Intelligent Automation & Soft Computing
DOI:10.32604/iasc.2021.016953
Article

Tech Science Press

Main Factor Selection Algorithm and Stability Analysis of Regional FDI Statistics

Juan Huang[1], Bifang Zhou[1], Huajun Huang[2,*], Dingwen Qing[1] and Neal N. Xiong[3]

[1]Centre for Innovation Research in Social Governance, Changsha University of Science and Technology, Changsha, 410114, China
[2]College of Information Technology and Management, Hunan University of Finance and Economics, Changsha, 410205, China
[3]Department of mathematics and computer science, Northeastern State University, OK, 74464, USA
[*]Corresponding Author: Huajun Huang. Email: huanghuajun@hufe.edu.cn

Received: 16 January 2021; Accepted: 04 May 2021

Abstract: There are various influencing factors in regional FDI (foreign direct investment) and it is difficult to identify the main influencing factors. For this reason, a main factor selection algorithm is proposed in this article for the main factors affecting regional FDI statistics by analyzing the regional economic characteristics and the possible influencing factors in the regional FDI. Then, an example is used to illustrate its effectiveness and its stability. Firstly, the characteristics of regional economy and the regional FDI data are introduced to develop the main factor selection algorithm based on the adaptive Lasso problem for the regional FDI and to establish the corresponding computing procedure. Then, based on the regional FDI statistical data of six provinces in the central China, the main factor selection algorithm is used to filter out the insignificant factors and identify the main influencing factors for the different regional FDI statistics, including the mean values, the median values, the maximum values, and the minimum values. Finally, the proposed algorithm is validated through an accuracy test experiment performed in central China. On this basis, its corresponding stability with the noise error case is analyzed and the control stability range of the algorithm is determined.

Keywords: Lasso problem; adaptive Lasso problem; main factor selection algorithm; regional FDI statistics; main factors affecting regional FDI Statistics

1 Introduction

Economic globalization is essential for economic development, but requires as well the rapid and balanced development among all regions [1]. In the new pattern of global economy, as the overall scale increases and the regional economic cooperation deepens, it is necessary to carry out innovation in the form of regional cooperation [2].

Due to historical and geographical reasons, there are many differences in the regional economies between various regions in China. Since the reforms and the opening up to the world, China has achieved success in implementing the policies aimed to revitalize its economy by taking the lead in the east, develop the western and central regions and revive the northeast. The level of economic development in the east is significantly higher than in the central and western regions while the level of economic development in central China is also higher than in the west. Liu et al. argued that the main reasons for the gap of regional economic development include development strategies and policies, regulations, and the degree of openness [3]. Whether economic development is constrained by the local market conditions of various regions [4], the priority of Chinese economy has been shifted in recent years. Take the economic data of various regions in the first half of 2018 as an example. The overall trend is positive. Based on the data of different regions, the fastest growth is achieved in the central region, with an average growth rate of 7.91%, which is higher than 7.23% for the west, while the slowest is linked to the northeast where the average growth rate is 4.53%. Overall, China has performed well in attracting FDI. In 2015, China became the largest country of foreign capital inflow around the world, the growth of which varies between different regions. Therefore, it is important to analyze the influencing factors in regional FDI for improving the growth rate of regional FDI.

For the analysis of influencing factors in regional FDI, there have been many researchers around the world publishing papers. Wang et al. applied the grey correlation theory to confirm and analyze the relationship between different influencing factors and the variation in regional economic development [5]. Xu et al. [6] conducted the co-integration test and constructed the error correction model to investigate the influencing factors in the destination of FDI in China. Zhou et al. [7] adopted the index DEA model to construct the indicators of high-quality economic development, and then performed the panel quantile regression to explore the impact of FDI on high-quality economic development. Lu et al. [8] relied on stepwise regression, co-integration and error correction models to analyze the influencing factors in FDI. Li et al. [9] conducted the grey relational analysis to compare and analyze the significance of correlation between the influencing factors for region FDI in Henan province from 1990 to 2009. Wei Zhou [10] analyzed the effects of variables related to industrial transfer on the FDI from the perspective of the driving factors in the industrial transfer of Beijing-Tianjin-Hebei urban agglomeration. With the data collected from 31 provinces as the samples, Zhou et al. [11] applied the static panels and spatial measurement models to analyze the correlation of FDI with the conditions of regional economic growth, financial development as well as the impact on industry development and industry optimization [12]. Currently, the global economy is getting more and more interdependent. In the meantime, the focus of research has shifted to exploring how to use the mathematical models to analyze the influencing factors in regional FDI, predict the trends of FDI reasonably and accurately, and promote regional economic development through foreign capital for China.

Although the influencing factors in regional FDI are complex and changeable, most scholars both in China and abroad relied on multivariate statistical or economic methods to analyze their patterns using time series data or panel data. It was found out that the impact of influencing factors on FDI was significant. However, it is inevitable for the above methods to be affected by the random choice of variables and multicollinearity. In this paper, a main factor selection algorithm is proposed to address the adaptive Lasso problem for computing the main influencing factors in FDI. This algorithm can not only eliminate multicollinearity, but also achieve the selection [13–16] and estimation of variables, thus providing effective reference for the regional FDI statistics in central China.

2 Description of Main Factor Selection Algorithm

2.1 Explanation of Lasso and Related Methods

Before the model of the Lasso problem is introduced, a simple linear regression model is presented as follows.

$$\mathrm{y} = \mathrm{X}\beta + \varepsilon. \tag{1}$$

where $y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$, $x_j = \begin{bmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{nj} \end{bmatrix}$, $X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} = \left[x_1, x_2, \cdots, x_p\right]$, $\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}$, $\varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$, p denotes the number of the characteristic parameters and n denotes the number of samples. In this paper, the following marks will also be used: $\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$, $\|\beta\|_2^2 = \sum_{j=1}^{p} \beta_j^2$ and $\|\beta\|_2 = (\sum_{j=1}^{p} \beta_j^2)^{1/2}$.

The RSS (residual sum of squares) of this model is

$$\mathrm{RSS} = (y - X\beta)^2. \tag{2}$$

The optimized objective function of β can be represented to minimize RSS as follows.

$$\min_{\beta} (y - X\beta)^2. \tag{3}$$

If the inverse of the matrix $X^T X$ exists, the least square estimation of β can be computed as $\widehat{\beta}_{ols} = (X^T X)^{-1} X^T y$.

If $n < p$ the matrix $X^T X$ is not a full-rank matrix, which means the above formula can't be used. Therefore, it is found out that there are an infinite number of solutions to Eq. (3) according to the knowledge related to the matrix, which indicates the difficulty in determining an optimal solution with the data provided, so that the problem is over-fitting.

