DOI: 10.37317/biul-2011-0054

ZYGMUNT KACZMAREK [1]
DARIUSZ R. MAŃKOWSKI [2]
[1] Instytut Genetyki Roślin Polskiej Akademii Nauk w Poznaniu, Pracownia Biometrii
[2] Instytut Hodowli i Aklimatyzacji Roślin, Państwowy Instytut Badawczy w Radzikowie
Pracownia Ekonomiki Nasiennictwa i Hodowli Roślin

Wprowadzenie do statystycznych analiz wielozmiennych∗ Część I. Podstawy teoretyczne

An introduction to multivariate statistical analyses Part I. Theoretical background

Analizy wielocechowe są coraz szerzej stosowane w badaniach rolniczych. Powszechna dostępność pakietów statystycznych realizujących tego typu analizy pozwala na ich powszechne wykorzystywanie. Problemem staje się więc umiejętność właściwego wykorzystania tych analiz i poprawnej interpretacji uzyskanych z nich wyników. W pracy omówiono podstawowe pojęcia analizy wielozmiennej, opisano wielozmienny model liniowy obserwacji, wielocechową analizę wariancji, niezbędne statystyki testowe a także wielocechowe metody oceny podobieństwa obiektów.

Słowa kluczowe: analizy wielocechowe, grupowanie wielocechowe obiektów, macierz korelacji, macierz kowariancji, MANOVA, wielocechowe miary podobieństwa

Multivariate analyses are increasingly used in agricultural research. The widespread availability of statistical packages pursuing this type of analysis allows for their use. Then, the ability of appropriate application of the analysis and correct interpretation of its results becomes problematic. The paper discusses basic concepts of the multivariate analysis, the multivariate model of observations, the multivariate analysis of variance, the appropriate statistical tests and the multivariate methods for estimation of the objects' similarity.

Key words: multivariate analysis, multivariate grouping, correlation matrix, covariance matrix, MANOVA, multivariate similarity measures

WSTĘP

Wielowymiarowy charakter zagadnień stanowiących przedmiot badań biologiczno-rolniczych wymaga stosowania w tych badaniach metod wielocechowych. Wynika to

∗ Praca była prezentowana w ramach I Warsztatów Biometrycznych, które odbyły się w IHAR-PIB w Radzikowie w dniach 14-15 września 2010 r.

ze złożoności zjawisk. Badanie porównawcze obiektów prowadzone tylko w jednym aspekcie nie jest wystarczające do ich wyjaśnienia. Również badanie porównawcze obiektów pod względem wielu cech, ale każdej rozpatrywanej oddzielnie, nie może przyczynić się do pełnego wyjaśnienia zachodzących zjawisk. Dopiero równoczesne uwzględnienie wszystkich obserwowanych cech stanowić może podstawę do wyciągnięcia adekwatnych wniosków. Z tych względów naturalnym postępowaniem w analizie porównawczej obiektów przyrodniczych wydaje się zastosowanie metod wielozmiennych.

Celem niniejszego opracowania jest przedstawienie jednoczynnikowego wielozmiennego modelu liniowego obserwacji i opartej na nim wielozmiennej analizy wariancji (MANOVA). Analizę tę poprzedzą rozważania dotyczące obserwacji wielu zmiennych i korelacji między nimi. Analizę wariancji omówimy dla dwóch najczęściej stosowanych układów eksperymentalnych: układu o klasyfikacji pojedynczej i układu o klasyfikacji podwójnej. W tym ostatnim skupimy się na analizie wariancji układu dwuczynnikowego z jedną obserwacją w podklasie, która w dużej mierze odpowiada analizie prowadzonej dla układu losowanych bloków. Prezentowane będą także możliwości testowania hipotez szczegółowych dotyczących zwłaszcza porównań (kontrastów) między obiektami. Omówione będą miary wielocechowego podobieństwa obiektów, takie jak odległość Euklidesa i odległość Mahalanobisa. Wskażemy na możliwość badania tzw. mocy dyskryminacyjnej cech. W celu graficznej prezentacji rozmieszczenia obiektów wielocechowych w nowej przestrzeni, dwu- lub trójwymiarowej, zasygnalizujemy metody składowych głównych i zmiennych kanonicznych.

W przygotowaniu niniejszego opracowania skorzystano z następujących publikacji: Caliński (1970), Anderberg (1973), Caliński i Kaczmarek (1973), McKeon (1974), Caliński i in. (1975), Kaczmarek (1975), Ceranka i in. (1975), Caliński i in. (1976), Morrison (1976), Seber (1984), Krzanowski (1988), Everitt i Dunn (1992), Krzyśko (2000), Sieczko i in. (2004), Srivastava (2004), Wu i in. (2006), Mądry (2007) Ukalska i in. (2007), Kaczmarek i in. (2008), Ukalska i in. (2008).

1. PRÓBA Z WIELOWYMIAROWEJ POPULACJI O ROZKŁADZIE NORMALNYM

Podstawą teorii metod statystycznych jest założenie, że istnieje zbiorowość generalna elementów rzeczywistych (tzw. populacja generalna), której zbadanie całości jest niemożliwe i dlatego dokonywany jest wybór jej mniejszego ale reprezentatywnego podzbioru zwanego próbą. Analizy oparte na określonych modelach statystycznych pozwalają wyciągać wnioski dotyczące populacji generalnej.

