DOI: 10.37317/biul-2012-0078

ZBIGNIEW LAUDAŃSKI [1]
DARIUSZ R. MAŃKOWSKI [2]
MAŁGORZATA FLASZKA [3]
[1] Zakład Biometrii, Wydział Zastosowań Informatyki i Matematyki, SGGW Warszawa
[2] Pracownia Ekonomiki Nasiennictwa i Hodowli Roślin, Zakład Nasiennictwa i Nasionoznawstwa IHAR — PIB Radzików
[3] Katedra Doświadczalnictwa i Bioinformatyki, Wydział Rolnictwa i Biologii, SGGW Warszawa

Eksploracyjna analiza czynnikowa w badaniach struktury zespołu zmiennych obserwowanych

Exploratory factor analysis in studying the structure of multivariate observations

W pracy przedstawiono podstawy i zastosowania praktyczne w naukach rolniczych eksploracyjnej analizy czynnikowej, EFA. Opisano zagadnienia numeryczne związane z prowadzeniem niezbędnych obliczeń. Na przykładach przedstawiono sposób interpretacji wyników EFA oraz przykłady jej wykorzystania w badaniach rolniczych.

Słowa kluczowe: analiza czynnikowa EFA, analiza składowych głównych PCA, korelacje

The paper presents the basic principles and practical applications in agricultural sciences of exploratory factor analysis, EFA. We described the numerical issues related to conducting the necessary calculations. Presented examples demonstrated how to interpret the results of EFA and examples of its use in agricultural research.

Key words: factor analysis EFA, principal component analysis PCA, correlations

WSTĘP

Nauki empiryczne opierają się na doświadczeniach, w których badane są wpływy różnych czynników istotnych dla badanego problemu — zjawiska lub procesu. Często w wielu dziedzinach życia czynniki istotne dla danego problemu musimy dopiero identyfikować. Są one określane jako czynniki wspólne lub latentne (ang. common factors, latent factors). Do tego celu przydatne są metody specjalnego typu, metody badające bezpośrednie związki (korelacje) między wieloma zmiennymi (cechami) obserwowalnymi i na podstawie tych związków pozwalające na identyfikację czynników wspólnych.

Wiadomym jest, że istnienie korelacji między dwoma pomiarami (cechami) nie daje jeszcze bezpośrednio żadnych informacji o powiązaniach przyczynowo-skutkowych. Podając jednak wzajemne korelacje między wieloma zmiennymi, możemy snuć pewne przypuszczenia o układzie czynników przyczynowych. Otóż operując korelacjami między

wieloma cechami możemy oczekiwać pewnej nadmiarowości, tzn. pokrywania się informacji zawartych w poszczególnych cechach, gdyż jedne korelacje wyznaczają wartości innych, a zawartą w tych korelacjach informację można wyrazić przy użyciu mniejszej liczby zmiennych, tzw. hipotetycznych wpływów przyczynowych — inaczej czynników. Przykładowo, jeżeli stwierdzamy, że można odtworzyć korelacje między piętnastoma cechami przy pomocy trzech czynników wspólnych (latentnych), to mamy prawo przypuszczać, że są one odpowiedzialne za całość, bądź zdecydowaną większość zmienności i współzmienności badanego zespołu cech. Tak określone postępowanie nosi nazwę eksploracyjnej analizy czynnikowej (EFA, ang. Exploratory Factor Analysis).

W przygotowaniu niniejszego opracowania skorzystano z następujących publikacji: Thurstone (1931, 1947), Kaiser (1958), Rao (1964), Gower (1966), Caliński i in. (1975), Cureton i Mulaik (1975), Gnanadesikan (1977), Szczotka (1977), Kim i Mueller (1978), Mardia i in. (1979), Seber (1984), Krzanowski (1988), Wójcik i Laudański (1989), Morrison (1990), Everitt i Dunn (1992), Hatcher (1994), Khattre i Naik (1999, 2000), Krzyśko (2000, 2009), Timm (2002), Sieczko i in. (2004, 2008), Armitage i Colton (2005), O'Rourke i in. (2005), Mądry (2007), Ukalska i in. (2007, 2008), Mańkowski i in. (2009), Walesiak i Gatnar (2009), Mądry i in. (2010).

Prezentowane w pracy przykłady obliczeniowe analizowano z wykorzystaniem Systemu SAS 9.2 (SAS Institute Inc. 2009) [Przykład 1] oraz pakietu SPSS (IBM SPSS Statistics) [Przykład 2].

ANALIZA SKŁADOWYCH GŁÓWNYCH

Analiza składowych głównych jest to metoda obliczeniowa analizowania struktury zbioru danych prezentowanych przez wektory zmiennych (cech) na podstawie współzależności między nimi. Cała informacja potrzebna do wyznaczenia składowych głównych zawiera się w macierzy kowariancji (**C**) analizowanych zmiennych. Nie mniej jednak, często analizę przeprowadzamy na macierzy kowariancji zmiennych standaryzowanych. Przy czym standaryzacja zmiennych realizowana jest przez standaryzację obserwacji: odjęcie odpowiedniej wartości średniej i podzielenie przez odchylenie standardowe. Praktycznie, jest to macierz korelacji analizowanych zmiennych (**R**). Składowe główne otrzymane dla macierzy **C** i dla macierzy **R** nie muszą być takie same, a przejście od jednych składowych do drugich w prosty sposób najczęściej nie jest możliwe. Własności składowych głównych konstruowanych na podstawie macierzy korelacji są o wiele bardziej kompleksowe niż składowych wyznaczanych z macierzy kowariancji. Jednakże niektórzy autorzy (Gnanadesikan, 1977; Mardia i in., 1979, Seber, 1984; Krzyśko, 2009) wskazują, że standaryzacja zmiennych polegająca na użyciu macierzy korelacji **R** w miejsce macierzy kowariancji **C** może powodować pewne komplikacje w interpretacji, wnioskowaniu oraz w określaniu rozkładu prawdopodobieństwa uzyskiwanych składowych głównych. Mardia i in. (1979) zaznaczają jednak, że powyższe komplikacje nie wpływają na jakość analizy czynnikowej metodą składowych głównych.

Celem analizy składowych głównych jest określenie kolejnych składowych głównych. Pierwsza składowa powstaje przez znalezienie takiego wektora $\mathbf{a}_1 = [a_{11}, a_{21}, \cdots, a_{p1}]$ o długości jednostkowej $\left(\sum_{i=1}^{p} a_{i1}^2 = 1\right)$, by kombinacja liniowa postaci:

$$\mathbf{Z}_1 = a_{11}\mathbf{X}_1 + a_{21}\mathbf{X}_2 + \cdots + a_{p1}\mathbf{X}_p \quad 1$$

miała największą wariancję wśród wszystkich takich kombinacji liniowych, gdzie $\mathbf{X}_i$ $(i = 1,2,\cdots,p)$ to wektory obserwowanych wartości zmiennych losowych.

