EKONOMETRIA ECONOMETRICS 4(42) • 2013

ISSN 1507-3866

Ewa Genge
Uniwersytet Ekonomiczny w Katowicach

POCZUCIE ŚLĄSKOŚCI WŚRÓD ŚLĄZAKÓW – ANALIZA EMPIRYCZNA Z WYKORZYSTANIEM MODELI KLAS UKRYTYCH

Streszczenie: Modele klas ukrytych (*latent class models*), zwane również analizą klas ukrytych (*latent class analysis*), zaliczane są do tzw. modeli ze zmiennymi ukrytymi (*latent variable models*), w których ukrytą zmienną jest klasa. Modele te można zaliczyć również do tzw. podejścia modelowego w taksonomii (*model-based clustering*), gdzie wykorzystywana jest idea mieszanek rozkładów. Celem artykułu będzie dokonanie podziału Ślązaków w różny sposób postrzegających śląskość. Badania przeprowadzone zostaną za pomocą modelu klas ukrytych dla danych zgromadzonych przez Katedrę Pedagogiki Społecznej Uniwersytetu Śląskiego w Katowicach. Obliczenia zostaną przeprowadzone za pomocą pakietów `poLCA` i `flexmix` programu **R**.

Słowa kluczowe: model klas ukrytych, mieszanka rozkładów, podejście modelowe w taksonomii, dane jakościowe.

1. Wstęp

Modele klas ukrytych (*latent class models*), zwane również analizą klas ukrytych (*latent class analysis*), zaliczane są do tzw. modeli ze zmiennymi ukrytymi (*latent variable models*), w których ukrytą zmienną jest klasa. Modele te można zaliczyć również do tzw. podejścia modelowego w taksonomii (*model-based clustering*), gdzie wykorzystywana jest idea mieszanek rozkładów (zob. [Domański, Pruska 2000; Witek 2009]). W odróżnieniu od heurystycznych metod taksonomicznych (tj. metod hierarchicznych, iteracyjno-aglomeracyjnych), w których podstawą klasyfikacji obiektów do klas są różnego rodzaju miary odległości, w podejściu modelowym obiekty klasyfikowane są na podstawie prawdopodobieństw.

Istotą modelowania klas ukrytych jest badanie związków między kategoriami zmiennych nominalnych i porządkowych. Wykorzystuje się dane zawarte w tablicy kontyngencji. Metoda ta została wprowadzona przez Lazarsfelda w latach pięćdziesiątych [Lazarsfeld 1950], w kolejnych latach rozwijana była przez Goodmana [1970], który przyczynił się do rozwinięcia algorytmu pozwalającego otrzymać pa-

rametry funkcji największej wiarygodności, Habermana [1979], który pokazał związek pomiędzy modelami klas ukrytych oraz modelami logarytmiczno-liniowymi. Metoda ta nadal cieszy się dużym zainteresowaniem i rozwijana jest przez uczonych, takich m.in. jak Hagenaars, Vermunt, Linzer i Lewis [Hagenaars, McCutcheon 2002; Vermunt 2010; Linzer, Lewis 2011].

2. Model klas ukrytych ze zmiennymi towarzyszącymi – definicja

W modelu klas ukrytych rozważa się 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 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. Przyjmując, że liczba wszystkich kategorii jest równa $l = \sum_{j=1}^{m} l_j$, zbiór można określić za pomocą macierzy o wymiarach $n \times m$.

Model klas ukrytych, oprócz zmiennych obserwowanych, może zawierać również 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. Bardzo często zmienne towarzyszące wykorzystywane są w badaniach marketingowych, ekonomicznych, psychologicznych, w których pozyskanie zmiennych obserwowanych jest bardzo kosztowne (por. [Witek 2011a, s. 223-241]).

