PRACE NAUKOWE
Uniwersytetu Ekonomicznego we Wrocławiu

RESEARCH PAPERS
of Wrocław University of Economics

Nr 426

Taksonomia 26
Klasyfikacja i analiza danych – teoria i zastosowania

Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2016

Redaktor Wydawnictwa: Agnieszka Flasińska

Redaktor techniczny: Barbara Łopusiewicz

Korektor: Barbara Cibis

Łamanie: Beata Mazur

Projekt okładki: Beata Dębska

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

Informacje o naborze artykułów i zasadach recenzowania
znajdują się na stronach internetowych
www.pracenaukowe.ue.wroc.pl
www.wydawnictwo.ue.wroc.pl

ISSN 1899-3192 (Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu)
e-ISSN 2392-0041
ISSN 1505-9332 (Taksonomia)

Wersja pierwotna: publikacja drukowana

Zamówienia na opublikowane prace należy składać na adres:
Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
ul. Komandorska 118/120, 53-345 Wrocław
tel./fax 71 36 80 602; e-mail:econbook@ue.wroc.pl
www.ksiegarnia.ue.wroc.pl

Druk i oprawa: TOTEM

Spis treści

Wstęp

W dniach 14–16 września 2015 r. w Hotelu Novotel Gdańsk Marina w Gdańsku odbyła się XXIV Konferencja Naukowa Sekcji Klasyfikacji i Analizy Danych PTS (XXIX Konferencja Taksonomiczna) „Klasyfikacja i analiza danych – teoria i zastosowania", zorganizowana przez Sekcję Klasyfikacji i Analizy Danych Polskiego Towarzystwa Statystycznego oraz Katedrę Statystyki Wydziału Zarządzania Uniwersytetu Gdańskiego. Przewodniczącymi Komitetu Organizacyjnego konferencji byli prof. dr hab. Mirosław Szreder oraz dr hab. Krzysztof Najman, prof. nadzw. UG, sekretarzami naukowymi dr hab. Kamila Migdał-Najman, prof. nadzw. UG oraz dr hab. Anna Zamojska, prof. nadzw. UG, a sekretarzem organizacyjnym Anna Nowicka z Fundacji Rozwoju Uniwersytetu Gdańskiego.

Konferencja Naukowa została dofinansowana ze środków Narodowego Banku Polskiego.

Zakres tematyczny konferencji obejmował takie zagadnienia, jak:

a) teoria (taksonomia, analiza dyskryminacyjna, metody porządkowania liniowego, metody statystycznej analizy wielowymiarowej, metody analizy zmiennych ciągłych, metody analizy zmiennych dyskretnych, metody analizy danych symbolicznych, metody graficzne),

b) zastosowania (analiza danych finansowych, analiza danych marketingowych, analiza danych przestrzennych, inne zastosowania analizy danych – medycyna, psychologia, archeologia, itd., aplikacje komputerowe metod statystycznych).

Zasadniczymi celami konferencji SKAD były prezentacja osiągnięć i wymiana doświadczeń z zakresu teoretycznych i aplikacyjnych zagadnień klasyfikacji i analizy danych. Konferencja stanowi coroczne forum służące podsumowaniu obecnego stanu wiedzy, przedstawieniu i promocji dokonań nowatorskich oraz wskazaniu kierunków dalszych prac i badań.

W konferencji wzięło udział 81 osób. Byli to pracownicy oraz doktoranci następujących uczelni i instytucji: AGH w Krakowie, Politechniki Łódzkiej, Politechniki Gdańskiej, Politechniki Opolskiej, Politechniki Wrocławskiej, Szkoły Głównej Gospodarstwa Wiejskiego w Warszawie, Szkoły Głównej Handlowej w Warszawie, Uniwersytetu im. Adama Mickiewicza w Poznaniu, Uniwersytetu Ekonomicznego w Katowicach, Uniwersytetu Ekonomicznego w Krakowie, Uniwersytetu Ekonomicznego w Poznaniu, Uniwersytetu Ekonomicznego we Wrocławiu, Uniwersytetu Gdańskiego, Uniwersytetu Jana Kochanowskiego w Kielcach, Uniwersytetu Łódzkiego, Uniwersytetu Mikołaja Kopernika w Toruniu, Uniwersytetu Przyrodniczego w Poznaniu, Uniwersytetu Szczecińskiego, Uniwer-

sytetu w Białymstoku, Wyższej Szkoły Bankowej w Toruniu, a także przedstawiciele NBP i PBS Sp. z o.o.

W trakcie dwóch sesji plenarnych oraz trzynastu sesji równoległych wygłoszono 58 referatów poświęconych aspektom teoretycznym i aplikacyjnym zagadnienia klasyfikacji i analizy danych. Odbyła się również sesja plakatowa, na której zaprezentowano 14 plakatów. Obradom w poszczególnych sesjach konferencji przewodniczyli profesorowie: Józef Pociecha, Eugeniusz Gatnar, Tadeusz Trzaskalik, Krzysztof Jajuga, Marek Walesiak, Barbara Pawełek, Feliks Wysocki, Ewa Roszkowska, Andrzej Sokołowski, Andrzej Bąk, Tadeusz Kufel, Mirosław Krzyśko, Krzysztof Najman, Małgorzata Rószkiewicz, Mirosław Szreder.

