EKONOMETRIA ECONOMETRICS 4(38) • 2012

ISSN 1507-3866

Agnieszka Sompolska-Rzechuła
Zachodniopomorski Uniwersytet Technologiczny w Szczecinie

EFEKTYWNOŚĆ KLASYFIKACJI A PARAMETRYCZNA METODA DOBORU CECH DIAGNOSTYCZNYCH

Streszczenie: W pracy przedstawiono dwa warianty parametrycznej metody doboru cech: z sumą oraz medianą elementów kolumny macierzy współczynników korelacji. Klasyczna wersja (z sumą) tej metody doboru cech ma pewne niedogodności, które mogą być zniwelowane przez zastąpienie sumy elementów kolumny macierzy współczynników korelacji ich medianą. Powoduje to zmniejszenie wrażliwości na wartości odstające współczynników korelacji. Celem artykułu było określenie wpływu wyników dwóch podejść w parametrycznej metodzie doboru na efektywność klasyfikacji obiektów. W każdej klasyfikacji wyłoniono, metodą Warda, po trzy klasy województw i zbadano efektywność otrzymanych podziałów, wykorzystując wskaźniki homogeniczności, heterogeniczności oraz poprawności grupowań, w których role środków ciężkości odgrywała mediana Webera.

Słowa kluczowe: parametryczna metoda doboru cech, klasyfikacja, efektywność klasyfikacji.

1. Wstęp

Zadaniem klasyfikacji jest badanie podobieństwa lub odrębności obiektów i ich zbiorów, chodzi zatem o podział zbioru obiektów na klasy zawierające obiekty podobne ze względu na obserwacje na zmiennych [Gatnar, Walesiak 2004]. W procesie klasyfikacji obiektów wyróżnia się kilka etapów postępowania. Jednym z pierwszych jest wybór cech charakteryzujących poszczególne obiekty. Etap ten jest bardzo ważnym, a jednocześnie najtrudniejszym zagadnieniem, ponieważ od jakości zestawu cech zależy wiarygodność ostatecznych wyników i trafność podejmowanych decyzji. Niezbędna jest kompleksowa znajomość analizowanego zagadnienia oraz specyfiki powiązań pomiędzy zjawiskami społeczno-gospodarczymi. Najbardziej właściwą procedurą doboru cech diagnostycznych jest wykorzystanie dwóch podejść, zarówno pozastatystycznych (merytorycznych i formalnych), jak i statystycznych. W pierwszym za cechy diagnostyczne uważane są te cechy, które w świetle wiedzy merytorycznej o badanym zjawisku są najważniejsze dla dokonania analizy porównawczej obiektów. Drugie podejście wykorzystuje odpowiednie procedury statystyczne [Panek 2009].

Celem artykułu jest rozważenie dwóch podejść w parametrycznej metodzie doboru cech diagnostycznych i określenie wpływu wyników na efektywność klasyfikacji obiektów. Postawiony problem zilustrowano badaniem o charakterze regionalnym, na przykładzie analizy województw pod względem poziomu życia ludności. Badanie dotyczyło roku 2009. Jako kryterium klasyfikacji wybrano kategorię poziom życia, ponieważ jest ona jedną z podstawowych kategorii badawczych w statystyce społecznej i począwszy od lat 90. XX wieku, zaobserwowano występowanie, z rosnącą siłą, nowych zjawisk związanych ze wzrostem i rozwojem gospodarczym oraz poziomem życia ludzi. Początek lat 90. XX wieku to również okres, w którym Polska wchodziła w coraz ściślejsze związki z krajami Unii Europejskiej. Dzięki procesowi integracji otworzyła się przed Polską możliwość znacznego przyspieszenia likwidacji wszelkiego rodzaju opóźnień technicznych, technologicznych i organizacyjnych. Zaistniała także szansa poprawy poziomu życia całego społeczeństwa, ponieważ Unia Europejska dąży do zmniejszenia dysproporcji w tym zakresie. Poziom życia jest kategorią nie do końca sprecyzowaną. Nadal w literaturze spotyka się wiele różnorodnych prób zdefiniowania poziomu życia, co stanowi zasadniczą trudność pojawiającą się przed badaczami. Jedna z pierwszych propozycji definiowania poziomu życia została zaproponowana przez Komisję Ekspertów ONZ i przedstawiona na początku lat 50. XX wieku. Według niej poziom życia obejmuje całokształt rzeczywistych warunków życia ludzi oraz stopień ich materialnego i kulturalnego zaspokojenia przez strumień dóbr i usług odpłatnych, a także pochodzących z funduszy społecznych [Zeliaś 2000]. W literaturze przedmiotu kategoria poziomu życia jest różnie definiowana, początkowo dotyczyła warunków życia ludzi i stopnia materialnego i kulturalnego zaspokajania ich potrzeb. Podejście to jednak nie zawierało ocen subiektywnych. Można przytoczyć następujące definicje poziomu życia, które są ściśle związane z podstawowym pojęciem badań społecznych, czyli potrzebą [Panek 2007]:

