ՀԱՅԿԱԿԱՆ ՍՍՀ ԳԻՏՈՒԹՅՈՒՆՆԵՐԻ ԱԿԱԴԵՄԻԱՅԻ ԶԵԿՈՒՅՑՆԵՐ
ДОКЛАДЫ АКАДЕМИИ НАУК АРМЯНСКОЙ ССР

L 1970 4

УДК 551.49.08

ГИДРОГЕОЛОГИЯ

Д. Г. Асатрян, Э. С. Халатян

Опыт применения алгоритмов классификации при гидрогеохимическом прогнозировании

(Представлено чл.-корр. АН Арм. ССР А. Т. Асланяном 21/VII 1969)

В гидрогеохимии особый интерес представляет задача классификации вод по концентрации микроэлементов. Эта задача сводится, как правило, к нахождению порогов для фоновых и аномальных содержаний микроэлемента. После того, как эти пороги выбраны, отнесение каждого источника к тому или иному классу тривиально, если содержание данного микроэлемента в воде известно. Однако весьма частой является ситуация, когда это содержание не определено и возникает задача отнесения источника к тому или иному классу по косвенным признакам. Основанием для такого подхода служит тот факт, что содержание отдельных микроэлементов зависит от физико-химической характеристики воды, определяемой макрокомпонентами. При этом может оказаться, что заранее определяя классы вод, выбрав пороги для фоновых и аномальных значений макроэлемента, не удастся хорошо различать эти классы по составу макроэлементов. Поэтому мы попытались так определить пороги для фоновых и аномальных содержаний микроэлемента, чтобы получившиеся классы хорошо разделились по комплексу косвенных показателей (содержанию макрокомпонентов) и вместе с тем соответствовали существу задачи.

В настоящем сообщении впервые сделана попытка по анализам химического состава минеральных вод Армении выделить классы с низкими, средними и сравнительно высокими относительными содержаниями бора. При этом нами использовались два различных вида описания химического состава вод, основанные на: 1) величинах содержаний макроэлементов, в *мг/экв*, 2) процентных содержаниях макроэлементов к величине общей минерализации.

Рассмотрим сначала задачу выделения двух классов относительных содержаний бора (т. е. величин $B \cdot 10^4/M$), имея информацию о содержании макрокомпонентов $(Na+K)$, Mg, Ca, Cl', SO_4'' и HCO_3' в *мг/экв*.

Всего рассматривалось множество из 169 источников, которое нужно было разбить на два подмножества соответственно с низким и высоким содержанием бора.

Первый вид описания, часто используемый при гидрогеохимических исследованиях, основан на упорядочении величин макрокомпонентов следующим образом. Пронумеруем макрокомпоненты в порядке, в котором они написаны выше, числами $1, \cdots, 6$. Каждому источнику поставим в соответствие перестановку $i_1, i_2, \cdots, i_6$ из чисел $1, 2, \cdots, 6$, у которой первые три места (i_1, i_2, i_3) отведены убывающему ряду катионов, а последние три места—(i_4, i_5, i_6)—убывающему ряду анионов. Например, описанием источника, у которого катионы упорядочены в убывающем порядке эквивалент-процентных содержаний как (Na + K), Ca, Mg, а анионы—HCO_3, Cl′, SO_4, является перестановка 132645 *.

Наиболее просто задачу можно было бы решить следующим образом. Перестановку π будем считать описанием источника первого класса (с низким содержанием бора) при данном b, если большинство источников, описываемых перестановкой π имеют концентрацию бора, не превосходящую b. Остальные описания относятся ко второму классу. b называется пороговым значением для бора.

Таким образом, с помощью перестановок можно классифицировать источники по содержанию бора. Если π—перестановка, описывающая источник из первого класса, то те источники, которые описываются перестановкой π, но у них содержание бора больше b, будем называть ошибочно классифицированными. Аналогично определяются ошибочно классифицированные источники для второго класса. Таким образом, с порогом b связано некоторое решающее правило, разбивающее множество описаний источников Π на два подмножества Π_1 и Π_2, причем $\Pi_1 + \Pi_2 = \Pi$. Однако при таком подходе мы не сможем классифицировать источники, описания которых не встречались ранее.

