# CLUSTER ANALYSIS TO IDENTIFY POTENTIAL EMPLOYERS OF UNIVERSITY GRADUATES

A.G. Karamzina
Department of Computer Science and Robotics, Ufa State Aviation Technical University, Ufa, Russia
e-mail: karamzina@tc.ugatu.ac.ru

***Abstract:*** *The results of cluster analysis are presented, the purpose of which was the verification of the base of industrial enterprises of the Republic of Bashkortostan to identify potential employers of graduates by 27.03.00 on the enlarged group of specialties and directions "Control in technical systems". The solution of the problem will allow to interact with potential graduates' consumers at the stage of developing higher education educational programs, which will ensure high-quality training of highly qualified personnel in accordance with the requests of enterprises and organizations of the region and, accordingly, further employment of graduates. The work analyzes the conditions for the formation of clusters, presents a sample of data for intellectual analysis, results of component and cluster analysis.*

## 1. Introduction

As part of the transformation of Russian education into the Bologna system and, first of all, the adoption of federal standards, one of the most important directions of the activity of higher education institution was the formation of effective models for interaction between the university and a potential employer.

Interaction of the university with a potential employer should be conducted already at the stage of developing the basic professional educational program. According to the methodological recommendations for the development and implementation of higher education educational programs at the bachelor's level [7], when developing and implementing educational programs of applied bachelor's degree, it is recommended to use various models of interaction with potential employers, including cluster models.

The use of cluster models has already taken a key position in research [1]. The relevance of cluster models is due to the fact that the cluster approach is considered the most effective tool in the management system.

To implement high-quality interaction with employers in the process of implementing educational programs of higher education, it is necessary to solve the problem of identifying potential employers. To solve this problem, the author proposes to apply the method of cluster analysis. Clustering will allow you to break up multiple objects (potential employers of the region) into groups - clusters. Within each cluster will be those enterprises and organizations whose interaction with them will have the most positive effect in the development and implementation of undergraduate programs, depending on the needs of the region in graduates in three areas of training, within the enlarged group of specialties and directions “Management in technical systems”:

- Quality control (QC);
- System analysis and management (SAM);
- Management in technical systems (MTS).

Analysis of the conditions for the formation of clusters is performed by the method of principal components and by cluster analysis.

## 2. Component analysis

Method of main component (MC) refers to a group of data visualization methods that allow you to identify data structures based on clarifying the relationships between objects and their characteristics. The result of the application of the method is the graphical mapping of a set of objects to a new coordinate space, and the mapping maximally reflects the features of the distribution of objects in the original multidimensional space [2, 4]. The application of linear methods of diminishing the dimension allows you to visualize multidimensional data by building a new coordinate space in which each coordinate axis is a linear combination of the original characteristics that correspond to the hidden characteristics present in the data. In addition to solving the visualization problem, linear methods of diminishing the dimensionality allow solving the problem of classification of initial features - to determine the interrelations between the characteristics by combining significant characteristics into groups and forming new integral features.

Method of main component allows, based on the analysis of $n$ objects, each of which is characterized by $p$ initial characteristics $x_1,\ldots, x_p$, to construct the $p$ major components $F_1, F_2, \ldots, F_p$ (the coordinate axes of the new space), the new coordinate system being a system of orthonormal linear combinations . Each $i$-th coordinate axis is a linear combination of the original features and is written as:

$$F_1 = \omega_{11}(x_1 - m_1) + \omega_{21}(x_2 - m_2) + \ldots + \omega_{p1}(x_p - m_p),$$

where $\omega_{ii}$ – the weight coefficient determining the contribution of the $i$-th features to the formation of the $j$-th component, $m_i$ – the mathematical expectation of the $i$-th features.

Advantages of the method are the least distortion of the structure of the original objects when they are projected into a space of smaller dimensions, and also the possibility of using data for analysis in combination with other methods of data exploration. The disadvantage is the possibility of a situation when the weight coefficients have values close to each other, which leads to a weak interpretation of the result obtained. This problem is solved by applying other types of analysis.

