Wstęp

Klasyfikacja i procedury taksonomiczne znajdują zastosowanie w wielu dziedzinach współczesnych badań. Wszędzie tam, gdzie pojawia się systematyczny podział przedmiotów lub zjawisk na klasy, podklasy, działy i poddziały dokonywany według określonej zasady, wykorzystywane są metody klasyfikacji danych. Klasyfikacja w sensie teoriomnogościowym to podział zupełny (suma zbiorów daje całą przestrzeń) danego zbioru na pewną liczbę rozłącznych podzbiorów. Przedmiotem klasyfikacji są zbiory obserwacji dowolnej natury. Każdy obiekt takiego zbioru jest opisany zwykle wieloma cechami zarówno ilościowymi, jak i jakościowymi. Zbiór cech (atrybutów lub własności) nazywa się przestrzenią klasyfikacji. Często we współczesnych badaniach dane zbierane do analizy zależą od jednostek czasu. Stosowana wtedy tzw. kostka danych tworzona jest przez zbiór obiektów, zbiór cech i zbiór jednostek czasu. Obiekty poddawane analizie nazywane są operacyjnymi jednostkami taksonomicznymi (OTU – *operational taxonomic unit*).

Podział jednostek taksonomicznych jest dokonywany na podstawie relacji podobieństwa, a otrzymane podzbiory nazywa się klasami abstrakcji, klasami podobieństwa czy klasami jednorodności. Opisem zasad klasyfikacji oraz metod klasyfikacji zajmuje się taksonomia. „Taksonomia jest [...] tą dziedziną statystycznej analizy wielowymiarowej, która zajmuje się teoretycznymi zasadami i regułami klasyfikacji obiektów wielocechowych”[1].

Jedną z wielu metod klasyfikacji obiektów jest analiza skupień. Pozwala ona wyodrębnić, wewnętrznie jednorodne, względem pewnej miary podobieństwa obiektów, podgrupy danego zbioru. Stosowanie tej metody wymaga kilku podstawowych decyzji badawczych.

Podstawą grupowania obiektów jest prawidłowy wybór zmiennych diagnostycznych. Ich rodzaj i liczba zależy od celu analizy. Jeśli przewiduje się redukcję wyjściowej listy zmiennych, to wstępny ich zestaw należy dobierać z nadmiarem. Zmienne mogą być bezpośrednio związane z celem wykonywanych badań, ale mogą także stanowić tak zwane zmienne symptomatyczne, wyrażające pewne szersze, niemierzalne aspekty. Formalnym założeniem w doborze zmiennych jest postulat ich niezależności. Jest to założenie bardzo atrakcyjne przy rozważaniach teoretycznych, ale w praktyce niespotykane. W takich przypadkach stosuje się rozwiązania pośrednie. Wstępną listę zmiennych diagnostycznych ustala się na podstawie analizy merytorycznej, a następnie, jeśli tylko lista jest dostatecznie liczna, prowadzi się poszukiwania optymalnego podzbioru poprzez odpowiednie metody redukcji nadmiarowych informacji. W podejściu, które łączy wytyczne merytoryczne z przesłankami statystycznymi,

[1] Patrz E. Nowak (1990)

w przypadkach konfliktowych, należy dać pierwszeństwo celom badanego zjawiska. Przy określaniu wstępnego zbioru cech diagnostycznych należy mieć na uwadze moc zbioru wartości tych cech oraz możliwą skalę pomiaru. Jednym ze wstępnych etapów badań powinno być doprowadzenie do porównywalności zmiennych diagnostycznych, bowiem zakres wartości cech wpływa na wyniki badań taksonomicznych.

Relacje podobieństwa obiektów określa się przede wszystkim na podstawie miary odległości między obiektami, dlatego postuluje się, aby przestrzeń klasyfikacji była przestrzenią metryczną. Jeszcze lepiej, gdy miara podobieństwa ma własność metryki. Niektóre miary odległości należy stosować ostrożnie i mieć na uwadze ich ograniczenia. W pracy porównane zostaną wyniki efektywności zaproponowanego testu przy wykorzystaniu dwóch najczęściej stosowanych miar odległości – odległości euklidesowej i kwadratu odległości euklidesowej.