Другой подход к решению задачи связан с введением в пространстве описаний расстояния. Естественно определить расстояние $\rho(\pi_i, \pi_j)$ между перестановками $\pi_i = (i_1, \cdots, i_6)$ и $\pi_j = (j_1, \cdots, j_6)$ как минимальное число транспозиций вида $(i, i+1)$, с помощью которых перестановка π_i переходит в перестановку π_j, а в качестве меры различимости классов рассматривать величину

$$R(b) = (\rho_{11} + \rho_{22})/2\rho_{12}, \qquad (1)$$

где

$$\rho_{kk} = \frac{1}{m_k^2} \sum_{\pi_i \in \Pi_k} \sum_{\pi_j \in \Pi_k} \rho(\pi_i, \pi_j) \quad (k = 1, 2)$$

$$\rho_{12} = \rho_{21} = \frac{1}{m_1 m_2} \sum_{\pi_i \in \Pi_1} \sum_{\pi_j \in \Pi_2} \rho(\pi_i, \pi_j).$$

---

* Очевидно, указанный способ описания является одним из вариантов описания минеральных вод по Курлову.

Здесь m_k — число элементов множества Π_k, ρ_{kk} — среднее расстояние между описаниями источников, входящих в k-й класс, ρ_{12} — среднее расстояние между классами. Мы должны выбрать θ^* так, чтобы $R(\theta^*)$ была минимальной (1), т. е. определить разбиение на классы таким образом, чтобы они оказались наиболее различимыми. Порог θ^* можно найти прямым перебором, проверяя лишь значения θ, совпадающие со значениями бора, которые встречаются в материале.

При прогнозировании источник, описываемый перестановкой π, будем относить к k-му классу, если

$$\frac{1}{m_k} \sum_{\pi_j \in \Pi_k} \rho(\pi, \pi_j) \leqslant \frac{1}{m_{3-k}} \sum_{\pi_j \in \Pi_{3-k}} \rho(\pi, \pi_j) \quad (k = 1, 2) \qquad (2)$$

Материал обучения состоял из 100 источников. Минимальное значение $R(\theta)$, вычисленное по формулам (1), достигалось при $\theta = \theta^* = 7$. Применение решающего правила (2) ко всем 169 источникам дало 41 ошибку, что составляет 24,3% *.

Столь высокий процент ошибок при прогнозировании, по всей вероятности, обусловлен грубостью способа описания источников, при котором теряется некоторая часть информации, доставляемой химическим составом воды. При выбранном способе описания близкими оказываются не только источники, у которых близки концентрации всех макрокомпонентов, но и источники, у которых сохраняются соотношения между анионами или катионами, хотя общая концентрация тех или других резко отлична. При таком способе описания не учитываются также весовые соотношения между анионами и катионами.

Несколько расширим теперь описание минеральных вод, приведенное выше. Рассмотрим перестановку $i_1, i_2, \ldots, i_6$, соответствующую последовательности номеров мест, занимаемых макрокомпонентами в убывающем ряду их значений в *мг/экв*. Применив критерий (1) и решающее правило (2) к тем же источникам, но при расширенном описании, получили вновь $\theta^* = 7$, но в этом случае количество ошибок на всем материале составило около 18%.

Анализ описаний источников в полученных классах показывает, что вряд ли удастся существенно улучшить этот результат путем применения других решающих правил. Действительно, большинство ошибок приходится на источники, описание которых совпадает для нескольких источников с разнообразным содержанием бора. Поэтому при принятом описании будут существовать источники, классифицированные неправильно при любом пороге и любом решающем правиле. Кстати, отметим, что число ошибочно классифицированных источников достигает минимума также при $\theta^* = 7$.

Решим теперь эту же задачу, используя второй вид описания, когда заданы процентные отношения макрокомпонентов к общей ми-

* Фактически % ошибок может оказаться заниженным, так как часть информации использовалась для обучения.

нерализации, прибавив к описанию и минерализацию. Как известно, бор относится к накапливаемым элементам и содержание его зависит от состава и величины общей минерализации воды (*). Нам представляется, что этот вид описания предпочтительнее при выделении аномалий.

Расположим все n источников в порядке возрастания бора. Пусть x_k^j — концентрация j-го компонента k-го источника ($k = 1, 2, \ldots, n$). Рассмотрим те компоненты, концентрация которых растет с ростом концентрации бора* и заменим числа x_k^j номерами r_k^j мест, занимаемыми ими в ряду $x_1^j \leqslant \ldots \leqslant x_n^j$. Например, источник с минимальной концентрацией j-го компонента получит номер 1, а источник с максимальной концентрацией — номер n, причем номера не повторяются. Для компонент, концентрация которых убывает с ростом концентрации бора, нумерация проводится в обратном порядке, т. е числа x_k^j заменяются номерами мест, занимаемыми ими в ряду $x_1^j \geqslant \ldots \geqslant x_n^j$. Обозначим через r_k суммы номеров, получаемых k-м источником, $r_k = \sum_{j=1}^{l} r_k^j$. Пусть $1 \leqslant \nu \leqslant n-1$, $i < \nu$, $m > \nu$, где ν — целое число. Введем следующую систему величин

$$Z_{im}(\nu) = \begin{cases} 1, & \text{если } r_i > r_m \\ 0, & \text{если } r_i < r_m \end{cases}$$

