Praca doktorska

Estymatory kalibracyjne w badaniu budżetów gospodarstw domowych

Marcin Szymkowiak

Promotor
prof. dr hab. Jan Paradysz

Uniwersytet Ekonomiczny w Poznaniu
Wydział Informatyki i Gospodarki Elektronicznej
Katedra Statystyki

Poznań 2009

Spis treści

Wstęp

Problem

Do najtrudniejszych problemów w badaniach reprezentacyjnych należy zaliczyć brak możliwości otrzymania odpowiedzi od wszystkich jednostek wylosowanych do próby. W literaturze przedmiotu, zaproponowano szereg rozwiązań niedogodności związanych z nieobecnością w badaniu, por. G. Kalton, I. Flores-Cervantes (2003). Skuteczność proponowanych metod wiąże się jednak ze sporym zwiększeniem kosztów, co kłóci się ze współczesnymi tendencjami w zakresie ekonomiczności badań oraz maksymalnym wykorzystaniem wszystkich dostępnych źródeł informacji spoza próby.

Uzasadnienie wyboru tematu

Konieczność redukcji kosztów, a także fakt, że tradycyjne estymatory, znane z klasycznej metody reprezentacyjnej, charakteryzują się, w przypadku istnienia braków odpowiedzi, zbyt dużym obciążeniem i wariancją, wymusiła potrzebę intensywnego rozwoju odpowiednich technik estymacji. Stosowanie estymatorów bezpośrednich na niskim poziomie agregacji przestrzennej, ze względu na niewielką liczebność próby oraz istniejące braki odpowiedzi, prowadzić będzie do niewłaściwych wyników, por. C-E. Särndal, S. Lundström (2005). W rezultacie ich wykorzystanie w praktycznych badaniach prowadzonych przez urzędy statystyczne obarczone będzie znacznymi błędami nielosowymi. Posiadanie właściwych danych jest ponadto szczególnie ważne z punktu widzenia samorządów terytorialnych, które zainteresowane są informacjami dotyczącymi terytorium, na którym prowadzą swoją politykę – również w kontekście warunków życia, sytuacji materialnej i bytowej gospodarstw domowych.

Znaczenie kalibracji dla teorii i praktyki estymacji

W teorii estymacji duże znaczenie przywiązuje się do błędów losowych. Jednakże, jak zobaczymy także w tej pracy, udział błędów nielosowych, często o charakterze systematycznym jest dużo większy od tych pierwszych. Zatem rozwój takich technik jak imputacja i kalibracja ma duże znaczenie dla teorii estymacji statystycznej, por. J-C. Deville, C-E. Särndal, (1992), N.T Longford (2005).

Wiele informacji niezbędnych w prowadzeniu racjonalnej działalności związanej

z planowaniem i zarządzaniem, jest publikowanych przez Główny Urząd Statystyczny w różnego rodzaju opracowaniach, dla całego kraju bądź w przekroju województw. Dla prowadzenia właściwej polityki społecznej niezbędne są jednak aktualne informacje dotyczące warunków życia, sytuacji materialnej i dochodowej gospodarstw domowych na najniższym poziomie terytorialnym – powiatu lub gminy. Ponadto duże odstępy między badaniami pełnymi (spisami), które nie zawsze dostarczają niezbędnych decydentom danych, a także długie okresy opracowywania wyników, nie są adekwatne do bieżących potrzeb samorządów terytorialnych. Z kolei dane otrzymywane z badań reprezentacyjnych, w odróżnieniu od tych z badań pełnych, nie zawsze mogą być wykorzystane ze względu na niewielką liczebność próby oraz istniejące braki odpowiedzi do oszacowania różnych parametrów na niskim poziomie agregacji z wykorzystaniem estymacji bezpośredniej.

Rosnący popyt na informacje, w przekroju powiatów i gmin oraz niedostatki klasycznej estymacji bezpośredniej, zmuszają do poszukiwania nowych metod i technik estymacji, które mogą zostać wykorzystane w badaniach statystycznych, w których istnieje problem małej liczebności próby oraz braków danych.

Odpowiedzią na rosnące z jednej strony, zapotrzebowanie na informacje na niskim poziomie agregacji, a wymogiem poprawy dokładności i precyzji szacunków z drugiej strony, jest rozwijająca się od kilku lat intensywnie metoda oparta o system wag – kalibracja.

Jest to najnowsza technika poświęcona zagadnieniu estymacji parametrów w populacji generalnej w badaniach z brakami odpowiedzi. Jej idea polega na tym, aby na podstawie zawartych w próbie danych i przy maksymalnym wykorzystaniu wszystkich dostępnych informacji dodatkowych, składających się na wektor zmiennych pomocniczych, dokonać estymacji — mając przy tym na uwadze, że nie od wszystkich jednostek biorących udział w badaniu uzyskano odpowiedź. Informacja ta wykorzystywana jest w taki sposób, aby spełnione były odpowiednie równania kalibracyjne. Wykorzystanie wszelkich dostępnych informacji, poprzez jednoczesne spełnienie odpowiednich warunków, ma na celu zniwelowanie negatywnego wpływu braków odpowiedzi. Wykorzystując tę ideę, w odpowiedni sposób konstruuje się estymatory kalibracyjne, które powinny być w mniejszym stopniu obciążone i być bardziej efektywne w porównaniu z estymatorami znanymi z klasycznej metody reprezentacyjnej, por. C-E. Särndal (2007) oraz C-E. Särndal, S. Lundström (2008).

Warto ponadto podkreślić, że w polskiej literaturze z zakresu metody reprezentacyjnej, nie było dotąd poważniejszych opracowań poświęconych skutkom nieobecności w badaniu. Także w światowej literaturze przedmiotu „nonresponse" przeszedł znamienną ewolucję od wysiłków uzyskania odpowiedzi „za wszelką cenę" do estymacji wykorzystującej alternatywne źródła informacji pośredniej.

Techniki estymacji parametrów w postaci estymatorów kalibracyjnych, które szczegółowo rozważamy w pracy, nie są u nas powszechnie znane i wykorzystywane. Jest to nie tylko pierwsza tego typu praca w Polsce, ale i w sporej części Europy, poza Skandynawią. Celem pracy jest zatem – między innymi – wskazanie możliwości i korzyści jakie dałoby zastosowanie estymatorów kalibracyjnych w badaniach z brakami odpowiedzi, co miałoby niepoślednią wartość praktyczną dla Głównego Urzędu Statystycznego.

Cele i charakter pracy

Praca ma charakter metodologiczno-poznawczy, a dane służą głównie do weryfikacji metod statystycznych. W **warstwie metodologicznej** poświęcona jest najnowszej technice estymacji, którą stosuje się w warunkach, gdy w badaniu wystąpią braki odpowiedzi. W **warstwie teoretycznej** — oprócz przeglądu występujących w literaturze estymatorów kalibracyjnych — zaproponowano własne estymatory kalibracyjne kwantyli, których konstrukcja opiera się na zbiorze informacji o kwantylach zmiennych pomocniczych. W **warstwie empirycznej**, w oparciu o autorskie programy na drodze symulacji komputerowej, dokonano kompleksowej oceny rozważanych w pracy estymatorów z wykorzystaniem rzeczywistych danych pochodzących z NSP'2002 oraz danych wygenerowanych z rozkładu logarytmiczno-normalnego. Wygenerowanie danych z tego rozkładu wynikało z faktu, że charakteryzuje się on asymetrią prawostronną – podobnie jak rozkłady wielu cech w badaniu budżetów gospodarstw domowych. Rozważania, w ramach warstwy empirycznej, odnoszą się ponadto do podjętej próby zastosowania estymatorów kalibracyjnych na potrzeby oszacowania wybranych kategorii wydatków gospodarstw domowych w przekroju powiatów województwa wielkopolskiego w 2002r. Przeprowadzona analiza empiryczna miała charakter pionierskiej aplikacji estymatorów kalibracyjnych w badaniu budżetów gospodarstw domowych na tak niskim poziomie agregacji przestrzennej i służyła w dużej mierze ocenie przydatności zastosowanej metodologii.

Problemem badawczym, podjętym w pracy, jest poszerzenie liczby szacowanych parametrów o kwantyle rozkładu, weryfikacja założeń i dowodów własności estymatorów kalibracyjnych oraz ocena ich przydatności do badań społeczno-ekonomicznych w przekroju powiatowym na przykładzie budżetów gospodarstw domowych. Realizacja głównego celu pracy zdeterminowała sformułowanie następujących hipotez badawczych:

- estymatory kalibracyjne charakteryzują się mniejszym obciążeniem i wariancją w porównaniu z estymatorami znanymi z klasycznej metody reprezentacyjnej,
- do oszacowania kategorii wydatków gospodarstw domowych, w sytuacji dużej frakcji braków odpowiedzi, niezbędne jest podejście kalibracyjne,
- estymatory kalibracyjne dostarczają bardziej wiarygodnych ocen wydatków gospodarstw domowych w przypadku znacznego odsetka braków odpowiedzi,
- estymacja kalibracyjna kwantyli poszerza instrumentarium w badaniach budżetów gospodarstw domowych, w przekrojach regionalnych – w warunkach znacznego odsetka odmów i innych błędów nielosowych.

Weryfikacja przedstawionych powyżej hipotez badawczych podjęta została w kolejnych rozdziałach pracy.

W pierwszym rozdziale przedstawiono charakterystykę dwóch najważniejszych metod statystycznych wykorzystywanych w badaniach statystycznych z brakami odpowiedzi – imputację i kalibrację. Wskazano na dotychczasowe obszary zastosowań tych metod na świecie. Poruszone tam zagadnienia w znacznej części odnoszą się do doświadczeń Węgier, Kanady, Stanów Zjednoczonych oraz państw skandynawskich, por. Ö. Éltető, M. László (2002), C-E. Särndal, S. Lundström (2005) . Szczególnie zwrócono uwagę na różnorodność zastosowań kalibracji w Szwecji, a także próby aplikacji

podejścia kalibracyjnego i imputacji w Polsce. Z racji nowej filozofii NSP'2011 opartej na modelu mieszanym, wykorzystującym rejestry administracyjne, wskazano również na możliwe obszary zastosowań kalibracji w planowanych badaniach reprezentacyjnych towarzyszących spisowi[1].

W rozdziale drugim przedstawiono teoretyczne podstawy kalibracji w badaniach statystycznych z brakami odpowiedzi. Wskazano zalety tej metody estymacji, jak i pewne zagrożenia związane z jej stosowaniem. Duży nacisk został położony na wyprowadzenie postaci wag kalibracyjnych estymatorów wartości globalnej, zakładając różne warianty postaci wektorów wartości globalnych zmiennych pomocniczych. Udowodniliśmy[2] przy tym, nie tylko najważniejsze twierdzenia odnoszące się do postaci wag kalibracyjnych, ale również pewnych pożądanych własności samych estymatorów. Wykorzystując najnowszą literaturę dokonano ponadto przeglądu najczęściej wykorzystywanych w praktyce estymatorów kalibracyjnych, por. P.S. Kott (2006), C-E. Särndal (2007) oraz C-E. Särndal, S. Lundström (2008). Zwróciliśmy też uwagę na korzyści płynące z zastosowania tzw. podejścia funkcyjnego, opartego na koncepcji wektora zmiennych instrumentalnych, które umożliwia konstrukcję wielu estymatorów kalibracyjnych mogących mieć zastosowanie w badaniach statystycznych z brakami odpowiedzi.

W rozdziale trzecim, w oparciu o metodologię konstruowania estymatorów kwantyli zaproponowaną przez T. Harmsa i P. Duchesne (2006), sformułowaliśmy system twierdzeń o postaci wag i własnościach estymatorów kalibracyjnych kwantyla rzędu α. Uchylono przy tym krępujące założenia, które przyjęli autorzy. Przy konstrukcji estymatorów kalibracyjnych założyliśmy, że:

- nie od wszystkich jednostek wylosowanych do próby uzyskaliśmy odpowiedzi,
- znane mogą być kwantyle różnych rzędów zmiennych pomocniczych bądź ich oszacowania.

W szczególności udowodniliśmy twierdzenie o:

1. wagach estymatora kalibracyjnego dla różnych wektorów zmiennych pomocniczych,
2. postaci estymatora kalibracyjnego, w przypadku gdy istnieje liniowa zależność między zmienną badaną a zmienną pomocniczą.

Punktem wyjścia do rozważań w rozdziale czwartym była konieczność zbadania podstawowych własności estymatorów kalibracyjnych. Ponieważ na drodze analitycznej w zasadzie jest niemożliwe, aby porównać obciążenie i wariancję rozpatrywanych estymatorów, dlatego podjęto próbę przeprowadzenia badań symulacyjnych w oparciu o dwa zbiory danych, celem empirycznej oceny skutków braków odpowiedzi. Korzystając z rzeczywistych danych, pochodzących z NSP'2002, a dotyczących sytuacji mieszkaniowej w powiatach województwa wielkopolskiego, poddano empirycznej ocenie

[1] W roku 2010 przygotowywany jest Powszechny Spis Rolny (PSR'2010), a rok później Narodowy Spis Powszechny (NSP'2011). W obu spisach planuje się szerokie wykorzystanie estymacji pośredniej opartej na rejestrach administracyjnych i badaniach sondażowych. Bierze się także pod uwagę wykorzystanie estymatorów kalibracyjnych, por. J. Dygaszewicz (2007).

[2] W rozprawie stosujemy formę bezosobową w odniesieniu do prac innych autorów lub własnych, mniej znaczących rozważań. Natomiast pierwszą osobę liczby mnogiej rezerwujemy sobie dla podkreślenia własnych, istotnych osiągnięć teoretycznych i poznawczych.

obciążenie i wariancję estymatorów średniej powierzchni mieszkań, opierając się przy tym na rozważanych w drugim rozdziale estymatorach kalibracyjnych wartości globalnej. Z kolei wykorzystując dane wygenerowane z rozkładu logarytmiczno-normalnego, podjęto próbę weryfikacji przydatności proponowanych w rozdziale trzecim estymatorów kalibracyjnych mediany. Badania symulacyjne wskazują, że nieobecność w badaniu, w pewnych przypadkach, może spowodować obciążenie estymatora nawet do 25%. W obydwu rozważanych badaniach symulacyjnych rozpatrywano różne warianty braków odpowiedzi. Dokonana na podstawie symulacji ocena obciążenia i efektywności estymatorów, w powiązaniu z analizą dostępnych źródeł danych statystycznych, miała również na celu ułatwić podjęcie decyzji o wyborze zastosowania konkretnych estymatorów w badaniu budżetów gospodarstw domowych.

W rozdziale piątym, podjęto próbę empirycznej oceny rozważanych w rozdziale drugim i trzecim estymatorów kalibracyjnych, na potrzeby estymacji wybranych kategorii wydatków gospodarstw domowych w przekroju powiatów województwa wielkopolskiego. Analiza przeprowadzona została w oparciu o dane pochodzące z badania budżetów gospodarstw domowych, w którym frakcja braków odpowiedzi często przekracza 50%, a które jest badaniem, mającym bardzo duży wpływ na konstrukcję różnych wskaźników gospodarczych, jak na przykład indeksu cen towarów i usług konsumpcyjnych. Jako źródła zmiennych pomocniczych wykorzystano informacje z samego badania budżetów gospodarstw domowych i z NSP'2002 o liczbie gospodarstw domowych w poszczególnych powiatach. Uzyskane w rozdziale wyniki poddano merytorycznej ocenie. Warto podkreślić, że przeprowadzona analiza empiryczna miała charakter pierwszych zastosowań i służyła w dużej mierze określeniu przydatności wybranych estymatorów kalibracyjnych. W rozdziale tym, szacowano nie tylko poziom średnich wydatków gospodarstw domowych, ale również ich medianę z wykorzystaniem zaproponowanych w rozdziale trzecim estymatorów kalibracyjnych kwantyli.

Źródła danych statystycznych

Wybór okresu analizy podyktowany został dostępnością danych statystycznych w przekroju powiatów województwa wielkopolskiego. Wynikało to z faktu, że zastosowanie niektórych rozważanych w pracy estymatorów kalibracyjnych, wymagało połączenia informacji z dwóch źródeł, badania budżetów gospodarstw domowych i spisu, który odbył się właśnie w 2002 roku.

Pierwsze z rozważanych źródeł jest jednym z najważniejszych zbiorów informacji o przychodach gospodarstw domowych, ich rozchodach, spożyciu oraz wyposażeniu w dobra trwałego użytkowania. Wyniki z badania budżetów gospodarstw domowych przedstawiane są według różnych grup społeczno-ekonomicznych, wielkości gospodarstwa domowego, typów biologicznych gospodarstw domowych, klasy miejscowości zamieszkania, kwintyli dochodów, makroregionów i województw. Badanie to jest przeprowadzane corocznie przez Główny Urząd Statystyczny, a jego wyniki umożliwiają wszechstronną analizę warunków życia ludności w wielu ich aspektach. Z racji tego, że jest to badanie reprezentacyjne, możliwe są uogólnienia wyników na całą populację wszystkich gospodarstw domowych.

Jako drugie źródło informacji przyjęto w pracy dane z NSP'2002, który dostarcza możliwie najpełniejszej informacji o stanie i strukturze ludności gospodarstw domo-

wych i rodzin, a także informacji o mieszkaniach i warunkach mieszkaniowych – na najniższym poziomie podziału teryterionalnego kraju. Uwzględnia również charakterystykę zmian, jakie zaszły w czasie w podstawowych strukturach demograficzno- społecznych ludności, gospodarstw domowych i rodzin oraz zmian w wielkości i standardzie zasobów mieszkaniowych. Zawiera także informacje niezbędne dla potrzeb tworzenia operatów do badań reprezentacyjnych prowadzonych w gospodarstwach domowych w latach pospisowych. Wyniki ze spisu uwzględniono w ramach przeprowadzonych w rozdziale czwartym badań symulacyjnych, celem empirycznej oceny skutków braków odpowiedzi. Informacje z tego źródła wykorzystano także w rozdziale ostatnim, chcąc zapewnić sobie sumowalność wag kalibracyjnych do liczebności całej populacji. W odniesieniu do badań symulacyjnych wykorzystano informacje o zasobach mieszkaniowych w przekroju powiatów województwa wielkopolskiego. W przypadku rozdziału piątego, wykorzystano natomiast informacje o liczbie gospodarstw domowych w poszczególnych powiatach, celem spełnienia odpowiedniego równania kalibracyjnego.

Intencją autora było również zwrócenie uwagi na możliwości, jakie daje podejście kalibracyjne na niskich poziomach agregacji danych. Nabierze to szczególnego znaczenia w kontekście zbliżajacego się NSP'2011, w którym po raz pierwszy na szeroką skalę wykorzystane zostaną rejestry administracyjne, które mogą stanowić potencjalne źródło zmiennych wspomagających. Wykorzystanie kalibracji, w warunkach tak kompleksowej infrastruktury statystycznej, uwiarygodni wyniki w różnych przekrojach terytorialnych i w znacznym stopniu przyczynić się może do zniwelowania ujemnego wpływu występującego w każdym badaniu problemu braków odpowiedzi.

Gorąco chciałbym podziękować Promotorowi pracy Panu prof. dr hab. Janowi Paradyszowi za życzliwą opiekę podczas jej pisania, za wiele cennych wskazówek i rad oraz za duchowe wspieranie mnie w chwilach zwątpienia.

Rozdział 1

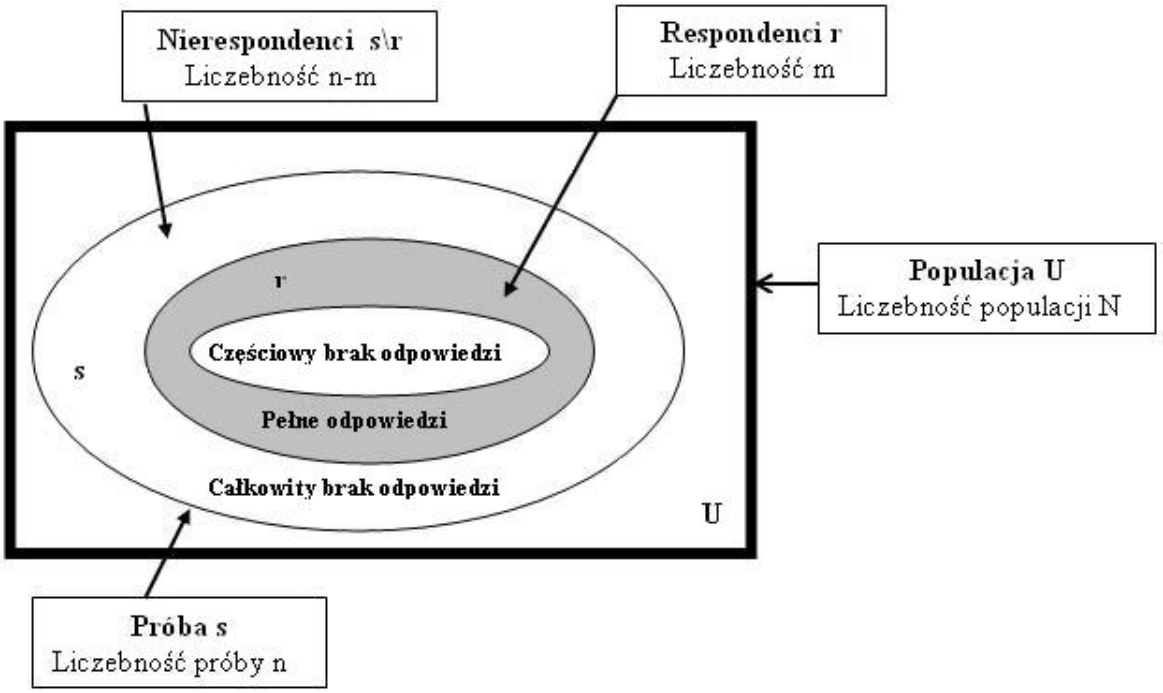
Przegląd metod estymacji w badaniach statystycznych z brakami odpowiedzi

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.

Rozdział 2

Estymatory kalibracyjne wartości globalnej

2.1. Podstawowe definicje i oznaczenia

W literaturze przedmiotu można znaleźć różne definicje kalibracji, które czynią z niej metodę w istotny sposób odróżniającą ją od innych podejść wykorzystywanych w badaniach statystycznych z brakami odpowiedzi. Wydaje się, że najpełniej sformułował kalibrację, a dokładniej „podejście kalibracyjne" jeden z jej twórców — C-E. Särndal (2007).

Definicja 2 (C-E.Särndal, 2007). *Podejście kalibracyjne w estymacji parametrów w odniesieniu do skończonych populacji składa się z :*

a) *obliczenia wag z uwzględnieniem informacji dodatkowych, tak aby spełnione było odpowiednie równanie — tzw. równanie kalibracyjne,*

b) *wykorzystania tych wag do estymacji wartości globalnej bądź innych parametrów, przy czym wartość zmiennej mnożona jest przez wagę, a sumowanie odbywa się po zbiorze wszystkich respondentów,*

c) *uzyskania w ten sposób nieobciążonych oszacowań parametrów, w przypadku gdyby w badaniu nie wystąpiły braki odpowiedzi oraz inne błędy nielosowe.*

Powyższa definicja wymaga pewnego komentarza. Kalibracja jako metoda obliczania wag jest wykorzystywana w badaniach częściowych, w których jednostki są dobierane do próby w oparciu o wcześniej określony schemat losowania. Oznacza to, że znane są wagi d_i wynikające ze schematu losowania próby, które są odwrotnościami prawdopodobieństw π_i dostania się i – tej jednostki do próby. W gruncie rzeczy kalibracja, zgodnie z definicją podaną przez Särndala, jako metoda obliczania wag, polega na korygowaniu wyjściowych wag wynikających ze schematu losowania próby. Do obliczania nowych wag w_i (tak zwanych wag kalibracyjnych) wykorzystuje się informacje zawarte w wektorze zmiennych pomocniczych. Zakłada się przy tym, że informacje dodatkowe

znane są co najmniej na poziomie próby s. Jest to więc metoda, w której wykorzystanie dodatkowych informacji, zawartych w wektorze zmiennych pomocniczych, ma przyczynić się do polepszenia uzyskiwanych szacunków, w przypadku gdy w badaniu występują braki danych. Jest to konsekwencją tego, że braki odpowiedzi od jednostek badania powodują, że stosowanie tradycyjnych estymatorów, znanych z klasycznej metody reprezentacyjnej, prowadzi zazwyczaj do dużych obciążeń oraz ich dużej wariancji. Należy przy tym podkreślić, że wykorzystanie informacji dodatkowych, celem poprawy oszacowań parametrów, znane było na długo przed tym, jak idea kalibracji w ujęciu prezentowanym przez Särndala i Deville'a (1992) została sformułowana. To co odróżnia kalibrację od innych metod ważenia, to takie wykorzystanie tych informacji, aby spełnione było odpowiednie równanie kalibracyjne. W zależności od poziomu na jakim dostępne są informacje pomocnicze (poziom próby bądź całej populacji) i od rodzaju szacowanego parametru, równanie to przyjmuje różną postać (na przykład wagi kalibracyjne, w przypadku estymatora kalibracyjnego wartości globalnej, są tak ustalane w równaniu kalibracyjnym, aby po ich przemnożeniu przez wartości zmiennych pomocniczych i dokonaniu sumowania na zbiorze wszystkich respondentów uzyskać jej wartość globalną). Rodzaj posiadanej informacji dodatkowej ma również wpływ na to, czy i w jaki sposób, wagi kalibracyjne mogą poprawić oszacowanie parametru oraz skutecznie „chronić" przed brakami odpowiedzi.

Pewnego wyjaśnienia wymaga również ostatni z podpunktów w definicji (2). Występowanie błędów nielosowych, w postaci braków odpowiedzi, jest w każdym badaniu zjawiskiem niepożądanym lecz normalnym. Dlatego bez względu na to czy braki odpowiedzi występują w badaniu czy nie, obciążenie estymatorów szacowanych parametrów jest w zasadzie nieuniknione, bez względu na to czy zastosowana będzie kalibracja czy też inna metoda. Wykluczenie tego podpunktu z definicji (2) nie stałoby na przeszkodzie w uzyskaniu wag kalibracyjnych, jednakże ze względu na fakt, że w literaturze zwraca się na ten aspekt uwagę, została ona przez Särndala włączona do sformułowanego przez niego podejścia.

Dokonując podsumowania zawartych powyżej rozważań, kalibrację można zdefiniować w następujący sposób:

Definicja 3. *Kalibracja to metoda polegająca na skorygowaniu wyjściowych wag wynikających ze schematu losowania próby, celem redukcji obciążenia wynikającego z istnienia braków odpowiedzi, tak aby spełnione było odpowiednie równanie kalibracyjne. Wagi te obliczane są w oparciu o wykorzystanie informacji dodatkowych z lub spoza próby.*

Istnieje kilka powodów, dla których kalibracja jest wykorzystywana w praktyce. Poza oczywistym argumentem redukcji obciążenia i zwiększenia precyzji szacunków poprzez włączenie dodatkowych informacji, wskazuje się również na inne aspekty powodujące, że w badaniach statystycznych odgrywać zaczęła istotną rolę. Do innych czynników można zaliczyć, por. J. Gambino (1999):

1. „Równowagę" – rozumianą w ten sposób, że po zastosowaniu kalibracji próba „wyglądem" jest zbliżona do całej populacji. Należy rozumieć to w ten sposób, że jeżeli zastosujemy wagi kalibracyjne do zmiennej, w oparciu o którą szacujemy intere-

sujący nas parametr, to powinniśmy uzyskać wyniki bliskie „prawdziwej" wartości tego parametru w populacji generalnej.

2. „Wygodę" – którą można rozumieć dwojako. Po pierwsze, zastosowanie kalibracji umożliwia wykorzystanie klasycznych metod statystycznych opartych na systemie wag (na przykład regresji ważonej), które nie dałaby poprawnych wyników, gdyby nie uwzględniono ujemnego wpływu braków odpowiedzi. Po drugie, w oparciu o pewien ustalony zbiór zmiennych pomocniczych można wyznaczyć wagi kalibracyjne, które będą użyteczne przy szacowaniu parametrów dla wielu interesujących badacza zmiennych. Może to znacznie skrócić proces estymacji.
3. „Logiczne szacunki" – po zastosowaniu kalibracji wyniki są bardziej logiczne, na przykład poszczególne wartości zmiennych pomocniczych przemnożone przez wagi kalibracyjne i zsumowane na zbiorze wszystkich respondentów, sumują się do ich wartości globalnych.

W dalszej części pracy przedstawione zostaną wybrane estymatory kalibracyjne wartości globalnej. W tym celu, w pierwszej kolejności, wprowadzimy niezbędne pojęcia i stosowane oznaczenia.

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:

$$Y = \sum_{i=1}^{N} y_i, \tag{2.1}$$

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

Klasycznym estymatorem wartości globalnej (2.1) jest znany z metody reprezentacyjnej estymator Horvitza-Thompsona, który wyraża się wzorem:

$$\hat{Y}_{HT} = \sum_{s} d_i y_i = \sum_{i=1}^{n} d_i y_i. \tag{2.2}$$

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 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.

Niech $r \subseteq s$ oznacza zbiór respondentów, dla których znana jest wartość zmiennej y. Załóżmy, że jest to zbiór m – elementowy, przy czym $m \leqslant n$. W szczególnym przypadku, gdy $r = s$, oznacza to, że wszystkie jednostki badania wylosowane do próby s udzieliły odpowiedzi i znana jest dla nich wartość zmiennej y na poziomie całej próby. Sytuacja taka w praktyce zdarza się jednak niezwykle rzadko (na skutek odmowy, braku czasu, choroby, nieobecności w domu, drażliwości pytania itp.), w związku z czym wagi d_i powinny zostać odpowiednio skorygowane (skalibrowane), aby zniwelować obciążenie wynikające z braków odpowiedzi.

W przypadku gdy w badaniu, w odniesieniu do zmiennej y, wystąpią braki odpowiedzi estymator (2.2) przyjmuje postać:

$$\hat{Y}_{HT} = \sum_{r} d_i y_i = \sum_{i=1}^{m} d_i y_i. \tag{2.3}$$

Ważona suma (2.3) jest zazwyczaj niedoszacowana w stosunku do wartości globalnej (2.1). Zgodnie z ideą kalibracji, wagi wynikające ze schematu losowania jednostek do próby, powinny być zwiększone, aby zrekompensować utratę informacji związaną z brakami danych. Poszukujemy nowych wag, dla wszystkich jednostek w próbie, dla których mamy informację o zmiennej y. 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. Funkcja taka powinna nie tylko posiadać odpowiednie własności w sensie matematycznym, ale również powinna umożliwić przedstawienie wektora wag $\mathbf{w} = (w_1, \ldots, w_m)^T$ w postaci jawnej[27]. W literaturze proponuje się szereg różnych postaci funkcji odległości D między wektorem wag wynikającym ze schematu losowania próby $\mathbf{d} = (d_1, \ldots, d_m)^T$, a wektorem wag kalibracyjnych $\mathbf{w} = (w_1, \ldots, w_m)^T$, por. J-C. Deville, C-E. Särndal, O. Sautory (1993), D.M. Stukel, M.A. Hidiroglou, C-E. Särndal (1996).

Na potrzeby pracy przyjmiemy funkcję odległości postaci:

$$D(\mathbf{w}, \mathbf{d}) = \frac{1}{2} \sum_{i=1}^{m} \frac{(w_i - d_i)^2}{d_i}. \tag{2.4}$$

Przyjęto więc, że funkcja $G(\cdot)$ występująca w warunku (W1 – 1.7) określającym postać funkcji odległości jest kwadratowa tj. wyraża się wzorem (1.10). Istotnie, w tym przypadku mamy, że[28]:

$$D(\mathbf{w}, \mathbf{d}) = \sum_{i=1}^{m} d_i G\left(\frac{w_i}{d_i}\right) = \sum_{i=1}^{m} d_i \frac{1}{2}\left(\frac{w_i}{d_i} - 1\right)^2 = \frac{1}{2} \sum_{i=1}^{m} \frac{(w_i - d_i)^2}{d_i}. \tag{2.5}$$

Funkcja ta posiada następujące własności:

- $D(\mathbf{w}, \mathbf{d}) \geqslant 0$,
- $D(\mathbf{d}, \mathbf{d}) = 0$,
- jest różniczkowalna względem w_i, $i = 1, \ldots, m$,
- jest ściśle wypukła,
- $\frac{\partial D(\mathbf{w}, \mathbf{d})}{\partial \mathbf{w}}$ jest ciągła.

[27] Nie jest to jednak istotne ograniczenie, gdyż w przypadku gdy nie jest możliwe poszukanie wektora wag kalibracyjnych w jawnej postaci dla pewnej funkcji odległości, należy zastosować odpowiednie algorytmy iteracyjne, które umożliwią znalezienie przybliżonego rozwiązania.

[28] Ze względu na założenie, że w odniesieniu do zmiennej y nie dysponujemy wszystkimi jej wartościami na poziomie próby s, sumowanie odbywa się po m – elementowym zbiorze respondentów.

Spełnia ona wszystkie własności nakładane na funkcję odległości, a ponadto umożliwia uzyskanie wektora wag kalibracyjnych w jawnej postaci.

Efektywne poszukiwanie wektora wag kalibracyjnych $\mathbf{w}$ jest uzależnione od informacji, jaką niesie wektor zmiennych pomocniczych $\mathbf{x}$. Niech k oznacza liczbę zmiennych pomocniczych, a $\mathbf{X_r}$ macierz utworzoną z wartości wszystkich zmiennych pomocniczych na poziomie zbioru respondentów r, dla których znana jest wartość zmiennej y, tj.:

$$\mathbf{X_r} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1k} \\ x_{21} & x_{22} & \dots & x_{2k} \\ \dots & \dots & \dots & \dots \\ x_{m1} & x_{m2} & \dots & x_{mk} \end{pmatrix}, \tag{2.6}$$

gdzie x_{ij} oznacza wartość zmiennej j dla jednostki badania i, przy czym $i = 1, \dots, m$, $j = 1, \dots, k$. Macierz (2.6) zawiera informacje na temat wszystkich zmiennych pomocniczych dla wszystkich respondentów, tzn. dla wszystkich jednostek wylosowanych do próby s, dla których znana jest wartość zmiennej y.
Niech ponadto

$$\mathbf{X_s} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1k} \\ x_{21} & x_{22} & \dots & x_{2k} \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{nk} \end{pmatrix}, \tag{2.7}$$

gdzie x_{ij} oznacza wartość zmiennej j dla jednostki badania i, przy czym $i = 1, \dots, n$, $j = 1, \dots, k$. Macierz (2.7) zawiera informacje na temat wszystkich zmiennych pomocniczych na poziomie całej próby s, tzn. również dla tych jednostek, dla których nie jest znana wartość zmiennej y.
Niech ponadto $\mathbf{X}$ oznacza wektor utworzony z wartości globalnych każdej zmiennej pomocniczej:

$$\mathbf{X} = \left(\sum_{i=1}^{N} x_{i1}, \sum_{i=1}^{N} x_{i2}, \dots, \sum_{i=1}^{N} x_{ik} \right)^T. \tag{2.8}$$

Poszczególne składowe wektora (2.8) odnoszą się do wartości globalnych kolejnych zmiennych pomocniczych. Informacje na temat takich zmiennych można pozyskać ze spisów bądź z odpowiednich rejestrów administracyjnych. W przypadku nieznajomości odpowiednich wartości globalnych stanowiących elementy wektora (2.8) zastępujemy je ocenami estymatora Horvitza-Thompsona ze wzoru (2.2), przy czym zmienną y zastępujemy odpowiednią zmienną pomocniczą x_i, $i = 1, \dots, k$. Uzyskujemy w ten sposób wektor postaci:

$$\breve{\mathbf{X}} = \left(\sum_{i=1}^{n} d_i x_{i1}, \sum_{i=1}^{n} d_i x_{i2}, \dots, \sum_{i=1}^{n} d_i x_{ik} \right)^T. \tag{2.9}$$

Na potrzeby konstrukcji estymatora kalibracyjnego wartości globalnej Y rozpatrzone zostaną trzy przypadki. W pierwszym, zakładać będziemy, że znany jest wektor wartości globalnych zmiennych pomocniczych. W drugim, przyjmiemy założenie, że nie dysponujemy takim wektorem, a posiadamy jedynie informacje w odniesieniu do zmiennych pomocniczych tylko na poziomie próby s. W ostatnim przypadku zakładać

będziemy, że dysponujemy informacją o wartościach globalnych zmiennych pomocniczych na poziomie całej populacji oraz, że możliwe jest oszacowanie innych wartości globalnych w oparciu o dane pochodzące z próby s.

