Геохимия

УДК 550.4.01:519.233.5

КОРРЕЛЯЦИОННЫЙ АНАЛИЗ В ГЕОХИМИИ: ТРИ ПРОБЛЕМЫ ПРАКТИЧЕСКОГО ПРИМЕНЕНИЯ

Ю.А. Ткачев

Институт геологии Коми НЦ УрО РАН, г. Сыктывкар
E-mail: Tkachev@geo.komisc.ru

Анализируются типичные ошибки применения корреляционно-регрессионного анализа в геохимии и получаемые в результате ложные выводы, имеющие характер артефактов. По каждому виду ошибок дан теоретический анализ, затем результаты компьютерного моделирования ситуации, порождающей ошибку, в сравнении с правильными результатами. Наконец, как типичный пример проанализирована одна работа (докторская диссертация), содержащая все три вида рассмотренных ошибок.

Введение

Трудами корифеев математической статистики корреляционно-регрессионный анализ был внедрён в биологию (в особенности – в агробиологию). В геологии и геохимии внедрение и широкое использование его связано с именами А.Б. Вистелиуса [1], Ф. Чейза [2, 3]. Однако практика применения корреляционного анализа в геологии и геохимии осложняется несколькими обстоятельствами:

а) геологи не всегда правильно понимают суть математических методов, о чём давно писали упомянутые выше авторы; это обстоятельство субъективное;

б) идентификация однородных, сопоставимых геологических объектов, которые составили бы вполне определенную генеральную совокупность или выборку из неё, является не тривиальной задачей;

в) в применении корреляционно-регрессионного анализа и его интерпретации имеется ряд нерешенных проблем.

Два последних обстоятельства – объективные. Всё это, вместе взятое, приводит к серьезным ошибкам, которые можно встретить в опубликованных работах, (например, [4]).

При значительной разработанности теоретической стороны методов корреляционно-регрессионного анализа, в их применении имеется большой пробел, связанный с интерпретацией результатов, в зависимости от того, что из себя представляет совокупность исходных данных. В частности, каков будет результат корреляционно-регрессионного анализа в подсовокупности исходной совокупности, в которую мы отобрали пробы (элементы совокупности) по какому-либо признаку, например, по значениям X и/или Y, как теперь интерпретировать полученные результаты, что в них содержательного, а что является *артефактами* (лат. *arte* искусственно + *factus* сделанный), вольно или невольно порожденными процедурно-математическими свойствами корреляционно-регрессионного анализа? Эта проблема весьма актуальна, так как геологи и геохимики часто меняют состав изучаемых генеральных совокупностей и способ получения выборок из них, не всегда отдавая себе отчет в том, правильно ли они потом интерпретируют получаемые результаты.

Статья посвящена исследованию аналитическими методами и методами компьютерного моделирования следующих вопросов:

1. Как изменится коэффициент корреляции и параметры уравнения регрессии в выборке, образованной из исходной совокупности селекцией проб:
   а) с помощью случайного выбора;
   б) с помощью неслучайного выбора – в заданном интервале содержаний одного из двух коррелирующих компонентов (например, селекцией проб, содержание компонента в которых или выше заданного, или ниже заданного, или близко к среднему);
   в) в заданном интервале (как в п. «б») суммы содержаний обоих коррелирующих компонентов.
2. Каким будет коэффициент корреляции и параметры уравнения регрессии, если коррелировать среднее содержание какого-либо компонента с его средним квадратическим отклонением? Как эти параметры изменятся, если распределение компонента подчиняется:
   а) нормальному (гауссовскому) закону;
   б) логарифмически-нормальному закону.

3. Как будет изменяться коэффициент корреляции и уравнение регрессии между двумя положительно или отрицательно коррелирующими или совсем не коррелирующими величинами в выборках, элементы которых содержат ряд других компонентов, находящихся между собой в зависимостях, определяемых наперёд заданной корреляционной матрицей. Как будет изменяться коэффициент корреляции при изменении содержаний этих других («фоновых») компонентов?

Итак, первая проблема математически формируется следующим образом: что произойдет с коэффициентом корреляции, если из исходной генеральной совокупности объема N производить *неслучайные выборки* объема n. Останется ли коэффициент корреляции r_i в этих выборках неизменным (колеблясь около истинного коэффициента корреляции в генеральной совокупности ρ случайным образом), или же r_i подвергнется систематическому изменению.

Вторая (извечная) проблема связана с тем, что большая часть обрабатываемых данных является результатами анализов, то есть процентными величинами – величинами с постоянной суммой, равной 100 %. Как будет изменяться коэффициент корреляции между какими-нибудь компонентами, если их доля в пробах будет увеличиваться на фоне уменьшения доли других компонентов?

Наконец, третья проблема связана с интерпретацией коэффициента корреляции между параметрами распределения одной случайной величины, а именно – между средним значением X и его средним квадратическим отклонением.

