

PRACE NAUKOWE

Uniwersytetu Ekonomicznego we Wrocławiu

RESEARCH PAPERS

of Wrocław University of Economics

Nr 327

Taksonomia 22

**Klasyfikacja i analiza danych –
teoria i zastosowania**

Redaktorzy naukowci

Krzysztof Jajuga, Marek Walesiak



Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2014

Redaktor Wydawnictwa: Barbara Majewska

Redaktor techniczny: Barbara Łopusiewicz

Korektor: Barbara Cibis

Łamanie: Beata Mazur

Projekt okładki: Beata Dębska

Publikacja jest dostępna w Internecie na stronach:

www.ibuk.pl, www.ebscohost.com,

w Dolnośląskiej Bibliotece Cyfrowej www.dbc.wroc.pl,

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

Kopiowanie i powielanie w jakiegokolwiek formie

wymaga pisemnej zgody Wydawcy

© Copyright by Uniwersytet Ekonomiczny we Wrocławiu

Wrocław 2014

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

ISSN 1505-9332 (Taksonomia)

Wersja pierwotna: publikacja drukowana

Druk: Drukarnia TOTEM

Spis treści

Wstęp	9
Eugeniusz Gatnar , Balance of payments statistics and external competitiveness of Poland.....	15
Andrzej Sokolowski, Magdalena Czaja , Efektywność metody k -średnich w zależności od separowalności grup.....	23
Barbara Pawelek, Józef Pocięcha, Adam Sagan , Wielosektorowa analiza ukrytych przejść w modelowaniu zagrożenia upadłością polskich przedsiębiorstw	30
Elżbieta Gołata , Zróżnicowanie procesu starzenia i struktur demograficznych w Poznaniu i aglomeracji poznańskiej na tle wybranych dużych miast Polski w latach 2002-2011.....	39
Aleksandra Łuczak, Feliks Wysocki , Ustalanie systemu wag dla cech w zagadnieniach porządkowania liniowego obiektów	49
Marek Walesiak , Wzmacnianie skali pomiaru dla danych porządkowych w statystycznej analizie wielowymiarowej	60
Paweł Lula , Identyfikacja słów i fraz kluczowych w tekstach polskojęzycznych za pomocą algorytmu <i>RAKE</i>	69
Mariusz Kubus , Propozycja modyfikacji metody złagodzonego LASSO.....	77
Andrzej Bąk, Tomasz Bartłomowicz , Wielomianowe modele logitowe wyborów dyskretnych i ich implementacja w pakiecie <i>DiscreteChoice</i> programu R.....	85
Justyna Brzezińska , Wykorzystanie modeli logarytmiczno-liniowych do analizy bezrobocia w Polsce w latach 2004-2012.....	95
Andrzej Bąk, Marcin Pelka, Aneta Rybicka , Zastosowanie pakietu <i>dcMNM</i> programu R w badaniach preferencji konsumentów wódki	104
Barbara Batóg, Jacek Batóg , Analiza stabilności klasyfikacji polskich województw według sektorowej wydajności pracy w latach 2002-2010	113
Małgorzata Markowska, Danuta Strahl , Klasyfikacja europejskiej przestrzeni regionalnej ze względu na filary inteligentnego rozwoju z wykorzystaniem referencyjnego systemu granicznego.....	121
Kamila Migdał-Najman, Krzysztof Najman , Formalna ocena jakości odwzorowania struktury grupowej na mapie Kohonena	131
Kamila Migdał-Najman, Krzysztof Najman , Graficzna ocena jakości odwzorowania struktury grupowej na mapie Kohonena	139
Beata Basiura, Anna Czapkiewicz , Badanie jakości klasyfikacji szeregow czasowych	148
Michał Trzęsiok , Wybrane metody identyfikacji obserwacji oddalonych.....	157

