STAT 5361 PROJECT REPORT

Portfolio optimization using Efficient frontier and applying ARIMA on the portfolio

Anand Joy
MSc Applied Financial Mathematics, Actuarial Science
Anand.joy@uconn.edu

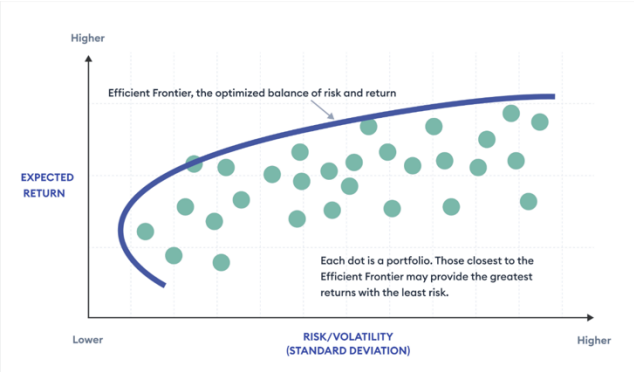
Introduction

In financial world, everyone wants to grow their wealth and they do that investing their capital into the market. People invest into different asset class such as real estate, bonds (Corporate and Sovereign), equity, and so on. But the risk of loosing the returns on investment is high if every penny is invested into single asset class, even when dealing with safer fixed income assets such as bonds. In order to lower the risk and improve investment, investors follow a principle of diversification of assets. This is one of the core principles in asset management especially in equity. Harry Markowitz proposed a theory in 1952 called the "Modern Portfolio theory (MPT)" which gives us the mathematical toolkit to get the maximum return for a set of stocks with the given risk tolerance.
However, past performance is not necessarily indicative of future performance, and there is no guarantee that a portfolio created using MPT will perform well in the future. The returns of a portfolio may change over time due to various factors, including changes in market conditions, changes in the underlying assets, and changes in investor sentiment.
The purpose of this project is to evaluate the performance of an efficient portfolio in future where the future values are forecasted using a time series model.

Background

Modern Portfolio Theory and Efficient frontier

MPT, or Modern Portfolio Theory, is a mathematical approach to investing that was developed by economist Harry Markowitz in the 1950s. It is based on the idea that investors can create portfolios of investments that offer the highest expected return for a given level of risk, or the lowest level of risk for a given level of expected return.



At its core, MPT is about finding the optimal balance between risk and reward. It suggests that investors can diversify their portfolios by including a mix of assets that are not perfectly correlated with each other, which can help to reduce overall risk. MPT also suggests that investors can use mathematical tools, such as the variance and covariance of different assets, to measure and compare the risk and return of different portfolios.
MPT has had a significant influence on the way that investors and financial professionals think about portfolio construction and risk management. It is widely taught in business schools and is used by investment professionals around the world. However, it has also faced criticism for its assumptions about investor behavior and its inability to fully capture the complexity of real-world financial markets.

Sharpe Ratio

The Sharpe ratio is a measure of risk-adjusted return that compares the return of an investment with its risk. It is calculated by subtracting the risk-free rate (such as the return on a risk-free asset such as

a Treasury bond) from the return of the investment and dividing the result by the standard deviation of the investment's returns.

$$Sharpe\ Ratio = \frac{R_p - R_f}{\sigma_p}$$

where:

R_p = return of portfolio

R_f = risk-free rate

σ_p = standard deviation of the portfolio's excess return

The Sharpe ratio can be used to compare the risk-adjusted returns of different investments or portfolios, with higher Sharpe ratios indicating a better trade-off between risk and return.

ARIMA

ARIMA models are a type of statistical model that are used to analyze and forecast time series data. They are a generalization of autoregressive moving average (ARMA) models, which are also used to analyze and forecast time series data. They are often used when the data show evidence of non-stationarity, which means that the mean of the data is not constant over time. In such cases, an initial differencing step (corresponding to the "integrated" part of the model) can be applied to eliminate the non-stationarity of the mean function (i.e., the trend). If the time series data show evidence of seasonality, the seasonal-differencing step can be applied to eliminate the seasonal component.