As objects $n$, regional industrial enterprises (potential employers) are considered in the amount of 58 pcs.

The following data are considered as feature of $p$:

- $x_1$ – demand for graduates;
- $x_2$ – professional growth (further training, training at the enterprise);
- $x_3$ – production sites (areas not occupied by main production, used directly for training students);
- $x_4$ – equipment (production equipment for training and honing skills in practice for students);
- $x_5$ – software;
- $x_6$ – technologies;
- $x_7$ – human resources (availability of competent managers of practices).

The first main component of $F_1$ ($X$) corresponds to the largest eigenvalue and is calculated as a linear combination of the original features, which has the largest variance. Thus, the first principal component is taken along the direction with the maximum variance. The second main component of $F_2$ ($X$) lies in a subspace perpendicular to where the first main component is located. Within this subspace, the second major component is taken along the direction with the maximum variance. Then the third main

component $F_3$ ($X$) is in the direction of the greatest dispersion in the subspace perpendicular to the first two, and so on.

The component analysis is carried out by means of the package STATGRAPHICS, which has the following advantages [7]:

- a combination of scientific methods for processing heterogeneous data with the possibility of creating modern high-quality interactive graphics;
- wide possibilities of interaction with other software products (spreadsheets, databases);
- high-quality two-dimensional and three-dimensional graphics,
- integrated graphics (all elements of graphical representations of analysis results can be converted).

The initial data for the intellectual analysis of data on seven feature ($x_1$-$x_7$) in three areas of preparation (QC, SAM, MTS), a STATGRAPHICS spreadsheet was introduced. To determine the value of each criterion for industrial enterprises, the deputy heads of the department for academic work in the relevant area of training for bachelors: 27.03.02 "Quality control" (QC), 27.03.03 "System analysis and management" (SAM), 27.03.04 "Management in technical systems" (MTS), chairman and members of the scientific and methodological council on the enlarged group of specialties and directions 27.00.00 "Management in technical systems" (50% of the size of which are representatives of employers).

The obtained data indicate that already two main components describe about 92% of the variance of the initial data. The third main component adds about 3% of the dispersion. As a result, a total of 95% of the variance.

The dispersion diagram of the entire set of objects on the plane of the two main components is shown in Fig. 1-3.

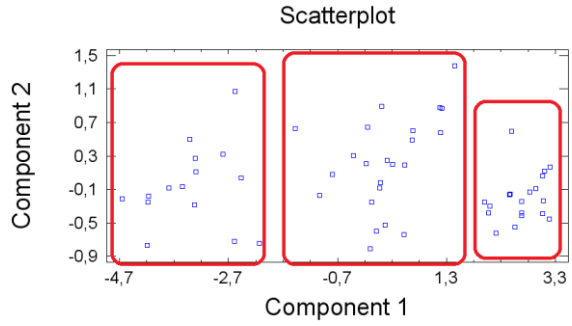


***Fig. 1.*** *Projection of the investigated objects on the space of two MC for QC*

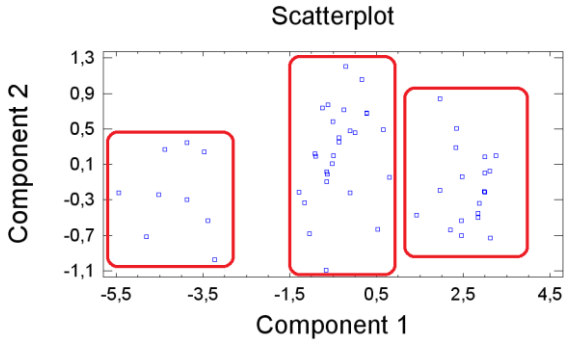


***Fig. 2.*** *Projection of the investigated objects on the space of two MC for SAM*

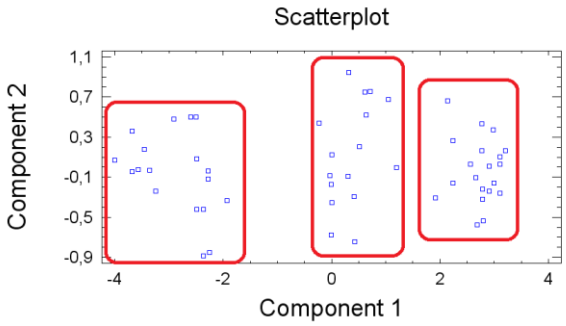


***Fig. 3.*** *Projection of the investigated objects on the space of two MC for MTS*

On the presented projections, three classes can be clearly traced.

