

*Tomasz Niedoba**

OCENA ZAWARTOŚCI POPIOŁU W POKŁADACH WĘGLA ZA POMOCĄ NIEPARAMETRYCZNYCH METOD STATYSTYCZNYCH**

1. Wprowadzenie

W procesie przeróbki węgla niezwykle ważnym problemem jest dokładne oszacowanie parametrów (np. zawartości popiołu, wilgotności, miąższości itp.) w badanej partii węgla. Aby tego dokonać, należy jak najdokładniej określić rozkład badanej cechy (czyli gęstość rozkładu, dystrybuantę). Węgiel wchodzący jako nadawa do procesu przeróbczego jest zazwyczaj mieszaniną węgli pochodzących z różnych pokładów danej kopalni. Każdy pokład charakteryzuje się odmiennymi cechami i generuje różne typy rozkładów badanych parametrów. Tak więc rozkład prawdopodobieństwa badanej zmiennej w całej populacji (partii węgla) należy traktować jako mieszaninę rozkładów badanej cechy w poszczególnych frakcjach [10, 11].

W artykule oszacowano zawartość popiołu w węglu na przykładzie jednej z kopalń. W celu określenia gęstości rozkładu zmiennej losowej, którą jest zawartość popiołu, zastosowano nieparametryczne metody aproksymacji, tj. metodę jądrową oraz metodę Fouriera.

2. Rozkład zawartości popiołu w węglu jako mieszanina rozkładów

Na przykładzie kopalni Y (istniejącej w rzeczywistości) wyróżniono 6 pokładów węgla, z których każdy charakteryzuje się innymi wartościami interesujących nas wielkości. Do przeróbki trafia węgiel będący mieszaniną węgli z wszystkich pokładów. Zawartość popiołu w węglu będącym nadawą będzie więc zróżnicowana i zależy od tego z jakiego pokładu pochodzi dana partia węgla. Rozważmy każdy z 6 pokładów, z których pochodzi wydobywany węgiel.

* Wydział Górnictwa i Geoinżynierii, Akademia Górniczo-Hutnicza, Kraków

** Artykuł powstał w ramach pracy statutowej nr 11.11.100.238

Określmy, że:

X_i — zawartość popiołu w węglu pochodzącym z i -tego pokładu, gdzie $i = 1, \dots, 6$.

X — zawartość popiołu w węglu wydobywanym w zakładzie Y.

Przyjmijmy, że rozkład zmiennej losowej X jest mieszaniną rozkładów zmiennych losowych X_1, \dots, X_6 , czyli ogólna postać dystrybuanty zmiennej X jest równa:

$$F(x) = \sum_{i=1}^6 p_i F_i(x),$$

gęstość rozkładu zmiennej X zadana jest wzorem:

$$f(x) = \sum_{i=1}^6 p_i f_i(x)$$

gdzie:

$f_i(x)$ — gęstość rozkładu zmiennej X_i ,

p_i — procentowy udział i -tej frakcji w populacji generalnej.

TABELA 1

Przykładowe zestawienia procentowych udziałów węgla z sześciu pokładów w nadawie trafiającej do przerobu

Numer i -tego pokładu	Pokład 1	Pokład 2	Pokład 3	Pokład 4	Pokład 5	Pokład 6
Średnia zawartość popiołu w węglu, %	13	12	15	14	12	15
Średnie odchylenie standardowe σ	2,2	1,5	3	4,3	3,5	1,9
M1	0	0	0,5	0	0	0,5
M2	0,5	0,5	0	0	0	0
M3	0	0,5	0,5	0	0	0
M4	0,1	0,5	0,1	0,1	0,1	0,1
M5	0,1	0,1	0,5	0,1	0,1	0,1
M6	0,2	0,2	0,2	0,2	0,1	0,1
M7	0	0,15	0,6	0	0,15	0
M8	0	0,3	0,7	0	0	0
M9	0,1	0,2	0,65	0	0,05	0

W tabeli 1 podano przykładowe zestawienia ($M_k, k = 1, \dots, 9$) udziałów procentowych węgla w nadawie trafiającej do przerobu oraz przykładowe rozkłady zmiennej losowej X

w nadawie przy założeniu log-normalności rozkładów zmiennych losowych X_i ($i = 1, \dots, 6$). Dane te pochodzą z istniejącej kopalni węgla B na Górnym Śląsku.

