ANLY 560 Time Series Final Project
Stock Market Index Analysis and Prediction

Jiaqi Tang, Xinyi Ye, Yue Han
MS in Analytics, Georgetown University
3700 O St NW, Washington, DC 20057
{Jt1215, xy142, yh584} @georgetown.edu

Abstract

A market index acts like a barometer which shows the overall condition of a segment of the financial market. It facilitates the investors to gauge the general moving pattern of the market. Understanding different market indexes is necessary to decide about which stocks to go for investing. In this paper, we explore three most broadly followed indexes -- Dow Jones Industrial Average (DJI), S&P 500 (SP500) and Nasdaq Composite(Nasdaq) -- in the United States. We study these three indexes with technical analysis, or exploratory data analysis. Following that, we apply two time series models-ARIMA model and Volatility model-to make prediction in order to make more informed and accurate investment decisions. The ARIMA model is used for forecasting prices, and the Volatility model is built for analyzing volatility. The result suggests that 1. time series can be used to statistically predict short-term stock trends, 2. the ARIMA models with the drift item perform the best, 3. the Nasdaq is most volatile compared with the other two, etc.

Introduction

Prediction of both stock price and its investment risk is the key issue for investors to form investment decisions within the stock market. An insightful understanding of the market index with a successful prediction could facilitate yield significant profit. In the financial market, time series data tracks the movement of the chosen stock or index. Time series forecasting makes use of statistics regarding historical values and associated patterns to predict future activity. The objective of this paper is to study the stock market index with Exploratory Data Analysis and predict the stock market with time series technique so as to make more informed and accurate investment decisions. Three stock market index--Dow Jones Industrial Average, S&P 500 and Nasdaq-in the past ten years are surveyed. In the United States, these are the three most widely followed indexes by both the media and investors. The media usually reports based on the movement of these three indexes throughout the day, along with vital news serving as contributors and detractors.

We will begin with studying and comparing the three indexes technically with Exploratory Data Analysis. We will conduct some baseline analysis, including introducing what each index represented, analyzing historical data, and comparing their trading volume, then we will dive deeper to study the risk, by calculating returns, comparing volatilities, and finding the correlation between each index. Then we build two time series models --ARIMA model and Volatility model-- to predict the index. ARIMA is used for the forecast of closing prices. For each index, several ARIMA models with different parameters are compared based on

methodologies, efficiency and prediction results, and then it is represented in the form of a Graph. Different types of Volatility models are built to analyze volatility, such as GARCH and stochastic volatility, Along with it, we get some new insights from these indexes.

Dataset

This project includes three datasets from Yahoo Finance, and each of them are the past ten years' (01/01/2009- 12/31/2019) daily data of our three selected market index--Dow Jones Industrial Average, S&P 500 and Nasdaq Composite. Each dataset has 7 columns, including Date, Open, High, Low, Close, Adj Close price and Volume, as well as 2767 rows of record.

Exploratory Data Analysis

I. Baseline Analysis

We start with analyzing the time series data for the three market indexes in the past ten years, visualized with an interactive plot you can play with (Fig.1.1, see Appendix A for interactive plot). You can zoom in to see details of stock price within a smaller period range, and zoom out to see the whole trend of the stock history. If you hover your mouse on it, you can also see the specific price of each stock each day. In a broader picture, we can observe an upward trend of each stock index within the past 10 years. And DJI has a significantly higher index level than that of Nasdaq and SP500. That's because the "Dow" includes only 30 stocks, all of which are among the largest, richest and most heavily traded companies in the United States, and also because DJI is price-weighted, therefore, DJ is affected only by changes in the stock prices, so companies with a higher share price or a more extreme price movement have a greater effect on the Dow. SP 500 is more covering, because it is calculated based on a larger sample of total U.S. stocks. Stocks within the SP 500 are weighted by their market value rather than their stock prices. In this way, the SP 500 attempts to ensure that a 10% change in a $20 stock will affect the index in the same way as a 10% change in a 50 dollar stock will. The Nasdaq represents the largest non-financial companies listed on the Nasdaq exchange and is generally regarded as a technology index given the heavy weighting given to tech-based companies.

