

# Kesra Nermend

---

## Eliminacja zmiennych na podstawie rankingu

---

Ekonomiczne Problemy Usług nr 106, 229-238

---

2013

Artykuł został opracowany do udostępnienia w internecie przez Muzeum Historii Polski w ramach prac podejmowanych na rzecz zapewnienia otwartego, powszechnego i trwałego dostępu do polskiego dorobku naukowego i kulturalnego. Artykuł jest umieszczony w kolekcji cyfrowej [bazhum.muzhp.pl](http://bazhum.muzhp.pl), gromadzącej zawartość polskich czasopism humanistycznych i społecznych.

Tekst jest udostępniony do wykorzystania w ramach dozwolonego użytku.

KESRA NERMEND

Uniwersytet Szczeciński

## ELIMINACJA ZMIENNYCH NA PODSTAWIE RANKINGU

### Wprowadzenie

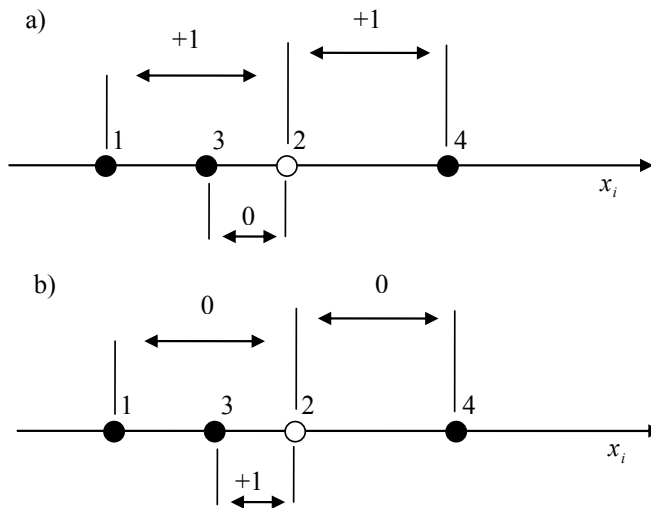
Określenie istotności zmiennych jest bardzo ważnym zagadnieniem. Przeprowadzając rangowanie, należy określić, które zmienne są ważne dla analizowanego zjawiska. Następnie wybrane zmienne należy zweryfikować, eliminując te zmienne, które nie różnicują obiektów albo różnicują je w sposób, który nie jest istotny z punktu widzenia przeprowadzanych badań<sup>1</sup>. Przy eliminacji zmiennych można wspomóc się rankingiem przeprowadzonym przez ekspertów mogących wskazać, które obiekty uważają za słabe, które za średnie, a które za dobre. Obserwując wartości zmiennych dla obiektów wzorcowych, można na ich podstawie określić, które zmienne są skorelowane z rankingiem. W klasycznym podejściu najczęściej używa się korelacji<sup>2</sup>. Ze względu na to, że ranking i wartości zmiennych mają różne skale pomiarowe, nie można w tym przypadku wykorzystać korelacji. Ale można wyliczyć współczynnik, który będzie określał zgodność wartości zmiennej z rankingiem.

<sup>1</sup> M. Kolenda, *Taksonomia numeryczna. Klasyfikacja, porządkowanie i analiza obiektów wielocechowych*, Wydawnictwo Akademii Ekonomicznej we Wrocławiu, Wrocław, 2006; K. Kukuła, *Metoda unitaryzacji zerowanej*, PWN, Warszawa 2000; M. Sobczyk, *Analiza porównawcza produkcji rolniczej w województwie lubelskim*, „Wiadomości Statystyczne” 1977, nr 2, s. 677–680; B. Pawełek, *Metody normalizacji zmiennych w badaniach porównawczych złożonych zjawisk ekonomicznych*, Wydawnictwo Uniwersytetu Ekonomicznego w Krakowie, Kraków 2008.

<sup>2</sup> I. Bronsztajn, K. Siemiendajew, *Matematyka. Poradnik encyklopedyczny*, PWN, Warszawa 2000.

