ИНФОРМАТИКА

UDC 519.233.5
MSC 62P12

Research of investment attractiveness based on cluster analysis

D. Qi, V. M. Bure

St. Petersburg State University, 7–9, Universitetskaya nab., St. Petersburg, 199034, Russian Federation

For citation: Qi D., Bure V. M. Research of investment attractiveness based on cluster analysis. *Vestnik of Saint Petersburg University. Applied Mathematics. Computer Science. Control Processes*, 2023, vol. 19, iss. 2, pp. 199–211. https://doi.org/10.21638/11701/spbu10.2023.206

The continued economic development of various countries or regions has resulted in increased competition in global markets, leading to a concentration of investors and skilled labour in locations with high investment attractiveness. The investment attractiveness of a given country or region is determined by its investment potential and risk, which are characterized by a combination of various significant factors.This paper seeks to develop an econometric model to estimate the amount of investment in fixed capital in a specific region, taking into consideration the linear relationship between the observed results, in order to determine the main conditions that are necessary for achieving stable and high economic growth. These conditions include the acceleration of investment activity and the implementation of major national reforms to ensure the effectiveness of the investment process. To assess the overall influence of the financial and economic indicators studied on the volume of investment, multiple regression analysis was utilized as the primary mathematical tool of the study. Furthermore, assumptions were made regarding the rank of the observations. To validate this hypothesis, a cluster analysis was conducted, grouping the observations into four clusters based on their results, depending on the volume of investment or the geographical characteristics of the region.

Keywords: investment attractiveness, cluster analysis, hierarchical regression model, multiple regression models, correlation analysis, least squares method.

1. Introduction. Over the past decade, the world's political and economic power balance has undergone significant changes, impacting not only individual national economies, but also the economies of countries and regions on a global scale. Consequently, the 'attractiveness' of a country has become a key indicator used by international investors to determine the level of investment that is appropriate for a given nation, based on its external business environment [1]. China has demonstrated a strong level of resilience, leading

© St. Petersburg State University, 2023

to a steady economic recovery and sustained prosperity. The economic outlook for China is promising, providing a key foundation for the attraction of global investors. China's appeal to businesses is primarily due to its sizeable market and high return on investments; however, in the current context of heightened risks and uncertainties, the security of investments is also a crucial consideration in corporate decision-making, not just in terms of economic returns. Enhancing the business environment and making policy formation more evidence-based and predictable will help to bolster the assurance of enterprises to operate within China [2].

The empirical data used in this analysis and modelling is taken from the World Bank website (The World Bank. Available at: http://data.worldbank.org/indicator, accessed on January 6, 2023) and the China Statistical Yearbook (National Bureau of Statistics. Available at: http://www.stats.gov.cn/tjsj/ndsj/, accessed on January 30, 2023 (in Chinese)), and provides numerical data on the level of investment in China by region and the values of various factors.

2. The foundation for building the model. In order to ensure the validity of the results, a restriction must be placed on the factors of the econometric model utilized for this statistical study, requiring that they be quantifiable. The main purpose of this study is to construct models that can evaluate the efficacy of investment in the region, accounting for geographical disparities. Investment activity is assumed to be contingent on the presence of investment conditions, and so it is recommended that an indicator of the volume of fixed capital investment in the region be used as a dependent factor. It is essential to consider that the volume of fixed capital investment is a multifaceted economic characteristic which is impacted by a wide array of socio-economic characteristics. The initial step is to determine the exogenous factors that may be present in the model. It is necessary to include indicators that describe both financial and physical geographical, legal, socio-cultural and environmental characteristics. Therefore, the following factors are taken into account: income per capita, cost of fixed assets, the amount of work carried out by the type of activity "Construction", Gross national product (GNP) per capita and unemployment rate.

In order to construct the model, additional research must be conducted under the following assumptions. The first supposition is that the amount of investment in the region dictates the level of investment activity, and the second is that the appeal of investment in the region is primarily contingent on the financial climate. The objective of the study is to build an econometric model that can estimate the amount of investment in fixed capital in the region, accounting for the linear relationship that exists between the observed results. It is assumed that the amount of fixed capital investment in the region is dependent on a number of socio-economic indicators, which can be represented by a function. Construct a function of the form:

$$y_i = f(x_{1i}, x_{2i}, x_{3i}, x_{4i}, x_{5i}), \quad i = 1, \ldots, 31. \tag{1}$$

In (1) i — region; y_i — estimate of the volume of investments in the current year; x_{1i} — average per capita money income in the current year; x_{2i} — the cost of fixed assets in the current year; x_{3i} — the amount of work performed by the type of activity "Construction" in the current year; x_{4i} — GNP per capita in the current year; x_{5i} — unemployment rate in the current year (in percent) .