Parametry zmiennych towarzyszących szacowane są zazwyczaj wraz z pozostałymi parametrami modelu klas ukrytych (jednocześnie). Ten sposób estymacji nazywany 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, zakłada się, że mają one wpływ na prawdopodobieństwa *a priori*. W klasycznym modelu klas ukrytych (bez zmiennych towarzyszących) zakłada się, 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:

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

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, $\mathbf{z}_i = [z_{i1,\ldots,}z_{im_2}]$,

$\mathbf{\Theta}_s$ – wektor parametrów ukrytej klasy P_s,

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

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

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

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

3. Estymacja parametrów oraz wybór liczby klas ukrytych

Najczęściej stosowaną 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, metody Warda) istnieje kilka statystycznych miar służących wyborowi i ocenie ich jakości dopasowania. W badaniach empirycznych na początku zwykle sprawdza się dopasowanie dla $s = 1$. W kolejnych krokach zwiększa się liczbę klas o jeden tak długo, aż model osiągnie najlepsze dopasowanie. Wraz z dodatkową liczbą klas liczba szacowanych parametrów wzrasta o $1 + \sum_j (l_j - 1)$. Dlatego też bardzo często 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 Akaikego AIC (*Akaike Information Criterion*) [Akaike 1974]. Kryteria te mogą dawać niejednoznaczne wskazania co do oceny modeli klas ukrytych. Porównania różnych kryteriów informacyjnych można znaleźć m.in. w pracach [McLachlan, Peel 2000; Biernacki i in. 1999; Bozdogan 2000; Witek 2011b].

4. Analiza empiryczna

Analizę klas ukrytych przeprowadzono na podstawie danych zebranych przez Katedrę Pedagogiki Społecznej Uniwersytetu Śląskiego w Katowicach, prowadzącą badania naukowe na temat funkcjonowania różnych grup społecznych w warunkach demokracji. Badanie to dotyczyło mieszkańców Górnego Śląska, a jego celem

było zdiagnozowanie różnych grup społecznych na tym terenie. Warunkiem udziału w badaniu była odpowiednia data urodzenia (pomiędzy 1960 a 1970) oraz identyfikacja z przynależnością do grupy Górnoślązaków. Analiza została przeprowadzona z uwzględnieniem dziesięciu wybranych zmiennych obserwowanych oraz 4 zmiennych towarzyszących, dotyczących różnych aspektów śląskości[1].

W przykładzie wykorzystano dziesięć zmiennych obserwowanych $X_1 - X_{10}$. W nawiasie podano oryginalne numery pytań przeprowadzonego badania ankietowego.

1) X_1 (pyt. 34): Czy Ślązacy Pana/i zdaniem powinni posiadać status mniejszości narodowej? (1 – tak; 2 – nie; 3 – nie wiem);

2) X_2 (pyt. 35): W Polsce toczą się spory o przyznanie mowie śląskiej statusu języka regionalnego, co skutkuje podkreśleniem odrębności śląskiej oraz możliwością pozyskiwania funduszy na promocję języka śląskiego. Jakie jest Pana/i zdanie na ten temat? (1 – jestem za, uważam, że to dobry pomysł; 2 – jestem przeciw, uważam, że to zły pomysł; 3 – nie mam zdania);

3) X_3 (pyt. 36a): Posługuje się Pan/i gwarą śląską w domu rodzinnym? (1 – tak; 2 – nie);

4) X_4 (pyt. 36b): Posługuje się Pan/i gwarą śląską w pracy? (1 – tak; 2 – nie);

5) X_5 (pyt. 36c): Posługuje się Pan/i gwarą śląską w rozmowach towarzyskich? (1 – tak; 2 – nie);

6) X_6 (pyt. 58): Najlepiej czuję się wśród swoich znajomych, przyjaciół, osób, które mają podobne poglądy, są tego samego wyznania, identyfikują się ze Śląskiem. Zgadza się Pan/i z tym zdaniem? (1 – zgadzam się w 100%; 2 – zgadzam się w 75%; 3 – zgadzam się w 50%; 4 – zgadzam się w 25%; 5 – nie zgadzam się);

