2. Кириченко Е.А. Механика глубоководных гидротранспортных систем в морском горном деле: [монография] / Евгений Алексеевич Кириченко. – Д.: Национальный горный университет, 2009. – 344 с.
3. К вопросу разработки способа автоматизированного управления переходными режимами в глубоководных эрлифтах / В.Е. Кириченко // Науковий вісник НГУ. – Дніпропетровськ: НГУ. – 2008. – № 11. – С. 71 - 75.
4. «Armadillo» C++ библиотека для решения задач линейной алгебры и матричных вычислений [Электронный ресурс] – Режим доступа: http://arma.sourceforge.net/ – Название с домашней страницы интернет.
5. «Blitz++». C++ библиотека для быстрых математических расчетов [Электронный ресурс] – Режим доступа: http://blitz.sourceforge.net/ – Название с домашней страницы интернет.
6. «GiNaC». Мощная математическая библиотека с богатыми возможностями символьных вычислений [Электронный ресурс] – Режим доступа: http://www.ginac.de/ – Название с домашней страницы интернет.
7. «MONO». Кроссплатформенный .Net фреймворк с открытым исходным кодом [Электронный ресурс] – Режим доступа: http://www.mono-project.com – Название с домашней страницы интернет.
8. «QT». Кроссплатформенный C++ фреймворк с открытым исходным кодом. [Электронный ресурс] – Режим доступа: http://qt.digia.com, http://qt-project.org – Название с домашней страницы интернет.

Рекомендовано до публікації д.т.н. Мещеряковим Л.І.
Надійшла до редакції 03.11.2014

УДК 550.428:553.93

© В.В. Ишков, Е.С. Козий

О КЛАССИФИКАЦИИ УГОЛЬНЫХ ПЛАСТОВ ПО СОДЕРЖАНИЮ ТОКСИЧНЫХ ЭЛЕМЕНТОВ С ПОМОЩЬЮ КЛАСТЕРНОГО АНАЛИЗА

Рассмотрены особенности использования кластерного анализа для классификации угольных пластов по содержанию токсичных элементов.

Розглянуто особливості використання кластерного аналізу для класифікації вугільних шарів щодо вмісту токсичних елементів.

There were considered features to apply cluster analysis to classify coal seams as for toxic elements content.

Изучение концентраций токсичных и потенциально токсичных элементов в углях пластов Красноармейского геолого-промышленного района обусловлено ужесточением требований к охране окружающей среды. Актуальность таких исследований обусловлена рядом Законов Украины, постановлениями Кабинета Министров, а также требованиями ГКЗ к качеству и содержанию геологических материалов при разведке угольных месторождений.

Научный и практический интерес вызывает установление возможностей классификации угольных пластов района по содержанию токсичных и потенциально токсичных элементов на основании результатов различных методов

кластерного анализа. К сожалению, эта проблема оставалась до настоящего времени практически не исследованной.

Использование кластерного анализа в целях классификации имеет ряд преимуществ, так как позволяет выполнить разбиение множества исследуемых объектов и признаков на однородные в соответствующем понимании группы или кластеры, а также выявить внутреннюю структуру (на разных иерархических уровнях) изучаемой выборочной совокупности. В то же время, как и любой другой метод, кластерный анализ имеет определенные недостатки. В частности, состав и количество кластеров зависит от выбираемых критериев группировки («стратегии классификации»), а применение различных методов, соответствующих различным концептуальным подходам к выделению таксонов, к одним и тем же выборкам, может привести к существенно отличающимся результатам. Таким образом, характерной особенностью кластерного анализа, в отличии от других методов многомерной статистики, служит сильная зависимость получаемых результатов от априорных установок исследователя на содержательном уровне. В связи с этим в данной работе основными задачами являлись: анализ результатов кластеризации угольных пластов различными методами, реализованными в одной из наиболее популярных профессиональных статистических программ «STATISTICA 6.0» [1] и выбор наиболее оптимального из них.

Для выполнения кластерного анализа в программе предлагается семейство иерархических агломеративных методов, двувходового объединения и итеративный дивизимный метод K-средних.

Метод двухвходового объединения используется при одновременной кластеризации как наблюдений так и переменных. В этом случае ожидается, что и наблюдения и переменные одновременно вносят вклад в выявление кластеров которые дальше интерпретируются в геологических понятиях. Главным недостатком метода являются проблемы с понятийной интерпретацией результатов, которые являются следствием того, что расстояние между разными кластерами может определяться различиями в переменных.

Проблематичность понятийной интерпретации результатов анализа не дает возможности для его использования в качестве оптимального метода решения поставленной задачи классификации.

Использование **итеративного дивизимного метода K-средних** в целях оптимальной классификации угольных пластов по содержанию токсичных и потенциально токсичных элементов в угле имеет существенные недостатки. Для него характерна присущая всем итеративным дивизимным методам проблема субоптимальных решений, которые заключаются в неудачных исходных разбиениях выборочных совокупностей. Необходимо отметить, что итерации по методу K-средних очень чувствительны к неудачным случайным разбиениям, к тому же все еще более усложняется при случайном выборе начального разбиения реализованного в программе. Его использование подразумевает существование априорных гипотез относительно числа кластеров (по наблюдениям или по переменным) а результат кластеризации представленный в виде системы таблиц, не позволяет наглядно и однозначно выявить и визуализировать структуру классификации.

Перечисленные недостатки метода не позволяют рассматривать его в качестве оптимального для классификации угольных пластов по содержанию токсичных и потенциально токсичных элементов в угле.

Семейство иерархических агломеративных (объединяющих) методов реализованных в программе относится к наиболее часто используемым группам кластерного анализа. Все они заключаются в последовательном объединении наиболее схожих объектов, которое можно визуализировать в виде древовидной диаграммы – дендрограммы (которая графически отображает иерархическую структуру матрицы сходства объектов). Такая наглядность результатов кластеризации является существенным достоинством этих методов. Как правило, в дендрограмме по горизонтали указываются кластеризуемые объекты, а по вертикали - значения межклассовых расстояний, при которых происходит их объединение (коэффициент слияния или дистанция объединения). При этом в результате анализа формируются группы не перекрывающихся кластеров, причем каждый кластер является элементом более широкого кластера на более высоком уровне сходства. Такой подход не требует профессиональной подготовки исследователя в области матричной алгебры или многомерной статистики, и в то же время позволяет однозначно интерпретировать результаты в геологических понятиях.

