ISSN 0868–5886 НАУЧНОЕ ПРИБОРОСТРОЕНИЕ, 2023, том 33, № 1, с. 86–94

МАТЕМАТИЧЕСКИЕ МЕТОДЫ И МОДЕЛИРОВАНИЕ В ПРИБОРОСТРОЕНИИ

УДК 543.612/ 621/ 684

© Л. В. Новиков, А. Г. Кузьмин, Ю. А. Титов, 2023

АНАЛИЗ МНОГОМЕРНЫХ ДАННЫХ ПО СОСТАВУ ГАЗОВ, ВЫДЕЛЯЕМЫХ ИЗ РАЗЛОМОВ ЗЕМНОЙ ПОВЕРХНОСТИ

Предлагается метод экспресс-прогноза состояния земной коры (в том числе прогноза землетрясений и извержений вулканов) по составу и интенсивности газов, регистрируемых в местах разломов земной поверхности. Метод основан на обучении без учителя с использованием большого объема предварительно собранных данных о составе и концентрации газов, выделяемых в зоне разломов земной коры. Состав и концентрация этих газов содержат информацию о процессах, происходящих в глубине Земли, что позволяет с некоторой вероятностью предсказать землетрясения или другие катастрофические события. Собранные данные служат для обучения системы распознавания вновь полученных данных путем формирования системы кластеров, каждый из которых является маркером того или иного процесса в земной коре. Близость в многомерном пространстве новых данных к ядру кластера является вероятностной мерой события, вызвавшего выброс газовой смеси, аналогичной кластеру.

Кл. сл.: экспресс-диагностика, кластерный анализ, многомерная плотность вероятности, обработка многомерных данных

ВВЕДЕНИЕ

Газовые смеси, выделяемые из разломов земной коры, как правило, содержат несколько компонентов. Чаще всего встречаются: CO_2, CH_4, He, H_2S, H_2, N_2, O_2 и другие в зависимости от места расположения разлома [1]. В многомерном пространстве данных, например, семь компонентов представлены в виде точки в семимерном пространстве. Множество замеров, выполняемых в течение некоторого промежутка времени, образует "облако", которое может состоять из нескольких тысяч точек данных. При этом концентрация каждого компонента в "облаке" и в целом спектр зависят от внутренних процессов, происходящих в земной коре, т.е. содержат информацию о разломе как геологическом объекте. Если эти процессы близки и повторяются, также близки интенсивности выделяемых газов, и в многомерном пространстве каждый из таких процессов образует группу (кластер) близко расположенных точек. В результате длительного наблюдения с использованием различных приборов может быть установлена статистическая связь между спектральным составом газов и внутренними физико-химическими процессами в земной коре. В дальнейшем, используя полученные результаты, по единичным замерам состава газов можно с некоторой вероятностью предсказать характер происходящих в земной коре процессов.

Этот подход может быть положен в основу метода экспресс-прогноза текущего состояния земной коры на основе масс-спектрометрического мониторинга состава газов. Процедура прогнозирования производится в три этапа.

Первый этап:

– наблюдение в течение продолжительного времени с регистрацией спектров газов — формирование обучающей выборки;

– формирование кластеров спектров и установление их связи с физико-химическими процессами в земной коре.

Второй этап:

текущее измерение спектра газов и определение его принадлежности тому или иному кластеру по минимуму расстояния между его центром (центроидой) и точкой спектра в многомерном пространстве.

Третий этап:

выводы о вероятности процессов, происходящих в земной коре.

ОБРАБОТКА ДАННЫХ

Теория

Для накопления данных о составе и интенсивности газов, выделяемых в разломах земной поверхности, целесообразно проводить параллельные замеры в нескольких разломах в одном геологическом районе в течение длительного промежутка времени с одновременной регистрацией процессов, происходящих в земной коре, с помощью других

приборов. Эти данные назовем обучающей группой, т.к. в дальнейшем они необходимы для сопоставления с текущими замерами и принятия решения о возможных последствиях.

В результате одного замера регистрируется несколько компонентов газа, образуя вектор-строку $[x_{i,1}, x_{i,2}, ..., x_{i,j}]$, где $x_{i,j}$ — интенсивность *j*-го компонента *i*-го замера обучающей группы, причем $i = 1,2,...,I$. Набор из *I* замеров *J* регистрируемых компонент газа образует (I,J) обучающую матрицу $\mathbf{X}$, столбцы которой обозначим как $\mathbf{X}_j$: $\mathbf{X} = [\mathbf{X_1}, ..., \mathbf{X}_j, ..., \mathbf{X}_J]$ [2].

