Humanities and Social Sciences **2015**
HSS, vol. XX, 22 (4/2015), pp. 159-168 **October - December**

Marek SOBOLEWSKI[1]

AUTOMATYZACJA ANALIZY SKUPIEŃ W PROGRAMIE STATISTICA

W artykule opisano aplikację, która pozwala zautomatyzować proces analizy skupień za pomocą programu STATISTICA. Głównym celem artykułu było zarysowanie pewnej uniwersalnej koncepcji systemu analizy danych, który umożliwiałby efektywne wykonywanie licznych i złożonych analiz statystycznych. W pierwszej kolejności dokonano przeglądu publikacji, w których wykorzystano taksonomiczne metody grupowania danych. Opisano zakres faktycznych i możliwych zastosowań metod taksonomicznych przez badaczy z różnych dziedzin nauki (w szczególności w obszarze nauk społecznych i humanistycznych). Następnie wskazano na pewne niedoskonałości w implementacji metod grupowania danych w pakietach statystycznych. Wychodząc naprzeciw oczekiwaniom praktyków, opracowano program Automatyzacja grupowania, będący rozszerzeniem pakietu STATISTICA, umożliwiający efektywne zastosowanie metod taksonomicznych w praktyce. Spośród wielu korzyści, jakie przynosi zastosowanie programu Automatyzacja grupowania, warto wymienić: możliwość równoległego wykonywania grupowania według kilku alternatywnych algorytmów i porównywania uzyskanych wyników, ustalanie wielu sposobów podziału dla każdej zdefiniowanej metody, automatyczne generowanie tabel zawierających charakterystykę utworzonych skupień w dokumencie programu Word. Schemat analizy może zostać zapisany i wykorzystany do kolejnych obliczeń na zbiorze danych o podobnej strukturze (na przykład po uzyskaniu nowych wyników badań eksperymentalnych i zwiększeniu liczby przypadków w bazie danych). Wyniki działania programu zaprezentowano na przykładzie z dziedziny nauk społecznych, który dotyczył analizy poziomu życia w miastach na prawach powiatu w latach 2003–2012. Ponieważ nadrzędnym celem powstałej pracy jest popularyzacja stosowania metod taksonomicznych w praktyce, przedstawiona w niej aplikacja będzie udostępniana drogą mailową wszystkim zainteresowanym osobom.
Słowa kluczowe: analiza skupień, program STATISTICA, Visual Basic

1.WSTĘP

Podstawowym celem niniejszej publikacji jest przedstawienie naukowcom i praktykom, wykorzystującym w swojej pracy narzędzia grupowania danych, rozszerzenia programu STATISTICA, za którego pomocą można znacząco usprawnić i wzbogacić proces analizy danych. Sama praca powinna być jednak postrzegana nieco szerzej. Autor dążył także do ukazania możliwości i korzyści płynących z budowy automatycznych systemów analizy danych oraz podkreślenia faktu, że jest to możliwe bez dodatkowych nakładów finansowych, w ramach już istniejącego i powszechnie wykorzystywanego oprogramowania.

Współczesne programy statystyczne umożliwiają wykonanie obliczeń, ale naukowiec czy praktyk, wykonujący regularnie pewne powtarzalne czynności, do sprawnej analizy

[1] Dr Marek Sobolewski, Katedra Metod Ilościowych, Wydział Zarządzania, Politechnika Rzeszowska, msobolew@prz.edu.pl

statystycznej potrzebuje narzędzia zautomatyzowanego, pozwalającego prowadzić analizy wielowariantowe wraz automatycznym raportowaniem wyników, najlepiej bezpośrednio w edytorze tekstów.

Automatyzacja zapewnia powtarzalność całego procesu obliczeń, raportowania, tworzenia wykresów, dzięki czemu można powtórzyć analizy dla innego zbioru obiektów, innego zestawu zmiennych czy z powodu wykrycia błędów w danych.

Częstość stosowania metod taksonomicznych (w szczególności procedur grupowania) jest niewielka w relacji do użyteczności tych metod – upowszechnienie narzędzi taksonometrii wpłynie stymulująco na ich dalszy rozwój.

