

# PRACE NAUKOWE

Uniwersytetu Ekonomicznego we Wrocławiu

# RESEARCH PAPERS

of Wrocław University of Economics

Nr 327

**Taksonomia 22**

**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](http://www.ibuk.pl), [www.ebscohost.com](http://www.ebscohost.com),

w Dolnośląskiej Bibliotece Cyfrowej [www.dbc.wroc.pl](http://www.dbc.wroc.pl),

The Central and Eastern European Online Library [www.ceeol.com](http://www.ceeol.com),

a także w adnotowanej bibliografii zagadnień ekonomicznych BazEkon

[http://kangur.uek.krakow.pl/bazy\\_ae/bazekon/nowy/index.php](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](http://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

<b>Wstęp</b> .....	9
<b>Eugeniusz Gatnar</b> , Balance of payments statistics and external competitiveness of Poland.....	15
<b>Andrzej Sokolowski, Magdalena Czaja</b> , Efektywność metody $k$ -średnich w zależności od separowalności grup.....	23
<b>Barbara Pawelek, Józef Pocięcha, Adam Sagan</b> , Wielosektorowa analiza ukrytych przejść w modelowaniu zagrożenia upadłością polskich przedsiębiorstw .....	30
<b>Elżbieta Gołata</b> , Zróżnicowanie procesu starzenia i struktur demograficznych w Poznaniu i aglomeracji poznańskiej na tle wybranych dużych miast Polski w latach 2002-2011.....	39
<b>Aleksandra Łuczak, Feliks Wysocki</b> , Ustalanie systemu wag dla cech w zagadnieniach porządkowania liniowego obiektów .....	49
<b>Marek Walesiak</b> , Wzmacnianie skali pomiaru dla danych porządkowych w statystycznej analizie wielowymiarowej .....	60
<b>Paweł Lula</b> , Identyfikacja słów i fraz kluczowych w tekstach polskojęzycznych za pomocą algorytmu <i>RAKE</i> .....	69
<b>Mariusz Kubus</b> , Propozycja modyfikacji metody złagodzonego LASSO.....	77
<b>Andrzej Bąk, Tomasz Bartłomowicz</b> , Wielomianowe modele logitowe wyborów dyskretnych i ich implementacja w pakiecie <i>DiscreteChoice</i> programu R.....	85
<b>Justyna Brzezińska</b> , Wykorzystanie modeli logarytmiczno-liniowych do analizy bezrobocia w Polsce w latach 2004-2012.....	95
<b>Andrzej Bąk, Marcin Pelka, Aneta Rybicka</b> , Zastosowanie pakietu <i>dcMNM</i> programu R w badaniach preferencji konsumentów wódki .....	104
<b>Barbara Batóg, Jacek Batóg</b> , Analiza stabilności klasyfikacji polskich województw według sektorowej wydajności pracy w latach 2002-2010 .....	113
<b>Małgorzata Markowska, Danuta Strahl</b> , Klasyfikacja europejskiej przestrzeni regionalnej ze względu na filary inteligentnego rozwoju z wykorzystaniem referencyjnego systemu granicznego.....	121
<b>Kamila Migdał-Najman, Krzysztof Najman</b> , Formalna ocena jakości odwzorowania struktury grupowej na mapie Kohonena .....	131
<b>Kamila Migdał-Najman, Krzysztof Najman</b> , Graficzna ocena jakości odwzorowania struktury grupowej na mapie Kohonena .....	139
<b>Beata Basiura, Anna Czapkiewicz</b> , Badanie jakości klasyfikacji szeregów czasowych .....	148
<b>Michał Trzęsiok</b> , Wybrane metody identyfikacji obserwacji oddalonych.....	157