Grażyna Dehnel, Tomasz Klimanek , Taksonomiczne aspekty estymacji pośredniej uwzględniającej autokorelację przestrzenną w statystyce gospodarczej.....	167
Michał Bernard Pietrzak, Justyna Wilk , Odległość ekonomiczna w modelowaniu zjawisk przestrzennych z wykorzystaniem modelu grawitacji.....	177
Maciej Beręsewicz , Próba zastosowania różnych miar odległości w uogólnionym estymatorze Petersena.....	186
Marcin Szymkowiak, Tomasz Józefowski , Konstrukcja i praktyczne wykorzystanie estymatorów typu SPREE na przykładzie dwuwymiarowych tabel kontyngencji.....	195
Marcin Pelka , Klasyfikacja pojęciowa danych symbolicznych w podejściu wielomodelowym.....	202
Małgorzata Machowska-Szewczyk , Ocena klas w rozmytej klasyfikacji obiektów symbolicznych.....	210
Justyna Wilk , Problem wyboru liczby klas w taksonomicznej analizie danych symbolicznych.....	220
Andrzej Dudek , Metody analizy skupień w klasyfikacji markerów map Google.....	229
Ewa Roszkowska , Ocena ofert negocjacyjnych w słabo ustrukturyzowanych problemach negocjacyjnych z wykorzystaniem rozmytej procedury SAW.....	237
Marcin Szymkowiak, Marek Witkowski , Zastosowanie analizy korespondencji do badania kondycji finansowej banków spółdzielczych.....	248
Bartłomiej Jefmański , Budowa rozmytych indeksów satysfakcji klientów z zastosowaniem programu R.....	257
Karolina Bartos , Odkrywanie wzorców zachowań konsumentów za pomocą analizy koszykowej danych transakcyjnych.....	266
Joanna Trzęsiok , Taksonomiczna analiza krajów pod względem dzietności kobiet oraz innych czynników demograficznych.....	275
Beata Bal-Domańska , Próba identyfikacji większych skupisk regionalnych oraz ich konwergencja.....	285
Beata Bieszk-Stolorz, Iwona Markowicz , Wpływ zasiłku na proces poszukiwania pracy.....	294
Marta Dziechciarz-Duda, Klaudia Przybysz , Wykształcenie a potrzeby rynku pracy. Klasyfikacja absolwentów wyższych uczelni.....	303
Tomasz Klimanek , Problem pomiaru procesu dezagrarnizacji wsi polskiej w świetle wielowymiarowych metod statystycznych.....	313
Małgorzata Sej-Kolasa, Mirosława Sztemberg-Lewandowska , Wybrane metody analizy danych wzdluznych.....	321
Artur Zaborski , Zastosowanie miar odległości dla danych porządkowych do agregacji preferencji indywidualnych.....	330
Mariola Chrzanowska, Nina Drejerska, Iwona Pomianek , Zastosowanie analizy korespondencji do badania sytuacji mieszkańców strefy podmiejskiej Warszawy na rynku pracy.....	338

Katarzyna Wawrzyniak , Klasyfikacja województw według stopnia realizacji priorytetów Strategii Rozwoju Kraju 2007-2015 z wykorzystaniem wartości centrum wierszowego	346
---	-----

