Grupowanie zmiennych w procesach eksploracji danych (*Data Mining*)
Variable clustering in exploration data processes

Mirosława Lasek, Marek Pęczkowski
Katedra Informatyki Gospodarczej i Analiz Ekonomicznych, Wydział Nauk Ekonomicznych, Uniwersytet Warszawski, ul. Długa 44/50, 00-241 Warszawa, e-mail: mlasek@wne.uw.edu.pl, mpeczkowski@wne.uw.edu.pl

Abstract
Variable clustering is a useful tool for data reduction. It removes collinearity, decreases variable redundancy and helps to interpret results of an analysis. In the paper, Variable Clustering Node of SAS Enterprise Miner is described. An example of clustering of households expenditures on food, alcohol and tobacco is presented.
Keywords*: variable clustering, exploration data*

Wstęp

Grupowanie obiektów, znane pod nazwą analizy skupień, jest jedną z najczęściej stosowanych metod eksploracyjnych *Data Mining*. Używane są metody hierarchicznego, jak i niehierarchicznego skupiania, wykorzystujące różne algorytmy i wersje metody, dające w wyniku podział obiektów rozłączny i zupełny, jak również wersja rozmyta, oparta na zastosowaniu teorii zbiorów rozmytych. Znacznie rzadziej, zarówno w opisach literaturowych, jak i w praktyce, spotyka się zastosowania metod grupowania w odniesieniu do zmiennych, ze względu na które są charakteryzowane obiekty (Anderberg, 1973). Nie odzwierciedla to faktu, że grupowanie zmiennych jest bardzo przydatne w analizach danych zawierających dużą liczbę zmiennych. W przedstawianym artykule chcielibyśmy zaprezentować niektóre możliwości, jakie może dać przeprowadzenie skupiania zmiennych.

1. Podstawowe cele grupowania zmiennych

W eksploracji danych zajmujemy się zazwyczaj obiektami, charakteryzowanymi za pomocą bardzo dużej liczby zmiennych (cech). Liczba ta dochodzi nierzadko do kilkuset. Część z nich wprowadza redundancję informacji, opisując te same lub zbliżone właściwości obiektów i utrudnia prowadzenie analizy danych, np. wykrycie współzależności, która może zachodzić między zmiennymi objaśniającymi a zmienną objaśnianą w budowanym modelu. Pogrupowanie zmiennych w skupienia może ułatwić analizę dzięki zastąpieniu grupy zmiennych jednym komponentem (*cluster component*), będącym kombinacją liniową tych zmiennych albo przez wybór jednej zmiennej jako reprezentanta grupy zmiennych (ta druga możliwość jest szczególnym przypadkiem pierwszej).

Skupianie zmiennych pozwala usunąć współliniowość zmiennych i wprowadzić większą przejrzystość w wykorzystywanym zbiorze obiektów. Zmniejszenie liczby zmiennych umożliwia budowę modelu o mniejszej złożoności niż w przypadku uwzględniania wszystkich zmiennych, ukazującego w sposób bardziej czytelny związek między zmiennymi objaśniającymi a zmienną objaśnianą. Skraca też czas potrzebny na zbudowanie modelu, a także ułatwia interpretację uzyskiwanych wyników, przy zaledwie niewielkiej utracie informacji.

Dodatkową zaletą grupowania zmiennych jest możliwość budowania oddzielnych modeli, z których każdy uwzględnia inne charakterystyki obiektów, reprezentowane przez zmienne pochodzące z różnych skupień.

2. Założenia grupowania zmiennych

W prowadzonych przez nas zastosowaniach i przykładzie opisanym w niniejszym artykule wykorzystywaliśmy algorytm grupowania zmiennych, opracowany przez *SAS Institute Inc.* i realizowany przez program *SAS Enterprise Miner* (Reference Help …, 2007).

Algorytm umożliwia uzyskiwanie skupień, zarówno rozłącznych, jak i hierarchicznych. Jest przeznaczony do grupowania zmiennych numerycznych, choć możliwe jest także specjalne postępowanie dla uwzględnienia zmiennych nienumerycznych.