<b>Grażyna Dehnel, Tomasz Klimanek</b> , Taksonomiczne aspekty estymacji pośredniej uwzględniającej autokorelację przestrzenną w statystyce gospodarczej.....	167
<b>Michał Bernard Pietrzak, Justyna Wilk</b> , Odległość ekonomiczna w modelowaniu zjawisk przestrzennych z wykorzystaniem modelu grawitacji.....	177
<b>Maciej Beręsewicz</b> , Próba zastosowania różnych miar odległości w uogólnionym estymatorze Petersena .....	186
<b>Marcin Szymkowiak, Tomasz Józefowski</b> , Konstrukcja i praktyczne wykorzystanie estymatorów typu SPREE na przykładzie dwuwymiarowych tabel kontyngencji .....	195
<b>Marcin Pelka</b> , Klasyfikacja pojęciowa danych symbolicznych w podejściu wielomodelowym .....	202
<b>Małgorzata Machowska-Szewczyk</b> , Ocena klas w rozmytej klasyfikacji obiektów symbolicznych.....	210
<b>Justyna Wilk</b> , Problem wyboru liczby klas w taksonomicznej analizie danych symbolicznych.....	220
<b>Andrzej Dudek</b> , Metody analizy skupień w klasyfikacji markerów map Google .....	229
<b>Ewa Roszkowska</b> , Ocena ofert negocjacyjnych w słabo ustrukturyzowanych problemach negocjacyjnych z wykorzystaniem rozmytej procedury SAW	237
<b>Marcin Szymkowiak, Marek Witkowski</b> , Zastosowanie analizy korespondencji do badania kondycji finansowej banków spółdzielczych.....	248
<b>Bartłomiej Jefmański</b> , Budowa rozmytych indeksów satysfakcji klientów z zastosowaniem programu R.....	257
<b>Karolina Bartos</b> , Odkrywanie wzorców zachowań konsumentów za pomocą analizy koszykowej danych transakcyjnych .....	266
<b>Joanna Trzęsiok</b> , Taksonomiczna analiza krajów pod względem dzietności kobiet oraz innych czynników demograficznych .....	275
<b>Beata Bal-Domańska</b> , Próba identyfikacji większych skupisk regionalnych oraz ich konwergencja.....	285
<b>Beata Bieszk-Stolorz, Iwona Markowicz</b> , Wpływ zasiłku na proces poszukiwania pracy .....	294
<b>Marta Dziechciarz-Duda, Klaudia Przybysz</b> , Wykształcenie a potrzeby rynku pracy. Klasyfikacja absolwentów wyższych uczelni.....	303
<b>Tomasz Klimanek</b> , Problem pomiaru procesu dezagrarnizacji wsi polskiej w świetle wielowymiarowych metod statystycznych.....	313
<b>Małgorzata Sej-Kolasa, Mirosława Sztemberg-Lewandowska</b> , Wybrane metody analizy danych wzdluznych.....	321
<b>Artur Zaborski</b> , Zastosowanie miar odległości dla danych porządkowych do agregacji preferencji indywidualnych .....	330
<b>Mariola Chrzanowska, Nina Drejerska, Iwona Pomianek</b> , Zastosowanie analizy korespondencji do badania sytuacji mieszkańców strefy podmiejskiej Warszawy na rynku pracy.....	338

<b>Katarzyna Wawrzyniak</b> , Klasyfikacja województw według stopnia realizacji priorytetów Strategii Rozwoju Kraju 2007-2015 z wykorzystaniem wartości centrum wierszowego .....	346
---	-----