Druga składowa główna powstaje przez znalezienie takiego wektora jednostkowego $\mathbf{a}_2 = [a_{12}, a_{22}, \cdots, a_{p2}]$, który jest ortogonalny do wektora pierwszej składowej $\mathbf{a}_1 = [a_{11}, a_{21}, \cdots, a_{p1}]$, to znaczy spełnia równość $\sum_{i=1}^{p} a_{i1} \cdot a_{i2} = 0$, a utworzona przez niego kombinacja

$$\mathbf{Z}_2 = a_{12}\mathbf{X}_1 + a_{22}\mathbf{X}_2 + \cdots + a_{p2}\mathbf{X}_p \quad 2$$

daje maksymalną wariancję wśród wszystkich takich kombinacji liniowych.

Warunek ortogonalności tych wektorów zapewnia nam niezależność ocen, czyli sumowanie się wariancji kolejnych składowych głównych do całkowitej wariancji układu. Ogólnie, kolejna składowa główna odpowiadająca kombinacji liniowej o największej wariancji wśród takich kombinacji jest ortogonalna z wcześniejszymi składowymi głównymi, tzn. pierwszą, drugą, itd.

Wektor $\mathbf{a}_j = [a_{1j}, a_{2j}, \cdots, a_{pj}]$ dla $j = 1,2,\cdots,p$ nazywamy *j*-tym wektorem ładunków lub współczynników *j*-tej składowej.

Matematycznie rozwiązanie niniejszego problemu uzyskujemy przez znalezienie wartości własnych i wektorów własnych macierzy korelacji. Wektor $\mathbf{a}_1 = [a_{11}, a_{21}, \cdots, a_{p1}]$, który daje maksymalną wariancję przy dodatkowym warunku $\mathbf{a'}_1\mathbf{a}_1 = \sum_{i=1}^{p} a_{i1}^2 = 1$ jest wektorem własnym odpowiadającym największej wartości własnej macierzy korelacji **R**. Wariancja nowej zmiennej

$$\mathbf{Z}_1 = a_{11}\mathbf{X}_1 + a_{21}\mathbf{X}_2 + \cdots + a_{p1}\mathbf{X}_p$$

jest zatem równa największej wartości własnej analizowanej macierzy.

Druga i kolejne składowe główne są kolejnymi wektorami własnymi analizowanej macierzy, podobnie jak kolejne wartości własne są wariancjami tych składowych.

ANALIZA CZYNNIKOWA

Podstawą zastosowania analizy czynnikowej jest przypuszczenie, że jeżeli mamy dużą liczbę powiązanych wewnętrznie wskaźników (cech), to związki między nimi mogą wynikać z istnienia jednego lub wielu czynników wspólnych, które są powiązane z poszczególnymi cechami analizowanego zespołu. Możemy stwierdzić, że u podstaw analizy czynnikowej leży założenie, że w zespole p cech $\{\mathbf{X}_i; i = 1,2,\cdots,p\}$ są ukryte czynniki, a w najprostszym przypadku jeden, będące źródłem wspólnej informacji tkwiącej w nich. Celem eksploracyjnej analizy czynnikowej jest wykrycie tych wspólnych czynników (nowego zbioru zmiennych), odpowiedzialnych za zachowanie się poszczególnych cech, czy też poszczególnych grup cech. Tak więc analiza czynnikowa

służy także do określania (poszukiwania) grup cech podobnie zachowujących się według ustalonych ocen związków między cechami, na przykład współczynników korelacji. Można założyć, że w poszukiwaniu wspólnych czynników najczęściej wykorzystujemy macierz korelacji (**R**) między poszczególnymi cechami analizowanego zespołu. Najbardziej upowszechnioną metodą wyznaczania czynników jest metoda składowych głównych, polegająca na przypisaniu czynnika $\mathbf{Z}_j$ wektorowi własnemu dla *j*-tej wartości własnej macierzy korelacji. Często udaje się „znalezionym" czynnikom nadawać sensowną interpretację merytoryczną. Należy jednak pamiętać o tym, że są to umowne wielkości pozwalające na opis matematyczny badanego zbioru danych. Należy nadmienić, że do najpopularniejszych metod wyodrębniania czynników głównych, poza analizą składowych głównych, należą między innymi metoda osi głównych oraz metoda największej wiarygodności.

Zakładamy więc, że między czynnikami $\mathbf{Z}_j$ $(j = 1,2,\cdots,q < p)$ i zmiennymi $\mathbf{X}_i$ zachodzą związki liniowe dla $i = 1,2,\cdots,p$:

$$\mathbf{X}_i = a_{i1}\mathbf{Z}_1 + a_{i2}\mathbf{Z}_2 + \cdots + a_{iq}\mathbf{Z}_q + b_i\mathbf{U}_i = \sum_{j=1}^{q} a_{ij}\mathbf{Z}_j + b_i\mathbf{U}_i\,, \qquad 3$$

a zapisane w notacji macierzowej:

$$\mathbf{X}_{p\times 1} = \mathbf{A}_{p\times q}\mathbf{Z}_{q\times 1} + \mathbf{B}_{p\times p}\mathbf{U}_{p\times 1}\,, \text{gdzie } \mathbf{B} = \mathrm{diag}\left(b_1, b_2, \cdots, b_p\right). \qquad 4$$

Współczynniki a_{ij} noszą nazwę ładunków czynnikowych czynnika $\mathbf{Z}_j$ (tzw. czynnik wspólny) na cechę $\mathbf{X}_i$ (nasilenie czynnika *j* w zmiennej *i*). Zmienne $\mathbf{U}_i$ są składnikami (czynnikami) specyficznymi w każdej zmiennej $\mathbf{X}_i$. Czynniki $\mathbf{Z}_j$ i $\mathbf{U}_i$, traktowane jako zmienne losowe, są wewnętrznie i między sobą nieskorelowane. Oznacza to, że $\mathrm{kor}\left(\mathbf{Z}_j; \mathbf{Z}_m\right) = 0$ i $\mathrm{kor}\left(\mathbf{U}_j; \mathbf{U}_m\right) = 0$ dla $j \neq m$ oraz $\mathrm{kor}\left(\mathbf{Z}_j; \mathbf{U}_m\right) = 0$ dla $j = 1,2,\cdots,q$ i $m = 1,2,\cdots,p$. Przy tych założeniach macierz korelacji **R** między p zmiennymi można przedstawić w postaci

$$\mathbf{R} = \mathbf{A}_{p\times q}\mathbf{A}^T_{q\times p} + \mathbf{B}^2_{p\times p}\,. \qquad 5$$

Jest to podstawowe równanie analizy czynnikowej.

