Comparative Performance of ARIMA and ARCH/GARCH Models on Time Series of Daily Equity Prices for Large Companies

John J. Sparks and Yuliya V. Yurova

Department of Information and Decision Sciences
University of Illinois at Chicago

ABSTRACT

This study provides a comparison of the performance of out-of-sample forecasts from ARMA vs. ARCH/GARCH models, especially relative to the utility of the non-constant estimate of the volatility provided by ARCH/GARCH methods. 31 large company stocks were selected and their daily log returns computed for a 10-year time period. ARIMA and a variety of ARCH/GARCH models were run against these log daily returns. The forecasts were analyzed using magnitude measures and a rough distributional measure. Analysis of the results showed that for one-step ahead forecasts ARCH/GARCH models outperform ARIMA models in modeling financial time series in terms of the most applied measure—the MAPD (mean absolute percentage deviation). Relative to the measures which accounted for the non-constant volatility estimate provided by ARCH/GARCH models, we saw that the mediocre performance of ARIMA on the MASE (mean absolute standardized error) measure came at a great expense in terms of the normality of the residuals. Therefore in the confines of this experiment ARIMA did not provide point estimates that were as accurate and that did not produce a relatively normal distribution of residuals.

INTRODUCTION

One of the primary benefits of ARMA models is their ability to correct for local trends in the data and "smooth away" patterns. One of the distinct features of financial time series is the non-constant volatility of the data (Tsay, 2002, p.80). The relative benefit of ARCH/GARCH modeling compared to ARMA modeling is that ARCH/GARCH provides a non-constant estimate of the volatility of the series. The aim of this research is to analyze the utility of the non-constant variance estimate when applied to log returns for a variety of U.S. equities using daily data and up to 6 step-ahead forecasts.

Two different classes of evaluation were analyzed for this study. In terms of magnitude measures, the models were evaluated using four different loss functions (Wei, 2005, p. 181 and Tsay, 2002, p. 163). As a rough distributional measure we assessed the normality of the forecast residuals (Tsay, 2002, p. 164). In order to minimize the affect of any outliers in the forecast or residual distributions, we used a moderately large sample of stocks that had robust daily trading.

This document is organized as follows. In Section 1, we describe the data selected for the study. Section 2 describes the 104 potential time series models under consideration. The forecast procedure and its measures are defined in Section 3. We present our empirical results in Section 4. Section 5 contains some concluding remarks.

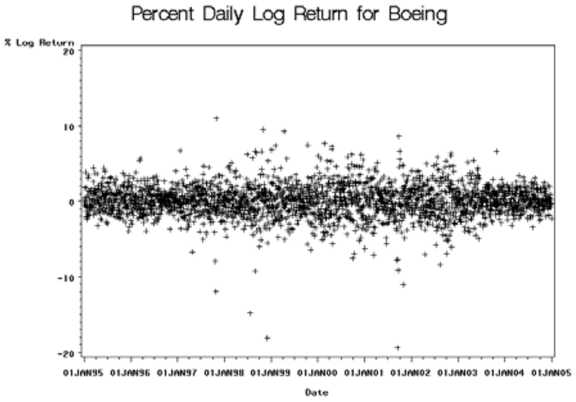
1. DATA

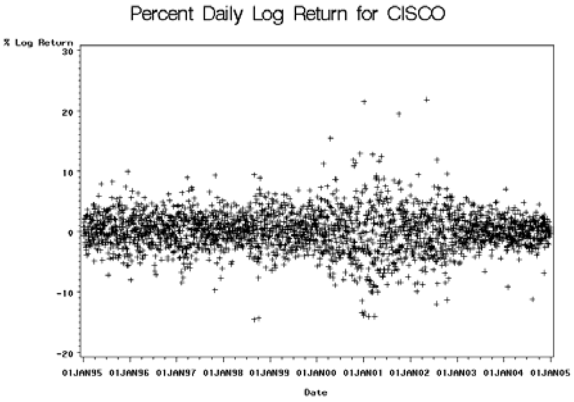
We obtained stock price data from the Center for Research of Security Prices (CRSP) of the University of Chicago. We chose to use daily data for selected large companies whose stocks are frequently traded to avoid missing information or other anomalies in the time series data. Daily data was pulled from the first trading day of 1995 to the last trading day of 2004. The typical company therefore had a series of 2,519 observations.