- poziom życia to stopień zaspokojenia potrzeb wynikający z konsumpcji wytworzonych przez człowieka dóbr materialnych i usług,
- poziom życia to stopień zaspokojenia potrzeb materialnych i kulturalnych przy istniejącej infrastrukturze umożliwiającej to zaspokojenie.

2. Opis metody

Prawidłowo przeprowadzony dobór cech diagnostycznych powinien składać się z dwóch etapów: doboru merytorycznego oraz doboru formalnego. W wielu badaniach empirycznych autorzy podają zestawy cech, opierając się na kryteriach merytorycznych lub formalnych bez szerszej dyskusji problemu. Spośród metod formalnych wykorzystywany jest współczynnik zmienności, jako narzędzie do oceny skuteczności dyskryminacji obiektów oraz metoda parametryczna doboru cech[1]. Ta

[1] Opis metody można znaleźć np. w pracach: [Panek 2009, s. 21-22; Młodak 2006, s. 29-30].

ostatnia stosowana jest bardzo często, ponieważ jest wygodna w użyciu i prosta rachunkowo.

Metoda parametryczna ma jednak dwie zasadnicze wady [Młodak 2006]:

1) 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,

2) 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 przez zastosowanie *metody odwróconej macierzy współczynników korelacji* [Panek 2009]. Kolejnym etapem, po wyodrębnieniu zbioru cech diagnostycznych, jest ich normalizacja, która może być przeprowadzona za pomocą jednego z trzech przekształceń normalizacyjnych, zwanego standaryzacją [Panek 2009]:

$$z_{ij} = \frac{x_{ij} - \overline{x}_j}{s(x_j)} \qquad i = 1, \ \ldots, \ n; \quad j = 1, \ \ldots, \ m,$$

gdzie: n – liczba obiektów, m – liczba cech.

Następnym krokiem jest wybór metody klasyfikacji. W literaturze przedmiotu istnieje wiele propozycji podziałów metod klasyfikacji. W pracy [Gatnar, Walesiak 2004] przedstawiono podział metod klasyfikacji na trzy grupy:

1) metody hierarchiczne (aglomeracyjne i deglomeracyjne);

2) metody podziału;

3) metody prezentacji graficznej.

W ocenie poziomu życia szczególne znaczenie mają hierarchiczne metody aglomeracyjne, które są dobrze opracowane pod względem metodologicznym i mają wiele zalet, do których można zaliczyć graficzną prezentację wyników klasyfikacji w postaci dendrogramu wskazującego na kolejność połączeń między klasami. Spośród wielu metod hierarchicznych do badania wybrano metodę Warda[2]. Została ona zaproponowana w roku 1963 i 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 mię-

[2] Opis metod analizy skupień, w tym metody Warda, można znaleźć np. w pracy [Balicki 2009].

dzy 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]:

$$hm_6^{*}\mathrm{m}x = \max_{k=1,\ldots,p} hm_6^{*}(P_k),$$

gdzie:

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

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

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

jest wektorem medianowym Webera, k – liczbą klas, $k = 1,2, \ldots, p$, p – liczbą skupień otrzymanych na danym poziomie grupowania.