и обозначим

$$Z(\nu) = \frac{1}{\nu(n-\nu)} \sum_{i=1}^{\nu} \sum_{m=\nu+1}^{n} Z_{im}(\nu).$$

В качестве меры статистической различимости классов, определяемых числом ν (или, что то же самое — порогом $\theta = B_\nu$), используется $Z(\nu)$, показывающая долю источников, расположенных не в „своем" классе при данном разбиении. При правильном выборе ν должна наблюдаться максимальная упорядоченность относительного расположения классов(4)**. Поэтому нужно найти значение ν^*, при котором $Z(\nu)$ принимает минимальное значение и положить $\theta^* = B_{\nu^*}$.

При прогнозировании применяем следующее решающее правило: имеем значения семи параметров $x^1, \ldots, x^7$, и желаем узнать в какой класс попадает соответствующий источник. Для этого x^j заменяем номером r_k^j значения x_k^j j-го компонента того источника из обучающего множества, для которого $|x^j - x_k^j|$ минимальна. Вычисляем

---

* Характер зависимости бора от компонент макросостава можно установить, например, по знаку коэффициента ранговой корреляции.

** Идея нахождения разделяющей границы использованием свойств упорядоченности классов принадлежит Ш. А. Губерману (3).

$\sum_{i=1}^{3} r_i^2$. Если $r^2 < r_*^2$, то источник малобороносный, в противном случае—высокобороносный

Вычисления проводились на материале, составленном из исходного описания тех же 169 источников. Для обучения была сделана случайная выборка из 60 источников. Вычислено $Z(*)$ и получено $b^* = 18$, $\theta^* = 7$. Число ошибок на всем материале составило меньше 12 %. Значительная доля ошибок приходилась на источники с содержанием бора, близким к пороговому. В частности, число ошибок для источников с содержанием бора от 5,0 до 12,0 составило 5. Возможно, что при более точных определениях бора часть ошибок устранится. Другая часть ошибок относилась к источникам, которые попадали в ошибку и при других описаниях.

Прилагаемые графики иллюстрируют существенное смещение распределений отдельных макроэлементов внутри разных классов (исключая сульфат-ион), что подтверждает содержательность проведенного нами разбиения на классы (рис. 1).