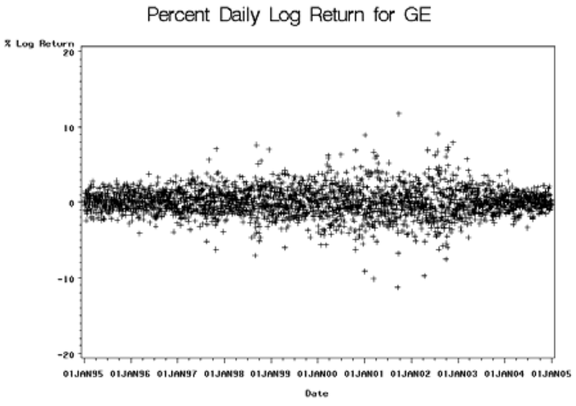
We selected a diverse set of companies in order to better represent the behavior of stock prices of large companies in the general market. A total of 31 companies ended up on our list. The companies, grouped in accordance with the Standard Industrial Classification (SIC) system, were selected with the intent of generally representing all major sectors of the economy. The 31 were composed of 18 manufacturing companies, 5 transportation and communication services companies, 3 financial services companies, and 5 trade and services companies.

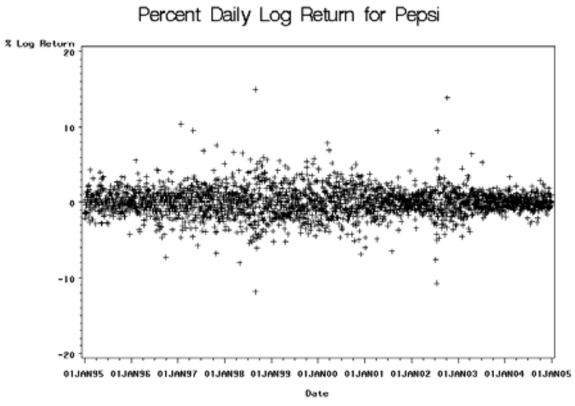
The price series were adjusted for stock splits and dividends to avoid atypical price changes and smooth the series. For completeness, please note that we defined our variable for study as follows: Given a price process, p_t, the percentage daily log return is defined by $r_t = (\log(p_t) - \log(p_{t-1})*100$, for $t = 2$ to $N + n$, where N is the number of observations used to fit the model and n is the number of forecasts.

Note that even though the percent daily log return produces a stable series in terms of a constant mean, the series still exhibit periods of non-constant volatility. This is demonstrated with graphs of four of the 31 stocks used in our study. Periods of increased volatility are seen in all four of these series as evidenced by periods of greater dispersion in the individual data points.


Percent Daily Log Return for Boeing


Percent Daily Log Return for CISCO





The table below shows summary statistics for the log of percent returns. The log returns moderately deviate about zero and reasonably symmetric, but kurtosis figures are generally large. This is in agreement with common findings in the literature regarding stock prices (Tsay, 2002, p.17, Campbell, Lo and MacKinlay 1997). Included on this table are the Hinich measures for linearity and normality. The implications of the Hinich tests are discussed next.