Методика компьютерного моделирования

Основой компьютерного моделирования в рассматриваемом случае является получение совокупности n проб, содержащих m компонентов с заданными средними, средними квадратическими отклонениями и с заданной корреляционной матрицей. Никакой проблемы не возникает, если компонентов два: n=2, x и y. Тогда с помощью датчика псевдослучайных чисел генерируется три нормально (или логнормально) распределенных независимых случайных числа: u, ε_1 и ε_2. С их помощью образуется пара чисел

$$x=u+\varepsilon_1, \qquad (1)$$

$$y=u+\varepsilon_2. \qquad (2)$$

Коэффициент корреляции r_{xy} в этом случае будет, согласно [5], равен

$$r_{xy}=\sum_{i=1}^{2}\sum_{j=1}^{2} r_{ij}k_{xi}k_{yi},$$

где r_{ij} – коэффициент корреляции между слагаемыми, составляющими x и y; в нашем случае

$$r_{uu}=1,\quad r_{u\varepsilon_1}=0,\quad r_{u\varepsilon_2}=0,\quad r_{\varepsilon_1\varepsilon_2}=0;$$

$$k_{x_i}=\frac{\sigma_{x_i}}{\sigma_x},\quad k_{y_i}=\frac{\sigma_{y_i}}{\sigma_y},$$

откуда $r_{xy}=r_{uu}\cdot\frac{\sigma_u}{\sigma_x}\cdot\frac{\sigma_u}{\sigma_y}$. Остальные слагаемые равны нулю, т. к. равны нулю коэффициенты корреляции между ними. Далее

$$r_{xy}=1\cdot\frac{\sigma_u^2}{\sigma_x\sigma_y},$$

где $\sigma_x^2=\sigma_u^2+\sigma_{\varepsilon_1}^2$, $\sigma_y^2=\sigma_u^2+\sigma_{\varepsilon_2}^2$.

Если мы моделируем величины с нулевыми средними и единичными дисперсиями, то

$$\sigma_u^2+\sigma_{\varepsilon_1}^2=1 \quad \text{и} \quad \sigma_u^2+\sigma_{\varepsilon_2}^2=1,$$

притом $\varepsilon_1=\varepsilon_2=\varepsilon$, откуда $\sigma_u^2+\sigma_\varepsilon^2=1$, $\sigma_\varepsilon^2=1-\sigma_u^2$.

Таким образом, для моделирования двух случайных величин x и y необходимо получить три случайные величины u, ε_1, ε_2 с дисперсиями σ_u^2, σ_ε^2, σ_ε^2 соответственно.

Описанное получение пары случайных величин x_i, y_i производится n раз (по числу проб). Затем средние и дисперсии преобразуются в величины, распределенные со средними m_x, m_y и средними квадратическими отклонениями s_x и s_y, по формулам

$$x_i'=x_i\cdot s_x+m_x,$$

$$y_i'=y_i\cdot s_y+m_y.$$

Если требуется моделировать *содержания*, т. е. величины *положительные*, то m_x и m_y следует выбирать достаточно большими, например $m_x>3s_x$, $m_y>3s_y$, и цензурировать моделируемые величины по условию $x_i'>0$, $y_i'>0$, т. е. отбрасывать отрицательные величины. При $m_x>3s_x$, $m_y>3s_x$ таких величин будет сравнительно немного, 0,13 %.

Замоделировать совокупность с m произвольно коррелирующими компонентами значительно сложнее. *Не всякая придуманная корреляционная матрица является непротиворечивой.* Уже при трех переменных два коэффициента корреляции могут быть выбраны произвольно, а третий предопределен, точнее – ограничен интервалом, который сужается по мере увеличения абсолютных значений двух первых коэффициентов.

При большом числе компонентов дело значительно усложняется. Произвольно составленная корреляционная матрица имеет весьма малую вероятность быть непротиворечивой. В связи с этим для решения некоторых задач (как наша) компьютерным моделированием лучше выбрать за основу какую-либо реальную корреляционную матрицу и определять остаточные дисперсии, т. е. дисперсии некоррелирующих слагаемых типа ε_1, ε_2 в равенствах (1), (2) решением системы уравнений

$$r_{x_tx_q}=\sum_{i=1}^{n}\sum_{j=1}^{m} r_{ij}k_{x_i}k_{y_j},\quad (t,q=1...m)$$

относительно k_x, k_y при r_{ij}=1 для общего слагаемого суммы и r_{ij}=0 – для остальных.

Полученная исходная двух- или многокомпонентная совокупность при этом еще не будет зам-

кнутой системой процентных величин. При необходимости анализа замкнутой системы процентных величин она преобразуется в таковую простым пересчетом так, чтобы сумма компонентов в каждой пробе составляла 100 %. Далее из этой выборки формируются подвыборки по одной из намеченных выше схем и по этим подвыборкам стандартными процедурами корреляционно-регрессионного анализа определяются коэффициенты корреляции и параметры уравнений регрессии.

Результаты теоретического анализа и компьютерного статистического моделирования

1. *Изучение динамики r, a и b в **случайных** выборках из генеральной совокупности двумерных элементов* (*система не является замкнутой процентной системой*).

Очевидно, что и коэффициент корреляции *r*, и параметры уравнений регрессии *a* и *b* в таких выборках будут несмещенными оценками этих величин в генеральной совокупности. В соответствии с теорией, флуктуации значений *r, a,* и *b* будут тем больше, чем меньше объем выборки, а именно

$$s_\theta^2 = \sigma_\theta^2 \frac{n}{n_0},$$

где s_ε^2 – дисперсия какого-либо из перечисленных выше параметров θ в выборке объема n_0, σ_θ^2 – её значение в конечной генеральной совокупности объёма *n*.

Приведенные в табл. 1 результаты полностью следуют теории. Это обстоятельство настолько очевидно и было предсказуемо, что, можно сказать, этот эксперимент мы проводили скорее для того, чтобы убедиться (и убедить читателей) в корректности алгоритма и программы моделирования.