Weźmy pod uwagę wektory n obserwacji p cech w próbie reprezentatywnej rozpatrywanej populacji. Wartości te można zestawić w postaci tzw. $n \times p$-wymiarowej macierzy danych:

$$\underset{(n\times p)}{\mathbf{Y}} = \begin{bmatrix} y_1^{(1)} & y_1^{(2)} & \cdots & y_1^{(p)} \\ y_2^{(1)} & y_2^{(2)} & \cdots & y_2^{(p)} \\ \vdots & \vdots & \ddots & \vdots \\ y_n^{(1)} & y_n^{(2)} & \cdots & y_n^{(p)} \end{bmatrix} = \begin{bmatrix} y^{(1)} & y^{(2)} & \cdots & y^{(p)} \end{bmatrix}. \tag{1.1}$$

Zakładamy, że dane te są obserwacjami *p*-wymiarowej zmiennej losowej w populacji generalnej posiadającej wielowymiarowy rozkład normalny z wektorem średnich **μ** i nieosobliwą macierzą kowariancji **Σ**. Wektor średnich z próby dla jednej (*r*-tej) cechy można zapisać jako:

$$\bar{y}^{(r)} = \frac{1}{n}\sum_{h=1}^{n} y_h^{(r)}, \tag{1.2}$$

natomiast macierz sum kwadratów i iloczynów (SSCP — sum of squares and cross-products) można zapisać w postaci:

$$\mathbf{A} = \{a_{ij}\} = \left\{\sum_{h=1}^{n}\left(y_h^{(r)} - \bar{y}^{(r)}\right)\left(y_h^{(s)} - \bar{y}^{(s)}\right)'\right\}. \tag{1.3}$$

Wówczas:

$$\mathbf{S} = \frac{1}{n-1}\cdot\mathbf{A} \tag{1.4}$$

jest macierzą oceny kowariancji z próby.

Współczynnik korelacji liniowej

Zauważmy, że w przypadku dwóch zmiennych $y^{(1)}$ i $y^{(2)}$ macierz kowariancji **S** można zapisać w postaci:

$$\mathbf{S} = \begin{bmatrix} \mathrm{Var}\left(\mathrm{y}^{(1)}\right) & \mathrm{Cov}\left(\mathrm{y}^{(1)},\mathrm{y}^{(2)}\right) \\ \mathrm{Cov}\left(\mathrm{y}^{(1)},\mathrm{y}^{(2)}\right) & \mathrm{Var}\left(\mathrm{y}^{(2)}\right) \end{bmatrix}.$$

Na podstawie elementów macierzy S można obliczyć współczynnik korelacji liniowej między tymi zmiennymi, będący miarą stopnia współzależności liniowej między nimi. Współczynnik ten ma postać:

$$r_{1,2} = \left[\frac{\mathrm{Cov}\left(\mathrm{y}^{(1)},\mathrm{y}^{(2)}\right)}{\sqrt{\mathrm{Var}(\mathrm{y}^{(1)})\mathrm{Var}\,(\mathrm{y}^{(2)})}}\right].$$

Dla *p* zmiennych $y^{(1)}$, $y^{(2)}$,...,$y^{(p)}$ można utworzyć macierz współczynników korelacji o postaci:

$$\mathbf{R} = \begin{bmatrix} 1 & r_{1,2} & r_{1,3} & \cdots & r_{1,p} \\ r_{2,1} & 1 & r_{2,3} & \cdots & r_{2,p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{p,1} & r_{p,2} & r_{p,3} & \cdots & 1 \end{bmatrix}, \tag{1.5}$$

przy czym $r_{1,2} = r_{2,1}$, $r_{1,3} = r_{3,1}$,...,$r_{1,p} = r_{p,1}$, itd.

2. WIELOZMIENNA ANALIZA WARIANCJI DLA KLASYFIKACJI POJEDYNCZEJ

Rozszerzeniem jednozmiennej analizy wariancji na przypadek *p* cech jest wielozmienna analiza wariancji — MANOVA.

Weźmy pod uwagę *k* obiektów o liczebnościach n_i(i = 1,2,...,k) obserwowanych pod względem *p* cech ($n = \sum_{i=1}^{k} n_i$). Przez $y_{ij}^{(r)}$ oznaczmy obserwację *i*-tego obiektu w *j*-tym powtórzeniu (j = 1,2,...,n_i) uzyskaną dla *r*-tej cechy (r = 1,2,...,p).

Obserwacje te można zestawić w postaci tzw. tablicy danych:

$$\underset{(n \times p)}{\mathbf{Y}} = \begin{bmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \\ \vdots \\ \mathbf{Y}_k \end{bmatrix} \quad \text{gdzie} \quad \underset{(n_i \times p)}{\mathbf{Y}_i} = \begin{bmatrix} y_{i1}^{(1)} & y_{i1}^{(2)} & \cdots & y_{i1}^{(p)} \\ y_{i2}^{(1)} & y_{i2}^{(2)} & \cdots & y_{i2}^{(p)} \\ \vdots & \vdots & \ddots & \vdots \\ y_{in_i}^{(1)} & y_{in_i}^{(2)} & \cdots & y_{in_i}^{(p)} \end{bmatrix} \tag{2.1}$$

Obserwacje dla tak zestawionej macierzy danych można opisać za pomocą modelu:

$$y_{ij}^{(r)} = \mu_i^{(r)} + e_{ij}^{(r)} = \mu^{(r)} + t_i^{(r)} + e_{ij}^{(r)}, \tag{2.2}$$

gdzie $y_{ij}^{(r)}$ jest *j*-tą obserwacją *i*-tego obiektu dla *r*-tej cechy; (i = 1,2,...,k;j = 1,2,...,n_i; r = 1,2,...,p); $\mu^{(r)}$ jest średnią ogólną dla *r*-tej cechy; $t_i^{(r)}$ oznacza efekt główny *i*-tego obiektu dla *r*-tej cechy; zaś $e_{ij}^{(r)}$ są błędami eksperymentalnymi dla *r*-tej cechy.

