УДК 004

# РАЗВИТИЕ И ПРИМЕНЕНИЕ ЦИФРОВОЙ КЛАССИФИКАЦИОННОЙ ТЕХНОЛОГИИ. G-МЕТОД

**Гавришин А.И.**

*Южно-Российский государственный политехнический университет имени М.И. Платова, Новочеркасск, e-mail: agavrishin@rambler.ru*

Цель статьи – кратко описать обоснование критерия Z-квадрат и процедуру оригинальной цифровой технологии классификации многомерных наблюдений (G-метод). Классификационные методы играют ведущую роль в познании жизни, общества и окружающего мира. На основе способа преобразования числа зависимых наблюдений (N) в статистически эквивалентное число независимых (n), предложенный А.А. Баговым (1969), выполнен вывод формулы критерия $Z^2$ для зависимых наблюдений и признаков. На основе оригинального критерия Z-квадрат развита новая цифровая компьютерная технология – G-метод классификации многомерных наблюдений, который обладает следующими важными свойствами: построение классификации в условиях неопределенности (задача без учителя); использование зависимых признаков; выделение таксонов различных уровней; оценку информативности отдельных признаков и исключение неинформативных; оценку сходства-различия между однородными таксонами по каждому признаку и по сумме признаков. Цифровая компьютерная технология – G-метод классификации многомерных наблюдений – успешно применена при изучении объектов, явлений и процессов на Земле, Луне, Марсе, Юпитере, кометах, астероидах и в дальнем космосе. В статье приведены примеры применения классификационной технологии при построении и анализе классификации лунных пород и стекол по химическому составу, по изучению связи особенностей поверхности Марса с его геологическим строением, по анализу связи поверхности спутника Сатурна Феба с его составом, по описанию закономерностей формирования химического состава шахтных вод Восточного Донбасса.

**Ключевые слова: статистический критерий, классификационная технология, без учителя, однородные таксоны, зависимые признаки**

# DEVELOPMENT AND APPLICATION OF DIGITAL CLASSIFICATION TECHNOLOGY. G-METHOD

**Gavrishin A.I.**

*South-Russian State Polytechnic University, Novocherkassk, e-mail: agavrishin@rambler.ru*

The purpose of the paper is to briefly describe the rationale for the criterion of Z-squared and guaranteed original digital technology classification of multivariate observations (G-method). Classification methods play a key role in the knowledge of life, society and the world around them. Classified means. Based on how the number of dependent observations (N) statistically equivalent to the number of independent (n) proposed by A.A. Bagrov (1969) made criterion formula $Z^2$ for dependent observations and signs. Based on the original Z-squared criteria developed new digital computer technology – G-classification method of multivariate observations (separation of homogeneous sets), which has the following important properties: building classification in the absence of a priori information about the taxonomic structure of the observations (a task without a teacher); the use of dependent traits; selection of taxa at various levels (types, classes, subclasses, etc.); assessment of the descriptiveness of the individual signs the classification and exclusion of no informative signs; evaluation of similarity-difference between homogeneous taxa in each basis and on the amount of signs. Digital computer technology – G-classification method of multivariate observations successfully applied when examining objects, phenomena and processes on Earth, Moon, Mars, Jupiter, comets, asteroids and deep space. The article gives examples of classification of technology when building and classification analysis of Lunar rocks and glass by chemical composition, for the study of the communication features of the surface of Mars with its geological structure, communication analysis the surface of Saturn's moon Phoebe since its composition, according to the description of the regularities of the formation of the chemical composition of mine waters in Eastern Donbass.

**Keywords: statistical criteria, classification technology, without a teacher, similar taxa, dependent traits**

В познании жизни, окружающего мира и общества классификационные методы играют ведущую роль. В определенном смысле можно утверждать, что классифицировать – значит познать. Хорошо известна роль гениальных классификаций в развитии человеческих знаний, например, стратиграфической шкалы, периодической системы химических элементов, классификаций биологических видов, горных пород, подземных вод и многих других.

Получая новую информацию, исследователь часто сталкивается с проблемой построения новой классификационной структуры наблюдений. Для решения этой задачи необходима эффективная цифровая компьютерная технология построения классификации многомерных наблюдений. При выделении однородных частей объектов, явлений и процессов появляется возможность обоснованно применить математико-статистические методы их объективной

характеристики, а сравнением однородных частей – открывать новые закономерности их формирования.

