

PRACE NAUKOWE

Uniwersytetu Ekonomicznego we Wrocławiu

RESEARCH PAPERS

of Wrocław University of Economics

Nr 328

Taksonomia 23

**Klasyfikacja i analiza danych –
teoria i zastosowania**

Redaktorzy naukowci

Krzysztof Jajuga, Marek Walesiak



Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2014

Redaktor Wydawnictwa: Barbara Majewska

Redaktor techniczny: Barbara Łopusiewicz

Korektor: Barbara Cibis

Łamanie: Beata Mazur

Projekt okładki: Beata Dębska

Publikacja jest dostępna w Internecie na stronach:

www.ibuk.pl, www.ebscohost.com,

w Dolnośląskiej Bibliotece Cyfrowej www.dbc.wroc.pl,

The Central and Eastern European Online Library www.ceeol.com,

a także w adnotowanej bibliografii zagadnień ekonomicznych BazEkon

http://kangur.uek.krakow.pl/bazy_ae/bazekon/nowy/index.php

Informacje o naborze artykułów i zasadach recenzowania znajdują się
na stronie internetowej Wydawnictwa

www.wydawnictwo.ue.wroc.pl

Tytuł dofinansowany ze środków Narodowego Banku Polskiego
oraz ze środków Sekcji Klasyfikacji i Analizy Danych PTS

Kopiowanie i powielanie w jakiegokolwiek formie
wymaga pisemnej zgody Wydawcy

© Copyright by Uniwersytet Ekonomiczny we Wrocławiu
Wrocław 2014

ISSN 1899-3192 (Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu)

ISSN 1505-9332 (Taksonomia)

Wersja pierwotna: publikacja drukowana

Druk: Drukarnia TOTEM

Spis treści

Wstęp	11
Małgorzata Rószkiewicz , Wykorzystanie metaanalizy w budowaniu modelu pomiarowego w przypadku braku niezmienniczości zasad pomiaru na przykładzie pomiaru zadowolenia z życia.....	13
Elżbieta Sobczak , Harmonijność inteligentnego rozwoju regionów Unii Europejskiej	21
Ewa Roszkowska, Renata Karwowska , Analiza porównawcza województw Polski ze względu na poziom zrównoważonego rozwoju w roku 2010.....	30
Tadeusz Kufel, Magdalena Osińska, Marcin Błażejowski, Paweł Kufel , Analiza porównawcza wybranych filtrów w analizie synchronizacji cyklu koniunkturalnego.....	41
Marcin Salamaga , Próba konstrukcji tablic „wymierania scenicznego” spektakli operowych na przykładzie Metropolitan Opera.....	51
Iwona Foryś , Wykorzystanie analizy dyskryminacyjnej do typowania rynków podobnych w procesie wyceny nieruchomości niemieszkalnych	59
Jerzy Korzeniewski , Selekcja zmiennych w klasyfikacji – propozycja algorytmu	69
Sabina Denkowska , Testowanie wielokrotne przy weryfikacji wieloczynnikowych modeli proporcjonalnego hazardu Coxa.....	76
Ewa Chodakowska , Teoria równań strukturalnych w klasyfikacji zmiennych jawnych i ukrytych według charakteru ich wzajemnych oddziaływań	85
Iwona Konarzewska , Model PCA dla rynku akcji – studium przypadku	94
Katarzyna Wójcik, Janusz Tuchowski , Dobór optymalnego zestawu słów istotnych w opiniach konsumentów na potrzeby ich automatycznej analizy	106
Aleksandra Łuczak , Zastosowanie metody AHP-LP do oceny ważności determinant rozwoju społeczno-gospodarczego w jednostkach administracyjnych	116
Aleksandra Witkowska, Marek Witkowski , Klasyfikacja pozycyjna banków spółdzielczych według stanu ich kondycji finansowej w ujęciu dynamicznym	126
Adam Depta , Zastosowanie analizy korespondencji do oceny jakości życia ludności na podstawie kwestionariusza SF-36v2	135
Marek Lubicz, Maciej Zięba, Konrad Pawelczyk, Adam Rzechonek, Marek Marciniak, Jerzy Kołodziej , Indukcja reguł dla danych niekompletnych i niezbalansowanych: modele klasyfikatorów i próba ich zastosowania do predykcji ryzyka operacyjnego w torakochirurgii	146

