PRACE NAUKOWE
Uniwersytetu Ekonomicznego we Wrocławiu

RESEARCH PAPERS
of Wrocław University of Economics

Nr 426

Taksonomia 26
Klasyfikacja i analiza danych – teoria i zastosowania

Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2016

Redaktor Wydawnictwa: Agnieszka Flasińska

Redaktor techniczny: Barbara Łopusiewicz

Korektor: Barbara Cibis

Łamanie: Beata Mazur

Projekt okładki: Beata Dębska

Tytuł dofinansowany ze środków Narodowego Banku Polskiego
oraz ze środków Sekcji Klasyfikacji i Analizy Danych PTS

Informacje o naborze artykułów i zasadach recenzowania
znajdują się na stronach internetowych
www.pracenaukowe.ue.wroc.pl
www.wydawnictwo.ue.wroc.pl

ISSN 1899-3192 (Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu)
e-ISSN 2392-0041
ISSN 1505-9332 (Taksonomia)

Wersja pierwotna: publikacja drukowana

Zamówienia na opublikowane prace należy składać na adres:
Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
ul. Komandorska 118/120, 53-345 Wrocław
tel./fax 71 36 80 602; e-mail:econbook@ue.wroc.pl
www.ksiegarnia.ue.wroc.pl

Druk i oprawa: TOTEM

Spis treści

Wstęp

W dniach 14–16 września 2015 r. w Hotelu Novotel Gdańsk Marina w Gdańsku odbyła się XXIV Konferencja Naukowa Sekcji Klasyfikacji i Analizy Danych PTS (XXIX Konferencja Taksonomiczna) „Klasyfikacja i analiza danych – teoria i zastosowania", zorganizowana przez Sekcję Klasyfikacji i Analizy Danych Polskiego Towarzystwa Statystycznego oraz Katedrę Statystyki Wydziału Zarządzania Uniwersytetu Gdańskiego. Przewodniczącymi Komitetu Organizacyjnego konferencji byli prof. dr hab. Mirosław Szreder oraz dr hab. Krzysztof Najman, prof. nadzw. UG, sekretarzami naukowymi dr hab. Kamila Migdał-Najman, prof. nadzw. UG oraz dr hab. Anna Zamojska, prof. nadzw. UG, a sekretarzem organizacyjnym Anna Nowicka z Fundacji Rozwoju Uniwersytetu Gdańskiego.

Konferencja Naukowa została dofinansowana ze środków Narodowego Banku Polskiego.

Zakres tematyczny konferencji obejmował takie zagadnienia, jak:

a) teoria (taksonomia, analiza dyskryminacyjna, metody porządkowania liniowego, metody statystycznej analizy wielowymiarowej, metody analizy zmiennych ciągłych, metody analizy zmiennych dyskretnych, metody analizy danych symbolicznych, metody graficzne),

b) zastosowania (analiza danych finansowych, analiza danych marketingowych, analiza danych przestrzennych, inne zastosowania analizy danych – medycyna, psychologia, archeologia, itd., aplikacje komputerowe metod statystycznych).

Zasadniczymi celami konferencji SKAD były prezentacja osiągnięć i wymiana doświadczeń z zakresu teoretycznych i aplikacyjnych zagadnień klasyfikacji i analizy danych. Konferencja stanowi coroczne forum służące podsumowaniu obecnego stanu wiedzy, przedstawieniu i promocji dokonań nowatorskich oraz wskazaniu kierunków dalszych prac i badań.

W konferencji wzięło udział 81 osób. Byli to pracownicy oraz doktoranci następujących uczelni i instytucji: AGH w Krakowie, Politechniki Łódzkiej, Politechniki Gdańskiej, Politechniki Opolskiej, Politechniki Wrocławskiej, Szkoły Głównej Gospodarstwa Wiejskiego w Warszawie, Szkoły Głównej Handlowej w Warszawie, Uniwersytetu im. Adama Mickiewicza w Poznaniu, Uniwersytetu Ekonomicznego w Katowicach, Uniwersytetu Ekonomicznego w Krakowie, Uniwersytetu Ekonomicznego w Poznaniu, Uniwersytetu Ekonomicznego we Wrocławiu, Uniwersytetu Gdańskiego, Uniwersytetu Jana Kochanowskiego w Kielcach, Uniwersytetu Łódzkiego, Uniwersytetu Mikołaja Kopernika w Toruniu, Uniwersytetu Przyrodniczego w Poznaniu, Uniwersytetu Szczecińskiego, Uniwer-

sytetu w Białymstoku, Wyższej Szkoły Bankowej w Toruniu, a także przedstawiciele NBP i PBS Sp. z o.o.

W trakcie dwóch sesji plenarnych oraz trzynastu sesji równoległych wygłoszono 58 referatów poświęconych aspektom teoretycznym i aplikacyjnym zagadnienia klasyfikacji i analizy danych. Odbyła się również sesja plakatowa, na której zaprezentowano 14 plakatów. Obradom w poszczególnych sesjach konferencji przewodniczyli profesorowie: Józef Pociecha, Eugeniusz Gatnar, Tadeusz Trzaskalik, Krzysztof Jajuga, Marek Walesiak, Barbara Pawełek, Feliks Wysocki, Ewa Roszkowska, Andrzej Sokołowski, Andrzej Bąk, Tadeusz Kufel, Mirosław Krzyśko, Krzysztof Najman, Małgorzata Rószkiewicz, Mirosław Szreder.