Wielkość $h_i^2 = \sum_{j=1}^{q} a_{ij}^2$ nazywamy zasobem wspólnej zmienności cechy $\mathbf{X}_i$ determinowanej czynnikami $\mathbf{Z}_j$ (część wariancji zmiennej $\mathbf{X}_i$, odpowiadająca czynnikowi $\mathbf{Z}_j$). Wielkość $b_i^2 = 1 - h_i^2$, odpowiadająca czynnikowi specyficznemu $\mathbf{U}_i$, nazywamy wariancją specyficzną. Suma zasobów $h_i^2 = \sum_{j=1}^{q} a_{ij}^2$ daje łączną determinację zmienności wszystkich $\mathbf{X}_i$, $i = 1,2,\cdots,p$, przez czynniki $\mathbf{Z}_j$, $j = 1,2,\cdots,q$.

Ponieważ suma wariancji zmiennych $\mathbf{X}_i$ jest równa p (suma elementów głównej przekątnej macierzy korelacji), więc współczynnik: jest zespołowym współczynnikiem determinacji.

$$R^2_{\mathbf{X}\cdot\mathbf{Z}} = \frac{1}{p}\sum_{i=1}^{p} h_i^2 = \frac{1}{p}\sum_{i=1}^{p}\sum_{j=1}^{q} a_{ij}^2 \,, \tag{6}$$

Suma kwadratów ładunków może być rozdzielona na części i przypisana poszczególnym uzyskanym czynnikom $\mathbf{Z}_j$, tzn.

$$\lambda_j = \sum_{i=1}^{p} a_{ij}^2 \;\; (j = 1,2,\cdots,q) \,, \tag{7}$$

Wielkość ta określa wagę j-tego czynnika wspólnego w determinacji zmienności zbioru $\{\mathbf{X}_i\}$.

ORTOGONALNE ROTACJE CZYNNIKÓW

Niech $\mathbf{R}$ oznacza macierz korelacji zmiennych $\{\mathbf{X}_i\}$. Oznaczając przez $\mathbf{A}$ macierz ładunków o elementach a_{ij} oraz p wierszach i q kolumnach, możemy zapisać macierz korelacji (analogicznie jak w (5)):

$$\mathbf{R} = \mathbf{A}\cdot\mathbf{A}^{\mathrm{T}} + \mathbf{B}^2 \,, \tag{8}$$

gdzie: $\mathbf{B}^2 = \mathrm{diag}\left(b_1^2, b_2^2, \cdots, b_p^2\right)$.

Otóż, jeżeli macierz $\mathbf{D}$ jest macierzą ortogonalną (iloczyn $\mathbf{D}\cdot\mathbf{D}^{\mathrm{T}}$ jest macierzą jednostkową), to transformacja czynników $\mathbf{Z}^{\mathrm{T}} = \mathbf{Z}\cdot\mathbf{D}$ nie zmienia struktury macierzy korelacji $\mathbf{R}$, ponieważ:

$$(\mathbf{A}\cdot\mathbf{D})(\mathbf{A}\cdot\mathbf{D})^{\mathrm{T}} = \mathbf{A}\cdot\mathbf{D}\cdot\mathbf{D}^{\mathrm{T}}\cdot\mathbf{A}^{\mathrm{T}} = \mathbf{A}\cdot\mathbf{A}^{\mathrm{T}} \,. \tag{9}$$

Transformacji, danej macierzą $\mathbf{D}$, geometrycznie odpowiada obrót kierunków głównych określających składowe główne. Można dokonać obrotu tak, aby ładunki przy cechach maksymalnie się różnicowały, przez co otrzymuje się ich prostszą interpretację.

Stąd warunek by *warians* kwadratów ładunków był maksymalny:

$$\mathrm{var}a = \sum_{j=1}^{q}\left[\frac{1}{p}\sum_{i=1}^{p}\left(a_{ij}^2 - \bar{a}_j^2\right)^2\right] = \max! \,, \tag{10}$$

gdzie: $\bar{a}_j^2 = \frac{1}{p}\sum_{i=1}^{p} a_{ij}^2$, prowadzi do metody *varimax* (maximum of the variance), sformułowanej przez Kaisera (1958) i dającej maksymalne zróżnicowanie ładunków w ramach czynnika.

Metoda varimax skupia się na prostej interpretacji kolumn macierzy czynników, natomiast metoda, która skupia się na prostej interpretacji wierszy tejże macierzy znana jest pod nazwą metody quartimax.

PRZYKŁAD 1

W swojej pracy o liniowych funkcjach dyskryminacyjnych Fisher (1936), jako przykład zastosowania uzyskanych wyników, badał trzy populacje (gatunki) kwiatów irysa — *setosa*, *versicolor* i *virginica*, każda po 50 obserwacji (kwiatów).

Dla trzech opisanych w doświadczeniu gatunków irysa wykonano odrębne analizy EFA metodą składowych głównych z rotacją varimax.

W tabeli 1 przedstawiono wartości własne wyznaczonych czynników głównych oraz ich udział w objaśnieniu obserwowanej zmienności. Dla *I. steosa*: pierwszy czynnik główny objaśniał ponad 51% obserwowanej zmienności, czynnik drugi — 25,6%, trzeci —16,7%, a czwarty — 6,3%. Dla *I. versicolor*: pierwszy czynnik główny objaśniał 73,2% obserwowanej zmienności, czynnik drugi — 13,7%, trzeci — 9,9%, a czwarty — 3,3%. Dla *I. virginica*: pierwszy czynnik główny objaśniał 61,4% obserwowanej zmienności, czynnik drugi — 24,2%, trzeci —11,3%, a czwarty — 3,2%.

