PRACE NAUKOWE
Uniwersytetu Ekonomicznego we Wrocławiu

RESEARCH PAPERS
of Wrocław University of Economics

Nr 384

Taksonomia 24
Klasyfikacja i analiza danych – teoria i zastosowania

Redaktorzy naukowi
Krzysztof Jajuga
Marek Walesiak

Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2015

Redaktor Wydawnictwa: Aleksandra Śliwka

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 stronie internetowej Wydawnictwa
www.pracenaukowe.ue.wroc.pl
www.wydawnictwo.ue.wroc.pl

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

Wersja pierwotna: publikacja drukowana

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

Druk i oprawa: TOTEM

Spis treści

Summaries

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS nr 384 • 2015
Taksonomia 24 ISSN 1899-3192
Klasyfikacja i analiza danych – teoria i zastosowania e-ISSN 2392-0041

Justyna Brzezińska
Uniwersytet Ekonomiczny w Katowicach
e-mail: justyna.brzezinska@ue.katowice.pl

ANALIZA KLAS UKRYTYCH W BADANIACH SONDAŻOWYCH[1]

Streszczenie: W badaniach sondażowych, obok zmiennych obserwowalnych, bardzo często występują zmienne nieobserwowalne. Zmienne te są abstrakcyjną kategorią, która wykorzystywana jest w celu syntezy lub agregacji właściwości zawartych w zmiennych obserwowalnych. Wykorzystuje się ją w przypadku, gdy zarówno zmienne obserwowalne, jak i zmienne ukryte mają charakter dyskretny. Metoda ta oparta jest na dwóch założeniach. Pierwsze to warunek lokalnej niezależności, drugie natomiast mówi o tym, że populacja składa się z rozłącznych i wyczerpujących jednorodnych podpopulacji, które łącznie tworzą klasę ukrytą. Celem artykułu jest klasyfikacja respondentów biorących udział w badaniu sondażowym na rozłączane grupy oraz identyfikacja otrzymanych klas. W niniejszym badaniu analizę klas ukrytych wykorzystano do badania zachowań społecznych wśród studentów. Obliczenia przeprowadzone zostaną w programie **R** dzięki funkcji `poLCA {poLCA}`.

Słowa kluczowe: analiza klas ukrytych, analiza danych jakościowych, badania sondażowe.

DOI: 10.15611/pn.2015.384.04

1. Wstęp

Zmienne ukryte stanowią podstawę modeli ze zmiennymi ukrytymi, które składają się na szerzej rozumiane metody struktur ukrytych (*latent structure methods*). Podziału tych metod dokonuje się ze względu na charakter zmiennej obserwowalnej oraz zmiennej ukrytej. Gdy zarówno zmienna obserwowalna, jak i zmienna ukryta mają charakter dyskretny, metoda ta nazywana jest analizą klas ukrytych (*latent class analysis*) [Hagenaars 1990; 1993]. Gdy zmienna obserwowalna jest ciągła, a zmienna ukryta dyskretna, mamy do czynienia z analizą profili ukrytych (*latent profile analysis*). W przypadku gdy zmienna obserwowalna jest zmienną dyskretną, a zmienna ukryta zmienną ciągłą, metoda nosi wówczas nazwę analizy cech

[1] Projekt został sfinansowany ze środków Narodowego Centrum Nauki przyznanych na podstawie decyzji numer DEC-2012/05/N/HS4/00174.

ukrytych (*latent trait analysis*). W sytuacji gdy zarówno zmienna obserwowalna, jak i zmienna ukryta mają charakter ciągły, mamy wówczas do czynienia z analizą czynnikową (*factor analysis*).