Teksty 25 recenzowanych artykułów naukowych stanowią zawartość prezentowanej publikacji z serii „Taksonomia" nr 26. Pozostałe recenzowane artykuły znajdują się w „Taksonomii" nr 27.

W pierwszym dniu konferencji odbyło się posiedzenie członków Sekcji Klasyfikacji i Analizy Danych Polskiego Towarzystwa Statystycznego, któremu przewodniczył prof. dr hab. Józef Pociecha. Ustalono plan przebiegu zebrania obejmujący następujące punkty:

A. Sprawozdanie z działalności Sekcji Klasyfikacji i Analizy Danych PTS.
B. Informacje dotyczące planowanych konferencji krajowych i zagranicznych.
C. Organizacja konferencji SKAD PTS w latach 2016 i 2017.
D. Wybór przedstawiciela Rady Sekcji SKAD PTS do IFCS.
E. Dyskusja nad kierunkami rozwoju działalności Sekcji.

Prof. dr hab. Józef Pociecha otworzył posiedzenie Sekcji SKAD PTS. Sprawozdanie z działalności Sekcji Klasyfikacji i Analizy Danych PTS przedstawiła sekretarz naukowy Sekcji dr hab. Barbara Pawełek, prof. nadzw. UEK. Poinformowała, że obecnie Sekcja liczy 231 członków. Przypomniała, że na stronie internetowej Sekcji znajdują się regulamin, a także deklaracja członkowska. Poinformowała, że zostały opublikowane zeszyty z serii „Taksonomia" nr 24 i 25 (PN UE we Wrocławiu nr 384 i 385). W „Przeglądzie Statystycznym" (zeszyt 4/2014) ukazało się sprawozdanie z ubiegłorocznej konferencji SKAD, która odbyła się w Międzyzdrojach, w dniach 8–10 września 2014 r. Prof. Barbara Pawełek przedstawiła także informacje dotyczące działalności międzynarodowej oraz udziału w ważnych konferencjach członków i sympatyków SKAD.

W konferencji Międzynarodowego Stowarzyszenia Towarzystw Klasyfikacyjnych (IFCS – International Federation of Classification Societies) w dniach 6–8 lipca 2015 r. w Bolonii, zorganizowanej przez Università di Bologna, udział wzięło 19 osób z Polski (w tym 17 członków Sekcji), które wygłosiły 15 referatów (wkład członków SKAD – 79,0%). Ponadto prof. Józef Pociecha był członkiem Komitetu Naukowego Konferencji z ramienia SKAD, członkiem Międzynarodowego Komitetu Nagród IFCS oraz organizatorem i przewodniczącym sesji nt. „Classification models for forecasting of economic processes".

W konferencji „European Conference on Data Analysis" (Colchester, 2–4 września 2015 r.) zorganizowanej przez The German Classification Society (GfKl) we współpracy z The British Classification Society (BCS) i Sekcją Klasyfikacji i Analizy Danych PTS (SKAD) udział wzięło 18 osób z Polski (w tym 14 członków Sekcji), które wygłosiły 15 referatów (wkład członków SKAD – 66,0%). Ponadto profesorowie Krzysztof Jajuga oraz Józef Pociecha byli członkami Komitetu Naukowego konferencji, prof. Andrzej Dudek został poproszony przez organizatorów o przygotowanie referatu i wygłoszenie na Sesji Plenarnej „Cluster analysis in XXI century, new methods and tendencies", prof. Krzysztof Jajuga był przewodniczącym sesji plenarnej, przewodniczącym sesji nt. „Finance and economics II" oraz organizatorem i przewodniczącym sesji nt. „Data analysis in finance", prof. Józef Pociecha był organizatorem i przewodniczącym sesji nt. „Outliers in classification procedures – theory and practice", prof. Andrzej Dudek był przewodniczącym sesji nt. „Machine learning and knowledge discovery II".

Kolejny punkt posiedzenia Sekcji obejmował zapowiedzi najbliższych konferencji krajowych i zagranicznych, których tematyka jest zgodna z profilem Sekcji. Prof. dr hab. Józef Pociecha poinformował o dwóch wybranych konferencjach krajowych (były to XXXIV Konferencja Naukowa „Multivariate Statistical Analysis MSA 2015", Łódź, 16–18 listopada 2015 r. i X Międzynarodowa Konferencja Naukowa im. Profesora Aleksandra Zeliasia nt. „Modelowanie i prognozowanie zjawisk społeczno-gospodarczych", Zakopane, 10–13 maja 2016 r.) oraz o trzech wybranych konferencjach zagranicznych. Konferencja „European Conference on Data Analysis" odbędzie się na Uniwersytecie Ekonomicznym we Wrocławiu w dniach 26–28 września 2017 r. W przeddzień tej konferencji, tj. 25.09.2017 r., odbędzie się Niemiecko-Polskie Sympozjum nt. „Analizy danych i jej zastosowań GPSDAA 2017". Następna konferencja Międzynarodowego Stowarzyszenia Towarzystw Klasyfikacyjnych (IFCS) odbędzie się w 2017 r. w Tokio. W 2019 r. Niemiecko-Polskie Sympozjum nt. „Analizy danych i jej zastosowań GPSDAA 2019" organizuje prof. Andreas Geyer-Schultz w Karlsruhe.

W następnym punkcie posiedzenia podjęto kwestię organizacji kolejnych konferencji SKAD. SKAD 2016 zorganizuje Katedra Metod Statystycznych Wydziału Ekonomiczno-Socjologicznego Uniwersytetu Łódzkiego.