Summaries

Eugeniusz Gatnar , Statystyka bilansu płatniczego a konkurencyjność gospodarki Polski	22
Andrzej Sokółowski, Magdalena Czaja , Cluster separability and the effectiveness of k -means method	29
Barbara Pawelek, Józef Pocięcha, Adam Sagan , Multisectoral analysis of latent transitions in bankruptcy prediction models.....	38
Elżbieta Golata , Differences in the process of aging and demographic structures in Poznań and the agglomeration compared to selected Polish cities in the years 2002-2011	48
Aleksandra Łuczak, Feliks Wysocki , Determination of weights for features in problems of linear ordering of objects	59
Marek Walesiak , Reinforcing measurement scale for ordinal data in multivariate statistical analysis	68
Paweł Lula , Automatic identification of keywords and keyphrases in documents written in Polish.....	76
Mariusz Kubus , The proposition of modification of the relaxed LASSO method.....	84
Andrzej Bąk, Tomasz Bartłomowicz , Microeconomic multinomial logit models and their implementation in the <code>DiscreteChoice</code> R package .	94
Justyna Brzezińska , The analysis of unemployment data in Poland in 2004-2012 with application of log-linear models	103
Andrzej Bąk, Marcin Pelka, Aneta Rybicka , Application of the MMLM package of R software for vodka consumers preference analysis.....	112
Barbara Batóg, Jacek Batóg , Analysis of the stability of classification of Polish voivodeships in 2002-2010 according to the sectoral labour productivity	120
Małgorzata Markowska, Danuta Strahl , Classification of the European regional space in terms of smart growth pillars using the reference limit system.....	130
Kamila Migdał Najman, Krzysztof Najman , Formal quality assessment of group structure mapping on the Kohonen's map	138
Kamila Migdał Najman, Krzysztof Najman , Graphical quality assessment of group structure mapping on the Kohonen's map	147
Beata Basiura, Anna Czapkiewicz , Validation of time series clustering	156
Michał Trzęsiok , Selected methods for outlier detection.....	166

Grażyna Dehnel, Tomasz Klimanek , Taxonomic aspects of indirect estimation accounting for spatial correlation in enterprise statistics	176
Michał Bernard Pietrzak, Justyna Wilk , Economic distance in modeling spatial phenomena with the application of gravity model.....	185
Maciej Beręsewicz , An attempt to use different distance measures in the Generalized Petersen estimator	194
Marcin Szymkowiak, Tomasz Józefowski , Construction and practical using of SPREE estimators for two-dimensional contingency tables.....	201
Marcin Pelka , The ensemble conceptual clustering for symbolic data.....	209
Małgorzata Machowska-Szewczyk , Evaluation of clusters obtained by fuzzy classification methods for symbolic objects.....	219
Justyna Wilk , Problem of determining the number of clusters in taxonomic analysis of symbolic data	228
Andrzej Dudek , Clustering techniques for Google maps markers.....	236
Ewa Roszkowska , The evaluation of negotiation offers in ill structure negotiation problems with the application of fuzzy SAW procedure	247
Marcin Szymkowiak, Marek Witkowski , The use of correspondence analysis in analysing the financial situation of cooperative banks.....	256
Bartłomiej Jefmański , The construction of fuzzy customer satisfaction indexes using R program.....	265
Karolina Bartos , Discovering patterns of consumer behaviour by market basket analysis of the transactional data.....	274
Joanna Trzęsiok , Cluster analysis of countries with respect to fertility rate and other demographic factors	284
Beata Bal-Domańska , An attempt to identify major regional clusters and their convergence	293
Beata Bieszk-Stolorz, Iwona Markowicz , The influence of benefit on the job finding process	302
Marta Dziechciarz-Duda, Klaudia Przybysz , Education and labor market needs. Classification of university graduates	312
Tomasz Klimanek , The problem of measuring deagrarianisation process in rural areas in Poland using multivariate statistical methods.....	320
Małgorzata Sej-Kolasa, Mirosława Sztemberg-Lewandowska , Selected methods for an analysis of longitudinal data.....	329
Artur Zaborski , The application of distance measures for ordinal data for aggregation individual preferences	337
Mariola Chrzanowska, Nina Drejerska, Iwona Pomianek , Application of correspondence analysis to examine the situation of the inhabitants of Warsaw suburban area in the labour market	345
Katarzyna Wawrzyniak , Classification of voivodeships according to the level of the realization of priorities of <i>the National Development Strategy 2007-2015</i> with using the values of centroid of the rows	355

Marcin Szymkowiak

Uniwersytet Ekonomiczny w Poznaniu

Tomasz Józefowski

Urząd Statystyczny w Poznaniu, Ośrodek Statystyki Małych Obszarów

KONSTRUKCJA I PRAKTYCZNE WYKORZYSTANIE ESTYMATORÓW TYPU SPREE NA PRZYKŁADZIE DWUWYMIAROWYCH TABEL KONTYNGENCJI

Streszczenie: Głównym celem artykułu jest przedstawienie estymatorów typu SPREE, wykorzystujących technikę iteracyjnego proporcjonalnego dopasowania na przykładzie dwuwymiarowych tabel kontyngencji. Estymatory te wykorzystywane są do korekty wejściowych wartości w tabeli kontyngencji, tak aby odtworzone były znane oszacowania brzegowe z badania reprezentacyjnego. W artykule wskazano również na praktyczne zastosowanie estymacji typu SPREE i algorytmu IPF w dwuwymiarowej tabeli kontyngencji.

