ARIMA-PCA and therefore were also omitted.

5.3 Optimized Tracking Results

In response to the nonutilitarian results produced from individual stock tracking, a second methodology was devised and developed for investigating economic significance. Initial efforts focused on implementation of an algorithm for tracking forecast accuracy for each stock over time, selecting and trading a single stock with the highest accuracy leading up to each sampling point. Results for each test year yielded terminal account balances with losses for most test intervals with both ARIMA and SSA methods. Various perturbations

156

of averaging accuracies over different-length time windows were investigated, all yielding poor account returns. Based on these observed outcomes, it was concluded that a measure of past accuracy was neither robust nor an adequate indicator of future performance.

With the goal of positive economic significance unreached, an additional approach was developed using accuracy as an analytical enhancement as opposed to a central focus as administered in the previous methods. The improved algorithm consisted of trading a single stock that exhibited the largest normalized, absolute value, forecasted change (Δ) in price for the succeeding step, hence an optimal tracking of expected performance of the stock pool at any given sampling time. Moreover, each trade was augmented with a stop-loss exit to mitigate effects of forecast inaccuracies. This scheme operated on the same forecasting models previously computed, but administered a different post-processing algorithm, saving a significant amount of computational time.[*] Application of this post-processing technique yielded remarkable results described hereinafter.

Model forecast results of the maximum changes for ARIMA-PCA in test interval 2015 are given in Figure 5-5.[†] While this plot was generally representative of forecasted maximum changes for most of the test intervals, Table 5-2 provides specific statistics for all test cases. Mean and dispersion (i.e., standard deviation) varied noticeably between intervals.

---

[*] Note that model generation/estimations consumed $> 98\%$ of the required processing time.

[†] Delta, Δ, is defined as the normalized change in price, and $|\Delta|$ is its absolute value.

Positive-negative symmetry was typically present, but not always, probably a consequence of the accuracy results discussed next.
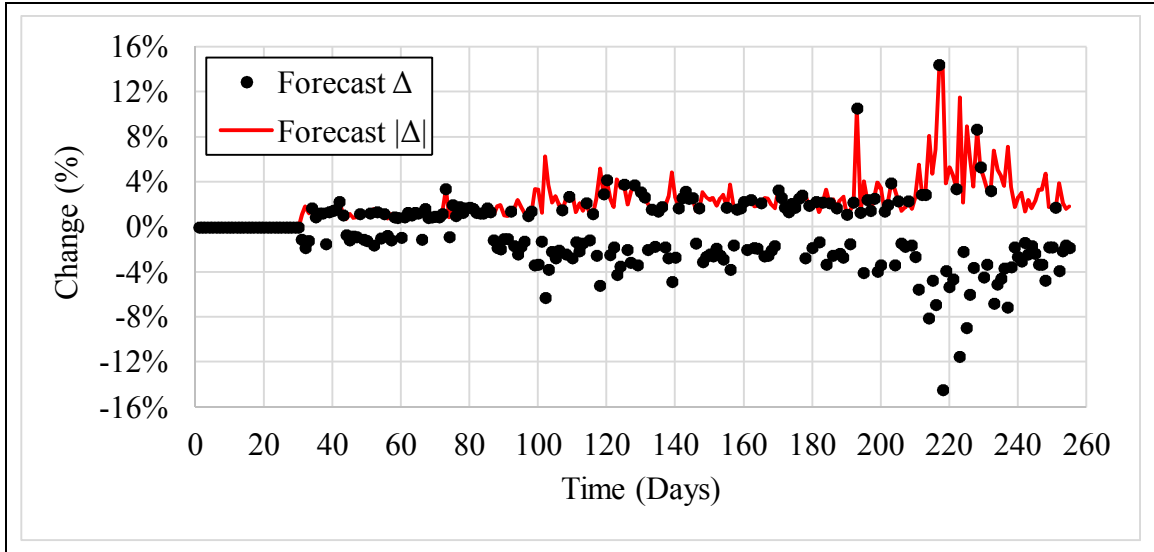
Figure 5-5. ARIMA-PCA Forecasts of Maximum-Change (2015)

Table 5-2. One-Step-Forecast-Change Statistics

Test Interval	ARIMA-PCA (%)		SSA-PCA (%)	
	|Mean|	Std Dev	|Mean|	Std Dev
2015	2.6	2.0	3.9	2.6
2014	1.6	0.8	2.6	1.7
2010	2.7	3.0	3.5	3.3
2009	5.3	5.4	5.6	4.6
2008	11.9	19.8	12.2	21.9

Note: \|Mean\| is defined as the mean of the absolute changes |Δ|.

A graph of forecasting accuracy results for test interval 2015 is shown in Figure 5-6. At any time point during the test interval, a one (on the vertical scale) represents a correct

forecast while a zero signifies an incorrect prediction. A summary of mean forecast accuracies for the other test intervals is provided in Table 5-3. Although the range varied from 47.1 to 57.3%, forecast accuracy appeared to have significantly-reduced impact to trading results with the use of the stop-loss.
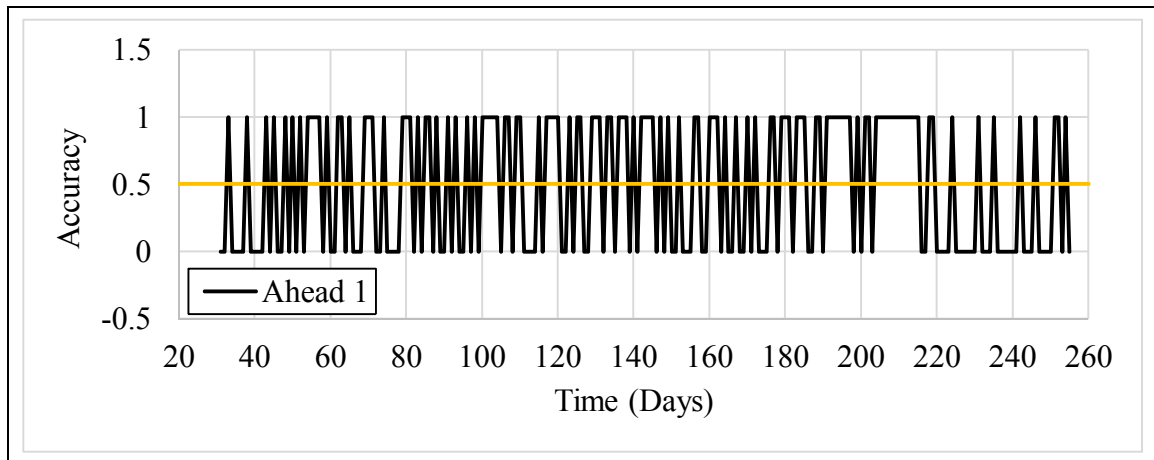
Figure 5-6. ARIMA-PCA Forecast Accuracy (2015)

Table 5-3. One-Step Mean Forecast Accuracies

Test Interval	ARIMA-PCA (%)	SSA-PCA (%)
2015	50.7	51.1
2014	57.3	54.2
2010	47.1	48.9
2009	52.9	50.2
2008	49.8	50.2

A plot showing a representative illustration of forecast direction over time is given in Figure 5-7. A forecasted increase is displayed as a one and a predicted decrease is shown as a negative one. Clearly, there was a continual change in expected direction of movement

159

as time progressed. Examination of the other four year-long test intervals exposed similar

irregular oscillatory motions of predicted directions for both ARIMA and SSA models.