Fig.1.1 DJI, Nasdaq, SP500 Index Price 2009-2019

Then we investigate the trading volume of each index and stack them over for comparison. This stacked area chart shows the stacked trading volume of the stocks under those three indexes. SP500 represents the broadest measure of the U.S. economy among the three major indices.

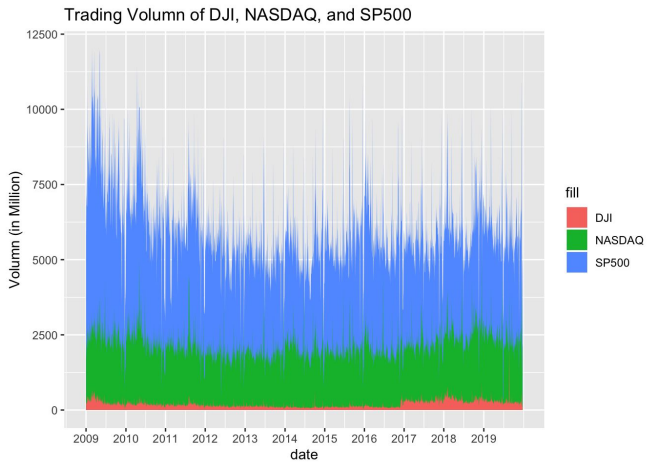

Fig.1.2 Trading Volume

The index includes 500 companies from all sectors of the economy with stocks listed on various stock exchanges. So combined, the companies in the SP500 account for about 75 percent of all U.S. stock. Since the Dow only represents 30, its total trading volume obviously takes a much lower portion on the market.

II. Risk Analysis

Now that we've finished conducting some baseline analysis, let's go ahead and dive a little deeper to analyze the risk of the stock. To investigate the risk, we'll need to take a closer look at the daily changes, ie. daily returns, of the stock, and not just its absolute value. Once returns are generated, we will be able to understand the growth and volatility of the stock. The mean or average of the daily returns is a measure of growth, and the variance of the daily returns is a measure of volatility. These are important because we can use historical return performance to screen and evaluate stocks using a risk-reward profile.

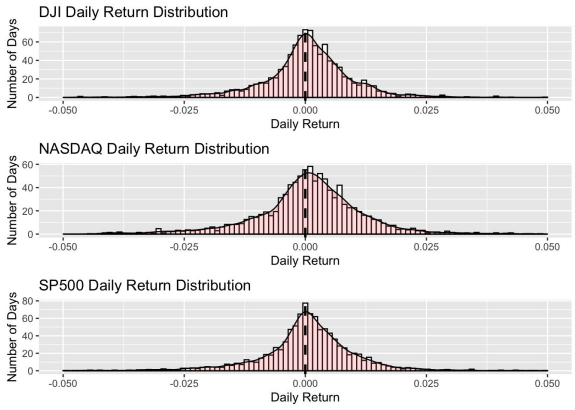

Fig. 1.3 Return Distribution

From the return time series plot (fig.1.4), we can observe the volatility of each stock. With calculation, the Dow Jones is the least volatile (var = 0.96) of the three major indexes as many of its composing companies are blue-chip but slower moving companies such as Walmart, Boeing, and the Coca-Cola Company. The Nasdaq is the most volatile (var = 1.16) among the three indexes, mainly because of its high concentration in riskier, emerging but high growth technology companies such as Google, Facebook, and Amazon. The volatility of the SP500 (var = 1.03) is typically somewhere between the two.