Таблица 1. *Иллюстрация постоянства коэффициента корреляции при формировании выборок случайным образом*

n	*r*	*a*	*b*	*n*	*r*	*a*	*b*
300	**0,86**	**0,81**	**–0,06**	**1000**	**0,50**	**0,50**	**0,01**
90	0,85	0,79	-0,06	300	0,44	0,45	0,03
72	0,84	0,75	-0,13	240	0,49	0,49	0,13
54	0,80	0,77	-0,05	180	0,50	0,49	0,01
36	0,76	0,72	-0,18	150	0,41	0,44	0,03
18	0,85	0,96	0,01	90	0,52	0,54	0,19
1000	0,79	0,81	0,00	1000	0,32	0,32	-0,01
300	0,80	0,80	-0,03	300	0,32	0,32	0,04
270	0,78	0,77	0,03	240	0,26	0,25	0,09
210	0,80	0,84	-0,08	180	0,29	0,28	0,05
150	0,78	0,85	0,08	150	0,38	0,38	-0,14
90	0,81	0,82	0,04	90	0,21	0,22	-0,11
300	0,57	0,57	-0,01	300	0,29	0,29	0,11
90	0,61	0,63	-0,14	90	0,40	0,37	0,10
81	0,42	0,42	0,04	72	0,44	0,42	0,03
63	0,61	0,61	-0,01	63	0,27	0,26	0,02
36	0,56	0,56	0,10	54	0,28	0,22	0,20
27	0,54	0,50	0,08	45	0,34	0,30	0,04

Первая строка каждой секции таблицы – заданная генеральная совокупность

2. *Изучение динамики r, a и b в **неслучайных** выборках из той же генеральной совокупности, что и в п.* 1.

В частности, в выборку отбирались: а) сначала все пробы, значения величины *x* в которых превышали некоторое установленное значение, б) затем все пробы, в которых значения *x* были ниже некоторого значения, в) пробы, значения *x* в которых заключены в некотором интервале вблизи среднего значения, г) наконец, пробы, значения суммы *x*+*y* в которых были выше, ниже или около середины заданных значений этой суммы в генеральной совокупности. Необходимость изучать неслучайные выборки по критерию суммы *x*+*y* возникла для того, чтобы в минимальной степени искажать корреляционный эллипс отсечением части проб: отсечение должно также производиться прямой, перпендикулярной длинной оси эллипса.

Во всех этих случаях по теории должно наблюдаться уменьшение коэффициента корреляции *r* и соответствующее уменьшение наклона графика уравнения регрессии (т. е. уменьшение *a* в уравнении *y*=*ax* +*b*). Это следует из того, что коэффициент корреляции можно выразить через дисперсии следующим образом:

$$r_{xy}^2 = 1 - \frac{\sigma_{остx}^2}{\sigma_x^2} = 1 - \frac{\sigma_{остy}^2}{\sigma_y^2},$$

где σ_x^2 – дисперсия величины *x* в совокупности (или в подсовокупности), т. е. величина, характеризующая «длину» корреляционного эллипса; $\sigma_{остx}^2$ – остаточная дисперсия величины *x*, т. е. величина, характеризующая «толщину» корреляционного эллипса; s_y^2, $\sigma_{остy}^2$ – те же параметры, но в отношении величины *y*.

Поскольку в любой выборке из заданного интервала значений *x* (или *y*) длина корреляционного эллипса уменьшается, а толщина – нет, то *r* будет уменьшаться (и даже менять знак с плюса на минус). Так как $a_{yx} = r\frac{\sigma_y}{\sigma_x}$, то с уменьшением *r* уменьшится и *a* – угловой коэффициент графика уравнения регрессии.

Моделирование выборок селекцией проб с высокими значениями величины *x* из генеральных совокупностей с различными исходными значениями r_{xy} подтвердило правильность теории (табл. 2). Оно наглядно показало, что уменьшение r_{xy} и a_{yx} является типичным артефактом, который нельзя интерпретировать в геохимическом или каком-либо другом содержательном смысле.

Исследование выборок, полученных селекцией из генеральной совокупности по сумме *x*+*y* показало ещё более яркую картину (см. табл. 3 – селекция по значениям *x*+*y*, близким к среднему их значению в совокупности табл. 4 – селекция по низким значениям суммы *x* и *y*).

3. *Исследование коэффициента корреляции между средним значением случайной величины и её средним квадратическим отклонением.*

Вообще говоря, теоретически в различных совокупностях между этими величинами не должно быть

никакой связи. Большие содержания могут слабо флуктуировать, а в других случаях наоборот – малые содержания могут варьировать значительнее. Но здесь речь пойдёт об одной совокупности заданным образом (нормально или логнормально) распределенных величин. Эта совокупность разбивается на ряд подсовокупностей по увеличению значения *x*. В каждой такой подсовокупности определяется среднее значение m_x и среднее квадратическое отклонение σ_x. Исследуется вопрос о том, каков будет коэффициент корреляции между m_x и σ_x и как он будет зависеть от вида распределения *x*. Рассмотрим нормальное распределение. Разделим трех-четырехсигмовый диапазон значений *x* на 9–11 равных интервалов. Средние значения переменной в каждом интервале близки к середине интервала (а в среднем интервале совпадает с ним). Дисперсии в интервалах, равно удаленных от среднего, равны, так как распределения в этих интервалах симметричны. Этого достаточно, чтобы убедиться в отсутствии корреляционной зависимости между m_x и σ_x: с увеличением m_x среднее квадратическое отклонение (относительно интервального среднего!) сначала увеличивается при движении к левому «односигмовому» интервалу, затем убывает к серединному интервалу, затем снова увеличивается к правому односигмовому интервалу и вновь уменьшается к правому хвосту распределения. При такой «раскладке» никакой корреляции между m_x и σ_x быть не может: σ_x при всех значениях m_x приблизительно сохраняет свое значение на уровне $\sqrt{\frac{h^2}{12}}$, где h – ширина интервала.

