The 2022 International Conference on Financial Technology and Business Analysis
DOI: 10.54254/2754-1169/5/20220108

Econometrics and Machine Learning Approach on Correlation in Stock Return between US Firms and Chinese Suppliers

Huiyan Xiao[1,a,*]

[1]*Imperial College Business School, Imperial College London, London, United Kingdom*
a. huiyan.xiao22@imperial.ac.uk
**corresponding author*

Abstract: This study investigates machine learning on influential factors on the correlation of the stock return movements between US firms and their Chinese suppliers, given that the overall correlation has already been proven in the previous study. The present investigation was conducted by performing econometrical analysis and applying a variety of regression models to US firms' stock returns (independent variable) and their Chinese suppliers' stock returns (dependent variable) with various factors involved. From the results, it was evident that industry differences are a factor that caused variations in the degree of stock return correlations. Furthermore, firm size and stock trading volume yielded positive impacts on the significance of the correlation. Subsequently, predicated on said findings, a prediction model was generated using the Random Forest machine learning approach. Using US customer firms' monthly stock return data, as well as the three aforementioned factors, the monthly stock return value of their Chinese supplier firms can be predicted.

Keywords: US-China stock relation, Fama-MacBeth method, random forest

1. Introduction and Background

1.1. Motivation

As the world's two current leading economic powers, the US and China possess relations in numerous areas. The relationship between the US and Chinese market is highly influential to the macro-economy and other fields, such as sociology, politics, and finance, specifically stock markets. The strong economic relationship will create other relationships formed on the US and Chinese stock markets, in other words, co-movements may arise. If we compare the Dow Jones Index (DJI) with the Shanghai Stock Exchange (SSE), there are traces of co-movements concentrating around the economic crisis in 2008. However, after 2010, the Dow Jones Index rose substantially, while the Shanghai Stock Exchange failed to rise significantly.

There are a multitude of reasons behind the “blur” co-movements. Firstly, not all Chinese firms possess a relationship with US firms. Also, China's stock markets suffer many limitations compared to US stock markets. Mainly because of China's stock markets have a “daily limit”, as well as the “T+1” schema. Although China's policies reduce the risk involved with investment, the activity in China's stock markets is highly decreased due to non-marketization. The aforementioned limitations, along with many others, are affecting the overall sync-movements of the two markets. There are not

DOI: 10.54254/2754-1169/5/20220108

many actions that can be taken in regard to China's policies. However, regarding the first limitation, if it can be concluded which firms have a relationship with US firms, along with the industry, firm size, popularity, and time period that are likely to produce co-movements, then the stock return changes based on another market may be predictable.

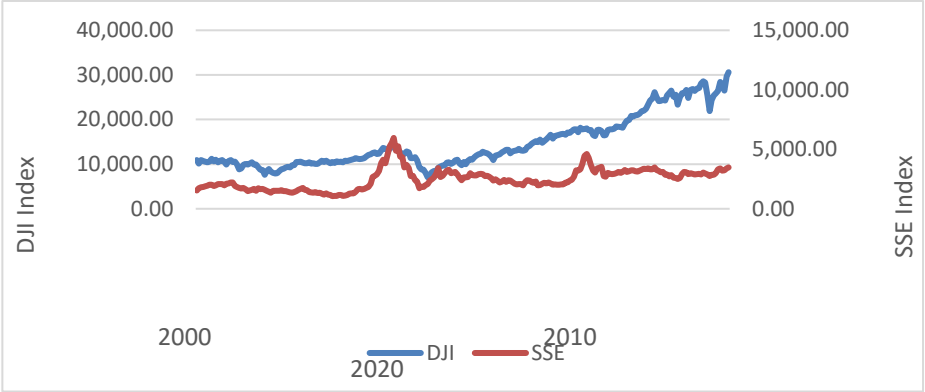

Figure 1: Monthly stock price of Dow Jones Industrial Index compared with Shanghai Stock Exchange Index from 2000 to 2020.

The relationship scheme of US “customer” firms with Chinese “supplier” firms was stated and explained in a previous study – “Co-relations Between US Firms with Their Chinese 'Supplier' Firms and the Co-movement of Their Stock Returns”conducted by the present author [1]. Furthermore, the correlation of stock returns of firms in the US and Chinese markets was evidenced in the previous study. In the present research, all the monthly data of the US “customer” firms' stock return between 2000 to 2019 were grouped based on the stock return, hence, grouped arbitrary data was used as the independent variable, and all the Chinese “supplier” firms' stock return was the dependent variable. Each Chinese supplier had its corresponding group of US “customer” firms. After conducting a Fama-MacBeth regression, a positive coefficient was produced, with an acceptable t-value evidencing the significance of the results [2].

However, since the number of stocks and the total value of the investment portfolio is not infinite, the overall correlation of stock returns is not enough to produce profit. Consequently, the factors that influence the degree of correlation had to be investigated. In order to enhance the practicality of the present research, the aim was to formulate a model, based on the influential factors, that reveals which firms have a stronger degree of correlation.