Analiza klas ukrytych jest stosunkowo nową metodą analizy wielowymiarowej, która powstała w połowie XX wieku. Jako pierwszy Lazerfeld [1950a; 1950b] użył jej do budowy pewnej typologii segmentów na podstawie obserwowalnych zmiennych dychotomicznych. W miarę pogłębiania się stopnia zaawansowania prowadzonych badań metoda ta przekształciła się z metody analizy zmiennych dychotomicznych na analizę zmiennych zawierających większą liczbę kategorii. Goodman [1974] uczynił analizę klas ukrytych możliwą do praktycznego zastosowania przez zastosowanie metody największej wiarygodności do estymacji parametrów modelu. Ponadto zaproponował on zastosowanie tej metody do analizy zmiennych politomicznych (zmiennych o wielu kategoriach) oraz wielokrotnych zmiennych ukrytych. Haberman [1979] pokazał natomiast związek analizy klas ukrytych z analizą logarytmiczno-liniową dla tablic z brakującymi danymi. Metodę tą najczęściej wykorzystuje się w badaniach społecznych o charakterze sondażowym [Dayton 1998; Vermunt 2003; Colins, Lanza 2010]. W niniejszym artykule analiza klas ukrytych wykorzystana zostanie w badaniu społecznym do analizy etycznych zachowań społecznych wśród studentów amerykańskich. Obliczenia przeprowadzone zostaną w programie **R** dzięki funkcji `poLCA {poLCA}`.

2. Analiza klas ukrytych

Głównym celem analizy klas ukrytych jest redukcja liczby zmiennych przy jak najmniejszej utracie informacji o badanym zjawisku, a także odkrycie nieobserwowalnej heterogeniczności w populacji. Klasy ukryte pełnią wówczas funkcję nieobserwowalnych czynników, które wpływają na zależność pomiędzy jednostkami w danej klasie.

Analiza klas ukrytych oparta jest na dwóch założeniach. Pierwsze założenie mówi o tym, że populacja składa się z rozłącznych (*mutually exclusive*) i spójnych (*exhaustive*) jednorodnych podpopulacji, które łącznie tworzą klasę ukrytą. Oznacza to, że jeden obiekt może należeć tylko do jednej klasy ukrytej. Drugie założenie nazywane jest warunkiem lub aksjomatem lokalnej niezależności (*local independence assumption*), zgodnie z którym związek między zmiennymi obserwowalnymi zależy od relacji pomiędzy zmiennymi obserwowalnymi a zmiennymi ukrytymi. Oznacza to, że jeśli zmienna ukryta jest stała, zmienne obserwowalne powinny być statystycznie niezależne. Warunek ten spełniony jest we wszystkich rodzajach modeli struktur ukrytych.

Analiza klas ukrytych ma na celu znalezienie oraz zidentyfikowanie odpowiedniej liczby klas, w których zmienne obserwowalne są od siebie niezależne. Inaczej mówiąc, metoda ta umożliwia rozwarstwienie tablicy kontyngencji zawierającej zmienne

obserwowalne przez zmienną ukrytą, przy czym poszczególne klasy stanowią kategorie zmiennej ukrytej o charakterze dyskretnym. Model taki w efekcie przydziela obserwacje do klas ukrytych, a w dalszym etapie pozwala na przypuszczenie, jak zmienne obserwowalne zachowają się pod wpływem zmiennych ukrytych.

W modelach klas ukrytych wyróżnia się następujące rodzaje zmiennych, w zależności od rodzaju skali pomiaru:

- zmienne ukryte (*latent variables*), które mogą być mierzone na skalach nominalnych lub porządkowych,
- zmienne obserwowalne (*manifest vaviables, response variaables*) lub zmienne objaśniane (*dependent variables*), które mogą być mierzone na różnych skalach pomiaru,
- zmienne towarzyszące (*concomitant variables, covariates*) i zmienne objaśniające (*predictor variables*), które mogą być mierzone na różnych skalach pomiaru.

Model musi zawierać przynajmniej jedną zmienną ukrytą i jedną zmienną obserwowalną; może on także zawierać zmienne towarzyszące. Zmienna ukryta jest zatem statyczna i dzieli populację na podpopulacje, zwane klasami ukrytymi.

