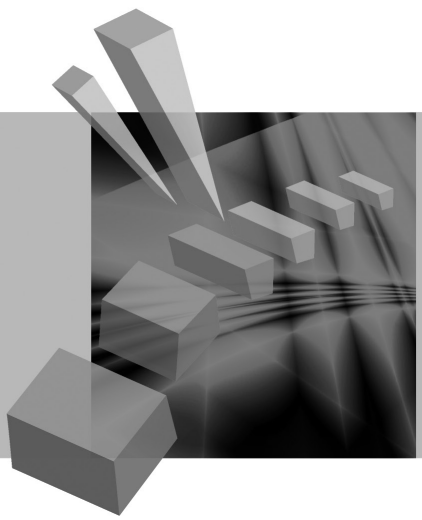
PRACE NAUKOWE
Uniwersytetu Ekonomicznego we Wrocławiu
RESEARCH PAPERS
of Wrocław University of Economics

279

Taksonomia 21

Klasyfikacja i analiza danych – teoria i zastosowania



Redaktorzy naukowi
Krzysztof Jajuga
Marek Walesiak

Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2013

Redaktor Wydawnictwa: Aleksandra Śliwka
Redaktor techniczny: Barbara Łopusiewicz
Korektor: Barbara Cibis
Łamanie: Małgorzata Czupryńska
Projekt okładki: Beata Dębska

Publikacja jest dostępna w Internecie na stronach:
www.ibuk.pl, www.ebscohost.com,
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 279 • 2013

Taksonomia 21. Klasyfikacja i analiza danych – teoria i zastosowania ISSN 1899-3192

Marcin Szymkowiak
Uniwersytet Ekonomiczny w Poznaniu

KONSTRUKCJA ESTYMATORÓW KALIBRACYJNYCH WARTOŚCI GLOBALNEJ DLA RÓŻNYCH FUNKCJI ODLEGŁOŚCI

Streszczenie: W badaniach statystycznych jedną z metod umożliwiających redukcję obciążenia i zwiększenie precyzji szacunku na skutek występowania braków informacji jest kalibracja, której podstawy teoretyczne zostały zaproponowane przez Devilla i Särndala [1992]. W klasycznym ujęciu wyznaczanie wag kalibracyjnych oparte jest na odpowiednio dobranej funkcji odległości, która minimalizuje odległość między wyjściowymi wagami wynikającymi ze schematu losowania próby a tzw. wagami kalibracyjnymi. W artykule przedstawione zostały różne funkcje odległości, które można wykorzystać na etapie konstrukcji wag kalibracyjnych. W części empirycznej, z wykorzystaniem programu R i funkcji calib dostępnej w pakiecie sampling, pokazane zostało, w jaki sposób wyznaczać wagi kalibracyjne w badaniach z brakami odpowiedzi dla różnych funkcji odległości.

Słowa kluczowe: kalibracja, wagi kalibracyjne, estymatory kalibracyjne, braki odpowiedzi, funkcja odległości.

1. Wstęp

W badaniach statystycznych prowadzonych przez urzędy statystyczne braki odpowiedzi stanowią jeden z istotnych problemów, który wpływa na jakość zebranych danych, a w konsekwencji na cały proces estymacji. Jedną z metod umożliwiających redukcję obciążenia i zwiększenie precyzji szacunku na skutek występowania braków informacji jest kalibracja, której podstawy teoretyczne zostały zaproponowane przez Devilla i Särndala [1992].

Zgodnie z definicją zaproponowaną przez Lundströma i Särndala [Särndal, Lundström 2005; Särndal 2007] kalibracja to metoda polegająca na korygowaniu wag wyjściowych wynikających ze schematu losowania próby, tak aby spełnione były odpowiednie równania kalibracyjne w odniesieniu do zmiennych pomocniczych. W wyniku jej zastosowania najczęściej udaje się zredukować obciążenie i wariancję wykorzystywanych w uogólnianiu wyników estymatorów.

W klasycznym ujęciu wyznaczanie wag kalibracyjnych oparte jest na odpowiednio dobranej funkcji odległości, która minimalizuje odległość między wyjściowymi

wagami wynikającymi ze schematu losowania próby a tzw. wagami kalibracyjnymi. Wykorzystuje się przy tym funkcję odległości opartą na tzw. metryce chi-kwadrat. W artykule przedstawione zostaną inne funkcje odległości, które można wykorzystać na etapie konstrukcji wag kalibracyjnych. W części empirycznej, z wykorzystaniem pakietu R, przedstawiona zostanie metoda ich wyznaczania wraz z ich empiryczną oceną.