2.2 Lasso Problem

The most common method used to solve the over-fitting equation is to regularize the parameters, so that l_1 regularization is introduced to avoid over-fitting. The optimized objective function is shown as follows.

$$\min_{\beta} \left\{ (y - X\beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}. \tag{4}$$

The above problem is called the Lasso problem, where λ is a penalty parameter and $\lambda \geq 0$, which is equivalent to

$$\min_{\beta} (y - X\beta)^2, s.t \sum_{j=1}^{p} |\beta_j| \leq C. \tag{5}$$

where $C \geq 0$ is the penalty parameter. Assume $\widehat{\beta_j^0}$ is the least square estimation of β_j, $C_0 = \sum_{j=1}^{p} \left|\widehat{\beta_j^0}\right|$. If $C \geq C_0$, the solution worked out using the least square method is the optimal solution of (4). If $C < C_0$, it indicates that part of coefficients β_j will be reduced or close to 0, or even equal to 0. Then, these coefficients will be eliminated for variable selection, so as to obtain the main factors.

In order for the Lasso problem to learn the mapping relationship [17] of the feature parameters in regional FDI, the near-end gradient descent algorithm [18] is applied to solve the Lasso problem.

Before the near-end gradient descent algorithm is introduced, some preliminary knowledge is presented briefly.

The optimization goal in Lasso problem [19] is

$$\min_{\beta}\left\{(\mathrm{y}-\mathrm{X}\beta)^2+\lambda\sum_{\mathrm{j}=1}^{\mathrm{p}}\left|\beta_{\mathrm{j}}\right|\right\}. \tag{6}$$

It can also be described as

$$\min_{\beta}\{\mathrm{f}(\beta)+\lambda\beta_1\}. \tag{7}$$

where $\mathrm{f}(\beta)=\|\mathrm{y}-\mathrm{X}\beta\|^2$, $\|\beta\|_1=\sum_{\mathrm{j}=1}^{\mathrm{p}}\left|\beta_{\mathrm{j}}\right|$, λ is a penalty parameter and $\lambda\geq 0$. When $\lambda=0$, Eq. (6) will be converted into the ordinary least square estimation.

The process of the near-end gradient descent algorithm is detailed as follow.

1) The optimized problem is

$$\min_{\beta}\{\mathrm{f}(\beta)+\lambda\mathrm{g}(\beta)\}. \tag{8}$$

where $\beta\in\mathrm{R}^{\mathrm{p}\times 1}$, $\mathrm{f}(\beta)$ represents a differentiable convex function and $\mathrm{g}(\beta)$ refers to a convex function. For $\mathrm{f}(\beta)$, if $\mathrm{f}(\beta)$ can be derived, $\nabla\mathrm{f}(\beta)$ satisfies the Lipsitz continuity condition [19]. For β_1 and β_2, there always exists a constant $\mathrm{L}>0$,

$$\nabla\mathrm{f}(\beta_2)-\nabla\mathrm{f}(\beta_1)_2^2\leq \mathrm{L}\beta_2-\beta_{12.}^2 \tag{9}$$

2) When $\beta\to\beta_{\mathrm{k}}$, the approximation of $\mathrm{f}(\beta)$ is calculated by means of Taylor expansion as follows.

$$\begin{aligned}\mathrm{f}(\beta)&\approx\mathrm{f}(\beta_{\mathrm{k}})+\nabla\mathrm{f}^{\mathrm{T}}(\beta_{\mathrm{k}})\cdot(\beta-\beta_{\mathrm{k}})+\tfrac{\mathrm{L}}{2}\beta-\beta_{\mathrm{k}2}^2\\&=\tfrac{\mathrm{L}}{2}\beta-(\beta_{\mathrm{k}}-\tfrac{1}{\mathrm{L}}\nabla\mathrm{f}(\beta_{\mathrm{k}})_2^2+\mathrm{const.}\end{aligned} \tag{10}$$

where β_{k} represents the kth iteration result of β, const is a constant, ∇ denotes the differential operator and T indicates the transpose.

3) The original solution at k + 1 iterations is

$$\beta_{\mathrm{k}+1}=\min_{\beta}\{\mathrm{f}(\beta)+\lambda\mathrm{g}(\beta)\}. \tag{11}$$

In the near-end gradient descent algorithm, the following formula can be solved to obtain $\beta_{\mathrm{k}+1}$ as follows.

$$\beta_{\mathrm{k}+1}=\min_{\beta}\{\mathrm{f}(\beta)+\lambda\mathrm{g}(\beta)\}$$

$$=\min_{\beta}\left\{\frac{\mathrm{L}}{2}\beta-(\beta_{\mathrm{k}}-\frac{1}{\mathrm{L}}\nabla\mathrm{f}(\beta_{\mathrm{k}})_2^2+\mathrm{const}+\lambda\mathrm{g}(\beta)\right\}. \tag{12}$$

If Due to $\mathrm{g}(\mathrm{x})=\|\beta\|_1$, let $\mathrm{z}=\beta_{\mathrm{k}}-\frac{1}{\mathrm{L}}\nabla\mathrm{f}(\beta_{\mathrm{k}})$, const can be omitted as a constant, so $\beta_{\mathrm{k}+1}=$

$$\min_{\beta}\left\{\frac{\mathrm{L}}{2}\beta-\mathrm{z}_2^2+\lambda\beta_1\right\}. \tag{13}$$

4) With (12) solved, set $\mathrm{F}(\mathrm{x}) = \frac{\mathrm{L}}{2}\sum_{\mathrm{j}=1}^{\mathrm{p}}\left(\beta_{\mathrm{j}} - \mathrm{z}_{\mathrm{j}}\right)^2 + \lambda\sum_{\mathrm{j}=1}^{\mathrm{p}}\left|\beta_{\mathrm{j}}\right|$, for β_{j} (the jth element in β), let

$$\frac{\partial \mathrm{F}(\mathrm{x})}{\partial \beta_{\mathrm{j}}} = \mathrm{L}\left(\beta_{\mathrm{j}} - \mathrm{z}_{\mathrm{j}}\right) + \lambda \mathrm{sgn}\left(\beta_{\mathrm{j}}\right) = 0. \tag{14}$$

Get:

$$\mathrm{z}_{\mathrm{j}} = \beta_{\mathrm{j}} + \frac{\lambda}{\mathrm{L}}\mathrm{sgn}\left(\beta_{\mathrm{j}}\right). \tag{15}$$

where sgn($*$) represents a symbolic function, $\mathrm{sgn}(*) = \begin{cases} 1 & * > 0 \\ 0 & * = 0 \\ -1 & * < 0 \end{cases}.$