W pierwszej części niniejszej pracy opisano wybrane przykłady wykorzystania metod grupowania w badaniach naukowych z zakresu nauk społecznych i humanistycznych. Następnie podjęto kwestię znaczenia automatyzacji procesu analizy danych w codziennej pracy naukowców i praktyków, wykorzystujących metody statystyczne. W rozdziale trzecim opisano pokrótce sposób zaimplementowania metod grupowania danych w pakiecie STATISTICA, sygnalizując w ten sposób konieczność stworzenia narzędzia usprawniającego możliwości programu. Kolejne dwa punkty stanowią najważniejszą część pracy – zawierają opis aplikacji Automatyzacja grupowania i przykładowe wyniki jej zastosowania.

2. METODY TAKSONOMICZNE W NAUKACH SPOŁECZNYCH

Metody analiz taksonomicznych odpowiadają najbardziej pierwotnej potrzebie poznawczej człowieka, jaką jest porządkowanie otaczającej go rzeczywistości. W tym sensie oczywiście każdą metodę analizy danych można określić jako „taksonomiczną", jednak w statystyce terminem tym zwykło się określać procedury grupowania i porządkowania danych wielowymiarowych[2].

Metody taksonomiczne wykorzystywane są przez naukowców i praktyków z wielu dziedzin. Szczególnie często w badaniach z zakresu nauk społecznych – najbardziej popularne są tu klasyfikacje jednostek administracyjnych względem poziomu czy jakości życia ich mieszkańców. W ramach taksonomii numerycznej wyodrębnia się zwyczajowo dwie grupy metod, przy czym podział ten związany jest z celem prowadzonych analiz. Istnieją więc metody grupowania – służące do wyodrębniania skupień podobnych obiektów ze względu na wiele cech statystycznych – oraz metody klasyfikacji albo porządkowania, które służą do tworzenia rankingu obiektów ze względu na wiele cech statystycznych. Metody grupowania dzielą się na hierarchiczne (w których skład wchodzą metody podziałowe i aglomeracyjne) oraz niehierarchiczne. W obrębie porządkowania liniowego można wydzielić dwie podstawowe grupy metod – wzorcowe i bez wzorca. Narzędzie automatyzacji tworzenia rankingów opisano już we wcześniejszej pracy[3], natomiast tu przedstawione jest narzędzie automatyzacji grupowania hierarchicznego, metoda ta bowiem jest jedną z najbardziej popularnych w badaniach naukowych i zastosowaniach praktycznych.

[2] T. Grabiński, *Metody taksonometrii*, Wydawnictwo Akademii Ekonomicznej w Krakowie, Kraków 1990.

[3] M. Sobolewski, *Automatyzacja analiza taksonomicznych w programie STATISTICA*, StatSoft, Kraków 2014.

Nawet skrótowy przegląd publikacji, w których wykorzystywane są metody grupowania hierarchicznego, prowadzi do przekonania o ogromnej uniwersalności tych narzędzi analizy danych. Oto przykłady takich zastosowań:

- metody analizy skupień stosowano w badaniach czystości środowiska naturalnego do klasyfikacji jakości wody pitnej[4];
- interesującym i niebanalnym obszarem zastosowań metod grupowania jest historia sztuki – na przykład analizę skupień wykorzystano do klasyfikacji ikon według wieku ich powstawania na podstawie danych o ich składzie chemicznym[5];
- inny przykład praktycznego wykorzystania taksonomii numerycznej dotyczy klasyfikacji zagrożeń zawodowych i wyodrębnienia jednorodnych grup pracowników według tych zagrożeń[6];
- metody analizy skupień wykorzystuje się także w psychologii, zarówno w badaniach ogólnych, jak i w bardziej szczegółowych działach[7].

We wszystkich wymienionych publikacjach do obliczeń wykorzystano narzędzia grupowania danych zaimplementowane w programie STATISTICA. Jednakże prezentacja wyników (zarówno jeśli chodzi o formę, jak i ich zakres) w tych pracach może budzić pewne zastrzeżenia. Dla badaczy przydatne byłoby niewątpliwie narzędzie ułatwiające wykonywanie analiz i raportowanie uzyskanych wyników. Więcej przykładów zastosowań metod taksonomicznych, zwłaszcza w naukach humanistycznych i społecznych, można znaleźć w czasopismach branżowych, a także czasopismach statystycznych (np. „Wiadomości Statystyczne"), w monografiach dotyczących metod taksonomicznych[8]oraz w licznych publikacjach elektronicznych[9].