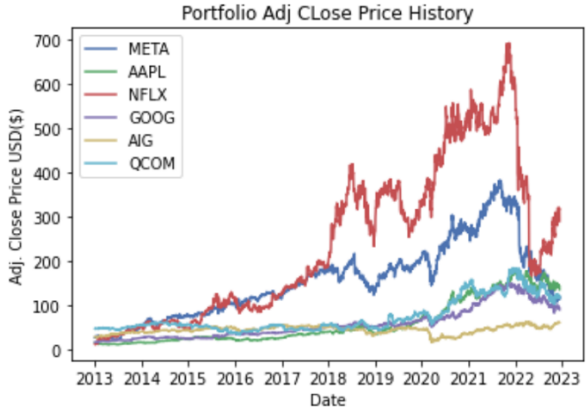
$$y'_t = c + \phi_1 y'_{t-1} + \cdots + \phi_p y'_{t-p} + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t$$

The syntax of these models are denoted by the notation ARIMA(p,d,q), where p is the order (number of time lags) of the autoregressive model, d is the degree of differencing (the number of times the data have had past values subtracted), and q is the order of the moving-average model.
Project

The project was conducted in two main stages: financial modeling using modern portfolio theory (MPT), and time series modeling and analysis.
First stage was focused on preparing the data for analysis and using MPT to construct portfolios that seek to maximize expected return for a given level of risk, or minimize risk for a given level of expected return. This involved analyzing historical data and making assumptions about the expected returns and risks of different assets and portfolios.
In the second stage, you likely focused on using time series modeling techniques to forecast future values in a series and analyze the results. This may have involved using techniques such as autoregressive integrated moving average (ARIMA) models or other statistical models to fit the data and make predictions about future values.

<u>Stage One</u>

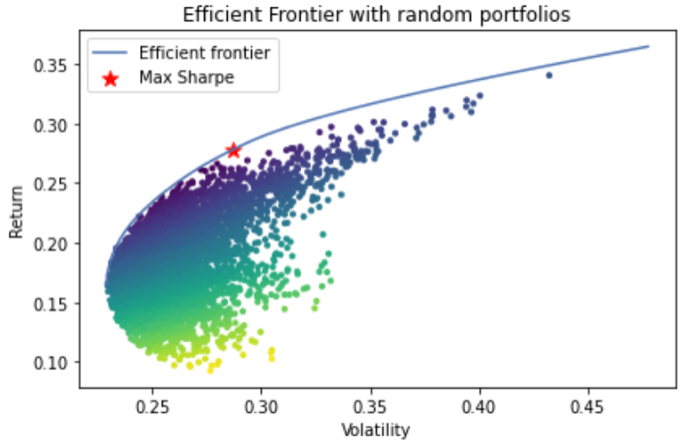
The real time financial data is imported from Yahoo Finance using the pandas reader.



After obtaining the necessary datasets of our required assets, convert them into daily returns using the pct_change() functionality. This will calculate the percentage change in price between each day, and can be useful for identifying mean-reverting behavior in the data

Quadratic optimization technique is used to find the efficient frontier and decided to go with the portfolio having maximum Sharpe ratio, which is a standard theoretical practice in portfolio construction. Quadratic technique used is called "PyPortfolioOpt" which is a subclass of convex optimization library(CVXPY).

To implement the asset allocation and leftover amount in your portfolio, you can use the weights and number of stocks of each asset from the optimization output to calculate the amount of capital to invest in each asset.



For example, if you want to invest $10000 in the optimal portfolio, you can use the weights and number of stocks to calculate the amount of capital to invest in each asset. Based on the output you provided, you would invest $6130.68 in Apple (AAPL) stocks, $3613.90 in Netflix (NFLX) stocks, and $260.50 in Google (GOOG) stocks, for a total of $10000. The remaining $184 would be left in cash or a risk-free asset.

It's important to note that the optimal asset allocation and leftover amount will depend on the expected returns and covariance matrix of the assets, as well as the objective and constraints of the optimization. The portfolio may perform differently in the future based on changes in the market and the underlying assets.

Stage Two

To make a time series forecast using an **ARIMA** model in Python, you can use the ARIMA function from the **statsmodels** library. The ARIMA function fits an autoregressive integrated moving average (ARIMA) model to the time series data and makes a forecast for a specified number of steps ahead. The ARIMA function takes in the time series data as the first argument and the order argument, which specifies the values of p, d and q for the model. The fit method of the ARIMA object fits the model to the data, and the forecast method makes a forecast for the specified number of steps ahead (n).

