

Tomasz Józefowski

Ośrodek Statystyki Małych Obszarów – Urząd Statystyczny w Poznaniu

Marcin Szymkowiak

Uniwersytet Ekonomiczny w Poznaniu, Wydział Informatyki i Gospodarki Elektronicznej, Katedra Statystyki, Ośrodek Statystyki Małych Obszarów – Urząd Statystyczny w Poznaniu

Autor do korespondencji: m.szymkowiak@ue.poznan.pl

ZASTOSOWANIE ESTYMATORA TYPU SPREE W SZACOWANIU LICZBY OSÓB BEZROBOTNYCH W PRZEKROJU PODREGIONÓW

Streszczenie: W literaturze przedmiotu wskazuje się, że estymatory klasy SMO (Statystyka Małych Obszarów – SMO) mają przewagę nad estymatorami znanymi z klasycznej metody reprezentacyjnej, gdyż umożliwiają dostarczenie potrzebnych informacji w sytuacji niewielkiej liczebności lub nawet braku obserwacji w próbie dla danego przekroju [Longford 2005]. Uzyskane w ten sposób oszacowania dla niższych poziomów przestrzennych bądź subpopulacji różnią się często po zsumowaniu od szacunków uzyskanych za pomocą metody reprezentacyjnej dla wyższego poziomu, który jest możliwy ze względu na wystarczającą liczebność próby. Jednym ze sposobów poradzenia sobie z powyżej opisaną niezgodnością jest zastosowanie estymatora typu SPREE [Zhang i Chambers 2004].

Głównym celem artykułu jest zaprezentowanie możliwości, jakie oferuje estymator typu SPREE do oszacowania liczby osób bezrobotnych na poziomie podregionów województwa wielkopolskiego przy wykorzystaniu danych pochodzących z rejestru bezrobotnych oraz Badania Aktywności Ekonomicznej Ludności.

Słowa kluczowe: statystyka małych obszarów, estymator typu SPREE, BAEL, kalibracja.

Klasyfikacja JEL: C8.

USING A SPREE ESTIMATOR TO ESTIMATE THE NUMBER OF UNEMPLOYED PEOPLE ACROSS SUBREGIONS

Abstract: The methodology of small area estimation (SAE) plays an important role in the field of modern information gathering, which aims to cut survey costs while lowering the respond-

ent burden. SAE methods have an advantage over classical methods since they enable reliable estimates at lower levels of spatial aggregation and with more domains, where the representative approach displays too much variability. This means that small area estimation can be used to handle cases with few or no observations for a given domain in the sample. However, cell total estimates for lower levels of spatial aggregation or subpopulations tend to differ from estimates calculated by means of higher levels of representation, which is possible due to their adequate sample size. One way of coping with this incompatibility is by applying a SPREE estimator. This is used to adjust the values in the cells of an estimated contingency table to the totals obtained by means of the representative method. Internal cells can initially be filled with data from previous censuses, or current administrative registers. The method seems to be particularly useful for estimating the parameters of the labour market, since the methodology used in the Labour Force Survey can only yield data at the level of a province. The users of statistical data, however, expect information which is more geographically disaggregated. Considering the above, the aim of the present paper is to demonstrate the potential of the SPREE estimator for estimating the number of unemployed at the level of subregions in the Wielkopolska province using data from the unemployment register and the Labour Force Survey.

Keywords: small area statistics, SPREE estimator, Labour Force Survey, calibration.

Wstęp

Metody statystyki małych obszarów odgrywają istotną rolę w kształtowaniu nowoczesnych technik pozyskiwania informacji, które są ukierunkowane na obniżenie kosztów badań przy jednoczesnym zmniejszeniu obciążeń respondentów. Dzięki swoim własnościom umożliwiają uzyskiwanie wiarygodnych szacunków na niższych poziomach agregacji przestrzennej oraz bardziej szczegółowych domen, dla których klasyczne metody estymacji charakteryzują się zbyt dużą wariancją estymatorów. Mają one przewagę nad estymatorami znanymi z klasycznej metody reprezentacyjnej, gdyż umożliwiają dostarczenie potrzebnych informacji w sytuacji niewielkiej liczebności lub nawet braku obserwacji w próbie dla danego przekroju. Uzyskane w ten sposób oszacowania dla niższych poziomów przestrzennych bądź subpopulacji po zsumowaniu różnią się często od szacunków uzyskanych za pomocą metody reprezentacyjnej dla wyższego poziomu, który jest możliwy ze względu na wystarczającą liczebność próby. Jednym ze sposobów poradzenia sobie z powyżej opisaną niezgodnością jest zastosowanie estymatora typu SPREE [Swanson i Tayman 2012]. Umożliwia on dostosowanie wartości w poszczególnych komórkach szacowanej tabeli kontyngencji do wartości brzegowych otrzymanych przy użyciu metody reprezentacyjnej. Komórki wewnętrzne tabeli początkowo mogą być wypełniane danymi z poprzednich spisów bądź też z bieżących re-

jestrów administracyjnych. Metoda ta jest szczególnie atrakcyjna w kontekście estymacji parametrów charakteryzujących rynek pracy, gdyż techniki użyte w Badaniu Aktywności Ekonomicznej Ludności pozwalają na publikowanie danych jedynie na poziomie województwa. Odbiorcy danych statystycznych oczekują jednak informacji dla bardziej szczegółowych przekrojów geograficznych. W związku z powyższym głównym celem artykułu jest zaprezentowanie możliwości, jakie oferuje estymator typu SPREE do oszacowania liczby osób bezrobotnych na poziomie podregionów województwa wielkopolskiego przy wykorzystaniu danych pochodzących z rejestru bezrobotnych oraz Badania Aktywności Ekonomicznej Ludności.

