

ZASTOSOWANIE ANALIZY WYRÓŻNIAJĄCEJ W GEOLOGII

UKD 550.8.011:519.2

Analiza wyróżniająca, zwana też analizą dyskryminacyjną (discriminant analysis), jest jedną z technik statystycznych używanych do testowania przynależności obiektów, w tym obiektów geologicznych do wstępnie zdefiniowanych populacji (w geologii do populacji geologicznych). W badaniach geologicznych jednym z typowych do rozwiązania jest następujące zadanie: na podstawie wiedzy merytorycznej są ustalone i porównywane dwie (lub kilka) populacje geologiczne przedmiotowe — mogą to być określone rodzaje skał, serie litofacyjne itp.; każda z rozpatrywanych populacji jest opisana przez pomiar tego samego zestawu cech. Jako cechy mogą być przyjęte dowolne własności fizyczne, chemiczne lub inne charakterystyczne dane, uznane przez geologa za reprezentatywne dla danych populacji. Jeżeli badane populacje, dane przez zbiory obiektów (prób) opisanych tym samym zestawem cech, porównuje się między sobą, przydatne byłoby mieć test pozwalający odpowiedzieć na następujące pytania:

1. Czy obie (kilka) populacje różnią się między sobą?

2. O ile populacje nie różnią się, to czy opisujące je zestawy cech są trafnie dobrane?

3. W jaki sposób przydzielić do jednej z dwu badanych populacji obiekt bezwzględnie należący do jednej z nich, ale taki, że zaklasyfikowanie go nie jest oczywiste?

Tego rodzaju testem jest tzw. funkcja wyróżniająca (dyskryminująca). Analiza wyróżniająca jako metoda oraz funkcja wyróżniająca stosowana była w naukach geologicznych od 1949 r. W paleontologii zastosował ją B. H. Burma (1), a następnie w różnych dziedzinach geologii wielu uczonych amerykańskich, kanadyjskich, radzieckich i angielskich. Opis matematyczny metody można znaleźć u W. C. Krumbeina (1965). Na jego podstawie wprowadzono tę metodę do praktyki geologii w Polsce. W Polsce liniową funkcję wyróżniającą stosowano przede wszystkim do rozdzielania osadów czwartorzędowych. J. Rzechowski (4) oraz N. Lipińska zastosowali ją do rozdzielania glin zwałowych, K. Wuttke (6) — do testowania rozdzielczości powstałych w różnych warunkach populacji piasków eolicznych oraz I. Wiatr (7) — do rozdzielania osadów limnoglacialnych. Badania polskie były wykonywane na podstawie programów obliczeniowych P. Stenzla, K. Uścińskiej (5) oraz prac metodycznych I. Wiatra, P. Stenzla, A. Kowalskiego (8). Poniżej zostaną opisane założenia geologiczne, procedura obliczeniowa i wnioski dotyczące metodyki stosowania analizy wyróżniającej w geologii.

ZAŁOŻENIA GEOLOGICZNE

Zakłada się, że na podstawie wiedzy geologicznej możliwe są do ustalenia przedmiotowe populacje geologiczne, rozumiane jako zestawienia wyników pomiarów ilościowych lub ocen jakościowych (przekształconych operacyjnie), tj. cech charakteryzujących obiekty tworzące populacje. Rozpatrywany będzie przypadek, gdy określono dwie populacje geologiczne.

Przykład 1. Badaniu podlegają: populacja gliny zwałowej złodowacenia bałtyckiego oraz populacja gliny zwałowej złodowacenia środkowopolskiego. Populacje te są reprezentowane przez określone ilości

próbek pobranych z obu rodzajów glin zwałowych. Z kolei dla każdej próbki pomierzono np. procentowe zawartości frakcji: piaskowej i ilowej, wilgotność naturalną, ciężar objętościowy, edometryczny moduł ścisłości, wskaźnik Skempton. Liczbowe zestawienie wyników pomiarów procentowych zawartości frakcji oraz wybranych cech fizycznych będzie opisywało populacje glin zwałowych, czyli populacja gliny zwałowej złodowacenia bałtyckiego:

