INEZ WIATR, PRZEMYSŁAW STENZEL

Uniwersytet Warszawski, Instytut Geologiczny

ANALIZA CZYNNIKOWA – SPOSÓB R

UKD 550.8;519.2

Środowisko geologiczne, przez które rozumie się wydzielony wycinek realnej przestrzeni trójwymiarowej, może być opisane za pomocą zestawienia wielkości szeregu cech (miąższość, głębokość stropu, spągu, rodzaje skał, własności fizyczne skał itp.). Użycie słowa „zestawienie" ma ten sens, że w większości przypadków nie jesteśmy w stanie podać postaci modelu środowiska uwzględniając jednocześnie wszystkie jego cechy. Ponadto nie jesteśmy pewni, które i ile z rozpatrywanych cech są „wiodącymi", tzn. decydują o aktualnej postaci środowiska. Powyższe stwierdzenia mają swoje źródło w złożoności natury procesów geologicznych, a także w fakcie, że badane środowisko zostało ukształtowane przez szereg procesów, działających w jednym lub różnym czasie. Reasumując środowisko geologiczne ma ukrytą strukturę. Rozszyfrowanie tej struktury może m.in. być dokonane na drodze ustalenia, które z mierzonych cech mają decydujący wpływ na jego obecną postać, oraz na ustaleniu wzajemnego oddziaływania tych cech. Idea postępowania sprowadza się do budowy modelu statystycznego środowiska. Zakłada się, że w samych wartościach cech zawarta jest struktura modelu. Założenie to wymaga więc rozpatrywania cech wzajemnie niezależnych, przynajmniej w sposób jawny, lub też do doprowadzenia do budowy modelu opartego na zmiennych istotnie nieskorelowanych. Zagadnienie tak postawione może być rozwiązane za pomocą tzw. analizy czynnikowej sposób R. Przedstawiony artykuł jest popularyzatorską próbą opisu tej metody w zastosowaniu do analizy środowisk geologicznych.

ZAGADNIENIE ANALIZY CZYNNIKOWEJ

W rozdziale podaje się geologiczne podstawy zastosowania metody oraz schematyczny opis matematycznej idei zagadnienia.

Założenia:

1. Analizowane środowisko jest środowiskiem jednorodnym. Może to być określony wycinek przestrzeni geologicznej lub populacja cech określonego rodzaju skały.
2. Środowisko jest poznane przez pobranie z niego odpowiedniej ilości prób (jako próbę można tu też traktować cały profil wiertniczy lub jego określoną część). Próby te w dalszej części tekstu zwane są obiektami.
3. W każdej próbie (obiekcie) mierzy się określoną (dowolną) ilość cech. Są to dane ilościowe pozwalające w swoim zestawieniu na wyodrębnienie badanego środowiska z szeregu innych. Do badania używać się będzie cech wzajemnie od siebie w sposób jawny niezależnych (określonym wzorem).
4. Zakłada się, że istnieją hipotetyczne czynniki zwane też metacechami takie, że: a) ilość ich jest mniejsza od ilości cech, b) w jednym czynniku „działają" jedna, dwie lub więcej cech, c) cechy „działające" w czynniku są z nim liniowo skorelowane, d) czynniki opisują strukturę modelu, a że są oparte na realnych pomierzonych cechach, opisują strukturę środowiska traktowanego jako model statystyczny. Wyznaczone czynniki muszą być nieskorelowane (niezależne od siebie).
5. Ilość obiektów jest większa od ilości cech mierzonych w obiekcie.

Zadaniem, jakie stoi przed nami, jest:

1) określenie ilości czynników,
2) interpretacja czynników w sensie geologicznym,
3) weryfikacja możliwości zastosowania tej techniki postępowania dla badania konkretnego środowiska.

Poniżej omawia się szerzej zarówno założenia, jak i sposób postępowania. Sprecyzowanie celu badania, jako analizy struktury, ogranicza rozważania do środowisk jednorodnych w sensie geologicznym, co za W. C. Kowalskim (4) sprowadza się do tego, by środowisko takie było przestrzennie ograniczone, powstało jako masa skalna w jednym okresie czasu, stanowiło tę samą litofację i podlegało w całej swej objętości działaniu tych samych procesów geodynamicznych. Są to założenia bardzo ostre, gdyż badacz musi zdawać sobie sprawę, że środowisko takie powinno być także możliwie częściowo jednorodne w sensie fizycznym, co nie zawsze pokrywa się z kryteriami geologicznymi podziału na środowiska jednorodne. Przykładowo określona masa skalna zaliczana do litofacji „a" na obszarze A może w istotny sposób różnić się od tej samej (w sensie opisu litologicznego) masy skalnej z takiej samej litofacji, ale na obszarze B, co wynikło np. z innej intensywności działania procesów geologicznych na obu obszarach.

