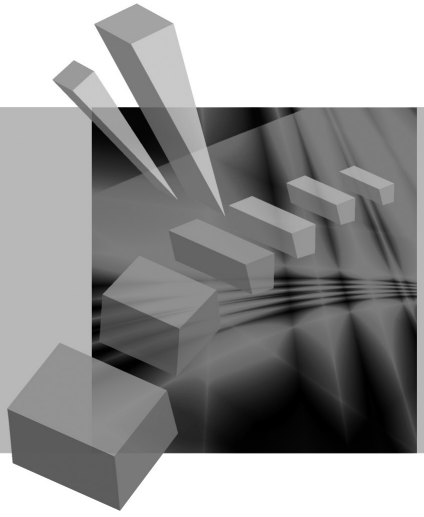
PRACE NAUKOWE
Uniwersytetu Ekonomicznego we Wrocławiu
RESEARCH PAPERS
of Wrocław University of Economics

278

Taksonomia 20

Klasyfikacja i analiza danych – teoria i zastosowania



Redaktorzy naukowi
Krzysztof Jajuga
Marek Walesiak

Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2013

Redaktor Wydawnictwa: Aleksandra Śliwka

Redaktor techniczny: Barbara Łopusiewicz

Korektor: Barbara Cibis

Łamanie: Małgorzata Czupryńska

Projekt okładki: Beata Dębska

Publikacja jest dostępna w Internecie na stronach:
www.ibuk.pl, www.ebscohost.com,
The Central and Eastern European Online Library www.ceeol.com,
a także w adnotowanej bibliografii zagadnień ekonomicznych BazEkon
http://kangur.uek.krakow.pl/bazy_ae/bazekon/nowy/index.php

Informacje o naborze artykułów i zasadach recenzowania znajdują się
na stronie internetowej Wydawnictwa
www.wydawnictwo.ue.wroc.pl

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

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

Wersja pierwotna: publikacja drukowana

Druk: Drukarnia TOTEM

Spis treści

Summaries

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS nr 278 • 2013

Taksonomia 20. Klasyfikacja i analiza danych – teoria i zastosowania ISSN 1899-3192

Ewa Genge
Uniwersytet Ekonomiczny w Katowicach

SEGMENTACJA UCZESTNIKÓW INDUSTRIADY Z WYKORZYSTANIEM ANALIZY KLAS UKRYTYCH

Streszczenie: Modele mieszanek, których składowe charakteryzowane są przez rozkłady prawdopodobieństw (tzw. rozkłady składowe mieszanki), reprezentują tzw. podejście modelowe w taksonomii. W ostatnim czasie na popularności coraz bardziej zyskują modele mieszanek rozkładów dla zmiennych jakościowych (mierzonych na skalach słabych), zwane również modelami lub analizą klas ukrytych (*latent class analysis*). Celem artykułu będzie segmentacja uczestników Industriady, tj. imprez organizowanych przy obiektach leżących na Szlaku Zabytków Techniki Województwa Śląskiego. Obliczenia zostaną przeprowadzone za pomocą pakietu `poLCA` programu R.

Słowa kluczowe: analiza klas ukrytych, model mieszanek, dane jakościowe.

1. Wstęp

Modele mieszanek, których składowe charakteryzowane są przez rozkłady prawdopodobieństw (tzw. rozkłady składowe mieszanki) od dawna znajdują swoje zastosowanie w taksonomii. W ostatnim czasie na popularności coraz bardziej zyskują modele mieszanek rozkładów dla zmiennych jakościowych (mierzonych na skalach słabych), zwane również modelami lub analizą klas ukrytych (*latent class analysis*). W modelach tych liczba rozkładów składowych jest nieznana (zmienna ukryta). Modele klas ukrytych reprezentują tzw. podejście modelowe, w którym podstawą klasyfikacji obserwacji do klas są oszacowane na podstawie modelu prawdopodobieństwa przynależności. Parametry modelu szacowane są metodą największej wiarygodności z wykorzystaniem algorytmów, tj. EM (*Expectation-Maximization*) czy algorytm Newtona-Raphsona. Celem referatu będzie segmentacja uczestników Industriady, tj. wyodrębnienie klas o podobnych wzorcach zachowań i postaw dla śląskich respondentów, a także dokonanie oceny wpływu zmiennych demograficznych na ich przynależność do klas. Osiągnięte wyniki mogą stanowić przesłankę przy podejmowaniu decyzji co do zasadności organizacji kolejnych imprez tego typu (tj. imprez organizowanych przy obiektach leżących na Szlaku Zabytków Techniki Województwa Śląskiego) oraz umiejętnym przeprowadzeniu akcji promocyjnej.

