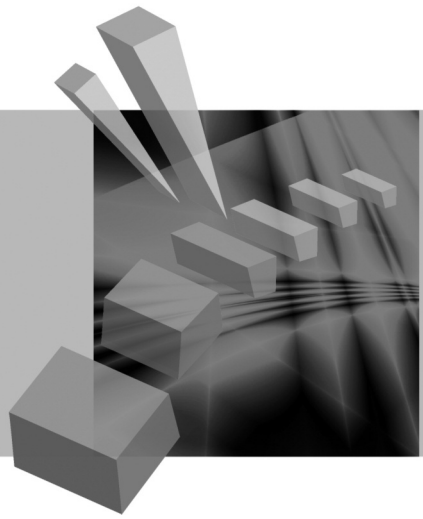
PRACE NAUKOWE
Uniwersytetu Ekonomicznego we Wrocławiu
RESEARCH PAPERS
of Wrocław University of Economics

242

Taksonomia 19.

Klasyfikacja i analiza danych – teoria i zastosowania



Redaktorzy naukowi
Krzysztof Jajuga
Marek Walesiak

Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2012

Recenzenci: Eugeniusz Gatnar, Elżbieta Gołata, Tadeusz Kufel, Józef Pociecha,
Mirosław Szreder, Feliks Wysocki

Redaktor Wydawnictwa: Aleksandra Śliwka

Redaktor techniczny: Barbara Łopusiewicz

Korektor: Barbara Cibis

Łamanie: Małgorzata Czupryńska

Projekt okładki: Beata Dębska

Tytuł sfinansowano ze środków Sekcji Klasyfikacji i Analizy Danych PTS
i Uniwersytetu Ekonomicznego we Wrocławiu

Publikacja jest dostępna na stronie www.ibuk.pl

Streszczenia opublikowanych artykułów są dostępne w międzynarodowej bazie danych
The Central European Journal of Social Sciences and Humanities http://cejsh.icm.edu.pl
oraz w 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

ISSN 1899-3192 (Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu)
ISSN 1505-9332 (Taksonomia)

Wersja pierwotna: publikacja drukowana

Druk: Drukarnia TOTEM
Nakład: 320 egz.

Spis treści

Summaries

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 242
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS 2012

Taksonomia 19. Klasyfikacja i analiza danych – teoria i zastosowania ISSN 1899-3192

Stanisława Bartosiewicz
Wyższa Szkoła Bankowa we Wrocławiu

JESZCZE RAZ O SKUTKACH SUBIEKTYWIZMU W ANALIZIE WIELOWYMIAROWEJ

Streszczenie: Artykuł jest uzupełnieniem artykułu pt. *Opowieść o skutkach subiektywizmu w analizie wielowymiarowej.* Obecnie prezentowany artykuł informuje o skutkach subiektywizmu w wyborze metod klasyfikacji obiektów ze względu na zjawisko złożone. Skutkami subiektywizmu są oczywiście zróżnicowane wyniki klasyfikacji. Wnioski z przeprowadzonych badań są następujące: 1) różnice w klasyfikacji obiektów zależą od liczby założonych klas: im więcej klas, tym większe różnice; 2) skutki subiektywizmu najwyraźniej występują przy wyborze metody klasyfikacji (taksonomia wrocławska znacznie mniej różnicuje wyniki niż metoda środków ciężkości), na drugim miejscu częściej notuje się skutki subiektywizmu przy wyborze rodzaju odległości, na trzecim (ostatnim) miejscu niekiedy (rzadziej niż przy wyborze rodzaju odległości) pojawiają się skutki subiektywizmu przy wyborze metody normalizacji cech dla zjawisk złożonych.

Słowa kluczowe: metoda środków ciężkości i taksonomii wrocławskiej, miary odległości, normalizacja.

1. Wstęp

Rok temu wystąpiłam na XIX Konferencji Sekcji Klasyfikacji i Analizy Danych, prezentując słuchaczom problemy odbiorców naszych badań ze względu na subiektywizm reprezentowany przez ekspertów w tworzeniu **rankingów** badanych obiektów w zakresie złożonych zjawisk społecznych, demograficznych lub ekonomicznych opisywanych za pomocą wielu cech (referat nosił tytuł: *Opowieść o skutkach subiektywizmu w analizie wielowymiarowej*) [Bartosiewicz, s. 17-20].

Po przeprowadzonych badaniach na przykładzie empirycznym doszłam wtedy do następujących dwu wniosków: 1) jest to wniosek banalny: *o kłopotach odbiorców badań decyduje subiektywny dobór zbioru cech opisujących badane zjawisko złożone;* 2) w drugim wniosku stwierdziłam, że właściwie *stosowane różne metody pozbawiania mian wybranych cech mają minimalny wpływ na rankingi obiektów ze względu na badane zjawisko złożone*, przeto pochopnie wyraziłam zdanie, iż poszukiwanie nowych metod w tym zakresie jest zajęciem jałowym.