1. Teoretyczne podstawy estymatora SPREE

Estymatory SPREE (Structure Preserving Estimation)¹ stanowią uogólnioną klasę estymatorów syntetycznych w tym znaczeniu, że wykorzystują pełną informację o ocenach estymatora bezpośredniego. W metodzie tej dokonujemy korekty liczebności znajdujących się w komórkach wielowymiarowej tabeli kontyngencji tak, aby skorygowane wartości sumowały się do znanych liczebności brzegowych. Wyjściowe liczebności w poszczególnych komórkach tabeli kontyngencji mogą na przykład pochodzić z ostatniego spisu, podczas gdy liczebności brzegowe odpowiadają rzetelnym ocenom uzyskanym z wykorzystaniem estymatora bezpośredniego na podstawie danych z badania reprezentacyjnego. Estymatory typu SPREE można wykorzystać na potrzeby szacunków wartości globalnych dla małych obszarów w okresach międzyspisowych [Berg i Fuller 2009].

W niniejszym punkcie przedstawiono teoretyczne podstawy konstrukcji „jednokrokowych” estymatorów typu SPREE. W odróżnieniu od „dwukrokowych” estymatorów typu SPREE nie jest konieczne stosowanie tzw. metody proporcjonalnego iteracyjnego dopasowywania (*iterative proportional fitting* – IPF), a liczebności końcowe w tabelach kontyngencji można wyznaczać wprost ze wzoru. Ideę wyznaczania skorygowanych liczebności w tabeli kontyngencji i ich dopasowania do znanych liczebności brzegowych uzyskanych z wykorzystaniem danych z badania reprezentacyjnego i estymatora bezpośredniego przedstawiono dla trójwymiarowych tabel.

¹ W polskiej literaturze brak tłumaczenia tego typu estymatora. Ponieważ estymacja typu SPREE jest techniką zachowującą strukturę, można byłoby tłumaczyć SPREE jako „estymator zachowujący strukturę”.

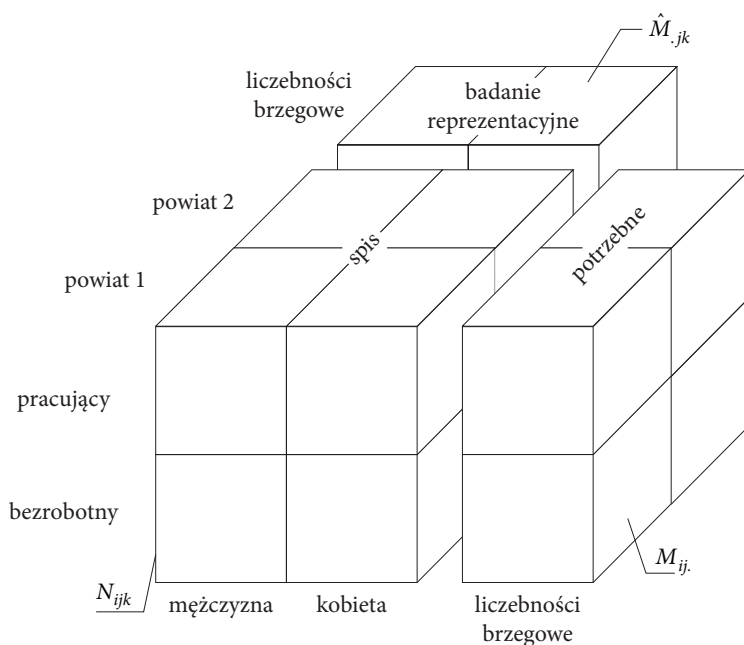
Niech N_{ijk} oznacza znane liczebności w trójwymiarowej tabeli kontyngencji pochodzące ze spisu bądź rejestru administracyjnego, gdzie $i = 1, \dots, D$ oznacza mały obszar (domenę), j oznacza j -ty wariant ($j = 1, \dots, J$) zmiennej y , dla której są dokonywane szacunki (na przykład y może oznaczać liczbę bezrobotnych, zatrudnionych itd.), a k oznacza k -ty wariant ($k = 1, \dots, K$) pewnej dodatkowej zmiennej związanej ze zmienną y (na przykład może to być płeć bądź klasa miejscowości zamieszkania respondenta – por. tabela 1). Ponadto zakładamy, że istnieją pewne bieżące oszacowania niektórych liczebności brzegowych – na podstawie danych pochodzących z badania reprezentacyjnego. Niech $\hat{M}_{.jk}$ oznaczają „rzetelne” oszacowania liczebności brzegowych $M_{.jk}$, które otrzymujemy, wykorzystując znany z metody reprezentacyjnej estymator bezpośredni wartości globalnej. Liczebności brzegowe $N_{.jk} = \sum_i N_{ijk}$ można