## Summaries

<b>Eugeniusz Gatnar</b> , Statystyka bilansu płatniczego a konkurencyjność gospodarki Polski .....	22
<b>Andrzej Sokółowski, Magdalena Czaja</b> , Cluster separability and the effectiveness of $k$ -means method .....	29
<b>Barbara Pawelek, Józef Pocięcha, Adam Sagan</b> , Multisectoral analysis of latent transitions in bankruptcy prediction models.....	38
<b>Elżbieta Golata</b> , Differences in the process of aging and demographic structures in Poznań and the agglomeration compared to selected Polish cities in the years 2002-2011 .....	48
<b>Aleksandra Łuczak, Feliks Wysocki</b> , Determination of weights for features in problems of linear ordering of objects .....	59
<b>Marek Walesiak</b> , Reinforcing measurement scale for ordinal data in multivariate statistical analysis .....	68
<b>Paweł Lula</b> , Automatic identification of keywords and keyphrases in documents written in Polish.....	76
<b>Mariusz Kubus</b> , The proposition of modification of the relaxed LASSO method.....	84
<b>Andrzej Bąk, Tomasz Bartłomowicz</b> , Microeconomic multinomial logit models and their implementation in the <code>DiscreteChoice</code> R package .	94
<b>Justyna Brzezińska</b> , The analysis of unemployment data in Poland in 2004-2012 with application of log-linear models .....	103
<b>Andrzej Bąk, Marcin Pelka, Aneta Rybicka</b> , Application of the MMLM package of R software for vodka consumers preference analysis.....	112
<b>Barbara Batóg, Jacek Batóg</b> , Analysis of the stability of classification of Polish voivodeships in 2002-2010 according to the sectoral labour productivity .....	120
<b>Małgorzata Markowska, Danuta Strahl</b> , Classification of the European regional space in terms of smart growth pillars using the reference limit system.....	130
<b>Kamila Migdał Najman, Krzysztof Najman</b> , Formal quality assessment of group structure mapping on the Kohonen's map .....	138
<b>Kamila Migdał Najman, Krzysztof Najman</b> , Graphical quality assessment of group structure mapping on the Kohonen's map .....	147
<b>Beata Basiura, Anna Czapkiewicz</b> , Validation of time series clustering .....	156
<b>Michał Trzęsiok</b> , Selected methods for outlier detection.....	166

<b>Grażyna Dehnel, Tomasz Klimanek</b> , Taxonomic aspects of indirect estimation accounting for spatial correlation in enterprise statistics .....	176
<b>Michał Bernard Pietrzak, Justyna Wilk</b> , Economic distance in modeling spatial phenomena with the application of gravity model.....	185
<b>Maciej Beręsewicz</b> , An attempt to use different distance measures in the Generalized Petersen estimator .....	194
<b>Marcin Szymkowiak, Tomasz Józefowski</b> , Construction and practical using of SPREE estimators for two-dimensional contingency tables.....	201
<b>Marcin Pelka</b> , The ensemble conceptual clustering for symbolic data.....	209
<b>Małgorzata Machowska-Szewczyk</b> , Evaluation of clusters obtained by fuzzy classification methods for symbolic objects.....	219
<b>Justyna Wilk</b> , Problem of determining the number of clusters in taxonomic analysis of symbolic data .....	228
<b>Andrzej Dudek</b> , Clustering techniques for Google maps markers.....	236
<b>Ewa Roszkowska</b> , The evaluation of negotiation offers in ill structure negotiation problems with the application of fuzzy SAW procedure .....	247
<b>Marcin Szymkowiak, Marek Witkowski</b> , The use of correspondence analysis in analysing the financial situation of cooperative banks.....	256
<b>Bartłomiej Jefmański</b> , The construction of fuzzy customer satisfaction indexes using R program.....	265
<b>Karolina Bartos</b> , Discovering patterns of consumer behaviour by market basket analysis of the transactional data.....	274
<b>Joanna Trzęsiok</b> , Cluster analysis of countries with respect to fertility rate and other demographic factors .....	284
<b>Beata Bal-Domańska</b> , An attempt to identify major regional clusters and their convergence .....	293
<b>Beata Bieszk-Stolorz, Iwona Markowicz</b> , The influence of benefit on the job finding process .....	302
<b>Marta Dziechciarz-Duda, Klaudia Przybysz</b> , Education and labor market needs. Classification of university graduates .....	312
<b>Tomasz Klimanek</b> , The problem of measuring deagrarianisation process in rural areas in Poland using multivariate statistical methods.....	320
<b>Małgorzata Sej-Kolasa, Mirosława Sztemberg-Lewandowska</b> , Selected methods for an analysis of longitudinal data.....	329
<b>Artur Zaborski</b> , The application of distance measures for ordinal data for aggregation individual preferences .....	337
<b>Mariola Chrzanowska, Nina Drejerska, Iwona Pomianek</b> , Application of correspondence analysis to examine the situation of the inhabitants of Warsaw suburban area in the labour market .....	345
<b>Katarzyna Wawrzyniak</b> , Classification of voivodeships according to the level of the realization of priorities of <i>the National Development Strategy 2007-2015</i> with using the values of centroid of the rows .....	355