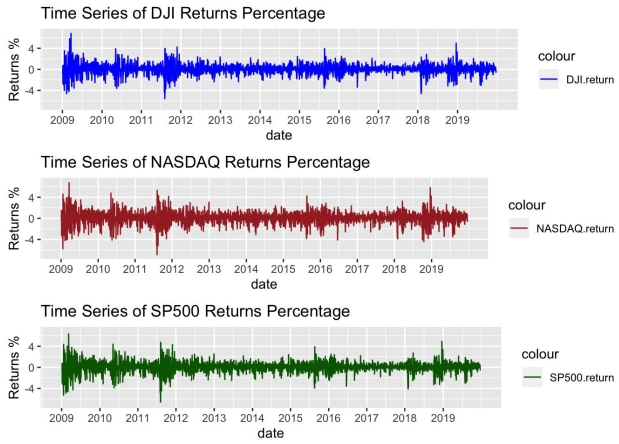

Fig. 1.4 Returns Time Series

Correlation can be applied to gain perspective on the overall condition of the larger market. We applied Pearson's Correlation to calculate correlation between those three stock indexes. Formulas that calculate correlation can predict how two stock indexes might perform relative to each other in the future. Applied to historical prices, correlation can help determine if stocks' index prices tend to move with or against each other. Our results (fig. 1.5) show significant correlation between those three stock indexes. Each pair exhibits 91%, 95% and 97% degree of correlation, which means that they all moved basically in lockstep with each other. It indicates that the market in general moves in the same direction, because they track companies impacted by the same business cycle and other important macroeconomic factors.

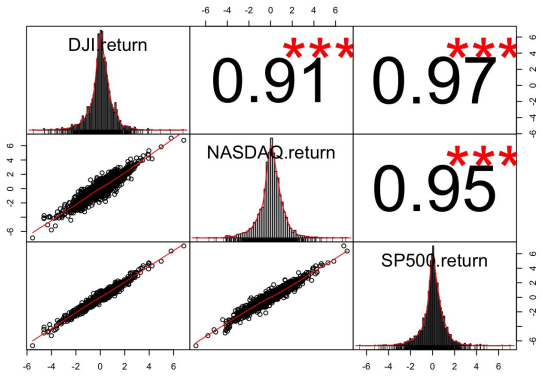

FIG.1.5 Correlation

Time Series Methodology

I. Data Preparation

Historical data of Dow Jones Index, S&P 500 and NASDAQ Composite during 2009 to 2019, retrieved from Yahoo Finance (https://finance.yahoo.com/). They are split to training set and test set as below:

Name	**Time Period**	
	Training Set	**Test Set**
Dow Jones Industrial Average	2009-01-01 to 2018-12-31	2019-01-01 to 2019-12-31
S&P 500		
NASDAQ Composites		

These datasets do not require cleaning since they are cautiously collected. There are no missing values, incorrect values, values with improper formatting, inconsistent values and outliers in the datasets.

Since the magnitude is large in all datasets, a *logarithm* transformation is applied to the data.

II. Model Description

Autoregressive Integrated Moving Average (ARIMA) Model:

ARIMA is the most widely used model in the finance sector. It was firstly introduced by Box and Jenkins in 1976[1]. A typical ARIMA model is characterized by 3 terms: p, d and q. p is the order of the autoregressive (AR) terms that represent the time series as a function of p past observations. d is the order of the integrated (I) terms model the differencing needed for the time series to become stationary. q is the order of the moving average (MA) terms that represent the number of moving average lags, controlling the past noise of the original time series. If $d = 0$, the model is reduced to an ARMA(p, q) model[2].

The autoregressive (AR) part,

$y_t = {}_1 y_{t-1} + \ldots + {}_p y_{t-p}$,

with p past observations and AR coefficients ${}_1 \ldots {}_p$.

The moving average (MA) part,

$y_t = {}_1 + \theta_1 {}_{t-1} + \ldots + \theta_q {}_{t-q}$,

with q moving average lags, MA coefficients $\theta_1 \ldots \theta_q$ and errors ${}_t \ldots {}_{t-q}$.

Finally, the integrated (I) part,

$y^*_t = y_t - y_{t-1} - \ldots - y_{t-d}$,

with the order of differencing of d.

When the model has differencing, a drift term can be included.

$$\phi(B)(1 - B)(X_t - ct) = \theta(B)e_t$$

,

Where $(B) = 1 - {}_1 B - \ldots - {}_p B^p$,

$(B) = 1 + {}_1 B + \ldots + {}_p B^p$ and c is the "drift" parameter.

To fit an ARIMA model, the time series should be stationary. To make the data stationary, the data is manipulated by the following steps (illustrated with DJI data only).

