*Zeszyty Naukowe*

*Instytutu Gospodarki Surowcami Mineralnymi i Energią*
*Polskiej Akademii Nauk*

*rok 2017, nr 100, s. 157–168*

Jacek MUCHA*, Monika WASILEWSKA-BŁASZCZYK**

# *Problemy opróbowania pośredniego złóż metodami korelacyjno-regresyjnymi na przykładzie złoża wapieni i margli*

Streszczenie: Metoda korelacyjno-regresyjna jako jedna z metod opróbowania pośredniego jest wykorzystywana w praktyce geologiczno-górniczej jedynie sporadycznie. Teoretycznie powinna ona być szczególnie przydatna do prognozowania zawartości niektórych składników chemicznych w złożach wapieni i margli z uwagi na łączące je silne zależności korelacyjne. W artykule przedstawiono wyniki analizy korelacji i regresji prostej oraz wielokrotnej (wielorakiej) dla 5 wytypowanych składników (CaO, $SiO_2$, $Al_2O_3$, MgO, $SO_3$) oznaczonych w próbach z otworów wiertniczych rozpoznawczych i otworów strzałowych wykonanych w złożu Barcin-Piechcin-Pakość. Jako miarę siły korelacji i jakości modeli regresyjnych wykorzystano współczynniki determinacji. Stwierdzono bardzo silną korelację liniową zawartości CaO i $SiO_2$ oraz silną korelację liniową zawartości CaO z $Al_2O_3$ i $SiO_2$ z $Al_2O_3$. Związki korelacyjne pozostałych par tlenków są słabe lub bardzo słabe i nie stwarzają podstaw do predykcji ich zawartości opartej na modelach regresyjnych wiążących je z zawartościami innych składników. Wykorzystanie modeli nieliniowych dla rozpatrywanych par składników przynosi jedynie niewielkie polepszenie jakości regresji, nieznaczące z praktycznego punktu widzenia. Do podobnych wniosków prowadzi także zastosowanie modeli regresji wielokrotnej, wiążącej zawartości kolejnych składników (z wyjątkiem CaO) ze wszystkimi pozostałymi. Uzyskany w dwóch przypadkach silny wzrost współczynników determinacji w porównaniu ze współczynnikami determinacji dla prostej korelacji liniowej okazał się być sztuczny i spowodowany występowaniem współliniowości pomiędzy zawartościami niektórych składników pełniących rolę zmiennych niezależnych. Z punktu widzenia praktyki geologiczno-górniczej uzyskane wyniki analizy wskazują na możliwość w pełni wiarygodnej predykcji jedynie zawartości $SiO_2$ oraz ograniczonej wiarygodności predykcji zawartości $Al_2O_3$, gdy znana jest zawartość CaO przy wykorzystaniu prostych, liniowych modeli regresji.

Słowa kluczowe: złoże wapieni i margli, składniki chemiczne, korelacje proste i wielorakie

* Dr hab. inż., prof. AGH, ** Dr inż., Akademia Górniczo-Hutnicza, Wydział Geologii, Geofizyki i Ochrony Środowiska, Kraków; e-mail: jacekm@agh.edu.pl; wasilews@agh.edu.pl

## *Problems related to indirect sampling methods using correlation and regression methods on the example of marl and limestone deposits*

Abstract: The correlation-regression method, as one of the indirect sampling methods, is only sporadically used in geological and mining activities. Theoretically, it should be particularly useful for predicting the content of some chemical components in limestone and marl deposits due to the correlation between them. The results of simple and multiple correlation and regression analysis for 5 selected components (CaO, $SiO_2$, $Al_2O_3$, MgO, and $SO_3$), determined in samples from exploratory boreholes and blast holes carried out in the Barcin-Piechcin-Pakość deposit, are presented in the article. The determination coefficients were used as a measure of the correlation power and the quality of the regression models. A very strong linear correlation between CaO and $SiO_2$ content and strong linear correlations between CaO and $Al_2O_3$ and $SiO_2$ with $Al_2O_3$ have been found. The correlation relationships of the remaining pairs of oxides are weak or very weak and do not provide a basis for prediction of their content based on regression models binding them with the content of other components. The use of nonlinear models for these pairs of oxides results in only a slight improvement in the quality of regression, insignificant from a practical point of view. The application of multiple regression models, linking the content of the mentioned components (with the exception of CaO), leads to similar conclusions. Compared to the determination coefficients of a simple linear correlation, a strong increase in determination coefficients obtained in two cases was found to be artificial and caused by a correlation between the content of the selected components acting as independent variables. From the geological and mining point of view, the results of the analysis indicate the possibility of a fully reliable prediction of $SiO_2$ content and the limited reliability of the $Al_2O_3$ content prediction when the CaO content is determined using simple linear regression models.

