Advances in Economics, Business and Management Research, volume 215
Proceedings of the 2022 7th International Conference on Social Sciences and Economic Development (ICSSED 2022)

Research on MSCI Barra CNE5 Model and Stock Selection

Bingjie Cai

[1]*The Chinese University of Hong Kong, Shenzhen, School of Data Science, Shenzhen, China, 518172*
[*]*Corresponding author. Email: 119010004@link.cuhk.edu.cn*

ABSTRACT

With the rapid development of the country's economy, people's disposable income is accumulating and the demand for financial management is growing. At the same time, the number of listed companies in China has exceeded 4,400 and the total market capitalization of the stock market. With a total market capitalization of over 91 trillion dollars, China has become the second largest stock market in the world [1]. Therefore, "how to select stocks scientifically and effectively" and "how to consistently beat the market in the stock market" are questions that stock market investors cannot avoid, and are also questions that people need to think about to improve their financial returns and control their investment risks. It is also a question that people need to think about to improve their financial returns and control investment risks. This paper focuses on stock selection strategies based on the style factors of the MSCI Barra China Equity Model (CNE5). The research target is the China A-share market. The logic of the study is to analyze the past market data by using the MSCI Barra China Equity Model (CNE5) style factors to identify the five factors that have a high degree of explanation and long-term validity in attributing performance to the Chinese equity market, and to verify the effectiveness of the stock selection strategy. The effectiveness of the stock selection strategy is verified by comparing the stock pool returns with the market indexes. The final results show that the strategy achieved 60.39%, with an annualized return of 10.17% and an excess return of 43.93% compared to the performance benchmark.

Keywords: *MSCI Barra CNE5 model, multi-factor model, stock selection, model back-testing*

1. INTRODUCTION

1.1. Literature Review

The factor model has been developed over a long period of time. The American economist Markowitz proposed the Portfolio Theory, which states that the risk of a single stock consists of systematic and unsystematic risk, and that investors can effectively reduce investment risk without changing expected returns [2].

In 1963, Nobel laureate William Shape proposed the CAPM model, which states the relationship between the expected return on assets and risky assets in the securities market, laying the foundation of modern financial market price theory [3]. Since the CAPM model is a single-factor model, it has been challenged by other scholars because of its relative simplicity and inability to explain value effects.

Scholars Eugene F. Fama and Kenneth R. French proposed the Fama-French three-factor model in 1993, pointing out that the excess return of assets in the securities market can be effectively explained by the market capitalization factor SML, book-to-market ratio factor HML and market excess return (Rm-Rf) of listed companies, which is a supplement to the CAPM model [4]. They also introduced the earnings factor RMW and the investment level factor CMA on top of the original three-factor model in 2013 and 2015, respectively, expanding the original three-factor to a five-factor model [5].

1.2. Overall Thoughts

This paper is a quantitative empirical study based on secondary data. The core part is to construct a multi-factor risk model based on the style factor of MSCI Barra China Equity Model (CNE5) model and a stock selection model with ROE factor to verify the effectiveness of stock selection.

This paper first collects the sensitivity and return data of 10 style factors of MSCI Barra China Equity Model (CNE5) model for each trading day from 2015 to 2019, and then analyzes the validity and pure factor return of 10 style factors to find out the five factors with high explanatory degree and long-term validity, and builds a model for back-testing.

With the increasing demand of people's financial management and the development of computer technology, quantitative investment is getting more and more attention and importance from investors. This paper is based on the construction of a multi-factor risk model to test the validity and explanatory degree of the relevant factors in the Chinese A-share market, which can provide a reference for the application of the relevant model in A-share.

2. INTRODUCTION OF BARRA MODEL AND MSCI BARRA CHINA EQUITY MODEL (CNE5)

2.1. Barra Model

In 1974, Barra Rosenberg, an American scholar, first proposed the use of multi-factor risk model to analyze the risk and return of investment portfolios based on Markowitz's portfolio theory and capital asset pricing model theory, etc. [6]. Barra Rosenberg then established Barra and released the first risk-return attribution model for the U.S. market in 1975. The Barra USE1 model was released in 1975, and updated versions of USE2, USE3 and USE4 were released in 1985, 1997 and 2011, respectively, to optimize risk and return attribution in the U.S. market [6].