	próbka nr 1	próbka nr
x_1 piasek	62	
x_2 il	15	
x_3 wilgotn. nat.	11,9	
x_4 ciężar objęt.	2,26	
x_5 wskaźnik Skempton	0,567	
x_6 ścisłość	190	
a populacja gliny zwałowej złodowacenia środkowopolskiego		

	próbka nr 1,	próbka nr
x_1 piasek	67	
x_2 il	13	
x_3 wilgotn. nat.	10,5	
x_4 ciężar objęt.	2,20	
x_5 wskaźnik Skempton	0,800	
x_6 ścisłość	920	

W tym przykładzie uznano, że obie populacje są wystarczająco dobrze opisane przez zestawienie procentowych zawartości wybranych frakcji w glinach oraz niektórych cech fizycznych. Jak widać do opisanie populacji użyto tu cech, w których znalazła odbicie geneza powstania skał tworzących populacje. Zestaw tych cech może być dobierany dowolnie, w paleontologii będą to np. wymiary badanych skamieniałości lub charakterystyczne stosunki tych wymiarów. Przy rozdzielaniu lagunowych osadów wapiennych Link (2) użył jako cech m. in. wartości potencjału utleniająco-redukcyjnego i pH mierzonego w lagunie; przy badaniu kolektorów roponośnych używa się jako cech: przepuszczalności, porowatości, charakterystycznych wielkości amplitud anomalii pomiarowych z karotazu wykonanego przy badaniu np. piaskowców zawierających i nie zawierających ropy naftowej itd. Użycie cech jakościowych do analizy wyróżniającej jest możliwe pod warunkiem, że dadzą się one operacyjnie zmienić (zdefiniować) do postaci liczbowej. Bardzo często oceniamy np. częstość występowania danej cechy jakościowej (częstość występowania określonych skamieniałości w badanej serii skalnej, częstość występowania danej warstwy w określonej litofacji itd.).

W odniesieniu do skał populacje jednorodne mogą być określone kryteriami homogeniczności, a więc są to skały:

a) powstałe jako określona masa skalna w jednym okresie w wyniku jednego procesu lub jednego zespołu procesów geodynamicznych,

b) stanowiące jedną litofację,

c) stanowiące tę samą strukturę podlegającą w całej swojej objętości tym samym procesom geodynamicznym,

d) charakteryzujące się tą samą historią geologiczną.

Reasumując, wszędzie tam, gdzie dwie populacje geologiczne dają się opisać tym samym zestawem cech, w których ma odbicie geneza powstania tych populacji, można próbować dodatkowo do celów

rozdzielania populacji i celów klasyfikacyjnych użyć metody analizy wyróżniającej. Celem jej jest dostarczenie równania funkcji (powierzchni), za pomocą której badane populacje będą: a) rozdzielane; b) nowe próbki będą przyporządkowane do jednej z populacji czyli klasyfikowane.

Przed omówieniem samej procedury obliczeniowej należy przeanalizować sam problem podziału na populacje. Powyżej założono, że jest to podział aprioryczny oparty na wiedzy geologicznej. Podział ten może być pewny lub dyskusyjny, zależnie od stopnia rozpoznania problemu. Przydział próbek do populacji może także być pewny lub niepewny. Bardzo często ma miejsce sytuacja, gdy część próbek jest bezdyskusyjnie przydzielona do populacji, natomiast sklasyfikowanie części próbek jest niepewne. W końcu cechy pomierzone w próbkach mogą mieć dla obu populacji zdecydowanie różne wartości, czyli są zdolne do rozdzielania populacji, bądź też mogą mieć wartości podobne, co sprawia, że ich zdolność rozdzielczą jest mała. Reasumując można mieć pewny lub niepewny „pierwotny” podział na populacje, pewny lub niepewny przydział próbek do populacji oraz pełną lub niepełną zdolność rozdzielczą cech.