Keywords: limestone and marl deposits, chemical components, simple and multiple correlations

## *Wprowadzenie*

W odróżnieniu od klasycznego, bezpośredniego opróbowania złóż, opróbowanie pośrednie nie wymaga fizycznego pobierania porcji materiału skalnego. Oceny zawartości składników użytecznych w wyrobiskach górniczych lub rdzeniach z otworów wiertniczych dokonuje się przy zastosowaniu metod wizualnych, geofizycznych lub korelacyjno-regresyjnych. Opróbowanie pośrednie uważa się, nie zawsze słusznie, za znacznie mniej dokładne od opróbowania bezpośredniego. Zapewne ten pogląd był jedną z ważniejszych przyczyn bardzo ograniczonego zastosowania metod opróbowania pośredniego w praktyce geologiczno-górniczej, pomimo pozytywnych rezultatów uzyskanych w trakcie testowania niektórych z nich. Przykładowo, metoda rentgenofluorecencyjna (XRF) dała obiecujące wyniki przy oznaczaniu zawartości Zn i Pb w wyrobiskach górniczych olkuskich złóż Zn-Pb i zawartości Cu w złożach rud Cu-Ag LGOM, jak również zawartości Sn w rdzeniach z otworów wiertniczych wykonanych w złożu Gierczyn (Nieć 1990). Również dobre wyniki uzyskano przy wizualnej ocenie zawartości Zn w wyrobiskach kopalni rud Zn-Pb Bolesław (Blajda i Niedzielski 1979).

Metoda korelacyjno-regresyjna sprowadza się zasadniczo do prognozy zawartości jednego składnika (nie oznaczanego w próbie), traktowanego jako zmienna zależna (wyjaśniana), opierając się na znajomości zawartości innego składnika, traktowanego jako zmienna niezależna (objaśniana) lub zawartości zespołu innych składników przy wykorzystaniu wyznaczonego modelu łączących ich zależności korelacyjnych. Ocena zawartości może być szczególnie przydatna w odniesieniu do pierwiastków śladowych, których oznaczanie w sposób

bezpośredni jest kłopotliwe i kosztowne. Postać matematyczną modelu (funkcji regresji) określa się z reguły metodą najmniejszych kwadratów. Praktyczne zastosowanie zbudowanego modelu uwarunkowane jest jego statystyczną istotnością wykazaną odpowiednim testem statystycznym i wysokim stopniem dokładności predykcji zmiennej zależnej. Metoda ta wydaje się być predestynowana do prognozowania zawartości niektórych składników chemicznych w złożach wapieni i margli. Wzrostowi zailenia złóż wapieni z oczywistych powodów towarzyszy zmniejszanie się zawartości CaO, przy jednoczesnym zwiększaniu się zawartości $SiO_2$ i $Al_2O_3$. Metoda korelacyjno-regresyjna jak dotychczas nie znalazła szerszego zastosowania zarówno w złożach wapieni jak i innych złożach.

Celem przedstawianych badań była analiza przydatności różnych wariantów metod korelacyjno-regresyjnych do prognozy zawartości niektórych składników w złożu wapieni i margli Barcin-Piechcin-Pakość.

## 1. Materiał podstawowy

Materiał podstawowy badań stanowiły wyniki opróbowania fragmentu złoża wapieni i margli Barcin–Piechcin–Pakość, a w szczególności dobrze udokumentowane oznaczenia w próbach pięciu składników chemicznych: CaO, $SiO_2$, $Al_2O_3$, MgO, $SO_3$ (Mucha i in. 2017). W analizie statystycznej wykorzystano trzy rodzaje zbiorów danych, które tworzyły:

A. Oznaczenia składników w próbach pobranych z rdzeni wiertniczych otworów rozpoznawczych; z uwagi na bardzo zróżnicowane długości opróbowanych odcinków rdzeni dokonano regularyzacji (ujednolicenia) długości próbek do 1m, przypisując im oznaczenia zawartości wyliczone przy zastosowaniu algorytmu na średnią ważoną z oznaczeń dla próbek oryginalnych; łącznie wykorzystano 30 253 oznaczenia dla każdego składnika w próbach zregularyzowanych.
B. Oznaczenia składników w próbach pobranych z materiału skalnego uzyskanego w trakcie wykonywania otworów strzałowych z poziomów eksploatacyjnych o długości rzędu 20 m (35 914 oznaczeń).
C. Uśrednione arytmetycznie oznaczenia zawartości składników z otworów strzałowych wykonanych w obrębie umownych elementarnych jednostek wydobywczych obejmujących średnio 30 otworów strzałowych odpalanych jednocześnie dla uzyskania jednorazowej porcji urobku (299 danych).

Wykorzystane w analizie zbiory danych są liczebnościowo wyjątkowo bogate, zwłaszcza w przypadku danych z otworów rozpoznawczych i strzałowych.

