TELKOMNIKA Indonesian Journal of Electrical Engineering
Vol.12, No.3, March 2014, pp. 2390 ~ 2397
DOI: http://dx.doi.org/10.11591/telkomnika.v12i3.4801

Using Genetic Algorithm for Index Tracking-Evidence from Shanghai 50

Jian-he Liu*[1], Xin Wu[2], Qing-Song Fang[3]
Finance Department, Zhejiang University of Finance and Economics,
No. 18 of XueYuan Street, XiaSha University-town, Hangzhou, 310018, P.R.China
*Corresponding author, email: bruceanne@163.com[1], wuxin@yeah.net[2], hexinstock@21cn.com[3]

Abstract

This paper uses historical data of Shanghai 50 Index as sample data, replaces traditional optimization methods with genetic algorithm, uses clustering analysis method to build tracking portfolio, and compares the empirical results with those of traditional optimization methods. The empirical results of Shanghai 50 index show that using genetic algorithm for index tracking can get better performance with lower volatility.

Keywords: *index tracking, tracking error, genetic algorithm, clustering analysis*

1. Introduction

In 2012, China's A share index funds with same target index had significantly different tracking errors. The tracking error differentiation of CSI 300 index funds was up to 5 times, that of ETF was even up to 10 times. Obviously, the tracking technology of A share index funds are different, which leads to different index error. This also means some index funds have limited competence to track index and the return of tracking index they provided can not meet the need of passive investors. So, it is necessary to improve index tracking performance.

2. Literature

Existing achievements of index tracking can be divided into two categories: optimization algorithms and tracking error minimization.

2.1. Optimization Algorithms

The relevant literature of optimization algorithms mainly uses optimization methods of Markowitz mean-variance model, factor model and neural network, etc. to get better optimal solution.

(1) Optimization method with mean-variance model

Perold (1984) [1] suggested using mean-variance model for index tracking. Haugen and Baker (1990) [2] thought that beta, determination coefficients and volatility can be used to measure tracking capability of mean-variance model. Fang and Wang (2005) [3] transformed multiple index tracking problem into a standard single index tracking. Gaivoronoski et al. (2005) [4] tested tracking difference between tracking portfolio and benchmark index with several methods. Li et al. (2005) [5] thought that index tracking is a secondary deviation to minimize the weighted return of benchmark index and tracking portfolio. Yu et al. (2006) [6] gave an improved Markowitz model for index tracking. And Yao et al. (2006)[7] considered a portfolio which contains only a few assets to track a financial benchmark portfolio, and designed a stochastic linear quadratic programming for index tracking and solved with semidefinite programming.

(2) Optimization method with factor model

Factor model can be used to track index. Rudd (1980) [8] proposed portfolio selection method and created a portfolio with beta which is relevant to index. Corielli and Marcellino (2006) [9] assumed that stock prices were driven by factor model. Canakgoz and Beasley (2009) [10] used mixed integer programming model based on the objective function of the linear

Received September 16, 2013; Revised October 5, 2013; Accepted October 28, 2013

regression. Liu Jianhe et al. (2012, 2013) [11, 12] used factor model to track Shanghai 50 and found that it can get better performance with factor model.

(3) Other optimization methods

Some scholars used cointegration and other optimization methods to track index, such as Alexander and Dimitriu (2006) [13], Dunis and Ho (2006) [14]. They used cointegration to optimize the weight of tracking portfolio.

2.2. Tracking Error Measurement

Tracking error is used to measure the difference between tracking portfolio's return and target index's return, and evaluate the performance of index tracking. Meade and Salkin (1990) [15] used quadratic Programming to minimize tracking error. Rudolf et al. (1999) [16] studied the results of MAD model, MADD model, MinMax model and DMinMax model. Jansen and van Dijk (2002) [17] studied tracking error minimization by limiting the number of constituent stocks in tracking portfolio. And Gaivoronski et al. (2005) discussed several measurement methods of tracking error.

In conclusion, firstly, most scholars used traditional optimization algorithms include mean-variance model, factor model and cointegration optimization method. Only some scholars studied other methods, such as neural network algorithm, etc. Secondly, most scholars built the tracking models based on stock prices or returns, but did not compare tracking model based on stock prices with that based on stock returns. How about the results if we consider both stock prices and returns? How about the optimization weights of them? Thirdly, most scholars did not consider the sample stock selection of tracking portfolio, they always used market value rank or random sampling to select sample stocks. We think these simple ways may neglect some valuable information. Focardi and Fabozzi (2004) [18] used clustering analysis method to select stocks for tracking portfolio. Unfortunately, they did not give the empirical results.