1. Differencing once and plot,

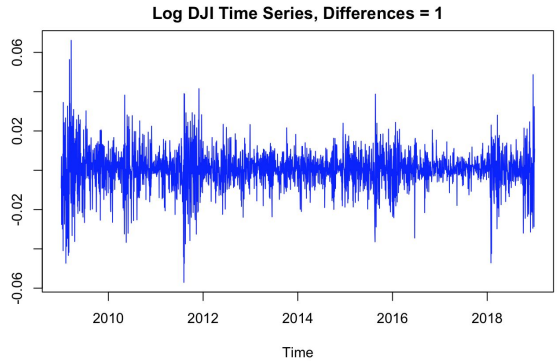

Fig. 2.1 Differenced log DJI training data, $d = 1$

Conduct Augmented Fuller-Dickey Test (fig. 2.2) to check the stationarity,

```
p-value smaller than printed p-value
        Augmented Dickey-Fuller Test

data:  dji.tsdiff1
Dickey-Fuller = -14.395, Lag order = 13, p-value = 0.01
alternative hypothesis: stationary
```

Fig. 2.2 Result of Augmented Dickey-Fuller test

With a p-value less than 0.01, the null hypothesis of non-stationary is rejected and the new time series can be considered as stationary.

2. ACF and PACF Plots

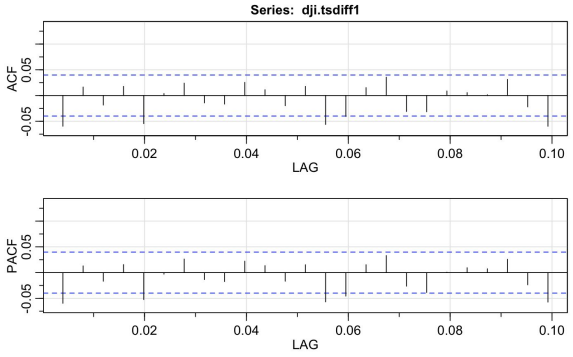

Fig. 2.3 ACF & PACF plots of stationary time series

According to the ACF and PACF plots (fig. 2.3), there is no cutoff. Thus, ARIMA(1,1,1) becomes the baseline model.

3. Drift Term

From the forecast plots, we can see the actual value is closer to the mean predicted value from the ARIMA(1,1,1) model with drift term, and lies completely within the 80% confidence range. With the results from analysis with other ARIMA models and parameter combinations, the models with a drift term always seem to have a better prediction in the forecast plots (fig. 2.4).

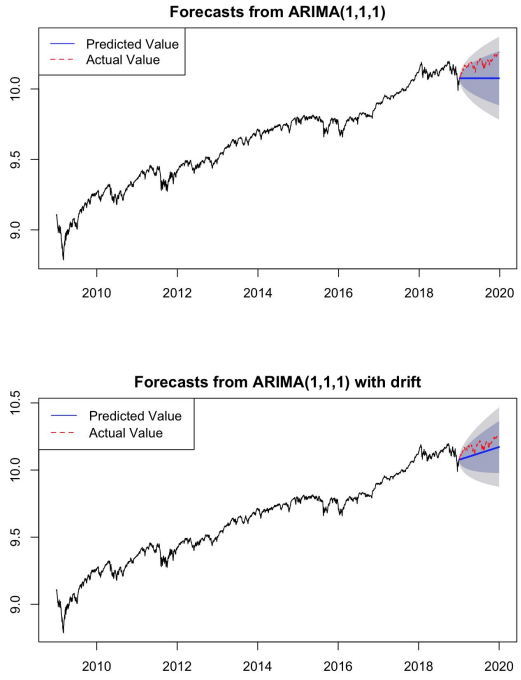

Fig. 2.4 Forecast plots of ARIMA(1,1,1) with/without drift

4. Residuals

Checking residuals is an important step to evaluate if the ARIMA model fits well.