Skupienia otrzymane w wyniku zastosowania algorytmu mogą być traktowane jako kombinacje liniowe zmiennych występujących w skupieniu. Każda taka liniowa kombinacja zmiennych jest pierwszą główną składową skupienia. Podobnie jak w analizie głównych składowych (*PCA*), pierwsza główna składowa jest śred-

nią ważoną zmiennych z tak dobranymi wagami, aby wyjaśnić możliwie najwięcej wariancji. Jednak w odróżnieniu od metody *PCA* rozważane składowe mogą być ze sobą skorelowane. W zwykłej metodzie głównych składowych kolejne komponenty (pierwsza, druga itd. składowa) są budowane na podstawie tego samego zbioru zmiennych. Tutaj bierzemy pod uwagę tylko pierwsze składowe główne, ale każda z nich jest budowana na podstawie innych zmiennych.

Dla zbudowania skupień, podobnie jak w analizie głównych składowych, wykorzystywana jest macierz korelacji lub kowariancji. Jeżeli jest wykorzystywana macierz korelacji, wszystkie zmienne są traktowane jako jednakowo ważne. Jeżeli jest wykorzystywana macierz kowariancji, zmienne o większej wariancji są traktowane jako istotniejsze w przeprowadzanej analizie.

3. Algorytm grupowania zmiennych

Algorytm skupiania zmiennych szuka takiego podziału zmiennych, aby maksymalizować wariancję, która jest wyjaśniona przez komponenty skupień, zsumowaną po wszystkich skupieniach.

Na ogół wszystkie komponenty skupień wyjaśniają mniej wariancji wszystkich rozważanych zmiennych niż taka sama liczba głównych składowych wyodrębnionych przez *PCA* na podstawie wszystkich zmiennych. Jednak komponenty skupień mają łatwiejszą interpretację. Główne składowe w *PCA* są na ogół trudne do interpretacji, nawet po zastosowaniu rotacji zmiennych.

Algorytm skupiania zmiennych jest podziałowy, tzn. punktem wyjścia jest zbiór wszystkich zmiennych traktowany jako jedno skupienie, a w kolejnych krokach następuje podział danego skupienia na podzbiory. Podział może być hierarchiczny lub niehierarchiczny, w zależności od wybranej opcji programu.

W algorytmie podziału powtarzane są następujące kroki:

1) wybierane jest skupienie, które będzie dzielone na dwa podzbiory. Kryterium wyboru jest albo najmniejszy udział wyjaśnionej zmienności przez komponent skupienia (gdy użytkownik wybierze opcję *Variation Proportion*) albo największa wartość własna odpowiadająca drugiej składowej głównej skupienia (gdy użytkownik wybierze opcję *Maximum Eigenvalue*);

2) po wybraniu skupienia w kroku 1. są znajdywane dwie pierwsze składowe główne stosując rotację *orthoblique*. Przypisuje się zmienne do tej z dwóch składowych, z którą ma większą wartość kwadratu współczynnika

korelacji (R^2). Składowe wyznaczają podział skupienia zmiennych na dwie części;

3) zmienne są na nowo przyporządkowywane do skupień w ten sposób, żeby maksymalizować wariancję określoną przez składowe skupień. Użytkownik programu może wybrać opcję *Keep Hierarchies* zapewniającą zachowanie struktury hierarchicznej skupień.

Krok 3. zawiera dwa etapy:

- najpierw są obliczane składowe skupień i każda zmienna zostaje przypisana do składowej, z którą ma największą wartość kwadratu współczynnika korelacji (R^2);
- następnie dla każdej zmiennej sprawdza się, czy przypisanie jej do innego skupienia zwiększy wartość wyjaśnionej wariancji. Jeżeli przesunięcie zmiennej do innego skupienia zwiększy wartość wyjaśnionej wariancji, to na nowo obliczane są składowe obu tych skupień, zanim następna zmienna będzie sprawdzana.

Wybór opcji *Keep Hierarchies* ogranicza zmianę przyporządkowania zmiennych do skupień w ten sposób, że podział zbioru może być tylko hierarchiczny. Oznacza to, że jeżeli w danym kroku podzielimy skupienie A na skupienia A_1 i A_2, to zmienna może przejść tylko z A_1 do A_2 albo z A_2 do A_1, ale nie do innych skupień. Użycie podziału hierarchicznego redukuje czas obliczeń i ułatwia interpretację skupień.