		Number of		Standard		Excess			Hinich Test:	Hinich Test:
Company Name	Start Date	Observations	Average	Deviation	Skewness	Kurtosis	Minimum	Maximum	Linearity	Gaussianity
Oracle Systems Corp	1/4/1995	2,519	0.074	3.7	-0.19	7.47	-34.5	27.1	9.8	27.2
Coca Cola Co	1/4/1995	2,519	0.019	1.7	-0.14	3.60	-11.1	9.2	10.4	25.8
Coca Cola Bottling Co Cons	1/4/1995	2,519	0.031	1.8	0.00	4.09	-9.9	10.8	9.9	24.7
Comcast Corp	1/4/1995	2,519	0.058	2.7	0.15	2.80	-15.2	14.8	12.2	33.0
General Electric Co	1/4/1995	2,519	0.058	1.9	0.03	3.24	-11.3	11.7	9.2	26.5
General Motors Corp	1/4/1995	2,519	0.008	2.0	-0.06	2.67	-14.5	9.8	10.5	22.6
International Business Machs Cor	1/4/1995	2,519	0.067	2.2	0.02	5.94	-16.9	12.4	8.9	21.3
Pepsico Inc	1/4/1995	2,519	0.045	1.8	0.34	6.07	-11.8	15.0	11.5	32.9
Apple Computer Inc	1/4/1995	2,519	0.048	3.7	-2.73	62.58	-73.1	28.7	21.8	80.3
Hershey Foods Corp	1/4/1995	2,519	0.060	1.6	0.97	19.92	-12.8	22.5	12.4	35.1
Boeing Co	1/4/1995	2,519	0.032	2.2	-0.62	7.24	-19.4	11.0	12.3	26.1
Abbott Laboratories	1/4/1995	2,519	0.044	1.9	-0.31	5.51	-17.6	11.7	13.6	28.8
Dow Chemical Co	1/4/1995	2,519	0.032	1.9	0.08	3.62	-11.2	10.8	9.4	29.3
Pfizer Inc	1/4/1995	2,519	0.057	2.0	-0.20	2.12	-11.8	9.3	9.0	14.0
Merck & Co Inc	1/4/1995	2,519	0.023	1.9	-1.80	28.84	-31.2	9.2	15.0	44.4
Hilton Hotels Corp	1/4/1995	2,519	0.028	2.3	-0.31	9.82	-26.9	14.6	12.1	33.8
Ford Motor Co Del	1/4/1995	2,515	-0.007	2.5	-2.44	50.74	-46.1	14.5	13.0	64.0
Cooper Tire & Rubber Co	1/4/1995	2,519	-0.005	2.2	0.23	4.30	-12.0	14.3	8.1	13.5
Xerox Corp	1/4/1995	2,519	0.002	3.3	-0.57	15.86	-29.8	33.0	20.6	75.5
Donnelley R R & Sons Co	1/4/1995	2,519	0.007	1.7	-0.23	3.93	-13.0	11.8	8.6	21.0
Mcdonalds Corp	1/4/1995	2,519	0.031	1.9	-0.08	4.28	-13.7	10.3	7.6	19.2
Host Marriott Corp	1/4/1995	2,519	0.030	2.1	-1.18	17.41	-28.3	12.2	16.5	56.7
Wal Mart Stores Inc	1/4/1995	2,519	0.064	2.1	0.10	2.17	-10.3	9.0	11.0	25.2
Southwest Airlines Co	1/4/1995	2,519	0.061	2.6	-0.30	6.90	-27.5	14.4	8.0	10.3
Century Telephone Entrprs Inc	1/4/1995	2,519	0.039	1.9	-0.42	7.45	-17.7	12.5	10.3	32.0
Federal Express Corp	1/4/1995	2,519	0.074	2.1	0.12	3.70	-15.5	11.0	8.6	14.5
Toys R Us Inc	1/4/1995	2,519	-0.016	2.6	0.37	5.51	-16.3	20.1	8.5	17.5
New Germany Fund Inc	1/4/1995	2,519	-0.009	1.8	-1.69	23.24	-26.7	8.4	13.4	52.4
Cisco Systems Inc	1/4/1995	2,519	0.092	3.3	0.11	3.50	-14.5	21.8	10.9	35.7
Delta & Pine Land Co	1/4/1995	2,519	0.068	2.7	-0.74	11.98	-30.2	15.9	9.7	29.8
Universal Holding Corp	1/4/1995	2,519	0.070	5.7	0.19	4.49	-31.8	34.8	32.1	42.2

One of the assumptions of the ARIMA model is the presence of linear dependence in the observations of the series. The violation of this assumption will lead to false conclusions and must be tested additionally. We applied the Hinich procedure to the percentage log return series to tests whether the series have a linear structure (Hinich 1982).