После достаточно длительного наблюдения с выполнением одного-двух замеров в сутки обучающая матрица $\mathbf{X}$ содержит ряд групп (кластеров), близких по времени измерений. Кластеры отличаются интенсивностью, а иногда и составом газов в зависимости от характера процессов, происходящих в земной коре в момент измерений.

Обозначим через $\mathbf{X}^k$ *k* -й кластер, *k* = (1, 2, 3, …, *K*), где *K* — число кластеров. Кластер представляет собой матрицу из I_k строк ($I_k << I$) и J столбцов. Каждая строка матрицы образует в *J*-мерном пространстве точку, а все строки — "облако" из I_k точек. Центр этого "облака", центроида кластера, имеет координаты в виде средних значений по столбцам: $\bar{\mathbf{X}}^k = [\bar{X}_1^k, \bar{X}_2^k, ..., \bar{X}_J^k]$ и дисперсий $\boldsymbol{\sigma}_k^2 = [\sigma_{1k}^2, \sigma_{1k}^2, ..., \sigma_{Jk}^2]$.

Идентификация новых замеров

Разбиение обучающих замеров на кластеры и их идентификация с определенным процессом в коре завершает обучение системы обработки данных. Для надежной идентификации новых замеров по обучающим данным число элементов в кластере должно быть как минимум больше трех.

Пусть очередной замер воздуха в разломе — $\mathbf{X}_d = \left[x_{d,1}, x_{d,2}, ..., x_{d,J}\right]$, где $x_{d,j}$ — текущая интенсивность компонента газовой смеси в разломе. Задача состоит в том, чтобы по этому замеру с некоторой вероятностью можно было судить о физико-химических процессах в недрах Земли, для чего оценим расстояние в многомерном пространстве между точкой $\mathbf{X}_d$ и центроидой *k*-го кластера $\bar{\mathbf{X}}^k$. Это расстояние в евклидовой метрике определяется как

$$\rho_k = \left(\mathbf{X}_d - \bar{\mathbf{X}}^k\right)' \left(\mathbf{X}_d - \bar{\mathbf{X}}^k\right), \qquad (1)$$

$'$ — символ транспонирования матрицы.

Принадлежность замера $\mathbf{X}_d$ тому или иному кластеру k_0 определяется по минимуму расстояния ρ_k:

$$\rho_{k_0} = \min_{(k)}(\rho_k). \qquad (2)$$

Очевидно, что величина ρ_k определяет также вероятность определенного процесса, происходящего в момент замера: при $\rho_k \equiv 0$ эта вероятность должна быть равна единице, а при удалении точки $\mathbf{X}_d$ от центроиды — уменьшаться до нуля. Этим свойством обладает функция плотности вероятности $P(\mathbf{X}_d)$ случайной величины $\mathbf{X}_d$ со средним значением $\bar{\mathbf{X}}^k$ и дисперсией данных, входящих в этот кластер, $\boldsymbol{\sigma}_k^2$:

$$P(\mathbf{X}_d) = W \cdot \exp\left\{-\frac{1}{2}\left(\mathbf{X}_d - \bar{\mathbf{X}}^k\right)' \mathbf{K}^{-1}\left(\mathbf{X}_d - \bar{\mathbf{X}}^k\right)\right\}, \qquad (3)$$

где $\mathbf{K}$ — ковариационная матрица: $\mathbf{K} = \mathrm{E}\left[\left(\mathbf{X} - \bar{\mathbf{X}}\right)\cdot\left(\mathbf{X} - \bar{\mathbf{X}}\right)'\right]$, E — символ математического ожидания, *W* — нормирующий множитель. Из условия равенства вероятности $P(\mathbf{X}_d)$ единице при $\rho \equiv 0$, должно быть *W* = 1. Условие принадлежности замера $\mathbf{X}_d$ кластеру $\mathbf{X}^k$ имеет вид: $P(\mathbf{X}_d) < \alpha$, где величина α выбирается методом экспертной оценки.

Метод главных компонент

Однако непосредственное использование приведенных формул для разделения данных на кластеры и расчета величины *P* сопряжено с ошибками, вызванными наличием большого числа параметров *J* и корреляционных связей между столбцами матрицы $\mathbf{X}$. Для сжатия данных, сокращения размерности пространства измерений используют ортогональное преобразование данных в пространство главных компонент — метод главных компонент (МГК) [3].

