ZESZYTY NAUKOWE POLITECHNIKI ŚLĄSKIEJ 2003
Seria: BUDOWNICTWO z. 98 Nr kol. 1574

Małgorzata HYB
Szkoła Główna Gospodarstwa Wiejskiego

WYBRANE METODY STATYSTYCZNE STOSOWANE W DEFINIOWANIU CHARAKTERYSTYK KONSOLIDACYJNYCH

Streszczenie. W pracy przedstawiono metodykę analizy statystycznej stosowaną przy wyznaczaniu wybranych charakterystyk konsolidacyjnych jako funkcji pewnych parametrów. Przedstawiono metody wyznaczania współczynników funkcji regresji w oparciu o różne miary dopasowania predykcji do danych empirycznych wraz z algorytmem optymalizacji wyboru wartości parametrów modelu przy minimalizacji średniego błędu względnego. Opisano na przykładach wpływ miary dopasowania na wyniki analizy statystycznej, wykorzystując wyniki badań laboratoryjnych na próbkach torfu i gytii pobranych z poligonu doświadczalnego Antoniny Katedry Geoinżynierii SGGW.

SELECTED STATISTICAL METHODS APPLIED IN DETERMINATION OF CONSOLIDATION PARAMETERS

Summary. The analysis of statistical methods used in the determination of consolidation characteristics as a function of other parameters is a very important task in the determination of settlement. Selected methods for estimation of coefficients of a regression function and an algorithm to calculate these coefficients are presented in this paper. The results obtained from calculation examples have shown that the reliability of statistical analysis depends on the choice of criteria to assess the fitness of experimental prediction.

1. Wstęp

Wyznaczanie charakterystyk odkształceniowych gruntów organicznych (torfu i gytii) jako funkcji wybranych parametrów jest niezbędne w prognozie osiadań podłoża pod obciążeniem. Na całkowity proces osiadań podłoża organicznego składają się trzy etapy odkształceń, nakładających się na siebie i przebiegających równocześnie: początkowe, konsolidacyjne i wtórne. Jako parametr stosowany do obliczeń osiadań początkowych zaproponowano moduł odkształcenia objętościowego E_u, który jest funkcją dwóch

zmiennych: dewiatora naprężenia q i naprężenia konsolidacyjnego σ_k, to znaczy $\hat{E}_u = f(q, \sigma_k)$. Parametrem zastosowanym w pracy do obliczeń osiadań konsolidacyjnych jest moduł odkształcenia E - moduł Younga, który jest funkcją dwóch składowych naprężenia σ_1 i σ_3, to znaczy $\hat{E} = f(\sigma_1, \sigma_2)$. Jako parametr stosowany do obliczeń osiadań wtórnych zaproponowano odkształcenie ε_s, które jest funkcją dewiatora naprężenia q, czasu t i średniego naprężenia $\sigma_{śr}$, to znaczy $\hat{\varepsilon}_s = f(q, t, \sigma_{śr})$.

Przeprowadzone badania laboratoryjne odkształceń na próbkach torfu i gytii pobranych z podłoża nasypu doświadczalnego Katedry Geoinżynierii SGGW w Antoninach pozwoliły na wyznaczenie empirycznych związków konstytutywnych opisanych powyżej przy zastosowaniu metod analizy regresji.

2. Wybór zależności funkcyjnej między badanymi cechami

Analiza regresji polega na wyznaczaniu związków funkcyjnych między kilkoma badanymi cechami i zbadaniu istotności tych związków. Pozwala to na wyznaczenie empirycznych zależności, które umożliwiają obliczanie wartości jednej cechy (np. moduł odkształcenia E_u) jako funkcji kilku innych cech (np. dewiatora q i naprężenia σ_k). Metody analizy regresji są ściśle zobiektywizowane jedynie w przypadku modelu liniowego lub typu liniowego (np. model wielomianowy), gdzie można dokładnie wyznaczyć współczynniki i zweryfikować hipotezy dotyczące istotności współczynnika korelacji. W przypadku innych modeli, które często lepiej opisują rzeczywistość, można uzyskać jedynie wyniki przybliżone i dlatego bardzo ważna jest właściwa metoda wyznaczania współczynników takiego modelu.