It's common to use model evaluation techniques, such as the Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC), to assess the quality of an ARIMA model and help select the optimal parameters. These criteria are based on the likelihood function of the model and penalize the model for the number of parameters it uses, with the goal of finding the model that balances model fit and parsimony. While the AIC tries to approximate models towards the reality of the situation, the BIC attempts to find the perfect fit.

$$\mathrm{BIC} = \mathrm{AIC} + ((\log T) - 2)(p + q + k)$$

$$\mathrm{AIC} = -2\log(L) + 2(p + q + k)$$

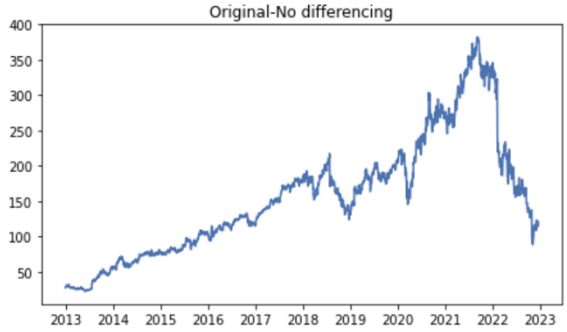
Where k is the intercept of ARIMA, and L is the likelihood.
To compute the AIC and BIC for an ARIMA model, you can use the aic and bic attributes of the ARIMA Results object returned by the fit method of the ARIMA class, or from ARIMAResults. Summary.

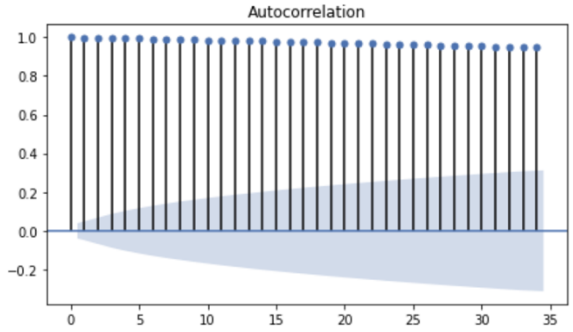
Other common evaluation metrics for time series models include the root mean squared error (RMSE) and residuals, which are the differences between the observed values and the predicted values of the model and the resid attribute of the ARIMAResults or ARIMAResults.summary() object to access the residuals.

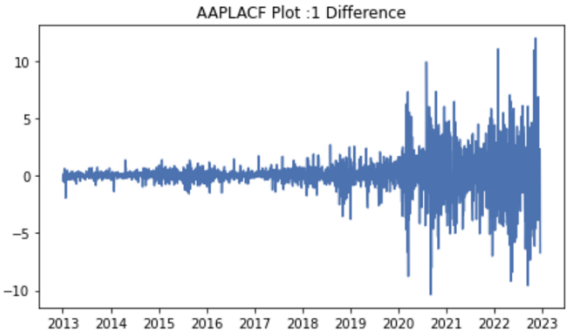
Differencing (d) parameter:

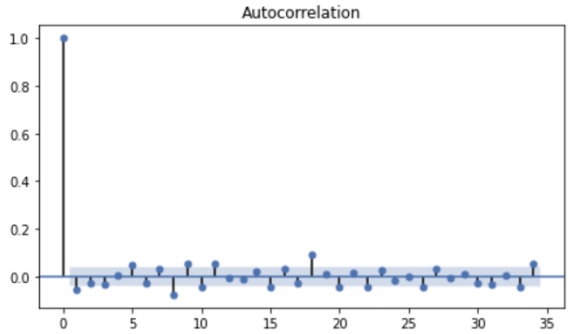
Differencing is a common technique used to make a time series stationary, which is a necessary condition for many time series modeling techniques, including ARIMA. By removing changes in the level of the time series, differencing can eliminate or reduce trend and seasonality, which can make it easier to model the remaining patterns in the data.
To determine the appropriate degree of differencing for a time series, you can use a combination of statistical tests and visual tools, such as the autocorrelation function (ACF) plot. The ACF plot shows the correlation between the time series and lagged versions of itself and can help identify whether the series is stationary or not.

It's important to be careful not to over-difference or under-difference the data, as this can lead to problems with the model. Over-differencing can cause increased standard deviation and make the series non-stationary, while under-differencing can also make the series non-stationary.