Для перехода в пространство ГК формируется новая матрица, состоящая из всех строк матрицы $\mathbf{X}$ и строки $\mathbf{X}_d$. Обозначим эту матрицу как $\mathbf{X}1$. Тогда в новой системе координат:

$$\mathbf{X}1 = \mathbf{T}\cdot\mathbf{P}' + \mathbf{e} = \sum_{j=1}^{A} \mathbf{t}_j \mathbf{p}'_j + \mathbf{e}, \qquad (4)$$

где $\mathbf{p}_j$ — собственные функции ковариационной матрицы $\mathbf{K}$. Матрицу $\mathbf{T}$ называют матрицей

счетов $\mathbf{T} = [\mathbf{T}_1, \mathbf{T}_2, ..., \mathbf{T}_A]$, ее размерность — $(I \times A)$; матрицу $\mathbf{P}$ называют матрицей *нагрузок*, ее размерность — $(I \times A)$; $\mathbf{e}$ — это матрица *остатков* (шумов) размерности $(I \times J)$; векторы-столбцы $\mathbf{T}_j$ $(j = (1, 2, ..., A))$ называют *главными компонентами* (ГК), A — число главных компонент. Величина A значительно меньше числа переменных J. Это означает, что основная информация сосредоточена в нескольких первых ГК. Последняя строка этой матрицы, вектор $\mathbf{T}_d$ — координаты параметров тестируемого состава воздуха в пространстве ГК: $\mathbf{T}_d = [t_{d,1}, t_{d,2}, ..., t_{d,A}]$.

Из данных в новой системе координат формируются кластеры $\mathbf{T}^k$ — матрицы из I_k строк ($I_k << I$) и A столбцов. Центроида кластера имеет координаты в виде средних значений по столбцам $\vec{\mathbf{T}}^k = [\vec{T}_1^k, \vec{T}_2^k, ..., \vec{T}_A^k]$ и дисперсий $\boldsymbol{\sigma}_k^2 = [\sigma_{1k}^2, \sigma_{1k}^2, ..., \sigma_{Ak}^2]$.

Свойство разложения по ГК таково, что дисперсия быстро уменьшается уже к четвертой ГК, а столбцы матрицы $\mathbf{T}^k$ не коррелированы, т.е.

$$\mathbf{T}_m^k \left(\mathbf{T}_n^k\right)' = \begin{cases} 0 & \text{при } n \neq m, \\ \sigma_k^2 & \text{при } n = m. \end{cases}$$

Учитывая это обстоятельство, в новой системе координат вероятность принадлежности нового замера кластеру k вычисляется по формуле:

$$P(\mathbf{T}_d) = \exp\left\{-\frac{1}{2}\left(\mathbf{T}_d - \vec{\mathbf{T}}^k\right)' \boldsymbol{\sigma}_k^{-2} \left(\mathbf{T}_d - \vec{\mathbf{T}}^k\right)\right\} =$$

$$= \exp\left\{-\frac{1}{2}\sum_{j=1}^{A} \frac{\left(t_{d,j} - \vec{t}_j^k\right)^2}{\sigma_{k,j}^2}\right\}. \quad (5)$$

Евклидово расстояние от замера с индексом d до центроиды k-го кластера равно:

$$\rho_k = \left(\mathbf{T}_d - \vec{\mathbf{T}}^k\right)' \left(\mathbf{T}_d - \vec{\mathbf{T}}^k\right) = \left\{\sum_{j=1}^{A} \left(t_{d,j} - \vec{t}_j^k\right)^2\right\}^{1/2}. \quad (6)$$

ОПИСАНИЕ АЛГОРИТМА

Обработка данных состоит из двух этапов: *обучение и диагностика*.

На этапе *обучения* формируется обучающая матрица $\mathbf{X}$ путем набора данных о составе и концентрации компонентов воздуха в местах выделения газа и привязки этих данных к происходящим там процессам. Затем, используя подходящий алгоритм кластеризации [4], разбивают накопленные данные на кластеры $\mathbf{X}^k$, каждый из которых отображает определенный физико-химический процесс в земной коре. Может быть использован другой вариант разбиения накопленных данных на кластеры: сортировка данных по принадлежности к процессу, происходящему в земной коре в момент замера состава выделяемых газов.

На этапе *диагностики* выполняется следующая последовательность операций:

1. Измеряются состав и концентрация диагностируемого источника газа и формируется вектор–строка $\mathbf{X}_d = \{x_{d,1}, x_{d,2}, ..., x_{d,J}\}$.

2. Отображение данных матрицы $\mathbf{X}$ и замера $\mathbf{X}_d$, т.е. матрицы $\mathbf{X}1 = [\mathbf{X}; \mathbf{X}_d]$, в пространство ГК (4).

3. Вычисляется расстояние по формуле (6) и определяется ближайший кластер по минимуму расстояния (2).