### 3. Cluster analysis

The next step is a cluster analysis, which is applied to the partitioning of a set of objects (industrial enterprises) into a given number of clusters (three) obtained on the basis of the method of main component.

Cluster analysis is designed to split a set of objects into a previously unknown or, in rare cases, a given number of clusters on the basis of some mathematical criterion for clustering [5].

Hierarchical methods of cluster analysis are designed to obtain a visual representation of the stratification structure of the whole set of objects under study. The rules are the criteria used to solve the question of the "similarity" of objects when they are combined into a group - agglomeration methods that are constructed on the basis of a consecutive combination of objects into groups and a corresponding decrease in the number of clusters. At the beginning of the algorithm, all objects are separate clusters, then sequentially at each step of the algorithm operation the most similar ones are combined into a cluster until all objects form one cluster. The result of hierarchical methods of clustering is a dendrogram - a tree diagram containing several levels, each of which corresponds to one of the steps in the process of sequential cluster.

The most common agglomerative algorithms are: the single-link method, the method of complete communication, the mean-coupling method, the Ward method [3]. In this paper, the Ward method was used as the clustering method, since in our case it is desirable that the clustering algorithm works well with a small number of observations.

The results of constructing dendrograms displaying the hierarchical structure of grouping of objects in three directions of preparation (QC, SAM, MTS) for three clusters are presented in Fig. 4-6.

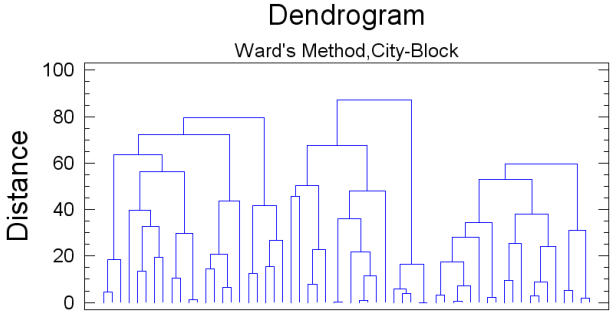


***Fig. 4.*** *Dendrogram for three clusters for QC*

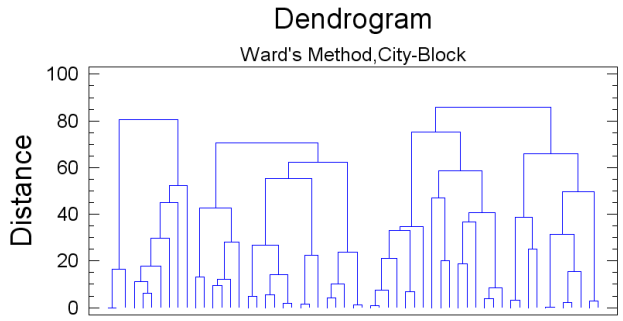


***Fig. 5.*** *Dendrogram for three clusters for SAM*

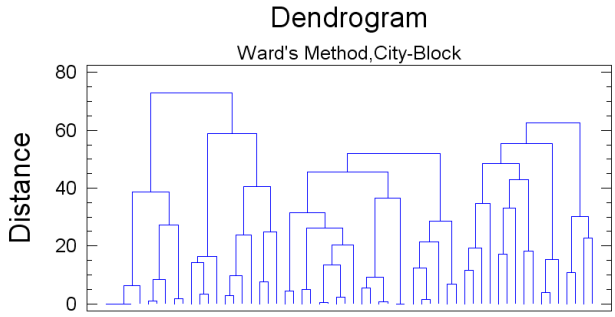


***Fig. 6.*** *Dendrogram for three clusters for MTS*

Based on the summary of cluster analysis, the rules for the formation of clusters.