Obliczenia zostaną przeprowadzone za pomocą pakietów: `poLCA, flexmix`, programu **R**.

2. Model klas ukrytych ze zmiennymi towarzyszącymi – definicja

W podejściu modelowym w taksonomii, w odróżnieniu od klasycznych metod analizy skupień wykorzystujących miary odległości jako podstawę klasyfikacji obiektów, szacuje się parametry modelu i oblicza prawdopodobieństwa przynależności obiektów do klas. Na podstawie wartości tych prawdopodobieństw klasyfikuje się obiekty. Rozważa się zatem zbiór *n* obiektów, charakteryzowanych za pomocą zmiennych dychotomicznych lub politomicznych, zwanych zmiennymi obserwowanymi (*manifest variables*) (zob. [Bąk 2011, s. 204-222]) o wielu kategoriach $l_1, \ldots, l_m$. Zbiór wszystkich obiektów można więc zapisać za pomocą wektora $\mathbf{x}_i = (x_{ijh}; j = 1, \ldots, m;\ h = 1, \ldots, l_j; i = 1, \ldots, n)$, gdzie $x_{ijh} = 1$ oznacza *i*-tą obserwację na *j*-tej zmiennej o *h*-tej kategorii. Jeżeli liczba wszystkich kategorii jest równa $l = \sum_{j=1}^{m} l_j$, wtedy zbiór określany jest za pomocą macierzy o wymiarach $n \times m$.

Model klas ukrytych, oprócz zmiennych obserwowanych, może zawierać jeszcze tzw. zmienne towarzyszące (*covariates* lub *concomitant variables*), mające wpływ na przynależność obiektów do klas (wpływ na prawdopodobieństwa *a priori*) (zob. np. [Dayton, Macready 1988, s. 173-178; Hagenaars, McCutcheon 2002]). Zmienne towarzyszące wraz ze zmiennymi $X_1, \ldots X_m$ biorą udział w szacowaniu parametrów modelu klas ukrytych, na podstawie którego można będzie dokonać klasyfikacji nowych obiektów bez udziału zmiennych obserwowanych. Zmienne towarzyszące wykorzystywane są często w badaniach marketingowych, ekonomicznych, psychologicznych, w których pozyskanie zmiennych obserwowanych jest bardzo kosztowne (por. [Witek 2011a, s. 223-241]).

Najczęściej parametry zmiennych towarzyszących szacowane są wraz z pozostałymi parametrami modelu klas ukrytych (jednocześnie). Ten sposób estymacji zwany jest jednokrokową techniką estymacji parametrów zmiennych towarzyszących (*one-step technique for estimating the effects of covariates*) (zob. np. [Dayton, Macready 1988, s. 173-178; Hagenaars, McCutcheon 2002]). Włączając do modelu klas ukrytych zmienne towarzyszące, zakładamy, że mają one wpływ na prawdopodobieństwa *a priori*. W klasycznym modelu klas ukrytych (bez zmiennych towarzyszących) zakładamy, że każda obserwacja ma takie samo prawdopodobieństwo przynależności do klasy ukrytej.