2. Teoretyczne podstawy kalibracji

Niech dana będzie *N*-elementowa populacja $U = \{1, \ldots, N\}$. Z populacji tej losujemy zgodnie z określonym schematem losowania *n*-elementową próbę $s \subseteq U$. Niech π_i oznacza prawdopodobieństwo inkluzji *i*-tej jednostki do próby, tzn. $\pi_i = P(i \in s)$ dla $i = 1, \ldots, N$, a $d_i = \frac{1}{\pi_i}$ będzie wagą odpowiadającą jednostce *i*.

Załóżmy, że celem badania jest oszacowanie wartości globalnej pewnej zmiennej *y*, określonej wzorem:

$$Y = \sum_{i=1}^{N} y_i, \tag{1}$$

gdzie y_i oznacza wartość zmiennej y dla *i*-tej jednostki badania, $i = 1, \ldots, N$.

Klasycznym estymatorem wartości globalnej (1) jest znany z metody reprezentacyjnej estymator Horvitza-Thompsona, który wyraża się wzorem:

$$\hat{Y}_{HT} = \sum_{s} d_i y_i = \sum_{i=1}^{n} d_i y_i. \tag{2}$$

Jeżeli nie są znane wszystkie wartości zmiennej *y* dla jednostek wylosowanych do próby (na przykład na skutek braków odpowiedzi), estymator Horvitza-Thompsona charakteryzuje się znacznym obciążeniem i dużą wariancją. Wynika to na ogół z faktu, że braki odpowiedzi nie mają charakteru czysto losowego, a powstałe błędy wynikają z różnic pomiędzy respondentami i nierespondentami. Zmniejsza się ponadto efektywna liczebność próby, co w konsekwencji powoduje, że sumowanie we wzorze (2) nie odbywa się po zbiorze wszystkich jednostek, które miały wziąć udział w badaniu, a tylko po zbiorze respondentów $r \subseteq s$. Zakładać przy tym będziemy, że jest to zbiór *m*-elementowy, przy czym $m \leq n$. W efekcie ważona suma (2) jest najczęściej niedoszacowana w stosunku do prawdziwej wartości (1). W związku z tym wagi d_i powinny zostać odpowiednio skorygowane skalibrowane), aby zniwelować obciążenie wynikające z braków odpowiedzi.

Oznaczmy przez w_i poszukiwaną wagę (tzw. wagę kalibracyjną) odnoszącą się do *i*-tego respondenta, $i = 1, \ldots, m$. Naszym celem jest poszukanie wag w_i w taki sposób, aby były możliwie jak najbliższe co do wartości wyjściowym wagom d_i i aby niwelowały obciążenie będące konsekwencją występowania braków odpowiedzi. Konstrukcja wag kalibracyjnych uzależniona jest od wyboru odpowiedniej funkcji odległości. W literaturze przedmiotu na potrzeby wyznaczania wag kalibracyjnych przyjmuje się najczęściej tzw. funkcję odległości chi-kwadrat

$$D(\mathbf{w}, \mathbf{d}) = \frac{1}{2} \sum_{i=1}^{m} \frac{(w_i - d_i)^2}{d_i}, \tag{3}$$

gdyż umożlia to uzyskanie wektora wag kalibracyjnych w jawnej postaci. Można pokazać [Szymkowiak 2007], że dla tej funkcji odległości wektor wag kalibracyjnych wyraża się wzorem:

$$w_i = d_i + d_i \left(\mathbf{X} - \hat{\mathbf{X}} \right)^T \left(\sum_{i=1}^{m} d_i \mathbf{x}_i \mathbf{x}_i^T \right)^{-1} \mathbf{x}_i, \tag{4}$$

przy czym $\mathbf{X}$ to wektor utworzony z wartości globalnej każdej zmiennej pomocniczej $x_1, \ldots, x_k$ tj.

$$\mathbf{X} = \left(\sum_{i=1}^{N} x_{i1}, \sum_{i=1}^{N} x_{i2}, \ \ldots, \ \sum_{i=1}^{N} x_{ik} \right)^T, \tag{5}$$