Małgorzata Misztal , Wybrane metody oceny jakości klasyfikatorów – przegląd i przykłady zastosowań.....	156
Anna M. Olszewska , Wykorzystanie wybranych metod taksonomicznych do oceny potencjału innowacyjnego województw	167
Iwona Bąk , Porównanie jakości grupowań powiatów województwa zachodniopomorskiego pod względem atrakcyjności turystycznej.....	177
Agnieszka Kozera, Joanna Stanisławska, Romana Głowicka-Wołoszyn , Segmentacja gospodarstw domowych według wydatków na turystykę zorganizowaną.....	186
Agnieszka Wałęga , Podejście syntetyczne w analizie spójności ekonomicznej gospodarstw domowych.....	196
Joanna Banaś, Małgorzata Machowska-Szewczyk, Bożena Mroczek , Zastosowanie analizy korespondencji do badania wpływu elektrowni wiatrowych na jakość życia ludności	205
Joanna Banaś, Krzysztof Małecki , Klasyfikacja punktów pomiarów ankietowych kierowców na granicy Szczecina z wykorzystaniem zmiennych symbolicznych.....	214
Aneta Becker , Wykorzystanie informacji granularnej w analizie wymagań rynku pracy.....	222
Katarzyna Cheba, Joanna Holub-Iwan , Wykorzystanie analizy korespondencji w segmentacji rynku usług medycznych.....	230
Adam Depta, Iwona Staniec , Identyfikacja czynników decydujących o jakości życia studentów łódzkich uczelni.....	238
Katarzyna Dębowska, Jarosław Kilon , Reguły asocjacyjne w analizie wyników badań metodą Delphi.....	247
Anna Domagała , O wykorzystaniu analizy głównych składowych w metodzie <i>Data Envelopment Analysis</i>	254
Alicja Grześkowiak , Analiza wykluczenia cyfrowego w Polsce w ujęciu indywidualnym i regionalnym.....	264
Anna M. Olszewska, Anna Gryko-Nikitin , Pomiar postrzegania jakości kształcenia uczelni wyższej na danych porządkowych z wykorzystaniem środowiska R.....	273
Karolina Paradysz , Hierarchiczna metoda grupowania powiatów jako podejście benchmarkowe w ocenie bezrobocia według BAEL-u w wybranych typach małych obszarów	282
Radosław Pietrzyk , Porównanie metod pomiaru efektywności zarządzania portfelami funduszy inwestycyjnych.....	290
Agnieszka Przedborska, Małgorzata Misztal , Wybrane metody statystyki wielowymiarowej w ocenie skuteczności terapeutycznej głębokiej stymulacji elektromagnetycznej u pacjentów z chorobą zwyrodnieniową stawów.....	299

Wojciech Roszka, Marcin Szymkowiak , Podejście kalibracyjne w statystycznej integracji danych	308
Iwona Skrodzka , Zastosowanie wybranych metod klasyfikacji do analizy kapitału ludzkiego krajów Unii Europejskiej	316
Agnieszka Stanimir , Wielowymiarowa analiza czynników sprzyjających włączeniu społecznemu	326
Dorota Strózik, Tomasz Strózik , Przestrzenne zróżnicowanie poziomu życia w województwie wielkopolskim.....	334
Izabela Szamrej-Baran , Identyfikacja przyczyn ubóstwa energetycznego w Polsce przy wykorzystaniu modelowania miękkiego.....	343
Janusz Tuchowski, Katarzyna Wójcik , Klasyfikacja obiektów w systemie Krajowych Ram Kwalifikacji opisanych za pomocą ontologii	353
Aleksandra Matuszewska-Janica , Grupowanie krajów Unii Europejskiej ze względu na poziom feminizacji sektorów gospodarczych	361
Monika Rozkrut, Dominik Rozkrut , Identyfikacja strategii innowacyjnych przedsiębiorstw usługowych w Polsce	369

Summaries

Małgorzata Rószkiewicz , The use of meta-analysis in building the measurement model in case of the absence of measurement invariance on the example of measuring of life satisfaction.....	20
Elżbieta Sobczak , Harmonious smart growth of European Union regions.....	29
Ewa Roszkowska, Renata Karwowska , The comparative analysis of Polish voivodeships with respect to sustainable development in 2010	40
Tadeusz Kufel, Magdalena Osińska, Marcin Błażejowski, Paweł Kufel , Comparative analysis of chosen filters in business cycles analysis	50
Marcin Salamaga , The attempt of construction of the life tables for opera works on the example of the Metropolitan Opera	58
Iwona Foryś , Using discriminant analysis to select similar markets in non-residential property valuation process.....	68
Jerzy Korzeniewski , Variable selection in classification – algorithm proposal	75
Sabina Denkowska , Multiple testing in the verification process of multifactorial Cox proportional hazards models	84
Ewa Chodakowska , The theory of structural equations modelling in the classification of observed variables and latent constructs according to the character of their relationship.....	93
Iwona Konarzewska , Modelling stock market by PCA factor model – case study	105