The Barra model decomposes risk-return sources into common factors and idiosyncratic return factors. The common factors can be further decomposed into risk-index factors and sector factors. The risk-index factors represent some commonalities of assets, such as common market environment, style (growth/value), size (large cap/small cap), etc. The trait return factor is determined by each stock on its own, and can be obtained by comparing the return of that stock relative to the market return and normalizing it. The overall decomposition of the factors is shown in Figure 1.

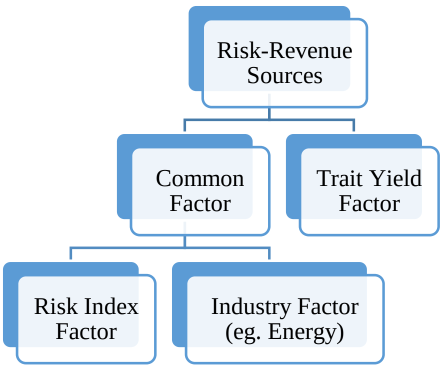

Figure 1. Risk-Revenue Sources

The Barra model has both advantages and disadvantages. In terms of advantages, since the number of stocks N in the asset market is much larger than the number of common factors K, quantitative analysis of the stock market with the help of the Barra model can have the effect of reducing the dimensionality and improving the accuracy of performance attribution and forecasting while greatly reducing the workload. On the downside, the Barra model is based on common factors and involves a lot of data cleaning and processing, so the work required to build a Barra model is very large and the estimation and calibration of factor exposure is difficult to perform efficiently.

2.2. MSCI Barra China Equity Model (CNE5)

According to the Barra China Equity Model (CNE5) model, the return of any stock can be attributed to several different risk factors, including country factor (market factor), style factor and sector factor. The formula [7] is as follows,

$$\mathrm{r_i} = f_c + \sum_s X_{ns} f_s + \sum_i X_{ni} f_i + u_n \tag{1}$$

where $\mathrm{r_i}$ is the return of stock i, f_c is the return of country (market) factor, f_s is the return on the style factor s, f_i is the return of industry factor i, X_{ns} is the exposure of each style factor, X_{ni} is the exposure of each industry factor, and u_n is the idiosyncratic return of each stock.

The CNE5 factor structure consists of 1 country factor, 10 style factors and 32 industry factors. A factor is a factor that describes a particular aspect of a stock's characteristics, such as the style factor Beta Beta factor, which reflects the volatility of a stock relative to the market as a whole. According to the Barra China Equity Model (CNE5) Descriptor Details, the details of the factors are as follows.

- **Country factor**: The essence of the country factor is a market portfolio weighted by market capitalization, i.e., for the country factor, all individual stocks are exposed to 1. The combination of country factors is approximately equivalent to the market portfolio, with individual stocks weighted very close to their market capitalization weights.
- **Style factors**: The MSCI Barra CNE5 model style factors include Beta, Momentum, Size, Earnings Yield, Residual Volatility, Growth, Book-to-Price, Leverage, Liquidity, and Non-linear Size [7]. They will be briefly introduced in the next part.
- **Industry factors**: The industry factor describes whether a stock belongs to this industry. In general, the 32 industry factors are mutually exclusive, i.e., assuming that Stock A belongs to the "Energy" industry, its factor exposure to the industry factor Energy is 1, and its factor exposure to other industries is 0.

The MSCI Barra CNE5 model is a new generation of special risk models built using a time series approach and volatility state adjustment approach. Overall, the Barra risk model explains an average of 28% of all market stocks [8].

3. FACTOR SELECTION AND MODEL BUILDING

The research ideas of the stock selection strategy based on the MSCI Barra CNE5 model style factors conducted in this paper are as follows.

(1) Preliminary establishment of a base factor library based on the MSCI Barra CNE5 model style factors.

(2) Data acquisition and pre-processing.

(3) Single factor validity test and pure factor return analysis of the factors in the base factor library.

(4) Model building.