There did not seem to be a bias toward either direction regardless of general market trend.
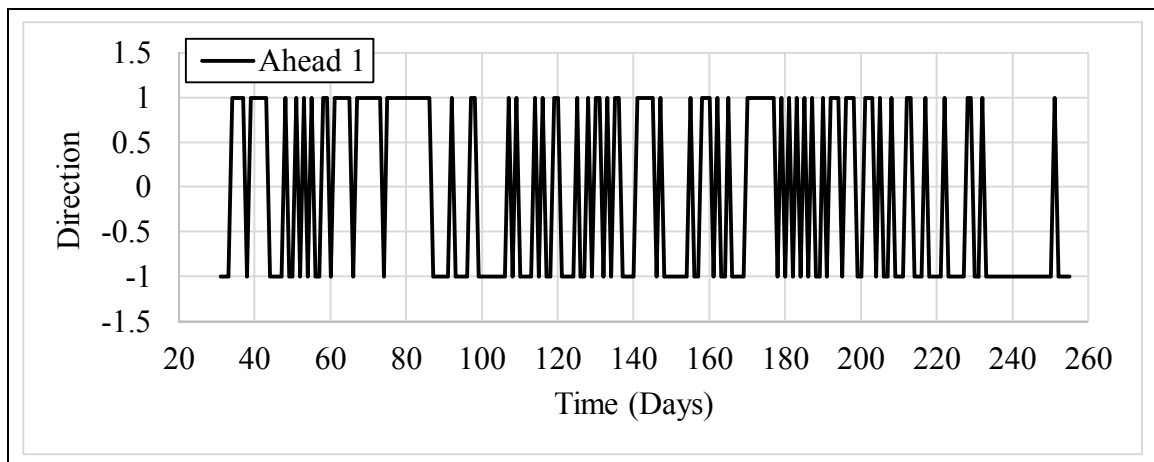
Figure 5-7. ARIMA-PCA Forecast Direction (2015)

Separate tracking was maintained for the one-step account and the five shifted (i.e.,

overlapped) five-step accounts. An example of the dynamics of daily returns for the 2015

test interval is shown in Figure 5-8.[*] A summary of the average daily returns is presented

in Table 5-4.[†] Results showed that ARIMA-PCA models typically performed better than

or comparable to the SSA-PCA models for the one-step accounts. In contrast, the five-step

accounts yielded mostly superior performance for SSA models, whereas only interval 2014

---

[*] Ahead 5-4 refers to the fourth of the five-step-forecast-overlapped intervals; it was a
representative selection, falling in the middle of the five shifted test cases. Note also
that the five-step returns are simply held at the previous level until the subsequent step
is completed. No additional gain/loss should be implied during these hold periods.

[†] For purposes of comparison, resultant five-step-forecast returns were divided by 5 to
reflect equivalent average daily returns.

showed better returns with ARIMA-PCA. In summary, one-step daily returns were always greater than the five-step (average daily) returns with the ARIMA models outperforming the SSA implementations.
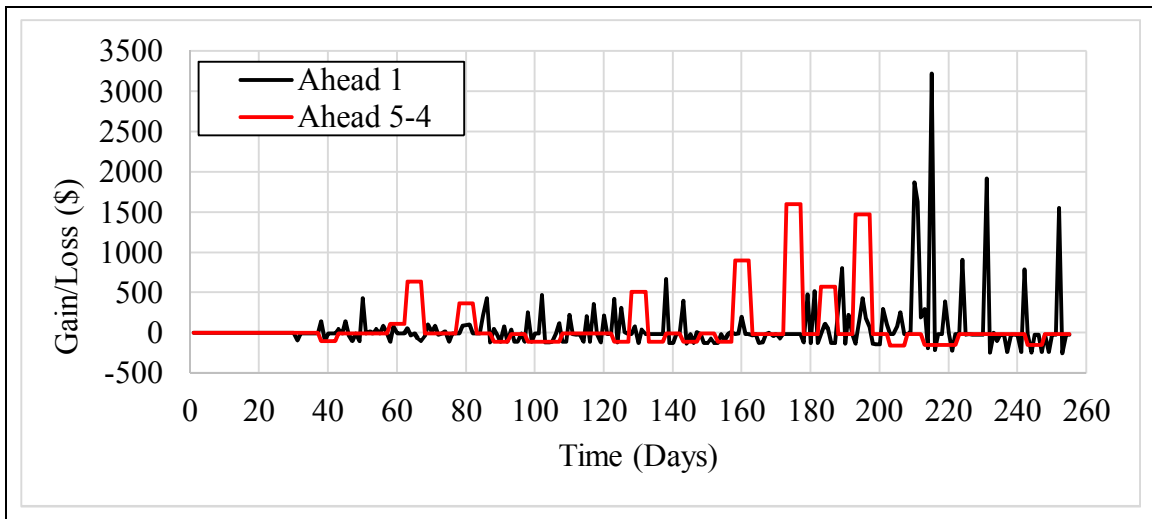
Figure 5-8. ARIMA-PCA Daily Returns (2015)

Table 5-4. Mean Daily Return without Fees

Test Interval	ARIMA-PCA ($)		SSA-PCA ($)	
	Ahead 1	Ahead 5	Ahead 1	Ahead 5
2015	63.9	18.3	35.4	19.5
2014	36.4	8.7	13.7	4.2
2010	27.0	6.9	36.5	11.4
2009	631.1	17.2	194.7	60.5
2008	1066.0	17.7	1301.7	42.5

Note: Ahead 5 is the mean of five overlapped forecast periods.

Figure 5-9 shows a typical example of stock selection dynamics. (Refer to Appendix 8.1 for the stock number-name association.) While no organized pattern was apparent (or expected), there were a few stocks that appeared to be chosen repeatedly by the decision algorithm. Evidence showed that any particular stock would fall into and out of favor over the year-long interval. A histogram of the frequency of occurrence for the same test interval is provided in Figure 5-10. While the distribution was spread across the field of stocks, there were two stocks identified for trading more than ten times during the interval and some that were never traded at any time. Similar dynamics were observed in the other four test intervals and for SSA-PCA models, as well.
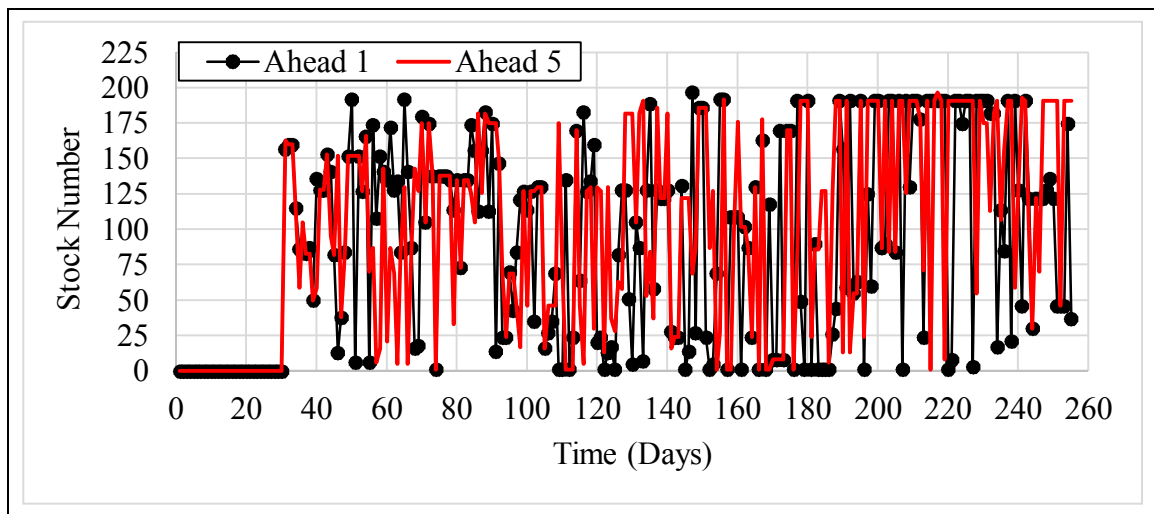
Figure 5-9. ARIMA-PCA Stock Selection Dynamics (2015)

