Methodology of optimal classification of regions by the level of industry development

Obidjon Khamidov 1, Tolqin Farmanov 2, Sokhibmalik Khomidov 1, and Asliddin Abdulloyev 1

1 Bukhara State University, M. Iqbol, 11, 200100 Bukhara City, Uzbekistan
2 Tashkent State Agrarian University, Universitet Ko‘chasi, 100700 Tashkent, Uzbekistan

Abstract. The article describes the methodology of classification of multidimensional observations, according to which, in theory, development indices of regional industrial production were developed. According to this index, conditional regions are divided into lower, middle and upper classes. In the study, the development tendencies of the regions classified by the development index into stratified groups were also determined. At the end of the article, the theoretical and empirical conclusions obtained as a result of the research are presented.

1 Introduction

Classification of multidimensional observations is a complex process that directly requires a lot of work and complex calculations. In turn, the issue of classification is carried out using methods of cluster analysis. The main purpose of clustering is to form similar groups between objects and analyze the relationships between them. The importance of clustering is determined by the nature and economic relevance of the practical problems it can solve. In our study, the issue of classification of multidimensional observations was highlighted, according to which conditional regions were divided into lower, middle and upper classes according to the “level of development of industrial production”. It is an important issue to sort these regions according to a certain level of development and to determine the laws in this regard, in which each region is characterized by its own set of indicators. Therefore, they cannot be sorted at the same level. This raises new scientific problems for the classification of objects, and the research focuses on the theoretical aspects of this problem.

The existence of a scientific need for the proposed methodology is that it considers not only the issue of classification of objects (regions), but also the issue of determining the development tendencies of objects into classified groups. In the study, the parallel clustering method was used in the classification of objects. The essence of parallel operations is that all indicators are compared and calculated at the same time at each step of the algorithm. Because it is more difficult to make a selection even when dividing into small number of classes. Therefore, the main essence of the development of various parallel classification algorithms is to determine a method that can provide a quick solution to the set goals and reduce the selection options.
2 The main part

The main part of this document is divided into several sections discussing the development of cluster analysis, its applications in various fields, and the theoretical aspects of categorizing regions based on compositional changes in industrial production. The document also introduces a new approach to classifying regions by the level of development of industrial production, utilizing specific indicators.

The document starts by mentioning the beginning of cluster analysis publications in the late 30s of the last century, but it became actively developed in the late 60s and early 70s. The development of information technologies led to the emergence of new methods, modifications, and algorithms of clustering, providing opportunities to process large volumes of data.

One unique aspect of cluster analysis is that it is a general research method for almost all fields of study. Today, this method is widely used in medicine to classify organisms and diagnose hereditary diseases, in chemistry to classify materials with similar properties, and in ecology to classify changes and events in nature.

According to theoretical analyses, hierarchical, parallel, and non-hierarchical clustering operations have been widely studied in most scientific sources. In scientific literature related to the theoretical aspects of our research, cluster methods and their specific features, interpretation of cluster analysis results, cluster analysis problems, cluster methods and algorithms, especially cluster analysis problems under unique conditions, problems of selecting variables and measuring distances between objects as well as evaluating cluster quality, algorithms that are useful for solving problems related to optimal grouping of objects and elimination of such problems, hierarchical and non-hierarchical cluster methods, optimization of the number of clusters, and contemporary and classical approaches to cluster analysis, including the formulation and algorithms of cluster analysis problems under unique conditions have been extensively investigated.

According to the scientific literature related to our practical research, the theoretical aspects of categorizing regions based on compositional changes in industrial production have been studied. In contrast to the aforementioned studies, our research has two unique aspects. Firstly, our research not only addresses the problem of classifying individual objects, but also identifies the potential for the development of classified groups. Secondly, in our classification problem, we have utilized algorithms that provide the possibility of "object-to-class transfer".

We select 6 important economic indicators to categorize regions by the level of development of industrial production, and then we create specific indicators from them. These indicators not only characterize the level of development of industrial production in regions, but also directly indicate the economic development of countries in this area.

To classify regions by the level of development of industrial production, we establish the following criteria:

- \( t_i q_{igdp} \) - Current added value of industrial production in the region during the period \( t \);
- \( t_i q_{igdp} \) - Gross regional product in the region during the period \( t \);
- \( t_i q_{iqe} \) - Export of industrial products in the region during the period \( t \);
- \( t_i q_{ite} \) - Total export volume in the region during the period \( t \);
- \( t_i q_{wq} \) - Added value of the country's industrial production during the period \( t \);
- \( t_i q_{wqe} \) - Volume of exports of industrial products in the country during the period \( t \).

