

OCENA REPREZENTATYWNOŚCI LOSOWYCH PRÓBEK EMPIRYCZNYCH

UKD 519.2:551.782.1.022.4:550.83(438—12 przedgórze Karpat)

Jednym z istotnych zagadnień, którego rozwiązanie warunkuje efektywność określanych na etapie treningu (nauki) statystycznych reguł klasyfikacji badanych obiektów, jest formowanie jednorodnych i dostatecznie reprezentatywnych próbek losowych klasyfikowanych obiektów o wiarygodności znanej *a priori* ich przynależności poszczególnych klas.

Przy formowaniu empirycznych próbek losowych należy pamiętać, że analizowane próbki powinny być w zasadzie statystycznie jednorodne, tj. obejmować obiekty charakteryzujące się takimi wartościami opisujących je parametrów, które podlegają tym samym lub nieistotnie różniącym się dla przyjętego poziomu prawdopodobieństwa prawom rozkładu. Pewne

sprawdzenie jednorodności próbki jest możliwe, gdy znany jest charakter rozkładu mierzonych parametrów dla obiektów jednego typu lub zadany jest związek empirycznej próbki z podziałem na określone typy (16). Wówczas w pierwszym przypadku sprawdzenie sprowadza się do porównania obserwowanego empirycznego rozkładu ze znanym teoretycznym, w drugim przypadku zaś — do porównania kilku rozkładów empirycznych. Często wystarczającym sprawdzeniem jednorodności statystycznej próbki może być weryfikacja hipotezy o jednowierchołkowości (jednomodalności) empirycznych krzywych rozkładu. Wynika to stąd, że geneza krzywych rozkładu wielowierchołkowych związana jest z połączeniem dwu lub więcej samodzielných, odrębnych statystycznie ziorów w jeden niejednorodny (6, 16).

W statystyce matematycznej przyjmuje się (2, 7), że dla zapewnienia reprezentatywności próbki, co jest warunkiem koniecznym formułowania znaczących wniosków odnośnie do badanej zbiorowości, wystarcza w praktyce spełnienie dwóch warunków, a mianowicie losowości i dostatecznej liczności próbki. Pierwszy warunek zostaje spełniony, gdy znamy losową naturę danych geologiczno-geofizycznych oraz zestawiamy próbkę empiryczną w sposób, zapewniający każdemu obiektowi spośród wszystkich będących do dyspozycji jednakową szansę dostania się do próbki. Zresztą ze względu na specyfikę analizowanego eksperymentalnego materiału geologiczno-geofizycznego, który w zasadzie uzyskiwany jest w wyniku pasywnych (biernych) eksperymentów, możliwości większego wpływu na losowość próbki empirycznej są w znacznym stopniu ograniczone. W praktyce przyjmuje się, że aby uzyskiwane reguły klasyfikacji statystycznej były dostatecznie efektywne do badań powinno się brać obiekty o jak najbardziej zróżnicowanych charakterystykach i z otworów leżących w różnych częściach analizowanego obszaru.

W odniesieniu do warunku dostatecznej liczebności próbki empirycznej w praktyce przyjmuje się pogląd, że im próbka jest liczniejsza, tym bardziej jest reprezentatywna. Ten punkt widzenia jest w określonym sensie słuszny, gdyż wnioski uzyskiwane na podstawie badania próbki empirycznej generalnie są tym dokładniejsze i pewniejsze, im liczniejsza jest próbka. Jednak poczynając od pewnej granicznej ilości obserwacji N_0 , dokładność i pewność uzyskiwanych z próbki wniosków okazują się praktycznie niezależne od jej liczności N . Innymi słowy, wprowadzenie dodatkowych (ponad N_0) obserwacji w zasadzie nowych danych o badanej zbiorowości nie wnosi. W konsekwencji nakłady na te dodatkowe obserwacje (o ile ich zadaniem jest dostarczenie informacji) są niecelowe. Można wykazać i co wyczuwa się intuicyjnie, że im bardziej zróżnicowana, zmienna jest badana zbiorowość, tym liczniejsza powinna być reprezentatywna próbka z niej pochodząca. Dotychczasowe praktyczne doświadczenia różnych autorów (1, 3, 8, 11, 13, 19), a także analiza oparta na teorii informacji (5, 18) wykazują, że dostatecznie reprezentatywne empiryczne próbki treningowe zależnie od stopnia niejednorodności i zróżnicowania własności skał powinny obejmować po 25—50 obiektów każdej klasy.

