# Multidimensional Classification Method in the Study of Natural and Anthropogenic Systems

A.I. Gavrishin*

*South- Russian State Polytechnic University named after M.I. Platov, Novocherkassk, Russia*

**Abstract:** In this article the problem of build and use classification of multivariate observations; this issue has played a leading role in the knowledge of the world around us. When studying natural- anthropogenic systems have the greatest value probably-statistical simulation methods of objects, phenomena and processes. Among classification technologies to focus on situations of constructing classifications if there is no a priori information about the taxonomic structure of the observations. In the work described the rationale for the original statistical criterion $Z^2$ and G-classification methods of multivariate observations. This used quantity conversion method dependent traits in an equivalent number of independent. G-method lets you select homogeneous classes and sub-classes of observations, evaluate differences between taxons, determine the informative signs and much more. The application of the classification shown in the case study method regularities of formation of chemical composition of mine waters in the Eastern Donbass. It is four main types of changes in the composition of mine waters: the first type has the most significant impact on environmental degradation in the region; the second and third types reflects the influence composition of groundwater chloride in the composition of mine water; the fourth type is the soda water, which may indicate the presence of the region's oil and gas fields.

**Keywords:** Criterion $Z^2$, Classification, G- method, Taxonomic structure, Modeling, Mine water, Eastern Donbass.

## INTRODUCTION

Classifications are very important in studying the world around us. In the development of civilization is well known for the leading role of such classifications as the periodic system of chemical elements, stratigraphic scale, classification of species and many others.

When exploring natural and anthropogenic systems of the highest recognition to probabilistic and statistical models of objects, phenomena and processes. There are three most important parts of such models: deterministic part due to main controlled factors; random part due to numerous uncontrollable factors; random part due to errors in measurement. The first part describes the laws and require the construction of classifications, distorts the real situation [1].

Researcher, receiving new information on natural and human systems, faces the need to classify the observations described many features. Most likely to occur the task of constructing this new classification. The task can be formulated as follows. There is a set of observations (N), each of which characterized M signs; you need to find such a taxonomic structure in which observations are interconnected inside homogeneous taxon, and taxons differ as much as possible.

*Address correspondence to this author at the South- Russian State Polytechnic University named after M.I. Platov, Novocherkassk, Russia;
Tel: +07 9094251601; Fax: +07 8635255390;
E-mail: _agavrishin@rambler.ru

Among technologies of classification can be divided into three main types: building classification in the absence of a priori information about the taxonomic structure (a task without a teacher); build classification, when you know some of the information about the taxonomic structure (a task without a teacher); classification of observations on known classification. The first type of classification is the most important and least developed. This is the situation addressed in this work: set out the rationale for the original criterion $Z^2$ (Z-squared) and considered building a classification method based on it multidimensional observation (G-method). Application of classification method shown in the example, the study of the regularities of the formation of the chemical composition of mine waters in the Eastern Donbass.

## JUSTIFICATION THE CRITERION OF $Z^2$

In the derivation of the criterion $Z^2$ -Gavrishin for dependent observations and signs of us used reception dependent number of observations (N) statistically equivalent to the number of independent (n) proposed by A.A. Bagrov [1, 2], which revealed the following: N dependent observations

$$X = \left\{x_1, x_2, \cdots x_N\right\}$$

$x \in N\left(O, S_j\right)$ in addition covariance matrix $C = \bar{X}\bar{X}'$.

You can find linear orthogonal operator *U* such that $X = UZ$ vector, $Z = \left\{z_1, z_2, \cdots z_N\right\}$ there are

E-ISSN: 2409-5761/18 

independent observers. However, $U^{-1} = U^1$ and then $C = \overline{ZZ}' U^{-1}$.

To determine the equivalent number of (n) independent observations for the original vector X is equivalent to the same for a summer residence for the vector $Z : X' X = x_1^2 + x_2^2 + \cdots + x_N^2 = z_1^2 + z_2^2 + \cdots + z_N^2$.

