Joanna Kisielińska[1]
Katedra Ekonomiki Rolnictwa i Międzynarodowych Stosunków Gospodarczych
Szkoła Główna Gospodarstwa Wiejskiego
Warszawa

Bezwzorcowa klasyfikacja obiektów w ekonomice rolnictwa

Cluster analysis in the agricultural economics

Synopsis. W artykule przedstawiono elementy analizy skupień, zwracając szczególną uwagę na ocenę uzyskanej klasyfikacji. Dokonano ponadto przeglądu publikacji, w których prezentowano badania wykorzystujące analizę skupień w ekonomice rolnictwa.

Słowa kluczowe: analiza skupień, ekonomika rolnictwa

Abstract. Elements of cluster analysis are presented in the article. Particular attention was paid to the appraisal of obtained classification. Moreover the research publications using cluster analysis in agricultural economics are reviewed.

Key words: cluster analysis, agricultural economics

Wstęp

Właściwa typologia obiektów opisanych wieloma cechami może mieć istotne znaczenie w przypadku dokonywania oceny badanej zbiorowości bądź poszukiwania zależności w obrębie niej obowiązujących. Do jej wykonania wykorzystywane są metody klasyfikacji bezwzorcowej, zwane także analizą skupień. Prowadząc ją jednak trzeba mieć na uwadze trudności w określeniu nieznanej struktury zbiorowości, zwłaszcza gdy wziętych pod uwagę cech diagnostycznych jest wiele. Nie ma wówczas możliwości graficznego przedstawienia danych, co w istotny sposób utrudnia interpretację i ocenę otrzymanej klasyfikacji. Stosowane metody matematyczno-numeryczne dokonują podziału zbiorowości na podgrupy, nawet jeżeli skupienia w niej nie występują.

Celem niniejszego artykułu jest sprawdzenie skuteczności wybranych metod analizy skupień dla celowo opracowanych danych liczbowych. Hipotetyczne dwie zbiorowości reprezentują przypadek klasowej struktury zbioru z wyraźnie odrębnymi skupieniami oraz przypadek całkowitego ich braku, czyli zbioru jednorodnego. Ponieważ znana jest struktura obydwu zestawów danych, więc wiadomo, jaka klasyfikacja jest poprawna. Uzyskane różnymi metodami podziały zostaną następnie ocenione przy pomocy wskaźnika Rousseeuwa stanowiącego miarę jakości klasyfikacji. Obliczenia wykonano przy użyciu pakietu Statistica i Statistica Neural Networks.

Artykuł ma jednak nie tylko cel metodologiczny. W ostatniej jego części dokonany zostanie przegląd publikacji z zakresu ekonomiki rolnictwa, w których wykorzystano metody klasyfikacji bezwzorcowej. Celem jego jest sprawdzenie, jakie metody preferują

[1] Dr hab., e-mail: joanna_kisielinska@sggw.pl

badacze, do jakich zagadnień je stosują, jak prowadzą ocenę uzyskanych rezultatów. Autorka ma nadzieję, że zarówno część metodologiczna jak i przeglądowa może pomóc osobom pragnącym stosować analizę skupień.

Metody bezwzorcowej klasyfikacji obiektów

Bezwzorcowa klasyfikacja obiektów należy do metod wielowymiarowej analizy danych, zajmującej się badaniem obiektów opisanych wieloma cechami. Informacje o obiektach umieszczane są w macierzy zwanej macierzą obserwacji, o liczbie wierszy równej liczbie obiektów i liczbie kolumn równej liczbie cech. Każdy z n obiektów może być traktowany jako punkt w k wymiarowej przestrzeni cech.

Tak zdefiniowana zbiorowość ma zostać podzielona na podgrupy. Podział należy przeprowadzić tak, aby obiekty z jednej grupy (klasy) były do siebie jak najbardziej podobne, a należące do klas odmiennych jak najbardziej różne. Miary podobieństw bądź różnic opierają się na odległościach pomiędzy jednostkami. Aby poszczególne wymiary (cechy) miały taki sam wpływ na odległości, na wstępie cechy są normalizowane.

Postuluje się zwykle, aby klasyfikacja była zupełna, rozłączna i niepusta. Zupełność oznacza, że każdy obiekt przypisano do jakiejś klasy. Rozłączność, że należy on jedynie do jednej klasy. Niepustość wymaga natomiast, aby do każdej klasy należał przynajmniej jeden obiekt.

Warunek niepustości spełnić jest najprościej. Wystarczy pominąć ewentualne klasy puste. Problem z rozłącznością pojawić się może, gdy występują jednostki podobne do obiektów nie tylko z jednej kasy. Naprzeciw problemom w takim przypadku wychodzi klasyfikacja rozmyta, którą traktować można jako oddzielny dział metod klasyfikacyjnych. Zapewnienie zupełności staje się kłopotliwe, gdy w badanej zbiorowości są jednostki odrębne, niepodobne do innych. Najprostszym rozwiązaniem jest tworzenie jednoelementowych klas, które w istocie można również interpretować jako swoiste wykluczenie takich obiektów.

Chcąc przeprowadzić klasyfikację obiektów należy dokonać szeregu rozstrzygnięć mających wpływ na uzyskane rezultaty. Należy do nich dobór obiektów i cech je opisujących, wybór formuł normalizacyjnych, wybór metody klasyfikacji, określenie liczby klas i wreszcie ocena przeprowadzonej klasyfikacji. W artykule niniejszym autorka chce ograniczyć się jedynie do wybranych zagadnień. Bardziej kompletne rozważania odnaleźć można w bogatej literaturze przedmiotu. Wymienić tu można prace Pociechy i współautorów [1988], Sokołowskiego [1992], pracę zbiorową redagowaną przez Gatnara i Walesiaka [Metody.. 2004] i wiele innych.