1.2. Research Objectives

There were two objectives in the present research. The first was to discover the factors that decide the power of correlation in stock returns. Firms were grouped predicated on the characteristics of the factors. Regressions were conducted on the different groups of firms in order to test the degree of correlation. The second objective in the present research was to establish a prediction model which can predict the monthly stock return changes of Chinese firms based on their US “customer” firms and any other influential factors. Hence, the influential factors had to be quantized in order to serve as features and produce an accurate model.

The prediction model should be able to predict the monthly stock return of a Chinese firm, while considering the influential factors. More importantly, the firm has to be a Chinese supplier for one or several US “customer” firms, and the stock return data, as well as the required characteristics, of the US firms should be accessible.

DOI: 10.54254/2754-1169/5/20220108

1.3. Research Methods

To discover the factors that define the degree of stock return correlation, regression was conducted on firms using variety of factors. For instance, if a regression test is applied to both the top 25% and the bottom 25% of the largest firms, the results will likely differ. Also, by comparing the coefficient and the t-value of the two tests, the group that possesses a stronger stock return correlation power can be concluded. A two-regression method was adopted in the present study, with Ordinary Least Squares (OLS) and Fama-MacBeth (FM) being the regressions. Compared to a common OLS regression, FM regression was suitable for financial data since the error of data dependency can be eliminated. OLS regression was conducted on firms in different industries, and the t-value of the regression results served as "scores" for the industries, which reflected the degree of correlation. Subsequently, predicated on the stock return value, the firm size, and stock trading volume, firms were classified into 5 groups, and FM regression was conducted on each group. The difference in the power of correlation was observed in the regression results.

By implementing the related factors (i.e., industry, size, and trading volume) into a well-formed model, the Chinese firms' stock return can be predicted based on their US "customer" firms' stock return. Random Forest was utilised to train the model. The main dataset was divided into two parts, with 90% of it being used for training and 10% being used for testing purposes. The prediction model is articulated so that it can be reused or potentially improved in the future.

2. Literature Review

2.1. Customer-supplier Firm Relation

Customer firms and supplier firms are influenced by the relationship they share. Wang believed that a firm's relationship with its non-financial stakeholders, such as principal customers/suppliers, is an important determinant of its shareholders' income [3]. Said belief was investigated by looking into the dependence on customer-supplier relationships, as well as financial distress surrounding the relationship. Moreover, using the data of 500 customer firms in the US and Germany, Cannon & Homburg argued that, for customer and supplier firms, increased communication frequency, different forms of supplier accommodation, product quality, and the geographic closeness of the supplier's facilities to the customer's buying location lowered the customer firms' costs [4]. Customer firms also tend to increase their purchases from suppliers that provide value by lowering each cost.

For the present study, it was necessary to mention the existence of the customer-supplier relationships, which potentially triggers the co-movements of customer and supplier firms' stock returns as they may share business conditions in some contexts. The real situation is more complicated since most of the firms have more than one supplier, and most supplier firms offer goods and services to more than one customer firm.

2.2. Stock Return Correlation

In Singh and Kaur's research, following the efficient tests of causality, that were inspired by Hill, an indirect impact of the US market volatility on the Chinese market was observed [5]. The portfolio managers should discount said observation in advance, and maintain the portfolio values by taking positions in futures and options markets [6]. In response to the magnitude of the recent trade war, Shi et al. examined the stock market co-movements between the US and China from the 3rd January 2017 to 23rd January 2020 [7]. They argued that co-movements among mainland China, Hong Kong, and US stock markets are positively affected by news releases and, after the 6th July 2018, were significantly enhanced. Moreover, there is empirical evidence that demonstrates positive announcement effects of stock market co-movements between the US and mainland China in specific sectors

DOI: 10.54254/2754-1169/5/20220108

(particularly, industrials and information technology). In respect to international investors, from said evidence, it can be observed that the US-China Trade War has reduced the benefits of portfolio diversification in managing risk. In the present study, the existence of co-movements on the stock market are elaborated, mainly in regard to recently released news. Additionally, an insight into how correlations in the stock market vary across different sectors was provided, which is one of the key factors discussed in the present research. In the research done by Alanyali et al., the relationship between financial news and the stock market was evidenced [8]. Said relationship was evidenced through the positive correlation between a company's daily number of mentions in the Financial Times and the daily transaction volume of the company's stock, both on the day before the news is released and on the same day the news is released. The reason for highlighting the relationship between the news and the stock market was that news is a major channel for investors to gather information and then make decisions on stocks based on the firm's customer firms.

2.3. Machine Learning Approach on Stock Market

In research pertaining to the stock market, there have been an abundance of examples relating to machine learning algorithms. Zhong & Enke introduced a comprehensive big data analytics process for predicting the daily return direction of the SPDR S&P 500 ETF based on 60 features [9]. In order to predict the daily direction of future stock market index returns, DNNs and traditional artificial neural networks were deployed over the entire pre-processed, but untransformed, dataset and two datasets transformed via principal component analysis (PCA). As a result, insight into using machine learning on stock returns was obtained. Compared to Zhong and Enke's study, there are far less features and a simpler model in the present research, hence, multiple advanced models were not necessary.