The Hinich procedure uses the property of bispectrum that, when properly normalized, is constant over all frequencies and is zero under normality. Stokes and Neurburger (1998) suggested performing this test over a range of frequencies to see if the results of the test are robust to different bandwidths and use the average value as a final test statistic. For further details on the bispectral test, see Tsay (1986), Priestley (1988), and Wei (2005). The figures provided in Table 2 in the appendix present the averages of the linearity and Gaussianity tests over the grid of admissible frequency values.

Both the Hinich test for normality (or Gaussianity) and test for linearity are much greater than the critical value of 1.96 for all admissible frequency values and for the mean of these statistics. Therefore, we cannot accept the assumption about normality and linearity of the daily log returns series. This result agrees with our initial hypothesis about strong non-linearity of financial series that is better modeled with conditional heteroscedastic models than with linear ARMA models.

2. MODEL BUILDING AND ESTIMATION

The six types of models evaluated in this study were: autoregressive moving-average ARMA, conditional heteroscedastic ARCH, generalized conditional heteroscedastic GARCH, integrated IGARCH, exponential EGARCH and MGARCH (garch-in-mean). Numerous criteria for model comparison have been introduced in the literature. Bayesian extension of the minimum Akaike's information criterion suggested by Schwartz, or SBC, has become a standard tool in time series fitting as it assesses the quality of the model fitting by suggesting the number of parameters in the model among all adequate models. It was introduced by Schwartz (1978) and takes the following form: $\text{SBC} = \ln(|\sum|) + r\ln(N)/N$; where r denotes the number of parameters estimated, N is the number of observations used to fit the model, and $\sum$ is the maximum likelihood estimate of the covariance matrix. In each model class a variety of models were run and the model with the lowest Schwartz's Bayesian criterion (SBC) was selected as the 'winner' for that type (Wei, 2005, p. 156).

For ARMA, AR orders 0 through 3 were run against MA orders 0 through 3 for a total of 16 models. In addition, AR terms for the original series were allowed to vary from 0 through 3 with a backwards elimination of non-significant terms. For the GARCH-type models, generally models of order (1,1), (1,2), (2,1) and (2,2) were run for the estimate of the conditional variance. These numbers of terms for the GARCH models were based on the studies of Engle (1990), Day and Lewis (1992), Tsay (2002, p. 95), which suggest that the GARCH models do not have to be complicated. A simple model with $r<=2$ and $s<=2$ is often sufficient to provide significant improvement over the traditional homoscedastic models.

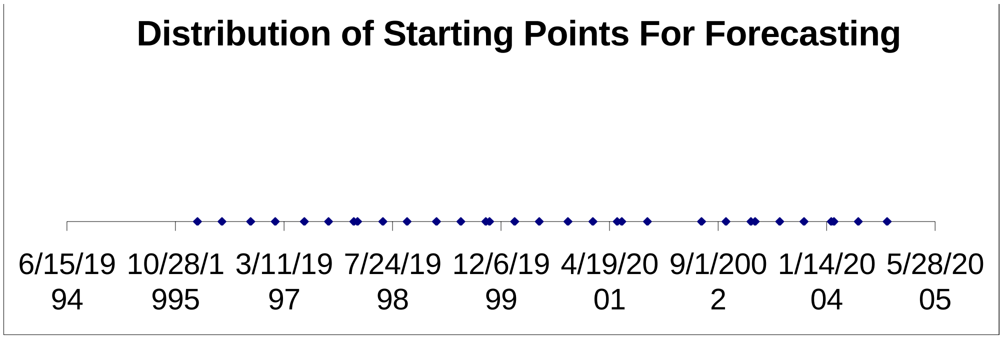
The total number of models built is detailed in the table below:

Model Type	**Number of Models**
ARMA	16
GARCH	32
ARCH	8
MGARCH	16
EGARCH	16
IGARCH	16
Total	**104**

Obviously, this large number of models made it necessary to automate the model selection process. This was accomplished using SAS. The programming effort to automate these procedures was significant. With the automation completed, however, we were able to run all models against all 31 stocks—and could run them against additional stocks in the future. As mentioned previously, the model within each model-type with the lowest BIC was retained and used in the forecasting stage. Note that inspection of the number of parameters included in the final models showed that the selected models generally contained an adequately small number of parameters. As such, we do not believe that these models over-fit the data.