По способу группировки или «стратегии классификации» все иерархические агломеративные методы реализованные в программе подразделяются на: метод одиночной связи («ближайшего соседа») [2], метод полной связи («наиболее удаленного соседа») [3], разновидности метода «средней связи»: невзвешенный метод «средней связи» («невзвешенное попарное среднее») и взвешенный метод «средней связи», взвешенный центроидный метод и метод Уорда [4]. Кроме того, во всех перечисленных методах могут быть использованы в качестве межклассовых расстояний: евклидово расстояние (или его квадрат) манхэттенское расстояние («расстояние городских кварталов»), метрики Чебышева и Миньковского, линейный коэффициент корреляции (точнее, 1 - линейный коэффициент корреляции), простой коэффициент совстречаемости (точнее, 1 – коэффициент совстречаемости). Применительно к особенностям решаемой задачи наиболее оптимальным является использование в качестве меры сходства евклидового расстояния.

Рассмотрим возможности применения иерархических агломеративных методов кластеризации к задаче классификации угольных пластов Красноармейского геолого-промышленного района по содержанию токсичных и потенциально токсичных элементов более подробно.

Метод одиночной связи формирует кластеры из принципа наличия хотя бы одной связи между объектами. Несмотря на то что, его результаты инвариантны к монотонным преобразованиям матрицы сходства и использование метода не ограничивает присутствие «совпадения» в данных, практическое его применение в целях классификации вызывает затруднения.

На примере результатов кластеризации угольных пластов по содержанию Ni (рис. 1) и Be (рис. 2) в углях видно, что по мере приближения к завершению процесса кластеризации формируется один большой кластер, в который остав-

шиеся объекты включаются один за другим. Окончательный результат является тривиальным следствием наличия одного кластера, включающего n-1 объектов и одного кластера, содержащего один объект (см. рис. 1 и 2).

Анализ рис. 1 и рис. 2 не позволяет определить количество и структуру кластеров содержащихся в исходных данных. Аналогичные результаты получены и при кластеризации этим методом угольных пластов по концентрациям других токсичных и потенциально токсичных элементов.

Метод полной связи, в отличии от рассмотренного выше метода одиночной связи накладывает более жесткие требования к объединению объектов в один кластер. В данном случае (рис. 3 и рис. 4), появляется тенденция к выявлению относительно компактных гиперсферических (в многомерном пространстве) кластеров, объединяющих схожие объекты.

Сопоставление рис.1 с рис. 3 и рис. 2 с рис. 4 позволяет выявить ряд преимуществ кластеризации с использованием метода полных связей. В тоже время, если при кластеризации по содержанию Ni полученная дендрограмма достаточно убедительно указывает на наличие трех кластеров (см. рис. 3), то при кластеризации по содержанию Be, выявление окончательного количества кластеров (см. рис. 4) не так очевидно. Кроме того, в обоих случаях использование только дендрограммы без привлечения первичных данных затрудняет отнесение отдельных пластов к тем или иным кластерам. Причем, если на рис. 3 это только пласт l_3, то на рис. 4 это уже ряд пластов: m_6^1, $l_7^н$, l_3, k_5.

Анализ результатов кластеризации угольных пластов по содержанию Ni показывает, что первый кластер формируют пласты с максимальным средним содержанием (42–27г/т), второй объединяет пласты со средними концентрациями (22–14г/т), а третий – с минимальными (13–10г/т).

Метод «средней связи» разработан Сокэлом и Минченером в 1958 г. как компромисс между методами одиночной и полной связи. В программе «STATISTICA 6.0» реализованы две разновидности метода: невзвешенный метод «средней связи» («невзвешенное попарное среднее») и взвешенный метод «средней связи». В первой разновидности метода расстояние между двумя кластерами вычисляется как среднее расстояние между всеми парами объектов в них, а во второй – кроме того, размер кластеров (т.е. количество содержащихся в них объектов) используется в качестве весового коэффициента. Использование количества содержащихся в кластере объектов в качестве весового коэффициента предполагает «хорошее качество» анализа при наличии в выборке кластеров неравного размера.

Сравнение результатов кластеризации угольных пластов по содержанию Ni невзвешенным (рис. 5) и взвешенным методами (рис. 7) «средней связи» показывает, что в обоих случаях четко проявляется наличие трех групп пластов: с аномально высокими содержаниями, средними и низкими концентрациями. Причем, если состав кластера содержащего пласты с аномально высокими содержаниями (среднепластовые значения 42–27г/т) остается постоянным, то структура и состав кластеров состоящих из пластов со средними и низкими концентрациями Ni в углях изменяется.

В первом случае (см. рис. 5) оба кластера примерно равных размеров, II кластер объединяет пласты со средними концентрациями Ni от 22 до 16г/т, а III кластер – от 15 до 10г/т. Во втором случае (см. рис. 7) размер II кластера увеличивается и он уже включает пласты со средними содержаниями Ni от 22 до 14г/т. Рассмотрение дендрограммы на рис. 7 без привлечения первичных данных затрудняет отнесение пластов $l_4^в$, $l_8^н$, m_2 и m_4^0 ко II или III кластеру.

Сопоставление результатов кластеризации угольных пластов по содержанию Be невзвешенным (рис. 6) и взвешенным методами (рис. 8) «средней связи» не позволяет визуально однозначно установить количество результирующих кластеров. В тоже время, структура кластеров и их размеры практически не изменяются.

Взвешенный центроидный метод использует в качестве расстояния между кластерами (объектами) расстояние между их центрами тяжести. На рис. 9 и рис. 10 приведены дендрограммы результатов кластеризации взвешенным центроидным методом угольных пластов соответственно по содержанию Ni и Be в угле. В первом случае достаточно уверенно выделяется наличие трех кластеров, а во втором – четырех. В обоих случаях видна четкая и однозначная структура кластеров, а также их существенно отличающиеся размеры.

На дендрограмме кластеризации пластов по содержанию Ni (см. рис. 9) первый кластер составляют пласты с аномально высокими концентрациями (от 42 до 27г/т), второй кластер – пласты со средним содержанием (от 22 до 14г/т) и третий кластер – пласты характеризующиеся минимальным содержанием Ni в угле (от 13 до 10г/т).

При кластеризации пластов по концентрациям Be (см. рис. 10) первый кластер образуют пласты с максимальным средним содержанием (4г/т), второй – с повышенным (2,9–2,4г/т), третий – с минимальным (1,5–1г/т) и четвертый – с относительно пониженным (2,1–1,6г/т).

Метод Уорда отличается от рассмотренных ранее иерархических агломеративных методов использованием методов дисперсионного анализа при оценки межкластерных расстояний. К недостаткам метода относится сильное воздействие профильного сдвига на результаты и стремление создавать кластеры малого и приблизительно равного размера. На рис. 11 и рис. 12 в качестве примеров реализации метода приведены дендрограммы результатов кластеризации угольных пластов соответственно по содержанию Ni и Be в угле.

Если в первом случае (см. рис. 11) при визуальном анализе дендрограммы достаточно уверенно устанавливается наличие трех результирующих кластеров, то во втором (см. рис. 12) их выявление представляется проблематичным. На дендрограмме кластеризации пластов по концентрациям Be отчетливо проявлены все названные выше недостатки метода.