In order to solve $\beta_{\mathrm{j}}{}^{*}$, use the function of z_{j} to express $\beta_{\mathrm{j}}{}^{*}$, so swap the axis $\beta - \mathrm{z}$ and

$$\begin{aligned} \beta_{\mathrm{j}} &= \mathrm{sgn}\left(\mathrm{z}_{\mathrm{j}}\right)\left(\left|\mathrm{z}_{\mathrm{j}}\right| - \frac{\lambda}{\mathrm{L}}\right) \\ &= \mathrm{sgn}\left(\mathrm{z}_{\mathrm{j}}\right) \cdot \max\left(\left|\mathrm{z}_{\mathrm{j}}\right| - \frac{\lambda}{\mathrm{L}}, 0\right). \end{aligned} \tag{16}$$

5) Iteration, otherwise, if $\left|\left(\mathrm{f}\left(\beta_{\mathrm{k}+1}\right) + \lambda \mathrm{g}\left(\beta_{\mathrm{k}+1}\right)\right) - \left(\mathrm{f}\left(\beta_{\mathrm{k}}\right) + \lambda \mathrm{g}\left(\beta_{\mathrm{k}}\right)\right)\right| < 10^{-4}$, stop.

2.3 Main Factor Selection Algorithm for Adaptive Lasso Problem

Based on the Lasso problem [20], the variable selection method assigns the same weight to the different coefficients in the Lasso solving method. As for the adaptive Lasso problem, the basic idea is to give the small weight to the small coefficients for punishment and then convert the adaptive Lasso problem into a Lasso problem through suitable transformation for obtaining their solution. In fact, the policy selection methods [21] can be used to achieve the purpose. However, they are usually subjected to some limitations. It is demonstrated that the adaptive Lasso method is the most ideal choice, which relies on penalty function to compress the variable coefficients. As suggested by Zou et al. [22], an effective penalty function is supposed to have three characteristics: continuity, unbiasedness and sparsity. Therefore, the solution to adaptive Lasso problem has a wider scope of applications as compared to the traditional mathematical statistics and the Lasso problem. According to the above requirements, this paper adopts the main factor selection algorithm to analyze FDI using regional FDI statistics.

The optimization goal of the adaptive Lasso problem is

$$\min_{\beta}\left\{\{\mathrm{y} - \mathrm{X}\beta\}^2 + \lambda\sum_{\mathrm{j}=1}^{\mathrm{p}}\widehat{\omega}_{\mathrm{j}}\left|\beta_{\mathrm{j}}\right|\right\}. \tag{17}$$

where, $\widehat{\omega}$ represents a known weight vector. The selection algorithm is based on some assumptions as follows. An estimate of the real model β is $\widehat{\beta}$, and the weight vector is defined as $\widehat{\omega} = 1/\left|\widehat{\beta}\right|^{\gamma}(\gamma > 0)$. Herein, for $\widehat{\beta}$, it is appropriate to choose the ordinary least squares estimation $\widehat{\beta}_{(\mathrm{ols})}$, the ridge estimation $\widehat{\beta}_{(\mathrm{Ridge})}$ and so on. In this paper, the ridge estimation is selected as $\widehat{\beta}_{(\mathrm{Ridge})}$ and $\gamma = 1$, then the weight vector $\widehat{\omega} = 1/\left|\widehat{\beta}_{(\mathrm{ols})}\right|$.

Main Factor Selection Algorithm:

1) Set $\mathrm{x}_{\mathrm{j}}{}^{*} = \mathrm{x}_{\mathrm{j}}/\widehat{\omega}_{\mathrm{j}},\ \mathrm{j} = 1, 2, \ldots, \mathrm{p}$.

2) Solve the following Lasso problem:

$$\min_{\beta}\left\{ \{y - \sum_{j=1}^{p} x_j{}^{*}\beta\}^{2} + \lambda \sum_{j=1}^{p} \widehat{\omega}_j \left|\beta_j\right| \right\}. \tag{18}$$

3) Since the adaptive Lasso problem (18) is a convex optimization problem, the near-end gradient descent algorithm is applied to solve the following Lasso problem as follows:

$$\widehat{\beta}^{*} = \underset{\beta}{\operatorname{argmin}} \{y - \sum_{j=1}^{p} x_j{}^{*}\beta\}^{2} + \lambda \sum_{j=1}^{p} \omega_j \left|\beta_j\right|. \tag{19}$$

4) Get $\widehat{\beta_j}^{*} = \frac{\widehat{\beta}^{*}}{\widehat{\omega}_j}, j = 1, 2, \ldots, p$.

3 Application Cases

For the main factor selection algorithm, the regional FDI in central China is exemplified to demonstrate its effectiveness and stability. The six provinces located in central China have different status of development and their respective advantages. To some extent, there is variation in the ability to attract those large and medium foreign-funded enterprises [23]. Therefore, it is necessary to analyze the six central provinces for regional FDI using the relevant data, so as to figure out the differences between these six provinces and develop targeted regional policies.

Upon investigation and statistics, there werev14 characteristic parameters selected from these six provinces located in central China, including the regional GDP (x_1), the average wage (x_2), the gross fixed asset formation (x_3), the road mileage (x_4), TIAE(total value of import and export) (x_5), the ratio of the industrial added value increment of the secondary and tertiary industry to the GDP increment (x_6), the *per capita* fiscal expenditure of government personnel (x_7), the total freight (x_8), TRS(total retail sales of consumer goods) (x_9), the number of patents granted to the region (x_{10}), the proportion of fiscal expenditure in GDP (x_{11}), the number of the industries above designated size (x_{12}), the number of the high education students (x_{13}) and the amount of FDI inflows in the previous five years (x_{14}) for these six provinces from 2006 to 2018. Then, the average, median, maximum, and minimum of these 14 factors are inputted into the model, with 80% of the samples randomly selected. That is to say, the 10-year data is taken as the training sample, and the remaining 20% is treated as the test sample.

The data used in this article is sourced from "Hunan Province Statistical Yearbook”, “Hubei Province Statistical Yearbook”, “Henan Province Statistical Yearbook”, “Shanxi Province Statistical Yearbook”, “Anhui Province Statistical Yearbook”, “Jiangxi Province Statistical Yearbook”, NBS (National Statistics Bureau), and SAFE (State Administration of Foreign Exchange). Due to the large amount of data and the space limit of this article, the data cannot be further detailed here.

3.1 Convergence Analysis of Main Factor Selection Algorithm for Regional FDI statistics

In order to illustrate the convergence of the algorithm proposed in this paper, the average, the median, the maximum and the minimum of the above 14 factors are taken as the input, while those of the annual actual utilization of FDI in these six provinces in central China are taken as the output to illustrate the characteristics of regional FDI.