It is customary in economics to use logarithms when constructing econometric models due to the belief that this will improve the statistical properties of the estimates. Without taking logarithms, the errors may be heteroskedastic, which can reduce the efficiency

of least squares estimates. This makes it difficult to draw accurate statistical conclusions about the quality of the estimates obtained. Taking logarithms ensures that the variance is constant, meaning that the error variance does not increase as the value of the independent variable increases. As such, taking the logarithm of the dependent variable with respect to e is necessary to ensure the accuracy of the statistical analysis. The new function takes the following form:

$$\ln y_i = f(x_{1i}, x_{2i}, x_{3i}, x_{4i}, x_{5i}), \quad i = 1, \ldots, 31.$$

All calculations for this study were carried out in the "RStudio" statistical data processing environment, which ensures efficient handling of large datasets and provides numerous opportunities for data visualization.

3. Cluster analysis of the Chinese region. Prior to constructing the assessment model, a hierarchical clustering analysis was performed on the data due to its small number of observations. Ward's method was employed for this purpose, as it is based on the idea that, when properly classified, the sum of squared differences between samples or variables within a class should be minimized, while the sum of squared differences between classes should be maximized. This method is well suited for the task of classifying indicators in this particular study. The basic idea behind Ward's method is to let the n samples or variables form one class, and then reduce one class at a time, with the sum of squared deviations increasing for each reduced class, select the two classes that increase S the least and merge them, and so on, until all samples or variables are grouped into one class [3].

Dividing the n samples into k classes, denoted G_1, G_2, G_3, $\ldots$, G_k, the sum of squares of the deviations of the samples in class t is

$$S_t = \sum_{i=1}^{n_t} (X_{it} - \bar{X}_t)^T (X_{it} - \bar{X}_t),$$

where X_{it}(m-dimensional vector) denotes the i-th sample in G_t, n_t denotes the number of samples in G_t, and X_t is the centre of gravity of G_t. If class G_p and class G_q are merged into a new class G_r, there are three intra-class divergence sums of squares S_p, S_q and S_r. The increased divergence sum of squares is $D_{pq}^2 = S_r - S_p - S_q$. If the two classes G_p and G_q are close to each other, D_{pq}^2 should be smaller and the classification is more reasonable; otherwise, D_{pq}^2 is larger and the classification would be unreasonable. Thus, when the sum of the squares of the deviations added by combining the two classes is viewed as a squared distance, there is the distance formula

$$D_{pq}^2 = \frac{n_p n_q}{n_r} \left(X_p - \bar{X}_q\right)^T \left(X_p - \bar{X}_q\right)$$

and recurrence formula

$$D_{kr}^2 = \frac{n_k + n_p}{n_r + n_k} D_{kp}^2 + \frac{n_k + n_q}{n_r + n_k} D_{kq}^2 - \frac{n_k}{n_r + n_k} D_{pq}^2.$$

The Ward method uses the Euclidean squared distance as its classification statistic, such that for any two samples i and j the Euclidean squared distance can be defined as

$$\begin{aligned} d_{ij}^2 = &\left(X_{i1} - X_{j1}\right)^2 + \left(X_{i2} - X_{j2}\right)^2 + \cdots + \\ &+ \left(X_{im} - X_{jm}\right)^2 = \sum_{n=1}^{\mathrm{m}} \left(X_{in} - X_{jn}\right)^2. \end{aligned} \tag{2}$$

In formula (2) X_{in} and X_{jn} denote the n-th variable value for the i-th sample and the n-th variable value for the j-th sample, respectively. In order to reduce the impact of variable magnitude on the distance measurements between samples, it is common to standardize the variables and use the standardized values for cluster analysis.

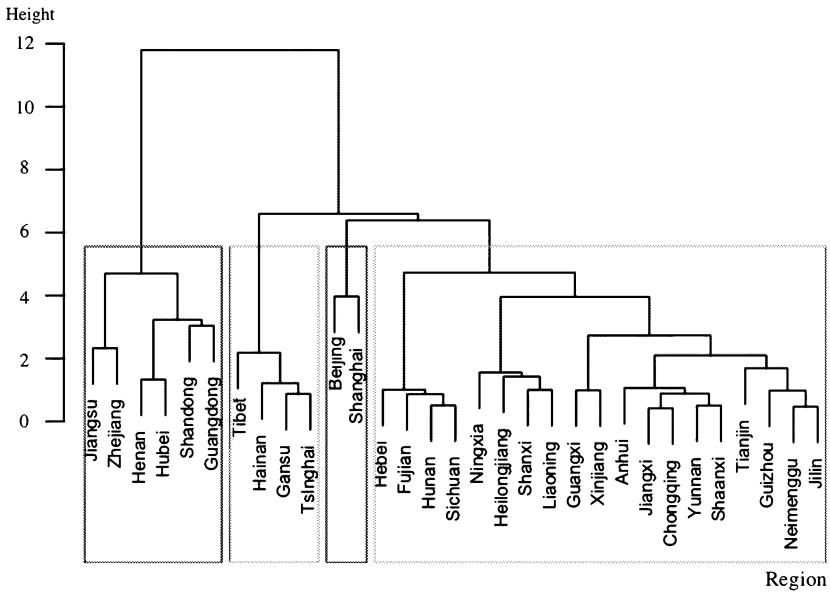

Figure 1. Cluster dendrogram

The method is implemented in R as follows [4]:
library("factoextra")
$d < -dist(my\ data,\ method = "euclidean")$
$res.hc < -hclust(d,\ method = "ward.D2")$.

