PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS nr 507 • 2018

Taksonomia 30
Klasyfikacja i analiza danych – teoria i zastosowania

ISSN 1899-3192
e-ISSN 2392-0041

Marcin Szymkowiak
Uniwersytet Ekonomiczny w Poznaniu, Urząd Statystyczny w Poznaniu
e-mail: m.szymkowiak@ue.poznan.pl

Tomasz Klimanek
Urząd Statystyczny w Poznaniu
e-mail: t.klimanek@stat.gov.pl

ANALIZA KLAS UKRYTYCH W BADANIU NIEPEŁNOSPRAWNOŚCI[1]

LATENT CLASS ANALYSIS IN DISABILITY SURVEY

DOI: 10.15611/pn.2018.507.24
JEL Classification: C38

Streszczenie: Analiza klas ukrytych jest jedną z metod wielowymiarowej analizy danych, której podstawy teoretyczne zostały sformułowane w połowie XX wieku [Lazarsfeld 1950]. Jej głównym celem jest redukcja liczby zmiennych przy jak najmniejszym poziomie utraty informacji o badanym zjawisku. Dzięki jej zastosowaniu możliwe jest odkrycie ukrytych klas w analizowanej populacji, które pełnią funkcję nieobserwowalnych czynników wpływających na zależności pomiędzy jednostkami w danej klasie [Brzezińska 2015]. Głównym celem artykułu jest zastosowanie analizy klas ukrytych na gruncie badań prowadzonych przez statystykę publiczną w Polsce w kontekście zjawiska niepełnosprawności. Autorzy, wykorzystując dane z Narodowego Spisu Powszechnego Ludności i Mieszkań 2011 (NSP 2011), procedurę LCA programu SAS oraz pakiet poLCA programu R, dedykowane tej technice wielowymiarowej analizy danych, podejmują próbę stworzenia „profilu demograficznego" osoby niepełnosprawnej w Polsce oraz charakterystyki grup osób, które odmawiały udzielenia odpowiedzi na pytania dotyczące tego zjawiska społecznego.

Słowa kluczowe: analiza klas ukrytych, niepełnosprawność, proc LCA, pakiet poLCA, Narodowy Spis Powszechny Ludności i Mieszkań 2011.

Summary: The main aim of the article is to present the application of latent class analysis for surveys conducted by the Central Statistical Office in Poland in the context of disability. Using data from the National Census of Population and Housing 2011 and software tools, such as the proc LCA procedure in SAS and the poLCA package in R, which implement this multivariate technique of data analysis, the authors describe an attempt to create a demographic profile of the disabled person in Poland. Moreover, the method is used as a way of

[1] Artykuł powstał w ramach grantu „Estymacja pośrednia w zakresie badania niepełnosprawności na podstawie NSP 2011", który został sfinansowany ze środków Narodowego Centrum Nauki przyznanych na podstawie decyzji numer DEC-2013/11/B/HS4/01472.

identifying different groups of people who refused to answer questions about this important social phenomenon.

Keywords: latent class analysis, disability, proc LCA, poLCA package, National Census of Population and Housing 2011.

1. Wstęp

Zjawisko niepełnosprawności należy współcześnie do jednych z najpoważniejszych problemów społecznych. Według danych Światowej Organizacji Zdrowia [WHO 2016] na całym świecie ponad miliard osób cierpi na różnego rodzaju dysfunkcje, które wpisują się w definicję osoby niepełnosprawnej. Szacuje się przy tym, że blisko 200 milionów osób doświadcza niepełnosprawności w znacznym stopniu. Również prognozy w tym zakresie są niepokojące i przewiduje się, że liczba osób niepełnosprawnych na całym świecie będzie wzrastać. Będzie to głównie konsekwencją zjawiska starzenia się społeczeństw. Z tego powodu w Milenijnych Celach Rozwoju Narodów Zjednoczonych kwestię niepełnosprawności uczyniono jednym z najważniejszych problemów, który wymagać będzie kompleksowego rozwiązania [UN 2015].

