METODY ILOŚCIOWE W BADANIACH EKONOMICZNYCH
Tom XIII/3, 2012, str. 180 – 190

WPŁYW METODY DOBORU CECH NA EFEKTYWNOŚĆ KLASYFIKACJI NA PRZYKŁADZIE ANALIZY JAKOŚCI ŻYCIA W ŚWIETLE ZRÓWNOWAŻONEGO ROZWOJU

Agnieszka Sompolska-Rzechuła
Katedra Zastosowań Matematyki w Ekonomii
Zachodniopomorski Uniwersytet Technologiczny w Szczecinie
e-mail: asompolska@zut.edu.pl

Streszczenie: W artykule podjęto próbę odpowiedzi na pytanie: Czy wyniki otrzymane za pomocą różnych metod doboru cech mają wpływ na efektywność klasyfikacji? Do badania wykorzystano dwie metody doboru cech: parametryczną metodę coraz metodę odwróconej macierzy współczynników korelacji. Skuteczność grupowań zweryfikowano za pomocą wskaźników homogeniczności, heterogeniczności i poprawności skupień. W ocenie efektywności grupowań wykorzystano podejście z medianą Webera. Badanie dotyczyło powiatów województwa zachodniopomorskiego pod względem jakości życia w świetle zrównoważonego rozwoju.

Słowa kluczowe: metoda doboru cech, efektywność klasyfikacji, jakość życia, rozwój zrównoważony

WPROWADZENIE

Celem artykułu jest próba odpowiedzi na pytanie: Czy wyniki otrzymane za pomocą różnych metod doboru cech mają wpływ na efektywność klasyfikacji? Dobór cech jest jednym z najważniejszych, a jednocześnie najtrudniejszych zagadnień. Niezbędna jest kompleksowa znajomość analizowanego zagadnienia oraz specyfiki powiązań pomiędzy zjawiskami społeczno-gospodarczymi. Od jakości zestawu cech zależy wiarygodność ostatecznych wyników i trafność podejmowanych decyzji [Gatnar, Walesiak 2004]. Próbę odpowiedzi na postawione pytanie podjęto na podstawie badania taksonomicznego powiatów ziemskich województwa zachodniopomorskiego, dokonując klasyfikacji obiektów

na podstawie zbioru cech otrzymanych metodami: parametryczną oraz metodą odwróconej macierzy współczynników korelacji. Podziału powiatów dokonano pod względem obiektywnej jakości życia mieszkańców w świetle zrównoważonego rozwoju w roku 2010.

Istnieje bardzo wiele określeń i klasyfikacji jakości życia, a problemami z jej zakresu zajmują się przedstawiciele wielu dyscyplin naukowych (filozofii, socjologii, psychologii, ekonomii oraz statystyki). Jako kategoria wyrażająca stopień samorealizacji człowieka jakość życia powinna być podstawowym przedmiotem zainteresowania społeczeństwa. Jak podaje Tadeusz Borys [Borys 2008] trzy kategorie: jakość życia, rozwój społeczny, gospodarczy i środowiskowy oraz instrumentarium tego rozwoju tworzą hierarchiczny układ pojęć ściśle ze sobą związanych i powinny być przedmiotem zintegrowanego pomiaru wskaźnikowego. Wspólne cechy zmian rozwojowych jakości życia i rozwoju zrównoważonego[1] znajdują odzwierciedlenie w powiązaniu opisu wskaźnikowego. Duże znaczenie mają wskaźniki rozwoju zrównoważonego w opisie pośredniej jakości życia oraz przy tworzeniu pośrednich wskaźników jakości życia. Większość wskaźników zrównoważonego rozwoju tworzy pośredni obraz obiektywnej jakości życia. Trwały i zrównoważony rozwój, w większości definicji, postrzegany jest jako taki sposób gospodarowania, który prowadzi do poprawy jakości życia.

Obliczenia zostały wykonane w arkuszu kalkulacyjnym Excel oraz programach: Statistica i **R**.

OPIS METODY

Zastosowanie metod wielowymiarowej analizy porównawczej wymaga wyboru obiektów oraz zbioru cech diagnostycznych charakteryzujących poszczególne obiekty.