W kolejnej części zebrania dokonano wyboru przedstawiciela Rady Sekcji SKAD PTS do IFCS na kadencję 2016–2019. Powołano Komisję Skrutacyjną, której przewodniczącym został prof. Tadeusz Kufel, a członkami dr hab. Iwona Konarzewska i dr Dominik Rozkrut. Profesor Józef Pociecha poprosił zebranych o proponowanie kandydatur zgłaszając jednocześnie prof. Andrzeja Sokołowskiego. Wobec braku następnych kandydatur listę zamknięto. Komisja Skrutacyjna przeprowadziła głosowanie tajne. W głosowaniu uczestniczyło 41 członków Sekcji. Profesor Andrzej Sokołowski został przedstawicielem Rady Sekcji SKAD PTS do

IFCS na kadencję 2016–2019, uzyskując następujący wynik: 39 głosów na „tak", 1 głos na „nie", 1 głos był nieważny.

W ostatnim punkcie zebrania dyskutowano nad kierunkami rozwoju działalności Sekcji obejmującymi następujące problemy: udział w międzynarodowym ruchu naukowym (wspólne granty, publikacje), umiędzynarodowienie konferencji SKAD (uczestnicy zagraniczni, dwujęzyczność konferencji), wydawanie własnego czasopisma.

Profesor Józef Pociecha zamknął posiedzenie Sekcji SKAD.

Krzysztof Jajuga, Marek Walesiak

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS nr 426 • 2016

Taksonomia 26 ISSN 1899-3192
Klasyfikacja i analiza danych – teoria i zastosowania e-ISSN 2392-0041

Jacek Batóg
Uniwersytet Szczeciński
e-mail: batog@wneiz.pl

IDENTYFIKACJA OBSERWACJI ODSTAJĄCYCH W ANALIZIE SKUPIEŃ

INFLUENCE OF OUTLIERS ON RESULTS OF CLUSTER ANALYSIS

DOI: 10.15611/pn.2016.426.01

Streszczenie: W ramach przeprowadzonego badania dokonano analizy porównawczej metod identyfikujących obserwacje odstające w zbiorze danych przestrzennych. Wykorzystano w tym celu metodę *k*-średnich oraz dane charakteryzujące gminy województwa zachodniopomorskiego pod względem poziomu dochodów i zadłużenia. Ocenie poddano wyniki uzyskane za pomocą wybranych metod wykrywania obserwacji odstających typu *false positive*: metody zaproponowanej przez Wanga, Zhanga, Li i Songa, jej autorskiej modyfikacji, metody Kandogana oraz metody *Outlier Removal Clustering*. Jako miarę homogeniczności podziału zastosowano miarę stopnia zróżnicowania obiektów wewnątrz skupiska. Uzyskane rezultaty pozwalają stwierdzić, że wszystkie zastosowane metody generują praktycznie identyczne wyniki. Występujące różnice polegają wyłącznie na odmiennej kolejności wskazywania obserwacji odstających.

Słowa kluczowe: analiza skupień, metoda *k*-średnich, obserwacje odstające.

Summary: The research concerns comparison of methods that enable identifying spatial outliers. The analysis was based on the statistical data describing income and public debt of gminas of Zachodniopomorskie voivodship. All considerations were applied to partitions made by *k*-means method. Identification of *false positive* outliers was provided by means of Wang, Zhang, Li and Song method, author's modification of this method and additionally methods proposed by Kandogan and Hautamäki. The level of objects' differentiation within group was used as a measure of homogeneity of partitions. The received results were very similar for all considered methods. Some differences occur only in order of indicated outliers.

Keywords: Cluster analysis, *k*-means method, outliers.

1. Wstęp

Prawidłowa interpretacja wyników prowadzonych analiz uzależniona jest w dużym stopniu nie tylko od kompletności i braku błędów pomiaru danych statystycznych, lecz także od występowania w zbiorze obserwacji obiektów uznawanych za nietypowe. Obiekty te nazywane są również obserwacjami odstającymi (*outlier)* i często określane też jako: szum, błąd, innowacja, wyjątek, osobliwość, niedoskonałość, zanieczyszczenie czy aberracja [Chandola, Banerjee, Kumar 2009]. Obserwacje odstające mogą być wynikiem rzeczywistego procesu lub błędu pomiaru i często nie są zauważane przez badaczy. Dzieje się tak, ponieważ obecnie duże zbiory obserwacji są przetwarzane za pomocą komputerów, co uniemożliwia ich precyzyjną kontrolę [Rousseeuw, Leroy 1987]. W literaturze występują zróżnicowane definicje pojęcia „obserwacja odstająca". Najczęściej spotykaną jest definicja sformułowana przez D.M. Hawkinsa „Obserwacją odstającą jest obserwacja, która na tyle znacząco odróżnia się od pozostałych obserwacji, że istnieje podejrzenie o odmiennym charakterze procesu, który ją wygenerował" [Hawkins 1980]. Niektórzy autorzy wskazują, że obserwacją odstającą jest obserwacja, która nie tylko znacząco odróżnia się od pozostałych danych, lecz przede wszystkim zakłóca relacje występujące między zmiennymi [Ghosh-Dastidar, Schafer 2006; Batóg, Batóg 2014]. Obserwacje nietypowe są nieco odmiennie definiowane w zagadnieniach grupowania obiektów. Wskazuje się, że są to obserwacje, które nie pasują do ogólnego podziału na klasy [Zhang, Ramakrishan, Livny 1997], obserwacje, które powinny być usunięte, aby podział na jednorodne klasy był bardziej wiarygodny [Guha, Rastogi, Shim 1998] lub obserwacje niebędące elementami skupisk ani szumu pojawiającego się w ich otoczeniu i zachowujące się niezgodnie z pewną normą [Aggarwal, Yu 2001]. Specyficznym rodzajem obserwacji nietypowych są obserwacje o charakterze przestrzennym (*spatial outlier*), czyli obserwacje, które znacząco różnią się od swojego sąsiedztwa, ale niekoniecznie wyróżniają się na tle całej populacji – mają identyczną lokalizację w przestrzeni jak inny podzbiór obiektów, lecz wyróżniają się w stosunku do nich wartością atrybutów pozaprzestrzennych [Shekhar, Lu, Zhang 2003, s. 140]. Obserwacje tego typu określane są często mianem *false-positive*.