Również w Polsce zjawisko niepełnosprawności stanowi bardzo poważny problem wymagający odpowiednich rozwiązań. Według danych z ostatniego NSP 2011 liczba osób niepełnosprawnych w Polsce wynosiła około 4,7 mln. Stanowiło to 12,2% ludności kraju, przy czym 4,1% stanowiły osoby niepełnosprawne wyłącznie biologicznie, 6,9% niepełnosprawne biologicznie i prawnie oraz 1,2% niepełnosprawne wyłącznie prawnie [GUS 2012].

Głównym celem artykułu jest zastosowanie analizy klas ukrytych w kontekście niepełnosprawności z wykorzystaniem danych pochodzących z NSP 2011. Do stworzenia swego rodzaju „profilu demograficznego" osoby niepełnosprawnej autorzy wykorzystają pakiet statystyczny SAS i procedurę LCA oraz program R i pakiet poLCA. Podjęta zostanie również dyskusja na temat braków odpowiedzi w odniesieniu do pytania spisowego dotyczącego niepełnosprawności i odkrycia najważniejszych charakterystyk grup osób, które w tym zakresie odmówiły jej podania.

2. Analiza klas ukrytych

Analiza klas ukrytych (*Latent Class Analysis* – LCA) jest jedną z metod wielowymiarowej analizy danych, której celem jest redukcja liczby zmiennych przy jak najmniejszej utracie informacji o badanym zjawisku oraz wykrycie nieobserwowalnej heterogeniczności w populacji poprzez znalezienie tzw. klas ukrytych, pełniących rolę nieobserwowalnych czynników wpływających na zależność pomiędzy obiektami przypisanymi do danej klasy [Brzezińska 2015].

Podstawy teoretyczne tej metody zostały sformułowane w 1950 roku przez Lazarsfelda, który szczegółowo omówił koncepcję tej wielowymiarowej techniki analizy danych [Lazarsfeld 1950]. Prace podjęte przez Lazarsfelda były następnie kontynuowane przez niego [Lazarsfeld 1959], jak i wielu innych statystyków [Anderson 1959; Clogg 1988; Heinen 1996]. Szczególną rolę odegrały jednak prace Goodmana, który wdrożył metodę największej wiarogodności na potrzeby estymacji parametrów modelu w analizie klas ukrytych [Goodman 1974]. Do współczesnych opracowań dogłębnie poruszających problematykę analizy klas ukrytych można zaliczyć prace Collinsa i Lanzy [2010] czy w języku polskim Sikorskiej [2012]. Stanowić będą one punkt wyjścia w niniejszym artykule do zapisu modelu analizy klas ukrytych.

W dalszej części zakładać będziemy, że dysponujemy pewnymi zmiennymi obserwowalnymi $j = 1,2, ..., J$ podlegającymi bezpośredniemu pomiarowi. Zakładamy ponadto, że każda ze zmiennych obserwowalnych ma $r_j = 1,2, ..., R_j$ kategorii. Punktem wyjścia w analizie klas ukrytych jest odpowiednia tablica kontyngencji, która powstaje poprzez zestawienie wariantów wszystkich J zmiennych obserwowalnych. Tablica ta ma $W = \prod_{j=1}^{J} R_j$ komórek. Każda z $w = 1,2, ..., W$ komórek tabeli kontyngencji odpowiada pewnemu wzorcowi odpowiedzi $\boldsymbol{y} = (r_1, ... r_J)$ respondentów dla J zmiennych obserwowalnych. Zakładamy, że $\boldsymbol{P}(\boldsymbol{Y} = \boldsymbol{y})$ oznacza prawdopodobieństwo uzyskania wzorca odpowiedzi $\boldsymbol{y} = (r_1, ... r_J)$, przy czym $\Sigma P(Y = y) = 1$. Niech ponadto L oznacza zmienną ukrytą o c kategoriach, przy czym $c = 1,2, ..., C$ oznacza liczbę wyodrębnionych klas ukrytych.