Teksty 25 recenzowanych artykułów naukowych stanowią zawartość prezentowanej publikacji z serii „Taksonomia" nr 26. Pozostałe recenzowane artykuły znajdują się w „Taksonomii" nr 27.

W pierwszym dniu konferencji odbyło się posiedzenie członków Sekcji Klasyfikacji i Analizy Danych Polskiego Towarzystwa Statystycznego, któremu przewodniczył prof. dr hab. Józef Pociecha. Ustalono plan przebiegu zebrania obejmujący następujące punkty:

A. Sprawozdanie z działalności Sekcji Klasyfikacji i Analizy Danych PTS.
B. Informacje dotyczące planowanych konferencji krajowych i zagranicznych.
C. Organizacja konferencji SKAD PTS w latach 2016 i 2017.
D. Wybór przedstawiciela Rady Sekcji SKAD PTS do IFCS.
E. Dyskusja nad kierunkami rozwoju działalności Sekcji.

Prof. dr hab. Józef Pociecha otworzył posiedzenie Sekcji SKAD PTS. Sprawozdanie z działalności Sekcji Klasyfikacji i Analizy Danych PTS przedstawiła sekretarz naukowy Sekcji dr hab. Barbara Pawełek, prof. nadzw. UEK. Poinformowała, że obecnie Sekcja liczy 231 członków. Przypomniała, że na stronie internetowej Sekcji znajdują się regulamin, a także deklaracja członkowska. Poinformowała, że zostały opublikowane zeszyty z serii „Taksonomia" nr 24 i 25 (PN UE we Wrocławiu nr 384 i 385). W „Przeglądzie Statystycznym" (zeszyt 4/2014) ukazało się sprawozdanie z ubiegłorocznej konferencji SKAD, która odbyła się w Międzyzdrojach, w dniach 8–10 września 2014 r. Prof. Barbara Pawełek przedstawiła także informacje dotyczące działalności międzynarodowej oraz udziału w ważnych konferencjach członków i sympatyków SKAD.

W konferencji Międzynarodowego Stowarzyszenia Towarzystw Klasyfikacyjnych (IFCS – International Federation of Classification Societies) w dniach 6–8 lipca 2015 r. w Bolonii, zorganizowanej przez Università di Bologna, udział wzięło 19 osób z Polski (w tym 17 członków Sekcji), które wygłosiły 15 referatów (wkład członków SKAD – 79,0%). Ponadto prof. Józef Pociecha był członkiem Komitetu Naukowego Konferencji z ramienia SKAD, członkiem Międzynarodowego Komitetu Nagród IFCS oraz organizatorem i przewodniczącym sesji nt. „Classification models for forecasting of economic processes".

W konferencji „European Conference on Data Analysis" (Colchester, 2–4 września 2015 r.) zorganizowanej przez The German Classification Society (GfKl) we współpracy z The British Classification Society (BCS) i Sekcją Klasyfikacji i Analizy Danych PTS (SKAD) udział wzięło 18 osób z Polski (w tym 14 członków Sekcji), które wygłosiły 15 referatów (wkład członków SKAD – 66,0%). Ponadto profesorowie Krzysztof Jajuga oraz Józef Pociecha byli członkami Komitetu Naukowego konferencji, prof. Andrzej Dudek został poproszony przez organizatorów o przygotowanie referatu i wygłoszenie na Sesji Plenarnej „Cluster analysis in XXI century, new methods and tendencies", prof. Krzysztof Jajuga był przewodniczącym sesji plenarnej, przewodniczącym sesji nt. „Finance and economics II" oraz organizatorem i przewodniczącym sesji nt. „Data analysis in finance", prof. Józef Pociecha był organizatorem i przewodniczącym sesji nt. „Outliers in classification procedures – theory and practice", prof. Andrzej Dudek był przewodniczącym sesji nt. „Machine learning and knowledge discovery II".

Kolejny punkt posiedzenia Sekcji obejmował zapowiedzi najbliższych konferencji krajowych i zagranicznych, których tematyka jest zgodna z profilem Sekcji. Prof. dr hab. Józef Pociecha poinformował o dwóch wybranych konferencjach krajowych (były to XXXIV Konferencja Naukowa „Multivariate Statistical Analysis MSA 2015", Łódź, 16–18 listopada 2015 r. i X Międzynarodowa Konferencja Naukowa im. Profesora Aleksandra Zeliasia nt. „Modelowanie i prognozowanie zjawisk społeczno-gospodarczych", Zakopane, 10–13 maja 2016 r.) oraz o trzech wybranych konferencjach zagranicznych. Konferencja „European Conference on Data Analysis" odbędzie się na Uniwersytecie Ekonomicznym we Wrocławiu w dniach 26–28 września 2017 r. W przeddzień tej konferencji, tj. 25.09.2017 r., odbędzie się Niemiecko-Polskie Sympozjum nt. „Analizy danych i jej zastosowań GPSDAA 2017". Następna konferencja Międzynarodowego Stowarzyszenia Towarzystw Klasyfikacyjnych (IFCS) odbędzie się w 2017 r. w Tokio. W 2019 r. Niemiecko-Polskie Sympozjum nt. „Analizy danych i jej zastosowań GPSDAA 2019" organizuje prof. Andreas Geyer-Schultz w Karlsruhe.

W następnym punkcie posiedzenia podjęto kwestię organizacji kolejnych konferencji SKAD. SKAD 2016 zorganizuje Katedra Metod Statystycznych Wydziału Ekonomiczno-Socjologicznego Uniwersytetu Łódzkiego.