Słowa kluczowe: statystyka małych obszarów, estymator SPREE, algorytm iteracyjnego proporcjonalnego dopasowania.

1. Wstęp

W badaniach prowadzonych przez krajowe urzędy statystyczne, w tym Główny Urząd Statystyczny, ze względu na sposób doboru jednostek do próby, wykorzystanie klasycznych metod estymacji pozwala na publikowanie wyników jedynie na dość wysokim poziomie agregacji, na przykład całego kraju czy na poziomie województwa. Odbiorcy danych statystycznych oczekują jednak informacji na niższych poziomach agregacji przestrzennej bądź bardziej szczegółowych domen (na przykład na poziomie podregionów czy powiatów bądź województwa, ale z uwzględnieniem klasy miejscowości zamieszkania). Remedium na rosnące zapotrzebowanie informacyjne oraz na organizację badania jest statystyka małych obszarów, która umożliwia estymację parametrów w sytuacji niewielkiej liczebności próby dla wyróżnionych domen [Rao 2003, s. 3].

Jedną z technik estymacji pośredniej, która może okazać się szczególnie przydatna w tego typu badaniach i która umożliwi uzyskanie wyników na niższych

poziomach agregacji przestrzennej, jest estymacja zachowująca strukturę (tzw. estymatory typu SPREE – *Structure Preserving Estimation*). Jest to technika, która przez odpowiednie połączenie danych z badania reprezentacyjnego z informacjami pochodzącymi z badań pełnych, np. spisów czy spisów opartych na rejestrach, umożliwia uzyskanie szacunków na niższych poziomach szczegółowości z akceptowalną precyzją [Józefowski, Szymkowiak 2014]. Metoda ta jest wykorzystywana na przykład w badaniu siły roboczej w Nowej Zelandii [Haslett, Noble, Zabala 2008, s. 14]. Mogłaby również stanowić cenną technikę estymacji w Badaniu Aktywności Ekonomicznej Ludności prowadzonym przez GUS. Połączenie informacji pochodzących z BAEL z danymi spisowymi umożliwiłoby bieżącą estymację wybranych charakterystyk z zakresu rynku pracy na niższych, aniżeli dotychczas, poziomach agregacji danych.

Głównym celem artykułu jest przedstawienie estymatorów typu SPREE wykorzystujących algorytm iteracyjnego proporcjonalnego dopasowania do dwuwymiarowej tablicy kontyngencji. Rozważania teoretyczne zostaną zilustrowane przykładem zastosowania omawianego algorytmu w badaniach z zakresu rynku pracy z wykorzystaniem programu R.

2. Teoretyczne podstawy estymatorów typu SPREE dla dwuwymiarowej tabeli kontyngencji

Estymatory typu SPREE wykorzystywane są w badaniach, w których zachodzi potrzeba korekty liczebności znajdujących się w komórkach wielowymiarowej tabeli kontyngencji tak, aby skorygowane wartości sumowały się do znanych liczebności brzegowych. Przykładowo liczebności w tabeli kontyngencji mogą pochodzić ze spisu, natomiast liczebności brzegowe odpowiadać będą rzetelnym oszacowaniom uzyskanym w wyniku zastosowania estymatora bezpośredniego i danych pochodzących z badania reprezentacyjnego. Technika ta może być szczególnie przydatna w okresach międzyspisowych. Ze względu na incydentalny charakter spisu dane te dezaktualizują się, mogą jednak stanowić punkt wyjścia do konstrukcji tabeli kontyngencji, w których liczebności brzegowe uzyskiwane są z wykorzystaniem aktualnych danych z badania reprezentacyjnego.