Kluczową rolę w analizie klas ukrytych odgrywa prawdopodobieństwo przynależności jednostki do klasy ukrytej c oznaczane jako γ_c tj. $P(L = c) = \gamma_c$, gdzie $\sum_{c=1}^{C} \gamma_c = 1$, oraz prawdopodobieństwo odpowiedzi r_j na pytanie odpowiadające zmiennej obserwowalnej j pod warunkiem przynależności danej jednostki do klasy ukrytej c, oznaczane jako $\rho_{j,r_j|c}$, przy czym $\sum_{r_j=1}^{R_j} \rho_{j,r_j|c} = 1$. Jeżeli y_j jest odpowiedzią na j-tą zmienną obserwowalną podlegającą pomiarowi, to prawdopodobieństwo $\boldsymbol{P}(\boldsymbol{Y} = \boldsymbol{y})$ można wyrazić następującym wzorem:

$$P(\boldsymbol{Y} = \boldsymbol{y}) = \sum_{c=1}^{C} P(L = c) P(\boldsymbol{Y} = \boldsymbol{y} | L = c), \tag{1}$$

gdzie $\boldsymbol{P}(\boldsymbol{L} = c) = \gamma c$ oraz

$$P(\boldsymbol{Y} = \boldsymbol{y} | L = c) = \prod_{j=1}^{J} \prod_{r_j=1}^{R_j} \rho_{j,r_j|c}^{I(y_j = r_j)}. \tag{2}$$

Kluczową kwestią w analizie klas ukrytych jest ustalenie liczby klas. W literaturze przedmiotu [Collins, Lanza 2010] zaproponowano wiele mierników, które można wykorzystać w procesie ustalania optymalnej liczby klas ukrytych. Punkt wyjścia stanowi statystyka G^2 określona poniższym wzorem:

$$G^2 = 2\sum_{i=1}^{W} n_i ln\left(\frac{n_i}{\hat{n}_i}\right), \tag{3}$$

gdzie n_i to rzeczywista liczebność w odpowiedniej komórce wielowymiarowej tablicy kontyngencji, a $\hat{n}_i$ to liczebność teoretyczna. Statystyka G^2 ma rozkład χ^2_{df}, gdzie liczba stopni swobody określona jest jako $df = W - K - 1$, a K to liczba niezależnych parametrów zdefiniowana jako:

$$K = C - 1 + C\sum_{j=1}^{J}(R_j - 1). \tag{4}$$

W przypadku wyboru modeli o różnej liczbie klas ukrytych należy wybrać ten, dla którego zaobserwuje się duży spadek wartości statystyki G^2, a dalsze spadki będą nieznaczne [Sikorska 2012]. W procesie ustalania liczby klas ukrytych można również skorzystać z odpowiednich kryteriów informacyjnych wykorzystujących w swojej konstrukcji statystykę G^2:

$$AIC = G^2 + 2K, \tag{5}$$

$$BIC = G^2 + K \times ln(n), \tag{6}$$

$$CAIC = BIC + K, \tag{7}$$

gdzie n oznacza liczebność próby. Ustalając liczbę klas ukrytych w oparciu o przedstawione powyżej kryteria informacyjne, należy wybrać taką liczbę klas ukrytych, dla której odpowiednie kryterium informacyjne przyjmie wartość najmniejszą. W praktyce najczęściej korzysta się przy tym z kryterium *BIC* określonego wzorem (6).

3. Niepełnosprawność w NSP 2011

Narodowy Spis Powszechny Ludności i Mieszkań 2011 przeprowadzony został metodą mieszaną, tzn. dane były pozyskiwane ze źródeł administracyjnych (rejestrów i systemów informacyjnych) oraz zbierane bezpośrednio od ludności w ramach badania reprezentacyjnego, a także tzw. badania pełnego. Zgodnie z przyjętą w NSP 2011 definicją zbiorowość osób niepełnosprawnych została podzielona na dwie podstawowe grupy:

- osoby niepełnosprawne prawnie, tj. takie, które posiadały odpowiednie, aktualne orzeczenie wydane przez organ do tego uprawniony;
- osoby niepełnosprawne tylko biologicznie, tj. takie, które nie posiadały orzeczenia, ale miały (odczuwały) całkowicie lub poważnie ograniczoną zdolność do wykonywania czynności podstawowych stosownie do swojego wieku.

Warto podkreślić, że kwestia niepełnosprawności w NSP 2011 należała do szczególnie wrażliwych problemów. Miało to swoje odzwierciedlenie w liczbie odmów

na pytania związane ze zjawiskiem niepełnosprawności. Blisko 1,5 mln respondentów nie udzieliło odpowiedzi na pytania dotyczące niepełnosprawności, co pozwala wysnuć wniosek, że braki danych w tym zakresie mogą rzutować na ostateczną ocenę tego zjawiska społecznego.