Dla określenia koniecznych (racjonalnych) liczebności N_0 próbek empirycznych można wykorzystać entropię, wielkość znaną z teorii informacji. Entropia $H(x)$, jako miara stopnia nieokreśloności systemu z dyskretnym zbiorem stanów, obliczona jest według wzoru [9].

$$H(x) = - \sum_{i=1}^n P_i \cdot \log_2 P_i \quad [1]$$

gdzie P_i — prawdopodobieństwo i -tego stanu systemu;

n — liczba możliwych stanów systemu.

W zależności od podstawy logarytmu, a mamy różne skale pomiaru entropii i różne jednostki, w dal-

szych rozważaniach będziemy stosowali podstawę $a=2$, której odpowiada jednostka entropii — bit.

Na systemy z ciągłym zbiorem stanów pojęcia entropii bezpośrednio się nie uogólnia. Jednak, oceniwszy dokładność pomiaru ciągłej wielkości losowej x , charakteryzującej system możemy twierdzić, że w interwale, którego rozmiar Δx odpowiada tej dokładności, poszczególne stany systemu praktycznie są nierozróżnialne. W efekcie układ z ciągłym zbiorem stanów można rozpatrywać jako system z dyskretnym zbiorem stanów, przy czym ilość możliwych dyskretnych stanów jest równa ilości interwałów, w których poszczególne ciągle stany układu w związku z istnieniem błędów pomiarów praktycznie są nierozróżnialne. Dla oceny $H(x)$ takiego systemu można wykonywać wzór analogiczny do [1]

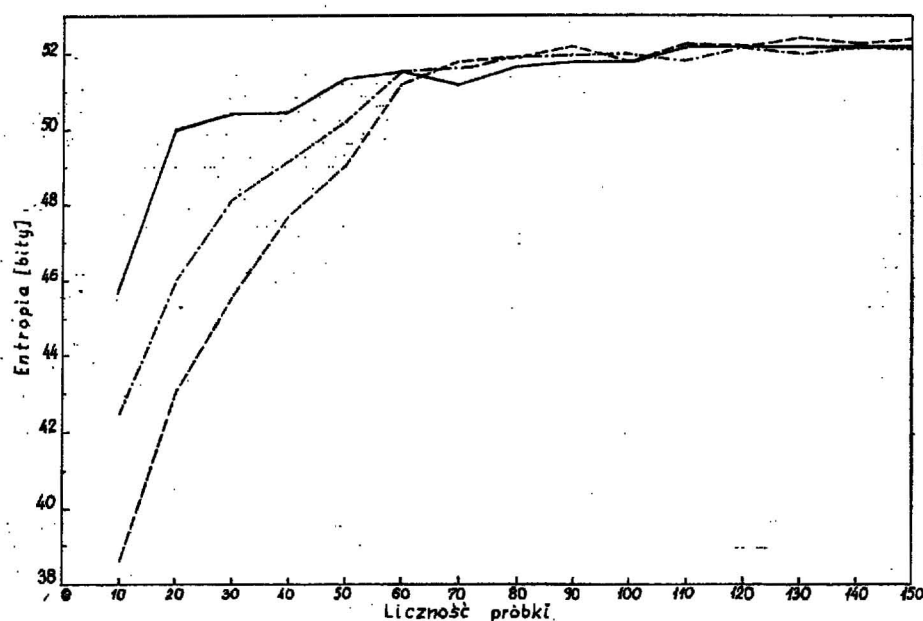
$$H(x) = - \sum_{i=1}^n [f(x) \cdot \Delta x] \cdot \log_2 [f(x) \cdot \Delta x], \quad [2]$$