Środowisko geologiczne reprezentowane będzie przez zestaw prób o odpowiedniej wielkości i jakości. Rozumie się przez to, że zestawem prób będzie np. zbiór wierceń odpowiednio liczny (w każdej badanej populacji ilość wierceń musi być większa od 30), a przez jakość zestawu prób rozumiec się będzie w tym przypadku takie rozłożenie wierceń w przestrzeni, że charakteryzują one główne cechy badanego środowiska. W próbie dokonujemy pomiarów cech, czyli każde wiercenie lub warstwa w wierceniu jest charakteryzowane pewnym zestawem cech. Dobór tych cech musi być taki, by odzwierciedlały one genezę środowiska, a jednocześnie by nie były one powiązane ze sobą jawnymi zależnościami funkcyjnymi. Przykładowo przy badaniu gruntów spoistych można w wierceniu określić jako cechy np.: miąższość utworów, wartość frakcji uziarnienia, granicę płynności, granicę plastyczności i wilgotność naturalną, natomiast takie parametry, jak wskaźnik plastyczności jako określone jawną formułą funkcyjną nie są tu brane pod uwagę.

W geologii zagadnienie niezależności cech nie jest zagadnieniem oczywistym. Można zaryzykować stwierdzenie, że środowisko geologiczne jako pewnego rodzaju układ zamknięty tworzy całość wzajemnie uwarunkowaną co oznacza, że zawsze istnieje jakaś zależność choćby niejawna lub nam nie znana pomiędzy dowolnymi parami cech. W naszym przypadku ograniczymy się do wyeliminowania tych cech, które w jawny sposób wynikają z innych. W przypadkach wątpliwych można na tym etapie koncepcyjnej budowy modelu przeprowadzić testowanie statystyczne (test Blomquista, test χ^2) na niezależność badanych par cech w populacji prób (obiektów). Tak więc środowisko geologiczne będzie reprezentowane przez zestaw pomiarów, który można zapisać w postaci:

$$
\begin{array}{l}
\text{Cechy} \\
\text{1 obiekt} \\
\text{2 obiekt} \\
\\
\text{i-ty obiekt} \\
\\
\\
n\text{-ty obiekt}
\end{array}
\begin{array}{c}
\quad 1 \qquad 2 \qquad\quad \text{j-ota} \qquad\quad \text{m-ta} \\
\begin{bmatrix}
X_{11}, & X_{12}, & \ldots, & X_{1j}, & \ldots\ldots, & X_{1m} \\
X_{21}, & X_{22}, & \ldots, & X_{2j}, & \ldots\ldots, & X_{2m} \\
\vdots & \vdots & & \vdots & & \vdots \\
X_{i1}, & X_{i2}, & \ldots, & X_{ij}, & \ldots\ldots, & X_{im} \\
\vdots & \vdots & & \vdots & & \vdots \\
\vdots & \vdots & & \vdots & & \vdots \\
X_{n1}, & X_{n2}, & \ldots, & X_{nj}, & \ldots\ldots, & X_{nm}
\end{bmatrix}
\end{array}
\qquad [1]
$$

numer cechy numer obiektu

Mamy więc n-obiektów np. wierceń, a każde z nich charakteryzuje m-cech tych samych dla wszystkich wierceń, ale o różnych wielkościach liczbowych. Z założenia $n \geqslant m$.

Zestawienie tego typu nazywa się macierzą.

Gdybyśmy teraz przeprowadzili badanie w jaki sposób następuje „klasyfikacja" obiektów, gdy weźmiemy pod uwagę zmiany wielkości odrębnie cechy 1, cechy 2 itd., to może się okazać, że np. cecha 1 i cecha 5 różnicują obiekty podobnie. Dla jednych i tych samych obiektów wielkości cechy 1 np. są duże, a cechy 5 małe itd. Stwierdzenie takie leży u podstaw rozumowania że, po pierwsze cechy te są ze sobą skorelowane, a po drugie, że jest do pomyślenia jakaś metacecha obejmująca w sobie działanie tych dwu cech jednocześnie. W geologii taka metacecha będzie reprezentowała określony proces geologiczny lub wynik jego działania wtedy, gdy wprowadzone do analizy cechy opisują genezę środowiska. Analiza czynnikowa obejmuje więc grupę metod matematyczno-statystycznych pozwalających na ustalenie hipotetycznych metacech zwanych czynnikami. Czynniki te z jednej strony mogą opisać strukturę środowiska w aspekcie jego genezy, z drugiej służą do ustalania przyczyn zmienności środowiska. Ponadto zawierają one informację opartą o oryginalne cechy, z tym że opierają się na założonych korelacjach pomiędzy cechami w zbiorze obiektów oraz na założeniu, że istnieje liniowa korelacja pomiędzy cechami a czynnikami (w naszym przykładzie cecha 1 i cecha 5 są liniowo skorelowane z metacechą, np. nr 1). Powyższe stwierdzenie można zapisać następująco:

$$X_{ji} = a_{ji}F + e_{ji} \qquad [2]$$

gdzie: X_{ji} — wielkość cechy i-tej, w próbie j-tej,
F — czynnik np. pierwszy (metacecha nr 1),
e_{ji} — błąd losowy,
a_{ji} — współczynnik, zwany ładunkiem czynnikowym dla cechy j-tej, w próbie i-tej,
i — 1, 2,..., n
j — 1, 2,..., m

Współczynnik a_{ji} stanowi więc pewnego rodzaju wagę, jaką przypisujemy dla związku łączącego wartość cechy X_{ji} z czynnikiem F. Gdy jest on duży w sensie liczbowym związek jest silny i odwrotnie.