162
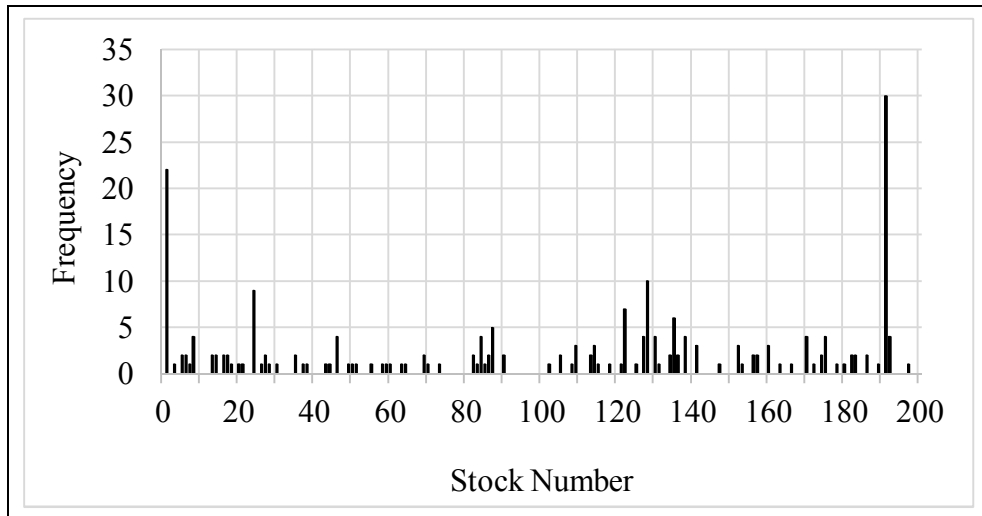
Figure 5-10. ARIMA-PCA Stock Selection Distribution (2015)

An example of the time-dependent change in account balances without fees and with fees included for the one-step ARIMA-PCA forecast model is shown in Figure 5-11. As seen in the figure, fees, although modest, did noticeably affect the final balance after a year of accumulation. A second plot including both one- and five-step balances is provided in Figure 5-12. This comparative plot revealed that one-step forecasts clearly outperformed the five-step models. This was typical, but not exclusive, for all tested models. A summary of account balances for all models is given in Table 5-5.

163
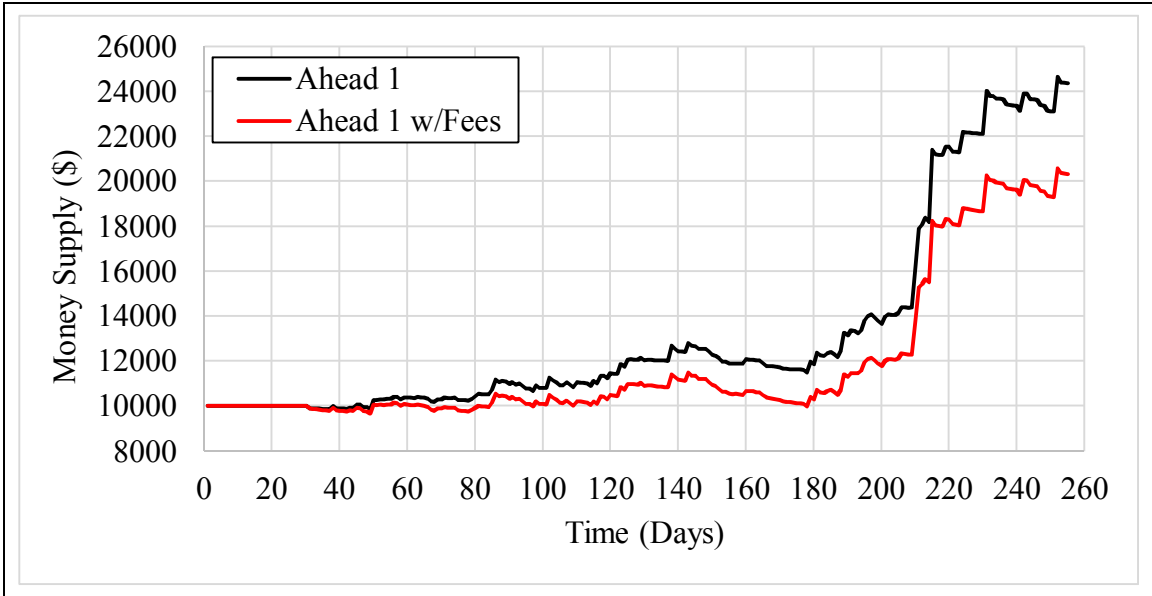
Figure 5-11. ARIMA-PCA Account Balance (2015)
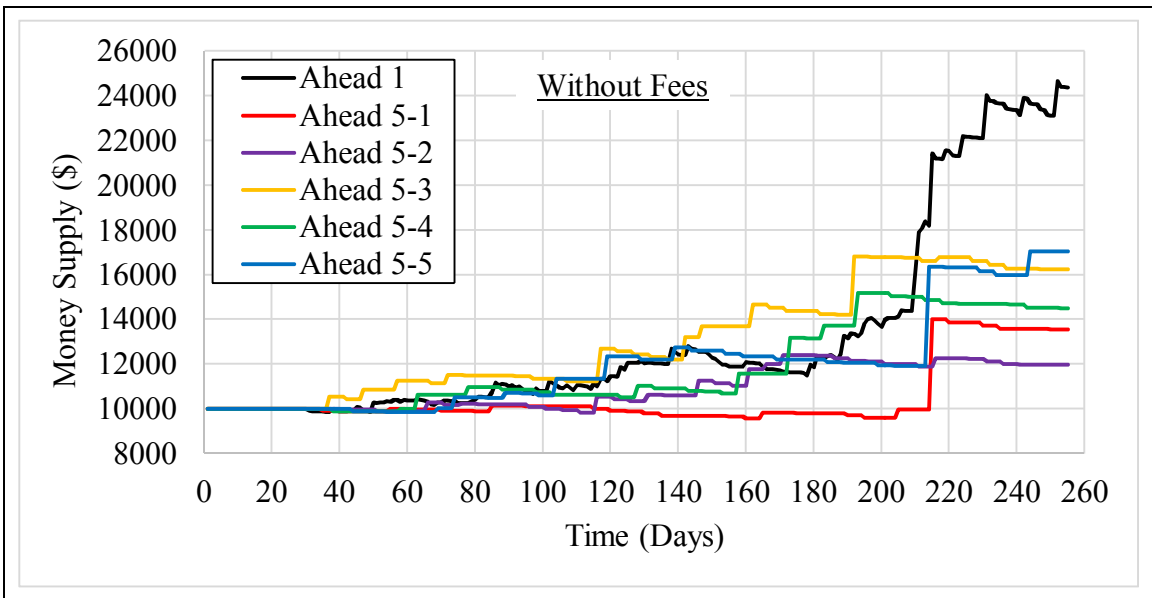
Figure 5-12. ARIMA-PCA Performance Comparison (2015)