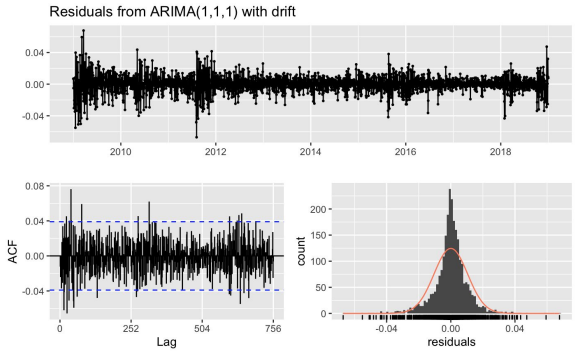

Fig. 2.5 Residuals plots of ARIMA(1,1,1) with drift

The residual plot (fig. 2.5) has 3 parts: the line plot of the residuals, the ACF plot and the histogram. As is shown in the residuals plot for the ARIMA(1,1,1) model with drift term, most

of the data are not significantly correlated and are normally distributed. Hence we can say that this model is a good fit.

Volatility Model:

Besides analyzing and forecasting price, understanding stock volatility with time series models is important. Volatility is a statistical measure of the dispersion of returns for a given security or market index. In most cases, the higher the volatility, the riskier the security.[10] Volatility is often measured as either the standard deviation or variance between returns from that same security or market index (Justin Kupper,2020). Thus, the volatility model and forecasting play a crucial role in measuring investment risks of various assets. In this project, there are two types of volatility models conducted for analyzing volatility, and there is an optimal one that has been selected as the final model.

Standard GARCH (1,1) :

In this project, first, the GARCH model has been utilized for volatility measurement. In the first part of the Volatility model, a standard GARCH model is estimated. [10]The GARCH (1,1) equations are described as:

$$Y_t = \mu + \delta\, Y_{t-1} + \varepsilon_t$$
$$h_t = \omega + \alpha\, \varepsilon_{t-1}^2 + \beta h_{t-1}$$

In the equation, Yt represents the return series with error term t being normally distributed having a mean of zero. Moreover, ht describes the conditional variance, and the mean volatility level, ht-1, the variance from the previous period. [10] The values of coefficients tend to be added to determine the persistence of shocks to volatility of each return series.

Stochastic Volatility:

Secondly, Stochastic volatility(SV) was processed to explain the shape of the implied volatility curve. Furthermore, with the assumption that a stochastic element in the time evolution of the conditional variance process. [9] The model introduced White (1987), and Heston (1993) assumes the following stochastic processes:

$$dS(t) = \mu S dt + \sigma(t) S dW_1(t),$$
$$v(t) = \sigma^2(t)$$
$$dv(t) = \phi v dt + \xi v dW_2(t),$$

where W1 and W2 are two correlated Wiener processes, μ, φ, and ξ are unknown parameters, and σ(t) is the return volatility process.[12] The Heston model assumes that the spot asset S(t) follows the diffusion process given by (1) and that the volatility σ(t) follows an Ornstein-Uhlenbeck process (OU), or that the variance process v(t) follows the square-root process (Cox, Ingersoll, and Ross)

$$dv(t) = k(\theta - v)dt + \xi\sqrt{v}dW_2(t),$$

where W1 and W2 are two correlated Wiener processes, and k (the speed of mean reversion), θ (long-run mean), and ξ are unknown parameters. Heston provides a closed-form solution for European options using Fourier transform methods, while Hull and White value European options in a series form for the case in which W1 and W2 are independent. Another example of diffusion models is the class of exponential linear models (Tauchen, 2004).[12] In this project, the basic model is given by :

$$\sigma_t = \exp(\xi_0 + \xi_1 v(t))$$
$$dv(t) = (\alpha_0 + \alpha_1 v)dt + \alpha_2 dW_2(t),$$

where ξ1, ξ2, α0, α1, and α2 are unknown parameters.