W identyfikacji przestrzennych obserwacji odstających stosowane są trzy podstawowe podejścia [Duan i in. 2009, s. 153 i n.]. Pierwsze z nich oparte jest na analizie rozkładów zmiennych (*distribution based*). Drugie podejście wykorzystuje odległości między obiektami (*distance ba*sed). Obiekt p ze zbioru D jest w tym przypadku uznawany za odstający, jeśli co najmniej określony procent obiektów zbioru D leży w odległości dalszej niż $d_{\min}$ od obiektu p[1]. Trzecie podejście

[1] Spotykana jest też interpretacja „mniej obiektów niż" występujących w pewnym promieniu r od obiektu p [Cherednichenko 2005, s. 10 i n.]. Zob. propozycję wykrywania obserwacji nietypowych wykorzystującą odległość Mahalanobisa [Jayakumar, Thomas 2103, s. 72].

uwzględnia gęstość poszczególnych skupień (*density based*). W jego ramach stosowana może być na przykład metoda wykorzystująca lokalną gęstość sąsiednich obiektów, oparta na analizie odległości (*Local Outlier Factor*) lub algorytm wykorzystujący spójność grafu (*outlier detection using indegree number*) (zob. [Hodge, Austin 2004, s. 4 i n.]).

2. Stosowane metody badawcze, cel pracy i hipoteza badawcza

Obserwacje odstające zniekształcają wyniki analizy skupień. Występuje to zwłaszcza w przypadku metod opartych na odległości euklidesowej, takich jak np. metoda *k*-średnich. Ma na to wpływ m.in. fakt, że algorytmy stosowane w analizie skupień są optymalizowane przede wszystkim pod kątem uzyskiwania jednorodnych grup obiektów. Wykrywanie za ich pomocą obserwacji odstających jest więc często utrudnione. Na szczególną uwagę zasługują w związku z tym metody grupowania, które pozwalają łączyć oba te cele. Zaliczają się do nich np. metody wykorzystujące odległości między obiektami, które pozwalają identyfikować obserwacje odstające typu *false-positive*. W przypadku tych metod pierwotny podział na skupiska dokonany za pomocą metody *k*-średnich jest następnie modyfikowany z wykorzystaniem: odległości obiektów od środków ciężkości skupień z wykorzystaniem punktu odniesienia *T* (*threshold*) w postaci z góry ustalonej liczby obiektów nietypowych [Wang i in. 2010, s. 34, 35], odchylenia standardowego odległości od środków ciężkości skupień [Kandogan 2012, s. 76] lub subiektywnego poziomu odniesienia ustalanego w oparciu o miarę nietypowości obliczaną na podstawie maksymalnych odległości od środków ciężkości skupień – *Outlier Removal Clustering* [Hautamäki i in. 2005, s. 981–983].

Kolejne etapy metody zaproponowanej przez H. Wanga, X. Zhanga, S. Li i X. Songa to: ustalenie wyjściowych skupień metodą *k*-średnich, obliczenie w poszczególnych skupieniach odległości obiektów od ich środków ciężkości, uporządkowanie obiektów według malejących odległości dla wszystkich skupisk łącznie, identyfikacji obiektów odstających, które stanowi 5% ogólnej liczby obserwacji charakteryzujących się największymi odległościami, ponowne zastosowanie metody *k*-średnich i powtórzenie poprzednich kroków. Niestety autorzy metody nie zaproponowali reguły stopu pozwalającej na ustalenie momentu, w którym uzyskane podzbiory obiektów charakteryzują się zadowalającą jednorodnością.

W metodzie Kandogana po dokonaniu podziału na skupienia metodą *k*-średnich obliczane jest dla każdego skupienia odchylenie standardowe odległości obiektów od środków ciężkości, a za obserwacje odstające przyjmuje się obiekty, dla których ich odległość od środka ciężkości danego skupienia jest większa niż wielokrotność odchylenia standardowego odległości – z reguły przyjmuje się trzykrotną wartość tej miary.