3. DLACZEGO WARTO EFEKTYWNIE ANALIZOWAĆ DANE?

Pytanie postawione w tytule tego podrozdziału brzmi może nieco retorycznie. Cóż bowiem złego może być w efektywnej analizie danych statystycznych (oczywiście zakładając, że w danym wypadku jakaś analiza statystyczna jest w ogóle potrzebna). Niemniej jednak pojęcie efektywności budzi nie tylko pozytywne skojarzenia. W swojej znakomitej książce *Technopol* Neil Postman sugeruje, by technologie zwiększające

[4] K. Rymuza, E. Radzka, *Zastosowanie analiz wielowymiarowych do oceny jakości wody pitnej*, „Żywność. Nauka. Technologia. Jakość" 91/6 (2013), s. 165–174.

[5] M. Pańczyk, E. Pańczyk, L. Giro, E. Gaździcka, J. Giemza, J. Świetlik-Olszewska, *Zastosowanie skaningowej mikroskopii elektronowej i instrumentalnej neutronowej analizy aktywacyjnej do identyfikacji pigmentów z ikony chrzest Chrystusa*, „Biuletyn Państwowego Instytutu Geologicznego" 2010/439, s. 459–468.

[6] M. Lotko, A. Lotko, *Zastosowanie analizy skupień do oceny zagrożeń zawodowych pracowników wiedzy i ich postaw wobec charakteru pracy*, „Eksploatacja i Niezawodność" 17/1 (2015), s. 80–89.

[7] E. Aranowska, T. Rogala, *Użyteczność analizy skupień (Cluster Analysis) w psychoakustyce*, 44 Otwarte Seminarium z Akustyki, Polskie Towarzystwo Akustyczne, Komitet Akustyki PAN, Gdańsk, s. 113–116.

[8] T. Grabiński, *Metody taksonometrii*, Wydawnictwo Akademii Ekonomicznej w Krakowie, Kraków 1990; Pociecha J., Podolec B., Sokołowski A., Zając K., *Metody taksonomiczne w badaniach społeczno-ekonomicznych*, PWN, Warszawa 1988.

[9]Liczne publikacje, w których wykorzystano metody analizy skupień w nauce i praktyce, znaleźć można w czytelni udostępnionej użytkownikom programu na stronie www.statsoft.pl.

efektywność funkcjonowania naszych społeczności oceniać nie tylko pod kątem problemów, jakie rozwiązują, lecz także brać pod uwagę problemy, jakie stwarzają[10].

Pamiętając jednak o zastrzeżeniach wysuwanych przez Postmana, można stwierdzić, że efektywna analiza danych nie niesie za sobą jakichś poważnych zagrożeń dla rzetelności i celowości prowadzenia analiz statystycznych. Może się oczywiście zdarzyć, że ktoś wykona więcej analiz tylko dlatego, że dysponował odpowiednim, efektywnym narzędziem programistycznym, ale nie wpłynie to na ocenę wartości tych dokonań.

Podczas realizacji każdego etapu pracy z danymi wykonuje się wiele powtarzalnych, żmudnych czynności, które czynią analizę statystyczną czasochłonną, stwarzając dodatkowo ryzyko popełnienia licznych błędów i omyłek. Tymczasem wiele z tych czynności można w prosty sposób zautomatyzować, czyniąc pracę z komputerem efektywniejszą, powtarzalną i bezbłędną.

4. IMPLEMENTACJA METOD GRUPOWANIA W PROGRAMIE STATISTICA

W programie STATISTICA zaimplementowano metody grupowania hierarchicznego i metodę *k*-średnich. Przedmiotem rozważań są te pierwsze. W grupie analiz określonych jako STATYSTYKI WIELOWYMIAROWE użytkownik programu STATISTICA znajdzie moduł ANALIZA SKUPIEŃ. Z jego pomocą można dokonać procesu grupowania hierarchicznego na podstawie kilku rodzajów odległości oraz kilku zasad wyznaczania odległości pomiędzy skupieniami. Z punktu widzenia użytkownika, który zainteresowany jest przede wszystkim efektywnym przeprowadzeniem podziału badanej zbiorowości oraz opisem uzyskanych skupień, program ma jednak kilka braków. Wymienić tu należy przede wszystkim: brak opcji standaryzacji w ramach analizy, brak możliwości wyboru punktu odcięcia na podstawie diagramu połączeń (dendrogramu) i automatycznego przypisania analizowanych przypadków do poszczególnych skupień w wyjściowym arkuszu danych. Ręczne przypisywanie numerów grup jest jeszcze możliwe do przeprowadzenia dla kilku czy kilkunastu obiektów, staje się jednak niemal niewykonalne przy kilkudziesięciu czy kilkuset przypadkach.