Natomiast w ocenie heterogeniczności zastosowano miernik:

$$ht_6^{*}\mathrm{mn} = \min_{k=1,\ldots,p} ht_6^{*}\left(P_k\right),$$

gdzie:

$$ht_6^{*}\left(P_k\right) = \operatorname*{med}_{\substack{i=1,\ldots,p \\ i \neq k}} \delta\left(\Gamma_{\theta i}, \Gamma_{\theta k}\right)$$

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

W ocenie poprawności grupowania wykorzystano kompleksowy miernik o postaci:

$$ct_6^{*} = \frac{hm_6^{*}\mathrm{m}x}{ht_6^{*}\mathrm{mn}}.$$

3. Materiał badawczy

Źródło danych w badaniu stanowiły informacje dotyczące województw Polski pod względem przyjętego kryterium, którym był poziom życia ludności. W badaniu wykorzystano dane statystyczne udostępnione przez Główny Urząd Statystyczny w Banku Danych Lokalnych (http://www.stat.gov.pl/bdl/app/portret.dims). Do analizy przyjęto następujący zestaw cech diagnostycznych:

X_1 – liczba ludności na 1 km^2,
X_2 – udział ludności w wieku przedprodukcyjnym w ogólnej liczbie ludności,
X_3 – udział ludności w wieku produkcyjnym w ogólnej liczbie ludności,
X_4 – udział ludności w wieku poprodukcyjnym w ogólnej liczbie ludności,
X_5 – ludność w wieku nieprodukcyjnym na 100 osób w wieku produkcyjnym,
X_6 – ludność w wieku poprodukcyjnym na 100 osób w wieku przedprodukcyjnym,
X_7 – ludność w wieku poprodukcyjnym na 100 osób w wieku produkcyjnym,
X_8 – liczba kobiet na 100 mężczyzn,
X_9 – zgony na 1000 ludności,
X_{10} – przyrost naturalny na 1000 ludności,
X_{11} – urodzenia żywe na 1000 ludności,
X_{12} – liczba małżeństw zawartych w ciągu roku na 1000 ludności,
X_{13} – liczba rozwodów na 1000 ludności,
X_{14} – zgony niemowląt na 1000 urodzeń żywych,
X_{15} – przeciętne miesięczne wydatki na 1 osobę,
X_{16} – stopa bezrobocia w %,
X_{17} – liczba ofert pracy ogółem na 1 bezrobotnego,
X_{18} – wskaźnik zatrudnienia ogółem w %,
X_{19} – przeciętne miesięczne wynagrodzenie brutto w relacji do średniej krajowej (Polska = 100),
X_{20} – przeciętna powierzchnia użytkowa mieszkania w m^2 na 1 osobę,
X_{21} – liczba mieszkań na 10 tys. ludności,
X_{22} – liczba studentów na 10 tys. ludności,
X_{23} – liczba praktyk lekarskich w miastach na 10 tys. ludności,
X_{24} – liczba praktyk lekarskich na wsi na 10 tys. ludności,
X_{25} – liczba ludności na 1 aptekę ogólnodostępną,
X_{26} – liczba osób korzystających ze świadczeń pomocy społecznej na 10 tys. ludności,
X_{27} – stopień wykorzystania miejsc noclegowych w %,
X_{28} – czytelnicy bibliotek publicznych na 1000 ludności,
X_{29} – wypożyczenia księgozbioru na 1 czytelnika,
X_{30} – liczba ludności na 1 miejsce w kinach stałych,
X_{31} – liczba widzów i słuchaczy w teatrach i instytucjach muzycznych na 1000 ludności,