Jeżeli założymy, że rozkład zawartości popiołu w poszczególnych pokładach jest rozkładem logarytmiczno-normalnym o parametrach (m_i, σ_i) , czyli:

$$f_i(x) = \frac{1}{x\sigma_i\sqrt{2\pi}} e^{-\frac{(\ln x - m_i)^2}{2\sigma_i^2}} \quad \text{dla } x > 0 \quad (1)$$

to funkcja gęstości zmiennej losowej X (czyli zawartości popiołu w badanej partii węgla) zadana jest wzorem:

$$f(x) = \frac{1}{\sqrt{2\pi}} \sum_{i=1}^6 p_i \frac{1}{\sigma_i x} e^{-\frac{(\ln x - m_i)^2}{2\sigma_i^2}} \quad (2)$$

Dla zmiennej losowej o rozkładzie logarytmiczno-normalnym zachodzą wzory:

$$E(X) = e^{\frac{1}{2}\sigma^2 + m}$$

oraz

$$V(X) = e^{\sigma^2 + 2m} (e^{\sigma^2} - 1).$$

Stosując metodę momentów do wyznaczenia estymatorów nieznanych parametrów rozkładu, otrzymujemy:

$$m_i = \ln \bar{x}_i - \frac{1}{2} \ln \frac{s_i^2 + \bar{x}_i^2}{\bar{x}_i^2} \quad (3)$$

$$\sigma_i^2 = \ln \frac{s_i^2 + \bar{x}_i^2}{\bar{x}_i^2} \quad \text{dla } i = 1, 2, \dots, 6 \quad (4)$$

Uwzględniając wartości zawarte w tabeli 2, można zapisać, że funkcja gęstości zmiennej X (czyli zawartości popiołu w nadawie) ma postać

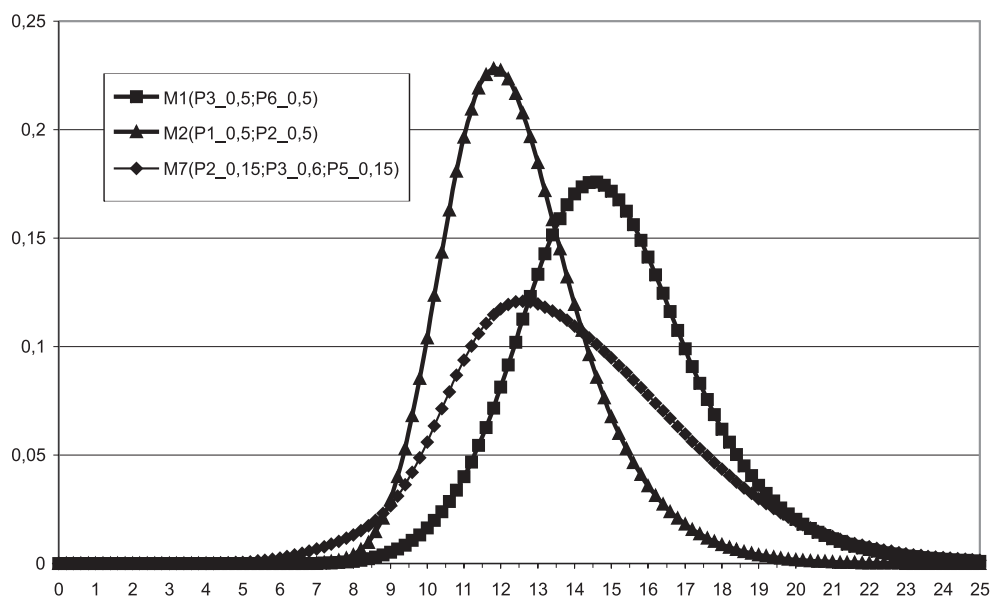
$$f(x) = \left(p_1 \frac{1}{0,167\sqrt{2\pi}} e^{-\frac{(\ln x - 2,551)^2}{0,056}} + p_2 \frac{1}{0,122\sqrt{2\pi}} e^{-\frac{(\ln x - 2,47)^2}{0,03}} + p_3 \frac{1}{0,197\sqrt{2\pi}} e^{-\frac{(\ln x - 2,688)^2}{0,078}} \right. \\ \left. + p_4 \frac{1}{0,3\sqrt{2\pi}} e^{-\frac{(\ln x - 2,594)^2}{0,18}} + p_5 \frac{1}{0,286\sqrt{2\pi}} e^{-\frac{(\ln x - 2,444)^2}{0,1632}} + p_6 \frac{1}{0,127\sqrt{2\pi}} e^{-\frac{(\ln x - 2,7)^2}{0,0326}} \right) \quad (5)$$