The results of the calculations are presented on a tree diagram (Figure 1). Data obtained from the World Bank [5] and the China Statistical Yearbook [6] takes inflation into account (all quoted at comparable prices). The figure illustrates the number of clusters into which the data has been divided, with each cluster represented by a distinct color. The data was divided into four clusters and the results were evaluated visually, demonstrating that the data displays hierarchical characteristics. Additionally, the descriptive statistics of the clusters support this conclusion.

Each cluster will be analyzed independently.

Group 1 encompasses six regions, which are characterized by the highest investment volumes. Although there are notable disparities in the values of individual factors, the differences in the investment attractiveness differentials between regions are not statistically significant. The descriptive statistics regarding the classification of the studied indicators are presented in the Table 1.

Group 2 encompasses four regions that have comparatively low investment levels. The descriptive statistics of this group are more consistent in comparison to Group 1.

Group 3 comprises two regions, Beijing and Shanghai, which are the most affluent regions in the whole nation, and the disparities between the two regions in each factor are relatively minor.

Table 1. **Descriptive statistics of clusters 1–4**

Parameter	x_1	x_2	x_3	x_4	x_5	y
Cluster 1						
Min	2 693	20 854	10 087	35 478	2.5	10.36
1st Qu.	21 067	24 835	11 398	46 357	2.625	10.42
Median	28 380	33 730	12 434	62 201	2.750	10.62
Mean	26 116	31 348	16 920	63 335	2.833	10.63
3rd Qu.	34 519	37 118	23 775	82 561	2.950	10.84
Max	42 046	39 658	27 957	89 705	3.40	10.92
Cluster 2						
Min	11 547	1 376	147.9	1 311	2.3	7.589
1st Qu.	12 395	2 456	279.1	2 296	2.6	8.096
Median	17 506	3 310	364.8	3 544	2.7	8.309
Mean	14 778	2 973	675.8	3 965	2.7	8.219
3rd Qu.	19 889	3 828	761.6	5 212	:2.8	8.433
Max	22 553	3 897	1 825.4	7 460	3.1	8.670
Cluster 3						
Min	57 230	10 946	6 426	28 015	1.400	8.888
1st Qu.	57 669	11 258	7 254	28 669	2.025	8.924
Median	58 109	11 570	8 082	29 324	2.650	8.960
Mean	58 109	11 570	8 082	29 324	2.650	8.960
3rd Qu.	58 548	11 881	8 909	29 978	3.275	8.996
Max	58 988	12193	9 737	30 633	3.900	9.032
Cluster 4						
Min	16 704	3 807	549.2	3 444	2.200	8.224
1st Qu.	20 571	10 003	2 675.8	15 999	3.250	9.366
Median	21 484	10 467	4 726.4	19 425	3.500	9.772
Mean	24 843	12 285	5 074.6	21 459	3.479	10.130
3rd Qu.	25 183	14 951	6 628.0	28 825	3.900	10.130
Max	58 988	19 083	11 400.3	36 980	4.200	10.417

Group 4 is the largest, containing nineteen regions. Not only do the regions in this group widely differ in their investment amounts, but their other factors also vary greatly.

Evaluating the relevance of choosing the number of clusters k is important for the application of the "k-means" method [7]. The dependence of the intra-group scattering on the number of clusters is plotted using the "elbow method", which reveals a significant decrease at $k = 2$ and a stabilization at $k = 4$ (Figure 2). This suggests that the clusters divided by $k = 4$ are more consistent with the geographic and economic situation. As such, the Chinese regions were divided into four clusters that differ in terms of investment level and geographical characteristics. The results of the cluster analysis show that there is a relationship between the investment attractiveness of regions and factors such as per capita income, the cost of fixed assets, GNP, the amount of work performed by the type of activity "Construction" and unemployment rate. It is therefore meaningful to use the method of multiple regression analysis to build models based on the observations in these clusters.

4. Building econometric models. For large clusters a multiple regression analysis will be performed for each year of data. The generated models will be evaluated as a whole and general conclusions will be drawn. Given that separate models will be developed for each group of observations, it is meaningful to assume that the relationship between the factors is linear. For small clusters correlation analysis is used to assess the degree of influence of the factors.