W oparciu o wiedzę dotyczącą badanych cech i odpowiednie wykresy wykonane na podstawie danych empirycznych można dokonać wyboru postaci funkcji regresji

$$\hat{y} = f(x_1, x_2, ..., x_m), \tag{1}$$

gdzie y oznacza zmienną zależną (opisywaną), a $x_1, x_2, ..., x_m$ są zmiennymi opisującymi. Wzór tej funkcji zależy od pewnej liczby parametrów $a_1, a_2, ..., a_k$ zwanych współczynnikami regresji. Dlatego też wprowadzono dalej oznaczenie $f(x_1, x_2, .., x_m) = f(x_1, x_2, ..., x_m; a_1, a_2, ..., a_k)$. Liczba tych parametrów (czyli wartość k) nie powinna być za duża, chociaż przy większej liczbie parametrów dopasowanie funkcji jest lepsze.

Należy zaznaczyć, że w przypadku zależności od wielu zmiennych wybór odpowiedniej postaci funkcji jest trudny i w dużym stopniu subiektywny. Dlatego też proponuje się zbadanie kilkunastu modeli w celu wybrania funkcji regresji najlepiej dopasowanej do posiadanych danych empirycznych. Dla zapewnienia porównywalności zgodności poszczególnych modeli z danymi empirycznymi należy dokonać wyboru właściwej miary dopasowania.

3. Różne miary dopasowania modelu do danych empirycznych

Po ustaleniu postaci funkcji regresji $\hat{y} = f(x_1, x_2, ..., x_m; a_1, a_2, ..., a_k)$ należy wyznaczyć wartości parametrów $a_1, a_2, ..., a_k$, tak aby predykcje $\hat{y}_i = f(x_1, x_2, ..., x_m; a_1, a_2, ..., a_k)$ obliczone dla danych wartości x_{ki}, k=1,2,...m były w odpowiednim sensie bliskie wartościom empirycznym y_i dla i=1, 2, ..., n. Wśród stosowanych miar zgodności modelu z danymi należy wymienić przede wszystkim:

1. współczynnik zgodności φ^2 dany wzorem:

$$\varphi^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{2}$$

gdzie $\bar{y}$ jest wartością średnią cechy y,

2. współczynnik korelacji krzywoliniowej R, który dla $\varphi^2 \leq 1$ jest dany wzorem:

$$R = \sqrt{1-\varphi^2} = \sqrt{1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}}, \tag{3}$$

3. miary bezwzględne:
   a. suma kwadratów reszt (residual) lub suma odchyleń kwadratowych SSD (sum of square deviation) dana wzorem:

$$SSD = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(\Delta y_i)^2, \tag{4}$$

b. średnie odchylenie kwadratowe MSD (mean square deviation) dane wzorem:

$$MSD = \sqrt{\frac{1}{n} SSD} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \bar{y}_i\right)^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\Delta y_i\right)^2}, \tag{5}$$

4. miary względne:

a. średnie kwadratowe odchylenie względne MSRD (mean square relative deviation) dane wzorem:

$$MSRD = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{y_i}\right)^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\delta y_i\right)^2}, \tag{6}$$

b. maksymalne odchylenie względne lub maksymalny błąd względny MRD (maximal relative deviation) dane wzorem:

$$MRD = \max_{i=1,2,\ldots,n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| = \max_{i=1,2,\ldots,n}\delta y_i, \tag{7}$$

c. średnie odchylenie względne lub średni błąd względny MRE (mean relative error) dany wzorem:

$$MRE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| = \frac{1}{n}\sum_{i=1}^{n}\delta y_i \tag{8}$$

Podstawową miarą dopasowania badanego modelu do danych empirycznych stosowaną dotychczas w analizie regresji jest współczynnik korelacji krzywoliniowej R. Wartości R bliskie 1 świadczą o silnej zależności między zmienną objaśnianą y i zmiennymi objaśniającymi $(x_1, x_2, \ldots, x_m)$ (Elandt 1964, Kaczmarek 1970). Zatem współczynniki funkcji regresji można wyznaczać maksymalizując współczynnik korelacji, czyli minimalizując sumę odchyleń kwadratowych SSD. Postępowanie takie prowadzi do klasycznej metody najmniejszych kwadratów (MNK) szeroko opisywanej w literaturze.