Po określeniu i zgromadzeniu danych dotyczących wstępnego zestawu cech należy podjąć odpowiednie działania weryfikacyjne według dwóch najistotniejszych kryteriów [Młodak 2006]:

1. Zmienność – cechy powinny wykazywać odpowiednią zmienność, czyli skutecznie dyskryminować obiekty. Do oceny zmienności wartości cech wykorzystuje się współczynnik zmienności:

$$v_j = \frac{s_j}{\bar{x}_j} \tag{1}$$

[1] W literaturze można znaleźć wiele określeń zrównoważonego rozwoju, przykładem ujęcia ogólnego jest definicja trwałego i zrównoważonego rozwoju, według której: „istotą rozwoju zrównoważonego i trwałego jest zapewnienie trwałej poprawy jakości życia współczesnych i przyszłych pokoleń poprzez kształtowanie właściwych proporcji między trzema rodzajami kapitału: ekonomicznym, ludzkim i przyrodniczym" Piontek F. (2001) Ekonomia a rozwój zrównoważony, Ekonomia i środowisko, str. 19.

gdzie: $\bar{x}_j$ to średnia arytmetyczna wartości cechy X_j, zaś s_j jest odchyleniem standardowym j-tej cechy, $j = 1, \ldots, m$, m – liczba cech.

2. Korelacja – dwie cechy silnie ze sobą skorelowane są nośnikami podobnej informacji, zatem jedna z nich jest zbędna. Do oceny siły związku między cechami stosuje się współczynnik korelacji. Punktem wyjścia jest macierz współczynników korelacji między wszystkimi parami cech:

$$\mathbf{R} = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1m} \\ r_{21} & 1 & \cdots & r_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ r_{m1} & r_{m2} & \cdots & 1 \end{bmatrix} \tag{2}$$

gdzie: r_{jk} to współczynniki korelacji liniowej Pearsona j-tej i k-tej cechy.

Metodą wykorzystywaną do dyskryminacji cech bazującą na macierzy współczynników korelacji jest metoda parametryczna, która jest wygodna w użyciu, ponieważ jest prosta rachunkowo.

Metoda parametryczna posiada jednak dwie zasadnicze wady [Młodak 2006, Panek 2009]:

- jest wrażliwa na wartości odstające, co oznacza, że na wysoką wartość współczynnika korelacji może, w dużym stopniu, wpływać jej wysokie skorelowanie nawet z jedną z cech,
- uwzględnia wyłącznie bezpośrednie powiązania cechy z innymi cechami, nie uwzględniając powiązań pośrednich.

Skutecznym sposobem zniwelowania pierwszej niedogodności jest zastąpienie w pierwszym kroku sumy elementów kolumny (wiersza) macierzy **R** przez ich medianę. Pozwala to uodpornić analizę na zaburzenia spowodowane przez obserwacje odstające.

Druga wada może być wyeliminowana poprzez zastosowanie metody odwróconej macierzy współczynników korelacji [Panek 2009, Malina, Zeliaś 1997]. Procedura eliminacji jest następująca: korzystając z macierzy współczynników korelacji **R**, wyznacza się macierz $\mathbf{R}^{-1} = [r^{ij}]$, gdzie wartości r^{ij} są elementami macierzy odwrotnej $\mathbf{R}^{-1}$. Element diagonalny r^{ii} macierzy $\mathbf{R}^{-1}$ jest równy jedności, jeśli zmienna X_j jest ortogonalna względem pozostałych zmiennych. W przypadku nieortogonalności $r^{ii} \in (1, +\infty)$, gdy zmienna jest nadmiernie skorelowana z pozostałymi, elementy diagonalne macierzy odwrotnej $\mathbf{R}^{-1}$ są znacznie większe od jedności, co jest symptomem złego uwarunkowania macierzy **R**.