7) X_7 (pyt. 59): Potrafię się odnaleźć w różnych sytuacjach, nawet wśród obcych ludzi, szybko nawiązuję kontakty, przynależność regionalna nie ma dla mnie znaczenia. Zgadza się Pan/i z tym zdaniem? (1 – zgadzam się w 100%; 2 – zgadzam się w 75%; 3 – zgadzam się w 50%; 4 – zgadzam się w 25%; 5 – nie zgadzam się);

8) X_8 (pyt. 82g): Lubię tradycyjne obyczaje, pielęgnuję kontakty i celebruję uroczystości świąteczno-rodzinne (1 – tak; 2 – nie);

9) X_9 (pyt. 86a): Jako Ślązak uważa Pan/i, że w Polsce Ślązacy to grupa dyskryminowana? (1 – tak; 2 – nie);

10) X_{10} (pyt. 86b): Jako Ślązak uważa Pan/i, że w Polsce Ślązacy to grupa uprzywilejowana? (1 – tak; 2 – nie).

---

[1] Badanie przeprowadzono na próbie 1400 osób. Prawidłowo wypełniono 900 ankiet. W analizie empirycznej wyłączono ankiety z brakami danych, tj. parametry modelu szacowano na podstawie 896 obserwacji. Dobór próby był celowy. Badanie przeprowadzono w niespełna stu miastach Górnego Śląskach. Najwięcej ankiet zgromadzono w miastach, tj. w Katowicach, Bytomiu, Chorzowie, Mysłowicach, Siemianowicach Śląskich, Rudzie Śląskiej, Rybniku, Tychach, Zabrzu. Warunkiem udziału w badaniu była przynależność badanej osoby do pokolenia transformacji (data urodzenia pomiędzy rokiem 1960 a 1970), deklaracja w spisie powszechnym, że dana osoba czuje się Ślązakiem oraz jest zaliczana do grupy autochtonów (jej rodzina od trzech pokoleń mieszka na Śląsku).

Uwzględniono również następujące zmienne towarzyszące:

a) Z_1: płeć respondenta (1 – mężczyzna, 2 – kobieta);

b) Z_2: wykształcenie: (1 – nieukończone podstawowe, 2 – podstawowe, 3 – zawodowe, 4 – średnie ogólnokształcące, 5 – średnie zawodowe, 6 – policealne, 7 – wyższe, ze stopniem inżyniera, 8 – wyższe ze stopniem magistra);

c) Z_3: wiek;

d) Z_4: identyfikacja z klasą społeczną (1 – wyższą; 2 – średnią; 3 – niższą);

W badaniach wykorzystano pakiet `poLCA` programu **R**.

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 3, AIC zaś liczbę klas równą 5. Kryteria te nie zawsze dają wyniki jednoznaczne. W licznych pracach (zob. np. [Biernacki i in. 1999; Witek 2011b]) kryterium BIC w porównaniu z innymi kryteriami informacyjnymi 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 trzy.

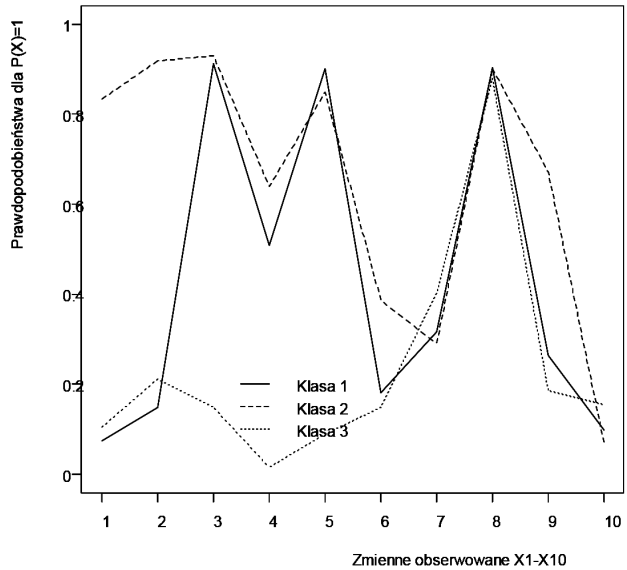
Następnie oszacowano kilka modeli klas ukrytych, różniących się zbiorem zmiennych towarzyszących (np. $Z_1 + Z_2$, $Z_1 + Z_3, Z_1 + Z_2 + Z_3 + Z_4$). Rozważano również interakcje pomiędzy zmiennymi towarzyszącymi (np. $Z_1 \times Z_2$, $Z_1 \times Z_2 \times Z_3$, $Z_1 \times Z_2 \times Z_3 \times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