TABELA 2

Zestawienie wartości oczekiwanych oraz wariancji dla poszczególnych pokładów węgla w kopalni B

Nr pokładu	m_i	σ_i^2
Pokład I	2,551	0,028
Pokład II	2,47	0,015
Pokład III	2,688	0,039
Pokład IV	2,594	0,09
Pokład V	2,444	0,0816
Pokład VI	2,7	0,0163

Przebiegi tej funkcji dla różnych udziałów M_k ilustruje rysunek 1.



Rys. 1. Przykładowe funkcje gęstości mieszanin rozkładów zawartości popiołu w węglu

Widać więc, że wynikowe funkcje gęstości dla zawartości popiołu w węglu różnią się znacząco postacią w zależności, jaki jest udział procentowy poszczególnych pokładów węgla w nadawie. Co więcej, krzywa gęstości zawartości popiołu w nadawie nie musi być krzywą gęstości rozkładu log-normalnego, a postać wzoru opisującego tę krzywą nie jest postacią dotyczącą typu rozkładu zmiennej losowej opisanego w literaturze.

3. Estymacja jądrowa

Estymatorem jądrowym gęstości $f(x)$ rozkładu zmiennej losowej X nazywa się funkcję

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x-x_i}{h}\right) \quad (6)$$

gdzie funkcja $K(x)$, zwana jądrem, jest funkcją nieujemną, spełniającą warunek [2]

$$\int_{-\infty}^{\infty} K(x) dx = 1 \quad (7)$$

natomiast $h > 0$ jest parametrem zwanym szerokością pasma; (x_1, \dots, x_n) oznacza realizację próbki losowej.

Przedstawiono wyniki estymacji opartej na [5]:

— jądrze Epanechnikova

$$K(x) = \begin{cases} \frac{3}{4\sqrt{5}} \left(1 - \frac{1}{5}x^2\right) & \text{dla } |x| \leq \sqrt{5} \\ 0 & \text{dla } |x| > \sqrt{5} \end{cases} \quad (8)$$

— jądrze Gaussa

$$\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right), x \in R \quad (9)$$

Doboru wartości szerokości pasma h dokonuje się [3, 6] następująco:

— dla jądra Epanechnikova

$$h_{opt} = 1,056n^{-\frac{1}{5}}s \quad (10)$$

— dla jądra Gaussa

$$h_{opt} = 1,066n^{-\frac{1}{5}}s \quad (11)$$

gdzie:

n — liczebność próbki losowej,

s — estymator odchylenia standardowego $\left(s = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}\right)$ obliczony na podstawie tej próbki.

