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

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

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

Tomasz Józefowski
Urząd Statystyczny w Poznaniu
e-mail: t.jozefowski@stat.gov.pl

ANALIZA KOSZYKOWA W BADANIU ZJAWISKA NIEPEŁNOSPRAWNOŚCI BIOLOGICZNEJ[1]

APPLICATION OF MARKET BASKET ANALYSIS IN RESEARCH ON BIOLOGICAL DISABILITY

DOI: 10.15611/pn.2018.507.09
JEL Classification: C1, C38

Streszczenie: W badaniach prowadzonych przez Główny Urząd Statystyczny analiza koszykowa, która raczej była wykorzystywana w badaniach marketingowych, nie była do tej pory stosowana. Odchodząc od tradycyjnego rozumienia „koszyka zakupów" i zastępując konkretne produkty wariantami określonych cech społeczno-demograficznych osób czy gospodarstw domowych, można na danych z badań reprezentacyjnych czy spisów dokonać jej implementacji celem poszukiwania odpowiednich reguł i zależności. Głównym celem pracy jest wykorzystanie analizy koszykowej na gruncie badań prowadzonych przez statystykę publiczną w Polsce. Autorzy, wykorzystując dane z ostatniego spisu – NSP 2011, programy SAS, R i pakiety statystyczne *arules* oraz *arulesViz*, a także wybrane zmienne z tego badania, wskazują obszary przydatności analizy koszykowej w identyfikacji odpowiednich reguł występujących w tym zbiorze w kontekście zjawiska niepełnosprawności biologicznej.

Słowa kluczowe: analiza koszykowa, Narodowy Spis Powszechny Ludności i Mieszkań 2011, niepełnosprawność biologiczna.

Summary: Commonly used as a tool in marketing studies, market basket analysis (MBA) has not been applied in surveys conducted by the Central Statistical Office so far. If one modifies

[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.

the traditional sense of a "market basket" by replacing specific products with variates of socio-economic variables of persons or households, it is possible to apply association analysis to data collected in surveys or censuses in order to look for association rules and patterns of co-occurrences. The main goal of the article is the application of market basket analysis to official statistical data. The authors demonstrate the usefulness of MBA by using SAS and two R packages (*arules* and *arulesViz*) to identify association rules between variables from the last census of 2011 that are related to the phenomenon of biological disability.

Keywords: market basket analysis, National Census of Population and Housing 2011, biological disability.

1. Wstęp

Analiza koszykowa wywodzi się z analizy danych transakcyjnych pochodzących z sieci dużych sklepów, a jej głównym celem jest identyfikacja współzależności cech oraz wykrycie logicznych reguł wiążących zmienne w zbiorze danych, poprzez wskazanie pozycji, które występują razem. Również sama jej nazwa odnosi się do koszyka zakupów klientów. Z biegiem czasu analiza koszykowa zaczęła odgrywać coraz większą rolę w analizie transakcji finansowych i ubezpieczeniowych, telekomunikacji, logistyce, a także farmaceutyce. Było to konsekwencją odejścia od tradycyjnego rozumienia pojęcia koszyka zakupów i znajdujących się w nim produktów, przykładowo na rzecz usług, taryf czy odwiedzanych na stronach internetowych podstron.

W badaniach prowadzonych przez Główny Urząd Statystyczny analiza koszykowa nie była do tej pory stosowana. Jednak zastępując tradycyjne rozumienie produktów w „koszyku zakupów" wariantami określonych cech społeczno-demograficznych osób czy gospodarstw domowych, można również na danych z badań reprezentacyjnych bądź spisów dokonać implementacji tej metody. Przykładem może być tutaj wykorzystanie analizy koszykowej do poszukiwania reguł występujących w relacjach między tzw. stanem cywilnym faktycznym a stanem cywilnym prawnym [Szymkowiak i in. 2018].