Model klas ukrytych dla danych jakościowych można zapisać jako mieszankę rozkładów wielomianowych, w której zakłada się, że każda obserwacja $\mathbf{x}_i$ pochodzi z mieszanki wielowymiarowych rozkładów wielomianowych (*mixture of multivariate multinominal distributions*) określonej jako:

$$f(\mathbf{x}_i, \mathbf{z}_i \mid \boldsymbol{\Theta}) = \sum_{s=1}^{u} \tau_s(\mathbf{z}_i, \boldsymbol{\alpha}) f_s(\mathbf{x}_i \mid \boldsymbol{\Theta}_s), \tag{1}$$

gdzie: f_s – funkcja gęstości ukrytej klasy P_s (*s*-tego rozkładu składowego mieszanki),

$\mathbf{x}_i$ – wektor realizacji zmiennych obserwowanych $\mathbf{x}_i = [x_{i1,\ldots,} x_{im_1}]$,

$\mathbf{z}_i$ – wektor realizacji zmiennych towarzyszących, $\mathbf{z}_i = [z_{i1,\ldots,} z_{im_2}]$,

Θ_s – wektor parametrów ukrytej klasy P_s,

Θ – wektor wszystkich parametrów mieszanki rozkładów, $\boldsymbol{\Theta} = (\tau_s, \boldsymbol{\Theta}_s)$

τ_s – prawdopodobieństwo *a priori* – wartość prawdopodobieństwa, że dana obserwacja należy do klasy

$$(\tau_s(\mathbf{z}_i, \boldsymbol{\alpha}) \geq 0 \wedge \sum_{s=1}^{u} \tau_s(\mathbf{z}_i, \boldsymbol{\alpha}) = 1), \boldsymbol{\Theta}_\mathbf{s} \neq \boldsymbol{\Theta}_l \forall s \neq l.$$

Wpływ zmiennych towarzyszących na prawdopodobieństwa *a priori* wyrażany jest za pomocą wielomianowej funkcji logitowej [Agresti 2002].

3. Estymacja parametrów oraz wybór liczby klas ukrytych

Popularną metodą szacowania parametrów największej wiarygodności jest algorytm EM [Dempster i in. 1977, s. 1-38]. W pakiecie `poLCA` wykorzystywana jest zmodyfikowana wersja algorytmu EM (zob. [Bandeen-Roche i in. 1997, s. 123-135]). Jedną z głównych zalet modeli klas ukrytych jest to, że w odróżnieniu od popularnych metod taksonomicznych (tj. *k*-średnich, metoda Warda), istnieje kilka statystycznych miar służących wyborowi i ocenie ich jakości dopasowania. Najczęściej w różnego rodzaju badaniach empirycznych na początku sprawdza się dopasowanie dla $s = 1$. W kolejnych krokach zwiększa się liczbę klas o jeden tak długo, aż model osiągnie najlepsze dopasowanie. Należy jednak pamiętać, że wraz z dodatkową liczbą klas liczba szacowanych parametrów wzrasta o $1 + \sum_j (l_j - 1)$. Dlatego najczęściej wykorzystywane są kryteria informacyjne, będące wyrazem kompromisu pomiędzy jakością dopasowania a złożonością modelu. Do najbardziej popularnych kryteriów informacyjnych zaliczane są: Bayesowskie kryterium informacyjne Schwarza BIC (*Bayesian Information Criterion*) [Schwarz 1978], kryterium informacyjne Akaike AIC (*Akaike Information Criterion*) [Akaike 1974].

4. Analiza empiryczna

W analizie wykorzystano dane z badania ankietowego, przeprowadzonego przez Katedrę Marketingu UE w Katowicach, dotyczące różnego nastawienia do Święta

Szlaku Zabytków Techniki Województwa Śląskiego, czyli Industriady[1]. W badaniu zgromadzono 552 ankiety. W przykładzie wykorzystano następujące pytania:

1. X_1 (Pyt. 1): Czy w ostatnim czasie zetknął(a) się Pan/-i z jakimikolwiek reklamami zabytków techniki województwa śląskiego?

2. X_2 (Pyt. 2): Czy w obiekcie, w którym się znajdujemy, jest Pan/-i po raz pierwszy w życiu?

3. X_3 (Pyt. 3): Czy w ostatnich 12 miesiącach odwiedził/a Pan/i jakieś inne zabytki techniki znajdujące się w województwie śląskim?