W różnych kombinacjach opisanej sytuacji analiza wyróżniająca ma różne zadania i inaczej jest interpretowana. W tabeli zestawiono powyższe sytuacje wraz z podaniem celu jaki można osiągnąć stosując analizę wyróżniającą. W dalszych częściach artykułu powyższe sytuacje będą omówione bardziej szczegółowo.

Przykład 1

Sytuacja	Podział na populacje	Przydział próbek do populacji	Zdolność rozdzielcza cech	Zadanie analizy wyróżniającej
I	pewny	pewny	pełna	klasyfikacja nowych próbek
II	pewny	pewny	niepełna	wyбір najbardziej „rozdzielczych” cech
III	pewny	częściowo niepewny	pełna	ocena wiarygodności próbek
IV	pewny	częściowo niepewny	niepełna	brak możliwości zastosowania

ANALIZA WYRÓŻNIAJĄCA

Dane są dwie populacje U i V . Liczebność próbek w populacji U wynosi n , a liczebność próbek w populacji V wynosi m . Na każdej z próbek dokonano pomiaru k — cech. Populacje U i V mają postać (patrz przykład 1)

$$U = \begin{bmatrix} U_{11} & \dots & U_{1n} \\ U_{21} & \dots & U_{2n} \\ \vdots & \ddots & \vdots \\ U_{k1} & \dots & U_{kn} \end{bmatrix} \quad \begin{matrix} \text{próbki} \\ \text{cechy } x_1, x_2, x_3 \dots x_k \end{matrix}$$

$$V = \begin{bmatrix} V_{11} & \dots & V_{1m} \\ V_{21} & \dots & V_{2m} \\ \vdots & \ddots & \vdots \\ V_{k1} & \dots & V_{km} \end{bmatrix} \quad \begin{matrix} \text{próbki} \\ \text{cechy } x_1, x_2, x_3 \dots x_k \end{matrix}$$

Cechy $x_1, x_2, \dots, x_k$ są niezależne i mają rozkład normalny w obu populacjach, czyli wartości np. dla cechy x_1 (w przykładzie 1 są to procentowe zawartości frakcji płaskowej w glinie zwałowej ze złodowacenia bałtyckiego i środkowopolskiego) mają rozkład normalny. Populacje U i V tworzą w przestrzeni k — wymiarowej (k — ilość cech) skupiska punktów. Należy znaleźć równanie powierzchni rozdzielającej w przestrzeni k — wymiarowej oba skupiska.

Oblicza się: macierz kowariancji

$$B = [b_{ij}] \quad i, j = 1 \dots k \quad [1] \\ [k \times k]$$

$$\text{gdzie } b_{ij} = \frac{1}{m+n-2} \left[\sum_{i=1}^n (U_{ii} - \bar{U}_i)(U_{ij} - \bar{U}_j) + \sum_{j=1}^m (V_{ji} - \bar{V}_i)(V_{jj} - \bar{V}_j) \right] \quad [2]$$

macierz odwrotną

$$\bar{U}_i = \frac{1}{n} \sum_{i=1}^n U_{ii} \quad \bar{V}_i = \frac{1}{m} \sum_{i=1}^m V_{ii} \quad [3]$$

$$B^{-1} = O = [O_{ij}] \\ [k \times k]$$

oraz k współczynników $a_1, a_2, \dots, a_k$

$$a_p = \sum_{j=1}^k O_{pj} (\bar{U}_j - \bar{V}_j) \quad [4]$$

Liniowa funkcja wyróżniająca $D(x_1, x_2, x_3, \dots, x_k)$ jest dana równaniem:

$$p = 1, 2, \dots, k$$

$$D(x_1, x_2, \dots, x_k) = \sum_{p=1}^k a_p x_p = a_1 x_1 + a_2 x_2 + \dots + a_k x_k \quad [5]$$

a jej wyróżniona wartość:

$$D_0 = \frac{1}{2} \sum_{p=1}^k a_p (\bar{U}_p + \bar{V}_p) \quad [6]$$

Jeżeli wyniki $x_1, x_2, \dots, x_k$ (w przykładzie 1 — procentowe zawartości poszczególnych frakcji) z jakiegokolwiek nowej próbki podstawimy do równania [5] i otrzymamy się $D > D_0$, to próbka należy do populacji V , w przypadku przeciwnym próbka należy do populacji U .