## 2. Metodyka badań

W analizie przyjęto, że zawartości CaO jako podstawowego składnika złóż wapieni i margli są określane każdorazowo na podstawie klasycznego opróbowania bezpośredniego, natomiast przedmiotem predykcji będą zawartości $SiO_2$, $Al_2O_3$, MgO i $SO_3$ przy zastosowaniu modelu zależności (funkcji regresji) wyznaczonego metodą najmniejszych

kwadratów. Do tego celu wykorzystano statystyczne techniki korelacyjno-regresyjne, a w szczególności:

- prostą korelację liniową (**KL**) – w tym przypadku badano zależności między każdą parą składników,
- korelację nieliniową (**KN**) między każdą parą składników – wykorzystano do tego celu zlinearyzowane modele nieliniowe otrzymywane przez podstawowe transformacje matematyczne zmiennych niezależnych lub zależnych, prowadzące do przekształcenia pierwotnie nieliniowej zależności zmiennych w zależność liniową,
- korelację wieloraką (wielokrotną) (**KW**), wiążącą liniowo zawartości każdego ze składników (z wyjątkiem zawartości CaO) ze wszystkimi pozostałymi.

Wymienione metody zastosowano oddzielnie dla każdej z wymienionych wcześniej trzech grup utworzonych zbiorów danych (A, B, C).

Jako miarę siły korelacji obliczono wartości współczynnika determinacji ($R^2$), który określa udział zmienności zmiennej zależnej wyjaśnionej przez model zależności (funkcję regresji) w jej całkowitej zmienności. Współczynnik determinacji może przyjmować wartości z zakresu od 0 do 1, przy czym w geologii często wyraża się je w procentach od 0 do 100%. Siłę zależności łatwo jest w praktyce zinterpretować wykorzystując klasyfikację Niecia (Nieć i in. 2012). W myśl tej klasyfikacji zależność uznaje się za:

- bardzo silną (bardzo wyraźną), gdy $80\% < R^2 < 100\%$,
- silną (wyraźną), gdy $50\% < R^2 \leq 80\%$,
- słabą, gdy $25\% < R^2 \leq 50\%$,
- bardzo słabą, gdy $10\% < R^2 \leq 25\%$.

Gdy $R^2 \leq 10\%$ przyjmuje się, że korelacja nie występuje, natomiast gdy $R^2 = 100\%$ zależność jest pełna i ma charakter zależności funkcyjnej (deterministycznej).

W praktyce zastosowanie modelu regresji do predykcji zmiennych jest uzasadnione, kiedy jest on statystycznie istotny i dodatkowo cechuje się wysoką wartością współczynnika determinacji $R^2 > 80\%$ lub rzadziej, z zachowaniem pewnej rezerwy i ograniczonego zaufania do uzyskanych wyników, gdy $R^2 > 50\%$. W przypadku prostego modelu liniowego pierwiastek ze współczynnika determinacji jest równy popularnemu współczynnikowi korelacji liniowej zmiennych. W przypadku, gdy $R^2 \leq 50\%$, wyznaczone modele zależności uznawano za nieatrakcyjne z punktu widzenia ich możliwości predykcyjnych.

Wiarygodność uzyskiwanych metodą najmniejszych kwadratów liniowych modeli zależności uwarunkowana jest spełnieniem szeregu założeń (Stanisz 2007). Spośród nich do ważniejszych należą:

- Normalność (przynajmniej przybliżona) rozkładu składnika losowego, tzn. różnic (reszt) między pomierzonymi i odczytanymi z modelu wartościami zmiennej zależnej dla wszystkich obserwacji, którą można najprościej zweryfikować za pomocą normalnego wykresu kwantylowego reszt; spełnienie tego warunku umożliwia weryfikację istotności modelu za pomocą testów statystycznych w szczególności dla małych zbiorów danych; dla dużych zbiorów danych, z których korzystano w analizie, założenie to ma mniejsze znaczenie z uwagi na możliwość korzystania z rozkładów asymptotycznych, a ponadto analizy regresji są odporne na niewielkie odstępstwa od normalności rozkładów.

- Stałość wariancji składnika losowego modelu; właściwość tę zwaną homoscedastycznością można zweryfikować za pomocą wykresu zależności standaryzowanych reszt względem prognozowanych wartości zmiennej zależnej. Brak nawet przybliżonej stałości wariancji reszt, określany jako heteroscedastyczność, może być spowodowany źle dobraną postacią funkcji regresji, nieuwzględnieniem innych istotnych zmiennych niezależnych lub zróżnicowaną jakością zbioru danych podstawowych. Jej występowanie może powodować błędną ocenę wartości parametrów regresji (ich przeszacowanie lub niedoszacowanie, niewłaściwy znak parametru kłócący się z teorią lub zdrowym rozsądkiem) i brak ich stabilności przejawiający się znaczącymi zmianami wartości parametrów modelu przy powiększaniu zbioru danych związanego z dodawaniem nowych pomiarów. Heteroscedastyczność skutkuje ponadto błędami systematycznymi oceny wariancji składnika losowego (najczęściej jej zaniżeniem) i w konsekwencji przeszacowaniem wartości współczynnika determinacji oraz nieprawidłowymi przedziałami ufności dla prognozowanych wartości zmiennej zależnej.
- Brak współliniowości zmiennych niezależnych w przypadku regresji wielokrotnej (brak silnej zależności korelacyjnej zmiennych niezależnych). Występowanie współliniowości zmiennych niezależnych może powodować błędną ocenę wartości parametrów regresji (ich przeszacowanie lub niedoszacowanie, niewłaściwy znak parametru sprzeczny z teorią opisującą dane zjawisko) i brak ich stabilności przejawiający się znaczącymi zmianami wartości parametrów modelu przy powiększaniu zbioru danych związanym z dodawaniem nowych pomiarów.