4. X_4 (Pyt. 4): Czy Pan/-i zdaniem zabytki techniki województwa śląskiego są tym, co ten region wyróżnia pozytywnie w porównaniu do innych regionów Polski[2]?

5. X_5 (Pyt. 5): Czy uważa Pan/-i, że należy kontynuować coroczną organizację Święta Szlaku Zabytków Techniki, polegającą na przygotowaniu w jedną sobotę czerwca wielu różnego rodzaju imprez odbywających się w tych zabytkach jednocześnie[3].

6. X_6 (Pyt. 6): Jak podoba się Panu/-i impreza, w której aktualnie uczestniczymy[4]?

7. X_7 (Pyt. 7): Z kim przybył(-a) Pan(-i) na dzisiejszą imprezę[5]?

W badaniu uwzględniono również następujące zmienne towarzyszące:

a) Z_1 : płeć respondenta,

b) Z_2 : wiek – mniej niż 18, 18-25 lat, 26-40 lat, 41-60 lat, więcej niż 60,

c) Z_3 : wykształcenie – podstawowe, zawodowe, średnie, wyższe,

d) Z_4 : obiekt – 14 obiektów znajdujących się na szlaku zabytków (np. EC Szombierki, Radiostacja Gliwicka, Kopalnia Guido w Zabrzu, ZK Ignacy w Rybniku).

Aby wybrać optymalną liczbę klas ukrytych (ukrytą liczbę składowych modelu), obliczono wartości kryteriów informacyjnych AIC oraz BIC dla liczby klas $s = 1,\ldots,u$ dla tzw. modelu podstawowego, tj. bez udziału zmiennych towarzyszących (*base model*) (zob. np. [Collins, Lanza 2011]). Kryterium BIC jako optymalną wskazało liczbę klas równą 2, AIC zaś liczbę klas równą 3. Kryteria te nie zawsze dają wyniki jednoznaczne. W licznych pracach (zob. np. [Biernacki i in. 1999; Witek 2011b]) kryterium BIC w porównaniu do innych kryteriów informacyjnych

---

[1] W ramach tej imprezy organizowane są specjalne iluminacje, mappingi, pokazy laserowe, koncerty, spektakle (np. podziemny happening z możliwością kąpieli w basenie w kopalni Guido), wystawy, warsztaty i konkursy. Wstęp na wszystkie imprezy jest bezpłatny. Celem tego śląskiego święta jest zapoczątkowanie Europejskiej Nocy Dziedzictwa Industrialnego (święta zabytków techniki odbywającego się tego samego dnia w Zagłębiu Ruhry oraz na Ukrainie, w Doniecku).

[2] Możliwe odpowiedzi na pyt. 1-4 to: 1 – nie, 2 – tak.

[3] Możliwe odpowiedzi na pyt. 5 to: 1 – nie należy kontynuować, 2 – należy kontynuować.

[4] Możliwe odpowiedzi na pyt. 6 to: 1 – bardzo mi się nie podoba, 2 – nie podoba mi się, 3 – ani mi się podoba, ani mi się nie podoba, 4 – podoba mi się, 5 – bardzo mi się podoba.

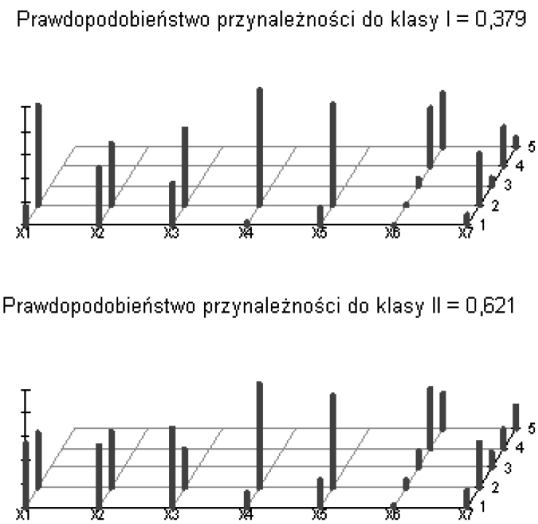
[5] Możliwe odpowiedzi na pyt. 7 to: 1 – sam, 2 – z osobą towarzyszącą, 3 – z dzieckiem (z dziećmi), 4 – z całą rodziną, 5 – z grupą znajomych.

dało bardzo dobre wyniki. Ponadto często w takich sytuacjach wybierane są modele mniej złożone (zob. np. [Collins, Lanza 2011]). W dalszej części pracy za stosowne uznano więc przyjęcie liczby klas równej dwa.