Większość z wymienionych niedogodności rozwiązano pakiecie STATISTICA PLUS[11]opracowanym przez firmę StatSoft Polska. Wprowadzono w nim opcję standaryzacji danych, możliwość definiowania punktu odcięcia, jak również zapisywania informacji o dokonanym podziale w wejściowym arkuszu danych[12].

Natomiast opisana w kolejnych punktach autorska propozycja uzupełnienia możliwości pakietu STATISTICA idzie o krok dalej, przede wszystkim pod kątem komponowania całych schematów analiz. Wzbogacone są również pewne elementy graficznej prezentacji danych w postaci dendrogramów. Automatycznie wyliczane są charakterystyki grup. Istotne jest również to, że w nakładce Automatyzacja grupowania użytkownik uzyskuje wszystkie wyniki w starannie sformatowanych tabelach dokumentu programu Word. Szczegółowy opis programu znajduje się w następnym punkcie.

---

[10] N. Postman, *Technopol*, Warszawskie Wydawnictwo Literackie MUZA, Warszawa 2004.

[11] G. Migut, *Nowe możliwości analizy danych – STATISTICA zestaw plus*, [w:] *Analiza danych w programie STATISTICA – przegląd*, StatSoft, Kraków 2012.

[12] *STATISTICA Zestaw Plus – instrukcja instalacji oraz podstawowe informacje o programie*, StatSoft, Kraków 2012.

5. MOŻLIWOŚCI PROGRAMUAUTOMATYZACJA GRUPOWANIA

Program STATISTICA ma wbudowany język programowania STATISTICA Visual Basic, który pozwala tworzyć własne aplikacje, w których wykorzystywane są możliwości programu STATISTICA. Ponieważ Visual Basic stanowi również język programowania w programie Word(czy Excel), możliwe jest korzystanie także z opcji dostępnych w tych programach[13].Interfejs programu umożliwia dobór listy metod grupowania hierarchicznego, które będą wykorzystane podczas analizy. Dodatkowo do każdej kombinacji metody wiązania oraz pomiaru odległości, które standardowo oferuje program STATISTICA, dołączona jest możliwość zdefiniowania odrębnej metody normalizacji[14]. Program wykona obliczenia dla wszystkich wskazanych metod i umieści wyniki w zbiorczym arkuszu oraz raporcie programu Word.

Dla każdej metody grupowania można określić jedną lub kilka metod „cięcia" dendrogramu. Użytkownik będzie mógł uczynić to przed rozpoczęciem obliczeń, na podstawie swoich wcześniejszych założeń lub też określać je dla każdego dendrogramu interaktywnie. Możliwe więc będzie ustalenie, czy wszystkie dendrogramy mają generować podział na przykładna cztery grupy, ale można określać to też indywidualnie dla każdego dendrogramu. Dostępne są również automatyczne metody dokonywania podziału (np. na podstawie maksymalnego skoku). Po zdefiniowaniu dowolnej liczby podziałów analizowanego zbioru danych zostaną one zapisane w formie dodatkowych kolumn w wyjściowym arkuszu programu STATISTICA.

Raport w programie Word może obejmować dendrogramy, tabele średnich grupowych dla zmiennych diagnostycznych, wskaźniki średnich grupowych oraz tabele zgodności różnych podziałów. Lista najważniejszych zalet programu Automatyzacja grupowania jest następująca:

- interaktywne określenie punktu odcięcia lub liczby wyodrębnianych grup;
- możliwość automatycznego określenia poziomu odcięcia (liczby grup);
- formatowanie dendrogramu;
- zestawienie średnich grupowych i wskaźników grupowych w formie tabel programu Word;
- możliwość wyboru kilku metod grupowania i porównanie wyników;
- możliwość zapisu całego schematu analizy (metody, punkty odcięcia, opcje wyników) i wykorzystanie do ponownego przeprowadzenia obliczeń na oryginalnych lub zmienionych danych.

6. PRZYKŁADOWE WYNIKI

Na rysunku1pokazano interfejs programu Automatyzacja grupowania z przykładowo zdefiniowanymi metodami grupowania, punktami odcięcia i innymi opcjami.