Cechy nadmiernie skorelowane, którym odpowiadają elementy diagonalne r^{ii} o wartościach większych niż 10, są eliminowane z pierwotnego zbioru cech. Jeżeli takie elementy nie występują, to procedurę uznaje się za zakończoną. Ponownie wyznacza się macierz odwrotną $\mathbf{R}^{-1}$ dla zredukowanego zbioru cech i analizuje jej

elementy diagonalne. Procedurę powtarza się do momentu osiągnięcia stabilności macierzy $\mathbf{R}^{-1}$, czyli pojawienia się elementów diagonalnych, których wartości nie przekraczają znacząco 10.

Otrzymany zbiór cech diagnostycznych stanowił podstawę klasyfikacji obiektów. Spośród wielu metod hierarchicznych do badania wybrano metodę Warda, która różni się od wszystkich pozostałych metod tym, że do oszacowania odległości między skupieniami wykorzystuje się podejście analizy wariancji. Metoda ta zmierza do minimalizacji sumy kwadratów odchyleń dowolnych dwóch hipotetycznych skupień, które mogą zostać uformowane na każdym etapie analizy. Ważną cechą tej metody jest zapewnienie minimalizacji kryterium wariancyjnego, które głosi, że wariancja wewnątrz skupień jest minimalna. Metoda Warda zapewnia zatem homogeniczność wewnątrz skupień i heterogeniczność między skupieniami, przez co uznawana jest za bardzo efektywną [Ward 1963].

Ostatnim etapem analizy taksonomicznej obiektów jest sprawdzenie jakości uzyskanych podziałów. Do oceny jakości klasyfikacji stosuje się mierniki homogeniczności oraz heterogeniczności skupień, wykorzystując koncepcję środka ciężkości grupy i odległości od niego. W badaniu wykorzystano podejście, w którym środek ciężkości danej grupy zastąpiony został medianą Webera jej elementów. Mediana Webera stanowi wielowymiarowe uogólnienie klasycznego pojęcia mediany. Chodzi o wektor, który minimalizuje sumę euklidesowych odległości od danych punktów reprezentujących rozpatrywane obiekty, a więc znajduje się niejako „pośrodku" nich, ale jest jednocześnie uodporniony na występowanie obserwacji odstających [Młodak 2006].

W ocenie homogeniczności otrzymanych grup wykorzystano miernik o następującej postaci [Młodak 2006]:

$$hm_{6\,\max}^{*} = \max_{k=1,\ldots,p} hm_6^{*}\left(P_k\right) \tag{3}$$

gdzie:

$$hm_6^{*}\left(P_k\right) = \operatorname*{med}_{i:O_i \in P_k} \delta\left(O_i, \Gamma_{\theta k}\right) \tag{4}$$

jest medianą odległości obiektów grupy P_k od jej wektora medianowego Webera,

$$\Gamma_{\theta k} = \left(\theta_{1P_k}, \theta_{2P_k}, \ldots, \theta_{mP_k}\right) \tag{5}$$

jest wektorem medianowym Webera, δ - odległość obiektów grupy P_k od jej wektora medianowego Webera, O_i - obiekty, θ_{mP_k} - mediana Webera rozpatrywanego układu m cech diagnostycznych, k – liczba klas, $k = 1,2,\ldots,p$, p – liczba skupień otrzymanych na danym poziomie grupowania.

Natomiast w ocenie heterogeniczności zastosowano miernik:

$$ht_{6\,\mathrm{mn}}^{*} = \min_{k=1,\ldots,p} ht_{6}^{*}(P_k) \tag{6}$$

gdzie:

$$ht_{6}^{*}(P_k) = \operatorname*{med}_{\substack{i=1,\ldots,p \\ i\neq k}} \delta(\Gamma_{\theta i}, \Gamma_{\theta k}) \tag{7}$$

jest medianą odległości pomiędzy medianą Webera danej grupy z analogicznymi wektorami dla pozostałych grup.