W analizie klas ukrytych modele różnią się jedynie liczbą klas ukrytych. Modele zawierające więcej parametrów (większą liczbę klas ukrytych) zapewniają lepsze dopasowanie do danych niż te, które opisane są przez mniejszą liczbę klas.

3. Analiza klas ukrytych dla tablic kontyngencji

Do modelowania klas ukrytych wykorzystywane są dwa podejścia: probabilistyczne oraz logarytmiczno-liniowe. Podejście pierwsze zaproponowane zostało przez Lazarsfelda [1950a; 1950b], a kontynuowane było przez Goodmana [1974]; model ze zmienną ukrytą przedstawiony jest w postaci prawdopodobieństwa.

W przypadku tablicy kontyngencji z pięcioma zmiennymi obserwowanymi: A ($h = 1, \ldots, H$), B ($j = 1, \ldots, J$), C ($k = 1, \ldots, K$), D ($l = 1, \ldots, L$), E ($m = 1, \ldots, M$), model ze zmienną ukrytą X o kategoriach T ($t = 1, \ldots, T$) jest przedstawiony jako prawdopodobieństwo przynależności do klasy ukrytej $X = t$:

$$\pi_{hjklm}^{ABCDE} = \sum_{t=1}^{T} \pi_{hjklmt}^{ABCDEX}, \tag{1}$$

przy czym:

$$\pi_{hjklmt}^{ABCDEX} = \pi_t^X \pi_{hjklmt}^{\bar{A}\bar{B}\bar{C}\bar{D}\bar{E}X} = \pi_t^X \pi_{ht}^{\bar{A}X} \pi_{jt}^{\bar{B}X} \pi_{kt}^{\bar{C}X} \pi_{lt}^{\bar{D}X} \pi_{mt}^{\bar{E}X}. \tag{2}$$

W równaniu (1) π_t^X oznacza prawdopodobieństwo tego, że obserwacja należy do klasy ukrytej $X = t$, co można zapisać symbolicznie jako $P(X = t)$. Symbolem $\pi_{hjklmt}^{\bar{A}\bar{B}\bar{C}\bar{D}\bar{E}X}$ oznaczone są prawdopodobieństwa warunkowe tego, że *hjklm*-ta katego-

ria zmiennej A,B,C,D,E znajdzie się w opisie klasy ukrytej *t*, natomiast symbolem $\pi_{ht}^{\overline{A}X}$ – prawdopodobieństwo warunkowe tego, że *h*-ta kategoria zmiennej *A* znajdzie się w opisie klasy ukrytej $X = t$, itd. Wszystkie elementy w równaniach (1) oraz (2) są prawdopodobieństwami, zatem ich wartości nie mogą być mniejsze od 0 ani przekraczać 1, a suma względem danego indeksu wszystkich prawdopodobieństw wynosi 1. Kategorie zmiennej ukrytej *X* są nieuporządkowane i traktowane jak kategorie zmiennej nominalnej.

Warunek (1) oznacza, że populacja może zostać podzielona na rozłączne i wyczerpujące się klasy ukryte, a każda obserwacja może należeć tylko do jednej klasy ukrytej. Równanie (2) nosi nazwę warunku lokalnej niezależności.

Prawdopodobieństwa warunkowe związane z każdą ze zmiennych wymagają spełnienia założenia:

$$\sum_{t=1}^{T}\pi_t^X = \sum_{h=1}^{H}\pi_{ht}^{\overline{A}X} = \sum_{j=1}^{J}\pi_{jt}^{\overline{B}X} = \sum_{k=1}^{K}\pi_{kt}^{\overline{C}X} = \sum_{l=1}^{L}\pi_{lt}^{\overline{D}X} = \sum_{m=1}^{M}\pi_{mt}^{\overline{E}X} = 1. \tag{3}$$

Warunek ten jest związany z faktem, iż populacja składa się z wyczerpujących podpopulacji, zatem suma wszystkich prawdopodobieństw w podpopulacjach sumuje się do 1.