Rule 1 for cluster 1: if the human resource = low; technology = low; PO = low; equipment = low; production site = low; professional growth = low; demand = low, then cluster 1.

Rule 2 for cluster 2: if the human resource = high; technology = high; software = high; equipment = high; production site = high; professional growth = high; demand = high, then cluster 2.

Rule 3 for cluster 3: if the human resource = average; technology = average; software = average; equipment = medium; production site = medium; professional growth = average; demand = average, then cluster 3.

The dispersion diagram, showing how the observations under study are grouped on the plane of two variables $x_4$ and $x_7$, for each areas of preparation is presented in Fig.7-9. Each cluster is indicated in the diagram by its own symbol.

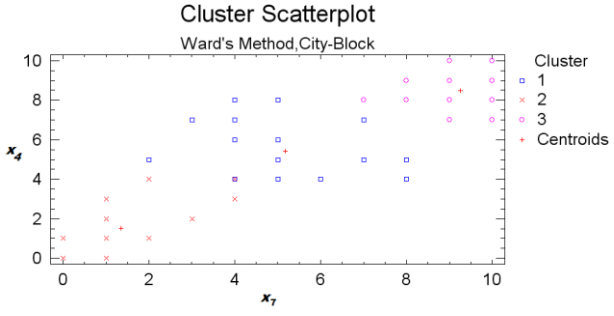


***Fig. 7.*** *The two-dimensional dispersion diagram for QC*

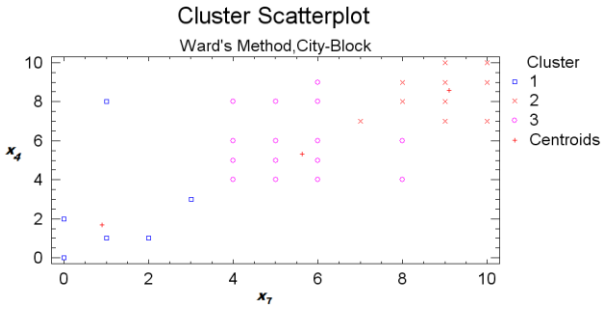


***Fig. 8.*** *The two-dimensional dispersion diagram for SAM*

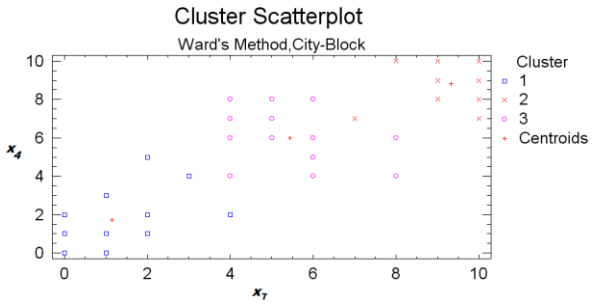


***Fig. 9.*** *The two-dimensional dispersion diagram for MTS*

A three-dimensional dispersion diagram showing how the observations under study are grouped on the plane of the three variables $x_4$, $x_5$ and $x_7$, and allowing a more detailed trace of the cluster's membership is shown in Fig. 10-12.