(5) Model back-testing.

(6) Model optimization.

(7) Optimized model back-testing.

3.1. Preliminary Establishment of a Base Factor Library

First, this research includes all the style factors of Barra model: Beta, Momentum, Size, Earnings Yield, Residual Volatility, Growth, Book-to-Price, Leverage, Liquidity, and Non-linear Size. They are defined as follows:

(1) Beta: computed as the slope coefficient in a time-series regression of excess stock return against the cap-weighted excess return of the estimation universe [7].

(2) Momentum: equals to the relative strength [7].

(3) Size: given by the logarithm of the total market capitalization of the firm [7].

(4) Earnings Yield: equals to the sum of 0.68 times of predicted earnings-to-price ratio, 0.21 times of cash earnings-to-price ratio and 0.11 times of trailing earnings-to-price ratio [7].

(5) Residual Volatility: is the sum of 0.74 times of daily standard deviation, 0.16 times of cumulative range and 0.1 times of historical sigma [7].

(6) Growth: equals to the sum of 0.18 times of long-term predicted earnings growth, 0.11 times of short-term predicted earnings growth, 0.24 times of earnings growth (trailing 5 years) and 0.47 times of sales growth (trailing 5 years) [7].

(7) Book-to Price: is the last reported book value of common equity divided by current market capitalization [7].

(8) Leverage: is the sum of 0.38 times of market leverage, 0.35 times of debt-to-assets and 0.27 times of book leverage [7].

(9) Liquidity: equals to the sum of 0.35 times of share turnover (1 month), 0.35 times of average share turnover (trailing 3 months) and 0.3 times of average share turnover (trailing 12 months) [7].

(10) Non-linear Size: when the standardized Size exposure (i.e., log of market cap) is cubed, the resulting factor is then orthogonalized with respect to the Size factor on a regression-weighted basis. Finally, the factor is winsorized and standardized [7].

3.2. Data Acquisition and Pre-processing

This paper is based on Chinese A-shares, and the sample contains daily data of all listed companies listed in the market for 1219 trading days over 5 years from January 5, 2015 to December 31, 2019. This sampling method helps to extend the general applicability of the stock selection model in the Chinese market. The data are sourced from the MSCI Barra database and Wind Financial Terminal, where the exposure of each style factor of the MSCI Barra CNE5 model is obtained from "barra_dailystockexposure" and the factor return is obtained from "barra_factorreturn". The ROE factor data of the optimization model is exported from the Wind Information Financial Terminal.

Table 1. Raw sample data reflecting factor returns used for model building

FACTORNAME	RETURN	REPORTFREQUENCY	ENDDATE	MODELCODE
Beta	0.0317	Daily	20150105	CQ
Residual Volatility	-0.564	Daily	20150105	CQ
Size	0.8206	Daily	20150105	CQ
Growth	-0.064	Daily	20150105	CQ
Non-linear Size	-0.4336	Daily	20150105	CQ
Momentum	-0.4762	Daily	20150105	CQ
Earnings Yield	0.4517	Daily	20150105	CQ
Leverage	0.0776	Daily	20150105	CQ

Liquidity	0.1159	Daily	20150105	CQ
Book-to-Price	0.3171	Daily	20150105	CQ

Note: only the first 10 rows are shown)

For the treatment of extreme values, the median absolute value depolarization method is used, with the formula: MAD = median (|fi-Median|), from which the range of factor values is determined as [Median-nMAD, Median+nMAD] [9].

This paper used the Z-SCORE standardization method to standardize the data. The formula is:

$$z_i = (x_i - \bar{x}) / \sigma \quad (2)$$

where $\bar{x}$ is the mean of the original data, σ is the standard deviation of the original data. After the above processing, the original factor is transformed into z_i, a series of data with a mean of 0 and standard deviation of 1. The value fluctuates up and down around 0, and greater than 0 means above average.