The algorithm of regional FDI statistics with the adaptive Lasso problem used to compute its mean after 27 iterations shows an iteration error of 0.00009, which is less than the required iteration error1e-4, so that the iteration ends. The algorithm for the adaptive Lasso problem with the median values after 5 iterations shows an iteration error of 0.00005, which is less than the required iteration error1e-4, so that the iteration is terminated. The iteration error for the solution to the adaptive Lasso problem based on the maximum value after 36 iterations is 0.00007, which is less than the required iteration error1e-4, so that the iteration is terminated. After 16 iterations of the minimum-based algorithm for the adaptive Lasso problem, the iteration error is 0.00009, which is less than the required iteration error 1e-4, so that the iteration is terminated. The iteration error process is shown in Fig. 2.

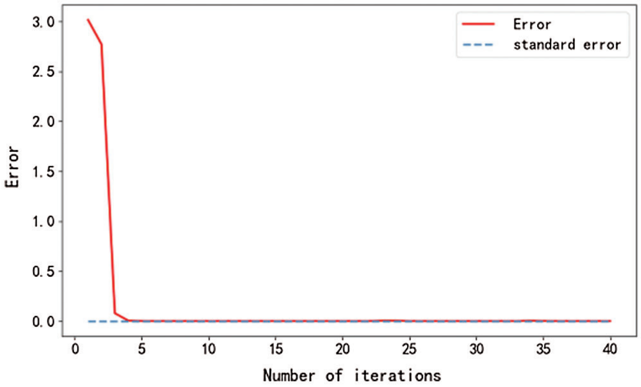

Figure 1: Mean value iterative errors

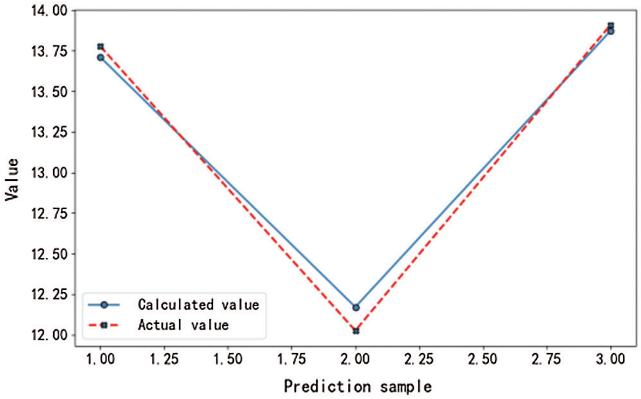

Figure 2: Computedmean value results

From the curves shown in Figs. 1–8, it can be seen that the main factor selection algorithm for the adaptive Lasso problem can converge rapidly and produce satisfactory results.