Based on the given initial indicators, we create the following specific indicators:

- \( t_i q_i z \) = \( \frac{q_i}{gdp} \) - the share of added value of industrial production in gross regional product in the region during the period \( t \);
\[ z_i^t = q_{e_i}^t \]
\[ z_i^t = q_{w_i}^t \]
\[ z_i^t = q_{w_i}^t \]
\[ t = \ldots \]
\[ \mathbf{i} = \ldots \]

\[ Z_i^t = (Z_i^t)_{\rho=\ldots} \]

\[ Z_i^t = \begin{pmatrix} z_i^t & z_i^t & z_i^t & z_i^t \\ z_i^t & z_i^t & z_i^t & z_i^t \\ z_i^t & z_i^t & z_i^t & z_i^t \\ z_i^t & z_i^t & z_i^t & z_i^t \end{pmatrix} \]

\[ S = \{ S^n, S^m, S^h \} \]

\[ S_i \cap S_j = \emptyset \] for \( i \neq j \)

\[ \bigcup_{i=1}^{k} S_i = \mathbf{i} \]

\[ k \]

\[ \mathbf{i} \]
have an aggregate character and be the sum of vector private indicators for all \( n_i, \ldots, n_2, n_1 \).

All private indicators are close or similar to each other in terms of their economic essence and construction. As the possible values for all \( n_i \) and \( \rho \) vary in a single interval \( 1 < \rho < 0 \), then separate measurements of specific indicators can be used to construct a generalized indicator.

So, when dividing into initial classes, we accept the generalized indicator (\( t_i \omega \)) representing the development of industrial production of country \( i \) for period \( t \), and we consider the procedure and options for its calculation.

Procedures and options for calculating the generalized indicator representing the development of regional industrial production:

1. All private indicators are of equal value, that is, their shares are the same. In this case, the generalized indicator representing the development of the industrial production of region \( i \) in period \( t \) will be equal to:

\[
\omega_i' = \sum_{j=1}^{n_i} z'_{i,j} \quad i = 1, 2, 3, \ldots, n
\]

2. For all private indicators, the percentage of importance \( \lambda_i \) has different values, then the generalized indicator will be equal to:

\[
\omega_i' = \sum_{j=1}^{n_i} \lambda_i z'_{i,j} \quad i = 1, 2, 3, \ldots, n
\]

3. If the percentage of importance of private indicators characterizing the development of industrial production of the region is \( \lambda_i \) and the percentage of importance of private indicators characterizing the place of this region in the country in this sector is \( \lambda_i - 1 \), then the cumulative indicator is calculated as follows:

\[
\omega_i' = \lambda_i \left( z'_{i,1} + z'_{i,2} \right) + \left( \lambda_i - \lambda_i \right) \left( z'_{i,3} + z'_{i,4} \right) \quad i = 1, 2, 3, \ldots, n
\]

here \( \lambda_i \) is selected using the expert evaluation method.

There are many ways to divide a given set of regions \( \{O_i \} \) into classes, and we introduce a quantity \( SQ \) as a criterion for choosing the best one. The quantity \( SQ \), regions \( i \) and \( j \) are determined by finding the distance \( d(Z'_i : Z'_j) \) between the vector indicators \( Z'_i \) and \( Z'_j \) characterizing the development of industrial production and regions \( O_i \) and \( O_j \) are included in the difference \( O_i - d(O_i : O_j) \).

There are various measures that characterize the distance between features, including the Euclidean distance, the Mahalonobis metric, the Hamming distance, and the Canberra metric.

Since the indicator of the vectors in the process under consideration is the same according to its economic content and calculation method, it is appropriate to use the simple Euclidean distance when measuring the specified distance:

\[
O = O, i = 1, 2, 3, \ldots, n
\]

\[
Q : S \quad i \quad j
\]

\[
d(Z'_i : Z'_j)
\]

\[
O_i \in O - d(O_i : O_j)
\]
\[ d_e(Z_i^t \cap Z_j^t) = \sqrt{\sum_{\rho=1}^{n} (z_{i,\rho}^t - z_{j,\rho}^t)^2} \]

\[ d_{be}(Z_i^t \cap Z_j^t) = \sqrt{\sum_{\rho=1}^{n} \lambda_{\rho} (z_{i,\rho}^t - z_{j,\rho}^t)^2} \]

\[ \lambda_{\rho} \leq \lambda_{\rho} \leq \rho = \ldots \]

We use an algorithm that sequentially performs “transfer of objects from class to class”