Tabela 1

Wartości własne czynników wspólnych i ich udział w objaśnianiu obserwowanej zmienności
Common factors eigenvalues and their participation in explanation of observed variability

Nr czynnika Factor number	*I. steosa*			*I. versicolor*			*I. virginica*		
	wartość własna eigenvalue	% objaśnianej zmienności % of explained variation	skumulowany % objaśnianej zmienności cumulated % of explained variation	wartość własna eigenvalue	% objaśnianej zmienności % of explained variation	skumulowany % objaśnianej zmienności cumulated % of explained variation	wartość własna eigenvalue	% objaśnianej zmienności % of explained variation	skumulowany % objaśnianej zmienności cumulated % of explained variation
1	2,0585	0,5146	0,5146	2,9263	0,7316	0,7316	2,4547	0,6137	0,6137
2	1,0222	0,2555	0,7702	0,5463	0,1366	0,8682	0,9647	0,2412	0,8549
3	0,6678	0,1670	0,9371	0,3950	0,0987	0,9669	0,4523	0,1131	0,9679
4	0,2515	0,0629	1,0000	0,1324	0,0331	1,0000	0,1283	0,0321	1,0000

Na rysunku 1 przedstawiono wykresy osypiska pozwalające na ocenę liczby właściwych, niosących największy ładunek informacyjny, czynników latentnych wystarczających do opisu obserwowanej zmienności cech rzeczywistych. W dalszej analizie skoncentrowano się na dwóch pierwszych czynnikach głównych, które łącznie odpowiadały za 77% zmienności *I. steosa*, 86,8% zmienności *I. versicolor* oraz 85,5% zmienności *I. virginica*.

W tabeli 2 przedstawiono wartości ładunków czynnikowych dla dwóch pierwszych czynników głównych trzech badanych gatunków irysa przed i po rotacji. Jak widać rotacja pozwoliła na wyraźniejsze zróżnicowanie struktury cech budujących wyznaczone czynniki główne. Przyjmując za wartość graniczą 0,5 można stwierdzić, iż dla gatunku *I. setosa* pierwszy czynnik główny budowały przede wszystkim długość działki kielicha i szerokość działki kielicha; drugi czynnik główny — długość płatka i szerokość płatka.

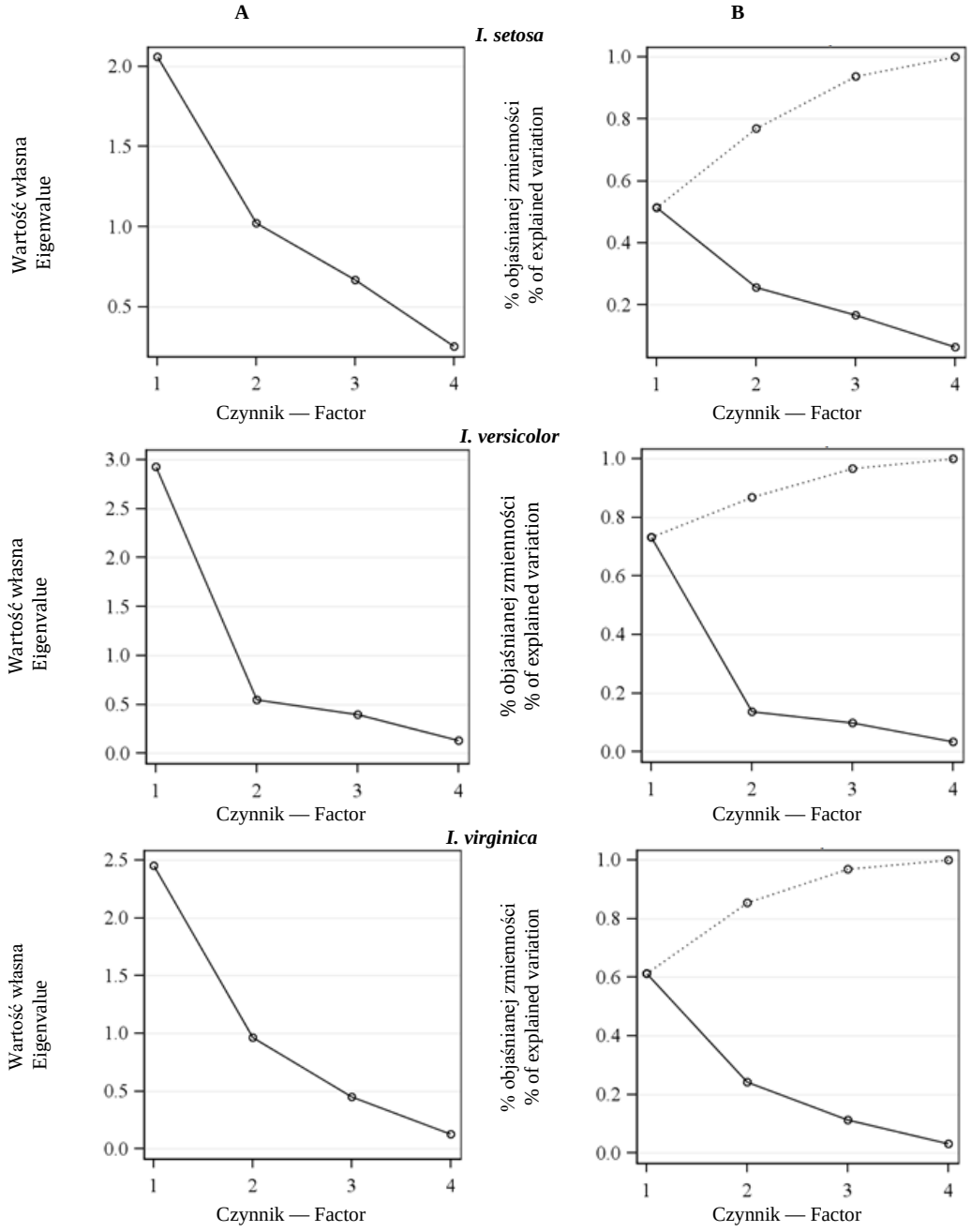
Tabela 2

Ładunki czynnikowe dla badanych gatunków irysa przed i po rotacji
Before and after rotation common factor loadings for studied species of iris