## 1. Sprawdzanie zgodności zmiennej z rankingiem

Współczynnik zgodności zmiennej z rankingiem można policzyć, sprawdzając, czy każdy z obiektów jest prawidłowo umiejscowiony w rankingiu. W tym celu liczy się współczynniki  $w'_i$  i  $w''_i$ . Współczynnik  $w'_i$  określa, ile obiektów będących przed danym obiektem w rankingiu ma mniejsze lub równe od niego wartości zmiennej i ile obiektów będących za danym obiektem w rankingiu ma większe lub równe wartości zmiennej (rys. 1a). Podobnie liczony jest współczynnik  $w''_i$ . Określa on, ile obiektów będących przed danym obiektem w rankingiu ma większe lub równe od niego wartości zmiennej i ile obiektów będących za danym obiektem w rankingiu ma mniejsze lub równe wartości zmiennej (rys. 1b). Liczenie tych dwóch współczynników jest koniecznością, gdyż nie jest wiadomo, czy ranking jest skorelowany dodatnio czy ujemnie z wartościami zmiennej. W przypadku rangowania nie jest istotny rodzaj korelacji, ale sam fakt jej istnienia.



Rys. 1. Zasada wyznaczania wartości współczynnika dla jednego obiektu:

a)  $w'_{i's}$ , b)  $w''_{j'i}$

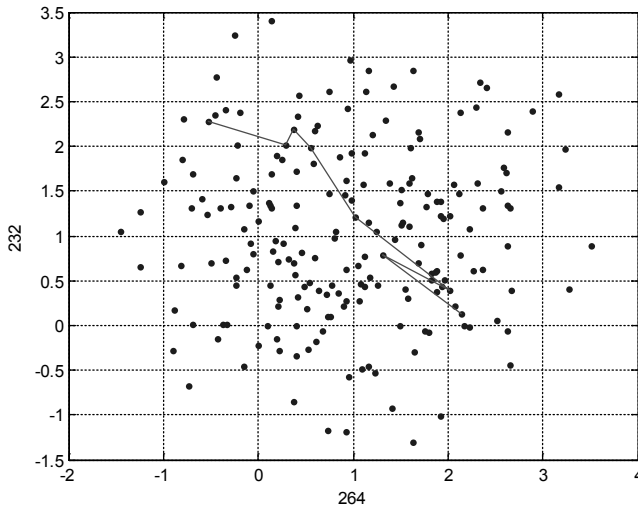
Źródło: opracowanie własne.

Na podstawie współczynników  $w'_j$  i  $w''_j$  wyznacza się właściwy współczynnik zgodności zmiennej z rankingiem:

$$w_i = \frac{\sum_{j=1}^M \max \left\{ \begin{matrix} w'_j \\ j \end{matrix} ; \begin{matrix} w''_j \\ j \end{matrix} \right\}}{M(M-1)} 100. \quad (1)$$

Współczynnik  $w_i$  określa procentową zgodność zmiennej z rankingiem. Wartość 100 oznacza pełną zgodność. Zmienna porządkuje wszystkie obiekty zgodnie z rankingiem. Najmniejszą możliwą wartością jest 50. Wartość 0 jest niemożliwa do osiągnięcia, gdyż oznaczałaby skorelowanie ujemne. Zmienna porządkowałaby wszystkie obiekty w kolejności odwrotnej niż w rankingu. W rangowaniu oznaczałoby to zmienną będącą destymulantą, która stanowi równie wartościową zmienną jak stymulantą.

Proponowana metoda może zostać użyta tylko do stymulant i destymulant. W przypadku nominant muszą one zostać najpierw zamienione na stymulanty lub destymulanty.



Rys. 2. Zmienne najlepiej odwzorowujące ranking obiektów

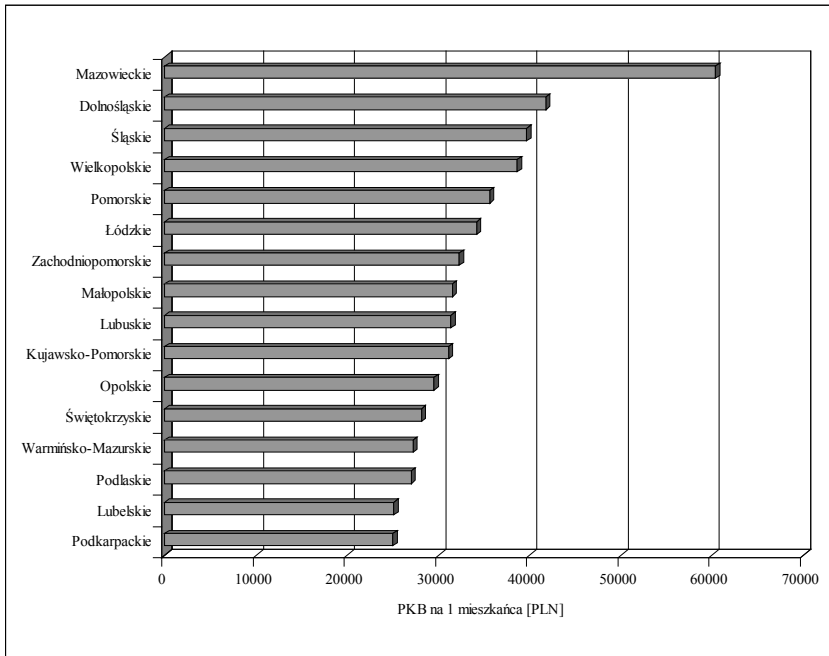
Źródło: opracowanie własne.

