Rozdział 1

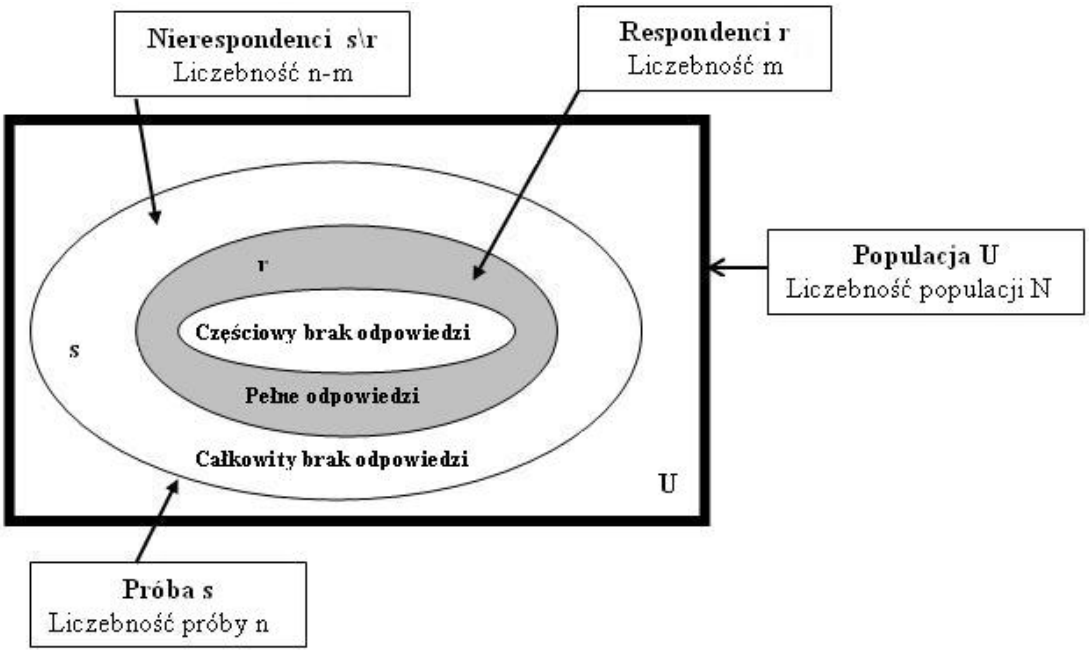
Przegląd metod estymacji w badaniach statystycznych z brakami odpowiedzi

1.1. Braki odpowiedzi jako główne źródło błędów nielosowych w badaniach statystycznych

W badaniach statystycznych jednym z głównych źródeł błędów nielosowych są braki odpowiedzi. Występują one zarówno w badaniach pełnych jak i częściowych. Chociaż może się wydawać, że w spisach ludności czy w sprawozdawczości statystycznej przypadki nieudzielania odpowiedzi są rzadsze aniżeli w badaniach dobrowolnych, takich jak na przykład, badanie budżetów gospodarstw domowych, to nie jest to jednak takie oczywiste. W badaniach towarzyszących spisowi, jak na przykład badanie dzietności kobiet NSP'1970 i NSP'1988 braki odpowiedzi sięgały 30%. Z kolei w badaniach budżetów gospodarstw domowych wskaźnik braku odpowiedzi wahał się w ostatnich latach od 30% do 50%.

Wśród braków danych można wyróżnić całkowity oraz częściowy brak odpowiedzi (por. rysunek 1.1)[3]. Z pierwszą sytuacją mamy do czynienia, gdy znane są tylko dane identyfikacyjne, a nie uzyskaliśmy żadnych informacji od badanej jednostki, na przykład na skutek odmowy bądź nieobecności w czasie przeprowadzenia badania. Z drugą sytuacją spotykamy się, gdy jednostki badania nie udzieliły odpowiedzi na niektóre pytania. Najczęściej niechęć podania informacji podyktowana jest drażliwością pytania lub obawą o ich wykorzystanie przeciwko respondentowi (pytania o plany lub zachowania prokreacyjne albo o wysokość dochodów). Istnieje zatem wiele powodów, dla których w badaniach występują braki odpowiedzi. Do najczęstszych należą niemożność wzięcia udziału w badaniu ze względu na wiek, chorobę, nieobecność w domu,

[3] W dalszym ciągu pracy będziemy stosować notację zgodną z oznaczeniami z rysunku 1.1. Przez N rozumieć będziemy liczebność populacji, n liczebność próby, a m liczebność zbioru respondentów. Same zbiory (populację, próbę oraz respondentów) oznaczać będziemy przez U, s i r odpowiednio.



Rysunek. 1.1. Zbiór respondentów i nierespondentów w badaniach statystycznych

Źródło: Uzupełnienie w oparciu o C-E. Särndal., S. Lundström (2005)

zmiana miejsca zamieszkania. Czynniki te mają charakter obiektywny. Istnieją również powody mające charakter subiektywny, a więc dana jednostka mogłaby wziąć udział w badaniu, ale ze względu na brak czasu czy niechęć do badania, odmawia udzielenia odpowiedzi.

Bez względu na przyczyny, jakie towarzyszą brakom odpowiedzi, ich występowanie jest źródłem wielu „zaburzeń". Jest tak dlatego, że osoby, które odmawiają wzięcia udziału w badaniu bądź nie udzielają na niektóre pytania odpowiedzi, na ogół różnią się od tych, co biorą w nim udział i dostarczają niezbędnych danych. Wskutek tego:

1. Zmniejsza się efektywny rozmiar badanej próby bądź populacji, co ma niekorzystny wpływ na wariancję estymatorów powodując ich zwiększenie.
2. Uzyskane wyniki obciążone są zbyt dużymi błędami. Wyznaczone oceny parametrów znacznie odbiegają od ich „prawdziwych" wartości, a skonstruowane na podstawie próby przedziały ufności różnych parametrów, koncentrują się wokół „złych" wartości.
3. Rozkłady wielu cech są zniekształcone i niemożliwe będzie zastosowanie wielu klasycznych metod statystycznych.
4. Zbyt niski wskaźnik udzielonych odpowiedzi nie wpływa korzystnie na pozytywne postrzeganie badania przez jego użytkowników i w skrajnych przypadkach może się ono okazać dla nich całkowicie bezużyteczne.

W praktyce badań statystycznych stosuje się różnego rodzaju metody, których celem jest zwiększenie frakcji udzielonych odpowiedzi. Mają one zarówno zastosowanie

na etapie zbierania danych (na przykład powtórne badanie jednostek, od których nie uzyskano danych, zastępowanie jednostek nie podejmujących badania innymi, stosowanie różnych bodźców — na przykład finansowych) oraz na etapie ich opracowywania (na przykład imputacja, kalibracja).

Generalnie metody te można podzielić na trzy zasadnicze grupy: prewencyjne, redukujące frakcję braków odpowiedzi oraz korygujące. Granica pomiędzy poszczególnymi technikami w ramach wyróżnionych grup nie zawsze jest ostra, przy czym można jednak przyjąć w ogólności, że podejście prewencyjne ma miejsce na etapie planowania badania przed zebraniem danych, redukcja braków odpowiedzi odbywa się na etapie ich zbierania, a korygowanie odbywa się w procesie estymacji, kiedy zebrano już niezbędne informacje, por. S. Tíngdahl (2004).