Omówione w tej części artykułu dane pochodzące z NSP 2011 w obszarze niepełnosprawności zostały wykorzystane na potrzeby egzemplifikacji analizy klas ukrytych. Miało to na celu stworzenie swego rodzaju „profilu demograficznego" osoby niepełnosprawnej oraz wskazanie głównych grup osób, które odmawiały udzielenia odpowiedzi na pytania dotyczące tego zjawiska.

4. Niepełnosprawność w NSP 2011 w świetle analizy klas ukrytych

W pierwszej kolejności z wykorzystaniem analizy klas ukrytych stworzono „profil demograficzny" osoby niepełnosprawnej w sensie biologicznym, wykorzystując dane pochodzące z NSP 2011. Na potrzeby analizy przyjęto następujące zmienne:

- płeć osoby niepełnosprawnej (1 – mężczyzna, 2 – kobieta),
- miejsce zamieszkania osoby niepełnosprawnej (1 – miasto, 2 – wieś),
- wiek osoby niepełnosprawnej (1 – do 29 lat, 2 – 30–49 oraz 3 – 50+),
- status na rynku pracy osoby niepełnosprawnej (1 – pracująca, 2 – bezrobotna, 3 – bierna zawodowo).

Zbiór danych ograniczono przy tym do wszystkich tych osób, które zadeklarowały, że są osobami niepełnosprawnymi w sensie biologicznym i dla których znane były wartości wszystkich analizowanych cech. Zbiór ten liczył blisko 900 tys. rekordów. W obliczeniach wykorzystano procedurę LCA języka 4GL programu SAS, a na potrzeby wizualizacji uzyskanych wyników pakiet poLCA programu R [Linzer, Lewis 2011].

Tabela 1 przedstawia informacje na temat wartości opisanych wcześniej kryteriów informacyjnych oraz statystyki G^2. Na podstawie danych zawartych w tej tabeli można przyjąć, że liczba klas ukrytych powinna wynosić 3. Świadczy o tym najniższa wartość współczynnika BIC oraz CAIC, a także fakt, że dla tej liczby klas ukrytych odnotowano duży spadek wartości statystyki G^2w porównaniu z jej poziomem dla dwóch klas.

Tabela 1. Wartości statystyki G^2 oraz kryteriów informacyjnych w badaniu niepełnosprawności

C	G^2	*AIC*	*CAIC*	*BIC*
2	287,38	313,38	429,56	416,56
3	90,80	130,80	309,54	289,54
4	30,35	84,35	325,65	298,65
5	5,95	73,95	377,81	343,81

Źródło: opracowanie własne.

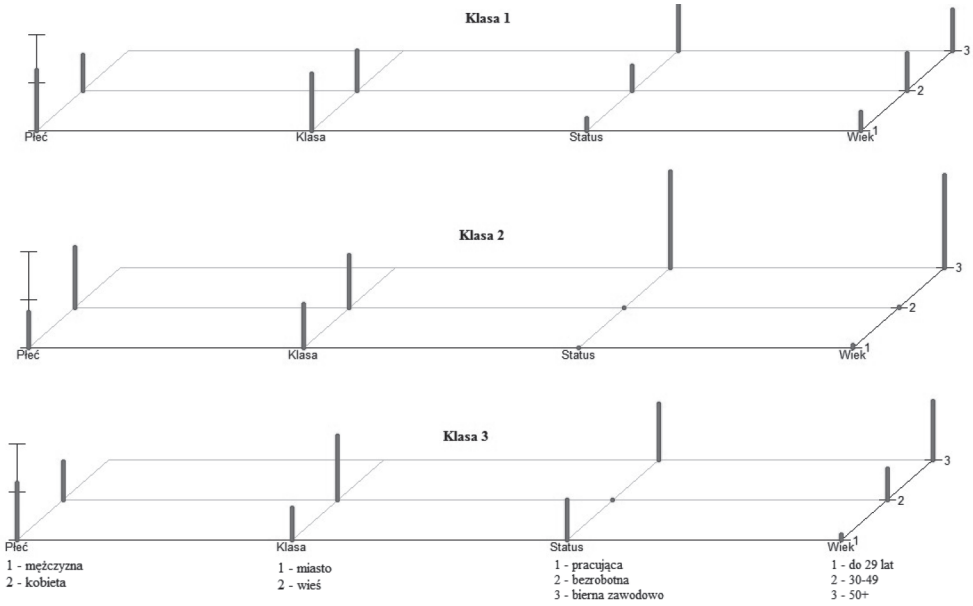
Tabela 2 przedstawia z kolei oceny prawdopodobieństw γ_c przynależności do odpowiednich klas ukrytych. Do klasy ukrytej 1 należy zatem blisko 62% obserwacji, do klasy ukrytej 2 ponad 27% obserwacji, a w klasie ukrytej 3 znalazło się nieco ponad 10% jednostek.