Of characteristic functions

$$\Phi(t) = \prod_{j=1}^{N} \left(1 - 2i\sigma_1^2 t\right)^{-1/2}.$$

For the first two moments of the distribution of values $K = z_1^2 + z_2^2 + \cdots + z_N^2$ get

$$m_1 = \sum_{j=1}^{N} \sigma_j^2,$$

$$m_2 = \left[\sum_j^N \sigma_j^2\right]^2 + 2\sum_j^N \sigma_j^4$$

Accordingly, the points for the distribution of values

$$\chi^2 = \frac{1}{\sigma_0^2}\left(\xi_1^2 + \xi_2^2 + \cdots \xi_N^2\right)$$

are equal

$$m_1 = \sigma_0^2 n, m_2 = \left(\sigma_0^2 n\right) + 2n\sigma_0^4.$$

Now you can find

$$n = \left[\sum_j^N \sigma_j^2\right] / \sum_j^N \sigma_j^4,$$

$$\sigma_0^2 = \sum_j^N \sigma_j^4 / \sum_j^N \sigma_j^2.$$

Using $\overline{ZZ}' = U^{-1} C U$, for *n* and $\sigma_0^2$ get

$$n = \left[trC\right]^2 / trC^2, \sigma_0^2 = \frac{trC^2}{trC},$$

where tr *C* -trace matrix *C.*

If the covariance matrix *C take* the innovative correlation matrix R, then

$$n = \left[trR\right]^2 / trR^2 = N^2 / \sum_{ph} r_{ph}^2$$

where $r_{ph}$ is the correlation coefficient between observations $p$ and $h$.

This formula (for a more detailed conclusion makes A.A. Bagrov, [2]) define an equivalent number of independent observers alike $n$ for the number of dependent $N$ we used in long-distance.

Condition of equivalence can be written

$$\sum_j^N \left(x_j / S_j\right)^2 / N = \sum_k^n \left(\xi_k / \sigma_0\right)^2 / n,$$

$$\sum_k^n \left(\xi_k / \sigma_0\right)^2 = n\sum_j^N \left(x_j / S_j\right)^2 / N = N\sum_j^N \left(x_j / S_j\right)^2 / \sum_{ph} r^2{}_{ph}.$$

Because

$$\sum_k^n \left(\xi_k / \sigma_0\right)^2$$

is the distribution of $\chi^2$ with the number of degrees of freedom *(f) = (n),* therefore the value of

$$Z^2 = N\sum_j^N \left(x_j / S_j\right)^2 / \sum_{ph} r^2{}_{ph}$$

also has a $\chi^2$ distribution with

$$f = n = N^2 \sum_{ph} r^2{}_{ph}.$$

Turning to the more general case where the average non-zero, you can write a similar procedure for the dependent traits (s and k) using the equivalence formula

$$Z^2 = \left(M / \sum_{sk} r^2{}_{sk}\right) \sum_i^M \frac{\left(x_{ij} - \bar{x}_i\right)^2}{s_i^2}.$$

On the other hand, for dependent traits and independent observations are [1]:

$$Z^2 = \frac{M}{\sum_{sk} r^2{}_{sk}} \cdot \sum_{ij}^{MN} \frac{\left(x_{ij} - \bar{x}_i\right)^2}{s_i} = K\sum_{ij}^{MN} Z_{ij}^2,$$

$$f = KMN, \quad K = \frac{M}{\sum_{sk} r^2{}_{sk}}, \quad G = \sqrt{2z^2} - \sqrt{2f - 1}.$$

here *Xij* is the value of trait $j$ in observation $I$; $\bar{X}_j$, *Sj* are the mean and the standard deviation of the $j$ th

trait; $r_{sk}$ is the coefficient of correlation between the *s*th and the $k$ th traits; $M$ is the number of traits; *N* is the number of observations; *f* is the number of degrees of freedom; *G* is a transformation of distribution $\chi 2$ to the normal distribution with parameters (0,1). If the calculated $G > Gq$, then the observation (or *N* observations) do not belong to the given homogeneous class of observations with losses level $q$.

Thus, we are convinced that the proposed criterion $Z^2$ has distribution close to $\chi^2$ if there are links between of signs.