W kolejnej części zebrania dokonano wyboru przedstawiciela Rady Sekcji SKAD PTS do IFCS na kadencję 2016–2019. Powołano Komisję Skrutacyjną, której przewodniczącym został prof. Tadeusz Kufel, a członkami dr hab. Iwona Konarzewska i dr Dominik Rozkrut. Profesor Józef Pociecha poprosił zebranych o proponowanie kandydatur zgłaszając jednocześnie prof. Andrzeja Sokołowskiego. Wobec braku następnych kandydatur listę zamknięto. Komisja Skrutacyjna przeprowadziła głosowanie tajne. W głosowaniu uczestniczyło 41 członków Sekcji. Profesor Andrzej Sokołowski został przedstawicielem Rady Sekcji SKAD PTS do

IFCS na kadencję 2016–2019, uzyskując następujący wynik: 39 głosów na „tak", 1 głos na „nie", 1 głos był nieważny.

W ostatnim punkcie zebrania dyskutowano nad kierunkami rozwoju działalności Sekcji obejmującymi następujące problemy: udział w międzynarodowym ruchu naukowym (wspólne granty, publikacje), umiędzynarodowienie konferencji SKAD (uczestnicy zagraniczni, dwujęzyczność konferencji), wydawanie własnego czasopisma.

Profesor Józef Pociecha zamknął posiedzenie Sekcji SKAD.

Krzysztof Jajuga, Marek Walesiak

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS nr 426 • 2016

Taksonomia 26 ISSN 1899-3192
Klasyfikacja i analiza danych – teoria i zastosowania e-ISSN 2392-0041

Adam Sagan
Uniwersytet Ekonomiczny w Krakowie,
e-mail: adam.sagan@uek.krakow.pl

Marcin Pełka
Uniwersytet Ekonomiczny we Wrocławiu
marcin.pelka@ue.wroc.pl

ANALIZA WIELOPOZIOMOWA Z WYKORZYSTANIEM DANYCH SYMBOLICZNYCH
MULTILEVEL ANALYSIS WITH APPLICATION OF SYMBOLIC DATA

DOI: 10.15611/pn.2016.426.18

Streszczenie: W artykule zaproponowano zastosowanie analizy danych symbolicznych w analizie danych wielopoziomowych. Podejście wielopoziomowe pozwala na uwzględnienie heterogeniczności populacji oraz przynależności respondentów do hierarchicznych układów funkcjonalnych. Celem artykułu jest budowa i estymacja wielopoziomowych modeli regresyjnych z wykorzystaniem danych symbolicznych na przykładzie danych o inwestycjach gospodarstw domowych w Polsce. W części empirycznej przedstawiono wyniki autorskich badań na podstawie ogólnopolskich badań ankietowych. Przeprowadzone analizy pozwoliły na zbadanie wpływu wieku respondenta, wieku najmłodszego dziecka oraz wieku decydenta na podejmowane decyzje inwestycyjne. Wyniki te wskazują także, że analiza danych symbolicznych może znaleźć zastosowanie w badaniach wielopoziomowych, a jej zastosowanie pozwala na szerszy opis obiektów zarówno na poziomie indywidualnym oraz grupowym.

Słowa kluczowe: analiza danych symbolicznych, analiza wielopoziomowa, dane symboliczne interwałowe.

Summary: The paper proposes the application of symbolic data analysis in the context of multilevel analysis. Multilevel analysis allows for taking into account heterogeneity of the data and the fact that respondents may be "nested" in many different functional elements. The main aim of the paper is the estimation of multilevel regression models for investment strategies in Polish households. The empirical part shows the results of the research for Polish households. The results suggest that symbolic data analysis may be a useful tool for multilevel analysis of data. It allows for describing objects in more detailed way – both at individual and aggregate level.

Keywords: multilevel analysis, symbolic data analysis, symbolic interval-valued data.

1. Wstęp

We współczesnych badaniach społecznych i marketingowych coraz większe znaczenie mają analizy pozwalające na uwzględnienie heterogeniczności populacji wynikającej z przyjętego schematu losowania próby oraz przynależności respondentów do hierarchicznych układów instytucjonalnych. Z tego punktu widzenia uzyskiwane dane mają często charakter zagnieżdżony. Przykładem jest zespołowy dobór próby i prowadzenie badań w układach instytucjonalnych (pracownicy w przedsiębiorstwach, pacjenci w szpitalach, uczniowie w klasach itp.). Z drugiej strony dane uzyskiwane w badaniach ankietowych są analizowane w postaci zmiennych z interwałami, wielowariantowych lub nominalnych, które mogą być przedmiotem agregacji na wyższym poziomie uogólnienia.

Dane tego typu są przedmiotem analiz głównie w zakresie analizy wielopoziomowej, niemniej jednak w zakresie analizy danych symbolicznych opracowano modele i metody analizowania danych na poziomie indywidualnym oraz zagregowanym, co może stanowić użyteczne narzędzie w zakresie analiz wielopoziomowych.

Celem artykułu jest budowa i estymacja wielopoziomowych modeli regresyjnych z wykorzystaniem danych symbolicznych na przykładzie danych o inwestycjach gospodarstw domowych w Polsce. Są one zbudowane na podstawie wyników ogólnopolskich badań ankietowych. Analiza danych symbolicznych doczekała się zastosowań w wielu metodach wielowymiarowych, jednak wykorzystanie ich w modelowaniu wielopoziomowym stanowi ciągle istotne pole dalszych poszukiwań.