Tabela 2. Oceny prawdopodobieństw γ_c przynależności do odpowiednich klas ukrytych w badaniu niepełnosprawności

Klasa	1	2	3
γ_c	0,6191	0,2759	0,1050

Źródło: opracowanie własne.

Na podstawie wyników estymacji prawdopodobieństw $\rho_{j,r_j|c}$ przedstawionych na rys. 1 można stworzyć profil osoby niepełnosprawnej w sensie biologicznym. Wysokość słupka oznacza odpowiednie prawdopodobieństwo $\rho_{j,r_j|c}$.



Rys. 1. Profil demograficzny osób niepełnosprawnych biologicznie – oceny prawdopodobieństw $\rho_{j,r_j|c}$

Źródło: opracowanie własne.

Do klasy 1 należą głównie osoby niepełnosprawnie w sensie biologicznym, mieszkające w mieście, płci męskiej, w wieku 50+ oraz bierne zawodowo. Do klasy 2 z kolei należą głównie kobiety, mieszkające na wsi, w wieku 50+ oraz bierne zawodowo. W klasie 3 natomiast znajdują się głównie osoby niepełnosprawnie biologicznie płci męskiej, mieszkające na wsi, bierne zawodowo lub pracujące, również

w grupie wiekowej 50+. Z powyższej analizy wyłania się swego rodzaju portret osoby niepełnosprawnej jako osoby biernej zawodowo oraz starszej. Stanowi to niejako potwierdzenie zauważalnego zjawiska występowania niepełnosprawności na skutek starzenia się populacji.

Analizę klas ukrytych wykorzystano również na potrzeby scharakteryzowania osób, które w NSP 2011 na pytania dotyczące niepełnosprawności nie chciały udzielić odpowiedzi, traktując je jako bardzo wrażliwe. Podobnie jak wyżej, w charakterze zmiennych obserwowalnych wzięto pod uwagę płeć, miejsce zamieszkania, status na rynku pracy oraz wiek. Przedstawiona analiza dotyczy osób, co do których nie wiadomo, czy były one w pełni sprawne, czy też cierpiały na jakiś rodzaj niepełnosprawności. Innymi słowy, przeprowadzono ją na zbiorze osób odmawiających udzielenia odpowiedzi w kontekście niepełnosprawności. Oznaczenia wariantów poszczególnych zmiennych są analogiczne do wcześniej przyjętych.

Tabela 3 przedstawia informacje na temat wartości najważniejszych kryteriów informacyjnych oraz statystyki G^2.

Tabela 3. Wartości statystyki G^2 oraz kryteriów informacyjnych w badaniu odmów

C	G^2	*AIC*	*CAIC*	*BIC*
2	2881,16	2907,16	3055,80	3042,80
3	956,18	996,18	1224,86	1204,86
4	334,70	388,70	697,42	670,42
5	97,47	165,47	554,23	520,23
6	23,08	105,08	573,87	532,87
7	0,0015	96,00	644,83	596,83

Źródło: opracowanie własne.

Dokonując analizy danych zawartych w tabeli 3 można uznać, że liczba klas ukrytych powinna wynosić 5. Świadczy o tym najniższa wartość współczynnika BIC oraz CAIC.

Tabela 4 przedstawia z kolei oceny prawdopodobieństw γ_c przynależności do odpowiednich klas ukrytych.