Model Construction and Evaluation Time Period. If evaluation of the forecasts from the models took place during a time of unusually high volatility, then this could potentially bias the results in favor of the ARCH/GARCH models, which specifically model for changes in the volatility. In order to circumvent this problem the terminal day for the construction of the model and the beginning of the forecasting varied randomly for each stock. That is, all models were built for one stock using data from January 4, 1995 to a terminal date. The forecasting time period began on the next trading day after this terminal date, and 1-step to 6-step ahead forecasts were computed for the subsequent 6 trading days. Note that a minimum of 250 trading days (approximately 10% of the available data) was enforced for the construction of the model.

The graph below gives a visual representation of the starting days of the forecasting period (represented with the dots). They are spread relatively evenly across the eligible time period.



3. FORECAST PROCEDURE AND FORECASTS POWER MEASURES

For purposes of producing forecasts we split the sample into an estimation period (the first N observations) and an evaluation period (the next n observations, where n was 6 for this study). As noted previously, for each time series and for each model we generated 1- to 6-step ahead forecasts, totaling 1,116 forecast points.

There are many ways to evaluate the forecasting performance of a model and there is no widely accepted single measure to compare models. We employed two types of methods commonly used in the literature: magnitude measures and a rough distributional measure. The magnitude measures are based on various loss functions, and are used when the main purpose of a model is to forecast future values (see for example Tsay, 2002, p. 163, Wei 2005, p. 181). Magnitude measure or loss functions involve evaluating how far the evaluated forecasts differ from the observed values.

We used four loss functions to measure performance of point forecasts. They are the mean squared error (MSE), mean absolute deviation (MAD), mean absolute percentage error (MAPE), and mean absolute percentage deviation (MAPD) defined below:

mean square error or $MSE = 1/M\sum (r_{N+l} - r_N(l))^2$
mean absolute deviation $MAD = 1/M \sum | r_{N+l} - r_N(l)|$
mean absolute standardized error $MASE = 1/M \sum (|r_{N+l} - r_N(l)|) / \sigma_a$
mean absolute percentage deviation or bias $MAPD = 1/M \sum (|r_{N+l} - r_N(l)|) / r_{N+l}$,

where r_N is the percentage log return at time N, N is the number of observations used to fit the model, l is the forecast horizon, and M is the number of the l-step ahead forecasts available in the forecasting subsample.

The model with the smallest magnitude on the measure is regarded as the best *l*-step ahead forecasting model. Clements and Hendry (1993) discussed in detail the trade offs of using these measures when comparing forecasting performance of statistical models. One of the major shortcomings of magnitude measures is that it is possible that different forecast horizons may result in selecting different models. Thus, we complimented our analysis with a distributional measure, which tests whether the residuals from the forecasts follow a normal distribution (Tsay, 2002, p. 164 and Stokes, 1997, p. 286). We therefore examined the normality of the residuals for the different methods for the different forecasts.

The use of ARCH/GARCH modeling plays a role in both of these types of evaluation measures. GARCH type models provide a varying estimate for the volatility of the series for every forecast point, as opposed to ARMA models that use a constant variance estimate. If the estimate of the volatility from the GARCH model is superior, then we would expect to see larger residuals for the point estimate associated with larger variance estimates. As such, any standardized residual from the GARCH estimation will be smaller than the standardized residual from the ARMA model. Thus we would expect the t-scores of the residuals to be smaller in magnitude for the GARCH models as opposed to the ARMA model. Also, if larger residuals are associated with larger variance estimates then the standardized residuals from the GARCH models will contain fewer outliers and therefore be more standard.[1]