X_{32} – PKB na 1 mieszkańca,
X_{33} – nakłady inwestycyjne na 1 mieszkańca w zł,
X_{34} – drogi publiczne o twardej nawierzchni w km na 100 km^2 powierzchni,
X_{35} – liczba samochodów osobowych zarejestrowanych na 1000 ludności,
X_{36} – liczba ofiar śmiertelnych na 100 wypadków drogowych,
X_{37} – emisja przemysłowych zanieczyszczeń powietrza pyłowych w tonach na 100 km^2,
X_{38} – emisja przemysłowych zanieczyszczeń powietrza gazowych w tonach na 100 km^2,
X_{39} – udział parków narodowych w ogólnej powierzchni w %,
X_{40} – plony z 1 ha zbóż ogółem w dt,
X_{41} – dochody budżetu województwa ogółem na 1 mieszkańca w zł,
X_{42} – wydatki z budżetu województwa ogółem na 1 mieszkańca w zł,
X_{43} – liczba podmiotów gospodarczych na tys. mieszkańców.

Zbiór potencjalnych cech diagnostycznych został utworzony po przeprowadzeniu formalno-merytorycznej analizy badanego zjawisk oraz wynikał z dostępności danych. Cechy dotyczyły różnych obszarów poziomu życia, wśród których można wymienić np.: sytuację demograficzną, rynek pracy, warunki mieszkaniowe, ochronę zdrowia, edukację, kulturę i turystykę, komunikację, ochronę środowiska, dochody i wydatki budżetów.

Tabela 1. Wyniki wyboru cech diagnostycznych za pomocą metody parametrycznej w dwóch wariantach

Wariant I		Wariant II	
Cechy centralne	Cechy satelitarne	Cechy centralne	Cechy satelitarne
X_1	X_{16}, X_{17}, X_{26}, X_{28}, X_{30}, X_{34}, X_{36}, X_{37}, X_{38}, X_{42}	X_{16}	X_1, X_{17}, X_{19}, X_{22}, X_{23}, X_{26}, X_{30}, X_{32}, X_{34}
X_{43}	X_{13}, X_{15}, X_{19}, X_{21}, X_{24}, X_{31}, X_{32}, X_{33},	X_{38}	X_6, X_{10}, X_{36}, X_{37},
X_9	X_6, X_{10}, X_{25}	X_{15}	X_{21}, X_{31}, X_{33}, X_{43}
X_{23}	X_{22}	X_{24}	X_{13}
X_{40}	X_{39}	X_{41}	X_{42}
X_{14}	–	X_9	–
X_{27}	–	X_{14}	–
X_{30}	–	X_{25}	–
X_{41}	–	X_{27}	–
		X_{28}	–
		X_{39}	–
		X_{40}	–

Źródło: obliczenia własne.

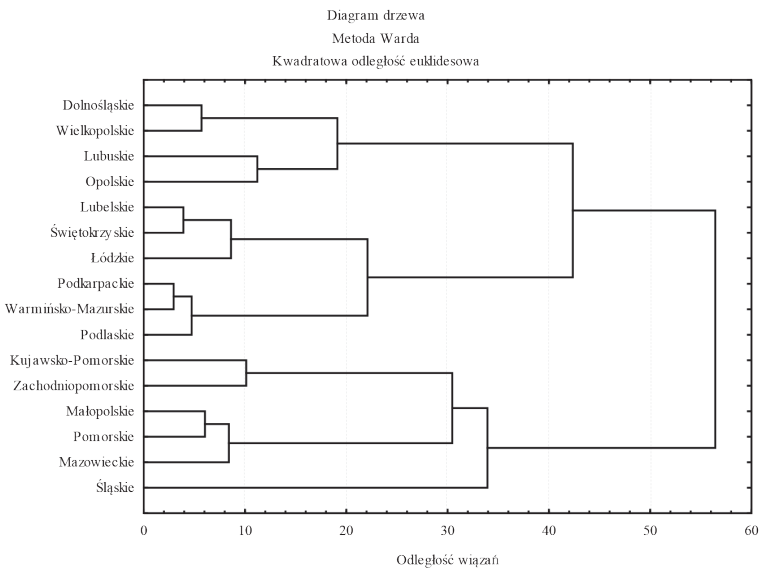
W doborze cech do badania taksonomicznego, w pierwszym kroku, uwzględniono kryterium dyskryminacyjne obiektów, wyrażone za pomocą współczynnika zmienności. Z badania wykluczono te cechy, dla których współczynnik zmienności przyjął wartości nie większe niż 10%. Były to cechy o następujących numerach: 2, 3, 4, 5, 7, 8, 11, 12, 18, 20, 29 i 35. Następnie przeprowadzono normalizację cech dla obu otrzymanych wariantów cech diagnostycznych. Cechy poddano dalszej weryfikacji, badając ich pojemność informacyjną. W tym celu wykorzystano metodę parametryczną w jej klasycznej wersji, z sumą elementów kolumny (lub wiersza) macierzy współczynników korelacji, oraz wariant, w którym sumę zastąpiono medianą.