164

Table 5-5. Terminal Account Balances

Test Interval	ARIMA-PCA ($)			SSA-PCA ($)		
	Ahead 1	Ahead 5 (Lowest)	Ahead 5 (Highest)	Ahead 1	Ahead 5 (Lowest)	Ahead 5 (Highest)
Without Fees						
2015	**24,373**	11,974	17,045	17,962	12,102	17,919
2014	**18,199**	10,972	13,860	13,088	9,526	13,003
2010	16,065	8,706	17,336	**18,218**	10,498	15,275
2009	**151,991**	9,969	17,229	53,799	10,702	68,255
2008	249,853	8,896	16,104	**302,878**	10,223	39,113
With Fees						
2015	**20,321**	11,512	16,440	14,853	11,820	17,437
2014	**14,897**	10,510	13,321	10,710	9,271	12,633
2010	13,219	8,299	**16,767**	15,234	10,195	14,930
2009	**143,932**	9,529	16,734	48,758	10,411	67,737
2008	221,659	8,490	15,692	**264,442**	9,939	38,232

Notes: 1. Values are rounded to the nearest dollar.
2. Bold values indicate highest value in each test interval.

Observing Table 5-5, the top performing model without fees also produced the best returns with fees, with the exception of 2010. The best consistency appeared to emanate from the one-step ARIMA-PCA models. On the other hand, no particular model dominated the field. Five-step forecast results revealed fundamental challenges to utility in that there were multiple overlapping forecast periods from which to select the best (or at least avoid the worst) starting day. Inasmuch as it was impossible to know ahead of time which starting day to select, results advanced the conclusion that the five-step forecast model was

generally less profitable. The single redeeming aspect regarding consistency is that all one-step models produced positive returns. In contrast; five-step model results were usually significantly lower and losses at termination were not uncommon. A comparison of the time progression of ARIMA-PCA and SSA-PCA account balances for all test intervals is shown in Figure 5-13, each referenced to its start day.
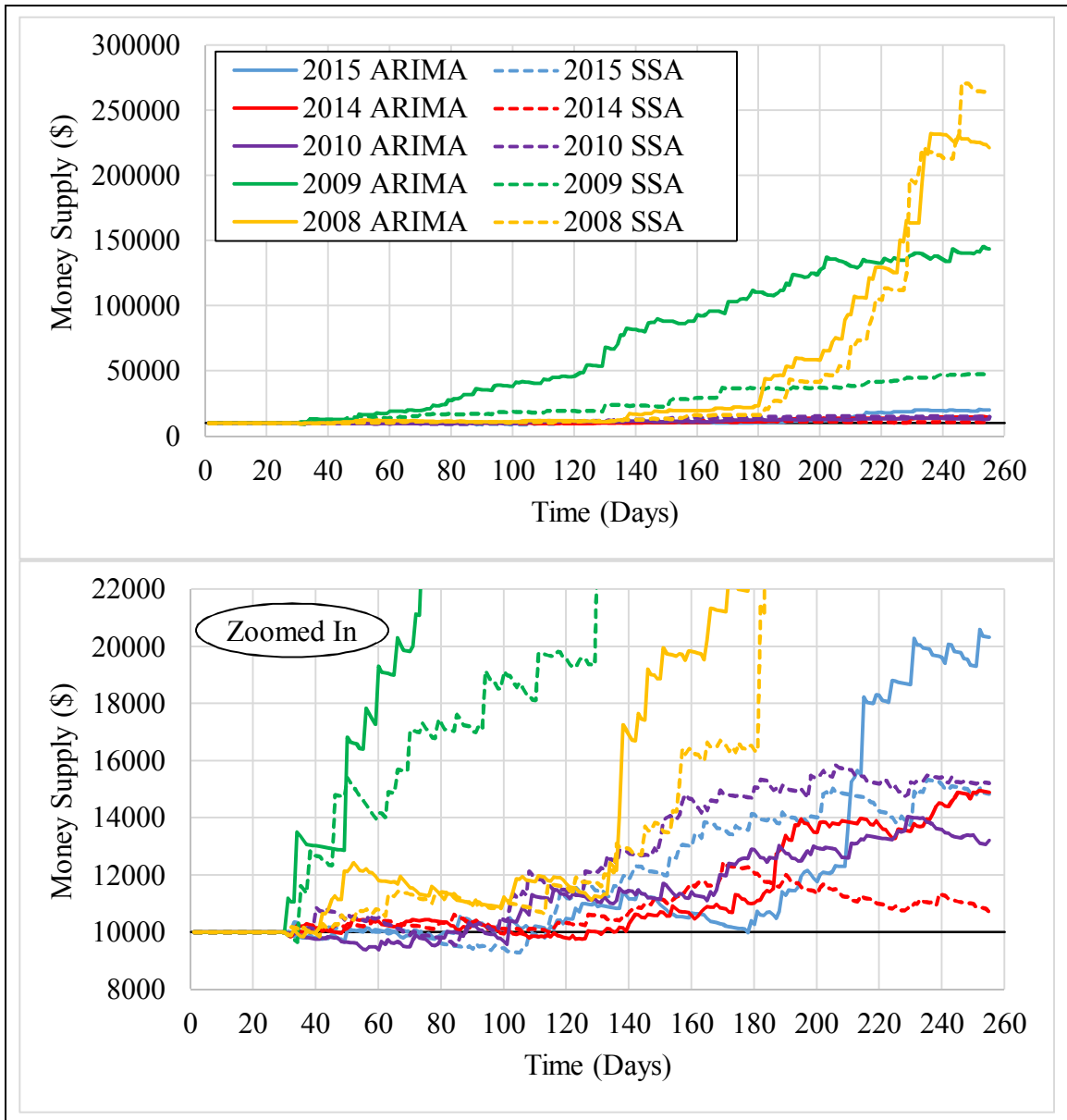
Figure 5-13. ARIMA-SSA Performance Comparison (including Fees)

Lastly, the optimized-tracking results for the one-step ARIMA-based model[*] were evaluated for risk vs. reward in a comparison with the performance of the S&P 500 Index during the same five test intervals. A greater standard deviation, i.e., dispersion, reflects a greater risk of an investment, but the comparison of investments requires the same expected return. To allow valid comparisons of different investments, the coefficient of variation (CoV) is a simple method that quantifies dispersion relative to the expected return of an investment. This measure is defined as

$$\text{CoV} = \frac{\text{Standard Deviation}}{\text{Mean}} = \frac{\sigma}{\mu} \; . \tag{5-1}$$

A summary of computed risk profiles is provided in Table 5-6.[†] Computations exposed a lower CoV for the optimized algorithm, which reflected a lower risk than investing in the S&P 500 Index during the same period. While the standard deviations produced by the optimal algorithm were larger, the mean daily returns of the S&P Index were close to zero, substantially increasing the quantified risk. Moreover, negative means for the S&P 500 reflected a net loss for the analyzed period, obviously an undesirable investment outcome.