Rozważmy dwuwymiarową tabelę kontyngencji. Niech N_{ij} oznacza znane liczebności w dwuwymiarowej tabeli kontyngencji pochodzące ze spisu, gdzie i oznacza mały obszar (domenę) oraz $i = 1, \dots, D$, a j oznacza j -ty wariant ($j = 1, \dots, J$) zmiennej y , dla której dokonywane są szacunki (na przykład y oznaczać może liczbę bezrobotnych, zatrudnionych itd.). Zakładamy ponadto, że istnieją bieżące oszacowania liczebności brzegowych w oparciu o dane pochodzące z badania reprezentacyjnego. Niech \widehat{M}_i oraz \widehat{M}_j oznaczają „rzetelne” oszacowania liczebności brzegowych M_i oraz M_j , które otrzymujemy, wykorzystując znany z metody reprezentacyjnej estymator bezpośredni wartości globalnej. Tabela 1

w poglądowy sposób przedstawia powyżej opisaną sytuację. Zakładamy, że znane są ze spisu informacje na temat liczby bezrobotnych kobiet i mężczyzn w podregionach (celem uproszczenia przyjęto dwa podregiony). Wartości brzegowe pochodzą natomiast z badania reprezentacyjnego i uzyskano je przy zastosowaniu estymatora bezpośredniego wartości globalnej. Ze względu na fakt, że wartości w wyjściowej tabeli kontyngencji nie sumują się do oszacowanych wartości brzegowych, należy je skorygować celem zachowania zgodności struktur. Problem sprowadza się zatem do poszukania nowych liczebności \hat{N}_{ij} , które nieznacznie różnią się od wartości oryginalnych N_{ij} ze spisu i które sumować się będą do wartości brzegowych uzyskanych z badania reprezentacyjnego.

Tabela 1. Przykładowa struktura danych dla małych obszarów dla dwuwymiarowej tabeli kontyngencji

Podregion	Płeć		\hat{M}_i
	Mężczyzna	Kobieta	
Podregion 1	N_{11}	N_{12}	\hat{M}_1
Podregion 2	N_{21}	N_{22}	\hat{M}_2
\hat{M}_j	\hat{M}_1	\hat{M}_2	

Źródło: opracowanie własne.

Ponieważ nie jest możliwe wyprowadzenie analitycznego wzoru na nowe liczebności \hat{N}_{ij} , stosuje się z tzw. technikę iteracyjnego proporcjonalnego dopasowania (IPF – *Iterative Proportional Fitting*) celem ich znalezienia. \hat{N}_{ij} określa się mianem estymatora typu SPREE liczebności w tabeli kontyngencji. IPF jest metodą, której idea polega na odpowiednim dopasowaniu liczebności w wyjściowej tabeli kontyngencji do wartości brzegowych. Dopasowanie struktur w tabeli kontyngencji odbywa się w sposób iteracyjny. W każdym kolejnym kroku zapewnia się zgodność danych z tabeli kontyngencji z wartościami brzegowymi w wierszach, a następnie w kolumnach. Proces ten jest tak długo powtarzany, aż osiągnie się sumowalność danych z tabeli kontyngencji do wszystkich liczebności brzegowych. Odpowiednie wzory na korygowane wartości \hat{N}_{ij} w tabeli kontyngencji w poszczególnych krokach $n = 1, 2, \dots$ przedstawiają się następująco:

$$\hat{N}_{ij}^{(2n-1)} = \frac{\hat{N}_{ij}^{(2n-2)} N_i}{\sum_{k=1}^J \hat{N}_{ik}^{(2n-2)}}, \quad (1)$$

$$\hat{N}_{ij}^{(2n)} = \frac{\hat{N}_{ij}^{(2n-1)} N_j}{\sum_{k=1}^D \hat{N}_{kj}^{(2n-1)}}, \quad (2)$$