Jedną z częściej stosowanych metod oceny siły współliniowości jest tzw. czynnik inflacji wariancji (VIF – *Variance Inflation Factor*) obliczany dla *i*-tej zmiennej niezależnej ze wzoru:

$$\mathrm{VIF}_i = \frac{1}{1 - R_i^2} \tag{1}$$

gdzie:

$R_i^2$ – współczynnik determinacji dla modelu regresji wielorakiej między *i*-tą zmienną niezależną i wszystkimi pozostałymi zmiennymi niezależnymi.

VIF oblicza się dla każdej ze zmiennych niezależnych (predyktorów) oddzielnie, co pozwala ustalić, która lub które z nich wprowadzają do modelu współliniowość.

Uważa się, że gdy VIF > 10, współliniowość ma charakter zakłócający, a zmienna odpowiedzialna za to (nadmiarowa) powinna być usunięta z modelu (Stanisz 2007).

W przypadku silnej współliniowości zmiennych niezależnych zamiast klasycznej analizy regresji liniowej można zastosować tzw. analizę regresji grzbietowej, w której sztucznie zmniejsza się wartość współczynników korelacji między zmiennymi niezależnymi aż do osiągnięcia zadowalających efektów oszacowania. Uzyskuje się w ten sposób estymatory parametrów modelu liniowego regresji wielorakiej o mniejszej wariancji ale obciążone w różnym stopniu błędami systematycznymi. Z uwagi na arbitralny sposób obniżania korelacji zmiennych niezależnych metoda ta nie znalazła powszechnej akceptacji (Stanisz 2007).

## 3. Wyniki badań

W pierwszej kolejności dla trzech rodzajów zbiorów danych podstawowych obliczono w programie STATGRAPHICS wartości współczynników determinacji ($R^2$), określające siłę korelacji liniowej wiążącej wszystkie pary zawartości rozpatrywanych składników chemicznych (tab. 1). We wszystkich analizowanych przypadkach korelacje są statystycznie istotne na poziomie istotności $\alpha = 0{,}05$. Bardzo silną korelację liniową ($87\% < R^2 < 92\%$) stwierdzono jednak jedynie między zawartością $SiO_2$ i CaO, zwiększającą się nieznacznie przy przejściu od prób zregularyzowanych z otworów rozpoznawczych przez otwory strzałowe do uśrednionych wartości składników w otworach strzałowych w obrębie elementarnych jednostek wydobywczych. Silną korelację ($50\% < R^2 < 66\%$) stwierdzono w przypadku pary zawartości $Al_2O_3$ i CaO oraz nieco zaskakująco i w sposób trudny do wytłumaczenia w przypadku pary zawartości $Al_2O_3$ i $SiO_2$, ale jedynie dla danych uśrednionych w elementarnych jednostkach wydobywczych. We wszystkich pozostałych przypadkach pomimo statystycznej istotności korelacje liniowe są zbyt słabe i w praktyce nieprzydatne do predykcji zawartości składników na podstawie modeli regresyjnych.

Zastosowanie najlepszych modeli nieliniowych (tzn. o najwyższym współczynniku determinacji) aproksymujących empiryczną zależność między parami składników, prowadzi jedynie do nieznaczącego praktycznie wzrostu siły korelacji mierzonej kilkuprocentowym

TABELA 1. Współczynniki determinacji dla prostej korelacji liniowej składników na podstawie prób: zregularyzowanych z otworów rozpoznawczych (A), z otworów strzałowych (B), z zespołu otworów strzałowych w obrębie elementarnej jednostki eksploatacyjnej (C)

TABLE 1. Determination coefficients for a simple linear correlation of components based on samples collected from: regularized exploratory boreholes (A), blast holes (B), and a set of blast holes carried out within the elementary exploitation unit (C)

| Składnik | Zbiór | $SiO_2$ | $Al_2O_3$ | MgO | $SO_3$ |
|---|---|---|---|---|---|
| CaO | A | 87,0 (1,66) | 50,6 (0,65) | 24,2 (0,68) | 3,6 (0,23) |
| | B | 90,0 (0,81) | 56,2 (0,19) | 32,6 (0,25) | 4,1 (0,16) |
| | C | 91,7 (0,42) | 65,4 (0,10) | 17,5 (0,08) | 11,9 (0,11) |
| $SiO_2$ | A | | 36,6 (0,74) | 4,5 (0,77) | 1,8 (0,23) |
| | B | | 45,6 (0,20) | 16,9 (0,27) | 0,7 (0,32) |
| | C | | 63,4 (0,10) | 8,0 (0,16) | 13,6 (0,16) |
| $Al_2O_3$ | A | | | 3,4 (0,77) | 2,5 (0,23) |
| | B | | | 16,0 (0,25) | 2,9 (0,16) |
| | C | | | 8,0 (0,09) | 13,6 (0,11) |
| MgO | A | | | | 1,7 (0,23) |
| | B | | | | 0,7 (0,32) |
| | C | | | | 0,2 (0,12) |

zwiększeniem współczynników determinacji (tab. 2–4). Ograniczając się do związków korelacyjnych zawartości składników z zawartościami CaO, nie obserwuje się formalnego przejścia od korelacji słabej do silnej, a tym bardziej bardzo silnej. Wynika z tego, że prostsze modele liniowe są wystarczające do opisu zależności zawartości składników.