Wśród metod analizy skupień wyróżnić można następujące grupy:

- hierarchiczne metody aglomeracyjne,
- hierarchiczne metody deglomeracyjne,
- metody obszarowe,
- metody optymalizujące wstępny podział obiektów,
- sieci neuronowe.

W hierarchicznych metodach aglomeracyjnych wstępnie przyjmuje się liczbę klas równą liczbie obiektów, a następnie łączy się klasy najbardziej do siebie podobne, redukując w każdym kroku liczbę klas o 1, aż do uzyskania jednej klasy obejmującej wszystkie

obiekty. Miarą podobieństwa obiektów i klas są odległości między nimi. Różne odmiany omawianych metod różnią się przede wszystkim sposobem wyznaczania odległości między klasami.

Popularne metody z tej grupy to:

- metoda pojedynczego wiązania (zwana też metodą najbliższego sąsiedztwa); odległość między dwoma skupieniami jest określona przez odległość między dwoma najbliższymi obiektami należącymi do różnych skupień,
- metoda pełnego wiązania (najdalszego sąsiedztwa); odległością między skupieniami jest największa z odległości między dwoma dowolnymi obiektami należącymi do różnych skupień,
- metoda Warda; odległością pomiędzy skupieniami jest wartość, o jaką zwiększy się wariancja wewnątrzgrupowa po połączeniu grup.

W hierarchicznych metodach deglomeracyjnych na początku zakłada się istnienie jednej klasy. W każdym kolejnym kroku liczbę klas zwiększa się o jeden, aż do uzyskania liczby klas równej liczbie obiektów. Różne odmiany metod różnią się sposobem wyboru klasy dzielonej. W klasie tej określane są dwa obiekty leżące najdalej od siebie i na podstawie odległości od nich dokonuje podziału.

W metodach obszarowych przestrzeń dzielona jest na rozłączne podobszary. Obiekty znajdujące się w tych obszarach zalicza się do jednej klasy. Stosuje się różne rodzaje podobszarów. Mogą to być wielowymiarowe kule czy prostopadłościany.

Metody optymalizujące wstępny podział obiektów startują od zadanego podziału zbiorowości na klasy. Wymaga to założenia liczby skupień i przypisania do nich obiektów (np. w sposób losowy). Można również startować od podziału przeprowadzonego inną metodą. Zadaniem metod optymalizacyjnych jest poprawa tego podziału.

Najczęściej stosowaną z omawianej grupy jest metoda k-średnich. W metodzie tej dla każdego skupienia obliczany jest środek ciężkości (zwany centroidem). Następnie obiekty przenoszone są do klas o najbliższych środkach ciężkości. Powstają nowe klasy, dla których ponownie oblicza się środki ciężkości. Procedura kończy się, gdy nie następuje zmiana klas dla obiektów.

Do wykonania klasyfikacji bezwzorcowej można użyć również sieci neuronowych. Są to tzw. sieci samoorganizujące się, których odmianą jest sieć Kohonena. Sieć jest uczona takiej samoorganizacji. Uczenie sieci polega na takim doborze wag neuronów na wyjściu, aby reagowały (rozpoznawały) na określone typy wzorców (w naszym wypadku określony typ obiektów).

Weryfikacja uzyskanej klasyfikacji

Uzyskany podział zbioru obiektów na klasy musi podlegać weryfikacji. Jego jakość w prosty sposób ocenić można jedynie dla zadania dwuwymiarowego, kreśląc wykres. Przy większej wymiarowości problemu wizualizacja jest niemożliwa (bądź bardzo utrudniona).

Weryfikacje przeprowadzić można stosując różne metody i wybrać jedynie te, które dały wyniki zgodne [Metody... 2004]. Można także przeprowadzić walidację wyników przez replikację, polegającą na wylosowaniu ze zbioru obiektów dwóch prób i ocenie zgodności otrzymanej klasyfikacji. Sposób ten może sprawiać jednak kłopot, gdy badana zbiorowość jest mało liczna.

Prawidłowość klasyfikacji pozwala ocenić syntetyczny miernik zaproponowany przez Rousseeuwa [1987] i rekomendowany przez Gatnara i Walesiaka [Metody... 2004]. Wskaźnik Rousseeuwa obliczany jest na podstawie średnich odległości każdego obiektu od obiektów z klasy rodzimej oraz obiektów z klasy mu najbliższej. Niech liczba klas w ocenianym podziale równa się *N*. Dla każdej klasy *P* i każdego obiektu *i* należącego do tej klasy obliczany jest indeks[2] *S(i)* według wzoru:

$$S(i) = \frac{b(i) - a(i)}{max[a(i); b(i)]} \qquad (1)$$

gdzie:
a(i) jest średnią odległością obiektu *i* od pozostałych obiektów z klasy *P*,
b(i) jest średnią odległością obiektu *i* od obiektów z klasy *R* położonej najbliżej tego obiektu.
Formuła obliczenia odległości *a(i)* jest następująca:

$$a(i) = \frac{1}{(n_P - 1)} \sum_{\substack{k \in P \\ k \neq i}} d_{ik} \qquad (2)$$