So this paper chooses to use genetic algorithm instead of traditional algorithm, tests different empirical results of price time series and return time series, and uses the clustering analysis of Focardi and Fabozzi (2004) to select tracking portfolio's constituent stocks instead of full replication method. In the first part, this paper puts forward research ideas based on existing literature. In the second part, we design genetic algorithm optimization model of index tracking. Then based on the second part, we set the parameters of tracking model in detail, such as objective function, constituent stocks, the weights of constituent stock, etc. In the fourth part, we conduct an empirical test, and compare optimization results of price time series with those of return time series. We get conclusion in the fifth part.

3. Genetic Algorithm Optimization Model of Index Tracking

3.1 Genetic Algorithm

(1) Basic principle of genetic algorithm

Genetic algorithm simulates genetic mechanism of biological evolution. Each individual of optimization solution set (that is, the solution) is encoded with binary, decimal or hexadecimal, then optimization problem can be classified by types, and the portfolio with better optimal solution can be found by iterative approach.

The fitness function is the only standard to judge the individual of solution set, which is based on objective function. The genetic algorithm's core content is to select one subset in solution set randomly, and then carry on genetic operation such as selection, hybridization and mutation. Genetic operation can make higher fitness solution and get higher breeding probability, so that the average fitness value of new solution set generation is higher than that value of old generation. With the iterative genetic operation generation after generation, it is very likely that the last solution set will contain the optimal solution.

(2) Control parameters of genetic algorithm process

a. The size of initial set.

On one hand, the larger the size of initial set is, the wider the value range of each set generation covers, the greater calculation amount of selection, hybridization and mutation is, and the less the number of iterations is. On the other hand, if the size of solution set is too small, the efficiency of selection, hybridization and mutation will decrease. So, the first step is to set a fitful size of initial set. In this paper, solution set is the set of 50 solution vector and the size is 50. One solution vector, which is combined by n stock's weights w_i (w_1 , w_2 , ... , w_n), is one solution of the objective function. In other words, each solution vector is the individual one.

b. The number of evolutionary generation

The number of evolutionary generation is the number of genetic search iterations. If enough time is allowed, the greater the number of evolutionary generation is, the higher the possibility of obtaining optimization solution is. In this paper, the generation is the new set of 50 solution vectors generated by function stepGA.

c. Parent generation and child generation

New solution set generated by genetic algorithm is called child generation and old solution set is called parent generation. Because the size of solution set is 50, the solution set $[W_1, W_2, \ldots, W_{50}]$ evolves to generate new set $[W_1', W_2', \ldots, W_{50}']$. The set $[W_1, W_2, \ldots, W_{50}]$ is parent generation and the new set $[W_1', W_2', \ldots, W_{50}']$ is child generation.

d. Selection, hybridization and mutation

The target of selection is to select excellent individuals from the solution set. And excellent individuals have chances to be parent generation and generate individuals of the next generation. The fitness function values of excellent individuals are always small. In this paper, selection is to select the solution set with small tracking error as parent generation. Hybridization, also named crossover, is to select two individuals of parent generation to crossover and generate two new individuals. In this paper, hybridization refers to selecting tracking portfolio with lower tracking error, encoding weights of portfolio constituent, and crossbreeding to generate the new tracking portfolio with new weights of constituent. And mutation refers to the changes of constituents' weights. The purpose of mutation is to avoid some good genes lost, especially after genetic operation, some individual of solution set may cause some string lose diversity and result in the loss of excellent

e. Hybridization probability

Hybridization probability p_c is used to control the probability of hybridization. If the probability is too large and chromosome strings update too fast, the individuals with low fitness function values (minimum optimization) are quickly destroyed. If the probability is too small, genetic search speed will be too slow. The hybridization probability is always ranging from 0.4 to 0.9.This paper set this value 0.5.

f. Mutation probability

Mutation probability Pm can increase the diversity of optimization solution set. A low Pm can change the position of chromosomes. But if Pm is too small, new individuals will not be generated. If the probability is too large, it will replace genetic algorithm with random search algorithm. The value of Pm ranges from 0.1 to 0.3, and we set Pm 0.2 in this paper.

g. Elite number and hybridization fraction (crossover fraction)

Elites are individuals with the lowest fitness value. They are portfolios with the smallest tracking errors in this paper. Elites in parent generation will not go through crossover and mutation, and the genes of elites will be inherited to child generation. It is called elitist strategy. Hybridization fraction is the proportion of individuals in child generation generated by hybridization to individuals those are not elites in parent generation. In this paper, hybridization fraction is the proportion of individuals with weight re-encoded in child generation to those without weight re-encoded in parent generation.