$\hat{\mathbf{X}}$ jest wektorem złożonym z oszacowanych wartości globalnych zmiennych pomocniczych

$$\hat{\mathbf{X}} = \left(\sum_{i=1}^{m} d_i x_{i1}, \sum_{i=1}^{m} d_i x_{i2}, \ldots, \sum_{i=1}^{m} d_i x_{ik} \right)^T, \tag{6}$$

a

$$\mathbf{x}_i = \left(x_{i1}, \ldots, x_{ik} \right)^T \tag{7}$$

jest wektorem złożonym z wartości wszystkich k zmiennych pomocniczych dla *i*-tego respondenta, $i = 1, \ldots, m$. Estymator kalibracyjny wartości globalnej (1) wyraża się wówczas wzorem:

$$\hat{Y}_{\mathbf{X}} = \sum_{i=1}^{m} w_i y_i. \tag{8}$$

Jedną z zalet, wyznaczonych na podstawie funkcji odległości (3) wag kalibracyjnych (4), jest możliwość ich uzyskania wprost ze wzoru. Wagi te jednak w pewnych sytuacjach wykazują pewne niepożądane właściwości, tj. mogą być ujemne dla niektórych respondentów bądź odstające, tzn. znacznie różnić się od wag wyjściowych wynikających ze schematu losowania próby. Na wagi[1] nakłada się zatem czasami warunki ograniczające, tj. wymusza się, aby znajdowały się one w pewnym z góry określonym przedziale, co zapobiega występowaniu wag ujemnych i odstających. Takie podejście nie zapewnia zazwyczaj uzyskania wag w jawnej postaci i zachodzi potrzeba korzystania z metod iteracyjnych w poszukiwaniu wektora wag kalibracyjnych. Nie stanowi to oczywiście istotnej bariery, jednak w przypadku źle wyspecyfikowanych warunków ograniczających nie ma gwarancji, że algorytm poszukiwania wag kalibracyjnych, na które nałożono pewne restrykcje, osiągnie zbieżność. W związku z tym proponuje się wiele różnych funkcji odległości, uwzględniając przy tym (bądź nie) warunki ograniczające na iloraz wag $w_i \setminus d_i$. Prowadzi to, w zależności od przyjętej funkcji odległości, do wag kalibracyjnych w jawnej postaci bądź wymaga zastosowania metod numerycznych w poszukiwaniu przybliżonych rozwiązań. Poniżej przedstawiono omawiane najczęściej w literaturze przedmiotu funkcje odległości, które są wykorzystywane w badaniach statystycznych z brakami odpowiedzi w procesie poszukiwania wag kalibracyjnych [Devill, Särndal 1992; Pumputis 2005; Plikusas, Pumputis 2004].

W zależności od przyjętej funkcji odległości uzyskuje się różne postacie wag kalibracyjnych. W procesie ich poszukiwania wykorzystuje się przy tym metodę czynników nieoznaczonych Lagrange'a.

Końcową postać wag kalibracyjnych dla wybranych funkcji odległości (funkcja 1, 3, 6 i 7 w tab. 1) można znaleźć w pracy Pumputisa [2005] oraz Plikusasa i Pumputisa [2004]. W pracach tych można znaleźć również formuły na wariancję estymatorów kalibracyjnych w zależności od zastosowanej funkcji odległości. Szczegółowo omówione wyprowadzenie wag kalibracyjnych dla funkcji odległości chi-kwadrat można znaleźć również w pracy Szymkowiaka [2007]. Nie dla wszystkich jednak przedstawionych w tab. 1 funkcji odległości można wyznaczyć analityczną postać wag kalibracyjnych (na przykład dla funkcji 2). Należy ponadto podkreślić, że dla każdej z prezentowanych w tab. 1 funkcji odległości 1-9 można nałożyć ograniczenia na iloraz wag kalibracyjnych i wag wynikających ze schematu losowania próby. Przykładowo możemy przyjąć, że $0,8 \le w_i \setminus d_i \le 1,2$. Oznacza to, że wagi kalibracyjne w_i nie mogą się różnić od wag wynikających ze schematu losowania próby d_i o więcej niż 20%. Brak analitycznej postaci wag kalibracyjnych dla niektórych funkcji odległości oraz możliwość nałożenia na wagi pewnych ograniczeń powoduje, że należy stosować algorytmy iteracyjne w procesie

[1] Dokładniej na iloraz wag kalibracyjnych w_i i wag wynikających ze schematu losowania próby d_i, tj. $w_i \setminus d_i$.

poszukiwania optymalnych wag finalnych. Wykorzystuje się w tym celu najczęściej metodę Newtona-Raphsona bądź inne iteracyjne algorytmy rozwiązywania równań.