Zakładamy, że $\sum_{i=1}^{k} t_i^{(r)} = 0$ i przyjmujemy te same założenia jak w przypadku jednej cechy. Przyjmujemy zatem założenie, że $E(e_{ij}^{(r)} e_{i'j'}^{(r)}) = 0$ dla $(i,j) \neq (i'j')$. Dodatkowo zakładamy, że każda próba Y_i składa się z n_i obserwacji *p*-wymiarowej zmiennej losowej o wektorze średnich $\boldsymbol{\mu}_i$:

$$\boldsymbol{\mu}_i = \left[\mu_i^{(1)} \mu_i^{(2)} \ \ldots \ \mu_i^{(P)}\right] \qquad \text{gdzie} \qquad \mu_i^{(r)} = \mu^{(r)} + \mu_i^{(r)}$$

i macierzy kowariancji $\boldsymbol{\Sigma}$:

$$\Sigma = \begin{bmatrix} \sigma^{(11)} & \sigma^{(12)} & \cdots & \sigma^{(1p)} \\ \sigma^{(21)} & \sigma^{(22)} & \cdots & \sigma^{(2p)} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma^{(p1)} & \sigma^{(p2)} & \cdots & \sigma^{(pp)} \end{bmatrix}.$$

Dla potrzeb testowania przyjmujemy założenie, że obserwacje te dotyczą zmiennych o łącznym wielowymiarowym rozkładzie normalnym.

Model (2.2) umożliwia przeprowadzenie wielozmiennej analizy wariancji, w której macierz sum kwadratów i iloczynów dla zmienności całkowitej, $\mathbf{S}_G$, jest podzielona na macierz sum kwadratów i iloczynów dla obiektów, $\mathbf{S}_T$, i macierz sum kwadratów i iloczynów dla błędu, $\mathbf{S}_E$, (tab. 1). Każdej z tych macierzy odpowiada określona liczba stopni swobody. Macierze sum kwadratów i iloczynów podzielone przez odpowiadające im stopnie swobody nazywają się macierzami średnich kwadratów i iloczynów.

Tabela 1

Trójwymiarowa analiza wariancji dla klasyfikacji pojedynczej
Three-dimensional analysis of variance for the one-way classification

Źródło zmienności Source of variation	Suma kwadratów dla cechy Sum of squares for treatment			Macierz sum kwadratów i iloczynów Sum of squares and cross-products matrix	Stopnie swobody Degrees of freedom	Macierz średnich kwadratów i iloczynów Mean squares and cross-products matrix
	1	2	3			
Obiekty Objects	T_{11}	T_{22}	T_{33}	$\mathbf{S}_T = \begin{bmatrix} T_{11} & T_{12} & T_{13} \\ T_{21} & T_{22} & T_{23} \\ T_{31} & T_{32} & T_{33} \end{bmatrix}$	$k-1$	$\mathbf{M}_T = \frac{\mathbf{S}_T}{k-1}$
Błędy Residuals	E_{11}	E_{22}	E_{33}	$\mathbf{S}_E = \begin{bmatrix} E_{11} & E_{12} & E_{13} \\ E_{21} & E_{22} & E_{23} \\ E_{31} & E_{32} & E_{33} \end{bmatrix}$	$n-k$	$\mathbf{M}_E = \frac{\mathbf{S}_E}{n-k}$
Razem Total	G_{11}	G_{22}	G_{33}	$\mathbf{S}_G = \begin{bmatrix} G_{11} & G_{12} & G_{13} \\ G_{21} & G_{22} & G_{23} \\ G_{31} & G_{32} & G_{33} \end{bmatrix}$	$n-1$	

W celu przeprowadzenia analizy wariancji opartej na modelu (2.2) wprowadźmy następujące oznaczenia:

— $Y_{i.}^{(r)} = \sum_{j=1}^{n_i} y_{ij}^{(r)}$ — jest sumą obserwacji *i*-tego obiektu dla *r*-tej cechy,

— $\bar{y}_{i.}^{(r)} = \frac{1}{n_i} Y_{i.}^{(r)}$ — jest średnią *i*-tego obiektu dla *r*-tej cechy,

— $Y_{..}^{(r)} = \sum_{i=1}^{k} Y_{i.}^{(r)}$— jest sumą wszystkich obserwacji dla *r*-tej cechy,

— $\bar{y}_{..}^{(r)} = \frac{1}{n_i} Y_{..}^{(r)}$ — jest średnią ogólną dla *r*-tej cechy ($n = n_1 + n_2 + \ldots n_k$).