According to these definitions of control parameters, we can design related parameters and use computers to optimize the index tracking algorithm.

3.2. Model Construction

(1) Data encoding

One of the basic works of genetic algorithm is encoding. Because genetic algorithm is based on encoding mechanism, code selection is an important factor of genetic algorithm. Solutions of optimization problems can be encoded with binary or decimal. Here we use binary encoding.

(2) Fitness function

Fitness function fitness(x) is used to evaluate the individuals and their ability to adapt to environment. For optimization problems, we must establish a mapping relationship between objective function f(x) and individual fitness function fitness(x). There are two basic conditions to be met: the value of fitness function must be greater than 0 and changing direction of the fitness function must be consistent with that of the objective function during evolution process of optimization.

In this paper, the fitness function value is the function value of index tracking optimization model. Here we try to find the lowest value of objective function. The lower the objective function value is, the more the individual adapts to environment.

(3) Termination criteria of genetic algorithm

Genetic algorithms can be proved to converge with probability 1, but in practice there are certain algebraic iteration limits. Generally, it is required to set the termination criteria of iterative process in advance. There are three methods. The first one is to set the number of evolutionary generation in advance. When iteration times reach the number, the algorithm terminates. The second one is to observe the fitness value. If the fitness value does not change significantly, the algorithm terminates. The last one is to use a combination of these two methods.

4. Research Method: The Construction of Index Tracking Model

Index tracking model faces two problems. Firstly, we should choose what types of stocks to build tracking portfolio. Secondly, we must decide the weights of these stocks. In this part, we focus on these two problems in building tracking model.

4.1. The Objective Function

The key of index tracking is to establish tracking error optimization model. In this paper, the tracking model is as (1) and (2).

$$TE = \left(\sum \Delta_t \left(\left|\sum_{i=1}^{N} P_{it} w_i - I_t\right|\right)^{\alpha}\right)^{(1/\alpha)} \tag{1}$$

$$r^{*} = \sum_{t=1}^{T} (r_t - R_t) \tag{2}$$

If $\alpha=1$, TE will be a linear tracking error. If $\alpha=2$, TE will be a quadratic tracking error. Δ_t represents the contribution for tracking error in each period. We can give it different weights in different periods to achieve better tracking performance. As (2), this paper defines the difference of returns between tracking portfolio and target index as excess return. With the change of model parameter values, this model can represent linear or quadratic tracking error. It is a combination of index tracking models, such as quadratic error optimization model, mean absolute deviation model Min Max model, etc.

4.2. Model Parameters And Constraints

(1) Model parameters

N is the constituent stocks' number of target index; K is the stocks' number of tracking portfolio. e_i is the minimum holding ratio of i stock in tracking portfolio. d_i is the maximum holding ratio of i stock in tracking portfolio. X_i is the share number of i stock in tracking portfolio. T is the rebalance time point. At time t, we rebuild the tracking portfolio, and $(1, 2, \ldots, T)$ is the sample period that we test tracking portfolio and target index. P_{it} is the price per share of stock i at time t. I_t is the values of target index at time t (1,2,…,T). R_t is the returns of target index at time t (1,2,…,T), $R_t=\log(I_t/I_{t-1})$.

C_{cash} is the change of tracking portfolio's cash flow at time T. If C_{cash} is greater than 0, there is new funds inflow for tracking portfolio. If C_{cash} is less than 0, this means that funds outflow from tracking portfolio. And if C_{cash} is 0, the cash flow of tracking portfolio has no change. C_{T-h} is the total initial value of tracking portfolio, h is holding period, and $C_{T-h} = \sum_{i=1}^{N} P_{iT-h} X_i$. C is total value plus the change of tracking portfolio's cash flow at time T, $C = \sum_{i=1}^{N} P_{iT} X_i + C_{cash}$.

$F_j(\zeta,\theta,t)$ is the function of transaction cost, which is the holding cost of stock I from ζ units to θ units at time T. Because of random changes of stock price, it is very difficult to catch the best time of stock trading. But transaction cost is generally a nonlinear function and very relevant to stock price. If $\theta \geq \psi \geq \zeta$, $F_j(\zeta,\theta,t) \geq F_j(\zeta,\psi,t)$; if $\theta \leq \psi \leq \zeta$, $F_j(\zeta,\theta,t) \leq F_j(\zeta,\psi,t)$.