Drugie podejście wykorzystywane w analizie klas ukrytych zaproponowane zostało przez Habermana [1974] i nazywane jest podejściem logarytmiczno-liniowym. Do modelu logarytmiczno-liniowego wprowadzona jest zmienna ukryta, która wpływa na wszystkie analizowane zmienne. Model taki w efekcie zawiera nie tylko efekty główne zmiennych obserwowalnych, ale także efekty zmiennej ukrytej oraz interakcje pomiędzy zmiennymi obserwowalnymi oraz zmienną ukrytą.

Multiplikatywna reprezentacja modelu logarytmiczno-liniowego dla pięciu zmiennych obserwowalnych *A*, *B*, *C*, *D*, *E* oraz jednej zmiennej ukrytej *X* przedstawiona jest w postaci równania multiplikatywnego:

$$m_{hjklmt}^{ABCDEX} = \eta \cdot \tau_h^A \cdot \tau_j^B \cdot \tau_k^C \cdot \tau_l^D \cdot \tau_m^E \cdot \tau_t^X \cdot \tau_{ht}^{AX} \cdot \tau_{jt}^{BX} \cdot \tau_{kt}^{CX} \cdot \tau_{lt}^{DX} \cdot \tau_{mt}^{EX} \tag{4}$$

lub w postaci addytywnej:

$$\ln(m_{hjklmt}^{ABCDEX}) = \lambda + \lambda_h^A + \lambda_j^B + \lambda_k^C + \lambda_l^D + \lambda_m^E + \lambda_t^X + \lambda_{it}^{AX} + \lambda_{jt}^{BX} + \lambda_{kt}^{CX} + \lambda_{lt}^{DX} + \lambda_{mt}^{EX} \tag{5}$$

Goodman wykazał, że w analizie klas ukrytych istnieje możliwość przejścia pomiędzy zapisem parametrów w podejściu probabilistycznym oraz log-liniowym [Haberman 1979]. Prawdopodobieństwo warunkowe tego, że *h*-ta kategoria zmiennej *A* znajdzie się w opisie klasy ukrytej $X = t$, definiowane jest jako:

$$\pi_{ht}^{\overline{A}X} = \frac{\exp\left(\lambda_h^A + \lambda_{ht}^{AX}\right)}{\sum_{h=1}^{H}\exp\left(\lambda_h^A + \lambda_{ht}^{AX}\right)}. \tag{6}$$

Z pomocą tej formuły można dokonać transformacji parametrów modelu z podejścia logarytmiczno-liniowego do podejścia probabilistycznego. Ze względu na prostotę modelu oraz łatwość interpretacji wyników najczęściej wykorzystywane jest probabilistyczne podejście Lazarsfelda.

Estymacja parametrów w analizie klas ukrytych polega na oszacowaniu liczby oraz wielkości poszczególnych klas ukrytych. Jako pierwszy element szacuje się prawdopodobieństwo przynależności do klasy $X = t$ (prawdopodobieństwo bezwarunkowe) (*latent class membership probability*). Prawdopodobieństwo to oznacza odsetek populacji należący do danej klasy ukrytej. W następnej kolejności szacowane są prawdopodobieństwa wystąpienia danej kategorii zmiennej, pod warunkiem przynależności do klasy ukrytej $X = t$ (prawdopodobieństwa warunkowe) (*conditional response probabilities*). Prawdopodobieństwa te stanowią tym samym podstawę opisu danej klasy ukrytej.

W celu wyboru optymalnej liczby klas analiza klas ukrytych wykorzystuje współczynnik chi-kwadrat (χ^2) oraz kryteria informacyjne *AIC* [Akaike 1973] oraz *BIC* [Schwartz 1978].