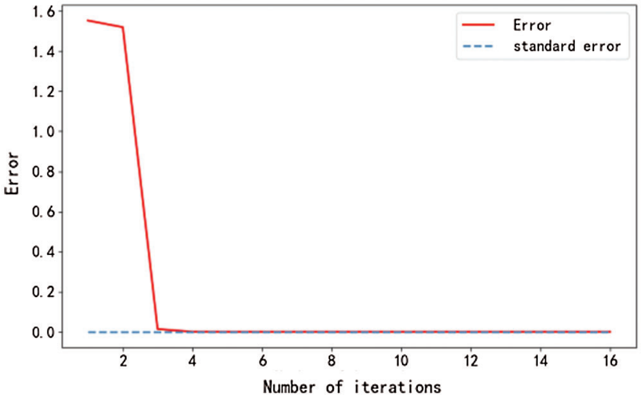

Figure 3: Median value iterative errors

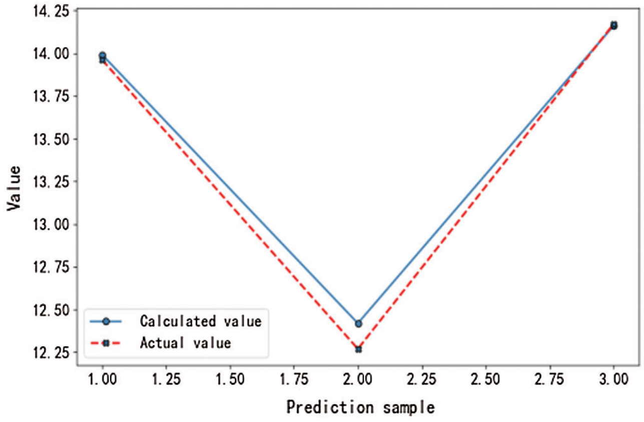

Figure 4: Computed median value results

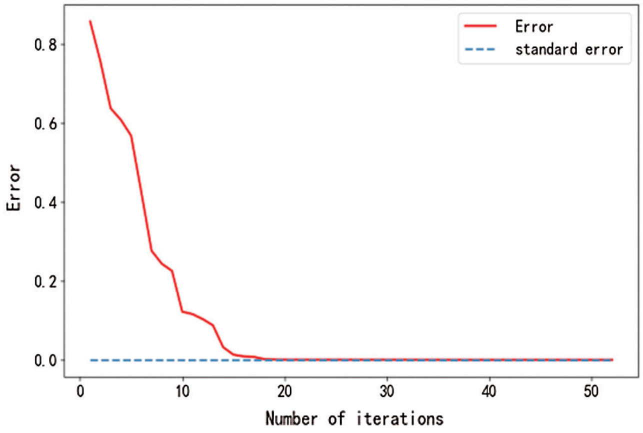

Figure 5: Maximum value iterative errors

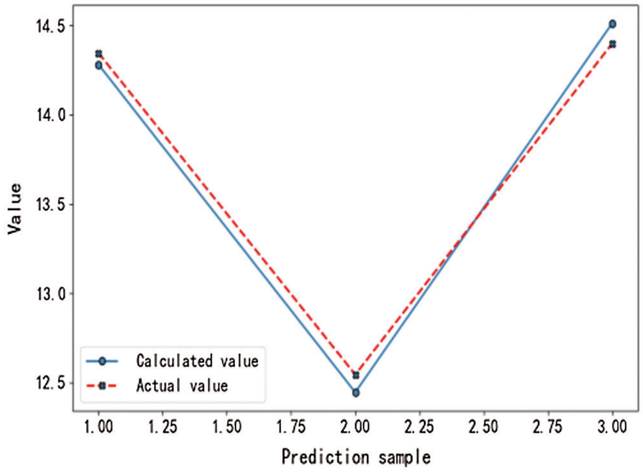

Figure 6: Computed maximum value results

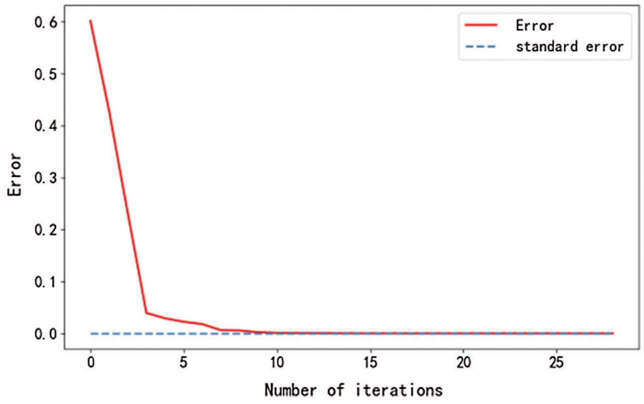

Figure 7: Minimum value iterative errors

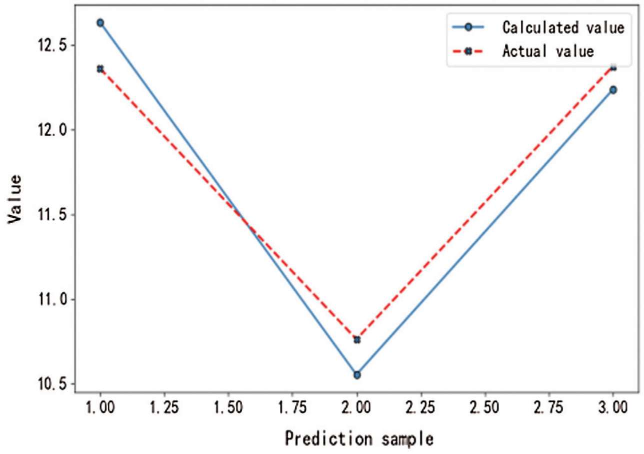

Figure 8: Computed minimum value results

After data training, the results of β_k are computed, as shown in Tab. 1.

Table 1: Adaptive Lasso coefficients of regional FDI statistics in six provinces of central China

	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}	β_{11}	β_{12}	β_{13}	β_{14}
Mean-based	0.0000	0.0798	0.4013	0.0000	0.7393	0.0000	0.0000	0.3314	0.0000	0.0000	0.4883	0.4537	0.0000	0.0000
Median-based	0.0000	0.1474	0.0000	1.2842	0.1035	–0.0877	0.3749	–0.1131	0.0000	0.0000	–0.6243	0.0000	1.5757	0.0000
Maximum-based	0.0726	1.0317	0.0000	–1.7003	0.0000	0.0000	0.0000	0.1097	0.0000	0.3890	0.0000	–0.2105	–0.2838	0.0000
Minimum-based	0.0000	–1.7700	0.0000	0.0000	2.4221	0.0000	0.0000	–1.4037	0.0000	0.0000	3.2232	–1.7540	8.8308	–1.9021

From Tab. 1, it can be seen clearly that the coefficients of the adaptive Lasso estimation (x_1), (x_4), (x_6), (x_7), (x_9), (x_{10}), (x_{13}) and (x_{14}) based on the mean values, (x_1), (x_3), (x_9), (x_{10}), (x_{12}) and (x_{14}) based on the median values,(x_3), (x_5), (x_6), (x_7), (x_9), (x_{11}) and (x_{14}) based on the maximum value and (x_1), (x_3), (x_4), (x_6), (x_7), (x_9) and (x_{10}) based on the minimum value are 0. That is to say, it is effective in eliminating the variables with multicollinearity when modeling is performed. In addition, it can be found out that the regional FDI in the six provinces based on the mean, the median, the maximum and the minimum are affected by 6, 8, 7 and 7 factors, respectively, suggesting that the main influencing factors are different. Except for the mean, the regional FDI in the six provinces based on the median, the maximum, and the minimum are negatively correlated with 3, 3, and 4 factors, respectively, indicating that the different statistics of regional FDI are affected by different factors, and that the extent of influence varies significantly. It is necessary to adjust the relevant economic policies and the economic factors according to the different target needs when the main factor selection algorithm is applied to analyze the regional FDI in the regulatory regions using different statistics.

3.2 Effectiveness of Main Factor Selection Algorithm and Its Stability

In this section, there are two kinds of errors used for the validity analysis: RMSE (root mean square error) and MAE (mean absolute error). The former can measure the deviation between the calculated value and the actual value, while the latter (mean absolute error) is the average of absolute errors, which can reflect the calculated value error faithfully. The RMSE and the MAE between the actual FDI statistics and the computed regional FDI statistics are shown in Tab. 2.

Table 2: Error between calculated results and actual values of regional FDI statistics

	Mean		Median		Maximum		Minimum	
	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
Adaptive Lasso	0.0949	0.0830	0.0893	0.0626	0.0944	0.0922	0.2135	0.2062

From Tab. 2, it can be found out that the RMSE and the MAE of the calculated values by the main factor selection algorithm for the regional FDI statistics are small, indicating that the algorithm is capable of selecting the main factors required to identify the influencing factors in regional FDI.

Then, in order to illustrate the stability of the established algorithm, the different degrees of the noise [24] are added into the data for evaluating the reliability of the data. In this section, we randomly select S sample feature data and add the different degrees of the noise. The values of S are 36 (accounting for 20% of the total number of data), 55 (accounting for 30%), and 72 (accounting for 40%). The noise is uniformly distributed in $U(-R, R)$, where R represents the upper bound of the ratio of the noise value to the original number, and the values are 2%, 10%, 20% and 40%. After the addition of noise to the data, A and setting $A' = A(1 + R)$, the noise-added data sample is used for the prediction of model training. The results are shown in Tab. 3 as follows.

Table 3: Error analysis of noise addition experiment based on average (R: Upper bound of the ratio of noise value to actual values)

R	Error class	S(Sample feature data of mean)				S(Sample feature data of median)			
		0%	20%	30%	40%	0%	20%	30%	40%
0%	RMSE	0.0949	0.0949	0.0949	0.0949	0.0893	0.0893	0.0893	0.0893
	MAE	0.083	0.083	0.083	0.083	0.0626	0.0626	0.0626	0.0626

Table 3 (continued).