γ means that the maximum allowable ratio of transaction costs and C. W_j is the weight of stock j in tracking portfolio. Z_j is a dummy variable, whose value is 1 if it is being held in tracking portfolio, otherwise, its value is 0. C_{trana} is total transaction cost when stock's weight changes from X_j to w_j at time T, $C_{trans} = \sum_{i=1}^{N} P_i(X_i, w_i, T)$.

r_t is the continuous return of new tracking portfolio at time t (t=1,2 , … , T) . The constituent stock's weight of new tracking portfolio is W_j. And $r_t = \log(\sum_{j=1}^{N} P_{jt} w_j)/(\sum_{j=1}^{N} P_{jt-1} w_j))$.

(2) Model constraints

The constraints of index tracking model are as following.

$$z_j \in [0,1] \text{ , i=1,…,N} \tag{3}$$

$$\sum_{j=1}^{N} z_j = K \tag{4}$$

$$\epsilon_j z_j \le P_{jT}/C \le \delta_j z_j \text{ ,} i = 1, \cdots, N \tag{5}$$

$$\sum_{j=1}^{N} P_{jt} w_t = C - C_{trans} \tag{6}$$

$$0 \le \sum_{i=1}^{N} P_i(X_i, w_i, T) \le c\gamma \tag{7}$$

$$W_i \ge 0, i=1,\ldots,N \tag{8}$$

Equation (4) can constrain the stock number of new tracking portfolio to K. When the stock j is not selected in new tracking portfolio, w_j is 0, (5) can let the value of z_j is 0. Otherwise, when the stock i is selected in new portfolio, w_i is not 0, and (5) can make z_i is 1. Equation (6) assumes the transaction cost is one composition part of tracking portfolio's value. In other words, total value of new portfolio at time T equals to the sum of current portfolio's value and cash flow's changes. Equation (7) sets upper and lower transaction costs. And (8) is a ban on short selling.

4.3. Tracking Portfolio Selection with Clustering Analysis

In the empirical part of this paper, we use squared deviation method to select cluster sample. Squared deviation method is also called Ward minimum variance method. If classification is reasonable, squared deviations between samples in same class should be small, squared deviation between different classes should be large. The observation distance between different classes is defined as $D_{pq}^2 = S_n^2 - S_p^2 - S_q^2$, where S_n^2 is the squared deviation between G_n classes, and G_n classes are merged with G_p and G_q. If the observation distance is as $d(x, y) = ||x - y||^2/2$, the observation distance between new class G_n and other class Gk is determined by (9).

$$D_{kn} = [(N_k + N_p)D_{kp} + (N_k + N_q)D_{kp} - N_k D_{pq}]/(N_k + N_p) \tag{9}$$

When using the Ward minimum variance method for classification, the first step is to set each sample as one class. In this paper, each sample is each constituent stock in target index. Then different classes can be merged to new classes until all samples are classified as one class. Squared deviation becomes larger after merging. It is necessary to choose two classes to be merged in case that squared deviation' increase is the smallest after merging. In this paper, we use price difference between constituent stocks of target index to calculate observation distance.

4.4. Weights of Constituent Stocks of Tracking Portfolio

This paper uses time series of prices and returns respectively to calculate constituent stocks' weights in tracking portfolio, combines two time series information with simple arithmetical method and gets new constituent stocks' weights for tracking.

5. Results and Discussion

5.1. Data Selection and Processing

We choose Shanghai 50 as target index. Sample period 1 is from July 7, 2009 to November 15, 2010 and has 330 trading days. Sample period 2 is from November 16, 2010 to March 23, 2011 and also has 330 trading days. In sample period 1 and sample period 2, observed variables are constituent stocks of Shanghai 50 on November 15, 2010. The data in this paper comes from GTA database. Because trading suspension results in closing price missing of corresponding stock, we use interpolation method to calculate stock return. And we also set the period from August 26, 2010 to November 15, 2010 as out of sample period in sample period 1, the period from January 6, 2012 to March 23, 2012 as out of sample period in sample period 2.