Stosując estymację jądrową przy użyciu jądra Epanechnikova, otrzymano następujące funkcje gęstości zawartości popiołu w węglu na poszczególnych pokładach ($f_i(x)$ — funkcja gęstości dla i -tego pokładu):

$$f_1(x) = 1,996 \sum_{i \in D_1(x)} n_i (1 - 1020,41(x - x_i)^2),$$

$$f_2(x) = f_3(x) = 6,708 \sum_{i \in D_1(x)} n_i (1 - 8000(x - x_i)^2),$$

$$f_4(x) = 1,156 \sum_{i \in D_1(x)} n_i (1 - 237,81(x - x_i)^2),$$

$$f_5(x) = 1,458 \sum_{i \in D_1(x)} n_i (1 - 378(x - x_i)^2),$$

$$f_6(x) = 1,795 \sum_{i \in D_1(x)} n_i (1 - 1338,89(x - x_i)^2),$$

gdzie $D_i(x) = \{i; x \in [x_i - h\sqrt{5}, x_i + h\sqrt{5}]\}$.

Używając w estymacji jądrowej jądra Gaussa, otrzymano następujące estymatory gęstości zawartości popiołu dla poszczególnych pokładów:

$$g_1(x) = 2,375 \sum_{i=1}^n e^{-\frac{(x-x_i)^2}{0,0008}},$$

$$g_2(x) = g_3(x) = 3,99 \sum_{i=1}^n e^{-\frac{(x-x_i)^2}{0,0002}},$$

$$g_4(x) = 1,735 \sum_{i=1}^n e^{-\frac{(x-x_i)^2}{0,0014}},$$

$$g_5(x) = 1,376 \sum_{i=1}^n e^{-\frac{(x-x_i)^2}{0,0017}},$$

$$g_6(x) = 3,325 \sum_{i=1}^n e^{-\frac{(x-x_i)^2}{0,0003}},$$

4. Estymacja gęstości metodą Fouriera

Jedną z bardziej znanych metod nowoczesnej teorii nieparametrycznej estymacji funkcji gęstości zmiennej losowej jest aproksymacja przy użyciu szeregu Fouriera [1, 4]. Niech ciąg funkcji $\{\varphi_j(x)\}$ jest ciągiem ortonormalnym na przedziale $[0, 1]$ gdzie

$$\varphi_j(x) = \begin{cases} 1 & \text{dla } j = 0 \\ \sqrt{2} \cos \pi j x & \text{dla } j = 1, 2, \dots \end{cases} \quad (12)$$

Jeżeli zmienna losowa przyjmuje wartości z przedziału $[a, b]$, to poprzez transformację liniową zamienia się ten przedział na $[0, 1]$. Zamianę zmiennych dokonuje się według poniższego wzoru

$$X' = \frac{1}{b-a} X - \frac{a}{b-a} \quad (13)$$

Estymatorem gęstości metodą Fouriera jest funkcja

$$f_J(x) = \sum_{j=0}^J Q_j \varphi_j(x) \quad (14)$$

gdzie J nazywa się parametrem obciążenia.

Estymatorami współczynników Fouriera Q_j są

$$\hat{Q}_j = \frac{1}{n} \sum_{i=1}^n \varphi_j(x_i) \quad (15)$$

gdzie x_1, \dots, x_n oznacza n -elementową realizację próbki losowej.

Za estymator parametru obciążenia J przyjmuje się

$$\hat{J} = \arg \min_{0 \leq J \leq J_n} \sum_{j=0}^J \left(\frac{2\hat{Q}_0}{n} - \hat{Q}_j^2 \right) \quad (16)$$

dla którego $\arg \min_{0 \leq J \leq J_n} (a_s)$ przyjmuje wartości równe indeksowi s najmniejszego elementu spośród (a_0, \dots, a_{J_n}) , natomiast J_n oznacza część całkowitą liczby $(4 + 0,5 \ln n)$.