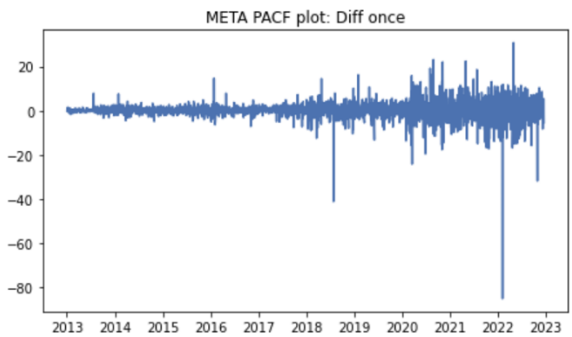


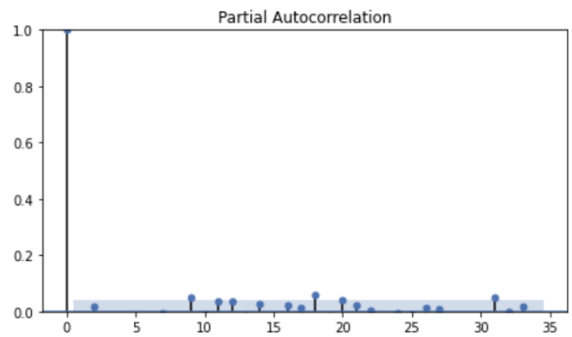




P

To determine the appropriate value of p for an ARIMA model you can use a variety of techniques, including visualizing the partial autocorrelation function (PACF) plot of the time series. The PACF plot shows the correlation between the time series and lagged versions of itself.
To find the order of p by inspecting the PACF plot, you can look for the lag number at which the correlation coefficient goes above the significance region, which is typically indicated by a line horizontal line on the plot. If multiple lags that go above the region, you may need to use a higher value of p to capture the relevant pattern in the data.

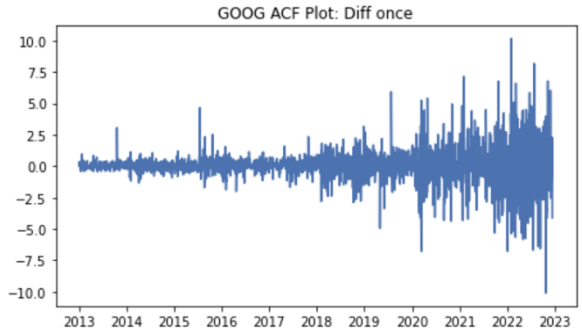


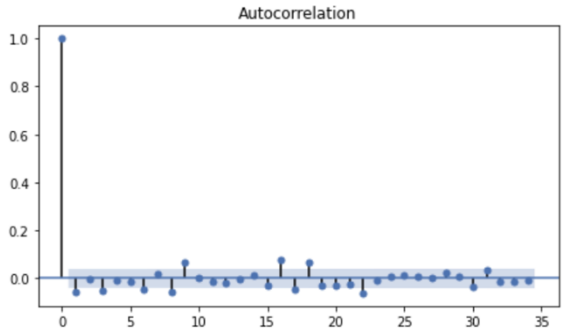


Q

To determine the appropriate value of q for an ARIMA model, you can use a varity of techniques including visual inspection of Auto Correlation function(ACF) plot of the time series. The ACF plot shows the correlation between the time series and the lagged version of itself.

To find the order of q by inspecting the ACF plot, you can look for the lag number at which the correlation coefficient goes above the significance region, which is typically indicated by a horizontal line on the plot. If there are multiple lags that go above the significance region, you may need to use a higher value of q to capture the relevant patterns in the data.





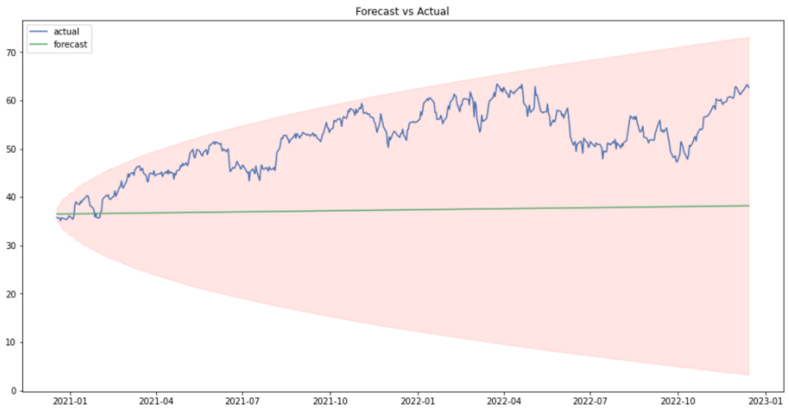
It's important to be careful not to select a value of q that is too high, as this can increase the risk of overfitting and lead to poor out-of-sample performance.