Po wyznaczeniu macierzy współczynników korelacji oraz przyjęciu wartości progowej współczynnika korelacji 0,5 wyznaczono zbiory cech diagnostycznych dla dwóch wariantów metody parametrycznej: z sumą (wariant I) oraz medianą (wariant II) elementów kolumny (lub wiersza) macierzy współczynników korelacji. Wyniki przedstawiono w tab. 1.

Jako ostateczne zbiory cech diagnostycznych przyjęto zestawy cech centralnych.

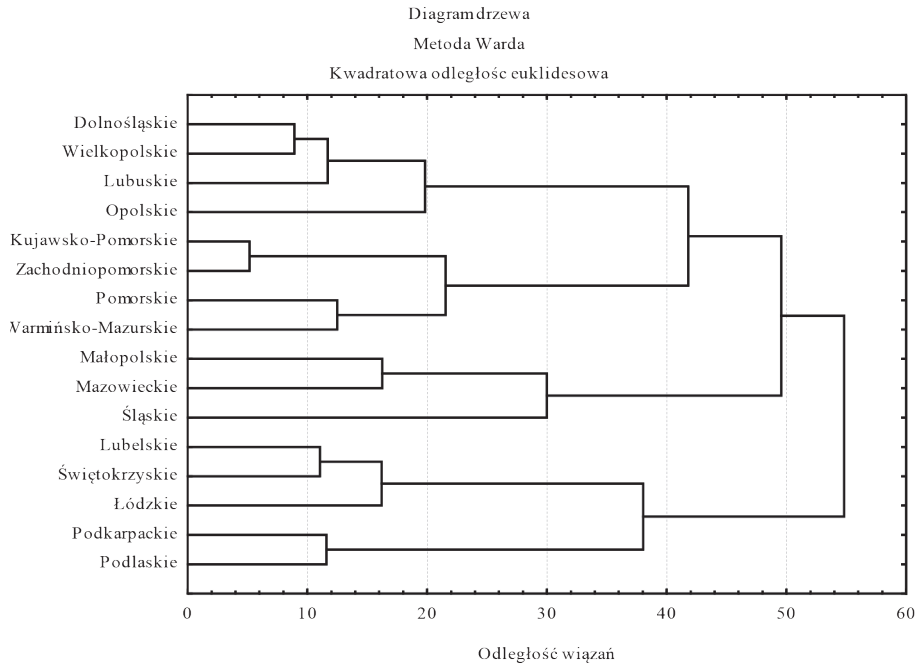
4. Wyniki badania

Wykorzystując otrzymane zbiory cech diagnostycznych, dokonano metodą Warda klasyfikacji województw Polski. Uzyskane dendrogramy zostały przedstawione na rys. 1 i 2.



Rys. 1. Diagram podziału województw Polski na podstawie zbioru cech uzyskanych metodą parametryczną – wariant I

Źródło: opracowanie własne.



Rys. 2. Diagram podziału województw Polski na podstawie zbioru cech uzyskanych metodą parametryczną – wariant II

Źródło: opracowanie własne.

Analizując dendrogramy, przyjęto trzy skupienia województw, a otrzymane grupy przedstawiono w tab. 2.