Tabela 1. Funkcje odległości w zagadnieniu wyznaczania wag kalibracyjnych

Nazwa	Postać funkcji odległości D(w,d)
Funkcja odległości chi-kwadrat	$D_1(\mathbf{w},\mathbf{d}) = \frac{1}{2}\sum_{i=1}^{m}\frac{(w_i - d_i)^2}{d_i}$
Zmodyfikowana funkcja odległości oparta na entropii	$D_2(\mathbf{w},\mathbf{d}) = \sum_{i=1}^{m}\left(w_i \log\frac{w_i}{d_i} - w_i + d_i\right)$
Funkcja odległości Hellingera	$D_3(\mathbf{w},\mathbf{d}) = 2\sum_{i=1}^{m}\left(\sqrt{w_i} - \sqrt{d_i}\right)^2$
Funkcja odległości oparta na entropii	$D_4(\mathbf{w},\mathbf{d}) = \sum_{i=1}^{m}\left(-d_i \log\frac{w_i}{d_i} + w_i - d_i\right)$
Zmodyfikowana funkcja odległości chi-kwadrat	$D_5(\mathbf{w},\mathbf{d}) = \frac{1}{2}\sum_{i=1}^{m}\frac{(w_i - d_i)^2}{w_i}$
Funkcja odległości Plikusasa 1	$D_6(\mathbf{w},\mathbf{d}) = \sum_{i=1}^{m}\left(\frac{w_i}{d_i} - 1\right)^2$
Funkcja odległości Plikusasa 2	$D_7(\mathbf{w},\mathbf{d}) = \sum_{i=1}^{m}\left(\frac{\sqrt{w_i}}{\sqrt{d_i}} - 1\right)^2$
Raking	$D_8(\mathbf{w},\mathbf{d}) = \sum_{i=1}^{m}\left(\frac{w_i}{d_i}\ln\frac{w_i}{d_i} - \frac{w_i}{d_i} + 1\right)$
Logitowa funkcja odległości	$D_9(\mathbf{w},\mathbf{d}) = A^{-1}\sum_{i=1}^{m}\left[\left(\frac{w_i}{d_i} - L\right)\ln\frac{\frac{w_i}{d_i} - L}{1 - L} + \left(U - \frac{w_i}{d_i}\right)\ln\frac{U - \frac{w_i}{d_i}}{U - 1}\right]$
	przy czym $A = \frac{U - L}{(U-1)(1-L)}$ a L i U oznaczają dolne i górne ograniczenie na iloraz wag $w_i \setminus d_i$.

Źródło: opracowanie własne.

Istnieje wiele programów, w których zaimplementowany został algorytm wyznaczania wag kalibracyjnych w zależności od przyjętej postaci funkcji odległości i warunków ograniczających. Większość z nich została napisana w języku 4GL w systemie SAS (CLAN 97, CALMAR, GES). Wyjątek stanowi G-CALIB, który został oprogramowany w programie SPSS. Duże możliwości w tym zakresie oferuje

również program R, w którym kalibracja została szeroko opisana w dwóch pakietach: *survey* i *sampling*.

3. Kalibracja w programie R

W programie R podejście kalibracyjne zostało zaimplementowane w dwóch pakietach: *survey* autorstwa Thomasa Lumleya oraz *sampling*, którego autorami są Yves Tillè i Alina Matei.

Tabela 2. Funkcje pakietu *sampling* związane z podejściem kalibracyjnym

calib(Xs, d, total, method, bounds, description, maxiter) – funkcja programu R wyznaczająca wagi kalibracyjne oraz umożliwiająca ich ocenę	
Argumenty funkcji	
Xs	macierz ze zmiennymi kalibracyjnymi (pomocniczymi)
d	wektor wag wejściowych (podlegających kalibracji)
total	wektor wartości globalnych zmiennych pomocniczych
method	metoda kalibracji (dostępne są cztery funkcje odległości: linear, raking, truncated, logit)
bounds	ograniczenia na wagi kalibracyjne
description	jeśli description=TRUE, tworzony jest raport podsumowujący wagi wejściowe i kalibracyjne, domyślnie description=FALSE
maxiter	liczba iteracji w algorytmie poszukiwania wag kalibracyjnych (niewymagana dla pierwszej metody)
checkcalibration(Xs, d, total, g, EPS) – funkcja programu R umożliwiająca sprawdzenie, czy algorytm wyznaczania wag kalibracyjnych osiągnął zbieżność. Argumenty podobne jak w funkcji calib, przy czym dodatkowy argument EPS określa dopuszczalne różnice między wartościami globalnymi zmiennych pomocniczych a oszacowanymi wartościami globalnymi tych zmiennych na podstawie wag kalibracyjnych.	