Następnie oszacowano kilka modeli klas ukrytych, różniących się zbiorem zmiennych towarzyszących (np. $Z_1 + Z_2$, $Z_1 + Z_3$, $Z_1 + Z_2 + Z_3 + Z_4$). Rozważano również interakcje pomiędzy zmiennymi towarzyszącymi (np. $Z_1 \times Z_2$, $Z_1 \times Z_2 \times Z_3$, $Z_1 \times Z_2 \times Z_3 \times Z_4$), ale żadna z nich nie okazała się istotna.

Na podstawie uzyskanych wyników (analiza kryteriów informacyjnych oraz badanie istotności parametrów za pomocą testu *t*-Studenta) przyjęto ostateczny podział badanej próby respondentów na dwie klasy z wykorzystaniem dwóch zmiennych towarzyszących, tj. wiek i wykształcenie.

Na rysunku 1 przedstawiono prawdopodobieństwa wyboru każdej z kategorii zmiennych obserwowanych (stosowna wysokość słupków) dla obu klas. Widoczne są także prawdopodobieństwa *a priori* (wagi) dla poszczególnych klas.



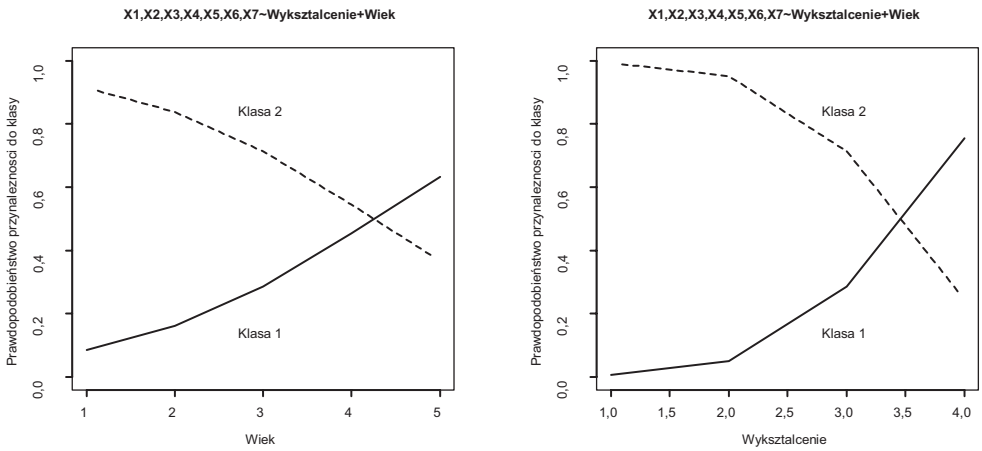
Rys. 1. Wyniki estymacji modelu klas ukrytych dla zmiennych obserwowanych

Źródło: obliczenia własne.

Na podstawie rys. 1 oraz badań empirycznych można zauważyć, że w klasie pierwszej, mniej licznej ($\tau_1 = 0{,}38$), aż 85% ankietowanych spotkało się z jakąkolwiek reklamą zabytków techniki województwa śląskiego. Ponad 52% badanych było w danym obiekcie po raz pierwszy. Aż 64% respondentów klasy pierwszej odwiedziło w ostatnim czasie jakikolwiek inny zabytek techniki województwa śląskiego. Prawie wszyscy ankietowani w tej klasie (98%) zgadzają się z opinią, że zabytki techniki województwa śląskiego są tym, co ten region wyróżnia pozytywnie w po-