Badania testowe przeprowadzono na zbiorze 100 obiektów, dla których wylosowano współrzędne o rozkładzie normalnym, o odchyleniu standardowym równym jeden i wartości średniej równej jeden. Z względu na losowy charakter zbioru do badań użyto 300 zmiennych. Z tego zbioru wylosowano 8 obiektów, które uporządkowano w sposób losowy. Stanowiły one losowy ranking. Następnie spośród 300 zmiennych wybrano dwie, które najlepiej odwzorowują ranking. Na rysunku 2 przedstawiono obiekty w układzie współrzędnych tych dwóch zmiennych. Linia połączono obiekty wzorcowe w takiej kolejności, w jakiej tworzyły ranking. Na osiach przedstawiono numery zmiennych.

Dla zmiennej numer 264 wartość współczynnika  $w_i$  wyniosła 96,43, a dla zmiennej numer 232 – 92,86. Widać, że zmienna 264 lepiej odwzorowuje ranking. Tylko jeden obiekt ma niewłaściwą kolejność. Zmienna 232 ma dwa takie obiekty. W praktyce wartość współczynnika  $w_i$  powinna być bardzo wysoka, najlepiej powyżej 90%.

## 2. Badania empiryczne

Dla przykładu stworzono ranking województw i zbadano, które zmienne zapewniają identyczną kolejność obiektów jak PKB. Wszystkie dane zaczerpnięto z GUS-u za rok 2010. Na podstawie zmiennej produkt krajowy brutto na 1 mieszkańca (ceny bieżące PKD 2007) stworzono ranking województw. Został on przedstawiony na rysunku 3. Z tego rankingu wybrano cztery województwa, dwa należące do najlepszych: mazowieckie i śląskie, jedno należące do przeciętnych: zachodniopomorskie, oraz jedno należące do słabych: lubelskie. Wybrano dwa województwa należące do najlepszych ze względu na to, że województw mazowieckie mocno odstaje od pozostałych, mając o około 30% większe PKB.



Rys. 3. Produkt krajowy brutto na 1 mieszkańca w 2010 roku (ceny bieżące PKD 2007)  
Źródło: GUS.

W badaniach użyto osiemnastu zmiennych z sześciu grup (tab. 1). Zmienne opisują liczbę ludności, powierzchnię województw, rynek pracy oraz strukturę wydatków i dochodów budżetów województw. Dla każdej zmiennej wyznaczono współczynnik dany wzorem (1) względem rankingu tworzonego przez cztery województwa.

Tabela 1

## Zmienne biorące udział w badaniach

Kategoria	Grupa	Podgrupa	Opis	Współczynnik
Podmioty gospodarcze	Podmioty gospodarki narodowej wpisane do REGON-u	Jednostki wpisane wg sektorów własnościowych	podmioty gospodarki narodowej ogółem	100
			sektor prywatny ogółem	100
Dochody i wydatki budżetów jednostek samorządu terytorialnego	Dochody budżetów województw	Dochody własne budżetów województw	razem	100
	Wydatki budżetów województw	Wydatki na ochronę zdrowia	ogółem	100
Ludność	Stan ludności i ruch naturalny	Ludność w wieku przedprodukcyjnym (17 lat i mniej), produkcyjnym i poprodukcyjnym wg płci	ogółem	83
			w wieku produkcyjnym	83
Rynek pracy	Pracujący w głównym miejscu pracy	Pracujący wg płci	ogółem	83
			kobiety	83
Podmioty gospodarcze	Podmioty gospodarki narodowej wpisane do rejestru regon	Jednostki wpisane wg sektorów własnościowych	sektor publiczny ogółem	83
Dochody i wydatki budżetów jednostek samorządu terytorialnego	Dochody budżetów województw	Dochody majątkowe (z ogółem)	ogółem	83
	Wydatki budżetów województw	Wydatki na oświatę i wychowanie	ogółem	83
Rynek pracy	Bezrobocie	Udział bezrobotnych zarejestrowanych w liczbie ludności w wieku produkcyjnym	ogółem	67
Leśnictwo	Lasy niestanowiące własności skarbu państwa	Powierzchnia gruntów leśnych	ogółem	67

Kategoria	Grupa	Podgrupa	Opis	Współczynnik
Dochody i wydatki budżetów jednostek samorządu terytorialnego	Dochody budżetów województw	Subwencje ogólne	razem	67
		Dotacje	dotacje celowe	67
		Dochody na 1 mieszkańca	ogółem	67
	Wydatki budżetów województw	Wydatki na 1 mieszkańca	ogółem	67
Podział terytorialny	Podział administracyjny, sieć osadnicza	Powierzchnia	ogółem w km kw.	50

Źródło: GUS.