Таблица 2. *Динамика коэффициента корреляции по данным компьютерного моделирования выборок из верхней части совокупности 1000 проб*

n	*r*	*a*	*b*	*n*	*r*	*a*	*b*
1000	**0,90**	**0,90**	**1,04**	**1000**	**0,79**	**0,80**	**0,01**
300	0,55	0,31	0,86	300	0,47	0,61	0,24
240	0,47	0,22	1,05	270	0,53	0,70	0,11
180	0,37	0,15	1,22	210	0,49	0,64	0,21
150	0,21	0,08	1,35	180	0,46	0,58	0,30
90	0,20	0,04	1,51	120	0,44	0,51	0,43
1000	**0,80**	**0,82**	**–0,00**	**1000**	**0,49**	**0,48**	**–0,63**
300	0,48	0,61	0,24	300	0,52	1,11	-0,84
240	0,46	0,56	0,33	240	0,51	1,10	-0,84
180	0,47	0,59	0,28	150	0,53	1,14	-0,90
90	0,51	0,66	0,17	60	0,43	0,91	-0,46
30	0,23	0,28	0,92	30	0,33	0,59	0,16

Первая строка каждой секции таблицы – генеральная совокупность

При логарифмически-нормальном распределении это утверждение относится к логарифмам величин: среднему значению логарифма и логарифмической дисперсии, что при потенцировании приводит к сильнейшей корреляции между частным (интервальным) средним и антилогарифмом дисперсии (соответственно, и средним квадратическим). Это и понятно: равные в логарифмах интервалы при потенцировании сильно растягиваются при движении по числовой оси вправо, приводя к увеличению дисперсии. Этот факт, впрочем, был давно замечен как в отношении природной изменчивости [5], так и в отношении погрешностей измерений [6].

Таблица 3. *Динамика коэффициента корреляции по данным компьютерного моделирования из совокупности 1000 проб селекцией по сумме значений коррелирующих величин*

n	*r*	*a*	*b*	*n*	*r*	*a*	*b*
1000	**0,90**	**0,92**	**0,05**	**1000**	**0,45**	**0,46**	**–0,01**
300	0,28	0,29	0,05	300	-0,74	-0,84	-0,01
270	0,17	0,17	0,06	270	-0,78	-0,89	-0,01
240	0,08	0,08	0,06	240	-0,82	-0,91	-0,01
210	-0,06	-0,06	0,07	210	-0,86	-0,93	-0,01
180	-0,21	-0,22	0,08	180	-0,90	-0,99	-0,01
150	-0,38	-0,37	0,09	150	-0,93	-1,02	-0,01
120	-0,55	-0,49	0,09	120	-0,96	-1,01	-0,01
90	-0,74	-0,65	0,09	90	-0,97	-1,04	-0,01
60	-0,86	-0,85	0,10	60	-0,99	-1,04	-0,02
30	-0,96	-0,98	0,11	30	-1,00	-1,02	-0,02

Из средней части совокупности 1000 проб; первая строка каждой секции – таблицы генеральная совокупность

Таблица 4. *Динамика коэффициента корреляции по данным компьютерного моделирования выборок из совокупности 1000 проб селекцией по сумме значений коррелирующих величин*

n	*r*	*a*	*b*	*n*	*r*	*a*	*b*
1000	**0,90**	**0,92**	**0,05**	150	0,06	0,04	-1,34
300	0,34	0,26	-0,88	120	0,02	0,01	-1,41
270	0,28	0,20	-0,98	90	0,06	0,04	-1,39
240	0,25	0,17	-1,05	60	0,07	0,06	-1,43
210	0,18	0,12	-1,15	30	0,00	0,00	-1,61
180	0,12	0,08	-1,23				

Из нижней части совокупности 1000 проб; первая строка слева – генеральная совокупность

Моделирование показало, что при *нормальном* (гауссовском) распределении в полной согласии с теорией никакой зависимости между m_x и σ_x не наблюдается, тогда как при моделировании *логарифмически-нормального* распределения эта зависимость наблюдается, и притом сильная (*r* близок к единице):

Среднее	0,31	0,47	0,54	0,72	0,80	1,06	1,21	1,60	1,89	2,54	3,20	4,56	8,91
Ср. кв. откл.	0,02	0,02	0,03	0,02	0,03	0,04	0,05	0,09	0,07	0,15	0,25	0,57	3,03
Коэф. кор.	0,91												

Это наглядно видно и на рис. 1.

4. *Исследование динамики коэффициента корреляции в замкнутой системе процентных величин при увеличении содержаний двух избранных компонентов на фоне уменьшения содержаний остальных компонентов.*

Еще К. Пирсон [7] более ста лет назад указывал на появление ложной корреляции в системе процентных величин. Настоящий бум интереса к этой проблеме породили работы Ф. Чейза [2, 3], известного петрографа и минералога. Его исследования показали следующее: а) в закрытой системе из $n-1$ коэффициентов корреляции каждой строки корреляционной матрицы по крайней мере один отрица-

телен; б) из общего числа $\binom{n}{2}$ коэффициентов корреляции по крайней мере $n-1$ отрицательны; в) в трехкомпонентной системе, если никакая дисперсия не превышает суммы двух остальных, то все три r будут отрицательны; г) компонент с максимальной дисперсией имеет отрицательные коэффициенты корреляции по крайней мере с двумя из остальных компонентов. В дальнейшем эти результаты были развиты в работах [8–10], а также [11].