Since the single factor ranking is likely to select a pool of stocks with more stocks concentrated in the same industry, in order to remove the significant risk exposure from the factor data, industry neutralization is required. The neutralization process is as follows: the factor value is defined as y, the industry dummy variable is defined as x, a linear regression is conducted, and the residuals of the regression model are used as the industry neutralization model. The residuals of the regression model are used as factor values of the neutralized industry. The residuals of the regression model are used as the neutralized factor values. After the industry neutralization, the industry mean is adjusted to a value of 0 [10].

3.3. Single Factor Validity Test and Pure Factor Return Analysis

Since the variables of the multi-factor model in this paper are mainly time series. The time series may have unit roots and lead to series non-stationarity. The unstable time series may cause pseudo-regression after entering the regression model. Therefore, all the variables need to pass the unit root test (Augmented Dicky Fuller Test) to ensure the reliability of the regression results. This paper mainly used ADF test and PP test. When ADF test is large enough and p-value is 0 means that the factor is stable. The non-stationary variables in the model will be replaced by their smooth first-order difference series. The formula [10] is as follows:

$$R_{i,t} = \alpha + \sum_{j=1}^{10} \beta_j (F_{j,i,t} - F_{j,i,t-1}) + \epsilon_t \qquad 1 \le j \le 10, i = 1,2,3,\dots, t = 2,3,4,\dots \quad (3)$$

Table 2 shows the analysis of the IC values of the 10 style factors. IC value refers to the information coefficient, which represents the cross-sectional correlation coefficient between the factor value of the selected stock and the stock return in the next period. The greater the absolute value of the information coefficient, the more effective the factor is. From the data, we can see that the ICs of the RESVOL factor, BTOP factor, LIQUIDTY factor, and SIZENL factor are all has an absolute value greater than 0.05. The SIZE factor and EARNYILD factor are both has an absolute value slightly less than 0.05, and the GROWTH factor and LEVERAGE factor have the lowest absolute values, showing that RESVOL factor, BTOP factor, LIQUIDTY factor, and SIZENL factor have stronger stock selection ability, while GROWTH factor and LEVERAGE factor have weaker stock selection ability. The stability coefficients of all 10 factors are above 0.8, and the stability coefficients of SIZE, EARNYILD, GROWTH, BTOP, LEVERAGE, LIQUIDTY and SIZENL factors are above 0.95, indicating that the above factors are stable. For IC win rate, SIZE factor, EARNYILD factor, RESVOL factor, SIZENL factor, LIQUIDTY factor, and BTOP factor have a higher prevalence than 0.6.

Table 2. Analysis of IC values for 10 style factors

Factor Name	IC Value	Stability Coefficient	Win Rate
Beta	-0.02	0.83	0.53
Residual Volatility	-0.07	0.90	0.70
Size	-0.04	0.99	0.62
Growth	0.01	0.95	0.50
Non-linear Size	0.05	0.97	0.72
Momentum	-0.03	0.87	0.57
Earnings Yield	0.04	0.96	0.62
Leverage	0.01	0.99	0.58
Liquidity	-0.09	0.96	0.68
Book-to-Price	0.06	0.98	0.63

Table 3 and Figure 2 show the annual returns of the 10 pure style factor portfolios over the 5-year period from 2015 to 2019, and Figure 3 shows the net value trend of the 10 pure style factor portfolios over the 5-year period from 2015 to 2019. The graphs of the returns of the 10 pure style factor portfolios show that during the full sample period, the Beta factor and Earnings Yield factor have a significant positive impact, the Book-to-Price factor and Growth factor have a positive but weak impact, the Liquidity factor and Non-linear Size factor have a negative impact, and the Liquidity factor and Non-linear Size factor have a negative impact. The Liquidity factor and Non-linear Size factor have a significant negative impact, the Leverage factor has a negative impact but to a lesser extent, and the Momentum factor, Residual Volatility factor, and Size factor have no fixed bias and are style rotation factors.