Wówczas składowe macierzy sum kwadratów i iloczynów $\mathbf{S}_G$ dla zmienności całkowitej wyliczamy ze wzoru:

$$G_{rs} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij}^{(r)} y_{ij}^{(s)} - \frac{1}{n} Y_{..}^{(r)} Y_{..}^{(s)},$$

dla macierzy sum kwadratów i iloczynów dla obiektów, $\mathbf{S}_T$, ze wzoru:

$$T_{rs} = \frac{Y_{1.}^{(r)} Y_{1.}^{(s)}}{n_1} + \frac{Y_{2.}^{(r)} Y_{2.}^{S(s)}}{n_2} + \cdots + \frac{Y_{k.}^{(r)} Y_{k.}^{(s)}}{n_k} - \frac{Y_{..}^{(r)} Y_{..}^{(s)}}{n},$$

a dla macierzy sum kwadratów i iloczynów dla błędu, $\mathbf{S}_E$, ze wzoru:

$$E_{rs} = \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}^{(r)} y_{ij}^{(s)} - \frac{Y_{1.}^{(r)} Y_{1.}^{(s)}}{n_1} + \frac{Y_{2.}^{(r)} Y_{2.}^{(s)}}{n_2} + \cdots + \frac{Y_{k.}^{(r)} Y_{k.}^{(s)}}{n_k},$$

gdzie $r, s = 1,2,...,p$.

3. WIELOZMIENNA ANALIZA WARIANCJI DLA KLASYFIKACJI PODWÓJNEJ Z JEDNĄ OBSERWACJĄ W PODKLASIE

Oprócz eksperymentów, w których badany jest wpływ jednego czynnika na zmienne wynikowe, spotyka się często eksperymenty bardziej złożone, a mianowicie takie, w których badany jest wpływ dwóch lub większej liczby czynników. Jeżeli w eksperymencie działają dwa czynniki równocześnie na wielu poziomach, to obserwacje sklasyfikowane są według dwóch kryteriów, czyli tworzą klasyfikację podwójną. Można przy tym rozróżnić modele z jedną i wieloma obserwacjami w podklasie. Obecnie zajmiemy się analizą wariancji wielu cech dla klasyfikacji podwójnej z jedną obserwacją w podklasie.

Załóżmy, że wszystkie badane obiekty (poziomy) czynnika A i wszystkie obiekty (poziomy) czynnika B opisane są za pomocą p rożnych zmiennych (cech). Rozważamy zatem sytuację, w której $n = a \cdot b$ p-wymiarowych obserwacji sklasyfikowano według przynależności do jednego z a poziomów czynnika A i jednego z b poziomów czynnika B.

Obserwacje te można zestawić w postaci $(ab \times p)$-wymiarowej macierzy danych:

$$\mathbf{Y} = \begin{bmatrix} y_{11}^{(1)} & y_{11}^{(2)} & \cdots & y_{11}^{(p)} \\ \vdots & \vdots & \cdots & \vdots \\ y_{1b}^{(1)} & y_{1b}^{(2)} & \cdots & y_{1b}^{(p)} \\ y_{21}^{(1)} & y_{21}^{(2)} & \cdots & y_{21}^{(p)} \\ \vdots & \vdots & \cdots & \vdots \\ y_{2b}^{(1)} & y_{2b}^{(2)} & \cdots & y_{2b}^{(p)} \\ \vdots & \vdots & \cdots & \vdots \\ y_{a1}^{(1)} & y_{a1}^{(2)} & \cdots & y_{a1}^{(p)} \\ \vdots & \vdots & \cdots & \vdots \\ y_{ab}^{(1)} & y_{ab}^{(2)} & \cdots & y_{ab}^{(p)} \end{bmatrix}.$$

Model matematyczny dla $y_{ij}^{(r)}$, czyli obserwacji i-tego poziomu czynnika A ($i = 1,2,...,a$) oraz j-tego poziomu czynnika B ($j = 1,2,...,b$) dotyczącej r-tej cechy ($r = 1,2,...,p$), można przedstawić w postaci:

$$y_{ij}^{(r)} = \mu^{(r)} + \alpha_i^{(r)} + \beta_j^{(r)} + e_{ij}^{(r)}, \tag{3.1}$$

gdzie: $\mu^{(r)}$jest średnią ogólną dla *r*-tej cechy, $\alpha_i^{(r)}$jest efektem *i*-tego poziomu czynnika A dla *r*-tej cechy,$\beta_j^{(r)}$ jest efektem *j*-tego poziomu czynnika B dla *r*-tej cechy, $e_{ij}^{(r)}$ są błędami eksperymentalnymi dla *r*-tej cechy.

Podobnie jak w przypadku modelu (2.2) zakładamy, że zmienne losowe $y_{ij}^{(r)}$ są niezależne i posiadają rozkład normalny oraz, że wariancja wszystkich$y_{ij}^{(r)}$ jest jednakowa. Zakładamy ponadto, że:

$$\sum_{i=1}^{a} \alpha_i^{(r)} = \sum_{j=1}^{b} \beta_j^{(r)} = 0\,.$$

Stosując metodę najmniejszych kwadratów można uzyskać estymatory parametrów modelu: $\hat{\mu}^{(r)} = \bar{y}_{..}^{(r)}$; $\hat{\alpha}_i^{(r)} = \bar{y}_{i.}^{(r)} - \bar{y}_{..}^{(r)}$; $\hat{\beta}_j^{(r)} = \bar{y}_{.j}^{(r)} - \bar{y}_{..}^{(r)}$.