oczywiście uzyskać na podstawie wyjściowej tabeli kontyngencji z liczebnościami N_{ijk} . Z upływem czasu, ze względu na incydentalny charakter spisu, dane te dezaktualizują się, a istnieje potrzeba bieżącego zasilania informacyjnego. Stąd liczebności N_{ijk} w tabeli kontyngencji pochodzące ze spisu, poprzez odpowiednią korektę, są dopasowywane do znanych liczebności brzegowych, które stanowią rzetelne i bieżące oszacowania z badania reprezentacyjnego z wykorzystaniem estymatora bezpośredniego. Oczywiście odpowiednie liczebności brzegowe $N_{.jk} = \sum_i N_{ijk}$ nie będą się sumowały do oszacowanych

liczebności brzegowych $\hat{M}_{.jk}$ uzyskanych z badania reprezentacyjnego. Należy je zatem w taki sposób skorygować, aby były odtwarzane liczebności brzegowe $\hat{M}_{.jk}$. Ważne również, aby nowe liczebności nie różniły się za bardzo od liczebności N_{ijk} i nie zmieniały w istotny sposób struktury danych zawartych w tabeli kontyngencji.

Rysunek 1 prezentuje opisaną powyżej sytuację w sposób graficzny. Zakładamy przy tym, że znane są informacje o liczbie pracujących i bezrobotnych ($J = 2$) w przekroju powiatów ($D = 2$) i uwzględnieniem płci ($K = 2$). N_{ijk} oznacza zatem liczbę osób ze spisu bądź rejestru administracyjnego, które pochodzą z i -tego powiatu, mają j -ty wariant statusu na rynku pracy i k -ty wariant płci. Informacje te są również zawarte w tabeli 1. Zakładamy przy tym, zgodnie z uwagami poczynionymi powyżej, że znane są bieżące oszacowania z badania reprezentacyjnego liczby pracujących i bezrobotnych mężczyzn oraz kobiet, tj. $\hat{M}_{.jk}$. Zwróćmy przy tym uwagę, że nie zakładamy znajomości z badania reprezentacyjnego oszacowanych liczebności brzegowych $\hat{M}_{.jk}$. Ich uzyskanie, ze względu na niewielką lub w niektórych wypadkach zerową liczebność w pewnych przekrojach, uniemożliwia uzyskanie rzetelnych oszacowań

z wykorzystaniem estymatora bezpośredniego. Na przykład brak reprezentacji lub niewielka liczebność próby w pewnym powiecie w kategorii pracujących mężczyzn może być przyczyną niemożności uzyskania takich oszacowań lub będą się one odznaczały bardzo niską precyzją. W prezentowanym w artykule podejściu zakładamy zatem znajomość tylko jednych liczebności brzegowych, które można uzyskać z badania reprezentacyjnego z wykorzystaniem estymacji bezpośredniej. Stąd wcześniejsze określenie w artykule opisywanego podejścia jako „jednokrokowe”, w odróżnieniu od estymacji SPREE typu „dwukrokowego”, gdzie zakłada się znajomość obydwu liczebności brzegowych. Liczebności M_{ij} w podejściu „jednokrokowym” uzyska się po korekcie liczebności N_{ijk} w oczywisty sposób.



Rysunek 1. Struktura danych dla małych obszarów

Źródło: Opracowano na podstawie pracy: [Purcell i Kish 1980]

Niech M_{ijk} oznacza nieznane i poszukiwane liczebności w trójwymiarowej tabeli kontyngencji, które będą odtwarzać oszacowania brzegowe z badania reprezentacyjnego, tzn. $\hat{M}_{.jk} = \sum_i M_{ijk}$, i nie będą się znacznie różniły od liczebności wejściowych N_{ijk} . Problem poszukiwania nowych liczebności

Tabela 1. Struktura danych dla małych obszarów

Powiat	Status osoby na rynku pracy	Płeć		$M_{ij.}$
		mężczyzna	kobieta	
Powiat 1	pracująca	N_{111}	N_{112}	$M_{11.}$
	bezrobotna	N_{121}	N_{122}	$M_{12.}$
Powiat 2	pracująca	N_{211}	N_{212}	$M_{21.}$
	bezrobotna	N_{221}	N_{222}	$M_{22.}$
		$\hat{M}_{.j1}$	$\hat{M}_{.j2}$	

w trójwymiarowej tabeli kontyngencji można zapisać w postaci poniższego zadania optymalizacyjnego.

(W1) Minimalizacja funkcji odległości:

$$D(N_{ijk}, M_{ijk}) \rightarrow \min, \quad (1)$$

(W2) Równania kalibracyjne:

$$\sum_i M_{ijk} = \hat{M}_{.jk}. \quad (2)$$

Pierwszy z warunków (W1) oznacza, że wyznaczone nowe liczebności w trójwymiarowej tabeli kontyngencji powinny być bliskie, w sensie przyjętej funkcji odległości, liczebnościom wejściowym ze spisu bądź rejestru administracyjnego. Drugi z warunków oznacza z kolei, że liczebności te powinny być w taki sposób wyznaczone, aby uzyskane na ich podstawie jedne z liczebności brzegowych pokrywały się z oszacowaniami, które zostaną uzyskane z wykorzystaniem estymatora bezpośredniego i informacji pochodzących z bieżącego badania reprezentacyjnego. Z racji na podobieństwo do sposobu w jaki są poszukiwane wagi kalibracyjne w estymatorach kalibracyjnych warunków (W2) określono mianem „równań kalibracyjnych”, a sama konstrukcja estymatorów typu SPREE jest w dużej mierze zbliżona do podejścia kalibracyjnego [Szymkowiak 2007; Särndal 2007].