równaniu do innych regionów Polski. 86% respondentów jest przekonanych, że należy kontynuować coroczną organizację Święta Szlaku Zabytków Techniki, polegającą na przygotowaniu w jedną sobotę czerwca wielu różnego rodzaju imprez odbywających się w tych zabytkach jednocześnie. W klasie pierwszej najwięcej, bo aż prawie połowa, respondentów było zadowolonych z imprezy, w której aktualnie uczestniczyli (49% osób odpowiedziało, że impreza im się podoba). Niewiele mniej, bo 45%, ankietowanych było bardzo zadowolonych z organizowanej imprezy. W klasie tej nie było żadnych niezadowolonych uczestników. W głównej mierze uczestnikami były osoby, które przyszły z osobą towarzyszącą (44%) czy też z całą rodziną (34%).

Klasa druga jest klasą liczniejszą – należy do niej 62% ankietowanych. W klasie tej więcej niż połowa (54%) respondentów nie spotkała się z żadną reklamą, natomiast mniej niż połowa (46%) uczestników imprezy spotkała się przynajmniej z jednym z nośników reklamy zabytków techniki województwa śląskiego, a 48% znalazło się w odwiedzanym obiekcie po raz pierwszy. W grupie tej przeważają osoby (67%), które w ostatnim roku nie odwiedziły żadnego innego zabytku techniki. 12% osób w tej klasie uważa, że zabytki z pewnością nie wyróżniają Śląska, zaś o 10% mniej respondentów (w porównaniu do klasy pierwszej) twierdzi, że zabytki pozytywnie wpływają na wizerunek Śląska. 22% osób klasy drugiej nie widzi sensu kontynuacji imprezy. Łącznie 20% osób nie była zadowolonych z imprezy lub nie miało zdania na jej temat. Zdecydowanie mniej (o 16% mniej w porównaniu do klasy pierwszej) osób w tej klasie odpowiedziało, że „impreza bardzo mi się podoba". W grupie tej również największą część stanowią uczestnicy, którzy zabrali ze sobą osobę towarzyszącą. Jednak w odróżnieniu od klasy pierwszej znajduje się tu spora grupa osób, która przyszła ze znajomymi (20%), jak również osoby, które przyszły same (13%), z dzieckiem lub dziećmi (13%).



Rys. 2. Prawdopodobieństwa przynależności respondentów do klas dla zmiennych towarzyszących

Źródło: obliczenia własne.

W kolejnej części pracy dokonano analizy wpływu zmiennych towarzyszących na przynależność analizowanych obiektów do klas. Jeżeli chodzi o zmienną „wiek", okazuje się, że z biegiem lat (im wyższa kategoria wiekowa) wzrasta prawdopodobieństwo przynależności do klasy pierwszej, a spada do drugiej. Jeżeli chodzi o drugą ze zmiennych towarzyszących, prawdopodobieństwa przynależności do klas kształtują się bardzo podobnie, tj. im wyższe wykształcenie, tym prawdopodobieństwo przynależności do klasy pierwszej wzrasta, a do klasy drugiej spada. Ilustrację graficzną prawdopodobieństw przynależności respondentów do klas dla obu zmiennych towarzyszących pokazano na rys. 2[6].

5. Podsumowanie

W artykule przedstawiono przykład zastosowania modeli klas ukrytych do oceny zadowolenia uczestników Święta Szlaku Zabytków Techniki. Analiza klas ukrytych umożliwiła segmentację respondentów na podstawie odpowiedzi udzielonych w badaniu przeprowadzonym przez Katedrę Marketingu UE Katowice. Wyodrębniono dwie klasy o podobnych wzorcach zachowań i postaw dla śląskich respondentów. Dokonano również oceny wpływu zmiennych demograficznych na ich przynależność do klas.