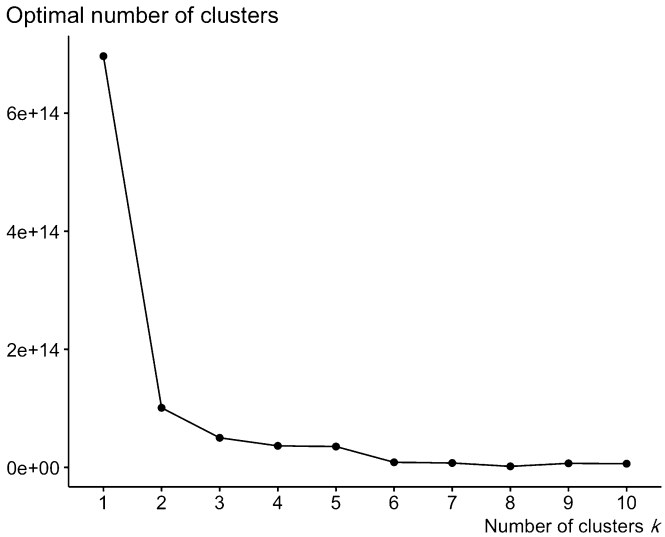

Figure 2. Choice relevance k of clusters

4.1. Multiple regression mode for group 4. The group 4 comprising of 19 regions was selected as the primary sample for analysis.

Based on the data collected from each of the 19 regions in group 4, the following observational model was employed:

$$\ln y_{2017i} = \alpha + \beta_1 \cdot x_{1i} + \beta_2 \cdot x_{2i} + \beta_3 \cdot x_{3i} + \beta_4 \cdot x_{4i} + \beta_5 \cdot x_{5i} + \epsilon_i, \tag{3}$$

where i — region; ϵ_i — the total effect of factors not taken into account by the model.

The estimates of the regression equation (3) were found and the results are presented in Table 2.

Table 2. **Coefficients estimates of the group 4**

Coefficient	Estimate	Std. Error	t-value	$\Pr(>\|t\|)$	Signif.
$\hat{\alpha}$	1.005e+01	4.828e-01	20.817	2.29e-11	***[1]
$\hat{\beta_1}$	−2.508e-05	8.018e-06	−3.128	0.00801	**[2]
$\hat{\beta_2}$	6.247e-05	3.399e-05	1.838	0.08901	.[3]
$\hat{\beta_3}$	3.220e-05	3.968e-05	0.812	0.43167	
$\hat{\beta_4}$	2.418e-05	1.956e-05	1.236	0.23822	
$\hat{\beta_5}$	−3.442e-01	1.334e-01	−2.580	0.02284	*[4]

[1] A p-value less than 0.001 indicates very strong evidence against the null hypothesis.
[2] A p-value less than 0.01 indicates even stronger evidence against the null hypothesis.
[3] A p-value less than 0.1 is considered weak evidence against the null hypothesis.
[4] A p-value less than 0.05 indicates a strong evidence against the null hypothesis and we can reject it in favor of the alternative hypothesis.

Therefore, the regression equation is as follows:

$$\begin{aligned}\ln \hat{y_{2017i}} =& 10.05 - 0.00002508 \cdot x_{1i} + 0.00006247 \cdot x_{2i} + 0.0000322 \cdot x_{3i} + \\ &+ 0.000002.418 \cdot x_{4i} - 0.3442 \cdot x_{5i}.\end{aligned}$$

According to the t-test, three of the five coefficient estimates ($\hat{\alpha}$, $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\beta}_5$) are statistically significant at the 5 % level of significance, with their p-value being less than 0.1. The values of R^2 and R^2_{adj} are 0.86 and 0.8062, respectively, indicating that approximately 80 % of the variation in the dependent variable is explained by the regression. Furthermore, the overall quality of the model is sufficient, as evidenced by Fisher's criteria: F-statistic = 15.98, with a corresponding p-value = $3.756e-05$, which is less than 0.05 and close to zero:

R^2	0.86
R^2_{adj}	0.8062
F	15.98
p-value (F)	3.756e-05

This result confirms that the average quality of the whole model is satisfactory.