W ocenie poprawności grupowania wykorzystano kompleksowy miernik o postaci:

$$ct_{6}^{*} = \frac{hm_{6\,\mathrm{mx}}^{*}}{ht_{6\,\mathrm{mn}}^{*}} \tag{8}$$

CHARAKTERYSTYKA MATERIAŁU BADAWCZEGO

Doboru cech diagnostycznych dokonano za pomocą dwóch metod: parametrycznej oraz odwróconej macierzy współczynników korelacji. Wstępna lista cech diagnostycznych obejmowała wskaźniki ujęte w grupy i kategorie wskaźników zrównoważonego rozwoju, które zostały przedstawione w tabeli 1.

Tabela 1. Grupy i kategorie wskaźników zrównoważonego rozwoju

Ład środowiskowy	Ład ekonomiczny	Ład społeczny
• Zmiany klimatu • Energia • Ochrona powietrza • Ekosystemy morskie • Zasoby słodkiej wody • Użytkowanie gruntów • Bioróżnorodność • Gospodarka odpadami	• Rozwój gospodarczy • Zatrudnienie • Innowacyjność • Transport • Zrównoważone wzorce produkcji	• Zmiany demograficzne • Zdrowie publiczne • Integracja społeczna • Edukacja • Dostęp do rynku pracy • Bezpieczeństwo publiczne • Zrównoważone wzorce konsumpcji

Źródło: opracowanie własne na podstawie [Wskaźniki zrównoważonego rozwoju Polski, GUS 2011]

Do opisu jakości życia przyjęto następujący zestaw cech [Województwo zachodniopomorskie, podregiony, powiaty, gminy 2011],:

X_1 - ludność w wieku nieprodukcyjnym na 100 osób w wieku produkcyjnym,
X_2 - małżeństwa zawarte na 1000 ludności,
X_3 - urodzenia żywe na 1000 ludności,
X_4 - zgony niemowląt na 1000 ludności,
X_5 - przyrost naturalny na 1000 ludności,
X_6 - rozwody na 1000 ludności,

X_7 - separacje na 100 tys. ludności,
X_8 - saldo migracji na 1000 ludności,
X_9 - liczba ludności na 1 placówkę biblioteczną,
X_{10} - liczba ludności na 1 instytucję kultury,
X_{11} - liczba ludności na 1 lekarza,
X_{12} - liczba ludności na 1 aptekę i punkt apteczny,
X_{13} - beneficjenci pomocy społecznej w % ogółu ludności,
X_{14} - osoby niepełnosprawne poniżej 16 roku życia na 1000 ludności poniżej 16 roku życia,
X_{15} - osoby niepełnosprawne powyżej 16 roku życia na 1000 ludności powyżej 16 roku życia,
X_{16} - korzystający z noclegów na 1000 ludności,
X_{17}- udzielone noclegi na 1000 ludności,
X_{18} - lesistość w %,
X_{19} - ludność korzystająca z oczyszczalni ścieków w % ludności ogółem,
X_{20} - emisja zanieczyszczeń pyłowych w tonach na km^2,
X_{21} - emisja zanieczyszczeń gazowych w tonach na km^2,
X_{22} - odpady wytworzone w ciągu roku w tys. t na km^2,
X_{23} - stopień redukcji wytworzonych zanieczyszczeń w %,
X_{24} - udział powierzchni o szczególnych walorach przyrodniczych prawnie chronionej w powierzchni powiatu (w%),
X_{25} - udział rezerwatów w powierzchni o szczególnych walorach przyrodniczych prawnie chronionej (w %),
X_{26} - pomniki przyrody na km^2,
X_{27} - nakłady na ochronę środowiska w tys. zł na km^2,
X_{28} - podmioty gospodarki narodowej w sektorze prywatnym na 1000 ludności,
X_{29} - stopa bezrobocia rejestrowanego (w %),
X_{30} - bezrobotne kobiety w liczbie bezrobotnych ogółem w %,
X_{31} - bezrobotni trwale bezrobotni w liczbie bezrobotnych ogółem w %,
X_{32} - długość sieci wodociągowej w km na 1 km^2,
X_{33} - długość sieci kanalizacyjnej w km na 1 km^2,
X_{34} - zasoby mieszkaniowe na 1000 ludności,
X_{35} - mieszkania w miastach wyposażone w łazienkę w % ogółu mieszkań,
X_{36} - mieszkania w miastach wyposażone w gaz z sieci w % ogółu mieszkań,
X_{37} - drogi publiczne powiatowe o twardej nawierzchni w km na 1 km^2,
X_{38} - wypadki drogowe na 10 tys. ludności,
X_{39} - śmiertelne ofiary wypadków drogowych na 10 tys. ludności.