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

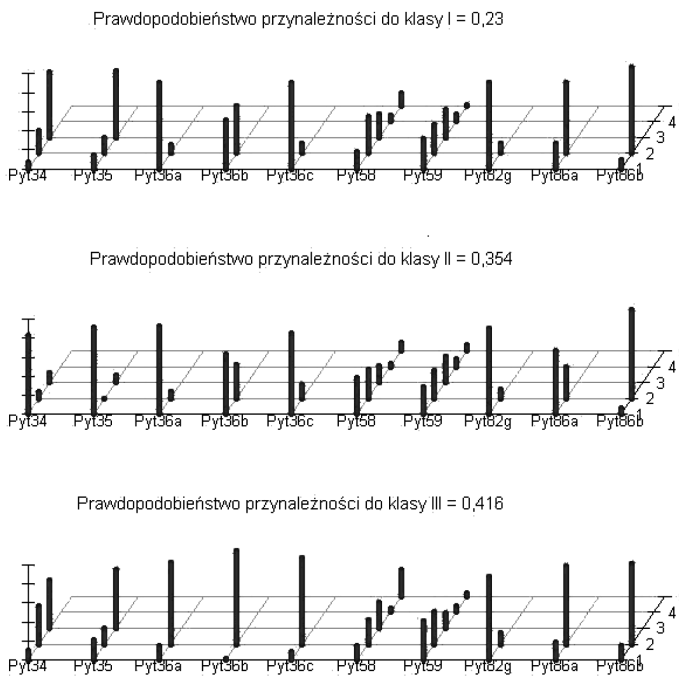
Dla wybranego modelu przedstawiono prawdopodobieństwa przyjmowania przez zmienne obserwowane wartości 1 (zgadzam się/tak) w klasie pierwszej, drugiej i trzeciej (rys. 1).

Na rysunkach 2 i 3 przedstawiono prawdopodobieństwa wyboru każdej kategorii dla zmiennych $X_1 - X_{10}$ dla trzech klas. Widoczne są także prawdopodobieństwa *a priori* (wagi) dla poszczególnych klas.

W klasie pierwszej, najmniej licznej ($\tau_1 = 0,23$), prawie 7,5% respondentów twierdzi, że Ślązacy powinni mieć status mniejszości narodowej, 24% jest przeciwnego zdania. Natomiast największa część respondentów (68,5%) jest niezdecydowana w tej kwestii. Stosunkowo niewielki odsetek (15%) w tej grupie stanowią respondenci, którzy uważają, że należy promować język śląski, prawie tak samo liczna część tej grupy (15,7%) jest przeciwna takiej promocji; największy odsetek stanowią również osoby niemające zdania na ten temat. 91% ankietowanych z tej grupy posługuje się gwarą śląską w domu, 51% w pracy, a 90% nie krępuje się posługiwać gwarą w towarzystwie. Największy odsetek (prawie 40%) w tej grupie stanowią respondenci, którzy z opinią, że najlepiej czują się wśród osób mających podobne poglądy



Rys. 1. Prawdopodobieństwo wyboru wartości 1 dla zmiennych $X_1 - X_{10}$
Źródło: opracowanie własne.