Concerning Random Forest, Tan et al. presented a Random Forest method for the purpose of investigating the excess return in Chinese stock markets [10]. In their research, a Random Forest model in the context of stock selection was implemented and two types of feature spaces - fundamental/technical feature space and pure momentum feature space - were included. Results evidenced that, although the excess return had weakened in recent years with respect to the multi-factor strategy, the market had lower efficiency and far from equilibrium.

In both of the aforementioned studies, it was argued that machine learning techniques are widely employed in the field of finance and the algorithms are becoming increasingly robust. However, researchers have failed to cover the international stock market correlation, especially for Sino-US relations. Consequently, a Random Forest regression on US and Chinese firms with a customer-supplier relationship was conducted in the present research.

3. Method Development

3.1. Dataset Selection

The meta-datasets employed in the present research were the same as in the previous study, with some extra information related to each stock. Essentially, there were 3 meta-datasets: US customer firms' monthly stock return, Chinese supplier firms' monthly stock return, and a dataset that pairs the customers and suppliers. In the previous study, an overall relational dataset was generated for regression analysis purposes from the 3 meta-datasets, and it was also being employed in the present study. Industrial category, trading volume of the stock during a particular month, and the size lag of the firm in the specific month were also used in the present research. Furthermore, other supplementary data were utilised, for instance, the overall stock market index of the US and China (Dow Jones, Nasdaq, Shanghai Stock Exchange, and Shenzhen Stock Exchange).

DOI: 10.54254/2754-1169/5/20220108

The majority of the utilised data stemmed from "Wharton Research Data Services" (wrds), which is a financial dataset platform established by The Wharton School, University of Pennsylvania [11]. Data pertaining to the stock markets as a whole was collected from Yahoo Finance [12]. The data concerning US customer firms was gathered from the Center for Research in Security Prices (CRSP), and the data regarding supplier firms in China was collected from the China Stock Market & Accounting Research Database (CSMAR) [13][14]. Lastly, the relational data stemmed from Factset Revere [15].

The dataset used in the present study was comprised of data from January 2000 to December 2019. Before 2000, trading between the US and China was scarce. Information transparency and information sharing at that time was also relatively low. Hence, the referentiality of the data before 2000 is inadequate. Furthermore, the market in 2020 was impacted significantly by COVID-19, making the market unpredictable to some degree due to substantial government intervention. Therefore, data within the specified 20 years was employed.

3.2. Methodology Comparison Analysis

3.2.1. OLS Regression vs. Fama-MacBeth Regression

The primary experimental method in the present research was Fama-MacBeth (FM) regression. FM regression is a method that is used to estimate parameters for asset pricing, etc., and was first established by Fama and MacBeth in 1973.

In the context of the present study, FM regression was the most suitable and applicable. The difference between OLS and Fama-MacBeth Regression is that FM was designed to address heteroskedasticity within time dimensions. For pooled OLS regression, data is assumed to occur in the form of independently identical distribution, which is inapplicable in the present context since the stock return between months were dependent. Although the results produced by OLS were superior to FM, employing FM in the present context was the correct decision, especially for data-driven experiments. However, for specific experiments that only require the difference in terms of the degree of stock return correlation instead of an accurate conclusion, OLS can be used to observe the existence and significance of the correlation.

3.2.2. Practicability of Random Forest

Random Forest regression was utilised to generate the ultimate model in the present research. Random Forest was first presented by Breiman, and is now a well-formed machine learning algorithm. Compared to manually implemented models, prediction models generated by machine learning algorithm are more precise and feature-included [16]. Random Forest is a supervised machine learning algorithm that employs ensemble learning for regression purposes. The term "ensemble learning" refers to the combination of machine learning algorithms that makes the prediction more accurate in comparison to using a single model. Essentially, trees are constructed during training time, with no interaction amongst each other. After the predictions are completed by all the decision trees, the algorithm uses the mean of the trees' predictions as the overall prediction.

Random Forest was chosen as it is suitable for processing relational data and conducting regression on said data. Random Forest regression reduces overfitting in decision trees in order to enhance accuracy. Additionally, Random Forest is highly efficient for large data sets, with no need to normalise the data. There is an abundance of studies that deal with financial data using Random Forest.

DOI: 10.54254/2754-1169/5/20220108

4. Correlation Experiments

4.1. Experiments Introduction

There were 2 major parts of the experiments in the present research. The first part focused on the industrial differences, while the second part aimed to explore the relationship between the power of correlation and firm size, as well as between the power of correlation and stock trading volume. In order to test the hypothesis, both parts were conducted using regression test. The first part only applied Ordinary Least Squares regression, while the second part used Fama-MacBeth regression since it needs more accuracy.

4.2. OLS Regression on Different Industry Sectors

4.2.1. Model and Method Outline

Determining a reliable source of industry sector classification was a significant aspect of the present experiments. After consideration, 2012 CSRC Industry Classification was selected [17]. This schema was formulated by China Securities Regulatory Commission (CSRC), with the scope of related laws in China. The guideline was formed by the official stock commission, and had all the necessary information and was most objective.

In addition to the base data (customer and supplier's stock returns) generated by the meta data, two additional pieces of data were needed in the present experiments: the industry name and industry code, which were both obtained from the 2012 CSRC Industry Classification. The class of industry are for supplier firms, and not for customer firms. The reason is that the data schema involved one supplier firm with several customer firms thereof, and thus, instead of considering all the customers' industries, it was preferable to focus on the supplier's division.