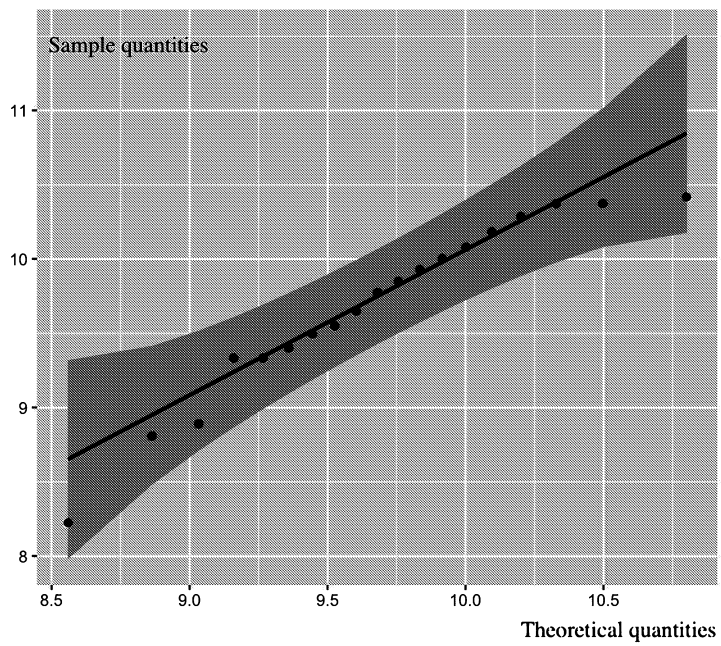

Figure 3. QQ plot for the group 4 residuals

The normality of the residual distribution needs to be evaluated, which can be done through the use of a QQ plot(Figure 3). The points on the plot generally lie along a single line, indicating that the distribution is normal, although there are some outliers present:

lag	Autocorrelation	DW statistic	p-value
1	0.15803	1.665274	0.194

χ^2	p-value (Breusch — Pagan)
2.945036	0.086142

The autocorrelation of the residuals was evaluated to verify that the selection of factors included in the model was accurate. The Durbin — Watson criterion was used to calculate DW = 1.6653, indicating that the residuals were not autocorrelated. Nevertheless, it was

still necessary to ensure that the residuals were homoscedastic, for which the Breusch—Pagan test was utilized, resulting in $\chi^2 = 2.945036$ and corresponding p-value$= 0.0861$. The analysis of the regression residuals yielded satisfactory results.

The stepwise regression algorithm is utilized to identify the factor with the least significant coefficient (maximum p-value) in each step, and this factor is then dropped in order to obtain the final model:

$$\ln \hat{y_{2017i}} = 9.666 - 0.00001741 \cdot x_{1i} + 0.000126 \cdot x_{2i} - 0.3164 \cdot x_{5i}.$$

Apply the same process to the data for other years 2009–2016:

2009:

$$\ln \hat{y_{2009i}} = 7.473 + 0.0001078 \cdot x_{2i} + 0.00000001253 \cdot x_{3i} + 0.00003056 \cdot x_{4i};$$

2010:

$$\ln \hat{y_{2010i}} = 7.988 - 0.00003046 \cdot x_{1i} + 0.0001484 \cdot x_{2i} + 0.00000001938 \cdot x_{3i};$$

2011:

$$\ln \hat{y_{2011i}} = 8.539 + 0.000000003164 \cdot x_{3i} + 0.00006634 \cdot x_{4i} - 0.1231 \cdot x_{5i};$$

2012:

$$\ln \hat{y_{2012i}} = 9.025 - 0.00001169 \cdot x_{1i} + 0.00006653 \cdot x_{4i} - 0.1351 \cdot x_{5i};$$

2013:

$$\ln \hat{y_{2013i}} = 9.452 - 0.00001755 \cdot x_{1i} + 0.00001159 \cdot x_{2i} + 0.00005484 \cdot x_{4i} - 0.1679 \cdot x_{5i};$$

2014:

$$\ln \hat{y_{2014i}} = 9.677 - 0.00001669 \cdot x_{1i} + 0.00005619 \cdot x_{4i} - 0.229 \cdot x_{5i};$$

2015:

$$\ln \hat{y_{2015i}} = 9.807 - 0.00002218 \cdot x_{1i} + 0.000000003823 \cdot x_{3i} + 0.00003945 \cdot x_{4i} - 0.16768 \cdot x_{5i};$$

2016:

$$\ln \hat{y_{2016i}} = 9.122 - 0.00002709 \cdot x_{1i} + 0.00007275 \cdot x_{2i} + 0.00000000645 \cdot x_{3i}.$$

4.2. Multiple regression mode for group 1. A multiple regression model was constructed for the group 1 using the analysis method in Table 3.

Table 3. **Coefficients estimates of group 1**

Coefficient	Estimate	Std. Error	t-value	$\Pr(>\|t\|)$	Signif.
$\hat{\alpha}$	9.130e+00	4.838e-01	18.873	0.0337	*
$\hat{\beta_1}$	−5.831e-06	4.699e-06	−1.241	0.4318	
$\hat{\beta_2}$	4.330e-05	1.367e-05	3.168	0.1947	
$\hat{\beta_3}$	1.585e-05	5.204e-06	3.045	0.2020	
$\hat{\beta_4}$	−9.283e-06	4.977e-06	−1.865	0.3133	
$\hat{\beta_5}$	2.178e-01	1.000e-01	2.178	0.2741	

Therefore, the regression equation is as follows:

$$\ln \hat{y_{2017i}} = 9.13 - 0.000005831 \cdot x_{1i} + 0.0000433 \cdot x_{2i} + 0.00001585 \cdot x_{3i} - \\ - 0.000009283 \cdot x_{4i} + 0.2178 \cdot x_{5i}.$$

The statistical significance of the coefficients is presented in Table 3. Only the estimates of the free coefficients being statistically significant at the 5 % level.