Zgodnie z postawionym zadaniem należy określić ile jest czynników F oraz przeprowadzić ich identyfikację. Przedstawiony poniżej szkic rozumowania matematycznego ma za zadanie wyjaśnić ideę postępowania i nie pretenduje do opisu matematycznej strony metody.

Został on zamieszczony dla czytelników pragnących zorientować się w sposobie postępowania. Operacje opisane w tej części tekstu są w całości wykonywane przez maszynę cyfrową, tak że specjaliście pozostaje interpretacja i identyfikacja czynników, co opisano w następnym rozdziale.

Mamy za zadanie scharakteryzować środowisko z uwzględnieniem jego zmienności. Środowisko reprezentowane jest przez macierz [1]. Każda kolumna tej macierzy reprezentuje cechę. Chcąc zbadać siłę związków zachodzących pomiędzy cechami, co pośrednio świadczy o strukturze modelu, obliczamy zgodnie z wzorem Pearsona macierz R współczynników korelacji pomiędzy cechami:

$$r_{jk} = \frac{\sum\limits_{i=1}^{n} (X_{ij} - \overline{X}_j)(X_{ik} - \overline{X}_k)}{\sqrt{\sum\limits_{i=1}^{n} (X_{ij} - \overline{X}_j)^2 \sum\limits_{i=1}^{n} (X_{ik} - \overline{X}_k)^2}} \qquad [3]$$

gdzie: r_{jk} — współczynnik korelacji pomiędzy cechą j a cechą k,
$\overline{X}_j$ — średnia arytmetyczna z j-tej kolumny macierzy [1],
$\overline{X}_k$ — analogicznie,
pozostałe oznaczenia jak we wzorze [1].

Macierz R ma więc postać:

$$\begin{bmatrix}
r_{11}, & r_{12}, & \ldots, & r_{1k}, & \ldots, & r_{1m} \\
r_{21}, & r_{22}, & \ldots\ldots\ldots\ldots, & & & r_{2m} \\
r_{j1}, & r_{j2}, & \ldots, & r_{jk}, & \ldots, & r_{jm} \\
r_{m1}, & r_{m2}, & \ldots, & r_{mk}, & \ldots, & r_{mm}
\end{bmatrix} \qquad [4]$$

gdzie m — ilość cech w macierzy.

Ponieważ macierz ta opisuje strukturę środowiska (modelu) zawiera w sobie i czynniki, które chcemy wyodrębnić, a przede wszystkim określić ich ilość. Na mocy założeń 4a, 4b i 4d możemy oznaczyć ilość czynników wykorzystując jedno z twierdzeń rachunku macierzowego. Mówi ono, że tzw. wektory rachunmacierzy są liniowo niezależne, gdy odpowiadające im wartości własne z tej macierzy są różne.

Przyjmując powyższe bez dowodu wyjaśniamy tylko, że w naszym przypadku ilość wartości własnych dla macierzy R wynosi m. Istnieją znane w metodach numerycznych sposoby obliczania zarówno wartości, własnych, jak i wektorów własnych macierzy. Nas interesuje (zakładając, że obliczenia wykonamy standardowym programem przy pomocy e.m.c.) tylko ilość wartości własnych różnych od siebie. Z reguły kilka z tych wartości ma wielkość równą lub zbliżoną do O. Pozostała więc (ich ilość), a jest ich mniej niż m wyznacza ilość czynników F (oznaczamy ją literą t) na mocy przytoczonego twierdzenia i założenia 4b. W ten sposób pozostaje do wyjaśnienia identyfikacja tych czynników.

Dla uproszczenia dalszych rozważań przyjęto, że macierz [1] zostanie przekształcona w macierz, której elementy są znormalizowane. Normalizacja ta polega na zastosowaniu następującego wzoru:

$$Z_{ji} = \frac{X_{ji} - \overline{X}_j}{S_j} \qquad [5]$$

gdzie

$$S_j = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_{ji} - \overline{X}_j)^2}$$

czyli kolejno oblicza się dla 1 kolumny macierzy [1] średnią arytmetyczną, odchylenie standardowe (wtedy $j = 1$) i wstawia do wzoru [3] otrzymując pierwszą kolumnę macierzy Z (wyrazy Z_{ji}) itd. Sens tej operacji jest taki, ażeby po pierwsze móc porównywać ze sobą cechy mierzone różnymi jednostkami (cm, g/cm³ itp.), a po drugie, by sprowadzić cechy do układów, w których średnia arytmetyczna z cechy jest równa O, a jej odchylenie standardowe równa się 1.

Pozwoli to nam bowiem na uproszczenie dalszego rozumowania.