R	Error class	S(Sample feature data of mean)				S(Sample feature data of median)			
		0%	20%	30%	40%	0%	20%	30%	40%
2%	RMSE	0.0949	0.1151	0.162	0.1262	0.0893	0.0962	0.1023	0.0946
	MAE	0.083	0.0742	0.1195	0.1034	0.0626	0.0588	0.0718	0.0658
10%	RMSE	0.0949	0.0926	0.0957	0.1422	0.0893	0.1115	0.1114	0.1131
	MAE	0.083	0.0846	0.0847	0.1269	0.0626	0.0729	0.1091	0.0719
20%	RMSE	0.0949	0.0825	0.1185	0.2006	0.0893	0.116	0.1153	0.127
	MAE	0.083	0.0754	0.0828	0.1964	0.0626	0.0824	0.093	0.1025
40%	RMSE	0.0949	0.1143	0.1382	0.2263	0.0893	0.1551	0.1591	0.167
	MAE	0.083	0.1141	0.1458	0.2028	0.0626	0.1312	0.138	0.1425
R	Error class	S(Sample feature data of maximum)				S(Sample feature data of minimum)			
		0%	20%	30%	40%	0%	20%	30%	40%
0%	RMSE	0.0944	0.0944	0.0944	0.0944	0.2135	0.2135	0.2135	0.2135
	MAE	0.0922	0.0922	0.0922	0.0922	0.2062	0.2062	0.2062	0.2062
2%	RMSE	0.0944	0.0897	0.098	0.1055	0.2135	0.2448	0.2848	0.2811
	MAE	0.0922	0.0615	0.0703	0.0945	0.2062	0.2636	0.2436	0.2613
10%	RMSE	0.0944	0.0883	0.0807	0.1159	0.2135	0.2651	0.2961	0.3113
	MAE	0.0922	0.1029	0.1024	0.1058	0.2062	0.2858	0.2859	0.2661
20%	RMSE	0.0944	0.1045	0.1227	0.1345	0.2135	0.3155	0.3817	0.4891
	MAE	0.0922	0.1261	0.131	0.1543	0.2062	0.3676	0.3835	0.47083
40%	RMSE	0.0944	0.1499	0.1765	0.2064	0.2135	0.4381	0.5780	0.6519
	MAE	0.0922	0.1658	0.2048	0.1849	0.2062	0.5694	0.6379	0.6154

From Figs. 9–12, it can be seen that the RMSE and the MAE of the prediction results show an evident increasing trend when the upper bound of noise R>20%, irrespective of how much noise is added to the sample. That is to say, when the experimental data deviates from the actual value by 20%, the overall performance of the main factor selection algorithm will deteriorate, the error value of the network will increase significantly, the reliability of the prediction result will be reduced, and its stability will be affected. Therefore, it is necessary to prevent the data error from exceeding 20% of the actual value when the mean and the median are computed.

From Figs. 13–16, it can be seen that the RMSE and the MAE of the prediction results show an obvious increasing trend when the upper bound of noise R>10%, regardless of how much noise is added to the sample. That is to say, when the experimental data deviates from the actual value by 10%, the overall performance of the main factor selection algorithm will decline, the error value of the network will increase significantly, the reliability of the prediction result will be reduced, and its stability will be affected. Therefore, it is necessary to prevent data error from exceeding 10% of the actual values when the maximum and the minimum are computed.

