PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS **nr 468 • 2017**

Taksonomia 28 ISSN 1899-3192
Klasyfikacja i analiza danych – teoria i zastosowania e-ISSN 2392-0041

Marcin Szymkowiak
Uniwersytet Ekonomiczny w Poznaniu
e-mail: m.szymkowiak@ue.poznan.pl

PODEJŚCIE KALIBRACYJNE WYKORZYSTUJĄCE ANALIZĘ SKŁADOWYCH GŁÓWNYCH W BADANIACH STATYSTYCZNYCH Z BRAKAMI ODPOWIEDZI

CALIBRATION APPROACH USING THE METHOD OF PRINCIPAL COMPONENT ANALYSIS IN SURVEYS WITH NONRESPONSE

DOI: 10.15611/pn.2017.468.24
JEL Classification: C83

Streszczenie: W badaniach statystycznych braki odpowiedzi stanowią jedno z głównych źródeł błędów nielosowych. W literaturze przedmiotu wskazuje się na dwie grupy metod statystycznych, których zastosowanie przyczynia się do eliminacji negatywnego wpływu braków odpowiedzi na proces estymacji nieznanych parametrów w populacji generalnej: imputacja i kalibracja. Szczególną rolę odgrywa tutaj kalibracja, która polega na skorygowaniu wag wynikających ze schematu losowania próby, tak aby odtworzone zostały wartości globalne wszystkich zmiennych pomocniczych. Głównym celem artykułu jest przedstawienie analizy składowych głównych jako metody doboru zmiennych pomocniczych w podejściu kalibracyjnym w badaniach statystycznych z brakami odpowiedzi. W artykule skoncentrowano uwagę na pokazaniu teoretycznych powiązań między podejściem kalibracyjnym a analizą składowych głównych z uwzględnieniem faktu, że w badaniu występować mogą braki odpowiedzi. W tym celu opisana zostanie konstrukcja odpowiednich wag kalibracyjnych wykorzystujących metodę składowych głównych.

Słowa kluczowe: analiza składowych głównych, podejście kalibracyjne, braki odpowiedzi.

Summary: Missing data are one of the major types of non-random errors in statistical surveys. Two main statistical techniques of eliminating the negative effect of nonresponse are considered in the literature: imputation and calibration. The main idea of calibration involves adjusting design weights in order to reproduce exactly known population totals of all auxiliary variables. The main purpose of this article is to present Principal Component Analysis (PCA) as a method for choosing auxiliary variables in the calibration approach in surveys with nonresponse. Special emphasis will be placed on showing theoretical links between the problem of finding calibration weights and the PCA methodology in the context of surveys with nonresponse. In particular, it will be shown how calibration weights can be constructed using PCA.

Keywords: principal component analysis, calibration approach, nonresponse.

1. Wstęp

W badaniach statystycznych występowanie błędów nielosowych w postaci braków odpowiedzi rodzi szereg problemów związanych z procesem estymacji nieznanych parametrów w populacji generalnej. Należy do nich zaliczyć przede wszystkim fakt znacznego wzrostu obciążenia i spadku precyzji stosowanych estymatorów. Jest to zazwyczaj konsekwencją różnic pomiędzy respondentami (osobami, które odpowiadają na pytania zawarte w ankiecie) a nierespondentami (osobami, które nie udzielają odpowiedzi na wszystkie lub wybrane pytania zawarte w ankiecie). Braki odpowiedzi nie mają bowiem na ogół charakteru losowego, a istnieje pewien ukryty mechanizm ich powstawania.