There are 90 classes overall for the 2012 CSRC industry classification, with only 44 appearing in the present dataset. In the pattern for the industry code of the classification, an alphabet is used to indicate the category, and a number is used to indicate the class. To illustrate, code “B06” is used for the coal mining and dressing industry, where “B” refers to the mining industry and 06 is the class number of the specific industry. The numerical number does not zero out for each category, for example, category B starts with B06, not B01.

Python script was used to match each supplier firm from the meta dataset with its industry code, and the data for 2012 CSRC industry classification were obtained from WRDS. Subsequently, firms in the same industry were grouped together, and OLS Regression was applied to generate the “industry score” for each sector, which was the t-value of the OLS Regression. In the present research, OLS Regression was used instead of Fama-MacBeth Regression since the absolute value of the score itself does not mean anything. The scores here are only used to act as an indicator for the potential degree of correlation for the specific industry. As such, the OLS Regression results are sufficient for determining the existence and significance of the correlation of stock return. Python was also used to conduct the OLS Regression. Finally, the degree of correlation of each industry could be determined by the value of the industry score, with a higher score indicating greater correlation.

4.2.2. Experiment Result

The industry score for different classes was generated, and the top five and bottom five industries are shown in Table 1. Four decimal places were kept for the score value. Essentially, the highest score was 6.6, down to -2.2 where the correlation was negative. The scores varied across different industries.

DOI: 10.54254/2754-1169/5/20220108

Table 2: Top five and bottom five industries according to regression t-value (score).

Top five industries			Bottom five industries		
Name	Code	Score	Name	Code	Score
Industry of rubber and plastic products	C29	6.6011	Agricultural and sideline food processing industry	C13	-2.2076
General equipment manufacturing	C34	6.1932	Press and publishing industry	R85	-2.1292
Manufacturing of computers, communications and other electronic equipment	C39	6.1139	Manufacturing of stationery, industrial arts, sports and entertainment supplies	C24	-1.7580
Exploitation auxiliary activities	B11	6.0019	Furniture manufacturing	C21	-1.5926
Industry of electric power and heat production and supply	D44	5.3341	Professional technical service industry	M74	-0.5027

There were 8 classes of industries where the score (t-value) was negative, indicating that there was a negative correlation of stock returns. According to the results, however, most of the t-values was not large enough to draw such a conclusion. In such financial data, Brownian Motion may exist, which makes the t-values negative. However, what can be concluded is that the coefficient was not significantly positive, which means there are no positive correlation in stock returns, so these particular classes should have a lower score.

Industries that have relatively high scores are mostly those involved in electric components manufacturing, which is to be expected since such industries belong to the field of technology, where the customer-supplier relationship is obvious. A higher degree of correlation is triggered by the fluent sharing of information and high news sensitivity. At the same time, for low score industries, several of which are not highly related across nations, such as press and publishing, as well as professional technical services (this industry is relatively region-independent). Surprisingly, several manufacturing industries were present on the list, which could potentially be attributed to the number of customer firms for the supplier being generally large, thereby rendering variation in the average customer firm stock return (major independent variable), and causing poor fitting. As for the agriculture industry, the Chinese government has hugely intervened in related industries, which has made the marketisation relatively low.

Overall, the hypothesis that “the degree of the correlation of stock return varies across different industries” is true, and the performances of most industries can be explained. Meanwhile, a quantitative dataset was produced to support the further Random Forest modelling.

4.3. Fama-MacBeth Regression on Different Firm Sizes and Trading Volumes

4.3.1. Model and Method outline

In the present experiment, the firm size and popularity of US customer firms were used as independent variable. This is because in reality, Chinese investors will buy in or out of stocks based on the movements of the US stock market, which is more powerful and has more influencing ability.

The firm size of the customer firms was averaged because a single supplier may have several customer firms. However, the sum of all the customers was taken for the trading volume, since more customer firms in the US means more impact by the US stock market, and trading volume reflects the stock market directly. Therefore, in this case, if a supplier has two customer firms in the data, summing all their trading volume partly indicates that the influence from the US market is higher

DOI: 10.54254/2754-1169/5/20220108

than those which only have one customer firm (assuming the trading volume for each firm is the same).

For each month, supplier firms with their customer firms were assigned to 5 groups based on their firm size or trading volume for that particular month. “Five” were selected because this is the normal scheme of grouping testing, and is known as quintile. A control group – a group of 3 were also tested to determine the differences. Ten groups were also tested, but the result was not as good as the five or three groups, and the ten groups were relatively redundant in showing the difference in degree of correlation since some groups' t-values were similar.

All the data belonging to the same groups were subsequently combined. After, through the result of FM Regression of each group, the degree of correlation could be detected. Python scripts were used to for sorting and grouping, then Stata was used to conduct Fama-MacBeth Regression. The results of this experiment require more accuracy than the previous experiment, since apart from showing the difference in the degree of correlation, a positive relationship between stock return correlation and the popularity of customer firms had to be evaluated. As such, Fama-MacBeth Regression was adopted here instead of simple OLS Regression.