przy czym $N_i = \sum_j N_{ij}$, $N_j = \sum_i N_{ij}$ oraz $\hat{N}_{ij}^{(0)} = N_{ij}$. Algorytm iteracyjnego wyznaczania liczebności \hat{N}_{ij} powtarza się tak długo, aż zachowana zostanie zgodność pomiędzy wartościami w tabeli kontyngencji pochodzącymi ze spisu i oszacowanymi wartościami brzegowymi wyznaczonymi w oparciu o aktualne dane pochodzące z badania reprezentacyjnego, przy czym $\hat{N}_{ij} = \lim_{n \rightarrow \infty} \hat{N}_{ij}^{(n)}$. Poniższy przykład ilustruje sposób działania algorytmu IPF w przypadku dwuwymiarowej tabeli kontyngencji.

Założmy w uproszczeniu, że dysponujemy informacją na temat liczby bezrobotnych w dwóch jednostkach terytorialnych (na przykład w podregionach) w przekroju płci (por. tabela 2). Zakładamy przy tym, że wartości wejściowe (pogrubione) pochodzą ze spisu, a wartości brzegowe stanowią oszacowania liczby bezrobotnych w każdej kategorii płci i w każdym podregionie. Zakładamy przy tym, co jest częstą w praktyce sytuacją, że nie jest możliwe ze względu na małe liczebności próby i w konsekwencji niską precyzję szacunków wyestymowanie liczby bezrobotnych w poszczególnych komórkach tabeli kontyngencji (na przykład bezrobotnych kobiet w podregionie 1). Stąd chcąc opublikować – z wykorzystaniem danych z bieżącego badania reprezentacyjnego – tak szczegółową tablicę na niskim poziomie agregacji przestrzennej, należy skorzystać z informacji z innych źródeł, na przykład z wcześniejszego spisu. Odpowiednie połączenie tych danych i skorzystanie z techniki estymacji typu SPREE, wykorzystującej algorytm IPF, zapewni spójność i zgodność struktur w konstruowanej tabeli kontyngencji. Szczegółowe obliczenia przeprowadzone na podstawie wzorów 1 i 2 zawarte są w tabeli 2 (wejściowa tabela kontyngencji).

W programie R istnieje możliwość przeprowadzenia algorytmu IPF w kilku pakietach. Do najczęściej wykorzystywanych należą pakiety `cat` (funkcja `ipf`) oraz `survey` (funkcja `raking`). Istnieją również dedykowane kody napisane w środowisku R do wyznaczania wartości w tabelach kontyngencji o różnych wymiarach. Przykładem może być kod napisany w Alaska Department of Labor and Workforce Development (ADLWD)¹, który został wykorzystany w przykładach zawartych w artykule. Poniższy kod służy do przeprowadzenia algorytmu IPF dla przykładu z tabeli 2 (wejściowa tabela kontyngencji).

```
source ("d:/ ipf2df. txt ") # wczytanie kodu implementującego IPF ze strony ADLWD
podregion <-c(" Podregion 1", " Podregion 2") # etykiety dla wierszy
plec<-c(" Mężczyzna ", " Kobieta ") # etykiety dla kolumn
dane <- matrix (c (100 ,150 ,150 ,20) ,nrow =2, ncol =2) # dane wejściowe do tabeli kontyngencji
rownames ( dane ) <- podregion # etykiety wierszy tabeli kontyngencji
colnames ( dane ) <-plec# etykiety kolumn tabeli kontyngencji
dane # wyświetlenie tabeli kontyngencji
rowc<- matrix (c (280 ,220) ,2 ,1) # deklaracja wartości brzegowych dla wierszy
colc<- matrix (c (300 ,200) ,2 ,1) # deklaracja wartości brzegowych dla kolumn
ipf2 (rowc , colc , dane ) # wywołanie algorytmu IPF
```

¹Kod można pobrać ze strony: <http://www.demog.berkeley.edu/~eddieh/datafitting.html>.

W wyniku zadziałania powyższego kodu otrzymujemy tabelę kontyngencji postaci: tabela 2, krok 11.