First, we rank the generalized indicators:

\[ S^u = \{ S^u, S^u, S^u \} \]

\[ \{ \omega^i, i = \ldots \} \]

\[ \omega^i_{1,1} \leq \omega^i_{1,1} \leq \omega^i_{1,1} \leq \ldots \omega^i_{1,1} \]

\[ \begin{bmatrix} \omega^i_{1,1} \omega^i_{1,1} + \omega^i_{1,1} - \omega^i_{1,1} \\ \omega^i_{1,1} + \omega^i_{1,1} - \omega^i_{1,1} \omega^i_{1,1} + \omega^i_{1,1} - \omega^i_{1,1} \omega^i_{1,1} + \omega^i_{1,1} - \omega^i_{1,1} \end{bmatrix} \]

\[ i = i \ldots \omega^i_{1,1} \]

\[ \omega^i_{1,1} \in \begin{bmatrix} \omega^i_{1,1} \omega^i_{1,1} + \omega^i_{1,1} \\ \omega^i_{1,1} + \omega^i_{1,1} \omega^i_{1,1} + \omega^i_{1,1} \end{bmatrix} \]
\[ S^i_q = \{ i, \tilde{i}, \tilde{i}_q \} \]

\[ i_q = \{ j, \tilde{j}, \tilde{j}_q \} \]

\[ \omega'_i \in \begin{cases} \omega'_i + \frac{\omega'_i - \omega'_{i-1}}{\omega'_i - \omega'_{i-1}} \omega'_i + \frac{\omega'_i - \omega'_{i+1}}{\omega'_i - \omega'_{i+1}} \omega'_i + \frac{\omega'_i - \omega'_{i-1}}{\omega'_i - \omega'_{i-1}} \omega'_i + \frac{\omega'_i - \omega'_{i+1}}{\omega'_i - \omega'_{i+1}} \omega'_i \end{cases} \]

\[ \omega'_q \not\in \begin{cases} \omega'_q + \frac{\omega'_q - \omega'_{q-1}}{\omega'_q - \omega'_{q-1}} \omega'_q + \frac{\omega'_q - \omega'_{q+1}}{\omega'_q - \omega'_{q+1}} \omega'_q + \frac{\omega'_q - \omega'_{q-1}}{\omega'_q - \omega'_{q-1}} \omega'_q + \frac{\omega'_q - \omega'_{q+1}}{\omega'_q - \omega'_{q+1}} \omega'_q \end{cases} \]

\[ S^i = \{ i, \tilde{i}, \tilde{i}_q \} \]

\[ S^i = \{ i_q, \tilde{i}_q \} \]

\[ \omega'_i \leq \omega'_q \]

\[ S^i_q = \{ i, \tilde{i}, \tilde{i}_q \} \]

\[ \omega'_i \leq \omega'_q \]

\[ \bar{\omega}' \leq \omega'_q \]

\[ \omega'_i = \omega'_i + k \left( \frac{\omega'_i - \omega'_{i-1}}{\omega'_i - \omega'_{i-1}} + \frac{\omega'_i - \omega'_{i+1}}{\omega'_i - \omega'_{i+1}} \right) \]

\[ S = \{ i_p, \tilde{i}_p \} \]

\[ S = \{ i_q, \tilde{i}_q \} \]

\[ f(\omega') \]

\[ [\omega'_i, \omega'_q] \]

\[ [\omega'_i, \omega'_q] \]

\[ [\omega'_i, \omega'_q] \]

\[ [\omega'_i, \omega'_q] \]

\[ [\omega'_i, \omega'_q] \]
The indicators represent the numbers of regions located in the interval \( t \in [0, \omega_t] \), \( m \omega \), and \( 0 \), respectively. The class represents the numbers of regions located in the interval \( t \in [0, \omega_t] \), \( m \omega \), and \( 0 \).

After selecting the initial separation option \( S_0 \), the value of the separation quality criterion \( Q(S) \) is determined. For a given number of classes, the quality of separation is the sum of the within-class variances:

\[
Q(S) = \sum_{k=0}^{K} \sum_{i \in S_k} d_i \left( \bar{Z}^i_k \right) \]

where \( d_i \left( \bar{Z}^i_k \right) \) represents the distance squared in Euclidean metric (or weighted Euclidean metric), \( S_k \) divides into \( k \) classes, the number of classes is fixed and equal to 3, \( k = 3, 2, 1 \).

\[
\bar{Z}^i_k = \left( \bar{z}^i_k \bar{z}^i_k \bar{z}^i_k \right)
\]

\[
z^i_{k,\rho} = \sum_{i \in S_k} z^i_{k,\rho} \quad (\rho = 0, \ldots, k = 3, 2, 1)
\]