Assess the overall quality of the model for the first cohort.

This result confirms that the average quality of the whole model is satisfactory:

R^2	0.9824
R^2_{adj}	0.8946
F	11.19
p-value (F)	0.0223

The F-statistic corresponding to the model is statistically significant with a p-value less than 0.05, indicating that the model is acceptable at a 5 % level of significance. The model demonstrates a high degree of explanatory power, as it explains approximately 90 % of the variation in the outcome variables:

lag	Autocorrelation	DW statistic	p-value
1	−0.4205415	2.354857	0.162

χ^2	p-value (Breusch − Pagan)
0.1263405	0.07222

The DW statistic of 2.35 falls within the interval of $1.5 < \text{DW} < 2.5$, indicating that the residuals are not autocorrelated. Furthermore, the Breusch − Pagan test is consistent with a p-value greater than 0.05, suggesting that the null hypothesis of residuals homoscedasticity is rejected. Consequently, the results of the statistical analysis demonstrate that the model quality of the first group is normal.

The stepwise regression algorithm is utilized to identify the factor with the least significant coefficient (maximum p-value) in each step, and this factor is then dropped in order to obtain the final model:

$$\ln \hat{y_{2017i}} = 9.179 + 0.0000371 \cdot x_{2i} - 0.000005542 \cdot x_{4i} + 0.2202 \cdot x_{5i}.$$

Apply the same process to the data for other years:

2009:

$$\ln \hat{y_{2009i}} = 8.518 + 0.00009.835 \cdot x_{2i} - 0.00001247 \cdot x_{4i};$$

2010:

$$\ln \hat{y_{2010i}} = 6.261 - 0.0001188 \cdot x_{1i} - 0.00006069 \cdot x_{2i} + 0.000000007946 \cdot x_{3i};$$

2011:

$$\ln \hat{y_{2011i}} = 8.072 + 0.00006173 \cdot x_{2i} + 0.1711 \cdot x_{5i};$$

2012:

$$\ln \hat{y_{2012i}} = 8.195 + 0.00004987 \cdot x_{2i} + 0.000000001065 \cdot x_{3i} + 0.1992 \cdot x_{5i};$$

2013:

$$\ln \hat{y_{2013i}} = 8.272 + 0.00004462 \cdot x_{2i} + 0.000000001145 \cdot x_{3i} + 0.231 \cdot x_{5i};$$

2014:

$$\ln \hat{y_{2014i}} = 9.788 - 0.00002668 \cdot x_{1i} + 0.00002101 \cdot x_{2i} + 0.000000002385 \cdot x_{3i};$$

2015:

$$\ln \hat{y_{2015i}} = 10 - 0.00002476 \cdot x_{1i} + 0.00002912 \cdot x_{2i} + 0.000000002255 \cdot x_{3i};$$

2016:

$$\ln \hat{y_{2016i}} = 10.19 - 0.00002719 \cdot x_{1i} + 0.00002742 \cdot x_{2i} + 0.000000002265 \cdot x_{3i}.$$

4.3. Cluster-specific examinations. In order to measure the influence of each factor on investment attractiveness, multiple regression analysis was chosen as the primary method of this study. In order to eliminate the effects of hierarchical structure, the observations were divided into four distinct groups and each group was analyzed separately [8]. As it was not feasible to use multiple regression analysis for small clusters, the multiple regression analyis was replaced by a correlation analysis in order to assess the degree of influence of the factors [9].

Correlation analysis was employed to conduct a statistical examination of the second and third groups with a small number of clusters.

For the second group, a corrplot correlation heat map was generated (Figure 4).

As depicted in Figure 4, the correlation between variable x_5 and the outcome variable is notably weak ($r = 0.08$), making it logical to exclude x_5 from the factor set. Eliminating x_5 from the explanatory factor set changes the correlation heat map to the following form (Figure 5).