gdzie:
d_{ik} jest odległością obiektu *i* od obiektu *k* należącego do klasy *P*,
n_P jest liczebnością klasy *P*.
Natomiast *b(i)* oblicza się jako:

$$b(i) = \min_R [d_{iR}] \quad \text{a} \quad d_{iR} = \frac{1}{n_R} \sum_{k \in R} d_{ik} \qquad (3)$$

gdzie:
d_{iR} jest średnią odległością obiektu *i* od obiektów z klasy $R \neq P$,
d_{ik} jest odległością obiektu *i* należącego do klasy *P* od obiektu *k* należącego do klasy *R*,
n_R jest liczebnością klasy *R*.
Indeks *S(i)* zostaje obliczony dla wszystkich elementów klasyfikowanej zbiorowości. Jego wartości mieszczą się w przedziale <-1;1>. Im bliżej jedynki tym silniejsza przynależność obiektu *i* do klasy *P*, czyli do klasy, do której obiekt przypisano.
Indeksy *S(i)* są potem uśredniane dla klas, a następnie dla całej zbiorowości. Średni indeks dla klasy *P* oznaczony zostanie jako *S(P)*, zaś uśredniony dla całej zbiorowością podzielonej zgodnie z ocenianą klasyfikacją na *N* klas, jako *S*. Formuły na ich wyznaczenie są następujące:

$$S(P) = \frac{1}{n_P} \sum_{i \in P} S(i) \quad \text{a} \quad S = \frac{1}{N} \sum_P S(P) \qquad (4)$$

Gatnar i Walesiak [Metody.. 2004] podają interpretację wskaźnika *S* zaproponowaną przez Kaufmana i Rousseeuwa [1990], którą przedstawiono w tabeli 1.

[2] Zwany w literaturze wskaźnikiem Silhouette.

Tabela 1. Interpretacja wartości wskaźnika *S* dla przeprowadzonej klasyfikacji

Table 1. Interpretation of the *S* score for a classification

Wartość *S*	Interpretacja
(0,70; 1,00>	silna struktura klas
(0,50; 0,70>	poważna struktura klas
(0,25; 0,50>	słaba struktura klas (należy zastosować inna metodę)
poniżej 0,25	nie wykryto struktury klas

Źródło: [Metody... 2004] za [Kaufman i Rousseeuw 1990].

Weryfikacja rezultatów uzyskanej klasyfikacji jest konieczna, zwłaszcza w problemach o wymiarowości większej niż 2. Niestety popularne gotowe oprogramowanie statystyczne nie daje miar pozwalających takiej oceny dokonać. Przedstawione w artykule indeksy *S* obliczone zostały przy pomocy prostej procedury napisanej w języku Visual Basic for Excel, przy zastosowaniu odległości euklidesowej.

Bezwzorcowa klasyfikacja przykładowych zbiorowości

W celu pokazania silnych i słabych stron poszczególnych metod przeprowadzona zostanie klasyfikacje dwóch przykładowych zestawów danych obejmujących po 30 obiektów opisanych 2 cechami (co umożliwia graficzne zobrazowanie struktury badanej zbiorowości). W pierwszym zestawie występują dwa wyraźne skupienia, drugi natomiast jest zbiorem w przybliżeniu jednorodnym, w którym skupienia nie występują. Obydwa zestawy podano w tabeli 2, zawierającej również wykresy. Do klasyfikacji wykorzystana została metoda Warda, metoda k-średnich i sieć Kohonena. Klasyfikacja prowadzono dla danych znormalizowanych poprzez standaryzację.

W tabeli 3 przedstawione zostały wyniki klasyfikacji obydwu zestawów danych uzyskane trzema metodami. Dla każdej metody podano wykres pokazujący sposób przypisania poszczególnych obiektów do klas oraz wartość wskaźnika Rousseeuwa dla otrzymanych klasyfikacji. Dla metody Warda podano ponadto dendrogram.

W przypadku stosowania metody Warda decyzja o założeniu istnienia określonej liczby skupień jest decyzją subiektywną. Pewnym wskazaniem może być tu wykres odległości wiązania dla poszczególnych etapów aglomeracji. Jeżeli na pewnym etapie wykres ten zaczyna stromo unosić się do góry, należy przyjąć, że połączone zostają skupienia leżące stosunkowo daleko od siebie. Graniczną odległość wiązania należy przyjąć poniżej tego gwałtownego wzrostu. Przenosząc poziom ten na dendrogram łatwo odczytać liczbę klas i sposób przypisania do nich obiektów.

W metodzie k-średnich należy z góry założyć liczbę klas. Dla zadań wielowymiarowych dobranie odpowiedniej liczby może być kłopotliwe. Gdyby liczba klas była znana, prawdopodobnie znana byłaby również przynależność do nich obiektów i niepotrzebne byłoby stosowanie analizy skupień. Wiele uwagi problemowi doboru liczby klas poświęcili Gatnar i Walesiak [Metody... 2004].

W przypadku używania do klasyfikacji sieci Kohonena należy również założyć z góry liczbę klas, która odpowiada liczbie neuronów wyjściowych. Można przyjąć wartość tę z pewnym nadmiarem, ponieważ nie wszystkie neurony muszą być wykorzystane. Inaczej mówiąc w sieci mogą pozostać klasy puste. Jest to zasadnicza przewaga sieci neuronowej

nad metodą k-średnich, która takich sytuacji nie dopuszcza. W wyniku użycia tej metody klas będzie tyle, ile zostało z góry założone[3].