Źródło: opracowanie własne.

Ze względu na podobne zastosowanie funkcji wyznaczających wagi kalibracyjne w obydwu pakietach oraz argumenty wejściowe przedstawiona zostanie jedynie funkcja *calib* z pakietu *sampling* i funkcje powiązane umożliwiające dodatkową analizę wag kalibracyjnych (por. tab. 2).

Jednym z bardzo ważnych argumentów funkcji calib jest method, który może przyjmować jedną z 4 wartości, tj. linear, raking, truncated oraz logit w zależności od przyjętej funkcji odległości. Ustalając method=„linear", przyjmujemy funkcję odległości chi-kwadrat, method=„raking" oznacza przyjęcie funkcji odległości typu raking, method=„truncated" oznacza przyjęcie funkcji odległości chi-kwadrat, jednak dodatkowo zakłada się, że iloraz wag $w_i \setminus d_i$ powinien znajdować się w góry określonym przedziale (określamy to za pomocą argumentu bounds), method=„logit" oznacza z kolei przyjęcie logitowej funkcji odległości.

4. Przykład zastosowania pakietu *sampling*

Załóżmy, że celem pewnego badania jest oszacowanie łącznego miesięcznego dochodu osób w badanej populacji. Na potrzeby przykładu przyjęto, że dysponujemy informacjami z hipotetycznego badania reprezentacyjnego (por. tab. 3), w którym zebrano m.in. informacje na temat płci (zmienna plec: k – kobieta, m – mężczyzna), klasy miejscowości zamieszkania (zmienna klasa: m – miasto, w – wieś) oraz miesięcznego dochodu (zmienna dochod). Zakładamy przy tym, że łączny miesięczny dochód wszystkich osób wynosi 2 700 000 zł. Przyjmijmy ponadto, że próbę o liczebności $n = 20$ wylosowano z populacji składającej się z $N = 1000$ osób zgodnie ze schematem losowania prostego ze zwracaniem. Stąd wagi wejściowe d_i są równe $N/n = 1000/20 = 50$. Załóżmy ponadto, że w badanej populacji jest 500 mężczyzn i 500 kobiet oraz 600 osób z miasta i 400 ze wsi. Ponieważ dla części osób nie posiadamy informacji o ich miesięcznym dochodzie (NA – not avaiable), ważona suma wynosząca 2 285 000 zł wyznaczona po zbiorze wszystkich respondentów, dla których znany jest dochód, zgodnie z formułą 2 jest niedoszacowana w stosunku do prawdziwej wartości. Zgodnie z ideą kalibracji należy skorygować wagi dla respondentów, którzy podali informacje o dochodzie, tak aby odtworzone zostały znane struktury dla zmiennych pomocniczych i zniwelowane zostało obciążenie będące konsekwencją braków odpowiedzi.

Tabela 3. Przykładowy zbiór danych[2]

Lp.	plec	klasa	dochod	di	x1	x2	x3	w1	w2	w3	w4
1	m	m	2000	50	0	1	1	63,39	63,45	62,50	62,96
2	k	w	2500	50	1	0	0	54,24	54,31	53,12	53,70
4	m	w	4000	50	0	1	0	61,02	60,92	62,50	61,73
5	m	m	1500	50	0	1	1	63,39	63,45	62,50	62,96
6	m	m	3500	50	0	1	1	63,39	63,45	62,50	62,96
7	k	m	3700	50	1	0	1	56,61	56,55	57,50	57,04
8	k	w	5500	50	1	0	0	54,24	54,31	53,12	53,70
9	k	m	2400	50	1	0	1	56,61	56,55	57,50	57,04
10	m	m	2200	50	0	1	1	63,39	63,45	62,50	62,96
11	k	w	2800	50	1	0	0	54,24	54,31	53,12	53,70
12	m	w	3200	50	0	1	0	61,02	60,92	62,50	61,73
13	k	m	1600	50	1	0	1	56,61	56,55	57,50	57,04
15	m	m	1900	50	0	1	1	63,39	63,45	62,50	62,96
16	k	w	2100	50	1	0	0	54,24	54,31	53,12	53,70
17	k	m	1400	50	1	0	1	56,61	56,55	57,50	57,04
18	m	w	2500	50	0	1	0	61,02	60,92	62,50	61,73
20	k	m	2900	50	1	0	1	56,61	56,55	57,50	57,04

Źródło: opracowanie własne.