4. EMPIRICAL RESULTS

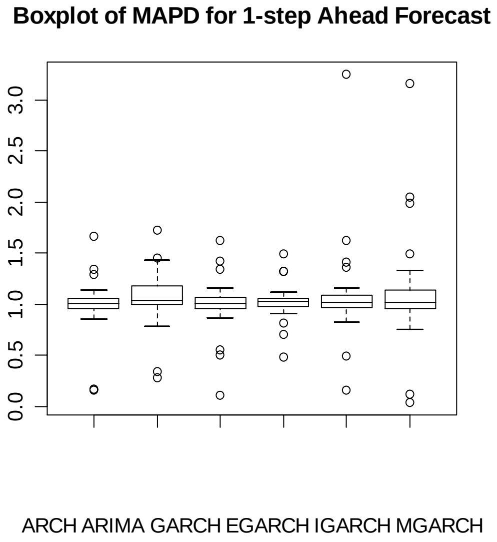
Different loss functions are more or less appropriate for different analyses depending on the business and research objectives of the study. Here we will focus on the MAPD because its interpretability is relatively straightforward—by what percent did we miss the target. Also, since we are focusing on the utility of the non-constant volatility estimate provided by ARCH/GARCH type of models we should also examine the MASE for reasons outlined in the previous paragraph. Additionally, as mentioned, we need to evaluate the normality of the residuals. This is done using the Jarque-Bera statistic for the standardized residuals, which uses the Lagrange multiplier procedure or score test on the Pearson family of distributions (Jarque and Bera, 1987).

Also, because the program that produced these figures runs in less than an hour on a standard PC, it could be incorporated into a process to produce estimates for these stocks on a daily basis.

[1] Assuming that both types of models are roughly equally unbiased.

Therefore, we will initially examine the 1-step ahead forecasts because they can be produced on a rolling basis if need be.

Summaries for the distribution of the 1-step ahead MAPDs are shown below.



Average MAPD for 1-step Ahead Forecasts

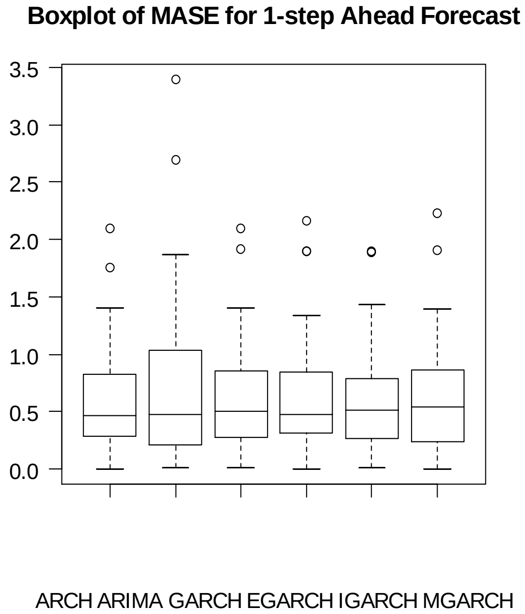
Model	MAPD
ARCH	**0.9838**
ARIMA	1.1345
EGARCH	1.0203
GARCH	0.9914
IGARCH	1.0848
MGARCH	1.1088

ARIMA model has the highest average MAPD and appears to have a relatively wide dispersion of MAPD figures as shown by the inter-quartile range. As such, it appears to have inferior MAPD performance for one-step ahead forecasts.

We now examine the measures that take the varying amounts of the estimate for the volatility into account. First, we analyze the mean of the absolute value of the standardized residuals (MASE).