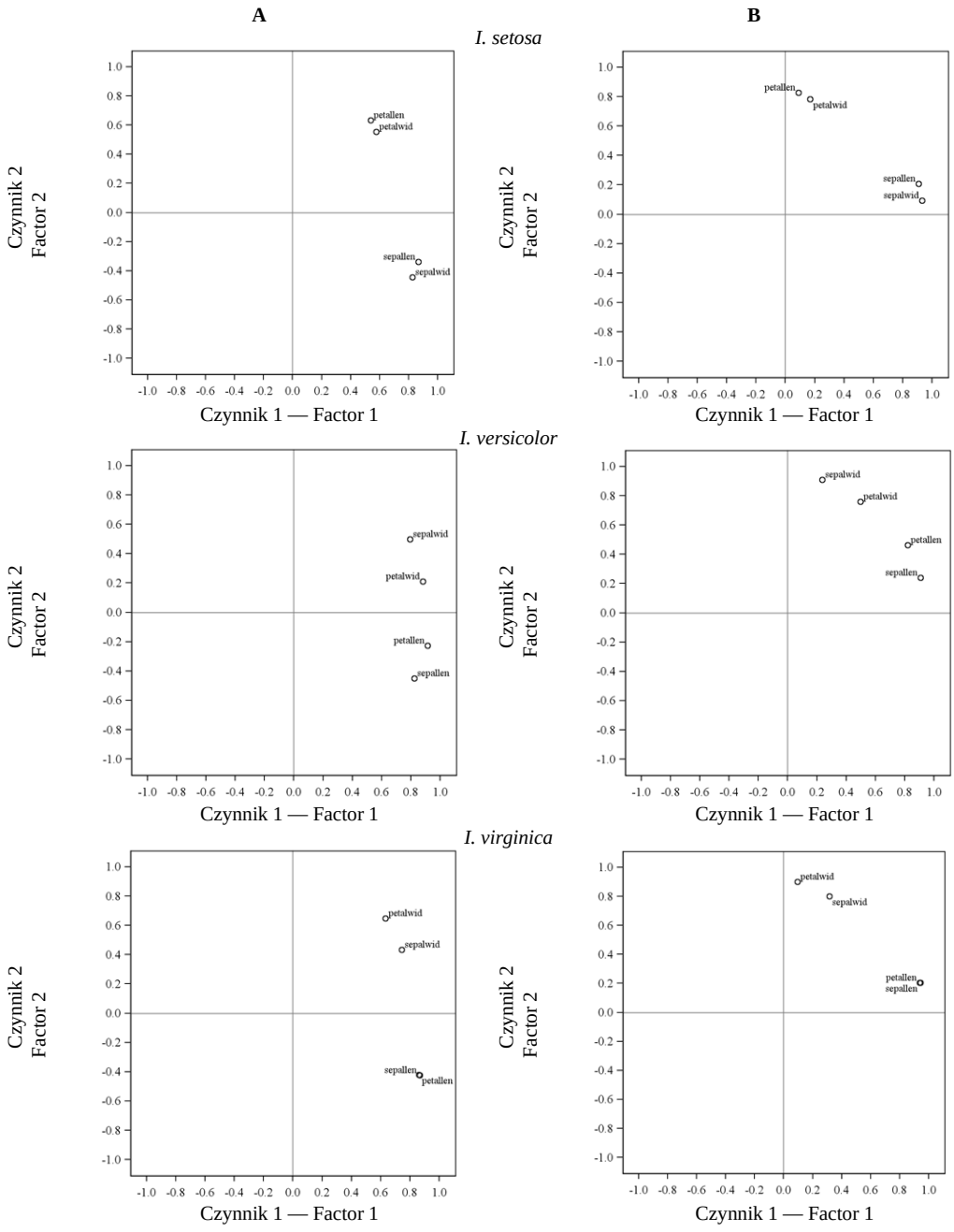
Gatunek Species	Cechy Feature	Ładunki czynnikowe przed rotacją Factor loadings before rotation		Ładunki czynnikowe po rotacji Factor loadings after rotation	
		Czynnik 1 Factor 1	Czynnik 2 Factor 2	Czynnik 1 Factor 1	Czynnik 2 Factor 2
I. steosa	Długość działki kielicha Sepal length	0,86719	–0,33869	**0,90797**	0,20572
	Szerokość działki kielicha Sepal width	0,82588	–0,44571	**0,93376**	0,09393
	Długość płatka Petal length	0,53866	0,63389	0,09067	**0,82689**
	Szerokośc płatka Petal width	0,57818	0,55408	0,16814	**0,78296**
I. versicolor	Długość działki kielicha Sepal length	0,82510	–0,45144	**0,90965**	0,23899
	Szerokość działki kielicha Sepal width	0,79519	0,49726	0,23602	**0,90768**
	Długość płatka Petal length	0,91437	–0,22679	**0,82014**	0,46354
	Szerokośc płatka Petal width	0,88156	0,20922	0,49670	**0,75777**
I. virginica	Długość działki kielicha Sepal length	0,86873	–0,42474	**0,94498**	0,20521
	Szerokość działki kielicha Sepal width	0,74504	0,43234	0,31664	**0,80109**
	Długość płatka Petal length	0,86189	–0,42202	**0,93793**	0,20310
	Szerokośc płatka Petal width	0,63411	0,64753	0,09621	**0,90118**

Dla gatunków *I. versicolor* oraz *I. virginica*: pierwszy czynnik główny budowały przede wszystkim długość działki kielicha i długość płatka; a drugi czynnik główny — szerokość działki kielicha i szerokość płatka. Pomimo podobnej konstrukcji wyznaczonych czynników głównych te dwa ostatnie gatunki różnią się strukturą szczegółową (ładunkami) dla omawianych czynników głównych. Różnice to potwierdza również wykres rozmieszczenia cech w układzie dwóch pierwszych czynników głównych (rys. 2). Na rysunku tym można również zauważyć, jakie zmiany w układzie analizowanych cech przyniosła przeprowadzona rotacja.

Podsumowując ten przykład obliczeniowy, można stwierdzić, że analiza czynnikowa pozwoliła na wskazanie różnic w strukturze wielocechowej charakteryzującej trzy badane w pracy Fishera (1936) gatunki irysa. Jest to przesłanką do stwierdzenia, że możliwe jest przeprowadzenie stosownej analizy dyskryminacyjnej pozwalającej z minimalnym błędem na rozdzielenie tych trzech gatunków na podstawi informacji o działkach kielicha i płatkach kwiatów irysa.