Metody prewencyjne, mające zapobiegać występowaniu braków odpowiedzi w badaniach statystycznych, wywodzą się z nauk o zachowaniu się jednostek (psychologii, socjologii) — co jest naturalną konsekwencją faktu, że proces zbierania danych wymaga kontaktu z respondentem. Niezbędna jest więc tutaj znajomość technik mających przełamać sceptycyzm i niechęć respondenta do udzielania informacji oraz promujących pozytywne nastawienie do całego badania. Dużą rolę odgrywają w ramach tej grupy metod, czynniki motywacyjne mające przekonać jednostkę do wzięcia udziału w badaniu[4]. Metody prewencyjne obejmują również zagadnienie konstrukcji kwestionariusza ankietowego, odpowiednie przeszkolenie ankietera, sposób zbierania danych oraz właściwe przygotowanie operatu losowania.

Metody redukujące frakcję braków odpowiedzi obejmują m.in. 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. W przypadku stosowania zastępowania, zwykle jednostki zastępcze wybiera się z próby rezerwowej kierując się zasadą, aby miały one podobne cechy podstawowe jak jednostki nie podejmujące badań. Nie jest to jednak regułą, gdyż w badaniu budżetów gospodarstw domowych, jednostki zastępcze losuje się, a więc mogą się one diametralnie różnić od jednostek wylosowanych pierwotnie do próby. Podobnie jak metody prewencyjne, w znacznej mierze wywodzą się one z nauk o zachowaniu się jednostek.

Trzecia grupa obejmuje różnego rodzaju metody estymacji i ważenia danych, których celem jest zniwelowanie obciążenia będącego konsekwencją wystąpienia w badaniu braków odpowiedzi. Z racji tego, że w każdym — nawet najlepiej zaplanowanym badaniu — występują braki danych, metody statystyczne, rozwijane w ramach tej grupy, odgrywają coraz większą rolę[5]. Szczególną rolę pełnią tutaj różnego rodzaju metody i techniki oparte o system wag – w tym podejście kalibracyjne.

[4] Przykładowo w badaniu budżetów gospodarstw domowych stosuje się bodźce finansowe dla gospodarstw, które biorą w nim udział.

[5] Metody prewencyjne i redukujące frakcję braków odpowiedzi znajdują się poza głównym nurtem rozważań pracy. Istnieje jednak bardzo bogata literatura, która poświęcona jest metodom zapobiegającym występowaniu braków danych na etapie planowania badania i ich zbierania, por. P. Campanelli (1997), C. F Cannell, P. V Miller i L. Oksenberg (1981), D.A. Dillman, J.J Eltinge, R.M Groves, i R.J.A Little (2002), B. Knäuper, R.F Belli, D.H Hill, A.R Herzog (1997), J. Kordos (1988), E.D Leeuw, J. Hox, i M. Huisman (2003). W pracy główny nacisk położony zostanie na metody estymacji w badaniach z brakami odpowiedzi, przy czym szczegółowo zajmujemy się tylko kalibracją, która stanowi najnowsze osiągnięcie metodologiczne w tym zakresie.

W światowej literaturze przedmiotu „nonresponse" przeszedł znamienną ewolucję – od podejścia polegającego na ograniczeniu się w procesie estymacji tylko do zbioru tych jednostek, dla których znane są wartości analizowanych cech[6], poprzez wysiłki uzyskania odpowiedzi „za wszelką cenę", aż do estymacji wykorzystującej alternatywne źródła informacji pośredniej.

W literaturze przedmiotu przedstawia się dwie podstawowe metody stosowane w przypadku wystąpienia braków odpowiedzi w badaniach statystycznych: imputację i kalibrację, por. C-E. Särndal, S. Lundström (2005). Pierwsza polega na zastąpieniu brakujących danych konkretnymi wartościami celem uzyskania kompletnego zbioru danych. Druga polega na odpowiednim ustaleniu wag, tak aby zredukować obciążenie wynikające z istnienia braków odpowiedzi. Metodom tym poświęcone będą kolejne podrozdziały pracy.

1.2. Imputacja

W badaniach statystycznych, metody imputacji rozwinęły się na potrzeby spisów przeprowadzanych w różnych państwach na przełomie lat 50–tych XX wieku. Przykładowo, w kanadyjskim spisie ludności z 1950 roku stosowano tzw. metodę Deminga dla szacowania brakujących danych, bazującą na rozkładzie częstości wartości cech w oparciu o wcześniej zgromadzone dane z poprzednich spisów.

Rozwój metod imputacji możliwy był dzięki postępowi, jaki miał miejsce w informatyce w latach 60–tych XX wieku. Jednym z pierwszych zastosowań komputera, w procesie szacowania brakujących danych, był spis powszechny w Stanach Zjednoczonych z 1960 roku. W spisie tym po raz pierwszy na szeroką skalę zastosowano imputację hot-deck w miejsce stosowanej wcześniej metody cold-deck. Podobnie poczyniono w kanadyjskim spisie powszechnym z 1961 roku, w którym wykorzystanie komputerów umożliwiło losowe uzupełnianie brakujących danych w oparciu o rekordy, dla których odpowiednie informacje istniały.

Intensywny rozwój teorii w zakresie imputacji miał miejsce w latach 80–tych XX wieku. Istotny wkład w tym zakresie miała pionierska praca Rubina (1976) oraz Little'a i Rubina (1987), w których przedstawiono po raz pierwszy w kompleksowy sposób metody imputacji, które następnie stosowano z powodzeniem w wielu badaniach statystycznych.

Jeszcze do niedawna niektórzy statystycy określali imputację mianem „ostatniej deski ratunku" w badaniach z brakami odpowiedzi, której stosowanie w praktyce może przynieść więcej szkód niż pożytku. W ostatnim czasie — dzięki intensywnemu rozwojowi teorii i odpowiedniego oprogramowania — panuje jednak powszechny pogląd, że

[6] Tylko w niektórych przypadkach można zastosować podejście polegające na pominięciu braków odpowiedzi i ograniczeniu się do jednostek, od których uzyskaliśmy niezbędne dane (na przykład gdy frakcja braków odpowiedzi jest niewielka bądź gdy istniał pewien losowy mechanizm generowania braków odpowiedzi). W badaniach statystycznych taki „zrandomizowany" mechanizm generowania braków odpowiedzi jednak nie występuje. Wynika to z faktu, że istnieją zazwyczaj istotne różnice między respondentami i nierespondentami. Dlatego przyjęcie założenia, że braki odpowiedzi mają charakter losowy, byłoby źródłem wielu błędów. Bardziej odpowiednie wydaje się więc uwzględnienie faktu, że dla niektórych obiektów brak jest całkowicie bądź częściowo danych i dokonanie próby ich wyszacowania bądź skorygowania uzyskanych wyników – w oparciu o odpowiednio dobrane wagi.

odpowiednio dobrana i zastosowana metoda imputacji może stanowić swego rodzaju remedium na występujące w badaniach braki danych, por. R. Ren (2002).

Poniżej przedstawione zostaną najczęściej wykorzystywane w praktyce metody szacowania brakujących danych. W pierwszej jednak kolejności zdefiniujemy imputację, wskazując jednocześnie na pewne pożądane jej własności.

Definicja 1 (C-E. Särndal, S. Lundström, 2005). *Imputacja jest to proces szacowania brakujących lub eliminowania niepoprawnych danych, oparty na wykrytych relacjach w zbiorze wartości tych samych lub innych zmiennych (lub obserwacji), dla których danych nie brakuje.*

Z powyższej definicji wynika, że zastosowanie imputacji prowadzi do przypisania każdej jednostce w miejsce brakujących lub nieważnych danych jakiejś wartości. Oznacza to, że brakujące dane uzupełniane są ich „substytutami" i są one z samej definicji „wartościami sztucznymi". Należy jednak podkreślić, że aby imputacja odegrała swoją rolę w badaniu muszą być spełnione trzy ważne założenia:

- imputacja nie powinna prowadzić do obciążeń bądź zmian rozkładów cech w zbiorze danych oraz do wzrostu wariancji stosowanych estymatorów,
- proces imputacji w większym stopniu powinien być uzależniony od danych pochodzących z próby aniżeli odwoływać się do założeń, co do natury brakujących danych,
- oszacowania ważnych statystyk z próby nie powinny „zbyt mocno" opierać się na imputowanych danych.