Używając nomenklatury z równania [5] i biorąc pod uwagę, że czynników F będzie pewnie więcej jak 1 oraz pamiętając o tym, że wszystkie cechy mają być rozważane, a nie tylko jedna jak w przypadku równania [2], zapiszmy układ równań będący sformalizowanym przedstawieniem zagadnienia analizy czynnikowej:

$$\begin{aligned} Z_1 &= a_{11}F_1 + a_{12}F_2 + \dots a_{1t}F_t \\ &\vdots \\ Z_m &= a_{m1}F_1 + a_{m2}F_2 + \dots a_{mt}F_t \end{aligned} \qquad [6]$$

gdzie: a, F — oznaczenia jak we wzorze [2]

czyli pierwsza i każda cecha może być przedstawiona jako kombinacja $t(F_1 \dots F_t)$ czynników (przy czym $t < m$), a związana jest z czynnikami za pomocą wielkości $a_{11}, a_{12} \dots a_{1t} \dots a_{n1} \dots a_{mt}$. W tym układzie równań jak narazie znane są tylko wartości $Z_1 \dots Z_m$ (na podstawie równania [5]) oraz wielkość t. Chcąc znaleźć wielkości $a_{11} \dots a_{mt}$, ażeby na ich podstawie móc interpretować co właściwie oznacza $F_1 \dots F_t$, należy znaleźć jakieś wyrażenie (wzór) na ich obliczanie. Bez dowodu i wyprowadzenia podamy, że prawdziwa jest zależność:

$$r_{jk} = a_{j1}\, a_{k1} + a_{j2}\, a_{k2} + \dots + a_{jt}\, a_{kt} \qquad [7]$$

gdzie: $k,j = 1,2 \dots m$

co można zapisać także:

$$r_{jk} = \sum_{p=1}^{t} a_{jp}\, a_{kp} \qquad [8]$$

Jest to podstawowe równanie analizy czynnikowej.

W równaniu [8], a właściwie całym układzie równań mamy t czynników do oznaczenia. Chcąc oznaczyć czynnik F_1 należy znaleźć wielkości liczbowe na wyrażenia (równania 6) $a_{11}, a_{21} \dots a_{m1}$. Szkic postępowania jest następujący: ponieważ F_1 reprezentuje jakąś część zmienności ogólnej w modelu, to przekształca się macierz R w ten sposób, że za wielkość $r_{11}, r_{22} \dots r_{ij} \dots r_{mn}$, które były poprzednio równe 1 podstawia się wyrażenia $h_j^2 - \lambda$.

gdzie: $$h_j^2 = \frac{r_{jk} r_{jl}}{r_{kl}}, \qquad j = 1, 2 \dots m \qquad [8a]$$

a r_{jk} i r_{kl} są to dwa największe współczynniki korelacji zmiennej Z_j. λ — określa pośrednio część zmienności dla F_1,

mamy więc (bez dowodu i wyprowadzenia) równanie:

$$\underset{R'}{\text{macierz}} \begin{bmatrix} h_1^2 - \lambda, & r_{12} & \dots & r_{1m} \\ r_{21}, & h_2^2 - \lambda & \dots & r_{2m} \\ \vdots & \vdots & & \\ r_{m1}, & r_{m2} & \dots & h_m^2 - \lambda \end{bmatrix} \times \begin{bmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{m1} \end{bmatrix} = 0 \qquad [9]$$

o nieznanych: λ i wektorze $[a_{11} \dots a_{m1}]$

Ażeby równanie [9] miało jako rozwiązanie $a_{11} \dots a_{m1}$ różne od O należy macierz R' przyrównać do O obliczając w ten sposób λ, a potem wektor $[a_{11} \dots a_{m1}]$. Sposoby obliczeń równań tego typu znane są z podręczników metod numerycznych. Analogiczne postępowanie prowadzi się dla każdego czynnika F operując zmienianą odpowiednio macierzą R'.

Po wykonaniu wszystkich obliczeń mamy w równaniu [6] określone wszystkie liczbowe wielkości na wyrażenia a_{ij} ($i = 1 \dots m$, $j = 1 \dots t$). Pełny matematyczny opis metody zamieszczono w publikacjach podanych w literaturze.

W zasadzie zadanie zostało wykonane, każdy czynnik w stosunku do każdej cechy jest liczbowo przyporządkowany. Pozostaje część najważniejsza dla specjalisty, tzn. identyfikacja i interpretacja czynników.

IDENTYFIKACJA I INTERPRETACJA CZYNNIKÓW

Proces identyfikacji czynników, które są zmiennymi hipotetycznymi — niemierzalnymi, oraz interpretację ich znaczenia można wykonać w dwu etapach. Etap pierwszy obejmuje analizę wielkości ładunków czynnikowych wiążących poszczególne cechy z czynnikami.

Rozpatrzmy to na przykładzie geologicznym. Przykład ten celowo został zapożyczony od Imbriego (1963 vide 5), gdyż dotyczy on typowego przypadku zastosowania analizy czynnikowej do badania budowy przestrzeni geologicznej w aspekcie rodzajów skał ją tworzących i wnioskowania o procesach geologicznych, które tę przestrzeń uformowały. W przytoczonym przykładzie dane liczbowe (macierze 10, 11, 12) oraz ostateczna identyfikacja procesów geologicznych została zaczerpnięta z Krumbeina (5). Opis interpretacji oraz wskazówki metodyczne dotyczące jej wykonywania są dziełem autorów.