**Andrzej Sokołowski, Magdalena Czaja**

Uniwersytet Ekonomiczny w Krakowie

---

## **EFEKTYWNOŚĆ METODY $k$ -ŚREDNICH W ZALEŻNOŚCI OD SEPAROWALNOŚCI GRUP**

---

**Streszczenie:** W pracy przedstawiono wybrane wyniki badań symulacyjnych dotyczących efektywności metody  $k$ -średnich, mierzonej procentem poprawnie zaklasyfikowanych obserwacji w zależności od separowalności grup, błędnej specyfikacji liczby skupień oraz obecności obserwacji odstających. Znalaziono analityczną postać funkcji opisującej efektywność metody  $k$ -średnich przy poprawnym ustaleniu liczby grup.

**Słowa kluczowe:** analiza skupień, metoda  $k$ -średnich, efektywność.

### **1. Wstęp**

Metoda  $k$ -średnich jest jedną z najpopularniejszych metod taksonomicznych. Nie wiele osób, nawet w Polsce, wie, że jej zasady jako pierwszy zaproponował Hugo Steinhaus [1956]. W 1957 r. metodę opisał S. Lloyd w wewnętrznym opracowaniu dla Bell Laboratories. Opublikował ją dopiero w 1982 r. Lloyd [Lloyd 1982]. Najczęściej nazwa metody:  $k$ -średnich kojarzona jest z Jamesem MacQueenem [MacQueen 1967]. Istnieje wiele wersji i modyfikacji metody  $k$ -średnich. Dobry przegląd historyczny zawierają prace [Bock 2007 oraz Jain 2010].

Standardowy algorytm metody  $k$ -średnich obejmuje następujące kroki (przy zadanym  $k$ ):

- 1) wybór wstępnych środków skupień,
- 2) przyporządkowanie każdego obiektu do najbliższego środka,
- 3) wyznaczenie nowych środków skupień,
- 4) powrót do punktu (2).

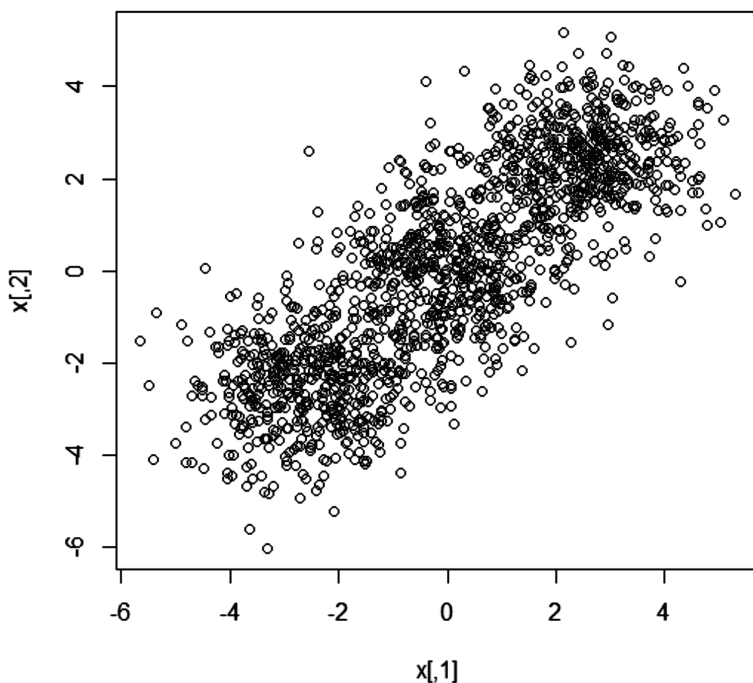
Przesuwanie punktów kończy się, gdy każdy punkt jest bliżej środka własnej grupy niż jakiegokolwiek innej.