4. Определяется вероятность P по формуле (5)

5. Анализ результата вычисления вероятности.

ПРОВЕРКА АЛГОРИТМА

Проиллюстрируем изложенную выше теорию на примере реальных 1024 замеров состава газа, выполненных в Ленинградской области.

На рис. 1 показан пример масс-спектра одного из замеров состава газовой смеси: CH_4, N_2, O_2, CO_2, Ar. Замеры выполнялись ежедневно один

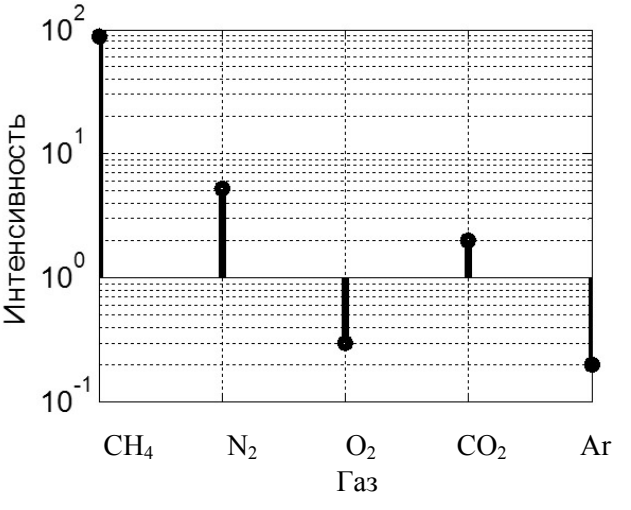

Рис. 1. Спектр одного из замеров состава смеси

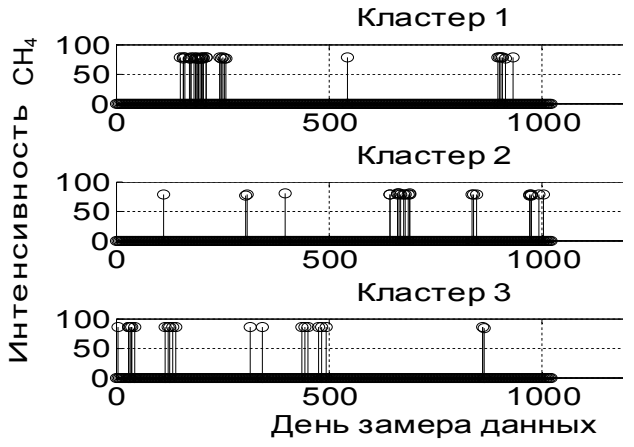

Рис. 2. Связь кластеров и времени наблюдений

раз в сутки. Данные, накопленные в течение продолжительного времени с использованием иерархического агломеративного алгоритма [4], разделены на кластеры, каждый из которых состоит из более чем десяти замеров близких по интенсивности компонентов газа. На рис. 2 показано расположение трех кластеров на временно́й шкале продолжительностью более 1000 дней наблюдений: каждый кластер связан с определенным временем и, соответственно, процессом, происходящим в земной коре. (На рисунке приведен только метан.)

В табл. 1 (столбец 2) приведено евклидово расстояние одного из замеров до девяти кластеров, вычисленных по формуле (6). Из таблицы следует, что замер принадлежит третьему кластеру с вероятностью 0.33, вычисленной по формуле (5) (столбец 3).

Табл. 1. Пример расположения замера относительно кластеров

№ кластера	Расстояние от замера до центроиды кластера	Вероятность
1	105.8	0
2	101	0
3	0.4	0.3371
4	136.91	0
5	4.5	0.007
6	664.4	0
7	28	0
8	67.6	0
9	217.4	0

Табл. 2. Характеристики точек-замеров кластера 3

№ замера из кластера 3	Расстояние от замера до центроиды кластера 3	Вероятность соответствующего геособытия
1	0.3442	0.0982
2	0.4309	0.1366
3	0.3153	0.1456
4	0.484	0.1709
5	0.7648	0.1825
6	1.0106	0.1833
7	0.7499	0.2241
8	0.4035	0.3371
9	0.7676	0.3864
10	0.303	0.428
11	0.143	0.4333
12	0.0725	0.4882
13	0.1344	0.546
14	0.1367	0.5643
15	0.0169	0.7377

Далее выберем пятнадцать замеров, принадлежащих третьему кластеру, но расположенных в разных точках "облака" этого кластера. В табл. 2 во втором столбце приведены расстояния до центроиды кластера, а в третьем столбце — вероятность события в земной коре, вызвавшего соответствующее выделение газов. Видно также, что чем ближе расположен замер к центру кластера, тем больше вероятность события. В приведенном примере замер номер 1 вероятнее всего принадлежит другому кластеру. Тогда по приведенной выше методике находят этот кластер и определяют соответствующую вероятность.