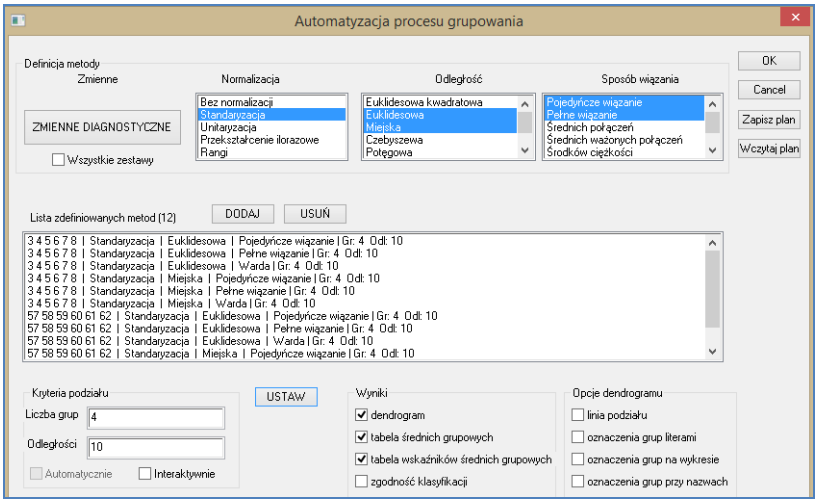
Przykładowe wyniki analiz przeprowadzonych za pomocą programu Automatyzacja grupowania dotyczyć będą klasyfikacji miast na prawach powiatu według poziomu życia w latach 2003–2012. Wejściowe zmienne diagnostyczne są następujące:

- wielkość oddziału w szkołach;

[13] G. Migut , *Automatyzacja raportowania w STATISTICA*, StatSoft Polska, Kraków 2009

[14] W programie uwzględniono kilka metod normalizacji danych. Opisywane narzędzie programistyczne można też wykorzystać do rozważań czysto metodologicznych, zajmując się na przykład kwestią wpływu sposobu normalizacji na wyniki grupowania.

- wskaźnik bezrobocia;
- mieszkania na 1000 mieszkańców;
- wskaźnik zgonów w wieku do 65 lat;
- przeciętne wynagrodzenie;
- wskaźnik motoryzacji.



Rys. 1. Interfejs programu Automatyzacja grupowania wraz z dwunastoma przykładowo zdefiniowanymi procedurami grupowania dla dwóch zestawów zmiennych (każdy z punktami odcięcia dla odległości aglomeracyjnej równej 10 lub czterech grup)

Źródło: opracowanie własne.

Podczas przykładowej analizy wykorzystano kilka alternatywnych metod grupowania, oceniono też zgodność uzyskanych podziałów. Przedstawiono charakterystykę uzyskanych grup za pomocą średnich i wskaźników średnich grupowych. Analiza będzie miała po części charakter dynamiczny – przeprowadzone zostanie grupowanie dla danych 2003 i 2012 roku wraz z oceną zgodności uzyskanych wyników.

Problemem dla praktyka może być fakt, że wyniki grupowania zależą od szczegółów rachunkowych zastosowanych metod. Oczywiście tego problemu nie da się rozwiązać w jednoznaczny sposób, choć w literaturze można znaleźć wiele sugestii na temat skuteczności poszczególnych metod grupowania hierarchicznego (w powszechnej opinii najlepsze wyniki można uzyskać, stosując algorytm Warda[15]).

Można jednak próbować ocenić, na ile wyniki grupowania uzyskanego różnymi metodami są ze sobą zgodne. Jeżeli ich zgodność jest wysoka, to problem doboru metody staje się nieistotny. Wspólną ideą proponowanych przez autora niniejszego artykułu narzędzi automatyzacji analiz danych jest możliwość zdefiniowania listy alternatywnych analiz oraz uzyskanie ocen zgodności uzyskanych wyników.

[15] T. Grabiński, *Metody taksonometrii*, Wydawnictwo Akademii Ekonomicznej w Krakowie, Kraków 1990.

W programie Automatyzacja grupowania użytkownik może zdefiniować cały zestaw metod i wyznaczyć współczynnik zgodności[16] pomiędzy różnymi klasyfikacjami.