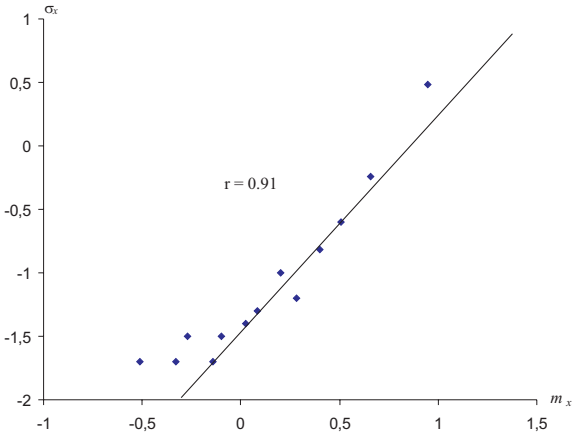

Рис. 1. *Зависимость между средним (m_x) и среднеквадратическим отклонением (σ_x) при моделировании логарифмически нормального распределения*

Если нас интересует динамика r только между двумя выделенными нами компонентами, то сумму всех оставшихся можно считать третьим компонентом. Если уменьшать содержание этого третьего (суммарного) компонента, то это автоматически приведет к уменьшению его дисперсии и, как следствие, к тому, что дисперсия одного из выделенных компонентов станет самой большой. По крайней мере с этого момента (но может, и значительно раньше) коэффициент корреляции между выделенными компонентами станет отрицательным.

Нам удалось найти наглядную графическую форму доказательства этого положения и неизбежности стремления коэффициента корреляции двух выделенных компонентов к значению –1 при уменьшении содержания остальных. Пусть между выделенными компонентами (например, Fe и Mn) существует положительная корреляционная зависимость. Тогда на графике (см. рис. 2) «облако» точек в координатах Fe – Mn будет представлять собой облако точек, вытянутое вправо-вверх. Проведем на этом графике дополнительную прямую, соединяющую точки (100 % Fe, 0 % Mn) и (0 % Fe, 100 % Mn).

При относительном увеличении содержания выделенных компонентов *без изменения их соотношения* (например, путем уменьшения суммы других (фоновых) компонентов) точки графика будут смещаться радиально от начала координат, приближаясь к прямой Fe+Mn=100 %. Скорость этого смещения (при одинаковом темпе увеличения содержания) будет уменьшаться, так что точки в конце концов лягут на прямую с координатами (100 % Fe, 0 % Mn), (0 % Fe, 100 % Mn), а это означает, что коэффициент корреляции между этими компонентами станет равным –1. При этом неважно, каким способом реализуется уменьшение содержания «фоновых» компонентов: специальной подборкой ли выборок из огромной природной генеральной совокупности объектов, химической ли обработкой проб одной и той же выборки (растворением карбонатной части проб, озолением проб углей, нефтей и горючих сланцев и т. д.). В зависимости от того, с какой скоростью относительно друг друга точки будут радиально удаляться от начала координат, изменение r может быть различным. Например, в процессе этого передвижения «положительный» эллипс может постепенно превратиться в «нейтральный» круг, а затем вытянуться длинной осью параллельно указанной прямой Fe+Mn=100 %, что будет соответствовать отрицательной корреляции, и, наконец, когда все точки лягут на эту прямую, коэффициент корреляции станет равным –1, угловой коэффициент прямой регрессии станет также равным –1. Такая модель реализуется в том наиболее вероятном случае, когда убывание фона будет пропорциональным его текущему содержанию. Может встретиться случай, когда убывание фона в абсолютных процентах одинаково во всех пробах. Тогда корреляционный эллипс сначала вытянется, и коэффициент корреляции увеличится, но по мере того, как всё большее число точек попадет на прямую Fe+Mn=100 %, коэффициент корреляции станет уменьшаться и быстро достигнет значения –1.

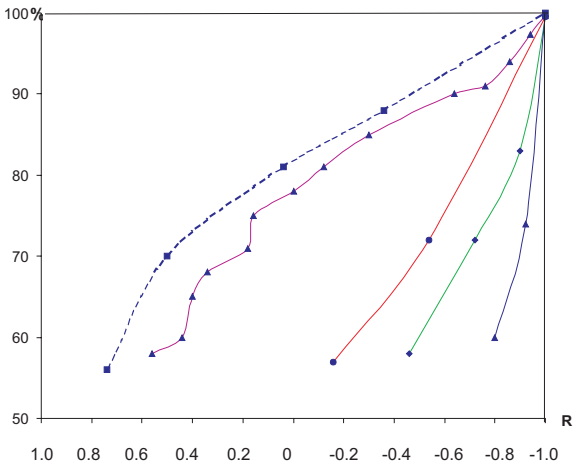

Рис. 2. *Радиальное от начала координат движение точек-проб при уменьшении содержания в них суммы «фоновых» компонентов. Квадратики – «начальное» положение точек, характеризуемое корреляционным эллипсом 1 (r=0,7); кружки – положение точек при уменьшении содержания фоновых компонентов в среднем с 45 до 25 %, корреляционный эллипс превратился в круг 2 (r=0,0)*

Единственный исключительный и маловероятный случай может возникнуть тогда, когда еще при малом содержании суммы выделенных компонентов коэффициент корреляции между ними был равен +1 при равенстве содержаний x и y. Тогда точки графи-

ка будут лежать в точности на прямой $x=y$, и при уменьшении содержания фоновых компонентов «стрела» точек будет перемещаться вдоль своего направления, и когда $x+y$ достигнет 100 %, все точки сольются в одну: вариации содержаний не будет, и коэффициент корреляции потеряет смысл (выродится).