Imbrie bada pewien rejon w stanie Kansas rozpoznany 31 otworami wiertniczymi. W każdym z nich pomierzono: głębokość całkowitą (do stropu permu) oraz sumaryczne miąższości: piaskowców, łupków, skał nieklastycznych, węglanowych oraz ewaporytów. Macierz [1] ma więc 31 wierszy i 6 kolumn. Nawiasem mówiąc dobór cech nie był najlepszy, gdyż można było zrezygnować z miąższości skał nieklastycznych równej miąższości skał węglanowych i ewaporytów oraz z głębokości całkowitej, równej sumie miąższości piaskowców, łupków i skał nieklastycznych. Są więc to cechy w sposób jawny związane z pozostałymi, ale przykład Imbriego ma na celu między innymi uwypuklenie i tej zasady, że nawet gdy wprowadzi się takie cechy do rozważań nie wpłynie to na jakość rozwiązania zagadnienia. Zgodnie z założeniami opisanymi powyżej lepiej jest jednak unikać wprowadzania do rozważań cech tego typu. Obliczona na podstawie macierzy [1] i wzoru [3] macierz korelacji ma wymiar 6 × 6 i postać:

Analizując macierz R można ustalić, że: głębokość całkowita jest silnie skorelowana z miąższościami łupków, skał nieklastycznych i ewaporytów, a słabo z miąższościami piaskowców i skał węglanowych, piaskowce są skorelowane tylko i to nienajsilniej z miąższościami skał węglanowych; łupki są silnie skorelowane z głębokością całkowitą i średnio z miąższościami skał nieklastycznych i ewaporytów, a słabo z pozostałymi cechami; skały nieklastyczne najsilniej skorelowane są z ewaporytami, a następnie z głębokością całkowitą i łupkami; skały węglanowe są wyłącznie skorelowane z piaskowcami, a ewaporyty ze skałami

$$R = \begin{bmatrix} 1.0000 & 0.2428 & 0.8869 & 0.8437 & 0.1451 & 0.8179 \\ 0.2428 & 1.0000 & -0.1233 & -0.0345 & 0.4615 & -0.1076 \\ 0.8869 & -0.1233 & 1.0000 & 0.6903 & -0.0531 & 0.6965 \\ 0.8437 & -0.0345 & 0.6903 & 1.0000 & 0.0587 & 0.9874 \\ 0.1415 & 0.4615 & -0.0531 & 0.0587 & 1.0000 & -0.10001 \\ 0.8179 & -0.1076 & 0.6965 & 0.9874 & -0.1001 & 1.0000 \end{bmatrix} \begin{array}{l} \text{głębokość całkowita} \\ \text{piaskowce} \\ \text{łupki} \\ \text{nieklastyczne} \\ \text{węglanowe} \\ \text{ewaporyty} \end{array} \qquad [10]$$

nieklastycznymi, łupkami i głębokością całkowitą. W sumie dopatrzyć się tu można 2 układów: łupki — ewaporyty i skały węglanowe — piaskowce. Więcej o tym środowisku można powiedzieć wykonując dalsze obliczenia.

Obliczone z macierzy R jej wartości własne są następujące:

3.465, 1.540, 0.560, 0.435, 0.000, 0.000

Ilość różnych, co do wielkości wartości własnych jest 4, czyli ilość czynników F wynosi 4. Rozwiązując układ równań [6] przy pomocy wzorów [7], [8], [9] otrzymuje się jako wynik macierz ładunków czynnikowych a_{ij}, $(i = 1 \dots m,\ j = 1 \dots t)$.

Czynniki

	1	2	3	4	
Z_1 głębokość całkowita	0.9536	0.2223	0.1266	0.1586	
Z_2 piaskowce	−0.0002	0.8695	0.4881	−0.0754	
Z_3 łupki	0.8748	−0.1049	0.0077	0.4370	
Z_4 skały nieklastyczne	0.9494	−0.0067	−0.1041	−0.2962	
Z_5 skały węglanowe	0.0215	0.8388	−0.5423	0.0426	[11]
Z_6 ewaporyty	0.9429	−0.1397	−0.0178	−0.3020	

ładunki czynnikowe a_{ij}

czyli np.

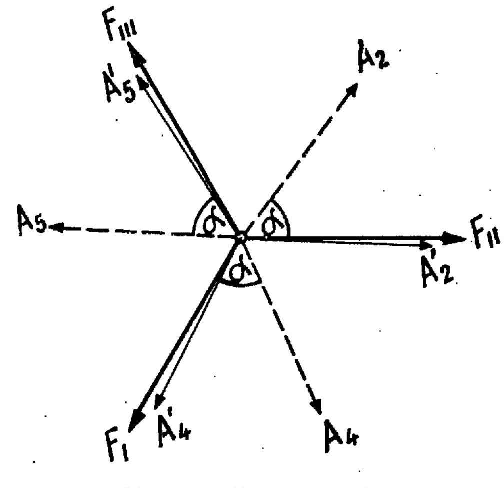
$$Z_1 = 0.9536\ F_1 + 0.2223\ F_2 + 0{\cdot}1266\ F_3 + 0.1586\ F_4$$