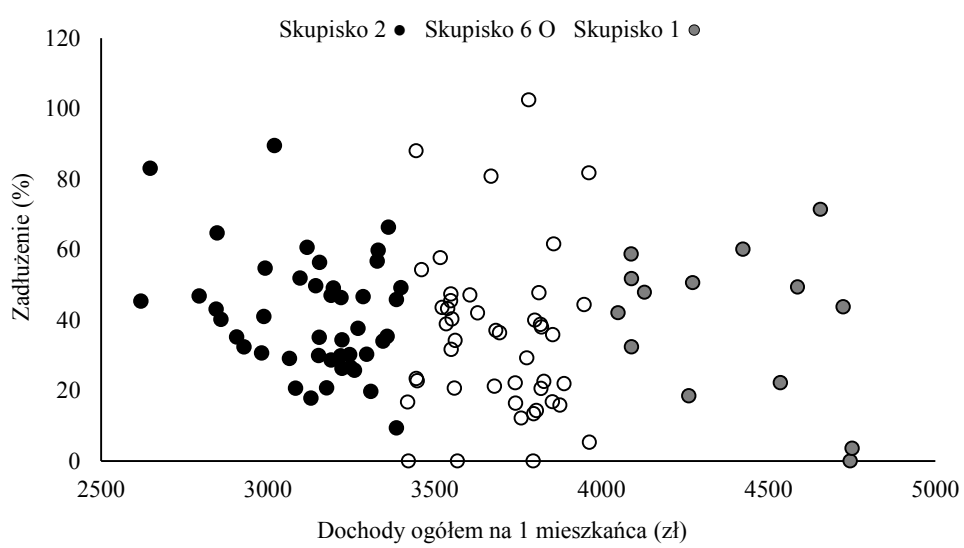
W przypadku metody *Outlier Removal Clustering* po uzyskaniu wstępnego podziału na skupienia metodą *k*-średnich znajdujemy maksymalną odległość od dowolnego środka ciężkości $d_{\max}$. Następnie dla każdego obiektu obliczana jest miara

nietypowości o_i, będącą stosunkiem odległości indywidualnej do odległości maksymalnej w danym skupieniu. W kolejnym kroku ustalany jest punkt odniesienia $T < 1$, a za obserwacje odstające uznawane są wszystkie, dla których wartości o_i są większe od T. Po ich usunięciu ponawiamy podział pozostałych obiektów na skupiska oraz powtarzamy wszystkie kroki, aż do momentu, w którym nie wystąpią już obserwacje uznawane za odstające.

Głównym celem przeprowadzonego badania jest analiza wyników identyfikacji obserwacji odstających uzyskanych dwoma wariantami metody pierwotnie zaproponowanej przez H. Wanga, X. Zhanga, S. Li i X. Songa. Rezultaty wykorzystania oryginalnej postaci algorytmu porównane zostaną z jego zmodyfikowaną wersją, określaną w pracy mianem wariantu lokalnego. Rozważania te zostaną uzupełnione odniesieniem do wyników identyfikacji obserwacji odstających uzyskanych za pomocą metod Kandogana i Hautamäkiego.

3. Wyniki empiryczne

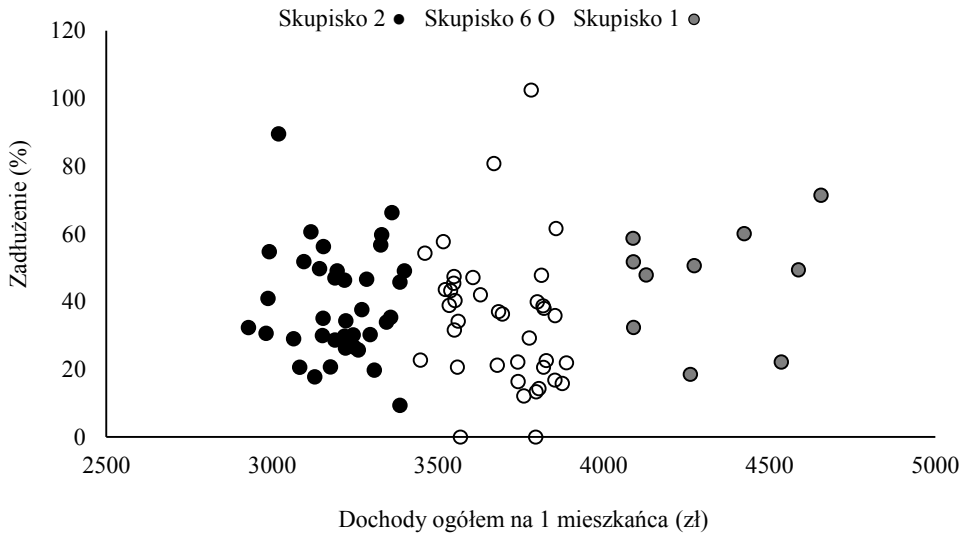
W analizie porównawczej wyników metod identyfikacji obserwacji odstających wykorzystane zostały dane charakteryzujące gminy województwa zachodniopomorskiego w 2014 r. ze względu na: poziom dochodu ogółem przypadający na 1 mieszkańca oraz relatywny stopień zadłużenia, obliczany jako stosunek długu do dochodu ogółem. Źródłem danych był Bank Danych Lokalnych Głównego Urzędu Statystycznego.