Stosując metodę Fouriera, otrzymano następujące estymatory funkcji gęstości:

$$h_1(x) = 1 - \sqrt{2}(0,0142 \cos \pi x + 0,6486 \cos 2\pi x),$$

$$h_2(x) = 1 + \sqrt{2}(0,2785 \cos \pi x + 1,0650 \cos 2\pi x),$$

$$h_3(x) = 1 - \sqrt{2}(0,4574 \cos \pi x + 0,5036 \cos 2\pi x),$$

$$h_4(x) = 1 - 0,9120 \cos \pi x,$$

$$h_5(x) = 1 + \sqrt{2}(0,1640 \cos \pi x - 0,4893 \cos 2\pi x),$$

$$h_6(x) = 1 - \sqrt{2}(0,4760 \cos \pi x + 0,6931 \cos 2\pi x - 0,7000 \cos 3\pi x).$$

Do oceny miary dopasowania zastosowane zostanie średnie odchylenie resztowe, obliczone według wzoru

$$s_r = \sqrt{\frac{\sum_{i=1}^n (F(x_i) - \hat{F}(x_i))^2}{n-2}} \quad (17)$$

gdzie:

- \hat{F} — dystrybuanta rozkładu uzyskana przez estymację,
- F — dystrybuanta empiryczna,
- n — liczba pomiarów.

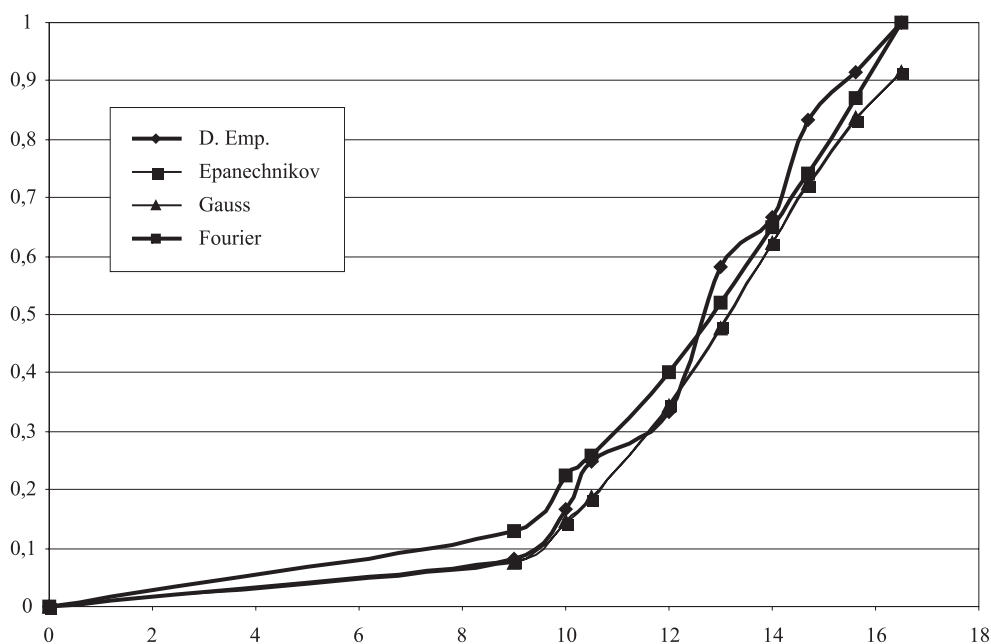
Wyniki dla stosowanych metod, jak i dla rozkładu logarytmiczno-normalnego przedstawia tabela 3.

TABELA 3

Zestawienie wyników odchyżeń resztowych dla zastosowanych metod aproksymacji

Pokład	Log-norm	Epanechnikov	Gauss	Fourier
I	0,0594	0,0809	0,0800	0,0487
II	0,0576	0,1164	0,1163	0,0422
III	0,0506	0,0472	0,0466	0,0636
IV	0,0840	0,0723	0,0702	0,0548
V	0,0511	0,0647	0,0637	0,0325
VI	0,0659	0,0608	0,0620	0,0490

Graficzną prezentację otrzymanych dystrybuant przedstawiono na rysunku 2.