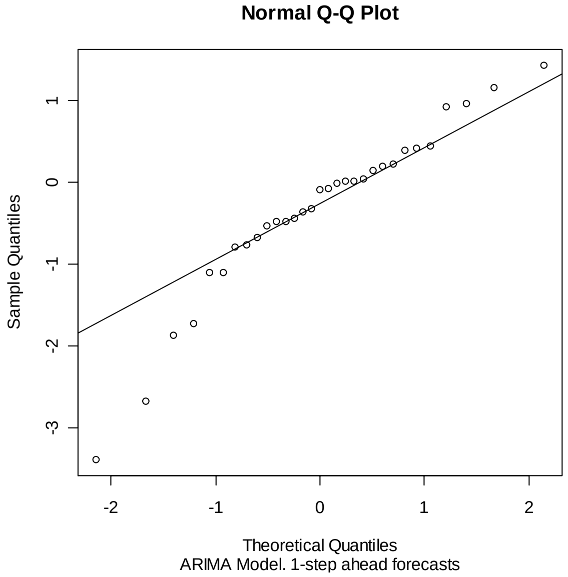
Average MASE for 1-step Ahead Forecasts

Model	MASE
ARCH	**0.4691**
ARIMA	0.4763
EGARCH	0.4702
GARCH	0.5010
IGARCH	0.5130
MGARCH	0.5448



From these we see that ARIMA model has an average MASE that is not particularly inferior or superior. The box plot shows, however, that the dispersion of MASE for ARIMA is the greatest as measured by the inter-quartile range. ARIMA also has the most significant outliers. So, in terms of the average MASE, the single estimate of volatility did not particularly hurt ARIMA's performance for this measure. The distribution of the MASE for ARIMA, however, did show a greater dispersion vs. the other model classes.

Next we examine the normality of the standardized residuals through the Jarque-Bera statistic.

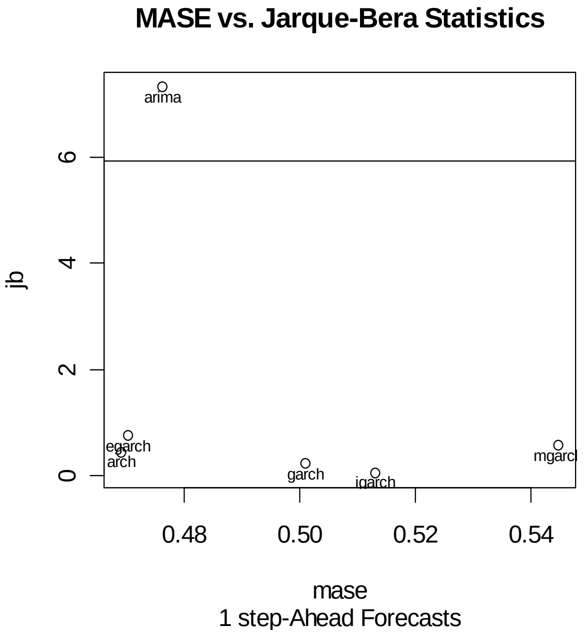


JB Statistic for 1-step Ahead Forecasts

Model	JB *p*-value
ARCH	0.8004
ARIMA	*0.0258**
EGARCH	0.6873
GARCH	0.8922
IGARCH	**0.9697**
MGARCH	0.7516

The table of p-values for the JB statistic shows that all ARCH/GARCH type models produced normal residuals while the ARIMA model did not. This non-normality can be seen on the Q-Q plot above for the ARIMA model. This plot shows outliers on the low side of the distribution. Therefore, ARIMA here distinguished itself in terms of inferior performance on the normality measure.

Finally we examine the MASE and normality measure simultaneously. This is demonstrated in the X-Y plot provided below.


MASE vs. Jarque-Bera Statistics

Note that the line on the graph above is the critical value for .05 level normality test for the JB statistic (points below this line pass the normality test). This graph shows that the middle-of-the-road performance produced by ARIMA on the MASE measure comes at a great expense in terms of normality of the residuals. Further we see that ARCH and E-GARCH have strong performance on both measures and that their performances are basically indistinguishable from each other. Therefore, these two are the best techniques using these two dispersion-based measures.

We provide a general description of some of the results for the 2-step to 6-step ahead forecasts. For 2-step ahead we saw that ARIMA was the only technique that actually produced normal residuals. In terms of the magnitude measures ARIMA did have the lowest MAPD figure. For the remaining magnitude measures, however, the ARCH/ GARCH type models were superior.

For the additional step-ahead forecasts we saw two general findings. First, normality of the residuals degraded as the forecast horizon increased. Also, ARCH/GARCH models tended to have superior magnitude measures except at the 4th step-ahead forecast where ARIMA had the best performance on all of the figures.