Głównym celem artykułu jest wykorzystanie analizy koszykowej do identyfikacji odpowiednich reguł charakteryzujących wybrane kategorie samooceny stopnia niepełnosprawności biologicznej na podstawie danych z ostatniego Narodowego Spisu Powszechnego Ludności i Mieszkań (NSP 2011). Warto w tym miejscu zauważyć, że w literaturze przedmiotu można znaleźć opracowania, w których podejmowana była próba stworzenia profilu demograficznego osób niepełnosprawnych oraz czynników wpływających na występowanie tego zjawiska. Przykładowo, w pracy [Kukulak-Dolata, Poliwczak 2015] omówione zostały determinanty społeczno-demograficzne różnicujące w największym stopniu poziom aktywności zawodowej osób niepełnosprawnych na polskim rynku pracy. Z kolei w raporcie [Bartkowski 2013] kompleksowo scharakteryzowano zbiorowość osób niepełnosprawnych, wskazując

m.in. wiek jako główny czynnik występowania niepełnosprawności oraz pokazując, że niepełnosprawni w większym stopniu koncentrują się w miastach aniżeli na wsi. Również [Slany 2014] dokonała opisu profilu demograficznego osób niepełnosprawnych w świetle danych z NSP 2011. Należy jednak podkreślić, że we wskazanych powyżej opracowaniach opis zbiorowości osób niepełnosprawnych bazował wyłącznie na wykorzystaniu publikowanych danych tabelarycznych ze spisu czy wybranych badań reprezentacyjnych. Podejście rozważane przez autorów w niniejszym artykule stanowi niestandardowe ujęcie problemu związane z identyfikacją czynników wpływających na ocenę stopnia niepełnosprawności biologicznej z wykorzystaniem metody analizy koszykowej szeroko stosowanej w badaniach marketingowych, a pomijanej w obszarze statystyki publicznej.

Obliczenia wykonano w programie SAS (przygotowanie danych) oraz R (analiza koszykowa), wykorzystując dedykowane pakiety *arules* oraz *arulesViz*.

2. Idea analizy koszykowej

W literaturze przedmiotu analiza koszykowa zaliczana jest do grupy metod dedykowanych wyszukiwaniu wzajemnych powiązań pomiędzy konkretnymi wartościami zmiennych kategorialnych, zazwyczaj w dużych zbiorach danych. Tradycyjnie jej głównym zastosowaniem jest znalezienie odpowiedzi na różnego rodzaju pytania dotyczące tego, jakie produkty są kupowane najczęściej razem, które produkty nie są kupowane w ogóle lub są kupowane rzadko oraz jakie jest prawdopodobieństwo, że klient, który zakupił pewien towar, zakupi również inny. Współcześnie obszar zastosowań analizy koszykowej nie ogranicza się jednak do transakcji w sklepie (por. [Łapczyński 2009]).

Punktem rozważań w analizie koszykowej są reguły skojarzeniowe, inaczej nazywane regułami asocjacyjnymi. Mają one postać A→B , gdzie A i B to zbiory tzw. atrybutów. W regule asocjacyjnej A to poprzednik reguły, a B to następnik reguły asocjacyjnej. Dla reguł asocjacyjnych typu A→B można zdefiniować trzy bardzo ważne miary określane mianem wsparcia, ufności oraz przyrostu. Na gruncie rachunku prawdopodobieństwa miary te można sformułować w następujący sposób:

Wsparcie (*support*) wyraża prawdopodobieństwo łącznego wystąpienia A oraz B i mówi, jaki jest odsetek transakcji w zbiorze wszystkich transakcji, które zawierają A i B:

$$P(A \cap B). \tag{1}$$

Ufność reguły (*confidence*) oznacza prawdopodobieństwo warunkowe P(B|A), tj. prawdopodobieństwo zajścia zdarzenia A pod warunkiem, że zaszło zdarzenie B, i mówi, jaki jest odsetek transakcji zawierających A, które również zawierają B:

$$P(B \mid A). \tag{2}$$

Przyrost (*lift*), będący ilorazem ufności i prawdopodobieństwa B, większy od 1 informuje, że kupno A zwiększa prawdopodobieństwa kupna B:

$$\frac{P(B|A)}{P(B)}. \tag{3}$$

Poszukiwanie reguł asocjacyjnych jest niezwykle czasochłonne ze względu na olbrzymią liczbę transakcji, zmiennych i analizowanych wariantów. W związku z powyższym niezbędne jest zbudowanie odpowiednich algorytmów poszukiwania reguł asocjacyjnych. Kamień milowy w opracowaniu odpowiednich algorytmów asocjacyjnych stanowiły prace Agrawala i Srikanta [1994], którzy zaproponowali dwa nowe algorytmy: Apriori oraz jego rozszerzenie AprioriTID. Algorytm Apriori jest algorytmem iteracyjnym, który w kolejnych krokach działania poszukuje tzw. zbiorów częstych będących reprezentantami zestawu elementów występujących wspólnie w transakcjach. Zakłada się przy tym, że wartość ich wsparcia jest nie mniejsza od pewnego minimalnego progu wsparcia. Następnie w oparciu o zbiory częste tworzy się reguły asocjacyjne, dla których współczynnik ufności jest większy lub równy od zadanej dla niego wartości progowej. Algorytm ten został zaimplementowany w wielu programach statystycznych, takich jak Statistica, SAS czy R.

3. Opis procedury badawczej

Celem przeprowadzonego postępowania badawczego, zaprezentowanego w niniejszym artykule, było sprawdzenie możliwości zastosowania analizy koszykowej na podstawie danych dotyczących niepełnosprawności biologicznej, a pochodzących z NSP 2011. Należy w tym miejscu wspomnieć, że definicja niepełnosprawności przyjęta w statystyce publicznej na potrzeby spisów powszechnych jest odzwierciedleniem postulatów zawartych w dwóch modelach niepełnosprawności – medycznym i społecznym (por. [Dehnel, Klimanek 2016]). Stąd jako niepełnosprawną traktuje się osobę, która posiada odpowiednie orzeczenie wydane przez organ do tego uprawniony (osoby niepełnosprawne prawnie), lub osobę, która takiego orzeczenia nie posiada, lecz odczuwa ograniczenie sprawności w wykonywaniu czynności podstawowych dla swojego wieku (osoby niepełnosprawne wyłącznie biologicznie). Ponadto ważnym uwarunkowaniem badawczym, który należy uwzględniać w analizach niepełnosprawności na podstawie NSP 2011, był fakt, że udzielanie odpowiedzi na pytania dotyczące niepełnosprawności odbywało się na zasadzie dobrowolności. Decyzja ta była podjęta ze względu na specyfikę i wrażliwość tematu [Slany 2014]. Wspomniana dobrowolność odpowiadania na pytania dotyczące niepełnosprawności skutkowała tym, że ponad 1,3 mln respondentów odmówiło udzielenia odpowiedzi na wszystkie pytania z tego obszaru. Można przypuszczać, że w tej zbiorowości znalazły się też osoby niepełnosprawne.

Sformułowano także następującą hipotezę badawczą: „Rozkład przestrzenny charakterystyk opisujących wykryte reguły asocjacyjne w układzie województw nie

jest równomierny, co może wskazywać na ważną rolę czynnika terytorialnego mającego wpływ na kształtowanie się związków między wybranymi wariantami samooceny niepełnosprawności biologicznej a uwzględnionymi zmiennymi społeczno-demograficznymi".

W celu weryfikacji powyższej hipotezy i osiągnięcia postawionego celu badawczego, na etapie przygotowania danych ograniczono populację do osób w wieku 15 lat i więcej. Ponadto ze względu na fakt, że badanie to po raz pierwszy w historii polskiej statystyki publicznej zostało przeprowadzone metodą mieszaną, tzn. wykorzystano dane pochodzące ze źródeł administracyjnych, a także dane zbierane od ludności w ramach przeprowadzonego na szeroką skalę badania reprezentacyjnego, zbiór ograniczono do tych respondentów, dla których można było określić jednoznacznie wszystkie kategorie analizowanych zmiennych (tj. bez pozycyjnych braków danych). Ostatecznie zbiór liczył 6 644 463 osoby i oprócz zmiennej, która była głównym przedmiotem zainteresowań badawczych, tj. samooceny stopnia niepełnosprawności[2], zawierał następujące charakterystyki osób:

- płeć (mężczyzna, kobieta),
- wiek (15–29 lat, 30–49 lat, 50 lat i więcej),
- miejsce faktycznego zamieszkania (miasto, wieś),
- wykształcenie (wyższe lub policealne, średnie, zasadnicze zawodowe, gimnazjalne lub podstawowe lub bez wykształcenia),
- aktywność ekonomiczna (pracujący, bezrobotny, bierny zawodowo).