2.2. Estymator kalibracyjny wartości globalnej ze znanym wektorem X

W pierwszym rozważanym podejściu zakładamy, że:

- dla każdego respondenta znana jest wartość dla każdej zmiennej pomocniczej, tj. znana jest macierz $\mathbf{X_r}$ określona wzorem (2.6),
- znany jest wektor wartości globalnych wszystkich zmiennych pomocniczych (2.8).

W podejściu tym nie wymaga się, aby znane były wartości zmiennych pomocniczych dla wszystkich jednostek wylosowanych do próby s, ale tylko w odniesieniu do respondentów, dla których znane są wartości zmiennej y.

Definicja 4. *Estymatorem kalibracyjnym wartości globalnej (2.1) przy znanym wektorze wartości globalnych* $\mathbf{X}$ *zmiennych pomocniczych jest:*

$$\hat{Y}_{\mathbf{X}} = \sum_{i=1}^{m} w_i y_i, \tag{2.10}$$

gdzie wektor wag $\mathbf{w} = (w_1, \ldots, w_m)^T$ *jest rozwiązaniem zadania optymalizacyjnego*

$$\mathbf{w} = \operatorname{argmin}_{\mathbf{v}} D(\mathbf{v}, \mathbf{d}), \tag{2.11}$$

przy warunku

$$\mathbf{X} = \tilde{\mathbf{X}}, \tag{2.12}$$

gdzie

$$\tilde{\mathbf{X}} = \left(\sum_{i=1}^{m} w_i x_{i1}, \sum_{i=1}^{m} w_i x_{i2}, \ldots, \sum_{i=1}^{m} w_i x_{ik} \right)^T, \tag{2.13}$$

a wektor wartości globalnych $\mathbf{X}$ *zdefiniowany jest wzorem (2.8).*

Równanie (2.12) jest tzw. równaniem kalibracyjnym. Zgodnie z nim, wektora wag $\mathbf{w}$ poszukujemy w ten sposób, aby dla każdej zmiennej pomocniczej jej wartość globalna równała się sumie odpowiednio przeważonych wartości zmiennych pomocniczych określonych na zbiorze respondentów r.

Zgodnie z ideą kalibracji, jeżeli będziemy dysponowali odpowiednio dobranym wektorem zmiennych pomocniczych, to wagi kalibracyjne spełniające równanie kalibracyjne (2.12), powinny umożliwić „lepsze" oszacowanie wartości globalnej zmiennej y. Innymi słowy, przy odpowiednio dobranym wektorze zmiennych pomocniczych — wagi, które będziemy wykorzystywać przy szacowaniu wartości globalnej zmiennej y — powinny „naśladować" zachowanie się wag, o których mowa w definicji (4). Oznacza to, że wektora wag kalibracyjnych szukamy w taki sposób, aby po ich przemnożeniu przez odpowiednie wartości poszczególnych zmiennych pomocniczych otrzymać ich wartości globalne — a w konsekwencji po zastosowaniu do zmiennej y uzyskać „lepsze" oszacowanie wartości globalnej (2.1). Informacja zawarta w wektorze zmiennych

pomocniczych, jest więc wykorzystywana w „ochronie" przed obciążeniem związanym z występowaniem braków odpowiedzi oraz celem redukcji wariancji estymatorów. Informacje tego typu można pozyskać z tego samego badania, na podstawie którego szacujemy wartość globalną zmiennej y bądź ze spisów, rejestrów administracyjnych lub z innych dostępnych źródeł danych.

Wybór odpowiednich cech do wektora zmiennych pomocniczych, jest więc kluczowym zagadnieniem z punktu widzenia procesu estymacji z wykorzystaniem podejścia kalibracyjnego. W literaturze panuje powszechny pogląd, że w przypadku odpowiednio dobranego wektora zmiennych pomocniczych, kalibracja jako metoda korygowania wyjściowych wag posiada kilka istotnych zalet, por. V.M. Estevao, C-E. Särndal (2000):

1. „Zgodność" — rozumianą w ten sposób, że system wag spełnia odpowiednie równanie kalibracyjne. Wagi są więc tak określane, aby „naśladowały" zachowanie się zmiennych pomocniczych tj. aby ich przeważone wartości sumowały się do wartości globalnych.
2. „Odległość" — wagi kalibracyjne konstruowane są w ten sposób, aby były bliskie (w sensie przyjętej funkcji odległości) wyjściowym wagom, będących odwrotnościami prawdopodobieństw inkluzji pierwszego rzędu. Wynika to z faktu, że estymatory konstruowane z użyciem wektora wag, wynikających ze schematu losowania próby, mają często bardzo pożądaną własność jaką jest nieobciążoność. Dlatego wszelkie zmiany wyjściowych wag powinny być „małe", aby zachować nieobciążoność.
3. „Kontrola wartości globalnych zmiennych pomocniczych" — która przejawia się tym, że czym więcej odpowiednio dobranych zmiennych pomocniczych i ich wartości globalnych wykorzystamy w procesie kalibracji, tym uzyskane wagi będą „lepiej" wykalibrowane. W efekcie wariancja estymatora i jego obciążenie będą wykazywały tendencję spadkową.

Obok istotnych zalet, należy zwrócić uwagę na fakt, że kalibracja przy źle dobranym wektorze zmiennych pomocniczych, może prowadzić do błędnych oszacowań. Jest to konsekwencją tego, że w niektórych przypadkach możliwe jest uzyskanie ujemnych lub bardzo dużych wag dodatnich. Istnieje oczywiście możliwość nałożenia na wagi kalibracyjne pewnych warunków nieujemności bądź zawierania się ich, w pewnym z góry określonym przedziale, wiąże się to jednak z trudnościami z ich wyznaczeniem w jawnej postaci oraz koniecznością zastosowania skomplikowanych algorytmów obliczeniowych.

Poniższe twierdzenie rozstrzyga postać wag estymatora kalibracyjnego określonego w definicji (4).

Twierdzenie 1. *Rozwiązaniem zadania minimalizacji funkcji odległości (2.11), przy warunku (2.12) jest wektor wag kalibracyjnych postaci* $\mathbf{w} = (w_1, \ldots, w_m)^T$*, którego składowe spełniają równanie:*

$$w_i = d_i + d_i \left(\mathbf{X} - \hat{\mathbf{X}}\right)^T \left(\sum_{i=1}^{m} d_i \boldsymbol{x}_i \boldsymbol{x}_i^T\right)^{-1} \boldsymbol{x}_i, \tag{2.14}$$

przy czym

$$\hat{\mathbf{X}} = \left(\sum_{i=1}^{m} d_i x_{i1}, \sum_{i=1}^{m} d_i x_{i2}, \ldots, \sum_{i=1}^{m} d_i x_{ik}\right)^T, \tag{2.15}$$

a

$$\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ik})^T, \tag{2.16}$$

jest wektorem złożonym z wartości wszystkich k zmiennych pomocniczych dla i – tego respondenta, $i = 1, \ldots, m$.

Dowód. Wagi kalibracyjne w_i dla $i = 1, \ldots, m$ znajdziemy, korzystając z metody czynników nieoznaczonych Lagrange'a.
Funkcja Lagrange'a ma postać:

$$L = \frac{1}{2}\sum_{i=1}^{m}\frac{(w_i - d_i)^2}{d_i} + \sum_{j=1}^{k}\lambda_j\left(\sum_{i=1}^{N}x_{ij} - \sum_{i=1}^{m}w_i x_{ij}\right), \tag{2.17}$$

gdzie $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_k)^T$ jest wektorem złożonym z czynników nieoznaczonych Lagrange'a. Pochodna funkcji L dla $i = 1, \ldots, m$ ma postać:

$$\frac{\partial L}{\partial w_i} = \frac{1}{2}\cdot\frac{2w_i - 2d_i}{d_i} - \sum_{j=1}^{k}\lambda_j x_{ij}. \tag{2.18}$$

Przyrównując obliczoną pochodną do zera, otrzymujemy następujące równanie:

$$\frac{w_i - d_i}{d_i} = \sum_{j=1}^{k}\lambda_j x_{ij}, \tag{2.19}$$

którego rozwiązaniem dla $i = 1, \ldots, m$ jest:

$$w_i = d_i\left(1 + \sum_{j=1}^{k}\lambda_j x_{ij}\right). \tag{2.20}$$

Ponieważ

$$\sum_{j=1}^{k}\lambda_j x_{ij} = \boldsymbol{x}_i^T\boldsymbol{\lambda} = \boldsymbol{\lambda}^T\boldsymbol{x}_i, \tag{2.21}$$

więc wagi (2.20) można przedstawić jako:

$$w_i = d_i\left(1 + \boldsymbol{x}_i^T\boldsymbol{\lambda}\right) = d_i\left(1 + \boldsymbol{\lambda}^T\boldsymbol{x}_i\right). \tag{2.22}$$

Mnożąc obustronnie równanie (2.22) przez $\boldsymbol{x}_i$, a następnie dokonując sumowania po zbiorze wszystkich respondentów r otrzymujemy

$$\sum_{i=1}^{m}w_i\boldsymbol{x}_i = \sum_{i=1}^{m}d_i\boldsymbol{x}_i\left(1 + \boldsymbol{x}_i^T\boldsymbol{\lambda}\right). \tag{2.23}$$

Dokonując przekształcenia ostatniego równania, otrzymujemy

$$\sum_{i=1}^{m}w_i\boldsymbol{x}_i - \sum_{i=1}^{m}d_i\boldsymbol{x}_i = \sum_{i=1}^{m}d_i\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{\lambda}. \tag{2.24}$$

Podstawiając do równania (2.24) wektory $\tilde{\mathbf{X}}$ i $\hat{\mathbf{X}}$, otrzymujemy równanie postaci

$$\tilde{\mathbf{X}} - \hat{\mathbf{X}} = \sum_{i=1}^{m} d_i \boldsymbol{x}_i \boldsymbol{x}_i^T \boldsymbol{\lambda}. \tag{2.25}$$

Korzystając z równania kalibracyjnego (2.12) i zakładając, że macierz $\sum_{i=1}^{m} d_i \boldsymbol{x}_i \boldsymbol{x}_i^T$ jest nieosobliwa, otrzymujemy następującą postać wektora $\boldsymbol{\lambda}$:

$$\boldsymbol{\lambda} = \left(\sum_{i=1}^{m} d_i \boldsymbol{x}_i \boldsymbol{x}_i^T \right)^{-1} \left(\mathbf{X} - \hat{\mathbf{X}} \right). \tag{2.26}$$

Ponieważ macierz $\sum_{i=1}^{m} d_i \boldsymbol{x}_i \boldsymbol{x}_i^T$ jest symetryczna, to:

$$\boldsymbol{\lambda}^T = \left(\mathbf{X} - \hat{\mathbf{X}} \right)^T \left(\sum_{i=1}^{m} d_i \boldsymbol{x}_i \boldsymbol{x}_i^T \right)^{-1}. \tag{2.27}$$

Podstawiając (2.27) do równania (2.22) otrzymujemy szukaną postać wag w_i.

Należy jeszcze sprawdzić, że w punkcie $\mathbf{w} = (w_1, \ldots, w_m)^T$ istnieje minimum (warunek dostateczny istnienia ekstremum warunkowego). Niech ξ będzie niezerowym wektorem takim, że $\xi \in \mathbb{R}^m$. Należy wykazać, że forma kwadratowa $d^2 L(\mathbf{w})(\xi)$ jest dodatnio określona. Mamy:

$$d^2 L(\mathbf{w})(\xi) = \sum_{i,j=1}^{m} \frac{\partial^2 L}{\partial w_i \partial w_j} \xi_i \xi_j. \tag{2.28}$$

Zauważmy, że:

$$\frac{\partial^2 L}{\partial w_i w_j} = \begin{cases} \pi_i & \text{dla } i = j, \\ 0 & \text{dla } i \neq j. \end{cases} \tag{2.29}$$

Podstawiając obliczone pochodne drugiego rzędu do formy kwadratowej (2.28) otrzymujemy, że:

$$d^2 L(\mathbf{w})(\xi) = \sum_{i,j=1}^{m} \frac{\partial^2 L}{\partial w_i \partial w_j} \xi_i \xi_j = \sum_{i=1}^{m} \pi_i \xi_i^2. \tag{2.30}$$

Jest to oczywiście forma kwadratowa dodatnio określona. Stąd wektor wag kalibracyjnych $\mathbf{w} = (w_1, \ldots, w_m)^T$, którego składowe spełniają warunek (2.14) jest poszukiwanym rozwiązaniem zadania minimalizacji funkcji odległości. ■

Jedną z bardzo ważnych własności estymatora kalibracyjnego wartości globalnej zmiennej y, w przypadku gdy znany jest wektor $\mathbf{X}$ złożony z wartości globalnych zmiennych pomocniczych, podaje poniższe twierdzenie. Zgodnie z nim, jeżeli związek pomiędzy zmienną y, a zmiennymi pomocniczymi $x_1, \ldots, x_k$ jest liniowy na poziomie całej populacji, to wówczas jesteśmy w stanie uzyskać dokładne oszacowanie wartości globalnej zmiennej y.

Oczywiście, w praktyce badań statystycznych, taka liniowa zależność na poziomie całej populacji nie istnieje. Twierdzenie to jednak pokazuje, że w przypadku gdy istnieje silna liniowa zależność między zmiennymi y i $x_1, \ldots, x_k$ to ocena punktowa estymatora wartości globalnej zmiennej y, powinna być bliska jej prawdziwej wartości.

Twierdzenie 2. *Niech między zmienną y i zmiennymi pomocniczymi $x_1, \ldots, x_k$ istnieje liniowa zależność na poziomie całej populacji, tzn. dla każdej jednostki $i \in U$ spełniona będzie zależność*

$$y_i = \boldsymbol{x}_i^T \boldsymbol{\beta}, \tag{2.31}$$

gdzie $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ik})^T$ jest wektorem złożonym z wartości wszystkich zmiennych pomocniczych, a $\boldsymbol{\beta}$ jest kolumnowym wektorem złożonym z współczynników regresji, $i = 1, \ldots, N$. Wtedy

$$\hat{Y}_{\mathbf{X}} = Y. \tag{2.32}$$

Dowód. Mamy

$$\hat{Y}_{\mathbf{X}} = \sum_{i=1}^{m} w_i y_i = \sum_{i=1}^{m} w_i \boldsymbol{x}_i^T \boldsymbol{\beta} = \left(\sum_{i=1}^{m} w_i \boldsymbol{x}_i \right)^T \boldsymbol{\beta} = \mathbf{X}^T \boldsymbol{\beta} = \sum_{i=1}^{N} y_i = Y. \tag{2.33}$$

Odpowiednie równości wynikają z przyjęcia liniowej zależności oraz z równania kalibracyjnego (2.12). ■

2.3. Estymator kalibracyjny wartości globalnej ze znanym wektorem $\breve{\mathbf{X}}$

W drugim rozważanym podejściu zakładamy, że:

- nie jest znany wektor wartości globalnych zmiennych pomocniczych (2.8),
- dla każdej jednostki wylosowanej do próby s znana jest wartość każdej zmiennej pomocniczej, tj. znana jest macierz $\mathbf{X_s}$ określona wzorem (2.7),
- znany jest wektor oszacowanych wartości globalnych $\breve{\mathbf{X}}$ określony wzorem (2.9).

W podejściu tym wymaga się, aby znane były wartości zmiennych pomocniczych dla wszystkich jednostek wylosowanych do próby s. Znajomość macierzy $\mathbf{X_s}$ jest niezbędna do oszacowania wartości globalnych poszczególnych zmiennych pomocniczych, które stanowią składowe wektora $\breve{\mathbf{X}}$.

Definicja 5. *Estymatorem kalibracyjnym wartości globalnej (2.1) przy znanym wektorze $\breve{\mathbf{X}}$ oszacowanych wartości globalnych zmiennych pomocniczych jest:*

$$\hat{Y}_{\breve{\mathbf{X}}} = \sum_{i=1}^{m} w_i y_i, \tag{2.34}$$

gdzie wektor wag $\mathbf{w} = (w_1, \ldots, w_m)^T$ jest rozwiązaniem zadania optymalizacyjnego

$$\mathbf{w} = \operatorname{argmin}_{\mathbf{v}} D(\mathbf{v}, \mathbf{d}), \tag{2.35}$$

przy warunku

$$\breve{\mathbf{X}} = \tilde{\mathbf{X}}, \tag{2.36}$$

gdzie $\breve{\mathbf{X}}$ i $\tilde{\mathbf{X}}$ zdefiniowane są wzorami (2.9) i (2.13) odpowiednio.

Równanie (2.36) jest w powyższej definicji tzw. równaniem kalibracyjnym. Zgodnie z nim, wektora wag $\mathbf{w}$ poszukujemy w ten sposób, aby dla każdej zmiennej pomocniczej, jej oszacowana wartość globalna równała się sumie odpowiednio przeważonych

wartości zmiennych pomocniczych określonych na zbiorze respondentów r. Definicja estymatora kalibracyjnego (5) różni się od definicji (4) użytym w równaniu kalibracyjnym wektorem. W definicji (4) zakładamy, że znane są wartości globalne zmiennych pomocniczych na poziomie całej populacji, a w definicji (5) zakładamy, że w ich miejsce podstawiamy oszacowania uzyskane w oparciu o dane pochodzące z próby s.

Postać estymatora kalibracyjnego określonego w definicji (5), rozstrzyga poniższe twierdzenie.

Twierdzenie 3. *Rozwiązaniem zadania minimalizacji funkcji odległości (2.35), przy warunku (2.36) jest wektor wag kalibracyjnych postaci* $\mathbf{w} = (w_1, \ldots, w_m)^T$, *którego składowe spełniają równanie:*

$$w_i = d_i + d_i \left(\breve{\mathbf{X}} - \hat{\mathbf{X}}\right)^T \left(\sum_{i=1}^{m} d_i \boldsymbol{x}_i \boldsymbol{x}_i^T\right)^{-1} \boldsymbol{x}_i, \tag{2.37}$$

dla $i = 1, \ldots, m$.

Dowód. Dowód twierdzenia jest analogiczny do dowodu twierdzenia (1). Zamiast wektora wartości globalnych $\mathbf{X}$ należy podstawić wektor oszacowanych wartości globalnych $\breve{\mathbf{X}}$, korzystając z równania kalibracyjnego (2.36). ■

2.4. Uogólniony estymator kalibracyjny wartości globalnej

W rozważanym podejściu zakładamy, że:

- znany jest wektor wartości globalnych $\mathbf{X}$ wszystkich zmiennych pomocniczych (2.8),
- dla każdej jednostki wylosowanej do próby s znana jest wartość każdej zmiennej pomocniczej, tj. znana jest macierz $\mathbf{X_s}$ określona wzorem (2.7),
- znany jest wektor $\breve{\mathbf{X}}$ oszacowanych wartości globalnych określony wzorem (2.9).

Nie będziemy przy tym wymagać, aby wektory $\mathbf{X}$ i $\breve{\mathbf{X}}$ były jednakowych wymiarów. Oznacza to, że w odniesieniu do k–elementowego zbioru zmiennych pomocniczych $x_1, \ldots x_k$, dysponować będziemy dla niektórych zmiennych ich wartościami globalnymi, a dla pozostałych ich oszacowaniami. W szczególnym przypadku może się zdarzyć, że znane będą wszystkie wartości globalne zmiennych pomocniczych bądź tylko ich oszacowania na podstawie próby s.

Wyznaczone wagi kalibracyjne umożliwią nam, po ich zastosowaniu do poszczególnych zmiennych pomocniczych, otrzymać ich wartości globalne oraz oszacowania (w przypadku zmiennych, dla których nie będą znane wartości globalne). Można się więc spodziewać, że wagi wyznaczone w ten sposób, jeszcze lepiej będą chroniły przed obciążeniem i wariancją wynikającymi z występujących w próbie braków danych.

Precyzując powyższe rozważania załóżmy, że dla k_1 zmiennych pomocniczych znane są wartości globalne. Dla k_2 zmiennych pomocniczych znane są natomiast oszacowania wartości globalnych w oparciu o dane pochodzące z próby s, przy czym $k_1, k_2 \leqslant k$ oraz $k_1 + k_2 = k$. W szczególnym przypadku, gdy $k_1 = k$ oraz $k_2 = 0$ zakładamy, że wyznaczamy postać wag kalibracyjnych przyjmując, że znane są wartości globalne wszystkich k zmiennych pomocniczych, por. estymator kalibracyjny ze znanym wektorem $\mathbf{X}$. Natomiast w sytuacji, gdy $k_1 = 0$ oraz $k_2 = k$ zakładamy, że wyznaczamy postać wag

kalibracyjnych przyjmując, że znane są tylko oszacowania wartości globalnych wszystkich k zmiennych pomocniczych, por. estymator kalibracyjny ze znanym wektorem $\breve{\mathbf{X}}$.

Definicja 6. *Uogólnionym estymatorem kalibracyjnym wartości globalnej (2.1) jest:*

$$\hat{Y}_{cal} = \sum_{i=1}^{m} w_i y_i, \tag{2.38}$$

gdzie wektor wag $\mathbf{w} = (w_1, \ldots, w_m)^T$ *jest rozwiązaniem zadania optymalizacyjnego*

$$\mathbf{w} = \operatorname{argmin}_{\mathbf{v}} D(\mathbf{v}, \mathbf{d}), \tag{2.39}$$

przy warunku

$$\begin{pmatrix} \mathbf{X_1} \\ \breve{\mathbf{X}}_\mathbf{2} \end{pmatrix} = \begin{pmatrix} \tilde{\mathbf{X}}_\mathbf{1} \\ \tilde{\mathbf{X}}_\mathbf{2} \end{pmatrix} \tag{2.40}$$

przy czym poszczególne wektory są postaci[29]*:*

$$\mathbf{X_1} = \left(\sum_{i=1}^{N} x_{i1}, \sum_{i=1}^{N} x_{i2}, \ldots, \sum_{i=1}^{N} x_{ik_1} \right)^T, \tag{2.41}$$

$$\breve{\mathbf{X}}_\mathbf{2} = \left(\sum_{i=1}^{n} d_i x_{i1}, \sum_{i=1}^{n} d_i x_{i2}, \ldots, \sum_{i=1}^{n} d_i x_{ik_2} \right)^T, \tag{2.42}$$

$$\tilde{\mathbf{X}}_\mathbf{1} = \left(\sum_{i=1}^{m} w_i x_{i1}, \sum_{i=1}^{m} w_i x_{i2}, \ldots, \sum_{i=1}^{m} w_i x_{ik_1} \right)^T, \tag{2.43}$$

$$\tilde{\mathbf{X}}_\mathbf{2} = \left(\sum_{i=1}^{m} w_i x_{i1}, \sum_{i=1}^{m} w_i x_{i2}, \ldots, \sum_{i=1}^{m} w_i x_{ik_2} \right)^T, \tag{2.44}$$

gdzie $k_1, k_2 \leqslant k$.

Równanie (2.40) jest w powyższej definicji równaniem kalibracyjnym. Zgodnie z nim, wektora wag $\mathbf{w}$ poszukujemy w ten sposób, aby dla zmiennych pomocniczych, dla których znane są wartości globalne, zsumowane na zbiorze respondentów r, a przeważone wagami kalibracyjnymi wartości tych zmiennych, równały się wartościom globalnym.

Podobnie ze zmiennymi, dla których nie są znane ich wartości globalne. W tym przypadku, zsumowane na zbiorze respondentów r, a przeważone wagami kalibracyjnymi wartości tych zmiennych, mają się równać oszacowanym wartościom globalnym.

Twierdzenie 4. *Rozwiązaniem zadania minimalizacji funkcji odległości (2.39), przy warunku (2.40) jest wektor wag kalibracyjnych postaci* $\mathbf{w} = (w_1, \ldots, w_m)^T$*, którego składowe spełniają równanie:*

$$w_i = d_i + d_i \left(\begin{pmatrix} \mathbf{X_1} \\ \breve{\mathbf{X}}_\mathbf{2} \end{pmatrix} - \begin{pmatrix} \hat{\mathbf{X}}_\mathbf{1} \\ \hat{\mathbf{X}}_\mathbf{2} \end{pmatrix} \right)^T \left(\sum_{i=1}^{m} d_i \boldsymbol{x}_i \boldsymbol{x}_i^T \right)^{-1} \boldsymbol{x}_i, \tag{2.45}$$

[29] W oznaczeniach poniższych wektorów, zgodnie z poczynionymi wcześniej uwagami, zakładamy, że jeżeli dla jakiejś zmiennej pomocniczej znana jest jej wartość globalna, to zmiennej tej nie uwzględniamy w konstrukcji wektora z oszacowanymi wartościami globalnymi.

gdzie $\boldsymbol{x}_i = (\boldsymbol{x}_1, \boldsymbol{x}_2)^T$ *jest wektorem, którego składowe są postaci* $\boldsymbol{x}_1 = (x_{i1}, \ldots, x_{ik_1})$, $\boldsymbol{x}_2 = (x_{i1}, \ldots, x_{ik_2})$,*a wektory* $\hat{\mathbf{X}}_\mathbf{1}$ *i* $\hat{\mathbf{X}}_\mathbf{2}$ *określone są w następujący sposób:*

$$\hat{\mathbf{X}}_\mathbf{1} = \left(\sum_{i=1}^{m} d_i x_{i1}, \sum_{i=1}^{m} d_i x_{i2}, \ldots, \sum_{i=1}^{m} d_i x_{ik_1} \right)^T, \tag{2.46}$$

$$\hat{\mathbf{X}}_\mathbf{2} = \left(\sum_{i=1}^{m} d_i x_{i1}, \sum_{i=1}^{m} d_i x_{i2}, \ldots, \sum_{i=1}^{m} d_i x_{ik_2} \right)^T. \tag{2.47}$$

Dowód. Wagi kalibracyjne w_i dla $i = 1, \ldots, m$ znajdziemy, korzystając z metody czynników nieoznaczonych Lagrange'a.
Funkcja Lagrange'a ma postać:

$$L = \frac{1}{2} \sum_{i=1}^{m} \frac{(w_i - d_i)^2}{d_i} + \sum_{j=1}^{k_1} \lambda_j \left(\sum_{i=1}^{N} x_{ij} - \sum_{i=1}^{m} w_i x_{ij} \right) + \sum_{j=1}^{k_2} \lambda_{k_1+j} \left(\sum_{i=1}^{n} d_i x_{ij} - \sum_{i=1}^{m} w_i x_{ij} \right), \tag{2.48}$$

gdzie $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_{k_1+k_2})^T$ jest wektorem złożonym z czynników nieoznaczonych Lagrange'a. Pochodna funkcji L dla $i = 1, \ldots, m$ ma postać:

$$\frac{\partial L}{\partial w_i} = \frac{1}{2} \cdot \frac{2w_i - 2d_i}{d_i} - \sum_{j=1}^{k_1} \lambda_j x_{ij} - \sum_{j=1}^{k_2} \lambda_{k_1+j} x_{ij}. \tag{2.49}$$

Przyrównując obliczoną pochodną do zera, otrzymujemy następujące równanie:

$$\frac{w_i - d_i}{d_i} = \sum_{j=1}^{k_1} \lambda_j x_{ij} + \sum_{j=1}^{k_2} \lambda_{k_1+j} x_{ij}, \tag{2.50}$$

którego rozwiązaniem dla $i = 1, \ldots, m$ jest:

$$w_i = d_i \left(1 + \sum_{j=1}^{k_1} \lambda_j x_{ij} + \sum_{j=1}^{k_2} \lambda_{k_1+j} x_{ij} \right). \tag{2.51}$$

Ponieważ

$$\sum_{j=1}^{k_1} \lambda_j x_{ij} + \sum_{j=1}^{k_2} \lambda_{k_1+j} x_{ij} = \boldsymbol{x}_i^T \boldsymbol{\lambda} = \boldsymbol{\lambda}^T \boldsymbol{x}_i, \tag{2.52}$$

więc wagi (2.51) można przedstawić jako:

$$w_i = d_i \left(1 + \boldsymbol{x}_i^T \boldsymbol{\lambda} \right) = d_i \left(1 + \boldsymbol{\lambda}^T \boldsymbol{x}_i \right). \tag{2.53}$$

Mnożąc obustronnie równanie (2.53) przez $\boldsymbol{x}_i$, a następnie dokonując sumowania po zbiorze wszystkich respondentów r otrzymujemy

$$\sum_{i=1}^{m} w_i \boldsymbol{x}_i = \sum_{i=1}^{m} d_i \boldsymbol{x}_i \left(1 + \boldsymbol{x}_i^T \boldsymbol{\lambda} \right). \tag{2.54}$$

Dokonując przekształcenia ostatniego równania, otrzymujemy

$$\sum_{i=1}^{m} w_i \boldsymbol{x}_i - \sum_{i=1}^{m} d_i \boldsymbol{x}_i = \sum_{i=1}^{m} d_i \boldsymbol{x}_i \boldsymbol{x}_i^T \boldsymbol{\lambda}. \tag{2.55}$$

Stąd otrzymujemy, że:

$$\begin{pmatrix} \sum_{i=1}^{m} w_i \boldsymbol{x}_1 \\ \sum_{i=1}^{m} w_i \boldsymbol{x}_2 \end{pmatrix} - \begin{pmatrix} \sum_{i=1}^{m} d_i \boldsymbol{x}_1 \\ \sum_{i=1}^{m} d_i \boldsymbol{x}_2 \end{pmatrix} = \sum_{i=1}^{m} d_i \boldsymbol{x}_i \boldsymbol{x}_i^T \boldsymbol{\lambda} \tag{2.56}$$

Korzystając z (2.43), (2.44), (2.46) i (2.47) równanie (2.56) może być przedstawione w następującej postaci:

$$\begin{pmatrix} \tilde{\mathbf{X}}_{\mathbf{1}} \\ \tilde{\mathbf{X}}_{\mathbf{2}} \end{pmatrix} - \begin{pmatrix} \hat{\mathbf{X}}_{\mathbf{1}} \\ \hat{\mathbf{X}}_{\mathbf{2}} \end{pmatrix} = \sum_{i=1}^{m} d_i \boldsymbol{x}_i \boldsymbol{x}_i^T \boldsymbol{\lambda} \tag{2.57}$$

Korzystając z równania kalibracyjnego (2.40) i zakładając, że macierz $\sum_{i=1}^{m} d_i \boldsymbol{x}_i \boldsymbol{x}_i^T$ jest nieosobliwa otrzymujemy następującą postać wektora $\boldsymbol{\lambda}$:

$$\boldsymbol{\lambda} = \left(\sum_{i=1}^{m} d_i \boldsymbol{x}_i \boldsymbol{x}_i^T \right)^{-1} \left(\begin{pmatrix} \mathbf{X}_{\mathbf{1}} \\ \breve{\mathbf{X}}_{\mathbf{2}} \end{pmatrix} - \begin{pmatrix} \hat{\mathbf{X}}_{\mathbf{1}} \\ \hat{\mathbf{X}}_{\mathbf{2}} \end{pmatrix} \right). \tag{2.58}$$

Ponieważ macierz $\sum_{i=1}^{m} d_i \boldsymbol{x}_i \boldsymbol{x}_i^T$ jest symetryczna, to:

$$\boldsymbol{\lambda}^T = \left(\begin{pmatrix} \mathbf{X}_{\mathbf{1}} \\ \breve{\mathbf{X}}_{\mathbf{2}} \end{pmatrix} - \begin{pmatrix} \hat{\mathbf{X}}_{\mathbf{1}} \\ \hat{\mathbf{X}}_{\mathbf{2}} \end{pmatrix} \right)^T \left(\sum_{i=1}^{m} d_i \boldsymbol{x}_i \boldsymbol{x}_i^T \right)^{-1}. \tag{2.59}$$

Podstawiając (2.59) do równania (2.53) otrzymujemy szukaną postać wag w_i.

Należy jeszcze sprawdzić, że w punkcie $\mathbf{w} = (w_1, \ldots, w_m)^T$ istnieje minimum (warunek dostateczny istnienia ekstremum warunkowego). Niech ξ będzie niezerowym wektorem takim, że $\xi \in \mathbb{R}^m$. Należy wykazać, że forma kwadratowa $d^2 L(\mathbf{w})(\xi)$ jest dodatnio określona. Mamy:

$$d^2 L(\mathbf{w})(\xi) = \sum_{i,j=1}^{m} \frac{\partial^2 L}{\partial w_i \partial w_j} \xi_i \xi_j. \tag{2.60}$$

Zauważmy, że:

$$\frac{\partial^2 L}{\partial w_i w_j} = \begin{cases} \pi_i & \text{dla } i = j, \\ 0 & \text{dla } i \neq j. \end{cases} \tag{2.61}$$

Podstawiając obliczone pochodne drugiego rzędu do formy kwadratowej (2.60) otrzymujemy, że:

$$d^2 L(\mathbf{w})(\xi) = \sum_{i,j=1}^{m} \frac{\partial^2 L}{\partial w_i \partial w_j} \xi_i \xi_j = \sum_{i=1}^{m} \pi_i \xi_i^2. \tag{2.62}$$

Jest to oczywiście forma kwadratowa dodatnio określona. Stąd wektor wag kalibracyjnych $\mathbf{w} = (w_1, \ldots, w_m)^T$, którego składowe spełniają warunek (2.45) jest poszukiwanym rozwiązaniem zadania minimalizacji funkcji odległości. ■

2.5. Estymator kalibracyjny wartości globalnej – podejście funkcyjne

Koncepcja wykorzystania podejścia funkcyjnego opierającego się na tzw. wektorze zmiennych instrumentalnych w badaniach statystycznych z brakami odpowiedzi, pochodzi od C-E. Särndala i S. Lundströma (2005). Wprowadzenie pojęcia wektora zmiennych instrumentalnych umożliwia uzyskanie wielu różnych zbiorów wag kalibracyjnych, które w dalszym ciągu spełniać będą odpowiednie równania kalibracyjne, a niekoniecznie warunek, w którym zakłada się, że wyznaczone wagi mają być bliskie – w sensie przyjętej funkcji odległości – wyjściowym wagom. Jest to więc podejście alternatywne w stosunku do opisanych w poprzednich podrozdziałach metod konstrukcji estymatorów kalibracyjnych, które polegały na minimalizacji funkcji odległości przy zadanych warunkach w postaci równań kalibracyjnych.

Załóżmy dalej, że celem badania jest znalezienie estymatora kalibracyjnego wartości globalnej. Niech $\mathbf{d}_p = (d_{p1}, \ldots, d_{pm})^T$ oznacza tzw. wektor wag początkowych, takich, że $d_{pi} > 0$ dla każdego $i \in r$. Składowa d_{pi} oznacza więc wagę początkową przypisaną i–temu respondentowi. Dla przykładu, moglibyśmy przyjąć, że $d_{pi} = d_i$ bądź $d_{pi} = \frac{n}{m} d_i$. W pierwszym przypadku zakładamy, że wagi początkowe w podejściu kalibracyjnym tożsame są z wagami wynikającymi ze schematu losowania próby. W drugim zaś zakładamy, że są one równe wagom wynikającym ze schematu losowania próby przemnożonym przez odwrotność frakcji braków odpowiedzi dla zmiennej y. W podobny sposób możemy określać wagi początkowe, w rozważanym w tym podrozdziale podejściu funkcyjnym, na potrzeby zagadnienia poszukiwania wag kalibracyjnych. Ważne jedynie, aby były one dodatnie dla każdego respondenta.

Zwróćmy uwagę, że wagi kalibracyjne w_i określone wzorem (2.45) można byłoby alternatywnie zapisać jako

$$w_i = d_i v_i, \tag{2.63}$$

gdzie

$$v_i = 1 + \boldsymbol{\lambda}^T \boldsymbol{x}_i, \tag{2.64}$$

a $\boldsymbol{\lambda}^T$ wyraża się wzorem (2.59). Czynnik v_i, przez który przemnażana jest waga d_i można byłoby określić mianem korygującego lub kalibrującego. Mnożąc bowiem wagi wynikjące ze schematu losowania próby przez niego, uzyskujemy wagi końcowe w postaci wag kalibracyjnych.