## CLASSIFICATION METHOD OF MULTIVARIATE OBSERVATIONS

Based on the original criteria for Z-square developed new G-method [1, 3] classification of multivariate observations (separation of homogeneous sets), which has the following important properties:

1) Construction of classification in the absence of a priori information about the taxonomic structure of the observations (a task without a teacher);

2) The use of dependent traits;

3) Allocation of taxonomic structures of various levels (classes, subclasses, etc.);

4) An assessment of the statistical parameters of each homogeneous taxa (mean, variance, correlation coefficients);

5) Unlimited ratio (M) and the number of observations (N);

6) Use of the information on changing average values, heterogeneity and the relationship between signs in homogeneous taxons;

7) Evaluation of informatively the individual signs in the classification and exclusion of no informatively signs;

8) Evaluation of similarity-difference between homogeneous taxons.

Procedure qualification (G-method) is reduced to the following main operations:

- -selecting a coordinate system in which the transformation of the multidimensional space of attributive to the distribution of $Z^2$;
- -finding the center of the first homogeneous taxon;
- -the transformation of coordinate systems and finding all the observations of the first homogeneous taxon;
- -repeat these operations for the observations, which were not included in previous similar taxons; -evaluation of similarity-difference between homogeneous taxons in each and all indications simultaneously;
- -evaluation of informatively signs the taxonomic structure;
- -repeat all operations for different levels of reliability allocation of homogeneous taxons.

Of different ways of finding the center of a homogeneous taxon turned out to be the most effective central method nearest points. The Centre is the three observations (points) the closest among themselves in multidimensional sign space. Largest G all observations found that belong to this homogeneous taxon. Changing the critical radius of a homogeneous taxon Gq, the classification can be obtained at various levels of detail and varying degrees of homogeneity of the taxon. The smaller the value of Gq, then the greater the homogeneity taxon, more detail of the classification, but a lower reliability validity of differences between taxons (level losses q).

Evaluation of similarity-difference between homogeneous taxons is also based on the criteria of $Z^2$ and boils down to the definition of trataksonomy variance for each and all together.

Largest $Z^2$ or G can be evaluated any number of membership new observations to fusion units, *i.e*. to produce a classification of observations. G-method is implemented in several models of computer programs (Optim, Anatf, G-mode, AGAT, etc.). The most popular and used effectively proved the program AGAT-2 (certificate about State registration No. 2008615215 dated October 29, 2008), allowing you to automatically build multidimensional classification observations of various the level of detail [4].

G-method successfully applied to build a classification and description of spatial-temporal patterns of formation of objects and systems of the Earth, Moon, Mars, Saturn, the asteroid, in deep space and using the astrophysical, kosmochemical, remote, geophysical, hydrogeological, geochemical and engineering-geological information with the construction of relevant maps [3, 5-8].

## REGULARITIES OF FORMATION OF CHEMICAL COMPOSITION OF MINE WATER

The coal industry caused significant changes in the environment Eastern Donbass. Long practicing coal deposits and functioning of water reducing systems have led to a significant transformation of the environment in the region. Changing balance and regime of groundwater occurs transformation of the chemical composition of natural waters with the formation of mineralized mine waters, pollution of surface watercourses, development of the processes of consolidation and seal rocks and many others phenomena and processes. Restructuring of the coal industry and the mass closure of coal mines in the Eastern Donbass are intensifying the subsidence of the Earth's surface and deformation of rocks, flooding of territories and heaps, the formation of anomalous composition of waters, intensive pollution of surface waters, the selection of "dead air" and other negative phenomena [8-10].

To study the basic types of directions of changes in the chemical composition of mine waters in the Eastern Donbass region used the results of water testing 84's coalmines. For the entire sample of mine, waters have a high heterogeneity (Table **1**). Salinity (M-mineralization) varies from 1400 to 11600 mg/l, Cl content from 46 to 2798, $SO_4$ - from 4 till 4327 mg/l, etc., for most components relative standard exceeds 60%. The average composition of water is sulphate magnesium-sodium, type two by classification of O.A. Alekin.

The most interesting results are obtained using a serial classification analysis. Total AGAT-2 computer technology severed 10 uniform hydrogeological and three species (A.1, A. 2 and A. 3) of abnormal values (Table **2**). Further analysis of the spatial disposition of the homogeneous hydrogeochemical species on the charts "component content - salinity" resulted in a certain allocation of four types of changes in the chemical composition of mine waters. Types include the following species of hydrogeochemical: first (46 observations) - 1.1, 2.1, 2.2, 2.3, 2.4; the second (39)-1.1, 1.3, A.1, A.2; third (45) - 1.1, 4.1, A.3; fourth (35) - 1.1, 1.2, 1.4, 3.1 (Table **2**). Hydrogeochemical 1.1 type characterizes the initial phase of the formation of the chemical composition of mine waters and is included in all directions.