Rys. 2. Wyniki segmentacji respondentów
Źródło: opracowanie własne.

```
Conditional item response (column) probabilities,
by outcome variable, for each class (row)

$Pyt34
 Pr(1) Pr(2) Pr(3)
class 1: 0.0736 0.2415 0.6849
class 2: 0.8327 0.0707 0.0966
class 3: 0.1044 0.3963 0.4994

$Pyt35
 Pr(1) Pr(2) Pr(3)
class 1: 0.1491 0.1572 0.6938
class 2: 0.9188 0.0028 0.0784
class 3: 0.2121 0.1681 0.6198

$Pyt36a
 Pr(1) Pr(2)
class 1: 0.9116 0.0884
class 2: 0.9297 0.0703
class 3: 0.1491 0.8509

$Pyt36b
 Pr(1) Pr(2)
class 1: 0.5079 0.4921
class 2: 0.6389 0.3611
class 3: 0.0159 0.9841

$Pyt36c
 Pr(1) Pr(2)
class 1: 0.9013 0.0987
class 2: 0.8487 0.1513
class 3: 0.0896 0.9104

$Pyt58
 Pr(1) Pr(2) Pr(3) Pr(4) Pr(5)
class 1: 0.1823 0.3803 0.2463 0.0628 0.1282
class 2: 0.3867 0.3099 0.1781 0.0346 0.0907
class 3: 0.1498 0.2532 0.2649 0.0458 0.2863

$Pyt59
 Pr(1) Pr(2) Pr(3) Pr(4) Pr(5)
class 1: 0.3174 0.2985 0.2969 0.0784 0.0088
class 2: 0.2901 0.2913 0.2810 0.0779 0.0598
class 3: 0.4039 0.3356 0.1576 0.0648 0.0382

$Pyt82g
 Pr(1) Pr(2)
class 1: 0.9024 0.0976
class 2: 0.9019 0.0981
class 3: 0.8799 0.1201

$Pyt86a
 Pr(1) Pr(2)
class 1: 0.2633 0.7367
class 2: 0.6726 0.3274
class 3: 0.1865 0.8135

$Pyt86b
 Pr(1) Pr(2)
class 1: 0.0994 0.9006
class 2: 0.0693 0.9307
class 3: 0.1543 0.8457
```

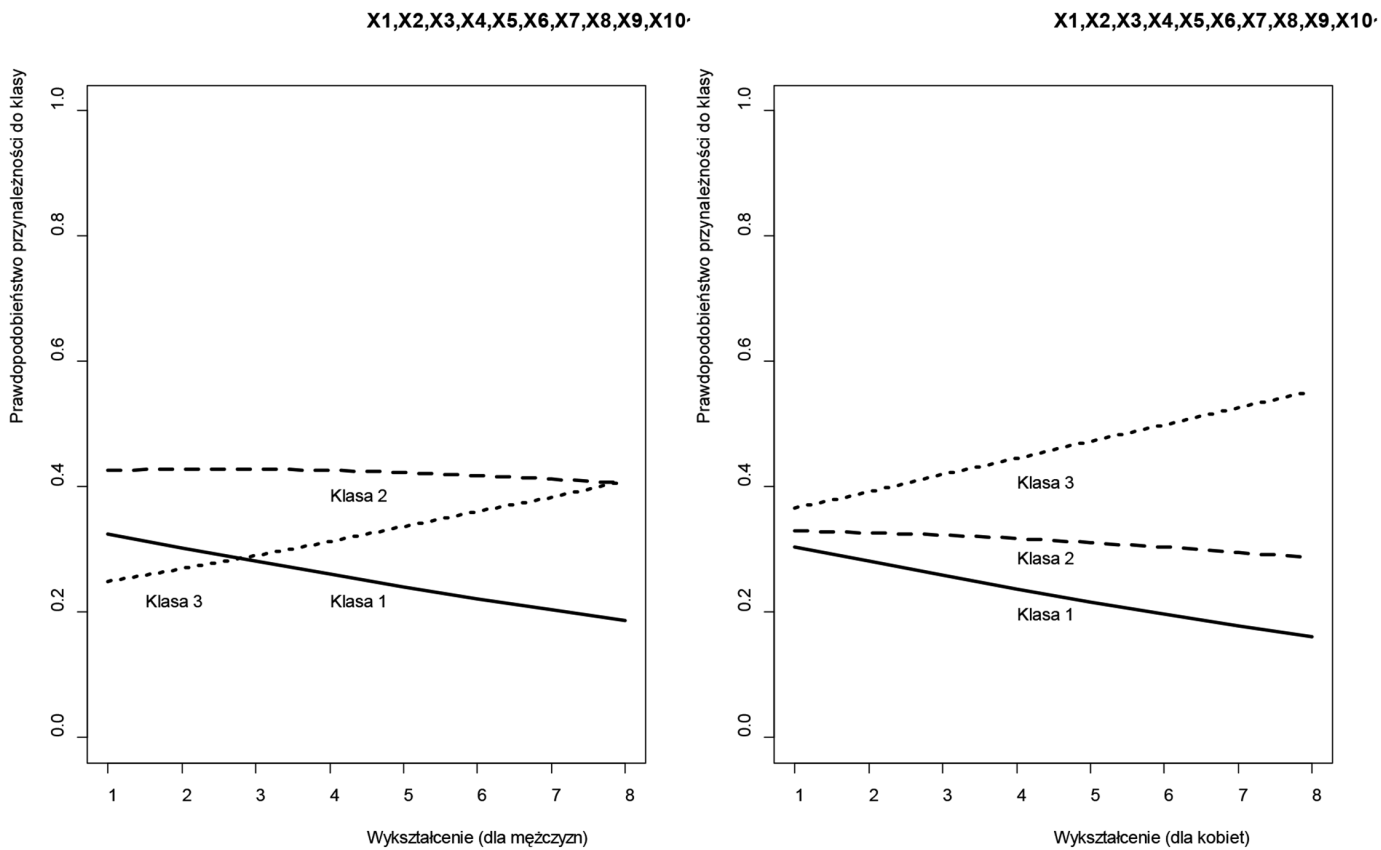
Rys. 3. Wyniki segmentacji respondentów – wydruk z programu R