По теории и методам классификации многомерных наблюдений имеются сотни работ, в которых предлагается использование кластерного анализа, применение алгоритмов классификации в условиях наличия априорной информации о классификационной структуре наблюдений (задача с учителем) и в меньшей степени задача самоорганизации (без учителя). Одной из первых работ в области распознавания образов считают исследования Ф. Розенблата по созданию персептрона в 1950-е гг. В России различным методам классификации наблюдений и распознавания образов посвящены сотни публикаций [1–3]; в области классификации геологических данных также выполнены многочисленные исследования [4–6].

Столкнувшись с проблемой обобщения геохимической информации по Уральскому региону (сотни тысяч анализов и тысячи геохимических аномалий), автор пришел к выводу о необходимости разработки новой классификационной технологии. Эта технология получила название G-метод, использована при обосновании многих научных и прикладных вопросов и опубликована более чем в 200 работах в России и за рубежом [6–8].

Одна из первых работ автора по проблемам построения классификации геологических данных опубликована в журнале «Геология и геофизика» [9]. Основная идеология нового критерия была представлена в кандидатской диссертации «Вопросы методики и некоторые результаты применения математической статистики и ЭЦВМ при гидрогеохимических исследованиях (на примере Урала)». В настоящей работе автор излагает краткий вывод критерия, основные положения классификационной технологии и некоторые примеры их использования.

*Вывод критерия*

При выводе формулы $Z^2$ для зависимых наблюдений и признаков нами использован прием преобразования числа зависимых наблюдений (N) в статистически эквивалентное число независимых (n), предложенный А.А. Багровым [3], который показал, что

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

где $r_{ph}$ – коэффициент корреляции между наблюдениями $p$ и $h$, tr R – след матрицы $R$.

Эту формулу определения эквивалентного числа независимых наблюдений $n$ по числу зависимых $N$ мы использовали для обоснования критерия $Z^2$ [5, 6, 8].

Воспользовавшись условием эквивалентности, можно записать

$$n\sum_{j}^{N}\left(x_j / S_j\right)^2 / N = N\sum_{j}^{N}\left(x_j / S_j\right)^2 / \sum_{ph} r_{ph}^2,$$

где $x_j$ – значение наблюдения $j$; $S_j$ – среднеквадратичное отклонение.

Но первая часть этого уравнения имеет распределение $\chi^2$ с числом степеней свободы $f = n$, следовательно, величина

$$Z^2 = N\sum_{j}^{N}\left(x_j / S_j\right)^2 / \sum_{ph} r_{ph}^2$$

также имеет распределение $\chi 2$ с

$$f = n = N^2 / \sum_{ph} r_{ph}^2.$$

Переходя к более общему случаю, когда среднее не равно нулю, можно записать аналогичную процедуру для зависимых признаков $M$ и независимых наблюдений $N$, воспользовавшись формулой эквивалентности

$$Z^2 = \frac{M}{\sum_{sk} r_{sk}^2} \cdot \sum_{ij}^{MN} \frac{\left(X_{ij} - \overline{X}_j\right)^2}{S_j^2} = K\sum_{ij}^{MN} Z_{ij}^2,$$

$$f = K \cdot M \cdot N,\ G = \sqrt{2z^2} - \sqrt{2\left(f-1\right)},$$

где $X_{ij}$ – значение признака $j$ в наблюдении $i$; $\overline{x_j}$, $S_j$ – среднее и стандартное отклонение признака $j$; $r_{sk}$ – коэффициент корреляции между признаками $s$ и $k$; $M$ – число признаков; $N$ – число наблюдений; $f$ – число степеней свободы; $G$ – преобразование распределения $\chi^2$ к нормальному с параметрами (0,1). Если вычисленное $G > G_q$, то наблюдение (или $N$ наблюдений) по $M$ признакам не принадлежит данному однородному классу наблюдений с уровнем значимости $q$.

Последняя ситуация (независимые наблюдения и зависимые признаки) очень часто встречается на практике и поэтому рассмотрим ее наиболее подробно.

Если признаки функционально зависимы, то все $r_{sk} = 1$ и коэффициент $K = 1/M$; $Z_{1j}^2 = Z_{2j}^2 = \ldots = Z_{Mj}^2$ и тогда

$$Z^2 = \sum_{j}^{N} Z_j^2,$$

с числом степеней свободы $f = N$.