Metody grupowania obiektów w analizie skupień można podzielić na hierarchiczne i niehierarchiczne. Wśród tych pierwszych wyróżniamy metody aglomeracyjne i podziałowe. Metody aglomeracyjne to przede wszystkim[2]: metoda pojedynczego wiązania (najbliższego sąsiedztwa - *Single linkage (Nearest neighbour)*), metoda pełnego wiązania (najdalszego sąsiedztwa - *Complete linkage (Furthest neighbour)*), metoda średnich połączeń (*Unweighted pair-group average*), metoda średnich połączeń ważonych (*Weighted pair-group average*), metoda środków ciężkości (*Unweighted pair-group centroid*), metoda ważonych środków ciężkości (mediany - *Weighted pair-group centroid*) i metoda Warda. W niniejszej pracy rozważano aglomerację zbioru obiektów, np. zbiór województw Polski, metodą aglomeracyjną H. Warda.

Metody aglomeracyjne prowadzą do uzyskania dendrogramu, czyli drzewka połączeń. Na jego podstawie ocenia się czy i na ile podgrup należy podzielić wybrany do analizy zbiór.W ten sposób, poprzez „cięcie" dendrogramu uzyskuje się wynikowy podział zbioru na podgrupy. Istnieją różne metody wyznaczania podziału wynikowego, ale niewiele z nich ma znaczenie praktyczne. W większości przypadków badacz intuicyjnie określa ostateczną liczbę podgrup. W literaturze przedmiotu można znaleźć metody wyznaczania ostatecznej liczby podzbiorów: mało wyrafinowane, takie jak „wzrokowa ocena drzewka połączeń"[3], lub pierwszy wyraźny przyrost odległości aglomeracyjnej[4] , jak również bardziej wyrafinowane np. oparte na macierzy zmienności wewntrzgrupowej i macierzy zmienności międzygrupowej. Wiele z nich zostało opisanych w pracy A. Sokołowskiego (1992). Niestety wydaje się, że żadna metoda nie jest

[2] Por. T. Grabiński (1992), A. Sokołowski (1992), B. Everit, S. Landau, M. Leese (2001), E. Gatnar, M. Walesiak (2004)
[3] Por. A. Malina (2004)
[4] Por. A. Sokołowski (1977), A. Sokołowski (1992)

w pełni zadowalająca. Podstawowym błędem niedoświadczonego badacza jest produkowanie klasyfikacji bez względu na to, czy zbiór powinien być dzielony na podzbiory, czy też nie. Mówi się wtedy o podziale na bardziej jednorodne podgrupy.

Głównym celem pracy jest zaprezentowanie procedury pozwalającej na przerwanie procesu aglomeracji metodą Warda na drodze odrzucenia hipotezy statystycznej. Hipoteza zerowa zakłada, że badana zbiorowość jest jednorodna i nie powinna być dzielona na podzbiory. Sposób wyznaczenia wartości krytycznych testu, na podstawie których jest weryfikowana hipoteza zerowa, oparty został na analizie rozkładu odległości aglomeracyjnej. Odległość ta wskazuje na każdym etapie łączenia, które obiekty lub grupy obiektów należy połączyć. Ponieważ wyprowadzenie analityczne tego rozkładu jest niemożliwe, zaproponowano wykorzystanie empirycznych testów jednorodności, a w dalszej części pracy wyznaczono wartości krytyczne metodami Monte Carlo.

Podzbiory obiektów są wyznaczane nie tylko na podstawie danych wybranych do analizy, ale też przez cel określonego badania i osobę która analizuje określone zjawisko. Mimo coraz bardziej pomysłowych prób automatyzacji procesu podziału nadal jednym z głównych wyróżników poprawności i trafności analizy danych jest doświadczenie i wyczucie badacza.

Ponieważ niewiele kryteriów zatrzymywania procesu aglomeracji ma znaczenie praktyczne, powstaje pytanie dotyczące efektywności zaproponowanej metody wyznaczania ostatecznej liczby podzbiorów. Oczywistym też wydaje się zwrócenie uwagi na różnice, które pojawiają się gdy wyjściowa macierz odległości między obiektami w zbiorze Ω mierzona jest różnymi sposobami: odległością euklidesową lub jej kwadratem. Duże znaczenie ma także pokazanie praktycznego zastosowania proponowanej procedury.

Wymienionym celom została podporządkowana konstrukcja pracy złożonej ze wstępu, sześciu rozdziałów, podsumowania i spisu literatury.

W pierwszym rozdziale omówiono podstawowe pojęcia związane z aglomeracyjnymi metodami klasyfikacji danych takie jak: podstawowe definicje przestrzeni obiektów, założenia modelu, dobór zmiennych, skale pomiaru i sposoby doprowadzania zmiennych do porównywalności, miara odległości między obiektami, a także najważniejsze metody aglomeracji obiektów. W szczegółowy sposób przedstawiono metodę grupowania, opartą na algorytmie J. H. Warda[5]. Na zakończenie rozdziału opisano wybrane kryteria stopu stosowane w hierarchicznej analizie skupień.