CONCLUSIONS

One of the distinct features of financial time series is the non-constant volatility of the data. ARCH/GARCH models were developed to better account for this non-constant behavior (compared to ARIMA models). So the non-constant volatility models should provide superior forecasting ability. That is, ARCH/GARCH models should provide residuals that are more normal and have smaller magnitude when scaled by the estimated volatility of the series. This was the general hypothesis of this research.

Here we compared a large number of models in terms of their ability to forecast the return in an out-of-sample setting. Our analysis used 31 U.S. stock return series and included 104 different ARCH/GARCH and ARIMA type models. The main finding is that for one-step ahead forecasts ARCH/GARCH models outperform ARIMA models in modeling financial time series in terms of the most applied measure—the MAPD. Relative to the measures which accounted for the non-constant volatility measure provided by ARCH/GARCH models, we saw that the mediocre performance of ARIMA on the MASE measure came at a great expense in terms of the normality of the residuals. Therefore in the confines of this experiment ARIMA did not provide point estimates that were as accurate and that did not produce a relatively normal distribution of residuals.

In terms of limitations and future research please note that while the sample used here was substantial and diverse amongst large companies, additional stocks and measures (e.g. monthly data) can also be examined to determine whether these results generalize beyond the specifics of this study. Given that a single program produces all the figures for this study, the number of stocks analyzed in this manner can be greatly increased without much additional effort from the researchers. In addition, the program can be amended to produce rolling one-step ahead forecasts, which would increase the number of data points available for analysis. We are aware that the results seen here are partly a result of limited time available for the study from the researchers for acquisition, inspection and cleansing of the data.

REFERENCES

Akaike, H. (1978). A Bayesian Analysis of the Minimum AIC Procedure, *Annals of the Institute of Statistical Mathematics*, 30A, 9-14.

Akaike, H. (1979). A Bayesian Extension of the Minimum AIC Procedure of Autoregressive Model Fitting. *Biometrica*, 66, 237-242.

Campbell, J.Y., Lo, A. W., and MacKinlay, A.C. (1997). *The Econometrics of Financial Markets*, Princeton University Press: New Jersey.

Clements, M.P., and Hendry, D.F. (1993). On the Limitations of Comparing Mean Square Forecast Errors. *Journal of Forecasting*, 12, 617-637.

Day, T.E., and Lewis, C.M. (1992). Stock Market Volatility and the Information Content of Stock Index Options, *Journal of Econometrics*, 52, 267-287.

Engle, R. F., Ng, V.K., and Rotschild, M. (1990). Asset Pricing with a Factor-arch Covariance Structure: Empirical Estimates for Treasury Bills, *Journal of Econometrics*, 45, 213-237.

Hinich, M. (1982). Testing for Gaussianity and Linearity of a Stationary Time Series. *Journal of Time Series Analysis*, 3, 169-176.

Jarque, C. M., and Bera, A. K. (1987) A Test for Normality of Observations and Regression Residuals, *International Statistical Review*, 55, 2, 163-172.

Priestley, M.B. (1988). *Non-linear and Non-stationary Time Series Analysis*, Academic Press: London.

Schwartz, G. (1987). Estimating the Dimension of a Model. *Ann. Statist*., 3, 13-41.

Stokes, H.H. (1997). Specifying and Diagnostically Testing Econometric Models, *Greenwood*

Stokes, H.H. and Nueburger, H.M.(1998). New Methods in Financial Modeling: Explorations and Applications, *Greenwood*

Tsay, R.S. (2002). Analysis of Financial Time Series, *Wiley Series in Probability and Statistics*

Tsay, R. S. (1986). Nonlinearity Tests for Time Series. *Biometrika*, 73, 461-466.

Wei, W.W.S. (2005). Time Series Analysis: Univariate and Multivariate Methods, *Pearson Education.*

Acknowledments. Authors are thankful to Dr. Ruey S. Tsay, University of Chicago, and Dr. Houston H. Stokes, University of Illinois at Chicago, for graciously providing their guidance and valuable research advice.