Źródło: opracowanie własne.

i identyfikujących się ze Śląskiem, zgadzają się w 75%, 25% ankietowanych zgadza się z tą opinią w 50%, natomiast 18% badanych zgadza się z tym stwierdzeniem aż w 100%. Najwięcej badanych z tej klasy (32%) jest całkowicie przekonanych, że potrafi się odnaleźć w różnych sytuacjach, nawet wśród obcych ludzi, szybko nawiązując kontakty, zaś przynależność regionalna nie ma dla nich znaczenia. Niewiele mniejszy odsetek badanych (30%) zgadza się z tym stwierdzeniem zarówno w 75%, jak i w 50%. Zaledwie 10% ankietowanych nie jest przywiązanych do śląskich tradycji i uroczystości świąteczno-rodzinnych. 26% badanych z tej grupy twierdzi, że Ślązacy są w Polsce grupą dyskryminowaną, a tylko 10% jest zdania, że jest to grupa w pewien sposób uprzywilejowana.

Klasa druga jest klasą liczniejszą – należy do niej 35% wszystkich ankietowanych. W klasie tej aż 83% respondentów uważa, że Ślązacy powinni mieć status mniejszości narodowej; tylko 7% ma odmienne zdanie na ten temat. W klasie tej również największy odsetek (92%) sądzi, że należy promować język śląski, a tylko nikły procent (0,02%) badanych jest przeciwny takiej inicjatywie. Klasa ta również cechuje się największym odsetkiem osób przyznających się do tego, że posługuje się gwarą śląską w domu rodzinnym (93%) oraz w pracy (64%). Ankietowani tej klasy wypadają nieco gorzej niż respondenci klasy pierwszej, jeżeli chodzi o posługiwanie się gwarą w rozmowach towarzyskich (85%). Prawie (40%) badanych w 100% zgadza się z opinią, że najlepiej czuje się wśród swoich znajomych, przyjaciół, osób, które mają podobne poglądy, są tego samego wyznania, identyfikują się ze Śląskiem, 31% zgadza się z tą opinią w 75%, a 18% w 50%. W grupie tej występuje najmniejszy odsetek osób, w ogóle niezgadzających się z tym twierdzeniem. W klasie drugiej ok. 30% osób zgadza się w pełni, w 75% oraz w połowie ze zdaniem, że potrafi odnaleźć się w różnych sytuacjach, nawet wśród obcych ludzi. Jednakże w grupie tej występuje największy odsetek respondentów (6%) przyznających się do tego, że przynależność regionalna ma dla nich znaczenie i w związku z tym szybko nie nawiązują kontaktów towarzyskich. Podobnie jak w klasie pierwszej 90% ankietowanych lubi i pielęgnuje śląskie tradycje i obyczaje. Grupę tę z pewnością wyróżnia największy odsetek osób (67%) twierdzących, że Ślązacy są w Polsce dyskryminowani, oraz najmniejszy odsetek (7%) uznających Ślązaków za grupę uprzywilejowaną.