Table 2.1 Coefficients

COEFFICIENTS							
Index	**Model**	**p-value**	**ar1**	**ar2**	**ma1**	**ma2**	**drift**
DJI	ARIMA(0,1,1)				-0.0581		4e-04
	ARIMA(1,1,1)		-0.6894		0.6413		4e-04
S&P 500	ARIMA(1,1,1)		-0.7506		0.7014		4e-04
	ARIMA(1,1,2)		-0.8064		0.7442	-0.0236	4e-04
	ARIMA(2,1,1)		-0.8362	-0.0240	0.7746		4e-04
NASDAQ	ARIMA(1,1,1)		-0.0241		-0.0239		6e-04
	ARIMA(1,1,2)		-0.0242		-0.0237	-0.0013	6e-04
	ARIMA(2,1,1)		-0.8648	-0.0308	0.8174		6e-04

Table 2.2 Model Performance

MODEL PERFORMANCE									
Index	**Model**	**p-value**	**Set**	**ME**	**RMSE**	**MAE**	**MPE**	**MAPE**	**MASE**
DJI	ARIMA(0,1,1) AIC = -16163.66 BIC =-16146.17	0.5908	Training	3.3e-06	0.0097	0.0066	-6.1e-06	0.070	0.054
			Test	-0.016	0.025	0.020	-0.15	0.19	0.16
	ARIMA(1,1,1) AIC = -16163.85 BIC = -16140.52	0.5796	Training	8.2e-06	0.0097	0.0066	-4.6e-05	0.070	0.054
			Test	-0.015	0.025	0.020	-0.15	0.19	0.16
S&P 500	ARIMA(1,1,1) AIC = -15817.77 BIC =- 15794.44	0.08802	Training	1.7e-06	0.010	0.0070	-2.6e-05	0.096	0.056
			Test	-0.075	0.080	0.075	-0.94	0.94	0.60
	ARIMA(1,1,2) AIC = -15816.65 BIC = 15787.49	0.09588	Training	-1.5e-06	0.010	0.0070	-7.1e-05	0.096	0.056
			Test	-0.076	0.080	0.076	-0.95	0.95	0.61
	ARIMA(2,1,1) AIC = -15816.59 BIC = -15787.43	0.09482	Training	8.2e-06	0.010	0.0070	6.0e-05	0.096	0.056
			Test	-0.076	0.080	0.076	-0.95	0.94	0.61
NASDAQ	ARIMA(1,1,1) AIC = -15242.05 BIC = -15218.72	0.08114	Training	2.6e-06	0.011	0.0081	3.0e-06	0.10	0.050
			Test	0.071	0.077	0.071	0.79	0.79	0.44
	ARIMA(1,1,2) AIC = -15240.0 BIC = -15210.89	0.07766	Training	2.6e-06	0.011	0.0081	3.0e-06	0.10	0.050
			Test	0.071	0.077	0.071	0.79	0.79	0.44
	ARIMA(2,1,1) AIC =-15240.55 BIC = -15730.96	0.07676	Training	2.6e-06	0.011	0.0081	3.0e-06	0.10	0.050
			Test	0.071	0.077	0.071	0.79	0.79	0.44

* All analysis is done in Log value.
** Drift is included in all

Results

I. ARIMA

Table 2.1 is the table of model coefficients that applied in the analysis. By comparing different ARIMA models to select the best model for each dataset, the performance will be analyzed in 6 measurements: ME (Mean Error), RMSE (Root Mean Squared Error), MAE (Mean Absolute Error), MPE (Mean Percentage Error), MAPE (Mean Absolute Percentage Error), and MASE (Mean Absolute Scaled Error). The AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) values are also estimators to compare models.

ME (Mean Error)[3]

ME is a term that refers to the average of all the errors in a set.

$$ME = \sum_{i=1}^{n} \frac{\hat{y}_i - y_i}{n}$$

RMSE (Root Mean Squared Error)[4]

It aggregates the magnitudes of the errors in predictions for various times into a single measure of predictive power.

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

MAE (Mean Absolute Error)[5]

MAE is simply the average absolute vertical or horizontal distance between each point in a scatter plot and the Y=X line. In other words, MAE is the average absolute difference between X and Y.