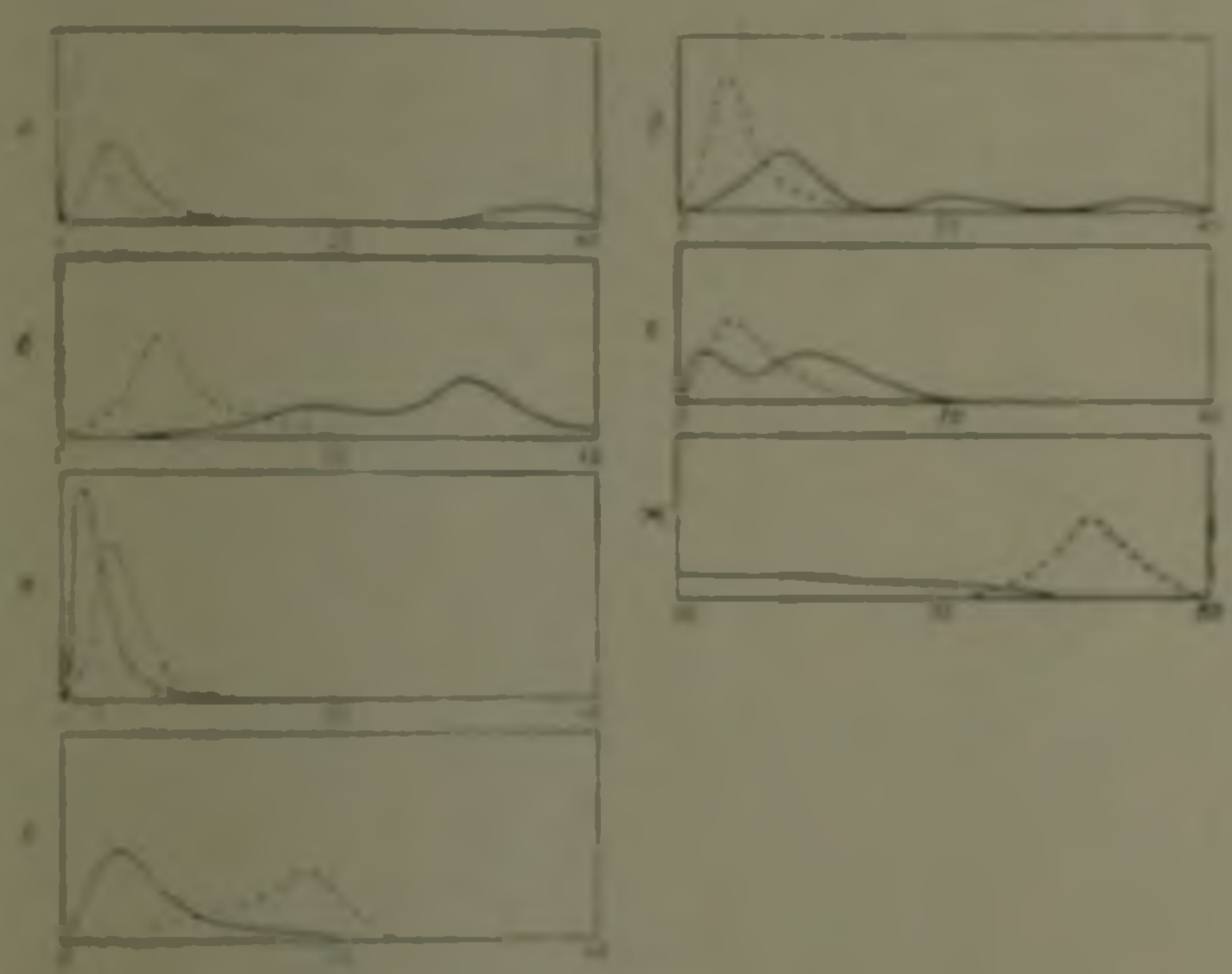

Рис. 1. Распределение макроэлементов в классах малобороносных источников (пунктир) и высокобороносных источников (сплошная линия): а—общая минерализация, б—натрий+калий, в—магний, г—кальций, д—хлор, е—сульфат-ион; ж—гидрокарбонат.

В заключение отметим, что порог b^* получился одинаковым вне зависимости от типа описания источников, в то время как для прогнозирования наиболее предпочтительным является второй тип описания и последний алгоритм. С их помощью нами определен второй

порог, который оказался равным 26. Общее количество ошибок при прогнозировании составило около 26 %.

Институт геологических наук
Академии наук Армянской ССР

Գ. Գ. ԱՍՈՏՅԱՆ, Է. Ս. ԽԱԼԱԹՅԱՆ

Հիդրոքիմիական կանխագուշակման մեջ դասակարգման ալգորիթմների կիրառման փորձ

Հոդվածում դիտարկվում է հանքային ջրերի մակրոտարրային բաղադրության հիման վրա հիդրոտարրերի քանակության փոփոխման տիրույթի՝ փոքր, միջին և համեմատաբար մեծ քանակությունների բաժանման խնդիրը, որը դրվում է հիդրոգեոքիմիական կանխագուշակման նպատակով։

Այն լուծելու նպատակով դիտարկվում է հանքային ջրերի քիմիական բաղադրության նկարագրման առավել տարածված երկու տարբերակ։ Ցույց է տրվում, որ գերդասական է երկրորդ տարբերակը և դրան համապատասխանող դասակարգման և կանխագուշակման ալգորիթմը։ Բերվում են խնդրի լուծումը բնորոշող քվանտ տվյալներ։

ЛИТЕРАТУРА — ԳՐԱԿԱՆՈՒԹՅՈՒՆ

[1] *Г. С. Себастиан*, Процессы принятия решений при распознавании образов. „Техника". Киев, 1965. [2] *С. Р. Крайнов*. Гидрогеохимический метод поисков месторождений бора. „Недра", М., 1964. [3] *Э. С. Халитян*. „Известии АН Арм ССР". Науки о Земле, т. XVIII, № 6 (1965). [4] *Д. Г. Агатрян*, Труды I конференции молодых специалистов ВЦАН Арм. ССР и ГрГУ, т. I, Ереван, 1969. [5] *Ш. А. Губерман, Э. М. Сидельникова, И. М. Чурикова*. в сб. Применение математических методов в геологии, Алма-Ата, 1968.