Tabela 2. Wyniki grupowania województw Polski metodą Warda

Grupowanie województw na podstawie zbioru cech uzyskanych metodą parametryczną					
Wariant I			Wariant II		
grupa I	grupa II	grupa III	grupa I	grupa II	grupa III
Wielkopolskie, Dolnośląskie, Opolskie, Lubuskie	Lubelskie, Świętokrzyskie, Łódzkie, Podkarpackie, Podlaskie, Warmińsko--mazurskie	Zachodniopomorskie, Kujawsko-Pomorskie, Pomorskie, Małopolskie, Pomorskie, Mazowieckie, Śląskie	Warmińsko-mazurskie, Pomorskie, Zachodniopomorskie, Kujawsko-pomorskie, Opolskie, Lubuskie, Wielkopolskie, Dolnośląskie	Małopolskie, Mazowieckie, Śląskie,	Lubelskie, Świętokrzyskie, Łódzkie, Podkarpackie, Podlaskie

Źródło: opracowanie własne.

Otrzymane grupy różnią się między sobą pod względem przynależności województw, składy poszczególnych klas częściowo pokrywają się. Aby określić sku-

teczność otrzymanych grupowań, zweryfikowano je, wyznaczając wartości wskaźników homogeniczności, heterogeniczności i poprawności skupień (tab. 3).

Tabela 3. Wartości wskaźników homogeniczności, heterogeniczności i poprawności skupień

Wskaźniki	Wariant I	Wariant II
Homogeniczności skupień	49,786	251,938
Heterogeniczności skupień	95,901	1334,685
Poprawności skupień	0,519	0,189

Źródło: obliczenia własne.

Analizując wyniki dotyczące efektywności grupowań przedstawione w tab. 3, można stwierdzić, że wykorzystując klasyczny wariant metody parametrycznej, otrzymano lepszy wynik w zakresie homogeniczności skupień, natomiast klasyfikacja otrzymana metodą Warda na podstawie zbioru cech uzyskanych metodą parametryczną wariant II (z medianą) dała dużo lepsze rezultaty pod względem zarówno heterogeniczności, jak i poprawności grupowania. Otrzymane wyniki potwierdza analiza dendrogramów.

Klasyfikacja oparta na metodzie parametrycznej (wariant II) wyodrębniła trzy skupienia województw pod względem poziomu życia. Do pierwszej grupy należy osiem województw. Klasa ta charakteryzuje się korzystnymi wielkościami średnich, w porównaniu do średnich ogólnych, odnoszących się do następujących cech: przeciętne miesięczne wydatki na 1 osobę, emisja przemysłowych zanieczyszczeń powietrza gazowych w tonach na 100 km^2, plony z 1 ha zbóż ogółem w dt, dochody budżetu województwa ogółem na 1 mieszkańca w zł. Ponadto w klasie tej zaobserwowano najniższą liczbę zgonów na 1000 ludności. Jednocześnie odnotowano najwyższy wskaźnik zgonów niemowląt na 1000 urodzeń żywych oraz najwyższą stopę bezrobocia. Do niekorzystnych wartości należą także średnie dotyczące: liczby ludności na 1 aptekę ogólnodostępną oraz stopnia wykorzystania miejsc noclegowych i udziału powierzchni parków narodowych w ogólnej powierzchni województwa. Grupa druga zawiera trzy województwa: małopolskie, mazowieckie i śląskie. Dobra sytuacja odnośnie do tej grupy występuje w przypadku takich średnich wartości cech, jak: przeciętne miesięczne wydatki na 1 osobę, stopa bezrobocia w %, stopień wykorzystania miejsc noclegowych w %. Negatywny wpływ na poziom życia w tej klasie ma dosyć wysoka średnia wartość odnosząca się do: zgonów niemowląt na 1000 urodzeń żywych, liczby ludności na 1 aptekę ogólnodostępną, emisji przemysłowych zanieczyszczeń powietrza gazowych w tonach na 100 km^2 (dwukrotnie wyższa średnia w porównaniu ze średnią ogólną i najwyższa wśród wszystkich klas). Trzecia klasa to pięć obiektów, których dobra sytuacja pod względem poziomu życia wynika z: niskiej wartości średniej dotyczącej wskaźnika zgonów niemowląt na 1000 urodzeń żywych, wysokiej średniej dla liczby praktyk lekarskich na wsi na 10 tys. ludności, dobrego dostępu do aptek oraz niskiej średniej emisji przemysło-

wych zanieczyszczeń powietrza gazowych w tonach na 100 km^2. Natomiast niekorzystnie na badane zjawisko wpływają: wysoki średni wskaźnik zgonów na 1000 ludności, wysoka stopa bezrobocia, niski stopień wykorzystania miejsc noclegowych w % oraz najniższy wskaźnik wysokości plonów.