Zastosowanie regresji wielorakiej (wielokrotnej) w przypadku danych z otworów strzałowych i uśrednionych danych dla grupy otworów strzałowych w elementarnej jednostce wydobywczej, podobnie jak i w przypadku korelacji nieliniowej nie przynosi (z wyjątkiem zawartości MgO dla otworów strzałowych) znaczącego podwyższenia mocy predykcyjnej modeli, o czym świadczą jedynie kilkuprocentowe wzrosty wartości współczynnika deter-

TABELA 2. Współczynniki determinacji dla prostej korelacji liniowej, korelacji nieliniowej i wielokrotnej składników na podstawie prób zregularyzowanych z otworów rozpoznawczych

TABLE 2. Determination coefficients for a simple linear correlation, nonlinear correlation, and multiple correlation of components based on the regularized exploratory borehole

| Składnik | Korelacja | $SiO_2$ | $Al_2O_3$ | MgO | $SO_3$ |
|---|---|---|---|---|---|
| CaO | liniowa | 87,0 | 50,6 | 24,2 | 3,6 |
| | nieliniowa | 88,0 | 51,0 | 32,8 | 4,3 |
| | wieloraka | 97,8 | 79,8 | 87,7 | 6,3 |

TABELA 3. Współczynniki determinacji dla prostej korelacji liniowej, korelacji nieliniowej i wielokrotnej składników na podstawie prób z otworów strzałowych

TABLE 3. Determination coefficients for a simple linear correlation, nonlinear correlation, and multiple correlation of components based on blast holes

| Składnik | Korelacja | $SiO_2$ | $Al_2O_3$ | MgO | $SO_3$ |
|---|---|---|---|---|---|
| CaO | liniowa | 90,0 | 53,1 | 32,6 | 4,1 |
| | nieliniowa | 90,4 | 56,2 | 37,9 | 7,1 |
| | wieloraka | 93,0 | 53,7 | 51,2 | 9,8 |

TABELA 4. Współczynniki determinacji dla prostej korelacji liniowej, korelacji nieliniowej i wielokrotnej składników na podstawie uśrednionych danych w grupach prób z otworów strzałowych w elementarnych jednostkach wydobywczych

TABLE 4. Determination coefficients for a simple linear correlation, nonlinear correlation, and multiple correlation of components based on the averaged data collected from sets of blast holes carried out within the elementary exploitation units

| Składnik | Korelacja | $SiO_2$ | $Al_2O_3$ | MgO | $SO_3$ |
|---|---|---|---|---|---|
| CaO | liniowa | 91,7 | 65,4 | 17,5 | 11,9 |
| | nieliniowa | 93,2 | 65,5 | 20,3 | 20,4 |
| | wieloraka | 92,8 | 67,1 | 27,2 | 17,4 |

minacji (tab. 3–4). Odmienne i silnie zaskakujące wyniki odnotowano natomiast dla prób zregularyzowanych. Korelacja wieloraka zawartości MgO i $Al_2O_3$ wyróżnia się potężnym wzrostem współczynnika determinacji (rzędu 40–60%) w porównaniu z prostą korelacją liniową i nieliniową tych składników z zawartością CaO (tab. 2). Znacznie mniejszy, ale znaczący wzrost współczynnika determinacji (o około 10%), stwierdzono również w przypadku korelacji wielorakiej zawartości $SiO_2$.

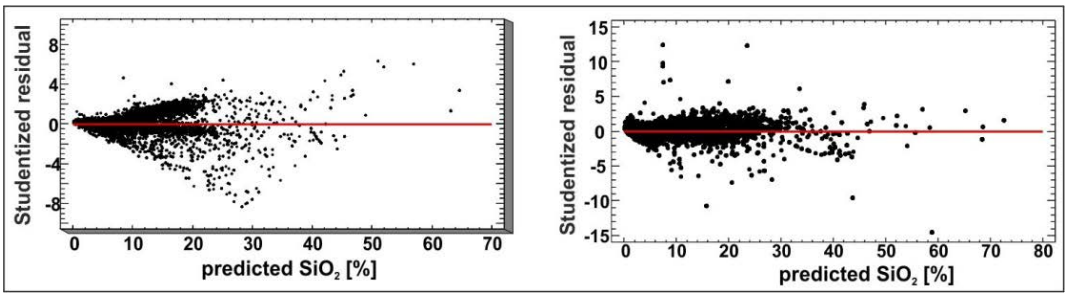
Dla wyjaśnienia tej nieoczekiwanie silnej korelacji obliczono współczynniki inflacji wariancji (VIF), które zestawiono w tabeli 5. W przypadku zmiennej zależnej $SiO_2$, nie stwierdza się zakłócającej model wieloraki współliniowości zmiennych niezależnych. Znajduje to potwierdzenie na wykresach zestandaryzowanych różnic (reszt) między zawartościami $SiO_2$ oznaczonymi w próbach i wyznaczonymi z modeli regresji (rys. 1). Dla prostej regresji liniowej wykres reszt ma formę rozszerzającego się lejka, co świadczy o heteroscedastyczności (niestałości wariancji składnika losowego) i niespełnieniu jednego z założeń poprawnego modelowania zależności. Dla modelu wielorakiego punkty charakteryzujące reszty zamknięte są z grubsza w granicach prostokąta, co pozwala przyjąć praktycznie założenie o stałości wariancji.