[2] W zbiorze danych ograniczono się jedynie do podania informacji o respondentach, tj. osobach, dla których znany był ich dochód. Brakuje więc osób o liczbie porządkowej 3, 14 i 19, dla których dochod=NA.

W rozważanym przykładzie w charakterze zmiennych pomocniczych wykorzystano płeć oraz klasę miejscowości. Utworzono przy tym 3 zmienne dychotomiczne (x_1 – 1, jeżeli osoba jest kobietą, 0 – w przeciwnym wypadku; x_2 – 1, jeżeli osoba jest mężczyzną, 0 – w przeciwnym przypadku; x_3 – 1, jeżeli osoba jest z miasta, 0 – w przeciwnym wypadku). Tak utworzone zmienne zagwarantują sumowalność odpowiednich wag kalibracyjnych w_i do faktycznej liczby kobiet i mężczyzn w populacji, a także do liczby osób zamieszkujących miasto oraz wieś[3].

Na potrzeby wyznaczania wag kalibracyjnych wykorzystano funkcję *calib* zaimplementowaną w programie R w pakiecie *sampling*. W pierwszej kolejności utworzono odpowiednie zbiory wejściowe (por. argumenty tej funkcji zawarte w tab. 2), a następnie wyznaczono wagi kalibracyjne dla wszystkich 4 funkcji odległości, które obsługiwane są przez funkcję *calib*. Składnia poleceń umożliwiająca uzyskanie wag kalibracyjnych jest następująca:

```
library(sampling)
# wczytanie danych
dane <- read.csv("d:/dane.csv", header = TRUE, sep = ";", dec =
",")
# Utworzenie zmiennych pomocniczych
dane$x1 <- ifelse(dane$plec == "k", 1, 0)
dane$x2 <- ifelse(dane$plec == "m", 1, 0)
dane$x3 <- ifelse(dane$klasa == "m", 1, 0)
# Ograniczenie zbioru danych do respondentów tj. osób, dla których
znany  jest dochód
dane_wej <- subset(dane, !is.na(dane$dochod))
# Utworzenie macierzy Xs
Xs <- cbind(dane_wej$x1, dane_wej$x2, dane_wej$x3)
# Utworzenie wektora wartości globalnych
total <- t(cbind(500, 500, 600))
# Utworzenie wektora wag kalibracyjnych dla 4 funkcji odległości
```

[3] Z formalnego punktu widzenia do opisu płci wystarczyłoby wziąć jedną zmienną dychotomiczną przyjmującą na przykład wartość 1 dla kobiet i 0 dla mężczyzn. W podejściu kalibracyjnym zagwarantowałoby to jedynie sumowalność wag kalibracyjnych do liczby kobiet w całej populacji, stąd wagi kalibracyjne dla mężczyzn nie musiałyby się sumować do liczby mężczyzn w populacji. Przyjęcie dwóch zmiennych dychotomicznych dla płci zagwarantuje sumowalność wag dla kobiet i mężczyzn do znanych wartości w populacji, a także zapewni sumowalność wag do łącznej liczby osób w populacji, tj. do 1000. Z tego względu dla zmiennej opisującej klasę miejscowości wystarczy utworzyć już tylko jedną zmienną dychotomiczną. Wówczas sumowalność wag kalibracyjnych do wszystkich osób w populacji i sumowalność wag dla osób z miasta do wszystkich osób zamieszkujących miasta wymusi sumowalność wag dla osób ze wsi do łącznej liczby osób mieszkających na wsi. Należy także podkreślić, że wagi korygowane są tylko dla tych respondentów, dla których znany jest miesięczny dochód (we wzorach na funkcje odległości D sumowanie odbywa się po zbiorze respondentów).

```
dane_wej$w1 <- dane_wej$di * calib(Xs, d = dane_wej$di, total, method = "linear") dane_wej$w2 <- dane_wej$di * calib(Xs, d = dane_wej$di, total, method = "raking") dane_wej$w3 <- dane_wej$di * calib(Xs, d = dane_wej$di, total, method = "truncated", bounds = c(0.75, 1.25))
dane_wej$w4 <- dane_wej$di * calib(Xs, d = dane_wej$di, total, method = "logit", bounds = c(0.7, 1.3))
print(dane_wej)
```

W wyniku zastosowanej składni poleceń otrzymano raport końcowy w programie R z wagami kalibracyjnymi dla wszystkich 4 funkcji odległości, który przedstawia tab. 3.