## 2. Cel pracy

Jednym z celów niniejszej pracy była ocena efektywności metody  $k$ -średnich w zależności od odległości skupień. Jest oczywiste, że metoda tym trafniej będzie rozpoznawała obserwacje i dzieliła zbiór, im dalej od siebie położone są środki grup. Dlatego naszym zadaniem było poszukiwanie postaci analitycznej funkcji, która będzie dobrze opisywała zmianę efektywności metody w zależności od separowalności grup. Efektywność jest tu mierzona procentem prawidłowo zidentyfikowanych obiektów. W pracy przedstawiono też niektóre wyniki badań wpływu obserwacji odstających na zachowanie się metody  $k$ -średnich oraz zakłócenia efektywności wynikające z niewłaściwej identyfikacji liczby grup.

## 3. Badania symulacyjne

Model symulacyjny przewidywał analizy w przestrzeni dwuwymiarowej dla zadanej liczby trzech grup. Obserwacje z tych grup generowane są przez dwuwymiarowe rozkłady normalne o niezależnych składowych. Rozważano rozkłady o brzegowych odchyleniach standardowych 1 oraz 2. Grupy oddalają się od siebie wzdłuż prostej  $y = x$ , a odległość między środkami grup oznaczona jest przez  $\alpha$ .

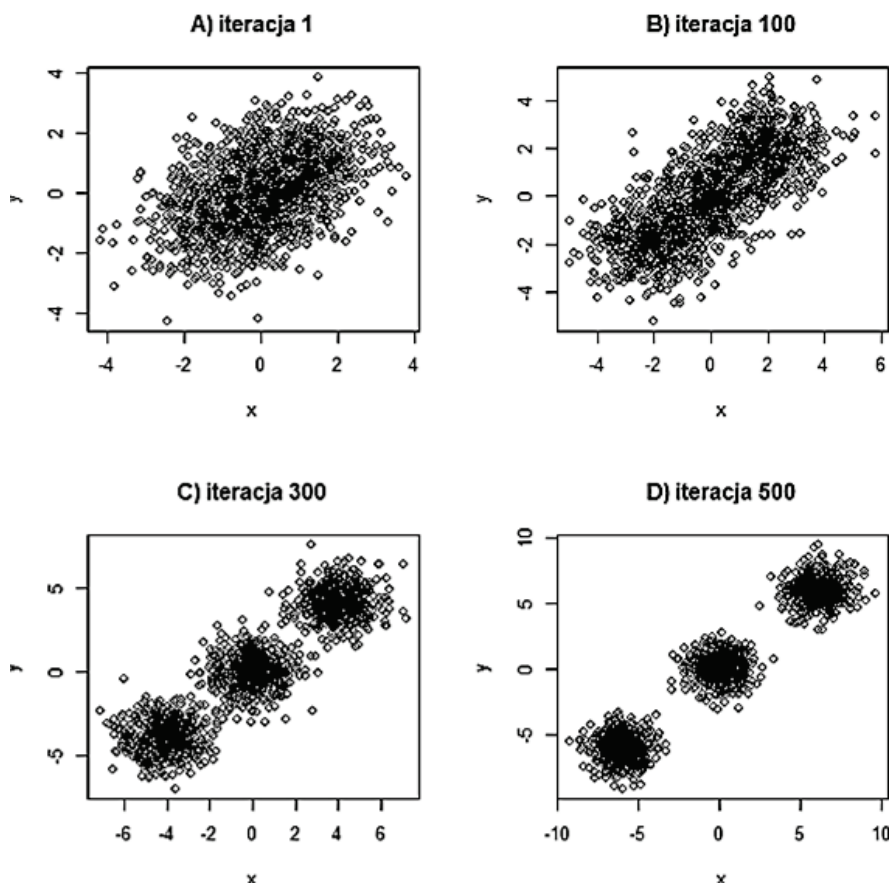


Rys. 1. Rozrzut punktów dla  $k = 3$  oraz  $\alpha = 2,5$

Źródło: opracowanie własne.