Jednakże analiza wielowymiarowa polega również na stosowaniu *klasyfikacji obiektów ze względu na zjawiska złożone.* Dziś prezentuję przeto wnioski uzyskane na podstawie empirycznego badania złożonego zjawiska pod nazwą: „nauka". Podmiotami badania są województwa Polski w roku 2007, a dane pochodzą z Rocznika Statystycznego GUS-u z działu „Ważniejsze dane o województwach" [„Rocznik Statystyczny" 2008]. Jeżeli GUS uważa, że wymienione tam cechy są **ważniejsze**, to chyba wystarczają one do opisu tego zjawiska. Oto spis tych cech:

- **X1** – nakłady na działalność badawczą i naukową (ceny bieżące w milionach zł).
- **X2** – pracownicy naukowo-badawczy zatrudnieni w działalności badawczej i rozwojowej na 1000 osób aktywnych zawodowo.
- **X3** – przedsiębiorstwa posiadające dostęp do Internetu w % ogółu przedsiębiorstw.

2. Opis procedury badań

Zrezygnowałam z weryfikowania tezy prowadzącej do banalnego wniosku, że zbiór cech opisujących złożone zjawisko wybierany subiektywnie prowadzi do zróżnicowanych wyników klasyfikacji.

Klasyfikację przeprowadziłam dwiema metodami:

I. Środków ciężkości, w skrócie **Ś.C.**

II. Taksonomii wrocławskiej, w skrócie **T.W.**

W każdej z metod zastosowałam trzy rodzaje odległości:

1) miejską, w skrócie **dm**,

2) euklidesową, w skrócie **de**,

3) na podstawie cechy syntetycznej, tj. sumy znormalizowanych cech, w skrócie **ds**.

W każdym rodzaju odległości zastosowałam dwa rodzaje normalizacji cech:

a) standaryzację, w skrócie **stand**,

b) unitaryzację, w skrócie **unit**.

Kolejno podzieliłam województwa kolejno na 4, 3 oraz 2 klasy.

Skutek subiektywizmu mierzyłam, porównując 9 par realizacji danych: opisujących poszczególne badane podmioty (województwa):

- stosując **stand**, mamy 3 pary: **dm**, **de**; **dm**, **ds**; **de**, **ds**;
- stosując **unit**, mamy 3 pary (jak wyżej).

W każdym rodzaju odległości stworzono parę **stand**, **unit** (tak powstają trzy pary: **stand**, **unit** w **dm**, **stand**, **unit** w **de oraz stand**, **unit** w **ds)**.

Porównania par dokonałam za pomocą dwu mierników: współczynnika korelacji rang Spearmana, w skrócie **r,** oraz mojego prostego wskaźnika w postaci ilorazu liczby zgodnych w poszczególnych parach pozycji województw przez ich ogólną liczbę (16), skrót nazwy wskaźnika to **m.w.** Wskaźnik **m.w.** daje na ogół informacje bardziej wyraziste w porównaniu ze współczynnikiem **r**.

3. Wyniki badań

Wnioski dotyczące skutków subiektywizmu w stosowaniu klasyfikacji złożonych obiektów wynikających z analizy danych w obu wersjach tab. 1 są następujące:

1) zróżnicowanie przydziału obiektów do klas zależy od liczby klas, a mianowicie: im ich więcej, tym większe zróżnicowanie;

2) największy wpływ na zróżnicowanie klasyfikacji mają jej metody: **T.W.** wywołuje znacznie mniej różnic w przydziale obiektów do klas niż **Ś.C.**;

3) drugie miejsce pod względem wpływu na zróżnicowanie klasyfikacji zajmują częściej rodzaje odległości;

4) najsłabiej częściej działają w tym zakresie różne sposoby normalizacji cech (zob. 3 ostatnie kolumny tab. 1).

Tabela 1. Wartości **r** i **m.w.** obliczone dla par odległości w podziale obiektów na 4, 3 i 2 klasy z uwzględnieniem obu zastosowanych metod klasyfikacji **(I) Ś.C.** oraz (**II) T.W.**

Pary odległości		dm de	dm de	dm ds	dm ds	de ds	de ds	dm	de	ds
klasy		stand	unit	stand	unit	stand	unit	st. unit	st. unit	st.unit
4	***r***	***0,8***	***0,8***	***0,6***	***0,7***	***0,6***	***0,5***	***0,9***	***0,9***	***1,0***
	m.w.	***0,8***	***0,8***	***0,4***	***0,8***	***0,6***	***0,6***	***0,7***	***0,9***	***1,0***
3	***r***	0,9	1,0	0,9	1,0	0,9	0,9	1,0	0,9	0,9
	m.w.	***0,8***	**0,9**	***0,8***	**0,9**	***0,6***	***0,8***	**0,9**	***0,8***	***0,8***
2	***r***	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0
	m.w.	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0
(II) T.W.										
Pary odległości		dm de	dm de	dm ds	dm ds	de ds	de ds	dm	de	ds
klasy		stand	unit	stand	unit	stand	unit	st. unit	st. unit	st.unit
4	***r***	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0
	m.w.	0,9	1,0	0,9	0,9	0,9	0,9	1,0	0,9	0,9
3	r	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0
	m.w.	1,0	0,9	0,9	0,9	0,9	1,0	1,0	0,9	0,9
2	***r***	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0
	m.w.	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0

Źródło: obliczenia własne.