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} = \frac{\sum_{i=1}^{n} |e_i|}{n}$$

MPE (Mean Percentage Error)[6]

MPE is the computed average of percentage errors by which forecasts of a model differ from actual values of the quantity being forecast.

$$\mathrm{MPE} = \frac{100\%}{n} \sum_{t=1}^{n} \frac{a_t - f_t}{a_t}$$

MAPE (Mean Absolute Percentage Error)[7]

MAPE expresses the accuracy as a ratio, the absolute value in this calculation is summed for every forecasted point in time and divided by the number of fitted points n.

$$MAPE = \frac{100\%}{N} \sum_{i=0}^{N} \frac{|y_i - \hat{y}_i|}{\hat{y}_i}$$

MASE (Mean Absolute Scaled Error)[8]

MASE is the mean absolute error of the forecast values, divided by the mean absolute error of the in-sample one-step naive forecast. It was proposed

$$\mathrm{MASE} = \mathrm{mean}\left(\frac{|e_j|}{\frac{1}{T-1}\sum_{t=2}^{T}|Y_t - Y_{t-1}|}\right) = \frac{\frac{1}{J}\sum_{j}|e_j|}{\frac{1}{T-1}\sum_{t=2}^{T}|Y_t - Y_{t-1}|}$$

in 2005 by statistician Rob J. Hyndman and Professor of Decision Sciences Anne B. Koehler, and has favorable properties when compared to other methods for calculating forecast errors.

The models for each parameter combination all have very similar performances.

II. Volatility Model selection

Through the implementation of Garch type models and stochastic volatility models to conduct volatility analysis, from Figure 3.1, it suggested that the stochastic volatility outperforms the others with the lowest errors at the testing set. In this case, the stochastic volatility model has been utilized as the model for analyzing the volatility model for these three indexes in this project.

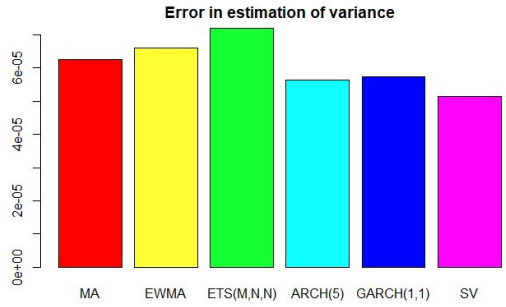

Figure 3.1 Errors comparison of different volatility models

Standard deviation forecasting results

According to figure 3.2,3.3,3.4 there are three envelope(standard deviation) plots that show the change of volatility varying time of these three stock indexes(Dow Jones, SP500, NASDAQ).

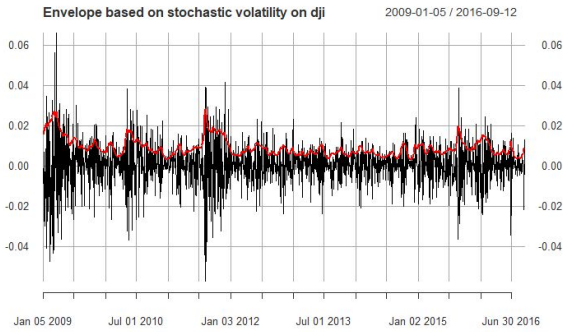

Figure 3.2 Envelope(standard deviation) forecasting based on stochastic volatility of Dow Jones

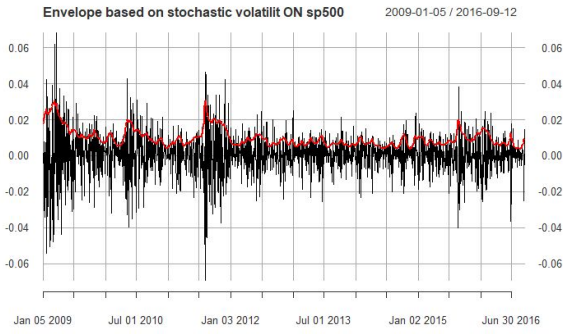

Figure 3.3 Envelope(standard deviation) forecasting based on stochastic volatility of SP500

Figure 3.4 Envelope(standard deviation) forecasting based on stochastic volatility of SP500