Metoda parametryczna pozwoliła na wyodrębnienie następującego zbioru cech diagnostycznych: X_4, X_{10}, X_{19}, X_{22}, X_{24}, X_{27}, X_{30}, X_{34}, X_{37}.

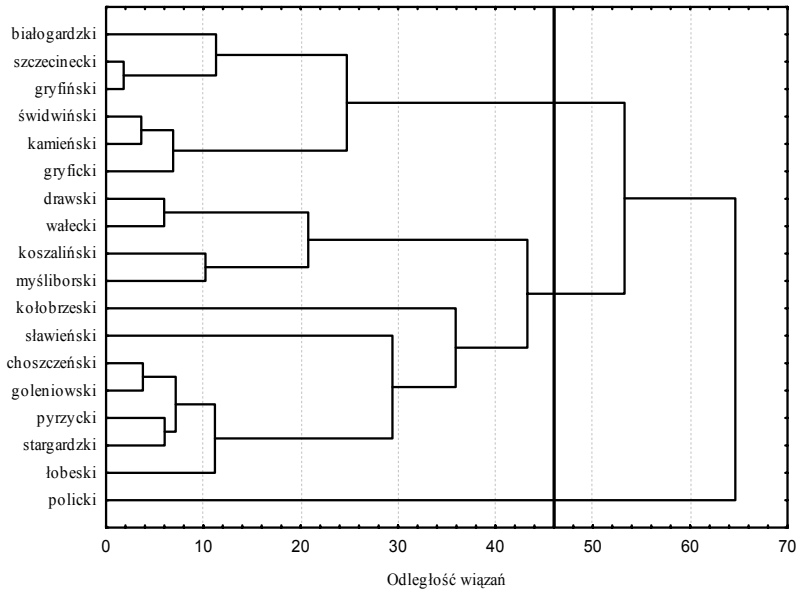
Natomiast wykorzystanie metody odwróconej macierzy współczynników korelacji doprowadziło do uzyskania zbioru cech: X_{10}, X_{23}, X_{24}, X_{32}, X_{33}, X_{34}, X_{37}, X_{38}, X_{39}.

WYNIKI BADANIA

Wykorzystując zbiór cech diagnostycznych uzyskany metodami: parametryczną i odwróconej macierzy współczynników korelacji dokonano klasyfikacji powiatów ziemskich województwa zachodniopomorskiego. Analizując dendrogramy otrzymano po trzy skupienia powiatów.

Dendrogram uzyskany metodą Warda na podstawie zbioru cech otrzymanych drugą metodą został przedstawiony na rysunku 1.

Rysunek 1. Dendrogram powiatów ziemskich województwa zachodniopomorskiego



Źródło: opracowanie własne

W tabeli 2 przedstawiono wyniki grupowania powiatów ziemskich metodą Warda na podstawie zbiorów cech uzyskanych metodami: parametryczną i odwróconej macierzy współczynników korelacji.

Tabela 2. Wyniki grupowania powiatów ziemskich metodą Warda na podstawie zbiorów cech uzyskanych metodami: parametryczną i odwróconej macierzy współczynników korelacji

Grupowanie powiatów metodą Warda na podstawie zbioru cech uzyskanych metodą					
parametryczną			odwróconej macierzy współczynników korelacji		
grupa I	grupa II	grupa III	grupa I	grupa II	grupa III
wałecki, drawski, goleniowski, choszczeński, kołobrzeski stargardzki, szczecinecki, myśliborski, pyrzycki, sławieński, policki, gryficki	świdwiński, kamieński, gryfiński,	białogardzki, koszaliński, łobeski	policki	łobeski, stargardzki, pyrzycki, goleniowski, choszczeński, sławieński, kołobrzeski, myśliborski, koszaliński, wałecki, drawski	gryficki, kamieński, świdwiński, gryfiński, szczecinecki, białogardzki

Źródło: opracowanie własne

Skuteczność grupowań zweryfikowano wyznaczając wartości wskaźników homogeniczności, heterogeniczności i poprawności skupień (tabela 3) [Młodak 2006].