gdzie $[f(x) \cdot \Delta x]$ — prawdopodobieństwo trafienia wartości losowej wielkości x , charakteryzującej system do interwału Δx . Interwały Δx można utożsamiać z przedziałami klasowymi, wykorzystywanymi przy konstrukcjach histogramów czy empirycznych krzywych rozkładu (poligonów, wieloboków częstotliwości). Wówczas wartości $[f(x) \cdot \Delta x]$ są wprost równe wartościom częstości, charakteryzującym poszczególne przedziały klasowe Δx . Entropia systemu uwarunkowana liczbą jego możliwych stanów, jest również prawdopodobieństwem tych stanów, jest kumulacyjną charakterystyką szeregu rozdziałczego empirycznego rozkładu losowej wielkości x , opisującej system.

Ze wzoru [1] widać, że jeśli prawdopodobieństwo jednego z możliwych stanów systemu jest równe 1, a wszystkich pozostałych 0, to entropia systemu $H(x) = 0$. Przypadek ten odpowiada systemowi o tej własności, że wielkość charakteryzująca system jest stała. Jeśli system może znajdować się w n różnych możliwych stanach z równym prawdopodobieństwem (tj. $P_i = 1/n$), to entropia systemu przyjmuje wartość maksymalną równą $\log_2 n$. Przypadek ten dotyczy sytuacji, gdy wielkość charakteryzująca system cechuje się rozkładem równomiernym (jednostajnym). Wobec powyższego entropia systemu, opisanego zmienną losową x , może przyjmować wartości z przedziału $[0, \log_2 n]$.

W swoich badaniach autorzy wykorzystali entropię do oceny granicznej (racjonalnej) liczebności N_0 próbek empirycznych, obejmujących opróbowane interwały w obrębie miocenu Przedgórze Karpat, a potrzebnych do rozwiązania zagadnienia typowania w profilach otworów skał zbiornikowych oraz oceny charakteru ich nasycenia. Oparto się przy tym na następujących zasadach (5). Zbiór interwałów można rozpatrywać jako pewien fizyczny system, charakteryzujący się pewną liczbą możliwych stanów i ich prawdopodobieństwem, a w efekcie — pewną w pełni określoną wartością entropii. System ten można przyjąć, jeśli określona jest jego entropia, tj. ustalona jest liczba i prawdopodobieństwo możliwych stanów systemu.

Pod entropią zbioru interwałów, jako systemu, rozumie się entropię generalnej zbiorowości wartości określonego parametru lub grupy parametrów, opisujących poszczególne interwały. Entropia próbki losowej będzie różnić się od entropii zbiorowości generalnej w ten sam sposób, jak entropia podsystemu od entropii systemu. Z teorii informacji wiadomo, że entropia podsystemu ze wzrostem liczby prób dąży do entropii systemu jako do granicy. W konsekwencji, przy powiększaniu w sposób losowy liczebności próbki empirycznej, jej entropia dąży do entropii zbiorowości generalnej jako do granicy, po osiągnięciu której praktycznie powinna się ustalić. Entropia kolejno rozpatrywanych podsystemów zmienia się, mając tendencję rosnącą, aż do momentu, kiedy doj-



Zmiany sumarycznej entropii losowych próbek opróbowanych interwałów mioceńskich w zależności od ich liczności dla strefy głębokościowej I ($H \leq 500$ m) i trzech wariantów obliczeń.

Changes in summative entropy of random samples representing different intervals of the Miocene profile depending on their number for a given interval I ($H \leq 500$ m) and three variants of calculations.

dziemy do podsystemu, którego entropia równa jest praktycznie entropii badanego systemu. W wyniku dalszego wzrostu liczby prób systemu (obserwacji) uzyskuje się podsystemy, ekwiwalentne sobie i systemowi w całości, tj. wielkość entropii pozostaje stała, równa entropii systemu. Liczba obserwacji, przy której entropia podsystemu staje się praktycznie równa entropii systemu, jest tą graniczną licznoscią próbki empirycznej N_{gr} , której dalsze przekraczanie jest niecelowe.