The first type of chemically modified mine water contains five homogeneous hydrogeochemical species (Table **2**); water from neutral low mineralized sulphate (1.1 type) go sour mineralized sulphate (2.2, 2.4) with high content of Fe (up to 0.1-0.2 g/l), Mn and Cu (up to 0.05-0.1 g/l), nitrates (up to 0.1 g/l), and other components. Well visible (as you move from one species to another) a legitimate reduction of $HCO_3$ and pH from neutral to highly acidic; salinity increases mainly due to the increase in $SO_4$, Na and Mg. Correlation coefficient of mineralization with the components has the following values: $SO_4$ (r = 0.97), Na (0.94), pH (-0.76), $HCO_3$ (-0.71), Mg (0.61), Ca (0.41). The average chemical composition of mine water hydrogeochemical first type is given in Table **3**; This marinated sulphate magnesium-sodium waters with mineralization 3.3 g/l, with the highest value relations ($rSO_4/rCl$) equal 11.3 and lowest ($rHCO_3/rCa + rMg$) - 0.04.

The second type of formation of chemical composition of mine water contains four hydrogeochemical species (Table **2**) and water vary from sulphate (1.1) to chloride-sulphate (a. 2). In the waters increase of the content of $SO_4$, Cl and Na. On the strength of the relationship with mineralization components are located in the following series: $SO_4$ (r = 0.94), Na (0.93), Cl (0.64), pH (0.47), Mg (0.45); a

**Table 1: Characteristics of Chemical Composition of Mine Water**

| Component | $X_m$ | Me | $X_{min}$ | $X_{max}$ | S |
|---|---|---|---|---|---|
| pH | 6.7 | 7.3 | 2.2 | 8.6 | 1.8 |
| $HCO_3$ | 272 | 285 | 0 | 991 | 211 |
| $SO_4$ | 1765 | 1558 | 390 | 4327 | 952 |
| Cl | 500 | 329 | 46 | 2798 | 507 |
| Ca | 100 | 72 | 15 | 363 | 80 |
| Mg | 217 | 220 | 14 | 510 | 105 |
| Na | 757 | 689 | 169 | 3558 | 476 |
| M | 3600 | 3200 | 1400 | 11600 | 1900 |

Note: $X_m$-arithmetic mean, Me-median, Xmin, Xmax - minimum and maximum value, S is the standard deviation.