Algorytm kończy się, gdy spełnione zostaną kryteria stopu podane w polu *Stopping Criteria*. Są to:

a) osiągnięto maksymalną liczbę skupień podaną w opcji *Maximum Clusters* (domyślnie: liczba zmiennych w analizie);

b) wartość własna odpowiadająca drugiej składowej głównej przekracza wartość podaną w opcji *Maximum Eigenvalue* (domyślnie: 1);

c) osiągnięto zadany udział wariancji wyjaśnionej wybrany w opcji *Variation Proportion* (domyślnie: 0).

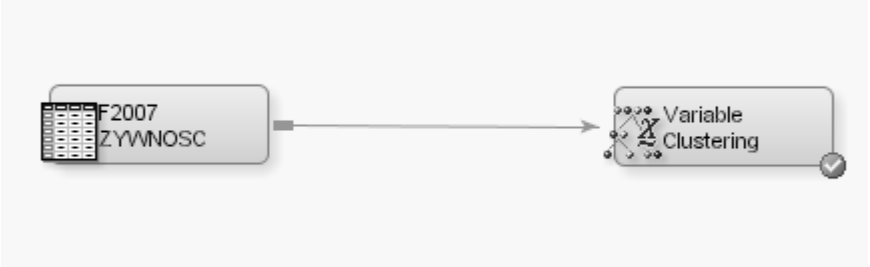
4. Przykład grupowania zmiennych

Przykład dotyczy grupowania zmiennych, opisujących wydatki na żywność, alkohol i tytoń w gospodarstwach domowych. Dane pochodzą z badania Budżetów Gospodarstw Domowych prowadzonego przez GUS w 2007 roku (Budżety gospodarstw …, 2008). Pozycje wydatków występujące w źródłowym zbiorze danych

zostały zagregowane, aby uniknąć wydatków mających znikomy udział w sumie wydatków na żywność, alkohol i tytoń.

Uwzględniono 31 pozycji wydatków (zmiennych) z 37121 gospodarstw domowych. Są to (w kolejności alfabetycznej): ciastka, cukier, drób, dżem-miód, herbata, jaja, kasza, kawa, makaron, masło, mąka, mięso, mleko, napoje, owoce, pieczywo, piwo, płatki, przyprawy, ryby, ryż, sery, słodycze, śmietana, tłuszcze, tytoń, warzywa, wędliny, wino, wódki, ziemniaki.

Zgodnie z wymaganiami programu *Enterprise Miner* zbudowano diagram przetwarzania danych. Składa się on z dwóch węzłów: wprowadzania danych i tworzenia skupień zmiennych. Pierwszy węzeł służy do określenia wykorzystywanego zbioru danych (*F2007ZYWNOSC*) i roli zmiennych występujących w analizie, drugi węzeł realizuje algorytm skupiania.



Źródło: opracowanie własne przy wykorzystaniu programu *Enterprise Miner.*

Rys. 1. Diagram na potrzeby skupiania zmiennych

W przykładzie wykorzystaliśmy macierz korelacji standaryzowanych zmiennych (opcja *Correlation* w polu *Clustering Source*). Wszystkie zmienne są zmiennymi numerycznymi, przedstawiającymi wielkości wydatków na poszczególne pozycje zakupów żywnościowych gospodarstw domowych.

Przyjęliśmy też domyślną opcję zachowania hierarchicznej struktury skupień.

Maksymalną liczbę skupień pozostawiono jako wielkość domyślną proponowaną przez program, co oznacza przyjęcie wielkości równej liczbie zmiennych wejściowego zbioru danych.

Węzeł *Variable Clustering* tworzy i eksportuje do dalszych węzłów wprowadzanych do diagramu (mogą nimi być np. węzły tworzenia modeli regresji, drzew decyzyjnych lub sieci neuronowych), liniową kombinację zmiennych każdego skupienia. Jest to ustalenie domyślne algorytmu (programu). Zamiast liniowej kombinacji można eksportować do dalszych węzłów „najlepszą zmienną" z każdego skupienia.

Jako „najlepsze zmienne" przyjmuje się takie zmienne skupień, które mają najmniejszą wartość parametru: *1- R^2 Ratio*, w skupieniach. *1- R^2 Ratio* jest to iloraz:

$$\frac{1-R_G^2}{1-R_I^2}$$

gdzie:

R_G^2 - współczynnik R^2 zmiennej ze składową główną jej skupienia,

R_I^2 - współczynnik R^2 zmiennej ze składową główną najbliższego skupienia.