Kolejnym krokiem była konwersja danych do tzw. zapisu macierzowego formatu danych, w którym poszczególne warianty zmiennych przekodowane są do postaci zmiennych binarnych (0/1), a ponadto na początku zbioru wprowadza się dodatkową zmienną oznaczającą numer transakcji (w naszym przypadku jest to numer osoby w zbiorze)[3]. Ten etap przetwarzania został wykonany za pomocą odpowiedniego kodu w języku 4GL w oprogramowaniu SAS. Fragment tak przygotowanego zbioru przedstawiony został poniżej (tab. 1).

---

[2] Warianty odpowiedzi w pytaniu spisowym: „Czy z powodu problemów zdrowotnych (kalectwa lub choroby przewlekłej) ma Pan(i) ograniczoną zdolność wykonywania zwykłych czynności (nauka w szkole, praca zawodowa, prowadzenie gospodarstwa domowego, samoobsługa) trwającą 6 miesięcy lub dłużej?", były następujące: tak, całkowicie ograniczoną; tak, poważnie ograniczoną; tak, umiarkowanie ograniczoną; nie, nie mam żadnych ograniczeń; nie chcę odpowiadać na to pytanie; nieustalona.

[3] Na przykład MAN to zmienna powstała z konwersji zmiennej płeć. Ma ona wartość 1 w przypadku kiedy osoba jest mężczyzną i 0 w przypadku gdy jest kobietą. Podobnie utworzone zostały pozostałe zmienne (z wyjątkiem zmiennej TRANS, oznaczającej numer transakcji – osoby). W przypadku zmiennej, która była przedmiotem analizy, tj. samooceny stopnia niepełnosprawności, utworzono trzy zmienne binarne: SAMOOC_CSM, która przyjmowała wartość 1 w przypadku kiedy osoba deklarowała całkowicie, poważnie lub umiarkowanie ograniczoną zdolność do wykonywania zwykłych czynności i 0 w pozostałych przypadkach, SAMOOC_NOLIM, która przyjmowała wartość 1 w przypadku kiedy osoba deklarowała brak ograniczeń i 0 w pozostałych przypadkach, SAMOOC_ODM, która przyjmowała wartość 1 w przypadku kiedy osoba nie chciała odpowiadać na to pytanie spisowe i 0 w pozostałych przypadkach.

Tabela 1. Macierzowy format danych w analizie koszykowej dla danych z NSP 2011

TRANS	MAN	WOMAN	WIEK1	WIEK2	WIEK3	CITY	RURAL	…
1	1	0	1	0	0	1	0	…
2	0	1	0	0	1	1	0	…
3	0	1	0	1	0	0	1	…
…	…	…	…	…	…	…	…	…
6644463	0	1	0	0	1	1	0	…

Źródło: opracowanie własne.

Przykładową składnię, która w kodzie R służy wyszukiwaniu reguł asocjacyjnych, związanych ze zmienną SAMOOC_NOLIM, przedstawia poniższy fragment:

```
asocjacje2 <- apriori(trans,
parameter = list(support=0.001,conf=0.05,minlen=5),
appearance = list(rhs=c('SAMOOC_NOLIM'),
                  lhs=c('MAN','WOMAN', 'WIEK1','WIEK2','WIEK3','RURAL',
                        'CITY','EDU_HIGH','EDU_MED','EDU_VOC','EDU_PRIM',
                        'WORK','UNEMPL','INACT')))
```

Pewnego wyjaśnienia wymagają ustawienia dla miar wsparcia (support=0.001) oraz ufności (conf=0.05). Domyślne ustawienia wynoszą odpowiednio 0,1 dla wsparcia oraz 0,8 dla ufności. Przyjęte w powyższym przykładzie wartości wynikają ze specyfiki analizy koszykowej. Podczas gdy w przypadku klasycznego wykorzystania tej techniki poszukuje się reguł częstych, inna musi być optyka w przypadku poszukiwania reguł dla zjawisk stosunkowo rzadkich bądź takich, gdzie w roli poprzednika występuje lista z dużą liczbą kategorii. Przyjęcie domyślnych, czyli wysokich wskaźników wsparcia i ufności prowadziłoby do tego, że reguły takie nie byłyby znajdowane. Przyjęcie parametru „minlen=5" ma zapobiegać tworzeniu się pustych reguł lub trywialnych reguł, na przykład postaci: {}=>SAMOOC_NOLIM, co miałoby miejsce w przypadku zastosowania domyślnego ustawienia „minlen=1". Parametry „lhs"i „rhs" mają natomiast służyć wskazaniu zmiennych binarnych, bądź ich kombinacji, które mają się pojawiać odpowiednio po lewej oraz prawej stronie reguły asocjacyjnej. Wybrane wyniki detekcji reguł asocjacji wraz z ich przestrzennym zróżnicowaniem są przedstawione w następnym punkcie.