Interpretacja: czynnik nr 1 obejmuje „działanie" cechy nr 1, 3, 4, 6, czynnik nr 2 składa się z cech nr 2 i 5 podobnie czynnik nr 3, a czynnik nr 4 jak gdyby jest opisany cechą nr 3. Jak widać z powyższego zestawienia określając „budowę" czynnika kierowano się liczbowymi wielkościami a_{ij}. Co oznaczają te czynniki? Sytuacja nie jest prosta, gdyż o ile czynniki nr 1 i nr 2 jak gdyby się uzupełniają, to czynnik nr 3 jest w zasadzie powtórzeniem czynnika nr 2 z tym, że jest słabiej niż on związany z cechami, a także czynnik nr 4 „zawiera" się w czynniku nr 1. W tym miejscu należy zaznaczyć, że w wielu innych przykładach, szczególnie dotyczących badania populacji określonego rodzaju skały już na tym etapie analizy udaje się jednoznacznie identyfikować czynniki i przypisać im określoną treść geologiczną. O ile taka sytuacja istnieje zadanie właściwie jest zakończone. W naszym przykładzie tak nie jest, gdyż identyfikacja czynników nie została w sposób jednoznaczny dokonana. Można jedynie stwierdzić, że działały tu 4 czynniki, z których: pierwszy obejmował procesy geologiczne, w wyniku działania których intensywnie powstawały skały ewaporytowe i łupki, co pośrednio miało silny wpływ na miąższość skał młodszych od utworów permskich; czynnik drugi objął działanie innych procesów geologicznych powodujących osadzenie się piaskowców i utworów węglanowych, co z kolei miało nikły wpływ na całkowitą miąższość badanych skał, a sądząc po znaku (—) ładunków czynnikowych (patrz kolumna 2 wiersze 3, 4, 6) czynnik ten nie daje się w ogóle związać z cechami „budującymi" czynnik nr 1. Czynniki nr 3 i nr 4 jak gdyby zawierają się w czynnikach odpowiednio: nr 2 i nr 1, nie są więc łatwe do identyfikacji. Nie zagłębiając się w dalsze rozważania geologiczne przedstawiamy teraz jeden ze sposobów pozwalających na lepszą identyfikację czynników w sytuacjach podobnych do opisanej powyżej. Sposób ten nosi nazwę rotacji czynników i stanowi II etap procesu identyfikacji. Podobnie jak w rozdziale poprzednim opisany zostanie jedynie szkic postępowania. Pełny opis matematyczny można znaleźć u H. Harmana (3). Proces rotacji jest w całości wykonywany przy pomocy e.m.c.

Danymi wyjściowymi są elementy a_{ij} macierzy rozpatrywanej poprzednio. Kolumny tej macierzy (odzwierciadlające czynniki) tworzą układ t wektorów liniowo niezależnych (na mocy założenia 4d). Macierz ma n-wierszy (w naszym przykładzie 6). Z własności macierzy a wynika, że gdy założymy, iż jej rząd wynosi t, to wśród n-wierszy tej macierzy można wybrać t wektorów niezależnych. Dokonuje się tego, najogólniej mówiąc, za pomocą obrotu w przestrzeni t-wymiarowej układu wektorów wierszowych macierzy a.

Wybór liniowo niezależnych wektorów wierszowych można przyrównać do ustalania cech wiodących. Chodzi nam także o związanie konkretnych cech (wiersze macierzy) z czynnikami (kolumny macierzy). Proces rotacji najłatwiej objaśnić za pomocą rysunku (ryc. 1).

Załóżmy (ze względów graficznych), że wyodrębniono tylko trzy czynniki F_1, F_2, F_3. Tworzą one w przestrzeni trójwymiarowej 3 osie układu (gdyż jako



Obraz graficzny rotacji.
Graphic picture of rotation

nieskorelowane ze sobą są do siebie prostopadłe). W tejże przestrzeni nasze rozpatrywane 6 cech tworzy układ wektorów oznaczonych jako A_1, $A_2 \dots A_6$ (*na rysunku umieszczono dla przejrzystości tylko wektory* A_2, A_4, A_5). Gdy układ wektorów $A_1 \dots A_6$ obrócimy o kąt α, to część z wektorów znajdzie się w pobliżu osi układu $F_1 \dots F_3$. Ze względu na ich położenie (są to wektory prawie prostopadłe względem siebie) mogą one być uważane za liniowo niezależne. Ponadto, gdy wektor np. A'_4 leży w pobliżu osi F_1, to współrzędna wektora A'_4 względem osi jest największa, co oznacza, że czynnik F_1 jest determinowany głównie przez cechę tworzącą wektor A_4. W naszym przykładzie po wykonaniu rotacji okazuje się (częściowo przedstawiono to na rysunku), że czynnik F_1 określają cechy: A_4 — miąższość skał nieklastycznych i A_6 — miąższość ewaporytów; czynnik F_2 w zasadzie jest określony przez miąższość piaskowców (cecha A_2): za czynnik F_3 jest „odpowiedzialna" miąższość skał węglanowych, natomiast czynnik F_4 określają: miąższość łupków i głębokość do stropu skał wieku permskiego. Ze względu na to, że czynniki $F_1 \dots F_4$ są uszeregowane względem ich udziału w zmienności modelu można domniemywać, że najbardziej intensywnie działały tu procesy zawarte w czynniku 1, następnie zawarte w czynniku 2 itd. W formie macierzy wynik rotacji przedstawia się następująco:

	1	2	3	4
całkowita głębokość	0.6418	0.2487	−0.0865	0.7202
piaskowce	−0.0268	0.9703	−0.2403	−0.0027
łupki	0.4119	−0.1050	0.0329	0.9045
skały nieklastyczne	0.9397	−0.0249	−0.0656	0.3348
skały węglanowe	−0.0070	0.2349	−0.9720	0.0042
ewaporyty	0.9367	−0.0620	0.0888	0.3330

[12]

Imbrie w następujący sposób interpretował procesy geologiczne na podstawie powyższej macierzy:

czynnik 1 — ograniczona cyrkulacja, wysoki udział parowania i gwałtowne osiadanie,
czynnik 2 — nagły dopływ materiału gruboklastycznego,
czynnik 3 — normalne poziomy cyrkulacji i parowania, powolne osadzanie (depozycja) i brak składników klastycznych,
czynnik 4 — szybkie osiadanie, pośrednie wartości cyrkulacji i parowania, obfite nagromadzenie drobnego detrytusu.