Minimalizacja błędów bezwzględnych predykcji Δy_i może dawać znaczne błędy względne δy_i, szczególnie dla małych wartości empirycznych y_i. Zatem ze względów inżynierskich warto poszukiwać funkcji regresji w oparciu o minimalizację względnych miar dopasowania przedstawionych powyżej. Jest to zmodyfikowana metoda najmniejszych kwadratów (ZMNK) stosowana w niektórych współczesnych programach obliczeniowych.

4. Metody wyznaczania współczynników funkcji regresji

Stosując przedstawione powyżej miary dopasowania modelu do danych empirycznych, zaproponowano następujące metody wyznaczania współczynników $a_1, a_2, ..., a_k$ funkcji regresji $y = f(x_1, x_2, ..., x_m; a_1, a_2, ..., a_k)$.

Sposób 1. Minimalizacja ze względu na parametry $a_1, a_2, ..., a_k$ wartości SSD, gdzie:

$$SSD = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - f(x_{1i}, x_{2i}, ..., x_{mi}; a_1, a_2, ..., a_k))^2 . \quad (9)$$

Jest to klasyczna metoda najmniejszych kwadratów prowadząca do minimalizacji współczynnika zgodności φ^2, czyli maksymalizacji współczynnika korelacji R.

Sposób 2. Minimalizacja ze względu na parametry $a_1, a_2, ..., a_k$ wartości sumy kwadratów błędów względnych danej wzorem:

$$\sum_{i=1}^{n} \left(\frac{y_i - \hat{y}_i}{y_i} \right)^2 = \sum_{i=1}^{n} \frac{1}{y_i^2} (y_i - f(x_{1i}, x_{2i}, ..., x_{mi}; a_1, a_2, ..., a_k))^2 . \quad (10)$$

Jest to metoda najmniejszych kwadratów z wagami (zmodyfikowana metoda najmniejszych kwadratów), gdzie wartościami wag są dane empiryczne y_i^2. Prowadzi ona do minimalizacji średniego kwadratowego odchylenia względnego MSRD określonego wzorem (6).

Sposób 3. Minimalizacja ze względu na parametry $a_1, a_2, ..., a_k$ procentowego średniego błędu względnego $MRE \cdot 100\%$ (czyli średniej arytmetycznej procentowych błędów względnych) określonego wzorem:

$$MRE \cdot 100\% = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \cdot 100\% = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{|y_i|} |y_i - f(x_{1i}, x_{2i}, ..., x_{mi}; a_1, a_2, ..., a_k)| \cdot 100\% . \quad (11)$$

Minimalizacja tej funkcji prowadzi do zmodyfikowanej metody najmniejszych kwadratów.

Sposób 4. Minimalizacja ze względu na parametry $a_1, a_2, ..., a_k$ wartości maksymalnego procentowego błędu względnego $MRE \cdot 100\%$ danego wzorem:

$$MRD \cdot 100\% = \max_{i=1,2,...,n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \cdot 100\% = \max_{i=1,2,...,n} \left| \frac{y_i - f(x_{1i}, x_{2i}, ..., x_{mi}; a_1, a_2, ..., a_k)}{y_i} \right| . \quad (12)$$

Należy zauważyć, iż przy zastosowaniu sposobów 2, 3 i 4 można uzyskać funkcje regresji, dla których współczynnik zgodności $\varphi^2 > 1$, co uniemożliwia obliczenie współczynnika korelacji R. Przy wyborze miary dopasowania stosowano przede wszystkim sposób drugi lub trzeci, a także ewentualnie sposób czwarty, ponieważ niewielkie

pogorszenie średniego błędu względnego może dać znaczne zmniejszenie błędu maksymalnego.