Model (3.1) umożliwia przeprowadzenie wielozmiennej analizy wariancji. Jeśli w modelu (3.1) przyjmiemy, że $\beta_j^{(r)}$ jest efektem *j*-tego bloku dla *r*-tej cechy to może on być modelem obserwacji dla układu losowanych bloków.

4. TESTOWANIE HIPOTEZ W OGÓLNYM WIELOZMIENNYM MODELU LINIOWYM

Testowanie hipotezy ogólnej

Weźmy pod uwagę ogólny wielozmienny model liniowy analizy wariancji dla danych z doświadczenia, w którym *k* obiektów obserwowanych jest pod względem *p* cech a liczba wszystkich obserwacji wynosi *n*.

Stosując symbolikę macierzową model ten można zapisać w postaci:

$$\mathbf{Y} = \mathbf{XT} + \mathbf{E}\,, \tag{4.1}$$

gdzie: $\mathbf{Y} = \{y_{\boldsymbol{hr}}\}$ jest $(n \times p)$-wymiarową macierzą obserwacji, $\mathbf{X} = \{x_{\boldsymbol{hi}}\}$ jest $(n \times k)$-wymiarową macierzą układu, rzędu $\leq n - p$, w której elementy są stałymi współczynnikami, $\mathbf{T} = \{t_{\boldsymbol{ir}}\}$ jest $(k \times p)$-wymiarową macierzą nieznanych parametrów określających efekty działania czynników kontrolowanych w doświadczeniu na występujące zmienne, $\mathbf{E} = \{e_{\boldsymbol{hr}}\}$ jest $(n \times p)$-wymiarową macierzą błędów losowych, (i = 1,2,...,k; h = 1,2,...,n; r = 1,2,...,p) .

Tak sformułowany model pozwala na weryfikację następującej hipotezy ogólnej:

$$H_{00}\colon \mathbf{CTM} = 0\,, \tag{4.2}$$

gdzie: macierz $\mathbf{C}$ o *g* wierszach i *k* kolumnach określa treść hipotezy dotyczącej kombinacji liniowych wierszy macierzy parametrów $\mathbf{T}$ (jest rzędu *g*), a macierz $\mathbf{M}$ o *p* wierszach i *u* kolumnach określa treść hipotezy dotyczącą kombinacji liniowych kolumn macierzy $\mathbf{T}$ (jest rzędu *u*).

Do testowania hipotezy ogólnej H_{00} stosowane są następujące statystyki testowe oparte na macierzach sum kwadratów i iloczynów, $\mathbf{S}_H$ (dla hipotezy) i $\mathbf{S}_E$ (dla błędu):

— statystyka Roya wykorzystująca maksymalną wartość własną iloczynu macierzy $\mathbf{S}_E^{-1}\mathbf{S}_H$,
— statystyka Hotellinga-Lawleya T^2, wykorzystująca ślad iloczynu macierzy $\mathbf{S}_E^{-1}\mathbf{S}_H$ pomnożony przez liczbę stopni swobody dla błędu,
— iloraz wiarogodności Λ Wilksa, będący ilorazem wyznaczników macierzy $|\mathbf{S}_E|$ i $|\mathbf{S}_E + \mathbf{S}_H|$,
— ślad Pillaia tj. ślad iloczynu $(\mathbf{S}_E + \mathbf{S}_H)^{-1}\mathbf{S}_H$.

W przypadku analizy jednozmiennej wszystkie powyższe statystyki testowe można sprowadzić do statystyki *F*, natomiast w przypadku analizy wielozmiennej każda ze statystyk (a) – (d) daje zwykle różne przybliżenie statystyki *F*, z wyjątkiem sytuacji gdy *g* = 1. W praktyce należy oczekiwać, że jeśli założenia modelu są w znacznym stopniu spełnione i jeśli któraś ze statystyk testowych w analizie wielozmiennej wyraźnie odrzuca hipotezę H_{00}, to pozostałe statystyki również ją odrzucą.

Przydatne w praktyce mogą być przekształcenia tych statystyk do statystyki *F*. Dla statystyk (b) i (c) przedstawiają się one następująco:

Ad (b): Do statystyki Hotellinga-Lawleya zastosowanie ma przekształcenie McKeona (1974) postaci:

$$\frac{1}{c}\text{ślad}\,(\mathbf{S}_E^{-1}\mathbf{S}_H) \sim F_{a,b}\,, \tag{4.3}$$

gdzie: $a = um_H$; $b = 4 + (a+2)/(d-1)$; $c = a(b-2)/[b(m_E - u - 1)]$; $d = (m_E + m_H - u - 1)(m_E - 1)/[m_E - u - 3)(m_E - u)]$, przy czym m_H i m_E oznaczają liczby stopni swobody odpowiednio dla hipotezy i dla błędu. Przedstawiona statystyka ma dokładnie rozkład *F*, gdy $m_H = 1$.