W praktyce badań statystycznych bardzo trudno jest dochować powyższych założeń. Należy ponadto zachować szczególną uwagę operując na zbiorze danych, wśród których znajdują się również dane imputowane. Nierozważne użycie imputacji może poważnie zniekształcić uzyskane wyniki, co z kolei może być źródłem źle wyciągniętych wniosków. W literaturze przedmiotu, jak i w badaniach statystycznych, wykorzystywanych jest wiele różnych metod imputacji. Imputowane wartości można zaklasyfikować do jednej z trzech głównych kategorii, por. C-E. Särndal, S. Lundström (2005):

- wartości imputowane z wykorzystaniem statystycznych reguł predykcyjnych,
- wartości imputowane uzyskiwane od jednostek badania mających podobne cechy,
- wartości imputowane w oparciu o opinię ekspertów.

Dwie pierwsze kategorie mogą być nazwane wspólnym terminem „imputacyjnych reguł statystycznych", ponieważ w procesie wyznaczania substytutów, wykorzystywane są różnego rodzaju narzędzia i techniki statystyczne. W ramach pierwszej kategorii wykorzystuje się relacje zachodzące pomiędzy zmienną imputowaną i innymi zmiennymi. Natomiast w drugim przypadku wykorzystywana jest technika „dawca-biorca", w której obiekt, dla którego imputujemy wartości jakiejś zmiennej, „pożycza" wartości od innych, bardzo podobnych obiektów. Trzecia kategoria obejmuje z kolei metody oparte na wiedzy i doświadczeniu specjalistów z zakresu danego badania.

Dokonując innego rozróżnienia możemy traktować imputowane wartości jako losowe, (gdy procedura imputacyjna przypisuje różne na ogół wartości zmiennym imputowanym dla różnych obiektów — na przykład imputacja typu hot-deck) oraz mające charakter deterministyczny, (kiedy różnym obiektom dla imputowanych zmiennych

przypisywane są te same wartości — na przykład imputacja z wykorzystaniem średniej).

Losowe przypisanie wartości jakiemuś obiektowi może nastąpić od dowolnej innej jednostki badania bądź od jednostki, która została wylosowana z utworzonej wcześniej tak zwanej homogenicznej grupy respondentów [HGR], utworzonej w oparciu o pewien zestaw cech wspólnych. Podobnie przypisanie średniej dla jakiegoś obiektu, dla którego brak informacji o jakiejś cesze może nastąpić w oparciu o wyliczoną średnią dla wszystkich innych jednostek badania, dla których takie dane posiadamy bądź można się ograniczyć do jej policzenia w ramach odpowiedniej grupy respondentów.

Należy jednak podkreślić, że zastosowanie imputacji nie daje gwarancji, że uzyskiwane wyniki będą mniej obciążone w porównaniu z wynikami, które uzyskalibyśmy, gdyby nie miało miejsce „fabrykowanie" danych. Dlatego należy imputację stosować z dużą rozwagą, kierując się przy tym doświadczeniem badawczym, intuicją oraz relacjami wykrytymi w zbiorze danych. Imputowane wartości powinny być bowiem „bliskie"prawdziwym, choć nieznanym na skutek braku odpowiedzi rzeczywistym wartościom.

W literaturze, jak i praktyce badań statystycznych, wykorzystywane są najczęściej następujące metody imputacji, por. C-E. Särndal, S. Lundström (2005), N.T Longford (2005):[7]

Imputacja dedukcyjna – metoda stosowana w przypadku, gdy brakujące dane można wyszacować w drodze dedukcji na podstawie innych informacji, które udało się uzyskać w wyniku badania. Jest ona bardzo popularna i często stosowana (na przykład, jeśli nie ma informacji o płci badanej osoby, to wiedząc że nosi ona imię żeńskie i jest zamężna, wiadomo że jest kobietą[8]).

Imputacja typu cold-deck – metoda polegająca na zastąpieniu brakujących danych wartościami spoza próby — ze źródeł zewnętrznych (rejestrów administracyjnych, spisu) lub z badań poprzednich.

Imputacja regresyjna – metoda zastępowania brakujących danych w oparciu o wartości uzyskane z odpowiednio dobranego modelu regresji. Imputowaną wartością może być wartość wprost z modelu bądź też wartość regresyjna z uwzględnieniem składnika resztowego. Wprowadźmy następujące oznaczenia:
y_k - wartość zmiennej y dla k-tej jednostki badania (zakładamy, że występuje brak danych dla tej zmiennej w k-tym obiekcie badania),
$\hat{y}_k$ - wartość zmiennej y dla k-tej jednostki badania po zastosowaniu imputacji,
$\mathbf{x}_k$ - macierz wartości zmiennych objaśniających (zakładamy, że dla wszystkich

[7] Niektórzy autorzy zaliczają do metod imputacji techniki, w których analizie poddawane są tylko te obiekty, dla których dostępne są wszystkie dane dla wszystkich zmiennych (innymi słowy dokonywana jest eliminacja obiektów, dla których występują dla jakichś zmiennych braki odpowiedzi) bądź biorą przy szacowaniu różnych parametrów dla jakiejś cechy te obiekty, dla których znane są wartości tej zmiennej. Ponieważ w tym przypadku nie występuje oszacowanie brakujących danych, więc tej techniki, nie będziemy zgodnie z przedstawioną definicją, zaliczać do metod imputacji.

[8] W przypadku badań prowadzonych w Polsce wystarczyłaby właściwie informacja, że badana osoba nosi imię żeńskie, żeby stwierdzić, że jest kobietą, niemniej jednak w niektórych krajach informacja o samym imieniu mogłaby być niewystarczająca do stwierdzenia płci badanej osoby ze względu na fakt, że to samo imię może nosić zarówno kobieta jak i mężczyzna.

jednostek badania znane są wartości wszystkich zmiennych objaśniających).
W metodzie tej wartość imputowana dla brakującej danej wyraża się wzorem:

$$\hat{y}_k = \boldsymbol{x}_k^T \hat{\boldsymbol{\beta}}_i, \tag{1.1}$$

gdzie:

$$\hat{\boldsymbol{\beta}}_i = \left(\sum_{r_i} a_k \boldsymbol{x}_k \boldsymbol{x}_k^T \right)^{-1} \sum_{r_i} a_k \boldsymbol{x}_k \boldsymbol{y}_k \tag{1.2}$$

Współczynniki regresji $\hat{\boldsymbol{\beta}}_i$ otrzymujemy stosując klasyczną ważoną metodę najmniejszych kwadratów w oparciu o dane uzyskane od respondentów, dla których znane są wartości zmiennej objaśnianej i zmiennych objaśniających.

Imputacja z wykorzystaniem średniej – jest to metoda zastąpienia brakujących danych przez średnią wartość cechy, która jest obliczana dla wszystkich jednostek, od których uzyskano odpowiedzi lub od części jednostek po wcześniejszym ich przydzieleniu do odpowiednich klas imputacyjnych według wartości zmiennych klasyfikujących[9]. Metoda ta jest bardzo prosta w zastosowaniach, wykazuje jednak pewne słabości. Rozkład cechy w wyniku zastąpienia brakujących danych wartościami średnimi zniekształca rzeczywisty rozkład cechy, parametry wyznaczone w oparciu o tę metodę mogą znacznie się różnić od ich prawdziwych wartości, a niektóre z nich mogą w ogóle stracić swoją wartość poznawczą (na przykład odchylenie standardowe czy współczynnik zmienności na skutek redukcji zmienności badanej cechy). Stosując ten rodzaj imputacji popełniamy błąd systematyczny.