5. Optymalizacja wyboru wartości parametrów modelu

Metoda wyznaczania współczynników funkcji regresji $\hat{y} = f(x_1, x_2, ..., x_m; a_1, a_2, ..., a_k)$ jest w każdym z wyżej przedstawionych sposobów taka sama. Prawą stronę wzorów (9), (10), (11) lub (12) traktuje się jak funkcję różniczkowalną współczynników $a_1, a_2, ..., a_k$. Obliczane są pochodne cząstkowe tej funkcji względem zmiennych $a_1, a_2, ..., a_k$ i po przyrównaniu ich do zera rozwiązywany jest odpowiedni układ k równań, na ogół nieliniowych, zwanych w literaturze układem równań normalnych. Rozwiązanie takiego układu jest możliwe w zasadzie tylko numerycznie. Obliczone w ten sposób wartości współczynników podstawiano do wzoru danej funkcji f i wyznaczano wartości $\hat{y}_i$, a następnie pozostałe wartości błędów.

Do wyznaczenia współczynników funkcji regresji można stosować pakiety statystyczne Statgraphics, SPSS, Statistica lub program Excel. Jednak stosowanie każdego z nich oddzielnie prowadzi do wyników, które nie są zadowalające. Dlatego też zaproponowano w pracy następujący algorytm optymalizacji wyboru wartości parametrów modelu, który przy założeniu iż ostatecznym celem jest minimalizacja funkcji (11) (funkcja celu), ma postać:

1. ustalenie postaci funkcji regresji $\hat{y} = f(x_1, x_2, ..., x_m; a_1, a_2, ..., a_k)$,
2. ustalenie początkowych wartości współczynników modelu ,
3. wyznaczenie za pomocą programu statystycznego SPSS 10.0 wartości parametrów modelu minimalizujących funkcję (9),
4. uzyskane w punkcie 3 wartości parametrów przenoszone są do programu Statgraphics Plus 4.1, w którym minimalizowana jest funkcja (10),
5. uzyskane w punkcie 4 wartości parametrów przenoszone są do programu Excel i za pomocą Solvera minimalizowana jest funkcja (11),
6. zapisywana jest najmniejsza wartość funkcji (11) (czyli wartość $MRE \cdot 100\%$) dla parametrów uzyskanych w punkcie 5. Następnie realizowany jest punkt 2,

7. z wartości otrzymanych w punkcie 6 wybierana jest najmniejsza, a odpowiadające jej wartości współczynników modelu (uzyskane w punkcie 5) uznawane są za parametry założonego w punkcie 1 modelu.

Za pomocą tego algorytmu można poddać analizie wiele modeli, czyli różnych postaci funkcji f.

6. Wpływ miary dopasowania na uzyskiwane wyniki analizy statystycznej

Poszukiwanie matematycznego modelu, który dobrze opisuje dane zjawisko, jest zagadnieniem bardzo trudnym i złożonym, szczególnie gdy chodzi o wybór postaci zależności funkcyjnej. Po dokonaniu wyboru takiej zależności należy wyznaczyć współczynniki modelu w oparciu o ustaloną miarę zgodności, której wybór powinien zostać dokonany przed rozpoczęciem obliczeń. Przedstawiona w pracy metoda wyznaczania współczynników funkcji regresji w oparciu o inne niż korelacja miary dopasowania ma ważny aspekt inżynierski, gdyż może prowadzić do ograniczenia błędów względnych, jednocześnie zachowując wysoki współczynnik korelacji.

Poniżej przedstawiono na przykładach wpływ wyboru różnych miar dopasowania na wyniki predykcji dla różnych rodzajów odkształceń i wybranych postaci funkcji regresji. Wyniki zamieszczono w tabelach, w których przyjęto następujące skróty:

Min. MRE - oznacza minimalizację średniego błędu względnego w %,

Min. MRD - oznacza minimalizację maksymalnego błędu w %,

Max R - oznacza maksymalizację współczynnika korelacji.

Tablica 1

Odkształcenia początkowe (torf). Wyniki porównawczej analizy statystycznej dla modelu $\hat{E}(q,\sigma_k) = a_1 \cdot q^{a_2} \cdot \sigma_k^{a_3}$

	Symbol	Min. MRE	Min. MRD	Max R
Suma odchyleń kwadratowych	SSD	36,44	36,08	23,68
Współczynnik korelacji w %	$R \cdot 100\%$	85,05	85,22	90,57
Średni błąd względny w %	$MRE \cdot 100\%$	20,57	24,39	25,42
Maksymalny błąd względny w %	$MRD \cdot 100\%$	243,91	52,87	134,3
Odchylenie standardowe bł. względnych	S	22,054	14,558	17,99
Mediana błędów względnych	Me	18,315	24,656	19,55
Liczba błędów powyżej średniej w %		39,8	50,8	41,11
Przedział ufności dla wartości MRE ·100%				
a) dolna granica		18,55	23,06	23,77
b) górna granica		22,60	25,73	27,07