Katarzyna Wójcik, Janusz Tuchowski , Selection of the optimal set of relevant words in consumers opinions in the context of the opinion mining ..	115
Aleksandra Łuczak , Application of AHP-LP to the evaluation of importance of determinants of socio-economic development in the administrative units	125
Aleksandra Witkowska, Marek Witkowski , A dynamic approach to the ranking of cooperative banks by their financial condition	134
Adam Depta , Application of correspondence analysis for the measurement of quality of life – questionnaire SF-36v2 based research	145
Marek Lubicz, Maciej Zięba, Konrad Pawelczyk, Adam Rzechonek, Marek Marciniak, Jerzy Kołodziej , Classification rules extraction for missing and imbalance data: models of classifiers and initial results in the rules-based thoracic surgery risk prediction.....	155
Małgorzata Misztal , Selected methods for assessing the performance of classifiers – an overview and examples of applications.....	166
Anna M. Olszewska , The application of selected quantitative methods to the evaluation of voivodeship innovation level potential.....	176
Iwona Bąk , The comparison of the quality of groupings of poviats of West Pomeranian Voivodeship in terms of tourism attractiveness	185
Agnieszka Kozera, Joanna Stanisławska, Romana Głowicka-Wołoszyn , Household segmentation with respect to the expenditure on organized tourism.....	195
Agnieszka Wałęga , Synthetic approach in the analysis of economic coherence of households	204
Joanna Banaś, Małgorzata Machowska-Szewczyk, Bożena Mroczek , Using the correspondence analysis to examine the impact of wind turbines on the quality of life.....	213
Joanna Banaś, Krzysztof Małecki , Classification of measurement survey points of drivers on the boundary of Szczecin using symbolic variables...	221
Aneta Becker , The use granular information in the analysis of the requirements of the labor market.....	229
Katarzyna Cheba, Joanna Hołub-Iwan , The application of the correspondence analysis of patients segmentation on the medical service market	237
Adam Depta, Iwona Staniec , Identification of the factors that determine the quality of students life at universities in Lodz.....	246
Katarzyna Dębkowska, Jarosław Kilon , Association rules in the analysis of research results the Delphi method	253
Anna Domagała , About using Principal Component Analysis in Data Envelopment Analysis	263
Alicja Grześkowiak , Analysis of the digital divide in Poland at the individual and regional level	272

Anna M. Olszewska, Anna Gryko-Nikitin , Assessment of perception of quality of teaching at an institution of higher learning based on the ordinal data with the utilization of R environment.....	281
Karolina Paradysz , The hierarchical method of grouping poviats as a benchmark approach in the assessment of unemployment by BAEL in selected types of small areas	289
Radosław Pietrzyk , Comparison of methods of measuring the performance of investment funds portfolios.....	298
Agnieszka Przedborska, Małgorzata Misztal , Selected multivariate statistical analysis methods in the evaluation of efficacy of deep electromagnetic stimulation in patients with degenerative joint disease	307
Wojciech Roszka, Marcin Szymkowiak , A calibration approach in statistical data integration	315
Iwona Skrodzka , Application of some methods of classification to the analysis of human capital in the European Union.....	325
Agnieszka Stanimir , Multivariate analysis of social inclusion factors.....	333
Dorota Strózik, Tomasz Strózik , Spatial differentiation of the standard of living in Great Poland Voivodeship	342
Izabela Szamrej-Baran , Identification of fuel poverty causes in Poland using soft modelling	352
Janusz Tuchowski, Katarzyna Wójcik , Classification of objects in the National Classification Framework described by the ontology.....	360
Aleksandra Matuszewska-Janica , Clustering of European Union states taking into consideration the levels of feminization of economic sectors..	368
Monika Rozkrut, Dominik Rozkrut , Identification of service sector innovation strategies in Poland.....	379

Anna Domagała

Uniwersytet Ekonomiczny w Poznaniu

O WYKORZYSTANIU ANALIZY GŁÓWNYCH SKŁADOWYCH W METODZIE *DATA ENVELOPMENT ANALYSIS*

Streszczenie: W artykule podjęto próbę porównania wybranych dwóch metod z grupy PCA-DEA, będących połączeniem analizy głównych składowych (PCA) z metodą *Data Envelopment Analysis* (DEA). Celem PCA-DEA jest poprawa rezultatów standardowej DEA, która w sytuacji zbyt małej liczby badanych obiektów i/lub zbyt dużej liczby cech opisujących obiekty traci moc dyskryminacyjną. Badanie polegało na porównaniu rezultatów uzyskiwanych przy zastosowaniu standardowego modelu DEA i modeli PCA-DEA w sytuacji prawidłowej oraz zbyt małej liczebności badanej grupy. Wykorzystano dane rzeczywiste i symulacyjne.

Słowa kluczowe: efektywność, DEA, PCA.

1. Wstęp

Metoda badania efektywności *Data Envelopment Analysis* (DEA), zaproponowana w 1978 roku przez Charnesa, Coopera i Rhodesa [1978], zyskała już wielu zwolenników zarówno w świecie nauki, jak i wśród praktyków, gdzie wykorzystywana jest jako narzędzie wsparcia w procesie zarządzania.

Z uwagi na względny charakter rezultatów metody DEA¹ kwestią niezwykle istotną jest prawidłowy dobór zmiennych opisujących badane obiekty. Nabiera to szczególnego znaczenia w przypadku, kiedy badaniu poddawana jest niewielka liczba obiektów opisana dużą liczbą zmiennych. Wskaźniki efektywności DEA ulegają wtedy przeszacowaniu, tworząc tzw. chmurę w bliskim otoczeniu granicy efektywności, co z kolei osłabia siłę dyskryminacyjną metody. Dlatego zaleca się, aby minimalna liczba badanych obiektów wynosiła [Cooper, Seiford, Tone 2007, s. 284]:

$$n_{\min} = \max \{ m \cdot s; 3 \cdot (m + s) \}. \quad (1)$$

¹ Efektywność danego obiektu w metodzie DEA obliczana jest w oparciu o obiekty uznane przez metodę za efektywne, a więc wzorcowe (*benchmarks*).