---

[*] All other models were excluded due to their generally inferior performance.

[†] Annualized returns may be approximated by multiplying daily returns by 252. Also note that the test intervals were not coincident with calendar years, but the S&P Index computations were configured to be coincident with the ranges of the test intervals.

Table 5-6. Risk vs. Reward for Daily Returns

Test Interval	Mean, μ (%)		Std Dev, σ (%)		CoV	
	S&P 500	**Optimized‡**	**S&P 500**	**Optimized‡**	**S&P 500**	**Optimized‡**
2015	–0.005	0.46	1.05	3.04	–203.4	6.6
2014	0.043	0.22	0.73	1.43	16.9	6.6
2010	0.051	0.14	1.13	1.52	22.1	10.6
2009	0.093	5.95	1.70	21.46	18.3	3.6
2008	–0.16	9.41	2.58	41.03	–16.2	4.4

‡ One-step ARIMA-PCA with fees

5.4 Performance Factors

A reader may question whether these results are credible since they demonstrated returns above those normally realized in reported financial literature.[*] This is addressed by the following points.

Statistics were computed to evaluate rates of return for all equites in the stock list for each forecast step in each test interval. As shown in Table 5-7, the averages of the positive and negative maximum daily returns for all stocks across each test period resulted in magnitudes greater than 4.3%. Furthermore, results revealed that the average number of stocks yielding (maximum) returns larger than ±1% was significant in each interval. The

---

[*] For example, in 2016, *Barron's Penta* top hedge fund posted a three-year (2013-2015) annualized return of 30%, with a 2015 return of 45% [52].

outcomes seen in the bottom two rows of the table exposed a small number of daily samples in each test year where no returns exceeded ±1%. Although detailed statistics are not included herein, it should be noted that analysis of mean daily return for each of the 197 stocks was very close to zero (like the S&P 500 Index). Key to achieving significant returns was identifying stocks with the largest daily changes.

Table 5-7. Test Interval Statistics for Stock List

Statistic		Test Interval				
		2008	2009	2010	2014	2015
Mean Value	**+ Maximum Daily Return (%)**	7.0	12.9	6.1	4.7	5.2
	– Maximum Daily Return (%)	-7.1	-12.0	-5.6	-4.3	-4.8
	Std Dev of Daily Returns (%)	1.7	3.0	1.4	1.0	1.2
	No. of Stocks with Returns > 1%	54	69	50	33	41
	No. of Stocks with Returns < –1%	57	71	46	30	39
No. of Days with No Return > 1%		1	4	5	2	7
No. of Days with No Return < –1%		0	1	2	3	1

The efficient market hypothesis (EMH), known also as the random-walk theory [4], continues to be controversial to the present. The frequency of occurrence displayed for ARIMA random-walk models in earlier chapters (and supported by forecast accuracy results in §5.3) indicated a combination of efficient markets and predictable markets, each surfacing at unknown and irregular points in time. Additionally, forecasted maximum change was found to be independent of forecasted accuracy. To maintain consistency, i.e., to avoid modifications of already-characterized forecast models, considering that forecast

accuracies occurred about the 50% range, some alternate method was required to achieve a positive trading return for any given test interval. Inclusion of a stop-loss limit was found to reduce the impacts of incorrect forecasts by restraining negative excursions. Stop-loss thresholds were set at 1% of account balance for all transactions and reduced to 0.075% for forecasts where the previous step accuracy was incorrect.

When assessing the magnitudes of account terminal balances of each test interval, the effect of compounding is a factor to be considered. Each test period included 225 (daily) samples wherein the available account balance at each step was applied toward the next forecast. The five-step models comprised 45 samples (225 ÷ 5). The effects of compounding for these two scenarios are displayed in Table 5-8 where the results of starting with $10,000 are presented. Hence, if a trading model can produce an average return rate of only 1% per day, a starting value of $10,000 will end near $93,000. Based on these compounding effects, the results obtained from this analysis do not suggest unreasonable returns, assuming the aforementioned forecast accuracies. Naturally, the 45-sample rate of return must be much higher to achieve similar results. Thus, unless they project significantly better forecast accuracies, five-step models incorporate a disadvantage to producing returns as large as the one-step models. This was evident in the test results.

In a final word about general performance, the back-testing algorithms were simple and intentionally designed to be autonomous, avoiding any subjective, user decision-driving input. These approaches were necessary for conducting the extensive evaluations for this project. In a real-world application, an investor would apply the autonomous trading

Table 5-8. Effects of Compounding

Interest Rate (%)	225-Sample Compounded Return ($)	Interest Rate (%)	45-Sample Compounded Return ($)
0.1	12,522	0.2	10,941
0.2	15,676	0.4	11,968
0.3	19,620	0.6	13,089
0.4	24,552	0.8	14,313
0.5	30,716	1	15,648
0.6	38,419	1.2	17,105
0.7	48,043	1.4	18,694
0.8	60,065	1.6	20,428
0.9	75,078	1.8	22,318
1.0	93,823	2	24,379
1.1	117,222	2.2	26,625
1.2	146,425	2.4	29,074
1.3	182,863	2.6	31,742
1.4	228,318	2.8	34,649
1.5	285,010	3	37,816

Note: Initial Value = $10,000

approach on a daily basis toward earning profits on the following day. With knowledge of market conditions, other information such as earnings announcements, news, etc., in all probability a careful investor would evaluate the next forecast prior to configuring a trade, applying judgement regarding potential outcome, and possibly select an alternate stock or

decide not to trade at all. In this manner, the input from a skilled investor/trader may be able to further improve on the algorithm's nominal return results.

6. CONCLUSION

"Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise."

– John W. Tukey (1915-2000), in "The Future of Data Analysis," 1962 [47]

This research project focused on forecasting using two very different methods of time series analysis and compared performance results. U.S. stock market equities were selected as the target data set due to perceived random-like characteristics, implied forecasting difficulty, and the general notion of a compelling interest to predict future prices for financial gain. The market was screened for candidate data meeting minimum requirements for trading liquidity, and historical longevity to allow a sufficient number of independent samples to be evaluated. Statistical analyses of the data found that the best stochastic fit was a (slightly-skewed) generalized t-distribution, definitely eliminating the historical, original working assumption of Gaussianity. Probability tests for goodness-of-fit were conducted to validate these findings.

Each stock data set was segmented into independent segments of non-overlapping data to support unconditional statistical evaluation. In order to perform coherent time series analyses of the data segments, preconditioning was applied to compensate for the occasional split inherent in equities. Univariate and multivariate analyses were carried out, generating ARIMA and SSA models for all. Model independence was evaluated. Out-of-sample forecasts were estimated and statistics were computed for every stock in every

173

independent segment. Statistical confidence intervals for all results were determined. Results were compared for univariate, bivariate, and trivariate combinations of highly-correlated stocks. Multivariate-univariate confidence interval shrinkages were examined.