2. Założenia modeli wielopoziomowych

Modele wielopoziomowe stanowią szeroki nurt analiz regresyjnych z wykorzystaniem prób złożonych. Modele te, w zależności od obszaru zastosowania, znane są w literaturze również jako hierarchiczne modele liniowe (HLM), modele z losowymi współczynnikami regresji (RCM), analiza kontekstowa, liniowe modele mieszane (LMEM) czy modele z efektami mieszanymi (MEM) (zob. m.in. [Bryk, Raudenbush 1992; Hox 2010; Goldstein 1995]). Warunkiem zastosowania modeli wielopoziomowych jest zagnieżdżony charakter danych, w którym losowane są z populacji jednostki zarówno z pierwszego, jak i z drugiego poziomu losowania.

Przykładem tego typu analiz są analizy osiągnięć uczniów w klasach szkolnych, firm w sektorach, głosowań wyborców w okręgach wyborczych, zadowolenia pacjentów w szpitalach czy kolejne powtarzane pomiary w badaniach panelowych. W przypadku danych tego typu nie można mówić o niezależności obserwacji Wynika to z zagnieżdżenia elementów niższych poziomów w elementach wyższego poziomu, np. uczniów w klasach, oraz z zależności jednostek z niższego poziomu od charakterystyk poziomu wyższego (np. wpływu charakterystyk nauczyciela na po-

ziom osiągnięć uczniów w danej klasie). W powtarzalnych badaniach panelowych niezależność obserwacji jest łamana poprzez występowanie efektów „halo" (sugerowanie się odpowiedziami z poprzednich fal badań). Prowadzić to może do obciążonych (*biased*) błędów standardowych oszacowań parametrów.

Ze względu na możliwe zróżnicowanie parametrów modelu w przekroju grup model wielopoziomowy składa się z co najmniej dwóch poziomów. Na pierwszym poziomie analizy zmienne z poziomu indywidualnego są wyjaśniane przez predyktory z tego samego poziomu:

$$y_{ij} = \beta_{0j} + \beta_{1j} X_{1ij} + \beta_{2j} X_{21ij} + \varepsilon_{ij}, \tag{1}$$

gdzie: y_{ij} – zmienna zależna w przekroju respondenta i należącego do klasy j; X_{1ij} – pierwsza zmienna niezależna w przekroju respondenta i należącego do klasy j; X_{2ij} – druga zmienna niezależna w przekroju respondenta i należącego do klasy j; ε_{ij} – reszta w modelu.

Symbole 0j, 1j i 2j przy wyrazach wolnych i współczynnikach nachyleń oznaczają, że są one zmiennymi losowymi o rozkładzie normalnym w przekroju klas j. Na drugim poziomie analizy są one wyjaśniane przez odpowiednie predyktory z wyższego poziomu (grupowego):

$$\begin{aligned} \beta_{0j} &= \gamma_{00} + \gamma_{01} W_{1j} + u_{oj}, \\ \beta_{1j} &= \gamma_{10} + \gamma_{11} W_{1j} + u_{1j}, \\ \beta_{2j} &= \gamma_{20} + \gamma_{21} W_{1j} + u_{2j}, \end{aligned} \tag{2}$$

gdzie: β_{kj} – zmienne zależne w przekroju klas j, W_{kj} – predyktory z poziomu zespołowego, u_{kij} – reszty w modelach II poziomu.

Łącząc oba poziomy analizy, można otrzymać formę zredukowaną równania wielopoziomowego:

$$\begin{aligned} y_{ij} &= \gamma_{00} + \gamma_{01} W_{1j} + \gamma_{10} X_{1ij} + \gamma_{11} X_{1ij} W_{1j} + \gamma_{20} X_{2ij} + \\ &+ \gamma_{21} X_{2ij} W_{1j} + u_{0j} + u_{1j} X_{2ij} + u_{2j} X_{2ij} + \varepsilon_{ij}. \end{aligned} \tag{3}$$

Z punktu widzenia dekompozycji efektów wewnątrzklasowych I międzyklasowych, zmienne z poziomu 1 mogą być ujmowane jako suma dwóch składowych: 1) komponentu wewnątrzklasowego (odchyleń indywidualnych obserwacji od średniej grupowej $y_{ij} - \hat{y}_{\cdot j}$) i 2) komponentu międzyklasowego (różnic między średnimi grupowymi $\hat{y}_{\cdot j}$). Stąd całkowita kowariancja zmiennych Σ_T może być dekomponowana na macierz kowariancji międzyklasowej Σ_B oraz kowariancji wewnątrzklasowej Σ_W, tak że $\Sigma_T \; \Sigma_B + \Sigma_W$. Własność ta jest podstawą szacowania równań strukturalnych dla modeli wielopoziomowych.

Wyróżnić można wiele podejść do szacowania modeli wielopoziomowych. Do najważniejszych należą dwuetapowe podejście Muthéna MUML [Muthén1989, 1994], dwuetapowa bezpośrednia estymacja Goldsteina [1995], dwuetapowa metoda Chou, Bentlera i Pentza [1998], metoda GLAMM Skrondala i Rabe-Hesketh [2004], metoda Raudenbusha (zob. np. [Paterson, Goldstein 1991]), metoda Ansari i Jedidi [2000], metoda Asparouhova i Muthéna [2009].