Oznacza to, iż zalecana liczba n badanych obiektów powinna zależeć od liczby m nakładów i s wyników opisujących dany obiekt. W sytuacji, gdy warunek ten nie jest spełniony należy rozważyć możliwość zwiększenia liczby badanych obiektów i/lub usunąć niektóre zmienne opisujące objekty. Jeżeli nie można już bardziej zredukować liczby zmiennych i nie ma też możliwości zwiększenia liczebności badanej grupy, z pomocą może przyjść analiza głównych składowych, która umożliwia zredukowanie wymiarów modelu DEA, a mimo to pozwala na ujęcie w badaniu wszystkich pożądanych zmiennych.

Celem opracowania jest przedstawienie oraz porównanie wybranych dwóch proponowanych w literaturze (wskazanej poniżej) sposobów połączenia analizy głównych składowych z DEA oraz próba potwierdzenia hipotezy badawczej, według której wykorzystanie analizy głównych składowych może znacząco poprawić rezultaty metody DEA.

2. Charakterystyka wykorzystanych metod

DEA (*Data Envelopment Analysis*)

Opracowana przez Charnesa, Coopera i Rhodesa [1978] wielowymiarowa, nieparametryczna, oparta na programowaniu liniowym metoda oceny względnej efektywności działania obiektów. Podstawowy model DEA to zorientowany na nakłady, radialny model CCR [Cooper, Seiford, Tone 2007, s. 43]. Obecnie pod nazwą *Data Envelopment Analysis* kryje się szeroki wachlarz modeli będących mniej lub bardziej zaawansowanymi modyfikacjami podstawowego modelu CCR².

PCA (*Principal Component Analysis*)

Analiza głównych składowych jest opartą na macierzy kowariancji lub macierzy korelacji wielowymiarową metodą ortogonalnej transformacji układu badanych cech, gdzie p obserwowalnych cech wejściowych przekształcanych jest w $r \leq p$ nieobserwowalnych i nieskorelowanych ze sobą cech zwanych głównymi składowymi, będącymi liniowymi kombinacjami cech wejściowych [Krzanowski 2008, s. 58]. Kolejne główne składowe charakteryzują się coraz mniejszą wariancją (będącą miarą ich zasobów informacyjnych), a więc pierwsza główna składowa wyjaśnia największy procent wariancji cech wejściowych. Jak pisze Krzanowski [2008, s. 66], podstawowym celem analizy głównych składowych jest redukcja liczby cech opisujących objekty.

² Np. radialny model BCC, uwzględniający zmienne efekty skali, ale także modele nieradialne, takie jak model addytywny lub model SBM (*Slacks-Based Measure*) – więcej o metodzie DEA zob. np. [Cooper, Seiford, Tone 2007], a w polskiej literaturze [Guzik 2009].

PCA-DEA

Jest to grupa metod będących mniej lub bardziej zaawansowanym połączeniem metod PCA i DEA. Mimo różnic w sposobie powiązania obu metod, wszystkie podejścia posiadają wspólną ideę – dążenie do poprawy mocy dyskryminacyjnej DEA poprzez redukcję liczby zmiennych opisujących badane obiekty. Poniżej omówiono wybrane dwa podejścia.

1. Shanmugama i Johnsona [2007] – przebiega dwuetapowo: najpierw wyznaczane są główne składowe osobno dla zbioru nakładów i wyników, a następnie w wybranym do badania modelu DEA zastępuje się oryginalne nakłady i wyniki pierwszymi głównymi składowymi. Decyzję o tym, ile głównych składowych wprowadzić do modelu DEA (a więc jaki przyjąć procent wariacji zmiennych wyjaśniany przez wybrane główne składowe), podejmuje badacz³.

2. Adler i Yazhensky [2010] – polega na włączeniu głównych składowych (wyznaczanych osobno dla nakładów oraz wyników) bezpośrednio do formalnego zapisu danego modelu DEA. Przykładowo dla modelu CCR-I [Adler, Yazhensky 2010, s. 275]:

$$\max U_o^T Y_o + U_{o,PC}^T Y_{o,PC} \quad (2)$$

przy warunkach:

$$(a) V_o^T X_o + V_{o,PC}^T X_{o,PC} = 1,$$

$$(b) V_o^T X_j + V_{o,PC}^T X_{j,PC} - U_o^T Y_j - U_{o,PC}^T Y_{j,PC} \geq 0,$$

$$(c) V_o, U_o \geq 0, \quad (d) V_{o,PC}^T L_x, U_{o,PC}^T L_y \geq 0, \quad (e) V_{o,PC}, U_{o,PC} - \text{bez ogr.}$$

gdzie: $X_o, X_j, Y_o, Y_j, V_o, U_o$ – tak jak w standardowym modelu CCR-I [por. Cooper, Seiford, Tone 2007, s. 43],

$X_{o,PC}, Y_{o,PC}$ – wektory głównych składowych związanych odpowiednio z nakładami i wynikami badanego obiektu o -tego⁴,

$V_{o,PC}, U_{o,PC}$ – wektory wag związanych odpowiednio z $X_{o,PC}$ oraz $Y_{o,PC}$,

L_x, L_y – wektory ładunków głównych składowych związane odpowiednio z wektorem głównych składowych $X_{o,PC}$ oraz $Y_{o,PC}$.