Ponieważ każdy z interwałów opisany jest szeregiem parametrów $\{x_1, x_2, \dots, x_m\}$, wobec tego zbiór interwałów powinien być traktowany jako system, którego pełna entropia wyraża się wzorem (5):

$$H(x_1, x_2, \dots, x_m) = -\sum \sum \dots \sum P \cdot \log_2 P, \quad [4]$$

gdzie P — prawdopodobieństwo kombinacji prostych stanów systemu. Jeśli zmienne $x_1, x_2, \dots, x_m$ są niezależne (nieskorelowane) między sobą, pełna entropia złożonego systemu równa jest wprost sumie entropii prostych systemów (sumie entropii każdego z parametrów):

$$H(x_1, x_2, \dots, x_m) = \sum_{i=1}^m H(x_i). \quad [5]$$

Biorąc pod uwagę fakt, że interesuje nas nie absolutna wartość ilości informacji (entropii) analizowanych zbiorów interwałów, a tylko jej przyrost ze wzrostem liczności próbek, możemy korzystać ze wzoru [5]. Oczywiście, jest to równoznaczne z zaniedbaniem istnienia zależności między poszczególnymi parametrami i prowadzi do istotnego zawyżenia pełnej entropii systemu. Nie ma to jednak znaczenia przy badaniu momentu stabilizacji entropii systemu w zależności od ilości prób (obserwacji).

Zgodnie z wcześniejszymi uwagami ocena entropii dowolnego systemu opisanego ciągłymi parametrami x , wymaga określenia przedziałów klasowych Δx tych parametrów. Problem wyboru Δx jak dotychczas jest słabo zbadany [5], a istniejące w literaturze zalecenia sprowadzają się do tego, że wielkość Δx należy wybierać wychodząc z dokładności pomiaru danego parametru, jego wartości średniej i stopnia wypełnienia klas. Przy analizie entropii losowych

próbek interwałów mioceńskich wartości Δx ustalono w oparciu o oceny standardowych bezwzględnych błędów $\sigma(\delta x)$ poszczególnych parametrów x [14]. Wartości Δx przyjmowano w przybliżeniu równe lub nieco mniejsze od podwojonych ocen błędów standardowych $\sigma(\delta x)$, co przy normalnym rozkładzie błędów odpowiada przybliżonemu 95-procentowemu jednostronnemu przedziałowi dokładności określania poszczególnych parametrów x .

Przed przystąpieniem do obliczeń entropii dla poszczególnych próbek empirycznych konieczne było również ustalenie typu rozkładów błędów pomiaru. W wyniku przeprowadzonych badań [15] stwierdzono, że zdecydowaną większość z analizowanych parametrów geofizyki wiertniczej, za pomocą których charakteryzowano opróbowane interwały cechuje się rozkładami błędów zbliżonymi do rozkładu logarytmiczno-normalnego.

Po ustaleniu typu rozkładów błędów pomiarowych oraz wartości przedziałów klasowych analizowanych parametrów przeprowadzono obliczenia zmian entropii poszczególnych parametrów geofizycznych, jak również sumarycznej entropii z uwzględnieniem pełnego zespołu tych parametrów, w oparciu o próbki losowe interwałów mioceńskich ze wzrastającą ich licznoscią N . Obliczenia zrealizowano na maszynie cyfrowej „Cyber 72” według programu ENTR ułożonego w języku FORTRAN na podstawie powyżej opisanego algorytmu.