W literaturze przedmiotu zaproponowano różne metody radzenia sobie z brakami odpowiedzi. Generalnie dzielą się one na trzy grupy. Pierwszą z nich stanowią metody prewencyjne wywodzące się z nauk o zachowaniu się jednostek (psychologia oraz socjologia). Związane jest to z faktem, że proces zbierania danych wymaga kontaktu z respondentem. Metody prewencyjne obejmują odpowiednie przygotowanie kwestionariusza ankietowego i przeszkolenie ankietera oraz właściwy sposób przygotowania operatu losowania. Drugą grupę stanowią techniki redukujące frakcję braków odpowiedzi. Można do nich zaliczyć: wysyłanie monitów z prośbą o wzięcie udziału w badaniu, ponowny kontakt telefoniczny, stosowanie bodźców finansowych, zastępowanie jednostek, które nie wyrażają chęci wzięcia udziału w badaniu, innymi itd. Ostatnią grupę metod wykorzystywanych w badaniach z brakami odpowiedzi stanowią techniki statystyczne, do których można zaliczyć imputację czy metody wykorzystujące korektę wag, w tym kalibrację. Metody te są często wykorzystywane, gdyż na ogół techniki prewencyjne czy redukujące frakcję braków odpowiedzi nie eliminują problemu braku danych. W zależności od charakteru braków odpowiedzi (pozycyjne czy jednostkowe) wykorzystuje się imputację lub kalibrację [Piasecki 2014]. W przypadku pozycyjnych braków odpowiedzi (*item nonresponse*), tj. gdy jednostka, która wzięła udział w badaniu nie udzieliła odpowiedzi na niektóre pytania, zazwyczaj wykorzystywana jest imputacja w różnych odmianach. W sytuacji gdy mamy do czynienia z jednostkowym brakiem odpowiedzi (*unit nonresponse*), tj. gdy nie mamy odpowiedzi na wszystkie pytania w kwestionariuszu, stosuje się zwyczajowo korektę wag w postaci kalibracji. Możliwe jest także łączne wykorzystanie imputacji i kalibracji celem eliminacji negatywnych skutków braków odpowiedzi w badaniach statystycznych [Särndal, Lundström 2005].

Kalibracja jako metoda korygowania wag stosowana jest na szeroką skalę w badaniach reprezentacyjnych prowadzonych przez krajowe urzędy statystyczne i ośrodki badania rynku. Jej idea polega na wykorzystaniu zmiennych pomocniczych w taki sposób, aby odtworzyć poprzez system wag kalibracyjnych ich znane wartości globalne (ze spisów, rejestrów administracyjnych czy oszacowanych na podstawie danych pochodzących z samego badania reprezentacyjnego). Wybór zmiennych pomocniczych wpływa jednak na końcową postać wag kalibracyjnych, a w efekcie na cały proces estymacji. Z tego punktu widzenia szczególnego znaczenia nabiera zagadnienie odpowiedniego doboru zmiennych pomocniczych w podejściu kalibracyjnym. Szczególną rolę w tym obszarze zaczęła odgrywać metoda składowych głównych, która stanowi obecnie nowy nurt w zakresie wyznaczania wag kalibracyjnych.

Głównym celem artykułu jest przedstawienie analizy składowych głównych jako alternatywnej metody doboru zmiennych pomocniczych w podejściu kalibracyjnym. W ujęciu teoretycznym pokazane zostaną powiązania pomiędzy podejściem kalibracyjnym a analizą składowych głównych z uwzględnieniem faktu, że w badaniu występować mogą braki odpowiedzi.

2. Kalibracja w badaniach statystycznych z brakami odpowiedzi

Teoretyczne podstawy kalibracji w badaniach statystycznych zostały sformułowane przez Devilla i Särndala [1992]. Koncepcja wykorzystania podejścia kalibracyjnego w badaniach statystycznych z brakami odpowiedzi została następnie szeroko omówiona w monografii Särndala i Lundströma [2005]. Zgodnie z definicją zaproponowaną przez Särndala i Lundströma [2005] 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 efekcie jej zastosowania najczęściej udaje się zredukować obciążenie i wariancję wykorzystywanych w uogólnianiu wyników estymatorów. Bardziej formalny zapis zagadnienia wyznaczania wag kalibracyjnych został przedstawiony poniżej.