Kluczową rolę w poszukiwaniu nowych liczebności M_{ijk} w trójwymiarowej tabeli kontyngencji odgrywa odpowiednio dobrana funkcja odległości. W literaturze przedmiotu wskazuje się na dwie najczęściej wykorzystywane w praktyce funkcje, które umożliwiają wyznaczenie liczebności M_{ijk} , tj. funkcję odległości χ^2 i dyskryminacyjną funkcję odległości. Wyrażają się one następującymi wzorami:

Funkcja odległości χ^2

$$D(N_{ijk}, M_{ijk}) = \frac{1}{2} \sum_{i,j,k} \frac{(N_{ijk} - M_{ijk})^2}{N_{ijk}}. \quad (3)$$

Dyskryminacyjna funkcja odległości

$$D(N_{ijk}, M_{ijk}) = \sum_{i,j,k} N_{ijk} \ln \frac{N_{ijk}}{M_{ijk}}. \quad (4)$$

Definicja 1. Estymatorem typu SPREE liczebności M_{ijk} w trójwymiarowej tabeli kontyngencji jest statystyka $\hat{M}_{ijk}$ będąca rozwiązaniem zadania optymalizacyjnego postaci:

$$\hat{M}_{ijk} = \arg \min_{M_{ijk}} D(M_{ijk}, N_{ijk}), \quad (5)$$

przy warunku

$$\sum_i M_{ijk} = \hat{M}_{.jk}. \quad (6)$$

Poniższe twierdzenia rozstrzygają postać estymatora typu SPREE określonego w definicji (1).

Twierdzenie 1. Rozwiązaniem zadania minimalizacji (5) dla funkcji odległości (3) i przy warunku (6) jest statystyka postaci:

$$\hat{M}_{ijk} = N_{ijk} \frac{\hat{M}_{.jk}}{N_{.jk}}. \quad (7)$$

Dowód. Na potrzeby dowodu tego twierdzenia wykorzystano metodę czynników nieoznaczonych Lagrange'a. Funkcja Lagrange'a ma postać:

$$L = \frac{1}{2} \sum_{i,j,k} \frac{(N_{ijk} - M_{ijk})^2}{N_{ijk}} - \sum_{j,k} \lambda_{jk} \left(\sum_i M_{ijk} - \hat{M}_{.jk} \right). \quad (8)$$

Pochodna funkcji L względem M_{ijk} ma postać:

$$\frac{\partial L}{\partial M_{ijk}} = \frac{1}{2} \cdot \frac{2M_{ijk} - 2N_{ijk}}{N_{ijk}} - \lambda_{jk}. \quad (9)$$

Przyrównując obliczoną pochodną do zera, otrzymujemy następujące równanie:

$$\frac{M_{ijk} - N_{ijk}}{N_{ijk}} = \lambda_{jk}, \quad (10)$$

którego rozwiązaniem jest:

$$M_{ijk} = N_{ijk}(1 + \lambda_{jk}). \quad (11)$$

Dokonując sumowania po wszystkich domenach, tj. po i , otrzymujemy następujące równanie:

$$\sum_i M_{ijk} = \sum_i N_{ijk}(1 + \lambda_{jk}). \quad (12)$$

Uwzględniając równanie (6), otrzymujemy, że:

$$\hat{M}_{\cdot,jk} = \lambda_{jk} \sum_i N_{ijk} + \sum_i N_{ijk}. \quad (13)$$

Ponieważ $N_{\cdot,jk} = \sum_i N_{ijk}$, więc ostatecznie otrzymujemy:

$$\lambda_{jk} = \frac{\hat{M}_{\cdot,jk} - N_{\cdot,jk}}{N_{\cdot,jk}}. \quad (14)$$

Podstawiając uzyskane powyżej λ_{jk} do równania (11), otrzymujemy poszukiwaną postać estymatora:

$$M_{ijk}^{opt} = \hat{M}_{ijk} = N_{ijk}(1 + \lambda_{jk}) = N_{ijk} \frac{\hat{M}_{\cdot,jk}}{N_{\cdot,jk}}. \quad (15)$$

Należy jeszcze sprawdzić, czy w punkcie $\hat{M}_{ijk}$ istnieje minimum (warunek dostateczny istnienia ekstremum warunkowego). W tym celu trzeba wykazać, że forma kwadratowa $d^2 L(\hat{M}_{ijk})(\xi)$ jest dodatnio określona dla pewnego niezerowego wektora ξ . Mamy:

$$d^2 L(\hat{M}_{ijk})(\xi) = \sum_{i,j,k} \sum_{l,m,n} \frac{\partial^2 L}{\partial M_{ijk} \partial M_{lmn}} \xi_{ijk} \xi_{lmn}. \quad (16)$$

Zauważmy, że:

$$\frac{\partial^2 L}{\partial M_{ijk} \partial M_{lmn}} = \begin{cases} \frac{1}{N_{ijk}} & \text{dla } i, j, k = l, m, n, \\ 0 & \text{dla } i, j, k \neq l, m, n. \end{cases} \quad (17)$$

Podstawiając obliczone pochodne drugiego rzędu do formy kwadratowej (16), otrzymujemy:

$$d^2 L(\hat{M}_{ijk}^{opt})(\xi) = \sum_{i,j,k} \sum_{l,m,n} \frac{\partial^2 L}{\partial M_{ijk} \partial M_{lmn}} \xi_{ijk} \xi_{lmn} = \sum_{i,j,k} \frac{1}{N_{ijk}^2} \xi_{ijk}^2. \quad (18)$$

Jest to oczywiście forma kwadratowa dodatnio określona. Stąd statystyka określona wzorem (7) jest poszukiwanym rozwiązaniem zadania minimalizacji funkcji odległości.