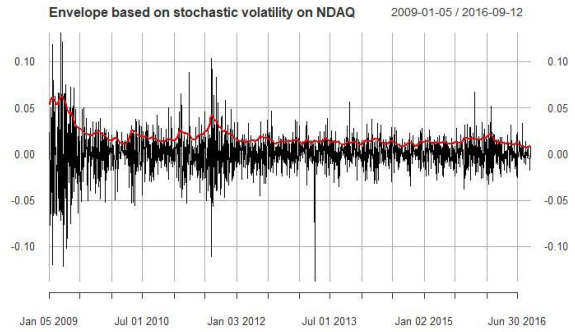

Figure 3.4 Envelope(standard deviation) forecasting based on stochastic volatility of Nasdaq

From the figure, it suggests that the volatility curves of Dow Jones and SP500 are extremely similar. In this case, the movement and volatility of returns of these two indexes are in the same trends among years. On the other hand, the volatility of NASDAQ has relatively higher volatility than other two indexes. However, from 2012, the volatility of NASDAQ shows a stable curve with less value of volatility. Therefore, NASDAQ are much less volatile and will be more stable after 2012. In addition, compared to the other two stock indexes, NASDAQ tends to be more stable from this figure with a more stable curve and less volatility.

Conclusions

In this project, there are different time series models utilized for analyzing stock price and predicting the price and volatility of stock indexes(Dow Jones, SP500, NASDAQ). Time series can be used to statistically predict short-term stock trends. For predictions of the price of stock indexes, the ARIMA models which include the drift item have better performance at forecast. However, the performances of models with different parameters are similar to each other in terms of AIC and BIC values as well as 6 measurements.

In addition, not only did this project predict the price of stock indexes, it also conducted different time series models to analyze and

froescate the volality of log return of these three stock indexes. With the lowest errors of the testing set, the Stochastic Volatility(SV) model overperformed than GARCH. Moveover, after performing volatility forecasting with the SV model, there are the same trends and volatile levels between Dow Jones and SP500. Compared with Dow Jones and SP500, NASDAQ are less stable. However, from 2012 , the forecasting volatility curve tends to be more stable , thus it will be less investment risk of NASDAQ. The results from this project are able to be used as references and information for investors.It will provide insights and guidance for them to make investment decisions with comprehensive quantitative time series analysis of price and volatility in this project.

References

1. Guitart, Anna, et al. "Forecasting player behavioral data and simulating in-game events." *Future of Information and Communication Conference*. Springer, Cham, 2018.
2. George EP, and Gwilym M. Jenkins. "Time series analysis. Forecasting and control." *Holden-Day Series in Time Series Analysis, Revised ed., San Francisco: Holden-Day, 1976*. 1976.
3. Mean Error: Definition https://lifehacker.com/insert-citations-in-multiple-formats-easily-with-google-1629738089
4. Root-mean-square deviation: https://en.wikipedia.org/wiki/Root-mean-square_deviatBoxion
5. Mean Absolute Error: https://en.wikipedia.org/wiki/Mean_absolute_error
6. Mean Percentage Error: https://en.wikipedia.org/wiki/Mean_percentage_error
7. Mean Absolute Percentage Error: https://en.wikipedia.org/wiki/Mean_absolute_percentage_error
8. Mean Absolute Scaled Error : https://en.wikipedia.org/wiki/Mean_absolute_scaled_error
9. Mihaela S¸erban. “Derivative Pricing under Multivariate Stochastic Volatility Models with Application to Equity Options.” 2006
10. S. Aun Hassan.“A Time Series Analysis of Major Indexes Using GARCH Model with Regime Shifts.” International Journal of Financial Research. 2017
11. Zhe Lin. “Modelling and forecasting the stock market volatility of SSE Composite Index using GARCH models”. Future Generation Computer Systems. 2018
12. C. Q. CAO,and R. S. TSAY. “Nonlinear Time-Series Analysis of Stock Volatilities.” JOURNAL OF APPLIED ECONOMETRICS. 1992

Appendices

Appendix A: Stock_Index_Analysis_EDA.html
Appendix B: ARIMA Models and Evaluations.
Appendix C: volatility_model.html
AppendixD:Stock_index_movement_correlation.html