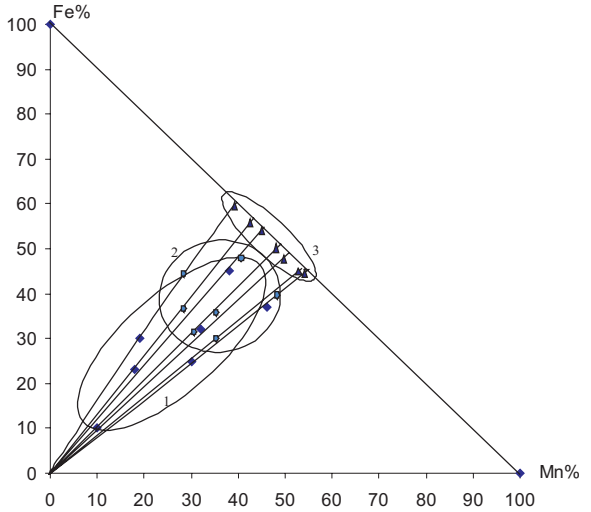

Рис. 3. *Зависимость коэффициента корреляции от суммы содержаний в процентной системе (исходный коэффициент корреляции между Fe и Mn в условно открытой системе равен +1,0)*

Результаты моделирования закрытой системы, представленные в табл. 5 полностью подтвердили приведенные теоретические рассуждения. На рис. 3 приведены кривые приближения коэффициентов корреляции к значению –1 в зависимости от «начального» коэффициента корреляции, т. е. от того, который наблюдался бы при очень сильном разбавлении системы фоновыми компонентами. Модель уменьшения «фона» выбрана пропорциональной текущему содержания фона.

Таким образом, превращение положительного коэффициента корреляции в отрицательный, в итоге равный –1 при уменьшении до нуля содержания всех остальных компонентов, является артефактом и не подлежит никакой содержательной интерпретации.

Промоделированная нами ситуация в точности отвечает той, с которой имел дело В.И. Силаев, изучавший связь между параметрами распределения Mn [4]. В его работе [4. С. 8, рис. 1] приведен корреляционный график, на котором для марганца изображено «облако» точек и уравнение регрессии y на x, где y – среднее квадратическое отклонение содержания марганца, x – среднее содержание марганца.

В основной части генеральной совокупности (более 60 групп объектов) содержания марганца охватывают шесть (!) десятичных порядков. Для неё автор приводит следующие данные: $y=0{,}928x-0{,}213$, $r=0{,}9$. Для значительно меньшей выборки из общей генеральной совокупности, полученной отбором объектов с содержанием Mn>1 % (около 25 точек) им приводятся такие данные: $y=0{,}477x+0{,}301$, $r=0{,}38$. Содержания в этой выборке охватывают чуть больше одного десятичного порядка. На основании такого «статистического» анализа автор делает следующий вывод: «*в природных водах и горных породах распределение Mn характеризуется весьма сильной* ($r=0{,}9$) *прямой корреляцией между средними содержаниями и стандартными отклонениями* ($S\overline{x}$), *а также строгой линейной регрессией в зависимости от* (рис. 1). *Это дает основание определить рассматриваемые объекты как множества с сугубо стохастическим, т. е. случайным распределением в них марганца. При переходе от горных пород к ископаемым марганцевым рудам, оруденелым илам и ЖМК корреляция между средними содержаниями и стандартными отклонениями сильно ослабевает вплоть до полного исчезновения в наиболее богатых оксидных рудах. График соответствующей регрессии сильно выполаживается с явной перспективой вырождения в горизонтальную линию. Все это свидетельствует о том, что при образовании марганцевых руд в отличие от процессов породообразования действует очень мощный фактор упорядочения, обеспечивающий эффект аномального концентрирования* Mn» [4. С. 8].

Таблица 5. *Динамика коэффициента корреляции при уменьшении суммы содержаний «фоновых» компонентов*

Содержание фоновых компонентов, %	№ моделирования	Mn, %	Fe, %	(Mn+Fe), %	r
~43	1	29,1	28,6	57,7	0,661
	2	23,1	27,9	51,0	0,734
	3	32,0	29,9	61,9	0,764
	4	28,3	28,9	57,2	0,472
	.....				
	10	29,1	27,5	56,6	0,610
Среднее из 10		**28,2**	**28,8**	**57,0**	**0,660**
~29	1	37,0	36,9	73,9	0,348
	2	34,2	32,6	66,8	0,343
	3	35,7	35,4	71,2	0,563
	4	38,7	37,7	76,5	-0,119
	.....				
	10	33,5	32,8	66,3	0,442
Среднее из 10		**35,3**	**35,9**	**71,3**	**0,287**
~11	1	43,6	46,1	89,7	-0,600
	2	45,4	43,1	88,5	-0,651
	3	43,7	47,4	91,2	-0,510
	4	44,1	43,5	87,6	-0,552
	.....				
	10	45,4	44,7	90,0	-0,423
Среднее из 10		**44,0**	**44,9**	**88,9**	**–0,549**
~4	1	46,7	49,3	96,1	-0,846
	2	49,6	48,0	97,6	-0,965
	3	48,9	48,0	96,9	-0,973
	4	47,1	47,7	94,8	-0,753
	.....				
	10	50,2	47,2	97,4	-0,980
Среднее из 10		**48,4**	**47,9**	**96,3**	**–0,874**
~1	1	47,6	51,6	99,2	-0,997
	2	49,5	49,7	99,2	-0,997
	3	50,2	48,8	99,0	-0,997
	.....				
	10	49,8	48,8	98,6	-0,918
Среднее из 10		**48,9**	**50,1**	**99,1**	**–0,987**