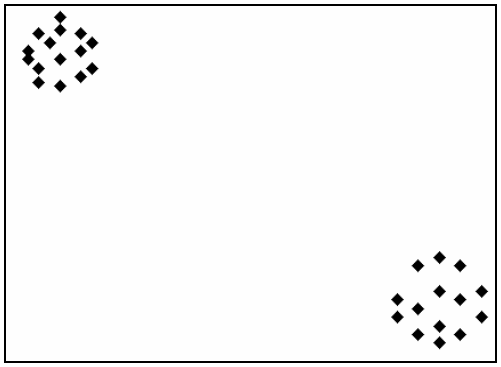
Dla metody k-średnich założono 9 klas i tyle samo neuronów wyjściowych w sieci Kohonena. Jest to wartość przykładowa, większa od liczby skupień w zestawie 1 i mniejsza od liczby obiektów w obydwu zestawach.

Tabela 2. Przykładowe zestawy danych wykorzystane do klasyfikacji

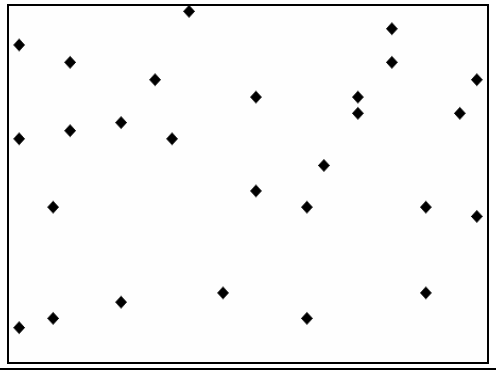
Table 2. Sample data sets used for classification

Zestaw 1		Zestaw 2	
x	y	x	y
3,55	4,4	3	1,1
3,55	4,3	3	3,3
3,6	4,2	3	4,4
3,6	4,6	3,2	1,2
3,6	4,03	3,6	1,4
3,65	4,5	3,2	2,5
3,7	4	3,3	3,4
3,7	4,8	3,3	4,2
3,7	4,3	3,6	3,5
3,7	4,65	3,8	4
3,8	4,6	3,9	3,3
3,8	4,1	4	4,8
3,8	4,4	4,2	1,5
3,85	4,2	4,4	2,7
3,85	4,5	4,4	3,8
5,3	1,3	4,7	1,2
5,3	1,5	4,7	2,5
5,4	1,1	4,8	3
5,4	1,9	5	3,6
5,4	1,4	5	3,8
5,5	1,2	5,2	4,2
5,5	1	5,2	4,6
5,5	2	5,4	1,5
5,5	1,6	5,4	2,5
5,5	1,6	5,6	3,6
5,6	1,1	5,7	4
5,6	1,5	5,7	2,4
5,6	1,9	5,9	1,1
5,7	1,3	6	3,5
5,7	1,6	6	4

Zestaw 1:



Zestaw 2:



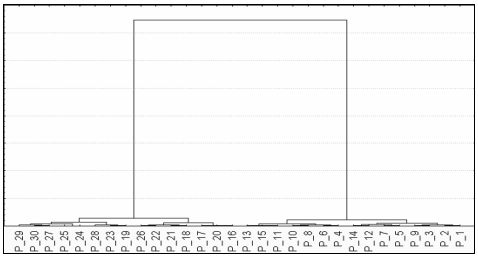
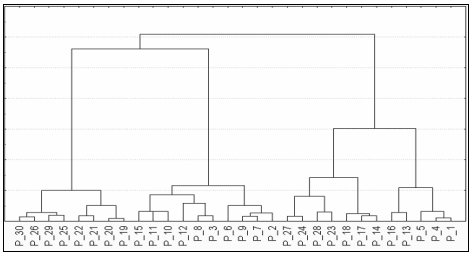
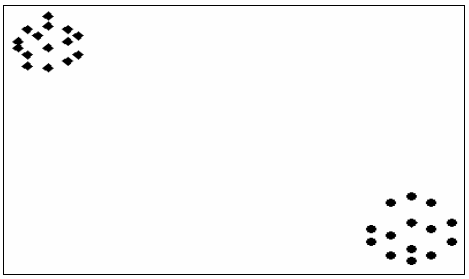
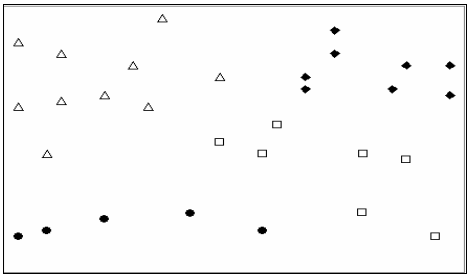
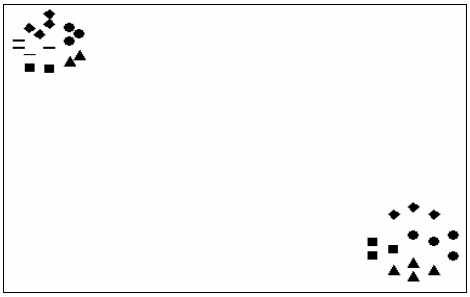
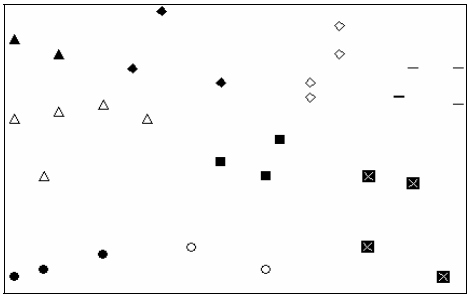
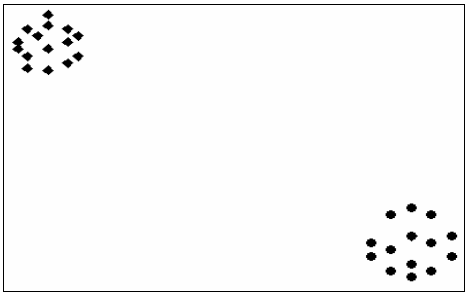
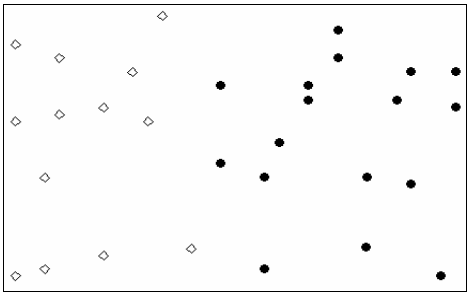
Źródło: opracowanie własne.