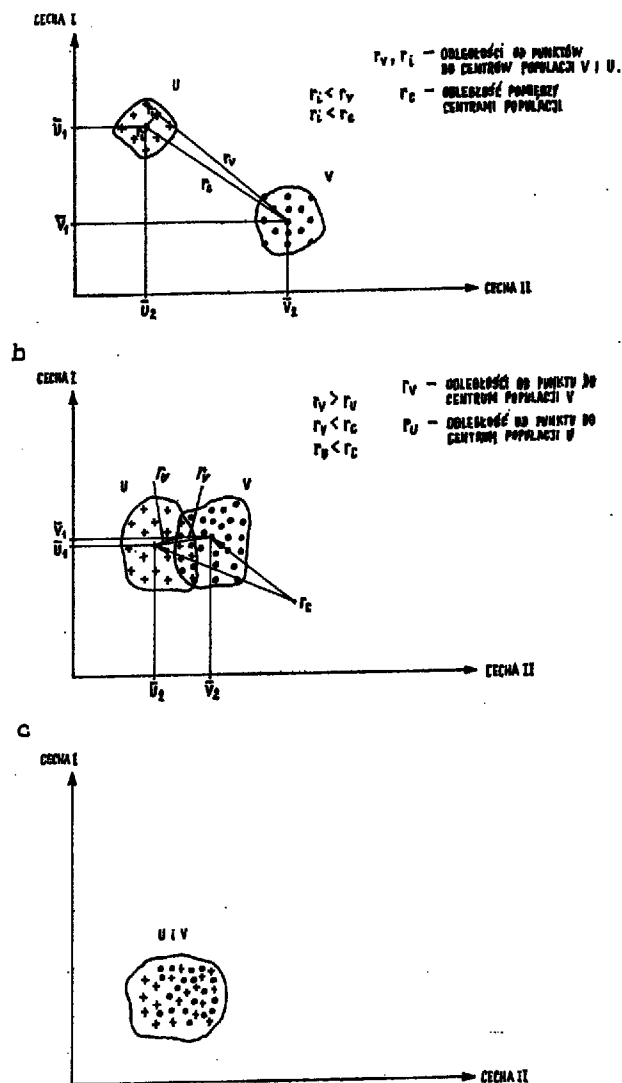
Jak widać na podstawie danych uznanych za pewne, opisujących populacje U i V wyznaczono współczynniki $a_1, a_2, \dots, a_k$, co pozwala stworzyć funkcję D , która z kolei rozdziela obie populacje oraz służy do klasyfikacji nowych próbek za pomocą wartości D_0 . Testowanie, czy funkcja wyróżniająca rozdziela prawidłowo lub nie rozdziela obu populacji, polega na postawieniu hipotezy zerowej $H_0: U_u = U_v$, co oznacza, że wektor średnich z populacji U oznaczony U_u jest równy wektorowi średnich z populacji V oznaczonego jako U_v . Używając danych z przykładu 1 H_0 oznacza, że średnie z np. procentowej zawartości frakcji płaskowej dla obu rodzajów glin zwałowych są sobie równe (lub prawie równe) itd. Gdy tak jest (hipoteza H_0 nie jest odrzucona) populacje

glin są nierozdzielcze, natomiast gdy średnie się różnią populacje można rozdzielić.