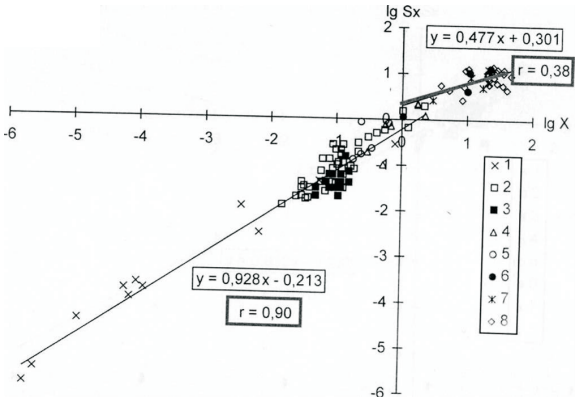

Рис. 4. *Зависимость между $S\bar{x}$ и $\bar{x}$ (по В.И. Силаеву [4]). Объекты: 1) природные воды и рассолы, 2) осадочные и метаморфические горные породы, 3) магматические и гидротермально-метасоматические горные породы, 4) диагенетические конкреции, 5) современные осадки, 6) оруденелые илы, 7) ЖМК и ЖМКО, 8) ископаемые марганцевые руды*

В действительности факт уменьшения коэффициента корреляции для выборки с наиболее высокими содержаниями Mn, выполаживания уравнения регрессии для неё связаны исключительно с тем, что диапазон содержаний в ней очень мал по сравнению с диапазоном содержаний в генеральной совокупности (более семи десятичных порядков величины!). Поэтому из какой бы части совокупности (по содержаниям Mn) ни была получена выборка, охватывающая такой же (чуть более одного порядка) диапазон содержаний, результат будет неизменным – значительное уменьшение коэффициента корреляции и выполаживание уравнения регрессии. Как видно из табл. 2–4, это явление – артефакт, никакого отношения не имеющий к геохимии Mn. Поэтому вся интерпретация В.И. Силаева грубо ошибочна.

Нельзя обойти молчанием и элементарную неграмотность в употреблении термина «регрессия». На цитированном рисунке (он приведён нами под номером 4) дано уравнение регрессии ***y* на *x***, т. е. $S\bar{x}$ на $\bar{x}$, а в цитированном тексте этот же факт преподносится как факт «строгой линейной регрессии $\bar{x}$ в зависимости от $S\bar{x}$», что далеко не одно и то же.

В связи с проблемой корреляции среднего содержания и его среднего квадратического отклонения, снова обратимся к графику из работы В.И. Силаева [4]. Если убрать с рисунка прямые уравнений регрессии, то видно, что в целом для исходной совокупности наблюдается высокий коэффициент корреляции. Подсчитанный нами по приведенным на этом рисунке точкам (не всегда различимым) он оказался равным +0,96. Это однозначно свидетельствует о том, что в генеральной совокупности объектов, подобранных В.И. Силаевым (воды, осадочные породы, магматические породы, оруденелые илы, железо-марганцевые руды) распределение марганца неплохо подчиняется логарифмически-нормальному закону. Этот вывод следует именно из высокой корреляции между $S\bar{x}$ и $\bar{x}$ и ничего другого не означает, в частности, это никак не связано с выводом о том, что якобы «при образовании марганцевых руд в отличие от процессов породообразования действует очень мощный фактор упорядочения, обеспечивающий эффект аномального концентрирования Mn». Рассматривая «статистический анализ» В.И. Силаева в аспекте корреляции среднего квадратического отклонения со средним значением, следует сделать еще несколько замечаний.

Первое: чтобы изучать корреляцию между $S\bar{x}$ и $\bar{x}$, необходимо иметь соответствующие **группы** проб с известной численностью n_i, так как среднее квадратическое отклонение среднего значения, т. е. является, кроме всего прочего, функцией объема группы *n*. Вероятно, автор имел в виду зависимость между *Sx*- в группе и в этой группе проб, а *Sx* и $S\bar{x}$ совсем не одно и то же.

Второе: на графике у В.И. Силаева оси обозначены $\lg S\bar{x} - \lg\bar{x}$, а в тексте логарифмы исчезли. Остается неясным, между чем и чем установлена «строгая линейная» связь: между $\bar{x}$ и $S\bar{x}$ или между их логарифмами. Что это: ошибки по небрежности или от наивной веры, что под знаком ли логарифма выступают величины или без него – не имеет значения.

И вновь мы находим иллюстрацию этого артефакта в работе В.И. Силаева [4]. Один из его главных выводов заключается в том, что в объектах с низкими содержаниями Fe и Mn (в водах, горных породах и др.) между ними наблюдается значимый коэффициент корреляции. В объектах с высокими содержаниями Mn этот коэффициент уменьшается, переходит через ноль, и далее уменьшается до –1.