W przypadku „dobrego" skupiania kwadrat współczynnika korelacji zmiennej ze swoją główną składową (R_G^2) powinien być duży. O dobrym wyodrębnieniu grup zmiennych świadczy też małe skorelowanie zmiennych z komponentami innych grup, zatem R_I^2 powinno być małe. Z tego wynika, że małe wartości ilorazu $\frac{1-R_G^2}{1-R_I^2}$ świadczą o dobrym grupowaniu.

Zmienna o najmniejszej wartości ilorazu jest wysoko skorelowana z komponentem swojej grupy i mało skorelowana z komponentami innych grup. Stąd wybrana zostaje jako najlepszy reprezentant swojej grupy.

Użytkownik programu może więc wybrać, która zmienna będzie eksportowana z węzła *Variable Clustering* do następnych węzłów diagramu. Powtórzmy i podsumujmy. Może to być:

- komponent skupienia utworzony jako pierwsza główna składowa zmiennych danego skupienia (*Cluster Component*);
- najlepsza zmienna traktowana jako reprezentant skupienia. Jest to zmienna, która ma najmniejszą wartość *1- R^2 Ratio* spośród wszystkich zmiennych w skupieniu (*Best Variable*).

5. Interpretacja wyników przykładu grupowania zmiennych

Wyniki są przedstawiane w postaci mapy skupień (rys. 3) albo w postaci dendrogramu (rys. 4).

W naszym przypadku zostało wyodrębnionych 6 skupień oznaczonych symbolami CLUS1, CLUS2,… , CLUS6.

Liczbę zmiennych i ich częstość (liczebność względną) w poszczególnych skupieniach przedstawiono na rysunku 2. (*Variable Frequency Table*).

Variable Frequency Table

Cluster	Frequency Count	Percent of Total Frequency
CLUS1	9	29.03226
CLUS2	10	32.25806
CLUS3	3	9.677419
CLUS4	3	9.677419
CLUS5	5	16.12903
CLUS6	1	3.225806

Źródło: opracowanie własne przy wykorzystaniu programu *Enterprise Miner* .

Rys. 2. Widok tablicy częstości skupień

Skupienie 1 (CLUS1) zawiera zmienne: pieczywo, wędliny, tłuszcze, jaja, cukier, mięso, ziemniaki, drób, mąkę.

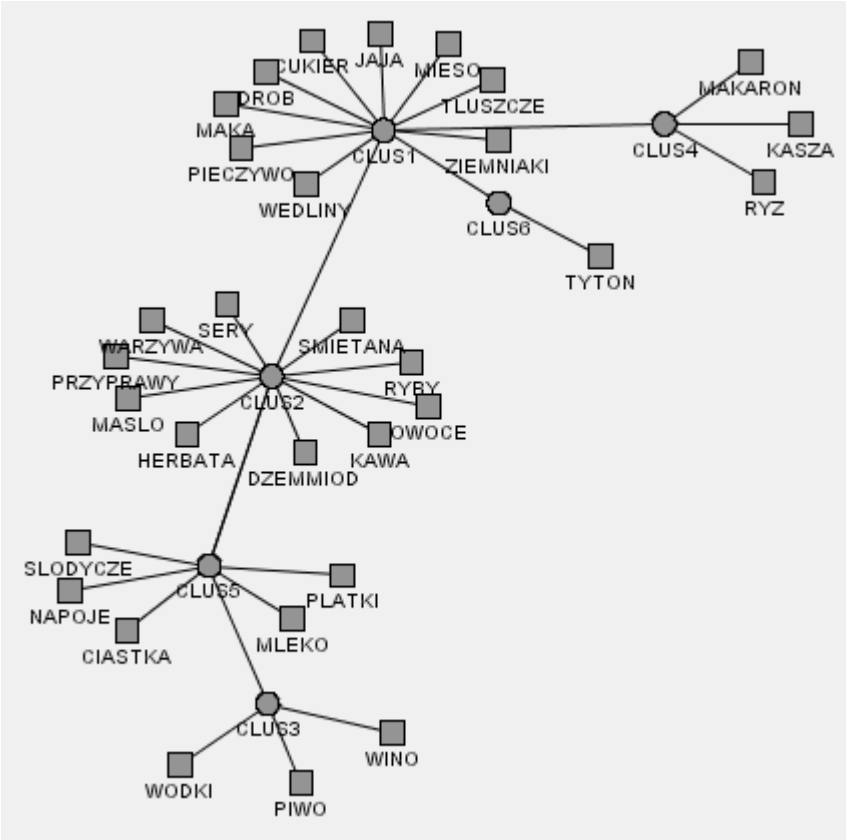
Skupienie 2 (CLUS2) zawiera zmienne: sery, owoce, warzywa, przyprawy, śmietanę, ryby, masło, herbatę, kawę, dżem-miód.