5.2. Clustering Analysis for Selecting Constituent Stocks of Tracking Portfolio

In order to build the optimal tracking portfolios, we use clustering analysis to select constituent stocks. This paper uses hierarchical clustering to divide constituent stocks of Shanghai 50 into different classes based on same characteristics, and measures the distance between classes with the “Ward” squared deviation. So we set the name stock1, stock2, …, stock50 for constituent stocks of Shanghai 50 based on their codes, and uses stock prices from May 31, 2011 to March 23, 2012 to adopt clustering analysis.

Based on the results of clustering analysis, the constituent stocks of Shanghai 50 can be divided into 9 classes. We can randomly select one stock from each class to build the tracking portfolio, shown as table 1. At the same time, we also select top 9 constituent stocks of Shanghai 50 with the average market capitalization in April, May and June, 2012 for comparison. We name tracking portfolio based on the results of clustering analysis with clustering tracking portfolio, and name comparison portfolio of top 9 constituent stocks with market capitalization portfolio.

Accordingly, we design nine empirical optimization programs for comparison. Program 1 uses traditional algorithm to track with return time series of market capitalization portfolio. Program 2 uses traditional algorithm to track with price time series of market capitalization portfolio. Program 3 uses traditional algorithm to track with return time series of clustering tracking portfolio. Program 4 uses traditional algorithm to track with price time series of clustering tracking portfolio. Program 5 uses genetic algorithm to track with return time series of market capitalization portfolio. Program 6 uses genetic algorithm to track with price time series of market capitalization portfolio. Program 7 uses genetic algorithm to track with return time series of clustering tracking portfolio. Program 8 uses genetic algorithm to track with price time series of clustering tracking portfolio. Program 9 is different from other programs, it combines information of price time series and return time series and uses genetic algorithm to track with two time series of clustering tracking portfolio.

5.3. Test Results

We use mean and variance of return, total return, cumulative excess return, tracking error and the correlation coefficient, etc. to compare the empirical tracking results of 9 programs.

From the results of Table 3, eight programs and the target index have high correlation, and genetic algorithm can obtain larger correlation coefficient than traditional algorithm, correlation coefficients of clustering tracking portfolio are larger than those of market capitalization portfolio. Program 7 may be the best because of its largest correlation coefficient.

From means and variances of return, it is easy to find that tracking results with price time series are better than those with return time series, and clustering tracking portfolio is better than market capitalization portfolio, genetic algorithm is better than traditional algorithm. Program 7 is also the best one.

But total return and cumulative excess return show that program 8, which uses genetic algorithm to track with price time series of clustering tracking portfolio, is better.

Table 1. Constituent Stocks of Clustering Tracking Portfolio

600050	600111	600188	600547	600585
601006	601169	601958	600519	

Table 2. Constituent Stocks of Market Capitalization Portfolio

600036	601268	600519	601088	601166
601398	601601	600028	601857	

Table 3. Correlation Coefficients of Tracking Portfolio and Target Index

Program	1	2	3	4	5	6	7	8
Sample period 1	0.9617	0.9516	0.8298	0.8991	0.9647	0.9691	0.9707	0.9569
Sample period 2	0.9702	0.9242	0.9423	0.9308	0.9781	0.9434	0.9572	0.9472

Tracking performance of program 7 also is better in aspect of average daily tracking errors on out of sample period.

Table 4. Average Daily Tracking Errors in Out of Sample Period

Program	1	2	3	4	5	6	7	8
Sample period 1	4.854	6.221	3.915	4.081	2.317	6.180	2.233	4.332
Sample period 2	2.671	3.093	8.511	9.030	2.145	3.007	2.114	8.008

Note: The results are multiplied by 1000000.

Clearly, program 7 and program 8 are better than other programs. The tracking performance of using genetic algorithm to track with clustering tracking portfolio is better. Because program 7 tracks with return time series and program 8 tracks with price time series, we use simple arithmetical method to combine constituent stocks' weights in program 7 and program 8, get new constituent stocks' weights of program 9. Now we can compare tracking performance of program 9 with that of program 7 and 8.

Table 5. Correlation Coefficients of Program 7, 8 and 9

	Program 7	Program 8	Program 9
Sample period 1	0.9707	0.9569	0.9751
Sample period 2	0.9572	0.9472	0.9611

Table 6. Return Means and Variances of Program 7, 8 and 9

		Program 7	Program 8	Program 9
Sample period 1	$\overline{R_p}$	1.9799	2.0722	1.9861
	σ	0.7149	0.7258	0.7082
	ADTR	2.233	4.332	1.915
Sample period 2	$\overline{R_p}$	1.4356	1.6646	1.5501
	σ	0.6885	0.8295	0.7540
	ADTR	2.114	8.008	1.832

Note: The results of mean and average daily tracking errors (ADTR) are multiplied by 1000;
The results of variance are multiplied by 100.