Jeśli chodzi o klasyfikację pierwszego zestawu (zestawu, w którym występują wyraźne dwa skupienia) zadanie klasyfikacji zostało wykonane dobrze metodą Warda i siecią Kohonenena. Dendrogram przedstawiony w tabeli 3 wyraźnie wskazuje, że występują dwie klasy i bez trudu odczytać można, które obiekty do nich należą. W sieci Kohonena wykorzystane zostały jedynie dwa neurony wyjściowe. Pozostałe nie rozpoznawały żadnego wzorca, co oznacza, że nie były potrzebne, bo są jedynie dwie klasy. Klasyfikacja przy użyciu sieci jest zgodna z klasyfikacją metodą Warda.

[3] Warto w tym miejscu przypomnieć, że metoda k-średnich należy do grupy metod optymalizujących wstępny podział zbiorowości.

Tabela 3. Wyniki klasyfikacji obydwu zestawów danych trzema wybranymi metodami

Table 3. Results of classification of the two sets of data by using three selected methods

Zestaw 1	Zestaw 2
klasyfikacja metodą Warda: *S*=0,896	klasyfikacja metodą Warda: *S*=0,434
	
	
klasyfikacja metodą k-średnich: *S*=0,421	klasyfikacja metodą k-średnich: *S*=0,414
	
klasyfikacja siecią Kohonena: *S*=0,896	klasyfikacja siecią Kohonena: *S*=0,348
	

Uwaga: na wykresach elementy różnych skupień oznaczone zostały różnymi symbolami.

Źródło: badania własne.

Na poprawność rozwiązania zadania wskazuje wskaźnik Rousseeuwa, przyjmujący dla obydwu metod wartość 0,896.

Klasyfikacja zestawu 1 uzyskana metodą k-średnich przy założonej nadmiarowo liczbie klas jest zdecydowanie niepoprawna. Wyraźnie występujące skupienia zostały podzielone na 5 i 4 klasy. Wskaźnika Rousseeuwa ma wartość 0,421, która oznacza słabą strukturę klas i wskazanie na zastosowanie innej metody. Oczywistym jest, że gdyby przyjęto liczbę klas równą 2, metoda k-średnich poprawnie by je rozpoznała. Przypomnieć jednak należy, że liczba skupień nie jest najczęściej z góry znana. W tym przypadku użycie metody k-średnich bez wstępnych analiz jest ryzykowne.

Wszystkie klasyfikacje zestawu 2 oceniane wskaźnikiem Rousseeuwa wykazują słabą strukturę klas, co jest zgodne ze stanem faktycznym. W metodzie Warda z wykresu odległości wiązania wynika, że należałoby wyróżnić 4 klasy, choć oczywiście wskazanie jest znacznie mniej jednoznaczne, niż w przypadku zestawu 1. Do 9 klas metodą k-średnich przypisane zostały obiekty leżące blisko siebie. Natomiast sieć Kohonena mając możliwość wyróżnienia maksymalnie 9 skupień wyodrębniła jedynie 2.

Dla zestawu 2 wskaźnik Rousseeuwa ma najwyższą wartość dla klasyfikacji otrzymanej metodą Warda, następnie metodą k-średnich, najniższą zaś dla sieci Kohonena. Należy jednak zwrócić uwagę, że są to wartości zbliżone, mimo że we wszystkich klasyfikacjach jest inna liczba klas. Może to wskazywać na jednorodność zbioru danych, choć w przypadku braku struktury klas wartości wskaźnika Rousseeuwa są poniżej 0,25 (tabela 1).

Wydaje się, że metoda Warda dzięki dendrogramowi jest najbezpieczniejsza zarówno w przypadku, gdy klasy są bardzo wyraźne, jak przy ich braku. Stosowanie jej może być jednak kłopotliwe dla bardzo licznych zbiorów danych. Metoda k-średnich przy niewłaściwie dobranej liczbie klas jest całkowicie zawodna. Dobrym narzędziem wydaje się sieć Kohonena. Znakomicie spisała się dla zestawu 1, w którym występują skupienia, w zestawie 2 natomiast przeprowadzona klasyfikacja ma najniższą wartość wskaźnika Rousseeuwa, co odpowiada rzeczywistości ze względu na brak klas.