Do testowania służy wyrażenie:

$$q = \frac{m n (m + n - k - 1)}{k (m + n) (m + n - 2)} \sum_{j=1}^k \sum_{i=1}^k d_{ij} (\bar{U}_i - \bar{V}_j) \times [7] \\ \times (\bar{U}_j - \bar{V}_j)$$

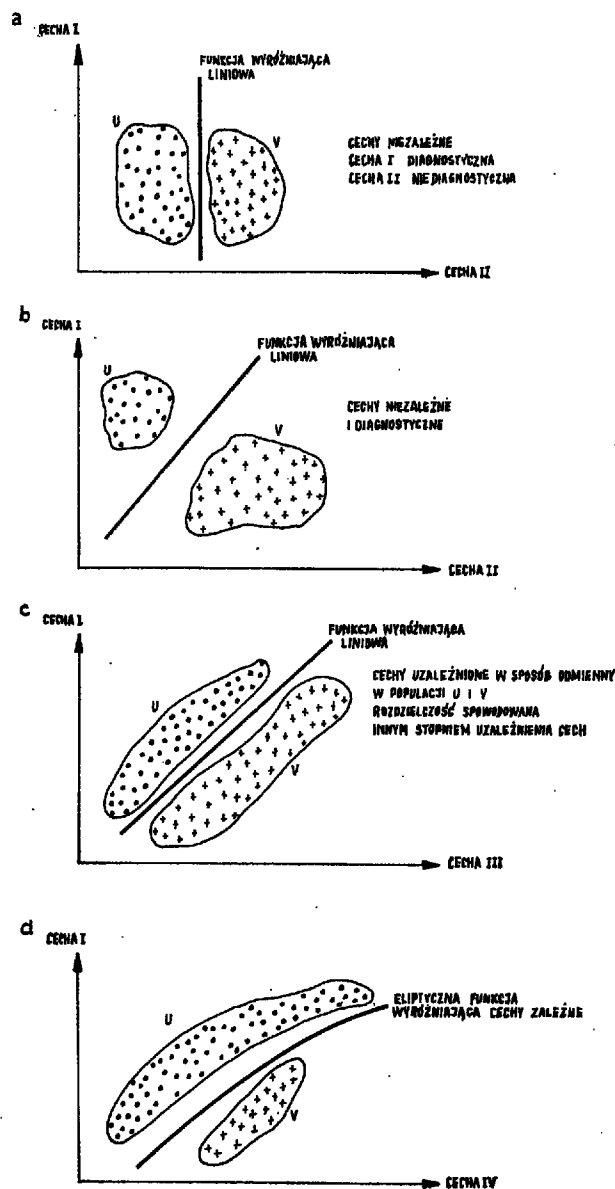
gdy $q > F$ (F — Snedecora dla k oraz $m+n-k-1$ stopni swobody) to H_0 jest odrzucona, czyli populacje można rozdzielać.



Ryc. 1. Graficzny obraz populacji rozdzielczych (a), częściowo rozdzielonych (b), nierozdzielonych (c).

Fig. 1. Graphic representation of: /a/ separated, /b/ partly separated, and /c/ not separated populations.

Poniżej przeprowadza się dyskusję dotyczącą właściwości funkcji wyróżniającej. Jak wspomniano populacje U i V tworzą w przestrzeni k — wymiarowej skupiska punktów. Dla ułatwienia dalszą dyskusję prowadzić się będzie przy założeniu, że są dwie cechy ($k=2$). Pierwszym zagadnieniem do przeanalizowania jest rozdzielczość populacji przy wykorzystaniu cech. Kiedy populacje dają się rozdzielić? Na ryc. 1 przedstawiono 3 sytuacje: a — dla populacji dających się rozdzielić, b — dla populacji częściowo rozdzielczych, c — dla populacji nierozdzielnych.