Do klasy pierwszej zaliczono przede wszystkim uczestników, którzy spotkali się z reklamą, świadomie zaplanowali swój czas, są zainteresowani zabytkami techniki (zwiedzali również inne obiekty na szlaku Industriady). Respondenci klasy drugiej Industriady są nastawieni nieco bardziej sceptycznie, być może przypadkowo znalazły się na imprezie. Są to głównie osoby młode, które najprawdopodobniej z ciekawości przyszły na ze swymi znajomymi. Nie są zainteresowane innymi zabytkami techniki, w związku z tym nie wszyscy respondenci tej klasy uważają, że kolejne edycje Industriady to dobry pomysł.

Literatura

Agresti A., *Categorical Data Analysis*, John Wiley&Sons, Hoboken 2002.

Akaike H., *A new look at statistical model identification*, "IEEE Transactionson Automatic Control" 1974, 19, s. 716-723.

Bandeen-Roche K., Miglioretti D.L., Zeger S.L., Rathouz P.J., *Latent variable regression for multiple discrete outcomes*, "Journal of the American Statistical Association" 1997, 92(40), s. 123-135.

Bąk A., *Modele klas ukrytych dla danych jakościowych*, [w:] E. Gatnar, M. Walesiak, *Analiza danych jakościowych i symbolicznych z wykorzystaniem programu R*, C.H. Beck, Warszawa 2011, s. 204-222.

[6] Dla zmiennej towarzyszącej „wykształcenie" sporządzono wykres, przyjmując, że zmienna jakościowa „wiek" jest równa kategorii występującej najczęściej. W podobny sposób sporządzono wykres i dokonano interpretacji dla zmiennej towarzyszącej „wiek" (zob. np. [Linzer i Lewis 2011; Witek 2011a]).

Biernacki C., Celeux G., Govaert G.: *Choosing models in model-based clustering and discriminant analysis*, "Journal of Statistical Computation and Simulation" 1999, 64, 49-71.

Collins L.M., Lanza S.T., *Latent Class and Latent Transition Analysis with Applications in the Social, Behavioral, and Health Sciences*, John Wiley&Sons, Wiley 2011.

Dayton C. M., Macready G.B., *Concomitant-variable latent-class models*, "Journal of the American Statistical Association" 1988, 83(401), s. 173-178.

Dempster A.P., Laird N.P., Rubin D.B., *Maximum likelihood for incomplete data via the EM algorithm (with discussion)*, "Journal of the Royal Statistical Society" 1977, 39, s. 1-38.

Hagenaars A.J., McCutcheon A.L., *Applied Latent Class Analysis*, Cambridge University Press, Cambridge 2002.

Linzer D., Lewis J., *poLCA: an R package for polytomous variable latent class analysis*, "Journal of Statistical Software" 2011, 42(10), s. 1-29.

Schwarz G., *Estimating the dimension of a model*, "The Annals of Statistics" 1978, 6, s. 461-464.

Witek E., *Modele mieszanek dla danych jakościowych*, [w:] E. Gatnar, M. Walesiak, *Analiza danych jakościowych i symbolicznych z wykorzystaniem programu R*, C.H. Beck, Warszawa 2011a, s. 223-241.

Witek E., *The Comparison of Model-Based Clustering with Heuristic Clustering Methods*, [w:] Cz. Domański, J. Białek, *Folia Oeconomica 255, Methodological Aspects of Multivariate Statistical Analysis, Statistical Models and Applications,* Wydawnictwo Uniwersytetu Łódzkiego, Łódź 2011b, s. 191-197.

CLUSTERING OF INDUSTRIAL HOLIDAY PARTICIPANTS WITH THE USE OF LATENT CLASS ANALYSIS

Summary: The paper focuses on latent class models and their application for quantitative data. Latent class modeling is one of multivariate analysis techniques of the contingency table and can be viewed as a special case of model-based clustering, for multivariate discrete data. It is assumed that every observation comes from one of the numbers of subpopulations, with its own probability distribution. We used latent class analysis for grouping and detecting homogeneity of participants of industrial holiday − "Industriada" using `poLCA` package of R. We analyzed data collected by the Marketing Department of University of Economics in Katowice.

Keywords: latent class analysis, mixture model, categorical data.