Rys. 1. Dominujące skupiska gmin województwa zachodniopomorskiego (krok 1, $n = 103$)

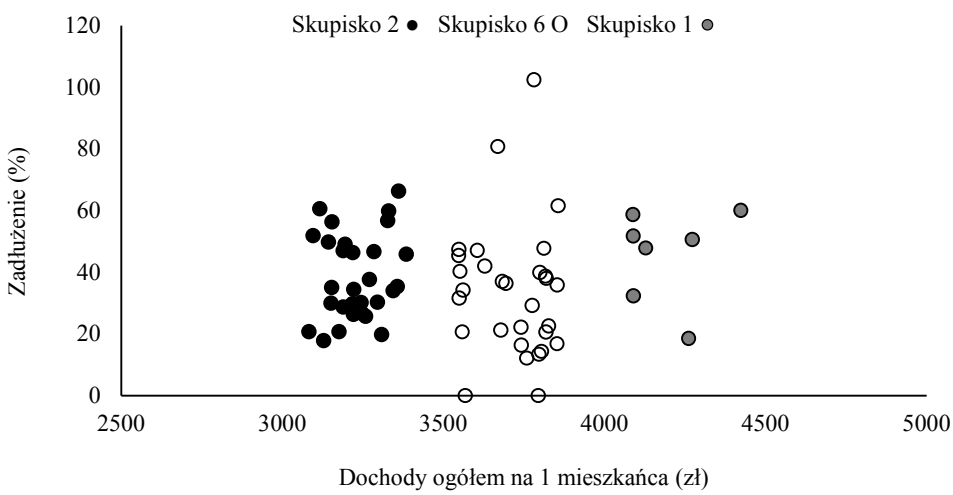
Źródło: obliczenia własne.

Na rysunku 1 przedstawione zostały trzy najliczniejsze skupiska obejmujące 103 obiekty. Ich liczebność wynosiła odpowiednio: 15 (skupienie 1), 44 (skupienie 2) oraz 45 (skupienie 6). Z rozważań wyeliminowane zostały trzy gminy stanowiące odrębne skupiska. Dwie z nich: Rewal i Ostrowice, charakteryzowały się znacząco



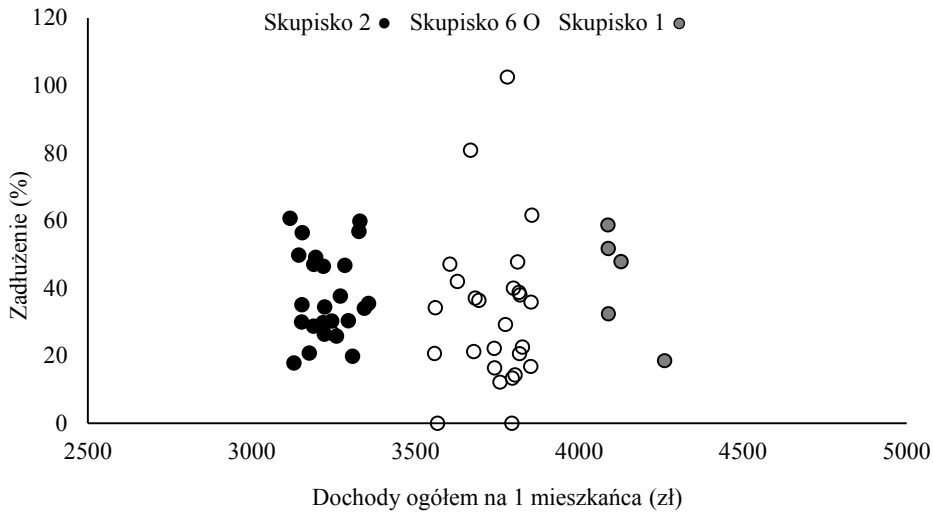
Rys. 2. Dominujące skupiska gmin województwa zachodniopomorskiego (krok 4, $n = 85$)

Źródło: obliczenia własne.



Rys. 3. Dominujące skupiska gmin województwa zachodniopomorskiego (krok 8, $n = 71$)

Źródło: obliczenia własne.



Rys. 4. Dominujące skupiska gmin województwa zachodniopomorskiego (krok 10, $n = 61$)

Źródło: obliczenia własne.

wyższym względnym poziomem zadłużenia, wynoszącym odpowiednio 266,8 oraz 304,2%. Trzecia natomiast wyróżniała się znacząco wyższym od pozostałych obiektów dochodem *per capita* równym 15 333 zł.

Analizując otrzymany podział obiektów, zauważyć można występowanie części wspólnych skupisk. Na kolejnych rys. 2–4 zaprezentowane zostały wyniki kolejnych iteracji, uzyskanych z wykorzystaniem modyfikacji metody Wanga, Zhanga, Li i Songa, która polegała na usuwaniu obserwacji nietypowych w ujęciu lokalnym, czyli w ramach poszczególnych skupisk.

Aby ocenić, czy kolejne kroki pozwalają uzyskać bardziej jednorodne grupy obiektów, obliczono miarę zróżnicowania dokonywanych podziałów w postaci zróżnicowania obiektów wewnątrz poszczególnych grup:

$$Q = \sum_{r=1}^{k} \frac{1}{n_r} \sum_{i=1}^{n_r} d_{ic_r}, \tag{1}$$

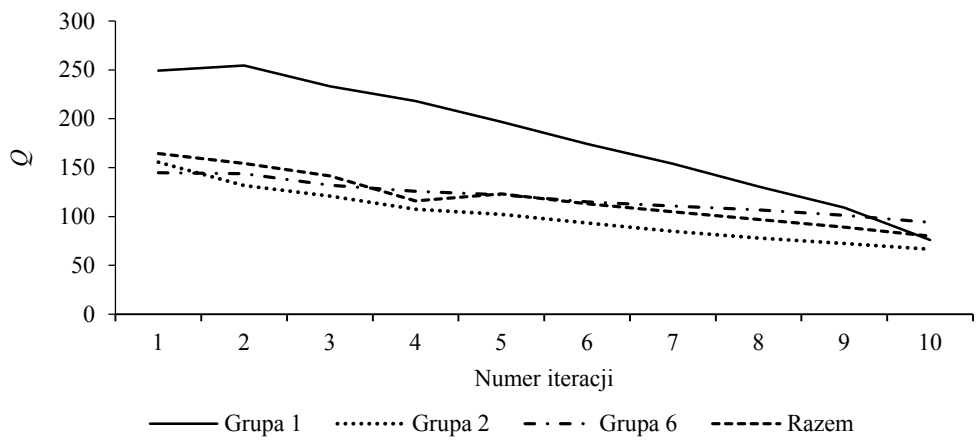
gdzie: r – numer grupy ($r = 1, 2, \ldots, k$); n_r – liczebność grupy r; i – numer obiektu; d_{ic_r} – odległość obiektu i od środka ciężkości w grupie r.

Jak można zauważyć, największy spadek zróżnicowania obliczany dla wszystkich analizowanych obiektów wystąpił po czwartej iteracji (o 17,9%), a po dokonaniu kolejnych dziesięciu kroków poziom zróżnicowania poszczególnych grup obiektów kształtował się na zbliżonym poziomie (zob. rys. 5). Jednocześnie, co zresztą nie jest zaskoczeniem, obserwujemy stały wzrost zróżnicowania tych grup.

Tabela 1. Miara zróżnicowania podziału *Q* w iteracjach 1–10

Nr iteracji	Grupa 1	Grupa 2	Grupa 6	Razem
1	249,1	155,6	144,7	164,3
2	254,6	131,6	143,6	154,2
3	233,1	120,9	132,1	141,3
4	218,0	107,4	125,6	116,0
5	196,9	102,2	122,2	123,2
6	174,2	93,2	115,0	112,9
7	154,0	84,9	110,8	104,7
8	130,8	77,9	106,6	96,8
9	109,0	72,2	101,1	89,2
10	75,9	66,4	93,9	79,9

Źródło: obliczenia własne.



Rys. 5. Miara zróżnicowania podziału *Q* w iteracjach 1–10

Źródło: tab. 1.

W przypadku zaproponowanej modyfikacji metody Wanga, Zhanga, Li i Songa, obiekty nietypowe usuwane są z każdej grupy proporcjonalnie do jej liczebności. Natomiast w podstawowym wariancie tej metody w pierwszych 3 iteracjach identyfikowane były tylko obserwacje odstające znajdujące się w grupach 1 i 2. Dopiero w czwartym kroku zidentyfikowano obserwacje nietypowe należące do grupy 6. Przykładowo w pierwszej iteracji oryginalny wariant rozważanej metody wskazał jako nietypowe gminy: Stargard Szczeciński, Lipiany (Skupisko 2) oraz Stara Dąbrowa, Nowe Warpno i Gościno (Skupisko 1). Natomiast w wariancie lokalnym zidentyfikowano jako obserwacje nietypowe gminy: Stargard Szczeciński, Lipiany (Skupisko 2), Stara Dąbrowa (Skupisko 1) oraz Kobylanka i Kołobrzeg (Skupisko 6). Ograniczeniem obu rozpatrywanych wariantów, oprócz braku

reguły stopu również brak miary stopnia nietypowości poszczególnych obserwacji. Wydaje się, że w pierwszym przypadku można zaproponować rozwiązanie w postaci: określonego spadku miary Q, osiągnięcia jej pewnego poziomu lub wyrównanie się niejednorodności wszystkich skupisk.

Wyniki uzyskane za pomocą metody Kandogana (dla *threshold* = 3σ) pozwoliły zidentyfikować w skupisku 1 te same obiekty odstające, jak dla zmodyfikowanej metody Wanga, Zhanga, Li i Songa, przy liczbie iteracji równej 4. Dla skupiska 2 identyczne obiekty odstające zostały wskazane już po dwóch iteracjach, a dla skupiska 6, podobnie jak dla pierwszego, po czterech krokach. Pojawiły się jednak dwa dodatkowe obiekty nietypowe (gminy Manowo i Wierzchowo), które w zmodyfikowanej metodzie Wanga, Zhanga, Li i Songa wystąpiły dopiero w piątej iteracji.

Wyniki metody *Outlier Removal Clustering* dla $T = 0{,}850$ były zgodne z wynikami uzyskanymi za pomocą metody Kandogana. Jedynie gmina Manowo występująca w skupisku 6 nie została zidentyfikowana jako nietypowa. Warto jednocześnie zauważyć, że ustalenie punktu odniesienia T na poziomie równym 0,900 umożliwiło identyfikację tylko trzech obiektów odstających. Występowały one wyłącznie w skupisku 1, zawierającym obiekty charakteryzujące się najwyższym stopniem zróżnicowania.