Tabela 1. Współczynniki zgodności podziału zbiorowości miast na prawach powiatu na cztery grupy, uzyskane za pomocą sześciu alternatywnych metod grupowania

Numer kolejnej metody	1	2	3	4	5	6
1	×	0,76	1,00	0,73	0,67	0,65
2	0,76	×	0,76	0,87	0,71	0,81
3	1,00	0,76	×	0,73	0,67	0,65
4	0,73	0,87	0,73	×	0,64	0,78
5	0,67	0,71	0,67	0,64	×	0,62
6	0,65	0,81	0,65	0,78	0,62	×

1 – odległość: euklidesowa kwadratowa metoda wiązania: zupełne połączenia; 2 – odległość: euklidesowa kwadratowa metoda wiązania: Warda; 3 – odległość: euklidesowa metoda wiązania: zupełne połączenia; 4 – odległość: euklidesowa metoda wiązania: Warda; 5 – odległość: miejska metoda wiązania: zupełne połączenia; 6 – odległość: miejska kwadratowa metoda wiązania: Warda

Źródło: opracowanie własne.

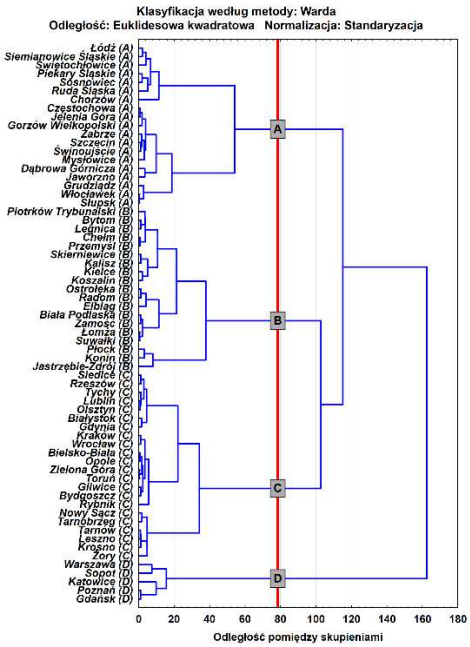
W tabeli1 przedstawiono współczynniki zgodności pomiędzy podziałem na cztery skupienia, dokonanym za pomocą sześciu alternatywnych metod – kombinacji trzech miar odległości oraz dwóch sposobów aglomeracji. Stosunkowo wysokie wartości tych wskaźników pozwalają stwierdzić, że wyniki grupowania nie zależą w znaczącym stopniu od rodzaju wykorzystanej metody.

Na podstawie dendrogramu (rys. 2) można zaproponować podział na cztery skupienia. Dendrogram został uzupełniony o odpowiednią linię odcięcia, miejsca podziału i przynależność poszczególnych obiektów do grup zaś zostały wyraźnie opisane na wykresie. Na przykład z wykresu bez trudu można odczytać, że grupę D tworzy pięć miast: Warszawa, Sopot, Katowice, Poznań i Gdańsk. Program Automatyzacja grupowania zakodował przynależność do grup w dodatkowej kolumnie wejściowego arkusza danych. Na tej podstawie w dokumencie programu Word zostały automatycznie stworzone tabele, zawierające charakterystykę wyodrębnionych skupień (tab. 2 i 3).

Aby scharakteryzować poszczególne grupy wyodrębnionych obiektów, wyznacza się średnie grupowe lub wskaźniki średnich grupowych. Program Automatyzacja grupowania generuje takie zestawienia w formie gotowych tabel w programie Word. Staranne sformatowanie wyników ułatwia ich interpretację – wartości najkorzystniejsze[17] zostały wyróżnione odcieniami zieleni, najmniej korzystne zaś – odcieniami koloru czerwonego. Dzięki temu łatwo opisać i wartościować wyodrębnione skupienia. Warto też podkreślić, że definiując listę analiz w wejściowym panelu programu, można dobierać różne zestawy zmiennych. W ten sposób dokonano porównania zgodności wyników grupowania (przy podziale na cztery skupienia) dla danych z lat 2003 i 2012. Wyznaczony współczynnik zgodności wynosił 0,64, co oznacza dość dużą zgodność podziałów uzyskanych na przestrzeni dziesięciu lat.

[16]W literaturze zaproponowano liczne miary zgodności dwóch wyników grupowania (czy bardziej naukowo – podziałów zbioru skończonego). W aplikacji Automatyzacja grupowania wykorzystano miarę bazującą na tabeli krzyżowej stworzonej dla par grupowanych obiektów. Metoda ta została szczegółowo opisana w pracy Edwarda Nowaka [7].