ЗАКЛЮЧЕНИЕ

Система диагностики процессов в земной коре, состоящая из прибора для количественного измерения состава воздуха в местах выделения газа и средства обработки данных с возможностью обучения и определения вероятности происходящих в земной коре процессов, обеспечивает экспресс-анализ состояния земной коры. Диагностика с определением вероятности принадлежности состава выделяемого газа одному из обучающих кластеров позволяет составить общую картину вероятных событий в земной коре и при необходимости принять соответствующие решения относительно вероятности определенных геологических процессов в данном регионе.

Благодарности

Авторы выражают благодарность К.Н. Котову за предоставленные образцы газовых смесей.

Работа выполнена в ИАП РАН в рамках государственного задания № 122040600002-3.

СПИСОК ЛИТЕРАТУРЫ

1. *Кузьмин Ю.Д., Кузьмин А.Г.* Масс-спектрометрический анализ состава газов на термальных площадках Камчатки в полевых условиях // Труды III научно-технической конференции "Проблемы комплексного геофизического мониторинга Дальнего Востока России", г. Петропавловск-Камчатский, 9–15 октября 2011 г. Обнинск: ГС РАН, 2011. С. 1–5.
2. *Большаков А.А., Каримов Р.Н.* Методы обработки многомерных данных и временных рядов. Учебное пособие для вузов. М.: Горячая линия-Телеком, 2007. 522 с.
3. *Айвазян С.А., Бухштабер В.М., Енюков И.С., Мешалкин Л.Д.* Прикладная статистика. Классификация и снижение размерности. М.: Финансы и статистика, 1989. 607 с.
4. *Мандель И.Д.* Кластерный анализ. М.: Финансы и статистика, 1988. 176 с.

Институт аналитического приборостроения РАН, г. Санкт-Петербург

Контакты: *Кузьмин Алексей Георгиевич*, agqz55@rambler.ru

Материал поступил в редакцию 24.10.2022

ISSN 0868–5886 NAUCHNOE PRIBOROSTROENIE, 2023, Vol. 33, No. 1, pp. 86–94

ANALYSIS OF MULTIDIMENSIONAL DATA ON THE COMPOSITION OF GASES EMITTED FROM FAULTS IN THE EARTH'S SURFACE

L. V. Novikov, A. G. Kuzmin, Yu. A. Titov

Institute for Analytical Instrumentation of RAS, Saint Petersburg, Russia

The composition and intensity of gases recorded at fault location on the earth's surface are used to make an express forecast of the state of the earth's crust (including the forecast of earthquakes and volcanic eruptions) by. The method is based on unsupervised learning using a large amount of pre-collected data on the composition and concentration of gases released in the fault zone of the earth's crust. The composition and concentration of these gases contain information about the processes occurring in the depths of the earth, which makes it possible to predict earthquakes or other catastrophic events with some probability. The collected data serve to train the recognition system for newly received data by forming a system of clusters, each of which is a marker of a particular process in the earth's crust. The proximity of new data in the multidimensional space to the core of the cluster is a probabilistic measure of the event that caused the release of a gas mixture similar to a cluster.

Keywords: express diagnostics, cluster analysis, multivariate probability density, multivariate data processing

INTRODUCTION

Gas mixtures released from faults in the earth's crust, as a rule, contain several components. Most often, there are: CO_2, CH_4, He, H_2S, H_2, N_2, O_2 and others, depending on the location of the crust [1]. In a multivariate data space, for example, seven components are represented as a point in a seven-dimensional space. Many measurements taken over a period of time form a "cloud," which can consist of several thousand data points. The concentration of each component in the "cloud" and the spectrum as a whole depend on the internal processes occurring in the earth's crust, i.e., contain information about the fault as a geological object. If these processes are close and repeated, the intensity of the gases released is also close, and in multidimensional space each of these processes forms a group (cluster) of closely spaced points. As a result of long-term observation using various instruments, a statistical relationship can be established between the spectral composition of gases and internal physicochemical processes in the earth's crust. Subsequently, using the results obtained, from sample measurements of the composition of gases, it is possible, with some probability, to predict the nature of the processes occurring in the earth's crust.

Based on mass spectrometric monitoring of gas composition, this methodology can be used to expressly forecast the current condition of the earth's crust. The forecasting procedure is carried out in three stages.

First stage:

– long-term observation with recording of gas spectra, i.e., formation of a training sample;

– the formation of spectral clusters and establishment of their relation to physicochemical processes in the Earth's crust.