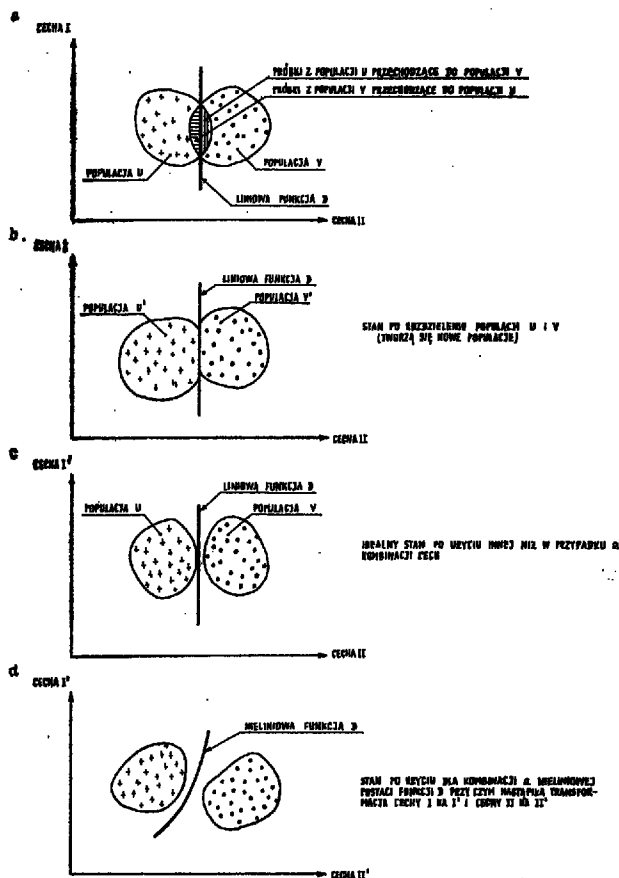


Ryc. 2. Rozdzielenie populacji a współzależność cech.

Fig. 2. Separation of populations versus interdependence of properties.

Z ryc. 1 wynika, że gdy dla obu badanych populacji ustalimy ich centra (będą to punkty o współrzędnych równych średnim arytmetycznym z cech), to każda z próbek z populacji U leży bliżej centrum populacji U niż centrum populacji V, wtedy gdy są one całkowicie rozdzielone. Przy częściowej rozdzielczości obu populacji część próbek z populacji U leży bliżej centrum populacji V niż centrum populacji U, podobnie część próbek z populacji V może leżeć bliżej centrum populacji U niż centrum populacji V. Dla populacji nierozdzielczych większość próbek zaliczonych do populacji U leży w obrębie populacji V i odwrotnie. Gdy populacje są rozdzielcze można przeprowadzić między nimi funkcję wyróżniającą (sytuacja I, tab. I), gdy populacje częściowo rozdzielcze można, na podstawie przeprowadzonej funkcji wyróżniającej analizować, czy fakt innego niż pierwotny zaliczenia próbki lub próbek jest spowodowany niepewnym przydziałem próbek do populacji, bądź też słabą rozdzielczością cech użytych do opisu populacji (sytuacje II i III z tab. I). W praktyce przyjmuje się, że gdy 5—10% próbek „nie zgadza się” z pierwotną klasyfikacją, populacje nadają się do rozdzielania.

Następnym problemem, który należy rozważyć jest niezależność cech opisujących populacje. Teoria



Ryc. 3. Graficzny obraz rozważań dotyczących możliwości rozdzielenia populacji geologicznych.

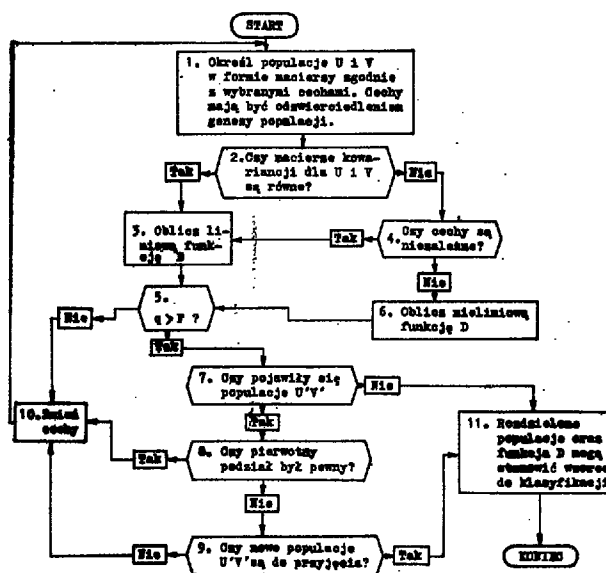
Fig. 3. Graphic representation of considerations of the possibility of separation of geological populations.

statystyczna wymaga by wszystkie cechy były względem siebie niezależne. W praktyce geologicznej cechy opisujące realne populacje geologiczne są często od siebie uzależnione. Czy oznacza to, że w większości przypadków użycie w geologii funkcji wyróżniającej nie jest możliwe?