**Table 2: Composition of Hydrogeochemical Kinds of Mine Water by Types (Components in mg/l and %-mol)**

| Component | Hydrogeochemical Types and Species | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | | | | | 2 | | | | 3 | | | 4 | | | |
| | 1.1 | 2.3 | 2.1 | 2.2 | 2.4 | 1.1 | 1.3 | A.1 | A.2 | 1.1 | 4.1 | A.3 | 1.1 | 1.4 | 1.2 | 3.1 |
| pH | 7.4 | 6.7 | 3.7 | 2.6 | 2.7 | 7.4 | 7.9 | 7.2 | 8.3 | 7.4 | 6.5 | 7.5 | 7.4 | 8.0 | 7.6 | 7.7 |
| $HCO_3$ | 297<br>14 | 151<br>4 | 1<br>0 | 0<br>0 | 0<br>0 | 297<br>14 | 364<br>12 | 336<br>8 | 308<br>6 | 297<br>14 | 114<br>2 | 540<br>7 | 297<br>14 | 431<br>18 | 417<br>16 | 827<br>26 |
| $SO_4$ | 113<br>65 | 2470<br>84 | 2350<br>92 | 3660<br>95 | 3660<br>89 | 1130<br>65 | 1700<br>66 | 2380<br>74 | 2880<br>70 | 1130<br>65 | 1660<br>46 | 2200<br>38 | 1130<br>65 | 763<br>41 | 1070<br>52 | 666<br>27 |
| Cl | 260<br>21 | 247<br>12 | 152<br>8 | 129<br>5 | 334<br>11 | 266<br>21 | 423<br>22 | 433<br>18 | 735<br>24 | 266<br>21 | 1400<br>52 | 2400<br>55 | 266<br>21 | 576<br>41 | 486<br>32 | 863<br>47 |
| Ca | 89<br>12 | 232<br>19 | 81<br>8 | 167<br>10 | 87<br>5 | 89<br>12 | 65<br>6 | 104<br>8 | 37<br>2 | 89<br>12 | 117<br>8 | 140<br>6 | 89<br>12 | 44<br>6 | 36<br>4 | 38<br>4 |
| Mg | 191<br>44 | 287<br>40 | 260<br>41 | 327<br>34 | 311<br>30 | 191<br>44 | 186<br>29 | 218<br>27 | 295<br>29 | 191<br>44 | 371<br>41 | 240<br>16 | 191<br>44 | 177<br>37 | 54<br>10 | 30<br>5 |
| Na | 363<br>44 | 575<br>41 | 621<br>51 | 1030<br>56 | 1270<br>65 | 363<br>44 | 809<br>65 | 1030<br>65 | 1360<br>69 | 363<br>44 | 883<br>51 | 2250<br>78 | 363<br>44 | 522<br>57 | 855<br>86 | 1110<br>91 |
| M | 2400 | 4000 | 3500 | 5500 | 5600 | 2400 | 3600 | 4500 | 5700 | 2400 | 4400 | 8100 | 2400 | 2600 | 2600 | 3600 |

special feature is the existence of a relationship with Cl. Average composition (Table **3**) is sulfate magnesium-sodium water of the second type with a relatively high concentration of Cl.

The third hydrogeochemical type contains three hydrogeochemical species (Table **2**) and is characterized by the transition from sulphate to sulphate-chloride waters. Salinity is formed mainly due to the growth in concentrations of Cl, $SO_4$, Na and $HCO_3$; the correlation coefficient of mineralization with the components are Na - 0.97, with Cl - 0.92, $SO_4$ - 0.85, $HCO_3$ - 0.47; in the first place on the strength of the connection (anions) Cl. On average, the water composition of the third type of sulfate-chloride magnesium-sodium (Table **3**).

The fourth hydrogeochemical type is formed four hydrogeochemical species (Table **2**) and testifies to move the composition of mine water from sulfate to hydrocarbonate-sulphate-chloride. It is original soda type. In salinity greatest role-play $HCO_3$, Cl and Na; on the strength of the relationship with mineralization components make up the next row: Na (r = 0.77), $HCO_3$ (0.62), Cl (0.54). A characteristic feature of the waters are very low concentrations of Ca and Mg. Average composition of sulfate-chloride sodium water

Listed facts allow us to draw conclusions about the genesis and processes of formation of the mine water allocated types. The first type is due to the development of intensive processes of oxidation of sulfur and sulfides, prisoners in coals and enclosing rocks (up to 3-5%). In the formation of chloride-sulphate waters of the second type are approximately equal to the role played by the processes of oxidation of sulfides and the processes of mixing water chloride composition originating in deep mines. The third type is further enhanced the influence of the inflow of groundwater chloride with increasing depth of refining coal seams, water become sulfate-chloride and the oxidation of sulphides fades into the background. The original is the fourth hydrogeochemical type to formed soda water with high content of HCO3 and very low Ca and Mg. The formation of these mine waters due to the influx of sodium groundwater and indicates the possible presence of oil and gas fields in the region [4, 8].

## CONCLUSION

When exploring natural and anthropogenic systems of the highest recognition to probabilistic and statistical models of objects, phenomena and processes. Among classification technologies to focus on situations of constructing classifications if there is no a priori information about the taxonomic structure of the observations. In the work described the rationale for the original statistical criterion $Z^2$ and G-classification methods of multivariate observations. This used quantity conversion method dependent traits in an equivalent number of independent. G-method lets you select homogeneous classes and sub-classes of