Rys. 2. Zestawienie aproksymant dystrybuant zawartości popiołu w węglu uzyskanych różnymi metodami dla węgla z pokładu I

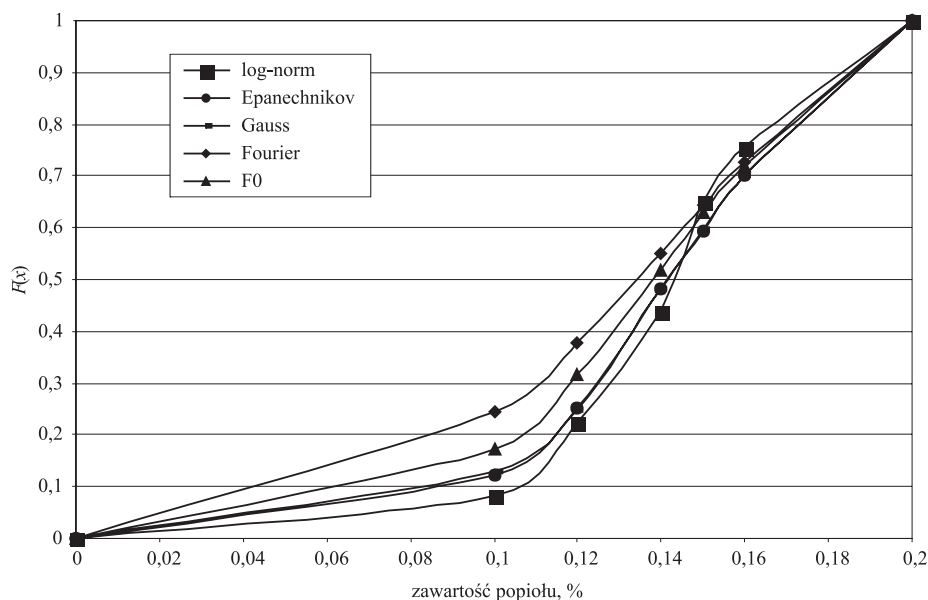
Dla większości pokładów najlepsze przybliżenie stanowi funkcja gęstości uzyskana metodą Fouriera, natomiast dla pokładu trzeciego najlepszy wynik daje metoda Gaussa. Analizując błędy, można stwierdzić, że dopasowane rozkłady logarymiczno-normalne w niektórych przypadkach dają lepsze wyniki niż rozkłady uzyskane metodami jądrowymi, co potwierdza celowość określania rozkładu zawartości popiołu w węglu jako rozkładu logarymiczno-normalnego.

Na podstawie otrzymanych wyników, najlepszą aproksymację funkcji gęstości zawartości popiołu w węglu w nadawie będzie stanowić mieszanina funkcji gęstości dla poszczególnych pokładów, dla których wartość s_r jest najmniejsza, czyli

$$\hat{f}_0(x) = \sum_{i=1}^n p_i f_{i0}(x) \tag{18}$$

gdzie $f_{i0}(x)$ oznacza estymator gęstości w i -tym pokładzie, dla którego wartość s_r jest najmniejsza [7–9].

Przykładowe porównanie dystrybuant mieszanin rozkładów otrzymanych różnymi metodami zaprezentowano na rysunku 3.



Rys. 3. Porównanie dystrybuant otrzymanych dla różnych typów aproksymacji na przykładzie mieszaniny rozkładów M5

TABELA 4

Zestawienie wartości prawdopodobieństwa, że zawartość popiołu w nadawie węgla nie przekracza 15% ($P(X < 15\%)$)

Typ mieszaniny	Log-norm	Epanechnikov	Gauss	Fourier	\hat{F}_0
MI	0,5641	0,4704	0,4721	0,5504	0,5338
MII	0,9176	0,8296	0,8326	0,9089	0,9089
MIII	0,7571	0,7055	0,7096	0,7742	0,7577
MIV	0,8258	0,7496	0,7522	0,8248	0,8315
MV	0,6519	0,5939	0,5968	0,6442	0,6277
MVI	0,7383	0,6601	0,6641	0,6887	0,6821
MVII	0,6734	0,6263	0,6304	0,6685	0,6488
MVIII	0,6702	0,6276	0,6319	0,6839	0,6608
MIX	0,6725	0,6249	0,6290	0,6762	0,6547

Tabela 4 przedstawia prawdopodobieństwa tego, że zawartość popiołu w węglu w nadawie nie przekracza 15% dla przykładowych mieszanin podanych w tabeli 1, przy założeniu, że dla wszystkich pokładów funkcje gęstości były wyznaczone tą samą metodą jak i dla funkcji gęstości, będącej mieszaniną optymalnych estymatorów na danym pokładzie.