Na podstawie doświadczeń ustalono, że o sukcesie w większym stopniu przesądza zdolność diagnostyczna cech do rozdzielania populacji niż ich wzajemne uzależnienie. Na ryc. 2a przedstawiono sytuację, gdy tylko jedna z dwu cech jest diagnostyczna, a obie są niezależne względem siebie, na ryc. 2b obie cechy są i diagnostyczne, i niezależne; na ryc. 2c i 2d cechy są uzależnione od siebie, ale uzależnienie to ma różny charakter (inną postać matematyczną) w populacji U niż w populacji V, co pozwala rozdzielić obie populacje na ryc. 2c za pomocą liniowej, a na ryc. 2d za pomocą eliptycznej funkcji wyróżniającej.

Reasumując, o ile cechy mają charakter wyraźnie diagnostyczny a więc gdy są niezależne, lub gdy ich uzależnienie jest tego rodzaju, że nie powoduje „nakładania” na siebie skupisk będących obrazem populacji, zastosowanie funkcji wyróżniającej jest możliwe. Na podstawie wzorów [1-7] oblicza się więc dla populacji U i V współczynniki a_p ($p=1, 2, \dots, k$) ustala postać funkcji, jej wyróżnioną wielkość D_0 oraz wielkość q . Gdy $q > F$, to populacje dają się rozdzielić. Postać druku z komputera (jako dane służą macierz U i V) jest następująca (wyniki obliczeń są fikcyjne, przykład 2 obok).

W przykładzie 2 (fikcyjnym) są dwie populacje: U i V; w populacji U było 11 próbek o numerach: 3, 4, 5, 6, 7, 8, 12, 15, 20, 16, 19, a w populacji V — 9 próbek o numerach: 1, 2, 9, 10, 11, 13, 14, 17, 18. Na każdej próbce zmierzono 3 cechy: x_1, x_2, x_3 .



Ryc. 4. Diagram czynności i decyzji przy stosowaniu analizy wyróżniającej.

Fig. 4. Diagram of procedure and decisions made in using of discriminant analysis.

Przykład 2

Kolejny numer	Wartości D		Nr próbki pierwotnie przyporządkowanej do:	
	populacja U	populacja V	populacji U	populacji V
1	4,0		4	
2	5,0		5	
3	6,0		3	
4	6,5		19	
5	7,0		6	
6	7,1		7	
7		8,0	—	1
8	8,0		12	
9		8,1	—	2
10	8,2		15	
11	8,3		8	
12	8,5		16	
	$D_0=9,0$			
13		9,1		9
14		9,2		10
15		10,0		11
16		11,2		13
17	11,5		20	
18		11,7		14
19		13,0		17
20		15,0		18

$q = \dots$

$$D = a_1x_1 + a_2x_2 + a_3x_3$$

W wyniku obliczeń otrzymano wartości liczbowe na: a_1, a_2, a_3 (w przykładzie nie podano liczb) oraz przydział próbek do populacji U i V. Z wyników jest widoczne, że próbki o numerach 1 i 2 przeszły z populacji V do U, a próbka nr 20 z populacji U do V. Populacja U' ma więc próbki o numerach: 1, 2, 3, 4, 5, 19, 6, 7, 12, 15, 8, 16, a populacja V' próbki o numerach: 9, 10, 11, 13, 14, 17, 18, 20.