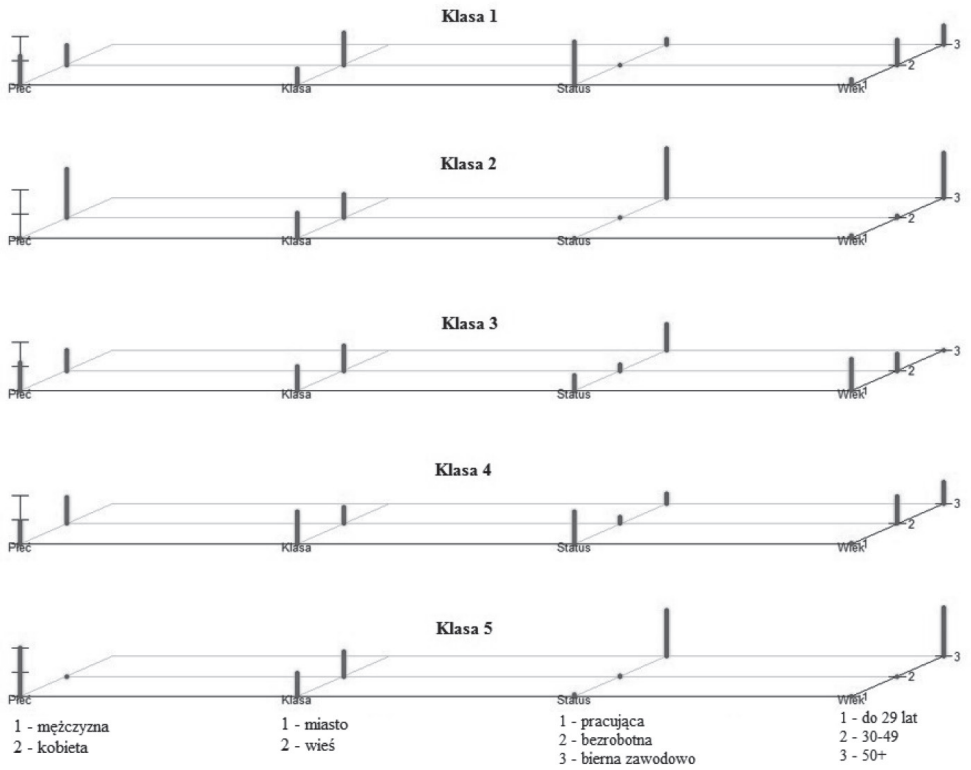
Tabela 4. Oceny prawdopodobieństw γ_c przynależności do odpowiednich klas ukrytych w badaniu odmów na temat niepełnosprawności

Klasa	1	2	3	4	5
γ_c	0,170	0,283	0,211	0,162	0,174

Źródło: opracowanie własne.

Do klasy ukrytej 1 należy zatem 17% obserwacji, do klasy ukrytej 2 ponad 28% obserwacji, w klasie ukrytej 3 znalazło się nieco ponad 21% jednostek, w klasie 4 ponad 16%, a w ostatniej klasie nieco ponad 17% obserwacji.

Na podstawie wyników estymacji prawdopodobieństw $\rho_{j,r_j|c}$ przedstawionych na rys. 2 można stworzyć profil osoby, która nie udzielała odpowiedzi w spisie na pytania dotyczące niepełnosprawności.



Rys. 2. Profil demograficzny osób odmawiających udzielenia odpowiedzi na tematy związane z niepełnosprawnością – oceny prawdopodobieństw $\rho_{j,r_j|c}$

Źródło: opracowanie własne.

W pierwszej z wyodrębnionych klas znaleźli się w głównie mężczyźni, mieszkający na wsi, pracujący, w wieku 30–49 lat. W drugiej klasie ukrytej znalazły się przede wszystkim kobiety, bierne zawodowo oraz w wieku 50+. Do trzeciej klasy można z kolei zaliczyć głównie mężczyzn, biernych zawodowo i wieku do 29 lat. Do czwartej klasy należą w znacznej mierze kobiety, mieszkające w mieście, pracujące oraz w wieku 30–49 lat. Wreszcie w klasie piątej znaleźli się przede wszystkim mężczyźni, bierni zawodowo, w wieku 50+.