Table 3. Returns for each of the 10 pure style factor portfolios during year 2015 to 2019

Factor Name	Year 2015	Year 2016	Year 2017	Year 2018	Year 2019
Beta	7.36%	0.81%	1.73%	10.33%	8.44%
Residual Volatility	10.50%	-3.35%	-5.15%	-7.50%	-8.48%
Size	-23.85%	-11.19%	4.79%	3.22%	1.32%
Growth	-0.02%	1.55%	1.01%	-0.40%	2.26%
Non-linear Size	-14.84%	-15.17%	-12.41%	-8.39%	-3.07%
Momentum	-11.33%	-2.41%	7.16%	4.19%	8.98%
Earnings Yield	6.22%	5.62%	8.15%	2.59%	-1.82%
Leverage	-0.57%	-0.73%	0.02%	-2.26%	-0.10%
Liquidity	-14.18%	-11.65%	-7.16%	-10.60%	-7.93%
Book-to-Price	-0.01%	3.76%	1.26%	0.91%	-1.73%

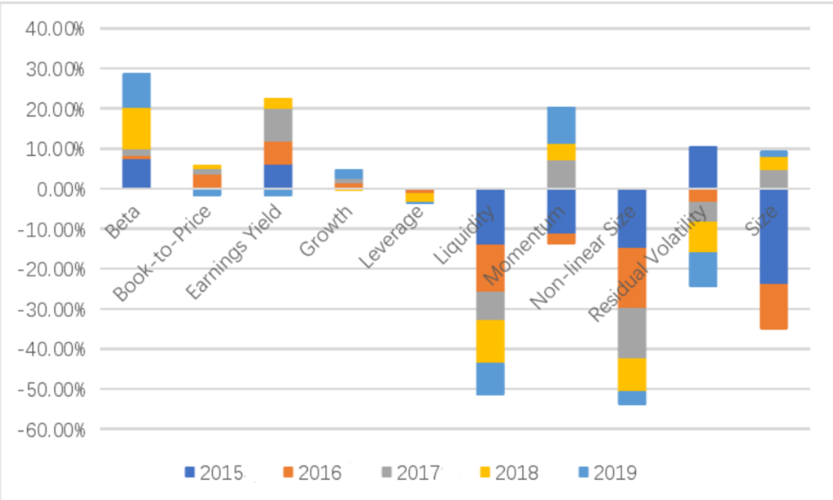

Figure 2. Returns for each of the 10 pure style factor portfolios during year 2015 to 2019

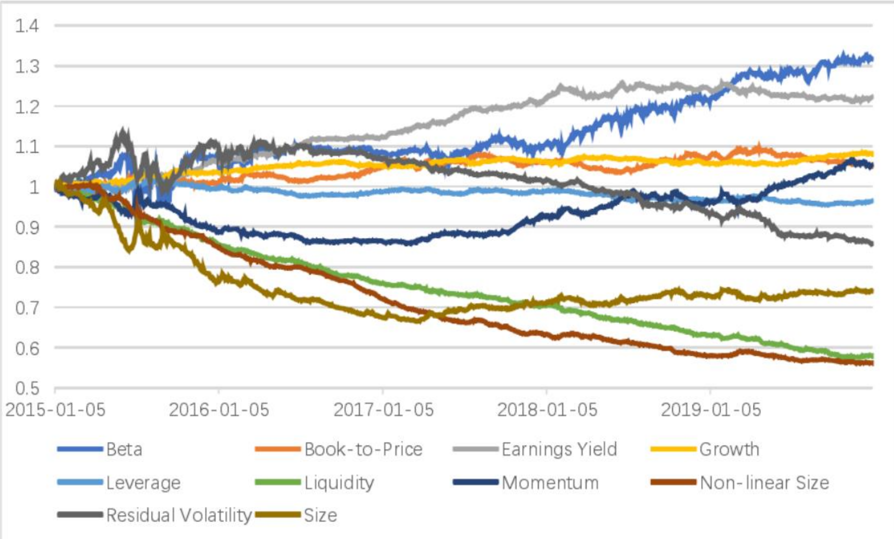

Figure 3. Net trend of 10 pure style factor portfolios during 2015 to 2019

3.4. Model Building

Through the above data processing, factor validity test and pure factor income analysis, this paper selects the following five factors that have high degree of explanation of income sources and are relatively effective for a long time as the effective factors of the model: Beta, Earnings Yield, Growth, Liquidity, and Non-linear Size. Their positive or negative correlations are summarized as follows.