4. Wyniki empiryczne

Spośród wykrytych w analizowanym zbiorze reguł, dotyczących zmiennych: SAMOOC_CSM, SAMOOC_NOLIM oraz SAMOOC_ODM, dalszej pogłębionej analizie poddano wymienione poniżej reguły (tab. 2). Przy wyborze tych reguł kierowano się kilkoma kryteriami rozpatrywanymi łącznie: wysoki poziom wszystkich

trzech miar stosowanych w analizie koszykowej: wsparcia, ufności oraz przyrostu na poziomie ogólnopolskim, ale także dążenie do jak największej liczby „produktów w koszyku", tzn. czynników wpływających na zmienne SAMOOC_CSM, SAMOOC_NOLIM, SAMOOC_ODMOWA .

Tabela 2. Postać reguły asocjacyjnej oraz jej charakterystyki na poziomie ogólnopolskim[4]

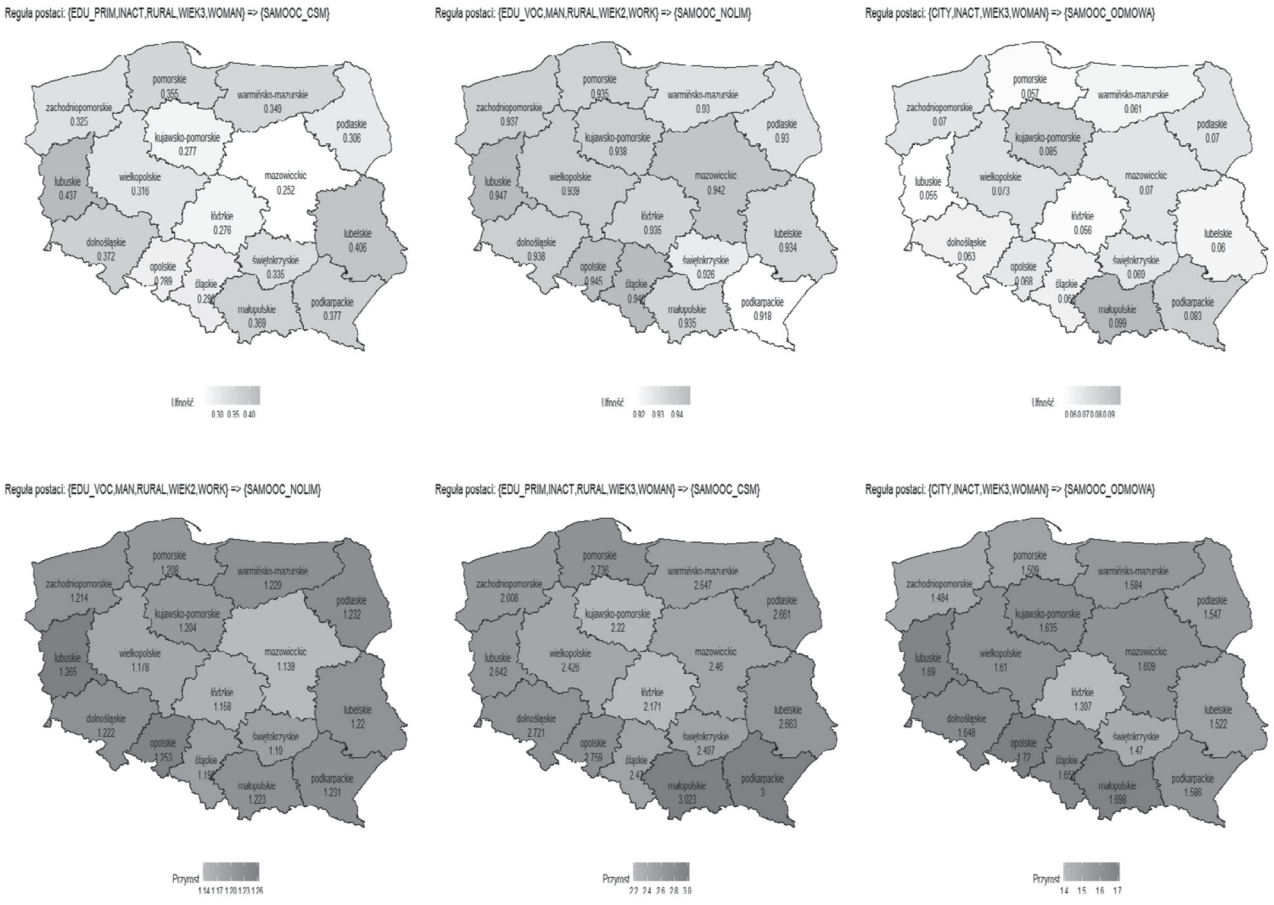
Reguła	Wsparcie	Ufność	Przyrost
{EDU_PRIM,INACT,RURAL,WIEK3,WOMAN} => {SAMOOC_CSM}	0,0162	0,3268	2.5952
{EDU_VOC,MAN,RURAL,WIEK2,WORK} => {SAMOOC_NOLIM}	0,0292	0,9360	1,1984
{CITY,INACT,WIEK3,WOMAN} => {SAMOOC_ODMOWA}	0,0066	0,0684	1,5632