Klasa trzecia jest klasą najliczniejszą ($\tau_3 = 0{,}42$). Klasę tę wyróżnia największy odsetek (40%) przeciwników mniejszości narodowej (w porównaniu z klasą pierwszą i drugą) oraz 50-procentowy odsetek osób niepotrafiących określić swego zdania na ten temat. W grupie tej występuje również największy odsetek respondentów przeciwnych promocji języka śląskiego (17%), a także spory procent osób (62%) niechcących wypowiadać swego zdania w tej kwestii. Jeżeli chodzi o posługiwanie się gwarą śląską, ankietowani klasy trzeciej posługują się nią w najmniejszym stopniu, tj. 15% przyznaje się do posługiwania gwarą w domu, niespełna 2% w pracy i 9% w towarzystwie. W klasie tej, w porównaniu z dwiema poprzednimi klasami, najmniejszy odsetek ankietowanych (15%) w pełni zgadza się z opinią, że najlepiej czuje się wśród osób identyfikujących się ze Śląskiem, największy zaś odsetek (prawie 30%) stanowią osoby, które nawet w najmniejszym stopniu nie zgadzają się z tą



Rys. 4. Wykres przynależności mężczyzn (strona lewa) i kobiet (strona prawa) do trzech klas

Źródło: opracowanie własne.

opinią. Najwięcej badanych tej klasy (40%) jest całkowicie przekonanych, że potrafi się odnaleźć w różnych sytuacjach, nawet wśród obcych ludzi, szybko nawiązuję kontakty, a przynależność regionalna nie ma dla nich znaczenia. Niewiele mniejszy odsetek badanych (34%) zgadza się z tym stwierdzeniem w 75%. Nieco mniej badanych niż w klasie pierwszej i drugiej, bo 88%, jest przywiązanych do śląskich tradycji i uroczystości świąteczno-rodzinnych. W odniesieniu do klasy pierwszej i drugiej w grupie trzeciej najmniejszy odsetek osób (18%) stanowią ankietowani będący zdania, że Ślązacy są w Polsce dyskryminowani. Zarazem w klasie tej aż 15% osób uważa Ślązaków za grupę uprzywilejowaną.

W kolejnej części pracy dokonano analizy wpływu zmiennych towarzyszących na przynależność analizowanych obiektów do klas. Jeżeli chodzi o zmienną towarzyszącą „wykształcenie", to dla mężczyzn o najniższym wykształceniu najwyższe jest prawdopodobieństwo przynależności do klasy drugiej, a najniższe do klasy trzeciej. Należy jednak zaznaczyć, że w przypadku klasy drugiej prawdopodobieństwo przynależności do tej klasy jest najwyższe i prawie takie samo dla osób o różnym poziomie wykształcenia. Prawdopodobieństwo przynależności do klasy pierwszej spada wraz z lepszym wykształceniem respondentów, prawdopodobieństwo zaś przynależności do klasy trzeciej wzrasta wraz z lepszym wykształceniem respondentów. Wpływ wykształcenia na przynależność do klas dla kobiet jest bardzo podobny (rys. 4), z tą różnicą, że prawdopodobieństwo przynależności do klasy trzeciej jest najwyższe dla kobiet o każdym poziomie wykształcenia.