Cztery zmienne zapewniły 100-procentową zgodność z rankingiem. Oznacza to, że ranking wykonany na podstawie tych czterech zmiennych zapewni kolejność obiektów określoną we wstępnym rankingu. Pierwsze dwie z tych zmiennych dotyczą liczby podmiotów gospodarczych, pierwsza wszystkich, a druga prywatnych. Oznacza to, że na PKB istotny wpływ ma liczba podmiotów gospodarczych, a przede wszystkim prywatnych. Podmioty sektora publicznego mają znacznie mniejszy wpływ, o czym świadczy współczynniki na poziomie 83% dla tych podmiotów, co oznacza, że ta zmienna nie zapewnia 100% z pierwotnym rankingiem. Dwie następne zmienne dotyczą przychodów i wydatków budżetów województw. Pierwsza – są to przychody własne województw, druga – wydatki na ochronę zdrowia. Wynika z tego, że bogatsze województwa, w których istnieje większe PKB na jednego mieszkańca, większe kwoty wydają na ochronę zdrowia swoich mieszkańców.

Na podstawie czterech zmiennych o 100-procentowym współczynniku wykonano ranking województw. W przeprowadzonych badaniach wykorzystano metodę porządkowania liniowego zwaną miarą wektorową<sup>3</sup>. Zaletą miary wektorowej jest jej duża elastyczność, szczególnie w jej odmianie wy-

<sup>3</sup> M. Kolenda, *op. cit.*; K. Nermend, *A synthetic measure of sea environment pollution*, „Polish Journal of Environmental Studies” 2006, Vol. 15, No. 4b, s. 127–129; *idem*, *Using average-variance representation in economic analyses*, „Polish Journal of Environmental Studies” 2006, Vol. 15, No. 4C, s. 123–126, *idem*, *Employing similarity measures to examine the development of technical infrastructure in Polish counties*, *Folia Oeconomica Stetinensia* 2008, Vol. 15, No. 7, s. 87–97, *idem*, *Zastosowanie rzutu wektora do budowy miernika syntetycznego*, „Przegląd Statystyczny” 2008, nr 3.



korzystującej wyłącznie iloczyn skalarny<sup>4</sup> Daje ona możliwość wykorzystania dowolnego iloczynu skalarnego do wyznaczenia miary syntetycznej. Wyniki zaprezentowano w tabeli 2. Można zauważyć, że cztery województwa z rankingu wstępnego mają kolejność zgodną z kolejnością w tym rankingu, wynika to ze 100% wartości współczynnika dla czterech wybranych zmiennych.

Tabela 2

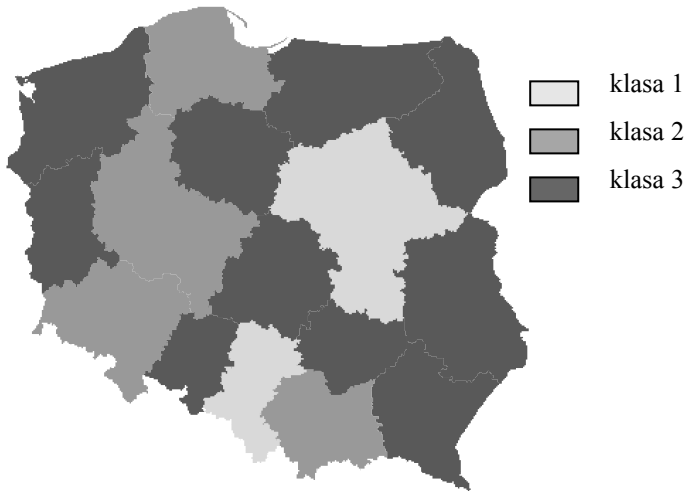
Ranking województw

Województwo	Miara	Klasa
Mazowieckie	2,36	1
Śląskie	0,89	1
Dolnośląskie	0,68	2
Wielkopolskie	0,50	2
Pomorskie	0,33	2
Małopolskie	0,19	2
Łódzkie	-0,07	3
Zachodniopomorskie	-0,18	3
Kujawsko-pomorskie	-0,31	3
Lubelskie	-0,40	3
Podkarpackie	-0,47	3
Świętokrzyskie	-0,64	3
Warmińsko-mazurskie	-0,69	3
Opolskie	-0,69	3
Lubuskie	-0,72	3
Podlaskie	-0,78	3

Źródło: opracowane własne.