Źródło: opracowanie własne.

Można zauważyć, że w grupie biernych zawodowo kobiet w wieku 50 lat i więcej, mieszkających na wsi i charakteryzujących się najniższym poziomem wykształcenia prawie 1/3 stanowią osoby niepełnosprawne biologicznie, oceniające swoją niepełnosprawność jako zupełną, poważną lub przynajmniej umiarkowaną. Co więcej, przynależność do tak określonej subpopulacji osób prawie 2,6-krotnie zwiększa prawdopodobieństwo własnej oceny stopnia niepełnosprawności biologicznej jako całkowite, poważne lub umiarkowane ograniczenie powszednich czynności. Z kolei przynależność do grupy pracujących mężczyzn w wieku 30–49 lat, mieszkających na wsi i legitymujących się wykształceniem zasadniczym w ponad 90% wiąże się z brakiem jakichkolwiek ograniczeń powszednich czynności. Jednocześnie posiadanie przez jednostki populacji takich charakterystyk nie zwiększa znacząco prawdopodobieństwa samooceny niepełnosprawności biologicznej jako „braku ograniczeń w wykonywaniu zwykłych czynności". Ostatnia z reguł wskazuje, że fakt bycia bierną zawodowo kobietą mieszkającą w mieście, w wieku 50 lat i więcej, zwiększa o nieco ponad połowę prawdopodobieństwo odmowy odpowiedzi na pytanie dotyczące samooceny stopnia niepełnosprawności biologicznej.

Na poniższych kartogramach (rys. 1) zamieszczono przestrzenne zróżnicowanie ufności (góra) oraz przyrostu (dół) analizowanych reguł asocjacyjnych. Przyjęto przy tym zasadę, że w ramach danego kartogramu intensywniejsza (ciemniejsza) barwa oznacza wyższe wartości odpowiednio: ufności i przyrostu.

[4] Użyte symbole oznaczają odpowiednio: SAMOOC_CSM – osoba deklarująca całkowicie, poważnie lub umiarkowanie ograniczoną zdolność do wykonywania zwykłych czynności, SAMOOC_NOLIM – osoba deklarująca brak ograniczeń w wykonywaniu zwykłych czynności, SAMOOC_ODMOWA – osoba odmawiająca odpowiedzi na pytanie spisowe stopnia samooceny niepełnosprawności biologicznej, MAN – mężczyzna, WOMAN – kobieta, CITY – osoba mieszkająca w mieście, RURAL – osoba mieszkająca na wsi, WORK – osoba pracująca, INACT – osoba bierna zawodowo, EDU_PRIM – osoba o wykształceniu gimnazjalnym lub podstawowym lub bez wykształcenia, EDU_VOC – osoba o wykształceniu zasadniczym zawodowym.