Tabela 3. Wartości wskaźników homogeniczności, heterogeniczności i poprawności skupień

Wskaźniki	Wariant oparty na zbiorze cech uzyskanych metodą odwróconej macierzy współczynników korelacji	Wariant oparty na zbiorze cech uzyskanych metodą parametryczną
homogeniczności skupień	390,81	733,80
heterogeniczności skupień	1846,72	732,30
poprawności skupień	0,21	1,00

Źródło: obliczenia własne

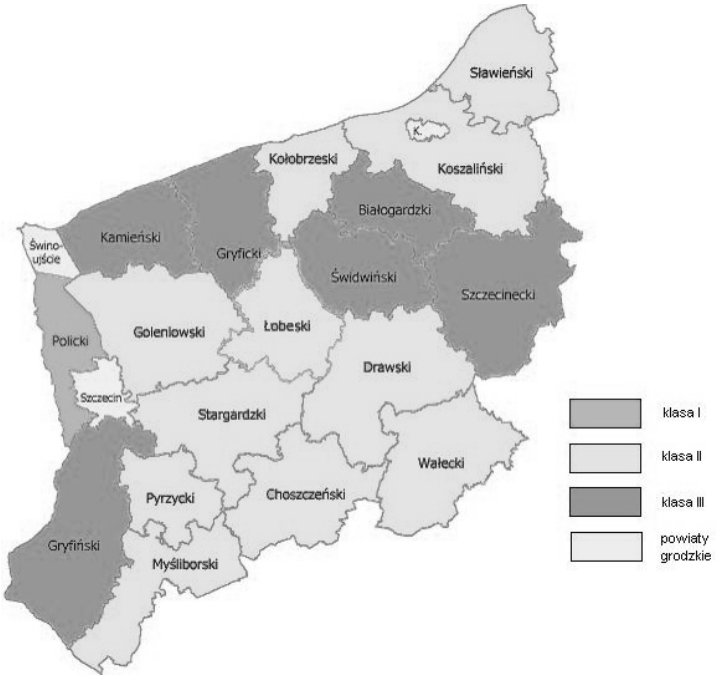
Analizując wyniki dotyczące efektywności grupowań, przedstawione w tabeli 3, można stwierdzić, że klasyfikacja otrzymana metodą Warda na podstawie zbioru cech uzyskanych metodą odwróconej macierzy współczynników korelacji dała lepsze rezultaty pod każdym względem, czyli zarówno homogeniczności i heterogeniczności, jak i poprawności grupowania w porównaniu z klasyfikacją otrzymaną z wykorzystaniem metody parametrycznej.

Klasyfikacja oparta na metodzie odwróconej macierzy współczynników korelacji wyodrębniła trzy skupienia, wśród których jest skupienie jednoelementowe – powiat policki. Powiat ten charakteryzuje się najwyższym

stopniem redukcji wytworzonych zanieczyszczeń (w %) i wysokim udziałem powierzchni o szczególnych walorach przyrodniczych prawnie chronionej w powierzchni powiatu, jak również korzystnymi wartościami wskaźników dotyczących infrastruktury technicznej (np. długość sieci kanalizacyjnej w km na 1 km^2 oraz drogi publiczne powiatowe o twardej nawierzchni w km na 1 km^2). Klasa druga z największą liczbą powiatów (11) wyróżnia się najniższym stopniem redukcji wytworzonych zanieczyszczeń (ponad siedmiokrotnie niższym w porównaniu ze średnią ogólną) ale wysokim udziałem powierzchni o szczególnych walorach przyrodniczych prawnie chronionej w powierzchni powiatu. Pozostałe wskaźniki oscylują wokół średnich ogólnych. Trzecia klasa wyróżnia się niekorzystnie pod względem bardzo niskiego stopnia redukcji wytworzonych zanieczyszczeń (w %) oraz najniższą wartością długości dróg publicznych powiatowe o twardej nawierzchni w km na 1 km^2.