**Table 3: The Average Composition of Mine Water by Type (Components in mg/l and %-mol)**

| Component | Hydrogeochemical Types | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 1 | | 2 | | 3 | | 4 | |
| pH | 4.5 | | 7.8 | | 6.9 | | 7.7 | |
| $HCO_3$ | 65 | 2 | 352 | 10 | 299 | 6 | 545 | 20 |
| $SO_4$ | 2900 | 90 | 1893 | 67 | 1700 | 42 | 856 | 39 |
| Cl | 195 | 8 | 483 | 23 | 1543 | 52 | 626 | 41 |
| Ca | 149 | 11 | 76 | 6 | 125 | 8 | 39 | 4 |
| Mg | 286 | 35 | 212 | 30 | 284 | 28 | 84 | 15 |
| Na | 830 | 54 | 876 | 64 | 1246 | 64 | 832 | 81 |
| M | 4390 | | 3940 | | 5240 | | 2920 | |
| $rHCO_3$/rCa+rMg | | 0.04 | | 0.28 | | 0.17 | | 1.1 |
| $rSO_4$/rCl | | 11.3 | | 2.9 | | 0.81 | | 0.95 |

observations, evaluate differences between taxons, determine the informative signs and much more.

The application of the classification shown in the case study method regularities of formation of chemical composition of mine waters in the Eastern Donbass. In the formation of the composition of mine waters revealed four main types. On the first type of acid sulfate waters are formed, enriched Fe, MN, Al and other metals. Such water formed by processes of intensive oxidation of sulphides. The second and third type of mine water is formed under the influence of sodium chloride groundwater. The fourth type is the soda water. The genesis of these waters is associated with efficient evaporation-condensation processes in the aquatic-carbon phase and may indicate the possible presence of oil and gas fields in the Eastern Donbass.

G-mode successfully applied when examining patterns of forming objects on Earth, Moon, Mars, Jupiter, asteroids, comets and outer space [1, 5-8].

## REFERANCES

[1] Gavrishin AI. Hydrogeochemical studies using mathematical statistics and computing. M.: Nedra 1974; 146 p.

[2] Bagrov AA. On an equivalent number of independent data. Works of Hydrometeorological Center 1969; 44: PP 3-11.

[3] Gavrishin AI. The Methodological Aspect of Development and Application Multivariate Classification G-Mode for Analyses Geochemical Trend // Journal of Advances in Applied & Computational Mathematics 2014; 1(1): 21-27. https://doi.org/10.15377/2409-5761.2014.01.01.4

[4] Gavrishin AI. Mine Waters of the Eastern Donbass and Their Effect on the Chemistry of Groundwater and Surface Water in the Region // Water Resources. 2018; 45(5): 785-794. https://doi.org/10.1134/S0097807818050081

[5] M.Fulchignoni, M.A.Barucci, E.F.Tedesco. A classification of G-mode and the three-parameter asteroids taxonomic classification techniques using a common data set. Planetary and Space science 1995; 43: 691-694. https://doi.org/10.1016/0032-0633(94)00195-W

[6] Tossi F, Orosei R, Coradini A. and 21 colleagues. Correlations between VIMS and RADAR data over the surfer of Titan: Implications for Titans surface properties. // Icarus 2010; 208(1): 366-384. https://doi.org/10.1016/j.icarus.2010.02.003

[7] Coradini A, Tossi F, Adriani A. and 32 colleagues. Identification of spectral units on Phoebe. // Icarus 2008; 193(1): 233-251. https://doi.org/10.1016/j.icarus.2007.07.023

[8] Gavrishin AI, Coradini A. The origin and the formation laws of groundwater and mine water chemistry in the Eastern Donets Basin. // Water Resources 2009; 36(5): 38-547. https://doi.org/10.1134/S0097807809050066

[9] Sakrutkin VE, Sklyarenko GY. The influence of coal mining on groundwater pollution (Eaestern Donbass). International Multidisciplinary Scientific GeoConference Surveying and Mining Ecology Management, SGEM, 15th, PP 927-932.

[10] Borisova VE, Serbinovskaya AM. Alternative methods of mine waters purification in Eastern Donbass. In the collection "Science. Education. Culture ". Novocherkassk: SRSTU 2016; 153-156.

Received on 7-11-2018 Accepted on 5-12-2018 Published on 11-12-2018

DOI: http://dx.doi.org/10.15377/2409-5761.2018.05.3