Z powyżej przeprowadzonej analizy wyłania się zatem obraz osoby odmawiającej udzielenia odpowiedzi na pytania dotyczące niepełnosprawności. Pierwszą kategorię opisuje osoba pracująca, w wieku 30–49 lat. W przypadku mężczyzn jest to

osoba mieszkająca głównie na wsi, a w odniesieniu do kobiet w mieście. Drugą kategorię stanowią z kolei osoby bierne zawodowo, w wieku 50+ (bez względu na płeć), bądź poniżej 29 lat (przede wszystkim mężczyźni), mieszkające zarówno w mieście, jak i na wsi.

5. Podsumowanie

W artykule przedstawiono możliwości wykorzystania analizy klas ukrytych w spisach powszechnych w kontekście ważnego zjawiska społecznego, jakim jest niepełnosprawność. Egzemplifikacja omawianej metody stanowi jedno z pierwszych jej zastosowań dla danych spisowych w Polsce. Technika ta okazała się niezwykle istotna z punktu widzenia stworzenia swego rodzaju profilu osoby niepełnosprawnej. Zjawisko niepełnosprawności biologicznej odnosi się głównie do osób starszych, w wieku 50+ oraz biernych zawodowo. Obraz ten wpisuje się niejako w dostrzegalny na całym świecie proces starzenia się społeczeństw i będące jego konsekwencją zjawisko niepełnosprawności.

Analiza klas ukrytych okazała się również przydatna w badaniu zjawiska odmów w kontekście niepełnosprawności, które w NSP 2011 roku obserwowane było na bardzo dużą skalę. Spora liczba odmów, będąca konsekwencją wrażliwości poruszanej w spisie tematyki, niewątpliwie mogła mieć wpływ na jakość procesu estymacji w odniesieniu do niepełnosprawności. Z tego punktu widzenia niezwykle ważne jest zatem poznanie nie tylko przyczyn tych odmów, ale również grup osób, które odmawiały udzielenia odpowiedzi na pytania dotyczące niepełnosprawności. Jak pokazało przeprowadzone w pracy badanie, analiza klas ukrytych stanowić może swego rodzaju remedium na problem odmów w tym sensie, że dzięki jej wykorzystaniu możliwe jest zdiagnozowanie grup osób, które odmawiają udzielenia odpowiedzi.

Literatura

Anderson T.W., 1959, *Some scaling methods and estimation procedures in the Latent Class Model*, [w:] *Probability and Statistics*, ed. U. Grenander, John Wiley & Sons, New York.

Brzezińska J., 2015, *Analiza logarytmiczno-liniowa. Teoria i zastosowania z wykorzystaniem programu R*, C.H. Beck, Warszawa.

Clogg C.C., 1988, *Latent Class Models for Measuring*, Langeheine and Rost.

Collins L., Lanza S., 2010, *Latent Class and Latent Transition Analysis: With Applications in the Social, Behavioral, and Health Sciences*, John Wiley & Sons.

Goodman L.A., 1974, *Exploratory latent structure analysis using both identifiable and unidentifiable models*, Biometrika, vol. 61, iss. 2, s. 215–231.

GUS, 2012, *Raport z wyników. Narodowy Spis Powszechny Ludności i Mieszkań 2011*, Główny Urząd Statystyczny, Warszawa.

Heinen T., 1996, *Latent Class and Discrete Latent Trait Models*, Sage, Thousand Oaks, CA.

Lazarsfeld P.F., 1950, *The interpretation and mathematical foundation of Latent Class Structure Analysis*, [w:] *Measurement and Prediction*, ed. S. Souffer, Princeton University Press, Princeton, NJ.

Lazarsfeld P.F., 1959, *Latent Structure Analysis*, [w:] *Psychology: A Study of a Science*, vol. 3, ed. S. Koch, McGraw-Hill, New York.

Linzer D.A., Lewis J.B., 2011, *poLCA: An R Package for Polytomous Variable Latent Class Analysis*, Journal of Statistical Software, vol. 42, iss. 10.

Sikorska I., 2012, *Analiza zmiennych ukrytych*, [w:] *Zaawansowane metody analiz statystycznych*, red. E. Frątczak, Oficyna Wydawnicza, Szkoła Główna Handlowa w Warszawie.

UN, 2015, *The Millenium Development Goals Report 2015*, United Nations, New York.

WHO, 2011, *World Report on Disability*, World Health Organization, Malta.