We wszystkich klasyfikacjach zestawu 2 wartość wskaźnika Rousseeuwa jest z przedziału, który należy interpretować jako słabą strukturę klas i wskazanie na użycie innej metody. W zestawie tym jednak klasy naprawdę nie występują, należałoby więc oczekiwać rezultatu poniżej 0,25. Być może granica równa 0,25 jest nieco zaniżona.

Podstawowy wniosek płynący z przedstawionych badań to konieczność stosowania miar jakości klasyfikacji. Jeżeli wskaźnik Rousseeuwa ma wartości wysokie (bliskie jedynki) przyjąć można, że zadanie wyodrębnienia skupień w rozpatrywanej zbiorowości wykonane zostało poprawnie. Niskie wartości (rzędu 0,4) mogą wskazywać na brak struktury klas bądź zawodność stosowanej metody dla danego przypadku. Rozstrzygnięcie, która sytuacja ma miejsce, wymaga zastosowania innych metod klasyfikacji. Słabe rezultaty otrzymane różnymi metodami można odczytywać jako brak skupień.

Ostateczne wnioski na temat słabych i silnych stron poszczególnych metod wymagają dalszych badań, zwłaszcza na zbiorach o wymiarach większych niż dwa. Godna polecenia wydaję się być sieć neuronowa, która potrafi wykryć skupienia bez konieczności zakładania ich liczby. W przypadku braku klas natomiast rezultat klasyfikacji wyraźnie na to wskazuje.

Przykłady zastosowania klasyfikacji bezwzorcowej w ekonomice rolnictwa

W tabeli 4 przedstawiono przykłady zastosowania bezwzorcowej klasyfikacji obiektów w zagadnieniach z zakresu ekonomiki rolnictwa, podanych w kolejności wynikającej z roku publikacji. W pierwszej z nich [Wysocki 1999] przeprowadzono identyfikację strategii marketingowych stosowanych przez firmy mleczarskie. Do klasyfikacji wykorzystano informacje ze 143 ankiet. Na wstępie przeprowadzono analizę korespondencji, której celem było przekształcenie 28 cech jakościowych w 22 cechy ilościowe. Następnie stosując aglomeracyjną metodę Warda wyodrębniono 3 klasy. Przyjęta liczba klas wynikała z analizy wskaźnika agregacji, pokazującego wzrost zmienności wewnątrzklasowej w procesie łączenia. Analiza wyników obejmowała wyodrębnienie cech stanowiących najważniejsze determinanty uzyskanej klasyfikacji. W tym celu porównywano frakcje i średnie wewnątrz klas z frakcjami i średnimi ogólnymi.

Kisielińska [2003] przeprowadziła klasyfikację zbiorowości gospodarstw indywidualnych opisanych wskaźnikami finansowymi stosując sieć neuronową Kohonena. Pełny zestaw obejmował 16 cech i nie pozwolił wyodrębnić sensownych skupisk gospodarstw (oceny dokonano badając rozstępy cech w klasach). Lepsze rezultaty uzyskano wykorzystując ograniczone zestawy wskaźników, oddzielnie dla wskaźników płynności, rentowności, obrotowości i sprawności.

Szczepaniak i Wigier [2003] identyfikowali czynniki wpływające na innowacyjność małych i bardzo małych firm przemysły spożywczego. Badaniami objęto 36 przedsiębiorstw. Analizę skupień przeprowadzono metodą k-średnich na podstawie 8 cech określających innowacyjność w zakresie wprowadzanych produktów i zmian organizacyjnych. Dla wyodrębnionych 3 skupień przeprowadzono ocenę trafności doboru zmiennych stosując analizę wariancji. Nie uzasadniono doboru liczby klas.

Błażejczyk-Majka i Kala [2004] stosowali aglomeracyjną metodę najdalszego sąsiedztwa w celu porównania zasobów siły roboczej w państwach UE i Polsce. Na podstawie standaryzowanych 3 cech przeprowadzono analizę skupień oddzielnie dla trzech lat (1990, 1995 i 1999). Liczbę skupień określono przyjmując podział na poziomie ½ największej odległości wiązania.

W publikacji kolejnej [Błażejczyk-Majka i Kala 2005] ten sam zespół autorów zastosował analizę skupień do badania struktury użytków rolnych dla 15 państw UE (Belgię połączono z Luksemburgiem) oraz Polski. Zasoby ziemi charakteryzowano przy pomocy 5 cech. Wykorzystując odległość miejską[4] zastosowano dwie metody aglomeracyjne, pojedynczego i pełnego wiązania.

Majewski [2005] badając regionalne zróżnicowanie skupu mleka przeprowadził klasyfikację 16 województw. Województwa opisano sześcioma cechami, do klasyfikacji zaś zastosowano aglomeracyjną metodę Warda. Uzyskane klasy scharakteryzowano średnimi wartościami cech.

Poczta [2005] natomiast stosując metodę Warda klasyfikowała kraje OECD biorąc pod uwagę poziom i strukturę wsparcia finansowego rolnictwa. Charakterystykę stanowiło 10 cech obejmujących 5 wskaźników wsparcia dla rolnictwa, opracowywanych i

[4] Odległością miejską między dwoma punktami jest pierwiastek z sumy wartości bezwzględnych różnic wszystkich współrzędnych (w przypadku odległości Euklidesowej obliczany jest pierwiastek z sumy kwadratów tych różnic).

publikowanych przez OECD, uzupełnionych o zmienne określające znaczenie rolnictwa w gospodarce danego kraju. Liczba klas wynikła z podziału dendrogramu na poziomie ½ maksymalnej odległości.