[17]Jeżeli w arkuszu danych wskazano destymulanty (czyli cechy, których niskie wartości uznaje się za korzystniejsze), zostanie to uwzględnione w zastosowanych schematach kolorystycznych.



Rys. 2. Wyniki grupowania – wykres przebiegu aglomeracji wraz z naniesioną linią odcięcia i nazwami grup, automatycznie zamieszczony w programie Word

Źródło: opracowanie własne.

Tabela 2. Charakterystyka utworzonych skupień – średnie wartości zmiennych diagnostycznych

Zmienne diagnostyczne	**Średnie w grupach**				***p***
	A	**B**	**C**	**D**	
Wielkość oddziału w szkołach 2003	22,3	24,7	22,5	21,2	0,0000***
Wskaźnik bezrobocia 2003	13,2	14,0	9,4	6,8	0,0000***
Mieszkania na 1000 mieszkańców 2003	376,2	344,5	349,1	401,3	0,0001***
Wskaźnik zgonów w wieku do 65 lat 2003	4,0	3,2	2,9	3,4	0,0000***
Przeciętne wynagrodzenie 2003	2100	2131	2170	2800	0,0028**
Wskaźnik motoryzacji 2003	253,3	264,2	281,4	387,2	0,0008***

p – wynik testu Kruskala-Wallisa

Źródło: opracowanie własne.

Tabela 3. Charakterystyka utworzonych skupień – wskaźniki średnich grupowych

Zmienne diagnostyczne	**Wskaźniki średnich grupowych**			
	A	**B**	**C**	**D**
Wielkość oddziału w szkołach 2003	0,97	1,07	0,98	0,92
Wskaźnik bezrobocia 2003	1,13	1,21	0,80	0,58
Mieszkania na 1000 mieszkańców 2003	1,05	0,96	0,97	1,12
Wskaźnik zgonów w wieku do 65 lat 2003	1,19	0,96	0,86	1,01
Przeciętne wynagrodzenie 2003	0,96	0,97	0,99	1,28
Wskaźnik motoryzacji 2003	0,92	0,96	1,02	1,40

Źródło: opracowanie własne.

Przeprowadzenie wszystkich opisanych analiz przy użyciu standardowych narzędzi dostępnych w programie STATISTICA byłoby bardzo czasochłonne. Ważna jest też forma prezentacji wyników, nawet wartościowe wyniki bowiem przedstawione w nieciekawy sposób nie zainteresują nikogo[18]. Cóż dopiero mówić o powtórzeniu tych analiz dla innego układu zmiennych lub konieczności poprawienia wyników, gdyby w wyjściowych arkuszu danych pojawiły się błędne wartości. Opcja Zapisz plan pozwala zapisać schemat wszystkich zdefiniowanych analiz i wykonać go ponownie za pomocą jednego kliknięcia.

7. PODSUMOWANIE

W artykule opisano potrzebę tworzenia kompleksowych systemów analizy danych, które oferowałyby nie tylko możliwość przeprowadzenia wyliczeń, ale także sprawne zarządzanie dużą ilością wyników (łącznie z automatycznym ich formatowaniem). Takie systemy analityczne, przy daleko posuniętej automatyzacji wykonywania prostych czynności, pozwalałyby także na wielowariantowe obliczenia oraz ich łatwe powtarzanie. Główna idea jest taka, że nowoczesne narzędzia analizy danych powinny sprowadzić do niemal tego samego wykonania jednej, kilku czy kilkuset analiz, powinny też oferować możliwość porównania uzyskiwanych wyników. ogram Automatyzacja grupowania spełnia te wymogi. W artykule opisano jego możliwości, ilustrując je dodatkowo konkretnymi wynikami. Będzie on oczywiście udoskonalany. Przykładowe kierunki dalszej jego rozbudowy, które znajdą się w drugiej, rozszerzonej wersji, są następujące:

- zwiększenie liczby dostępnych procedur normalizacyjnych, współczynników zgodności i innych opcjonalnych wyników;
- wykresy profili skupień w formie wykresu słupkowego lub radarowego;
- analiza grupami (na przykład grupowanie względem tych samych zmiennych dla różnych okresów czasowych);
- wprowadzenie metody eliminacji zmiennych na podstawie macierzy korelacji.