Warunki (d) i (e) rozwiązują ewentualny problem ujemnych wartości głównych składowych⁵. Decyzję o tym, ile głównych składowych wprowadzić do modelu DEA, podobnie jak poprzednio, podejmuje badacz.

W tabeli 1 podsumowano najważniejsze wady i zalety obu podejść z grupy PCA-DEA.

³ Zaleca się postępowanie zgodne z zasadami przyjętymi w analizie głównych składowych, a więc np. kryterium wartości własnej Kaisera czy wykres ospiska Cattella. Patrz: np. [Panek 2009, s. 181].

⁴ $X_{o,PC}$ oraz $Y_{o,PC}$ w modelu DEA traktowane są jak oryginalne nakłady i wyniki.

⁵ Ujemne wartości głównych składowych mogą wynikać tylko z ujemnych wartości ich ładunków, gdyż oryginalne zmienne wykorzystywane w modelu CCR-I są dodatnie.

Tabela 1. Wady i zalety PCA-DEA

Podejście/cechy	Wady	Zalety
Shanmugam i Johnson [2007]	brak równoważności rezultatów PCA-DEA i DEA przy 100% wyjaśnionej wariancji utrudnieniem mogą stać się ujemne ładunki głównych składowych ⁶	prostota obliczeniowa (dzięki dwuetapowości podejścia) możliwość zastosowania z dowolnym modelem DEA (także korzystając z dostępnych programów komputerowych dedykowanych standardowej DEA)
Adler i Yazhemyky [2010]	rozbudowana postać modelu DEA (pakiety komputerowe z oprogramowanymi standardowymi modelami DEA nie znajdują zastosowania) ⁷	równoważność rezultatów analizy DEA i PCA-DEA w przypadku ujęcia w modelu (2) wszystkich głównych składowych uwzględnienie ryzyka wystąpienia ujemnych głównych składowych

Źródło: opracowanie własne.

3. Warianty badania

W przeprowadzonym badaniu empirycznym porównano działanie standardowej DEA i omówionych powyżej podejść PCA-DEA. Dla ułatwienia pierwsze podejście PCA-DEA będzie nazywane w skrócie „Shan_John”, a drugie – „Adl_Yaz”. Przykładowo, symbol „Shan_John_2_1” będzie związany z rezultatami badania metodą PCA-DEA w wersji zaproponowanej przez Shanmugama i Johnsona [2007], w której wykorzystano dwie pierwsze główne składowe opisujące nakłady oraz (jedną) pierwszą główną składową opisującą wyniki danego obiektu.

Badanie przeprowadzono w dwóch wariantach, które szczegółowo opisano w tabeli 2.

W wariacie I wykorzystano dane rzeczywiste – przeżywalność kobiet i mężczyzn w chorobie nowotworowej (rak skóry) – podane przez Shanmugama i Johnsona [2007] i porównano rezultaty metody DEA oraz obu wersji PCA-DEA. Wykorzystano model BCC-O, gdyż model ten zastosowano w pracy Shanmugama i Johnsona [2007]. Badaniu poddano 45 krajów opisanych czterema nakładami i dwoma wynikami.

⁶ Wydaje się jednak, iż ewentualny problem można rozwiązać albo przez zmianę wszystkich znaków ładunków na przeciwne (co prawdopodobnie wykorzystali w swoim podejściu autorzy), albo poprzez rotację układu osi, która została wykorzystana w niniejszym badaniu.

⁷ Adler i Yazhemyky [2010] udostępniają swój autorski program komputerowy, jednak obejmuje on tylko trzy modele DEA – radialne modele CCR i BCC oraz model addytywny.

Tabela 2. Warianty badania empirycznego

Wariant / charakterystyka	Wariant I dane rzeczywiste	Wariant II dane symulacyjne
Dane	rzeczywiste, źródło: [Shanmugam, Johnson 2007]	symulacyjne: funkcja Cobba-Douglassa (stałe efekty skali) z nałożoną nieefektywnością na nakłady lub wyniki wybranych obiektów, źródło: badania własne
Liczba zmiennych (nakłady + wyniki)	4 + 2	6 + 3
Liczba obiektów i zalecana dla DEA minimalna liczba obiektów zgodnie ze wzorem (1)	45 krajów zalecana: $n_{\min} = 18$	15 obiektów zalecana: $n_{\min} = 27$
Zastosowany model DEA	BCC-O	CCR-I

Źródło: opracowanie własne.

Tabela 3. Liczba głównych składowych wykorzystana w wariantach badania

Wykorzystana w PCA-DEA liczba głównych składowych (procent wyjaśnionej wariancji nakładów oraz wyników)	
Wariant I	Shan_John_2_1 (99,53% oraz 92,94%), Adl_Yaz_2_1 (99,54% oraz 92,94%)
Wariant II	a) nieefektywność po stronie wszystkich nakładów: Shan_John_2_1 (98,41% oraz 93,94%), Adl_Yaz_2_1 (98,42% oraz 93,94%) b) nieefektywność dla dwóch podgrup nakładów: Shan_John_2_1 (98,51% oraz 93,94%), Adl_Yaz_2_1 (98,52% oraz 93,94%) c) nieefektywność po stronie wszystkich wyników: Shan_John_2_1 (97,82% oraz 92,00%), Adl_Yaz_2_1 (97,84% oraz 92,02%)

Źródło: opracowanie własne.