Цитируем: «*При переходе от природных вод и горных пород к ископаемым марганцевым рудам, оруденелым илам, ЖМК и ЖМКО корреляция между* Mn *и* Fe *скачкообразно* (? – *Ю.Т.*) *изменяется с прямой на обратную, а это свидетельствует о том, что марганцеворудный процесс по сравнению с породообразованием характеризуется противоположной тенденцией геохимического поведения* Mn *относительно* Fe *– следование за железом и рассеяние, характерные для марганца в природных водах и горных породах, сменяются в рудных месторождениях вытеснением железа и концентрированием.*

Изменение характера корреляции Mn *и* Fe *при переходе от горных пород к рудам явно определяется ростом концентрации марганца. В горных породах по мере их обогащения марганцем прямая связь между* Mn *и* Fe *испытывает тенденцию к ослаблению до полного ее «разрушения» в понимании Н.М. Страхова. Проведенные расчеты показали, что такое «разрушение» наступает при содержании* Mn *около* 1 *мас. %, т. е. при достижении уровня, считавшегося академиком Н.М. Страховым предельно именно для «седиментационных» обстановок. В марганцевых рудах, напротив, возникшая обратная корреляция между* Mn *и* Fe *с ростом содержания марганца неуклонно и резко возрастает. Проведенный анализ сви-*

детельствует, что при образовании марганцевых руд происходит радикальное обособление Mn *не только относительно* Fe [4. С. 9]».

Это привело цитируемого автора к следующему защищаемому положению докторской диссертации:

«*В процессах геохимической миграции наиболее близкими к* Mn *спутниками являются* Fe *и* Ca*, но образование марганцевых руд происходит только в результате радикального обособления марганца от этих элементов*» [4. С. 8].

Эти выводы автор считает непосредственно вытекающими из рис. 1, (у нас рис. 4) действительно свидетельствующем об изменении $r_{Mn,Fe}$ фактически до –1 в богатых Fe и Mn объектах. На деле же никакого изменения в геохимической природе коррелирующих Fe и Mn нет. Просто положительный коэффициент корреляции между ними при малых концентрациях испытывает искажение при том же соотношении Fe/Mn под влиянием закрытости системы: как только Fe+Mn начинают составлять существенную часть каждой пробы, эффект закрытости ощущается всё сильнее (см. табл. 5). В чисто Mn – Fe объектах r неизбежно станет равным –1. Эффект отрицательной корреляции, давно открытый, многократно обсужденный, и вновь «обнаруженный» В.И. Силаевым для пары Fe – Mn, является артефактом и не может подлежать какой-либо содержательной интерпретации. Это следует из теории корреляции в процентной системе величин, проиллюстрированной выше компьютерным моделированием (табл. 5, рис. 2, 3 настоящей работы). Подобными ошибками некорректной «содержательной» интерпретации статистических артефактов изобилуют и другие разделы работы В.И. Силаева [4].

При моделировании выявился еще один любопытный феномен. Если коэффициент корреляции в не закрытой, не процентной системе между выделенными компонентами близок или равен –1, то дисперсия их суммы равна

$$s^2_{x+y} = (s_x - s_y)^2,$$

т. е. весьма мала (или равна нулю при $s_x=s_y$), и пересчет на закрытую систему сильно увеличивает этот коэффициент при условии, что дисперсия суммы фоновых компонентов больше, чем $(s_x–s_y)^2$. Этот феномен известен геохимикам, занимающимся породами с большим количеством разбавителя, например углей, когда поведение элементов рассматривается либо в пересчете химических элементов на золу, либо в целом на уголь; известняков, когда то же самое делается в пересчете на нерастворимый остаток или на всю карбонатную породу. В золе или нерастворимом остатке часто наблюдаются отрицательные коэффициенты корреляции между компонентами. При пересчете на всю породу именно из-за большой дисперсии содержания органики (или карбонатной части) – отрицательные корреляции превращаются в положительные. Парадоксальный феномен получил теоретическое объяснение и подтверждён компьютерным моделированием, следовательно, перестал быть парадоксальным.

СПИСОК ЛИТЕРАТУРЫ

1. Вистелиус А.Б. Проблемы математической геологии // Геология и геофизика. – 1962. – № 12. – С. 3–9.
2. Chayes F. A petrographic criterion for the possible replacement origin of rocks // Amer. J. Sci. – 1948. – V. 246. – P. 413–420.
3. Chayes F. Detecting Nonrandom Associations Between Proportions by Tests of Remaining-Space Variables // Mathematical Geol. – 1983. – V. 15. – № 1. – P. 197–206.
4. Силаев В.И. Механизмы и закономерности эпигенетического марганцевого минералообразования: Автореф. дис. ... докт. геол.-мин. наук. – Сыктывкар, 2006. – 40 с.
5. Ткачев Ю.А., Юдович Я.Э. Статистическая обработка геохимических данных. – Л.: Наука, 1975. – 236 с.
6. Иванова Т.И., Ткачев Ю.А. Спектральный анализ в геологии и геохимии. – Екатеринбург: УрО РАН, 2003. – 297 с.
7. Pearson K. On lines and planes of closest tit to systems of points in space // Phil. Mag. – 1901. – Ser. 6. – V. 2. – № 11. – P. 559–572.
8. Darroch J.N., Ratcliff D. Null correlation for proportions ?? // Jour. Math. Geol. – 1970. – V. 2. – P. 307–312.
9. Kork J.O. Examination of the Chayes-Kruskal Procedure for Testing Correlations Between Proportions // Mathematical Geol. – 1977. – V. 9. – № 6. – P. 543–562.
10. Snow J.W. Association of proportions // Jour. Intern. Assoc. Math. Geol. – 1975. – V. 7. – № 1. – P. 63–73.
11. Ткачев Ю.А. Проблема процентных величин в минералогии, петрографии и геохимии. – Сыктывкар: Геопринт, 1999. – 27 с.

Поступила 30.10.2007 г.