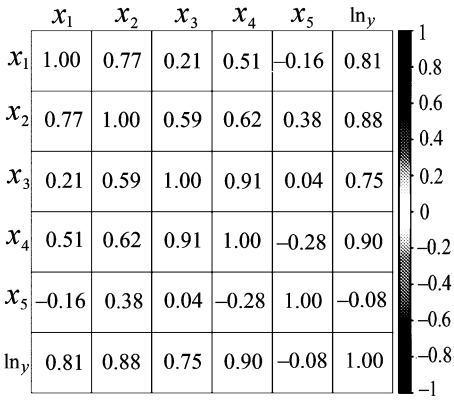

Figure 4. Thermal correlation map for group 2

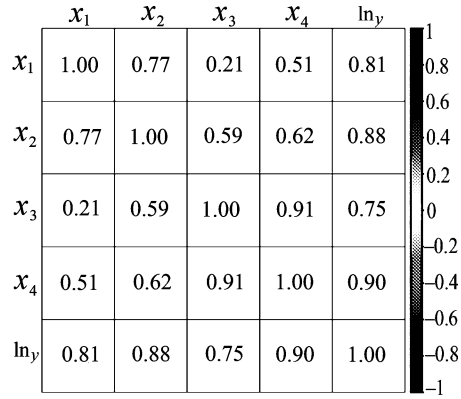

Figure 5. Map after remove x_5

As evidenced by Figure 5, x_4 is strongly correlated with $x_3 (r > 0.7)$, and moderately correlated with x_1 and x_2 $(0.4 < r < 0.7)$. Consequently, it is sensible to exclude x_4 from the factor set, even though it has a stronger correlation with the outcome variable (Figure 6).

In the subsequent step, x_2 is hypothesized to be removed due to its moderate correlation with both x_1 and x_3, while the correlation between x_1 and x_3 is notably low. The resulting outcomes are then obtained (Figure 7).

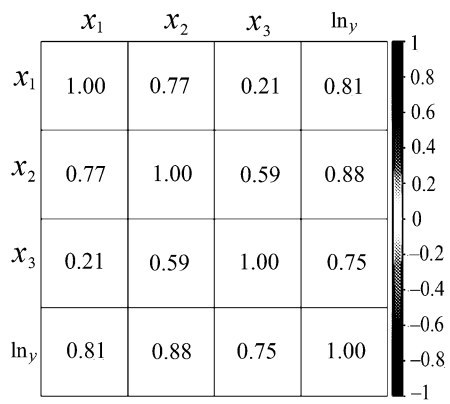

Figure 6. Map after remove x_5, x_4

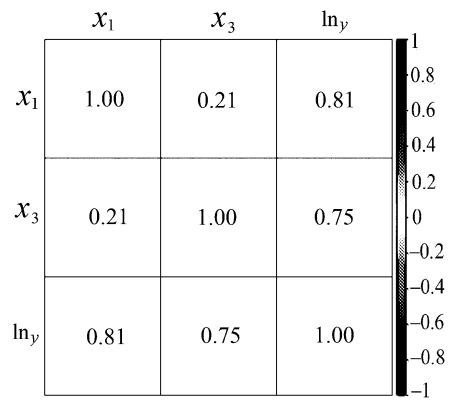

Figure 7. Map after remove x_5, x_4, x_2

As the covariates between the explanatory variables are successively eliminated, the most strongly correlated factors are eliminated to obtain a set of independent variables that includes only x_1 and x_3. In this instance, x_3 is less strongly associated with the outcome variable y than x_1. Thus, it can be determined that in the second cluster, the primary factor influencing the formation of y values is the factor x_1.

In the third cluster, the insufficient amount of data does not allow for a detailed presentation of the correlation heat map, and thus does not provide enough information to draw any conclusions about the statistical relationship between the studied indicators and investment attractiveness. The limited number of observations in this cluster likely indicates a deviation from the overall pattern.

5. Conclusion. In order to evaluate the influence of various factors on investment attractiveness, multiple regression analysis was chosen as the primary tool for the study. The observations were clustered into four groups and a separate analysis was conducted for each cluster. Due to the heterogeneity of cluster sizes, this phase of the research was conducted separately for both large and small clusters. For the smaller clusters, where the amount of data was insufficient for multiple regression analysis, correlation analysis was employed to assess the degree of influence of the factors [10]. In the first group (large cluster), which are the most attractive areas for investment, a strong correlation was observed between the volume of investment and the cost of fixed assets (x_2), and the amount of work performed by the type of activity "Construction" (x_3), with the strongest correlation for the cost of fixed assets (x_2). The fourth group (large cluster), which is characterized by low investment attractiveness, exhibited a strong correlation between investment volume and average per capita money income(x_1), and the amount of work performed by the type of activity "Construction" (x_3), with average per capita money income(x_1) being identified as the primary factor influencing the investment volume. The second group (small cluster) is characterized by the lowest investment attractiveness, with the investment volume being closely correlated with average per capita money income (x_1) and the amount of work performed by the type of activity "Construction" (x_3). In particular, average per capita money income (x_1) was identified as the primary factor influencing this cluster. Therefore, it is essential to analyze the investment attractiveness of these regions separately for each region, taking into account the respective economic characteristics, and develop relevant policies accordingly [11].