W przeciwieństwie do tego składnika modele wielorakie ze zmiennymi zależnymi $Al_2O_3$ i MgO wykazują silną współliniowość zmiennych niezależnych. W przypadku $Al_2O_3$ zmiennymi niezależnymi nadmiarowymi są zawartości CaO (VIF = 26,1) i $SiO_2$ (VIF = 20,5). Usunięcie tych zmiennych skutkuje jednak znaczącym obniżeniem wartości współczynnika determinacji do poziomu zbliżonego dla prostego modelu liniowego wiążącego ten składnik z CaO lub $SiO_2$. Do podobnych wniosków prowadzi analiza modelu wielorakiego z zawartością MgO jako zmienną zależną. Nadmiarową zmienną niezależną jest zawartość CaO, a jej usunięcie skutkuje drastycznym obniżeniem współczynnika determinacji do kilku procent (tab. 5). Wykresy reszt dla prostych modeli liniowych w obu przypadkach mają kształt rozszerzających się lejków, a dla modeli wielorakich ich zawężenie do prostokąta jest znacznie mniejsze niż w przypadku modelu wielorakiego z $SiO_2$ jako zmienną zależną (rys. 2–3).

TABELA 5. Badanie współliniowości zmiennych niezależnych w modelach wielorakich za pomocą czynnika inflacji wariancji (VIF) dla zbioru oznaczeń w próbach zregularyzowanych z otworów rozpoznawczych ($R^2$ – współczynnik determinacji modelu po wyeliminowaniu zmiennej niezależnej nadmiarowej z VIF > 10)

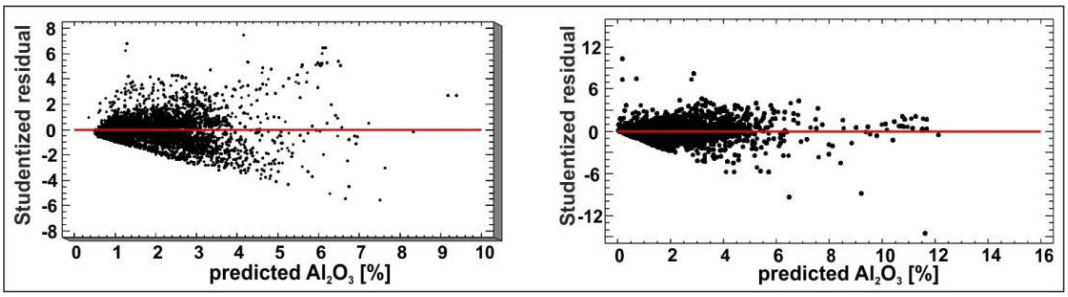
TABLE 5. Examination of the collinearity of independent variables in multiple models using the Variance Inflation Factor (VIF) for the set of values determined using regularized samples ($R^2$ – coefficient of determination of the model after elimination of independent superfluous variable with VIF > 10)

| $SiO_2$ | | $Al_2O_3$ | | MgO | |
|---|---|---|---|---|---|
| zmienne niezależne | czynnik inflacji wariancji (VIF) | zmienne niezależne | czynnik inflacji wariancji (VIF) | zmienne niezależne | czynnik inflacji wariancji (VIF) |
| CaO | 2,8 | CaO | 26,1 ($R^2$ = 38,0%) | CaO | 10,5 ($R^2$ = 5,8%) |
| $Al_2O_3$ | 2,2 | $SiO_2$ | 20,5 ($R^2$ = 54,8%) | $SiO_2$ | 8,1 |
| MgO | 1,4 | MgO | 3,5 | $Al_2O_3$ | 2,2 |
| $SO_3$ | 1,0 | $SO_3$ | 1,1 | $SO_3$ | 1,1 |