Podejście wielopoziomowe pociąga za sobą występowanie indywidualnego i grupowego poziomu analizy. Ten rodzaj modeli jest szczególnie istotny dla danych symbolicznych cechujących się podobnym układem zależności.

3. Dane symboliczne jako specyficzny rodzaj danych wielopoziomowych

Obiekty symboliczne, w przeciwieństwie do obiektów w ujęciu klasycznym, mogą być opisywane przez następujące rodzaje zmiennych [Bock, Diday (red.), 2000, s. 2, 3; Billard, Diday 2006, s. 7–30; Dudek 2013, s. 35, 36; Diday, Noirhomme-Fraiture 2008, s. 10–19]:

- zmienne nominalne, porządkowe, przedziałowe oraz ilorazowe,
- zmienne interwałowe – czyli przedziały liczbowe,
- zmienne wielowariantowe – czyli listy kategorii lub wartości,
- zmienne wielowariantowe z wagami – czyli listy kategorii z wagami,
- zmienne histogramowe – czyli listy wartości z wagami.

Więcej o obiektach i zmiennych symbolicznych, sposobach otrzymywania zmiennych symbolicznych z baz danych, różnicach i podobieństwach między obiektami symbolicznymi a klasycznymi piszą m.in.: [Bock, Diday (red.) 2000, s. 2–8; Dudek 2013, s. 42, 43; 2004; Billard, Diday 2006, s. 7–66; Noirhomme-Fraiture, Brito 2011; Diday, Noirhomme-Fraiture 2008, s. 3–30].

W analizie danych symbolicznych mamy do czynienia z dwoma głównymi typami obiektów – obiektami symbolicznymi pierwszego rzędu i obiektami symbolicznymi drugiego rzędu, które mogą być traktowane jako specyficzny rodzaj danych wielopoziomowych.

Obiekty symboliczne pierwszego rzędu (*first-order objects*, *single individuals*) są to pojedyncze obiekty w sensie klasycznym – respondent A, produkt T, firma X, itp., z tym że są one opisywane przez zmienne symboliczne. Obiekty te reprezentują poziom indywidualny w sensie analizy wielopoziomowej. Zmienne symboliczne opisujące te obiekty otrzymuje się albo bezpośrednio z wykorzystaniem kwestionariusza ankiety (gdzie odpowiedzi mają charakter zmiennych symbolicznych) albo z wykorzystaniem danych klasycznych i dokonując ich agregacji z wykorzystaniem informacji o czasie (*temporal aggregation*) [Noirhomme-Fraiture, Brito 2011, s. 158]. Dane klasyczne, będące podstawą tego typu agregacji, zawarto w tab. 1, a wynik agregacji (obiekty symboliczne pierwszego rzędu, dane indywidualne) w tab. 2.

Tabela 1. Dane klasyczne będące podstawą agregacji

Id klienta	Wydana kwota	Zakupione produkty	Kata płatnicza	Miejsce zamieszkania
001	100	napoje	Visa	Wrocław
001	400	elektronika	Mastercard	Wrocław
002	200	ubrania	Visa	Wrocław
003	250	żywność	Visa	Wałbrzych
002	500	elektronika	Visa	Wrocław
003	450	ubrania	Electron	Wałbrzych
001	150	napoje	Visa	Wrocław

Źródło: opracowanie własne (dane sztuczne).

Tabela 2. Obiekty symboliczne (dane indywidualne) – agregacja danych z uwzględnieniem informacji o czasie (*temporal aggregation*)

Id klienta	Wydana kwota	Zakupione produkty	Kata płatnicza	Miejsce zamieszkania
001	<100, 400>	{napoje (2/3); elektronika (1/3)}	{Visa, Mastercard}	Wrocław
002	<200, 500>	{ubrania (1/2); elektronika (1/2)}	{Visa}	Wrocław
003	<250, 450>	{żywność (1/2); ubrania (1/2)}	{Visa, Electron}	Wałbrzych

Źródło: opracowanie własne (dane sztuczne).

Obiekty symboliczne drugiego rzędu (*second-order objects*, *super-individuals*, *aggregate objects*) są to mniej lub bardziej homogeniczne klasy, grupy obiektów w sensie klasycznym lub obiektów symbolicznych pierwszego rzędu. Obiekty te reprezentują poziom zagregowany w sensie analizy wielopoziomowej.

Obiekty symboliczne drugiego rzędu są najczęściej wynikiem agregacji danych z uwzględnieniem innych czynników niż czas (*contemporary aggregation*) [Noirhomme-Fraiture, Brito, 2011, s. 158].

W tabeli 3 zawarto dane klasyczne, na podstawie których w tab. 4 i 5 zawarto dane symboliczne drugiego rzędu.

Tabela 3. Dane klasyczne będące podstawą agregacji

Id respondenta	Miejsce zamieszkania	Id gospodarstwa domowego	Zarobki	Czas spędzony na dojazd do pracy	Wydatki na żywność
1	Wałbrzych	001	1500	10	300
2	Wałbrzych	001	1800	20	350
3	Wrocław	002	2500	10	250
4	Wrocław	002	2900	15	380
5	Wałbrzych	003	1800	25	150
6	Wałbrzych	003	2550	45	170

Źródło: opracowanie własne (dane sztuczne).

Tabela 4. Obiekty symboliczne drugiego rzędu – poziom gospodarstwa domowego

Id gospodarstwa domowego	Zarobki	Czas spędzony na dojazd do pracy	Wydatki na żywność
001	<1500, 1800>	<15, 20>	<300, 350>
002	<2500, 2900>	<10, 15>	<250, 380>
003	<1880, 2550>	<25, 45>	<150, 170>

Źródło: opracowanie własne (dane sztuczne).