Second stage:

ongoing measurement of the spectrum of gases and determination of its belonging to a particular cluster by the minimum distance between its center (centroid) and the point of the spectrum in multidimensional space.

Third stage:

conclusions about the probability of processes occurring in the Earth's crust.

DATA PROCESSING

Theory

In order to accumulate data on the composition and intensity of gases released in faults on the earth's surface, it is advisable to carry out parallel measurements in several faults in one geological area for a long period of time with simultaneous recording of processes occurring in the earth's crust using other instruments. We will call this group of data a training group, since in the future they will be necessary for comparison with ongoing measurements and making decisions on possible consequences.

As a result of one measurement, several gas components are recorded, forming a vector row

$[x_{i,1}, x_{i,2}, ..., x_{i,j}]$, where $x_{i,j}$ is the intensity of the *j*-th component of the *i*-th measurement of the training group, wherein $i = 1, 2, ..., I$. A set of *I* measurements of *J* recorded gas components forms (I, J) a training matrix $\mathbf{X}$, which columns are denoted as $\mathbf{X}_j$: $\mathbf{X} = [\mathbf{X_1}, ..., \mathbf{X}_j, ..., \mathbf{X}_J]$ [2].

After a sufficiently long observation with one or two measurements per day, the training matrix $\mathbf{X}$ contains a number of groups (clusters) of measurements that are close in time. Clusters differ in intensity and sometimes in composition of gases, depending on the nature of the processes taking place in the earth's crust at the time of measurements. Let's use the $\mathbf{X}^k$ to denote *k*-th cluster, *k* = (1, 2, 3, ..., *K*), where *K* is the number of clusters. A cluster is a matrix of I_k rows ($I_k << I$) and J columns. Each row of the matrix forms a point in *J*-dimensional space, and all rows form a "cloud" of I_k points. The center of this "cloud," the cluster centroid, has coordinates in the form of average values in the columns $\bar{\mathbf{X}}^k = [\bar{X}_1^k, \bar{X}_2^k, ..., \bar{X}_J^k]$, the variance is $\boldsymbol{\sigma}_k^2 = [\sigma_{1k}^2, \sigma_{1k}^2, ..., \sigma_{Jk}^2]$.

Identification of new measurements

Splitting the training measurements into clusters and identifying them with a specific process in the earth's crust completes the training of the data processing system. The number of cluster elements must be greater than three in order to reliably identify new measurements using training data.

Let the next measurement of the air in the crust — $\mathbf{X}_d = \left[x_{d,1}, x_{d,2}, ..., x_{d,J}\right]$, where $x_{d,j}$ — the ongoing intensity of the gas mixture component in the crust. The question is whether this measurement can be used to assess the physicochemical processes in the bowels of the earth. For this purpose, we estimate the distance in the multidimensional space between the point $\mathbf{X}_d$ and the centroid $\bar{\mathbf{X}}^k$ of the *k*-th cluster. This distance in the Euclidean metric is defined as

$$\rho_k = \left(\mathbf{X}_d - \bar{\mathbf{X}}^k\right)' \left(\mathbf{X}_d - \bar{\mathbf{X}}^k\right), \quad (1)$$

$'$ is a matrix transpose symbol.

The minimum distance ρ_k determines whether the measurement $\mathbf{X}_d$ belongs to a particular cluster k_0:

$$\rho_{k_0} = \min_{(k)}(\rho_k). \quad (2)$$

Obviously, the value ρ_k also determines the probability of a certain process that occurs at the time of measurement: if $\rho_k \equiv 0$ this probability is equal to 1, and if a point $\mathbf{X}_d$ moves away from the centroid, it decreases up to zero. This property has a probability density function $P(\mathbf{X}_d)$ of a random variable $\mathbf{X}_d$ with an average value $\bar{\mathbf{X}}^k$ and variance $\boldsymbol{\sigma}_k^2$ of data in this cluster:

$$P(\mathbf{X}_d) = \\ = W \cdot \exp\left\{-\frac{1}{2}\left(\mathbf{X}_d - \bar{\mathbf{X}}^k\right)' \mathbf{K}^{-1} \left(\mathbf{X}_d - \bar{\mathbf{X}}^k\right)\right\}, \quad (3)$$

where $\mathbf{K}$ is the covariance matrix: $\mathbf{K} = \mathrm{E}\left[\left(\mathbf{X} - \bar{\mathbf{X}}\right) \cdot \left(\mathbf{X} - \bar{\mathbf{X}}\right)'\right]$, E is the expected value symbol, *W* is the normalizing factor. Under the condition of equality of probability $P(\mathbf{X}_d)$ to 1 if $\rho \equiv 0$, $W = 1$. The condition of measurement $\mathbf{X}_d$ belonging to the cluster $\mathbf{X}^k$ has the form: $P(\mathbf{X}_d) < \alpha$, where the value α is selected by the expert assessment method.