Bivariate modeling and forecasting with three different covariates were performed. Examination of results with covariates of trading volume, PCA first principal component, and volatility, revealed that PCA exhibited the best overall forecasting accuracy in the entire field of investigated elements, including the univariate models. The bivariate-PCA structures were applied at the next level to evaluate economic significance and robustness through a process known as back-testing.

Initial results of back-testing via tracking the performance of individual stocks yielded forecasting results similar to those obtained in earlier analyses. Unfortunately, the inconsistencies obtained over multiple testing intervals proved useless in attaining a sense of positive economic significance. The results were simply too random to produce reliable economic forecasting outcomes. A second back-testing setup using the best forecasting results across the field of stocks at each sample time yielded notable results and substantial economic significance. Robustness of this method was validated through evaluation in multiple market-trend conditions.

There are several possible areas to evaluate for continuing and extending this investigation in future research efforts. The main objective would be to improve economic returns by increasing forecast accuracies. Areas to investigate include:

1. Determining effects of various window lengths;

2. Varying SSA model thresholds;

3. Developing a better method to extract the optimal SSA groups;

4. Evaluation with financial assets other than equites;

5. Determining economic significance and robustness with alternate covariates;

6. Evaluation and possible modifications to the back-testing method to function with a smaller field of data;

7. Improving the computational speed of model generation.

In conclusion, results indicated that this project successfully reached its goal of finding a rigorous, reliable, mathematically-based method to forecasting asset movements in a tradeable economic market.

"The acquisition of any knowledge whatever is always useful to the intellect, because it will be able to banish the useless things and retain those which are good. For nothing can be either loved or hated unless it is first known."

– Leonardo da Vinci (1452-1519), in "The Notebooks of Leonardo da Vinci," vol. 1

7. REFERENCES

[1]   finviz - Financial Visualizations, [Online]. Available: http://finviz.com/screener.ashx. [Accessed March 2016].

[2]   A. W. Lo, C. Petrov and M. Wierzbicki, "It's 11PM–Do You Know Where Your Liquidity Is? The Mean-Variance-Liquidity Frontier," *Journal of Investment Management,* vol. 1, no. 1, pp. 55-93, 2003.

[3]   Yahoo! Finance, [Online]. Available: https://finance.yahoo.com/. [Accessed March 2016].

[4]   E. F. Fama, "The Behavior of Stock-Market Prices," *Journal of Business,* vol. 38, no. 1, pp. 34-105, 1965.

[5]   T. Doganoglu and S. Mittnik, "An Approximation Procedure for Asymmetric Stable Paretian Densities," *Computational Statistics,* vol. 13, no. 4, pp. 463-476, 1998.

[6]   F. M. Aparicio and J. Estrada, "Emprical Distributions of Stock Returns: Scandinavian Securities Markets, 1990-95," Carlos III University - Department of Statistics and Econometrics and Department of Business, Madrid, 1997.

[7]   J. B. Gray and D. W. French, "Empirical Comparisons of Distributional Models for Stock Index Returns," *Journal of Business Finance & Accounting,* vol. 17, no. 3, pp. 451-459, 2008.

[8]   S. T. Rachev, S. V. Stoyanov, A. Biglova and F. J. Fabozzi, "An Empirical Examination of Daily Stock Return Distributions for U.S. Stocks," in *Data Analysis and Decision Support*, Berlin-Heidelberg, Springer-Verlag, 2005, pp. 269-281.

[9]   R. S. Tsay, Analysis of Financial Time Series: Financial Econometrics, New York, NY: John Wiley & Sons, Inc., 2002.

[10] "Errors and Residuals," World Heritage Encyclopedia, [Online]. Available: http://worldheritage.org/articles/Errors_and_residuals_in_statistics. [Accessed 30 June 2016].

[11] J. D. Cryer and K.-S. Chan, in *Time Series Analysis with Applications in R*, New York, Springer, 2008, p. 42.

[12] P. Theodossiou, "Financial Data and the Skewed Generalized T Distribution," *Management Science,* vol. 44, no. 12, pp. 1650-1661, 1998.

[13] C. Hansen, J. B. McDonald and W. K. Newey, "Instrumental Variables Estimation With Flexible Distributions," *Journal of Business & Economic Statistics,* vol. 28, no. 1, pp. 13-25, 2010.

[14] R. Savage, "Probability Inequalities of the Tchebycheff Type," *Journal of Research of the National Bureau of Standards, Section B: Mathematics and Mathematical Physics,* vol. 65B, no. 3, pp. 211-226, 1961.

[15] L. J. Bain and M. Engelhardt, "Other Goodness-of-Fit Tests," in *Introduction to Probability and Mathematical Statistics, 2nd ed.*, Pacific Grove, Duxbury, 1992, pp. 457-460.

[16] S. Csorgo and J. J. Faraway, "The Exact and Asymptotic distributions of Cramer-von Mises Statistics," *Journal of the Royal Statistical Society,* vol. 58, no. 1, pp. 221-234, 1996.

[17] R. J. Hyndman and G. Athanasopoulos, "8.5 Non-seasonal ARIMA models," in *Forecasting: Principles and Practice*, Melbourne, OTexts, 2013, p. https://www.otexts.org/fpp/8/5.

[18] R. J. Hyndman and G. Athanasopoulos, "8.6 Estimation and Order Selection," in *Forecasting: Principles and Practice*, Melborne, OTexts, 2013, p. https://www.otexts.org/fpp/8/6.

[19] S. Makridakis and M. Hibon, "ARMA Models and the Box-Jenkins," *Journal of Forecasting,* vol. 16, pp. 147-163, 1997.

[20] G. Ljung and G. Box, "On a measure of lack of fit in time series models," *Biometrika,* vol. 65, no. 2, pp. 297-303, 1978.

[21] Y. Hong, "Serial Correlation and Serial Dependence," The Wang Yanan Institute for Studies in Economics, Xiamen, 2006.

[22] R. J. Hyndman, "HYNDSIGHT, A blog by Rob J Hyndman," 24 January 2014. [Online]. Available: http://robjhyndman.com/hyndsight/ljung-box-test/. [Accessed April 2016].

[23] K. Fokianos and M. Pitsillou, "Consistent testing for pairwise dependence in time series," *Technometrics,* 24 February 2016.

[24] Y. Hong, "Consistent Testing for Serial Correlation of Unknown Form," *Econometrica,* vol. 64, no. 4, pp. 837-864, 1996.

[25] R. S. Tsay, Multivariate Time Series Analysis with R and Financial Applications, Hoboken: John Wiley & Sons, Inc., 2014.

[26] A. Papoulis and S. U. Pillai, in *Probability, Random Variables, and Stochastic Processes*, New York, McGraw-Hill, 2002, pp. 398-402.

[27] K. Fokianos and M. Pitsillou, "On Multivariate Auto-Distance Covariance and Correlation," *Submitted for publication (TBD),* 2016.

[28] D. S. Broomhead and G. P. King, "Extracting Qualitative Dynamics from Experimental Data," *Physica D,* vol. 20, no. 2-3, pp. 217-236, 1986.

[29] D. S. Broomhead and G. P. King, "On the qualitative analysis of experimental dynamical systems," in *Nonlinear Phenomena and Chaos*, Bristol, Adam Hilger, 1986, pp. 113-144.