References

1. Wang Qian. Environmental regulation and foreign direct investment attractiveness: Evidence from China Provinces. *Review of Development Economics*, 2022, vol. 26, no. 2. https://doi.org/10.1111/rode.12871

2. Qi D., Bure V. M. Statistical analysis of investment attractiveness of China's regions. *Vestnik of Saint Petersburg University. Applied Mathematics. Computer Science. Control Processes*, 2022, vol. 18, iss. 1, pp. 189–195. https://doi.org/10.21638/11701/spbu10.2022.116

3. Granato D. Use of principal component analysis (PCA) and hierarchical cluster analysis (HCA) for multivariate association between bioactive compounds and functional properties in foods: A critical perspective. *Trends in Food Science and Technology*, 2018, vol. 72, no. 72, pp. 83–90. https://doi.org/10.1016/j.tifs.2017.12.006

4. Bure V. M., Parilina E. M., Sedakov A. A. *Applied statistics methods in R and Excel*. 3 ed. St. Petersburg, Lan' Publ., 2019, 196 p.

5. *The World Bank*. Available at: http://data.worldbank.org/indicator (accessed: January 18, 2023).

6. *National Bureau of Statistics*. Available at: http://www.stats.gov.cn/tjsj/ndsj/ (accessed: February 8, 2023). (In Chinese)

7. Govender P., Sivakumar V. Application of K-means and hierarchical clustering techniques for analysis of air pollution: A review (1980–2019). *Atmospheric Pollution Research*, 2020, vol. 11, no. 1, pp. 40–56. https://doi.org/10.1016/j.apr.2019.09.009

8. Olilingo F. Z., Aditya H. P., Kusuma P. How Indonesia economics works: correlation analysis of macroeconomics in 2010–2019. *The Journal of Asian Finance, Economics and Business*, 2020, vol. 7, no. 8, pp. 117–130.

9. Senthilnathan S. Usefulness of correlation analysis. *SSRN Electronic Journal*, 2019. https://doi.org/10.2139/ssrn.3416918

10. Iakushev V. P., Bure V. M., Mitrofanova O. A., Mitrofanov E. P. Theoretical foundations of probabilistic and statistical forecasting of agrometeorological risks. *Vestnik of Saint Petersburg University. Applied Mathematics. Computer Science. Control Processes*, 2021, vol. 17, iss. 2, pp. 174–182. https://doi.org/10.21638/11701/spbu10.2021.207

11. Iakushev V. P., Bure V. M., Mitrofanova O. A., Mitrofanov E. P. K voprosu avtomatizatsii postroeniia variogramm v zadachakh tochnogo zemledeliia [On the issue of semivariograms constructing

automation for precision agriculture problems]. *Vestnik of Saint Petersburg University. Applied Mathematics. Computer Science. Control Processes*, 2020, vol. 16, iss. 2, pp. 177–185. https://doi.org/10.21638/11701/spbu10.2020.209 (In Russian)

Received: February 22, 2023.
Accepted: April 25, 2023.

Authors' information:

Dongfang Qi — Postgraduate Student; st073409@student.spbu.ru

Vladimir M. Bure — Dr. Sci. in Technics, Professor; vlb310154@gmail.com

Исследование инвестиционной привлекательности на основе кластерного анализа

Д. Ци, В. М. Буре

Санкт-Петербургский государственный университет, Российская Федерация,
199034, Санкт-Петербург, Университетская наб., 7–9

Для цитирования: *Qi D., Bure V. M.* Research of investment attractiveness based on cluster analysis // Вестник Санкт-Петербургского университета. Прикладная математика. Информатика. Процессы управления. 2023. Т. 19. Вып. 2. С. 199–211. https://doi.org/10.21638/11701/spbu10.2023.206

Продолжающееся экономическое развитие различных стран или регионов привело к усилению конкуренции на мировых рынках, что обусловило концентрацию инвесторов и квалифицированной рабочей силы в местах с высокой инвестиционной привлекательностью. Инвестиционная привлекательность той или иной страны или региона определяется ее инвестиционным потенциалом и риском, которые характеризуются сочетанием разных значимых факторов. Предпринята попытка разработать эконометрическую модель для оценки объема инвестиций в основной капитал в конкретном регионе с учетом линейной зависимости между наблюдаемыми результатами, чтобы определить основные условия, необходимые для достижения стабильного и высокого экономического роста. К таким условиям относятся ускорение инвестиционной активности и проведение крупных национальных реформ для обеспечения эффективности инвестиционного процесса. Для оценки общего влияния изучаемых финансово-экономических показателей на объем инвестиций в качестве основного математического инструмента исследования был использован множественный регрессионный анализ. Кроме того, были сделаны предположения относительно ранга наблюдений. Для подтверждения этой гипотезы был проведен кластерный анализ, сгруппировавший наблюдения в четыре кластера на основе их результатов в зависимости от объема инвестиций или географических характеристик региона.

Ключевые слова: инвестиционная привлекательность, кластерный анализ, иерархическая регрессионная модель, модели множественной линейной регрессии, корреляционный анализ, метод наименьших квадратов.

Контактная информация:

Ци Дунфан — аспирант; st073409@student.spbu.ru

Буре Владимир Мансурович — д-р техн. наук, проф.; vlb310154@gmail.com