Fig. 1. An approximate theoretical view of the empirical distribution of the cumulative indicator of industrial production development.

\[
\sum_{i \in S_k} z^i_{k,\rho} = \sum_{i \in S_k} z^i_{k,\rho}
\]

\[
Q(S) = \sum_{k=0}^{K} \sum_{i \in S_k} d_i \left( \bar{Z}^i_k \right)
\]
The following condition can be obtained as a condition for the completion of the algorithm: if

\[ m \cdot S \cdot Q(m) \] and

\[ 1 + m \cdot S \cdot Q(1 + m) \]

are the values of the functions in successive steps, the classes obtained in these steps are

\[ \{m, m, m, m, m\} \] and

\[ \{1, 1, 1, 1, 1\} \].

If the following condition is fulfilled for small \( \epsilon > 0 \) (\( \epsilon \) is a calculation accuracy), then the classification process stops:

\[ \epsilon < 1 - m \cdot S \cdot Q(1 - m) \]

As a result of the performed classification, we get the separation

\[ 1 + m \cdot S \cdot Q(1) \] and
denote it by \( t \).

It should be noted that the sequence

\[ \{m, m, m, m, m\} \]

is monotonically decreasing:

\[ \{1, 1, 1, 1, 1\} \]

The constructed algorithm is applied several times to exactly the same set,

\[ \{n, i, Z, t, n, i, 2, 1, 1\} \]

and after different initial allocations, the best variant of

\[ \{m, m, m, m, m\} \]

is finally formed.

According to the abovementioned algorithm, a set of

\[ \{n, i, O, O, i, 2, 1, 1\} \]

regions for each period \( t \) can be divided into three classes according to the level of development of industrial production. If we have the level of development of industrial production in a cross-section of regions and the statistics of monitoring the economy of regions of length \( N \) (\( N \) - the length of the retrospective) and

\[ 0, 0, \ldots, 0, t, N, t, N, t, \ldots, 0, \ldots, 2, t, N, t, \ldots, 0 \]

is the base year), then the proposed algorithm allows us to build a family of classifications into

\[ \{n, i, O, O, i, 2, 1, 1\} \]

classes for each \( t \) from the retrospective period

\[ \{0, 0, \ldots, 0, t, N, t, N, t, \ldots, 0, \ldots, 2, t, N, t, \ldots, 0\} \].

It is possible to form the following using the theoretical operations of the calculation of sets:

\[ t \cdot k \cdot t \cdot N \cdot t \cdot k \cdot S \cdot 0 \cdot 0 = \bigcap_{t=0}^{t=0} \bigcup_{k=0}^{k=0} S \cdot k \cdot \bigcup_{t=0}^{t=0} S \cdot k \cdot \bigcup_{k=0}^{k=0} \bigcup_{t=0}^{t=0} S \cdot k \cdot \bigcup_{t=0}^{t=0} S \cdot k \]

Here,

- \( 1 \cdot S \cdot k \) - a set of regions characterized by stable low rates of industrial production development and maintaining such dynamics throughout the retrospective period.

- \( 2 \cdot S \cdot k \) - a set of regions characterized by medium and high development of industrial production, respectively, and maintaining such dynamics during the analyzed period. In that case, \( k \cdot S \cdot k \) can be called stable zones of different development of industrial production in the region.

It should be noted that the development of industrial production in a set of

\[ \{i, i, S, ~ S, ~ S, i, 2, 1, 1\} \]

regions has changing and unstable dynamics. Therefore, the set

\[ \{S, ~ S, ~ S, i, 2, 1, 1\} \]

can be called an area of unstable development of industrial production in the region. These regions "move" from class to class during the retrospective period. The displacement of each

\[ i \cdot k \cdot i \cdot S \cdot k \]

region can be determined from the condition

\[ t \cdot k \cdot t \cdot i \cdot k \cdot S \cdot k \cdot \bigcup_{i=0}^{i=0} S \cdot k \cdot \bigcup_{i=0}^{i=0} S \cdot k \cdot \bigcup_{i=0}^{i=0} S \cdot k \cdot \bigcup_{i=0}^{i=0} S \cdot k \]
\[ K'_i = \{ k(i') \mid t \in [t_i - N + t, t_i + t] \} \]

\[ K'_i = \{ k(i) = \text{const} = k \} \]

\[ N_i(i) = \sum k'_i(l) \mid l = \sum k_i(n) \]

\[ T_i(i) = \frac{N_i(i)}{N} \mid l = \sum n_i \]

\[ \sum_{i=1}^{n} T_i(i) = \frac{1}{N} \sum_{i=1}^{n} N_i(i) \]

\[ G \subset S \]

\[ G_i = G \cap S_i \]

\[ G = G \cap S \]

\[ G_1 = G \cup \bigcup_{k=1}^{n} G_k = G \cup (G \cap S_k) \]

\[ i \in G \]
The dynamics and character of migrations of regions can be analyzed according to relations (22) - (25).