Анализ дендрограммы кластеризации угольных пластов по содержанию Ni показывает, что первый кластер объединяет пласты с максимальным средним содержанием (42–27г/т), второй формируют пласты со средними концентрациями (22–17г/т), а третий – с минимальными (16–10г/т).

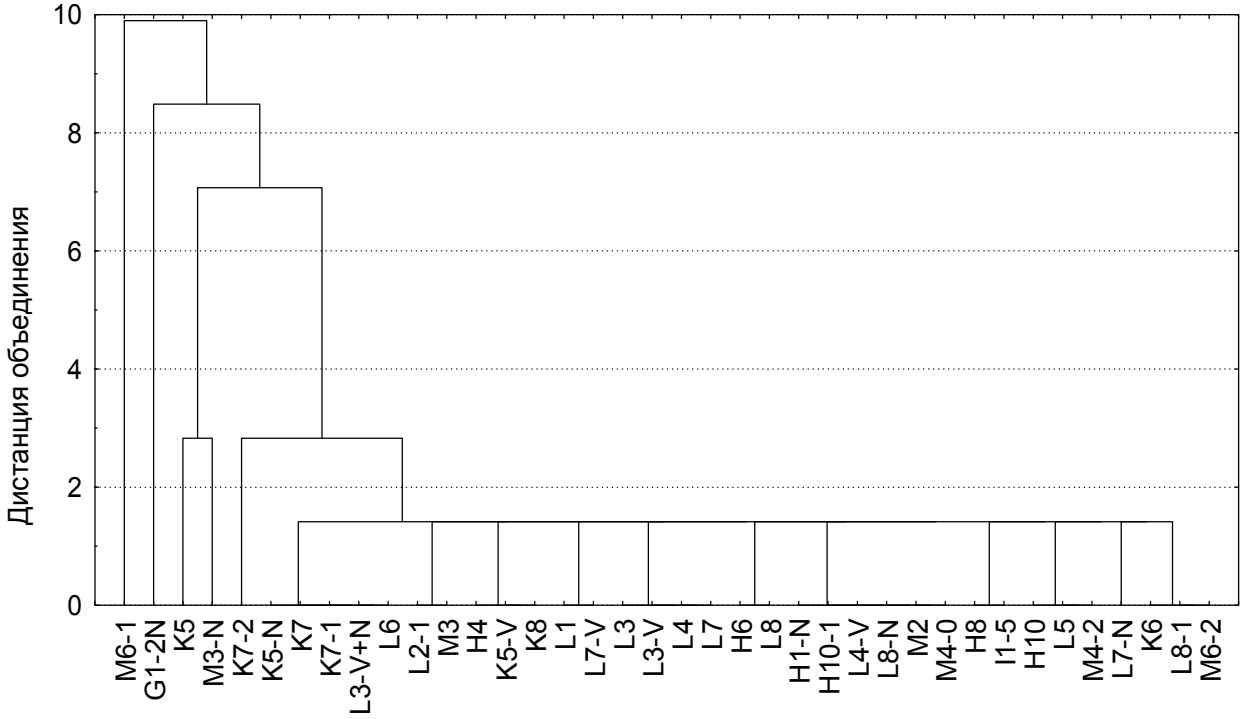

Рис. 1. Дендрограмма результатов кластеризации методом одиночной связи угольных пластов Красноармейского геолого-промышленного района по содержанию Ni в угле

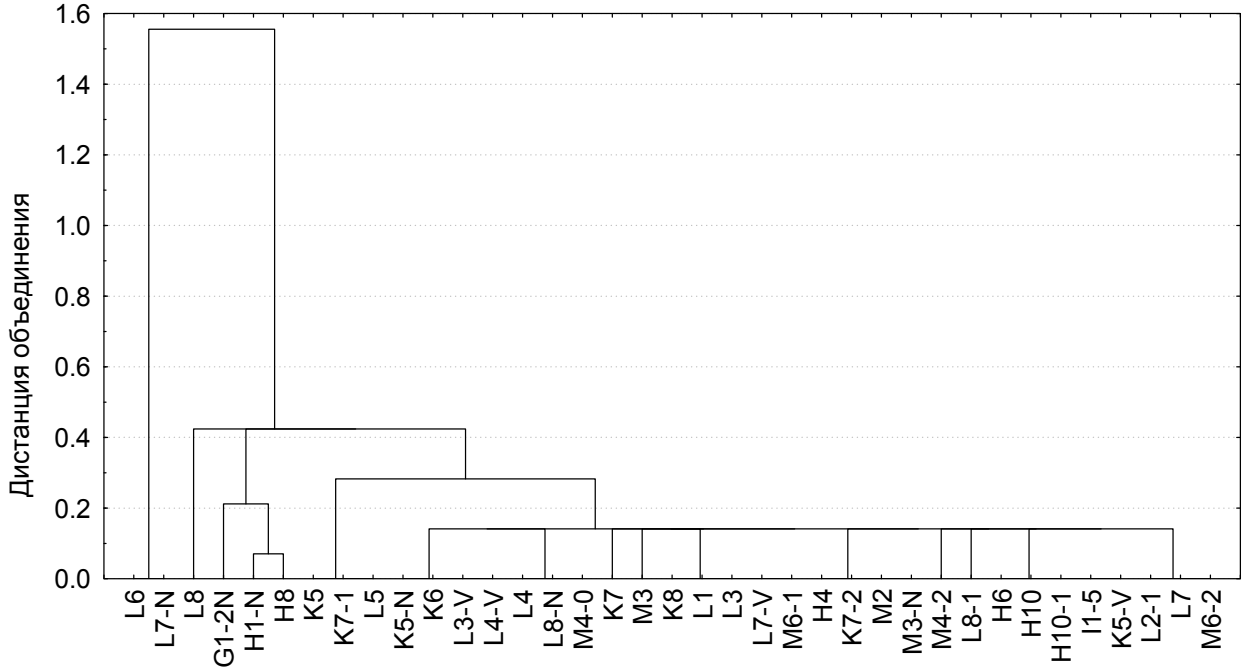

Рис. 2. Дендрограмма результатов кластеризации методом одиночной связи угольных пластов Красноармейского геолого-промышленного района по содержанию Be в угле