Skupienie 3 (CLUS3) zawiera zmienne: wódki, wino, piwo.

Skupienie 4 (CLUS4) zawiera zmienne: ryż, makaron, kaszę.

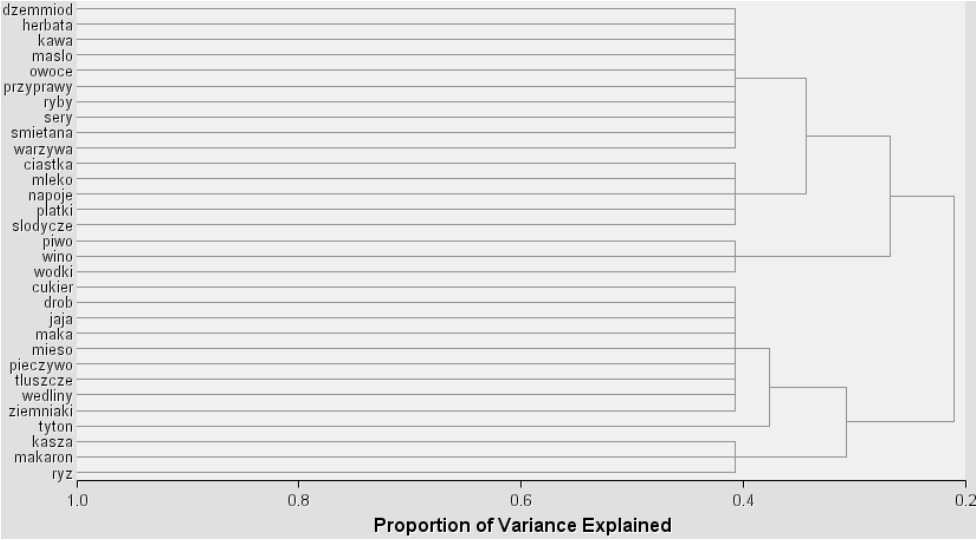
Skupienie 5 (CLUS5) zawiera zmienne: napoje, słodycze, ciastka, płatki, mleko.

Skupienie 6 (CLUS6) zawiera zmienną: tytoń.



Źródło: opracowanie własne przy wykorzystaniu programu *Enterprise Miner.*

Rys. 3. Wykres skupień zmiennych w postaci mapy skupień



Źródło: opracowanie własne przy wykorzystaniu programu *Enterprise Miner*.

Rys. 4. Wykres skupień zmiennych w postaci dendrogramu

Skupienie 1 (CLUS1) grupuje pozycje wydatków na żywność, charakterystyczne dla gospodarstw domowych o najbardziej tradycyjnej strukturze ponoszenia wydatków żywnościowych. Wydatki na pieczywo, wędliny, tłuszcze są powiązane z wydatkami na jaja, mięso, drób, a także takie pozycje jak ziemniaki, mąka i cukier.

W skupieniu 2 (CLUS2) znalazły się pozycje, które łączy się z przestrzeganiem lekkostrawnej, wegetariańskiej diety. Powiązane są tu wydatki na sery, owoce, warzywa, ryby, masło z wydatkami na przyprawy, śmietanę, herbatę, kawę, dżem - miód (te ostatnie dżem i miód traktowane przez nas łącznie).

Pozycje skupienia 3 (CLUS3) wskazują na łączenie przez pewne gospodarstwa wydatków na różne, rozmaite napoje alkoholowe. Jako pozycje wydatków zgrupowane zostały wydatki na wódki, wina i piwa.

Skupienie 4 (CLUS4) grupuje wydatki na ryż, makarony, kasze, wskazując na gospodarstwa opierające dietę na produktach zbożowych.

W skupieniu 5 (CLUS5) znalazły się napoje bezalkoholowe, słodycze, ciastka, płatki, mleko. Jest to połączenie pozycji w kierunku diety „lekkiej", ale „słodkiej".

Skupienie 6 (CLUS6) zawiera tylko jedną pozycję: tytoń, która nie jest łączona z wydatkami na inne rozpatrywane tu pozycje.