4. Zastosowanie analizy klas ukrytych w badaniach sondażowych

Analiza klas ukrytych wykorzystana została w badaniu sondażowym przeprowadzonym wśród studentów uczelni wyższej [Dayton 1998]. Zbiór `cheating` danych dostępny jest w pakiecie `poLCA` programu **R**. Studenci odpowiadali na pytania TAK lub NIE względem następujących obserwowalnych zmiennych dychotomicznych:

- X_1 – czy kłamali, by uniknąć egzaminu,
- X_2 – czy kłamali by uniknąć pisania testu,
- X_3 – czy zakupili lub pozyskali test przed egzaminem,
- X_4 – czy kopiowali odpowiedzi od studentów siedzących obok nich.

W kwestionariuszu uwzględniono także politomiczną zmienną towarzyszącą:

- C – średnia ocen ze studiów (do 2,99; 3,00-3,25; 3,26-3,50; 3,51-3,75; 3,76-4,00).

W badaniu udział wzięło 319 respondentów, przy czym średniej C nie uwzględniono dla czterech studentów, którzy zapewnili, że nigdy nie oszukiwali. Procedura estymacji każdego modelu wymaga określenia *ex ante* liczby klas. W pierwszej części zbudowano modele z jedną, dwiema, trzema i czterema klasami ukrytymi. Modele, w których uwzględniono zmienne obserwowalne, oceniono za pomocą odpowiednich współczynników oraz kryteriów informacyjnych. Wyniki przedstawiono w tab. 1.

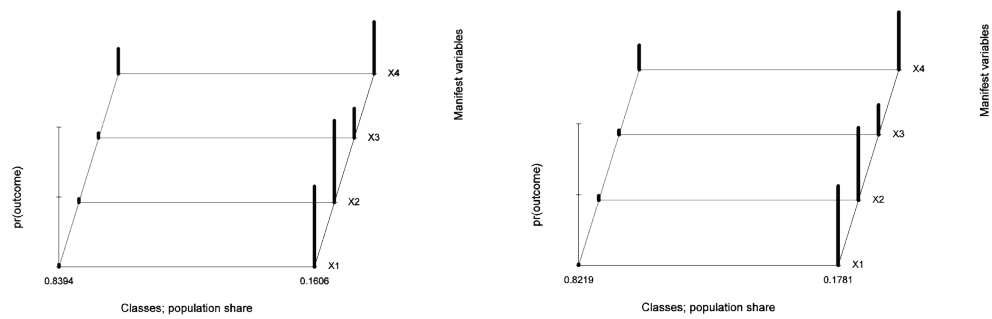
Modele maksymalizujące funkcję wiarygodności (*logL*) cechują się lepszym dopasowaniem modelu do danych. Jeśli chodzi o kryteria informacyjne, ich mniejsza wartość wskazuje na model lepiej dopasowany do danych. W analizie klas

Tabela 1. Wartości funkcji wiarygodności oraz kryteriów informacyjnych dla modelu bez zmiennej towarzyszącej

Liczba klas T	Log$L$	χ^2	AIC	BIC
$t = 1$	-467,438	136,342	942,876	957,937
$t = 2$	-440,027	8,323	898,054	931,941
$t = 3$	-438,209	4,276	904,417	957,130
$t = 4$	-436,145	0,002	910,291	981,829

Źródło: opracowanie własne w programie R.

ukrytych najczęściej wykorzystywanym kryterium jest kryterium Bayesa (*BIC*). Z analizy tab. 1 wynika, że kryteria informacyjne osiągają najmniejszą wartość dla dwóch klas (*AIC* = 898, 054, *BIC* = 931, 941). Estymatory prawdopodobieństw przynależności do każdej z dwóch klas wynoszą odpowiednio 0,8307 dla klasy pierwszej i 0,1606 dla klasy drugiej. Wartości te można przedstawić postaci słupków, których wysokości odpowiadają prawdopodobieństwom wyboru danej odpowiedzi (rys. 1a).



a) model bez zmiennej towarzyszącej b) model zawierający zmienną towarzyszącą

Rys. 1. Wyniki estymacji modelu klas ukrytych

Źródło: opracowanie własne.