Это означает, что мы имеем дело только с одним независимым признаком.

Если признаки независимы, т.е. $r_{sk} = 0$ при $s \neq k$ и $r_{sk} = 1$ при $s = k$, то $K = 1$ и тогда

$$Z^2 = \sum_{ij} Z_{ij}^2,\ f = MN.$$

Получен тривиальный случай для независимых признаков.

Таким образом, последняя формула справедлива и для крайних случаев, когда величины зависят функционально или независимы.

*Метод классификации многомерных наблюдений*

При изучении природных объектов, явлений и процессов широко применяются различные модификации кластерных построений, выделения однородных статистических совокупностей, факторного анализа, процедуры распознавания образов, которые позволяют выделить некоторые однородные по заданному комплексу признаков части объектов и далее проводить анализ закономерностей их формирования.

Среди методов классификации многомерных наблюдений выделяются три главные группы:

1) построение классификации в условиях неопределенности (задача без учителя);

2) построение классификации при наличии некоторых априорных данных;

3) классифицирование новых наблюдений по известной классификации.

Формально задача построения классификации может быть сформулирована следующим образом: множество наблюдений *N*, каждое из которых охарактеризовано *M* признаками, необходимо разделить на подмножества – однородные таксоны, внутри которых наблюдения близки между собой, а таксоны максимально различаются.

В основу рассматриваемой классификационной процедуры положен описанный критерий $Z^2$, позволяющий использовать зависимые признаки. Проверяется нулевая гипотеза (Но) об отсутствии различий между множеством многомерных наблюдений и данной статистической совокупностью, т.е. гипотеза о принадлежности многомерных наблюдений к данному однородному таксону.

На основе оригинального критерии Z-квадрат развит новый цифровой компьютерный G-метод [5, 6, 8] классификации многомерных наблюдений, в котором удалось реализовать большинство указанных выше важных особенностей: построение классификации в условиях самоорганизации (задача без учителя); использование зависимых признаков; неограниченное соотношение между числом наблюдений (*N*) и числом признаков (*M*); выделение таксонов различных уровней; оценка сходства-различия между однородными таксонами по каждому признаку и по сумме признаков и другие.

Процедура цифровой компьютерной классификационной технологии сводится к следующим операциям: выбирается исходная система координат; отыскивается центр первого однородного таксона и все наблюдения этого таксона; выполняется повторение указанных операций для наблюдений, не вошедших в предыдущие однородные таксоны.

Создано несколько вариантов компьютерных программ реализующих G-метод (Оптим, Анатф, G-mode, AGAT и др.). Наиболее эффективной оказалась программа AGAT-2 (Свидетельство о государственной регистрации программы № 2008615215 от 29 октября 2008 г.), позволяющая автоматически строить классификации многомерных наблюдений различного уровня детальности.

*Примеры применения классификационной технологии*

Цифровая компьютерная технология классификации многомерных наблюдений (G-метод) успешно применена при изучении природных и антропогенных систем на Земле, Луне, Марсе, Юпитере, кометах, астероидах и в дальнем космосе. По результатам этих исследований опубликовано более 200 работ, в том числе более 100 в России и за рубежом в соавторстве с итальянскими, французскими, немецкими и американскими коллегами [5, 6, 8].

Ниже рассмотрены отдельные примеры практического применения цифровой компьютерной технологии G-метод при построении классификаций объектов, явлений и процессов и описания закономерностей их формирования.

*Классификация лунных пород и стекол*

При построении классификации лунных пород и стекол по химическому составу использовано более 2500 анализов образцов, доставленных автоматическими станциями «Луна-16 и 20» и астронавтами кораблей «Аполлон-11, 12, 14, 15, 16 и 17». Для каждой из космических экспедиций были составлены классификации образцов по составу породообразующих окислов [5, 7]. Обобщение всех классификаций позволило уверенно выделить семь главных групп лунных пород:

1. Анортозиты (габбро-анортозиты).

2. Материковые базальты (полевошпатовые базальты, анортозитовые габбро).

3. Криповые базальты (KREEP, базальты Фра Мауро).

4. Габбро-базальты (морские базальты Моря Изобилия).

5. Морские базальты (базальты Океана Бурь и Моря Дождей).