Rys. 1. Wykresy osypiska (A) oraz wykresy stopnia objaśnianej zmienności przez wyznaczone czynniki główne (B) dla trzech gatunków irysa uzyskane w analizie EFA
Fig. 1. Obtained in the EFA analysis scree plots (A) and plots of the degree of variability explained by appointed common factors (B) for three tested iris species



Rys. 2. Wykresy rozmieszczenia analizowanych cech w układzie dwóch pierwszych czynników głównych przed (A) i po (B) rotacji

Fig. 2. Plots of distribution of analyzed features in the layout of the first two common factors before (A) and after (B) rotation

PRZYKŁAD 2

W Pracowni Ekonomiki Nasiennictwa i Hodowli Roślin Instytutu Hodowli i Aklimatyzacji Roślin — Państwowego Instytutu Badawczego w latach 1986–2003 prowadzono, obejmujące teren całego kraju, badania ankietowe gospodarstw indywidualnych. Zbierano informacje o uprawianych odmianach dziesięciu głównych ziemiopłodów oraz o rodzaju materiału siewnego, charakterystykę pól uprawnych, informacje o stosowanych czynnikach produkcji i agrotechnice. W ramach badań ankietowych, prowadzonych w ciągu 18 lat, zebrano informacje o około 8 500 polach na których uprawiano w tym czasie pszenicę ozimą. Z tej bazy danych odrzucono rekordy dotyczące najmniejszych gospodarstw i najmniejszych pól uprawnych, rekordy zawierające braki danych oraz rekordy dotyczące marginalnie uprawianych odmian pszenicy ozimej, pozostawiając do dalszych analiz około 4 000 rekordów charakteryzujących pola uprawne gospodarstw produkcyjnych. Spośród badanych w ankietach cech do dalszych analiz wybrano cechy będące znaczącymi czynnikami warunkującymi uzyskiwane plony (Laudański i in., 2007 a, 2007 b). Były to: liczba zabiegów środkami ochrony roślin, jakość przedplonu wyrażona w punktach, rodzaj stosowanego materiału siewnego, odczyn gleby, odmiana, jakość gleby wyrażona w punktach waloryzacji rolniczej przestrzeni produkcyjnej, liczba lat od ostatniego użycia obornika, ilość wysiewu, dawka nawożenia NPK w czystych składnikach oraz termin siewu. Wymienione czynniki agrotechniczne miały charakter cech jakościowych oraz ilościowych. Część z nich, (np. odmiany) przyjmowała wartości (poziomy czynnika) o konkretnych nazwach. Cechy te, określane jako nominalne, są ewidentnie jakościowe w tym sensie, że nie mają żadnych związków z cechami ilościowymi. Natomiast czynniki ilościowe, (np. nawożenie mineralne) reprezentowane przez wartości (kategorie) uporządkowane, mają w większym lub mniejszym stopniu związek z cechami ilościowymi. Czynniki te w analizach potraktowano, jako ewidentnie jakościowe poprzez dyskretyzację czynnika ilościowego — wartość cechy stała się kolejnym poziomem czynnika. W ten sposób wszystkie warianty czynników można ocenić odpowiadającą wielkością średniego plonu pszenicy, uzyskanego dla poszczególnych poziomów na podstawie nieobciążonych, ważonych estymatorów efektów czynnika z modelu obserwacji postaci

$$y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \varepsilon_{ijk} ,$$

gdzie drugi czynnik stanowiły lata badań. To przekształcenie pozwoliło na przyporządkowanie wartości $\mu + \alpha_i$ odpowiednim kategoriom analizowanych czynników.

Tak przygotowane dane poddano eksploracyjnej analizie czynnikowej, w celu wyodrębnienia głównych grup czynników produkcji — makroczynników determinujących uzyskiwane efekty w uprawie pszenicy ozimej w postaci plonu.

W tabeli 3 przedstawiono współczynniki korelacji liniowej Pearsona pomiędzy analizowanymi cechami i plonem pszenicy ozimej. Jak widać wyznaczone współczynniki korelacji nie są duże. Najwyższą wartość współczynnika korelacji odnotowano dla relacji nawożenie NPK — plon ($r = 0{,}541$).

Tabela 3

Współczynniki korelacji liniowej Pearsona pomiędzy plonem i analizowanymi czynnikami produkcji
Pearson linear correlation coefficient between yield and analyzed production factors

	Plon Yield	Liczba zabiegów pesty-cydami Number of pesticide usage	Wartość przed-plonu Forecrop value	Materiał siewny Seed material	Odczyn gleby Soil pH	Odmia-na Cultivar	Jakość gleby Soil quality	Liczba lat od użycia obornika Number of years since the use of manure	Ilość wysiewu Amount of seeding	Nawo-żenie NPK NPK fertali-zation	Termin siewu Date of sowing
Plon Yield	1	0,374	0,287	0,280	0,184	0,291	0,192	0,167	0,264	0,541	0,217
Liczba zabiegów pestycydami Number of pesticide usage	0,374	1	0,286	0,304	0,072	0,327	0,033	0,208	0,127	0,375	0,149
Wartość przedplonu Forecrop value	0,287	0,286	1	0,194	0,058	0,188	0,084	0,231	0,090	0,260	0,134
Materiał siewny Seed material	0,280	0,304	0,194	1	0,087	0,337	0,049	0,163	0,090	0,265	0,095
Odczyn gleby Soil pH	0,184	0,072	0,058	0,087	1	0,098	0,015	0,028	0,054	0,110	0,100
Odmiana Cultivar	0,291	0,327	0,188	0,337	0,098	1	0,071	0,106	0,081	0,284	0,109
Jakość gleby Soil quality	0,192	0,033	0,084	0,049	0,015	0,071	1	0,051	0,081	0,057	0,023
Liczba lat od użycia obornika Number of years since the use of manure	0,167	0,208	0,231	0,163	0,028	0,106	0,051	1	0,076	0,175	0,114
Ilość wysiewu Amount of seeding	0,264	0,127	0,090	0,090	0,054	0,081	0,081	0,076	1	0,191	0,129
Nawożenie NPK NPK fertilization	0,541	0,375	0,260	0,265	0,110	0,284	0,057	0,175	0,191	1	0,168
Termin siewu Date of sowing	0,217	0,149	0,134	0,095	0,100	0,109	0,023	0,114	0,129	0,168	1

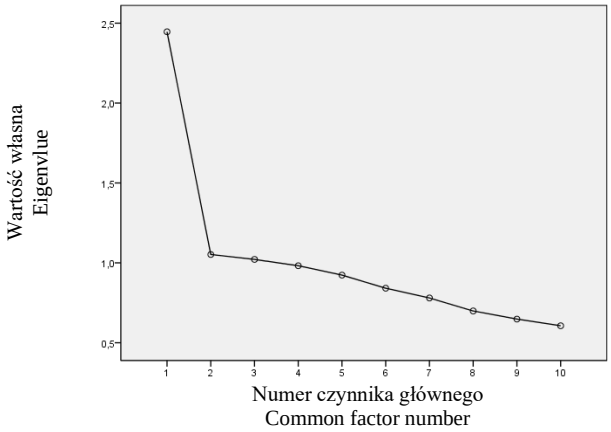
W tabeli 4 przedstawiono wartości własne wyznaczonych czynników głównych oraz ich udział w objaśnieniu obserwowanej zmienności. Wyodrębnienie istotnej liczby czynników głównych odbyło się na podstawie tzw. kryterium wartości własnej, które mówi, iż tylko te czynniki są znaczące, dla których wartości własne są powyżej 1 (kryterium to można stosować tylko i wyłączeni, gdy analizę opieramy na macierzy korelacji). Na podstawie tego kryterium, zwanego również kryterium Kaisera, wskazano cztery znaczące czynniki główne. Odpowiadały one łącznie za ponad 55% obserwowanej zmienności, w tym pierwszy czynnik, po rotacji — 19,5%, drugi — 12,7%, trzeci — 12,6%, a czwarty — 10,3% zmienności obserwowanej w źródłowym zbiorze danych. W praktyce, często można spotkać jeszcze dwa popularne kryteria wyodrębniania znaczącej liczby czynników głównych. Pierwsze to kryterium osypiska — polega na sporządzeniu wykresu osypiska (rys. 3), a następnie z tego wykresu odczytuje się liczbę czynników dla