Widoczne są dwa ugrupowania skupień. Pierwsze ugrupowanie, to blisko położone względem siebie skupienia 1, 4 oraz 6, a więc charakteryzowane jako skupienie „tradycyjnych wydatków", „wydatków na produkty zbożowe" oraz tytoń. Drugie ugrupowanie tworzą położone blisko siebie skupienia 2, 5 oraz 3. Są to skupienia „diety lekkostrawnej", diety określonej jako „lekka" i „słodka" oraz „napojów alkoholowych".

Variable Selection Table

Cluster	Variable	Label	R-Square With Own Cluster Component	Next Closest Cluster	R-Sqaure with Next Cluster Component	Type	1-R2 Ratio	Variable Selected
CLUS1	CLUS1	Cluster 1	1	CLUS2	0.339505	ClusterComp	0	YES
CLUS1	PIECZYWO	Pieczywo	0.546923	CLUS2	0.169446	Variable	0.545512	NO
CLUS1	WEDLINY	Wêdliny	0.503318	CLUS2	0.223538	Variable	0.639673	NO
CLUS1	TLUSZCZE	Margaryna, ...	0.443342	CLUS2	0.136769	Variable	0.644854	NO
CLUS1	JAJA	Jaja	0.432103	CLUS2	0.136994	Variable	0.658045	NO
CLUS1	CUKIER	Cukier	0.36192	CLUS2	0.083688	Variable	0.696357	NO
CLUS1	MIESO	Miêso	0.371393	CLUS2	0.151739	Variable	0.741054	NO
CLUS1	ZIEMNIAKI	Ziemniaki	0.242366	CLUS2	0.047621	Variable	0.795518	NO
CLUS1	DROB	Drób	0.289529	CLUS2	0.124662	Variable	0.811653	NO
CLUS1	MAKA	M'ka	0.212697	CLUS2	0.098761	Variable	0.873578	NO
CLUS2	CLUS2	Cluster 2	1	CLUS5	0.353894	ClusterComp	0	YES
CLUS2	SERY	Sery	0.481774	CLUS5	0.230852	Variable	0.673766	NO
CLUS2	OWOCE	Owoce	0.426394	CLUS5	0.173207	Variable	0.693772	NO
CLUS2	WARZYWA	Warzywa	0.453489	CLUS1	0.220203	Variable	0.700838	NO
CLUS2	PRZYPRAWY	Sól, sosy, z...	0.394438	CLUS1	0.196248	Variable	0.753419	NO
CLUS2	SMIETANA	Œmietana, ...	0.301601	CLUS1	0.113029	Variable	0.787397	NO
CLUS2	RYBY	Ryby, zwierz...	0.273127	CLUS1	0.091367	Variable	0.799963	NO
CLUS2	MASLO	Mas³o	0.232867	CLUS1	0.067192	Variable	0.822391	NO
CLUS2	HERBATA	Herbata	0.230925	CLUS5	0.080615	Variable	0.836511	NO
CLUS2	KAWA	Kawa	0.245229	CLUS5	0.102201	Variable	0.840691	NO
CLUS2	DZEMMIOD	D¿em, mar...	0.086589	CLUS4	0.017084	Variable	0.929288	NO
CLUS3	CLUS3	Cluster 3	1	CLUS5	0.098846	ClusterComp	0	YES
CLUS3	WODKI	Wódki, likiery	0.589788	CLUS5	0.043494	Variable	0.428865	NO
CLUS3	WINO	Wina, inne ...	0.456724	CLUS2	0.038527	Variable	0.565046	NO
CLUS3	PIWO	Piwo	0.403848	CLUS5	0.071725	Variable	0.642215	NO
CLUS4	CLUS4	Cluster 4	1	CLUS1	0.143433	ClusterComp	0	YES
CLUS4	RYZ	Ry¿	0.525784	CLUS2	0.058852	Variable	0.50387	NO
CLUS4	MAKARON	Makaron	0.478846	CLUS1	0.127525	Variable	0.597328	NO
CLUS4	KASZA	Kasza	0.418657	CLUS1	0.040231	Variable	0.605711	NO
CLUS5	CLUS5	Cluster 5	1	CLUS2	0.353894	ClusterComp	0	YES
CLUS5	NAPOJE	Napoje bez...	0.575829	CLUS2	0.228513	Variable	0.54981	NO
CLUS5	SLODYCZE	Czekolada, ...	0.544489	CLUS2	0.212149	Variable	0.578169	NO
CLUS5	CIASTKA	Ciastka, piz...	0.429171	CLUS2	0.136816	Variable	0.661307	NO
CLUS5	PLATKI	P³atki	0.345952	CLUS2	0.069144	Variable	0.702631	NO
CLUS5	MLEKO	Mleko	0.33472	CLUS2	0.154085	Variable	0.786462	NO
CLUS6	CLUS6	Cluster 6	1	CLUS1	0.021083	ClusterComp	0	YES
CLUS6	TYTON	Papierosy, t...	1	CLUS1	0.021083	Variable	0	NO