6. Высокотитанистые морские базальты (базальты Моря Спокойствия).

7. Калиевые граниты (высококремнистые породы).

Группы объединены в три типа. Первый тип – анортозитовые породы – является материковым веществом. Ко второму типу отнесены пять групп базальтов, которые образовались позднее первого типа при излиянии расплавов вещества на поверхность. Третий оригинальный по составу тип обнаружен впервые авторами и соответствует калиевым гранитам, которые могут быть остаточным материалом после отделения базальтов. Наличие значительного количества калиевых гранитов на Луне может сильно повлиять на существующие представления о ее элементарном составе, так как с указанными породами могут быть связаны повышенные концентрации многих элементов, которые сейчас относят к потерянным в процессе аккреции.

*Изучение поверхности Марса*

G-метод применен при построении классификации топографической поверхности Марса [10] по данным Марсианского орбитального лазерного альтиметра (MOLA). В качестве главных признаков использованы средние квадратичные отклонения высоты и уклонов и экспонента Харста (Hurst). Выявлено, что параметр «среднее квадратичное отклонение высот» резко различен для северного и южного полушарий Марса, в то время как экспонента Харста имеет выраженный широтный тренд. Классификация наблюдений выявила 29 однородных таксонов, некоторые из них имеют четкую корреляцию с геологическим строением планеты, другие – коррелируются с широтным трендом.

*Изучение спутника Сатурна Феба*

Для изучения состава поверхности спутника Сатурна Феба использованы спектральные данные, полученные космической экспедицией Кассини (Cassini) с помощью спектрометра VIMS. Феба, по мнению многих астрономов, является внешним телом, захваченным гравитацией Сатурна из пояса Купера, поэтому он характеризуется оригинальным составом и строением. По результатам классификации спектральных данных [11] установлено, что помимо крупных таксонов, характеризующих основную часть поверхности спутника, которая сложена каменно-ледяным материалом, выделены оригинальные классы, свидетельствующие о наличии участков, где коррелируются высокие содержания $H_2O$ и $CO_2$ (газ заключен во льду) и где $CO_2$ не коррелируется с $H_2O$. Это может свидетельствовать о наличии оригинальных процессов на Фебе.

*Изучение гидрогеохимических закономерностей*

Значительные объемы исследований и количество публикаций посвящены проблемам изучения закономерностей формирования химического состава шахтных, грунтовых и поверхностных вод Восточного Донбасса [5, 6, 8]. При обобщении гидрогеохимической информации использовано более 2000 анализов вод за столетний период (с 1920 по 2015 г.).

*Типы изменения состава шахтных вод.* С помощью G-метода выделено четыре главных гидрогеохимических типа изменения химического состава шахтных вод по результатам опробования в 1923, 1967, 1992 и в 2015 гг. [8]. Первый тип – это слабокислые (величина pH соответственно составила 4,4; 4,5; 6,0 и 5,7) сульфатные магниево-кальциево-натриевые воды с минерализацией 4,6; 4,4; 4,5 и 5,7 г/л. Второй – нейтральные (величина pH – 7,3; 7,8; 7,6 и 6,7) хлоридно-сульфатные натриевые воды с минерализацией 3,9; 3,9; 4,2 и 7,6 г/л. Третий тип (отсутствует в 2015 г.) – нейтральные (pH – 7,0; 6,9 и 7,8) сульфатно-хлоридные натриевые воды с минерализацией 3,2; 3,0 и 5,1 г/л. Четвертый – нейтральные (pH – 7,6; 7,7; 7,6 и 7,3) гидрокарбонатно-сульфатно-хлоридные натриевые (содовые) с минерализацией 3,2; 2,9; 4,5 и 2,7 г/л.

Наиболее интересные генетические выводы связаны с первым и четвертым типами. Первый тип характеризуется преобразованием исходных слабоминерализованных вод в кислые сульфатные воды с высокими содержаниями Fe, Al, Cu, Pb, Co и других металлов, что обусловлено интенсивным развитием процессов окисления серы и сульфидов. Воды первого типа формируют наиболее интенсивные потоки загрязнения природных вод региона [12, 13].