Na wykresie na rys. 1 widoczna jest także informacja o wielkości poszczególnych klas. Widać, że klasy 1 i 2 są wyraźnie odrębne, gdyż wysokości słupków w każdej z wyróżnionych klas są do siebie zbliżone.

Klasę pierwszą charakteryzują studenci, którzy kłamali, by uniknąć egzaminu (0,9834), kłamali, by uniknąć pisania testu (0,9708), zapłacili za zdobycie testu przed egzaminem (0,9629) oraz kopiowali odpowiedzi od kolegów (0,8181). Druga klasa obejmuje osoby, które nie kłamały, by uniknąć egzaminu (0,5769), nie kłamały, by uniknąć pisania testu (0,5891), podobnie jak w klasie pierwszej zapłaciły za zdobycie testu przed egzaminem (0,7840) i również kopiowały odpowiedzi od

kolegów (0,6236). Analiza klas ukrytych pozwoliła zatem na klasyfikację studentów na oszukujących, należących do klasy pierwszej z prawdopodobieństwem 0,8307 oraz na niekłamiących, by uniknąć pisania testu, lecz kopiujących odpowiedzi, by zdać go lepiej, należących do klasy drugiej z prawdopodobieństwem 0,1606.

Analiza klas ukrytych pozwala także na uwzględnienie w modelu zmiennej towarzyszącej. W niniejszym badaniu zbudowano model, w którym zmiennymi obserwowalnymi są cztery zmienne: X_1, X_2, X_3, X_4, oraz dodatkowa – zmienna towarzysząca *C*. Podobnie jak w poprzedniej części analizy, gdzie nie uwzględniono zmiennej towarzyszącej, zbudowano model z dwiema, trzema i czterema klasami ukrytymi, które oceniono za pomocą kryteriów informacyjnych (tab. 2).

Tabela 2. Wartości funkcji wiarygodności oraz kryteriów informacyjnych dla modelu zawierającego zmienną towarzyszącą

Liczba klas *T*	Log*L*	χ^2	*AIC*	*BIC*
t = 2	429,368	8,642	879,277	916,803
t = 3	nie znaleziono	5,379	876,369	936,409
t = 4	-443,349	22,453	930,697	1013,254

Źródło: opracowanie własne w programie R.

Najmniejszą wartość kryterium informacyjne *BIC* = 916, 803 osiąga w przypadku $t = 2$ klas ukrytych. Estymatory prawdopodobieństw przynależności do każdej z dwóch klas wynoszą odpowiednio 0,8508 dla klasy pierwszej i 0,1492 dla klasy drugiej. Wartości te można przedstawić na wykresie (rys. 1b). Interpretacja oraz wyniki są bardzo zbliżone do tych, które uzyskano w analizie klas ukrytych bez uwzględnienia zmiennej towarzyszącej.

Oszacowane parametry modelu zawierają dodatkową informację na temat zmiennej towarzyszącej C. Model bez zmiennej towarzyszącej w porównaniu z modelem zawierającym taką zmienną jest nieznacznie lepiej dopasowany do danych. Wobec tego charakterystyka klas ukrytych w przypadku uwzględnienia zmiennej towarzyszącej C jest bardzo zbliżona do tego, którą uzyskano w analizie bez tej zmiennej. Klasę pierwszą tworzą studenci, którzy kłamali, by uniknąć egzaminu (0,9903), kłamali, by uniknąć pisania testu (0,9647), zapłacili za zdobycie testu przed egzaminem (0,9655) oraz kopiowali odpowiedzi od kolegów (0,8257). Druga klasa to osoby, które nie kłamały, by uniknąć egzaminu (0,5611), nie kłamały, by uniknąć pisania testu (0,5142), podobnie jak w klasie pierwszej zapłaciły za zdobycie testu przed egzaminem (0,7850) i również kopiowały odpowiedzi od kolegów (0,5925). W wyniku przeprowadzenia analizy klas ukrytych w badaniu sondażowym uzyskano dwie rozłączne klasy studentów charakteryzujących się odmiennymi zachowaniami etycznymi podczas zdawania egzaminu.