Ad (c): Do statystyki Wilksa stosować można przekształcenie postaci:

$$F = \frac{1 - \Lambda^{1/w}}{\Lambda^{1/w}} \cdot \frac{dw - z}{um_H} \sim F_{um_H;dw-z}\,, \tag{4.4}$$

gdzie: $w = \sqrt{\frac{u^2 m_H^2 - 4}{u^2 + m_H^2 - 5}}$, $d = m_E - \frac{1}{2}(u - m_H + 1), z = \frac{um_H - 2}{2}$ (pamiętając, że $m_H = g$). Przekształcona statystyka ma dokładnie rozkład *F*, gdy *u* lub m_H jest równe 1 lub 2.

Testowanie hipotez szczegółowych

Odrzucenie wielozmiennej hipotezy liniowej H_{00} pociąga za sobą zwykle indywidualne sprawdzanie hipotez szczegółowych. Interesujące mogą być zarówno indywidualne kombinacje liniowe kolumn macierzy parametrów $\mathbf{T}$, jak również indywidualne kombinacje liniowe wierszy macierzy $\mathbf{T}$, a także indywidualne kombinacje równocześnie kolumn i wierszy macierzy $\mathbf{T}$. Weryfikowane mogą być następujące hipotezy szczegółowe:
— a)hipoteza dotycząca pojedynczej kombinacji liniowej kolumn macierzy $\mathbf{CT}$ postaci:

$$H_{0l} : \mathbf{CTm}_l = \mathbf{0}, l = 1,2,\ldots,u^*$$

testowana za pomocą statystyki $F = \frac{m_E}{m_H}\frac{SS_{H,l}}{SS_{E,l}}$ i porównywana z wartością krytyczną $F_{\alpha;m_H,m_E}$,

— b)hipoteza dotycząca pojedynczej kombinacji liniowej wierszy macierzy **TM** postaci:

$$H_{f0} : \mathbf{c}_f'\mathbf{TM} = \mathbf{0}, f = 1,2,\dots,g^*$$

testowana za pomocą statystyki $F = \frac{m_E - u + 1}{u}$ ślad $\left(\mathbf{S}_{E,f}^{-1}\mathbf{S}_{H,f}\right)$ i porównywana z wartością krytyczną $F_{\alpha;u,m_E-u+1}$,

— c) hipoteza dotycząca pojedynczej kombinacji liniowej wierszy i równocześnie kolumn macierzy **T** postaci:

$$H_{fl} : \mathbf{c}_f'\mathbf{Tm}_l = \mathbf{0},$$

testowaną za pomocą statystyki $F = m_E \frac{SS_{H,fl}}{SS_{E,fl}}$ i porównywana z wartością krytyczną $F_{\alpha;1,m_E}$.

Badanie mocy dyskryminacyjnej zmiennych

W analizie doświadczenia wielocechowego możemy być zainteresowani badaniem mocy dyskryminacyjnej dowolnego podzbioru zmiennych, rozpatrywanej z uwagi na pewną hipotezę *H*. Przez moc dyskryminacyjną zmiennych będziemy tu rozumieli ich udział w odrzucaniu hipotezy *H*.

Dla określenia mocy dyskryminacyjnej zmiennych możemy posłużyć się statystyką warunkową Wilksa postaci:

$$\Lambda_{s/t} = \frac{\Lambda_{s+t}}{\Lambda_t} \tag{4.5}$$

gdzie:Λ_{s+t} jest statystyką Wilksa zbudowaną dla wszystkich $p = s + t$ zmiennych, natomiast Λ_t jest taką statystyką zbudowaną dla ustalonych t ($<p$) zmiennych, których wpływ chcemy wyeliminować.

Przekształcenie statystyki $\Lambda_{\frac{s}{t}}$ w funkcję testową *F* można zapisać w postaci:

$$F = \frac{1 - \Lambda_{s/t}^{1/w}}{\Lambda_{s/t}^{1/w}} \cdot \frac{dw - z}{um_H}, \tag{4.6}$$

przy czym wielkości *u, m_H*, *d*, *w*, *z* zostały określone w statystykach (4.3) i (4.4).

Przyjmując $s = 1$ możemy za pomocą statystyki (4.6) zbadać moc dyskryminacyjną jednej zmiennej przy $p-1$ zmiennych ustalonych. Przeprowadzając to badanie dla każdej zmiennej możemy wyeliminować z dalszej analizy zmienne najsłabiej dyskryminujące.

5. MIARY WIELOCECHOWEGO PODOBIEŃSTWA OBIEKTÓW

Populacja wielocechowa, czyli populacja której elementy są scharakteryzowane przez wiele cech, może być matematycznie określona przez wektor średnich **μ** oraz macierz wariancji i kowariancji znaną jako macierz kowariancji ∑.

Weźmy pod uwagę doświadczenie, w którym każdy z *k* obiektów jest replikowany n_i razy ($i = 1,2,...,k$) i obserwowany pod względem *p* cech ($r=1,2,...,p$). Zapiszmy wektory średnich dla obiektów *i* oraz *j* ($i \neq j$) = 1,2,...,*k* w postaci:

$$\bar{\mathbf{y}}_i = \begin{bmatrix} \bar{y}_i^{(1)} \\ \bar{y}_i^{(2)} \\ \vdots \\ \bar{y}_i^{(p)} \end{bmatrix}; \; \bar{\mathbf{y}}_j = \begin{bmatrix} \bar{y}_j^{(1)} \\ \bar{y}_j^{(2)} \\ \vdots \\ \bar{y}_j^{(p)} \end{bmatrix}, \tag{5.1}$$

gdzie: $\bar{y}_i^{(r)}$ ($\bar{y}_j^{(r)}$) są średnimi obiektu *i* (obiektu *j*) cechy *r* (*r*=1,2,...,*p*).