Predykcyjne dopasowane według średniej – jest to pewna odmiana imputacji regresyjnej, w której nierespondentowi jest dobierany respondent na zasadzie najbliższej wartości uzyskanej w oparciu o wyznaczony model regresji, przy czym zamiast wartości regresyjnej nierespondentowi przypisuje się wartość respondenta.

Imputacja z wykorzystaniem innej zmiennej – dla danej zmiennej X, dla której brakuje odpowiedzi dla niektórych obiektów poszukiwana jest zmienna Y, która jest blisko związana ze zmienną X i może być uważana za „substytut" zmiennej X. Brakujące dane dla obiektów dla zmiennej X uzyskuje się w oparciu o wartości jakie ma zmienna Y.

Imputacja metodą najbliższego sąsiada – w metodzie tej imputowaną wartością dla cechy y dla k-tego obiektu badania jest $\hat{y}_k = y_{l(k)}$, gdzie $l(k)$ jest dawcą dla tego obiektu. Idea tej metody polega na tym, że skoro dwa obiekty mają zbliżone lub te same wartości dla pewnej grupy cech (to samo wykształcenie, płeć, wiek, itd.) to powinny mieć również zbliżone wartości dla cechy y. Dawcą jest obiekt $l(k)$, który należy do zbioru r wszystkich respondentów, i dla którego wybrana funkcja odległości przyjmuje wartość najmniejszą. Dawcą jest więc obiekt, dla którego wielowymiarowa funkcja odległości obliczona pomiędzy wszystkimi obiektami dawcami, a obiektem biorcą przyjmuje wartość minimalną. Zakładamy przy tym, że odległość tę liczymy w oparciu o pewne zmienne, które są znane dla wszystkich

[9] Pojęcie klasy (grupy) imputacyjnej odnosi się do rozłącznych zbiorów, których elementami są jednostki badania stanowiące obiekty o podobnych cechach. W ten sposób powstają klasy obiektów w miarę jednorodnych. Głównym celem tworzenia klas jest fakt, że inne relacje zachodzą w różnych podgrupach wylosowanej próby. W badaniach społecznych klasy takie często tworzy się w oparciu o takie zmienne klasyfikujące jak wiek, wykształcenie czy płeć.

dawców jak i biorcy. Biorca przyjmuje zatem wartość cechy y od dawcy, który minimalizuje odległość liczoną w oparciu o miarę odległości D_{lk}. W przypadku, gdy dawcy będziemy szukać posługując się tylko jedną zmienną, możemy wykorzystać funkcję odległości postaci:

$$D_{lk} = |x_l - x_k| . \tag{1.3}$$

W przypadku wielowymiarowym, gdy dawcy szukamy w oparciu o zbiór J zmiennych, wygodną w zastosowaniu jest metryka postaci:

$$D_{lk} = \sqrt{\sum_{j=1}^{J} (x_{jl} - x_{jk})^2}, \tag{1.4}$$

gdzie x_{ji} jest wartością j-tej cechy dla i-tego obiektu badania.

Imputacja typu hot-deck – w metodzie tej imputowaną wartością dla cechy y dla k-tego obiektu jest $\hat{y}_k = y_{l(k)}$, gdzie $l(k)$ jest dawcą dla tego obiektu losowo wybranym spośród wszystkich obiektów, z kompletnym rekordem danych bądź spośród takich obiektów, które należą do tej samej klasy imputacyjnej. Rozkład wartości cechy y po tak zastosowanym uzupełnieniu brakujących danych wygląda całkiem „naturalnie", ale wciąż może różnić się w znaczący sposób od rozkładu cechy jaki uzyskalibyśmy, gdyby wszystkie jednostki badania z próby s udzieliły odpowiedzi na pytanie odnoszące się do zmiennej y. Wynika to z faktu, że respondenci jak i nierespondenci mogą się różnić w odniesieniu do takich parametrów jak średnia, odchylenie standardowe itd.

Imputacja w oparciu o opinie ekspertów – w procesie imputacji brakujących informacji wykorzystuje się ekspertów, którzy mając wiedzę na temat badanej populacji oraz wartości jakie mogłyby przyjmować poszczególne zmienne, studiując uważnie poszczególne rekordy, są w stanie zaproponować realistyczne wartości w miejsce brakujących danych.

Główny Urząd Statystyczny wykorzystywał imputację w kilku prowadzonych przez siebie badaniach. Na szeroką skalę zastosowana ona została w prowadzonym przez państwa członkowskie Unii Europejskiej badaniu EU-SILC, a wdrożonym przez GUS w Polsce w 2005 roku, por. GUS (2008). EU-SILC jest europejskim badaniem dochodów i warunków życia, którego podstawowym celem jest dostarczenie porównywalnych dla Unii Europejskiej danych dotyczących dochodów, ubóstwa oraz zjawiska społecznego wykluczenia. W prowadzonym przez GUS badaniu, jednostką badania jest gospodarstwo domowe oraz wszyscy członkowie gospodarstwa, którzy do dnia 31 grudnia, w roku poprzedzającym badanie, ukończyli 16 lat. Informacje dotyczące sytuacji całego gospodarstwa domowego spisywane są na specjalnym kwestionariuszu gospodarstwa domowego (EU-SILC-1G), natomiast informacje dotyczące osób w wieku 16 lat i więcej – na kwestionariuszu indywidualnym (EU-SILC-1I).

Realizowane przez GUS europejskie badanie warunków życia ma charakter dobrowolny i prowadzone jest techniką bezpośredniego wywiadu z respondentem z zastosowaniem kwestionariuszy papierowych (tzw. metoda PAPI). W przypadku wywiadu indywidualnego, dopuszczało się ponadto realizację tzw. wywiadu zastępczego, przeprowadzonego z inną osobą z gospodarstwa domowego, która mogła udzielić wiarygodnych informacji o osobie objętej badaniem (dotyczy to osób zaliczonych do składu

gospodarstwa domowego, a nieobecnych w miejscu zamieszkania w okresie trwania badania).

W celu wyboru próby zastosowano schemat losowania dwustopniowego, warstwowego z różnymi prawdopodobieństwami wyboru na pierwszym stopniu — podobnie jak w badaniu budżetów gospodarstw domowych. Jako operat losowania wykorzystano Urzędowy Rejestr Podziału Terytorialnego Kraju TERYT, przy czym jednostkami losowania pierwszego stopnia były obwody spisowe, zaś na drugim stopniu losowano mieszkania, w ramach których badane były wszystkie gospodarstwa domowe (razem w badaniu udział wzięło 16 263 gospodarstw domowych).

Ze względu na dobrowolny charakter badania, jednym z głównych źródeł błędów były braki odpowiedzi. W przypadku kwestionariusza gospodarstwa domowego, wskaźnik całkowitego braku odpowiedzi wyniósł 30%, a w przypadku kwestionariusza indywidualnego 33%. Tak wysoka frakcja braków odpowiedzi wymusiła konieczność imputacji danych.

W zależności od rodzaju i charakteru brakujących informacji w badaniu EU-SILC, zrealizowanym w 2005 roku, zastosowano różne metody imputacji: metodę hot-deck, imputację regresyjną oraz imputację dedukcyjną[10].