Jak widać w tabeli 3⁸, w wariancie I w obu wersjach analizy PCA-DEA wykorzystano dwie pierwsze główne składowe dla nakładów (99,5% wyjaśnionej wariancji nakładów) i pierwszą główną składową dla wyników (92,9% wariancji wyników).

W wariancie II wygenerowano dane sztuczne⁹. Utworzono 15 obiektów opisanych 6 nakładami i 3 wynikami. Nakłady podzielono na dwie równoliczne podgrupy o wysokiej korelacji wewnętrznej¹⁰ i niskiej korelacji zewnętrznej¹¹. Wyniki

⁸ Różnice dla obu podejść PCA-DEA w zakresie stopnia wyjaśnionej wariancji wynikają z faktu, iż w obliczeniach metodą Shanmugama i Johnsona [2007] w przypadku ujemnych głównych składowych dokonywano rotacji osi. W podejściu Adler i Yazhemskey [2010] nie dokonuje się rotacji.

⁹ Z uwagi na ograniczoną objętość tekstu szczegóły sposobu symulacji danych nie zostaną omówione.

¹⁰ Współczynniki korelacji liniowej Pearsona w obu podgrupach przekraczały 0,9.

także były silnie ze sobą skorelowane. Otrzymano w ten sposób grupę 15 jednostek efektywnych w sensie DEA. Następnie dla wybranych obiektów wprowadzono: a) nieefektywność po stronie wszystkich nakładów, b) nieefektywność osobno w dwóch podgrupach nakładów oraz c) nieefektywność po stronie wszystkich wyników¹². W wariancie II przyjęto taką liczbę pierwszych głównych składowych, która wyjaśnia ponad 90% wariancji nakładów i wyników¹³.

4. Rezultaty badania i wnioski

W wariancie I zbadano 45 krajów, opisanych czterema nakładami i dwoma wynikami. Oznacza to, iż zastosowanie metody PCA-DEA nie jest tutaj konieczne¹⁴ i należy oczekiwać, iż jej rezultaty będą zbliżone do rezultatów standardowej DEA. Badania potwierdziły te przypuszczenia (wysokie wartości współczynników korelacji¹⁵ w tabeli 4), ale pokazały także coś więcej.

Tabela 4. Współczynniki korelacji pomiędzy rezultatami DEA i PCA-DEA (wariant I)

Metoda / współczynniki korelacji z rezultatami standardowej DEA (model BCC-O)	Współczynnik korelacji liniowej Pearsona	Współczynnik korelacji tau Kendalla
PCA-DEA w wersji Shan_John_2_1 (model BCC-O)	0,9218	0,7815
PCA-DEA w wersji Adl_Yaz_2_1 (model BCC-O)	0,9227	0,7568

Źródło: opracowanie własne.

Niższy współczynnik korelacji rang (tau Kendalla) wskazuje, iż nastąpiła zmiana pozycji rankingowych niektórych obiektów. Wynika to między innymi z faktu, iż standardowa DEA wskazała 9 obiektów efektywnych (w analizie Adl_Yaz_2_1 pojawiły się 3, a w Shan_John_2_1 tylko 2 obiekty wzorcowe). Wydawałoby się, iż PCA-DEA błędnie szacuje efektywność, jednak bliższe przyjrzenie się rezultatom obu analiz oraz danym wejściowym zdaje się wskazywać, iż to właśnie PCA-DEA lepiej oddaje efektywność obiektów. Warto dokładniej przyjrzeć się trzem krajom: Kostaryka i Trynidad oraz Singapur. Dane tych krajów (na-

¹¹ Współczynniki korelacji pomiędzy nakładami należącymi do różnych podgrup nie przekraczały 0,2.

¹² Wprowadzanie nieefektywności polegało na zwiększeniu nakładów przy niezmiennych wynikach (nieefektywność po stronie nakładów) lub zmniejszeniu wyników bez zmiany nakładów (nieefektywność po stronie wyników).

¹³ Patrz: tabela 3.

¹⁴ Warunek opisany wzorem (1) jest spełniony.

¹⁵ Oddających podobieństwo pomiędzy rezultatami danego podejścia PCA-DEA a rezultatami standardowej DEA. Wszystkie podane współczynniki są statystycznie istotne (wartość $p < 0,01$).

kłady oznaczono symbolem x_i , a wyniki y_r) i wskaźniki efektywności uzyskane w przypadku standardowej DEA oraz obu wersji PCA-DEA zawarto w tabeli 5.

Tabela 5. Dane wejściowe oraz rezultaty DEA oraz PCA-DEA (wariant I badania)

Dane i typ analizy / obiektu	Nakłady				Wyniki		Wskaźniki efektywności (model BCC-O)		
	x_1	x_2	x_3	x_4	y_1	y_2	DEA	PCA-DEA w wersji Shan_John_2_1	PCA-DEA w wersji Adl_Yaz_2_1
Kostaryka	10	63 600	2064	275	0,89	0,99	100%	94,95%	96,12%
Trynidad	10	63 600	2064	275	0,97	0,98	99,49%	98,48%	98,62%
Singapur	1	64 700	2127	225	0,97	0,99	100%	98,99%	100%

Źródło: opracowanie własne.

