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 naukowi
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

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

Wersja pierwotna: publikacja drukowana

Druk: Drukarnia TOTEM

Spis treści

Summaries

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS nr 327 • 2014
Taksonomia 22 ISSN 1899-3192
Klasyfikacja i analiza danych – teoria i zastosowania

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 proporcjonalego dopasowania do dwuwymiarowej tablicy kontyngencji. Rozważania teoretyczne zostaną zilutrowane 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 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 $\widehat{N}_{ij}$, 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		$\widehat{M}_{i.}$
	Mężczyzna	Kobieta	
Podregion 1	N_{11}	N_{12}	$\widehat{M}_{1.}$
Podregion 2	N_{21}	N_{22}	$\widehat{M}_{2.}$
$\widehat{M}_{.j}$	$\widehat{M}_{.1}$	$\widehat{M}_{.2}$	

Źródło: opracowanie własne.

Ponieważ nie jest możliwe wyprowadzenie analitycznego wzoru na nowe liczebności $\widehat{N}_{ij}$, stosuje się z tzw. technikę iteracyjnego proporcjonalnego dopasowania (IPF – *Iterative Proportional Fitting*) celem ich znalezienia. $\widehat{N}_{ij}$ 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 $\widehat{N}_{ij}$ w tabeli kontyngencji w poszczególnych krokach $n = 1, 2,\ldots$ przedstawiają się następująco:

$$\widehat{N}_{ij}^{(2n-1)} = \frac{\widehat{N}_{ij}^{(2n-2)} N_{i.}}{\sum_{k=1}^{J} \widehat{N}_{ik}^{(2n-2)}}, \tag{1}$$

$$\widehat{N}_{ij}^{(2n)} = \frac{\widehat{N}_{ij}^{(2n-1)} N_{.j}}{\sum_{k=1}^{D} \widehat{N}_{kj}^{(2n-2)}}, \tag{2}$$

przy czym $N_{i.} = \sum_j N_{ij}$, $N_{.j} = \sum_i N_{ij}$ oraz $\widehat{N}_{ij}^{(0)} = N_{ij}$. Algorytm iteracyjnego wyznaczania liczebności $\widehat{N}_{ij}$ 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 $\widehat{N}_{ij} = \lim_{n\to\infty} \widehat{N}_{ij}^{(n)}$. Poniższy przykład ilustruje sposób działania algorytmu IPF w przypadku dwuwymiarowej tabeli kontyngencji.

Załóżmy 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 Laborand 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
```

[1]Kod 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 $\widehat{N}_{ij}$ w wynikowej tabeli kontyngencji. Najczęściej przyjmuje się przy tym za „punkt startowy" algorytmu IPF macierz złożoną z samych jedynek.

Załóżmy, 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 obarczonych 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 Zeland.

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 fiting 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.