Punkty reprezentowane przez wektory o *p* składowych (*p* średnich obiektowych) mogą być przedstawione jako punkty w przestrzeni *p*-wymiarowej. Geometrycznie odległość między tymi punktami, czyli między obiektami i i j można określić za pomocą odległości Euklidesa d_{ij} postaci:

$$d_{ij} = \left[\sum_{r=1}^{P} \left(\bar{y}_i^{(r)} - \bar{y}_j^{(r)} \right)^2 \right]^{\frac{1}{2}}; \; i \neq j \, . \tag{5.2}$$

Ten sposób wyrażenia podobieństwa obiektów ma jednak poważne ograniczenie wynikające z całkowitej ignorancji korelacji między badanymi cechami. Fakt ten sprawia, że odległość Euklidesa może być wykorzystywana jako miara wielocechowego podobieństwa obiektów jedynie w przypadku cech w pełni niezależnych.

Z tych względów jako miarę wielocechowego podobieństwa obiektów bezpieczniej przyjąć uogólnioną odległość Mahalanobisa. Jej kwadrat można zapisać w postaci następującej formy kwadratowej:

$$D_{ij}^2 = \left(\bar{\mathbf{y}}_i - \bar{\mathbf{y}}_j\right)' \bar{\Sigma}^{-1} \left(\bar{\mathbf{y}}_i - \bar{\mathbf{y}}_j\right) = \left(\bar{\mathbf{y}}_i - \bar{\mathbf{y}}_j\right)' \mathbf{M}_E^{-1} \left(\bar{\mathbf{y}}_i - \bar{\mathbf{y}}_j\right) \tag{5.3}$$

gdzie: $\mathbf{M}_E^{-1}$ jest odwrotnością oszacowanej macierzy kowariancji (odwrotnością macierzy średnich kwadratów i iloczynów dla błędu w wielozmiennej analizie wariancji), a wektory $\bar{\mathbf{y}}_i$ i $\bar{\mathbf{y}}_j$ zostały zdefiniowane w (5.1).

Dla weryfikacji hipotezy orzekającej, że oba obiekty mają takie same średnie, czyli, że odległość między nimi wynosi zero, można zastosować następującą „odległość krytyczną”:

$$D_\alpha^2 = \left[\frac{1}{n_i} + \frac{1}{n_j} \right] \cdot \frac{p m_E}{m_E - p + 1} F_{\alpha;p,m_E-p+1} \tag{5.4}$$

gdzie: n_i (n_j) są liczbami replikacji dla *i*-tego (*j*-tego) obiektu (patrz układ całkowicie losowy), *p* oznacza liczbę cech, m_E jest liczbą stopni swobody odpowiadających macierzy $\mathbf{S}_E$, a $F_{\alpha;p,m_E-p+1}$ jest wartością krytyczną odczytaną z tablic Fishera-Snedecora na poziomie istotności α dla *p* i m_E –p +1 stopni swobody.

Dla układu losowanych bloków, czyli w sytuacji gdy liczba replikacji każdego obiektu jest równa liczbie bloków *b*, odległość krytyczna D_α przyjmuje postać:

$$D_\alpha^2 = \frac{2 p m_E}{b(m_E - p + 1)} F_{\alpha;p,m_E-p+1}. \tag{5.5}$$

Powyższe testowanie można przeprowadzić przy założeniu istnienia wspólnej macierzy kowariancji oraz faktu, że łączny rozkład obserwowanych *p* zmiennych jest normalny.

6. GRUPOWANIE OBIEKTÓW WIELOCECHOWYCH

Analiza wariancji wyników doświadczenia, którego obiekty obserwowane są pod względem określonych cech (zmiennych) kończy się z reguły testowaniem hipotezy o równości jedno- lub wielocechowych średnich obiektowych. Odrzucenie tej hipotezy nie daje jeszcze informacji o tym, które z badanych obiektów różnią się między sobą istotnie, a między którymi różnice te nie są istotne. Procedury obliczeniowe i pakiety statystyczne dotyczące analizy doświadczeń wielocechowych rzadko uwzględniają potrzebę grupowania obiektów. Zawarte w nich metody wieloimienne, takie jak analiza składowych głównych, czy też analiza zmiennych kanonicznych (Caliński i in., 1975) umożliwiają co prawda znalezienie graficznych obrazów rozmieszczenia obiektów na płaszczyźnie, pozwalają nawet wyznaczyć pewne ich skupienia, jednakże nie dokonują formalnego i obiektywnego podziału obiektów na grupy w maksymalnym stopniu wewnętrznie jednorodne pod względem badanego zespołu cech. Również odległości Mahalanobisa, a także odległości Euklidesa stanowią co najwyżej podstawę do wykreślenia dendrogramu lub dendrytu najkrótszych połączeń mimo, że uznawane są często za miarę wielocechowego podobieństwa obiektów.