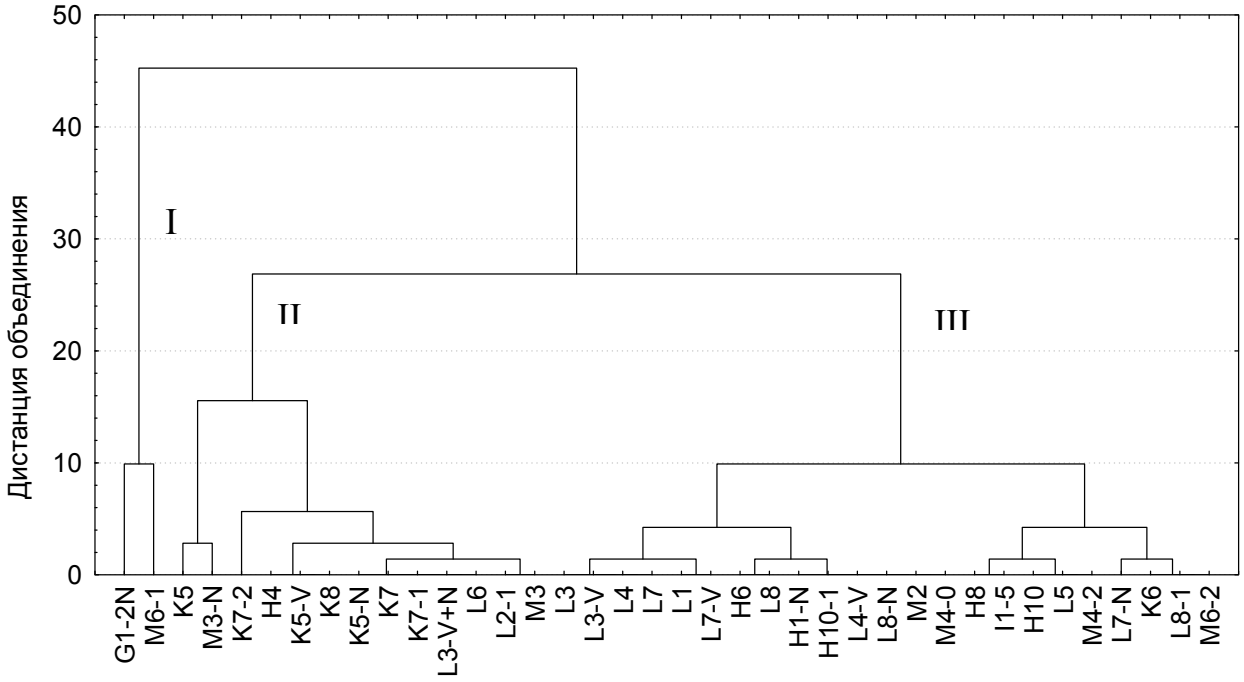

Рис. 3. Дендрограмма результатов кластеризации методом полных связей угольных пластов Красноармейского геолого-промышленного района по содержанию Ni в угле

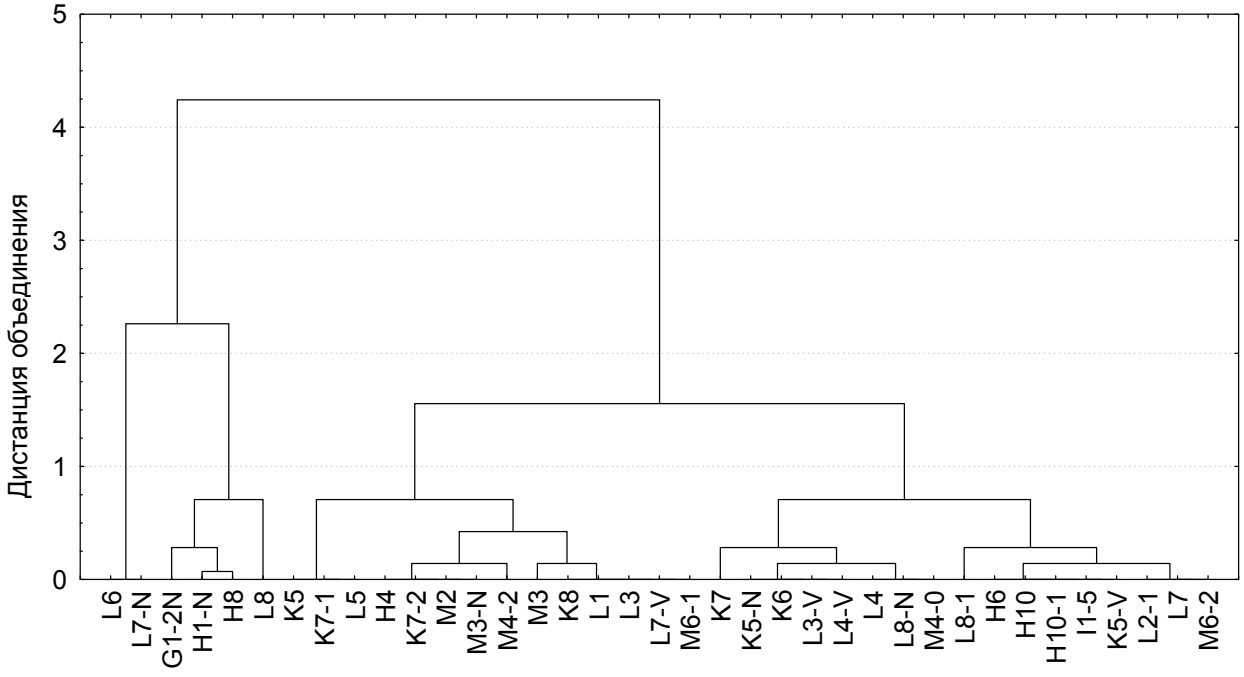

Рис. 4. Дендрограмма результатов кластеризации методом полных связей угольных пластов Красноармейского геолого-промышленного района по содержанию Be в угле