Четвертый тип формирования химического состава шахтных вод – это оригинальные содовые гидрокарбонатно-сульфатно-хлоридные и хлоридные натриевые воды с высокими содержаниями $HCO_3$ и очень низкими Ca и Mg. В горные выработки поступают содовые подземные воды, которые образуются в результате испарительно-конденсационных процессов в водоуглеродной газовой фазе (обратная вертикальная геохимическая зональность подземных вод региона). Автор делает прогноз [5, 6, 8], что в районе угольных шахт, где обнаружены содовые воды четвертого типа, наиболее высоки перспективы обнаружения нефтегазовых месторождений,

например, в структурах Гуково-Зверевского угленосного района.

**Заключение**

В познании жизни, общества и окружающего мира классификационные методы играют ведущую роль. Получая новую информацию, исследователь обычно сталкивается с проблемой построения новой классификационной структуры наблюдений для обнаружения и описания ранее не известных закономерностей.

Вывод формулы описанного критерия $Z^2$ для зависимых наблюдений и признаков выполнен с использованием способа преобразования числа зависимых наблюдений (N) в статистически эквивалентное число независимых (n), предложенного А.А. Багровым.

На основе оригинального критерии Z-квадрат развита и применена новая цифровая компьютерная технология – G-метод классификации многомерных наблюдений, который обладает такими важными свойствами, как построение классификации в условиях неопределенности (задача без учителя); использование зависимых признаков; оценка сходства-различия между однородными таксонами, выделение таксонов различного уровня и другими.

Цифровая компьютерная классификационная технология (G-метод) успешно применена при изучении объектов, явлений и процессов на Земле, Луне, Марсе, Юпитере, кометах, астероидах и в дальнем космосе. Например, применение указанной технологии при изучении лунных пород и стекол позволило построить их классификацию по химическому составу, изучить связь особенностей поверхности Марса с его геологическим строением, установить зависимость цвета поверхности спутника Сатурна Феба от его состава, описать закономерности формирования химического состава шахтных и подземных вод Восточного Донбасса.

*Автор выражает благодарность за помощь в проведении исследований Международной программе «Эразмус. Минерал+», «Модернизация геологического образования в российских и вьетнамских университетах».*

**Список литературы**

1. Акопов А.С. Имитационное моделирование. М.: Наука, 2018. 389 с.

2. Бухтояров В.В. Трехступенчатый эволюционный метод формирования коллективов нейронных сетей для решения задач классификации // Программные продукты и системы. 2012. № 4. С. 101–106.

3. Фомин Я.А. Распознавание образов: теория и применения. М.: ФАЗИС, 2012. 429 с.

4. Багров А.А. Об эквивалентном числе независимых данных // Труды Гидромет. науч.-исслед. центра. Л.: Гидрометиздат, 1969. Вып. 44. С. 3–11.

5. Гавришин А.И., Корадини А. Многомерный классификационный метод и его применение при изучении природных объектов. М.: Недра, 1994. 92 с.

6. Gavrishin A.I. Multidimensional Classification Method in the Study of Naturel and Anthropogenic Systems. Journal of Advances in Applied & Computational Mathematics. 2018. No. 5. P. 16–21.

7. Tossi F. and seven colleagues. G-mode classification of spectroscopic data. Earth, Moon and Planets. 2005. № 96. P. 165–197.

8. Gavrishin A.I. Mine Waters of the Eastern Donbass and Their Effect on the Chemistry of Groundwater and Surface Water in the Region. Water Resources. 2018a. vol. 45. No. 5. P. 785–794.

9. Гавришин А.И., Юшков Ю.Н. О математической интерпретации результатов геохимических поисков // Геология и геофизика. 1967. № 6. С. 67–72.

10. Orosei R. and six colleagues. Self-affine behavior of Martian topography at kilometer scale from Mars Orbiter Laser Altimeter data. Journal of Geophysical Research. 2003. Vol. 108. No. E4 (8023). P. GDS 4-1 – 4-10.

11. Coradini A. and 32 colleagues. Identification of spectral units on Phoebe. Icarus. 2008. vol. 193. No 1. P. 233–251.

12. Закруткин В.Е., Скляренко Г.Ю., Гибков Е.В. Особенности химического состава и степень загрязненности подземных вод углепромышленных районов Восточного Донбасса // Известия вузов. Северокавказский регион. Серия: Естественные науки. 2014. № 4. С. 73–77.

13. Мохов А.В. О растекании шахтных вод из затопленных угольных шахт в недрах // Доклады Академии наук. 2011. Т. 438. № 4. С. 494–496.