Zastosowanie metody hot-deck przeprowadzono w ramach klas imputacyjnych, które zostały stworzone w oparciu o wartości zmiennych kategoryzujących. W przypadku braków dawców, w obrębie danej klasy imputacyjnej, stosowano podejście sekwencyjne polegające na ograniczaniu zmiennych kategoryzujacych poprzez stopniową eliminację zmiennych najmniej ważnych, aż do momentu, w którym klasa imputacyjna była niepusta.

W przypadku imputacji regresyjnej, zastosowane zostały dwie jej odmiany: imputacja regresyjna z losowymi resztami rzeczywistymi i imputacja regresyjna deterministyczna. W odniesieniu do pierwszej z nich przyjmowano logarytmiczną bądź wykładniczą funkcję regresji. Wartością imputowaną dochodu była suma wartości teoretycznej otrzymanej z modelu i reszty wylosowanej spośród rzeczywistych reszt otrzymanych przy jego estymacji, przy czym zbiór rekordów, spośród których losowana była reszta ograniczany był do najbliższych rekordowi imputowanemu, ze względu na wartość teoretyczną uzyskaną z modelu. W przypadku imputacji regresyjnej deterministycznej zastosowano jej klasyczną wersję tj. za wartość imputacyjną przyjmowano wartość teoretyczną z modelu regresji.

Metoda hot-deck stosowana była zwłaszcza w tych sytuacjach, w których liczba rekordów do imputacji była stosunkowo niewielka bądź gdy nie udało się znaleźć odpowiednio dopasowanego do danych empirycznych modelu regresji. Z kolei imputacja dedukcyjna wykorzystywana była tylko w wyjątkowych sytuacjach, gdy uzupełnienie brakujących danych na podstawie stwierdzonych zależności było oczywiste[11].

Spośród innych badań, w których zastosowano imputację, na uwagę zasługuje Powszechny Spis Rolny 2002, który swym zasięgiem objął gospodarstwa rolne o powierzchni użytków rolnych powyżej 1ha, gospodarstwa indywidualne o powierzchni

[10] Imputacje w zakresie zmiennych dochodowych przeprowadzono w oparciu o ogólne zasady imputacji określone w rozporządzeniu Komisji Europejskiej nr 1981/2003 z 21 października 2003 roku.

[11] Szczegółowy opis zmiennych, w odniesieniu do których stosowana była imputacja oraz wyniki przeprowadzonych analiz, można znaleźć w raporcie z badania EU-SILC 2006r., por. GUS (2008).

użytków rolnych od 0,1 do 1 ha, osoby fizyczne będące właścicielami zwierząt gospodarskich, nie posiadające użytków rolnych lub posiadające użytki rolne o powierzchni mniejszej niż 0,1 ha oraz różnego typu pozostałe gospodarstwa rolne. W przypadku gospodarstw rolnych, których użytkownicy odmówili udziału w Powszechnym Spisie Rolnym 2002 r. oraz w tych sytuacjach, gdy kontakt z użytkownikami gospodarstw był niemożliwy, dokonano imputacji podstawowych danych o gospodarstwie, por. GUS (2003a).

Wykorzystanie imputacji na szeroką skalę planuje się również w zbliżających się spisach PSR'2010 i NSP'2011, które oparte będą po raz pierwszy na mieszanej metodzie zbierania danych. W Polsce dotychczas przeprowadzone spisy opierały się na podejściu klasycznym, który polegał na zatrudnianiu rachmistrzów spisowych, odwiedzających wszystkie zamieszkane jednostki i zapisujących uzyskane od respondentów informacje na tradycyjnych formularzach spisowych, przygotowanych w wersji papierowej.

Koncepcja modelu mieszanego — z powodzeniem realizowanego w państwach skandynawskich — będzie polegała na połączeniu metody wykorzystania danych z rejestrów administracyjnych z badaniami reprezentacyjnymi. Jedynie w sytuacjach, w których zidentyfikowane podmioty nie będą objęte rejestrami bądź w przypadku, gdy dane o nich będą szczątkowe, przewiduje się przeprowadzenie spisu uzupełniającego z wykorzystaniem rachmistrzów i formularzy elektronicznych. Planuje się przy tym wykorzystanie w maksymalnym stopniu rejestrów administracyjnych – tak, aby wszystkie zmienne obowiązkowe znalazły pełne pokrycie w istniejących źródłach administracyjnych[12]

Jednym z założeń planowanego spisu jest to, aby w przypadku zmiennych obowiązkowych (tzw. core topics), których wartości nie będzie można znaleźć w dostępnych rejestrach i systemach administracyjnych, już na poziomie mikro stosować metody imputacji brakujących danych, tak aby budowane w oparciu o kompletny zbiór rekordów modele matematyczno-statystyczne umożliwiały poprawną estymację na poziomie makro. Planuje się przy tym nie tylko zastępowanie braków w jednych rejestrach danymi z innych, ale również ich imputację w oparciu o informacje pozyskane z różnych badań reprezentacyjnych.

W celu opracowania modelu relizacji NSP'2011 powołana została specjalna podgrupa ds. metod statystyczno-matematycznych na rzecz spisu, której głównym celem będzie opracowanie koncepcji prac związanych z estymacją pośrednią dla małych obszarów, imputacją i kalibracją. Efektem prac tej podgrupy ma być wypracowanie metodyki umożliwiającej estymację brakujących danych (imputację) oraz zastosowanie kalibracji i metod estymacji pośredniej (statystyka małych obszarów[13]) na potrzeby NSP'2011 w różnych przekrojach terytorialnych[14].

[12] Szczegółowy opis założeń metodologicznych oraz kompleksowej wizji spisów powszechnych – PSR'2010 i NSP'2011 można znaleźć w opracowaniu J. Dygaszewicza (2007).

[13] Estymatory pośrednie znane ze statystyki małych obszarów nazywać będziemy zamiennie estymatorami klasy SMO.

[14] W skład powołanej podgrupy na rzecz NSP'2011 weszli również pracownicy Katedry Statystyki Wydziału Informatyki i Gospodarki Elektronicznej Uniwersytetu Ekonomicznego w Poznaniu. W ramach wykonanych do tej pory przez podgrupę prac w oparciu o dane rzeczywiste z NSP'2002 i PSR'2002 przeprowadzono szereg analiz i badań symulacyjnych nad własnościami estymatorów odpornych, kalibracyjnych i klasy SMO, por. G. Dehnel (2009), E. Gołata (2009), T. Klimanek (2009) i M. Szymkowiak (2009).

Planowana integracja danych z rejestrów administracyjnych, umożliwiłaby imputację brakujących informacji w badaniach reprezentacyjnych towarzyszących spisowi oraz w innych badaniach statystycznych, które odbędą się po spisie. Należy jednak podkreślić, że ze względu na skalę przedwsięzięcia i fakt, że imputacja danych ma się odbywać na poziomie jednostkowym, zadanie to z racji ograniczeń czasowych będzie trudne do zrealizowania bądź całkowicie niewykonalne. Dlatego zasadne wydaje się w planowanym spisie przypisanie większej roli — niż imputacji — metodom estymacji parametrów w przypadku występowania braków danych (kalibracja) oraz wykorzystanie estymacji pośredniej, która umożliwi oszacowanie różnego rodzaju parametrów na bardzo niskich poziomach agregacji przestrzennej[15].