Tabela 4. Przykłady zastosowania klasyfikacji bezwzorcowej w problemach z zakresu ekonomiki rolnictwa

Table 4. Examples of cluster analysis in agricultural economics

Autor	Rok publikacji	Problem	Obiekty	Cechy	Metoda
F. Wysocki	1999	strategie marketingowe w polskim przemyśle mleczarskim	143 polskich mleczarni	34 cechy	metoda Warda
J Kisielińska	2003	zróżnicowanie gospodarstw rolniczych	998 gospodarstw	16 wskaźników finansowych	sieć neuronowa Kohonena
I. Szczepaniak, M. Wigier	2003	innowacyjność małych i bardzo małych firm przemysłu spożywczego	36 przedsiębiorstw	8 cech	metoda k-średnich
L. Błażejczyk-Majka, R. Kala	2004	zasoby siły roboczej rolnictwa polskiego i krajów UE	15 państw członków UE w 2000 i Polska	3 cechy	metoda najdalszego sąsiada
L. Błażejczyk-Majka, R. Kala	2005	charakterystyka użytków rolnych wybranych państw Unii Europejskiej	15 państw członków UE w 2000 i Polska	5 cech opisujących zasoby ziemi	metoda pojedynczego i pełnego wiązania
J. Majewski	2005	regionalne zróżnicowanie skupu mleka w Polsce	16 województw	6 cech	metoda Warda
A. Poczta	2005	poziom i struktura wsparcia finansowego rolnictwa w krajach OECD	13 państw	10 cech	metoda Warda
M.Adamowicz, A Nowak	2006	typy wiejskich gospodarstw domowych	611 wiejskich gospodarstw domowych z województwa lubelskiego	4 cechy	metoda k-średnich
L. Osowska	2006	typologia funkcjonalna obszarów wiejskich Pomorza Środkowego	65 gmin wiejskich i wiejsko-miejskich Pomorza Zachodniego	7 cech	metoda Warda
R.Pietrzykowski P. Kobus	2008	analiza wyników ekonomiczno-produkcyjnych gospodarstw rolnych wybranych państw UE	25 państw	4 cechy	metoda pojedynczego wiązania, Warda i k-średnich
W. Poczta, J. Średzińska, K. Pawlak	2008	analiza sytuacji finansowej gospodarstw rolnych państw UE	23 państwa	14 cech	metoda Warda

Źródło: opracowanie własne.

Typy wiejskich gospodarstw domowych wyodrębniali Adamowicz i Nowak [2006]. Zbiorowość obejmującą 611 gospodarstw opisano 4 cechami (wykształcenie głowy rodziny,

powierzchnia gospodarstwa, dochód na 1 osobę, wiek głowy rodziny). Zastosowano metodę k - średnich przyjmując 5 skupień (porównano również klasyfikację z 3 i 7 skupieniami).

W pracy Osowskiej [2006] opracowano typologię obszarów gmin Pomorza Zachodniego. Badaniem objęto 65 gmin wiejskich i wiejsko-miejskich opisanych 7 cechami. Skupienia wyodrębniono metodą Warda i scharakteryzowano średnimi wartościami cech. Nie podano podstawy, na jakiej określono liczbę klas.

Pietrzykowski i Kobus [2008] klasyfikowali wybrane państwa UE na podstawie wyników ekonomiczno-produkcyjnych gospodarstw rolnych. Wzięto pod uwagę plony pszenicy, wydajność mleczną krów, wartość dodaną netto oraz dochód z rodzinnego gospodarstwa rolnego. Analizę skupień przeprowadzono trzema metodami: pojedynczego wiązania, Warda i k-średnich. Autorzy podkreślają, że podział na grupy był arbitralny. Zastosowane metody klasyfikacji dały różne wyniki, co wskazywać może na brak struktury klas w analizowanej zbiorowości.

Poczta, Średzińska i Pawlak [2004] również klasyfikowali państwa UE. Zastosowali metodę Warda dla 14 cech charakteryzujących wyniki ekonomiczno-produkcyjne gospodarstw rolnych. Dendrogram przecięto na poziomie określającym 5 skupień, które autorzy nazwali jednorodnymi. Nie podano jednak przesłanek, jakimi kierowano się dokonując takiego podziału, nie zbadano jednorodności klas. Skupienia charakteryzowano medianą wskaźników finansowych.

Podsumowując powyższy przegląd stwierdzić należy, że w żadnej publikacji nie dokonano oceny klasyfikacji przy pomocy obiektywnych miar. W większości przypadków stosowano jedynie jedną metodę. Rzadko starano się uzasadnić przyjęcie określonej liczby klas. Również analiza wartości cech diagnostycznych w obrębie klas była niewystarczająca. Często stosowane wartości średnie bez miar zmienności (np. rozstępu) nie mają dużych wartości informacyjnych co do własności klas, czy ich jednorodności.

Podsumowanie

Klasyfikacja wykonana trzema metodami (metodą Warda, k-średnich i siecią Kohonena) dwóch przykładowych zestawów danych, z których jeden charakteryzuje się istnieniem wyraźnych dwóch skupisk obiektów, drugi zaś jest jednorodny pozwala wyciągnąć następujące wnioski.