Źródło: opracowanie własne przy wykorzystaniu programu *Enterprise Miner*.

Rys. 5. Tablica zawierająca statystyki dotyczące skupień

Enterprise Miner wyświetla tablicę zawierającą statystyki dotyczące otrzymanych skupień (rys. 5). W kolejnych kolumnach tablicy są podane: nazwa skupienia (*Cluster*), nazwa zmiennej (*Variable*) i etykieta zmiennej (*Label*), wartość R_G^2 zmiennej ze składową główną jej skupienia (*R-Square With Own Cluster Component*), nazwa najbliższego skupienia do podanego w tym samym wierszu pierwszej kolumny tablicy (*Next Closest Cluster*), wartość R_I^2 zmiennej ze składową główną tego (najbliższego) skupienia (*R-Square With Next Cluster Component*), typ zmiennej – komponent skupienia lub pojedyncza zmienna (*Type*), wartość $1-R^2$ *Ratio* (*1-R2 Ratio*), zaznaczenie wybranego (najlepszego) reprezentanta skupienia – *YES* lub *NO* (*Variable Selected*).

Skupienia są dobrze wyodrębnione, o czym świadczą małe wartości R_I^2. O poprawnym skupianiu świadczą też małe wartości *1-R^2 Ratio*. W tablicy zmienne w skupieniach są uporządkowane rosnąco według wartości *1-R^2 Ratio*. Z tablicy możemy odczytać, że „najlepsze zmienne" (*best variables*), to: pieczywo, sery, wódki, ryż, napoje, tytoń, które mogą być wybrane jako reprezentanci grup.

Podsumowanie

Grupowanie zmiennych pozwala w znacznym stopniu ułatwić eksplorację danych dzięki możliwości ograniczenia liczby zmiennych. Analiza zyskuje na przejrzystości i czytelności.

Grupowanie zmiennych jest często pierwszym krokiem do dalszych analiz, w których stosujemy metody predykcyjne z mniejszą liczbą zmiennych objaśniających. Dzięki ograniczeniu liczby zmiennych możemy budować mniej złożone modele, w krótszym czasie, o przydatności podobnej do modeli z bardzo dużą liczbą zmiennych. Przydatne dla analiz może być także budowanie wielu modeli uwzględniających zmienne z różnych skupień.

W przedstawionym przez nas przykładzie rozważaliśmy 31 zmiennych – pozycji wydatków na żywność gospodarstw domowych w Polsce. Dzięki wykorzystaniu grupowania zmiennych otrzymaliśmy 6 skupień zmiennych, będących kombinacjami liniowymi zmiennych ze skupień, z których każda może być traktowana dalej w analizach eksploracyjnych danych jako jedna zmienna – komponent. Postępując nieco inaczej możemy wybrać z każdego skupienia jedną zmienną jako jego reprezentanta i tylko te wybrane zmienne przyjmować do dalszych analiz. W każdym przypadku zyskujemy dzięki ograniczeniu złożoności i uproszczeniu analizy, nie tracąc możliwości wystarczająco dogłębnego zbadania problemu. Po-

nadto skupianie zmiennych może pozwolić na odkrycie pewnej dodatkowej wiedzy, tak jak np. w naszym przypadku stało się z wiedzą o łączeniu wydatków gospodarstw domowych na różne pozycje żywnościowe.

Piśmiennictwo

1. Anderberg M. R. 1973. *Cluster Analysis for Applications,* Academic Press Inc., New York.
2. *Budżety gospodarstw domowych w 2007 r*., Informacje i opracowania statystyczne, GUS, Warszawa 2008.
3. *Reference Help – Enterprise Miner 5.3., Variable Clustering Node*, SAS Institute Inc., Cary, NC, USA 2007.