1.3. Kalibracja

Kalibracja, w swych różnych formach, stała się na przestrzeni ostatnich lat ważną metodą wykorzystywaną w estymacji różnych parametrów w badaniach statystycznych z brakami odpowiedzi. Kalibracja — jako nowy termin w metodzie reprezentacyjnej — a szerzej, w statystyce małych obszarów pojawił się w literaturze, w znaczeniu opisywanym w pracy, około 15 lat temu. Należy jednak podkreślić, że zastosowanie odpowiednich metod ważenia, z uwzględnieniem wag będących odwrotnościami prawdopodobieństw inkluzji jednostek do próby, znane było dużo wcześniej, por. M.H Hansen, W.N Hurwitz (1943). Kalibracja, w ujęciu prezentowanym w pracy, jest rozszerzeniem tych metod — w oparciu o odpowiednie wykorzystanie zmiennych pomocniczych.

Kalibracja jest zatem jedną z metod opartych na systemie wag, którą można wykorzystać w badaniach statystycznych z brakami odpowiedzi[16]. To co ją w istotny sposób odróżnia od innych metod, to takie wykorzystanie zmiennych pomocniczych, aby spełnione były odpowiednie równania kalibracyjne przy jednoczesnym minimalizowaniu odległości między wartościami wag wynikającymi ze schematu losowania próby, a wagami kalibracyjnymi.

Podstawy teoretyczne kalibracji zostały sformułowane w pionierskiej pracy Särndala i Deville'a (1992) z początku lat 90-tych XX wieku, w której autorzy przedstawili sposób konstrukcji estymatora kalibracyjnego wartości globalnej, w którym wagi (tak zwane wagi kalibracyjne) uzyskane zostały z wyjściowych wag — wynikających ze schematu losowania próby — w oparciu o wykorzystanie informacji zawartych w wektorze zmiennych pomocniczych. W podejściu tym autorzy przyjęli założenie, że znane są wartości globalne zmiennych pomocniczych oraz, że w odniesieniu do tych zmiennych, znane są ich wartości dla wszystkich jednostek na poziomie próby. Koncepcja ta stała się punktem wyjścia do zastosowania podobnego podejścia w odniesieniu do

[15] Jak pokazują doświadczenia holenderskie, zastosowanie imputacji na szeroką skalę, tylko w przypadku niewielu zmiennych, może prowadzić do wiarygodnych wyników. W wielu jednak sytuacjach są one mało przekonywujące, a czasami prowadzą do niewiarygodnych oszacowań różnych parametrów, por. B. Kroese, R.H Renssen, M. Trijssenaar (2000).

[16] Wśród innych metod, które były bądź w dalszym ciągu są wykorzystywane w badaniach z brakami odpowiedzi, można wymienić warstwowanie po wylosowaniu, raking, klasę uogólnionych estymatorów regresyjnych (GREG) oraz modele ważonej regresji logistycznej. Szczegółowy opis tych metod wraz z ich praktycznym zastosowaniem w badaniach z brakami odpowiedzi, można znaleźć w pracach Kaltona i Floresa-Cervantesa (2003), Särndala i Lundströma (2005) oraz Deville'a, Särndala i Sautorego (1993).

badań z brakami odpowiedzi. Idea wykorzystania zmiennych pomocniczych — znana już w wielu innych obszarach statystyki — nabrała nowego znaczenia w odniesieniu do tego typu badań. Dała również asumpt do rozwoju teorii w ramach szerszego nurtu — statystyki małych obszarów — w odniesieniu do tej jej części, która poświęcona jest wpływowi braków odpowiedzi na wyniki estymacji. Podobnie, jak w przypadku statystyki małych obszarów, w podejściu kalibracyjnym ważną rolę odgrywa koncepcja maksymalnego wykorzystania danych z wszelkich dostępnych źródeł informacji. To, co jednak odróżnia kalibrację od klasycznej estymacji pośredniej, to specyficzne jej wykorzystanie, tzn. w taki sposób, aby spełnione były odpowiednie równania kalibracyjne przy jednoczesnym minimalizowaniu funkcji odległości między wagami wynikającymi ze schematu losowania próby, a wagami kalibracyjnymi. Drugą istotną różnicą między estymatorami kalibracyjnymi, a estymatorami klasy SMO jest możliwość dokonywania szacunków w przypadku tzw. próby zerowej. W przeciwieństwie do niektórych estymatorów statystyki małych obszarów, nie ma możliwości zastosowania estymatorów kalibracyjnych w sytuacji, gdy nie dysponujemy żadną informacją z próby na temat badanej zmiennej.

Pionierskie dzieło Särndala i Deville'a stanowiło punkt wyjścia dla wielu prac, w których autorzy podejmowali próby konstrukcji estymatorów kalibracyjnych dla innych aniżeli wartość globalna, parametrów. W. Changbao i Y. Luan (2003) w oparciu o ideę Deville'a i Särndala, zaproponowali wykorzystanie kalibracji do konstrukcji estymatorów wariancji. Z kolei, T. Harms i P. Duchesne (2006) posłużyli się nią w procesie budowy estymatorów kalibracyjnych dla kwantyli[17].

Intensywny rozwój teorii oraz pojawiające się coraz częściej praktyczne zastosowania kalibracji w różnych dziedzinach życia, znalazły swoje odzwierciedlenie w postaci referatów, z czego znamienita większość publikowana jest w renomowanych czasopismach („Survey Methodology", „Journal of Official Statistics", „Journal of the American Statistical Association"). Jej rozwój stanowił również podstawę organizacji międzynarodowych konferencji poświęconych podejściu kalibracyjnemu. Wśród nich należy wymienić – przede wszystkim – zorganizowaną w 2007 roku przez Kanadyjski Urząd Statystyczny „Workshop on Calibration and Estimation in Surveys (WCES)„ oraz zorganizowaną w Kuusamo „Second Baltic-Nordic Conference on Survey Sampling".

Wraz z rozwojem metodologii pojawiło się odpowiednie oprogramowanie, które wykorzystywane jest w praktycznych zastosowaniach przez urzędy statystyczne różnych państw (na przykład, w Belgii G-Calib, Calmar we Francji, Bascula w Holandii, GES w Kanadzie i CLAN 97 w Szwecji). Programy te w większości napisane zostały w języku 4GL w systemie SAS. Wyjątek stanowi Bascula oprogramowana w Delphi

[17] Autorzy zaproponowali metodę wyznaczania estymatora kalibracyjnego kwantyla rzędu α, zakładając, że znane są wartości zmiennej y na poziomie całej próby s. Innymi słowy nie rozpatrywali sytuacji, w której mogą istnieć braki odpowiedzi dla tej zmiennej. W rozdziale 3 dokonujemy rozszerzenia rozpatrywanego przez autorów podejścia – uwzględniając wpływ braków odpowiedzi. Wyprowadziliśmy również postać wag uogólnionego estymatora kalibracyjnego kwantyla rzędu α zakładając, że znane są kwantyle dowolnych rzędów zmiennych pomocniczych na poziomie całej populacji bądź ich oszacowania z próby. Jest to daleko idące uogólnienie koncepcji T. Harmsa i P. Duchesne, którzy w odniesieniu do zmiennych pomocniczych zakładali, że do wyznaczenia estymatora kalibracyjnego kwantyla rzędu α zmiennej y znane są również kwantyle rzędu α dla każdej zmiennej pomocniczej – por. definicja (18) oraz twierdzenie (11).

oraz G-Calib, którego kod został zaimplementowany w pakiecie SPSS. W pakietach tych, zagadnienie poszukiwania estymatorów kalibracyjnych, sformułowane zostało jako nieliniowy problem optymalizacyjny polegający na poszukiwaniu wag kalibracyjnych $\mathbf{w} = (w_1, \ldots, w_n)^T$, które minimalizują pewną funkcję odległości tak, aby spełnione były odpowiednie równania kalibracyjne, a wyznaczone wagi znajdowały się w pewnym z góry ustalonym przedziale.