Generowano próby o liczebnościach 500 oraz 1000. W badaniach wykorzystano program napisany specjalnie w tym celu w środowisku R. Wstępne środki grup wybierane były w sposób losowy, a jako miarę odległości wzięto odległość euklidesową. Efektywność była mierzona frakcją obiektów poprawnie zakwalifikowanych. Przykładowy rozrzut punktów dla  $k = 3$ ,  $\alpha = 2,5$  przedstawiono na rys. 1. Efektywność wyniosła w tym przypadku 0,942.



Rys. 2. Oddalanie się grup w modelu symulacyjnym

Źródło: obliczenia własne.

W analizach symulacyjnych badano zmiany efektywności związane ze stopniowym oddalaniem się grup od siebie, z niewielkim skokiem, równym 0,01 (pojedynczy skok to tzw. iteracja). Środkowa grupa ma obydwie wartości przeciętne równe zeru. Pierwsza grupa ma wartości przeciętne równe  $\mu_1 = -1 - 0,01 \cdot j$ , gdzie  $j$  jest numerem iteracji. Wartości przeciętne trzeciej grupy są równe  $\mu_3 = +1 + 0,01 \cdot j$ . Oddalanie się grup ilustruje rys. 2.

#### 4. Wyniki analizy efektywności metody $k$ -średnich

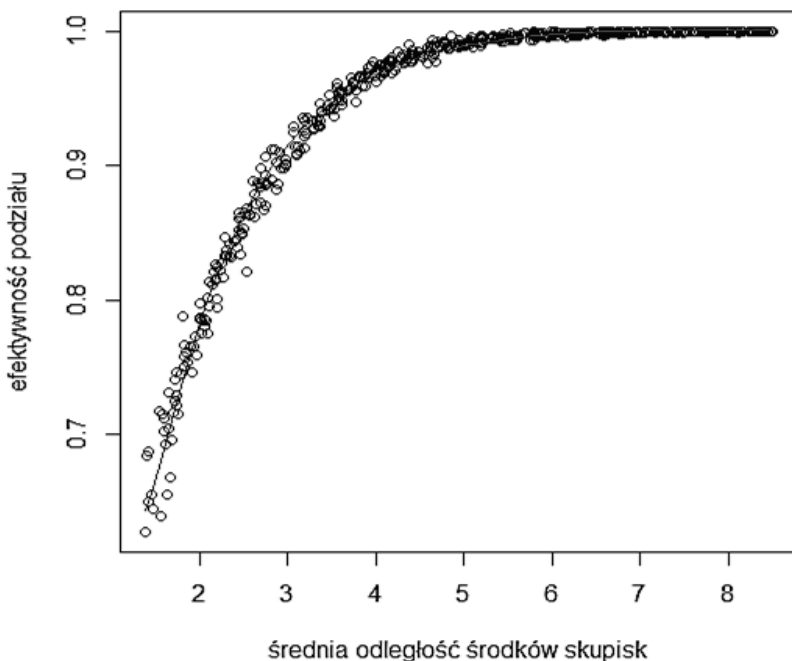
Na drodze analizy wyników symulacyjnych badań efektywności oraz prób dopasowywania różnych funkcji analitycznych stwierdzono, że efektywność metody  $k$ -średnich da się bardzo dobrze opisać za pomocą funkcji (przy założeniu prawidłowej specyfikacji liczby grup):

$$E = \beta_0 + \beta_1\alpha^{-1} + \beta_2\alpha^{-2} + \beta_3\alpha^{-3}.$$

Wykorzystując wyniki symulacji, oszacowano parametry tej funkcji metodą najmniejszych kwadratów, otrzymując:

$$\hat{E} = 0,953729 + 0,785544\alpha^{-1} - 3,568560\alpha^{-2} + 2,609153\alpha^{-3}.$$