Correlation coefficient of program 9 is larger than those of program 7 and 8. Tracking portfolio considering two kinds of time series is more relevant than tracking portfolio considering single time series. It means that more information results in stronger correlation.

Return mean and variance of program 9 is not the best one, but it is obvious that program 9 can decrease return volatility. And Table 7 also shows that the tracking error of program 9 is the least. It further illustrates that more information can improve index tracking performance.

Overall, genetic algorithm optimization model aided with clustering analysis and information of price and return time series can result in better tracking performance than traditional optimization model with information of single time series.

6. Conclusion

This paper replaces traditional stock selection method of top market capitalization or random sampling with clustering analysis to choose constituent stocks of tracking portfolio, uses genetic algorithm instead of traditional algorithm, and gets higher probability to find optimal solution. The empirical results of Shanghai 50 also show that genetic algorithm optimization model aided with clustering analysis can get better tracking performance with lower risk. So, the genetic algorithm optimization method is useful for improving index fund's performance and helpful for index fund's management.

Acknowledgements

This paper is sponsored by Chinese National Funds of Social Science (No. 11BJY033 and 12CJY022), Zhejiang colleges' creative team project of industrial development and public financial policy research (No.CCJFH104) and CSC project 201208330282.

References

[1] Perold AF, Large scale Portfolio optimization. *Journal of Management Science*. 1984; 30(10): 1143-1160.
[2] Haugen RA, Baker NL. Dedicated stock portfolios. *Journal of Portfolio Management*. 1990; 16(4): 17-22.
[3] Y Fang, SY Wang. A fuzzy index tracking portfolio selection model. *Lecture Notes in Computer Science*. 2005; 3(11): 554–561.
[4] Gaivoronoski, S Krylov, N Vander Wijst. Optimal portfolio selection and dynamic benchmark tracking. *European Journal of Operational Research*. 2005; 6(2): 115–131.
[5] J Li, Y Ma, Y Zeng. *Research on application of Stein rule estimation for index tracking problem*. Proceedings of the International Conference on Management Science and Engineering. 2005; 12(13): 434–439.
[6] L Yu, S Zhang, XY Zhou. A downside risk analysis based on financial index tracking models. *Stochastic Finance*. 2006; 17(11): 213–236.
[7] DD Yao, S Zhang, XY Zhou. Tracking a financial benchmark using a few assets. *Operations Research*. 2006; 11(54): 232–246.
[8] Rudd A. Optimal selection of passive portfolios. *Financial Management*. 1980; 9(1): 57-66.
[9] F Corielli, M Marcellino. Factor based index tracking. *Journal of Banking and Finance.* 2006; 32(30): 2215–2233.
[10] Canakgoz NA, Beasley JE. Mixed-integer programming approaches for index tracking and enhanced indexation. *European Journal of Operational Research*. 2009; 196: 384-399
[11] Liu Jianhe, Wu Xin, Chen Keyi. Index Tracking with Factor Model: Evidence from Shanghai 50. *Journal of Convergence Information Technology*. 2012; 7(19): 347-356.
[12] Liu Jianhe, Wu Xin, Chen Keyi. The Impact of Margin Trading on Shanghai 50 Index Tracking. *Advanced Science Letters*. 2013; 19(4): 1153-1157.
[13] Alexander Carol, Dimitriu Anca. Indexing and statistical arbitrage: tracking error or cointegration?. *Journal of Portfolio Management*. 2005; 31(2): 50–63.
[14] Dunis Christian L, Ho Richard. Cointegration portfolios of European equities for index tracking and market neutral strategies. *Journal of Asset Management*. 2005; 6(1): 33.
[15] Meade N, Salkin GR. Developing and maintaining equity index fund. *Journal of the Operational Research Society*. 1990; 41(7): 599–607.
[16] Rudolf Markus, Wolter Hans-Jurgen, Zimmermann Heinz. A linear model for tracking error minimization. *Journal of Banking and Finance*. 1999; 23(1): 85-103.
[17] Jansen Roel, van Dijk Ronald. Optimal benchmark tracking with small portfolios. *Journal of Portfolio Management*. 2002; 28(2): 33–9.
[18] Focardi SM, Fabozzi FJ. A methodology for index tracking based on time-series clustering. *Quantitative Finance*. 2004; 4(4): 417–25.