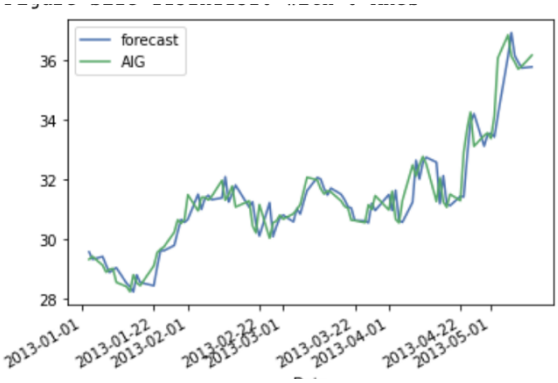
Next step involves running ARIMA model with the parameters that's been found using plots and fitting the model on the test data. For the train test purpose, I used an 80:20 split where 80% of the data was used for training the model and 20 % for testing.
After running the model on the training set, the model fit was analyzed in the test dataset. The model summary is an essential part of ARIMA toolkit in python. It shows the AR(p) and MA(q) terms and their relevance, coefficient, significance, and standard error. It has another important feature that displays the AIC, BIC and log likelihood of the model as well. We can decide p, d and q terms by running the model and logging the AIC and BIC for each set of parameters. After trial and error, the model with least BIC and AIC taken as the final model parameter.

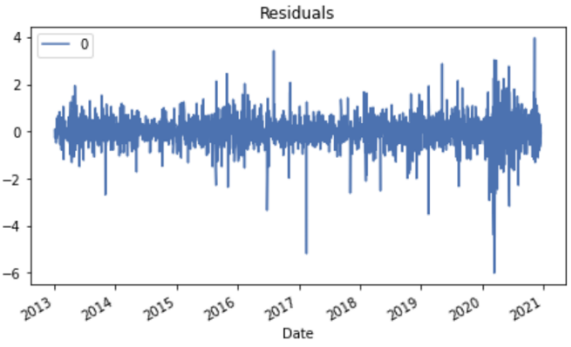
This is a collection of AIC and BIC values for AIG:

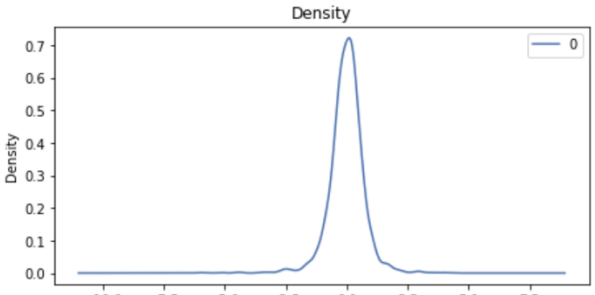
P	d	q	AIC	BIC
7	1	6	4322.681	4406.717
7	1	7	4306.149	4395.803
6	1	6	4307.146	4385.593
6	2	7	4318.376	4402.420

From the above table, it is evident that when (p,d,q) values are (6,1,6) the AIC and BIC are the lowest. So it is a set of optimum parameters.
It is always best to get a visual idea about how the forecast and actual value lie in the test and train dataset. The below figure is an illustration of AIG's ARIMA model in the test dataset. The colored region is the confidence interval. The figure below it represents the density curve of the residuals.

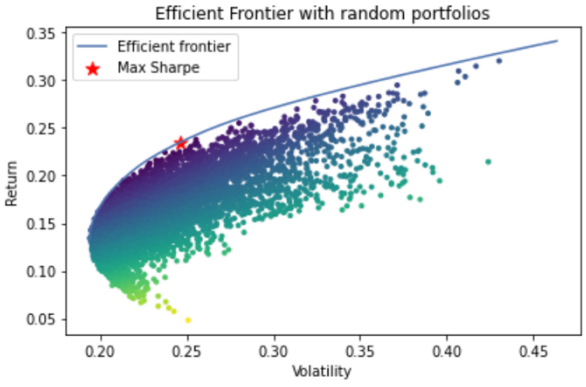








After repeating the same procedure for all the stocks in the list, append the forecast values with given number of steps into the future to the main data frame containing historical values from Yahoo Finance.
By doing so we can evaluate the expected return in future of the initial portfolio and its variance and calculate its variation from the previous expected return and variance.