[5] Patrz J.H Ward (1963)

Rozdział drugi zawiera metodę wyznaczania wartości krytycznych pozwalających stwierdzić zasadność przerwania procesu aglomeracji metodą Warda. Przerwanie procesu aglomeracji następuje na skutek odrzucenia hipotezy statystycznej głoszącej, że badana zbiorowość jest jednorodna i nie powinna być dzielona na podzbiory. Kryterium przerwania procesu aglomeracji opiera się na analizie rozkładu prawdopodobieństwa odległości aglomeracyjnej. Ponieważ wyprowadzenie analityczne tego rozkładu jest bardzo trudne, wyznaczano go przy pomocy symulacji komputerowej. Empiryczne testy jednorodności[6] wymagają przyjęcia pewnych założeń dotyczących zmiennych diagnostycznych. Przyjęto, że liczba obiektów w zbiorze wyjściowym wynosi szesnaście[7]; ilość cech jest liczbą naturalną z przedziału [2,20]; model populacji jednorodnej to wielowymiarowy rozkład normalny o niezależnych składowych; wyjściową macierz odległości między obiektami w zbiorze Ω mierzono dwoma sposobami: odległością euklidesową lub jej kwadratem.

Rozkłady statystyk testowych wyznaczano dwustopniowo. Najpierw wygenerowano wstępne oceny wartości krytycznych, a następnie dopasowano funkcję analityczną aproksymującą surowe wartości krytyczne. Wstępne oceny wartości krytycznych policzono osobno dla każdej odległości łączenia, a także z uwzględnieniem całych wektorów odległości aglomeracyjnych. Dopasowana funkcja zależy od rozmiaru przestrzeni klasyfikacji oraz liczby klas. W dalszej części rozdziału zastosowano różne metody optymalizacji parametrów szukanej funkcji.

Przedstawione w rozdziale drugim empiryczne testy jednorodności wymagają niezależności zmiennych diagnostycznych. W rozdziale trzecim do wyznaczenia wartości krytycznych zastosowano metody Monte Carlo. W tych metodach dodatkowe warunki, jak np. macierz kowariancji, obliczane są na podstawie danych empirycznych, a następnie przeprowadzane są symulacje z uwzględnieniem tych warunków. W trzecim rozdziale pracy przedstawiono wartości krytyczne odległości aglomeracyjnych dwóch cech skorelowanych zadanym współczynnikiem korelacji liniowej. Obliczenia przeprowadzono kolejno dla $r = 0{,}76$, $r = 0{,}86$ i $r = 0{,}96$.

Rozdział czwarty zawiera konstrukcję liczenia wartości krytycznych[8] przy dodatkowym założeniu, że cechy są skorelowane zadaną macierzą korelacji liniowej. Metoda generowania zmiennych o zadanej macierzy korelacji zaczerpnięta została z książki *Generatory liczb losowych* [9].

[6] Por. A. Sokołowski (1992); S. Denkowska, A. Sokołowski (2001)
[7] Liczba szesnaście została przyjęta dla potrzeb grupowania zbioru województw Polski
[8] również metodą Monte Carlo
[9] Por. R. Zieliński (1979), R. Zieliński, R. Wieczorkowski (1997),

Tematem rozdziału piątego jest próba oceny efektywności proponowanych testów. Efektywność metod taksonomicznych jest najczęściej rozumiana jako zdolność prawidłowego rozpoznania rzeczywistej struktury obiektów w wielowymiarowej przestrzeni klasyfikacji. Badania efektywności algorytmów są przeprowadzane na przykładach empirycznych lub sztucznie wygenerowanych. Wykorzystując podejście symulacyjne przeprowadzono kilka eksperymentów komputerowych, w celu sprawdzenia proponowanych testów w warunkach nieprawdziwości hipotezy o jednorodności badanej struktury danych. Badania te dotyczyły jedynie prawdopodobieństwa rozpoznania prawidłowej liczby grup, a nie przyporządkowania pojedynczych obserwacji do wyznaczonych grup.

W rozdziale szóstym podano empiryczne przykłady wnioskowania opartego na zaproponowanych procedurach kryterium stopu w procesie aglomeracji metodą Warda. W analizie wykorzystano dane obejmujące różne sfery zjawisk społecznych i ekonomicznych. Badany był zbiór województw pod kątem warunków mieszkaniowych ludności, wskaźników bezrobocia, aktywności kulturalnej ludności, produkcji żywca rzeźnego oraz ruchu naturalnego ludności. Ta różnorodność danych pozwala zweryfikować stosowanie zaproponowanych wartości krytycznych.