Niech $\boldsymbol{z}_i$ bedzie dowolnym wektorem kolumnowym tego samego wymiaru co $\boldsymbol{x}_i$. Liczba wierszy tego wektora jest więc uzależniona od wymiarów wektora $\mathbf{X}$ wartości globalnych pewnych zmiennych pomocniczych i wektora $\breve{\mathbf{X}}$ oszacowanych na podstawie próby s wartości globalnych pozostałych zmiennych. W przypadku, gdy znanych jest k_1 wartości globalnych zmiennych pomocniczych na poziomie całej populacji i k_2 oszacowanych wartości globalnych na poziomie próby s, wektor ten można przedstawić w następującej postaci:

$$\boldsymbol{z}_i = (\underbrace{z_{i1}, \ldots, z_{k_1}}_{k_1}, \underbrace{z_{k_1+1}, \ldots, z_{k_2}}_{k_2})^T = (z_{i1}, \ldots, z_k)^T. \tag{2.65}$$

Definicja 7. *Wektor $\boldsymbol{z}_i$ nazywamy wektorem zmiennych instrumentalnych, jeżeli wagi kalibracyjne można przedstawić w postaci*

$$w_i = d_{pi} F\left(\boldsymbol{\lambda}^T \boldsymbol{z}_i\right) \tag{2.66}$$

i spełniają one odpowiednie równanie kalibracyjne, przy czym $F(\cdot)$ jest funkcją o określonej postaci matematycznej.

Funkcja $F(\cdot)$ odgrywa taką samą rolę jak funkcja $D(\mathbf{v}, \mathbf{d})$ minimalizująca odległość między wektorem wag wynikających ze schematu losowania próby (stąd określenie „podejście funkcyjne"), a poszukiwanym wektorem wag kalibracyjnych. W zależności od wyboru postaci funkcji $F(\cdot)$ wagi kalibracyjne będą różnej postaci, a ich wyznaczenie wymagać będzie zastosowania metod numerycznych. Na potrzeby pracy rozważać będziemy funkcję liniową $F(u) = 1 + u$. Alternatywnie funkcję $F(\cdot)$ można byłoby, na przykład, określić jako $F(u) = exp(u)$.

Estymator kalibracyjny wartości globalnej – wykorzystujący podejście funkcyjne, można zdefiniować w następujący sposób.

Definicja 8. *Estymatorem kalibracyjnym wartości globalnej (2.1) skonstruowanym w oparciu o podejście funkcyjne jest:*

$$\hat{Y}_{Fcal} = \sum_{i=1}^{m} w_i y_i, \tag{2.67}$$

gdzie wektor wag $\mathbf{w} = (w_1, \ldots, w_m)^T$ postaci (2.66) spełnia równanie kalibracyjne

$$\mathbf{X_F} = \tilde{\mathbf{X}}_\mathbf{F}, \tag{2.68}$$

a wektory $\mathbf{X_F}$ i $\tilde{\mathbf{X}}_\mathbf{F}$ określone są w następujący sposób:

$$\mathbf{X_F} = \begin{pmatrix} \mathbf{X_{F_1}} \\ \mathbf{X_{F_2}} \end{pmatrix}, \quad \tilde{\mathbf{X}}_\mathbf{F} = \begin{pmatrix} \tilde{\mathbf{X}}_{\mathbf{F_1}} \\ \tilde{\mathbf{X}}_{\mathbf{F_2}} \end{pmatrix}, \tag{2.69}$$

przy czym

$$\mathbf{X_{F_1}} = \left(\sum_{i=1}^{N} x_{i1}, \sum_{i=1}^{N} x_{i2}, \ldots, \sum_{i=1}^{N} x_{ik_1} \right)^T, \tag{2.70}$$

$$\mathbf{X_{F_2}} = \left(\sum_{i=1}^{n} d_{pi} x_{i1}, \sum_{i=1}^{n} d_{pi} x_{i2}, \ldots, \sum_{i=1}^{n} d_{pi} x_{ik_2} \right)^T, \tag{2.71}$$

$$\tilde{\mathbf{X}}_{\mathbf{F_1}} = \left(\sum_{i=1}^{m} w_i x_{i1}, \sum_{i=1}^{m} w_i x_{i2}, \ldots, \sum_{i=1}^{m} w_i x_{ik_1} \right)^T, \tag{2.72}$$

$$\tilde{\mathbf{X}}_{\mathbf{F_2}} = \left(\sum_{i=1}^{m} w_i x_{i1}, \sum_{i=1}^{m} w_i x_{i2}, \ldots, \sum_{i=1}^{m} w_i x_{ik_2} \right)^T. \tag{2.73}$$

Zwróćmy uwagę, że rozważane w poprzednich podrozdziałach wagi kalibracyjne są szczególnym przypadkiem wag określonych w definicji (7). Aby się o tym przekonać, wystarczy podstawić $d_{pi} = d_i$ oraz $\boldsymbol{z}_i = \boldsymbol{x}_i$. Przy tym, jeżeli przyjmiemy, że

$k_1 = k$ oraz $k_2 = 0$, to otrzymamy postać wag odpowiadającą estymatorowi kalibracyjnemu wartości globalnej ze znanym wektorem $\mathbf{X}$ wartości globalnych zmiennych pomocniczych. Gdy $k_1 = 0$ oraz $k_2 = k$, to otrzymamy postać wag odpowiadającą estymatorowi kalibracyjnemu wartości globalnej ze znanym wektorem $\check{\mathbf{X}}$ oszacowanych wartości globalnych zmiennych pomocniczych. W innych przypadkach tj. gdy znany jest wektor wartości globalnych k_1 zmiennych pomocniczych określonych na poziomie całej populacji i wektor oszacowanych wartości globalnych k_2 zmiennych pomocniczych wyznaczonych w oparciu o dane pochodzące z próby s, $0 < k_1, k_2 \leqslant k$, to otrzymamy wagi odpowiadające uogólnionemu estymatorowi kalibracyjnemu wartości globalnej.

Poniższe twierdzenie pokazuje, jaka jest postać wag kalibracyjnych przy liniowej funkcji $F(u) = 1 + u$.

Twierdzenie 5. *Jeżeli macierz* $\sum_{i=1}^{m} d_{pi}\boldsymbol{z}_i\boldsymbol{x}_i{}^T$ *jest nieosobliwa, to wektor wag kalibracyjnych* $\mathbf{w} = (w_1, \ldots, w_m)^T$*, którego składowe są postaci (2.66) spełnia równania kalibracyjne (2.68) wtedy i tylko wtedy, gdy wektor* $\boldsymbol{\lambda}^T$ *jest postaci:*

$$\boldsymbol{\lambda}^T = \left(\mathbf{X_F} - \hat{\mathbf{X}}_\mathbf{F}\right)^T \left(\sum_{i=1}^{m} d_{pi}\boldsymbol{z}_i\boldsymbol{x}_i^T\right)^{-1}, \tag{2.74}$$

gdzie

$$\hat{\mathbf{X}}_\mathbf{F} = \begin{pmatrix} \hat{\mathbf{X}}_{\mathbf{F_1}} \\ \hat{\mathbf{X}}_{\mathbf{F_2}} \end{pmatrix}, \tag{2.75}$$

a

$$\hat{\mathbf{X}}_{\mathbf{F_1}} = \left(\sum_{i=1}^{m} d_{pi}x_{i1}, \sum_{i=1}^{m} d_{pi}x_{i2}, \ldots, \sum_{i=1}^{m} d_{pi}x_{ik_1}\right)^T, \tag{2.76}$$

$$\hat{\mathbf{X}}_{\mathbf{F_2}} = \left(\sum_{i=1}^{m} d_{pi}x_{i1}, \sum_{i=1}^{m} d_{pi}x_{i2}, \ldots, \sum_{i=1}^{m} d_{pi}x_{ik_2}\right)^T. \tag{2.77}$$

Dowód. Załóżmy w pierwszej kolejności, że wektor wag kalibracyjnych $\mathbf{w}$, którego składowe są postaci (2.66) spełnia równania kalibracyjne (2.68). Mamy więc:

$$w_i = d_{pi}F\left(\boldsymbol{\lambda}^T\boldsymbol{z}_i\right), \tag{2.78}$$

skąd

$$w_i = d_{pi}\left(1 + \boldsymbol{\lambda}^T\boldsymbol{z}_i\right). \tag{2.79}$$

Z powyższego równania wynika, że

$$w_i - d_{pi} = d_{pi}\boldsymbol{\lambda}^T\boldsymbol{z}_i. \tag{2.80}$$

Mnożąc ostatnie równanie obustronnie przez $\boldsymbol{x}_i^T$, a następnie dokonując sumowania po zbiorze respondentów r otrzymujemy, że:

$$\sum_{i=1}^{m} w_i\boldsymbol{x}_i^T - \sum_{i=1}^{m} d_{pi}\boldsymbol{x}_i^T = \boldsymbol{\lambda}^T \sum_{i=1}^{m} \left(d_{pi}\boldsymbol{z}_i\boldsymbol{x}_i^T\right). \tag{2.81}$$

Stąd otrzymujemy, że:

$$\tilde{\mathbf{X}}_\mathbf{F}^T - \hat{\mathbf{X}}_\mathbf{F}^T = \boldsymbol{\lambda}^T \sum_{i=1}^{m} \left(d_{pi}\boldsymbol{z}_i\boldsymbol{x}_i^T\right). \tag{2.82}$$

Korzystając z faktu, że macierz $\sum_{i=1}^{m} d_{pi}\boldsymbol{z}_i\boldsymbol{x}_i{}^T$ jest nieosobliwa oraz, że spełnione jest równanie kalibracyjne (2.68) otrzymujemy, że

$$\boldsymbol{\lambda}^T = \left(\mathbf{X_F} - \hat{\mathbf{X}}_\mathbf{F}\right)^T \left(\sum_{i=1}^{m} d_{pi}\boldsymbol{z}_i\boldsymbol{x}_i^T\right)^{-1}. \tag{2.83}$$

Na odwrót, załóżmy, że składowe wektora $\mathbf{w}$ określone są wzorem (2.66) oraz, że wektor $\boldsymbol{\lambda}^T$ jest postaci (2.74). Mamy wtedy

$$w_i = d_{pi} + d_{pi}\left(\mathbf{X_F} - \hat{\mathbf{X}}_\mathbf{F}\right)^T \left(\sum_{i=1}^{m} d_{pi}\boldsymbol{z}_i\boldsymbol{x}_i^T\right)^{-1}\boldsymbol{z}_i, \tag{2.84}$$

z czego wynika, że

$$w_i - d_{pi} = d_{pi}\left(\mathbf{X_F} - \hat{\mathbf{X}}_\mathbf{F}\right)^T \left(\sum_{i=1}^{m} d_{pi}\boldsymbol{z}_i\boldsymbol{x}_i^T\right)^{-1}\boldsymbol{z}_i. \tag{2.85}$$

Mnożąc ostatnie równanie obustronnie przez $\boldsymbol{x}_i^T$, a następnie dokonując sumowania po zbiorze respondentów r otrzymujemy, że:

$$\sum_{i=1}^{m} w_i\boldsymbol{x}_i^T - \sum_{i=1}^{m} d_{pi}\boldsymbol{x}_i^T = \left(\mathbf{X_F} - \hat{\mathbf{X}}_\mathbf{F}\right)^T \left(\sum_{i=1}^{m} d_{pi}\boldsymbol{z}_i\boldsymbol{x}_i^T\right)^{-1} \left(\sum_{i=1}^{m} d_{pi}\boldsymbol{z}_i\boldsymbol{x}_i^T\right). \tag{2.86}$$

Stąd mamy

$$\tilde{\mathbf{X}}_\mathbf{F}^\mathbf{T} - \hat{\mathbf{X}}_\mathbf{F}^\mathbf{T} = \left(\mathbf{X_F} - \hat{\mathbf{X}}_\mathbf{F}\right)^T, \tag{2.87}$$

z czego wynika, że $\mathbf{X_F} = \tilde{\mathbf{X}}_\mathbf{F}$. ■

Estymatory kalibracyjne, wyznaczone w oparciu o podejście funkcyjne, stanowią szeroką klasę estymatorów wartości globalnej zmiennej y. Wynika to z faktu, że istnieje duża dowolność przy wyborze wektora wag początkowych $\mathbf{d}_{pi}$ i wektora zmiennych instrumentalnych $\boldsymbol{z}_i$. W konsewencji, istnieje wiele możliwych sposobów określenia wag kalibracyjnych w_i odpowiadających estymatorowi kalibracyjnemu $\hat{Y}_{Fcal}$. Poniżej rozpatrzymy niektóre z możliwych do uzyskania estymatorów kalibracyjnych wartości globalnej zmiennej y wskazując postać wag początkowych d_{pi} i wektora $\boldsymbol{z}_i$.

Na początku przyjmijmy, że $k_1 = 1, k_2 = 0$. W tej sytuacji zakładamy, że dysponujemy jedną zmienną pomocniczą, dla której znana jest jej wartość globalna. W najprostszym przypadku załóżmy, że ta zmienna przyjmuje wartości równe 1 dla każdej jednostki $i \in U$, a więc również dla każdej jednostki $i \in s, i \in r$. W tym przypadku wektor $\mathbf{X}_F = N$, a wektor $\hat{\mathbf{X}}_\mathbf{F} = \sum_{i=1}^{m} d_i$ redukuje się do sumy wartości wag wynikających ze schematu losowania próby obliczonej dla wszystkich respondentów, dla których znana jest wartość zmiennej y. Przyjmujemy więc, że $\boldsymbol{z}_i = \boldsymbol{x}_i = 1$ oraz $d_{pi} = d_i$. W tym przypadku, jak łatwo pokazać, poszczególne składowe wektora wag kalibracyjnych redukują się do:

$$w_i = d_i + \frac{d_i\left(N - \sum_{i=1}^{m} d_i\right)}{\sum_{i=1}^{m} d_i} = \frac{d_i N}{\sum_{i=1}^{m} d_i}, \tag{2.88}$$

a estymator kalibracyjny wartości globalnej jest postaci:

$$\hat{Y}_{Fcal} = N\bar{y}_r = \hat{Y}_{EXP}, \tag{2.89}$$

gdzie:

$$\bar{y}_r = \frac{\sum_{i=1}^{m} d_i y_i}{\sum_{i=1}^{m} d_i}. \tag{2.90}$$

Tak określony wektor zmiennych pomocniczych, który nie różnicuje jednostek, jest jednak nieefektywny w badaniach z brakami odpowiedzi (indeks EXP odnosi się do często wykorzystywanego w metodzie reprezentacyjnej oznaczenia estymatora bezpośredniego (ekspansyjnego)). Obciążenie tego estymatora, jak i jego wariancja, są zazwyczaj bardzo duże. W badaniach statystycznych jest on jednak często wyznaczany, gdyż stanowi punkt odniesienia dla innych estymatorów, które są z nim porównywane. Jest on również stosowany w tych przypadkach, w których badacz ma przekonanie, że braki odpowiedzi mają charakter losowy. Z powyższych rozważań nasuwa się więc wniosek, że przy wyborze zmiennych do wektora zmiennych pomocniczych warto, aby obok zmiennej sztucznej gwarantującej nam sumowalność wag do liczebności populacji, umieścić inne zmienne – skorelowane ze zmienną Y.

Rozważmy obecnie sytuację, w której obok sztucznej zmiennej pomocniczej tożsamościowo równej 1, występują inne zmienne w wektorze zmiennych pomocniczych. Załóżmy, że $k_1 = k, k_2 = 0$, przy czym $k > 0$ i przyjmijmy, że ostatnia ze zmiennych pomocniczych przyjmuje wartość 1 dla wszystkich jednostek. Niech ponadto $d_{pi} = d_i$ oraz $\boldsymbol{z}_i = \boldsymbol{x}_i = (x_{i1}, \ldots, x_{ik-1}, 1)^T$. Wektor wartości globalnych $\mathbf{X}_F$ jest postaci

$$\mathbf{X_F} = \left(\sum_{i=1}^{N} x_{i1}, \ldots, \sum_{i=1}^{N} x_{ik-1}, N \right)^T. \tag{2.91}$$

Dla tak określonych wag początkowych i wektora zmiennych instrumentalnych $\boldsymbol{z}_i$, system wag posiada ważną własność wyrażającą się tym, że suma wag kalibracyjnych równa się liczebności populacji. Chcąc więc sobie zapewnić sumowalność wag do liczebności populacji N wystarczy przyjąć za jedną ze zmiennych pomocniczych 1.

Przyjmijmy teraz, że dysponujemy jedną zmienną pomocniczą, dla której znana jest jej wartość globalna, tzn. $k_1 = 1, k_2 = 0$. W przeciwieństwie jednak do rozważanego poprzednio estymatora bezpośredniego zakładamy, że nie jest ona równa tożsamościowo 1 dla wszystkich jednostek. W tej sytuacji $\boldsymbol{x}_i = x_i$. Niech dalej $d_{pi} = d_i$ oraz $\boldsymbol{z}_i = 1$ dla wszystkich jednostek i. Jest to przypadek, w którym wektor zmiennych pomocniczych nie jest równy wektorowi zmiennych instrumentalnych. W tej sytuacji mamy, że $\mathbf{X}_F = \sum_{i=1}^{N} x_{i1}$ oraz $\hat{\mathbf{X}}_F = \sum_{i=1}^{m} d_i x_{i1}$. W tym przypadku wektor $\boldsymbol{\lambda}^T$ przyjmuje postać

$$\boldsymbol{\lambda}^T = \frac{\sum_{i=1}^{N} x_{i1} - \sum_{i=1}^{m} d_i x_{i1}}{\sum_{i=1}^{m} d_i x_{i1}} = \frac{\sum_{i=1}^{N} x_{i1}}{\sum_{i=1}^{m} d_i x_{i1}} - 1. \tag{2.92}$$

Wagi kalibracyjne są więc postaci

$$w_i = d_i \frac{\sum_{i=1}^{N} x_{i1}}{\sum_{i=1}^{m} d_i x_{i1}}, \tag{2.93}$$

a estymator kalibracyjny wartości globalnej przyjmuje postać

$$\hat{Y}_{Fcal} = \sum_{i=1}^{m} w_i y_i = \sum_{i=1}^{N} x_{i1} \frac{\sum_{i=1}^{m} d_i y_i}{\sum_{i=1}^{m} d_i x_{i1}} = \hat{Y}_{RA}. \tag{2.94}$$

Jest to estymator ilorazowy wartości globalnej (subskrypt RA odnosi się do angielskiej nazwy tego estymatora – *ratio estimator*).

Gdyby przyjąć w powyższym rozumowaniu, że $k_1 = 0, k_2 = 1$, to wtedy w miejsce wartości globalnej zmiennej pomocniczej, należałoby podstawić jej oszacowanie na podstawie próby s. Estymator ilorazowy przyjąłby postać

$$\hat{Y}_{RA} = \sum_{i=1}^{n} d_i x_{i1} \frac{\sum_{i=1}^{m} d_i y_i}{\sum_{i=1}^{m} d_i x_{i1}}. \tag{2.95}$$

W ostatnim rozważanym przypadku załóżmy, że oprócz zmiennej x_i do wektora zmiennych pomocniczych włączymy zmienną tożsamościowo równą 1. Zakładamy więc, że $k_1 = 2, k_2 = 0$. Niech ponadto $d_{pi} = d_i$ oraz $\boldsymbol{x}_i = \boldsymbol{z}_i = (x_{i1}, 1)^T$. W tej sytuacji mamy, że

$$\mathbf{X}_F = \left(\sum_{i=1}^{N} x_{i1}, N \right)^T \text{ oraz } \quad \hat{\mathbf{X}}_F = \left(\sum_{i=1}^{m} d_i x_{i1}, \sum_{i=1}^{m} d_i \right)^T. \tag{2.96}$$

Estymator kalibracyjny jest więc postaci, por. C-E. Särndal, S. Lundström (2005):

$$\hat{Y}_{Fcal} = N \left[\bar{y}_r + (\bar{x} - \bar{x}_r) \beta_r \right] = \hat{Y}_{REG}, \tag{2.97}$$

gdzie

$$\beta_r = \frac{\sum_{i=1}^{m} d_i (x_{i1} - \bar{x}_r)(y_i - \bar{y}_r)}{\sum_{i=1}^{m} d_i (x_{i1} - \bar{x}_r)^2}, \tag{2.98}$$

przy czym

$$\bar{x} = \frac{\sum_{i=1}^{N} x_{i1}}{N}, \quad \bar{x}_r = \frac{\sum_{i=1}^{m} d_i x_{i1}}{\sum_{i=1}^{m} d_i}, \quad \bar{y}_r = \frac{\sum_{i=1}^{m} d_i y_i}{\sum_{i=1}^{m} d_i}. \tag{2.99}$$

Tabela. 2.1. Estymatory kalibracyjne wartości globalnej zmiennej y

$\boldsymbol{x}_i$	$\boldsymbol{z}_i$	k_1	k_2	$\mathbf{X}_F$	**Estymator kalibracyjny**
1	$\boldsymbol{x}_i$	1	0	N	$\hat{Y}_{EXP} = N\bar{y}_r$
x_{i1}	1	1	0	$\sum_{i=1}^{m} x_{i1}$	$\hat{Y}_{RA} = \sum_{i=1}^{N} x_{i1} \frac{\sum_{i=1}^{m} d_i y_i}{\sum_{i=1}^{m} d_i x_{i1}}$
$(x_{i1}, 1)^T$	$\boldsymbol{x}_i$	2	0	$\left(\sum_{i=1}^{N} x_{i1}, N\right)^T$	$\hat{Y}_{REG} = N\left(\bar{y}_r + (\bar{x} - \bar{x}_r) \frac{\sum_{i=1}^{m} d_i (x_{i1} - \bar{x}_r)(y_i - \bar{y}_r)}{\sum_{i=1}^{m} d_i (x_{i1} - \bar{x}_r)^2}\right)$
$(x_{i1}, \ldots, x_{ik})^T$	$\boldsymbol{x}_i$	k	0	$\mathbf{X}$	$\hat{Y}_{\mathbf{X}} = \sum_{i=1}^{m} y_i \left\{ d_i + d_i \left(\mathbf{X} - \hat{\mathbf{X}}\right)^T \left(\sum_{i=1}^{m} d_i \boldsymbol{x}_i \boldsymbol{x}_i^T\right)^{-1} \boldsymbol{x}_i \right\}$
$(x_{i1}, \ldots, x_{ik})^T$	$\boldsymbol{x}_i$	0	k	$\breve{\mathbf{X}}$	$\hat{Y}_{\breve{\mathbf{X}}} = \sum_{i=1}^{m} y_i \left\{ d_i + d_i \left(\breve{\mathbf{X}} - \hat{\mathbf{X}}\right)^T \left(\sum_{i=1}^{m} d_i \boldsymbol{x}_i \boldsymbol{x}_i^T\right)^{-1} \boldsymbol{x}_i \right\}$
$(\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2})^T$	$\boldsymbol{x}_i$	k_1	k_2	$\begin{pmatrix} \mathbf{X_1} \\ \breve{\mathbf{X}}_\mathbf{2} \end{pmatrix}$	$\hat{Y}_{cal} = \sum_{i=1}^{m} y_i \left\{ d_i + d_i \left(\begin{pmatrix} \mathbf{X_1} \\ \breve{\mathbf{X}}_\mathbf{2} \end{pmatrix} - \begin{pmatrix} \hat{\mathbf{X}}_\mathbf{1} \\ \hat{\mathbf{X}}_\mathbf{2} \end{pmatrix} \right)^T \left(\sum_{i=1}^{m} d_i \boldsymbol{x}_i \boldsymbol{x}_i^T \right)^{-1} \right\} \boldsymbol{x}_i$

Źródło: Opracowanie własne

Jest to estymator kalibracyjny regresyjny wartości globalnej (subskrypt REG odnosi się do angielskiej nazwy tego estymatora – *regression estimator*).

Tabela (2.1) przedstawia postać rozważanych do tej pory estymatorów kalibracyjnych wartości globalnej zmiennej y. Zawarto w niej ponadto informację o wektorze zmiennych pomocniczych $\boldsymbol{x}_i$, liczbie zmiennych pomocniczych, dla których znane są wartości globalne, liczbie zmiennych pomocniczych, dla których znane są oszacowania ich wartości globalnych w oparciu o dane pochodzące z próby s oraz wektorze $\boldsymbol{z}_i$. We wszystkich rozważanych estymatorach przyjęto przy tym założenie, że wektor wag początkowych spełnia zależność $d_{pi} = d_i$. Oznacza to, że procesowi kalibracji poddawane będą wagi wynikające ze schematu losowania próby.

2.6. Wnioski

W rozdziale drugim przedstawione zostały – w kompleksowy sposób – metody konstrukcji estymatorów kalibracyjnych wartości globalnej. Udowodniliśmy twierdzenia o postaci wag kalibracyjnych rozważanych estymatorów, a także pewne ich własności. Wyróżniliśmy przy tym wszystkie możliwe poziomy na jakich dostępna może być informacja o zmiennych pomocniczych, co ma wpływ na postać wag kalibracyjnych oraz na najważniejsze własności samych estymatorów.

Przedstawiono ponadto, opierając się na najnowszej literaturze przedmiotu, ideę podejścia funkcyjnego, która umożliwia konstrukcję szerokiej klasy estymatorów kalibracyjnych spełniających odpowiednie równania kalibracyjne bez konieczności minimalizowania odległości między wagami kalibracyjnymi, a tymi, które wynikają ze schematu losowania próby.

Zastosowanie w badaniach praktycznych – omawianych estymatorów – jest w dużej mierze uzależnione jednak od dostępności informacji o zmiennych pomocniczych. Innym bardzo ważnym kryterium wyboru odpowiedniego estymatora, które należy wziąć pod uwagę, jest jego obciążenie i wariancja.

Ponieważ na drodze analitycznej niemożliwe jest w zasadzie ustalenie, który z rozważanych estymatorów kalibracyjnych charakteryzuje się mniejszym obciążeniem i wariancją, w rozdziale czwartym – mającym charakter poznawczy – zaproponowane i przeprowadzone zostało badanie symulacyjne na rzeczywistych danych pochodzących z NSP'2002 celem szczegółowego omówienia wybranych charakterystyk, rozważanych w rozdziale drugim estymatorów. Wnioski płynące z wyników badań symulacyjnych, pozwolą ocenić, w przybliżony sposób, najważniejsze własności estymatorów, a w konsekwencji ułatwić wskazanie najodpowiedniejszego estymatora w praktycznych zastosowaniach.

Mając na uwadze dostępność informacji o zmiennych pomocniczych oraz wnioski płynące z rozważań zawartych w rozdziale czwartym, w ostatnim rozdziale pracy, podjęta zostanie próba empirycznej oceny rozpatrywanych – w rozdziale drugim – estymatorów kalibracyjnych w badaniu budżetów gospodarstw domowych.

Rozdział 3

Estymatory kalibracyjne kwantyla rzędu α

3.1. Podstawowe definicje i oznaczenia

W podejściu kalibracyjnym ze względu na fakt, że w wielu badaniach oprócz braków odpowiedzi, rozkłady analizowanych cech charakteryzują się silnymi asymetriami — dużą rolę zaczęto przywiązywać do problemu estymacji kwantyli, które są często bardziej pożądanymi miarami położenia aniżeli średnia. Dotyczy to zwłaszcza badań społeczno–ekonomicznych, w których kwantyle są bardzo często wykorzystywane.

Przykładowo, w badaniu budżetów gospodarstw domowych, podaje się oprócz informacji o średnich dochodach gospodarstw domowych i ich wydatkach na różnego rodzaju usługi i artykuły – jako podstawowych miarach określających zamożność, poziom życia oraz strukturę konsumpcji – również informacje o grupach kwintylowych dochodu rozporządzalnego na osobę w gospodarstwie oraz medianie spożycia niektórych artykułów żywnościowych.

Z kolei w europejskim badaniu dochodów i warunków życia (EU-SILC) wyznacza się grupy decylowe gospodarstw, celem określenia stopnia zróżnicowania społeczeństwa pod względem uzyskiwanych dochodów oraz podaje się różnego rodzaju wskaźniki oparte na decylach – wskaźnik zróżnicowania decylowego, wskaźnik wahania decylowego, wskaźnik dyspersji decylowej, por. GUS (2007). Informacje o kwantylach mogą być zatem użyteczne w określeniu stopnia zróżnicowania gospodarstw pod względem wielkości uzyskiwanych dochodów, spożycia artykułów żywnościowych czy ponoszonych wydatkach na różnego rodzaju dobra i usługi.

W związku z powyższym, w literaturze z zakresu metody reprezentacyjnej, estymacji kwantyli przypisuje się bardzo duże znaczenie. Szczególną rolę odgrywają przy tym te podejścia, które oparte są na wykorzystaniu zmiennych pomocniczych z różnych źródeł celem poprawy jakości uzyskanych szacunków. Należy przy tym podkreślić, że proces estymacji kwantyli związany jest bezpośrednio z poprzedzającym go etapem poszukiwania odpowiedniego estymatora dystrybuanty.

Próby konstrukcji estymatorów kwantyli, z użyciem zmiennych pomocniczych, znane były na długo przed tym, jak kalibracja stała się powszechnie wykorzystywaną metodą w badaniach częściowych. Jedną z pierwszych wzmianek na temat wykorzystania informacji dodatkowych w szacowaniu kwantyli można znaleźć w pracy Chambersa i Dunstana (1986). W swoim artykule autorzy, opierając się na podejściu wspomaganym modelem, rozważali metodę konstrukcji estymatorów kwantyli rzędu α bazując na estymatorze funkcji rozkładu z uwzględnieniem zmiennych pomocniczych. Metoda ta, jak wykazali Chambers, Dorfman i Hall (1992) może jednak dawać gorsze oszacowania w porównaniu z klasycznymi estymatorami – w przypadku źle dobranego modelu opisującego rozkład cechy w populacji. Do podobnych wniosków, na podstawie zastosowanego podejścia symulacyjnego, doszli Rao, Kovar i Mantel (1990). Jak wykazują autorzy, w przypadku niewłaściwego modelu opisującego rozkład badanej cechy w populacji, podejście modelowe może dawać gorsze wyniki w porównaniu z zastosowaniem klasycznych estymatorów kwantyli, których konstrukcja oparta jest tylko na informacjach pochodzących z próby.

W ciągu ostatnich lat coraz większą uwagę zaczęto przywiązywać do podejścia kalibracyjnego w odniesieniu do kwantyli. Pierwszym, który zwrócił uwagę na możliwość zastosowania podejścia kalibracyjnego, był Kovačević (1997). W zaproponowanej przez siebie metodzie wykorzystał kalibrację do konstrukcji estymatora dystrybuanty opierając się na odpowiednich momentach zmiennych pomocniczych. Podejście to znalazło następnie zastosowanie w badaniach panelowych europejskich gospodarstw domowych.

Z ostatnich osiągnięć z zakresu estymatorów kalibracyjnych kwantyli na uwagę zasługują prace Ren'a (2002), który — jak podają Harms i Duchesne (2006) — był pierwszym badaczem, który w kompleksowy sposób zastosował paradygmat kalibracji sformułowany przez Deville'a i Särndala dla estymatora kalibracyjnego dystrybuanty i kwantyla rzędu α. Dalsze prace z zakresu kalibracji nawiązywały do artykułu Ren'a. Na szczególną uwagę zasługuje przy tym metodologia wyznaczania estymatorów kalibracyjnych kwantyli zaproponowana przez Harmsa i Duchesne (2006), a oparta na koncepcji tzw. dystrybuanty interpolacyjnej.

Należy jednak podkreślić, że w żadnej z dotychczas cytowanych prac nie zwracano uwagi na wpływ braków odpowiedzi na proces estymacji kwantyla rzędu α. Okazuje się, że zaproponowaną przez Ren'a, Harmsa i Duchesne metodologię można przenieść na przypadek, gdy w badaniu nie dysponujemy kompletnym zbiorem informacji.

Rozdział ten poświęcony jest zatem konstrukcji estymatorów kalibracyjnych w sytuacji, gdy w badaniu występują braki danych. Proponowane podejście rozszerzymy dodatkowo o przypadek, gdy nie są znane — w przeciwieństwie do tego co zakładali Harms i Duchesne — kwantyle rzędu α dla wszystkich zmiennych pomocniczych na poziomie populacji U, a tylko ich oszacowania. Co więcej, uwzględniony zostanie przypadek estymatora kalibracyjnego kwantyla rzędu α, przy założeniu, że dla zmiennych pomocniczych znane są kwantyle dowolnego rzędu bądź ich oszacowania z próby.

Załóżmy — podobnie jak w odniesieniu do estymatora kalibracyjnego wartości globalnej, że dana jest N – elementowa populacja $U = \{1, \ldots, N\}$, z której losujemy zgodnie z określonym schematem losowania n – elementową próbę $s \subseteq U$. Niech π_i oznacza prawdopodobieństwo inkluzji i – tej jednostki do próby, natomiast $d_i = \frac{1}{\pi_i}$ wagę odpowiadającą jednostce i.

Zakładamy, że celem badania jest oszacowanie kwantyla $Q_{y,\alpha}$ rzędu α zmiennej y.

W szczególnym przypadku, gdy $\alpha = 0.5$ dokonujemy oszacowania mediany zmiennej y. Załóżmy w dalszym ciągu, że r oznacza m – elementowy zbiór respondentów, dla których znana jest wartość zmiennej y. W konsekwencji $s \backslash r$ oznacza zbiór nierespondentów tj. $n-m$ jednostek w próbie, które z różnych powodów nie udzieliły odpowiedzi i nie jest znana dla nich wartość zmiennej y.

Obecnie zdefiniujemy kilka podstawowych pojęć niezbędnych w dalszych rozważaniach.

Definicja 9. *Dystrybuantą zmiennej y nazywamy funkcję postaci:*

$$F_y(t) = \frac{\sum_{i=1}^{N} H(t - y_i)}{N}, \tag{3.1}$$

gdzie:

$$H(t - y_i) = \begin{cases} 1, & t \geqslant y_i, \\ 0, & t < y_i, \end{cases} \tag{3.2}$$

dla $i = 1, \ldots, N$ i $t \in \mathbb{R}$.

Definicja 10. *$Q_{y,\alpha}$ nazywamy kwantylem rzędu α zmiennej y jeżeli:*

$$Q_{y,\alpha} = \inf \{ t \,|F_y(t) \geqslant \alpha \}, \tag{3.3}$$

gdzie $\alpha \in (0, 1)$, $t \in \mathbb{R}$.

Wyznaczenie z powyższej definicji kwantyla rzędu α jest w badaniach prowadzonych metodą reprezentacyjną niemożliwe, ze względu na nieznajomość funkcji rozkładu zmiennej y. Dlatego niezbędne jest znalezienie odpowiedniego estymatora dystrybuanty, w oparciu o który będzie można wyznaczyć interesujący nas kwantyl.

Definicja 11. *Dystrybuantą interpolacyjną zmiennej y nazywamy funkcję postaci:*

$$\hat{F}_{y,cal}(t) = \frac{\sum_{i=1}^{m} w_i H_{y,r}(t, y_i)}{\sum_{i=1}^{m} w_i}, \tag{3.4}$$

gdzie funkcja $H_{y,r}(t, y_i)$ jest określona jako:

$$H_{y,r}(t, y_i) = \begin{cases} 1, & y_i \leqslant L_{y,r}(t), \\ \beta_{y,r}(t), & y_i = U_{y,r}(t), \\ 0, & y_i > U_{y,r}(t), \end{cases} \tag{3.5}$$

przy czym:

$$L_{y,r}(t) = max\{\{y_i, i \in r \,|y_i \leqslant t\} \cup \{-\infty\}\}, \tag{3.6}$$

$$U_{y,r}(t) = min\{\{y_i, i \in r \,|y_i > t\} \cup \{\infty\}\}, \tag{3.7}$$

$$\beta_{y,r}(t) = \frac{t - L_{y,r}(t)}{U_{y,r}(t) - L_{y,r}(t)}, \tag{3.8}$$

dla $i = 1, \ldots, m$, $t \in \mathbb{R}$.

W podobny sposób definiujemy dystrybuantę interpolacyjną poszczególnych zmiennych pomocniczych.

Definicja 12. *Dystrybuantą interpolacyjną zmiennej pomocniczej x_j nazywamy funkcję postaci:*

$$\hat{F}_{x_j,cal}(t) = \frac{\sum_{i=1}^{n} w_i H_{x_j,s}(t, x_{ij})}{\sum_{i=1}^{n} w_i}, \tag{3.9}$$

gdzie funkcja $H_{x_j,s}(t, x_{ij})$ jest określona jako:

$$H_{x_j,s}(t, x_{ij}) = \begin{cases} 1, & x_{ij} \leqslant L_{x_j,s}(t), \\ \beta_{x_j,s}(t), & x_{ij} = U_{x_j,s}(t), \\ 0, & x_{ij} > U_{x_j,s}(t), \end{cases} \tag{3.10}$$

przy czym:

$$L_{x_j,s}(t) = max\{\{x_{ij}, i \in s \,|x_{ij} \leqslant t\} \cup \{-\infty\}\}, \tag{3.11}$$

$$U_{x_j,s}(t) = min\{\{x_{ij}, i \in s \,|x_{ij} > t\} \cup \{\infty\}\}, \tag{3.12}$$

$$\beta_{x_j,s}(t) = \frac{t - L_{x_j,s}(t)}{U_{x_j,s}(t) - L_{x_j,s}(t)}, \tag{3.13}$$

dla $i = 1, \ldots, n$, $j = 1, \ldots, k$, i $t \in \mathbb{R}$.