When performing FM Regression, the impact of outliers should be eliminated. In certain months, the value of stock return for a supplier or for a group of customers are extreme (for example, -0.6 for a single month). Although in the assumptions of the present study, a huge shift in stock return is likely to trigger co-movements, the fitting of the regression model would be affected, and thus, should be minimised. One option is to just ignore the lines of data with outliers; however, as previously mentioned, extreme values are likely to have correlation, so data flattening is a preferred method for dealing with the outliers. Stata has a library that directly supports data flattening, which is called “Winsor”. For the data, the highest 5% and lowest 5% customer and supplier returns were distinguished, and the values were replaced with the value at exactly the threshold of the highest 95% and the lowest 95%. Such method is valid since the absolute value of the coefficient of regression is not that important in the present model, and data flattening does not make up data. The sign of the value as well as the trends are also not changed, and only the impact from outliers is eliminated. This scheme is the default operation of “Winsor”, and this pattern has been applied in a large number of studies.

4.3.2. Experiment Result

For both sub-experiments on firm size as well as trading volume, five groups were formed to determine the t-value produced by the FM Regression. Additionally, as a control group, the result of the 3-group scheme is shown as well.

4.3.3. Result of the Experiment on Firm Siz

The average firm sizes of the customer firms for each Chinese supplier were calculated. Table 2 shows the coefficient (keeping 4 decimal places) and the t-values of 5 groups and 3 groups, respectively. With Group 1 to 5 from small average firm size to large.

Table 2: Coefficients and t-values of different firm size groups.

Quintile Group number	Coefficient	t-value
1	-0.0293	-1.72
2	-0.0774	-2.07
3	0.0548	1.53
4	0.0665	1.54
5	0.1241	2.34

DOI: 10.54254/2754-1169/5/20220108

Table 2: (continued).

Tripartite Group number	Coefficient	t-value
1	-0.0337	-2.21
2	0.0774	3.04
3	0.0414	1.16

From the values from Table 2, a conclusion could be drawn that the degree of correlation varied across different size of firms. However, the pattern of change in the correlation for the 3-group schema was different from that of the 5-group schema. In this case, 3 groups may not be effective enough to distinguish the different firm size, since the pattern of the experiment with 10 groups matched the pattern of the quintile test. However, the tripartite test still shows that small firms had poor performance in terms of stock return correlation. Figure 2 shows the overall trends of coefficient and t-values for different sizes of firms.

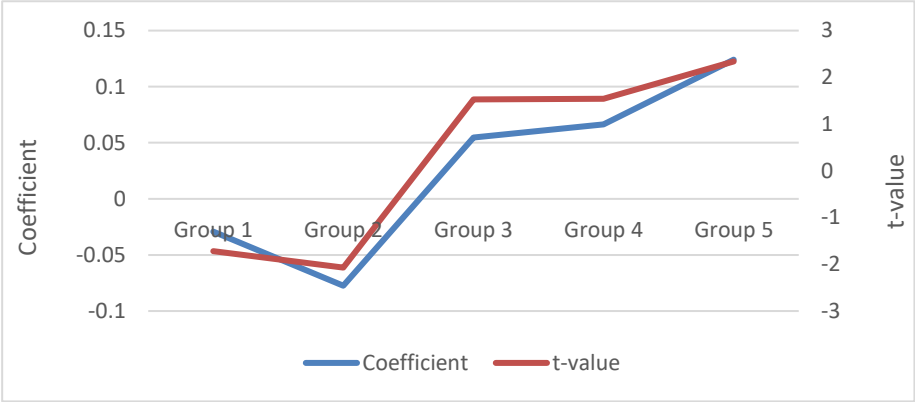

Figure 2: Trends of coefficient and t-values for different sizes of firm.

The trend shows that overall, the degree of stock return correlation increased as the size of the firm increased. There was a jump from Group 2 to Group 3, which may indicate that the firms were large enough to receive investor and press' attention, and may also indicate the existence of financial information exchange. The growth in degree of correlation subsequently became much smoother, which could be attributed to larger customer firms having a complex supplier relationship, and the simple direct correlation being harder to distinguish because more suppliers are involved, which results in the unexpectedly low growth in degree of correlation.

Based on the results, the assumption of larger firms having a higher degree of correlation of stock return is true. Also, with this information in hands, the firm size value was included in the Random Forest modelling as one of the features to generate the model.

4.3.4. Result of the Experiment on Trading Volume

The sum of the trading volume of the customer firms' stocks for each Chinese supplier were calculated. Table 3 shows the coefficient (keeping 4 decimal places) and the t-values of 5 groups and 3 groups, respectively. With Group 1 to 5 from small average firm size to large.

DOI: 10.54254/2754-1169/5/20220108

Table 3: Coefficients and t-values of different trading volume groups.

Quintile Group number	Coefficient	t-value
1	-0.0108	-0.55
2	0.0071	-0.27
3	0.0627	1.48
4	0.0979	2.27
5	0.0936	2.14
Tripartite Group number	Coefficient	t-value
1	-0.0257	-1.70
2	0.0577	2.03
3	0.0535	1.86

The coefficients and t-values of different groups varied, and the basic trend was similar to the firm size experiment. For the quintile test, Groups 4 and 5 had higher degrees of correlation based on the coefficients and t-values, and Groups 1 and 2 performed badly. The t-value of the third group of the tripartite grouping was lower than the second group, but the degree of correlation of the third group was not necessarily lower, since the difference was considerable low. However, the t-value of Group 1 was significantly smaller and became negative, which shows there was no correlation (even negative correlation) for this particular group. Figure 3 shows the overall trend of coefficient and t-value for firm clusters with different trading volumes.