Rys. 1. Przestrzenny rozkład ufności i przyrostu dla wybranych reguł asocjacyjnych

Źródło: opracowanie własne.

W stosunku do wszystkich trzech zaprezentowanych reguł asocjacyjnych można zauważyć wpływ czynnika przestrzennego na kształtowanie się ich charakterystyk, przy czym, jak się wydaje, w większym stopniu dotyczy to przyrostu niż ufności. W przypadku reguły o postaci {EDU_PRIM,INACT,RURAL,WIEK3,WOMAN} => {SA-MOOC_CSM} największą ufnością charakteryzują się województwa zachodnie (lubuskie i dolnośląskie) oraz województwa południowo-wschodnie (lubelskie, podkarpackie oraz małopolskie) – od 0,37 do 0,44. Natomiast w przypadku województw w środkowej części kraju miernik ten nie przekracza 0,28. Oznacza to w tych województwach dużo niższe prawdopodobieństwo warunkowe wystąpienia zupełnej, znacznej lub umiarkowanej samooceny niepełnosprawności biologicznej przy założeniu, że respondentem jest bierna zawodowo kobieta w wieku przekraczającym 50 lat, mieszkająca na wsi, o niskim poziomie wykształcenia. Natomiast analiza przyrostu wskazuje, że we wszystkich województwach są to charakterystyki, które znacznie zwiększają prawdopodobieństwo takiej odpowiedzi, w stosunku do prawdopodobieństwa bezwarunkowego. Reguła o postaci {EDU_VOC,MAN,RURAL,WIEK2,WORK} => {SAMOOC_NOLIM} charakteryzuje się dużą ufnością we wszystkich województwach – od 0,92 do 0,95. Oznacza to, że warunkowe prawdopodobieństwo wystąpienia odpowiedzi o braku ograniczeń w wykonywaniu codziennych czynności, przy założeniu, że respondentem jest pracujący mężczyzna w wieku 30–49 lat, mieszkający na wsi, posiadający wykształcenie zasadnicze zawodowe, charakteryzuje się niewielkim zróżnicowaniem przestrzennym. Kartogram prezentujący wartości przyrostów w układzie wojewódzkim dla tej reguły także wskazuje na niewielkie zróżnicowanie – od 1,14 do 1,27, co oznacza niewielki wpływ takich charakterystyk respondenta na udzielenie odpowiedzi o braku ograniczeń w wykonywaniu codziennych czynności.

Ostatnia para kartogramów dotyczy przypadku odmowy odpowiedzi na pytanie o samoocenę niepełnosprawności. Ufność analizowanej w tym przypadku reguły postaci: {CITY,INACT,WIEK3,WOMAN} => {SAMOOC_ODMOWA} w układzie wojewódzkim waha się od 0,05 do 0,10, osiągając najwyższą wartość w województwie małopolskim, a najniższą w województwie lubuskim. Charakterystyki przyrostu dla tej reguły we wszystkich województwach osiągają wartości przekraczające 1 i wahają się od 1,40 w województwie łódzkim, do 1,72 w województwie opolskim. Oznacza to, że prawdopodobieństwo odmowy na pytanie dotyczące samooceny stopnia niepełnosprawności znacznie zwiększa się w przypadku respondentów, którymi będą bierne zawodowo kobiety w wieku powyżej 50 lat i mieszkające w miastach. Widoczny jest przy tym nieco silniejszy wpływ tych charakterystyk na fakt odmowy odpowiedzi na pytanie kwestionariusza spisowego dotyczącego samooceny stopnia niepełnosprawności w przypadku osób mieszkających w województwach południowo-zachodniej Polski w porównaniu z osobami mieszkającymi w pozostałych województwach.

5. Zakończenie

Zdaniem autorów artykułu udało się osiągnąć postawiony cel badawczy, tzn. potwierdzić przydatność analizy koszykowej jako metody statystycznej służącej do identyfikacji reguł opisujących związki między samooceną stopnia niepełnosprawności biologicznej a zestawem wybranych zmiennych społeczno-demograficznych w dużych zbiorach danych. Ponadto daje się wyraźnie zauważyć przestrzenne wzorce reguł asocjacyjnych między wybranymi wariantami odpowiedzi na pytanie dotyczące samooceny stopnia niepełnosprawności a charakterystyką osób ze względu na płeć, wiek, wykształcenie, status na rynku pracy i miejsce zamieszkania. Były one analizowane w układzie wojewódzkim, który niekoniecznie musi pokrywać się z obszarami odmiennych postaw odnośnie postrzegania niepełnosprawności.