Różnica między dystrybuantą interpolacyjną zmiennej y, a zmienną x_j odnosi się w zasadzie do zbioru, na którym dokonujemy sumowania odpowiednich wartości. Dla zmiennej y, sumowanie ze względu na braki odpowiedzi, odbywa się po zbiorze respondentów r. W przypadku zmiennych pomocniczych, dla których zakładamy, że mamy informację na poziomie całej próby, sumowanie odbywa się po wszystkich jednostkach do niej należących.

Występujące w definicji dystrybuanty interpolacyjnej funkcje, mają bardzo intuicyjne interpretacje. $L_{y,r}(t)$ oznacza najbliższego sąsiada punktu t „z dołu" w uporządkowanym niemalejąco zbiorze wartości zmiennej y. W przypadku gdy $t = y_i$ dla pewnego $i \in r$ to wtedy $L_{y,r}(t) = L_{y,r}(y_i) = y_i$. $U_{y,r}(t)$ oznacza z kolei najbliższego sąsiada punktu t „z góry", w uporządkowanym niemalejąco zbiorze wartości zmiennej y, który jest jednocześnie większy od t. Funkcja $H_{y,r}(t, y_i)$ jest zmodyfikowaną wersją funkcji H określonej we wzorze (3.2). W szczególnym przypadku, gdy $t \in \{y_i, i \in r\}$ spełniony jest warunek $H(t - y_i) = H_{y,r}(t, y_i)$. Parametr $\beta_{y,r}(t)$ oznacza z kolei współczynnik liniowej interpolacji pomiędzy $L_{y,r}(t)$ i $U_{y,r}(t)$. Oznacza to, że w punktach t, dla których $t \neq y_i$ funkcja $H(t - y_i) \neq H_{y,r}(t, y_i)$.

3.2. Estymator kalibracyjny kwantyla rzędu α ze znanym wektorem $Q_{\boldsymbol{x},\alpha}$

Załóżmy, że znany jest wektor $Q_{\boldsymbol{x},\alpha} = (Q_{x_1,\alpha}, \ldots, Q_{x_k,\alpha})^T$ złożony z kwantyli rzędu α wszystkich zmiennych pomocniczych. Zakładamy, że informacje na temat kwantyli

posiadamy z innych niż badanie częściowe źródeł: spisów lub rejestrów administracyjnych. Podobnie, jak w przypadku estymatora kalibracyjnego wartości globalnej, rozważymy różne sytuacje, w zależności od informacji dodatkowych, jakie będziemy posiadali.

W pierwszym rozważanym podejściu zakładamy, że:

- dla każdego respondenta znana jest wartość każdej zmiennej pomocniczej, tj. znana jest macierz $\mathbf{X_r}$ określona wzorem (2.6),
- znany jest wektor $Q_{\boldsymbol{x},\alpha}$ kwantyli rzędu α wszystkich zmiennych pomocniczych.

W podejściu tym nie wymaga się, aby znane były wartości zmiennych pomocniczych dla wszystkich jednostek wylosowanych do próby s, ale tylko w odniesieniu do respondentów, dla których znane są wartości zmiennej y.

Definicja 13. *Estymatorem kalibracyjnym kwantyla $Q_{y,\alpha}$ rzędu α zmiennej y jest:*

$$\hat{Q}_{y,cal,\alpha} = \hat{F}^{-1}_{y,cal}(\alpha), \tag{3.14}$$

gdzie dystrybuanta interpolacyjna $\hat{F}_{y,cal}(t)$ określona jest wzorem (3.4), a wektor wag kalibracyjnych $\boldsymbol{w} = (w_1, \ldots, w_m)^T$ jest rozwiązaniem zadania minimalizacji

$$\mathbf{w} = \text{argmin}_{\mathbf{v}} D(\mathbf{v}, \mathbf{d}), \tag{3.15}$$

przy warunkach:

$$\sum_{i=1}^{m} w_i = N, \tag{3.16}$$

$$Q_{\boldsymbol{x},\alpha} = \hat{Q}_{\boldsymbol{x},cal,\alpha}, \tag{3.17}$$

gdzie:

$$\hat{Q}_{\boldsymbol{x},cal,\alpha} = \left(\hat{Q}_{x_1,cal,\alpha}, \ldots, \hat{Q}_{x_k,cal,\alpha}\right)^T, \tag{3.18}$$

a funkcja odległości $D(\mathbf{v}, \mathbf{d})$ określona jest wzorem (2.4).

Równania (3.16) i (3.17) są tzw. równaniami kalibracyjnymi. Pierwsze z nich odnosi się do naturalnego warunku, aby suma wag kalibracyjnych równała się liczebności populacji. Drugie z nich jest odpowiednikiem równania (2.12). Zgodnie z nim, wektora wag $\mathbf{w}$ poszukujemy w ten sposób, aby kwantyl każdej zmiennej pomocniczej, określony na poziomie całej populacji, równał się oszacowanemu kwantylowi na podstawie zbioru respondentów. Do estymacji tego kwantyla wykorzystujemy dystrybuantę interpolacyjną (3.9) oraz macierz $\mathbf{X_r}$. Z tego względu w definicji dystrybuanty (3.9) sumowanie odpowiednich wartości, odbywa się na zbiorze wartości wszystkich respondentów tj. w miejsce liczebności próby n należy wstawić liczebność zbioru respondentów m. Wynika to z samej idei kalibracji. Można tutaj bowiem zastosować podobne rozumowanie, jak w przypadku estymatora kalibracyjnego wartości globalnej ze znanym wektorem $\mathbf{X}$ wartości globalnych zmiennych pomocniczych.

Mianowicie, jeżeli będziemy dysponowali odpowiednio dobranym wektorem zmiennych pomocniczych, to wagi kalibracyjne spełniające równanie kalibracyjne (3.17), powinny umożliwić „lepsze" oszacowanie kwantyla rzędu α zmiennej y. Innymi słowy,

przy odpowiednio dobranym wektorze zmiennych pomocniczych — wagi, które będziemy wykorzystywać przy szacowaniu kwantyla rzędu α zmiennej y — powinny „odtwarzać" zachowanie się wag uzyskanych z równania kalibracyjnego (3.17). „Naśladowanie wag kalibracyjnych" oznacza przy tym sytuację, w której ich zastosowanie na poziomie próby do każdej zmiennej pomocniczej $i = 1, \ldots, k$ oddzielnie, umożliwia uzyskanie szacunku kwantyla równego dokładnej wartości kwantyla $Q_{x_i,\alpha}$ tej zmiennej na poziomie całej populacji, a w konsekwencji również przybliżonej wartości kwantyla $Q_{y,\alpha}$ zmiennej y.

Zwróćmy uwagę, że estymator kalibracyjny kwantyla rzędu α moglibyśmy równoważnie zdefiniować w następujący sposób.

Definicja 14. *Estymatorem kalibracyjnym kwantyla $Q_{y,\alpha}$ rzędu α zmiennej y jest:*

$$\hat{Q}_{y,cal,\alpha} = \hat{F}^{-1}_{y,cal}(\alpha), \tag{3.19}$$

gdzie dystrybuanta interpolacyjna $\hat{F}_{y,cal}(t)$ określona jest wzorem (3.4), a wektor wag kalibracyjnych $\boldsymbol{w} = (w_1, \ldots, w_m)^T$ jest rozwiązaniem zadania minimalizacji

$$\mathbf{w} = \operatorname{argmin}_{\mathbf{v}} D(\mathbf{v}, \mathbf{d}), \tag{3.20}$$

przy warunkach:

$$\sum_{i=1}^{m} w_i = N, \tag{3.21}$$

$$\hat{F}_{x_i,cal}(Q_{x_i,\alpha}) = \alpha, \tag{3.22}$$

gdzie $i = 1, \ldots, k$, a funkcja odległości $D(\mathbf{v}, \mathbf{d})$ określona jest wzorem (2.4).

Równoważność obydwu definicji wynika z uwagi poczynionej powyżej, że wagi dobierane są w ten sposób, że ich zastosowanie na poziomie próby do każdej zmiennej pomocniczej $i = 1, \ldots, k$ oddzielnie, umożliwia uzyskanie dokładnej wartości kwantyla $Q_{x_i,\alpha}$ tej zmiennej na poziomie całej populacji. Oznacza to, że wartość dystrybuanty interpolacynej dla i – tej zmiennej pomocniczej w punkcie $\hat{Q}_{x_i,cal,\alpha}$ przy wektorze wag kalibracyjnych $\mathbf{w}$ spełnia równanie $\hat{F}_{x_i,cal}\left(\hat{Q}_{x_i,cal,\alpha}\right) = \alpha$. Definicja (14) jest jednak wygodniejsza do wyprowadzenia wzoru na postać wag kalibracyjnych, o czym mówi poniższe twierdzenie.

Twierdzenie 6. *Rozwiązaniem zadania minimalizacji funkcji odległości (3.20), przy warunku (3.21) i (3.22) jest wektor wag kalibracyjnych $\mathbf{w} = (w_1, \ldots, w_m)^T$, którego składowe spełniają równanie:*

$$w_i = d_i + d_i \mathbf{a}_i^T \left(\sum_{i=1}^{m} d_i \mathbf{a}_i \mathbf{a}_i^T \right)^{-1} \left(\mathbf{T}_{\mathbf{a}} - \sum_{i=1}^{m} d_i \mathbf{a}_i \right), \tag{3.23}$$

gdzie:

$$\mathbf{T}_{\mathbf{a}} = (N, \underbrace{\alpha, \ldots, \alpha}_{k})^T, \tag{3.24}$$

$$\mathbf{a}_i = (1, a_{i1}, \ldots, a_{ik})^T, \tag{3.25}$$

przy czym:

$$a_{ij} = \begin{cases} N^{-1}, & x_{ij} \leqslant L_{x_j,r}(Q_{x_j,\alpha}), \\ N^{-1}\beta_{x_j,r}(Q_{x_j,\alpha}), & x_{ij} = U_{x_j,r}(Q_{x_j,\alpha}), \\ 0, & x_{ij} > U_{x_j,r}(Q_{x_j,\alpha}), \end{cases} \tag{3.26}$$

$i = 1, \ldots, m$, $j = 1, \ldots, k$.

Dowód. Wagi kalibracyjne w_i dla $i = 1, \ldots, m$ znajdziemy, korzystając z metody czynników nieoznaczonych Lagrange'a.
Funkcja Lagrange'a ma postać:

$$L = \frac{1}{2}\sum_{i=1}^{m}\frac{(w_i - d_i)^2}{d_i} + \lambda_0\left(N - \sum_{i=1}^{m} w_i\right) + \sum_{j=1}^{k}\lambda_j\left(\alpha - \hat{F}_{x_j,cal}\left(Q_{x_j,\alpha}\right)\right). \tag{3.27}$$

Z równania kalibracyjnego (3.21) wynika, że:

$$\hat{F}_{x_j,cal}\left(Q_{x_j,\alpha}\right) = \frac{\sum_{i=1}^{m} w_i H_{x_j,r}\left(Q_{x_j,\alpha}, x_{ij}\right)}{\sum_{i=1}^{m} w_i} = \frac{\sum_{i=1}^{m} w_i H_{x_j,r}\left(Q_{x_j,\alpha}, x_{ij}\right)}{N}. \tag{3.28}$$

Funkcję Lagrange'a można więc przedstawić w postaci:

$$L = \frac{1}{2}\sum_{i=1}^{m}\frac{(w_i - d_i)^2}{d_i} + \lambda_0\left(N - \sum_{i=1}^{m} w_i\right) + \sum_{j=1}^{k}\lambda_j\left(\alpha - \frac{\sum_{i=1}^{m} w_i H_{x_j,r}\left(Q_{x_j,\alpha}, x_{ij}\right)}{N}\right). \tag{3.29}$$

Pochodna funkcji L dla $i = 1, \ldots, m$ ma postać:

$$\frac{\partial L}{\partial w_i} = \frac{1}{2}\cdot\frac{2w_i - 2d_i}{d_i} - \lambda_0 - \sum_{j=1}^{k}\lambda_j\frac{H_{x_j,r}\left(Q_{x_j,\alpha}, x_{ij}\right)}{N}. \tag{3.30}$$

Ponieważ:

$$H_{x_j,r}\left(Q_{x_j,\alpha}, x_{ij}\right) = \begin{cases} 1, & x_{ij} \leqslant L_{x_j,r}(Q_{x_j,\alpha}), \\ \beta_{x_j,r}(Q_{x_j,\alpha}), & x_{ij} = U_{x_j,r}(Q_{x_j,\alpha}), \\ 0, & x_{ij} > U_{x_j,r}(Q_{x_j,\alpha}), \end{cases} \tag{3.31}$$

to przyjmując, że:

$$a_{ij} = \frac{H_{x_j,r}\left(Q_{x_j,\alpha}, x_{ij}\right)}{N} \tag{3.32}$$

otrzymujemy:

$$a_{ij} = \begin{cases} N^{-1}, & x_{ij} \leqslant L_{x_j,r}(Q_{x_j,\alpha}), \\ N^{-1}\beta_{x_j,r}(Q_{x_j,\alpha}), & x_{ij} = U_{x_j,r}(Q_{x_j,\alpha}), \\ 0, & x_{ij} > U_{x_j,r}(Q_{x_j,\alpha}). \end{cases} \tag{3.33}$$

Pochodną funkcji L można więc przedstawić w następującej postaci:

$$\frac{\partial L}{\partial w_i} = \frac{w_i - d_i}{d_i} - \lambda_0 - \sum_{j=1}^{k}\lambda_j a_{ij}. \tag{3.34}$$

Przyjmując, że:

$$a_{i0} = 1, \tag{3.35}$$

dla $i = 1, \ldots, m$, otrzymujemy, że:

$$\frac{\partial L}{\partial w_i} = \frac{w_i - d_i}{d_i} - \sum_{j=0}^{k} \lambda_j a_{ij}. \tag{3.36}$$

Przyjmując $\mathbf{a}_i = (1, a_{i1}, \ldots, a_{ik})^T$ i $\boldsymbol{\lambda} = (\lambda_0, \ldots, \lambda_k)^T$ dla $i = 1, \ldots, m$ oraz przyrównując do zera obliczoną pochodną otrzymujemy, że:

$$w_i = d_i \left(1 + \boldsymbol{\lambda}^T \mathbf{a}_i\right) = d_i \left(1 + \mathbf{a}_i^T \boldsymbol{\lambda}\right), \tag{3.37}$$

co wynika z faktu, że:

$$\sum_{j=0}^{k} \lambda_j a_{ij} = \boldsymbol{\lambda}^T \mathbf{a}_i = \mathbf{a}_i^T \boldsymbol{\lambda}. \tag{3.38}$$

Mnożąc obustronnie równanie (3.37) przez $\mathbf{a}_i$, a następnie dokonując sumowania po zbiorze wszystkich respondentów r otrzymujemy:

$$\sum_{i=1}^{m} w_i \mathbf{a}_i = \sum_{i=1}^{m} d_i \mathbf{a}_i \left(1 + \mathbf{a}_i^T \boldsymbol{\lambda}\right). \tag{3.39}$$

Dokonując przekształcenia ostatniego równania otrzymujemy:

$$\sum_{i=1}^{m} w_i \mathbf{a}_i - \sum_{i=1}^{m} d_i \mathbf{a}_i = \sum_{i=1}^{m} d_i \mathbf{a}_i \mathbf{a}_i^T \boldsymbol{\lambda}. \tag{3.40}$$

Zwróćmy uwagę, że:

$$\begin{aligned}
\sum_{i=1}^{m} w_i \mathbf{a}_i &= \left(\sum_{i=1}^{m} w_i, \sum_{i=1}^{m} w_i a_{i1}, \ldots, \sum_{i=1}^{m} w_i a_{ik} \right)^T = \\
&= \left(N, \sum_{i=1}^{m} w_i a_{i1}, \ldots, \sum_{i=1}^{m} w_i a_{ik} \right)^T = \\
&= \left(N, \hat{F}_{x_1,cal}\left(Q_{x_1,\alpha}\right), \ldots, \hat{F}_{x_k,cal}\left(Q_{x_k,\alpha}\right) \right)^T = \\
&= (N, \underbrace{\alpha, \ldots, \alpha}_{k})^T = \mathbf{T}_{\mathbf{a}}.
\end{aligned} \tag{3.41}$$

Zakładając, że macierz $\sum_{i=1}^{m} d_i \mathbf{a}_i \mathbf{a}_i^T$ jest nieosobliwa, otrzymujemy następującą postać wektora $\boldsymbol{\lambda}$:

$$\boldsymbol{\lambda} = \left(\sum_{i=1}^{m} d_i \mathbf{a}_i \mathbf{a}_i^T \right)^{-1} \left(\mathbf{T}_{\mathbf{a}} - \sum_{i=1}^{m} d_i \mathbf{a}_i \right). \tag{3.42}$$

Stąd, korzystając z równania (3.37) i wyznaczonej postaci wektora $\boldsymbol{\lambda}$ otrzymujemy, że:

$$w_i = d_i \left(1 + \mathbf{a}_i^T \boldsymbol{\lambda}\right) = d_i + d_i \mathbf{a}_i^T \left(\sum_{i=1}^{m} d_i \mathbf{a}_i \mathbf{a}_i^T \right)^{-1} \left(\mathbf{T}_{\mathbf{a}} - \sum_{i=1}^{m} d_i \mathbf{a}_i \right), \tag{3.43}$$

a więc, poszukiwaną postać składowej wektora wag kalibracyjnych.

Należy jeszcze sprawdzić, że w punkcie $\mathbf{w} = (w_1, \ldots, w_m)^T$ istnieje minimum (warunek dostateczny istnienia ekstremum warunkowego). Niech ξ będzie niezerowym wektorem takim, że $\xi \in \mathbb{R}^m$. Należy wykazać, że forma kwadratowa $d^2L(\mathbf{w})(\xi)$ jest dodatnio określona. Mamy:

$$d^2L(\mathbf{w})(\xi) = \sum_{i,j=1}^{m} \frac{\partial^2 L}{\partial w_i \partial w_j} \xi_i \xi_j. \tag{3.44}$$

Zauważmy, że:

$$\frac{\partial^2 L}{\partial w_i w_j} = \begin{cases} \pi_i & \text{dla } i = j, \\ 0 & \text{dla } i \neq j. \end{cases} \tag{3.45}$$

Podstawiając obliczone pochodne drugiego rzędu do formy kwadratowej (3.44) otrzymujemy, że:

$$d^2L(\mathbf{w})(\xi) = \sum_{i,j=1}^{m} \frac{\partial^2 L}{\partial w_i \partial w_j} \xi_i \xi_j = \sum_{i=1}^{m} \pi_i \xi_i^2. \tag{3.46}$$

Jest to oczywiście forma kwadratowa dodatnio określona. Stąd wektor wag kalibracyjnych $\mathbf{w} = (w_1, \ldots, w_m)^T$, którego składowe spełniają warunek (3.23), jest poszukiwanym rozwiązaniem zadania minimalizacji funkcji odległości. ■

Po wyznaczeniu wektora wag kalibracyjnych $\mathbf{w} = (w_1, \ldots, w_m)^T$, kolejnym etapem jest oszacowanie kwantyla $Q_{y,\alpha}$ rzędu α zmiennej y tj. $\hat{Q}_{y,cal,\alpha}$. Poniższe twierdzenie umożliwia konstrukcję estymatora kalibracyjnego takiego kwantyla.

Twierdzenie 7. *Jeżeli dla pewnego $p \in \{1, \ldots, m-1\}$ spełnione są następujące warunki:*

$$\hat{F}_{y,cal}(y_p) \leqslant \alpha, \quad \hat{F}_{y,cal}(y_{p+1}) > \alpha, \tag{3.47}$$

to estymator kalibracyjny $\hat{Q}_{y,cal,\alpha}$ kwantyla $Q_{y,\alpha}$ wyraża się następującym wzorem:

$$\hat{Q}_{y,cal,\alpha} = y_p + \frac{N\alpha - \sum_{i=1}^{p} w_i}{w_{p+1}} (y_{p+1} - y_p). \tag{3.48}$$

Dowód. Pokażemy, że $\hat{F}_{y,cal}\left(\hat{Q}_{y,cal,\alpha}\right) = \alpha$. Bez straty ogólności możemy założyć, że $y_1 \leqslant y_2 \leqslant \ldots \leqslant y_m$. W przeciwnym przypadku możemy uporządkować wartości zmiennej y w zbiorze respondentów r w ciąg niemalejący i dokonać przenumerowania indeksów odpowiadających poszczególnym jednostkom. Z założenia, że istnieje jednostka p, dla której $\hat{F}_{y,cal}(y_p) \leqslant \alpha$, oraz $\hat{F}_{y,cal}(y_{p+1}) > \alpha$ wynika, że $\hat{Q}_{y,cal,\alpha} \in \langle y_p, y_{p+1})$. W związku z powyższym zachodzą poniższe równania:

$$L_{y,r}\left(\hat{Q}_{y,cal,\alpha}\right) = y_p, \tag{3.49}$$

$$U_{y,r}\left(\hat{Q}_{y,cal,\alpha}\right) = y_{p+1}, \tag{3.50}$$

$$\beta_{y,r}\left(\hat{Q}_{y,cal,\alpha}\right) = \frac{\hat{Q}_{y,cal,\alpha} - y_p}{y_{p+1} - y_p}. \tag{3.51}$$

W konsekwencji:

$$H_{y,r}\left(\hat{Q}_{y,cal,\alpha}, y_i\right) = \begin{cases} 1, & y_i \leqslant y_p, \\ \frac{\hat{Q}_{y,cal,\alpha}-y_p}{y_{p+1}-y_p}, & y_i = y_{p+1}, \\ 0, & y_i > y_{p+1}. \end{cases} \tag{3.52}$$

Stąd otrzymujemy, że:

$$\hat{F}_{y,cal}\left(\hat{Q}_{y,cal,\alpha}\right) = \frac{\sum_{i=1}^{m} w_i H_{y,r}\left(\hat{Q}_{y,cal,\alpha}, y_i\right)}{\sum_{i=1}^{m} w_i} \underset{(3.21)}{=} \frac{\sum_{i=1}^{m} w_i H_{y,r}\left(\hat{Q}_{y,cal,\alpha}, y_i\right)}{N} =$$

$$= \frac{\sum_{i=1}^{p} w_i H_{y,r}\left(\hat{Q}_{y,cal,\alpha}, y_i\right) + w_{p+1} H_{y,r}\left(\hat{Q}_{y,cal,\alpha}, y_{p+1}\right) + \sum_{i=p+2}^{m} w_i H_{y,r}\left(\hat{Q}_{y,cal,\alpha}, y_i\right)}{N} =$$

$$\underset{(3.52)}{=} \frac{\sum_{i=1}^{p} w_i + w_{p+1}\frac{\hat{Q}_{y,cal,\alpha}-y_p}{y_{p+1}-y_p}}{N} = \frac{\sum_{i=1}^{p} w_i + w_{p+1}\frac{y_p + \frac{N\alpha - \sum_{i=1}^{p} w_i}{w_{p+1}}(y_{p+1}-y_p) - y_p}{y_{p+1}-y_p}}{N} =$$

$$= \frac{\sum_{i=1}^{p} w_i + w_{p+1}\frac{\frac{N\alpha - \sum_{i=1}^{p} w_i}{w_{p+1}}(y_{p+1}-y_p)}{y_{p+1}-y_p}}{N} = \frac{\sum_{i=1}^{p} w_i + w_{p+1}\frac{N\alpha - \sum_{i=1}^{p} w_i}{w_{p+1}}}{N} = \alpha, \tag{3.53}$$

co kończy dowód twierdzenia. ■

Jedną z bardzo ważnych własności estymatora kalibracyjnego kwantyla rzędu α zmiennej y, w przypadku gdy dysponujemy tylko jedną zmienną pomocniczą x, tj. gdy $k = 1$ podaje poniższe twierdzenie. Zgodnie z nim, jeżeli związek pomiędzy zmienną y, a zmienną pomocniczą x jest liniowy oraz znany jest kwantyl rzędu α zmiennej x na poziomie całej populacji, to wówczas jesteśmy w stanie uzyskać dokładne oszacowanie wartości kwantyla $Q_{y,\alpha}$ zmiennej y, w przypadku gdy współczynnik kierunkowy jest dodatni oraz $Q_{y,1-\alpha}$, w przypadku gdy współczynnik kierunkowy jest ujemny.

Oczywiście w praktyce badań statystycznych taka liniowa zależność na poziomie całej populacji nie istnieje. Twierdzenie to jednak pokazuje, że w przypadku, gdy istnieje silna liniowa zależność między zmiennymi y i x ocena punktowa estymatora kwantyla $Q_{y,\alpha}$ powinna być bliska jego prawdziwej wartości.

Twierdzenie 8. *Załóżmy, że dla wszystkich jednostek $i \in U$ spełniona jest zależność $y_i = a_0 + a_1 x_i$ oraz wartości zmiennej pomocniczej x tworzą ciąg niemalejący tj. $x_1 \leqslant \ldots \leqslant x_U$. Niech ponadto dla zmiennej x znany będzie kwantyl $Q_{x,\alpha}$ rzędu α, dla którego istnieją dwie jednostki $t_1, t_2 \in r$ takie, że $x_{t_1} < Q_{x,\alpha} < x_{t_2}$. Wówczas:*

- *jeżeli $a_1 > 0$ to $Q_{y,\alpha} = \hat{Q}_{y,cal,\alpha}$,*
- *jeżeli $a_1 < 0$ to $Q_{y,1-\alpha} = \hat{Q}_{y,cal,1-\alpha}$.*

Dowód. Rozważymy dwa przypadki $a_1 > 0$ oraz $a_1 < 0$. Przypadek, kiedy $a_1 = 0$ jest oczywisty, ponieważ wtedy dla każdej jednostki $i \in U$, $y_i = x_i$, a więc i odpowiednie kwantyle zmiennej y i x są sobie równe.

Niech $a_1 > 0$ oraz $t \in \mathbb{R}$. Ponieważ zależność $y_i = a_0 + a_1 x_i$ zachodzi dla wszystkich jednostek $i \in U$ oraz $a_1 > 0$, to wtedy spełnione są następujące równości:

$$\begin{aligned} &L_{y,r}\left(a_0+a_1 t\right) = \max\left\{\left\{y_i, i \in r \,|y_i \leqslant a_0 + a_1 t\right\} \cup \{-\infty\}\right\} = \\ &= \max\left\{\left\{a_0 + a_1 x_i, i \in r \,|a_0 + a_1 x_i \leqslant a_0 + a_1 t\right\} \cup \{-\infty\}\right\} = \\ &= a_0 + a_1 \max\left\{\left\{x_i, i \in r \,|x_i \leqslant t\right\} \cup \{-\infty\}\right\} = a_0 + a_1 L_{x,r}\left(t\right), \end{aligned} \tag{3.54}$$

$$\begin{aligned} &U_{y,r}\left(a_0+a_1 t\right) = \min\left\{\left\{y_i, i \in r \,|y_i > a_0 + a_1 t\right\} \cup \{\infty\}\right\} = \\ &= \min\left\{\left\{a_0 + a_1 x_i, i \in r \,|a_0 + a_1 x_i > a_0 + a_1 t\right\} \cup \{\infty\}\right\} = \\ &= a_0 + a_1 \min\left\{\left\{x_i, i \in r \,|x_i > t\right\} \cup \{\infty\}\right\} = a_0 + a_1 U_{x,r}\left(t\right) \end{aligned} \tag{3.55}$$

oraz

$$\begin{aligned} &\beta_{y,r}\left(a_0+a_1 t\right) = \frac{a_0 + a_1 t - L_{y,r}\left(a_0+a_1 t\right)}{U_{y,r}\left(a_0+a_1 t\right) - L_{y,r}\left(a_0+a_1 t\right)} = \\ &= \frac{a_0 + a_1 t - \left(a_0 + a_1 L_{x,r}\left(t\right)\right)}{a_0 + a_1 U_{x,r}\left(t\right) - \left(a_0 + a_1 L_{x,r}\left(t\right)\right)} = \frac{a_1\left(t - L_{x,r}\left(t\right)\right)}{a_1\left(U_{x,r}\left(t\right) - L_{x,r}\left(t\right)\right)} = \\ &= \frac{t - L_{x,r}\left(t\right)}{U_{x,r}\left(t\right) - L_{x,r}\left(t\right)} = \beta_{x,r}\left(t\right). \end{aligned} \tag{3.56}$$

W konsekwencji spełniona jest następująca równość:

$$H_{y,r}\left(a_0 + a_1 t, y_i\right) = H_{x,r}\left(t, x_i\right). \tag{3.57}$$

Mamy bowiem:

$$\begin{aligned} H_{y,r}\left(a_0 + a_1 t, y_i\right) &= \begin{cases} 1, & y_i \leqslant L_{y,r}\left(a_0 + a_1 t\right) \\ \beta_{y,r}\left(a_0 + a_1 t\right), & y_i = U_{y,r}\left(a_0 + a_1 t\right) = \\ 0, & y_i > U_{y,r}\left(a_0 + a_1 t\right) \end{cases} \\ &= \begin{cases} 1, & a_0 + a_1 x_i \leqslant a_0 + a_1 L_{x,r}\left(t\right) \\ \beta_{x,r}\left(t\right), & a_0 + a_1 x_i = a_0 + a_1 U_{x,r}\left(t\right) = \\ 0, & a_0 + a_1 x_i > a_0 + a_1 U_{x,r}\left(t\right) \end{cases} \\ &= \begin{cases} 1, & x_i \leqslant L_{x,r}\left(t\right) \\ \beta_{x,r}\left(t\right), & x_i = U_{x,r}\left(t\right) = H_{x,r}\left(t, x_i\right). \\ 0, & x_i > U_{x,r}\left(t\right) \end{cases} \end{aligned} \tag{3.58}$$

Ponadto

$$\hat{F}_{y,cal}\left(a_0 + a_1 t\right) = \hat{F}_{x,cal}\left(t\right). \tag{3.59}$$

Wynika to z faktu, że:

$$\hat{F}_{y,cal}\left(a_0 + a_1 t\right) = \frac{\sum_{i=1}^{m} w_i H_{y,r}\left(a_0 + a_1 t, y_i\right)}{\sum_{i=1}^{m} w_i} = \frac{\sum_{i=1}^{m} w_i H_{x,r}\left(t, x_i\right)}{\sum_{i=1}^{m} w_i} = \hat{F}_{x,cal}\left(t\right). \tag{3.60}$$

Dokonując podstawienia $t = Q_{x,\alpha}$ otrzymujemy, że[30]:

$$\hat{F}_{y,cal}\left(a_0 + a_1 Q_{x,\alpha}\right) = \hat{F}_{x,cal}\left(Q_{x,\alpha}\right) = \alpha. \tag{3.61}$$

Ponieważ jednak $a_0 + a_1 Q_{x,\alpha} = Q_{y,\alpha}$ to również:

$$\hat{F}_{y,cal}\left(Q_{y,\alpha}\right) = \alpha. \tag{3.62}$$

W konsekwencji $\hat{Q}_{y,cal,\alpha} = \hat{F}^{-1}_{y,cal}\left(\alpha\right) = Q_{y,\alpha}$.

Niech $a_1 < 0$ oraz $t \in \mathbb{R}$. W tym przypadku dodatkowo rozważymy następujące dwie sytuacje. W pierwszej, załóżmy, że istnieje taka jednostka $j \in r$, że $x_j = t$. Ponieważ zależność $y_i = a_0 + a_1 x_i$ zachodzi dla wszystkich jednostek $i \in U$ oraz $a_1 < 0$ i dla pewnego $j \in r, x_j = t$, to wtedy spełnione są następujące równości:

$$\begin{aligned}
&L_{y,r}\left(a_0 + a_1 t\right) = \max\left\{\left\{y_i, i \in r \,|y_i \leqslant a_0 + a_1 t\right\} \cup \left\{-\infty\right\}\right\} = \\
&= \max\left\{\left\{a_0 + a_1 x_i, i \in r \,|a_0 + a_1 x_i \leqslant a_0 + a_1 t\right\} \cup \left\{-\infty\right\}\right\} = \\
&= \max\left\{\left\{a_0 + a_1 x_i, i \in r \,|x_i \geqslant t\right\} \cup \left\{-\infty\right\}\right\} = \\
&= \max\left\{\left\{a_0 + a_1 x_i, i \in r \,|x_j = t\right\} \cup \left\{-\infty\right\}\right\} = a_0 + a_1 x_j = \\
&= a_0 + a_1 t = a_0 + a_1 \max\left\{\left\{x_i, i \in r \,|x_i \leqslant t\right\} \cup \left\{-\infty\right\}\right\} = a_0 + a_1 L_{x,r}\left(t\right),
\end{aligned} \tag{3.63}$$

$$\begin{aligned}
&U_{y,r}\left(a_0 + a_1 t\right) = \min\left\{\left\{y_i, i \in r \,|y_i > a_0 + a_1 t\right\} \cup \left\{\infty\right\}\right\} = \\
&= \min\left\{\left\{a_0 + a_1 x_i, i \in r \,|a_0 + a_1 x_i > a_0 + a_1 t\right\} \cup \left\{\infty\right\}\right\} = \\
&= \min\left\{\left\{a_0 + a_1 x_i, i \in r \,|x_i < t\right\} \cup \left\{\infty\right\}\right\} = a_0 + a_1 x_{j-1}
\end{aligned} \tag{3.64}$$

oraz

$$\begin{aligned}
\beta_{y,r}\left(a_0 + a_1 t\right) &= \frac{a_0 + a_1 t - L_{y,r}\left(a_0 + a_1 t\right)}{U_{y,r}\left(a_0 + a_1 t\right) - L_{y,r}\left(a_0 + a_1 t\right)} = \\
&= \frac{a_0 + a_1 t - \left(a_0 + a_1 t\right)}{a_0 + a_1 x_{j-1} - \left(a_0 + a_1 x_j\right)} = 0.
\end{aligned} \tag{3.65}$$

Z powyższych równości wynika, że:

$$\begin{aligned}
H_{y,r}\left(a_0 + a_1 t, y_i\right) &= \begin{cases} 1, & y_i \leqslant L_{y,r}\left(a_0 + a_1 t\right) \\ \beta_{y,r}\left(a_0 + a_1 t\right), & y_i = U_{y,r}\left(a_0 + a_1 t\right) = \\ 0, & y_i > U_{y,r}\left(a_0 + a_1 t\right) \end{cases} \\
&= \begin{cases} 1, \; a_0 + a_1 x_i \leqslant a_0 + a_1 x_j \\ 0, \; a_0 + a_1 x_i = a_0 + a_1 x_{j-1} \\ 0, \; a_0 + a_1 x_i > a_0 + a_1 x_{j-1} \end{cases} = \begin{cases} 1, \; x_i \geqslant x_j \\ 0, \; x_i \leqslant x_{j-1} \end{cases}.
\end{aligned} \tag{3.66}$$

Z drugiej zaś strony

$$H_{x,r}\left(t, x_i\right) = H_{x,r}\left(x_j, x_i\right) = \begin{cases} 1, & x_i \geqslant x_j \\ 0, & x_i \leqslant x_{j+1} \end{cases}. \tag{3.67}$$

[30] Założenie, że istnieją dwie jednostki $t_1, t_2 \in r$ takie, że $x_{t_1} < Q_{x,\alpha} < x_{t_2}$ gwarantuje, że $L_{y,r}\left(a_0 + a_1 Q_{x,\alpha}\right) \neq -\infty$ oraz $U_{y,r}\left(a_0 + a_1 Q_{x,\alpha}\right) \neq \infty$. Innymi słowy, założenie to gwarantuje nam istnienie kwantyla $\hat{Q}_{y,cal,\alpha}$.