Niech dana będzie N-elementowa populacja $\mathrm{U} = \{1, \dots, \mathrm{N}\}$. Z populacji tej losujemy zgodnie z określonym schematem losowania n-elementową próbę $\mathrm{s} \subseteq \mathrm{U}$. Niech π_k oznacza prawdopodobieństwo inkluzji k-tej jednostki do próby, a $\mathrm{d_k} = {}^1/_{\pi_\mathrm{k}}$ odpowiednią wagę wyjściową wynikającą ze schematu jej losowania dla $\mathrm{k} = 1, \dots, \mathrm{N}$, która informuje o tym, ile jednostek z populacji jest reprezentowanych przez daną jednostkę z próby. Celem badania jest oszacowanie wartości globalnej zmiennej y, określonej wzorem:

$$\mathrm{Y} = \sum_{k=1}^{\mathrm{N}} \mathrm{y_k}, \tag{1}$$

gdzie y_k oznacza wartość zmiennej y dla k-tej jednostki badania, $k = 1, \dots, N$. Do szacowania wartości globalnej (1) bardzo często wykorzystuje się estymator Horvitza-Thompsona dany poniższym wzorem:

$$\hat{Y}_{HT} = \sum_{k=1}^{n} d_k y_k. \tag{2}$$

Jeżeli w badaniu nie uzyskaliśmy odpowiedzi od wszystkich jednostek biorących w nim udział w odniesieniu do zmiennej y wówczas sumowanie we wzorze (2) odbywa się jedynie po zbiorze respondentów $r \subseteq s$, którego liczebność m spełnia warunek $m < n$. Wówczas tak określona ważona suma (2) jest zazwyczaj niedoszacowana w stosunku do wartości globalnej (1). W takim przypadku należy dokonać korekty (kalibracji) wag wynikających ze schematu losowania próby d_k.

Niech $\mathbf{x} = (x_1, x_2, \dots, x_p)^T$ oznacza wektor złożony ze zmiennych pomocniczych wziętych pod uwagę w procesie wyznaczania wag kalibracyjnych, a $\mathbf{x}_k = \left(x_{1k}, x_{2k}, \dots, x_{pk}\right)^T$ niech oznacza wektor wartości zmiennych pomocniczych dla *k*-tej jednostki badania, $k = 1, \dots, n$. Zakładamy ponadto, że znany jest wektor wartości globalnych **X** złożony z wartości globalnych poszczególnych zmiennych pomocniczych, tj.

$$\mathbf{X} = \left(\sum_{k=1}^{N} x_{1k}, \sum_{k=1}^{N} x_{2k}, \dots, \sum_{k=1}^{N} x_{pk}\right)^T = \sum_{k \in U} \mathbf{x}_k. \tag{3}$$

Wartości tego wektora znane są najczęściej z badań pełnych, takich jak spisy, czy z innych źródeł, takich jak rejestry administracyjne. W przypadku gdy dla jakiejś zmiennej pomocniczej l nie jest znana jej wartość globalna, można ją zastąpić jej oszacowaniem z wykorzystaniem wzoru (2), tj. $\hat{X}^l_{HT} = \sum_{k=1}^{n} d_k x_{lk}$.

Niech w_k oznacza nową wagę (kalibracyjną), która zostanie wykorzystana w procesie estymacji wartości globalnej (1) z zastosowaniem formuły (2). Zgodnie z ideą kalibracji wagi w_k powinny być bliskie (w sensie przyjętej funkcji odległości) wagom d_k oraz powinny odtwarzać wartości globalne wszystkich zmiennych pomocniczych wziętych pod uwagę w podejściu kalibracyjnym.

Zadanie poszukiwania wag kalibracyjnych można sformułować w następujący sposób:

$$\min_{w_k} \sum_{k \in r} G_k(w_k, d_k), \tag{4}$$

przy spełnieniu poniższego warunku:

$$\sum_{k\in r} w_k \mathbf{x}_k = \sum_{k\in U} \mathbf{x}_k \tag{5}$$

określanego mianem równania kalibracyjnego, przy czym G_k jest pewną funkcją ściśle wypukłą, różniczkowalną, dla której $G(d_k, d_k) = 0$ oraz $G_k(1) = G_k{}'(1) = 0$. Wykorzystując metodę czynników nieoznaczonych Lagrange'a, można pokazać, że wektor wag kalibracyjnych wyraża się wzorem [Deville, Särndal 1992]:

$$w_k = d_k F_k(\boldsymbol{\lambda}^T \mathbf{u}_k), \tag{6}$$

przy czym F_k jest funkcją odwrotną do funkcji $G_k{}'(w_k, d_k)$, gdzie:

$$G_k'(w_k, d_k) = \frac{\partial G_k(w_k, d_k)}{\partial w_k}, \tag{7}$$

$\mathbf{u}_k$ jest tzw. wektorem zmiennych instrumentalnych, który w najczęstszych zastosowaniach jest określany jako $\mathbf{u}_k = \mathbf{x}_k$. W praktyce bardzo często przy wyznaczaniu wag kalibracyjnych korzysta się z tzw. funkcji chi-kwadrat G opisanej wzorem:

$$G_k(w_k, d_k) = \frac{(w_k - d_k)^2}{2d_k}, \tag{8}$$

dla której końcowa postać wag kalibracyjnych wyraża się wzorem:

$$w_k = d_k + d_k \boldsymbol{\lambda}^T \mathbf{u}_k, \tag{9}$$

przy czym wektor $\boldsymbol{\lambda}^T$ określony jest jako:

$$\boldsymbol{\lambda}^T = \left(\mathbf{X} - \sum_{\mathbf{r}} d_k \mathbf{x}_k\right)^T \left(\sum_{\mathbf{r}} d_k \mathbf{u}_k \mathbf{x}_k^T\right)^{-1}. \tag{10}$$

Ostatecznie estymator kalibracyjny wartości globalnej (1) jest postaci:

$$\hat{Y}_{cal} = \sum_{k=1}^{m} w_k y_k. \tag{11}$$

Wzór na wariancję estymatora (11) oraz sposób jej konstrukcji przedstawiony jest szczegółowo w pracy Särndala i Lundströma [2005].

3. Teoretyczne podstawy analizy składowych głównych

Analiza składowych głównych (*principal component analysis* – PCA) jest wielowymiarową metodą transformacji obserwowalnych zmiennych pierwotnych w nowe, wzajemnie ortogonalne zmienne, tzw. składowe główne. W metodzie tej, która jest jedną z najczęściej wykorzystywanych technik eksploracyjnej analizy danych, dokonujemy tzw. redukcji wymiarowości. Zastępujemy w niej bowiem oryginalny zbiór, zazwyczaj skorelowanych cech, przez stosunkowo małą liczbę nieskorelowanych zmiennych – składowych głównych (składowe główne są kombinacjami liniowymi wejściowych zmiennych), które wyjaśniają prawie całą zmienność danych. Pierwsza składowa główna wyznaczana jest w taki sposób, aby wyjaśnić najwięcej całkowitej zmienności. Z matematycznego punktu widzenia pierwszą składową główną można przedstawić jako kombinację liniową oryginalnych zmiennych $x_1, x_2, \dots, x_p$ w postaci:

$$Z_1 = a_{11}x_1 + a_{12}x_2 + \cdots + a_{1p}x_p = \mathbf{a}_1^T\mathbf{x}. \tag{12}$$

Wektor $\mathbf{a}_1$, który maksymalizuje wariancję $s_{Z_1}^2 = \mathbf{a}_1^T\mathbf{S}\mathbf{a}_1$, gdzie $\mathbf{S}$ jest macierzą kowariancji z próby i spełnia warunek $\mathbf{a}_1^T\mathbf{a}_1 = 1$, jest wektorem charakterystycznym odpowiadającym największej wartości własnej $\lambda_1$macierzy $\mathbf{S}$, tj. jest największym co do wartości pierwiastkiem równania:

$$|\mathbf{S} - \lambda\mathbf{I}| = 0. \tag{13}$$

Druga składowa z kolei wybierana jest tak, aby nie była skorelowana z pierwszą i wyjaśniała jak najwięcej pozostałej zmienności. W celu wyznaczenia drugiej składowej głównej Z_2 konstruujemy zatem kombinację liniową postaci:

$$Z_2 = a_{21}x_1 + a_{22}x_2 + \cdots + a_{2p}x_p = \mathbf{a}_2^T\mathbf{x} \tag{14}$$

w taki sposób, że jest ona nieskorelowana z Z_1, ma maksymalną wariancję $s_{Z_2}^2 = \mathbf{a}_2^T\mathbf{S}\mathbf{a}_2$, i spełnia warunek $\mathbf{a}_2^T\mathbf{a}_2 = 1$. Wektor $\mathbf{a}_2$ jest wektorem własnym macierzy $\mathbf{S}$, odpowiadającym drugiej wartości własnej $\lambda_2 < \lambda_1$, ortogonalnym do wektora $\mathbf{a}_1$ i unormowanym, tj. $\mathbf{a}_2^T\mathbf{a}_2 = 1$.

Proces ten kontynuuje się tak długo, aż otrzyma się p składowych głównych wzajemnie nieskorelowanych o wariancjach $\lambda_1 \geq \lambda_2, \dots, \geq \lambda_p$, przy czym w praktycznych zastosowaniach bierze się pod uwagę kilka pierwszych składowych głównych, które wyjaśniają dużą zmienność oryginalnych danych.

4. Konstrukcja wag kalibracyjnych wykorzystująca metodę analizy składowych głównych

W podejściu kalibracyjnym kluczową rolę w procesie wyznaczania wag w_k stanowi odpowiednio dobrany zestaw zmiennych pomocniczych. W przypadku dużej liczby potencjalnych zmiennych pomocniczych istotnym zagadnieniem jest zatem redukcja tego zbioru do najważniejszych zmiennych. W procesie doboru zmiennych pomocniczych oraz konstrukcji odpowiednich równań kalibracyjnych można wykorzystać przy tym zarówno kryteria merytoryczne, jak i statystyczne. Można również w podejściu kalibracyjnym w zakresie doboru zmiennych pomocniczych wykorzystać opisaną powyżej metodą analizy składowych głównych. Poniżej przedstawiony zostanie sposób wyznaczania wag kalibracyjnych z wykorzystaniem analizy składowych głównych.

Załóżmy w dalszej części, że jesteśmy zainteresowani oszacowaniem wartości globalnej (1). W tym celu chcemy skorzystać z estymatora kalibracyjnego określonego wzorem (11). Ze względu na istniejące w badaniu braki odpowiedzi w odniesieniu do zmiennej y należy skorygować wagi $\mathrm{d_k}$ wynikające ze schematu losowania próby. Załóżmy, że $\mathrm{x_1, x_2, \dots, x_p}$ są zmiennymi pomocniczymi, które wykorzystane mogą zostać w procesie poszukiwania wag kalibracyjnych $\mathrm{w_k}$. W oryginalnie sformułowanym problemie poszukiwania wag kalibracyjnych spełniony musi być warunek (5). Oznacza to, że wyznaczone wagi kalibracyjne $\mathrm{w_k}$ po ich zastosowaniu do wszystkich zmiennych pomocniczych $\mathrm{x_1, x_2, \dots, x_p}$ odtwarzają dokładnie ich znane wartości globalne (3). W podejściu kalibracyjnym wykorzystującym analizę składowych głównych w procesie kalibracji wag bierzemy pod uwagę kombinacje liniowe zmiennych $\mathrm{x_1, x_2, \dots, x_p}$ uzyskane na bazie tej techniki. Załóżmy w dalszej części, że spośród wszystkich p składowych głównych w dalszej analizie wzięto pod uwagę l pierwszych składowych głównych $\mathrm{Z_1, Z_2, \dots, Z_l, l < p}$. Problem wyznaczania wag kalibracyjnych dla nowych zmiennych pomocniczych, które stanowią składowe główne, można sformułować w następujący sposób:

- (W1) minimalizacja funkcji odległości:

$$\mathrm{D}\left(\mathbf{w}^{\mathrm{PCA}}, \mathbf{d}\right) = \frac{1}{2}\sum_{\mathrm{k=1}}^{\mathrm{m}} \frac{\left(\mathrm{w_k^{PCA}} - \mathrm{d_k}\right)^2}{\mathrm{d_k}} \rightarrow \min, \tag{15}$$

- (W2) równania kalibracyjne:

$$\sum_{\mathrm{k=1}}^{\mathrm{m}} \mathrm{w_k^{PCA}} \mathbf{z}_\mathrm{k} = \sum_{\mathrm{k=1}}^{\mathrm{N}} \mathbf{z}_\mathrm{k}, \tag{16}$$

gdzie $\mathbf{z}_\mathrm{k} = \left(\mathrm{z_{1k}, z_{2k}, \dots, z_{lk}}\right)^\mathrm{T}$ oraz $\mathrm{w_k^{PCA}}$ to waga kalibracyjna przypisana k-temu respondentowi i uzyskana z wykorzystaniem metody składowych głównych. Jest to zatem problem poszukiwania wag kalibracyjnych sformułowany podobnie do zagadnienia (4) przy warunku ograniczającym (5). Zauważmy, że w zaproponowanym podejściu nie wymaga się, aby znane były wartości globalne oryginalnych zmiennych pomocniczych $\mathrm{x_1, x_2, \dots, x_p}$ dane wzorem (3). Wymaga się jednak, aby znane były wartości jednostkowe wszystkich zmiennych pomocniczych $\mathrm{x_1, x_2, \dots, x_p}$ na poziomie populacji celem wyznaczenia odpowiednich wartości globalnych wszystkich składowych głównych wziętych w procesie konstrukcji wag kalibracyjnych.

Twierdzenie o wagach kalibracyjnych $\mathbf{w_k^{PCA}}$.

Rozwiązaniem zadania minimalizacji zadania (15) przy warunku (16) jest wektor wag kalibracyjnych $\mathbf{w}^{PCA} = \left(w_1^{PCA}, w_2^{PCA}, \dots, w_m^{PCA}\right)^T$, *którego składowe spełniają równanie*:

$$w_k^{PCA} = d_k + d_k \left(\sum_{k=1}^{N} \mathbf{z}_k - \hat{\mathbf{Z}} \right)^T \left(\sum_{k=1}^{m} d_k \mathbf{z}_k \mathbf{z}_k^T \right)^{-1} \mathbf{z}_k, \tag{17}$$

przy czym:

$$\hat{\mathbf{Z}} = \left(\sum_{k=1}^{m} d_k z_{1k}, \sum_{k=1}^{m} d_k z_{2k}, \dots, \sum_{k=1}^{m} d_k z_{lk} \right)^T. \tag{18}$$

Wagi kalibracyjne w_k^{PCA}, o których mowa w powyższym twierdzeniu, można łatwo wyprowadzić, korzystając z metody czynników nieoznaczonych Lagrange'a. Co więcej, wariancję tak otrzymanego estymatora kalibracyjnego, wykorzystującego w konstrukcji wag metodę analizy składowych głównych, można uzyskać analogicznie jak w przypadku estymatora (11) [Särndal, Lundström 2005].

Wagi kalibracyjne, o których mowa w powyższym twierdzeniu, można również otrzymać, wykorzystując inne znane w literaturze przedmiotu funkcje odległości. Przegląd wykorzystywanych w praktyce funkcji odległości można znaleźć m.in. w pracy [Deville, Särndal 1992]. Wykorzystanie innych, aniżeli w przedstawionym artykule, funkcji odległości prowadzi na ogół do nieco innych rozwiązań. Co więcej, dla niektórych funkcji odległości nie jest możliwe wyznaczenie wag kalibracyjnych w jawnej postaci i do ich poszukiwania wykorzystuje się metody numeryczne. Zaletą przedstawionej funkcji odległości chi-kwadrat jest to, że umożliwia uzyskanie wektora wag kalibracyjnych wprost ze wzoru. Z tego powodu jest ona w praktycznych badaniach prowadzonych przez krajowe urzędy statystyczne na całym świecie najczęściej wykorzystywana (*de facto* jej przyjęcie prowadzi do uzyskania tzw. estymatora GREG). Pewną wadą jest to, że możliwe jest uzyskanie wag ujemnych bądź ekstremalnych. Przy odpowiednim doborze zmiennych pomocniczych problem ten udaje się jednak bardzo często wyeliminować. Jest ona również, oprócz innych funkcji odległości wskazanych w pracy Deville'a i Särndala [1992], oprogramowana w wielu programach statystycznych. Przykładowo program R na potrzeby kalibracji oferuje pakiety *laeken*, *survey* czy *sampling*. Z kolei w programie SAS dedykowane kalibracji jest makro CALMAR, w którym ta funkcja odległości jest również zaimplementowana.