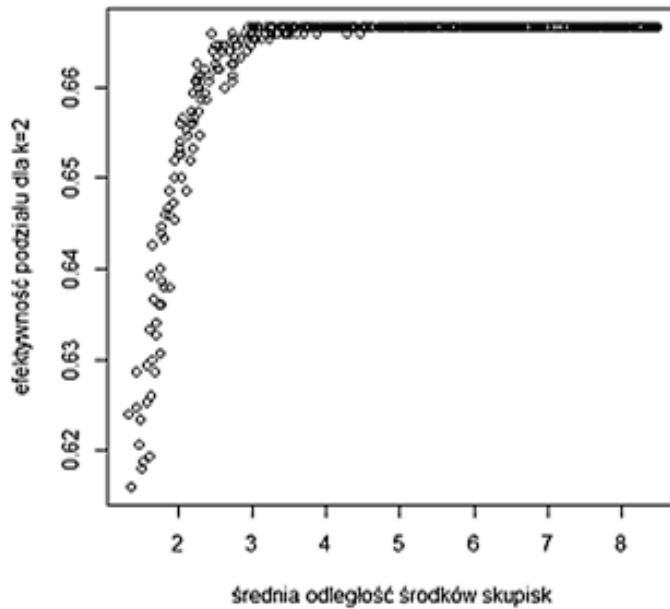
Parametry tej funkcji są wysoce istotne statystycznie (co stwierdzono testem istotności parametrów strukturalnych modelu regresji wielorakiej, wykorzystującym statystykę  $t$ -Studenta), a jej przebieg ilustruje rys. 3.



**Rys. 3.** Efektywność metody  $k$ -średnich w zależności od odległości skupisk

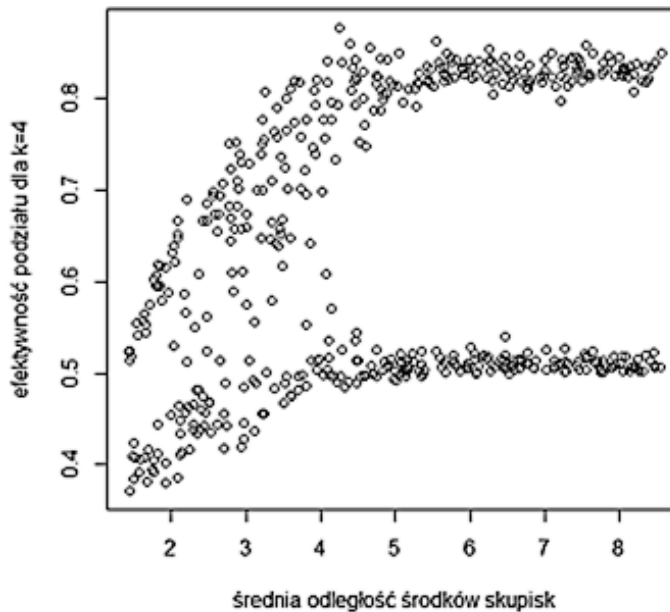
Źródło: obliczenia własne.

Wykonano szereg analiz symulacyjnych, pozwalających na ocenę efektywności metody  $k$ -średnich przy niewłaściwym ustaleniu liczby skupisk oraz w przypadku obecności obserwacji odstających. Tutaj prezentujemy tylko wybrane, typowe wy-



Rys. 4. Efektywność metody  $k$ -średnich w przypadku przyjęcia  $k = 2$  zamiast  $k = 3$

Źródło: obliczenia własne.

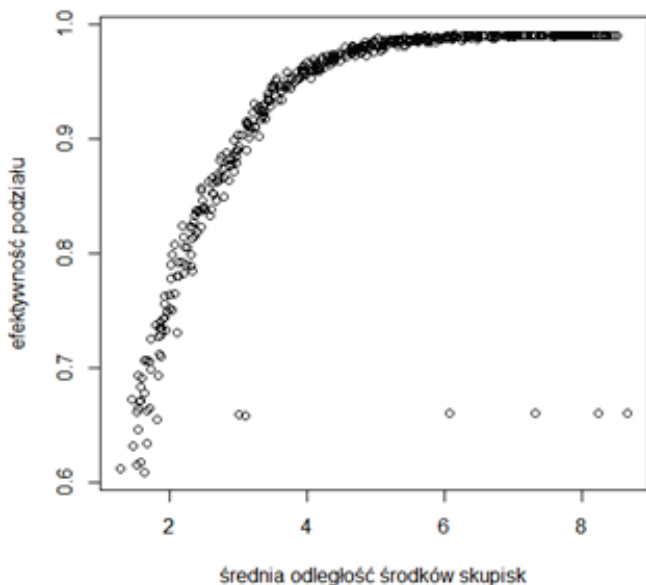


Rys. 5. Efektywność metody  $k$ -średnich przy błędnym podaniu  $k = 4$  zamiast  $k = 3$