Analizując przytoczoną powyżej interpretację na tle macierzy [12] widać, że Imbrie oceniał czynniki jako zespoły procesów geologicznych powodujących osadzanie się różnego rodzaju skał. Intensywność tych procesów wiązana jest z wielkością liczb w macierzy [12].

Wniosek następny, dotyczący opisanej interpretacji może sprowadzać się do pytania: czy potrzeba aż analizy czynnikowej, ażeby ustalić tego typu stwierdzenia geologiczne? Odpowiedź na to pytanie jest wg nas następująca: w przypadku „prostej" struktury układu (tak jak to było w zamieszczonym przykładzie) zastosowanie analizy czynnikowej pozwala „sprawdzić" tę technikę, ale nie daje wielu nowych stwierdzeń. W takich przypadkach jedyną korzyścią może być ocena intensywności lub długotrwałości działania procesów, trudna do ustalenia w inny sposób. Ocenę tę autorzy proponują wykonywać następująco: załóżmy, że czynnik I „wykorzystał" 60% zmienności ogólnej układu, czynnik II 20%, czynnik III 10% i czynnik IV 5%. Pozostałe 5% zmienności nie jest rozpatrywane. Obliczenie procentowych udziałów czynników w zmienności układu dokonane jest w e.m.c. (patrz λ we wzorze 9).

Na podstawie procentowych udziałów czynników w zmienności układu można więc używając liczb oraz ich kombinacji typu: sumy, różnice, ilorazy ustalać quasi ilościową charakterystykę modelu. I tak na czynnik I w macierzy [12] w głównej mierze, bo w ok. 2/3 składa się wpływ procesu parowania (ok. 40% zmienności) i w ok. 1/3 wpływ osiadania (ok. 20% zmienności). W czynniku II proces gwałtownego dopływu materiału gruboklastycznego stanowi główny (ok. 2/3) składnik, co daje szacunek ok. 13% zmienności. Podobnie analizując czynniki III i IV można szacować udział depozycji na ok. 7%, a osiadanie na 4% zmienności. Można więc, choćby w przybliżeniu, zaproponować ilościową gradację wpływu różnych procesów na aktualną postać środowiska, a stąd możliwe już są do wysunięcia przesłanki do np. oceny prawdopodobieństwa obecności np. określonych rodzajów skał lub też uściślenia chronologicznej kolejności działania procesów.

W sytuacji, gdy środowisko ma bardzo złożoną strukturę, może to być jak wiadomo spowodowane albo działalnością wielu procesów geologicznych, albo też złożonością działania tych procesów.

Przykładem z tej dziedziny mogą być np. gliny zwałowe. W takich przypadkach wnikliwa analiza macierzy [12] pozwala na przeprowadzenie interpretacji opartej na następujących pytaniach:

a. Czy są do pomyślenia procesy geologiczne (w danym przypadku) oparte na wybranych cechach z czynnika I? (z czynnika II, III, IV itd.).
b. Czy wybrane cechy mają duże w sensie liczbowym wartości w macierzy typu [12] i czy jest to zgodne z naturą procesu i może świadczyć np. o jego intensywności?
c. Czy o ile na proces składa się 2 lub więcej cech, to w przypadku gdy np. 1 cecha ma dużą wartość, a inne mają wartości małe świadczy to o odmianie procesu, i czy ta odmiana może mieć miejsce dla badanego środowiska?
d. Czy są do pomyślenia procesy „cząstkowe" składające się na proces geologiczny w sensie aktualnie używanym?

Odpowiedzi na powyższe pytania pozwalają niejednokrotnie na uwypuklenie działania tych procesów, o których specjalista wie, że „powinny" działać, ale brak jest jednoznacznych dowodów na stwierdzenie tego faktu. Z doświadczeń stosowania analizy czynnikowej, szczególnie do badania populacji skał wynika także wniosek, że prawidłowa interpretacja wymaga niejednokrotnie stosowania dla opisu działania procesu geologicznego pojęć z zakresu fizyki, gdyż wymagane jest niejednokrotnie rozczłonkowanie działania procesu geologicznego na szereg zjawisk fizycznych opisujących tylko pewne „części" jego działania. W tym miejscu jeszcze raz pragniemy podkreślić, że właściwy dobór cech rzutuje w decydujący sposób na możliwość przeprowadzenia interpretacji.

SYNTEZA

Kolejne etapy pracy specjalisty dają się sprowadzić do następujących punktów:

1. Należy stworzyć model koncepcyjny badanego środowiska. Tworząc model należy wybrać i pomierzyć te cechy, które ten model opisują w pełni, mając na uwadze cel, do jakiego się dąży.