Ujmując zagadnienie bardziej formalnie, problem poszukiwania wag kalibracyjnych można opisać w następujący sposób. Niech $\mathbf{d} = (d_1, \ldots, d_n)^T$ będzie wektorem wag wynikających ze schematu losowania próby, a $\mathbf{w} = (w_1, \ldots, w_n)^T$ poszukiwanym wektorem wag kalibracyjnych, gdzie n oznacza liczebność próby. Niech G będzie dowolną funkcją spełniającą następujące warunki:

- $G(\cdot)$ jest ściśle wypukła i dwukrotnie różniczkowalna,
- $G(\cdot) \geqslant 0$,
- $G(1) = 0$,
- $G'(1) = 0$,
- $G''(1) = 1$.

Załóżmy, że celem badania jest oszacowanie wartości globalnej zmiennej y, tj.

$$Y = \sum_{i=1}^{N} y_i, \tag{1.5}$$

gdzie N oznacza liczebność populacji, a y_i wartość zmiennej y dla i – tej jednostki, $i = 1, \ldots, N$. Niech ponadto $x_1, \ldots, x_k$ oznaczają zmienne pomocnicze, które wykorzystane zostaną w problemie wyznaczania wag kalibracyjnych, a $\mathbf{X}_j$ oznacza wartość globalną zmiennej x_j, $j = 1, \ldots, k$, tj.

$$\mathbf{X}_j = \sum_{i=1}^{N} x_{ij}, \tag{1.6}$$

gdzie x_{ij} oznacza wartość j – tej zmiennej pomocniczej dla i – tej jednostki badania.

Problem poszukiwania wag kalibracyjnych w ujęciu matematycznym można przedstawić w następujący sposób.

(W1) Minimalizacja funkcji odległości:

$$D(\mathbf{w}, \mathbf{d}) = \sum_{i=1}^{n} d_i G\left(\frac{w_i}{d_i}\right) \longrightarrow \min, \tag{1.7}$$

(W2) Równania kalibracyjne:

$$\sum_{i=1}^{n} w_i x_{ij} = \mathbf{X}_j, \quad j = 1, \ldots, k, \tag{1.8}$$

(W3) Warunki ograniczające:

$$L \leqslant \frac{w_i}{d_i} \leqslant U, \quad \text{gdzie: } L < 1 \text{ i } U > 1,\ i = 1, \ldots, n. \tag{1.9}$$

Pierwszy z warunków (W1) orzeka, że wagi kalibracyjne powinny być w taki sposób wyznaczone, aby były możliwie bliskie — w sensie przyjętej funkcji odległości — wagom wynikającym ze schematu losowania próby. Funkcja G mierzy odległość między ilorazem wag $\frac{w_i}{d_i}$, a 1. Warunek drugi (W2) stanowi istotę teorii kalibracji i orzeka, że wagi powinny być tak dobrane, aby po ich zastosowaniu do wszystkich zmiennych pomocniczych uzyskać ich wartości globalne. Jeżeli ten warunek będzie spełniony, to również po wykorzystaniu tych wag do zmiennej y, powinniśmy dostać ocenę wartości globalnej bliską jej prawdziwej wartości. Trzeci z warunków (W3) jest tzw. warunkiem ograniczającym, który zapobiegać ma sytuacjom, w których uzyskane wagi kalibracyjne przyjmują wartości ujemne bądź ekstremalne.

Warunki (W1) i (W3) mogą zostać uchylone. W podejściu funkcyjnym (por. podrozdział 2.5, str.40) nie zakłada się bowiem, przy wyznaczaniu estymatorów kalibracyjnych, aby wagi kalibracyjne były bliskie wartościom wag wynikających ze schematu losowania próby w sensie przyjętej funkcji odległości. Uchylenie tego założenia powoduje, że dobierając w odpowiedni sposób tzw. wektor zmiennych instrumentalnych można uzyskać szeroką klasę estymatorów o różnej postaci i własnościach. Z kolei pominięcie warunku trzeciego, jest w zastosowaniach praktycznych często wygodne, gdyż wówczas istnieje możliwość wyznaczenia wag kalibracyjnych wprost ze wzoru – bez konieczności stosowania skomplikowanych algorytmów numerycznych. Należy jednak pamiętać o tym, że przy niewłaściwie dobranym wektorze zmiennych pomocniczych, uchylenie tego założenia może prowadzić do powstania ujemnych bądź ekstremalnych wag[18].

Istnieje również pewna dowolność przy wyborze funkcji $G(\cdot)$. Najczęściej rozważa się w literaturze następujące jej postacie, por. J-C. Deville, C-E. Särndal (1992), C-E. Särndal (2007), A. Plikuskas (2007):

$$G_1(x) = \frac{1}{2}(x-1)^2, \tag{1.10}$$

$$G_2(x) = \frac{(x-1)^2}{x}, \tag{1.11}$$

$$G_3(x) = x(\log x - 1) + 1, \tag{1.12}$$

$$G_4(x) = 2x - 4\sqrt{x} + 2, \tag{1.13}$$

$$G_5(x) = \frac{1}{2\alpha}\int_1^x \sinh\left[\alpha\left(t - \frac{1}{t}\right)\right] dt, \tag{1.14}$$

gdzie α jest dodatnim parametrem, pozwalającym sterować stopniem rozrzutu wag kalibracyjnych w stosunku do wag wynikających ze schematu losowania próby (domyślnie parametr przyjmuje wartość 1), a sinh jest funkcją sinusa hiperbolicznego zdefiniowanego jako $\sinh(x) = \frac{e^x - e^{-x}}{2}$.

Kalibracja jest z powodzeniem wykorzystywana w praktyce badań statystycznych z brakami odpowiedzi przez urzędy statystyczne wielu państw. Węgierski Urząd Statystyczny stosuje ją w dwóch badaniach statystycznych: w badaniu budżetów gospodarstw domowych od 1994 roku i w badaniu aktywności ekonomicznej ludności od

[18] W pracy zakładać będziemy, że wyznaczane wagi kalibracyjne spełniać będą pierwsze dwa warunki bądź tylko drugi z nich (podejście funkcyjne). Umożliwi to, przy odpowiednio dobranej funkcji odległości, wyznaczenie jawnej postaci wektora wag kalibracyjnych.

1995 roku, por. – Ö. Éltető, M. László (2002). Mimo stosowania czynników motywujących (w postaci pieniędzy), a mających zachęcić gospodarstwa domowe do wzięcia udziału w badaniu, frakcja braków odpowiedzi wynosiła 40% i od 4% do 15% w BBGD i BAEL odpowiednio na przestrzeni lat 1994–2000. W badaniach prowadzonych przez Węgierski Urząd Statystyczny wykorzystywana jest tzw. uogólniona iteracyjna metoda skalowania (kalibracji) wyrażająca się tym, że jako funcję odległości przyjmuje się:

$$D(\mathbf{w}, \mathbf{d}) = \sum_{i=1}^{m} \left(w_i \log \frac{w_i}{d_i} - w_i + d_i \right), \tag{1.15}$$

a zatem funkcja $G(x)$ określona jest wzorem (1.12). W metodzie tej poszukiwanie wektora wag $\mathbf{w}$ będącego rozwiązaniem odpowiednich równań kalibracyjnych, odbywa się w oparciu o algorytm iteracyjny, który w kolejnych krokach dopasowuje wagi tak, aby znaleźć przybliżone rozwiązanie równań kalibracyjnych przy jednoczesnym zmniejszaniu wartości funckji odległości (1.15)[19].