Wynika to z faktu, że:

$$L_{x,r}(t) = L_{x,r}(x_j) = x_j, \tag{3.68}$$

$$U_{x,r}(t) = U_{x,r}(x_j) = x_{j+1}, \tag{3.69}$$

$$\beta_{x,r}(t) = \beta_{x,r}(x_j) = 0. \tag{3.70}$$

Z równości (3.66), (3.67) wynika, że

$$H_{y,r}(a_0 + a_1 t, y_i) = H_{y,r}(a_0 + a_1 x_j, y_i) = 1 - H_{x,r}(t, x_i) = 1 - H_{x,r}(x_j, x_i). \tag{3.71}$$

Załóżmy teraz, że dla każdej jednostki $j \in r$, $x_j \neq t$. Przyjmijmy, że dla pewnego $j \in r$, $x_j < t < x_{j+1}$. Ponieważ zależność $y_i = a_0 + a_1 x_i$ zachodzi dla wszystkich jednostek $i \in U$ oraz $a_1 < 0$ i dla każdej jednostki $j \in r, x_j \neq t$, to wtedy spełnione są następujące równości:

$$L_{x,r}(t) = x_j, \tag{3.72}$$

$$U_{x,r}(t) = x_{j+1}. \tag{3.73}$$

Stąd:

$$\begin{aligned} &L_{y,r}(a_0+a_1 t) = \max\{\{y_i, i \in r \,|y_i \leqslant a_0 + a_1 t\} \cup \{-\infty\}\} = \\ &= \max\{\{a_0 + a_1 x_i, i \in r \,|a_0 + a_1 x_i \leqslant a_0 + a_1 t\} \cup \{-\infty\}\} = \\ &= \max\{\{a_0 + a_1 x_i, i \in r \,|x_i \geqslant t\} \cup \{-\infty\}\} = a_0 + a_1 x_{j+1} = a_0 + a_1 U_{x,r}(t), \end{aligned} \tag{3.74}$$

$$\begin{aligned} &U_{y,r}(a_0+a_1 t) = \min\{\{y_i, i \in r \,|y_i > a_0 + a_1 t\} \cup \{\infty\}\} = \\ &= \min\{\{a_0 + a_1 x_i, i \in r \,|a_0 + a_1 x_i > a_0 + a_1 t\} \cup \{\infty\}\} = \\ &= \min\{\{a_0 + a_1 x_i, i \in r \,|x_i < t\} \cup \{\infty\}\} = a_0 + a_1 x_j = a_0 + a_1 L_{x,r}(t) \end{aligned} \tag{3.75}$$

oraz

$$\begin{aligned} \beta_{y,r}(a_0+a_1 t) &= \frac{a_0 + a_1 t - L_{y,r}(a_0+a_1 t)}{U_{y,r}(a_0+a_1 t) - L_{y,r}(a_0+a_1 t)} = \\ &= \frac{a_0 + a_1 t - (a_0 + a_1 U_{x,r}(t))}{a_0 + a_1 L_{x,r}(t) - (a_0 + a_1 U_{x,r}(t))} = \frac{t - U_{x,r}(t)}{L_{x,r}(t) - U_{x,r}(t)} = \\ &= \frac{U_{x,r}(t) - t}{U_{x,r}(t) - L_{x,r}(t)} = 1 - \frac{t - L_{x,r}(t)}{U_{x,r}(t) - L_{x,r}(t)} = 1 - \beta_{x,r}(t). \end{aligned} \tag{3.76}$$

$$\begin{aligned} H_{y,r}(a_0 + a_1 t, y_i) &= \begin{cases} 1, & y_i \leqslant L_{y,r}(a_0 + a_1 t) \\ \beta_{y,r}(a_0 + a_1 t), & y_i = U_{y,r}(a_0 + a_1 t) = \\ 0, & y_i > U_{y,r}(a_0 + a_1 t) \end{cases} \\ &= \begin{cases} 1, & a_0 + a_1 x_i \leqslant a_0 + a_1 U_{x,r}(t) \\ 1 - \beta_{x,r}(t), & a_0 + a_1 x_i = a_0 + a_1 L_{x,r}(t) = \\ 0, & a_0 + a_1 x_i > a_0 + a_1 L_{x,r}(t) \end{cases} \\ &= \begin{cases} 1, & x_i \geqslant U_{x,r}(t) \\ 1 - \beta_{x,r}(t), & x_i = L_{x,r}(t) = 1 - H_{x,r}(t, x_i). \\ 0, & x_i < L_{x,r}(t) \end{cases} \end{aligned} \tag{3.77}$$

W obydwu przypadkach, gdy $a_1 < 0$ otrzymaliśmy, że:

$$H_{y,r}\left(a_0 + a_1 t, y_i\right) = 1 - H_{x,r}\left(t, x_i\right). \tag{3.78}$$

W przypadku, gdy współczynnik kierunkowy $a_1 < 0$ spełniona jest następująca równość:

$$a_0 + a_1 Q_{x,\alpha} = Q_{y,1-\alpha}. \tag{3.79}$$

Ponadto

$$\hat{F}_{y,cal}\left(a_0 + a_1 t\right) = 1 - \hat{F}_{x,cal}\left(t\right). \tag{3.80}$$

Wynika to z faktu, że:

$$\hat{F}_{y,cal}\left(a_0 + a_1 t\right) = \frac{\sum_{i=1}^{m} w_i H_{y,r}\left(a_0 + a_1 t, y_i\right)}{\sum_{i=1}^{m} w_i} = \frac{\sum_{i=1}^{m} w_i \left(1 - H_{x,r}\left(t, x_i\right)\right)}{\sum_{i=1}^{m} w_i} = 1 - \hat{F}_{x,cal}\left(t\right). \tag{3.81}$$

Stąd, przyjmując, że $t = Q_{x,\alpha}$ otrzymujemy, że:

$$\hat{F}_{y,cal}\left(Q_{y,1-\alpha}\right) = \hat{F}_{y,cal}\left(a_0 + a_1 Q_{x,\alpha}\right) = 1 - \hat{F}_{x,cal}\left(Q_{x,\alpha}\right) = 1 - \alpha. \tag{3.82}$$

Ostatecznie $\hat{Q}_{y,cal,1-\alpha} = \hat{F}^{-1}_{y,cal}\left(1-\alpha\right) = Q_{y,1-\alpha}$. ■

3.3. Estymator kalibracyjny kwantyla rzędu α ze znanym wektorem $\hat{Q}_{\boldsymbol{x},\alpha}$

W sytuacji, gdy nie dysponujemy wektorem $Q_{\boldsymbol{x},\alpha} = \left(Q_{x_1,\alpha}, \ldots, Q_{x_k,\alpha}\right)^T$ kwantyli rzędu α, zachodzi konieczność oszacowania jego składowych na podstawie informacji pochodzących z próby. Jest to często spotykana sytuacja ze względu na fakt, że nie zawsze odpowiednie dane o kwantylach, możliwe są do pozyskania ze spisów bądź odpowiednich rejestrów administracyjnych.

W drugim rozważanym podejściu zakładamy, że:

- dla wszystkich jednostek z próby znana jest wartość każdej zmiennej pomocniczej, tj. znana jest macierz $\mathbf{X_s}$ określona wzorem (2.7),
- nie jest znany wektor $Q_{\boldsymbol{x},\alpha}$ kwantyli rzędu α wszystkich zmiennych pomocniczych,
- znany jest wektor $\hat{Q}_{\boldsymbol{x},\alpha} = \left(\hat{Q}_{x_1,\alpha}, \ldots, \hat{Q}_{x_k,\alpha}\right)^T$, którego poszczególne składowe są oszacowaniami odpowiednich składowych wektora $Q_{\boldsymbol{x},\alpha}$ w oparciu o dane pochodzące z próby s.

W podejściu tym nie wymaga się, aby znane były kwantyle rzędu α na poziomie populacji dla wszystkich zmiennych pomocniczych. Wymaga się jednak, aby znane były wartości tych zmiennych dla wszystkich jednostek wylosowanych do próby s, celem oszacowania ich kwantytli. Pytanie jakie przy tym pojawia się w tej sytuacji jest, jakiego estymatora użyć, aby móc znaleźć oszacowania kwantyli zmiennych pomocniczych. Naturalnym kandydatem jest estymator Horvitza-Thompsona, zdefiniowany poniżej.

Definicja 15. *Estymatorem Horvitza-Thompsona kwantyla rzędu α zmiennej pomocniczej x_j jest:*

$$\hat{Q}_{x_j,HT,\alpha} = \hat{F}_{x_j}^{-1}(\alpha), \tag{3.83}$$

gdzie:

$$\hat{F}_{x_j}(t) = \frac{\sum_{i=1}^{n} d_i H_{x_j,s}(t, x_{ij})}{\sum_{i=1}^{n} d_i}, \tag{3.84}$$

a funkcja $H_{x_j,s}(t, x_{ij})$ zdefiniowana jest wzorem (3.10), $j = 1, \ldots, k$, $t \in \mathbb{R}$.

Definicja estymatora kalibracyjnego kwantyla rzędu α zmiennej y przy znanym wektorze $\hat{Q}_{\boldsymbol{x},\alpha}$ oszacowanych na podstawie próby s kwantyli zmiennych pomocniczych, przyjmuje następującą postać.

Definicja 16. *Estymatorem kalibracyjnym kwantyla $Q_{y,\alpha}$ rzędu α zmiennej y jest:*

$$\hat{Q}_{y,cal,\alpha} = \hat{F}_{y,cal}^{-1}(\alpha), \tag{3.85}$$

gdzie dystrybuanta interpolacyjna $\hat{F}_{y,cal}(t)$ określona jest wzorem (3.4), a wektor wag kalibracyjnych $\boldsymbol{w} = (w_1, \ldots, w_m)^T$ jest rozwiązaniem zadania minimalizacji:

$$\mathbf{w} = \operatorname{argmin}_{\mathbf{v}} D(\mathbf{v}, \mathbf{d}), \tag{3.86}$$

przy warunkach:

$$\sum_{i=1}^{m} w_i = N, \tag{3.87}$$

$$\hat{F}_{x_i,cal}\left(\hat{Q}_{x_i,\alpha}\right) = \alpha, \tag{3.88}$$

gdzie $i = 1, \ldots, k$, a funkcja odległości $D(\mathbf{v}, \mathbf{d})$ określona jest wzorem (2.4).

Równanie (3.88) jest odpowiednikiem równania kalibracyjnego (2.36) wartości globalnej ze znanym wektorem $\breve{\mathbf{X}}$. Zgodnie z nim, wektora wag $\mathbf{w}$ poszukujemy w ten sposób, aby dla każdej zmiennej pomocniczej, ocena estymatora Horvitza-Thompsona kwantyla rzędu α wyznaczona na podstawie danych pochodzących z próby s, równała się oszacowanemu kwantylowi na podstawie zbioru respondentów r z uwzględnieniem wag kalibracyjnych.

Definicja estymatora kalibracyjnego kwantyla (14) różni się od definicji (16) użytym w równaniu kalibracyjnym wektorem. W definicji (14) zakładamy, że znane są kwantyle rzędu α zmiennych pomocniczych na poziomie całej populacji, a w definicji (16) zakładamy, że znane są tylko ich oszacowania uzyskane w oparciu o dane pochodzące z próby s.

Podobnie jak w twierdzeniu (6) można znaleźć wektor wag kalibracyjnych – korzystając z metody czynników nieoznaczonych Lagrange'a.

Twierdzenie 9. *Rozwiązaniem zadania minimalizacji funkcji odległości (3.86), przy warunku (3.87) i (3.88) jest wektor wag kalibracyjnych* $\mathbf{w} = (w_1, \ldots, w_m)^T$*, którego składowe spełniają równanie:*

$$w_i = d_i + d_i \mathbf{a}_i^T \left(\sum_{i=1}^{m} d_i \mathbf{a}_i \mathbf{a}_i^T \right)^{-1} \left(\mathbf{T}_{\mathbf{a}} - \sum_{i=1}^{m} d_i \mathbf{a}_i \right), \tag{3.89}$$

gdzie:

$$\mathbf{T}_{\mathbf{a}} = (N, \underbrace{\alpha, \ldots, \alpha}_{k})^T, \tag{3.90}$$

$$\mathbf{a}_i = (1, a_{i1}, \ldots, a_{ik})^T, \tag{3.91}$$

przy czym:

$$a_{ij} = \begin{cases} N^{-1}, & x_{ij} \leqslant L_{x_j,r}(\hat{Q}_{x_j,\alpha}), \\ N^{-1}\beta_{x_j,r}(\hat{Q}_{x_j,\alpha}), & x_{ij} = U_{x_j,r}(\hat{Q}_{x_j,\alpha}), \\ 0, & x_{ij} > U_{x_j,r}(\hat{Q}_{x_j,\alpha}), \end{cases} \tag{3.92}$$

$i = 1, \ldots, m$, $j = 1, \ldots, k$.

Dowód. Dowód twierdzenia jest analogiczny do dowodu twierdzenia (6). Wystarczy tylko w miejsce wektora $Q_{\boldsymbol{x},\alpha}$ kwantyli rzędu α zmiennych pomocniczych na poziomie całej populacji, podstawić wektor $\hat{Q}_{\boldsymbol{x},\alpha}$ ich oszacowań uzyskanych z wykorzystaniem estymatora Horvitza-Thompsona określonego w definicji (15). ■

3.4. Estymator kalibracyjny kwantyla rzędu α ze znaną macierzą $\mathbf{Q}_{\boldsymbol{x},\alpha}$

Rozważany w tym podrozdziale przypadek estymatora kalibracyjnego rzędu α jest uogólnieniem dwóch poprzednich. Zakładać będziemy, że znane są wektory $Q_{\boldsymbol{x},\alpha}$ i $\hat{Q}_{\boldsymbol{x},\alpha}$ kwantyla rzędu α. Przyjmujemy więc tutaj założenie, że dysponujemy kwantylami rzędu α dla części zmiennych pomocniczych na poziomie całej populacji oraz ich oszacowaniami na podstawie próby s dla pozostałych zmiennych. Zakładamy więc, że dla k_1 zmiennych pomocniczych znane są ich kwantyle rzędu α na poziomie całej populacji, a dla k_2 zmiennych pomocniczych tylko ich oszacowania na podstawie n – elementowej próby s, przy czym $k_1 + k_2 = k$.

Niech $\mathbf{Q}_{\boldsymbol{x},\alpha}$ oznacza macierz złożoną z kwantyli rzędu α wszystkich zmiennych pomocniczych x_i, $i = 1, \ldots, k$. Macierz tą możemy zapisać jako:

$$\mathbf{Q}_{\boldsymbol{x},\alpha} = \begin{pmatrix} Q_{\boldsymbol{x},\alpha} \\ \hat{Q}_{\boldsymbol{x},\alpha} \end{pmatrix}, \tag{3.93}$$

gdzie $Q_{\boldsymbol{x},\alpha} = \left(Q_{x_1,\alpha}, \ldots, Q_{x_{k_1},\alpha}\right)^T$ i $\hat{Q}_{\boldsymbol{x},\alpha} = \left(\hat{Q}_{x_1,\alpha}, \ldots, \hat{Q}_{x_{k_2},\alpha}\right)^T$. Zakładać tutaj będziemy, podobnie jak w przypadku uogólnionego estymatora kalibracyjnego wartości globalnej, że jeżeli dla jakiejś zmiennej pomocniczej znany jest jej kwantyl w populacji, to zmiennej tej nie uwzględniamy w budowie wektora $\hat{Q}_{\boldsymbol{x},\alpha}$.

W przypadku, gdy znane są tylko kwantyle rzędu α wszystkich zmiennych pomocniczych na poziomie całej populacji, to wtedy $\mathbf{Q}_{\boldsymbol{x},\alpha} = Q_{\boldsymbol{x},\alpha}$. Natomiast w przypadku, gdy znane są tylko ich oszacowania, to macierz $\mathbf{Q}_{\boldsymbol{x},\alpha}$ redukuje się do $\hat{Q}_{\boldsymbol{x},\alpha}$.

Naturalnym jest, że uwzględnienie tych dwóch przypadków razem powinno poprawić proces estymacji. Wagi kalibracyjne — po ich zastosowaniu do odpowiednich zmiennych pomocniczych — umożliwią nam bowiem uzyskanie znanych kwantyli na poziomie całej populacji bądź ich oszacowań na podstawie informacji zawartych w próbie s. W konsekwencji, gdy zastosujemy je do zmiennej y, to powinniśmy uzyskać oszacowanie kwantyla rzędu α tej zmiennej bliskie jego „prawdziwej" wartości (efekt „naśladowania wag"). Rozważane podejście można więc zastosować w sytuacji, gdy tylko dla niektórych branych pod uwagę zmiennych pomocniczych, znamy ich kwantyle w populacji.

Definicja 17. *Estymatorem kalibracyjnym kwantyla $Q_{y,\alpha}$ rzędu α zmiennej y jest:*

$$\hat{Q}_{y,cal,\alpha} = \hat{F}^{-1}_{y,cal}(\alpha), \tag{3.94}$$

gdzie dystrybuanta interpolacyjna $\hat{F}_{y,cal}(t)$ określona jest wzorem (3.4), a wektor wag kalibracyjnych $\boldsymbol{w} = (w_1, \ldots, w_m)^T$ jest rozwiązaniem zadania minimalizacji:

$$\mathbf{w} = \operatorname{argmin}_{\mathbf{v}} D(\mathbf{v}, \mathbf{d}), \tag{3.95}$$

przy warunkach:

$$\sum_{i=1}^{m} w_i = N, \tag{3.96}$$

$$\left(\hat{F}_{x_1,cal}(Q_{x_1,\alpha}), \ldots, \hat{F}_{x_{k_1},cal}\left(Q_{x_{k_1},\alpha}\right)\right)^T = (\underbrace{\alpha, \ldots, \alpha}_{k_1})^T \tag{3.97}$$

oraz

$$\left(\hat{F}_{x_1,cal}\left(\hat{Q}_{x_1,\alpha}\right), \ldots, \hat{F}_{x_{k_2},cal}\left(\hat{Q}_{x_{k_2},\alpha}\right)\right)^T = (\underbrace{\alpha, \ldots, \alpha}_{k_2})^T \tag{3.98}$$

przy czym funkcja odległości $D(\mathbf{v}, \mathbf{d})$ określona jest wzorem (2.4).

Znalezienie wektora wag kalibracyjnych $\mathbf{w} = (w_1, \ldots, w_m)^T$ wymaga minimalizacji funkcji odległości tak, aby spełnionych było $k+1$ równań kalibracyjnych. Znalezienie postaci wektora wag kalibracyjnych umożliwia poniższe twierdzenie.

Twierdzenie 10. *Rozwiązaniem zadania minimalizacji funkcji odległości (3.95), przy warunku (3.96) i (3.97) jest wektor wag kalibracyjnych $\mathbf{w} = (w_1, \ldots, w_m)^T$, którego składowe spełniają równanie:*

$$w_i = d_i + d_i \mathbf{c}_i^T \left(\sum_{i=1}^{m} d_i \mathbf{c}_i \mathbf{c}_i^T\right)^{-1} \left(\mathbf{T}_{\mathbf{c}} - \sum_{i=1}^{m} d_i \mathbf{c}_i\right), \tag{3.99}$$

gdzie:

$$\mathbf{T}_{\mathbf{c}} = (N, \underbrace{\alpha, \ldots, \alpha}_{k})^T, \tag{3.100}$$

$$\mathbf{c}_i = (c_{i0}, c_{i1}, \ldots, c_{ik})^T = (1, a_{i1}, \ldots, a_{ik_1}, b_{i1}, \ldots, b_{ik_2})^T, \tag{3.101}$$

przy czym dla $i = 1, \ldots, m$ *oraz* $j = 1, \ldots, k_1$

$$a_{ij} = \begin{cases} N^{-1}, & x_{ij} \leqslant L_{x_j,r}(Q_{x_j,\alpha}), \\ N^{-1}\beta_{x_j,r}(Q_{x_j,\alpha}), & x_{ij} = U_{x_j,r}(Q_{x_j,\alpha}), \\ 0, & x_{ij} > U_{x_j,r}(Q_{x_j,\alpha}), \end{cases} \tag{3.102}$$

a dla $i = 1, \ldots, m$ *oraz* $j = 1, \ldots, k_2$

$$b_{ij} = \begin{cases} N^{-1}, & x_{ij} \leqslant L_{x_j,r}(\hat{Q}_{x_j,\alpha}), \\ N^{-1}\beta_{x_j,r}(\hat{Q}_{x_j,\alpha}), & x_{ij} = U_{x_j,r}(\hat{Q}_{x_j,\alpha}), \\ 0, & x_{ij} > U_{x_j,r}(\hat{Q}_{x_j,\alpha}). \end{cases} \tag{3.103}$$

Dowód. Wagi kalibracyjne w_i dla $i = 1, \ldots, m$ znajdziemy, korzystając z metody czynników nieoznaczonych Lagrange'a.
Funkcja Lagrange'a ma postać:

$$\begin{aligned} L = \frac{1}{2}\sum_{i=1}^{m}\frac{(w_i - d_i)^2}{d_i} + \lambda_0\left(N - \sum_{i=1}^{m} w_i\right) + \sum_{j=1}^{k_1}\lambda_j\left(\alpha - \hat{F}_{x_j,cal}\left(Q_{x_j,\alpha}\right)\right) \\ + \sum_{j=1}^{k_2}\lambda_{k_1+j}\left(\alpha - \hat{F}_{x_j,cal}\left(\hat{Q}_{x_j,\alpha}\right)\right). \end{aligned} \tag{3.104}$$

Z równania kalibracyjnego (3.96) wynika, że:

$$\hat{F}_{x_j,cal}\left(Q_{x_j,\alpha}\right) = \frac{\sum_{i=1}^{m} w_i H_{x_j,r}\left(Q_{x_j,\alpha}, x_{ij}\right)}{\sum_{i=1}^{m} w_i} = \frac{\sum_{i=1}^{m} w_i H_{x_j,r}\left(Q_{x_j,\alpha}, x_{ij}\right)}{N} \tag{3.105}$$

oraz

$$\hat{F}_{x_j,cal}\left(\hat{Q}_{x_j,\alpha}\right) = \frac{\sum_{i=1}^{m} w_i H_{x_j,r}\left(\hat{Q}_{x_j,\alpha}, x_{ij}\right)}{\sum_{i=1}^{m} w_i} = \frac{\sum_{i=1}^{m} w_i H_{x_j,r}\left(\hat{Q}_{x_j,\alpha}, x_{ij}\right)}{N}. \tag{3.106}$$

Funkcję Lagrange'a można więc przedstawić w postaci:

$$\begin{aligned} L = \frac{1}{2}\sum_{i=1}^{m}\frac{(w_i - d_i)^2}{d_i} + \lambda_0\left(N - \sum_{i=1}^{m} w_i\right) \\ + \sum_{j=1}^{k_1}\lambda_j\left(\alpha - \frac{\sum_{i=1}^{m} w_i H_{x_j,r}\left(Q_{x_j,\alpha}, x_{ij}\right)}{N}\right) \\ + \sum_{j=1}^{k_2}\lambda_{k_1+j}\left(\alpha - \frac{\sum_{i=1}^{m} w_i H_{x_j,r}\left(\hat{Q}_{x_j,\alpha}, x_{ij}\right)}{N}\right). \end{aligned} \tag{3.107}$$

Pochodna funkcji L dla $i = 1, \ldots, m$ ma postać:

$$\frac{\partial L}{\partial w_i} = \frac{1}{2} \cdot \frac{2w_i - 2d_i}{d_i} - \lambda_0 - \sum_{j=1}^{k_1} \lambda_j \frac{H_{x_j,r}\left(Q_{x_j,\alpha}, x_{ij}\right)}{N} - \sum_{j=1}^{k_2} \lambda_{k_1+j} \frac{H_{x_j,r}\left(\hat{Q}_{x_j,\alpha}, x_{ij}\right)}{N}. \quad (3.108)$$

Ponieważ:

$$H_{x_j,r}\left(Q_{x_j,\alpha}, x_{ij}\right) = \begin{cases} 1, & x_{ij} \leqslant L_{x_j,r}(Q_{x_j,\alpha}), \\ \beta_{x_j,r}(Q_{x_j,\alpha}), & x_{ij} = U_{x_j,r}(Q_{x_j,\alpha}), \\ 0, & x_{ij} > U_{x_j,r}(Q_{x_j,\alpha}), \end{cases} \quad (3.109)$$

oraz

$$H_{x_j,r}\left(\hat{Q}_{x_j,\alpha}, x_{ij}\right) = \begin{cases} 1, & x_{ij} \leqslant L_{x_j,r}(\hat{Q}_{x_j,\alpha}), \\ \beta_{x_j,r}(\hat{Q}_{x_j,\alpha}), & x_{ij} = U_{x_j,r}(\hat{Q}_{x_j,\alpha}), \\ 0, & x_{ij} > U_{x_j,r}(\hat{Q}_{x_j,\alpha}), \end{cases} \quad (3.110)$$

to przyjmując, że:

$$a_{ij} = \frac{H_{x_j,r}\left(Q_{x_j,\alpha}, x_{ij}\right)}{N} \quad \text{oraz} \quad b_{ij} = \frac{H_{x_j,r}\left(\hat{Q}_{x_j,\alpha}, x_{ij}\right)}{N} \quad (3.111)$$

otrzymujemy:

$$a_{ij} = \begin{cases} N^{-1}, & x_{ij} \leqslant L_{x_j,r}(Q_{x_j,\alpha}), \\ N^{-1}\beta_{x_j,r}(Q_{x_j,\alpha}), & x_{ij} = U_{x_j,r}(Q_{x_j,\alpha}), \\ 0, & x_{ij} > U_{x_j,r}(Q_{x_j,\alpha}), \end{cases} \quad (3.112)$$

$$b_{ij} = \begin{cases} N^{-1}, & x_{ij} \leqslant L_{x_j,r}(\hat{Q}_{x_j,\alpha}), \\ N^{-1}\beta_{x_j,r}(\hat{Q}_{x_j,\alpha}), & x_{ij} = U_{x_j,r}(\hat{Q}_{x_j,\alpha}), \\ 0, & x_{ij} > U_{x_j,r}(\hat{Q}_{x_j,\alpha}). \end{cases} \quad (3.113)$$

Pochodną funkcji L można więc przedstawić w następującej postaci:

$$\frac{\partial L}{\partial w_i} = \frac{w_i - d_i}{d_i} - \lambda_0 - \sum_{j=1}^{k_1} \lambda_j a_{ij} - \sum_{j=1}^{k_2} \lambda_{k_1+j} b_{ij}. \quad (3.114)$$

Przyjmując, że:

$$c_{i0} = 1, c_{i1} = a_{i1}, \ldots, c_{ik_1} = a_{ik_1}, c_{ik_1+1} = b_{i1}, \ldots, c_{ik_2} = b_{ik_2} \quad (3.115)$$

dla $i = 1, \ldots, m$, otrzymujemy, że:

$$\frac{\partial L}{\partial w_i} = \frac{w_i - d_i}{d_i} - \sum_{j=0}^{k} \lambda_j c_{ij}. \quad (3.116)$$

Przyjmując $\mathbf{c}_i = (1, c_{i1}, \ldots, c_{ik})^T$ i $\boldsymbol{\lambda} = (\lambda_0, \ldots, \lambda_k)^T$ dla $i = 1, \ldots, m$, otrzymujemy, że:

$$w_i = d_i\left(1 + \boldsymbol{\lambda}^T \mathbf{c}_i\right) = d_i\left(1 + \mathbf{c}_i^T \boldsymbol{\lambda}\right), \quad (3.117)$$

co wynika z faktu, że:

$$\sum_{j=0}^{k} \lambda_j c_{ij} = \boldsymbol{\lambda}^T \mathbf{c}_i = \mathbf{c}_i^T \boldsymbol{\lambda}. \quad (3.118)$$

Mnożąc obustronnie równanie (3.117) przez $\mathbf{c}_i$, a następnie dokonując sumowania po zbiorze wszystkich respondentów r otrzymujemy:

$$\sum_{i=1}^{m} w_i \mathbf{c}_i = \sum_{i=1}^{m} d_i \mathbf{c}_i \left(1 + \mathbf{c}_i^T \boldsymbol{\lambda}\right). \tag{3.119}$$

Dokonując przekształcenia ostatniego równania otrzymujemy:

$$\sum_{i=1}^{m} w_i \mathbf{c}_i - \sum_{i=1}^{m} d_i \mathbf{c}_i = \sum_{i=1}^{m} d_i \mathbf{c}_i \mathbf{c}_i^T \boldsymbol{\lambda}. \tag{3.120}$$

Zwróćmy uwagę, że:

$$\begin{aligned}
\sum_{i=1}^{m} w_i \mathbf{c}_i &= \left(\sum_{i=1}^{m} w_i, \sum_{i=1}^{m} w_i c_{i1}, \ldots, \sum_{i=1}^{m} w_i c_{ik}\right)^T = \\
&= \left(N, \sum_{i=1}^{m} w_i a_{i1}, \ldots, \sum_{i=1}^{m} w_i a_{ik_1}, \sum_{i=1}^{m} w_i b_{i1}, \ldots, \sum_{i=1}^{m} w_i b_{ik_2}\right)^T = \\
&= \left(N, \hat{F}_{x_1,cal}\left(Q_{x_1,\alpha}\right), \ldots, \hat{F}_{x_{k_1},cal}\left(Q_{x_{k_1},\alpha}\right), \hat{F}_{x_1,cal}\left(\hat{Q}_{x_1,\alpha}\right), \ldots, \hat{F}_{x_{k_2},cal}\left(\hat{Q}_{x_{k_2},\alpha}\right)\right)^T = \\
&= (N, \underbrace{\alpha, \ldots, \alpha}_{k})^T = \mathbf{T}_{\mathbf{c}}.
\end{aligned} \tag{3.121}$$

Zakładając, że macierz $\sum_{i=1}^{m} d_i \mathbf{c}_i \mathbf{c}_i^T$ jest nieosobliwa otrzymujemy następującą postać wektora $\boldsymbol{\lambda}$:

$$\boldsymbol{\lambda} = \left(\sum_{i=1}^{m} d_i \mathbf{c}_i \mathbf{c}_i^T\right)^{-1} \left(\mathbf{T}_{\mathbf{c}} - \sum_{i=1}^{m} d_i \mathbf{c}_i\right). \tag{3.122}$$

Stąd, korzystając z równania (3.117) i wyznaczonej postaci wektora $\boldsymbol{\lambda}$ otrzymujemy, że:

$$w_i = d_i \left(1 + \mathbf{c}_i^T \boldsymbol{\lambda}\right) = d_i + d_i \mathbf{c}_i^T \left(\sum_{i=1}^{m} d_i \mathbf{c}_i \mathbf{c}_i^T\right)^{-1} \left(\mathbf{T}_{\mathbf{c}} - \sum_{i=1}^{m} d_i \mathbf{c}_i\right), \tag{3.123}$$

a więc poszukiwaną postać składowej wektora wag kalibracyjnych $\mathbf{w} = (w_1, \ldots, w_m)^T$.

Należy jeszcze sprawdzić, że w punkcie $\mathbf{w} = (w_1, \ldots, w_m)^T$ istnieje minimum (warunek dostateczny istnienia ekstremum warunkowego). Niech ξ będzie niezerowym wektorem takim, że $\xi \in \mathbb{R}^m$. Należy wykazać, że forma kwadratowa $d^2 L(\mathbf{w})(\xi)$ jest dodatnio określona. Mamy:

$$d^2 L(\mathbf{w})(\xi) = \sum_{i,j=1}^{m} \frac{\partial^2 L}{\partial w_i \partial w_j} \xi_i \xi_j. \tag{3.124}$$

Zauważmy, że:

$$\frac{\partial^2 L}{\partial w_i w_j} = \begin{cases} \pi_i \text{ dla } i = j, \\ 0 \quad \text{dla } i \neq j. \end{cases} \tag{3.125}$$

Podstawiając obliczone pochodne drugiego rzędu do formy kwadratowej (3.124) otrzymujemy, że:

$$d^2 L(\mathbf{w})(\xi) = \sum_{i,j=1}^{m} \frac{\partial^2 L}{\partial w_i \partial w_j} \xi_i \xi_j = \sum_{i=1}^{m} \pi_i \xi_i^2. \tag{3.126}$$

Jest to oczywiście forma kwadratowa dodatnio określona. Stąd wektor wag kalibracyjnych $\mathbf{w} = (w_1, \ldots, w_m)^T$, którego składowe spełniają warunek (3.99), jest poszukiwanym rozwiązaniem zadania minimalizacji funkcji odległości. ■

3.5. Uogólniony estymator kalibracyjny kwantyla rzędu α

W poprzednich podrozdziałach zakładaliśmy, że przy szacowaniu kwantyla $Q_{y,\alpha}$ rzędu α zmiennej y, dysponowaliśmy tylko kwantylami rzędu α zmiennych pomocniczych $x_1, \ldots, x_k$ na poziomie populacji bądź ich oszacowaniami uzyskanymi w oparciu o próbę s. Gdyby jednak dostępna była informacja na temat kwantyli innych rzędów aniżeli α w odniesieniu do zmiennych pomocniczych na poziomie całej populacji (lub ich oszacowań), to w konsekwencji powinniśmy uzyskać wagi kalibracyjne, dla których estymator kalibracyjny charakteryzowałby się mniejszym obciążeniem i wariancją.

Załóżmy, że $A = \{1, \ldots, M\}$ oznacza M – elementowy zbiór, gdzie $M < m - 1$. Niech ponadto $A_1, \ldots, A_k \subseteq A$. Załóżmy ponadto, że dla każdej zmiennej pomocniczej $x_1, \ldots, x_k$ dysponujemy kwantylami odpowiednich rzędów na poziomie całej populacji, tzn. znane są kwantyle postaci:

$$\begin{array}{ll} Q_{x_1,\alpha_i^1}, & i \in A_1, \\ Q_{x_2,\alpha_i^2}, & i \in A_2, \\ \vdots & \vdots \\ Q_{x_k,\alpha_i^k}, & i \in A_k. \end{array} \tag{3.127}$$

Na przykład, jeżeli $A_1 = \{1, 2, 3\}$, to oznacza to, że w odniesieniu do zmiennej pomocniczej x_1 znane są trzy kwantyle Q_{x_1,α_1^1}, Q_{x_1,α_2^1} i Q_{x_1,α_3^1} na poziomie całej populacji. Kwantylami tymi mogą być na przykład: trzy kwartyle, trzy wybrane decyle itd. Zauważmy, że w proponowanym podejściu nie zakładamy, aby dla każdej zmiennej pomocniczej znane były kwantyle tego samego rzędu. Oznacza to, że szacujemy kwantyl $Q_{y,\alpha}$ rzędu α zmiennej y, natomiast w odniesieniu do poszczególnych zmiennych pomocniczych zakładamy, że dysponujemy skończonymi zbiorami kwantyli różnych bądź tych samych rzędów.

Przyjmijmy ponadto, że dla każdej zmiennej pomocniczej $x_1, \ldots, x_k$ dysponujemy oszacowaniami kwantyli odpowiednich rzędów, uzyskanymi na podstawie danych pochodzących z próby s. Zakładamy zatem, że znane są oceny kwantyli zmiennych $x_1, \ldots, x_k$ postaci:

$$\begin{array}{ll} \hat{Q}_{x_1,\beta_i^1}, & i \in B_1, \\ \hat{Q}_{x_2,\beta_i^2}, & i \in B_2, \\ \vdots & \vdots \\ \hat{Q}_{x_k,\beta_i^k}, & i \in B_k, \end{array} \tag{3.128}$$

gdzie $B = \{1, \ldots, M\}$, $M < m-1$, a $B_1, \ldots, B_k \subseteq B$. Zmiana oznaczenia zbioru A na B oznacza, że dla części zmiennych znane mogą być kwantyle dowolnych rzędów na poziomie całej populacji, a dla pozostałych – należy je oszacować korzystając z informacji zawartych w próbie. Przykładowo, w odniesieniu do zmiennej x_1 możemy znać tylko medianę na poziomie całej populacji, a kwartyl pierwszy i trzeci oszacować na podstawie informacji zawartych w próbie s, dla zmiennej x_2 możemy znać wszystkie kwartyle w całej populacji, a dla zmiennej x_3 te kwartyle trzeba będzie oszacować.