W latach siedemdziesiątych i osiemdziesiątych zaproponowano wiele metod statystycznych i procedur obliczeniowych grupowania obiektów wielocechowych, lecz nie znalazły one częstego wykorzystania w praktyce. Jedną z propozycji grupowania obiektów opartą na analizie skupień jest metoda wykorzystująca odległości Mahalanobisa oraz odległości krytyczne. W metodzie tej, grupującej obiekty o wielowymiarowych rozkładach normalnych ze wspólną macierzą kowariancji, odległość Mahalanobisa jest traktowana jako miara podobieństwa między dwoma obiektami. Metoda może być stosowana zarówno dla obiektów jednocechowych, jak i wielocechowych. Dla obu sytuacji obowiązuje to samo kryterium grupowania obiektów. Proces tworzenia grup jednorodnych trwa do momentu, w którym kwadrat najmniejszej odległości między grupami obiektów, min D^2, będzie większy od kwadratu odległości krytycznej D_α^2.

PODSUMOWANIE

Omówione w niniejszej pracy zagadnienia można uznać za podstawy analiz wielocechowych. Od niedawna analizy te stają się coraz bardziej popularne w badaniach rolniczych, genetyce i hodowli roślin. Staje się tak z racji powszechnej dostępności pakietów statystycznych wyposażonych w narzędzia i analizy dla danych wielocechowych, oraz z racji możliwości obliczeniowych współczesnych komputerów. Powszechna dostępność wielocechowych metod analizy danych niestety często nie idzie w parze z umiejętnością poprawnego przeprowadzenia takich analiz i właściwą interpretacją uzyskanych wyników. Dlatego też niezmiernie ważnym jest by zrozumieć podstawy tych analiz, gdyż to pozwala na właściwe ich przeprowadzenie i poprawne wnioskowanie.

LITERATURA

Anderberg M. R. 1973. Cluster Analysis for Applications. Academic Press, New York.
Caliński T. 1970. Wielozmienna analiza wariancji i pokrewne metody wielowymiarowe. PAN, Warszawa.
Caliński T., Czajka S., Kaczmarek Z. 1975. Analiza składowych głównych i jej zastosowania. Algorytmy biometryczne i statystyczne (ABS-36). AR Poznań.
Caliński T., Dyczkowski A., Kaczmarek Z. 1976. Testowanie hipotez w wielozmiennej analizie wariancji i kowariancji. Algorytmy biometryczne i statystyczne (ABS-45). AR Poznań.
Caliński T., Kaczmarek Z. 1973. Metody kompleksowej analizy doświadczenia wielocechowego. Trzecie Colloquium Metodologiczne z Agro-Biometrii, PAN i PTB Warszawa: 257 — 320.
Ceranka B., Chudzik H., Kaczmarek Z., Krzyśko M. 1975. Wielozmienna analiza wyników doświadczeń w układach blokowych. Algorytmy biometryczne i statystyczne (ABS-35). AR Poznań.
Everitt B. S., Dunn G. 1992. Applied Multivariate Data Analysis. Oxford University Press. New York.
Kaczmarek Z. 1975. Wielozmienna analiza kowariancji i jej niektóre zastosowania. Matematyka Stosowana 5: 139 — 156.
Kaczmarek Z., Czajka S., Adamska E. 2008. Propozycja metody grupowania obiektów jedno- i wielocechowych z zastosowaniem odległości Mahalanobisa i analizy skupień. Biul. IHAR 249: 9 — 18.
Krzanowski W. J. 1988. Principles of multivariate analysis: a users's perspective. Oxford University Press.
Krzyśko M. 2000. Wielowymiarowa analiza statystyczna. Wydawnictwo Naukowe UAM, Poznań.
Mądry W. 2007. Metody statystyczne do oceny różnorodności fenotypowej dla cech ilościowych w kolekcjach roślinnych zasobów genowych. Zesz. Probl. Post.-Nauk Rol. 517: 21 — 41.
McKeon J. J. 1974. F approximations to the distribution of Hotelling's T^2_0. Biometrika 61: 381 — 383.
Morrison D. F. 1976. Multivariate Statistical Methods. McGraw-Hill. New York.
Seber G. A. F. 1984. Multivariate Observations. Wiley. New York.
Sieczko L., Mądry W., Zieliński A., Paderewski J., Urbaś-Szwed K. 2004. Zastosowanie analizy składowych głównych w badaniach nad wielocechową charakterystyką zmienności genetycznej w kolekcji zasobów genowych pszenicy twardej (*Triticum durum* L.). XXXIV Coll. Biometryczne: 223 — 239.
Srivastava M. S. 2004. Multivariate theory for analyzing high-dimensional data. Technical Report, University of Toronto, Toronto, Canada.
Ukalska J., Mądry W., Ukalski K., Masny A. 2007. Wielowymiarowa ocena różnorodności fenotypowej w kolekcji zasobów genowych truskawki. Cz. II. Grupowanie genotypów. Zeszyty Prob. Postępów Nauk Rolniczych 517: 759 — 766.
Ukalska J., Ukalski K., Śmiałowski T., Mądry W. 2008. Badanie zmienności i współzależności cech użytkowych w kolekcji roboczej pszenicy ozimej (*Triticum aestivum* L.) za pomocą metod wielowymiarowych. Cz. II. Analiza składowych głównych na podstawie macierzy korelacji fenotypowych i genotypowych. Biul. IHAR 249: 45 — 57.
Wu Y., Genton M. G., Stefanski L. A. 2006. A Multivariate two sample mean test for small sample size and missing data. Biometrics 62: 877 — 885.