Wszystkie wagi kalibracyjne wyznaczone na postawie 4 różnych formuł pozwalają odtwarzać znane struktury demograficzne na poziomie całej populacji w odniesieniu do zmiennych płeć i klasa miejscowości zamieszkania. Oszacowane na ich podstawie, z wykorzystaniem czterech różnych funkcji odległości, łączne dochody kształtowały się odpowiednio: $\hat{Y}_{\mathbf{X}}^{linear} = 2\ 674\ 475$ zł; $\hat{Y}_{\mathbf{X}}^{raking} = 2\ 674\ 422$ zł; $\hat{Y}_{\mathbf{X}}^{truncated} = 2\ 674\ 879$ zł; $\hat{Y}_{\mathbf{X}}^{\log it} = 2\ 674\ 475$ zł. Wagi te więc, bez względu na przyjętą funkcję odległości, pozwalają dodatkowo redukować obciążenie będące konsekwencją braków odpowiedzi.

5. Podsumowanie

W artykule zaprezentowano metody wyznaczania wag kalibracyjnych dla różnych funkcji odległości, w przypadku gdy w badaniu występują braki odpowiedzi. Przedstawiona technika umożliwia redukcję obciążenia w sytuacji niepełnych danych i w związku z tym może być użyteczna w działalności wszystkich tych instytucji, które na co dzień zajmują się opracowywaniem wyników na podstawie badań ankietowych, do których jednostki dobierane są zgodnie z określonym schematem losowania. Dotyczyć to będzie przede wszystkim ośrodków badania opinii publicznej i Głównego Urzędu Statystycznego, które wykorzystują w wielu badaniach metodę reprezentacyjną na etapie projektowania próby oraz uogólniania wyników na podstawie danych, w których występują braki odpowiedzi.

Literatura

Deville J-C., Särndal C-E. (1992), *Calibration estimators in survey sampling*, „Journal of the American Statistical Association", vol. 87, pp. 376-382.

Pumputis D. (2005), *Calibrated estimators under different distance measures,* Proceedings of the Workshop on Survey Sampling Theory and Methodology 2005, pp. 137-141.

Plikusas A., Pumputis D. (2004), *Calibrated estimators of totals under different distance measures*, "Lietuvos Matematikos Rinkinys" 2004, vol. 44, special issue, pp. 572-576.

Särndal C-E., Lundström S. (2005), *Estimation in Surveys with Nonresponse*, John Wiley & Sons, Ltd.

Särndal C-E. (2007), *The calibration approach in survey theory and practice*, "Survey Methodology", vol. 33, no. 2, pp. 99-119.

Szymkowiak M. (2007), *Przyczynek do kalibracji w badaniach statystycznych z brakami odpowiedzi*, [w:] *Kapitał ludzki i wiedza w gospodarce. Wyzwania XXI wieku*, E. Panek (red.), Zeszyty Naukowe nr 96, Wydawnictwo Akademii Ekonomicznej w Poznaniu, Poznań 2007, s. 194-204.

CONSTRUCTION OF CALIBRATION ESTIMATORS OF TOTALS FOR DIFFERENT DISTANCE MEASURES

Summary: Missing data are one of the major types of non-random errors in statistical surveys. One of the methods proposed by Deville and Särndal [1992] which is designed to offset the negative effect of missing data is calibration, which is successfully used in practice by statistical offices of many countries. In its classical form calibration is a method in which calibrated weights are computed by minimizing a distance measure between the initial sampling weights and new weights, which need to satisfy certain calibration constraints. The main goal of this paper is to present the construction of calibration estimators of totals for different types of distance measures. Its empirical part, based on the calib function, which is available in R program in the sampling package, is devoted to the method of finding calibration weights in surveys with nonresponse for different distance measures.

Keywords: calibration, calibration weights, calibration estimators, nonresponse, distance measures.