Niech l_1 oznacza liczbę wszystkich kwantyli na poziomie populacji, które znane są dla zmiennych pomocniczych $x_1, \ldots, x_k$. Niech ponadto l_2 oznacza liczbę wszyst-

kich oszacowanych kwantyli w oparciu o dane pochodzące z próby s dla zmiennych pomocniczych $x_1, \ldots, x_k$. Załóżmy dodatkowo, że $l = l_1 + l_2$. Mamy więc:

$$l_1 = \sum_{i=1}^{k} |A_i| \,, \tag{3.129}$$

$$l_2 = \sum_{i=1}^{k} |B_i| \,, \tag{3.130}$$

$$l = \sum_{i=1}^{k} \left(|A_i| + |B_i| \right). \tag{3.131}$$

Definicja 18. *Uogólnionym estymatorem kalibracyjnym kwantyla $Q_{y,\alpha}$ rzędu α zmiennej y jest:*

$$\hat{Q}_{y,cal,\alpha} = \hat{F}^{-1}_{y,cal}(\alpha), \tag{3.132}$$

gdzie dystrybuanta interpolacyjna $\hat{F}_{y,cal}(t)$ określona jest wzorem (3.4), a wektor wag kalibracyjnych $\boldsymbol{w} = (w_1, \ldots, w_m)^T$ jest rozwiązaniem zadania minimalizacji:

$$\mathbf{w} = \operatorname{argmin}_{\mathbf{v}} D\left(\mathbf{v}, \mathbf{d}\right), \tag{3.133}$$

przy warunkach:

$$\begin{array}{ll}
\hat{F}_{x_1,cal}\left(Q_{x_1,\alpha_i^1}\right) = \alpha_i^1, & i \in A_1, \\
\vdots & \vdots \\
\hat{F}_{x_k,cal}\left(Q_{x_k,\alpha_i^k}\right) = \alpha_i^k, & i \in A_k, \\
\hat{F}_{x_1,cal}\left(\hat{Q}_{x_1,\beta_i^1}\right) = \beta_i^1, & i \in B_1, \\
\vdots & \vdots \\
\hat{F}_{x_k,cal}\left(\hat{Q}_{x_k,\beta_i^k}\right) = \beta_i^k, & i \in B_k,
\end{array} \tag{3.134}$$

oraz

$$\sum_{i=1}^{m} w_i = N, \tag{3.135}$$

przy czym funkcja odległości $D\left(\mathbf{v}, \mathbf{d}\right)$ określona jest wzorem (2.4).

Znalezienie wektora wag kalibracyjnych $\mathbf{w} = (w_1, \ldots, w_m)^T$ wymaga minimalizacji funkcji odległości tak, aby spełnionych było $2l + 1$ równań kalibracyjnych. Znalezienie postaci wektora wag kalibracyjnych umożliwia poniższe twierdzenie.

Twierdzenie 11. *Rozwiązaniem zadania minimalizacji funkcji odległości (3.133), przy warunkach (3.134) i (3.135) jest wektor wag kalibracyjnych $\mathbf{w} = (w_1, \ldots, w_m)^T$, którego składowe spełniają równanie:*

$$w_i = d_i + d_i \mathbf{c}_i^T \left(\sum_{i=1}^{m} d_i \mathbf{c}_i \mathbf{c}_i^T \right)^{-1} \left(\mathbf{T}_{\mathbf{c}} - \sum_{i=1}^{m} d_i \mathbf{c}_i \right), \tag{3.136}$$

gdzie:

$$\mathbf{T}_{\mathbf{c}} = \left(N, \boldsymbol{\alpha}_{i_1}^1, \ldots, \boldsymbol{\alpha}_{i_k}^k, \boldsymbol{\beta}_{i_1}^1, \ldots, \boldsymbol{\beta}_{i_k}^k \right)^T, \tag{3.137}$$

$$\mathbf{c}_i = (1, \mathbf{a}_{i1}, \ldots, \mathbf{a}_{ik}, \mathbf{b}_{i1}, \ldots, \mathbf{b}_{ik})^T, \tag{3.138}$$

przy czym:

$$\boldsymbol{\alpha}_{i_p}^p = (\alpha_i^p, \ldots, \alpha_i^p), \quad i \in A_p, \tag{3.139}$$

$$\boldsymbol{\beta}_{i_p}^p = (\beta_i^p, \ldots, \beta_i^p), \quad i \in B_p, \tag{3.140}$$

$$\mathbf{a}_{ip} = \left(a_{ip}^1, \ldots, a_{ip}^{|A_p|}\right), \tag{3.141}$$

$$\mathbf{b}_{ip} = \left(b_{ip}^1, \ldots, b_{ip}^{|B_p|}\right), \tag{3.142}$$

natomiast:

$$a_{ip}^j = \begin{cases} N^{-1}, & x_{ip} \leqslant L_{x_p,r}(Q_{x_p,\alpha_j^p}), \\ N^{-1}\beta_{x_p,r}(Q_{x_p,\alpha_j^p}), & x_{ip} = U_{x_p,r}(Q_{x_p,\alpha_j^p}), \quad j \in A_p, \\ 0, & x_{ip} > U_{x_p,r}(Q_{x_p,\alpha_j^p}), \end{cases} \tag{3.143}$$

$$b_{ip}^j = \begin{cases} N^{-1}, & x_{ip} \leqslant L_{x_p,r}(\hat{Q}_{x_p,\beta_j^p}), \\ N^{-1}\beta_{x_p,r}(\hat{Q}_{x_p,\beta_j^p}), & x_{ip} = U_{x_p,r}(\hat{Q}_{x_p,\beta_j^p}), \quad j \in B_p, \\ 0, & x_{ip} > U_{x_p,r}(\hat{Q}_{x_p,\beta_j^p}). \end{cases} \tag{3.144}$$

dla $i = 1, \ldots, m$, $p = 1, \ldots, k$.

Dowód. Wagi kalibracyjne w_i dla $i = 1, \ldots, m$ znajdziemy, korzystając z metody czynników nieoznaczonych Lagrange'a.
Funkcja Lagrange'a ma postać:

$$L = \frac{1}{2}\sum_{i=1}^m \frac{(w_i - d_i)^2}{d_i} + \lambda_0\left(N - \sum_{i=1}^m w_i\right) + \sum_{p=1}^k \sum_{j\in A_p} \lambda_j^p\left(\alpha_j^p - \hat{F}_{x_p,cal}\left(Q_{x_p,\alpha_j^p}\right)\right)$$
$$+ \sum_{p=1}^k \sum_{j\in B_p} \gamma_j^p\left(\beta_j^p - \hat{F}_{x_p,cal}\left(\hat{Q}_{x_p,\beta_j^p}\right)\right). \tag{3.145}$$

Z równania kalibracyjnego (3.135) wynika, że:

$$\hat{F}_{x_p,cal}\left(Q_{x_p,\alpha_j^p}\right) = \frac{\sum_{i=1}^m w_i H_{x_p,r}\left(Q_{x_p,\alpha_j^p}, x_{ip}\right)}{\sum_{i=1}^m w_i} = \frac{\sum_{i=1}^m w_i H_{x_p,r}\left(Q_{x_p,\alpha_j^p}, x_{ip}\right)}{N} \tag{3.146}$$

oraz

$$\hat{F}_{x_p,cal}\left(\hat{Q}_{x_p,\beta_j^p}\right) = \frac{\sum_{i=1}^m w_i H_{x_p,r}\left(\hat{Q}_{x_p,\beta_j^p}, x_{ip}\right)}{\sum_{i=1}^m w_i} = \frac{\sum_{i=1}^m w_i H_{x_p,r}\left(\hat{Q}_{x_p,\beta_j^p}, x_{ip}\right)}{N}. \tag{3.147}$$

gdzie $p = 1, \ldots, k$, $j \in A_p$, $j \in B_p$. Funkcję Lagrange'a można więc przedstawić w postaci:

$$\begin{aligned} L = &\frac{1}{2}\sum_{i=1}^{m}\frac{(w_i - d_i)^2}{d_i} + \lambda_0\left(N - \sum_{i=1}^{m} w_i\right) + \\ &+ \sum_{p=1}^{k}\sum_{j\in A_p}\lambda_j^p\left(\alpha_j^p - \frac{\sum_{i=1}^{m} w_i H_{x_p,r}\left(Q_{x_p,\alpha_j^p}, x_{ip}\right)}{N}\right) + \\ &+ \sum_{p=1}^{k}\sum_{j\in B_p}\gamma_j^p\left(\beta_j^p - \frac{\sum_{i=1}^{m} w_i H_{x_p,r}\left(\hat{Q}_{x_p,\beta_j^p}, x_{ip}\right)}{N}\right). \end{aligned} \tag{3.148}$$

Pochodna funkcji L dla $i = 1, \ldots, m$ ma postać:

$$\frac{\partial L}{\partial w_i} = \frac{1}{2}\cdot\frac{2w_i - 2d_i}{d_i} - \lambda_0 - \sum_{p=1}^{k}\sum_{j\in A_p}\lambda_j^p\frac{H_{x_p,r}\left(Q_{x_p,\alpha_j^p}, x_{ip}\right)}{N} - \sum_{p=1}^{k}\sum_{j\in B_p}\gamma_j^p\frac{H_{x_p,r}\left(\hat{Q}_{x_p,\beta_j^p}, x_{ij}\right)}{N}. \tag{3.149}$$

Ponieważ:

$$H_{x_p,r}\left(Q_{x_p,\alpha_j^p}, x_{ip}\right) = \begin{cases} 1, & x_{ip} \leqslant L_{x_p,r}(Q_{x_p,\alpha_j^p}), \\ \beta_{x_p,r}(Q_{x_p,\alpha_j^p}), & x_{ip} = U_{x_p,r}(Q_{x_p,\alpha_j^p}), \\ 0, & x_{ip} > U_{x_p,r}(Q_{x_p,\alpha_j^p}), \end{cases} \tag{3.150}$$

oraz

$$H_{x_p,r}\left(\hat{Q}_{x_p,\beta_j^p}, x_{ip}\right) = \begin{cases} 1, & x_{ip} \leqslant L_{x_p,r}(\hat{Q}_{x_p,\beta_j^p}), \\ \beta_{x_p,r}(\hat{Q}_{x_p,\beta_j^p}), & x_{ip} = U_{x_p,r}(\hat{Q}_{x_p,\beta_j^p}), \\ 0, & x_{ip} > U_{x_p,r}(\hat{Q}_{x_p,\beta_j^p}), \end{cases} \tag{3.151}$$

to przyjmując, że:

$$a_{ip}^j = \frac{H_{x_p,r}\left(Q_{x_p,\alpha_j^p}, x_{ip}\right)}{N} \quad \text{oraz} \quad b_{ip}^j = \frac{H_{x_p,r}\left(\hat{Q}_{x_p,\beta_j^p}, x_{ip}\right)}{N} \tag{3.152}$$

otrzymujemy:

$$a_{ip}^j = \begin{cases} N^{-1}, & x_{ip} \leqslant L_{x_p,r}(Q_{x_p,\alpha_j^p}), \\ N^{-1}\beta_{x_p,r}(Q_{x_p,\alpha_j^p}), & x_{ip} = U_{x_p,r}(Q_{x_p,\alpha_j^p}), \\ 0, & x_{ip} > U_{x_p,r}(Q_{x_p,\alpha_j^p}), \end{cases} \tag{3.153}$$

$$b_{ip}^j = \begin{cases} N^{-1}, & x_{ip} \leqslant L_{x_p,r}(\hat{Q}_{x_p,\beta_j^p}), \\ N^{-1}\beta_{x_p,r}(\hat{Q}_{x_p,\beta_j^p}), & x_{ip} = U_{x_p,r}(\hat{Q}_{x_p,\beta_j^p}), \\ 0, & x_{ip} > U_{x_p,r}(\hat{Q}_{x_p,\beta_j^p}). \end{cases} \tag{3.154}$$

Pochodną funkcji L można więc przedstawić w następującej postaci:

$$\frac{\partial L}{\partial w_i} = \frac{w_i - d_i}{d_i} - \lambda_0 - \sum_{p=1}^{k}\sum_{j\in A_p}\lambda_j^p a_{ip}^j - \sum_{p=1}^{k}\sum_{j\in B_p}\gamma_j^p b_{ip}^j. \tag{3.155}$$

Przyjmując, że:

$$\underbrace{c_{i0}=1,\underbrace{c_{i1}=a_{i1}^{1},\ldots,c_{il_1}=a_{ik}^{|A_k|}}_{l_1 \text{ składników}},\underbrace{c_{il_1+1}=b_{i1}^{1},\ldots,c_{il}=b_{ik}^{|B_k|}}_{l_2 \text{ składników}}}_{l+1 \text{ składników}} \tag{3.156}$$

oraz $\lambda_1=\lambda_1^1,\ldots,\lambda_{l_1}=\lambda_{|A_k|}^k,\lambda_{l_1+1}=\gamma_1^1,\ldots,\lambda_l=\gamma_{|B_k|}^k$ dla $i=1,\ldots,m$, otrzymujemy, że:

$$\frac{\partial L}{\partial w_i}=\frac{w_i-d_i}{d_i}-\sum_{j=0}^{l}\lambda_j c_{ij}. \tag{3.157}$$

Przyjmując ponadto, że:

$$\mathbf{c}_i=(1,c_{i1},\ldots,c_{il})^T=(1,\mathbf{a}_{i1},\ldots,\mathbf{a}_{ik},\mathbf{b}_{i1},\ldots,\mathbf{b}_{ik})^T, \tag{3.158}$$

gdzie wektory $\mathbf{a}_{ip}$ i $\mathbf{b}_{ip}$ dla $p=1,\ldots,k$ określone są jako:

$$\mathbf{a}_{ip}=\left(a_{ip}^1,\ldots,a_{ip}^{|A_p|}\right), \tag{3.159}$$

$$\mathbf{b}_{ip}=\left(b_{ip}^1,\ldots,b_{ip}^{|B_p|}\right), \tag{3.160}$$

oraz $\boldsymbol{\lambda}=(\lambda_0,\ldots,\lambda_l)^T$ dla $i=1,\ldots,m$, otrzymujemy, że:

$$w_i=d_i\left(1+\boldsymbol{\lambda}^T\mathbf{c}_i\right)=d_i\left(1+\mathbf{c}_i^T\boldsymbol{\lambda}\right). \tag{3.161}$$

Jest to konsekwencją faktu, że:

$$\sum_{j=0}^{l}\lambda_j c_{ij}=\boldsymbol{\lambda}^T\mathbf{c}_i=\mathbf{c}_i^T\boldsymbol{\lambda}. \tag{3.162}$$

Mnożąc obustronnie równanie (3.161) przez $\mathbf{c}_i$, a następnie dokonując sumowania po zbiorze wszystkich respondentów r otrzymujemy:

$$\sum_{i=1}^{m}w_i\mathbf{c}_i=\sum_{i=1}^{m}d_i\mathbf{c}_i\left(1+\mathbf{c}_i^T\boldsymbol{\lambda}\right). \tag{3.163}$$

Dokonując przekształcenia ostatniego równania otrzymujemy:

$$\sum_{i=1}^{m}w_i\mathbf{c}_i-\sum_{i=1}^{m}d_i\mathbf{c}_i=\sum_{i=1}^{m}d_i\mathbf{c}_i\mathbf{c}_i^T\boldsymbol{\lambda}. \tag{3.164}$$

Zwróćmy uwagę, że:

$$\begin{aligned}
\sum_{i=1}^{m}w_i\mathbf{c}_i&=\left(\sum_{i=1}^{m}w_i,\sum_{i=1}^{m}w_ic_{i1},\ldots,\sum_{i=1}^{m}w_ic_{il}\right)^T=\\
&=\left(N,\sum_{i=1}^{m}w_ia_{i1}^1,\ldots,\sum_{i=1}^{m}w_ia_{ik}^{|A_k|},\sum_{i=1}^{m}w_ib_{i1}^1,\ldots,\sum_{i=1}^{m}w_ib_{ik}^{|B_k|}\right)^T=\\
&=\left(N,\hat{F}_{x_1,cal}\left(Q_{x_1,\alpha_1^1}\right),\ldots,\hat{F}_{x_k,cal}\left(\hat{Q}_{x_k,\beta_{|B_k|}^k},\right)\right)^T=\\
&=(N,\underbrace{\alpha_1^1,\ldots,\beta_{|B_k|}^k}_{l})^T=\mathbf{T}_{\mathbf{c}}.
\end{aligned} \tag{3.165}$$

Zakładając, że macierz $\sum_{i=1}^{m} d_i \mathbf{c}_i \mathbf{c}_i^T$ jest nieosobliwa, otrzymujemy następującą postać wektora $\boldsymbol{\lambda}$:

$$\boldsymbol{\lambda} = \left(\sum_{i=1}^{m} d_i \mathbf{c}_i \mathbf{c}_i^T \right)^{-1} \left(\mathbf{T}_{\mathbf{c}} - \sum_{i=1}^{m} d_i \mathbf{c}_i \right). \tag{3.166}$$

Stąd, korzystając z równania (3.161) i wyznaczonej postaci wektora $\boldsymbol{\lambda}$ otrzymujemy, że:

$$w_i = d_i \left(1 + \mathbf{c}_i^T \boldsymbol{\lambda} \right) = d_i + d_i \mathbf{c}_i^T \left(\sum_{i=1}^{m} d_i \mathbf{c}_i \mathbf{c}_i^T \right)^{-1} \left(\mathbf{T}_{\mathbf{c}} - \sum_{i=1}^{m} d_i \mathbf{c}_i \right), \tag{3.167}$$

a więc, poszukiwaną postać składowej wektora wag kalibracyjnych $\mathbf{w} = (w_1, \ldots, w_m)^T$.

Należy jeszcze sprawdzić, że w punkcie $\mathbf{w} = (w_1, \ldots, w_m)^T$ istnieje minimum (warunek dostateczny istnienia ekstremum warunkowego). Niech ξ będzie niezerowym wektorem takim, że $\xi \in \mathbb{R}^m$. Należy wykazać, że forma kwadratowa $d^2 L(\mathbf{w})(\xi)$ jest dodatnio określona. Mamy:

$$d^2 L(\mathbf{w})(\xi) = \sum_{i,j=1}^{m} \frac{\partial^2 L}{\partial w_i \partial w_j} \xi_i \xi_j. \tag{3.168}$$

Zauważmy, że:

$$\frac{\partial^2 L}{\partial w_i w_j} = \begin{cases} \pi_i \text{ dla } i = j, \\ 0 \text{ dla } i \neq j. \end{cases} \tag{3.169}$$

Podstawiając obliczone pochodne drugiego rzędu do formy kwadratowej (3.168) otrzymujemy, że:

$$d^2 L(\mathbf{w})(\xi) = \sum_{i,j=1}^{m} \frac{\partial^2 L}{\partial w_i \partial w_j} \xi_i \xi_j = \sum_{i=1}^{m} \pi_i \xi_i^2. \tag{3.170}$$

Jest to oczywiście forma kwadratowa dodatnio określona. Stąd wektor wag kalibracyjnych $\mathbf{w} = (w_1, \ldots, w_m)^T$, którego składowe spełniają warunek (3.136) jest poszukiwanym rozwiązaniem zadania minimalizacji funkcji odległości. ■

3.6. Wnioski

W rozdziale trzecim opisaliśmy metody konstrukcji estymatorów kalibracyjnych kwantyla rzędu α zmiennej y przy założeniu, że w badaniu występują braki odpowiedzi. Przedstawiliśmy własne propozycje wyznaczania wag kalibracyjnych w różnych sytuacjach tj. gdy znane są tylko kwantyle rzędu α wszystkich zmiennych pomocniczych w populacji bądź ich oszacowania na podstawie próby. Podsumowaniem rozdziału, dotyczącym rozważań teoretycznych nad estymatorami kalibracyjnymi kwantyli, było twierdzenie o postaci wag uogólnionego estymatora kalibracyjnego kwantyla rzędu α, w którym założono, że dysponujemy informacjami o kwantylach dowolnego rzędu zmiennych pomocniczych, zarówno na poziomie populacji jak i próby.

Praktyczne wykorzystanie estymatorów kalibracyjnych jest uzależnione od dostępności informacji o cechach pomocniczych i stopnia ich skorelowania z badaną zmienną. Wybór odpowiedniego estymatora jest ponadto uzależniony od jego wariancji i obciążenia. Ponieważ na drodze analitycznej, podobnie jak w przypadku estymatorów kalibracyjnych wartości globalnej, jest w zasadzie niemożliwe ustalenie, który estymator

charakteryzuje się najmniejszym obciążeniem i wariancją, dlatego bardzo pomocne, przy podjęciu decyzji o wyborze konkretnego estymatora, mogą okazać się wyniki badań symulacyjnych. W badaniach tych, korzystając z danych rzeczywistych bądź sztucznie wygenerowanych, można określić najważniejsze własności rozważanych estymatorów, co w konsekwencji może ułatwić wskazanie najodpowiedniejszego estymatora w praktycznych zastosowaniach.

Dalszym rozwinięciem wniosków płynących, z rodziału trzeciego, będzie rozdział czwarty poświęcony zbadaniu najważniejszych własności estymatorów kalibracyjnych mediany zmiennej y. Omówione zostaną wybrane charakterystyki rozważanych estymatorów kalibracyjnych, w oparciu o przeprowadzone badania symulacyjne, a także sformułowane zostaną wskazówki co do ich praktycznego wykorzystania.

Dostępność informacji o zmiennych pomocniczych oraz wnioski płynące z rozdziału czwartego, stanowić z kolei będą punkt wyjścia do empirycznej oceny estymatorów kalibracyjnych mediany w badaniu budżetów gospodarstw domowych. Będzie to przedmiotem zainteresowania w ostatnim rozdziale pracy.

Rozdział 4

Empiryczna ocena skutków braków odpowiedzi

4.1. Założenia badania symulacyjnego dla średniej

Praktyczne wykorzystanie estymatorów, w różnego rodzaju badaniach, jest zazwyczaj poprzedzone oceną ich własności na drodze eksperymentów symulacyjnych. Ponieważ dla wielu estymatorów kalibracyjnych, trudno jest wyznaczyć wartość wariancji czy obciążenia, miary te szacuje się na drodze badań symulacyjnych, co może przyczynić się do wskazania najlepszych estymatorów w praktycznych zastosowaniach. W rozdziale czwartym podjęta zostanie próba zbadania własności estymatorów kalibracyjnych dla średniej oraz kwantyla rzędu α, które w ujęciu teoretycznym omówiono szczegółowo w poprzednich rozdziałach pracy. W przypadku kwantyla zakładać będziemy, że $\alpha = 0,5$, tj. ocenie podlegać będzie mediana zmienej Y. Otrzymane wyniki porównane zostaną z ocenami uzyskanymi z wykorzystaniem estymatora Horvitza-Thompsona dla średniej i mediany odpowiednio.

W przypadku szacowania średniej cechy Y, analizy dokonano w oparciu o rzeczywiste dane pochodzące z NSP'2002. Rozpatrywane były dwie cechy: powierzchnia użytkowa mieszkania (Y) oraz liczba izb w mieszkaniu (X). Szacowaną wartością była przeciętna użytkowa powierzchnia mieszkania w przekroju powiatów województwa wielkopolskiego. Jako zmienną pomocniczą przyjęto liczbę izb w mieszkaniu, za czym przemawiała silna dodatnia zależność między tymi cechami. Ocena współczynnika korelacji liniowej Pearsona między powierzchnią użytkową, a liczbą izb w mieszkaniu w przekroju powiatów wynosiła od $\rho_{XY} = 0,84$ do $\rho_{XY} = 0,9$. Populację generalną stanowiły wszystkie mieszkania w ramach danego powiatu województwa wielkopolskiego. Na potrzeby oszacowania średniej powierzchni mieszkań, wykorzystano estymator bezpośredni i wybrane estymatory kalibracyjne.

- estymator Horvitza Thompsona

$$\hat{\bar{Y}}_{HT} = \frac{\sum_{i=1}^{m} d_i y_i}{\sum_{i=1}^{m} d_i}. \tag{4.1}$$

Z kolei w odniesieniu do estymatorów kalibracyjnych wybrano do dalszej analizy trzy z nich, przy czym w pierwszej kolejności dokonano oszacowania wartości globalnej powierzchni, a następnie wyznaczono średnią powierzchnię użytkową mieszkań w przekroju powiatów.

Na potrzeby oszacowania wartości globalnej powierzchni mieszkań w danym powiecie wykorzystano następujące estymatory kalibracyjne:

- estymator kalibracyjny ze znanym wektorem $\mathbf{X}$ wyrażający się wzorem (2.10).

$$\hat{Y}_{\mathbf{X}} = \sum_{i=1}^{m} w_i y_i, \tag{4.2}$$

gdzie wagi kalibracyjne są postaci:

$$w_i = d_i + d_i \left(\mathbf{X} - \hat{\mathbf{X}}\right)^T \left(\sum_{i=1}^{m} d_i \boldsymbol{x}_i \boldsymbol{x}_i^T\right)^{-1} \boldsymbol{x}_i. \tag{4.3}$$

Przyjęto przy tym, że jedyną znaną wartością globalną – stanowiącą składową wektora $\mathbf{X}$ – jest łączna liczba izb we wszystkich mieszkaniach w ramach powiatu. Wektor wartości globalnych $\mathbf{X}$ był zatem postaci:

$$\mathbf{X} = \sum_{i=1}^{N_j} x_i^j, \tag{4.4}$$

gdzie x_i^j oznaczało liczbę izb w i – tym mieszkaniu w j – tym powiecie województwa wielkopolskiego, a N_j liczbę wszystkich mieszkań w tym powiecie, $(j = 1, \ldots, 35)$.

- estymator kalibracyjny ze znanym wektorem $\breve{\mathbf{X}}$ wyrażający się wzorem (2.34).

$$\hat{Y}_{\breve{\mathbf{X}}} = \sum_{i=1}^{m} w_i y_i, \tag{4.5}$$

gdzie wagi kalibracyjne są postaci:

$$w_i = d_i + d_i \left(\breve{\mathbf{X}} - \hat{\mathbf{X}}\right)^T \left(\sum_{i=1}^{m} d_i \boldsymbol{x}_i \boldsymbol{x}_i^T\right)^{-1} \boldsymbol{x}_i. \tag{4.6}$$

W tym przypadku zakładano, że nie jest znana łączna liczba izb we wszystkich mieszkaniach w ramach danego powiatu, a jedynie jej oszacowanie na podstawie próby s. Jedyną składową wektora $\breve{\mathbf{X}}$ stanowiła oszacowana w oparciu o estymator Horvitza-Thompsona – łączna liczba izb w powiecie.

- estymator kalibracyjny regresyjny, który jest szczególnym przypadkiem estymatora kalibracyjnego ze znanym wektorem $\mathbf{X}$ (por. wzór. 2.10 oraz 2.97).

$$\hat{Y}_{REG} = \sum_{i=1}^{m} w_i y_i. \tag{4.7}$$

  Dla tego estymatora przyjęto, że wektor wartości globalnych $\mathbf{X}$ był postaci:

$$\mathbf{X} = \left(\sum_{i=1}^{N_j} x_i^j, N_j \right)^T. \tag{4.8}$$

  W konsekwencji, dla estymatora kalibracyjnego regresyjnego założono, że wektor zmiennych pomocniczych można zapisać jako:

$$\mathbf{x}_i = \left(x_i^j, 1 \right)^T. \tag{4.9}$$

  Dodatkowe dołączenie sztucznej zmiennej pomocniczej, przyjmującej wartość 1 dla każdego mieszkania w ramach danego powiatu (a zatem, również dla mieszkania wylosowanego do próby) powodowało, że wagi kalibracyjne sumowały się do łącznej liczby wszystkich mieszkań w ramach powiatu[31].

Na potrzeby badań symulacyjnych losowano z każdego powiatu próby 1%, 10% i 20% mieszkań. Zastosowano prz tym schemat losowania warstwowego z alokacją proporcjonalną do pierwiastka drugiego stopnia z liczby mieszkań w powiatach, które stanowiły warstwy[32]. Następnie po wylosowaniu n_j – elementowej próby mieszkań dla j – tego powiatu, zastępowano część informacji o powierzchni mieszkań (zmienna Y) brakami danych. W efekcie po takiej operacji, dysponowaliśmy w próbie informacją o liczbie izb dla każdego wylosowanego mieszkania, a tylko dla części mieszkań posiadaliśmy dane o ich powierzchni. Zastosowano przy tym trzy różne podejścia dla tworzenia

[31] Ponieważ w dalszym ciągu pracy mowa będzie o estymatorach kalibracyjnych średniej, wprowadzone zostaną ich następujące oznaczenia: $\hat{\bar{Y}}_{\mathbf{X}}$ dla estymatora kalibracyjnego średniej ze znanym wektorem $\mathbf{X}$, $\hat{\bar{Y}}_{\breve{\mathbf{X}}}$, dla estymatora kalibracyjnego średniej ze znanym wektorem $\breve{\mathbf{X}}$ i $\hat{\bar{Y}}_{REG}$, dla estymatora kalibracyjnego regresyjnego średniej.

[32] W badaniu budżetów gospodarstw domowych stosuje się terytorialny, warstwowy, dwustopniowy schemat losowania próby z różnymi prawdopodobieństwami wyboru na I stopniu. Jest to zatem bardziej skomplikowane podejście aniżeli rozważane w pracy losowanie warstwowe. Wybór takiego schematu losowania próby podyktowany był faktem, że wszystkie obliczenia możliwe były do przeprowadzenia w ramach prac podgrupy ds. statystyczno-matematycznych w Urzędzie Statystycznym w Poznaniu. W jego siedzibie znajdowały się tylko dwa stanowiska komputerowe z systemem SAS. Dostęp do danych spisowych umieszczonych na serwerze Głównego Urzędu Statystycznego był natomiast możliwy tylko online. Stanowiło to istotne „wąskie gardło" uniemożliwiające efektywne wykonywanie badań symulacyjnych, które nawet w schemacie losowania prostego, były czasochłonne i nie zawsze ze względu na ograniczoną pojemność dysków sieciowych, możliwe do przeprowadzenia, nawet dla małej liczby replikacji. Poza tym, pionierski charakter prowadzonych prac nad własnościami estymatorów kalibracyjnych w drodze badań symulacyjnych, które zazwyczaj przeprowadza się w pierwszej kolejności na prostszych schematach losowania próby, usprawiedliwia przyjęty w pracy schemat losowania warstwowego będącego pewnym uproszczeniem tego, który wykorzystywany jest w badaniu budżetów gospodarstw domowych. Przeprowadzenie badań symulacyjnych, dla bardziej złożonych schematów losowania próby, będzie jednak przedmiotem dalszych prac analitycznych w ramach przygotowań do NSP'2011.

braków danych. W pierwszym zakładano, że braki danych generowane były w sposób losowy (wariant 1). W drugim (wariant 2) i trzecim (wariant 3), założono, że braki danych przypisano mieszkaniom o najmniejszej i największej powierzchni odpowiednio. Przyjęto przy tym, że w każdym przypadku frakcja braków danych w próbie wynosiła: 1%, 10% i 20%. Dla różnych wariantów liczebności próby (trzy możliwości), frakcji braków (trzy możliwości) oraz sposobów ich generowania (trzy możliwości) przeprowadzono po 500 replikacji dla każdego powiatu i na tej podstawie dokonano oszacowania wartości oczekiwanej średniej powierzchni mieszkań w powiecie, wartości oczekiwanej obciążenia badanych estymatorów, ich wariancji empirycznej oraz względnych błędów szacunku. W sumie łączna liczba przeprowadzonych replikacji w różnych układach wynosiła 472 500.

Po przeprowadzeniu badania symulacyjnego dla każdego z testowanych estymatorów obliczono przybliżoną wartość oczekiwaną średniej powierzchni mieszkań, wartość oczekiwaną obciążenia, wariancje empiryczne oraz względne błędy szacunków. Poszczególne miary dla wszystkich testowanych estymatorów policzono z następujących wzorów:

- wartość oczekiwana średniej powierzchni mieszkań w j – tym powiecie

$$\hat{\bar{Y}}_j = \frac{1}{500}\sum_{i=1}^{500}\bar{Y}_{ji}, \tag{4.10}$$

  gdzie $\bar{Y}_{ji}$ oznacza ocenę średniej powierzchni mieszkań w j – tym powiecie w i – tej replikacji uzyskaną z wykorzystaniem odpowiedniego estymatora, ($j = 1, \ldots, 35$, $i = 1, \ldots, 500$).
- wartość oczekiwana obciążenia

$$B_j = \frac{1}{500}\sum_{i=1}^{500}\left|\bar{Y}_{ji} - \bar{Y}_j\right|, \tag{4.11}$$

  gdzie $\bar{Y}_j$ oznacza średnią powierzchnię mieszkań w j – tym powiecie.
- wariancja empiryczna

$$\hat{V}_j^2 = \frac{1}{499}\sum_{i=1}^{500}\left(\bar{Y}_{ji} - \hat{\bar{Y}}_j\right)^2. \tag{4.12}$$

- względny błąd szacunku

$$REE_j = \frac{\sqrt{\hat{V}_j^2}}{\hat{\bar{Y}}_j}\cdot 100\%. \tag{4.13}$$

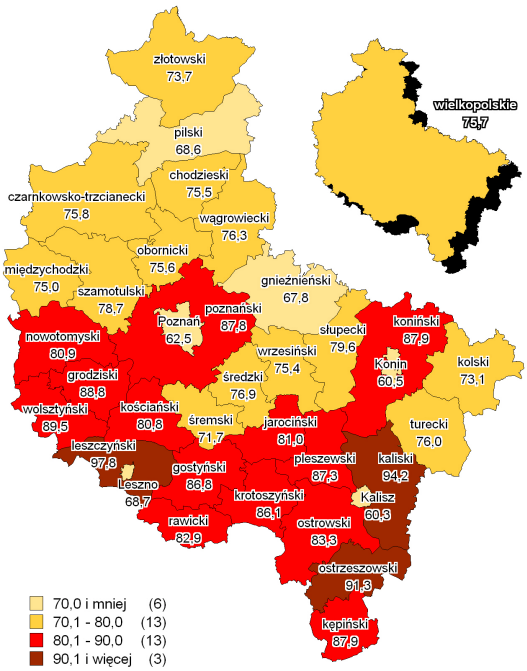
4.2. Wyniki badania symulacyjnego dla średniej

Zgodnie z opisaną w poprzednim podrozdziale procedurą, przeprowadzono badania symulacyjne celem oceny własności wybranych estymatorów kalibracyjnych[33]. Miary jakości estymacji wyznaczone zostały dla każdego powiatu oddzielnie, przy czym szczegółowe wyniki obliczeń dla różnych wielkości prób, frakcji braków danych i wariantów

[33] Na potrzeby prowadzonych badań symulacyjnych nad jakością rozważanych estymatorów został napisany specjalny program w języku 4GL w systemie SAS, którego kod jest dostępny w Urzędzie Statystycznym w Poznaniu.

ich generowania – ze względu na dużą liczbę tabel zaprezentowane będą tylko w odniesieniu do powiatu chodzieskiego[34]. Z kolei na wykresach mapowych przedstawione zostaną wyniki przeprowadzonych badań symulacyjnych dla wszystkich powiatów województwa wielkopolskiego[35].

Zgodnie z wynikami NSP'2002 średnia powierzchnia mieszkań w województwie wielkopolskim wynosiła 75,7m^2, przy czym w miastach była równa 67,3m^2, a na wsi 90,6m^2. Średnia liczba izb we wszystkich mieszkaniach wynosiła z kolei 3,95, przy czym średnia ta była większa dla mieszkań na wsi w porównaniu ze średnią liczbą izb w mieszkaniach zlokalizowanych w miastach[36].



Rysunek. 4.1. Średnia powierzchnia mieszkań (w m^2) w przekroju powiatów województwa wielkopolskiego w 2002r.

Źródło: Opracowanie własne na podstawie danych z NSP'2002

[34] Dokonując kompleksowej analizy uzyskanych wyników dla pozostałych powiatów można wyciągnąć bardzo podobne wnioski, co dodatkowo uzasadniało zamieszczenie w pracy szczegółowych wyników tylko dla jednego powiatu.

[35] Ze względu na dużą liczbę kombinacji, w tekście rozdziału 4 zamieszczone zostaną wyniki badań symulacyjnych dla 1% próby, z której generowano 10% braków danych zgodnie z wariantem 2. Dodatek A pracy, zawiera wykresy mapowe dla innych wariantów. Ograniczono się przy tym, celem uzupełnienia, do przypadku, gdy wielkość próby wynosiła 1%, frakcja braków danych 10%, a braki generowano zgodnie z wariantem 1 i 3 (wykresy mapowe dla wariantu 2 zaprezentowane zostały w tekście rozdziału 4). Ponadto zamieszczono mapy dla wszystkich trzech wariantów, gdy wielkość próby wynosiła 10%, a frakcja braków danych 20%.

[36] Rozkład średniej liczby izb w mieszkaniach w przekroju powiatów województwa wielkopolskiego przedstawia rysunek A.1.

Na rysunku 4.1 zauważalne jest zróżnicowanie powierzchni użytkowej mieszkań w przekroju powiatów województwa wielkopolskiego. Najmniejsza jest średnia powierzchnia mieszkań w miastach na prawach powiatu (Poznań, Konin, Leszno i Piła). W północnej i północno-wschodniej części województwa średnia ta dla wielu powiatów znajduje się w przedziale 70–80m^2. Natomiast dla powiatów zlokalizowanych w południowej części średnia powierzchnia użytkowa mieszkań przekracza 80m^2, a w przypadku trzech powiatów (leszczyńskiego, kaliskiego i ostrzeszowskiego) 90m^2. Dla powiatu chodzieskiego, dla którego prezentowane będą szczegółowe wyniki przeprowadzonych symulacji, średnia powierzchnia użytkowa mieszkań wyliczona w oparciu o dane z NSP'2002 wynosiła 75,5m^2.