Obliczone prawdopodobieństwa różnią się w zależności od rodzaju mieszaniny. Dla rozkładu logarytmiczno-normalnego przyjmują największe wartości, ale trzeba wziąć pod uwagę, że rozkład ten nie był optymalnie dopasowany do rzeczywistego rozkładu najbardziej prawdopodobne wartości przedstawia ostatnia kolumna.

Wartości tych prawdopodobieństw można wykorzystać przy ocenie ryzyka ekonomicznego przy podejmowaniu decyzji determinujących jakość oferowanego węgla. Błędne oszacowanie jego parametrów może prowadzić do wymiernych strat finansowych, jak również obniżyć wartość produktu.

5. Wnioski

Na podstawie przeprowadzonych badań można sformułować następujące wnioski:

- 1) rozkład zawartości popiołu w nadawie węgla należy rozpatrywać jako mieszaninę rozkładów tej cechy w poszczególnych pokładach;
- 2) zastosowanie nieparametrycznych metod estymacji pozwala precyzyjniej określić funkcję gęstości niż w przypadku założenia, że rozkład badanej cechy należy do rodziny rozkładów zadanych z góry (np. rozkład logarytmiczno-normalny);
- 3) w poszczególnych pokładach nie zawsze ta sama metoda daje najlepszy, możliwy efekt, dlatego celowe jest zastosowanie funkcji gęstości mieszaniny optymalnych estymatorów dla poszczególnych frakcji;
- 4) metody nieparametrycznej estymacji pozwalają dokładniej określić prawdopodobieństwo ryzyka ekonomicznego, związanego z założeniem określonej jakości węgla.

LITERATURA

- [1] *Chentsov N.N.*: Evaluation of an unknown distribution density from observations. *Soviet Math. Dokl.* 3, 1962, 1159–1562
- [2] *Efromovich S.*: Nonparametric curve estimation. New York, Springer-Verlag 1999
- [3] *Efron B.*: Nonparametric standard errors and confidence intervals. *Canad. J. Statist.*, 9, 1981
- [4] *Efron B., Tibshirani R.J.*: Using specially designed exponential families for density estimation. *Ann. Statist.* 24, 1996, 2431–2461
- [5] *Epanechnikov V.K.*: Non-parametric estimation of a multivariate probability density. *Theory Prob. Appl.*, 14, 1969
- [6] *Gajek L., Kaluszka M.*: Wnioskowanie statystyczne. Warszawa, WNT 2000
- [7] *Martyniak J.*: Niepewność wyników oznaczania właściwości produktu ziarnistego w badaniach wrywkowych. *Gospodarka Surowcami Mineralnymi*, 20, 2, 2004
- [8] *Niedoba T.*: The utilization of non classical statistical methods in the raw materials quality estimation — the estimation of economical risk. 8th Conference on Environment and Mineral Processing, Part II, VŠB-TU Ostrava, 2004, 29–37

- [9] *Niedoba T., Tumidajski T.*: The approximation of grain composition curves by non-parametric statistical methods. XXIII International Mineral Processing Congress, vol. 1, Stambuł, 2006, 203–209
- [10] *Tora B., Tumidajski T.*: Wpływ błędów opróbowania na dokładność bilansów składników materiałów uziarnionych. *Gospodarka Surowcami Mineralnymi*, 15, zeszyt specjalny, 1999
- [11] *Tumidajski T.*: Stochastyczna analiza własności materiałów uziarnionych i procesów ich rozdziału. Kraków, Wydawnictwo AGH 1997