Tablica 2

Odkształcenia wtórne (gytia). Wyniki porównawczej analizy statystycznej dla modelu

$$\dot{\varepsilon}_s(t,q,\sigma_{śr}) = a_1 \cdot arctg(t \cdot a_2 + a_3) + a_4 \cdot arctg(q \cdot a_5 + a_6) + a_7 \cdot arctg(\sigma_{śr} \cdot a_8 + a_9)$$

	Symbol	Min. MRE	Min. MRD	Max R
Suma odchyleń kwadratowych	SSD	1803,07	7949,23	785,92
Współczynnik zgodności	φ^2	0,294	1,295	0,128
Współczynnik korelacji w %	$R \cdot 100\%$	84,04	-	93,38
Średni błąd względny w %	$MRE \cdot 100\%$	28,99	50,46	67,24
Maksymalny błąd względny w %	$MRD \cdot 100\%$	179,25	93,48	1128,22
Odchylenie standardowe bł. względnych	S	29,227	25,440	178,771
Mediana błędów względnych	Me	24,200	54,844	14,487
Liczba błędów powyżej średniej w %		7,4	56,0	13,2
Przedział ufności dla wartości MRE ·100% a) dolna granica b) górna granica		 25,81 32,16	 47,69 53,23	 47,81 86,68

Tabela 2 zawiera także współczynnik zgodności φ^2, ponieważ w przypadku minimalizacji maksymalnego błędu względnego nie można wyznaczyć współczynnika korelacji (wzór (3)), gdyż φ^2=1,295.

7. Podsumowanie

Z przedstawionej powyżej porównawczej analizy statystycznej należy wnioskować, że przyjęcie konkretnej miary dopasowania ma istotny wpływ na uzyskiwane wyniki. Wydaje się, że rodzaj tego dopasowania należy określić przed przystąpieniem do obliczeń. Stosowana najczęściej miara dopasowania, jaką jest współczynnik korelacji, może prowadzić do dość zaskakujących wyników, np. w tabeli 2 maksymalizacja R prowadzi do średniego błędu względnego 67,24 % i maksymalnego błędu względnego 1128,22 %, a R=93,38 %. Jak pokazują wyniki przedstawione w tabeli 1, minimalizacja maksymalnego błędu względnego może znacznie ten błąd obniżać nie powodując zwiększenia innych parametrów (współczynnik korelacji, średni błąd względny).

LITERATURA

1. Elandt R.: Statystyka matematyczna w zastosowaniu do doświadczalnictwa rolniczego. PWN, Warszawa 1964.
2. Kaczmarek Z.: Metody statystyczne w hydrologii i meteorologii. Wyd. Komunikacji i Łączności, Warszawa 1970.
3. Lechowicz Z., Szymański A.: Prediction of consolidation of organic soil. Annals of Warsaw Agricultural University. No.20, 1984, p. 55-59.
4. Sas W.: Modelowanie odkształceń gruntów organicznych z uwzględnieniem zmian właściwości ośrodka. Rozprawa doktorska SGGW, Warszawa 2001.
5. Szymański A.: Czynniki warunkujące analizę odkształcenia gruntów organicznych obciążonych nasypem. Rozprawa habilitacyjna. Wyd. SGGW, Warszawa 1991.
6. Szymański A., Sas W.: Deformation characteristtics of organic soils. Annals of Warsaw Agricultural University –SGGW. Land Reclam 32. 2001, p. 3-6.

Recenzent: Prof. zw. dr hab. inż. Zbigniew MŁYNAREK

Abstract

The analysis of statistical methods used in the determination of consolidation characteristics as a function of other parameters is a very important task in the determination of settlement. Selected methods for estimation of coefficients of a regression function and an algorithm to calculate these coefficients are presented in this paper. The results obtained from calculation examples have shown that the reliability of statistical analysis depends on the choice of criteria to assess the fitness of experimental prediction.