Kostaryka (wraz z Singapurem) charakteryzuje się najwyższą wartością wyniku y_2 (na tle wszystkich 45 badanych obiektów) i z tego tylko względu osiąga w standardowym modelu DEA efektywność 100%¹⁶. Pozycja Trynidadu, który przy tych samych nakładach ma nieznacznie tylko niższą wartość tego wyniku, a dużo wyższą wartość wyniku y_1 jest już niższa. Wydawałoby się, iż to Trinidad powinien uzyskać wyższą wartość wskaźnika efektywności, a oceniony jest przez DEA słabiej. Jeżeli jednak spojrzymy na rezultaty analizy PCA-DEA zawarte w tabeli 5, okazuje się, iż tu obiekty zostały ocenione poprawnie. Najwyższą efektywnością charakteryzuje się zgodnie z oczekiwaniami Singapur, na drugiej pozycji znajduje się Trinidad, a na trzeciej Kostaryka.

Analizując rezultaty badania w wariacie I, można zatem spróbować wysunąć ostrożny wniosek, iż zastosowanie PCA-DEA w przypadku odpowiednio dużej liczby badanych obiektów¹⁷ nie zmienia znacząco rezultatów analizy efektywności (co potwierdzają współczynniki korelacji), ale pozwala pokonać wskazaną wyżej wadę radialnego modelu DEA¹⁸.

Wariant II zakładał porównanie działania omawianych metod w przypadku niespełnienia warunku (1) dotyczącego liczebności grupy. Osłabiona wtedy moc dys-

¹⁶ W przypadku modeli radialnych zorientowanych na wyniki (a takim jest model BCC-O) wystarczy, aby jeden z wyników obiektu miał wartość najwyższą (na tle pozostałych obiektów), a metoda wskaże ten obiekt jako efektywny. Stąd też standardowa DEA wskazuje Kostarykę jako efektywną, zrównując ją tym samym z Singapurem (który powinien uzyskać dużo wyższą pozycję rankingową niż Kostaryka).

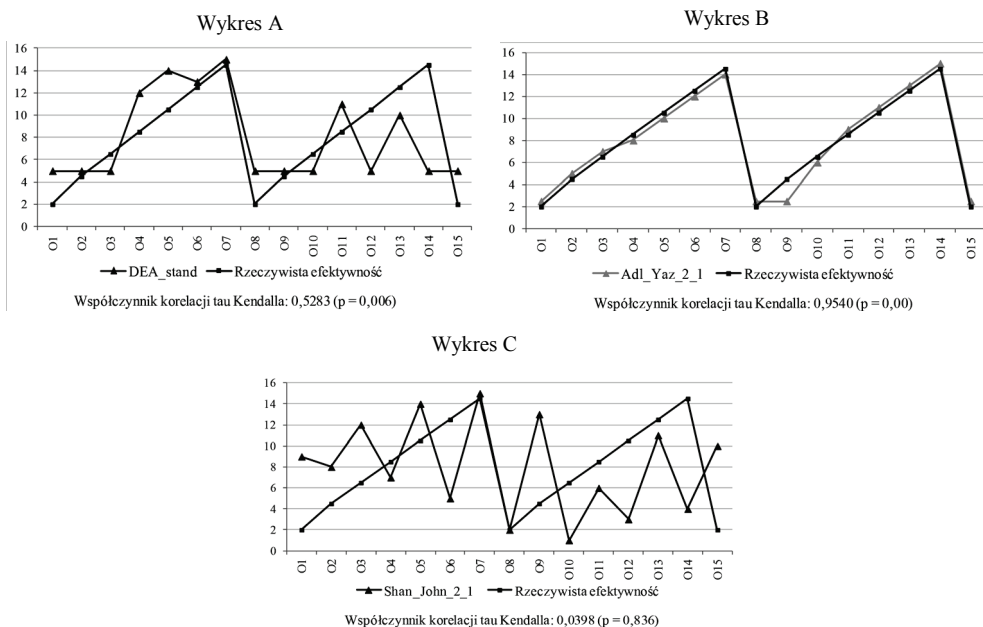
¹⁷ Spełniającej więc warunek określony wzorem (1).

¹⁸ Wniosek nazwano ostrożnym, gdyż wymaga on potwierdzenia w dodatkowych badaniach, które autorka planuje przeprowadzić.

kryminacyjna standardowej DEA powoduje, iż rezultaty tej metody nie są wiarygodne. Aby móc zweryfikować działanie metod z grupy PCA-DEA, trzeba zatem analizować obiekty sztuczne, których faktyczna efektywność jest z góry ustalona i znana.

Jak już wspomniano, badaniu poddano tu 15 obiektów, opisanych 6 nakładami i 3 wynikami, co oznaczało celowe niespełnienie warunku (1). Jak się okazało, bez względu na sposób wprowadzenia nieefektywności (po stronie nakładów lub wyników), rezultaty obliczeń były bardzo podobne – ocenę efektywności najbardziej zbliżoną do rzeczywistej każdorazowo uzyskiwano za pomocą podejścia PCA-DEA zaproponowanego przez Adler i Yazhemsy [2010]. Poniżej przedstawiono zatem dokładniej tylko rezultaty badania przy nieefektywności po stronie wszystkich nakładów (zob. wariant II w tabeli 3, pkt a).