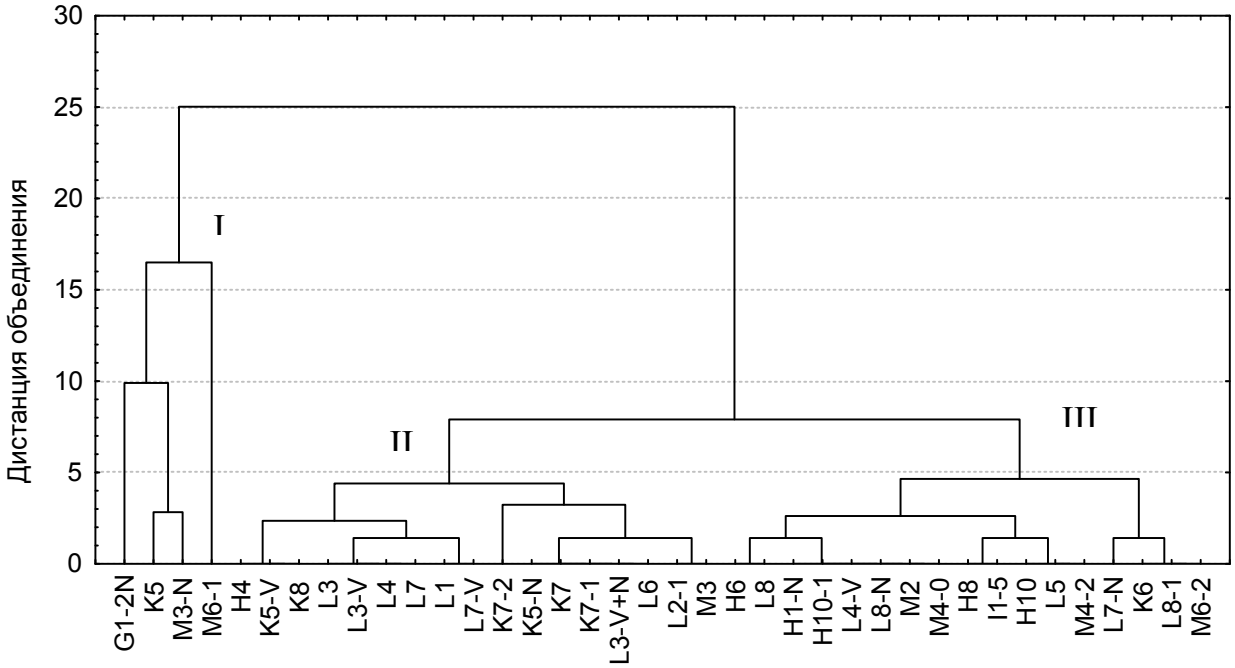

Рис. 5. Дендрограмма результатов кластеризации невзвешенным методом «средней связи» угольных пластов Красноармейского геолого-промышленного района по содержанию Ni в угле

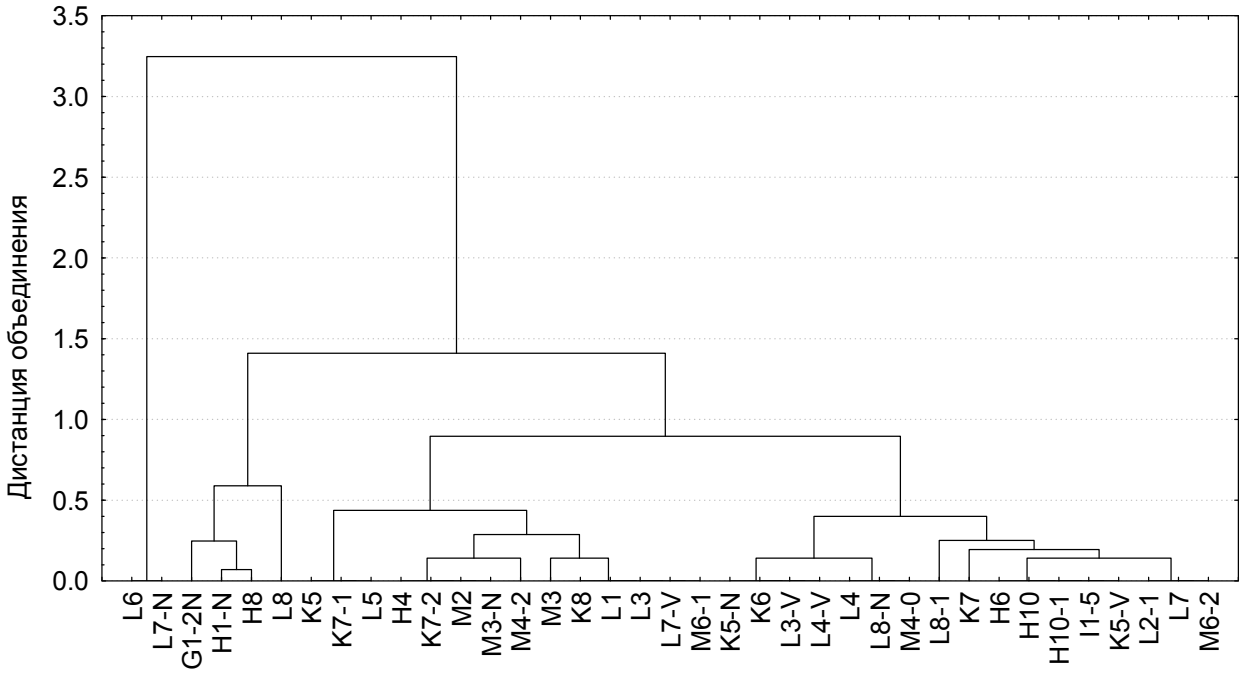

Рис. 6. Дендрограмма результатов кластеризации невзвешенным методом «средней связи» угольных пластов Красноармейского геолого-промышленного района по содержанию Be в угле

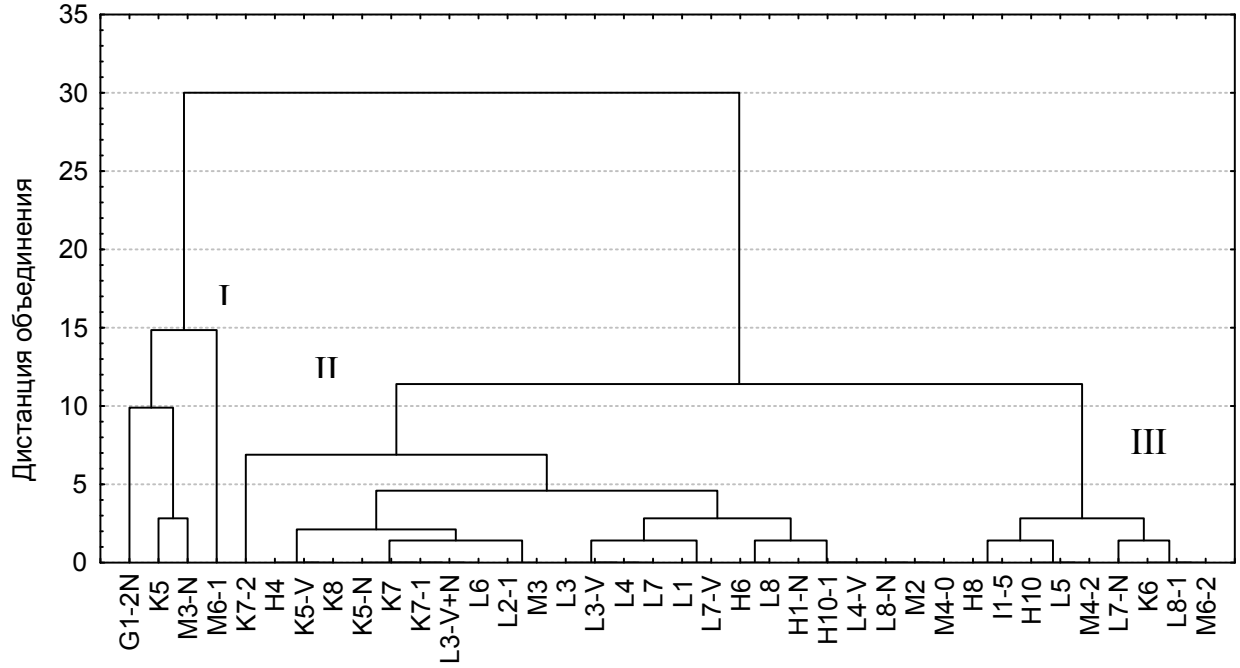

Рис. 7. Дендрограмма результатов кластеризации взвешенным методом «средней связи» угольных пластов Красноармейского геолого-промышленного района по содержанию Ni в угле