Principal Component Analysis

However, the direct use of the given formulas to divide the data into clusters and calculate the *P* value is associated with errors caused by the presence of a large number of parameters *J* and correlation relationships between the columns of the matrix $\mathbf{X}$. To compress data, reduce the dimension of the measurement space, an orthogonal transformation of the data into the space of the main components is used — the Principal Component Analysis (PCA) [3].

For the transition to the PC space, a new matrix is formed, consisting of all the rows of the matrix $\mathbf{X}$ and the row $\mathbf{X}_d$. Denote this matrix as $\mathbf{X}1$. Then in the new coordinate system:

$$\mathbf{X}1 = \mathbf{T} \cdot \mathbf{P}' + \mathbf{e} = \sum_{j=1}^{A} \mathbf{t}_j \mathbf{p}'_j + \mathbf{e}, \quad (4)$$

where $\mathbf{p}_j$ — eigenfunctions of the covariance matrix $\mathbf{K}$. The matrix $\mathbf{T}$ is called the matrix of accounts $\mathbf{T} = \left[\mathbf{T}_1, \mathbf{T}_2, ..., \mathbf{T}_A\right]$, its dimension — ($I \times A$); a matrix $\mathbf{P}$ is called a load matrix, its dimension is ($I \times A$); $\mathbf{e}$ — a matrix of residues (noise) of dimension $(I \times J)$; column vectors $\mathbf{T}_j$ $\left(j = (1, 2, ..., A)\right)$ are called *principal components* (PCs), A — the number of main components. The value A is significantly less than the number of variables *J*. This means that the main information is concentrated in the first few PCs.

The last row of this matrix, vector $\mathbf{T}_d$ is the coordinates of the parameters of the tested air composition in the PC space: $\mathbf{T}_d = \left[t_{d,1}, t_{d,2}, ..., t_{d,A} \right]$.

From the data in the new coordinate system, clusters $\mathbf{T}^k$ are formed — matrices of I_k rows ($I_k << I$) and A columns. The cluster centroid has coordinates in the form of average values in columns $\bar{\mathbf{T}}^k = [\bar{T}_1^k, \bar{T}_2^k, ..., \bar{T}_A^k]$. The variance $\boldsymbol{\sigma}_k^2 = \left[\sigma_{1k}^2, \sigma_{1k}^2, ..., \sigma_{Ak}^2 \right]$.

The PC decomposition property is such that the dispersion decreases rapidly by the fourth PC, and the columns of the matrix $\mathbf{T}^k$ are not correlated, i.e.

$$\mathbf{T}_m^k \left(\mathbf{T}_n^k \right)' = \begin{cases} 0 & \text{при } n \neq m, \\ \sigma_k^2 & \text{при } n = m. \end{cases}$$

Given this circumstance, in the new coordinate system, the probability of belonging to the new measurement to the cluster k is calculated using the formula:

$$P\left(\mathbf{T}_d\right) = \exp\left\{ -\frac{1}{2} \left(\mathbf{T}_d - \bar{\mathbf{T}}^k \right)' \boldsymbol{\sigma}_k^{-2} \left(\mathbf{T}_d - \bar{\mathbf{T}}^k \right) \right\} =$$

$$= \exp\left\{ -\frac{1}{2} \sum_{j=1}^{A} \frac{\left(t_{d,j} - \vec{t}_j^k \right)^2}{\sigma_{k,j}^2} \right\}. \quad (5)$$

The Euclidean distance from the measurement with index d to the centroid of the k-th cluster is:

$$\rho_k = \left(\mathbf{T}_d - \bar{\mathbf{T}}^k \right)' \left(\mathbf{T}_d - \bar{\mathbf{T}}^k \right) = \left\{ \sum_{j=1}^{A} \left(t_{d,j} - \vec{t}_j^k \right)^2 \right\}^{1/2}. \quad (6)$$

ALGORITHM DESCRIPTION

Data processing consists of two stages: *training* and *diagnostics*.

At *the training* stage, a training matrix $\mathbf{X}$ is formed by collecting data on the composition and concentration of air components at gas release points and linking this data to the processes taking place there. Then, using a suitable clustering algorithm [4], the accumulated data is divided into clusters $\mathbf{X}^k$, each of which maps a certain physicochemical process in the earth's crust. Another variant of splitting the accumulated data into clusters can be used: sorting the data according to the processes taking place in the earth's crust at the time of measuring the composition of the released gases.