5. Zakończenie

Analiza danych niemetrycznych w wielowymiarowych tablicach kontyngencji jest jedną z częściej wykorzystywanych metod w badaniach ekonomicznych, medycznych oraz psychologicznych. Modele klas ukrytych wykorzystywane są wówczas, gdy badane zmienne są bezpośrednio nieobserwowalne i mają charakter zmiennych skokowych. Analiza ma na celu znalezienie oraz zidentyfikowanie odpowiedniej liczby klas ukrytych, w których zmienne obserwowalne są od siebie niezależne.

W niniejszym artykule zaprezentowano zastosowanie analizy klas ukrytych w badaniach sondażowych. Przeprowadzono analizę klas ukrytych zarówno bez uwzględnienia, jak i z uwzględnieniem zmiennej towarzyszącej. Wyniki estymacji porównano za pomocą odpowiednich współczynników pozwalających na ocenę modelu do danych. W obu przypadkach, zarówno z uwzględnieniem, jak i bez uwzględnienia zmiennej towarzyszącej, najlepszą liczbą klas okazała się 2. Dzięki zastosowanej metodzie możliwe stało się opisanie tych klas i porównanie otrzymanych wyników.

Literatura

Akaike H. (1973), *Information theory and an extension of the maximum likelihood principle*, *Proceedings of the 2nd International Symposium on Information*, Akademiai Kiado, Budapest.

Collins L.M., Lanza S.T. (2010), *Latent Class and Latent Transition Analysis: With Applications in the Social*, Behavioral, and Health Sciences, Wiley.

Dayton C.M. (1998), *Latent Class Scaling Analysis*, Sage Publications, Thousand Oaks, CA.

Goodman L.A. (1974), *Exploratory latent structure analysis using both identifiable and unidentifiable models*, Biometrica, 61, 215-231.

Haberman S.J. (1974), *The ANALYSIS of Frequency Data*, University of Chicago Press, Chicago.

Haberman S.J. (1979), *Analysis of Qualitative Data. Vol 2. New Developments*, Academic Press, New York.

Hagenaars J.A. (1990), *Categorical Longitudinal Data: Loglinear Panel, Trend, and Cohort Analysis*, CA: Sage, Newbury Park.

Hagenaars J.A. (1993), *Loglinear models with latent variables*, *Sage University Paper series on Qualitative Applications in the Social Sciences, 07-094*, Newbury Park, CA.

Lazarsfeld P.F. (1950a), The Interpretation and Mathematical Foundation of Latent Class Structure Analysis, [w:] Souffer S. (ed.) *Measurement and Prediction*, Princeton University Press, Princeton, NJ.

Lazarsfeld P.F. (1950b), *The Logical and Mathematical Foundation of Latent Structure Analysis*, [w:] Souffer S. (ed.), *Measurement and Prediction*, Princeton University Press, Princeton, NJ.

Schwartz G. (1978), *Estimating the dimensions of a model*, Annals of Statistics **6**, 461-464.

Vermunt J.K. (2003), *Applications of Latent Class Analysis in Social Science Research, Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, Springer, Berlin Heidelberg, 22-36.

LATENT CLASS ANALYSIS IN SURVEY RESEARCH

Summary: In survey research, in addition to manifest variables, we deal with latent variables. They are an abstract category usually used for the aggregation of properties contained in manifest variables. Latent class analysis allows to analyze discrete, as well continuous variables. We can also analyze non response datasets. Latent class analysis is based on two assumption. The first assumption is the condition of local independence, the second one is that the population is divided into homogeneous and exhaustive subpopulations which together form a latent class. Latent class analysis is in a sample survey, which goal is usually market segmentation. The main goal of this paper is to present latent class analysis in the analysis of social behavior. All calculations will be conducted in R software with the use of `poLCA{poLCA}` function.

Keywords: latent class analysis, categorical data analysis, sample survey.