Podział powiatów województwa zachodniopomorskiego pokazuje rys. 2.

Rysunek 2. Podział powiatów województwa zachodniopomorskiego metodą Warda na podstawie zbioru cech otrzymanych metodą odwróconej macierzy współczynników korelacji



Źródło: opracowanie własne

PODSUMOWANIE

W pracy przedstawiono próbę odpowiedzi na pytanie dotyczące wpływu zbiorów cech diagnostycznych otrzymanych różnymi metodami doboru cech na efektywność klasyfikacji. W badaniu wykorzystano dwie metody doboru cech: parametryczną i odwróconej macierzy współczynników korelacji. Parametryczna procedura doboru cech posiada dwie niedogodności, które są niwelowane w metodzie odwróconej macierzy współczynników korelacji. Otrzymane zbiory posłużyły do klasyfikacji ziemskich powiatów województwa zachodniopomorskiego pod względem obiektywnej jakości życia w świetle zrównoważonego rozwoju. Badanie dotyczyło 2010 roku. Efektywność klasyfikacji zbadano wykorzystując wskaźniki homogeniczności, heterogeniczności oraz poprawności grupowań, w których role środków ciężkości odgrywała mediana Webera. Klasyfikacja wykorzystująca zbiór cech uzyskany metodą odwróconej macierzy współczynników korelacji dała lepsze rezultaty pod względem wszystkich trzech kryteriów.

Badanie wykazało, iż metody klasyfikacji mogą być skutecznym narzędziem w ocenie jakości życia mieszkańców, a wyniki metod doboru cech do badania taksonomicznego mają wpływ na jakość i na rezultaty klasyfikacji.

BIBLIOGRAFIA

Borys T., Rogala P. (red.) (2008) Jakość życia na poziomie lokalnym – ujęcie wskaźnikowe, Program Narodów Zjednoczonych ds. Rozwoju, Warszawa, str. 9-10

Gatnar E., Walesiak M. (2004) Metody statystycznej analizy wielowymiarowej w badaniach marketingowych, Wydawnictwo Akademii Ekonomicznej we Wrocławiu, str. 320

Młodak A. (2006) Analiza taksonomiczna w statystyce regionalnej, Difin, Warszawa, str. 31

Panek T. (2009) Statystyczne metody wielowymiarowej analizy porównawczej, Szkoła Główna Handlowa w Warszawie, str. 22

Malina A., Zeliaś A. (1997) O budowie taksonomicznej miary jakości życia, Taksonomia 4, Wydawnictwo Akademii Ekonomicznej we Wrocławiu, str. 238-263

Ward J. H. (1963), Hierarchical grouping to optimize an objective function, Journal of the American Statistical Association, No. 58

Piontek F. (2001) Ekonomia a rozwój zrównoważony, Ekonomia i środowisko, Białystok, str. 19

Województwo zachodniopomorskie, podregiony, powiaty, gminy (2011), Urząd Statystyczny w Szczecinie

Wskaźniki zrównoważonego rozwoju Polski (2011), Główny Urząd Statystyczny, Urząd Statystyczny w Katowicach

THE INFLUENCE OF THE METHOD OF THE FEATURE SELECTION ON THE CLASSIFICATION'S EFFICIENCY BASED ON THE QUALITY OF LIFE IN LIGHT ON THE SUSTAINABLE DEVELOPMENT

Abstract: In the article attempts to answer the question: Do the results, obtained by means of the various feature selection method, have any influence on the classification's efficiency? For the analysis two methods were used: parametric method and the matrix inverse method of the correlation coefficients. The effectiveness of classifications was checked by use of homogeneity, heterogeneity and correctness of clustering coefficients. The approach was used in the assessment of the classification's efficiency, with the center of gravity replaced with the Weber's median. The analysis was local and concerned the districts in zachodniopomorskie province in terms of the quality of life in the light of sustainable development.

Keywords: the feature selection method, the classification's efficiency, quality of life, sustainable development