Tabela. 4.1. Wartość oczekiwana estymatorów średniej powierzchni mieszkań w powiecie chodzieskim (w m^2)

Powiat chodzieski		Wielkość próby								
		1%			10%			20%		
		Frakcja braków			Frakcja braków			Frakcja braków		
		1%	10%	20%	1%	10%	20%	1%	10%	20%
Wariant 1	$\bar{Y}$	75.50	75.50	75.50	75.50	75.50	75.50	75.50	75.50	75.50
	$\hat{\bar{Y}}_{HT}$	75.01	75.37	75.98	75.47	75.48	75.62	75.54	75.48	75.45
	$\hat{\bar{Y}}_{\tilde{\mathbf{X}}}$	75.04	75.73	76.37	75.49	75.70	76.00	75.57	75.69	75.85
	$\hat{\bar{Y}}_{\mathbf{X}}$	75.57	75.89	76.02	75.54	75.76	75.96	75.52	75.74	75.93
	$\hat{\bar{Y}}_{REG}$	75.66	75.79	75.45	75.52	75.56	75.48	75.49	75.52	75.51
Wariant 2	$\bar{Y}$	75.50	75.50	75.50	75.50	75.50	75.50	75.50	75.50	75.50
	$\hat{\bar{Y}}_{HT}$	75.71	80.24	86.05	75.98	80.29	85.19	76.02	80.33	85.31
	$\hat{\bar{Y}}_{\tilde{\mathbf{X}}}$	75.43	76.54	79.15	75.56	76.52	78.24	75.59	76.56	78.33
	$\hat{\bar{Y}}_{\mathbf{X}}$	75.52	76.52	78.74	75.67	76.59	78.16	75.71	76.60	78.29
	$\hat{\bar{Y}}_{REG}$	75.43	75.54	76.84	75.57	75.62	76.57	75.62	75.64	76.69
Wariant 3	$\bar{Y}$	75.50	75.50	75.50	75.50	75.50	75.50	75.50	75.50	75.50
	$\hat{\bar{Y}}_{HT}$	74.40	64.30	58.57	73.26	63.91	58.21	73.46	64.23	58.21
	$\hat{\bar{Y}}_{\tilde{\mathbf{X}}}$	74.93	68.99	66.66	74.00	68.77	66.35	74.21	69.16	66.37
	$\hat{\bar{Y}}_{\mathbf{X}}$	74.68	69.16	66.30	74.18	68.94	66.40	74.20	69.14	66.34
	$\hat{\bar{Y}}_{REG}$	74.71	68.82	65.06	74.29	68.60	65.05	74.29	68.82	65.01

Źródło: Opracowanie własne na podstawie wyników badań symulacyjnych

Gdy braki danych dla powierzchni mieszkań generowane były w sposób losowy (wariant 1), to bez względu na liczebność próby oraz frakcję braków, zarówno wartość oczekiwana estymatora Horvitza-Thompsona jak i estymatorów kalibracyjnych średniej powierzchni mieszkań były bliskie wartości rzeczywistej średniego metrażu, por. tabela 4.1 oraz rysunek A.2 i A.10. Estymatory te charakteryzowały się niewielkimi obciążeniami. Analizując obciążenie estymatorów można jednak zauważyć, że znajomość wartości globalnej zmiennej pomocniczej – łącznej liczby izb we wszystkich mieszkaniach w powiecie – przyczyniła się do jego znacznej redukcji. Sytuacja ta dotyczy estymatora kalibracyjnego ze znanym wektorem $\mathbf{X}$, a zwłaszcza estymatora kalibracyjnego regresyjnego, w którym wagi kalibracyjne sumowały się do liczby wszystkich mieszkań w ramach powiatu, por. tabela 4.2.

Również w przekroju powiatów województwa wielkopolskiego, gdy zastosowano podejście randomizacyjne w procesie tworzenia braków danych, nie zauważono istotnych różnic między rzeczywistą średnią powierzchnią mieszkania, a jej wartością oczekiwaną dla wszystkich rozważanych estymatorów. Najmniejszym obciążeniem w rozpatrywanej klasie estymatorów charakteryzował się estymator kalibracyjny regresyjny, por. rysunek A.3 i A.11.

Tabela. 4.2. Wartość oczekiwana obciążenia średniej powierzchni mieszkań w powiecie chodzieskim (w m^2)

Powiat chodzieski		Wielkość próby								
		1%			10%			20%		
		Frakcja braków			Frakcja braków			Frakcja braków		
		1%	10%	20%	1%	10%	20%	1%	10%	20%
Wariant 1	$\hat{\bar{Y}}_{HT}$	3.19	2.63	3.05	0.93	0.80	1.15	0.61	0.59	0.65
	$\hat{\bar{Y}}_{\mathbf{X}}$	3.20	2.58	3.15	0.91	0.85	1.15	0.60	0.58	0.65
	$\hat{\bar{Y}}_{\mathbf{X}}$	1.84	1.75	2.08	0.55	0.60	0.79	0.34	0.40	0.53
	$\hat{\bar{Y}}_{REG}$	1.83	1.72	1.87	0.52	0.54	0.64	0.31	0.33	0.38
Wariant 2	$\hat{\bar{Y}}_{HT}$	2.56	5.02	10.65	1.02	4.79	9.69	0.70	4.83	9.81
	$\hat{\bar{Y}}_{\mathbf{X}}$	2.50	2.74	4.66	0.89	1.24	2.75	0.56	1.11	2.83
	$\hat{\bar{Y}}_{\mathbf{X}}$	1.64	1.96	3.52	0.55	1.13	2.66	0.36	1.10	2.79
	$\hat{\bar{Y}}_{REG}$	1.60	1.65	2.16	0.51	0.50	1.10	0.33	0.36	1.19
Wariant 3	$\hat{\bar{Y}}_{HT}$	2.98	11.20	16.93	2.24	11.59	17.29	2.04	11.27	17.29
	$\hat{\bar{Y}}_{\mathbf{X}}$	2.92	6.54	8.84	1.55	6.73	9.15	1.31	6.34	9.13
	$\hat{\bar{Y}}_{\mathbf{X}}$	1.87	6.34	9.20	1.33	6.56	9.10	1.30	6.36	9.16
	$\hat{\bar{Y}}_{REG}$	1.85	6.68	10.45	1.23	6.90	10.46	1.21	6.68	10.49

Źródło: Opracowanie własne na podstawie wyników badań symulacyjnych

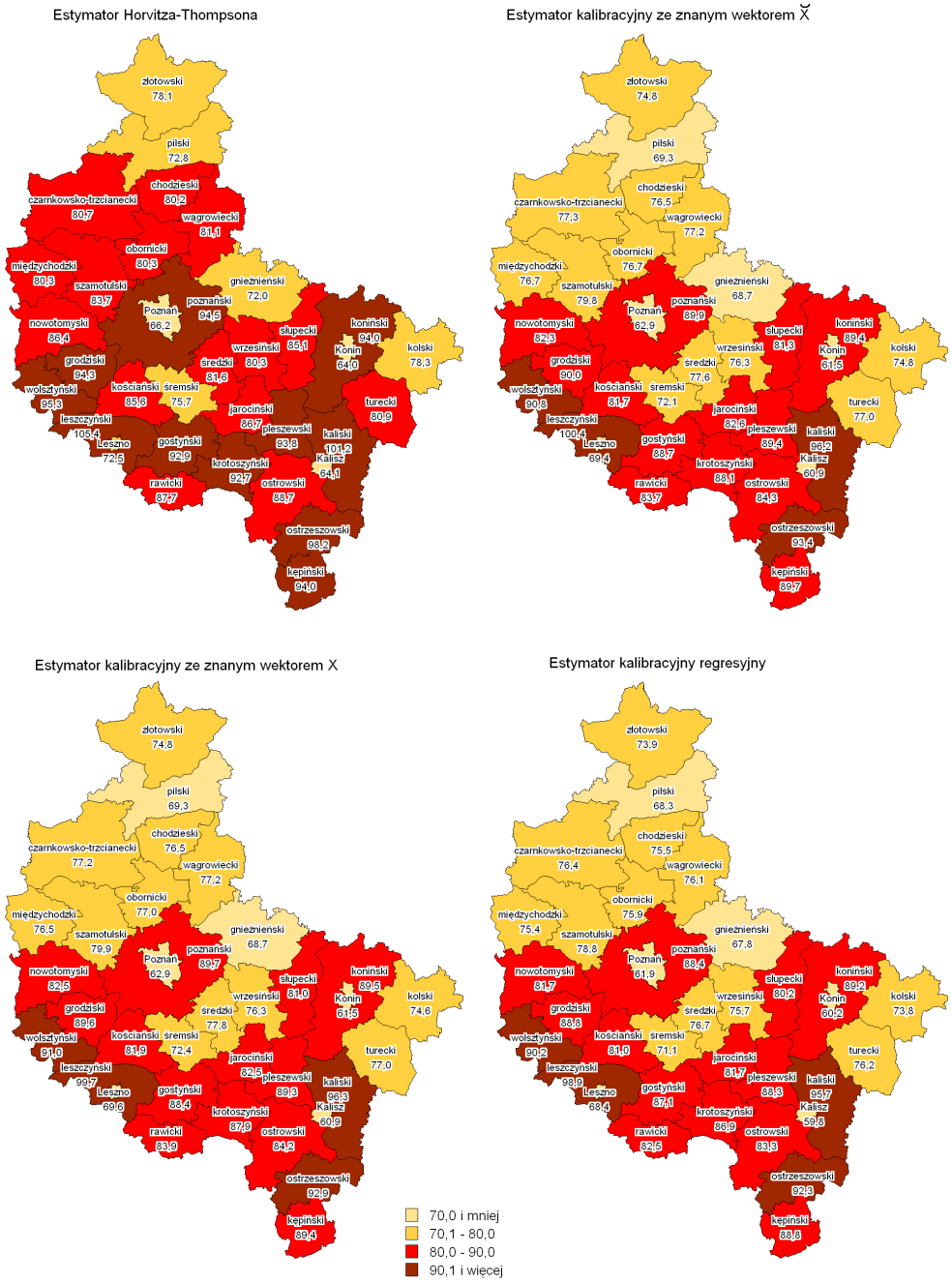
Przewaga estymatorów kalibracyjnych ujawniała się zwłaszcza w sytuacjach, gdy braki danych przypisane zostały mieszkaniom o najniższej – jak i najwyższej powierzchni (wariant 2 i 3 odpowiednio)[37].

Kiedy braki danych przypisane zostały mieszkaniom o najniższym metrażu, estymator bezpośredni wykazywał tendencję do zawyżania średniej powierzchni. W takim wariancie wartość oczekiwana poszczególnych estymatorów kalibracyjnych była zbieżna z „prawdziwą" wartością średniej powierzchni mieszkań. Spore rozbieżności można natomiast zauważyć dla estymatora bezpośredniego. Było to zwłaszcza widoczne, gdy frakcja braków danych wynosiła co najmniej 10% bez względu na liczebność próby, por. tabela 4.1 i rysunek 4.2.

Analizując obciążenie estymatorów w przekroju powiatów województwa wielkopolskiego można zauważyć, że estymator Horvitza-Thompsona dla każdego powiatu

[37] Są to pewne „skrajne" przypadki. W badaniach statystycznych zazwyczaj tak nie jest, że tylko jednostki o najniższych bądź o najwyższych wartościach badanej cechy nie udzielają odpowiedzi na interesujące pytanie. Jednak z reguły braki odpowiedzi nie mają charakteru losowego, a istnieje pewien ukryty mechanizm ich powstawania, który powoduje, że braki danych przyczyniają się do powstawania błędów systematycznych. Można jednak sądzić, że skoro w takich skrajnych przypadkach estymatory kalibracyjne dają lepsze oszacowania od estymatora bezpośredniego to również w innych wariantach będzie podobnie.

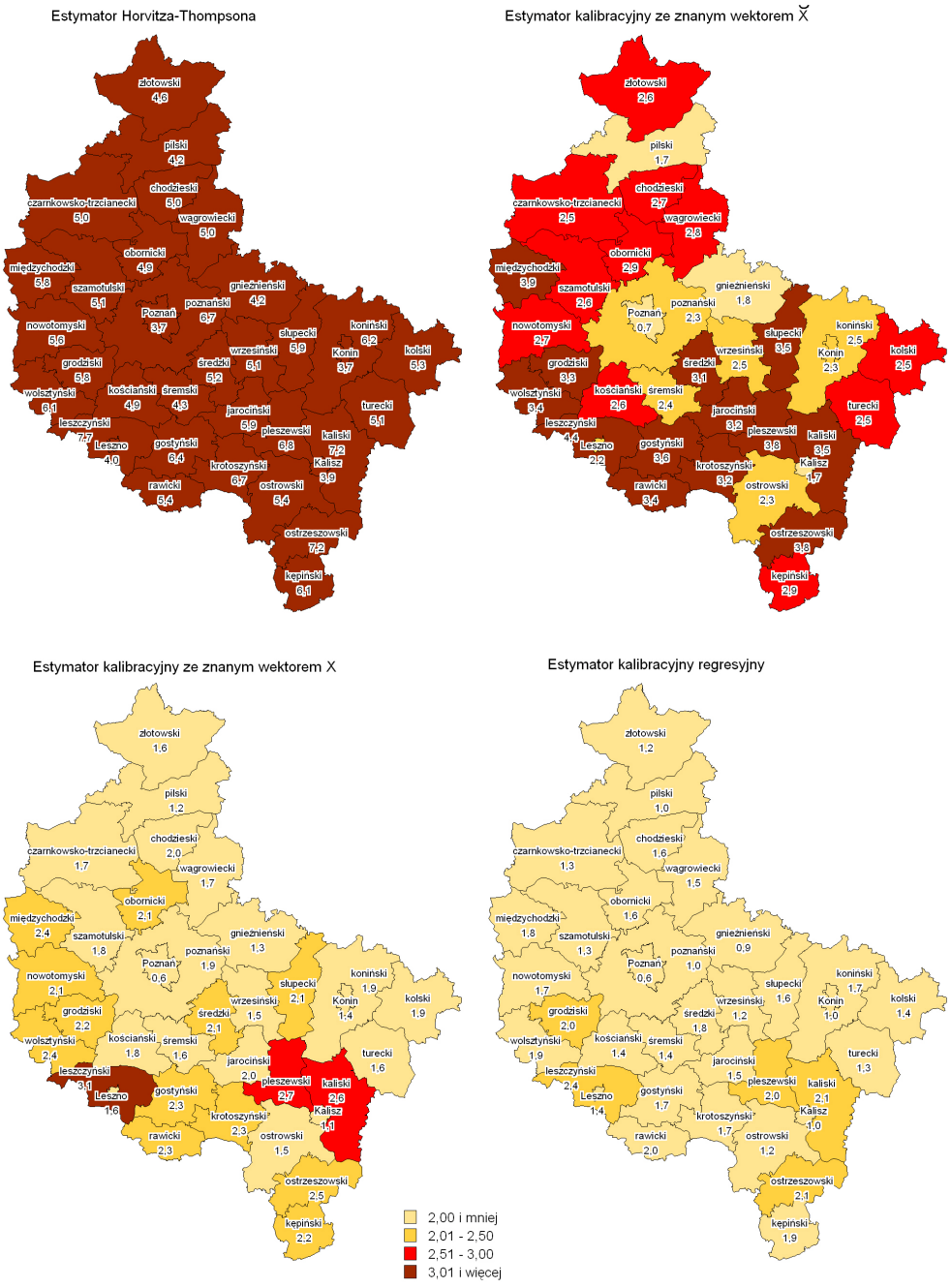
charakteryzował się znacznym obciążeniem. Jest to wyraźnie widoczne na rysunku 4.3.



Rysunek. 4.2. Wartość oczekiwana estymatorów średniej powierzchni mieszkań (w m^2) w przekroju powiatów województwa wielkopolskiego – wariant 2, próba 1%, frakcja braków danych 10%

Źródło: Opracowanie własne na podstawie wyników badań symulacyjnych

Podobnie jak w przypadku powiatu chodzieskiego, najmniejsze obciążenie dla pozostałych powiatów, można było zaobserwować dla estymatora kalibracyjnego $\hat{\bar{Y}}_{REG}$ i $\hat{\bar{Y}}_{\mathbf{X}}$.



Rysunek. 4.3. Wartość oczekiwana obciążenia estymatorów średniej powierzchni mieszkań (w m^2) w przekroju powiatów województwa wielkopolskiego – wariant 2, próba 1%, frakcja braków danych 10%

Źródło: Opracowanie własne na podstawie wyników badań symulacyjnych

Obciążenie było tym większe, im w ramach ustalonej liczebności próby, frakcja braków była wyższa. Spośród estymatorów kalibracyjnych najmniejszym obciążeniem charakteryzował się estymator kalibracyjny regresyjny. W niewiele większym stopniu obciążone były również estymatory kalibracyjne ze znanym wektorem $\mathbf{X}$ i $\breve{\mathbf{X}}$. Przy ustalonej liczebności próby, wzrost frakcji braków powodował znaczne zwiększenie obciążenia estymatora bezpośredniego, podczas gdy dla estymatorów kalibracyjnych zależność ta była mniej zauważalna.

Znajomość wartości globalnej zmiennej pomocniczej ma zatem kluczowe znaczenie dla jakości uzyskanych wyników. Kiedy nie jest ona znana, to jak pokazują wyniki badań symulacyjnych, zastąpienie jej oceną estymatora bezpośredniego także w istotny sposób wpływa na zmniejszenie obciążenia. Wyniki uzyskane z wykorzystaniem estymatora $\hat{\bar{Y}}_{\breve{\mathbf{X}}}$ nie są już jednak tak dobre – pod względem obciążenia – jak dla $\hat{\bar{Y}}_{\mathbf{X}}$ i $\hat{\bar{Y}}_{REG}$, ale w dalszym ciągu są lepsze od ocen, jakie daje estymator Horvitza-Thompsona.

Gdy braki danych przypisywane były mieszkaniom o największej powierzchni, estymator Horvitza-Thompsona wykazywał tendencję do zaniżania średniego metrażu. Była ona zwłaszcza widoczna w sytuacjach, gdy dla ustalonej liczebności próby frakcja braków była większa. Wartości oczekiwane wszystkich rozpatrywanych estymatorów nie były na ogół zbieżne z „prawdziwą" średnią powierzchnią mieszkań. Dotyczyło to przede wszystkim przypadku, gdy frakcja braków wynosiła co najmniej 10%. Wówczas estymatory kalibracyjne wykazywały tendencję do zaniżania średniej powierzchni mieszkań, w znacznie mniejszym jednak stopniu aniżeli estymator bezpośredni.

Rozpatrując wariancję estymatorów można zauważyć, że dla estymatora bezpośredniego jest ona znacznie większa niż dla estymatorów kalibracyjnych, por. tabela 4.3. Dotyczy to wszystkich rozpatrywanych wariantów generowania braków danych. Jest to szczególnie zauważalne, gdy wielkość próby wynosiła 1%. Wraz ze wzrostem liczebności próby, wariancja estymatora Horvitza-Thompsona wykazywała tendencję malejącą, w dalszym jednak ciągu estymatory kalibracyjne charakteryzowały się mniejszą od bezpośredniego wariancją. Jest to zauważalne przede wszystkim w przekroju powiatów województwa wielkopolskiego. Przykładowo, gdy braki danych przypisywane były mieszkaniom o najmniejszej powierzchni (wariant 2), estymatory $\hat{\bar{Y}}_{\mathbf{X}}$ i $\hat{\bar{Y}}_{REG}$ odznaczały się najmniejszą wariancją, por. rysunek 4.4.

Do podobnego wniosku można dojść analizując rozkład wariancji estymatorów w przekroju powiatów dla dwóch pozostałych wariantów. Zarówno, gdy braki danych generowano w sposób losowy, jak i gdy przypisywano je mieszkaniom o najwyższej powierzchni, najmniejszą wariancją charakteryzowały się estymatory kalibracyjne $\hat{\bar{Y}}_{\mathbf{X}}$ i $\hat{\bar{Y}}_{REG}$.

Znajomość wartości globalnej zmiennej pomocniczej odgrywa istotną rolę z punktu widzenia wariancji estymatora. Gdy nie jest ona znana, to jak pokazują wyniki badań symulacyjnych, zastąpienie jej oceną estymatora bezpośredniego w istotny sposób wpływa na zmniejszenie obciążenia estymatora kalibracyjnego $\hat{\bar{Y}}_{\breve{\mathbf{X}}}$, ale nie na jego wariancję, która jest porównywalna z wariancją estymatora Horvitza-Thompsona. Dotyczy to wszystkich trzech wariantów generowania braków danych, przy czym gdy wielkość próby wynosiła 1%, estymatory $\hat{\bar{Y}}_{HT}$ i $\hat{\bar{Y}}_{\breve{\mathbf{X}}}$ charakteryzowały się znacznie większą wariancją od estymatorów $\hat{\bar{Y}}_{\mathbf{X}}$ i $\hat{\bar{Y}}_{REG}$, por. tabela 4.3 i rysunki 4.4, A.4 oraz A.8.

Tabela. 4.3. Wariancja estymatorów średniej powierzchni mieszkań w powiecie chodzieskim

Powiat chodzieski		Wielkość próby								
		1%			10%			20%		
		Frakcja braków			Frakcja braków			Frakcja braków		
		1%	10%	20%	1%	10%	20%	1%	10%	20%
Wariant 1	$\hat{\bar{Y}}_{HT}$	15.36	11.28	15.78	1.41	1.01	1.94	0.58	0.56	0.64
	$\hat{\bar{Y}}_{\breve{\mathbf{X}}}$	15.50	11.06	16.14	1.38	1.03	1.75	0.55	0.53	0.54
	$\hat{\bar{Y}}_{\mathbf{X}}$	5.13	4.56	7.20	0.46	0.49	0.81	0.18	0.20	0.24
	$\hat{\bar{Y}}_{REG}$	4.89	4.37	5.98	0.41	0.48	0.65	0.17	0.18	0.22
Wariant 2	$\hat{\bar{Y}}_{HT}$	9.73	12.00	23.68	1.28	1.58	1.65	0.47	0.76	0.86
	$\hat{\bar{Y}}_{\breve{\mathbf{X}}}$	9.55	10.17	20.41	1.27	1.46	1.36	0.48	0.71	0.76
	$\hat{\bar{Y}}_{\mathbf{X}}$	4.63	5.02	8.53	0.44	0.58	0.54	0.17	0.25	0.31
	$\hat{\bar{Y}}_{REG}$	4.65	4.12	5.59	0.40	0.43	0.40	0.17	0.20	0.21
Wariant 3	$\hat{\bar{Y}}_{HT}$	12.49	7.85	6.13	0.86	0.60	0.43	0.55	0.29	0.34
	$\hat{\bar{Y}}_{\breve{\mathbf{X}}}$	12.70	9.42	7.79	0.89	0.75	0.68	0.55	0.36	0.44
	$\hat{\bar{Y}}_{\mathbf{X}}$	4.52	3.60	2.77	0.36	0.25	0.23	0.17	0.11	0.13
	$\hat{\bar{Y}}_{REG}$	4.63	4.08	4.06	0.36	0.31	0.34	0.16	0.14	0.18

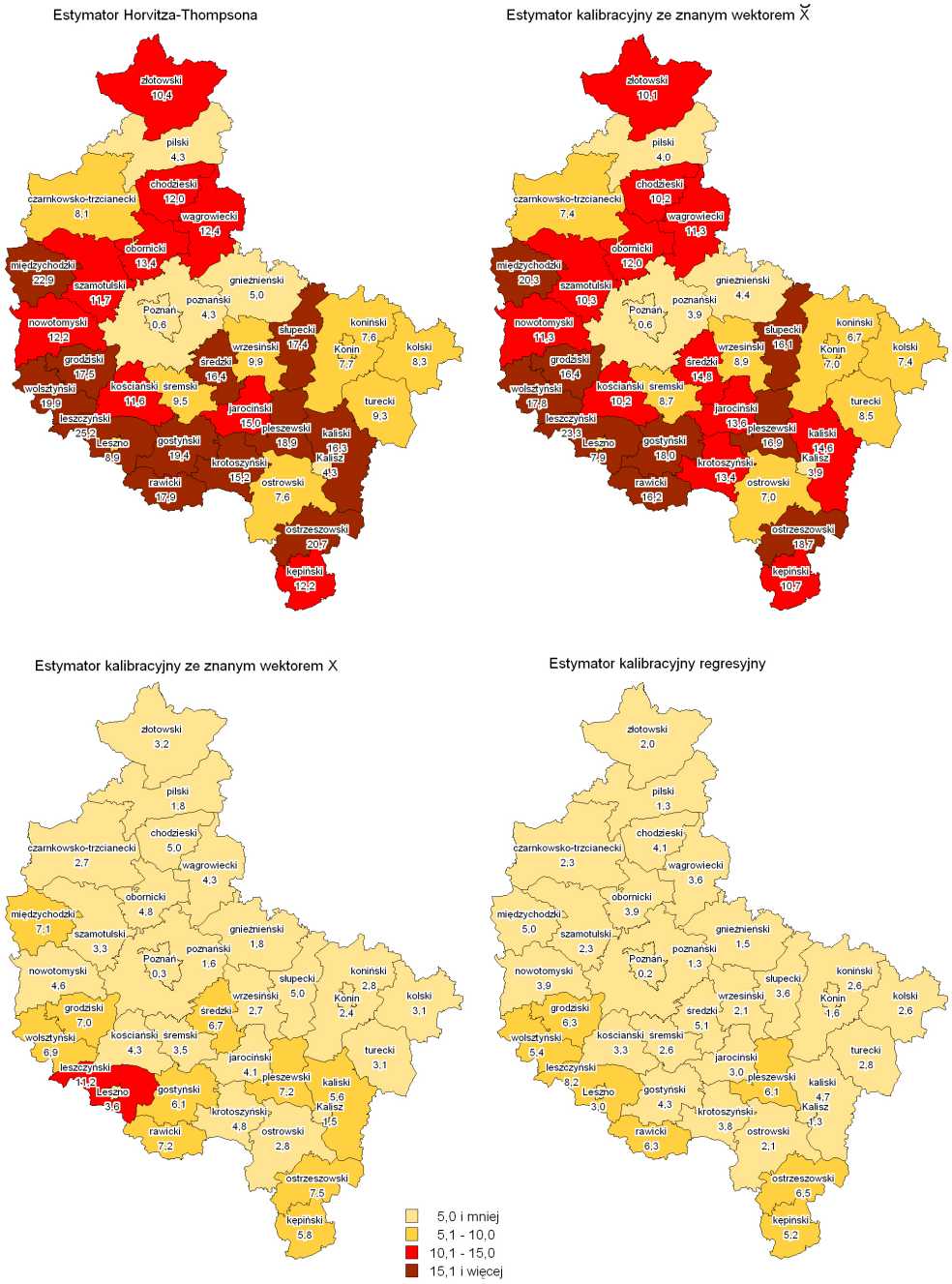
Źródło: Opracowanie własne na podstawie wyników badań symulacyjnych

O precyzji estymatora świadczy także względny błąd szacunku, por. tabela 4.4.

Tabela. 4.4. Względny błąd szacunku estymatorów średniej powierzchni mieszkań w powiecie chodzieskim (w %)

Powiat chodzieski		Wielkość próby								
		1%			10%			20%		
		Frakcja braków			Frakcja braków			Frakcja braków		
		1%	10%	20%	1%	10%	20%	1%	10%	20%
Wariant 1	$\hat{\bar{Y}}_{HT}$	5.23	4.46	5.23	1.57	1.33	1.84	1.00	0.99	1.06
	$\hat{\bar{Y}}_{\breve{\mathbf{X}}}$	5.25	4.39	5.26	1.56	1.34	1.74	0.98	0.96	0.97
	$\hat{\bar{Y}}_{\mathbf{X}}$	3.00	2.82	3.53	0.90	0.93	1.18	0.56	0.59	0.65
	$\hat{\bar{Y}}_{REG}$	2.92	2.76	3.24	0.85	0.91	1.07	0.54	0.56	0.62
Wariant 2	$\hat{\bar{Y}}_{HT}$	4.12	4.32	5.66	1.49	1.57	1.51	0.91	1.08	1.08
	$\hat{\bar{Y}}_{\breve{\mathbf{X}}}$	4.10	4.17	5.71	1.49	1.58	1.49	0.91	1.10	1.12
	$\hat{\bar{Y}}_{\mathbf{X}}$	2.85	2.93	3.71	0.88	1.00	0.94	0.55	0.65	0.71
	$\hat{\bar{Y}}_{REG}$	2.86	2.69	3.08	0.84	0.87	0.82	0.55	0.59	0.60
Wariant 3	$\hat{\bar{Y}}_{HT}$	4.75	4.36	4.23	1.27	1.21	1.12	1.01	0.84	1.01
	$\hat{\bar{Y}}_{\breve{\mathbf{X}}}$	4.76	4.45	4.19	1.28	1.26	1.24	1.00	0.87	1.00
	$\hat{\bar{Y}}_{\mathbf{X}}$	2.85	2.74	2.51	0.81	0.73	0.72	0.56	0.48	0.55
	$\hat{\bar{Y}}_{REG}$	2.88	2.94	3.10	0.81	0.81	0.90	0.55	0.54	0.64

Źródło: Opracowanie własne na podstawie wyników badań symulacyjnych



Rysunek. 4.4. Wariancja estymatorów średniej powierzchni mieszkań w przekroju powiatów województwa wielkopolskiego – wariant 2, próba 1%, frakcja braków danych 10%

Źródło: Opracowanie własne na podstawie wyników badań symulacyjnych

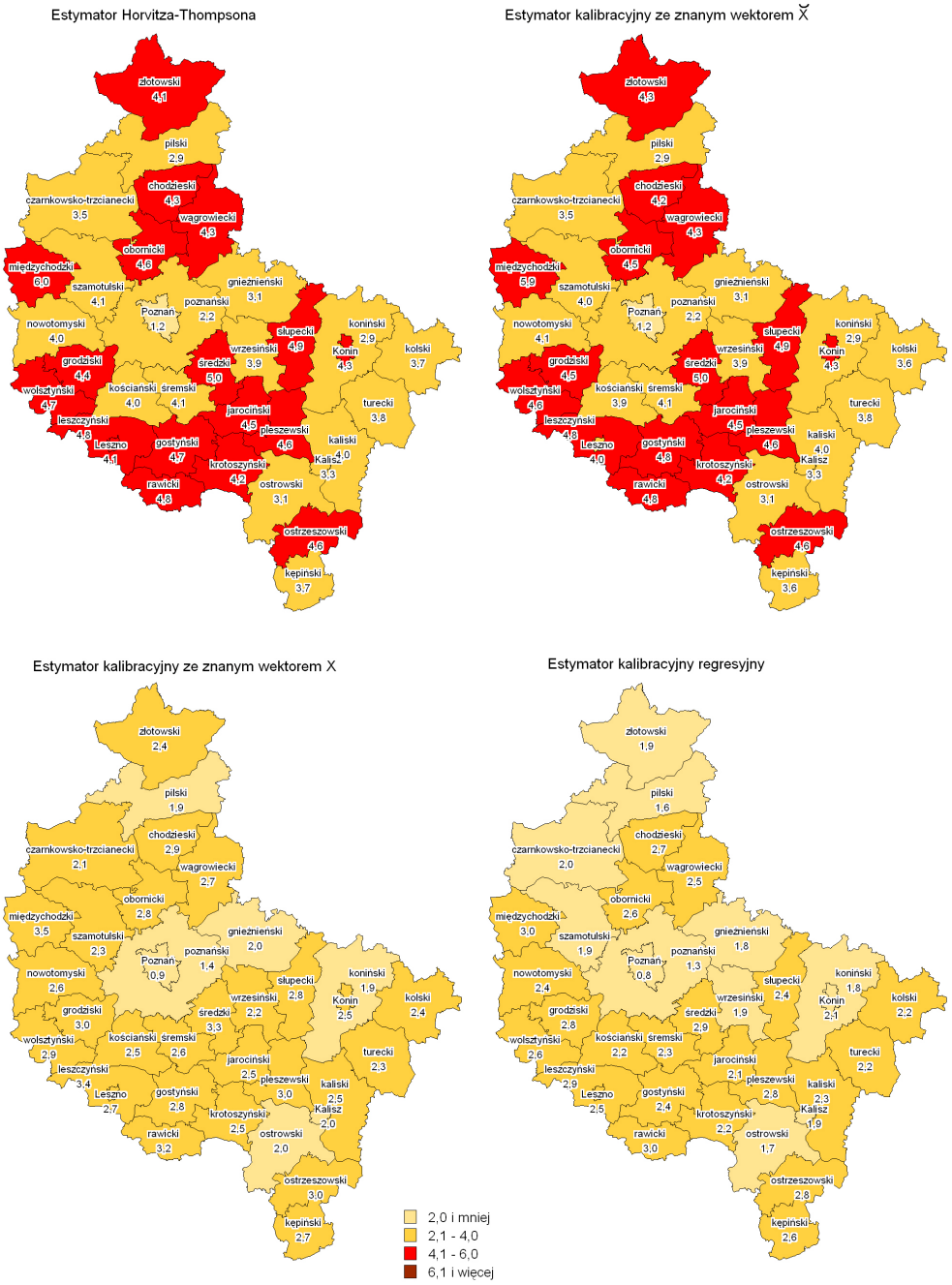
Wyniki badania symulacyjnego pokazują, że estymator Horvitza-Thompsona jest najmniej efektywny ze wszystkich przedstawionych. Z grupy estymatorów kalibracyjnych najwyższą precyzją szacunków charakteryzował się $\hat{\bar{Y}}_{\mathbf{x}}$ i $\hat{\bar{Y}}_{REG}$. Wykorzystanie informacji dodatkowej, w postaci wartości globalnej zmiennej pomocniczej, wpływa nie tylko na redukcję obciążenia, ale również względnego błędu szacunku. Zwiększanie efektywności rozważanych estymatorów wiąże się również bezpośrednio z zagadnieniem wielkości próby oraz frakcji braków danych. Precyzja szacunków średniej powierzchni mieszkań była zazwyczaj większa w sytuacjach, gdy frakcja braków była niższa. Ponadto, przy ustalonej na określonym poziomie frakcji braków danych, precyzja była wyższa dla większych prób. Oznacza to spadek względnego błędu szacunku przy wzroście liczby mieszkań wylosowanych do próby.

Największy względny błąd szacunku wszystkich rozpatrywanych estymatorów zaobserwowano, gdy braki danych generowano w sposób losowy. Dotyczyło to przede wszystkim sytuacji, kiedy wielkość próby wynosiła 1% – bez względu na liczebność frakcji braków danych.

Dokonując analizy precyzji szacunków, w przekroju powiatów województwa wielkopolskiego, można wyciągnąć bardzo podobne wnioski. Największe wartości błędu względnego otrzymano właściwie dla większości powiatów dla estymatora bezpośredniego, bez względu na liczebność próby, frakcję braków danych oraz mechanizm ich powstawania. Przykładowo, gdy wielkość próby wynosiła 1%, frakcja braków 10% i przypisywane były one mieszkaniom o najniższym metrażu (wariant 2) względny błąd szacunku estymatora bezpośredniego wynosił od 2,9% dla powiatu pilskiego do 6% dla powiatu międzychodzkiego. Dla porównania, względny błąd szacunku dla estymatora kalibracyjnego regresyjnego nie przekraczał dla każdego powiatu 3%, por. rysunek 4.5.

Dokonując podsumowania symulacyjnego badania obciążenia i efektywności estymatorów średniej powierzchni mieszkań można stwierdzić, że najlepsze są estymatory kalibracyjne wykorzystujące informacje dodatkowe w postaci wartości globalnej zmiennej pomocniczej. Charakteryzują się one najmniejszym obciążeniem, wariancją i względnym błędem szacunku. Możliwość wykorzystania dodatkowych informacji pozwala w istotny sposób zwiększyć poprawę precyzji, przy jednoczesnej redukcji obciążenia będącego skutkiem braków danych. Zależność taka była zauważalna we wszystkich rozpatrywanych wariantach oraz dla różnych wielkości prób i frakcji braków.

Warto podkreślić, że estymatory kalibracyjne wykazywały swoją wyższość nad estymatorem Horvitza-Thompsona zwłaszcza, gdy braki odpowiedzi nie miały charakteru losowego, a frakcja braków była duża. Jest to szczególnie ważne z punktu widzenia badań prowadzonych przez Główny Urząd Statystyczny, gdzie braki odpowiedzi, są źródłem dużych błędów systematycznych z racji skali na jaką występują i faktu, że nie mają one charakteru losowego.