Biorąc pod uwagę powyższe rozwiązanie oraz uwzględniając ilość będącego do dyspozycji materiału przemysłowego (ilość opróbowanych interwałów) dotyczącego osadów miocenu Przedgórze Karpat (wraz z jego rozkładem głębokościowym) całość mioceńskich osadów rozbito na pięć następujących stref głębokościowych:

- I. $H \leq 500$ m;
- II. $500 \text{ m} < H \leq 1000$ m;
- III. $1000 \text{ m} < H \leq 1500$ m;
- IV. $1500 \text{ m} < H \leq 2000$ m;
- V. $H > 2000$ m;

Otrzymano tym sposobem pięć odrębnych zbiorowości generalnych, przy czym ilość istniejącego materiału przemysłowego zapewniała zestawienie próbek empirycznych dla najmniej licznych klas skał

w poszczególnych strefach głębokościowych o licznosciach co najmniej 30–40 opróbowanych interwałów. Obliczenia realizowano odrębnie dla każdej z pięciu wydzielonych stref głębokościowych miocenu. W przypadku parametrów, których pozorne błędy pomiarowe cechują się rozkładami zbliżonymi do normalnego (15) operowano pierwotnymi wartościami x w pozostałych przypadkach posługiwano się wartościami przekształconymi $u(x)$. Umożliwiało to zachowanie stałej wartości pozornych błędów pomiarowych poszczególnych parametrów w pełnym zakresie ich zmienności, a w konsekwencji — również stałych wartości ich przedziałów klasowych. Obliczenia dla każdej strefy głębokościowej prowadzono w trzech wariantach — przy trzech różnych ciągach próbek empirycznych, zestawionych w sposób losowy niezależnie od siebie. Liczność próbek N zmieniała się w granicach od 10 do 150 ze stałym krokiem $N=10$.

Przykładowe wyniki badania zmian sumarycznej entropii losowych próbek opróbowanych interwałów miocennskich zależnie od ich licznosci dla strefy głębokościowej 0–500 m przedstawiono graficznie na rycinie.

Na podstawie uzyskanych wyników można stwierdzić, że sumaryczna entropia losowych próbek interwałów miocennskich w poszczególnych strefach głębokościowych generalnie ulega stabilizacji przy licznosciach próbek N rzędu 60–80. Dalszy wzrost ilościowy próbek nie powoduje znaczniejszych zmian ich entropii, wobec czego $N=60-80$ można uznać za licznosc w zupełności wystarczającą (graniczną). Tylko w przypadku II strefy głębokościowej (500 m $< H \leq 1000$ m) obserwuje się pewien dalszy wzrost entropii do $N=100$, co może być objaśnione stosunkowo dużą liczbą opróbowanych w tej strefie interwałów o dużych miąższościach (typu tzw. jaksmanickiego — 17), cechujący się znacznym zróżnicowaniem parametrów. Biorąc pod uwagę fakt, że analizowane próbki obejmowały łącznie trzy klasy skał (zbite oraz zbiornikowe gazo- i wodonośne) można sądzić, iż racjonalne losowe próbki reprezentujące poszczególne klasy skał powinny liczyć ok. 30–40 opróbowanych interwałów. Zgodnie jest to zarówno z praktycznymi doświadczeniami wielu autorów (3, 4, 8, 10, 11), jak również z obliczeniami teoretycznymi opartymi na teorii informacji (18).

LITERATURA

1. Choliń A.I., Elanski M.M. — Przemienienie matematycznych metod dla reszenia geologo-geofizycznych zadacz. Tr. MINCH i GP, wyp. 91. Moskwa, 1959.
2. Czermiński J., Iwasiewicz A. i in. — Metody statystyczne w doświadczalnictwie chemicznym. PWN, Warszawa, 1974.
3. Dachnow W.N., Churgin J.I. i in. — Matematyczne roboty po wniedrieniu wycislitelnoj techniki dla reszenia zadacz promyslowoj geofiziki i petrofiziki. Fondy MINCH i GP. Moskwa, 1960.

SUMMARY

The paper deals with the problem of estimating representativeness of random empirical samples. The estimations were made using entropy, i.e. a quantity known from the theory of information. They gave a limit value of the population of empirical samples, above which the accuracy and reliability of conclusions drawn from the analysis appear to be practically not effected by the number of samples. The representativeness of random empirical samples was evaluated on the basis of a population of objects comprising Miocene horizons from the Carpathian Foreland, sampled and described with a set of geophysical parameters. On the basis of the results obtained it may be stated that the summative entropy of random samples of Miocene horizons becomes generally stable when population of samples approaches 60–80.