[30] N. Golyandina, V. Nekrutkin and A. Zhigljavsky, Analysis of Time Series Structure, Boca Raton: Chapman & Hall/CRC, 2001.

[31] N. Golyandina and A. Zhigljavsky, Singular spectrum Analysis for Time Series, Heidelberg: Springer, 2013.

[32] R. Mahmoudvand, N. Najari and M. Zokaei, "On the Optimal Parameters for Reconstruction and Forecasting in Singular Spectrum Analysis," *Communications in Statistics—Simulation and Computation,* vol. 42, no. 4, pp. 860-870, 2013.

[33] M. Zokaei, R. Mahmoudvand and N. Najari, "Comparison of Singular Spectrum Analysis and ARIMA," in *Proceedings of the ISI 58th World Statistical Congress*, Dublin, 2011.

[34] D. S. Wilks, "Resampling Hypothesis Tests for Autocorrelated Fields," *Journal of Climate,* vol. 10, no. 1, pp. 65-82, 1997.

[35] K. E. Trenberth, "Some Effects of Finite Sample Size and Persistence on Meterological Statistics. Part I: Autocorrelations," *Monthly Weather Review,* vol. 112, pp. 2359-2368, 1984.

[36] H. Hassani and R. Mahmoudvand, "Multivariate Singular Spectrum Analysis: A General View and New Vector Forecasting Approach," *International Journal of Energy and Statistics,* vol. 1, no. 1, pp. 55-83, 2013.

[37] N. Golyandina and D. Stepanov, "SSA-based Approaches to Analysis and Forecast of Multidimensional Time Series," in *Proceedings of the 5th St.Petersburg Workshop on Simulation*, St. Petersburg, 2005.

[38] J. L. Person, A Complete Guide to Technical Trading Tactics, Hoboken: John Wiley & Sons, Inc., 2004.

[39] J. Shlens, "A Tutorial on Principal Component Analysis," 10 December 2005. [Online]. Available: http://www.cs.cmu.edu/~elaw/papers/pca.pdf. [Accessed 17 January 2017].

[40] R. A. johnson and D. W. Wichern, "Principal Components," in *Applied Multivariate Statistical Analysis, 6th ed.*, Upper Saddle River, Pearson Education, Inc., 2007, pp. 430-480.

[41] R. S. Tsay, "Conditiional Heteroscedastic Models," in *Analysis of Financial Time Series, 3rd ed.*, Hoboken, John Wiley & sons, Inc., 2010, pp. 109-173.

[42] T. Bollerslev, "Generalized Autoregressive Conditional Heteroskedasticity," *Journal of Econometrics,* vol. 31, pp. 307-327, 1986.

[43] R. F. Engle, "Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kindom Inflation," *Econometrica,* vol. 50, no. 4, pp. 987-1007, 1982.

[44] K. Li and D. Weinbaum, "The Empirical Performance of Alternative Extreme Value Volatility Estimators," Stern School of Business, New York University, New York, 2000.

[45] L. C. G. Rogers, S. E. Satchell and Y. Yoon, "Estimating the Volatility of Stock Prices: A Comparison of Methods that use High and Low Prices," *Applied financial Economics,* vol. 4, pp. 241-247, 1994.

[46] D. Yang and Q. Zhang, "Drift-Independent Volatility Estimation Based on High, Low, Open, and Close Prices," *Journal of Business,* vol. 73, no. 3, pp. 477-491, 2000.

[47] J. W. Tukey, "The Future of Data Analysis," *Annals of Mathematical Statistics,* vol. 33, no. 1, pp. 13-14, 1962.

[48] B. Mandelbrot, "The Variation of Certain Speculative Prices," *The Journal of Business,* vol. 36, no. 4, pp. 394-419, 1963.

[49] C. Davis, "The Skewed Generalized T Distribution Tree Package Vignette," Chicago, 2015.

[50]   "Available CRAN Packages By Name," The R Foundation, [Online]. Available: https://cran.r-project.org/web/packages/available_packages_by_name.html. [Accessed 2016].

[51]   A. Ghalanos, "Introduction to the rugarch Package (Version 1.3-1)," 16 August 2015. [Online]. Available: https://cran.cnr.berkeley.edu/web/packages/rugarch/index.html. [Accessed 8 February 2017].

[52]   E. Uhlfeder, "Top 100 Hedge Funds for 2016," Barron's, 18 June 2016. [Online]. Available: http://www.barrons.com/articles/our-top-100-hedge-funds-1466222376. [Accessed 17 April 2017].

8. APPENDICES

8.1 Appendix 1 – Stock List

No.	Stock	No.	Stock	No.	Stock	No.	Stock
1	AA	26	BEN	51	CTL	76	FDX
2	AAPL	27	BHI	52	CVS	77	FISV
3	ABT	28	BK	53	CVX	78	FITB
4	ADBE	29	BMY	54	D	79	GD
5	ADI	30	C	55	DD	80	GE
6	ADM	31	CA	56	DE	81	GIS
7	ADP	32	CAG	57	DHR	82	GLW
8	ADSK	33	CAH	58	DIS	83	GPS
9	AEP	34	CAT	59	DOV	84	HAL
10	AET	35	CCE	60	DOW	85	HCP
11	AFL	36	CCL	61	DTE	86	HD
12	AIG	37	CELG	62	DUK	87	HES
13	AMAT	38	CERN	63	EA	88	HON
14	AME	39	CHD	64	ECL	89	HOT
15	AMGN	40	CI	65	ED	90	HPQ
16	APA	41	CL	66	EIX	91	HRL
17	APC	42	CLX	67	EMC	92	HST
18	APD	43	CMCSA	68	EMR	93	HSY
19	AXP	44	CMI	69	EOG	94	IBM
20	BA	45	CMS	70	EQT	95	INTC
21	BAC	46	COP	71	ES	96	IP
22	BAX	47	COST	72	ETR	97	ITW
23	BBT	48	CPB	73	EXC	98	JCI
24	BBY	49	CSCO	74	F	99	JNJ
25	BDX	50	CSX	75	FAST	100	JPM

Stock List – continued

No.	Stock	No.	Stock	No.	Stock	No.	Stock
101	K	126	MU	151	PPL	176	TXN
102	KLAC	127	MYL	152	RF	177	UNH
103	KMB	128	NBL	153	ROK	178	UNP
104	KO	129	NEE	154	ROST	179	USB
105	KR	130	NEM	155	RTN	180	UTX
106	L	131	NKE	156	SCHW	181	VFC
107	LB	132	NOC	157	SLB	182	VLO
108	LLTC	133	NSC	158	SO	183	VMC
109	LLY	134	NTRS	159	STI	184	VNO
110	LMT	135	NUE	160	STJ	185	VZ
111	LOW	136	NWL	161	STT	186	WBA
112	LRCX	137	OMC	162	SWK	187	WDC
113	LUV	138	ORCL	163	SWKS	188	WEC
114	MAS	139	OXY	164	SYK	189	WFC
115	MAT	140	PAYX	165	SYMC	190	WHR
116	MCD	141	PCAR	166	SYY	191	WMB
117	MDT	142	PCG	167	T	192	WMT
118	MHFI	143	PEG	168	TAP	193	WY
119	MMC	144	PEP	169	TGT	194	XEL
120	MMM	145	PFE	170	TJX	195	XLNX
121	MO	146	PG	171	TMO	196	XOM
122	MOS	147	PGR	172	TROW	197	XRX
123	MRK	148	PH	173	TRV		
124	MSFT	149	PNC	174	TSN		
125	MSI	150	PPG	175	TSO		