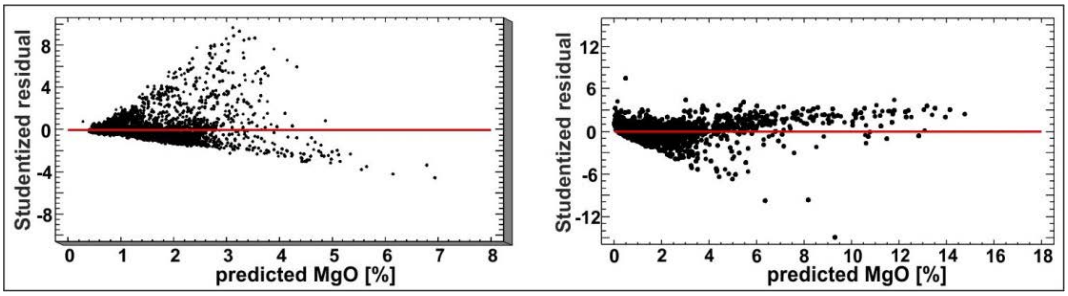

Rys. 1. Wykresy zależności reszt od prognozowanych zawartości $SiO_2$ przez prosty model liniowy $SiO_2 = f(CaO)$ (z lewej) i model wieloraki (z prawej)

Fig. 1. Graphs of dependence of regression residuals on the predicted $SiO_2$ content based on the simple linear model $SiO_2 = f(CaO)$ (left) and multiple model (right)




Rys. 2. Wykresy zależności reszt od prognozowanych zawartości $Al_2O_3$ przez prosty model liniowy $Al_2O_3 = f(CaO)$ (z lewej) i model wieloraki (z prawej)

Fig. 2. Graphs of dependence of regression residuals on the predicted $Al_2O_3$ content based on the simple linear model $Al_2O_3 = f(CaO)$ (left) and multiple model (right)




Rys. 3. Wykresy zależności reszt od prognozowanych zawartości MgO przez prosty model liniowy $MgO = f(CaO)$ (z lewej) i model wieloraki (z prawej)

Fig. 3. Graphs of dependence of regression residuals on the predicted MgO content based on the simple linear model $MgO = f(CaO)$ (left) and multiple model (right)

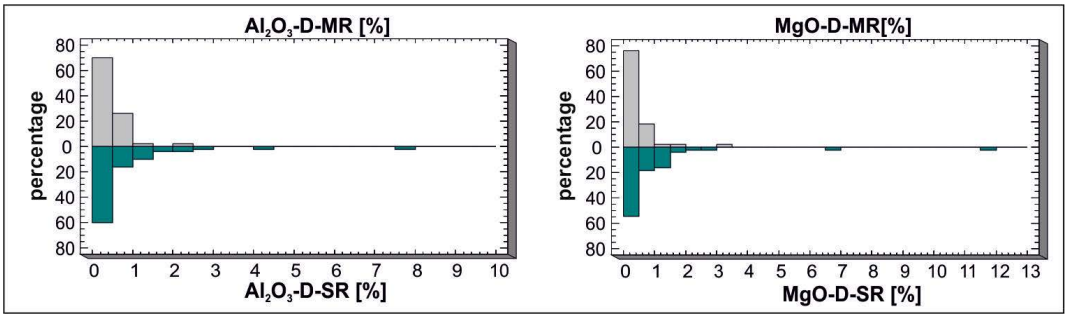
W pewnej sprzeczności z tymi wynikami pozostają rezultaty porównania zawartości $Al_2O_3$ i MgO stwierdzonych (oznaczonych w próbach) i obliczonych z modelu regresji wielorakiej w 50 dobranych losowo próbach zregularyzowanych, potraktowanych jako zbiór testowy. Średnie różnice absolutne są wyraźnie mniejsze dla modeli wielorakich niż dla modelu prostego (tab. 6). Ponadto w pierwszym przypadku są one zbliżone do średniego absolutnego błędu modelu natomiast w drugim przypadku wyraźnie go przewyższają.

TABELA 6. Porównanie średnich błędów absolutnych ocen zawartości $Al_2O_3$ i MgO w 50 losowo dobranych próbach zregularyzowanych wykonanych przy zastosowaniu prostego modelu liniowego, wiążącego je z zawartością CaO (jako zmienną niezależną) oraz modelu wielorakiego (w nawiasach przedstawiono średnie błędy absolutne modeli)

TABLE 6. The comparison of mean absolute errors for $Al_2O_3$ and MgO content predictions based on 50 randomly selected, regularized samples using a simple linear model linking them with the CaO content (as an independent variable) and a multiple model (the mean absolute errors of the models are shown in parentheses)

| Składnik | Średnie błędy absolutne ocen składników w zbiorze testowym [%] | |
|---|---|---|
| | model liniowy | model wieloraki |
| $Al_2O_3$ [%] | 0,80 (0,65) | 0,40 (0,41) |
| MgO [%] | 0,97 (0,68) | 0,39 (0,33) |

Graficznie uzyskane wyniki zilustrowano za pomocą histogramów średnich absolutnych błędów ocen (rys. 4). Na gorsze właściwości predykcyjne modeli liniowych w porównaniu z modelami regresji wielorakiej wpływa pojawianie się nielicznych anomalnie wysokich wartości błędów absolutnych i mniejszy udział małych błędów predykcji z pierwszego przedziału klasowego (0–0,5%).