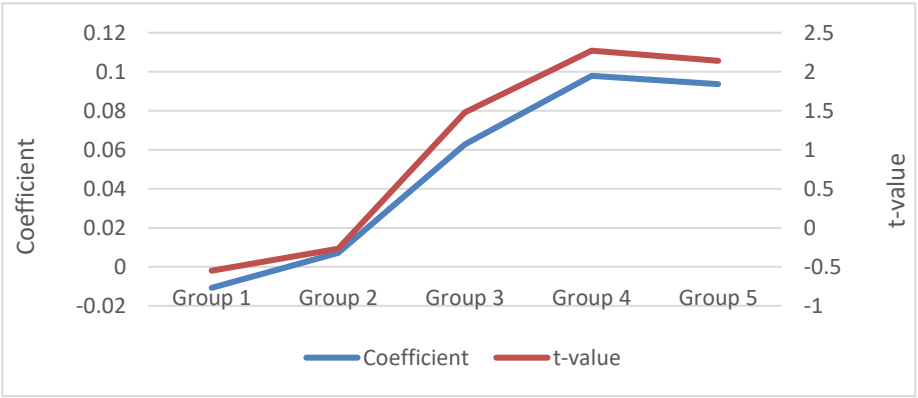

Figure 3: Trends of coefficient and t-values for different trading volumes of firms.

The trend shows that as the trading volume became larger, the degree of stock return increased. A rapid increase in the coefficient and the t-value existed between Groups 2, 3, and 4, indicating that the correlation became more stable and robust as the trading volume of customer firms increased. However, the growth stopped from Group 4 to Group 5. The reasons may be similar to those in the firm size experiment, that is, the relationship between customer and supplier became complex. Further, because the trading volume is the summation of all the customer firms for a specific supplier firm, Group 5 usually involved multiple customers, which render the average customer return unstable.

The hypothesis that supplier firms with customer firms with a larger trading volume have greater correlation in stock return is true. The trading volume value was included in the Random Forest modelling as well to generate the model.

DOI: 10.54254/2754-1169/5/20220108

5. Random Forest Predictive Modelling

5.1. Machine Learning Modelling

5.1.1. Machine Learning Approach

Machine Learning methods are not a necessary procedure for finding the correlation. General statistical methods such as correlation test and regression test sufficient for are determine the existence of correlation. However, a well-formed model can provide the ability of automatic judgements.

Quantitative analysis is a trending topic that has recently emerged, along with financial computing and financial technology. Despite such trend, human intervention cannot be completely replaced even for huge quantitative analysis companies. Even the finest machine learning algorithms are not accurate enough to be used for making money in a stable manner, and the gaps still exist. For the present research, the major aim was to prove the correlation of stock return, and the accuracy required was not that high. After browsing examples and other studies, Random Forest was selected, which is a powerful algorithm for data classification and regression. In the present study, “Random Forest Regression” was the function that was used.

5.1.2. Model Building and Outline

Random Forest is a well-known machine learning algorithm for classification, regression and prediction. More information about the Random Forest is provided in Section 3.3.3. Through the previous experiments discussed in Chapter 4, three factors were found to have an impact on the degree of correlation. Thus, in addition to the US customer firms' stock return, three more features were added, namely industry score, firm size and trading volume. The dependent variable was Chinese supplier firms' stock return. All the data were monthly data.

Python was used to conduct the Random Forest modelling, along with the “sklearn” library, which is a powerful machine learning packet with numerous ML algorithms. The model that was used in the present research was “RandomForestRegressor”.

Attempts were made to trim some of the dataset because of the impact of outliers or Brownian Motion, via removing the data with the absolute value supplier return greater than 0.3, or below 0.05 because of the random motion. However, the result was not any better than doing nothing. The Mean Absolute Error was nearly the same, and the Adjusted R-Square even decreased slightly. As such, I decided not to trim the dataset, since the Random Forest itself is an effective algorithm that could deal with the outliers and random errors in hand.

In terms of the dataset, 90% was used as training data, while the other 10% was used for testing purposes. At the end of training, the generated model was saved for future use, which was one of the final outputs of the present research. The model can be directly imported and reused for other data, although the schema of data needs to be consistent with the current dataset.

Adjusted R-Square was taken as the mean indicator to determine the goodness of fit of the model. Compared with other indicators like Mean Square Error (MSE), R-Square has the advantage of “readability” since no matter the value of data, the value is always between 0 to 1. Furthermore, R-Square is better for testing the fitting performance. As for Adjusted R-Square, in addition to R-Square, a “penalty” is applied on the additional variables if they do not improve the performance of the model.

After the model was generated, the testing group was used to test the performance of the model. Mean Absolute Error (MAE) was used to see the performance. Compared to MSE, MAE could eliminate the effect of dimension differences, which made the value of it easier to analyse.