Warto podkreślić, że wykorzystanie metody analizy koszykowej w takich zastosowaniach, gdzie celem jest detekcja reguł asocjacyjnych w małych subpopulacjach, wymaga sformułowania zupełnie innej hierarchii miar określanych jako: wsparcie, ufność i przyrost. W klasycznych zastosowaniach analizy koszykowej dużo większe znaczenie przypisuje się wysokim wartościom miernika określanego jako wsparcie. Wynika to oczywiście z dążenia do wygenerowania jak najwyższego zysku, czyli detekcji reguł najczęstszych. Natomiast te zastosowania, które dotyczą wykrywania reguł za pomocą analizy koszykowej w specyficznych, nielicznych subpopulacjach (osoby odczuwające ograniczenie związane z niepełnosprawnością czy odmawiające odpowiedzi na pytanie dotyczące niepełnosprawności), wymagają większego zwrócenia uwagi na ufność, a przede wszystkim przyrost. Należy także zwrócić uwagę na potencjalne obszary wykorzystania analizy koszykowej w statystyce publicznej w prezentacji wyników badań. Może się ona stać etapem preselekcji wyjściowego, licznego zestawu tablic statystycznych do tych układów, które będą odzwierciedlały zidentyfikowane, o dużej wartości poznawczej, reguły opisujące związki między analizowanymi zmiennymi.

Literatura

Agrawal R., Srikant R., 1994, *Fast Algorithms for Mining Association Rules*, Proceedings of the 20th VLDB Conference Santiago, Chile.

Bartkowski J., 2013, *Położenie społeczne i ekonomiczne zbiorowości osób niepełnosprawnych w Polsce na podstawie demograficznych danych zastanych i ustaleń badań przeprowadzonych w ostatnich pięciu latach*, Raport przygotowany w ramach projektu „Od kompleksowej diagnozy sytuacji osób niepełnosprawnych w Polsce do nowego modelu polityki społecznej wobec niepełnosprawności", AGH Kraków, manuskrypt.

Dehnel G., Klimanek T., 2016, *Disability in the National Censuses of 2002 and 2011 – a comparison of information scope*, Acta Universitatis Lodziensis. Folia Oeconomica, vol. 5(325), s. 127–141.

GUS, Kwestionariusz badania reprezentacyjnego w ramach NSP 2011, https://stat.gov.pl/download/gfx/portalinformacyjny/pl/defaultstronaopisowa/5781/1/1/nsp_2011_badanie_reprez_ludnosc_wykaz_pytan.pdf (3.11.2017).

Kukulak-Dolata I., Poliwczak I., 2015, *Profile społeczno-demograficzne osób niepełnosprawnych a ich aktywność zawodowa*, Polityka Społeczna, nr 3, s. 5–12.

Łapczyński M., 2009, *Analiza koszykowa i analiza sekwencji – wielki brat czuwa*, StatSotf Polska.

Package 'arules', 2017, https://cran.r-project.org/web/packages/arules/arules.pdf (9.02.2017), dokumentacja pakietu program R.

Package 'arulesViz', 2017, https://cran.r-project.org/web/packages/arulesViz/arulesViz.pdf (9.02.2017), dokumentacja pakietu program R.

Slany K., 2014, *Osoby niepełnosprawne w świetle Narodowego Spisu Powszechnego Ludności i Mieszkań z 2011 r. – wybrane aspekty*, Niepełnosprawność – Zagadnienia, Problemy, Rozwiązania, nr II/2014(11), Uniwersytet Jagielloński w Krakowie, AGH w Krakowie.

Szymkowiak M., Klimanek T., Józefowski T., 2018, *Zastosowanie analizy koszykowej w wybranych badaniach statystyki publicznej*, referat wygłoszony podczas XXI Warsztatów Metodologicznych im. Prof. S. Mynarskiego, 26.05.2017 r. (w druku).