8.2 Appendix 2 – Stock Pairs

Tier 1

Pair No.	Stock 1	Stock 2	Pair No.	Stock 1	Stock 2
1	BHI	NBL	26	ADM	CLX
2	COST	GPS	27	AET	MO
3	LLTC	SO	28	AMAT	BAC
4	LRCX	TJX	29	AME	CERN
5	ABT	CHD	30	BAC	C
6	BA	TSO	31	BAX	GIS
7	CTL	TROW	32	BHI	OMC
8	MAS	PGR	33	CMCSA	SWK
9	NWL	UNH	34	COST	ROST
10	OXY	PEG	35	DOW	HD
11	AA	OMC	36	DOW	PNC
12	AAPL	ADSK	37	DUK	FDX
13	AFL	MO	38	FITB	HD
14	BEN	D	39	GD	KR
15	CPB	D	40	NEE	TROW
16	ECL	F	41	OXY	RF
17	FDX	LB	42	PCAR	WEC
18	FISV	PH	43	STI	VMC
19	FITB	HOT	44	STI	VNO
20	LUV	PPL	45	SWKS	WBA
21	PAYX	SYMC			
22	SO	TGT			
23	STI	WDC			
24	STT	WMB			
25	TROW	WY			

Tier 2

Pair No.	Stock 1	Stock 2	Pair No.	Stock 1	Stock 2
46	ABT	C	71	FISV	TSN
47	ADSK	TXN	72	FISV	WHR
48	AFL	TSN	73	KLAC	SWK
49	AFL	VZ	74	LLTC	MYL
50	AIG	DE	75	LLTC	PEP
51	AMAT	CAG	76	ORCL	WMT
52	AME	IBM	77	PCAR	SO
53	AXP	HSY	78	PCAR	TJX
54	BA	ORCL	79	STT	UNP
55	BA	PNC	80	STT	WMT
56	CAH	CERN	81	SWK	TROW
57	CAH	ECL	82	SYK	WFC
58	CCE	HD	83	TJX	VNO
59	CI	CVS	84	TMO	VNO
60	CMI	LRCX	85	TSO	XEL
61	COP	DTE			
62	COP	FITB			
63	COST	HAL			
64	COST	IBM			
65	CPB	JNJ			
66	DOW	PCG			
67	DTE	PGR			
68	EA	XRX			
69	ES	LRCX			
70	FDX	WBA			

8.3   Appendix 3 – Stock Triplets

No.	Stock 1	Stock 2	Stock 3	No.	Stock 1	Stock 2	Stock 3
1	BHI	NBL	OMC	26	TROW	WY	NEE
2	COST	GPS	ROST	27	TROW	WY	SWK
3	COST	GPS	HAL	28	CTL	TROW	NEE
4	COST	GPS	IBM	29	CTL	TROW	SWK
5	COST	ROST	HAL	30	NEE	TROW	SWK
6	COST	ROST	IBM	31	MAS	PGR	DTE
7	COST	HAL	IBM	32	OXY	PEG	RF
8	LLTC	SO	MYL	33	AA	OMC	BHI
9	LLTC	SO	PEP	34	ADSK	TXN	AAPL
10	LLTC	MYL	PEP	35	AFL	MO	TSN
11	SO	TGT	LLTC	36	AFL	MO	VZ
12	SO	TGT	PCAR	37	AFL	TSN	VZ
13	LLTC	SO	PCAR	38	AFL	MO	AET
14	LRCX	TJX	CMI	39	BEN	D	CPB
15	LRCX	TJX	ES	40	CPB	D	JNJ
16	CMI	LRCX	ES	41	ECL	F	CAH
17	TJX	VNO	LRCX	42	FDX	LB	WBA
18	TJX	VNO	PCAR	43	FDX	LB	DUK
19	LRCX	TJX	PCAR	44	FDX	WBA	DUK
20	ABT	CHD	C	45	FISV	PH	TSN
21	BA	TSO	ORCL	46	FISV	PH	WHR
22	BA	TSO	PNC	47	FISV	TSN	WHR
23	BA	ORCL	PNC	48	FITB	HOT	HD
24	TSO	XEL	BA	49	FITB	HOT	COP
25	TROW	WY	CTL	50	FITB	HD	COP

Stock Triplets – continued

No.	Stock 1	Stock 2	Stock 3	No.	Stock 1	Stock 2	Stock 3
51	STI	WDC	VMC	76	STI	VNO	TMO
52	STI	WDC	VNO	77	TJX	VNO	TMO
53	STI	VMC	VNO	78	SWKS	WBA	FDX
54	STT	WMB	UNP	79	AFL	TSN	FISV
55	STT	WMB	WMT	80	AME	IBM	COST
56	STT	UNP	WMT	81	ORCL	WMT	BA
57	AMAT	BAC	CAG	82	CAH	CERN	ECL
58	BAC	C	AMAT	83	COP	DTE	FITB
59	AME	CERN	IBM	34	DTE	PGR	COP
60	AME	CERN	CAH	85	ORCL	WMT	STT
61	BAC	C	ABT				
62	SWK	TROW	CMCSA				
63	SWK	TROW	KLAC				
64	CMCSA	SWK	KLAC				
65	DOW	HD	PNC				
66	DOW	HD	PCG				
67	DOW	PNC	PCG				
68	DOW	HD	FITB				
69	DOW	HD	CCE				
70	FITB	HD	CCE				
71	DOW	PNC	BA				
72	PCAR	WEC	SO				
73	PCAR	WEC	TJX				
74	PCAR	SO	TJX				
75	STI	VNO	TJX				

8.4 Appendix 4 – Gauss Inequality

According to [14] equation (9'), the Gauss Inequality states

$$\Pr\left(\left|X-\mu\right|\geq\kappa\sigma\right)\leq\frac{4}{9}\frac{1+s^2}{\left(\kappa-s\right)^2} \quad \text{if } \kappa>s \ , \tag{8-1}$$

where X is a random variable having a density function with a single mode, μ_0, with mean

μ, variance σ^2, and $s=\left|\dfrac{\mu-\mu_0}{\sigma}\right|$. When the mode equals the mean, $s=0$ and (8-1)

simplifies to

$$\Pr\left(\left|X-\mu\right|\geq\kappa\sigma\right)\leq\frac{4}{9}\frac{1}{\kappa^2} \quad \text{if } \kappa>0 \ , \tag{8-2}$$

providing a tighter bound than the more general Tchebycheff Inequality [14]. Using (8-2), probability bands, Table 8-1, similar to the Normal standard deviations (68.3–95.5–99.7%) may be calculated with κ being the number of standard deviations, σ. Gaussian bands are provided for comparison. Clearly, the fatter tails of the non-Gaussian distribution extend the probabilities of data occurring farther away from the distribution peak.

Table 8-1. Probability Bands

κ	Non-Gaussian Probability (%)	Gaussian Probability (%)
1	55.56	68.27
2	88.89	95.45
3	95.06	99.73
4	97.22	99.99
5	98.22	100.00
6	98.77	100.00
7	99.09	100.00
8	99.31	100.00
9	99.45	100.00
10	99.56	100.00