Table 4. Effective factor information after rescreening

Factor Name	Linear Correlation
Beta	Positive
Growth	Positive
Non-linear Size	Negative
Earnings Yield	Positive
Liquidity	Negative

For the evaluation and screening of multi-factor model, this paper adopts the scoring method. Scoring method refers to scoring individual stocks according to the size of each factor, then calculating the total score according to the weight, and finally sorting the total score of individual stocks and realizing the screening of stocks [11]. For the evaluation of multi-factor model, the stock portfolio return measured by scoring method can evaluate the advantages and disadvantages of alternative stock selection model. Considering that the equal weight method is based on static weight, which is simpler and more stable [10], this paper selects the equal weight method to establish the multi factor stock selection model. The model is as follows:

$$Z_{Score} = X_{Beta} + X_{Earnings\ Yield} + X_{Growth} - X_{Liquidity} - X_{Non-linear\ Size} \quad (4)$$

The logic of stock selection in this paper: sort the five effective factors selected according to the degree of analysis, score the stocks under each effective factor by using the percentile integer scoring method, and select the 10 stocks with the highest score to build the portfolio.

3.5. Model Back-testing

This paper selects the historical data of five years from January 1, 2015 to December 31, 2019 for back test, so as to make an empirical analysis on the return rate of the portfolio selected based on the model.

Transaction fee: according to the current prevailing tax standard in the market, the specific fee is 0.03% of the commission when buying and 0.03% of the commission when selling, plus 0.1% stamp duty. The minimum commission for each transaction is 5 RMB.

From the back-testing results, it can be seen that the strategy is effective on the whole, with an annualized return of 3.19%, an excess return of 4.57% and a winning rate of 0.566. However, from the back test data, the sharp ratio and information ratio of the strategy show that the strategy still has a large room for optimization. The results of the combination back-testing trend chart are shown in the figure below.

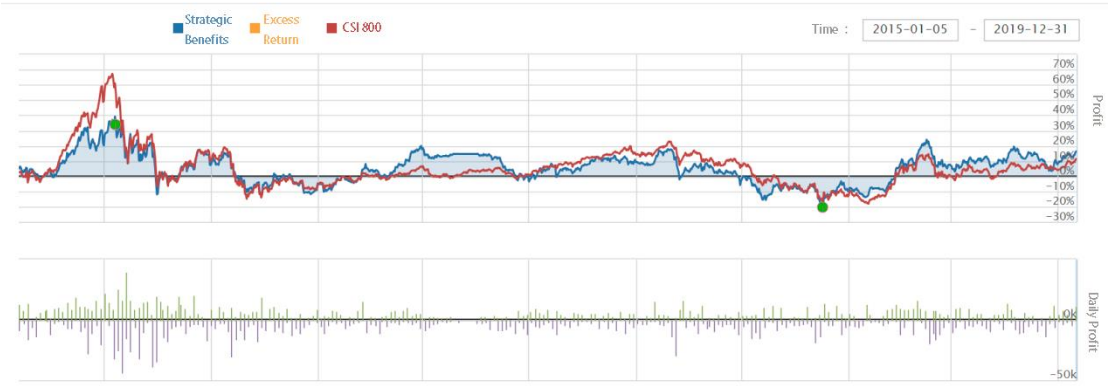

Figure 4. Model Back-testing Results

3.6. Model Optimization

ROE return on net assets is equal to the percentage of net profit and average shareholders' equity. It is one of the most important indicators that truly reflect the profitability of enterprises [12]. The higher the ROE, the stronger the profitability. From the factor architecture of MSCI Barra CNE5 model constructed in this paper, it can be seen that the ROE factor is not included in 10 style factors and components. Therefore, the following optimization is carried out: select the stocks with the top 50 scores according to formula (4), then sort them from high to low according to the latest ROE of the stocks, and select the stocks with the top 10 to build the stock pool.

3.7. Optimized Model Back-testing

From the back-testing results, it can be seen that the effectiveness and profitability of the optimization model are much higher than the original model. During the back-testing period, the strategic return increased from 16.53% to 60.39%, the excess return increased from 4.57% to 43.93%, the annualized return increased from 3.19% to 10.17%, and the maximum pullback value also decreased from 42.59% to 39.64%.