4. Omówienie wyników badań w zastosowanym przykładzie

Na zakończenie kilkanaście zdań o wybranym materiale doświadczalnym na podstawie podziału województw na 4 klasy. Zakładając, że wybrane cechy opisujące stan nauki w województwach są symulantami, numery klas wskazują równocześnie kolejność klas województw od najwyższego poziomu stanu nauki (klasa 1) do naj-

słabszego (klasa 4). Poziom ten zmierzyłam, obliczając w każdej klasie średnią cech syntetycznych utworzonych na podstawie standaryzacji i na podstawie unitaryzacji, oddzielnie dla obu wykorzystanych metod klasyfikacji (**Ś.C.** oraz **T.W.**). Wyniki tych obliczeń pokazuje tab. 2. W niej też znajdują się numery województw przydzielonych do poszczególnych klas przez użycie odległości **ds**.

Tabela 2. Stan poziomu nauki w klasach według **(I) Ś.C.** i **(II) T.W.**

(I) Ś.C.				
klasy 4	1	2	3	4
ds stand	6,96	1,72	–0,50	–3,34
nr woj	7	6,11,15	pozostałe	10,13
ds unit	2,86	1,49	0,95	0,25
nr woj	7	6,11,15	pozostałe	10,13
(II) T.W.				
klasy 4	1	2	3	4
ds stand	6,96	3,00	–0,58	–3,72
nr woj	7	6,11,	pozostałe	10
ds unit	2,86	1,39	–0,98	0,25
nr woj	7	6,11	pozostałe	10,13

Źródło: obliczenia własne.

Najwyższy poziom stanu **nauki** obserwujmy w województwie mazowieckim (7). Jest to w obu metodach klasa jednoelementowa. Drugie miejsce w poziomie stanu **nauki** przy zastosowaniu metody **Ś.C.** zajmują województwa małopolskie, pomorskie oraz wielkopolskie, a metoda **T.W.** tworzy tu grupę dwuelementową złożoną z województw małopolskiego i pomorskiego dla **ds unit** oraz dla **ds stand**. Najniższy poziom stanu **nauki** obserwujemy w klasie 4, gdzie sytuacja przedstawia się następująco: dwuelementowe grupy województw podlaskiego i świętokrzyskiego występują w obu metodach dla **ds unit** oraz w metodzie **Ś.C.** dla **ds stand**, natomiast w metodzie **T.W.** powstała klasa jednoelementowa reprezentująca województwo podlaskie dla **ds stand**. W klasie 3 zajmującej trzecie miejsce w rankingu poziomu rozwoju **nauki** mamy przy zastosowaniu metody **Ś.C.** województwa: dolnośląskie, kujawsko-pomorskie, lubelskie, lubuskie, łódzkie, opolskie, podkarpackie, śląskie, warmińsko-mazurskie oraz zachodniopomorskie. Zastosowanie metody **T.W.** dodaje do wyżej wymienionej listy jeszcze województwo wielkopolskie dla **ds unit** i dodatkowo świętokrzyskie dla **ds stand**.

Literatura

Bartosiewicz S., *Opowieść o skutkach subiektywizmu w analizie wielowymiarowej,* [w:] Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu, Taksonomia 18, *Klasyfikacja i analiza danych – teoria i zastosowania*, s. 17-20.

„Rocznik Statystyczny" 2008.

THE EFFECTS OF SUBJECTIVISM IN MULTIVARIATE ANALYSIS REVISITED

Summary: This article provides the completion of the in-press article *The story about consequences of subjectivism in multivariate analysis*, by delivering information on the effects of subjectivism in choosing the method of classification of complex objects. The obvious effect of subjectivism is the difference in results of the classification. Conducted research brings the following conclusions: 1) the difference in classification results depends on the number of clusters: the more the clusters the bigger the difference; 2) the effects of subjectivism are the most apparent in case of choosing the method of classification (Wroclaw numerical taxonomy tends to differentiate the results less than K-Means Method), less noticeable are the effects of subjectivism in choosing the type of distance measure, finally – occurring rarely than in case of choosing the distance measure – the effects of subjectivism when choosing the method of normalization of the characteristics of complex objects.

Keywords: *k*-Means Method and Wroclaw numerical taxonomy, distance measures, normalization.