\$fitted.table

Mężczyzna Kobieta

Podregion 1 107.4739 172.52606

Podregion 2 192.5279 27.47215

Tabela 2. Procedura algorytmu IPF

Wejściowa tabela kontyngencji				Krok 1			
Płeć				Płeć			
Podregion	Mężczyzna	Kobieta	Ogółem	Podregion	Mężczyzna	Kobieta	Ogółem
Podregion 1	100	150	280	Podregion 1	112,00	168,00	280,00
Podregion 2	150	20	220	Podregion 2	194,10	25,90	220,00
Ogółem	300	200	500	Ogółem	306,10	193,90	500,00
Krok 2				Krok 3			
Płeć				Płeć			
Podregion	Mężczyzna	Kobieta	Ogółem	Podregion	Mężczyzna	Kobieta	Ogółem
Podregion 1	109,80	173,30	283,10	Podregion 1	108,6	171,4	280
Podregion 2	190,20	26,70	216,90	Podregion 2	192,9	27,1	220
Ogółem	300,00	200,00	500,00	Ogółem	301,5	198,5	500
Krok 4				Krok 5			
Płeć				Płeć			
Podregion	Mężczyzna	Kobieta	Ogółem	Podregion	Mężczyzna	Kobieta	Ogółem
Podregion 1	108,00	172,70	280,80	Podregion 1	107,70	172,30	280,00
Podregion 2	192,00	27,30	219,20	Podregion 2	192,60	27,40	220,00
Ogółem	300,00	200,00	500,00	Ogółem	300,40	199,60	500,00
Krok 6				Krok 7			
Płeć				Płeć			
Podregion	Mężczyzna	Kobieta	Ogółem	Podregion	Mężczyzna	Kobieta	Ogółem
Podregion 1	107,60	172,60	280,20	Podregion 1	107,50	172,50	280,00
Podregion 2	192,40	27,40	219,80	Podregion 2	192,60	27,40	220,00
Ogółem	300,00	200,00	500,00	Ogółem	300,10	199,90	500,00
Krok 8				Krok 9			
Płeć				Płeć			
Podregion	Mężczyzna	Kobieta	Ogółem	Podregion	Mężczyzna	Kobieta	Ogółem
Podregion 1	107,51	172,54	280,05	Podregion 1	107,49	172,51	280,00
Podregion 2	192,49	27,46	219,95	Podregion 2	192,53	27,47	220,00
Ogółem	300	200,00	500	Ogółem	300,02	199,98	500,00
Krok 10				Krok 11 – osiągnięcie zbieżności			
Płeć				Płeć			
Podregion	Mężczyzna	Kobieta	Ogółem	Podregion	Mężczyzna	Kobieta	Ogółem
Podregion 1	107,48	172,53	280,01	Podregion 1	107,48	172,52	280,00
Podregion 2	192,52	27,47	219,99	Podregion 2	192,52	27,48	220,00
Ogółem	300,00	200,00	500,00	Ogółem	300,00	200,00	500,00

Źródło: opracowanie własne.

W praktycznych zastosowaniach może się zdarzyć, że znane są tylko liczebności brzegowe oszacowane na podstawie danych pochodzących z badania reprezentacyjnego. Oznacza to, że nie są znane wartości wejściowe w dwuwymiarowej tabeli kontyngencji. W takich przypadkach najczęściej ustala się pewien „punkt startowy”, który jest niezbędny do znalezienia wartości \hat{N}_{ij} w wynikowej tabeli kontyngencji. Najczęściej przyjmuje się przy tym za „punkt startowy” algorytmu IPF macierz złożoną z samych jedynek.