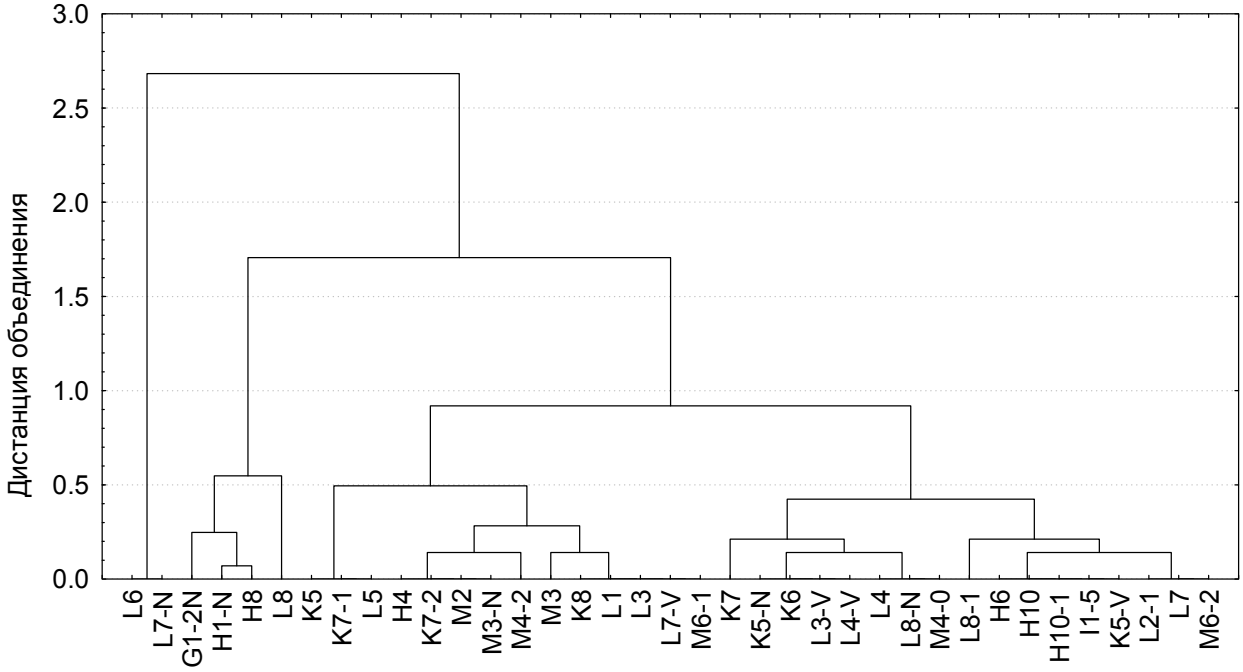

Рис. 8. Дендрограмма результатов кластеризации взвешенным методом «средней связи» угольных пластов Красноармейского геолого-промышленного района по содержанию Be в угле

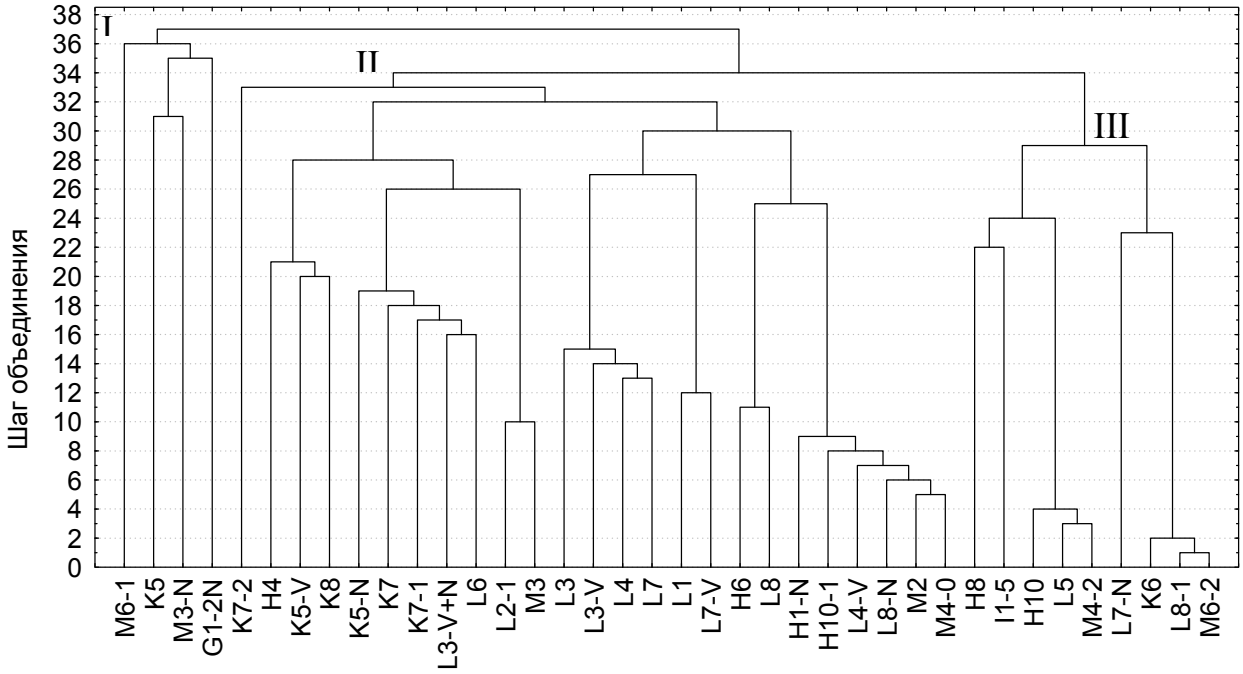

Рис. 9. Дендрограмма результатов кластеризации взвешенным центроидным методом угольных пластов Красноармейского геолого-промышленного района по содержанию Ni в угле