Źródło: obliczenia własne.

niki. Jeżeli błędnie przyjęto  $k = 2$  (zamiast  $k = 3$ ), to efektywność nieznacznie tylko przekracza  $2/3$ , gdyż jedna grupa (środkowa) jest sztucznie rozdzielana na dwie części (rys. 4).

Przy podaniu zbyt dużej liczby skupisk jedno z nich jest sztucznie rozdzielane, a efektywność (przy przyjęciu prawidłowej identyfikacji jako tej, która dobrze przyporządkowuje kod grupy) niejako rozdziela się i nie poprawia nawet przy dużej separowalności (rys. 5).



**Rys. 6.** Efektywność metody  $k$ -średnich w przypadku obecności 1% punktów odstających

Źródło: obliczenia własne.

Ciekawe, lecz raczej spodziewane wyniki przyniosły analizy efektywności przy obecności obserwacji odstających. W eksperymencie symulacyjnym zamieniano 1% losowo wybranych punktów na punkty o współrzędnych (15, 15). Jeżeli w przebiegu metody  $k$ -średnich taki punkt odstający zostaje wylosowany jako wstępny środek skupienia, wówczas metoda nie jest w stanie go już opuścić. Dla pozostałych punktów brakuje więc jednego środka i efektywność nie przekracza  $2/3$ . Gdy punkt odstający nie jest wybierany jako wstępny środek, to jest przyporządkowywany do trzeciej grupy, a punkty odstające obniżają tylko graniczną, maksymalną efektywność.

## 5. Podsumowanie

Badania symulacyjne wykazały oczywiste i spodziewane zjawisko poprawiania się efektywności metody  $k$ -średnich w miarę wzrostu separowalności grup. Osiągnięciem pracy jest zidentyfikowanie analitycznej postaci funkcji (i oszacowanie jej parametrów), która bardzo dobrze opisuje zmiany tej efektywności. Prawdopodobieństwo zakłócenia poprawności działania metody  $k$ -średnich przez obserwacje odstające wzrasta wraz ze wzrostem udziału tych obserwacji w ogólnej liczbie klasyfikowanych obiektów, gdyż wówczas wzrasta prawdopodobieństwo tego, że taki punkt izolowany zostanie wylosowany jako wstępny środek grupy. Eksperymenty symulacyjne potwierdziły też kluczowe znaczenie właściwego wyboru liczby grup.

## Literatura

- Bock H.-H. (2007), *Clustering methods: A history of  $k$ -means algorithms*, [w:] *Selected Contributions in Data Analysis and Classification*, Springer, Berlin – Heidelberg, s. 161-172.
- Jain A.K. (2010), *Data clustering: 50 years beyond  $k$ -means*, Pattern Recognition Letters, 31 sierpnia, s. 651-666.
- Lloyd S. (1982), *Least squares quantization in PCM*, IEEE Trans. Inform. Theory, 28, s. 129-137.
- MacQueen J. (1967), *Some methods for classification and analysis of multivariate observations*, Fifth Berkeley Symposium on Mathematics, Statistics and Probability. University California Press, s. 281-297.
- Steinhaus H. (1956), *Sur la division des corps materiel en parties*, Bull. Acad. Polon. Sci. Cl. III. 4, s. 801-804.

### CLUSTER SEPARABILITY AND THE EFFECTIVENESS OF $K$ -MEANS METHOD

**Summary:** Selected results of simulation analysis on  $k$ -means method effectiveness is presented in the paper. The effectiveness is measured by the percentage of correctly identified observations. The effectiveness has been studied depending on group separability, wrongly identified number of clusters and the presence of outliers. The analytical function describing the effectiveness has been found and estimated.

**Keywords:** cluster analysis,  $k$ -means, effectiveness.