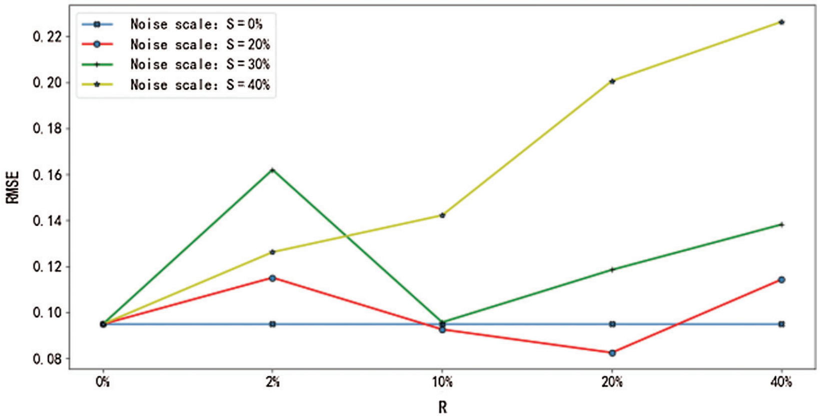

Figure 9: RMSE of mean values with noise

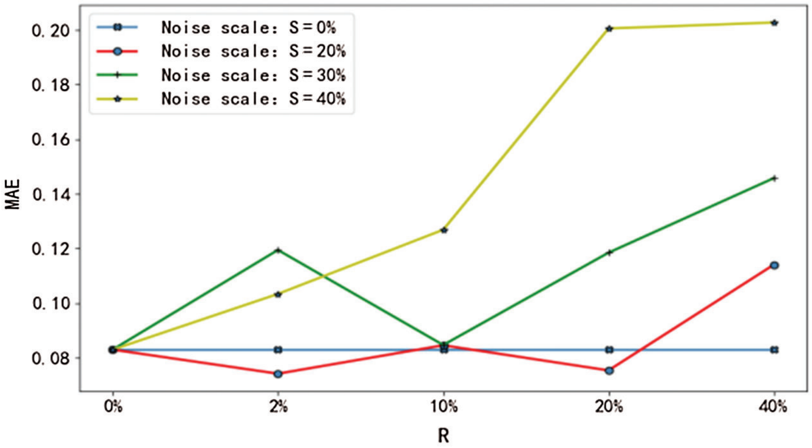

Figure 10: MAE of mean values with noise

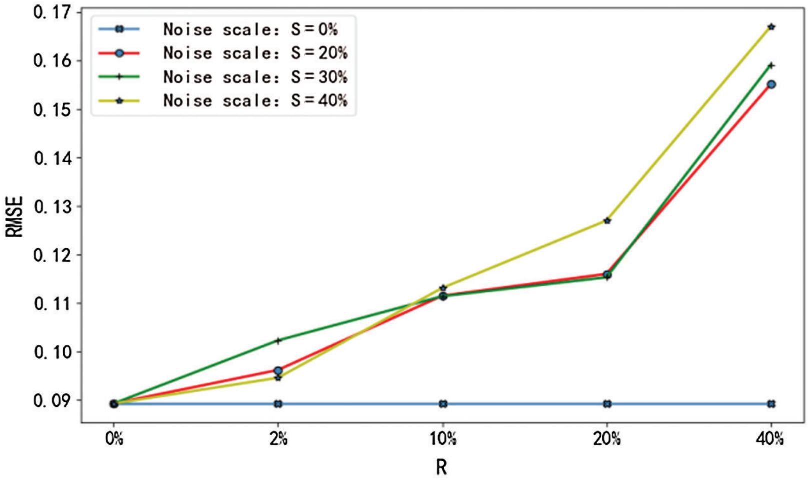

Figure 11: RMSE of median values with noise

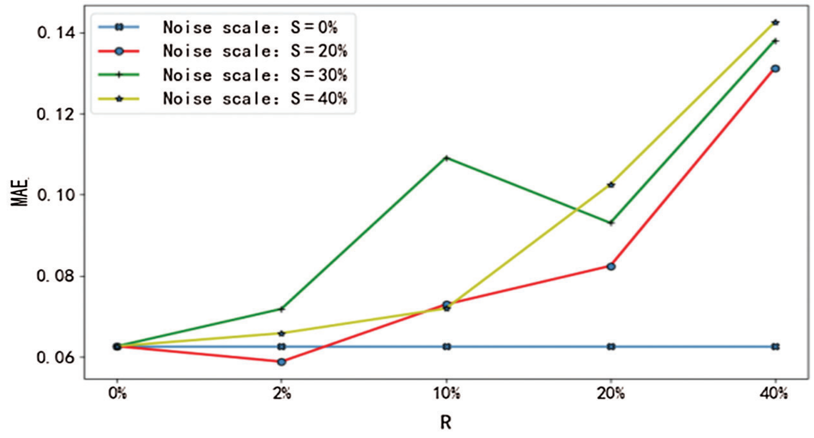

Figure 12: MAE of median values with noise

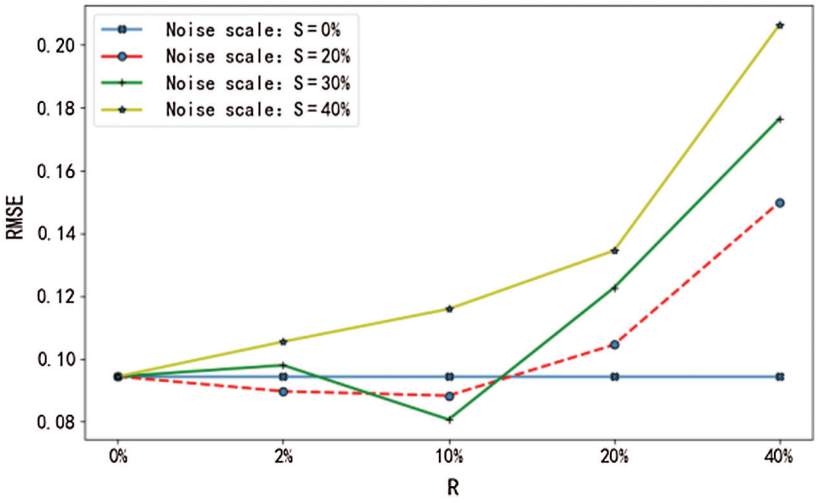

Figure 13: RMSE of maximum values with noise

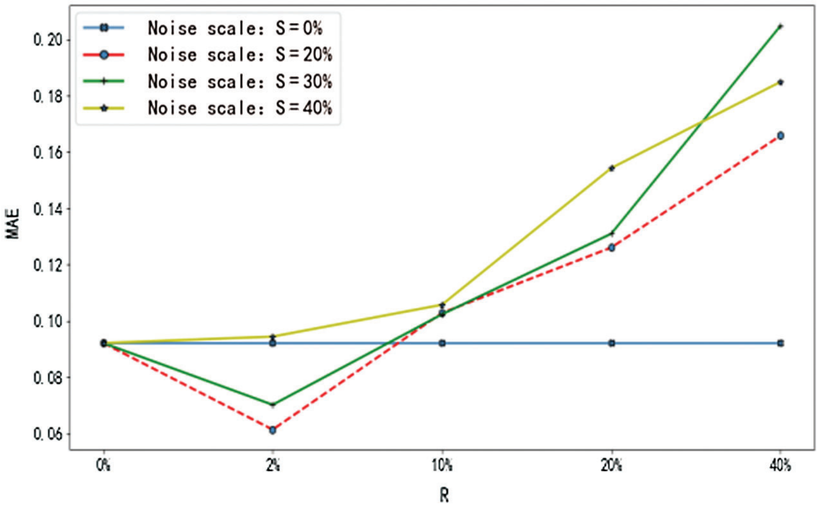

Figure 14: MAE of maximum values with noise

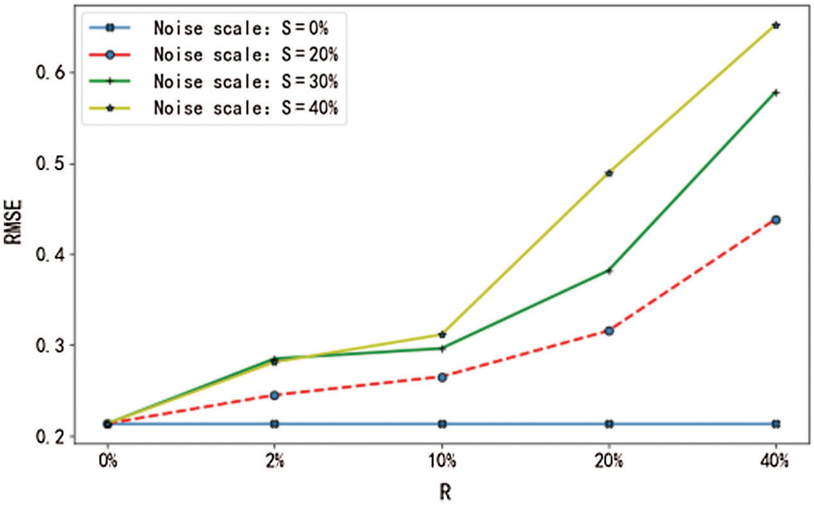

Figure 15: RMSE of minimum values with noise

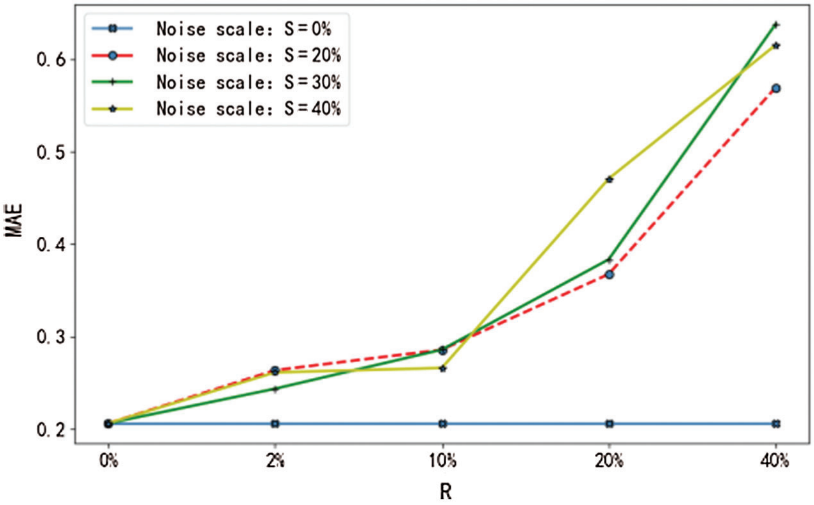

Figure 16: MAE of minimum values with noise

Therefore, the algorithm proposed in this paper performs more consistently in computing the regional FDI statistics when the parameter data error of the regional FDI falls within the stable range of the actual value.