Twierdzenie 2. Rozwiązaniem zadania minimalizacji (5) dla funkcji odległości (4) i przy warunku (6) jest statystyka postaci:

$$\hat{M}_{ijk} = N_{ijk} \frac{\hat{M}_{\cdot,jk}}{N_{\cdot,jk}}. \quad (19)$$

Dowód. W celu udowodnienia twierdzenia wykorzystano ponownie metodę czynników nieoznaczonych Lagrange'a. Funkcja Lagrange'a ma postać:

$$L = \sum_{i,j,k} N_{ijk} \ln \frac{N_{ijk}}{M_{ijk}} - \sum_{j,k} \lambda_{jk} \left(\sum_i M_{ijk} - \hat{M}_{\cdot,jk} \right). \quad (20)$$

Pochodna funkcji L względem M_{ijk} ma postać:

$$\frac{\partial L}{\partial M_{ijk}} = N_{ijk} \cdot \left(\frac{-N_{ijk}}{M_{ijk}^2} \right) \cdot \frac{M_{ijk}}{N_{ijk}} - \lambda_{jk}. \quad (21)$$

Przyrównując obliczoną pochodną do zera, otrzymujemy następujące równanie:

$$-\frac{N_{ijk}}{M_{ijk}} = \lambda_{jk}, \quad (22)$$

którego rozwiązaniem jest:

$$M_{ijk} = -N_{ijk}\lambda_{jk}. \quad (23)$$

Dokonując sumowania po wszystkich domenach, tj. po i , otrzymujemy następujące równanie:

$$\sum_i M_{ijk} = -\sum_i N_{ijk}\lambda_{jk}. \quad (24)$$

Uwzględniając równanie (6) oraz to, że $N_{.jk} = \sum_i N_{ijk}$, otrzymujemy:

$$\lambda_{jk} = -\frac{\hat{M}_{.jk}}{N_{.jk}}. \quad (25)$$

Podstawiając uzyskane powyżej λ_{jk} do równania (23), otrzymujemy poszukiwaną postać estymatora:

$$M_{ijk}^{opt} = \hat{M}_{ijk} = -N_{ijk}\lambda_{jk} = -N_{ijk}\left(-\frac{\hat{M}_{.jk}}{N_{.jk}}\right) = N_{ijk} \cdot \frac{\hat{M}_{.jk}}{N_{.jk}}. \quad (26)$$

Podobnie jak w twierdzeniu 1 należy jeszcze sprawdzić, czy w punkcie $\hat{M}_{ijk}$ istnieje minimum (warunek dostateczny istnienia ekstremum warunkowego). W tym celu trzeba wykazać, że forma kwadratowa $d^2L(\hat{M}_{ijk})(\xi)$ jest dodatnio określona dla pewnego niezerowego wektora ξ .

Zauważmy, że:

$$\frac{\partial^2 L}{\partial M_{ijk} \partial M_{lmn}} = \begin{cases} \frac{N_{ijk}}{M_{ijk}^2} & \text{dla } i, j, k = l, m, n, \\ 0 & \text{dla } i, j, k \neq l, m, n, \end{cases} \quad (27)$$

$$d^2L(\hat{M}_{ijk}^{opt})(\xi) = \sum_{i,j,k} \sum_{l,m,n} \frac{\partial^2 L}{\partial M_{ijk} \partial M_{lmn}} \xi_{ijk} \xi_{lmn} = \sum_{i,j,k} \frac{N_{ijk}}{\hat{M}_{ijk}^2} \xi_{ijk}^2. \quad (28)$$

Jest to oczywiście forma kwadratowa dodatnio określona. Stąd statystyka określona wzorem (19) jest poszukiwanym rozwiązaniem zadania minimalizacji funkcji odległości.

Z powyższych twierdzeń wynika, że – bez względu na wybór funkcji odległości – uzyskujemy tę samą postać estymatora liczebności w trójwymiarowej

tabeli kontyngencji. Mając wyznaczone nowe liczebności $\hat{M}_{ijk}$ w trójwymiarowej tabeli kontyngencji, które sumują się do liczebności brzegowych $\hat{M}_{.jk}$ uzyskanych z badania reprezentacyjnego, bardzo łatwo można uzyskać pozostałe wartości brzegowe $M_{ij.}$, dokonując sumowania po k , tj. po wszystkich wariantach dodatkowej zmiennej związanej ze zmienną y . Uzyskujemy w ten sposób wzór na liczebności brzegowe $M_{ij.}$:

$$M_{ij.} = \sum_k N_{ijk} \frac{\hat{M}_{.jk}}{N_{.jk}}. \quad (29)$$

Z racji tego, że liczebności wejściowe N_{ijk} w trójwymiarowej tabeli kontyngencji pochodzą z badania pełnego (spisów, rejestrów administracyjnych) i nie są w związku z tym obarczone błędami losowymi, wariancja estymatora typu SPREE jest uzależniona tylko od wariancji $\hat{M}_{.jk}$. Wariancja estymatora $\hat{M}_{ijk}$ wyraża się zatem wzorem:

$$V(\hat{M}_{ijk}) = V\left(N_{ijk} \frac{\hat{M}_{.jk}}{N_{.jk}}\right) = \left(\frac{N_{ijk}}{N_{.jk}}\right)^2 V(\hat{M}_{.jk}). \quad (30)$$

Celem zilustrowania omawianej metody wyznaczania liczebności w trójwymiarowej tabeli kontyngencji z wykorzystaniem „jednokrokowego” estymatora typu SPREE rozważmy następujący przykład – por. tabela 2. Załóżmy, że ze spisu dysponujemy informacją na temat liczby osób pracujących i bezrobotnych w przekroju powiatów i płci.