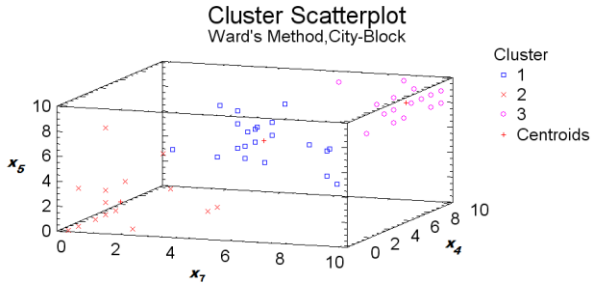


***Fig. 10.*** *The three-dimensional dispersion diagram for QC*

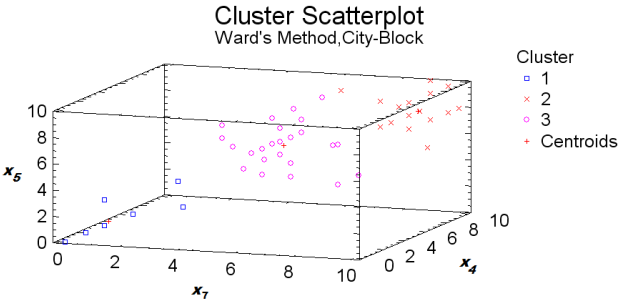


***Fig. 11.*** *The three-dimensional dispersion diagram for SAM*

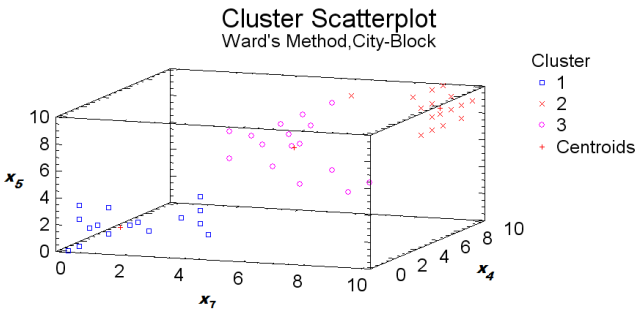


***Fig. 12.*** *The three-dimensional dispersion diagram for MTS*

Based on the analysis carried out, a model of the demand for graduates at industrial enterprises of the Republic of Bashkortostan (Fig. 13) is constructed. Enterprises are selected according to the second rule of cluster analysis and display the highest indicators of the selected characteristics.

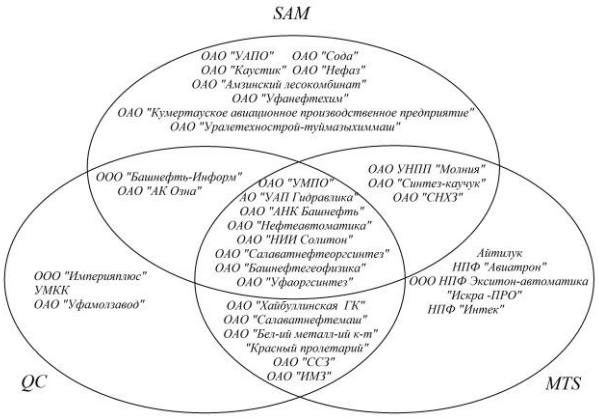


***Fig. 13.*** *Model of demand for graduates in industrial enterprises of the Republic of Bashkortostan*

## 4. Conclusion

The paper presents the results of the component and cluster analysis, which allowed:

- identify potential employers interested in training graduates in a large group of specialties 27.03.00 "Management in technical systems" in the areas of bachelor's training 27.03.02 "Quality management", 27.03.03 "System analysis and management", 27.03.04 "Management of informatics in technical systems";
- build a model of the demand for graduates in these areas at industrial enterprises of the Republic of Bashkortostan.

## References

1. Beljakin S.K., Teben'kova E.A. "Cluster model of university interaction with the professional community". *Vector of science TSU* 2013; 2 (24): 399-402.
2. Dubrov A.M., Mhitorjan V.S., Troshin L.I. "Multidimensional statistical methods", Moscow, Finance and Statistics Publ. House, 2000.
3. Djuk V., Samojlenko A. "Data Mining", StPetersburg: "Peter" Publ. House, 2001.
4. Kalinina V.N., Solov'ev V.I. "Introduction multivariate statistical analysis", Moscow, State University of Management Publ. House, 2003.
5. Korneev V.V., Garre A.F., Vasjutkin S.V., Rajh V.V. "Database. Intellectual information processing", Moscow, "Nolidzh" Publ. House, 2001.
6. Methodical recommendations for the development and implementation of higher education educational programs at the undergraduate level. Type of educational program "Applied Bachelor". Approved 11.09.2014 № AK-2916/03.
7. Simchera V.M. "Methods of multivariate analysis of statistical data", Moscow, Finance and Statistics Publ. House, 2008.