Na ryc. 3 powyższy przykład przedstawiono graficznie, używając dla uproszczenia tylko dwu cech, cecha I= x_1 i cecha II odpowiada x_2 . Ryc. 3a przedstawia stan obu populacji w chwili podziału za pomocą liniowej funkcji wyróżniającej D. Na ryc. 3b zobrazowano nowe populacje U', V'. W tym momencie należy przeprowadzić analizę merytoryczną

pod kątem możliwości zamiany populacji U w U' i V w V' . Innymi słowy należy odpowiedzieć na pytanie, czy jest dopuszczalne by ze zdefiniowanej uprzednio populacji U można było wziąć próbkę nr 20, a dołączyć próbkę nr 1 i nr 2 oraz czy z populacji V można wziąć próbkę nr 1 i nr 2, a dołączyć do niej próbkę nr 20? Jeżeli jest to dopuszczalne to cel został osiągnięty i populacje U' i V' oraz funkcja D mogą służyć jako wzorce klasyfikacyjne do tego problemu geologicznego. Gdy nie jest to dopuszczalne należy zrewidować albo dobór cech $x_1, x_2, \dots, x_n$, bo być może na innym zestawie cech osiągnie się sukces — ryc. 3c, bądź też zmienić klasę funkcji D , bo może przyjęta klasa funkcji nie jest optymalna do rozdzielania populacji. Na ryc. 3d przedstawiono stan, gdy zmieniono: cechy (I na I', II na II'') oraz klasę funkcji z liniowej na eliptyczną.

Czynności jakie należy wykonać stosując funkcję wyróżniającą można przedstawić w postaci diagramu (ryc. 4), który stanowi metodologiczną instrukcję dla geologa i dla ośrodka obliczeniowego. Tylko geolog może wykonać czynności nr: 1, 8, 9, 10, a tylko za pomocą komputera można wykonać efektywnie czynności nr: 2, 3, 4, 5, 6.

WNIOSKI

1. Analiza wyróżniająca jest techniką, która może być przydatna w geologii do rozdzielania obiektów, ciał, podprzestrzeni opisanych zestawami cech.
2. Efektywne zastosowanie tej techniki wymaga wyboru i pomiaru cech obrazujących genezę powstania populacji i mających własności diagnostyczne.

SUMMARY

This article deals with the application of discriminant function in geological sciences. Multivariate discriminant analysis is a statistical method of assigning samples to previously defined populations on the basis of a number of variables considered simultaneously.

LITERATURA

1. Burma B. H. — Studies in quantitative paleontology, II: multivariate analysis — a new analytical tool for paleontology and geology. Jour. Paleontol. vol. 23, 1949.
2. Link A. J. — A physico-chemical and textural study of carbonate sedimentation in a lagoonal environment, unpublished M. A. thesis. Northwestern University, 1964.
3. Lipińska N. — Zastosowanie wielozmiennej funkcji wyróżniającej do rozdzielania dwóch poziomów glin zwałowych obszaru miasta Szczecina i okolic. Badania fizjograficzne nad Polską Zachodnią, (w druku).
4. Rzechowski J. — Charakterystyka litofacjalna najważniejszych poziomów stratygraficznych mezopleistocenu Niżu Polskiego. Arch. IG, 1971.
5. Stenzel P., Uścińska K. — Zastosowanie metod matematycznych i ETO w geofizyce i geologii. Pr. Post. techn. Arch. PPG, 1970.
6. Wuttke K. — Zastosowanie funkcji wyróżniającej w analizie porównawczej wybranych osadów sedimentacji eolicznej. Arch. Inst. Hydrol. i Geol. Inż. UW. 1972.
7. Wiatr I. — Model statystyczny wybranych cech środowiska inżyniersko-geologicznego kopalnych dolin Przykory i Małgorzaty w okolicy Turka. Biul. geol. UW, 1973, nr 15.
8. Wiatr I., Stenzel P., Kowalski A. — Metodyka zastosowania funkcji wyróżniającej w populacjach przedmiotowych. Arch. Inst. Hydrol. i Geol. Inż. UW. 1971.

РЕЗЮМЕ

Статья касается применения дискриминантной функции в геологических исследованиях. Дискриминантный анализ представляет статистический метод определения принадлежности объектов, в том числе и геологических, к предварительно установленным совокупностям (к геологическим совокупностям в геологии).