Tabela 5. Obiekty symboliczne drugiego rzędu – poziom miasta

Miasto	Zarobki	Czas spędzony na dojazd do pracy	Wydatki na żywność
Wałbrzych	<1500, 2550>	<10, 45>	<150, 350>
Wrocław	<2500, 2900>	<10, 15>	<250, 380>

Źródło: opracowanie własne (dane sztuczne).

4. Korelacje, kowariancje i regresja liniowa danych symbolicznych interwałowych

W analizie wielopoziomowej możliwe jest wykorzystanie korelacji, kowariancji i regresji liniowej, dlatego konieczne jest zdefiniowanie tych miar na potrzeby danych symbolicznych interwałowych, które mają postać przedziału liczbowego o krańcach $[a_u, b_u]$, gdzie u – numer obiektu symbolicznego.

Średnia dla całej próby w przypadku zmiennych symbolicznych interwałowych wraża się wzorem [Billard, Diday 2006, s. 79]:

$$\overline{Z} = \frac{1}{2m}\sum_{u\in E}\left(b_u + a_u\right), \tag{4}$$

gdzie: m – liczba zmiennych, b_u, a_u – górny (dolny) kraniec zmiennej symbolicznej, u – numer obiektu symbolicznego, E – zbiór obiektów symbolicznych.

Wariancja dla całej próby obiektów symbolicznych wyraża się wzorem [Billard, Diday 2006, s. 80]:

$$\overline{S} = \frac{1}{4m^2}\left[\sum_{u\in E}\left(b_u + a_u\right)\right]^2, \tag{5}$$

gdzie oznaczenia jak we wzorze 4.

Korelacja pomiędzy dwiema zmiennymi symbolicznymi interwałowymi Z_1, Z_2 wyraża się wzorem [Billard, Diday 2006, s. 132]:

$$r(Z_1, Z_2) = \mathrm{Cov}(Z_1, Z_2) / S_{Z_1}, S_{Z_2}, \tag{6}$$

gdzie: $Cov(Z_1, Z_2)\,\mathrm{Cov}(Z_1, Z_2)$ – kowariancje pomiędzy zmiennymi Z_1, Z_2; S_{Z_1}, S_{Z_2} – wariancje pomiędzy zmiennymi Z_1, Z_2.

Kowariancje pomiędzy zmiennymi Z_1, Z_2 wyznaczane są zgodnie ze wzorem [Billard, Diday 2006, s. 132]:

$$Cov(Z_1, Z_2) = \frac{1}{3m} \sum_{u \in E} G_1 G_2 [Q_1 Q_2]^{\frac{1}{2}}, \tag{7}$$

$$Q_j = \left(a_{uj} - \overline{Z}_j\right)^2 + \left(a_{uj} - \overline{Z}_j\right)\left(b_{uj} - \overline{Z}_j\right) + \left(b_{uj} - \overline{Z}_j\right)^2, \tag{8}$$

$$G_j = \begin{cases} -1 & \text{dla} \quad \overline{Z}_{uj} \le \overline{Z}_j \\ 1 & \text{dla} \quad \overline{Z}_{uj} > \overline{Z}_j \end{cases}, \tag{9}$$

gdzie: $j = 1, 2$ – numer obiektu symbolicznego, pozostałe oznaczania jak we wzorach (4) oraz (5).

W przypadku regresji liniowej danych symbolicznych interwałowych stosowany jest standardowy model najmniejszych kwadratów (zob. np. [Pełka 2014]), w którym elementy macierzy $\mathbf{X}$ oraz $\mathbf{Y}$ w równaniu:

$$\hat{\mathbf{b}} = \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{y} \tag{10}$$

są zastępowane albo przez środki przedziałów liczbowych (metoda środków), albo przez środki i promienie tychże przedziałów liczbowych (metoda środków i promieni).

W części empirycznej artykułu zastosowana zostanie wyłącznie metoda środków (*center method*), gdzie wzór (10) przestawia się następująco [Lima-Neto, de Carvalho 2008, s. 1500–1515, 2010, s. 333–347; Billard, Diday, 2006, s. 198–201; Diday, Noirhomme-Fraiture 2008, s. 360, 361]:

$$\hat{\mathbf{b}} = \left(\left(\mathbf{X}^c\right)^T \left(\mathbf{X}^c\right)\right)^{-1} \left(\mathbf{X}^c\right)^T \mathbf{y}^c, \tag{11}$$

gdzie: $\mathbf{X}^c$ – macierz środków zmiennych objaśniających, $\mathbf{y}^c$ – macierz środków zmiennej objaśnianej.

Oszacowane wartości teoretyczne zmiennej objaśnianej oblicza się odrębnie dla krańców dolnych (oznaczanych indeksem L) i górnych (oznaczanych indeksem U) tej zmiennej, zgodnie ze wzorem:

$$\hat{\mathbf{y}}_L = \left(\mathbf{X}_L\right)^T \hat{\mathbf{b}} \text{ oraz } \hat{\mathbf{y}}_U = \left(\mathbf{X}_U\right)^T \hat{\mathbf{b}}. \tag{12}$$

Więcej o regresji liniowej danych symbolicznych interwałowych piszą m.in.: E.A. Lima-Neto i F.A.T. de Carvalho [2008, 2010], L. Billard i E. Diday [2006], E. Diday i M. Noirhomme-Fraiture [2008] oraz M. Pełka [2014].