Tabela 2. Bezrobotni i pracujący według powiatów i płci

Powiat	Status osoby na rynku pracy	Płeć		$M_{ij.}$
		mężczyzna	kobieta	
Powiat 1	pracująca	100	150	$M_{11.}$
	bezrobotna	50	60	$M_{12.}$
Powiat 2	pracująca	150	20	$M_{21.}$
	bezrobotna	80	100	$M_{22.}$
		400	380	

Na potrzeby przykładu, celem uproszczenia, przyjmijmy, że dane są tylko dwa powiaty tzn. $D = 2$. Załóżmy ponadto, że wartości brzegowe zostały oszacowane z wykorzystaniem estymatora bezpośredniego i danych pochodzących

z bieżącego badania reprezentacyjnego, na przykład Badania Aktywności Ekonomicznej Ludności. Przyjmujemy przy tym, że pracujących i bezrobotnych mężczyzn jest odpowiednio 280 i 120, a pracujących i bezrobotnych kobiet odpowiednio 200 i 180 tzn. $\hat{M}_{11} = 280$, $\hat{M}_{21} = 120$, $\hat{M}_{12} = 200$, $\hat{M}_{22} = 180$. Stąd oszacowana liczba mężczyzn i kobiet po wszystkich powiatach, bez względu na status na rynku pracy wynosi odpowiednio 400 i 380. Zwróćmy jednak uwagę, że liczebności spisowe N_{ijk} w poszczególnych komórkach tabeli kontyngencji nie sumują się do oszacowanych liczebności brzegowych z badania reprezentacyjnego. Należy więc je odpowiednio skorygować tak, aby zapewniona była sumowalność do oszacowanych wartości brzegowych. W tym celu należy skorzystać ze wzoru (7). Tabela 3 zawiera informacje na temat bezrobotnych i pracujących według powiatów i płci po zastosowaniu estymatora typu SPREE, tj. po odpowiedniej korekcie.

Tabela 3. Bezrobotni i pracujący według powiatów i płci – po zastosowaniu estymatora typu SPREE

Powiat	Status osoby na rynku pracy	Płeć		M_{ij}
		mężczyzna	kobieta	
Powiat 1	pracująca	112	176	288
	bezrobotna	46	68	114
Powiat 2	pracująca	168	24	192
	bezrobotna	74	112	186
		400	380	

Odpowiednie – skorygowane wartości $\hat{M}_{ijk}$ w wyznaczonej trójwymiarowej tablicy kontyngencji uzyskano z następujących wyliczeń²:

$$\begin{aligned}\hat{M}_{111} &= 100 \cdot \frac{280}{250} = 112, & \hat{M}_{121} &= 50 \cdot \frac{120}{130} = 46, & \hat{M}_{211} &= 150 \cdot \frac{280}{250} = 168, \\ \hat{M}_{221} &= 80 \cdot \frac{120}{130} = 74, & \hat{M}_{112} &= 150 \cdot \frac{200}{170} = 176, & \hat{M}_{122} &= 60 \cdot \frac{180}{160} = 68, \\ \hat{M}_{212} &= 20 \cdot \frac{200}{170} = 24, & \hat{M}_{222} &= 100 \cdot \frac{180}{160} = 112.\end{aligned}$$

Z kolei wartości brzegowe M_{ij} otrzymano ze wzoru (29).

² Wyniki zaokrąglono do wartości całkowitych.

2. Estymator typu SPREE w szacowaniu liczby bezrobotnych w przekroju podregionów

Podstawowym źródłem informacji o rynku pracy w Polsce jest Badanie Aktywności Ekonomicznej Ludności (BAEL). Jest to badanie reprezentacyjne, które dostarcza kompleksowych danych na temat sytuacji w zakresie aktywności ekonomicznej ludności, tzn. fakcie wykonywania pracy, pozostawania bezrobotnym lub biernym zawodowo. Najniższym poziomem podziału administracyjnego, na którym udostępniane są wyniki z BAEL jest województwo. Jest to konsekwencja reprezentacyjnego charakteru badania i wielkości próby. Oznacza to, że oszacowania na niższym poziomie podziału terytorialnego są obciążone zbyt dużym błędem losowym, podobnie jak dodatkowe przekroje w ujęciu wojewódzkim.

Istnieje jednak potrzeba pokrycia informacyjnego na niższych poziomach agregacji przestrzennej czy też bardziej szczegółowych domen. Dla władz powiatu, gminy czy miasta szczególnie istotna z punktu widzenia prowadzenia właściwej polityki rynku pracy jest informacja na temat bezrobocia w ich regionie czy jednostce urbanistycznej, a mniejszą rolę odgrywają dane na poziomie całego kraju bądź województwa. Powstaje zatem luka informacyjna, której wyniki z Badania Aktywności Ekonomicznej Ludności nie są w stanie wypełnić ze względu na niewystarczającą liczebność próby, na niższych aniżeli województwo, poziomach terytorialnych.