Figure 5. Optimized Model Back-testing Results

4. CONCLUSION

Based on the style factor of MSCI Barra China Equity Model (CNE5), this paper designs a set of feasible quantitative stock selection strategies for investors. The construction logic of the strategy is to analyze the data of the past market through the style factor of MSCI Barra China Equity Model (CNE5) model, and find out the factors with high and long-term effective explanation for the performance attribution of China's stock market. The factors include Beta factor, Earnings Yield factor, Growth factor, Liquidity factor and Non-linear Size factor. Then, the stock selection model is constructed through these factors, and the stock pool is constructed. The effectiveness of the stock selection strategy is verified by comparing the return of the stock pool with the trend of the market index. At the same time, in order to improve the performance of the stock selection model, we try to add ROE factor on the basis of the original model for secondary screening to further improve the effectiveness of the model. Through the back-testing data, we can see that the annual return of the optimized model exceeds 10% and a good excess return is obtained.

There are two deficiencies in this paper: first, building a model based on MSCI Barra China Equity Model (CNE5) involves a lot of data cleaning and processing, which requires a lot of workload and high professional requirements, and is not suitable for ordinary investors. Second, the MSCI Barra China Equity Model (CNE5) model has 10 style factors and 32 industry factors in addition to national factors, and the annual operation data exceeds 30 million. Therefore, when running the model, it cannot run out of verification results for a longer period of time due to machine performance.

AUTHORS' CONTRIBUTIONS

This paper is independently completed by Bingjie Cai.

ACKNOWLEDGMENTS

Thanks to The Barra China Equity Model (CNE5) website for giving such detailed introduction and plentiful resources relevant to the research in this paper. The JoinQuant platform also provides a large number of Python functions for me to program, as well as visual tools which make me more intuitive. All these things give the basic data support and flexible tools to the paper.

REFERENCES

[1] Ming Shuo, How many listed companies are there in China? Oct. 4, 2021. https://baijiahao.baidu.com/s?id=1712667104595117841&wfr=spider&for=pc.

[2] H. M. Markowitz, Foundations of portfolio theory [J] The journal of finance, 1991, 46(2): 469-477.

[3] W. F. Sharpe, A simplified model for portfolio analysis [J] Management science, 1963, 9(2): 277-293.

[11] E. W. Russell, A multiple scoring method for the assessment of complex memory functions [J] Journal of consulting and clinical psychology, 1975, 43(6): 800.

[12] Y. Cai, Q. Yin, Q. Su, et al., Prediction Method of Enterprise Return on Net Assets Based on Improved Random Forest Algorithm [C] Journal of Physics: Conference Series. IOP Publishing, 2020, 1682(1): 012083.

[4] E. F. Fama, K. R. French, Common risk factors in the returns on stocks and bonds [J] Journal of financial economics, 1993, 33(1): 3-56.

[5] E. F. Fama, K. R. French, A five-factor asset pricing model [J]. Journal of financial economics, 2015, 116(1): 1-22.

[6] E. Namvar, Barr Rosenberg and Barra [M] The Berkeley-Haas Case Series. University of California, Berkeley. Haas School of Business, 2005.

[7] D. J. Orr, I. Mashtaler, Supplementary Notes on the Barra China Equity Model (CNE5), 2012.

[8] X. Yue, J. Bai, Q. Liu, et al., Sportfolio: Stratified visual analysis of stock portfolios [J] IEEE transactions on visualization and computer graphics, 2019, 26(1): 601-610.

[9] S. Rudrapatna, A. M. Hamming, M. J. H. Wermer, et al., Measurement of distinctive features of cortical spreading depolarizations with different MRI contrasts [J] NMR in Biomedicine, 2015, 28(5): 591-600.

[10] J. Y. Liang, Research on stock selection strategy based on MSCI Barra CNE5 model factors [D] South China University of Technology, 2020. DOI: 10.27151/d.cnki.ghnlu.2020.004068.