5. Wyniki badań empirycznych

Oszacowanie modeli wielopoziomowych na podstawie danych empirycznych zostało dokonane na ogólnopolskiej próbie 1100 respondentów z 440 gospodarstw domowych[1]. Badania zostały przeprowadzone w 2012 r. za pomocą standaryzowanego wywiadu kwestionariuszowego przeprowadzonego wśród członków gospodarstwa domowego (ojca, matki i najstarszego dziecka powyżej 16. roku życia obecnego w domu).

Celem badania była identyfikacja preferencji i wartości gospodarstw domowych w obszarze kierunków alokacji zasobów na konsumpcję, oszczędzanie i inwestowanie. Jednym ze szczegółowych celów badania była ocena preferencji inwestycji w depozyty bankowe i inne instrumenty finansowe (akcje, obligacje).

W analizie wykorzystano wielopoziomowy model regresji liniowej danych symbolicznych interwałowych, gdzie zmienną zależną jest wielkość inwestycji w depozyty bankowe i inne instrumenty finansowe (y, zmienna symboliczna interwałowa), a zmiennymi niezależnymi są wiek respondenta (x_1), wiek najmłodszego dziecka w gospodarstwie domowym (x_2) oraz wiek decydenta (x_3).

Współczynnik ścieżkowy w przypadku modelu indywidualnego dla wieku respondenta, w przekroju całej próby, (y w zależności od x_1) wyniósł 0,011, a wyraz wolny 39,415. Współczynnik dopasowania R_L^2 wyniósł 0,65, a R_U^2 wyniósł 0,68[2]. Oznacza to, że wiek respondenta w niewielkim stopniu wpływa na podejmowane inwestycje (osoby starsze wiekiem skłonne są do podejmowania większych inwestycji niż osoby młodsze).

W przypadku dla modelu na poziomie zagregowanym, w przekroju całej próby, gdzie zbudowano jeden model regresji dla wszystkich zmiennych, współczynniki ścieżkowe (y w zależności od zmiennych x_1, x_2, x_3) wyniosły odpowiednio –0,029, 0,025 oraz 0,048, a wyraz wolny 28,859. Współczynnik dopasowania R_L^2 wyniósł 0,61, a R_U^2 wyniósł 0,60. Wyniki te sugerują, że wraz z rosnącym wiekiem (*ceteris paribus*) respondenci są mniej skłonni do inwestowania w instrumenty finansowe,

[1] Badania prowadzone były na reprezentatywnej próbie gospodarstw domowych w Polsce (w jego ramach ankietowani byli rodzice i najstarsze dziecko). Badanie dotyczyło wartości i preferencji członków gospodarstw domowych w zakresie konsumpcji, oszczędzania i inwestowania. Obszerniej badanie to opisuje publikacja A. Sagan [(red.) 2014].

[2] R_L^2 oraz R_U^2 są współczynnikami dopasowania modelu do danych, gdzie R_L^2 oznacza dopasowanie dla krańców dolnych, a R_U^2 dopasowanie dla krańców górnych zmiennej symbolicznej interwałowej.

natomiast zarówno wraz z rosnącym wiekiem najmłodszego dziecka (*ceteris paribus*) respondenci są bardziej skłonni do podejmowania tego typu inwestycji. Podobny wpływ na decyzje o inwestycjach ma tu wiek decydenta, którego oszacowany parametr ma największą wartość.

Dla poziomu zagregowanego zbudowano także trzy indywidualne modele, jeden dla każdej ze zmiennych niezależnych – wyniki zestawiono w tab. 6.

Tabela 6. Wyniki estymacji dla trzech różnych modeli

Współczynnik ścieżkowy	Wielkość współczynnika ścieżkowego	Wyraz wolny
$y = a_1 + b_1x_1$	$b_1 = 0{,}003$	$a_1 = 1{,}222$
$y = a_2 + b_2x_2$	$b_2 = 0{,}0564$	$a_2 = 12{,}6156$
$y = a_3 + b_3x_3$	$b_3 = 0{,}1478$	$a_3 = 44{,}5381$

Źródło: opracowanie własne (dane sztuczne).

Z tabeli 6 wynika, że w przypadku budowy różnych modeli wpływ poszczególnych zmiennych, mimo że szacowany osobno, jest zbliżony do ich wpływu w modelu uwzględniającym wszystkie zmienne. Modele zawarte w tab. 6 miały niewielkie wartości dopasowania do danych.

W badaniach podjęto także próbę połączenia danych z poziomu indywidualnego oraz poziomu gospodarstw domowych (zmienne x_1 oraz x_2) i oszacowano parametry modeli regresji w ramach każdego gospodarstwa domowego (419 odrębnych modeli). Niestety z powodu niewielkiej liczebności osób w gospodarstwach domowych, niemożliwe okazało się oszacowanie parametrów związanych ze zmienną x_2 (wiek najmłodszego dziecka w gospodarstwie domowym).

Oszacowania parametrów dla wybranych obserwacji są następujące:

$$\hat{y}_2 = 40{,}59 - 0{,}39x_1,$$
$$\hat{y}_6 = 50{,}16 - 0{,}50x_1,$$
$$\hat{y}_{15} = 40 + 0{,}72x_1,$$
$$\hat{y}_{418} = 24{,}72 - 0{,}09x_1,$$
$$\hat{y}_{419} = 30{,}50 - 0{,}45x_1.$$

W przypadku tych modeli dopasowanie do danych było bardzo niewielkie z uwagi na brak uwzględnienia w modelu zmiennej x_2.