W obydwu badaniach, na potrzeby kalibracji, wykorzystuje się w konstrukcji wektora wartości globalnej — pełniącego rolę kontrolną[20] — informacje o pewnych charakterystykach demograficznych. Przykładowo, w badaniu budżetów gospodarstw domowych bierze się pod uwagę informacje o liczbie gospodarstw domowych, liczbie osób o określonym poziomie wykształcenia, płci, wieku oraz aktywności ekonomicznej. Przyjęcie tak wielu zmiennych pomocniczych w wektorze wartości globalnych wymaga zazwyczaj rozwiązania wielu równań kalibracyjnych przy jednoczesnym poszukiwaniu minimum funkcji odległości, co nie zawsze musi prowadzić w przyjętej iteracyjnej metodzie do znalezienia wektora wag kalibracyjnych $\mathbf{w}$[21].

Na szeroką skalę podejście kalibracyjne wykorzystywane jest w państwach skandynawskich (zwłaszcza w Szwecji)[22]. Przykładowo, w Szwecji kalibracja zastosowana została w badaniach nad jakością życia i zdrowia, w których frakcja braków odpowiedzi wynosiła 35%. Jako zmienne wspomagające wykorzystano dane ze spisu oraz z rejestru edukacyjnego, a dotyczyły one informacji na temat płci, wieku, miejsca urodzenia, grupy dochodowej, poziomu wykształcenia oraz stanu cywilnego. Oszacowania wybranych parametrów, w odniesieniu do wielu zmiennych (stan zdrowia, sytuacja finansowa), z wykorzystaniem estymatorów kalibracyjnych, w różnych przekrojach — w ramach grup wyznaczonych przez zmienne pomocnicze — dokonano z zastosowaniem

[19] Zaletą wykorzystywanej w pracy funkcji odległości jest możliwość znalezienia wektora wag kalibracyjnych explicite, czego nie można uzyskać stosując funkcję (1.15).

[20] Więcej na temat funkcji kontrolnej wektora wartości globalnej można znaleźć w rozdziale 2 na stronie 32.

[21] Przykładowo, w badaniu budżetów gospodarstw domowych, prowadzonym przez Węgierski Urząd Statystyczny, znalezienie wag kalibracyjnych wymaga rozwiązania około 100 równań kalibracyjnych. Algorytm uogólnionej iteracyjnej metody kalibracji zaimplementowany został w IML-u w pakiecie SAS i tylko w nielicznych przypadkach nie był zbieżny, tj. gdy któreś z równań kalibracyjnych nie miało rozwiązania.

[22] Jest to zapewne związane z faktem, że twórca kalibracji, profesor Carl-Erik Särndal, będący wybitnym znawcą metody reprezentacyjnej i statystyki małych obszarów jest Szwedem. W ramach prowadzonych badań naukowych współpracował ze Szwedzkim Urzędem Statystycznym nad zastosowaniami kalibracji w badaniach z brakami odpowiedzi, co zaowocowało m.in. powstaniem programu CLAN oraz przyczyniło się do praktycznego wykorzystania kalibracji w wielu badaniach prowadzonych przez ten urząd.

programu CLAN97, por. C-E. Särndal, S. Lundström (2005). Wśród innych ważnych badań, w których wykorzystuje się estymatory kalibracyjne, należy zaliczyć badanie budżetów gospodarstw domowych (Szwecja, Finlandia, Dania)[23], badanie wykorzystania czasu (Finlandia), badanie dotyczące bezpieczeństwa i przestępczości (Szwecja).

W Polsce podejście kalibracyjne jest wykorzystywane w bardzo ograniczonym zakresie przez Główny Urząd Statystyczny. W zasadzie jedynym badaniem, w którym zastosowano estymatory kalibracyjne, z racji dużej frakcji braków odpowiedzi, było wspomniane już EU-SILC. W badaniu tym wykorzystano metodę kalibracji zintegrowanej — w wersji sinusa hiperbolicznego, ze względu na wykazaną w praktyce własność uzyskania wag kalibracyjnych bardzo blisko skupionych wokół wag wyjściowych. Wagi początkowe — wynikające z przyjętego dwustopniowego schematu doboru jednostek do próby — korygowane były w oparciu o informacje o liczbie gospodarstw domowych oraz o liczbie osób według płci i wieku[24]. Informacje te na poziomie województw (NUTS2) z dodatkowym podziałem na obszary miejski i wiejski pochodziły z Narodowego Spisu Powszechnego Ludności i Mieszkań 2002 oraz z bieżących szacunków demograficznych, por. – GUS (2008).

Podobnie, jak w przypadku imputacji, planuje się wykorzystanie kalibracji na potrzeby PSR'2010 i NSP'2011. W pierwszej kolejności kalibracja zastosowana zostanie w badaniu reprezentacyjnym towarzyszącym spisowi, a obejmującym swym zasięgiem około 15% populacji, co umożliwi estymację różnych parametrów w odniesieniu do wielu cech (zwłaszcza tzw. non-core topics) w różnych przekrojach, na dowolnym szczeblu podziału administracyjnego kraju. Po drugie, przewiduje się wykorzystanie kalibracji w planowanych do przeprowadzenia, tuż po spisie (w 2012 roku), dwóch dodatkowych badaniach reprezentacyjnych, poświęconych tematyce imigracji cudzoziemców do Polski oraz dzietności kobiet i powiązaniom generacyjnym rodzin, co ma zwiększyć precyzję estymacji[25].

Pierwsze próby testowania własności estymatorów kalibracyjnych, w oparciu o rzeczywiste dane pochodzące z NSP'2002, zostały już podjęte w ramach prowadzonych przez podgrupę ds. metod statystyczno-matematycznych prac na rzecz spisu. Jak pokazują wyniki badań symulacyjnych, estymatory kalibracyjne charakteryzowały się mniejszym obciążeniem i wariancją w porównaniu z innymi znanymi z klasycznej metody reprezentacyjnej estymatorami. Uzyskanie obiecujących wyników w ramach przeprowadzonych eksperymentów może być więc istotnym argumentem za ich zastosowaniem na potrzeby zbliżającego się spisu[26]. Dotyczyć to będzie zwłaszcza badań

[23] W badaniach budżetów gospodarstw domowych kalibrację wykorzystuje wiele innych państw: Litwa, wspomniane już Węgry oraz Szwajcaria. W tym ostatnim państwie frakcja braków odpowiedzi w badaniu jest szczególnie wysoka i według danych Szwajcarskiego Urzędu Statystycznego wynosi około 70%.

[24] W przeciwieństwie do badania budżetów gospodarstw domowych przyjęto cztery kategorie wielkości gospodarstwa: 1-osobowe, 2-osobowe, 3-osobowe, 4 i więcej osobowe. W odniesieniu do liczby osób według płci i wieku przyjęto 14 grup wieku: poniżej 16 lat, 16–19 lat, 11 5-letnich grup, grupa 75 i więcej lat.

[25] W badaniach towarzyszących spisowi, jak na przykład badanie dzietności kobiet NSP'1970 i NSP'1988 braki odpowiedzi sięgały 30%.

[26] Szczegółowe wyniki przeprowadzonych badań symulacyjnych zawarte są w raporcie przygotowanym dla Głównego Urzędu Statystycznego – por. M. Szymkowiak (2009).

przeprowadzonych metodą reprezentacyjną, a towarzyszących spisowi oraz innych prowadzonych przez GUS (na przykład BAEL).

Ponieważ wiele badań prowadzonych przez GUS jest obarczonych bardzo dużymi brakami odpowiedzi, kalibracja może stanowić swego rodzaju remedium na ten rodzaj błędów nielosowych. Jak pokazują bowiem doświadczenia państw wykorzystujących w praktyce podejście kalibracyjne, jej zastosowanie, może w znaczący sposób zredukować obciążenie i zmniejszyć wariancję stosowanych estymatorów.