DOI: 10.54254/2754-1169/5/20220108

5.2. Experiment Results

Customer firms and supplier firms' monthly stock returns, as well as other three features were included in the Random Forest Regressor, and the model was successfully generated. Table 4 shows the overall results of the model and tests.

Table 4: Overall results of the Random Forest Model.

Adjusted R-Square	Mean Absolute Error	Feature Importance
0.8241	0.1018	[0.3351, 0.2564, 0.2405, 0.1680]

The Adjusted R-Square of the model was 0.8241 indicating the prediction model was relatively good, and also that the fitting was relatively good, which refers to the existence of correlation in stock return. The MAE is 0.1018, which was tested under the 10% test dataset. After attempting possible optimisations towards the training dataset (for example, data flattening, removing data with the absolute value of return > 0.3, removing data with the absolute value < 0.05, and other means.), current schema (no trims on data) was the best option by the MAE value. Although in certain solutions, the MAE was smaller than the current value, the absolute value of the dependent variable was also increased for the purpose of optimisation, and in this case, the decrease in MAE could be ignored. Figure 4 shows the comparison of part of the testing set. From the mapping, an observation can be made that the overall shape was fine, with the majority of the perditions being valid. However, for some extreme return values, the prediction was not good.

The feature importance for customer stock return, industry score, firm size and trading volume was 0.3351, 0.2564, 0.2405 and 0.1680, respectively, which were to be expected. Figure 5 shows the weights of each feature. To conclude, aside from the customer firms' stock return, which was the major indicator, industry score and firm size had a significant impact on suppliers' return as well, with trading volume having less weight.

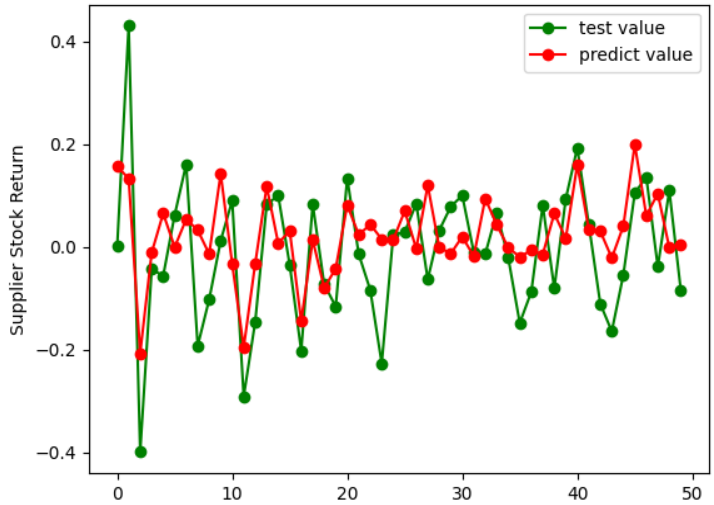

Figure 4: Predicted values vs. real values of part of the testing se.

DOI: 10.54254/2754-1169/5/20220108

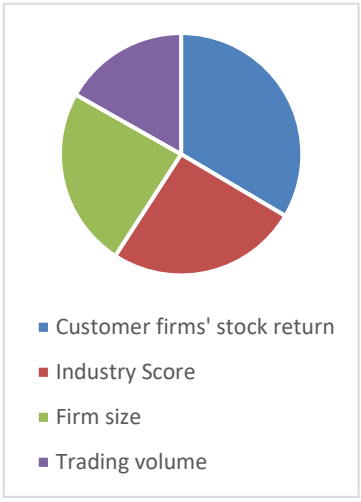

Figure 5: Weights of each feature.

The feature importance for customer stock return, industry score, firm size and trading volume was 0.3351, 0.2564, 0.2405 and 0.1680, respectively, which were to be expected. Figure 5 shows the weights of each feature. To conclude, aside from the customer firms' stock return, which was the major indicator, industry score and firm size had a significant impact on suppliers' return as well, with trading volume having less weight.

6. Conclusion

6.1. Summary

The present research is a follow-up study of a previous study “Co-relations Between US Firms with Their Chinese 'Supplier' Firms and the Co-movement of Their Stock Returns”, with a focus on the correlation or so-called co-movements of monthly stock returns between the US and Chinese markets. Compared with the previous study, more factors affecting the correlation were explored in the present study, and a prediction model was established.

The research consisted of two major parts, the correlation experiments and the machine learning approach. There were two sub-experiments of the correlation experiments. The first sub-experiment was conducted to investigate the degree of correlation in stock return for different industries based on the supplier firms' sector, while the second sub-experiment was conducted to explore whether there was a positive relationship between the degree of correlation and size of customer firms, as well as between the degree of correlation and trading volume of customer firms. Regression is the major method used to judge the degree of correlation. Ordinary Least Squares regression is used to see the correlation in difference industries that the supplier firms are in, and Fama-MacBeth regression is used to test the relationship between stock return correlation and size as well as the trading volume of the customer firms. Based on the result of the two sub-experiments, the degree of stock return varies across different industries, as well as different sizes of firms and stock trading volumes of firms. An “industry score” was generated during the first experiment, which was the t-value of the regression results for each industry, indicating the power of stock return correlation. For the second experiment, more specifically, firms with larger size and trading volume will have more correlation power in stock return.