Dokonując analizy wpływu zmiennej towarzyszącej płeć, można stwierdzić, że prawdopodobieństwo przynależności do klasy drugiej jest wyższe dla mężczyzn niż dla kobiet. Odwrotna sytuacja ma miejsce w przypadku klasy trzeciej, zaś prawdopodobieństwo przynależności do klasy pierwszej jest takie samo dla kobiet jak dla mężczyzn[2]. Ze względu na ograniczenia objętościowe na rys. 4 zamieszczono tylko wykres dla zmiennej towarzyszącej Z_2 (wykształcenie).

5. Podsumowanie

W artykule przedstawiono przykład zastosowania modeli klas ukrytych do oceny poczucia śląskości wśród Ślązaków. Analiza klas ukrytych umożliwiła segmentację respondentów na podstawie odpowiedzi udzielonych w badaniu przeprowadzonym przez Katedrę Polityki Społecznej Uniwersytetu Śląskiego w Katowicach. Wyodrębniono trzy klasy o podobnych wzorcach zachowań i postaw dla śląskich respondentów. Dokonano również oceny wpływu zmiennych demograficznych na ich przynależność do klas.

[2] Dla zmiennej towarzyszącej „płeć" dokonano interpretacji i sporządzono wykres, przyjmując, że zmienna jakościowa „wykształcenie" jest równa kategorii występującej najczęściej, tj. wykształcenie średnie ogólnokształcące (zob. np. [Linzer, Lewis 2011; Witek 2011a]).

Do klasy pierwszej zaliczono Ślązaków przywiązanych do gwary i tradycji, lecz nieco mało odważnych, niemających do końca poczucia własnej tożsamości i odrębności. Klasa druga to grupa osób postrzegających śląskość jako szansę na realizację wspólnych projektów i budowanie swojego miejsca w świecie – a zatem dążąca do zwiększonego wpływu na własną przyszłość. Największą grupę jednak stanowią osoby mieszkające na Górnym Śląsku, lecz o najbardziej sceptycznym nastawieniu do gwary i promocji języka śląskiego, niezainteresowane statusem mniejszości narodowej oraz nie do końca czujące się Ślązakami.

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.

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

Bozdogan H., *Akaike's information criterion and recent developments in information criterion*, "Journal of Mathematical Psychology" 2000, 44, s. 62-91.

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.

Domański C., Pruska K., *Nieklasyczne metody statystyczne*, PWE, Warszawa 2000.

Goodman L., *The multivariate analysis of qualitative data: interactions among multiple classification*, "Journal of the American Statistical Association" 1970, 65, s. 226-256.

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

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

Lazarsfeld P.F., *The Logical and Mathematical Foundations of Latent Structure Analysis*, [w:] S.A. Stouffer, *Measurement and Prediction*, John Wiley&Sons, New York 1950, s. 362-412.

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

McLachlan G.J., Peel D., *Finite Mixture Models*, Wiley, New York 2000, s. 81-116.

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

Vermunt J.K., *Latent class modeling with covariates: Two improved three-step approaches*, "Political Analysis" 2010, 18, s. 450-469.

Witek E., *Analiza skupień – podejście modelowe*, [w:] M. Walesiak, E. Gatnar, *Statystyczna analiza danych z wykorzystaniem programu R*, PWN, Warszawa 2009, s. 434-462.

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:] C. Domański, J. Białek, *Folia Oeconomica 255, Methodological Aspects of Multivariate Statistical Analysis, Statistical Models and Applications,* Wydawnictwo Uniwersytetu Łódzkiego, Łódź 2011b, s. 191-197.

A SENSE OF BEING SILESIAN – AN EMPIRICAL ANALYSIS WITH THE USE OF LATENT CLASS MODELS

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 each 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 Silesian people using `poLCA` package of R. We analyzed data collected by the Department of Social Pedagogy, University of Silesia in Katowice.

Keywords: latent class analysis, mixture model, model-based clustering, categorical data.