Założmy, że podobnie jak w tabeli 2 (wejściowa tabela kontyngencji) znane są oszacowania brzegowe z badania reprezentacyjnego, tj. znana jest oszacowana liczba bezrobotnych w podregionie 1 i podregionie 2, a także oszacowana liczba mężczyzn i kobiet. Zakładamy jednak, że nie są znane ze spisu wartości N_{ij} w wejściowej tabeli kontyngencji. Przyjmujemy ponadto, że ich oszacowanie ze względu na zbyt małe liczebności w odpowiednich przekrojach jest obarczone zbyt niską precyzją szacunku. Konstrukcja tabeli kontyngencji w takim przypadku jest również możliwa. Należy jednak, zgodnie z uwagą poczynioną powyżej, ustalić „punkt startowy” algorytmu IPF w postaci macierzy złożonej z samych jedynek.

W wyniku zadziałania kodu z domyślnie przyjętym „punktem startowym” otrzymujemy wynikową tabelę kontyngencji postaci:

```
$fitted.table  
Mężczyzna Kobieta  
Podregion 1 168 112  
Podregion 2 132 88
```

Wyniki uzyskane w oparciu o tak opisane postępowanie mogą odbiegać w znaczny sposób od rezultatów uzyskanych w sytuacji, gdy dane wejściowe pochodzą ze spisów. Należy więc traktować je ze szczególną ostrożnością. W takich bowiem przypadkach struktura wejściowej tabeli kontyngencji ulega znacznym zmianom, choć sam algorytm IPF osiąga zbieżność. Wykorzystanie informacji pochodzących z dodatkowych źródeł, takich jak spisy, jest ponadto bardziej uzasadnione z merytorycznego punktu widzenia. Zazwyczaj końcowa tabela kontyngencji „zachowuje strukturę” tabeli wejściowej, a zmiany liczebności są niewielkie i pozwalają odtwarzać wartości brzegowe uzyskane z badania reprezentacyjnego.

3. Podsumowanie

Zaprezentowana w artykule metoda wyznaczania estymatorów typu SPREE dla dwuwymiarowych tablic kontyngencji może być stosowana w każdym badaniu częściowym, w którym występuje problem uzyskania wiarygodnych informacji obarczonych niewielkimi błędami szacunku na niskich poziomach agregacji przestrzennej bądź szczegółowo zdefiniowanych domen. Metoda ta w przypadku znajomości liczebności brzegowych tabeli kontyngencji wykorzystuje technikę iteracyjnego proporcjonalnego dopasowywania. Należy jednak podkreślić, że jedynie

znajomość wstępnych wartości wejściowych ze spisu do tabeli kontyngencji nie zmienia jej struktury po zastosowaniu algorytmu IPF. Brak takiej informacji, jak to zostało pokazane w artykule, mimo zachowania zgodności z wartościami brzegowymi może w istotny sposób zmienić wyniki oszacowań.

Estymatory typu SPREE mogą zatem znaleźć zastosowanie przede wszystkim w badaniach prowadzonych przez Główny Urząd Statystyczny, w których wielkość próby i dotychczas stosowane estymatory uniemożliwiają uzyskanie wiarygodnych i obciążonych małymi błędami szacunków na niskich poziomach agregacji.

Literatura

- Haslett S., Noble A., Zabala F. (2008), *New Approaches to Small Area Estimation of Unemployment*, Statistics New Zealand.
- Józefowski T., Szymkowiak M. (2014), *Zastosowanie estymatora typu SPREE w szacowaniu liczby osób bezrobotnych w przekroju podregionów*, *Studia Oeconomica Posnaniensia*, w druku.
- Rao J.N.K. (2003), *Small Area Estimation*, John Wiley & Sons, Hoboken, New Jersey.

CONSTRUCTION AND PRACTICAL USING OF SPREE ESTIMATORS FOR TWO-DIMENSIONAL CONTINGENCY TABLES

Summary: The main aim of the article is to demonstrate the potential of the SPREE estimation based on iterative proportional fitting for two-dimensional contingency table. This technique is used to adjust values in the cells of an estimated contingency table to the totals obtained by means of the survey sampling. In the article some practical aspects of using IPF and SPREE estimation in the context of two-dimensional contingency table were also shown.

Keywords: small area estimation, SPREE estimator, iterative proportional fitting algorithm.