Podział województw Polski metodą Warda na podstawie zbioru cech diagnostycznych otrzymanych metodą parametryczną – wariant II pokazuje rys. 3.

Rys. 3. Podział województw Polski metodą Warda na podstawie zbioru cech diagnostycznych otrzymanych metodą parametryczną – wariant II

Źródło: opracowanie własne.

5. Podsumowanie

W pracy rozważono wykorzystanie popularnej metody doboru cech diagnostycznych – parametrycznej metody doboru cech – w badaniu taksonomicznym, w dwóch wariantach: w wariancie I przyjęto sumę elementów kolumny macierzy współczynników korelacji, natomiast wariant II uwzględniał medianę tych elementów. Zbadano także wpływ wyników otrzymanych w poszczególnych wariantach na efektywność klasyfikacji. Przedstawione podejście zilustrowano przykładem dotyczącym klasyfikacji województw Polski w roku 2009 pod względem poziomu życia mieszkańców. W każdej klasyfikacji dokonanej metodą Warda wyłoniono trzy klasy województw i zbadano efektywność otrzymanych podziałów, wykorzystując wskaźniki homogeniczności, heterogeniczności oraz poprawności grupowań, w których rolę środków

ciężkości odgrywała mediana Webera. Zastosowanie mediany Webera w ocenie jakości klasyfikacji pozwoliło na uzyskanie większej odporności na wpływ obserwacji odstających i traktowanie zbioru cech diagnostycznych jako całości w całej analizie. Klasyfikacja na podstawie II wariantu metody parametrycznej (z medianą) dała lepsze rezultaty w porównaniu z wariantem I (z sumą), które dotyczyły heterogeniczności i poprawności skupień.

Przeprowadzone badanie wykazało, iż metody klasyfikacji są skutecznym narzędziem w ocenie poziomu życia mieszkańców, a wyniki uzyskane za pomocą różnych metod doboru cech do badania taksonomicznego mają wpływ na jakość klasyfikacji.

Literatura

Balicki A., *Statystyczna analiza wielowymiarowa i jej zastosowania społeczno-ekonomiczne*, Wydawnictwo Uniwersytetu Gdańskiego, Gdańsk 2009.

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

Młodak A., *Analiza taksonomiczna w statystyce regionalnej*, Difin, Warszawa 2006.

Panek T., *Statystyczne metody wielowymiarowej analizy porównawczej*, Szkoła Główna Handlowa w Warszawie, Warszawa 2009.

Panek T. (red.), *Statystyka społeczna*, Polskie Wydawnictwo Ekonomiczne, Warszawa 2007.

Ward J.H., *Hierarchical grouping to optimize an objective function*, „Journal of the American Statistical Association" 1963, no 58.

Zeliaś A. (red.), *Taksonomiczna analiza przestrzennego zróżnicowania poziomu życia w Polsce w ujęciu dynamicznym*, Wydawnictwo Akademii Ekonomicznej w Krakowie, Kraków 2000.

THE CLASSIFICATION'S EFFICIENCY FOR THE PARAMETRIC METHOD OF FEATURE SELECTION

Summary: The article presents two variants of the parametric feature selection method: with the sum and median of the elements in the column of the correlation coefficients matrix. The classic version has some disadvantages. The application of the median in place of the sum gives the possibility of the elimination of these disadvantages. This reduces sensitivity to outliers correlation coefficients. The aim of the paper is to determine the effect of the results of two different approaches for parametric selection method for the classification's efficiency. Three groups were extracted in the classifications by means of the Ward's method. 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.

Keywords: parametric metod of feature selection, classification's efficiency.