4. Zakończenie

W artykule przedstawiono wyniki wybranych metod identyfikacji obiektów odstających, występujących w danych przestrzennych. Uzyskane wyniki pozwalają stwierdzić, że rezultaty otrzymane za pomocą wszystkich metod są praktycznie identyczne. W przypadku oryginalnej metody Wanga, Zhanga, Li i Songa, w porównaniu do jej modyfikacji, zauważyć można jedynie odmienną kolejność identyfikacji obserwacji nietypowych. Natomiast metoda zaproponowana przez E. Kandogana, wykazuje skłonność do wskazywania większej liczby obserwacji odstających, gdy punkt odniesienia zostanie ustalony na poziomie trzech odchyleń standardowych odległości. Wysoka zgodność wyników zaprezentowanych metod pozwala sformułować wartościowe wnioski dla władz samorządowych, zwłaszcza w odniesieniu do prowadzonej polityki spójności wewnątrzregionalnej. Warto jednak w tym celu wykorzystać większą liczbę cech diagnostycznych, chrakteryzujących odmienność poszczególnych jednostek samorządowych. W dalszych badaniach warto podjąć próbę określenia reguły stopu dla metody Wanga, Zhanga, Li i Songa. Można również rozważyć ocenę zastosowania wszystkich rozważanych metod na przykład w analizie skupień wykorzystującej medianę lub w przypadku grupowania z wykorzystaniem rozmytej metody k-średnich. Należy jednak mieć na uwadze, że metoda k-średnich, ze względu na wykorzystywanie średniej i odchyle-

nia standardowego, jest z założenia wrażliwa na występowanie obserwacji odstających, co może powodować trudności w prawidłowej identyfikacji obiektów tego typu.

Literatura

Aggarwal C., Yu P., 2001, *Outlier detection for high dimensional data*, Proceedings of the ACM SIGMOD International Conference on Management of Data, vol. 30, no. 2, s. 37–46.

Batóg J., Batóg B., 2014, *Analiza wpływu obserwacji nietypowych na wyniki modelowania regionalnej wydajności pracy*, Zeszyty Naukowe Uniwersytetu Szczecińskiego nr 811, Studia i Prace Wydziału Nauk Ekonomicznych i Zarządzania nr 36, Metody ilościowe w ekonomii, t. 1, Wydawnictwo Naukowe Uniwersytetu Szczecińskiego, s. 125–138.

Chandola V., Banerjee A., Kumar V., 2009, *Anomaly detection: A Survey*, ACM Computing Surveys (CSUR), vol. 41, no. 3, article no. 15, DOI: 10.1145/1541880.1541882.

Cherednichenko S., 2005, *Outlier Detection in Clustering*, Master's Thesis, University of Joensuu, Department of Computer Science, http://www.cs.uku.fi/pub/Theses/2005_MSc_Cherednichenko_Svetlana.pdf (22.06.2015).

Duan L., Xu L., Liu Y., Lee J., 2009, *Cluster-based outlier detection*, Annals of Operations Research, vol. 168, no. 1, s. 151–168.

Ghosh-Dastidar B., Schafer J.L., 2006, *Outlier detection and editing procedures for continuous multivariate data*, Journal of Official Statistics, vol. 22, no. 3, s. 487–506.

Guha S., Rastogi R., Shim K., 1998, *CURE an efficient clustering algorithm for large databases*, Proceedings of the ACM SIGMOD International Conference on Management of Data, vol. 27, no. 2, s. 73–84.

Hautamäki V., Cherednichenko S., Kärkkäinen I., Kinnunen T., Fränti P., 2005, *Improving k-means by outlier removal*, [w]: H. Kalviainen, J. Parkkinen, A. Kaarna (red.), *Image Analysis*, *14th Scandinavian Conference, SCIA 2005, Joensuu, Finland, June 19–22, 2005, Proceedings, Series Lecture Notes in Computer Science*, vol. 3540, Springer, Berlin–Heidelberg, s. 978–987, DOI: 10.1007/11499145_99.

Hawkins D.M., 1980, *Identification of Outliers*, Chapman and Hall, London.

Hodge V.J., Austin J., 2004, *A survey of outlier detection methodologies*, Artificial Intelligence Review, vol. 22, no. 2, s. 85–126, DOI: 10.1007/s10462-004-4304-y.

Jayakumar G.S.D.S., Thomas B.J., 2013, *A new procedure of clustering based on multivariate outlier detection*, Journal of Data Science, vol. 11, no. 1, s. 69–84.

Kandogan E., 2012, *Just-in-Time Annotation of Clusters, Outliers, and Trends in Point-based Data Visualizations*, IBM Center for Advanced Visualization, IBM Research, IEEE Conference on Visual Analytics Science and Technology, Seattle.

Rousseeuw P.J., Leroy A.M., 1987, *Robust Regression and Outlier Detection*, John Wiley &Sons, New York.

Shekhar S., Lu C., Zhang P., 2003, *A unified approach to detecting spatial outliers*, GeoInformatica, vol. 7, no. 2, s. 139–166, DOI: 10.1023/A:1023455925009.

Wang H., Zhang X., Li S., Song X., 2010, *Spatial clustering and outlier analysis for the regionalization of maize cultivation in China*, Proceedings of the 9th WSEAS International Conference on Applied Computer and Applied Computational Science, s. 31–36, http://www.wseas.us/e-library/conferences/2010/Hangzhou/Acacos/Acacos-04.pdf (14.04.2015).

Zhang T., Ramakrishnan R., Livny M., 1997, *BIRCH: A new data clustering algorithm and its applications*, Data Mining and Knowledge Discovery, vol. 1, no. 2, s. 141–182.