W tej części artykułu zaprezentowano praktyczne wykorzystanie estymatora typu SPREE w szacowaniu liczby bezrobotnych w przekroju podregionów województwa wielkopolskiego z uwzględnieniem płci oraz wieku³. W tym celu wykorzystano dane pochodzące z Badania Aktywności Ekonomicznej Ludności za II kwartał 2011 roku⁴. Ze względu na małe liczebności próby w odpowiednich przekrojach podregionów województwa wielkopolskiego wyznaczonych przez płeć i wiek, nie jest zasadne wykorzystanie estymatora Horvitz-Thompsona w szacowaniu liczby bezrobotnych. Ponieważ estymator typu SPREE wymaga danych wejściowych do trójwymiarowej tabeli kontyngencji⁵, więc wykorzystano informacje pochodzące z miesięcznej spra-

³ Podregiony stanowią poziom agregacji przestrzennej o jeden niżej aniżeli województwo. Wyniki z BAEL, jak to zostało zasygnalizowane, nie są publikowane na tym poziomie.

⁴ Dane te zaczerpnięto z Banku Danych Lokalnych.

⁵ Wymiary tabeli tworzą podregiony województwa wielkopolskiego, kategorie płci oraz wieku. Dane wejściowe pochodzą zazwyczaj z rejestrów administracyjnych, spisu czy z innych źródeł.

wzrostawczości Ministerstwa Pracy i Polityki Społecznej sporządzanej przez powiatowe urzędy pracy⁶.

Tabela 4 zawiera szczegółowe informacje na temat liczby bezrobotnych zarejestrowanych w podregionach województwa wielkopolskiego w II kwartale 2011 roku w przekroju płci oraz wieku. Dane te pochodzą z miesięcznej sprawozdawczości Ministerstwa Pracy i Polityki Społecznej (formularz MPiPS-01), które są sporządzane przez powiatowe urzędy pracy. W tabeli tej zawarto ponadto informacje o liczbie bezrobotnych w województwie wielkopolskim z uwzględnieniem płci i wieku, ale pochodzące z Badania Aktywności Ekonomicznej Ludności za II kwartał 2011 roku.

Tabela 4. Bezrobotni zarejestrowani w podregionach województwa wielkopolskiego w II kwartale 2011 roku

Podregion	Wiek	Płeć	
		mężczyzna	kobieta
Kaliski	poniżej 25 lat	2 317	4 191
	25 lat i więcej	7 672	11 487
Koniński	poniżej 25 lat	3 713	5 292
	25 lat i więcej	11 151	14 940
Leszczyński	poniżej 25 lat	1 766	3 155
	25 lat i więcej	5 690	7 768
Pilski	poniżej 25 lat	1 830	3 229
	25 lat i więcej	7 178	9 758
Poznański	poniżej 25 lat	1 341	2 167
	25 lat i więcej	5 060	7 243
Miasto Poznań	poniżej 25 lat	534	742
	25 lat i więcej	4 850	5 230
Województwo wielkopolskie	poniżej 25 lat	19 000	18 000
	25 lat i więcej	45 000	51 000

Źródło: Miesięczna sprawozdawczość Ministerstwa Pracy i Polityki Społecznej.

Zwróćmy uwagę, że wyjściowe liczebności w tabeli kontyngencji nie sumują się do odpowiednich liczebności brzegowych z Badania Aktywności Ekonomicznej Ludności. Zachodzi zatem potrzeba ich korekty celem zapewnienia

⁶ Należy mieć na uwadze to, że występują nieco inne definicje osoby bezrobotnej w obydwu źródłach, tj. w sprawozdawczości powiatowych urzędów pracy (bezrobocie rejestrowane) i BAEL. Szczegółowe informacje na temat występujących różnic można znaleźć w pracy Janukowicza [2010].

zgodności z wynikami z badania BAEL. Umożliwi to jednocześnie prezentację liczby bezrobotnych w ujęciu podregionów, płci i wieku zgodnie z definicją z badania BAEL. Celem odpowiedniego dopasowania struktur z dwóch różnych źródeł danych wykorzystano opisany w artykule estymator typu SPREE.

Tabela 5 przedstawia wyniki estymacji liczby bezrobotnych w II kwartale 2011 roku w przekroju podregionów województwa wielkopolskiego, z uwzględnieniem płci i wieku i z wykorzystaniem estymatora typu SPREE. Po jego zastosowaniu wartości wewnętrzne w utworzonej tabeli kontyngencji sumują się do odpowiednich liczebności brzegowych pochodzących z BAEL. Co więcej, uzyskano również informacje o liczbie bezrobotnych w poszczególnych grupach wieku w każdym z podregionów.