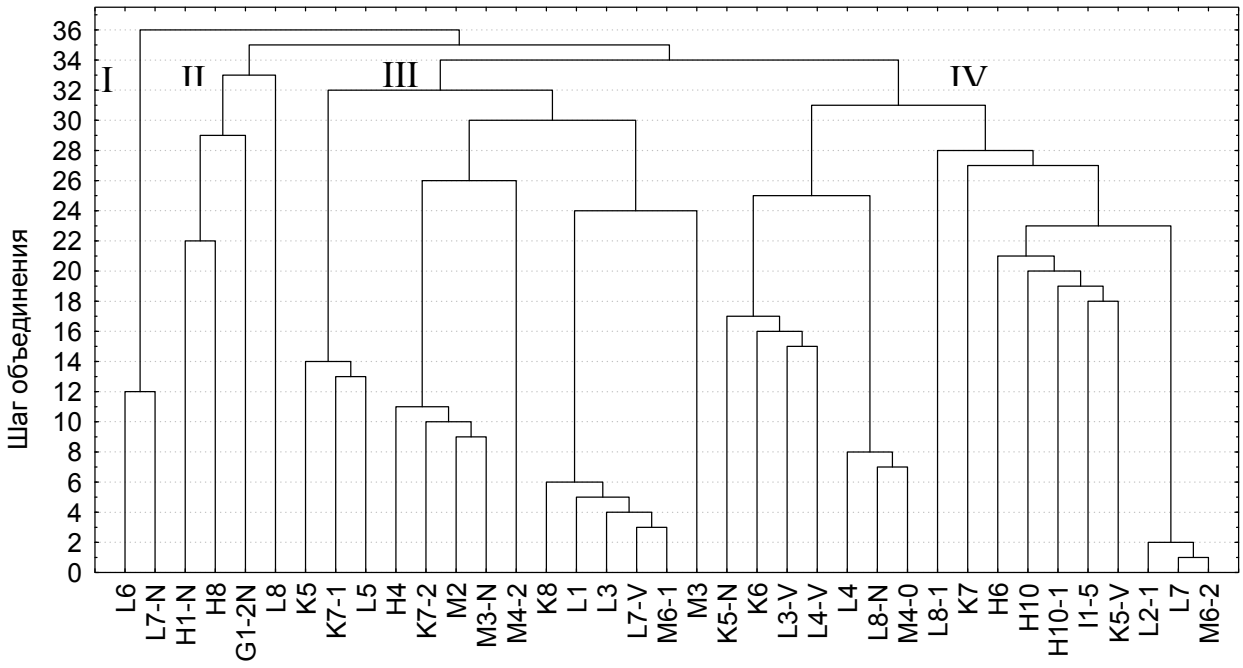

Рис. 10. Дендрограмма результатов кластеризации взвешенным центроидным методом угольных пластов Красноармейского геолого-промышленного района по содержанию Be в угле

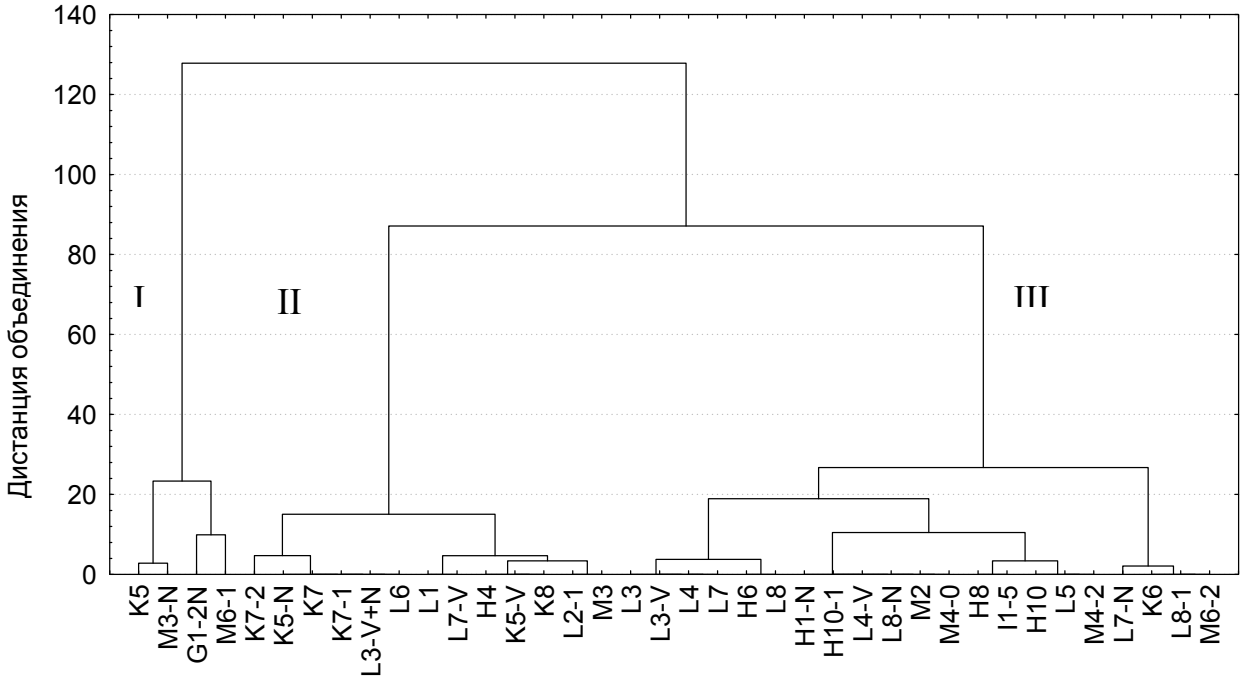

Рис. 11. Дендрограмма результатов кластеризации методом Уорда угольных пластов Красноармейского геолого-промышленного района по содержанию Ni в угле