In order to produce a prediction model on the Chinese firms' monthly stock returns, Random Forest was used in the present research. In addition to customer firms' stock return, the other three features of industry score, firm size and trading volume were included in the Random Forest regressor. Python

DOI: 10.54254/2754-1169/5/20220108

with the library “scikit-learn” was used to conduct the Random Forest, which directly generated a saveable model. The Adjusted R-Square of the model was 0.8241, indicating the fitting of the model is good. The MAE of the testing dataset was 0.1018, which was not perfect, but enough when showing the trends of the stocks. The four features of the model, customer firms' monthly stock return, industry score of the supplier firm, average size and sum of the stock trading volume of the customer firms have weights of 0.3351, 0.2564, 0.2405 and 0.1680, respectively.

6.2. Future Work

The outcome of the present research is satisfactory, with the hypothesis being proven and a model being generated. However, there are still a number of ways to improve the existing research. The data used in the present research were all monthly data, so the results were also monthly. However, daily data could be used to conduct all the experiments again. By using daily data instead of monthly, the conclusion will be more useable in real world trading. The presented method can be improved in a number of ways. More features can be added to the Random Forest, and the algorithm itself could also be changed to other advanced ML approaches, which are more powerful and robust than Random Forest. The content of news in the US and China for example, could be processed by the NLP algorithm, which will certainly have an impact on the degree of correlation. There are many other stock markets world-wide with significant relations as well, for instance US and Japan or Europe and China. The same approaches in this study and the previous study could be used to test the correlation power of other markets, and then generate a ML model based on the characteristics of the markets.

References

[1] *Xiao, H., 2021. Co-relations Between US Firms with Their Chinese 'Supplier' Firms and the Co-movement of Their Stock Returns. Advances in Economics, Business and Management Research, Volume Proceedings of the 2021 International Conference on Financial Management and Economic Transition (FMET 2021), pp. 531-536.*

[2] *Fama, E. F. & MacBeth, J. D., 1973. Risk, Return, and Equilibrium: Empirical Tests. Journal of Political Economy, 81(3), pp. 607-636.*

[3] *Wang, J., 2012. Do firms' relationships with principal customers/suppliers affect shareholders' income?. Journal of Corporate Finance, 18(4), pp. 860-878.*

[4] *Cannon, J. & Homburg, C., 2001. Buyer – Supplier Relationships and Customer Firm Costs. Journal of Marketing, 65(1), pp. 29-43.*

[5] *Hill J. B., 2007. Efficient tests of long-run causation in trivariate VAR processes with a rolling window study of the money–income relationship. Journal of Applied Econometrics, 22(4), pp. 747-765.*

[6] *Singh, A. & Kaur, P., 2015. Stock Market Linkages: Evidence From the US, China and India During the Subprime Crisis. Timisoara Journal of Economics and Business, 8(1), pp. 137-162.*

[7] *Shi, Y., Wang, L. & Ke, J., 2021. Does the US-China trade war affect co-movements between US and Chinese stock markets?. Research in International Business and Finance, Volume 58.*

[8] *Alanyali, M., Moat, H. S. & Preis, T., 2013. Quantifying the Relationship Between Financial News and the Stock Market. Scientific Reports, Volume 3.*

[9] *Zhong, X. & Enke, D., 2019. Predicting the daily return direction of the stock market using hybrid machine learning algorithms. Financial Innovation, 5(24).*

[10] *Tan, Z., Yan, Z. & Zhu, G., 2019. Stock selection with random forest: An exploitation of excess return in the Chinese stock market. Heliyon, 5(8).*

[11] *The Wharton School, University of Pennsylvania, 2022. Wharton Research Data Services. [Online] Available at: https://wrds-www.wharton.upenn.edu/ [Accessed November 2021].*

[12] *Yahoo, 2022. Yahoo Finance. [Online] Available at: https://uk.finance.yahoo.com/ [Accessed November 2021].*

[13] *The Wharton School, University of Pennsylvania, 2022. Center for Research in Security Prices, LLC (CRSP). [Online] Available at: https://wrds-www.wharton.upenn.edu/pages/about/data-vendors/center-for-research-in-security-prices-crsp/ [Accessed 20 February 2022].*

[14] *The Wharton School, University of Pennsylvania, 2022. China Stock Market & Accounting Research (CSMAR). [Online] Available at: https://wrds-www.wharton.upenn.edu/pages/about/data-vendors/china-stock-market-accounting-research-csmar/ [Accessed 20 February 2022]*

DOI: 10.54254/2754-1169/5/20220108

[15] The Wharton School, University of Pennsylvania, 2022. FactSet. [Online] Available at: https://wrds-www.wharton.upenn.edu/pages/about/data-vendors/factset/ [Accessed 21 February 2022].

[16] Breiman, L., 2001. Random Forests. Machine Learning, 45(1), pp. 5-32

[17] China Securities Regulatory Commission, 2012. Guidelines for the Industry Classification of Listed Companies. [Online] Available at: https://www.bourse.lu/documents/pdf-LGX-SSL-CSRC.pdf [Accessed 28 February 2022].