2. Należy sprawdzić czy badane środowisko spełnia założenia podane w punktach 1—5.

3. Obliczyć macierz R.

4. Obliczyć wartości własne macierzy R, co doprowadza do ustalenia ilości czynników.

5. Obliczyć macierz A.

6. Wykonać pierwszy etap analizy, biorąc pod uwagę wielkości ładunków czynnikowych. O ile doprowadza to do identyfikacji czynników przeprowadzić merytoryczną ocenę pod kątem wyznaczenia sensu geologicznego czynników w badanym środowisku. Na tym etapie możliwe są trzy odpowiedzi: TAK, NIE, BYĆ MOŻE. Gdy odpowiedź jest TAK zadanie się kończy, gdy NIE oznacza to, że technika analizy czynnikowej nie nadaje się do analizy badanego środowiska, co może mieć swoje źródło albo w niewłaściwym doborze cech, albo w niemożności przyjęcia za słuszne założenia, że cechy są liniowo skorelowane z czynnikami. Gdy odpowiedź jest BYĆ MOŻE należy wykonać rotację, co pozwala na wyjaśnienie sytuacji. Naturalnie po wykonaniu rotacji niezbędna jest także interpretacja merytoryczna.

TECHNIKA OBLICZEŃ

Program na analizę czynnikową opracowano pierwotnie na zlecenie Instytutu Hydrogeologii i Geologii Inż. U W w Przedsiębiorstwie Poszukiwań Geofizycznych (K. Uścińska, P. Stenzel i inni, 1970), a następnie ulepszono w trakcie prac problemowych w Instytucie Hydrogeologii i Geologii Inżynierskiej UW.

Program napisany jest w języku GIER ALGOL IV na e.m.c. GIER i znajduje się zarówno w PPG, jak i w IHiGI Uniwersytetu. Użytkownik przygotowuje następujące dane: ilość obiektów, macierz (1).

Program wyprowadza do wydruku: średnie wielkości dla poszczególnych cech, wariancje dla poszczególnych cech, macierz korelacji R, wartości własne macierzy korelacji, ładunki czynnikowe dla poszczególnych czynników w kolejnych cechach wraz z procentem wykorzystania zmienności całkowitej.

Program „rotacja" jest także napisany w języku GIER—ALGOL IV i wymaga przygotowania następujących danych: dowolny tekst (bez przecinka) zakończony przecinkiem, ilość cech, ilość wyodrębnionych czynników, macierz ładunków czynnikowych (wartości jak w macierzy [11]. Program wyprowadza do wydruku następujące wyniki: tekst objaśniający, macierz ładunków po rotacji (patrz macierz [12]). Czas pracy e.m.c. w większości przypadków jest mniejszy od 0,5 godziny.

LITERATURA

1. Bielecka K. — Metody określania elementów wiodących w strukturze. Prz. geogr., 1970, z. 3.
2. Czyż T. — Zastosowanie metody analizy czynnikowej do badań ekonomicznej struktury regionalnej Polski, Prz. geogr., 1971, nr 92.
3. Harman H. — Modern factor analysis. University of Chicago Press. Chicago, 1960.
4. Kowalski W. C. — Regionalna geologia inżynierska Polski. Praca złożona do druku w Wyd. Uniwers., Warszawa, 1970.
5. Krumbein W. C., Graybill F. A. — An introduction to statistical models in geology. Mc Graw-Hill Book Co. London—N. York, 1965.
6. Określenie poziomu rozwoju ekonomicznego powiatów. Opracowanie próbne metodą analizy czynnikowej. GUS, Warszawa, 1968.
7. Piasecki Z. — Analiza czynnikowa metodą Hotellinga. PWN, Warszawa, 1969.
8. Stenzel P., Uścińska K. — Zastosowanie metod matematycznych i ETO w geofizyce i geologii. Prace postępu technicznego. PPG, 1970.
9. Wiatr I. — Model statystyczny wybranych cech środowiska inżynersko-geologicznego kopalnych dolin Przykony i Małgorzaty w okolicy Turka. Maszynopis złożony w „Biuletynie Geologicznym UW", nr 15, 1971.
10. Wiatr I., Stenzel P., Kowalski A. — Metodyka zastosowania analizy czynnikowej do badania układu strukturalnego cech fizycznych populacji geologicznych. Praca wykonana w IH i GI UW w ramach problemu węzłowego 06.1.1.

SUMMARY

The article deals with the application of factor analysis-R in geological sciences.

Factor analysis comprises several mathematical-statistical techniques allowing a reduction of the initial set of variables that characterize the objects under observation to a smaller number of hypothetical variables, called factors.

The factors are latent, that is directly unobservable, variables. The factorial model belongs therefore to the class of latent structures. The identification of the factors is of decisive importance to an adequate description of reality of geological environment.

РЕЗЮМЕ

Статья посвящена применению факторного анализа — способ Р в геологических дисциплинах. Этот анализ охватывает несколько математико-статистических приемов, которые позволяют сократить исходное число переменных, характеризующих объекты и являющихся предметом наблюдений, до меньшего количества гипотетических переменных, названных факторами. Факторы являются скрытыми, не поддающимися непосредственному наблюдению переменными. Следовательно факторная модель относится к классу скрытых структур. Определение факторов имеет решающее значение в достоверной оценке действительных условий геологической среды.