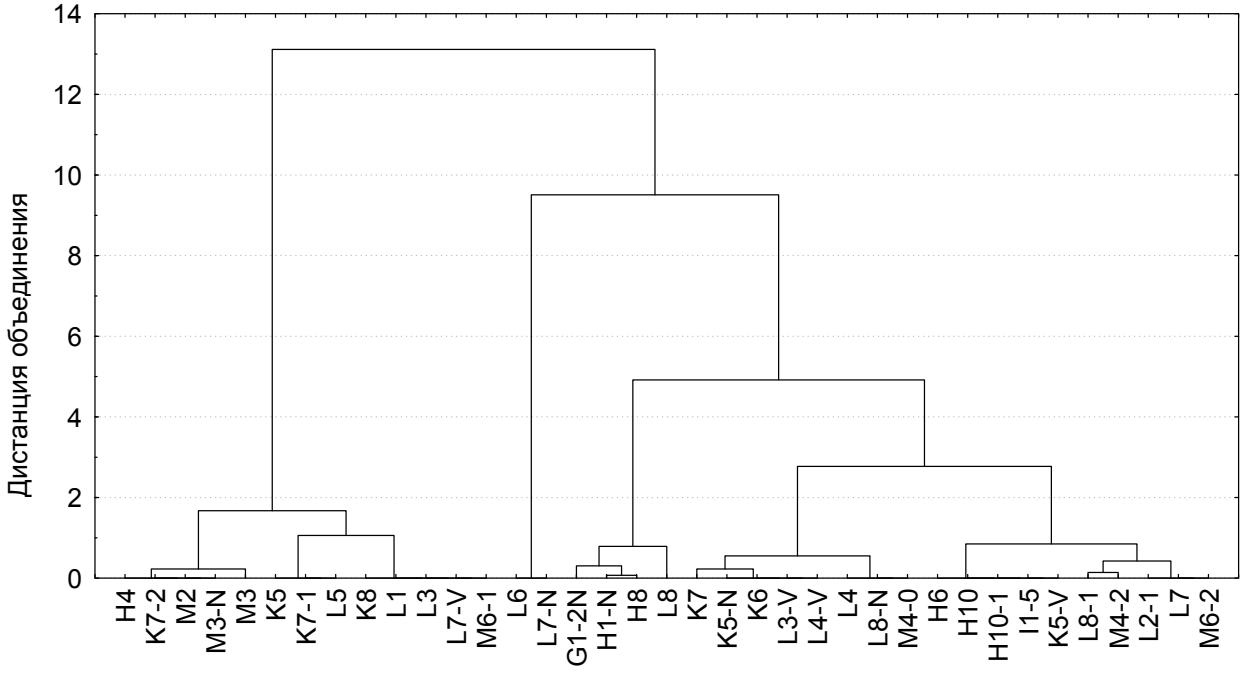

Рис. 12. Дендрограмма результатов кластеризации методом Уорда угольных пластов Красноармейского геолого-промышленного района по содержанию Be в угле

Обсуждение результатов. Результаты кластеризации угольных пластов Красноармейского геолого-промышленного района по содержанию в углях Ni и Be иерархическими агломеративными методами показали:

1. Использование метода одиночной связи приводит во всех случаях к результатам не позволяющим установить количество и структуру кластеров содержащихся в исходных данных. Следовательно, его применение в целях классификации угольных пластов по содержанию токсичных и потнциально токсичных элементов не рекомендуется.

2. Результаты кластеризации угольных пластов по содержанию Ni методами полной связи (относится к группе методов «расширяющих пространство»), взвешенным «средней связи» и взвешенным центроидным (оба принадлежат к группе методов оставляющих свойства исходного пространства без изменений) полностью идентичны. Во всех случаях в результате сформировано три существенно отличающихся по своим размерам кластера: I – объединяет четыре пласта с аномально высокими концентрациями: 42-27г/т ($g_1^{2н}$, k_5, $m_3^{н}$ и m_6^{1}), II – содержит двадцать пять пластов со средними концентрациями: 22-14г/т ($k_5^{н}$, $k_5^{в}$, k_7, k_7^{1}, k_7^{2}, k_8, l_1, l_2^{1}, l_3, $l_3^{в+н}$, $l_3^{в}$, l_4, $l_4^{в}$, l_6, l_7, $l_7^{в}$, l_8, $l_8^{н}$, m_2, m_3, m_4^{0}, h_4, h_6, $h_1^{н}$ и h_{10}^{1}) и III – включает девять пластов с минимальными содержаниями 13-10г/т (k_6, l_5, l_5^{1}, $l_7^{н}$, l_8^{1}, m_4^{2}, m_6^{2}, h_8 и h_{10}). В то же время использование взвешенного центроидного метода для визуальной оценки количества и структуры кластеров наиболе эффективно.

3. Кластеризация методами Уорда (относится к группе методов «расширяющих пространство») и невзвешенной «средней связи» (принадлежит к группе методов не меняющих свойства исходного пространства) угольных плстов по содержанию Ni так же выявила в исходных данных три кластера. Если в обоих случаях первый кластер содержит четыре пласта с аномально высокими концентрациями: 42-27г/т ($g_1^{2н}$, k_5, $m_3^{н}$ и m_6^{1}), то состав и структура второго и третьего кластеров несколько различаются. При применение метода Уорда во второй кластер входят тринадцать пластов со средними концентрациями 22-17г/т (h_4, $k_5^{н}$, $k_5^{в}$, k_7, k_7^{1}, k_7^{2}, k_8, l_1, l_2^{1}, $l_3^{в+н}$, l_6, $l_7^{в}$, и m_3), а в третий – двадцать один пласт с содержаниями 16-10г/т (m_4^{0}, m_4^{2}, m_6^{2}, m_2, l_3, $l_3^{в}$, l_4, $l_4^{в}$, l_5, l_5^{1}, $l_7^{н}$, l_7, l_8, $l_8^{н}$, l_8^{1}, k_6, $h_1^{н}$, h_6, h_8, h_{10}, и h_{10}^{1}). При кластеризации методом невзвешенной «средней связи» второй клстер формируют семнадцать пластов с концентрациями 22-16г/т (h_4, $k_5^{н}$, $k_5^{в}$, k_7, k_7^{1}, k_7^{2}, k_8, l_1, l_2^{1}, $l_3^{в+н}$, l_3, $l_3^{в}$, l_4, l_6, l_7, $l_7^{в}$, и m_3), а третий – семнадцать пластов с содержаниями 15-10 г/т (m_4^{0}, m_4^{2}, m_6^{2}, m_2, $l_4^{в}$, l_5, l_5^{1}, $l_7^{н}$, l_8, $l_8^{н}$, l_8^{1}, k_6, $h_1^{н}$, h_6, h_8, h_{10}, и h_{10}^{1}). Сопоставление результатов кластеризации этими методами показывает, что использование метода невзвешенной «средней связи» чаще приводит к созданию кластеров близких размеров, чем метод Уорда.

4. При кластеризации угольных пластов по содержанию Be только взвешенный центроидный метод позволяет однозначно установить количество и выявить структуру конечных кластеров.

Выводы. Анализ результатов кластеризации угольных пластов в целях их классификации по содержанию токсичных и потенциально токсичных элементов различными методами, реализованными программе «STATISTICA 6.0»

свидетельствет, что наиболее эффективным является применение взвешенного центроидного метода. Его использование позволяет не только установить количество результирующих кластеров, но и выявить их структуру. Дальнейшие исследования результатов кластеризации необходимо сосредоточить на их интерпретации в геологических понятиях.

Список литературы

1. Боровиков В.П.. STATISTICA: искусство анализа данных на компьютере. Для профессионалов. – СПб. Питер, 2001. – 658 с.
2. Sneath P. The application of computers to taxonomy // Journal of General Microbiology. – 1957. - №17. – P.201-226.
3. Sokal R., Michener C.D. A statistical method for evaluating systematic relationships // University of Kansas Scientific Bulletin. – 1958. - №38. P.1409-1438.
4. Ward J. Hierarchical grouping to optimize an objective function // Journal of the American Statistical Association. – 1963. - №58. P.236-244.

Рекомендовано до публикации д.г.н. Приходченком В.Ф.
Надійшла до редакції 20.09.2014