Rys. 4. Histogramy błędów absolutnych ocen zawartości $Al_2O_3$ (z lewej) i MgO (z prawej) w 50 punktach zbioru testowego na podstawie modelu liniowego (SR) i modelu wielorakiego (MR)

Fig. 4. Absolute errors histograms for $Al_2O_3$ (left) and MgO (right) content prediction based on 50 sample collection points and both linear (SR) and multiple regression models (MR)

Oznacza to, że modele regresji wielorakiej pomimo stwierdzonej współliniowości niektórych zmiennych niezależnych mogą dawać trafne prognozy zawartości składnika pełniącego rolę zmiennej zależnej. Rezultat ten należy jednak traktować z pewną rezerwą i zweryfikować jego prawdziwość na nowym zbiorze danych testowych pochodzących z opróbowania innych rejonów eksploatacji złoża.

## Podsumowanie i wnioski

1. Dla rozpatrywanego złoża Barcin–Piechcin–Pakość, przy znajomości zawartości CaO, przedmiotem wiarygodnej predykcji opartej na modelach korelacyjno-regresyjnych mogą być jedynie zawartości $SiO_2$ i w ograniczonym stopniu zawartości $Al_2O_3$.
2. Nie stwierdzono znaczących różnic w dokładności predykcji zawartości $SiO_2$ i $Al_2O_3$ na podstawie prób zregularyzowanych, prób z otworów strzałowych i grupie prób z otworów strzałowych chociaż obserwuje się pewien wzrost siły korelacji w podanej kolejności zbiorów danych.
3. Zastosowanie zlinearyzowanych modeli liniowych oraz modeli regresji wielokrotnej nie prowadzi z praktycznego punktu widzenia do wartego zainteresowania podwyższenia jakości predykcji.
4. Stwierdzone zaskakujące wzrosty dokładności predykcji $Al_2O_3$ i MgO w przypadku zastosowania regresji wielorakiej dla danych z prób zregularyzowanych są iluzoryczne i są następstwem występowania nadmiernej współliniowości zmiennych niezależnych.
5. W modelowaniu zależności składników chemicznych dla potrzeb predykcji z wykorzystaniem regresji wielorakiej należy zwrócić szczególną uwagę na zjawisko współliniowości zmiennych niezależnych, która może prowadzić do uzyskania błędnych wyników sugerujących możliwość wysokiej dokładności prognozowania zawartości składników. Zagadnienie to w świetle danych literaturowych należy jednak do trudnych i nie posiadających jednoznacznego rozwiązania, szczególnie gdy dysponuje się ubogimi liczebnościowo zbiorami danych. Nie ma ponadto jasnych wytycznych, co do wielkości siły korelacji między zmiennymi niezależnymi, począwszy od której należy traktować je jako silnie współliniowe i zniekształcające wyniki modelowania wielorakiego, a w konsekwencji dające podstawę wykluczenia ich ze zbioru danych podstawowych. Zdarza się bowiem (jak w opisanym w tekście przykładzie na zbiorze testowym), że zmienne skażone współliniowością mogą dawać zadowalającą predykcję (trafne prognozy zawartości). Przykładowo, według programu STATGRAPHICS na kwestie współliniowości należy zwrócić uwagę już wówczas, gdy współczynnik korelacji przekracza 0,5 (tzn. współczynnik determinacji przekracza 0,25) lecz według innych opinii w praktyce współliniowość należy uwzględniać, gdy współczynnik korelacji liniowej między dowolną parą zmiennych niezależnych jest wyższy od współczynnika korelacji liniowej między zmienną zależną i tymi zmiennymi niezależnymi (Stanisz 2007).

Praca zrealizowana częściowo w ramach badań statutowych Katedry Geologii Złożowej i Górniczej AGH (nr 11.11.140.320) w 2017 roku.

## *Literatura*

Blajda, R. i Niedzielski, B. 1979. Porównanie wyników oceny wizualnej z wynikami chemicznego opróbowania jednego ze złóż cynkowo-ołowiowych. *Przegląd Geologiczny* Vol. 27, No. 12, s. 665–668.

Mucha i in. 2017 – Mucha, J., Wasilewska-Błaszczyk, M., Cieniawska, M. i Chudzik, W. 2017. Ocena wiarygodności prognozy jakości kopaliny na podstawie modelu 3D (na przykładzie fragmentu złoża wapieni i margli Barcin-Piechcin-Pakość). *Górnictwo Odkrywkowe. Surface Mining* nr 4, Wrocław, s. 10–17.

Nieć M., 1990. *Geologia kopalniana*. Warszawa: Wyd. Geol., 504 s.

Nieć i in. 2012 – Nieć, M., Mucha, J., Bromowicz, J. i Wasilewska-Błaszczyk, M. 2012. *Metodyka dokumentowania złóż kopalin stałych.* Tom 3. *Opróbowanie złóż.* Kraków, 128 s.

Stanisz, A. 2007. *Przystępny kurs statystyki z zastosowaniem STATISTICA PL na przykładach z medycyny.* Tom 2. *Modele liniowe i nieliniowe*. StatSoft Polska sp. z o.o., 865 s.

STATGRAPHICS® Centurion XVII User Manual 2014, Statpoint Technologies, Inc.