- Uzyskana klasyfikacji musi być oceniona przy użyciu obiektywnych miar. Jedynie w przypadkach zbiorów dwuwymiarowych możliwa jest wizualizacja zadania, pozwalająca stwierdzić, czy zadanie podziału zbiorowości wykonane zostało poprawnie, czy też nie. W artykule do oceny wykorzystano wskaźnik Rousseeuwa, którego wartości mogą stanowić miary jakości klasyfikacji.
- Niskie wartości wskaźnika Rousseeuwa (z przedziału 0,25-0,5) pozwalają jedynie stwierdzić, że dana klasyfikacja jest niepoprawna, nie oznaczają jednak, że w zbiorowości skupisk wyróżnić się nie da. Należy użyć także innych metod. Jeżeli różne metody dają klasyfikacje o zbliżonych, lecz niskich wskaźnikach Rousseeuwa można przypuszczać, że struktura klasowa nie występuje.
- Do klasyfikacji warto stosować sieci Kohonena, ponieważ dobrze wykrywają strukturę klasową, jeśli taka występuje. Jeśli brak skupisk w badanej zbiorowości, klasyfikacja wykonana przy użyciu sieci wyraźnie na to wskazuje.

Dokonany w artykule przegląd przypadków stosowania klasyfikacji bezwzorcowej w zagadnieniach z zakresu ekonomiki rolnictwa wskazuje, że autorzy pomijają niezwykle ważny jej etap, jakim jest ocena uzyskanego podziału. Często stosują jedynie jedną metodę, co byłoby uzasadnione, gdyby otrzymana klasyfikacja została pozytywnie zweryfikowana przy użyciu obiektywnych miar. Jeśli jednak oceny nie dokonano, bezpieczne jest zastosowanie wielu metod, w celu porównania uzyskanych rezultatów. Postępowanie takie pozwoli ostatecznie rozstrzygnąć, czy skupienia w zbiorowości występują, czy też nie.

Literatura

Adamowicz M., Nowak A. [2006]: Charakterystyczne typy wiejskich gospodarstw domowych na przykładzie województwa lubelskiego. *Roczniki Naukowe Stowarzyszenia Ekonomistów Rolnictwa i Agrobiznesu* t. VII z. 4.

Błażejczyk-Majka L., Kala R.[2004]: Porównanie zasobów siły roboczej rolnictwa polskiego i krajów UE w latach 1990-1999. *Roczniki Naukowe Stowarzyszenia Ekonomistów Rolnictwa i Agrobiznesu* t. VI, z. 5.

Błażejczyk-Majka L., Kala R. [2005]: Metody analizy skupień do charakterystyki użytków rolnych wybranych państw Unii Europejskiej. *Roczniki Naukowe Stowarzyszenia Ekonomistów Rolnictwa i Agrobiznesu* t. VII z. 5.

Kaufman L., Rousseeuw P.J. [1990]: Finding groups in data: an introduction to cluster analysis. Wiley, Nowy Jork.

Kisielińska J. [2003]: Klasyfikacja gospodarstw rolniczych siecią neuronową Kohonena w oparciu o wybrane wskaźniki finansowe. *Zagadnienia Ekonomiki Rolnej* 2.

Majewski J. [2005]: Regionalne zróżnicowanie skupu mleka w Polsce oraz czynniki je determinujące. *Roczniki Naukowe Stowarzyszenia Ekonomistów Rolnictwa i Agrobiznesu* t. VII, z. 5.

Metody statystycznej analizy wielowymiarowej w badaniach marketingowych. [2004]. Gatnar E. Walesiak M. (red.). Wydawnictwo Akademii Ekonomicznej im. Oskara Langego we Wrocławiu, Wrocław.

Osowska L. [2006]: Typologia funkcjonalna obszarów wiejskich Pomorza Środkowego. *Roczniki Naukowe Stowarzyszenia Ekonomistów Rolnictwa i Agrobiznesu* t. VII, z. 4.

Pietrzykowski R., Kobus P. [2008]: Wielowymiarowe metody statystyczne w analizie wyników ekonomiczno-produkcyjnych gospodarstw rolnych wybranych państw Unii Europejskiej. *Problemy Rolnictwa Światowego* t. 4 (XIX).

Pociecha J., Podolec B., Sokołowski A., Zając K. [1988]: Metody taksonomiczne w badaniach społeczno-ekonomicznych. Państwowe Wydawnictwo Naukowe, Warszawa.

Poczta A. [2005]: Poziom i struktura wsparcia finansowego rolnictwa w krajach OECD po powstaniu WTO. *Roczniki Naukowe Stowarzyszenia Ekonomistów Rolnictwa i Agrobiznesu* t. VII, z. 7A.

Poczta W., Średzińska J., Pawlak K. [2008]: Sytuacja finansowa gospodarstw rolnych krajów UE sklasyfikowanych według ich wyników produkcyjno-ekonimicznych. *Problemy Rolnictwa Światowego* t. 4 (XIX), ss. 379-387.

Rousseeuw P.J. [1987]: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* 20.

Sokołowski A. [1992]: Empiryczne testy jednorodności w taksonomii. *Zeszyty Naukowe Akademii Ekonomicznej w Krakowie* 108.

Szczepaniak I., Wigier M. [2003]: Identyfikacja czynników wpływających na innowacyjność małych i bardzo małych firm przemysłu spożywczego. *Zagadnienia Ekonomiki Rolnej* 4.

Wysocki F. [1999]: Metody statystyczne w badaniu strategii marketingowych w polskim przemyśle mleczarskim. *Roczniki Nauk Rolniczych Seria G* t. 88, z. 1.