Tabela 5. Bezrobotni w podregionach województwach wielkopolskiego w II kwartale 2011 roku

Podregion	Wiek	Płeć		Ogółem
		mężczyzna	kobieta	
Kaliski	poniżej 25 lat	3 828	4 018	7 846
	25 lat i więcej	8 299	10 382	18 681
Koniński	poniżej 25 lat	6 134	5 073	11 207
	25 lat i więcej	12 062	13 503	25 565
Leszczyński	poniżej 25 lat	2 917	3 025	5 942
	25 lat i więcej	6 155	7 021	13 176
Pilski	poniżej 25 lat	3 023	3 096	6 119
	25 lat i więcej	7 764	8 820	16 584
Poznański	poniżej 25 lat	2 215	2 077	4 293
	25 lat i więcej	5 473	6 547	12 020
Miasto Poznań	poniżej 25 lat	882	711	1 594
	25 lat i więcej	5 246	4 727	9 973
Województwo wielkopolskie	poniżej 25 lat	19 000	18 000	37 000
	25 lat i więcej	45 000	51 000	96 000

Źródło: Opracowanie własne z wykorzystaniem estymatora SPREE.

Ze względu na to, że w Banku Danych Lokalnych, z którego zaczerpnięto informacje o liczbie bezrobotnych (definicja zgodna z definicją stosowaną w BAEL), są publikowane wskaźniki precyzji jedynie dla wybranych kategorii, nie było możliwości uzyskania wskaźników precyzji szacunków liczby bezrobotnych w przekroju podregionów, płci i wieku.

Analiza danych zawartych w tabelach 4 i 5 wskazuje, że dla wielu przekrojów różnice w liczbie bezrobotnych są na zadowalającym poziomie. Jest to

zgodne z filozofią konstrukcji estymatora typu SPREE, w którym poszukuje się nowych liczebności w tabeli kontyngencji, mających zapewnić zgodność z wartościami brzegowymi, i które nie będą się znacznie różniły od wartości wejściowych pochodzących ze spisu, rejestrów administracyjnych czy innych źródeł. Należy jednak podkreślić, że w przypadku mężczyzn w grupie wiekowej poniżej 25 lat różnice te były rzędu 65%. Na przykład w podregionie kaliskim liczba bezrobotnych zarejestrowanych mężczyzn poniżej 25 roku życia wynosiła 2317. Po korekcie, z wykorzystaniem estymatora typu SPREE, wzrosła do 3828. Różnice te w wielu wypadkach są zrozumiałe i wynikają między innymi z innych definicji osoby bezrobotnej w analizowanych źródłach. Oszacowana wartość liczby bezrobotnych ujmuje bowiem nie tylko osoby zarejestrowane, ale i te, które pracują w tzw. szarej strefie.

Podsumowanie

W artykule przedstawiono w kompleksowy sposób metodę konstrukcji jednokrokowych estymatorów typu SPREE. Podano najważniejsze twierdzenia dotyczące wyprowadzania postaci estymatorów liczebności w trójwymiarowych tabelach kontyngencji, tak aby zapewnić zgodność z wartościami brzegowymi pochodzącymi z badania reprezentacyjnego. Rozważania teoretyczne zilustrowano praktycznym wykorzystaniem estymatora typu SPREE w szacowaniu liczby bezrobotnych osób w przekroju płci i wieku dla podregionów województwa wielkopolskiego.

Zaprezentowana metodologia wyznaczania estymatorów typu SPREE może być stosowana w każdym badaniu częściowym, w którym występuje problem uzyskania wiarygodnych informacji obciążonych niewielkimi błędami szacunku na niskich poziomach agregacji przestrzennej. Możliwe jest również zastosowanie w badaniach praktycznych tzw. dwukrokowego estymatora typu SPREE. Jak wskazuje literatura przedmiotu [Rao 2003], można zakładać polepszenie uzyskanych wyników w stosunku do podejścia jednokrokowego. Wymaga to jednak znajomości wszystkich wartości brzegowych z badania reprezentacyjnego. Estymatory typu SPREE mogą znaleźć zatem szczególnie zastosowanie w badaniach prowadzonych przez Główny Urząd Statystyczny, w których wielkość próby i dotychczas stosowane estymatory uniemożliwiają uzyskanie wiarygodnych i obciążonych małymi błędami szacunków na poziomie niższym aniżeli województwo.

Bibliografia

- Berg, E., Fuller, W.A., 2009, *A SPREE Small Area Procedure for Estimating Population Counts*, SSC Annual Meeting, Proceedings of the Survey Methods Section.
- Janukowicz, P., 2010, *Bezrobocie rejestrowane a bezrobocie według BAEL*, Polityka Społeczna, nr 1.
- Longford, N., 2005, *Missing Data and Small Area Estimation: Modern Analytical Equipment for the Survey Statistician*, Series: Statistics for Social and Behavioral Sciences, Springer.
- Purcell, N.J., Kish, L., 1980, *Postcensal Estimates for Local Areas (or Domains)*, International Statistical Review, 48, s. 3–18.
- Rao, J.N.K., 2003, *Small Area Estimation*, John Wiley & Sons, INC., Publication.
- Särndal, C.-E., 2007, *The Calibration Approach in Survey Theory and Practice*, Survey Methodology, vol. 33, no. 2, s. 99–119.
- Swanson, D.A., Tayman, J., 2012, *Subnational Population Estimates*, The Springer Series on Demographic Methods and Population Analysis, New York.
- Szymkowiak, M., 2007, *Przyczynek do kalibracji w badaniach statystycznych z brakami odpowiedzi*, w: Panek, E. (red.), *Kapitał ludzki i wiedza w gospodarce: wyzwania XXI wieku*, Akademii Ekonomicznej w Poznaniu, Poznań, s. 194–204.
- Zhang, L., Chambers, R.L., 2004, *Small Area Estimates for Cross-classifications*, Journal of the Royal Statistical Society B, 66, s. 479–496.