4. Dementiew L.F., Koskow W.N. i in. — O wydelenii plastow — kolektorow po promyslowo-geofizycznym dannym s ispolzowanijem ECVN. Nefit i gaz Tiumeni, wyp. 8, 1970.
5. Dementiew L.F. — Reszenije zadacz nefitiegazopromyslowoj geologii s primienieniem wierojatnostno-statisticheskich metodow. Dokt. dissent., Tiumen, 1971.
6. Goldin S.W. — O prowierke odnorodnosti sowokupnostej geologiczieskich obiektow. Tr. ZapSibNIGNI, wyp. 18, Tiumen, 1968.
7. Hellwig Z. — Elementy rachunku prawdopodobienstwa i statystyki matematycznej. PWN, Warszawa, 1972.
8. Kompleksnaja interpretacija geologiczieskich i geofizicheskich dannych na wycislitelnych maszynach. Tr. MINCH i GP, wyp. 62. Moskwa, 1966.
9. Kulinkowicz A.J., Basin J.N. — Primienienie teorii informacii k woprosam radioaktivnogo karotaza. Tr. WNIJGG, wyp. 3. Moskwa, 1968.
10. Kuzniecowa G.S. — Analiz effektivnosti metodik wydelenija glinistych produktivnych kolektorow. Sb. Geofiz. issl. nefit. i gaz. skwazn. „Niedra”, Moskwa, 1971.
11. Matematicheskie metody w gazoniefitianoj geologii i geofizike. „Niedra”, Moskwa, 1972.
12. Rodionow D.A. — Statisticheskije metody razgranichenija geologiczieskich obiektow po kompleksu priznakow. „Niedra”, Moskwa, 1968.
13. Rozowski L.B. — Wwiedienie w teoriu geologiczieskiego podobija i modelirowanija. „Niedra”, Moskwa, 1969.
14. Rychlicki S., Twardowski K. — Ocena dokladnosc pomiarow geofizyki wiertniczej w obrebie osadow miocennskich Przedgorza Karpat. Techn. Poszuk. geol. (w druku).
15. Rychlicki S., Twardowski K. — Zagadnienie typu rozkladow pozornych błędow pomiarow geofizyki wiertniczej. Zesz. nauk. AGH (w druku).
16. Szimanski A.A., Uszakin J.M. i in. — Ispolzowanije funkcii raspriedielenija pri reszenii niekotorych poiskowo-razwiedocznych zadacz. Sb. Matemat. metody w polsk. razwied. praktiki. Irkutski polit. inst., Irkutsk, 1970.
17. Twardowski K. — Statystyczny model interpretacji materialow geofizyki wiertniczej dla celow zlozowych rejonu Przemyśl — Jaksmanice. Pr. dokt. IWN AGH, Krakow, 1973.
18. Walujski A.A. — Oprzedelenije optimalnogo objema obuczajuszczich posledowatielnosti pri ispolzowanii niekotorych programow raspodawaniija obrazow dla reszenija geologiczieskich zadacz. Tr. KIVNII, wyp. 21, Moskwa, 1971.
19. Zukow N.N. — Metody reszenija zadacz optimalnogo oprobowanija, statisticheskogo analiza i klassifikacii pri geologo-geofizicheskich issledowanijach. Awto-ref. Kand. dissent., MGRI, Moskwa, 1969.

РЕЗЮМЕ

В статье рассуждается оценка представительности случайных эмпирических проб. При помощи неэнтропии — величины из области теории информации — был определен предельный объем эмпирических проб. Выше предельного объема точность и надежность данных получаемых из проб являются практически независимые от её объема. Оценка представительности случайных эмпирических проб была проведена на основании множества объектов, которые составляли горизонты опробованные в пределах миоценовых отложений, лежащие на Предгорьях Карпат и описанные комплексом геофизических параметров. На основании полученных результатов авторы приходят к выводу, что суммарная неэнтропия случайных проб миоценовых интервалов в общем подвергается стабилизации при объеме проб порядка 60–80.