4 Conclusion

Based on variable selection and statistical theory, the main factor selection algorithm is applied to identify the main influencing factors in regional FDI statistics. Then, the regional FDI statistics examples obtained from central China are used to perform verification. Besides, the random noise data experiments are conducted on the characteristic data. Finally, the stability range of regional FDI prediction is determined for the six provinces in central China. According to the results, the algorithm is effective in identifying the main influencing factors in regional FDI. In addition, it is also revealed that the main influencing factors in different regional FDI statistics are quite different, indicating that the FDI statistics in regional economy are affected by different factors. Therefore it is necessary to formulate different policies as reference for the control and development of regional FDI statistics when FDI statistics change in the control area. The algorithm proposed in this paper is also applicable for other economic statistics.

Funding Statement: This project is supported by the National Key R&D Program of China under grant 2019YFE0122600; Open Fund of Centre for Innovation Research in Social Governance of Changsha University of Science and Technology under grant 2017ZXB07; National Social Science Foundation of China under grant16BKS024; Major projects of Hunan Social Science Fund under grant 16ZD02 and Hunan 13th five-year Education Planning Project under grant XJK19CGD011.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] P. Lu, "Research on the relationship between economic globalization and regional economic integration," *Shanxi Agricultural Economics*, vol. 21, pp. 19–20, 2018.

[2] Q. Zeng, H. Wu and F. Liu, "New features of global economic pattern and significance of 'Belt and Road' initiative," *Research on Technical Economy and Management*, vol. 8, pp. 113–117, 2018.

[3] Y. Fang, "A study on the imbalance of regional economic development in China," *Productivity Research*, vol. 7, pp. 69–72, 2020.

[4] X. Li, Y. Wei and X. Liu, "Convergence or divergence? A literature review," *Economic Research Journal*, vol. 7, no. 39, pp. 70–81, 2004.

[5] Y. Wang, "Research on the gap between FDI and regional economic development of China based on grey relational theory," *Systems Engineering*, vol. 30, no. 3, pp. 426–430, 2010.

[6] T. Xu, F. Hong and L. Zhang, "A demonstration on influencing factors of FDI location choice: Based on co-integration and error correction model," in *Proc. of the 2010 Int. Conf. on Logistics Engineering and Intelligent Transportation Systems*, pp. 1–4, 2010.

[7] Z. Zhou, L. Deng, H. Xiao, S. Wu and W. Liu, "The impact of foreign direct investment on high-quality economic development in China-An analysis based on Index DEA and panel quantile regression," *Management Science in China*, vol. 10, no. 4, pp. 1–12, 2020.

[8] S. Lu and M. Xin, "An empirical study on regional influencing factors of FDI and countermeasures in liaoning province," in *Proc. of the 2010 Int. Conf. on Management Science & Engineering 17th Annual Conf. Proc.*, pp. 859–864, 2010.

[9] H. Y. Li and D. M. Li, "An empirical study of influencing factors of absorbing FDI in Henan," in *Proc. of the 2011 Int. Conf. on E-Business and E-Government (ICEE)*, pp. 1–3, 2011.

[10] W. Zhou, "The effect of the driving factors of Jingjinji industry transfer on FDI," *Statistics and Decision Making*, vol. 36, no. 17, pp. 110–114, 2020.

[11] B. Zhou and H. Shao, "FDI, financial development and regional economic growth: A spatial econometric analysis based on provincial panel data," *Economic Restructuring*, vol. 4, pp. 150–157, 2020.

[12] Z. Zhou, J. H. Qin, X. Y. Xiang, Y. Tan, Q. Liu *et al.,* "News text topic clustering optimized method based on TF-IDF algorithm on spark," *Computers, Materials & Continua*, vol. 62, no. 1, pp. 217–231, 2020.

[13] Z. He, "Building a nest to attract a phoenix: Infrastructure and foreign direct investment," *Modern Management Science*, vol. 12, pp. 45–47, 2017.

[14] Z. D. Wang, J. H. Qin, X. Y. Xiang and Y. Tan, "A privacy-preserving and traitor tracking content-based image retrieval scheme in cloud computing," *Multimedia Systmes,* 2021.

[15] T. Q. Zhou, B. Xiao, Z. P. Cai and M. Xu, "A utility model for photo selection in mobile crowd sensing," *IEEE Transactions on Mobile Computing*, vol. 20, no. 1, pp. 48–62, 2021.

[16] T. Xu, M. Zhao, X. Yao and K. He, "An adjust duty cycle method for optimized congestion avoidance and reducing delay for WSNS," *Computers Materials & Continua*, vol. 65, no. 2, pp. 1605–1624, 2020.

[17] Q. Liu, X. Y. Xiang, J. H. Qin, Y. Tan, J. S. Tan *et al.,* "Coverless steganography based on image retrieval of DenseNet features and DWT sequence mapping," *Knowledge-Based Systems*, vol. 192, pp. 105375–105389, 2020.

[18] S. Yu and J. Zhang, "Lasso-based study on the factors influencing foreign direct investment," *Journal of the Hunan University (Social Science Edition)*, vol. 28, no. 2, pp. 53–56, 2014.

[19] J. Li and C. Xin, "Economic growth effect and regional heterogeneity of foreign direct investment," *Urban Problem*, vol. 4, pp. 51–61, 2020.

[20] W. T. Ma, J. H. Qin, X. Y. Xiang, Y. Tan and Z. B. He, "Searchable encrypted image retrieval based on multi-feature adaptive late-fusion," *Mathematics*, vol. 8, no. 1019, pp. 1–15, 2020.

[21] L. Y. Xiang, S. H. Yang, Y. H. Liu, Q. Li and C. Z. Zhu, "Novel linguistic steganography based on character-level text generation," *Mathematics*, vol. 8, pp. 1558, 2020.

[22] H. Zou, "The adaptive lasso and its oracle properties," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1418–1429, 2006.

[23] X. Liang, Y. Luo and D. Peng, "Comprehensive evaluation of industrial carrying capacity in central China," *Finance and Economics*, vol. 7, pp. 91–96, 2020.

[24] Y. J. Luo, J. H. Qin, X. Y. Xiang and Y. Tan, "Coverless image steganography based on multi-object recognition," *IEEE Transactions on Circuits and Systems for Video Technology,* 2021.