At *the diagnostic* stage, the following sequence of operations is performed:

1. The composition and concentration of the diagnosed gas source are measured, and a string vector $\mathbf{X}_d = \left\{ x_{d,1}, x_{d,2}, ..., x_{d,J} \right\}$ is formed.
2. Mapping of the matrix $\mathbf{X}$ data and measurement result $\mathbf{X}_d$, i.e., the matrix $\mathbf{X}1 = \left[\mathbf{X}; \mathbf{X}_d \right]$ in the PC space (4).
3. The distance is calculated using the formula (6) and the nearest cluster is determined by the minimum distance (2).
4. Probability P is determined by formula (5)
5. Analysis of the probability calculation result is carried out.

ALGORITHM CHECK

Let's illustrate the above theory using the example of real 1024 gas composition measurements taken in the Leningrad Region, Russia.

Fig. 1 shows an example of the mass spectrum of one of the gas mixture composition measurements: CH_4, N_2, O_2, CO_2, Ar. The measurements were taken once a day. The data accumulated over a long period of time using the hierarchical agglomerative algorithm [4] are divided into clusters, each of which consists of more than ten measurements of gas components similar in intensity.

Fig. 1. Spectrum of one of the mixture composition measurements

Fig. 2 shows the location of three clusters on a time scale lasting more than 1000 days of observations: each cluster is associated with a certain time and, accordingly, with a process occurring in the earth's crust. (The figure shows only methane).

Fig. 2. Example of distribution of methane concentrations assigned to 3 clusters on the time axis

Tab. 1 (column 2) shows the Euclidean distance of one of the measurements to nine clusters calculated using formula (6). It follows from the table that the measurement belongs to the third cluster with probability 0.33, calculated by formula (5) (column 3).

Tab. 1. An example of the location of the measurement result relative to clusters

Next, we select fifteen measurement results belonging to the third cluster, and located at various points in this cluster's cloud. In Tab. 2, the second column shows the distances to the cluster centroid, and in the third column — the probability of an event in the earth's crust that caused the corresponding release of gases. You can also see that the closer the measurement result is to the center of the cluster, the more likely the event is. In the example shown, measurement result number 1 most likely belongs to another cluster. Then, using the above method, this cluster is found and the corresponding probability is determined.

Tab. 2. Characteristics of cluster 3 measurement results points

CONCLUSION

The system for diagnosing processes in the earth's crust, consisting of a device for quantitative measurement of the composition of air in places of gas release and a data processing tool capable of learning and determining the probability of processes occurring in the earth's crust, provides an express analysis of the state of the earth's crust. Diagnostics with the determination of the probability that the composition of the released gas belongs to one of the training clusters makes it possible to compile a general picture of probable events in the earth's crust and, if necessary, make appropriate decisions regarding the likelihood of specific geological processes in a given region.

REFERENCES

1. Kuzmin Yu.D., Kuzmin A.G.[Mass spectrometry analysis of gas composition at Kamchatka thermal sites in the field]. *Trudy III nauchno-technicheskoy konferenzii "Problemy kompleksnogo geofizicheskogo monitoringa Dal'nego Vostoka Rossii", g. Petropavlovsk-Kamchatskiy, 9–15 oktyabrya 2011 g.* [Proceedings of the III Scientific and Technical Conference "Problems of Integrated Geophysical Monitoring of the Russian Far East", Petropavlovsk-Kamchatsky, October 9–15, 2011], Obninsk, GS RAN Publ., 2011, pp. 1–5. (In Russ.).
2. Bolshakov A.A., Karimov R.N. *Metody obrabotki mnogomernych dannych i vremennych ryadov. Uchebnoe posobie dlya vuzov* [Methods for processing multidimensional data and time series. A textbook for universities]. Moscow, Goryachaya liniya-Telekom Publ., 2007. 522 p. (In Russ.).
3. Ayvazyan S.A., Buchshtaber V.M., Enyukov I.S., Meshalkin L.D. *Prikladnaya statistika. Klassifikaziya i snizhenie razmernosti* [Applied statistics. Classification and dimensioning]. Moscow, Finansy i statistika Publ., 1989. 607 p. (In Russ.).
4. Mandel I.D. *Klasternyy analiz* [Cluster analysis]. Moscow, Finansy i statistika Publ., 1988. 176 p. (In Russ.).

Contacts: *Kuzmin Aleksey Georgievich,*
agqz55@rambler.ru

Article received by the editorial office on 24.10.2022