której krzywa przestaje cechować się gwałtownym spadkiem. Drugie kryterium to tak zwane kryterium stopnia objaśnianej zmienności. Polega ono na tym, że z góry ustala się stopień zmienności jaki ma być objaśniany przez czynniki główne, a następnie wybiera się taką liczbę czynników, by skumulowany procent objaśnianie zmienności nie był mniejszy od założonego poziomu.

Tabela 4

Wartości własne czynników wspólnych i ich udział w objaśnianiu obserwowanej zmienności
Common factors eigenvalues and their participation in explanation of observed variability

Numer czynnika Factor number	Początkowe wartości własne Initial eigenvalues			Wartości własne po wyodrębnieniu Extracted eigenvalues			Wartości własne po rotacji Eigenvalues after rotation		
	ogółem total	% wariancji % of variation	% skumulowany cumulated %	ogółem total	% wariancji % of variation	% skumulowany cumulated %	ogółem total	% wariancji % of variation	% skumulowany cumulated %
1	2,446	24,461	24,461	2,446	24,461	24,461	1,945	19,449	19,449
2	1,052	10,523	34,984	1,052	10,523	34,984	1,272	12,717	32,166
3	1,022	10,217	45,200	1,022	10,217	45,200	1,255	12,552	44,718
4	0,982	9,820	55,021	0,982	9,820	55,021	1,030	10,303	55,021
5	0,923	9,233	64,254						
6	0,841	8,413	72,667						
7	0,780	7,801	80,468						
8	0,699	6,988	87,456						
9	0,648	6,481	93,937						
10	0,606	6,063	100,000						



Rys. 3. Wykresy osypiska dla analizowanych czynników produkcji
Fig. 3. Scree plot for analyzed production factors

Dla wyodrębnionych czterech czynników głównych wyznaczono wartości ładunków czynnikowych (tab. 5). Stosując, podobnie jak to miało miejsce w pierwszym omawianym przykładzie, graniczną wartość 0,5, można zaobserwować, że pierwszy czynnik główny budowały przede wszystkim: odmiana, rodzaj materiału siewnego, liczba zabiegów ochrony roślin pestycydami oraz nawożenie NPK. Wszystkie ładunki czynnikowe dla tych

cech były dodatnie, co można interpretować jako wprost proporcjonalny wpływ tych cech na kształtowanie się wartości czynnika głównego. Z racji, że wszystkie wymienione cechy powiązane są z nakładami finansowymi, które bezpośredni rolnik musi ponieść (np. dobór i zakup nasion konkretnej odmiany, wybór rodzaju materiału siewnego, zakup i użycie pestycydów, zakup i wysiew nawozów mineralnych) pierwszy czynnik główny możemy określić jako **czynnik nakładowy**.

Tabela 5

Ładunki czynnikowe po rotacji dla analizowanych czynników produkcji
After rotation common factor loadings for analyzed production factors

Czynniki produkcji Production factors	Czynnik główny Common factor			
	1	2	3	4
Odmiana Cultivar	**0,755**	-0,036	-0,005	0,068
Materiał siewny Seed material	**0,707**	0,074	-0,026	0,020
Liczba zabiegów pestycydami Number of pesticide usage	**0,623**	0,333	0,121	-0,043
Nawożenie NPK NPK fertilization	**0,542**	0,250	0,300	0,047
Liczba lat od użycia obornika Number of years since the use of manure	0,107	**0,714**	0,112	-0,024
Wartość przedplonu Forecrop value	0,321	**0,554**	0,128	0,072
Termin siewu Date of sowing	0,015	0,231	**0,686**	-0,115
Ilość wysiewu Amount of seeding	0,012	0,118	**0,575**	0,359
Odczyn gleby Soil pH	0,273	-0,456	**0,565**	-0,100
Jakość gleby Soil quality	0,066	0,009	-0,003	**0,929**

Drugi czynnik główny budowały: liczba lat do ostatniego użycia obornika oraz wartość przedplonu. Cechy te wpływały wprost proporcjonalnie na kształtowanie się drugiego czynnika głównego. Obydwie cechy związane były z wydarzeniami ‘historycznymi’ dotyczącymi pola uprawnego, dlatego też czynnik ten można określić jako **historię pola**.

Trzeci czynnik główny budowany był przede wszystkim przez: termin siewu, ilość wysiewu oraz odczyn gleby. Wszystkie te cechy wykazywały wpływ wprost proporcjonalny na kształtowanie się trzeciego czynnika głównego. Cechy te charakteryzowały siew pszenicy ozimej oraz warunki pola związane z jego właściwym utrzymaniem (wapnowanie), dlatego też trzeci czynnik główny możemy określić jako **przygotowanie pola i siew**.

Ostatni, czwarty czynnik główny budowany był przede wszystkim przez jedną cechę — jakość gleby. Czynnik ten przyjmował tym większe wartości im wyższa była jakość gleby, dlatego można go określić mianem **czynnika siedliskowego**.

Podsumowując, w powyższym przykładzie przedstawiono szeroką interpretację uzyskanych wyników danych włącznie z możliwością merytorycznej interpretacji wyznaczonych czynników głównych.

PODSUMOWANIE

Jednym z często poruszanych i ważnych zagadnień jest możliwie obiektywna ocena informacji jaka tkwi w określonym zbiorze wielkości (danych) opisujących interesujące nas zjawisko lub proces. Pojedyncze cechy na ogół nie oddają adekwatnego odbicia badanego zjawiska lub procesu wielowymiarowego. Każda z tych cech tylko w pewnym stopniu je odzwierciedla. Dlatego też omawiane w niniejszej pracy metody obliczeniowe statystyki matematycznej dają istotne możliwości analizowania struktur tkwiących w zespole wielocechowych danych przyrodniczych. Metody te pozwalają przede wszystkim na przybliżenie wielocechowych zjawisk poprzez sprowadzenie ich do prostszego opisu i interpretacji we wzajemnym porównywaniu populacji wybranych roślin. Przykład 1 wyraźnie artykułuje zróżnicowanie analizowanych wielocechowych obiektów w pierwszej fazie na dwie odrębne grupy, by w następnej fazie uzyskać potwierdzenie zróżnicowania kolejnych dwóch grup obiektów.