In order to test the proposed methodology, conditional economic indicators for conditional regions presented in Appendixes 1 - 4 were used. The obtained empirical results are recorded in Table 1. According to the analysis, the first group of low-developed industrial production in the period T1 - T4 (uS 1) included the conditional region No.13 (the average value of the development index of region No.13 in T1 - T4 is 8.63), in T1 this group has values of 13.41 and in T3 with values of 7.25 included in the region No.1.

The second group (uS 2) with moderately developed industrial production in T1 - T4 included the region No.8 and No.3, and the region No.12 in T1 and T2. In particular, in the region No.12, this index was 23.17 in T1, and 21.26 in T2.

Table 1. Indexes of development of industrial production in the conditional regions (in percentages)

<table>
<thead>
<tr>
<th>Conditional regions</th>
<th>Periods</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(T1)</td>
</tr>
<tr>
<td>No1</td>
<td>13.41</td>
</tr>
<tr>
<td>No2</td>
<td>26.83</td>
</tr>
<tr>
<td>No3</td>
<td>22.27</td>
</tr>
<tr>
<td>No4</td>
<td>29.52</td>
</tr>
<tr>
<td>No5</td>
<td>28.91</td>
</tr>
<tr>
<td>No6</td>
<td>27.16</td>
</tr>
<tr>
<td>No7</td>
<td>32.68</td>
</tr>
<tr>
<td>No8</td>
<td>21.13</td>
</tr>
<tr>
<td>No9</td>
<td>27.32</td>
</tr>
<tr>
<td>No10</td>
<td>28.38</td>
</tr>
<tr>
<td>No11</td>
<td>29.37</td>
</tr>
<tr>
<td>No12</td>
<td>23.17</td>
</tr>
<tr>
<td>No13</td>
<td>8.54</td>
</tr>
<tr>
<td>No14</td>
<td>26.50</td>
</tr>
<tr>
<td>No15</td>
<td>25.07</td>
</tr>
<tr>
<td>No16</td>
<td>28.95</td>
</tr>
<tr>
<td>No17</td>
<td>28.67</td>
</tr>
<tr>
<td>No18</td>
<td>26.01</td>
</tr>
<tr>
<td>No19</td>
<td>27.50</td>
</tr>
<tr>
<td>No20</td>
<td>28.54</td>
</tr>
</tbody>
</table>

The analysis shows that most conditional regions belong to the third group (uS 3) with highly developed industrial production. According to the level of development of industrial production, the tendency of conditional regions to develop into stratified groups is illustrated in the Table 2 below.

Table 2. Development propensities of conditional regions into stratified groups according to the level of development of industrial production (in coefficients)
According to the analysis, during the years T1 - T4, the industrial production of region №13 tends to develop in the first group with a 100 percent probability, and the industrial production of the region №1 with a 50 percent probability.

Indicators of development of industrial production of the regions №3 and №8 with 100 percent probability and industry of the region №12 with 50 percent probability tend to develop in the second group. Industrial production development indicators of all other conditional regions tend to develop in the third group with 100 percent probability.

### 3 Conclusion

1. Empirical results obtained in the course of the research provide opportunities to develop a specific strategy for the long-term economic development of industrial production in regions with different development trends, to identify regions with active and slow development of the processing industry, and to clarify its reasons.

2. On the basis of the research, development indexes (RI) of industrial production of conditional regions were developed, according to which this index varies within the range of $0 \leq RI \leq 100$. Which means that the index value approaching 100 indicates the stable development of the industrial production of the region.

3. According to the obtained empirical results, the industrial production of the region №13 has a 100 percent probability, and the industry of the region №1 has a lower level with a 50 percent probability, the industries of the regions №3 and №8 have a 100 percent probability, and the industry of region №13 has a 50 percent probability. Industrial production of all other conditional regions tend to develop at a medium level with 100 percent probability.
References

1. L. A. Soshnikova, V. N. Tamashevich, G. Uebe, M. Shefer, Multidimensional statistical analysis in economics, Moscow, 469 (1999)