Wyniki te, mimo braku możliwości oszacowania parametrów dla drugiej zmiennej, pozwalają w pewnym (choć ograniczonym stopniu) poznać wpływ wieku respondenta na podejmowane decyzje w ramach każdego z gospodarstw domowych. Wyniki wskazują, że wraz z wiekiem respondenta znaczna część gospodarstw domowych jest mniej skłonna do podejmowania ryzyka inwestycyjnego.

6. Podsumowanie

Dane symboliczne mogą być traktowane jako pewien rodzaj danych wielopoziomowych. Dane tego typu pozwalają na opisywanie obiektów na poziomie indywidualnym oraz zagregowanym. W analizie danych symbolicznych obiekty mogą być opisywane w bardziej złożony sposób z wykorzystaniem różnego typu zmiennych symbolicznych – interwałowych, wielowariantowych, histogramowych itd. Niemniej jednak taki opis danych wymaga zastosowania odpowiednich metod i narzędzi. Dla różnych typów zmiennych symbolicznych literatura przedmiotu zaproponowała różne rozwiązania w zakresie obliczania macierzy korelacji, kowariancji, średnich oraz wariancji Dane symboliczne mogą zostać zastosowane w regresji liniowej danych symbolicznych. Pozwala to na analizowanie zależności na poziomie indywidualnym i zagregowanym. Autorzy zaprezentowali, w jaki sposób dane symboliczne mogą być analizowane w kontekście analiz wielopoziomowych oraz zaproponowali trzy różne rozwiązania – model dla danych zagregowanych, model dla danych indywidualnych a także model łączący dane indywidualne i zagregowane.

Wadą proponowanego podejścia jest to, że nie jest to pełny model wielopoziomowy, którego budowa pozostaje zagadnieniem otwartym. Do innych ograniczeń można zaliczyć ograniczenia związane z regresją liniową danych symbolicznych – m.in. brak możliwości weryfikacji założeń modelu liniowego oraz brak rozwiązań w zakresie estymacji modeli dla danych symbolicznych różnych typów.

Literatura

Ansari A., Jedidi K., 2000, *Bayesian factor analysis for multilevel binary observations*, Psychometrika, vol. 65, no. 4, s. 475–496.

Asparouhov T., Muthén B., 2009, *Exploratory structural equation modeling*, Structural Equation Modeling, vol. 16, no. 3, s. 397–438.

Billard L., Diday E., 2006, *Symbolic Data Analysis. Conceptual Statistics and Data Mining*, John Wiley & Sons, Chichester.

Bock H.-H., Diday E. (red.), 2000, *Analysis of Symbolic Data. Explanatory Methods for Extracting Statistical Information from Complex Data*, Springer-Verlag, Berlin–Heidelberg.

Bryk A.S., Raudenbush S.W., 1992, *Hierarchical Linear Models: Applications and Data Analysis Methods*, Sage Publications, Thousand Oaks.

Chou C.P., Bentler P.M., Pentz M.A., 1998, *Comparisons of two statistical approaches to study growth curves: The multilevel model and the latent curve analysis*, Structural Equation Modeling, vol. 5, no. 3, s. 247–266.

Diday E., Noirhomme-Fraiture M., 2008, *Symbolic Data Analysis. Conceptual Statistics and Data Mining*, Wiley, Chichester.

Dudek A., 2013, *Metody analizy danych symbolicznych w badaniach ekonomicznych*, Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu, Wrocław.

Goldstein H., 1995, *Multilevel Statistical Models*, Edward Arnold, New York.

Hox J.J., 2010, *Multilevel Analysis: Technique and Applications*, Routledge, New York.

Lima-Neto E.A., de Carvalho F.A.T., 2008, *Centre and range method to fitting a linear regression model on symbolic interval data*, Computational Statistics and Data Analysis, vol. 52, no. 1, s. 1500–1515.

Lima-Neto E.A., de Carvalho F.A.T., 2010, *Constrained linear regression models for symbolic interval-valued variables*, Computational Statistics and Data Analysis, vol. 54, s. 333–347.

Muthén B., 1989, *Latent variable modeling in heterogeneous populations*, Psychometrika, vol. 54, no. 4, s. 557–585.

Muthén B., 1994, *Multilevel covariance structure analysis*, Sociological Methods & Research, vol. 22, no. 3, s. 376–398.

Noirhomme-Fraiture M., Brito P., 2011, *Far Beyond the Classical Data Models: Symbolic Data Analysis*, Statistical Analysis and Data Mining, vol. 4, no. 2, s. 157–170.

Pełka M., 2014, *Podejście wielomodelowe w regresji danych symbolicznych interwałowych*, Ekonometria 4 (46), s. 211–220.

Paterson L., Goldstein H., 1991, *New statistical methods for analysing social structures: An introduction to multilevel models*, British Educational Research Journal, vol. 17, no. 4, s. 387–393.

Sagan A. (red.), 2014, *Values and Preferences in Income Allocation of Polish Households*, Fundacja Uniwersytetu Ekonomicznego w Krakowie, Kraków.

Skrondal A., Rabe-Hesketh S., 2004, *Generalized Latent Variable Modeling: Multilevel, Longtidual and Structural Equation Models*, Chapman & Hall, Boca Raton.