5. Podsumowanie

W artykule przedstawiono od strony teoretycznej konstrukcję wag estymatora kalibracyjnego wykorzystującego metodę analizy składowych głównych. Opisana koncepcja może być szczególnie interesująca w badaniach statystycznych z brakami odpowiedzi, gdzie ze względu na istnienie tego rodzaju błędów nielosowych, zachodzi potrzeba odpowiedniej korekty wag wynikających ze schematu losowania próby. Może być ona również przydatna w sytuacji, gdy dysponujemy „dużym" zbiorem potencjalnych zmiennych pomocniczych i zachodzi konieczność jego redukcji, a także gdy nie dysponujemy dla wszystkich rozważanych zmiennych pomocniczych ich wartościami globalnymi. Pewną wadę stanowi natomiast konieczność posiadania danych jednostkowych dla wszystkich oryginalnych zmiennych pomocniczych, tak aby możliwe było wyznaczenie wartości globalnych wszystkich składowych głównych.

Rozważania teoretyczne przedstawione w artykule warto zastosować w praktycznych badaniach prowadzonych przez Główny Urząd Statystyczny, gdzie braki odpowiedzi stanowią poważny problem (np. Badanie Budżetów Gospodarstw Domowych, Badanie Aktywności Ekonomicznej Ludności czy Europejskie Badanie Dochodów i Warunków Życia). Stanowić to będzie jednak przedmiot rozważań autora w kolejnych etapach i artykułach o charakterze naukowym. Jak pokazują nieliczne prace w tym zakresie, zastosowanie wag kalibracyjnych wykorzystujących koncepcję metody analizy składowych głównych może w badaniach z brakami odpowiedzi przyczynić się do redukcji obciążenia i wariancji estymatorów, będących skutkiem istnienia tych błędów nielosowych [Kassaye, Yigit 2012]. W tym miejscu warto podkreślić jednak, że prace te były prowadzone na sztucznie wygenerowanych zbiorach danych, które nie uwzględniają całej specyfiki danych rzeczywistych. Tym bardziej warto rozważyć w realnych badaniach opisaną koncepcję estymatorów kalibracyjnych wykorzystujących wagi otrzymane na bazie metody analizy składowych głównych. Będzie to przedmiotem prac badawczych autora w przyszłości.

Literatura

Deville J.-C., Särndal C.-E., 1992, *Calibration Estimators in Survey Sampling*, Journal of the American Statistical Association, vol. 87, s. 376-382.

Józefowski T., Szymkowiak M., 2012, *Estymatory kalibracyjne w badaniach statystycznych*, Wiadomości Statystyczne, nr 1, s. 31-43.

Kassaye M.H., Yigit D., 2012, *Calibration based on principal components*, Örebro Univeristy, Örebro Univeristy School of Business.

Piasecki T., 2014, *Metody imputacji w badaniach gospodarstw domowych*, Wiadomości Statystyczne, nr 9 (640), s. 1-20.

Särndal C.E., Lundström S., 2005, *Estimation in Surveys with Nonresponse*, Wiley, New York.

Szymkowiak M., 2009, *Imputacja i kalibracja – nowe możliwości estymacji w badaniach statystycznych z brakiem odpowiedzi*, Zeszyty Naukowe, nr 116, Wydawnictwo Uniwersytetu Ekonomicznego w Poznaniu.