Należy podkreślić, że program będzie udostępniany wszystkim naukowcom i badaczom zainteresowanym prowadzeniem analiz skupień, stworzenie tej aplikacji ma na celu upowszechnienie i ułatwienie stosowania w praktycetej bardzo ciekawej techniki analizy danych.

LITERATURA

[1] Aranowska E., Rogala T., *Użyteczność analizy skupień (Cluster Analysis) w psychoakustyce*, 44 Otwarte Seminarium z Akustyki, Polskie Towarzystwo Akustyczne, Komitet Akustyki PAN, Gdańsk, s. 113–116.

[2] Kozak M., *Statystyka: wizualizacja, korelacja, publikacja*, [w:] *Zastosowania statystyki i data mining w badaniach naukowych*, StatSoft, Kraków 2012.

[3] Lotko M., Lotko A., *Zastosowanie analizy skupień do oceny zagrożeń zawodowych pracowników wiedzy i ich postaw wobec charakteru pracy*, „Eksploatacja i Niezawodność" 17/1 (2015), s. 80–89.

[4] Migut G., *Automatyzacja raportowania w STATISTICA*, StatSoft Polska, Kraków 2009.

[18] M. Kozak, *Statystyka: wizualizacja, korelacja, publikacja*, [w:] *Zastosowania statystyki i data mining w badaniach naukowych*, StatSoft, Kraków 2012.

[5] Migut G., *Nowe możliwości analizy danych – STATISTICA zestaw plus*, [w:] *Analiza danych w programie STATISTICA – przegląd*, StatSoft, Kraków 2012.
[6] Nowak E., *Metody taksonomiczne w klasyfikacji obiektów społeczno-gospodarczych*, PWE, Warszawa 1990.
[7] Pańczyk M., Pańczyk E., Giro L., Gaździcka E., Giemza J., Świetlik-Olszewska, *Zastosowanie skaningowej mikroskopii elektronowej i instrumentalnej neutronowej analizy aktywacyjnej do identyfikacji pigmentów z ikony chrzest Chrystusa*, „Biuletyn Państwowego Instytutu Geologicznego" 2010/439, s. 459–468.
[8] Pociecha J., Podolec B., Sokołowski A., Zając K., *Metody taksonomiczne w badaniach społeczno-ekonomicznych*, PWN, Warszawa 1988.
[9] Postman N., *Technopol*, Warszawskie Wydawnictwo Literackie MUZA, Warszawa 2004.
[10] Rymuza K, Radzka E., *Zastosowanie analiz wielowymiarowych do oceny jakości wody pitnej*, „ŻYWNOŚĆ. Nauka. Technologia. Jakość"91/6 (2013), s. 165–174.
[11] Sobolewski M., *Automatyzacja analiza taksonomicznych w programie STATISTICA*, StatSoft, Kraków 2014.
[12] *STATISTICA Zestaw Plus – instrukcja instalacji oraz podstawowe informacje o programie*, StatSoft, Kraków 2012.

AUTOMATION OF CLUSTERING IN STATISTICA

This paper describes an application that allows to automate the process of cluster analysis by using STATISTICA. The more general purpose of the article was, furthermore, to outline a universal concept of data analysis system, which would enable the efficient execution of multiple and complex statistical analysis.First, a review of the publications which used data taxonomic grouping method was done. The range of actual and potential applications of taxonomic methods by researchers from various fields of science (in particular in the field of social sciences and humanities) was described. Then there were identified some shortcomings in the implementation of clustering methods of data in statistical packets.To meet the expectations of practitioners, a program called Automation grouping was developed, which is an extension of STATISTICA, enabling effective use of taxonomic methods in practice. Among many benefits of the use of this program one can mention the following: the possibility of parallel execution of several alternative clustering algorithms and comparing the results, setting a number of divisions for each defined method to automatically generate tables that contain the characteristics of clusters created in a Worddocument. An analysis diagram can be saved and used for further calculations on the data set of similar structure (e.g. after getting new experimental results and the increase in the number of cases in the database).

The results of the program operation are presented on the example of the social sciences, which dealt with the analysis of the standard of living in towns with county rights in 2003–2012. Since priority of the resulting work is to promote the use of taxonomic methods in practice, the application described will be made available to all interested users.

Keywords: Cluster analysis, STATISTICA, Visual Basic

DOI:10.7862/rz.2015.hss.54

Przesłano do redakcji: marzec 2015
Przyjęto do druku: grudzień 2015