Na zamieszczonych na rysunku 1 wykresach zestawiono rzeczywiste miejsca rankingowe badanych 15 obiektów z miejscami rankingowymi wskazywanymi przez standardową DEA (wykres A) oraz oba podejścia PCA-DEA (wykresy B oraz C). Występują rangi wiązane¹⁹. Podano również wartości współczynników korelacji tau Kendalla.



Rys. 1. Rankingi obiektów – nieefektywność po stronie nakładów

Źródło: opracowanie własne.

¹⁹ Z tego względu przykładowo obiekty efektywne (których było więcej niż jeden) nie miały rang równych 1. Obiekty efektywne, a więc zajmujące najwyższe pozycje rankingowe, odzwierciedlają punkty leżące na rysunkach najniżej.

Duża liczba obiektów efektywnych²⁰ wskazanych przez standardową metodę DEA potwierdza, iż metoda ta w przypadku niespełnionego warunku (1) traci moc dyskryminacyjną (choć nadal poprawnie rozpoznaje obiekty faktycznie efektywne). Jeśli chodzi o metodę PCA-DEA, widać wyraźnie, iż tylko w podejściu Adler i Yazhemsy [2010] efektywność została oceniona właściwie. W metodzie Shanmugama i Johnsona [2007] zaobserwowano zaskakująco słabą zbieżność między efektywnością wskazywaną przez tę metodę a faktyczną efektywnością badanych obiektów.

Powyższe wnioski należy traktować jako wstępne i bardzo ostrożne. Badania z wykorzystaniem danych symulacyjnych będą przez autorkę kontynuowane, w celu zweryfikowania hipotezy o przewadze podejścia Adler i Yazhemsy [2010] nad podejściem zaproponowanym przez Shanmugama i Johnsona [2007].

5. Podsumowanie oraz kierunki dalszych badań

Pomysł wykorzystania analizy głównych składowych w metodzie DEA wydaje się być bardzo obiecujący, jednak różnice pomiędzy proponowanymi w literaturze metodami z grupy PCA-DEA wskazują konieczność wnikliwego porównania generowanych przez nie rezultatów.

Ponadto należy podkreślić, iż niniejsza praca miała na celu jedynie porównanie i sprawdzenie samego działania danego podejścia PCA-DEA. Kolejnym obszarem, w którym należy prowadzić badania, to obszar praktycznego wykorzystania uzyskanych rezultatów. Wprowadzenie do modeli DEA głównych składowych zamiast²¹, lub obok²², oryginalnych nakładów i wyników opisujących obiekty może utrudniać interpretację rezultatów badania.

Warto również przyjrzeć się możliwościom oprogramowania poszczególnych modeli PCA-DEA, gdyż dostępne programy komputerowe pozwalają na wykorzystanie tylko modeli podstawowych. Duży potencjał wydaje się drzeć w środowisku typu „R”.

Literatura

- Adler N., Yazhemsy E. (2010), *Improving discrimination in data envelopment analysis: PCA-DEA or variable reduction*, „European Journal of Operational Research” No. 202, s. 273-284.
- Charnes A., Cooper W.W., Rhodes E. (1978), *Measuring the Efficiency of Decision Making Units*, „European Journal of Operational Research” 2, s. 429-444.
- Cooper W.W., Seiford L.M., Tone K. (2007), *Data Envelopment Analysis. A Comprehensive Text with Models, Applications, References and DEA-Solver Software*, Springer, New York.

²⁰ Na rysunku 1 są to punkty położone najniżej, na poziomie „5” (z uwagi na występujące rangi związane).

²¹ Jak to się dzieje w podejściu Shanmugama i Johnsona [2007].

²² Podejście Adler i Yazhemsy [2010].

- Guzik B. (2009), *Podstawowe modele DEA w badaniu efektywności gospodarczej i społecznej*, Wydawnictwo Uniwersytetu Ekonomicznego w Poznaniu, Poznań.
- Krzyszowski W.J. (2008), *Principles of Multivariate Analysis. A User's Perspective*, Oxford University Press, New York.
- Panek T. (2009), *Statystyczne metody wielowymiarowej analizy porównawczej*, Szkoła Główna Handlowa w Warszawie – Oficyna Wydawnicza, Warszawa.
- Shanmugam R., Johnson C. (2007), *At a crossroad of data envelopment and principal component analyses*, „Omega”, No. 35, s. 351-364.

ABOUT USING PRINCIPAL COMPONENT ANALYSIS IN DATA ENVELOPMENT ANALYSIS

Summary: The article presents a comparison of two selected methods of PCA-DEA which are a connection of Principal Component Analysis (PCA) and Data Envelopment Analysis (DEA). The aim of PCA-DEA methods is to improve the results of a traditional DEA which discriminatory power weakens when the number of variables that describe objects increases and/or when the number of objects decreases. The results of a traditional DEA and PCA-DEA were compared in case of correct and too small group of studied objects. Real and simulated data sets were used.

Keywords: efficiency, DEA, PCA.