By using convex optimization on the portfolio and using efficient frontier, we can find the portfolio with max Sharpe ratio at that time point in future. This can provide us insights into asset allocation and to get a balanced return in present and in future.

Summary

This project was a starting step into my idea on making a balanced portfolio or rather a portfolio made using MPT theorem but with a insight of how the market perform in future. This project is a collaboration of finance and time series modelling- where core concepts in finance such as diversification, MPT, risk adjusted returns, risk tolerance, and Sharpe ratio are used along with statistical instruments such as ARIMA model, BIC and AIC criteria, log likelihood, ACF, PACF plots are used to understand the behavior of the model.
From my experience with the project, I found that the shift in asset allocation that could potentially happen in future to get maximum Sharpe ratio portfolio is huge, and the expected returns get suboptimal for the previous optimal portfolio if the forecasts were to come true. This could potentially open up opportunities for buying options such as call or put options, and also exercise non vanilla options such as Asian and barrier option. This opens up other options such as short selling and rebalancing portfolio at certain time steps in order to get more returns than the initial one.
Another finding is that the expected return of initial optimized portfolio reduces drastically in future according to time series forecast.

After this project, I will work on making a hybrid approach that has a balanced return in present and in future, where the portfolio selected from the efficient frontier, with respect to risk tolerance of the investor and Sharpe ratio as well. Using a reinforcement learning on the return and risk of two portfolios(present and time series portfolio) along with employing other financial tool kits such as P/E ratio, book value of the asset, dividends, and other attributes in prediction and as rewards in RL would provide higher accuracy and may help to provide a balanced portfolio.

Conclusion

A portfolio based on historical data using modern portfolio theory can fail to give the optimal return and may even cause higher risk- if not balanced correctly and intermittently. By using time series forecasting and other statistical instruments, it's possible to get an understanding of what could happen and how much impact it would have on the portfolio.

In addition to time series forecasting, Monte Carlo simulation and importance sampling methods can also be useful tools for analyzing the potential risk and return of a portfolio. This can help investors understand the potential range of outcomes for their portfolio and make more informed investment decisions.
Importance sampling is a statistical technique that can be used to estimate the value of a complex financial instrument or portfolio. It involves sampling the distribution of possible outcomes in a way that emphasizes the most important or relevant scenarios.
In summary, using time series forecasting, Monte Carlo simulation, and importance sampling methods can help investors get a better understanding of the potential risk and return of their portfolio and make more informed investment decisions.

References

1. Arima models: https://otexts.com/fpp2/non-seasonal-arima.html
2. ARIMA wiki: https://en.wikipedia.org/wiki/Autoregressive_integrated_moving_average
3. An Open-Source Implementation of the Critical-Line Algorithm for Portfolio Optimization by David H Bailey and Marcos Lopez de Prado : https://www.mdpi.com/1999-4893/6/1/169
4. Mean-Variance vs. Mean-Absolute Deviation: A Performance Comparison of Portfolio Optimization Models by Geoffrey Kasenbacher, Jordan Lee, and Klod Euchukanonchai : https://www.researchgate.net/profile/Geoffrey-Kasenbacher/publication/330358884_Mean-Variance_vs_Mean-Absolute_Deviation_A_Performance_Comparison_of_Portfolio_Optimization_Models/links/5c3c4239a6fdccd6b5ab3e4e/Mean-Variance-vs-Mean-Absolute-Deviation-A-Performance-Comparison-of-Portfolio-Optimization-Models.pdf
5. Yahoo Finance, URL: https://finance.yahoo.com/
6. Price forecasting and risk portfolio optimization by V. Centano, I. R. Georgiev,V. Mihova and V. Pavlov
7. Prediction of the Best Portfolio for Bitcoin and Gold based on the ARIMA Model, by Qi Zhou, Zixuan Chen, Zhuoying Cai, Ziwei Xia
8. Comaprision on forecasting ability of Arima models by Simon Stevenson, ISSN: 1463-578X
9. ARIMA and finance : https://www.investopedia.com/terms/a/autoregressive.asp , https://www.investopedia.com/articles/trading/07/stationary.asp
10. "Price forecasting and risk portfolio optimization", AIP Publishing, 2019 by V. Centeno, I. R. Georgiev, V. Mihova, V. Pavlov
11. Tezuran, Mustafa. "Cagri Merkezi Performans Analizi Ve cagri sayilarinin Mevsimsel Tahminlemesi", Marmara Universitesi (Turkey), 2020