Inny obszar zastosowań niniejszych metod to wykorzystanie możliwości grupowania poszczególnych cech w zespoły pozwalające na ich praktyczną interpretację (przykład 2). Ma to szczególne znaczenie przy analizowaniu zbioru cech egzogenicznych (tj. cech o charakterze zmiennych niezależnych) w analizach zjawisk przyczynowo-skutkowych. Często pojedyncze cechy przyczynowe nie wskazują na ich istotny cząstkowy wpływ na badany skutek, ale w zespole z innymi cechami może okazać on się bardzo wyraźny i znaczący.

LITERATURA

Armitage P., Colton T. (eds.) 2005. Encyclopedia of Biostatistics. 2nd Edition. John Wiley & Sons Inc., Hoboken: 2055 — 2067, : 4588 — 4592.

Caliński T., Czajka S., Kaczmarek Z. 1975. Analiza składowych głównych i jej zastosowania. Algorytmy biometryczne i statystyczne (ABS-36). AR Poznań.

Cureton E. E., Mulaik S. A. 1975. The weighted varimax rotation and the promax rotation. Psychometrika 40: 183 — 195.

Everitt B. S., Dunn G. 1992. Applied Multivariate Data Analysis. Oxford University Press, New York.

Fisher R. A. 1936. The Use of Multiple Measurements in Taxonomic Problems. Annals of Eugenics, 7: 179 — 188.

Gnanadesikan R. 1977. Methods for statistical data analysis of multivariate observations. Wiley, New York.

Gower J. C. 1966. Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika, 53 (3/4): 325 — 338.

Hatcher L. 1994. A Step-by-step Approach to Using SAS for Factor Analysis and Structural Equation Modeling. SAS Publishing, SAS Institute Inc., Cary, NC, USA.

Kaiser, H. F. 1958. The varimax criterion for analytic rotation in factor analysis. Psychometrika 23: 187 — 200.

Khattre R., Naik D. N. 2000. Multivariate Data Reduction and Discrimination with SAS Software. SAS Publishing, SAS Institute Inc., John Wiley & Sons Inc., New York, USA.

Khattree R., Naik D. N. 1999. Applied Multivariate Statistics with SAS Software. Second Edition. SAS Publishing, SAS Institute Inc., John Wiley & Sons Inc., New York, USA.
Kim J. O., Mueller C. W. 1978. Factor analysis. Statistical Methods and Practical Issues. Sage Publishing, Beverly Hills.
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.
Krzyśko M. 2009. Podstawy wielowymiarowego wnioskowania statystycznego. Wydawnictwo Naukowe UAM, Poznań.
Laudański Z., Mańkowski D. R., Sieczko L. 2007 a. Próba oceny technologii uprawy pszenicy ozimej na podstawie danych ankietowych gospodarstw indywidualnych. Część 1. Metoda wyodrębniania technologii uprawy. Biul. IHAR 244: 33 — 43.
Laudański Z., Mańkowski D. R., Sieczko L. 2007 b. Próba oceny technologii uprawy pszenicy ozimej na podstawie danych ankietowych gospodarstw indywidualnych. Część 2. Ocena technologii uprawy. Biul. IHAR 244: 45 — 57.
Mardia K. V., Kent J. T., Bibby J. M. 1979. Multivariate analysis. Academic Press, London.
Mądry W. 2007. Metody statystyczne do oceny różnorodności fenotypowej dla cech ilościowych w kolekcjach roślinnych zasobów genowych. Zeszyty Probl. Post.-Nauk Rol., Nr 517: 21 — 41.
Mądry W., Pluta S., Sieczko L., Studnicki M. 2010. Phenotypic diversity in a sample of blackcurrant (*Ribes nigrum* L.) cultivars maintained in the fruit breeding department at the Research Institute of Pomology and Floriculture in Skierniewice, Poland. Journal of Fruit and Ornamental Plant Research, 18 (2): 23 — 37.
Mańkowski D. R., Laudański Z., Martyniak D., Flaszka M. 2009. Struktura wielocechowej zmienności odmianowej wiechliny łąkowej (*Poa pratensis* L.). Biul. IHAR 254: 189 — 200.
Morrison D. F., 1990. Wielowymiarowa analiza statystyczna. PWN, Warszawa.
O'Rourke N., Hatcher L., Stepanski E. J. 2005. A step-by-step approach to using SAS for univariate & multivariate statistics. Second edition. SAS Publishing, SAS Institute Inc., Cary, NC, USA.
Rao C. R. 1964. The use and interpretation of principal component analysis in applied research. Sankhyã, A26: 329 — 358.
SAS Institute Inc. 2009. SAS/STAT 9.2 User's Guide, Second Edition. SAS Publishing, SAS Institute Inc., Cary, NC, USA.
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.
Sieczko L., Masny A., Mądry W., Żurawicz E. 2008. Analiza podobieństwa rodzin mieszańcowych truskawki powtarzającej owocowanie pod względem wielkości i jakości plonów owoców. Biul. IHAR 250: 297 — 307.
Szczotka F. A. 1977. Podstawy analizy czynnikowej. Listy Biometryczne, Nr 55–59: 1 — 69.
Thurstone L. L. 1931. Multiple factor analysis. Psychological Review, 38: 406 — 427.
Thurstone L. L. 1947. Multiple factor analysis. Chicago, USA: The University of Chicago Press.
Timm N. H. 2002. Applied multivariate analysis. Springer Verlag Inc., New York, USA.
Ukalska J., Mądry W., Ukalski K., Masny A. 2007. Wielowymiarowa ocena różnorodności fenotypowej w kolekcji zasobów genowych truskawki. Cz. II. Grupowanie genotypów. Zeszyty Probl. Post. Nauk Rol. 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.
Walesiak M., Gatnar E. 2009. Statystyczna analiza danych z wykorzystaniem programu R. Wydawnictwa Naukowe PWN, Warszawa.
Wójcik A. R., Laudański Z. 1989. Planowanie i wnioskowanie statystyczne w doświadczalnictwie. PWN, Warszawa.