Na podstawie uzyskanej wartości miary województwa podzielono na klasy, co pozwoliło na ich zwizualizowanie na mapce (rys. 4). Można zauważyć, że uzyskany ranking nie jest dokładnie zgodny z rankingiem z rysunku 3. Wpłynęła na to mała liczba obiektów w rankingu wstępnym. Ogranicza to ilość informacji, którą dysponuje proponowana metoda, co zwiększa wartość współczynnika. Im ranking wstępny zawiera więcej obiektów, tym dopasowanie powinno być dokładniejsze. W przypadku kiedy ranking tworzą eksperci, trudno jednak oczekiwać, że będzie on zawierał dużą liczbę obiektów, gdyż przy ich dużej liczbie człowiekowi trudno jest określić właściwą kolejność obiektów.

<sup>4</sup> K. Nermend, *Vector Calculus in Regional Development Analysis*, Series: Contributions to Economics, Springer 2009.



Rys. 4. Ranking województw

Źródło: opracowanie własne.

Jako docelową liczbę klas wybrano cztery, jednak ostatnia klasa jest klasą pustą, stąd na rysunku 4 są tylko trzy klasy. Wynika to z silnej niejednorodności województw. Z jednej strony jest jedno województwo silnie odbiegające od pozostałych (mazowieckie), którego wartość miary jest 2,5 razy większa od następnego województwa w rankingu. Z drugiej strony jest wiele województw o bardzo małej, zbliżonej wartości miary (świętokrzyskie, warmińsko-mazurskie, opolskie, lubuskie, podlaskie).

## Podsumowanie

Przedstawiona metoda sprawdzania zgodności z rankingiem pozwala na wybranie zmiennych tak, aby ranking tworzony na ich podstawie był zgodny z rankingiem zaproponowanym wstępnie. Jej własności są nieco zbliżone do korelacji liniowej, ale może działać także w przypadkach nieliniowych. Jej skuteczność zależy od liczby obiektów w rankingu wstępnym.

**Literatura**

- Bronsztejn I., Siemiendiajew K., *Matematyka. Poradnik encyklopedyczny*, PWN, Warszawa 2000.
- Kolenda M., *Taksonomia numeryczna. Klasyfikacja, porządkowanie i analiza obiektów wielocechowych*, Wydawnictwo Akademii Ekonomicznej we Wrocławiu, Wrocław, 2006.
- Kukuła K., *Metoda unitaryzacji zerowanej*, PWN, Warszawa 2000.
- Nermend K., *A synthetic measure of sea environment pollution*, „Polish Journal of Environmental Studies” 2006, Vol. 15, No. 4b.
- Nermend K., *Using average-variance representation in economic analyses*, „Polish Journal of Environmental Studies” 2006, Vol. 15, No. 4C.
- Nermend K., *Employing similarity measures to examine the development of technical infrastructure in Polish counties*, *Folia Oeconomica Stetinensia* 2008, Vol. 15, No. 7.
- Nermend K., *Zastosowanie rzutu wektora do budowy miernika syntetycznego*, „Przegląd Statystyczny” 2008, nr 3.
- Nermend K., *Vector Calculus in Regional Development Analysis*, Series: Contributions to Economics, Springer, 2009.
- Pawełek B., *Metody normalizacji zmiennych w badaniach porównawczych złożonych zjawisk ekonomicznych*, Wydawnictwo Uniwersytetu Ekonomicznego w Krakowie, Kraków 2008.
- Sobczyk M., *Analiza porównawcza produkcji rolniczej w województwie lubelskim*, „Wiadomości Statystyczne” 1977, nr 2.

**THE ELIMINATION OF VARIABLES BASED ON THE RANKING****Summary**

In the article a method that enables elimination of variables on the basis of previously created ranking is presented. Preliminary ranking shows expected order of objects and specially determined coefficient informs, to which extent the given variables guarantee the required ranking. For maximum compatibility with the preliminary ranking its order must be maintained at the rankings created on the basis of selected variables. This paper presents research on simulated data and on data from the Polish Central Statistical Office.

*Translated by Kesra Nermend*