Rysunek. 4.5. Względny błąd szacunku estymatorów średniej powierzchni mieszkań (w %) w przekroju powiatów województwa wielkopolskiego – wariant 2, próba 1%, frakcja braków danych 10%

Źródło: Opracowanie własne na podstawie wyników badań symulacyjnych

4.3. Założenia badania symulacyjnego dla mediany

W przypadku szacowania mediany cechy Y, z wykorzystaniem estymatorów kalibracyjnych, analizę przeprowadzono w oparciu o sztucznie utworzoną populację[38]. W tym celu wygenerowano 10 000 obserwacji dla zmiennej Y, która podlegała rozkładowi logarytmiczno-normalnemu z prawdziwą wartością mediany równą $Q_{y,0.5} = 2\,356$. Podobnie postąpiono w przypadku zmiennej X, dla której również wygenerowano 10 000 obserwacji z rozkładu logarytmiczno-normalnego ze znanymi kwartylami wynoszącymi odpowiednio: $Q_{x,0.25} = 2\,612$, $Q_{x,0.5} = 2\,988$ i $Q_{x,0.75} = 3\,423$. Otrzymano w ten sposób 10 000 par liczbowych $(x_1, y_1), \ldots, (x_{10\ 000}, y_{10\ 000})$, które można traktować jako wartości dwuwymiarowej zmiennej (X, Y), przy czym Y jest zmienną badaną w populacji, a X zmienną pomocniczą.

Na potrzeby oszacowania mediany zmiennej Y wykorzystano estymator bezpośredni i trzy estymatory kalibracyjne:

- estymator Horvitza-Thompsona

$$\hat{Q}_{y,HT,0.5} = \hat{F}_y^{-1}(0.5), \tag{4.14}$$

gdzie:

$$\hat{F}_y(t) = \frac{\sum_{i=1}^{m} d_i H_{y,r}(t, y_i)}{\sum_{i=1}^{m} d_i}. \tag{4.15}$$

- estymator kalibracyjny mediany ze znanym wektorem $\hat{Q}_{\boldsymbol{x},0.5}$

$$\hat{Q}^1_{y,cal,0.5} = \hat{F}^{-1}_{y,cal}(0.5). \tag{4.16}$$

W przypadku tego estymatora założono, że nie jest znana mediana $Q_{x,0.5}$ zmiennej pomocniczej w populacji generalnej. Dlatego zastąpiono ją oceną estymatora Horvitza-Thompsona opierając się na danych pochodzących z wylosowanej z populacji próby s. Jedyną składową wektora $\hat{Q}_{\boldsymbol{x},0.5}$ była zatem oszacowana mediana zmiennej pomocniczej.

- estymator kalibracyjny mediany ze znanym wektorem $Q_{\boldsymbol{x},0.5}$

$$\hat{Q}^2_{y,cal,0.5} = \hat{F}^{-1}_{y,cal}(0.5). \tag{4.17}$$

W przypadku tego estymatora założono, że znana jest mediana $Q_{x,0.5}$ zmiennej pomocniczej w populacji generalnej, która stanowiła jedyną składową wektora $Q_{\boldsymbol{x},0.5}$.

[38] Wynikało to głównie z faktu, że obliczenia na danych rzeczywistych z NSP'2002 mogły być wykonywane tylko w Urzędzie Statystycznym w Poznaniu, przy czym dane znajdowały się na serwerze Głównego Urzędu Statystycznego w Warszawie, co przy niewielkiej przepustowości łącza, dużych zbiorach danych i dużej liczbie niezbędnych do przeprowadzenia replikacji, w zasadzie uniemożliwiało testowanie wszystkich omawianych estymatorów kalibracyjnych. Dlatego w badaniach symulacyjnych zdecydowano się poddać weryfikacji własności estymatorów kalibracyjnych średniej powierzchni mieszkań w oparciu o dane spisowe. Drugi powód dotyczył rozkładu wielu cech w badaniu budżetów gospodarstw domowych, które charakteryzują się asymetrią prawostronną. Dlatego w odniesieniu do estymatorów kalibracyjnych mediany, zdecydowano się na przeprowadzenie badań symulacyjnych w oparciu o dane wygenerowane z rozkładu logarytmiczno-normalnego, który również charakteryzuje się asymetrią prawostronną i często wykorzystywany jest do modelowania rozkładu dochodów czy wydatków gospodarstw domowych.

- uogólniony estymator kalibracyjny mediany

$$\hat{Q}^{3}_{y,cal,0.5} = \hat{F}^{-1}_{y,cal}(0.5). \tag{4.18}$$

W przypadku tego estymatora założono, że znane są wszystkie kwartyle $Q_{x,0.25}$, $Q_{x,0.5}$ i $Q_{x,0.75}$ zmiennej pomocniczej w populacji generalnej.

Dystrybuanta interpolacyjna zmiennej Y, która stanowiła podstawę wyznaczania ocen wszystkich rozważanych w badaniu symulacyjnym estymatorów kalibracyjnych mediany wyrażała się wzorem:

$$\hat{F}_{y,cal}(t) = \frac{\sum_{i=1}^{m} w_i H_{y,r}(t, y_i)}{\sum_{i=1}^{m} w_i}. \tag{4.19}$$

W dalszym ciągu postępowano w sposób zbliżony do procedury opisanej w przypadku badania symulacyjnego dla średniej powierzchni mieszkań[39]. Losowano zatem, zgodnie ze schematem losowania zależnego, próby wielkości 1%, 10% i 20% z całej populacji.

Następnie, po wylosowaniu n – elementowej próby, zastępowano część informacji o wartościach zmiennej Y brakami danych. Zastosowano przy tym trzy opisane wcześniej różne podejścia dla tworzenia braków danych. W pierwszym zakładano, że braki danych generowane były w sposób losowy (wariant 1). W drugim (wariant 2) i trzecim (wariant 3), założono, że braki danych przypisano jednostkom o najmniejszych i największych wartościach cechy Y odpowiednio. Przyjęto przy tym, że w każdym przypadku frakcja braków danych w próbie wynosiła: 1%, 10% i 20%. Dla różnych wariantów liczebności próby (trzy możliwości), frakcji braków (trzy możliwości) oraz sposobów ich generowania (trzy możliwości) przeprowadzono po 500 replikacji i na tej podstawie dokonano oszacowania wartości oczekiwanej mediany zmiennej Y, wartości oczekiwanej obciążenia badanych estymatorów, ich wariancji empirycznej oraz względnych błędów szacunku według poniższych wzorów:

- wartość oczekiwana estymatora mediany zmiennej Y

$$\hat{\bar{Q}}_{y,0.5} = \frac{1}{500} \sum_{i=1}^{500} \hat{Q}^{(i)}_{y,0.5}, \tag{4.20}$$

gdzie $\hat{Q}^{(i)}_{y,0.5}$ oznacza ocenę mediany zmiennej Y w i – tej replikacji uzyskaną z wykorzystaniem jednego z czterech rozważanych w badaniu estymatorów , $i = 1, \ldots, 500$.

- wartość oczekiwana obciążenia

$$B = \frac{1}{500} \sum_{i=1}^{500} \left| \hat{Q}^{(i)}_{y,0.5} - Q_{y,0.5} \right|, \tag{4.21}$$

gdzie $Q_{y,0.5}$ oznacza medianę zmiennej Y w populacji.

[39] Różnica w zasadzie dotyczyła przyjętego schematu losowania próby.

- wariancja empiryczna

$$\hat{V}^2 = \frac{1}{499}\sum_{i=1}^{500}\left(\hat{Q}_{y,0.5}^{(i)} - \hat{\bar{Q}}_{y,0.5}\right)^2. \tag{4.22}$$

- względny błąd szacunku

$$REE = \frac{\sqrt{\hat{V}^2}}{\hat{\bar{Q}}_{y,0.5}} \cdot 100\%. \tag{4.23}$$

4.4. Wyniki badania symulacyjnego dla mediany

Podobnie, jak w przypadku badań symulacyjnych dla średniej, gdy braki danych dla zmiennej Y generowane były w sposób losowy (wariant 1), to bez względu na liczebność próby oraz frakcję braków, zarówno wartość oczekiwana estymatora bezpośredniego jak i estymatorów kalibracyjnych mediany, były bliskie „prawdziwej" wartości mediany w populacji, por. tabela 4.5. Estymatory te charakteryzowały się w tej sytuacji niewielkimi obciążeniami.

Tabela. 4.5. Wartość oczekiwana estymatorów mediany zmiennej Y

Wartość oczekiwana estymatorów		Wielkość próby								
		1%			10%			20%		
		Frakcja braków			Frakcja braków			Frakcja braków		
		1%	10%	20%	1%	10%	20%	1%	10%	20%
Wariant 1	$Q_{y,0.5}$	2356.0	2356.0	2356.0	2356.0	2356.0	2356.0	2356.0	2356.0	2356.0
	$\hat{Q}_{y,HT,0.5}$	2352.1	2350.4	2345.2	2354.5	2354.2	2357.4	2355.6	2355.0	2355.6
	$\hat{Q}^1_{y,cal,0.5}$	2351.3	2352.2	2347.3	2354.7	2354.6	2357.1	2355.7	2354.9	2355.4
	$\hat{Q}^2_{y,cal,0.5}$	2352.8	2353.4	2356.9	2354.6	2355.6	2356.9	2356.1	2355.7	2356.1
	$\hat{Q}^3_{y,cal,0.5}$	2355.1	2355.6	2358.6	2354.9	2355.7	2356.5	2356.2	2355.8	2355.9
Wariant 2	$Q_{y,0.5}$	2356.0	2356.0	2356.0	2356.0	2356.0	2356.0	2356.0	2356.0	2356.0
	$\hat{Q}_{y,HT,0.5}$	2359.7	2437.5	2521.6	2362.1	2438.9	2527.1	2363.6	2439.8	2527.2
	$\hat{Q}^1_{y,cal,0.5}$	2353.6	2380.1	2409.7	2356.3	2378.6	2410.1	2357.7	2378.7	2410.5
	$\hat{Q}^2_{y,cal,0.5}$	2354.2	2380.5	2410.1	2356.9	2378.9	2411.1	2358.1	2377.9	2410.7
	$\hat{Q}^3_{y,cal,0.5}$	2355.6	2365.6	2380.5	2355.4	2363.4	2377.1	2356.2	2362.9	2376.7
Wariant 3	$Q_{y,0.5}$	2356.0	2356.0	2356.0	2356.0	2356.0	2356.0	2356.0	2356.0	2356.0
	$\hat{Q}_{y,HT,0.5}$	2346.7	2265.8	2179.7	2346.8	2275.8	2190.8	2347.2	2274.5	2193.1
	$\hat{Q}^1_{y,cal,0.5}$	2353.1	2324.9	2289.1	2352.9	2333.5	2301.3	2353.3	2332.4	2304.0
	$\hat{Q}^2_{y,cal,0.5}$	2355.1	2326.5	2295.1	2353.5	2333.6	2301.9	2353.3	2332.5	2303.9
	$\hat{Q}^3_{y,cal,0.5}$	2357.3	2349.5	2326.6	2355.5	2350.6	2329.5	2355.1	2350.3	2330.6

Źródło: Opracowanie własne na podstawie wyników badań symulacyjnych

Przewaga estymatorów kalibracyjnych ujawniała się zwłaszcza w sytuacjach, gdy braki danych nie miały charakteru losowego, a przypisane zostały jednostkom o najniższych jak i najwyższych wartościach zmienej Y (wariant 2 i 3 odpowiednio).

Gdy braki danych przypisane zostały jednostkom o najniższych wartościach zmiennej Y, estymator bezpośredni wykazywał tendencję do zawyżania mediany $Q_{y,0.5}$. W takim wariancie wartość oczekiwana poszczególnych estymatorów kalibracyjnych była

w dużym stopniu zbieżna z „prawdziwą" wartością mediany zmiennej Y w populacji. Spore rozbieżności można natomiast zauważyć dla estymatora bezpośredniego. Było to zwłaszcza widoczne, gdy frakcja braków danych wynosiła co najmniej 10% bez względu na liczebność próby, por. tabela 4.5.

Kiedy braki danych przypisywane były jednostkom o najwyższych wartościach zmiennej Y, estymator Horvitza-Thompsona wykazywał tendencję do zaniżania mediany $Q_{y,0.5}$. Była ona zwłaszcza widoczna, gdy dla ustalonej liczebności próby frakcja braków była większa. Wartości oczekiwane wszystkich rozpatrywanych estymatorów kalibracyjnych i w tym wariancie były na ogół zbieżne z „prawdziwą" medianą zmiennej Y w populacji. W takich przypadkach estymatory kalibracyjne wykazywały oczywiście tendencję do zaniżania wartości mediany, w znacznie mniejszym jednak stopniu aniżeli estymator bezpośredni.

Obciążenie rozpatrywanych estymatorów było tym większe, im w ramach ustalonej liczebności próby, frakcja braków była wyższa, por. tabela 4.6. Spośród estymatorów kalibracyjnych najmniejszym obciążeniem charakteryzował się uogólniony estymator kalibracyjny mediany. W nieco większym stopniu obciążone były estymatory kalibracyjne ze znanym wektorem $Q_{\boldsymbol{x},0.5}$ i $\hat{Q}_{\boldsymbol{x},0.5}$. Przy ustalonej liczebności próby, wzrost frakcji braków powodował znaczne zwiększenie obciążenia estymatora bezpośredniego, podczas gdy dla estymatorów kalibracyjnych zależność ta była w mniejszym stopniu zauważalna.

Tabela. 4.6. Wartość oczekiwana obciążenia estymatorów mediany zmiennej Y

Wartość oczekiwana obciążenia estymatorów		Wielkość próby								
		1%			10%			20%		
		Frakcja braków			Frakcja braków			Frakcja braków		
		1%	10%	20%	1%	10%	20%	1%	10%	20%
Wariant 1	$\hat{Q}_{y,HT,0.5}$	65.95	69.63	72.28	18.49	19.86	20.61	12.33	13.40	14.57
	$\hat{Q}^1_{y,cal,0.5}$	65.81	68.19	67.17	18.38	19.29	19.01	12.22	13.05	13.52
	$\hat{Q}^2_{y,cal,0.5}$	46.15	45.66	47.32	12.87	13.07	13.68	8.44	8.93	9.26
	$\hat{Q}^3_{y,cal,0.5}$	42.34	43.39	44.73	12.01	12.37	13.12	7.94	8.33	8.84
Wariant 2	$\hat{Q}_{y,HT,0.5}$	65.57	93.54	167.44	20.80	82.93	171.02	14.26	83.80	171.12
	$\hat{Q}^1_{y,cal,0.5}$	65.30	66.60	78.03	20.34	26.56	54.18	13.16	23.55	54.51
	$\hat{Q}^2_{y,cal,0.5}$	41.10	46.97	61.40	13.63	23.66	55.01	8.64	21.99	54.63
	$\hat{Q}^3_{y,cal,0.5}$	39.14	42.42	41.58	12.12	12.48	21.82	7.95	9.24	20.70
Wariant 3	$\hat{Q}_{y,HT,0.5}$	66.97	100.81	177.66	21.31	80.17	165.14	14.67	81.49	162.85
	$\hat{Q}^1_{y,cal,0.5}$	66.87	69.13	90.14	20.09	26.88	54.88	12.96	24.55	52.02
	$\hat{Q}^2_{y,cal,0.5}$	42.72	50.97	69.45	13.14	23.75	54.11	8.53	23.57	52.08
	$\hat{Q}^3_{y,cal,0.5}$	38.97	41.66	55.57	11.95	12.79	27.58	7.66	9.18	25.51

Źródło: Opracowanie własne na podstawie wyników badań symulacyjnych

Znajomość mediany (bądź kwartyli w przypadku rozpatrywanego w badaniu uogólnionego estymatora kalibracyjnego mediany) zmiennej pomocniczej X ma zatem fundamentalne znaczenie dla jakości uzyskanych wyników. Gdy nie jest ona znana, to jak

pokazują wyniki badań symulacyjnych, zastąpienie jej oceną estymatora bezpośredniego w dalszym ciągu – w istotny sposób – wpływa na zmniejszenie obciążenia.

Rozpatrując wariancję estymatorów mediany można zauważyć, że dla estymatora bezpośredniego jest ona znacznie większa niż dla estymatorów kalibracyjnych, por. tabela 4.7. Dotyczy to w zasadzie wszystkich rozpatrywanych wariantów generowania braków danych. Jest to szczególnie zauważalne, kiedy wielkość próby wynosiła 1%. Wraz ze wzrostem liczebności próby, wariancja estymatora Horvitza-Thompsona wykazywała tendencję malejącą, w dalszym jednak ciągu estymatory kalibracyjne charakteryzowały się mniejszą od bezpośredniego wariancją.

Tabela. 4.7. Wariancja estymatorów mediany zmiennej Y

Wariancja estymatorów		Wielkość próby								
		1%			10%			20%		
		Frakcja braków			Frakcja braków			Frakcja braków		
		1%	10%	20%	1%	10%	20%	1%	10%	20%
Wariant 1	$\hat{Q}_{y,HT,0.5}$	6856.0	7605.9	8308.8	537.3	612.6	673.6	238.9	281.0	330.2
	$\hat{Q}^1_{y,cal,0.5}$	6875.5	7466.9	7136.8	533.4	581.8	582.4	234.8	264.0	285.5
	$\hat{Q}^2_{y,cal,0.5}$	3320.7	3391.4	3604.0	260.1	277.2	292.5	112.2	125.8	132.3
	$\hat{Q}^3_{y,cal,0.5}$	2771.1	2970.2	3141.6	219.7	246.6	265.5	99.2	107.2	118.8
Wariant 2	$\hat{Q}_{y,HT,0.5}$	6967.8	7134.2	7673.2	651.6	682.7	694.8	258.5	281.8	307.7
	$\hat{Q}^1_{y,cal,0.5}$	6949.4	6641.8	6470.9	647.8	557.7	611.5	266.8	225.0	289.7
	$\hat{Q}^2_{y,cal,0.5}$	2595.1	2869.9	2504.5	281.9	218.9	318.0	111.2	82.6	135.9
	$\hat{Q}^3_{y,cal,0.5}$	2419.8	2775.3	2239.6	230.9	183.0	198.8	98.6	78.0	73.1
Wariant 3	$\hat{Q}_{y,HT,0.5}$	6702.1	6170.1	6128.3	605.9	617.1	504.9	252.5	280.2	262.1
	$\hat{Q}^1_{y,cal,0.5}$	6775.7	6705.8	7394.3	610.6	572.5	593.4	245.3	248.1	315.6
	$\hat{Q}^2_{y,cal,0.5}$	2864.3	3252.5	3490.5	261.2	260.7	309.6	107.2	104.0	139.8
	$\hat{Q}^3_{y,cal,0.5}$	2542.2	2733.2	4741.2	225.1	232.6	317.8	92.2	99.6	140.8

Źródło: Opracowanie własne na podstawie wyników badań symulacyjnych

Spośród wszystkich rozważanych estymatorów najbardziej efektywny był uogólniony estymator kalibracyjny mediany. Znajomość większej liczby kwantyli zmiennej pomocniczej odgrywa zatem, podobnie jak w przypadku obciążenia, istotną rolę z punktu widzenia zmniejszenia wariancji estymatora. Jest to zauważalne we wszystkich rozpatrywanych wariantach generowania braków danych, bez względu na ich frakcję oraz wielkość próby. Kiedy nie jest znany żaden kwantyl zmiennej pomocniczej, to jak pokazują wyniki badań symulacyjnych, zastąpienie go oceną estymatora bezpośredniego również przyczynia się do zmniejszenia wariancji we wszystkich rozpatrywanych wariantach.

Podobnie, jak w przypadku analizy własności estymatorów średniej powierzchni mieszkań, wyniki badania symulacyjnego dla mediany pokazują, że estymator bezpośredni jest najmniej efektywny ze wszystkich przedstawionych. Z grupy estymatorów kalibracyjnych najwyższą precyzją szacunków charakteryzowały się $\hat{Q}^2_{y,cal,0.5}$ i $\hat{Q}^3_{y,cal,0.5}$. Wykorzystanie informacji dodatkowej w postaci mediany zmiennej pomocniczej (bądź większej liczby znanych kwantyli tej zmiennej na poziomie populacji) wpływa zatem

nie tylko na redukcję obciążenia, ale również względnego błędu szacunku. Precyzja szacunków mediany zmiennej Y była zazwyczaj większa w sytuacjach, gdy frakcja braków była niższa. Ponadto, przy ustalonej na określonym poziomie frakcji braków danych, precyzja była wyższa dla większych prób. Oznacza to spadek względnego błędu szacunku przy wzroście liczby jednostek wylosowanych do próby, por. tabela 4.8.

Tabela. 4.8. Względny błąd szacunku estymatorów mediany zmiennej Y (w %)

Względny błąd szacunku		Wielkość próby								
		1%			10%			20%		
		Frakcja braków			Frakcja braków			Frakcja braków		
		1%	10%	20%	1%	10%	20%	1%	10%	20%
Wariant 1	$\hat{Q}_{y,HT,0.5}$	3.52	3.71	3.89	0.98	1.05	1.1	0.66	0.71	0.77
	$\hat{Q}^1_{y,cal,0.5}$	3.53	3.67	3.6	0.98	1.02	1.02	0.65	0.69	0.72
	$\hat{Q}^2_{y,cal,0.5}$	2.45	2.47	2.55	0.68	0.71	0.73	0.45	0.48	0.49
	$\hat{Q}^3_{y,cal,0.5}$	2.24	2.31	2.38	0.63	0.67	0.69	0.42	0.44	0.46
Wariant 2	$\hat{Q}_{y,HT,0.5}$	3.54	3.47	3.47	1.08	1.07	1.04	0.68	0.69	0.69
	$\hat{Q}^1_{y,cal,0.5}$	3.54	3.42	3.34	1.08	0.99	1.03	0.69	0.63	0.71
	$\hat{Q}^2_{y,cal,0.5}$	2.16	2.25	2.08	0.71	0.62	0.74	0.45	0.38	0.48
	$\hat{Q}^3_{y,cal,0.5}$	2.09	2.23	1.99	0.65	0.57	0.59	0.42	0.37	0.36
Wariant 3	$\hat{Q}_{y,HT,0.5}$	3.49	3.47	3.59	1.05	1.09	1.03	0.68	0.74	0.74
	$\hat{Q}^1_{y,cal,0.5}$	3.5	3.52	3.76	1.05	1.03	1.06	0.67	0.68	0.77
	$\hat{Q}^2_{y,cal,0.5}$	2.27	2.45	2.57	0.69	0.69	0.76	0.44	0.44	0.51
	$\hat{Q}^3_{y,cal,0.5}$	2.14	2.23	2.96	0.64	0.65	0.77	0.41	0.42	0.51

Źródło: Opracowanie własne na podstawie wyników badań symulacyjnych

Największy względny błąd szacunku wszystkich rozpatrywanych estymatorów zaobserwowano, gdy braki danych generowano w sposób losowy. Dotyczyło to przede wszystkim sytuacji, gdy wielkość próby wynosiła 1%, a frakcja braków danych wynosiła 20%.

Dokonując podsumowania symulacyjnego badania obciążenia i efektywności estymatorów mediany zmiennej Y można stwierdzić, że najlepsze są estymatory kalibracyjne wykorzystujące informacje dodatkowe w postaci wartości kwantyli zmiennej pomocniczej na poziomie populacji. Charakteryzują się one najmniejszym obciążeniem, wariancją i względnym błędem szacunku. Możliwość wykorzystania dodatkowych informacji pozwala zatem – w istotny sposób – zwiększyć precyzję przy jednoczesnej redukcji obciążenia będącego skutkiem braków danych. Zależność taka była zauważalna we wszystkich rozpatrywanych wariantach oraz dla różnych wielkości prób i frakcji braków.

Warto podkreślić, że estymatory kalibracyjne mediany, podobnie jak estymatory kalibracyjne średniej, wykazywały swoją wyższość nad estymatorem bezpośrednim w sytuacjach, gdy braki odpowiedzi nie miały charakteru losowego, a frakcja braków była duża.

4.5. Wnioski

Podsumowując wyniki obydwu przeprowadzonych w rozdziale czwartym badań symulacyjnych nad własnościami rozważanych w pracy estymatorów kalibracyjnych, można zauważyć pewne wspólne elementy:

- Obciążenie estymatorów spada wraz ze wzrostem liczebności próby.
- Przy ustalonej wielkości próby oraz frakcji braków danych najmniejszą tendencję do przeszacowywania lub niedoszacowywania analizowanego parametru zauważyć można w sytuacji, gdy braki danych mają charakter losowy. W innym przypadku obciążenie rośnie – w mniejszym jednak stopniu dla estymatorów kalibracyjnych.
- W sytuacji, gdy braki danych przypisywano jednostkom o najniższych wartościach analizowanej cechy, oceny uzyskane dla rozważanych estymatorów były przeszacowane w stosunku do „prawdziwej" wartości parametru – w znacznie mniejszym jednak stopniu dla estymatorów kalibracyjnych. Obciążenie to rosło wraz ze wzrostem frakcji braków danych.
- W sytuacji, gdy braki danych przypisywano jednostkom o najwyższych wartościach analizowanej cechy, oceny uzyskane dla rozważanych estymatorów były niedoszacowane w stosunku do „prawdziwej" wartości parametru – w znacznie mniejszym jednak stopniu dla estymatorów kalibracyjnych. Niedoszacowanie to było tym większe, im wyższa była frakcja braków danych.
- Estymatory kalibracyjne charakteryzują się mniejszą wariancją aniżeli estymator bezpośredni. Największą wariancję mają rozważane estymatory, gdy braki danych mają charakter losowy. Wariancja estymatorów rośnie wraz ze wzrostem frakcji braków danych, w mniejszym jednak stopniu dla estymatorów kalibracyjnych.
- Przewaga estymatorów kalibracyjnych nad bezpośrednim jest szczególnie zauważalna, kiedy mechanizm generujący braki danych nie jest przypadkowy. Jest to szczególnie ważne, gdyż w badaniach społecznych przeprowadzanych przez GUS, często uchylają się od udzielania odpowiedzi jednostki o najniższych i najwyższych wartościach badanej cechy.

Wyniki badań symulacyjnych potwierdziły, że na obciążenie estymatorów i ich wariancję wpływ ma odpowiednie wykorzystanie informacji dodatkowych w postaci wartości globalnych bądź kwantyli zmiennych pomocniczych.

Wśród rozważanych estymatorów kalibracyjnych najbardziej pożądanymi własnościami – w przypadku szacowania średniej – charakteryzował się estymator kalibracyjny regresyjny, a dla mediany uogólniony estymator kalibracyjny.

Dokonana w rozdziale czwartym, na podstawie symulacji ocena własności estymatorów kalibracyjnych, w połączeniu z informacją o dostępnych źródłach danych, stanowiła podstawę decyzji wyboru najodpowiedniejszych metod szacunku wybranych kategorii wydatków gospodarstw domowych, w przekroju powiatów województwa wielkopolskiego, w przeprowadzonym – w ostatnim rozdziale badaniu.

Rozdział 5

Empiryczna ocena rozważanych estymatorów kalibracyjnych w badaniu budżetów gospodarstw domowych

5.1. Badanie budżetów gospodarstw domowych jako pole zastosowań estymatorów kalibracyjnych

Po przedstawieniu własności estymatorów kalibracyjnych oraz analitycznych rozważaniach, zawartych w rozdziale drugim i trzecim, przejdziemy do ich empirycznej weryfikacji na przykładzie budżetów gospodarstw domowych w województwie wielkopolskim. Zaznaczmy przy tym, że jest to pierwsze zastosowanie estymatorów kalibracyjnych w odniesieniu do budżetów gospodarstw domowych w Polsce, co stwierdzamy dla podkreślenia pionierskości zastosowań. Nie ma bowiem innych przykładów zastosowań kalibracji, które mogłyby wzmocnić bądź osłabić wynikające z pracy wnioski.

Badanie budżetów gospodarstw domowych stanowi jedno z podstawowych źródeł danych o warunkach i sytuacji bytowej poszczególnych grup ludności. Dostarcza wielu informacji o przychodach, rozchodach, spożyciu i wyposażeniu w dobra trwałego użytkowania gospodarstw domowych. Umożliwia przeprowadzanie różnego rodzaju analiz nad zróżnicowaniem warunków bytowych grup społeczno-ekonomicznych ludności oraz wskazanie przyczyn, które determinują powstawanie tych różnic.

Wyniki badań budżetów gospodarstw domowych znajdują szerokie zastosowanie. Stanowią nie tylko podstawę oceny sytuacji bytowej poszczególnych grup ludności, ale są również niezbędne w procesie kreowania oraz monitorowania efektów polityki społeczno-ekonomicznej w skali kraju – jak i w kontekście przynależności Polski do Unii Europejskiej. Wykorzystywane są w opracowywaniu prognoz ekonomicznych dotyczących spożycia indywidualnego, opracowywania modeli dotyczących obciążeń podatkowych gospodarstw domowych oraz ustalania minimalnego wynagrodzenia. Służą

również jako wagi do obliczania indeksów cen towarów i usług konsumpcyjnych oraz stanowią ważne źródło informacji na temat skali ubóstwa w Polsce.

Jednostką badania jest gospodarstwo domowe jedno- lub wieloosobowe. Przedmiotem badania jest jego budżet rozumiany jako systematyczne zestawienie przychodów i rozchodów (pieniężnych i niepieniężnych) za dany okres. W badaniu zbiera się również informacje o wielkości spożycia wybranych artykułów oraz korzystania z różnych usług, cechach demograficznych i społeczno-ekonomicznych osób wchodzących w skład gospodarstwa domowego, ich aktywności ekonomicznej, rodzaju wykonywanej pracy, użytkowaniu gruntów, wyposażeniu w dobra trwałego użytkowania i nieruchomości, a także o subiektywnej ocenie sytuacji materialnej gospodarstwa domowego.

Badanie budżetów gospodarstw domowych ma w Polsce ponad 50-letnią tradycję. Było już przeprowadzane w okresie międzywojennym i powojennym, ale dopiero począwszy od 1957 roku prowadzone jest co rok i obejmuje większą liczbę respondentów. Przykładowo, w 1957 roku udział w badaniu wzięło 1 200 gospodarstw, podczas gdy w 2006 roku było ich już 37 508. Zmianie ulegały również metody samego badania. Początkowo stosowano podejście branżowe, które obejmowało wybrane grupy zawodowe, a jedno gospodarstwo domowe brało w nim udział nawet przez okres 4–5 lat. Obecnie stosowane jest podejście terytorialne, obejmujące swym zasięgiem wszystkie województwa i różne typy gospodarstw domowych, przy czym jedno gospodarstwo bierze w nim udział przez okres jednego miesiąca. Podstawową metodą zbierania informacji jest wywiad osobisty i prowadzenie książeczek budżetowych. Do zbierania danych w ramach wywiadu wykorzystuje się kwestionariusze: kartę statystyczną gospodarstwa domowego, informacje o gospodarstwie nieprzystępującym do badania oraz informacje uzupełniające o gospodarstwie domowym.

Badanie budżetów gospodarstw domowych jest przeprowadzane metodą reprezentacyjną. Istnieje zatem możliwość uogólniania uzyskanych wyników na wszystkie gospodarstwa domowe w kraju. W 2002 roku zastosowano terytorialny, warstwowy, dwustopniowy schemat losowania z różnymi prawdopodobieństwami wyboru na I stopniu. Jednostkami losowania pierwszego stopnia były terenowe punkty badań, natomiast na drugim stopniu losowane były mieszkania[40].

Operat losowania jednostek pierwszego stopnia stanowiły wykazy rejonów statystycznych, które zostały opracowane dla potrzeb Narodowego Spisu Powszechnego, a aktualizowane każdego roku o zmiany wynikające z podziału administracyjnego kraju, z nowego budownictwa, wyburzania starych budynków itd.

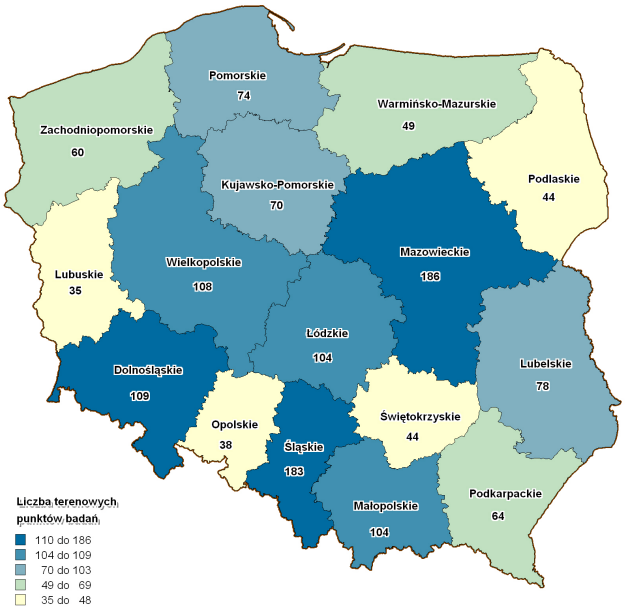
W 2002 roku w badaniu budżetów gospodarstw domowych wykorzystywane były dwie podpróbki liczące po 675 terenowych punktów badań, por. rysunek 5.1:

- podpróbka 1 - wylosowana w 2001 roku do badania na lata 2002–2003,
- podpróbka 2 - wylosowana w 2000 roku do badania na lata 2001–2002.

W badaniu budżetów gospodarstw domowych przyjęto zasadę, że terenowy punkt badań w mieście powinien liczyć co najmniej 250 mieszkań, zaś wiejski 150. Wymagało to łączenia niektórych rejonów statystycznych z sąsiednimi – celem spełnienia tego założenia. Przed losowaniem mieszkań, terenowe punkty badań zostały powarstwowane,

[40] Metodologia losowania próby przedstawiona została w oparciu o publikowane przez GUS opracowania dotyczące badania budżetów gospodarstw domowych, por. GUS (2003b).

w ramach każdego województwa, według klas miejscowości stanowiących miasta pogrupowane według liczby mieszkańców oraz wieś. Przyjęto również założenie, że każde losowane mieszkanie (a więc i gospodarstwo domowe) powinno mieć jednakowe prawdopodobieństwo dostania się do próby. W związku z tym, liczba losowanych terenowych punktów badań z danej warstwy, była proporcjonalna do liczby mieszkań w warstwie. Z każdego terenowego punktu badań do próby losowano jednakową liczbę mieszkań.



Rysunek. 5.1. Liczba terenowych punktów badań według województw w 2002r.

Źródło: Opracowanie własne

Operat losowania drugiego stopnia stanowiły z kolei wykazy zamieszkanych mieszkań w ramach wylosowanych terenowych punktów badań. Losowanie mieszkań przeprowadzono w oparciu o następujące założenia:

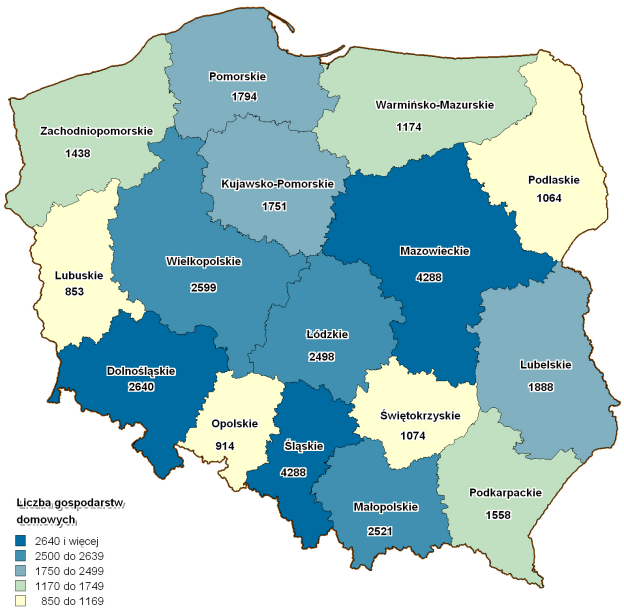
1. W badaniu zastosowano model rotacji całkowitej z miesięcznym okresem wymiany próby[41].
2. W każdym miesiącu losuje się po dwa mieszkania, w ramach danego terenowego punktu badań, a w wylosowanych mieszkaniach w badaniu udział biorą wszystkie zamieszkujące je gospodarstwa domowe.

[41] Do 1982 roku stosowano w badaniu budżetów gospodarstw domowych metodę ciągłą, co oznaczało, że te same gospodarstwa domowe podlegały badaniu przez rok i dłużej. W latach 1982–1993 wykorzystano metodę rotacji kwartalnej. Stosowana począwszy od 1993 roku metoda rotacji miesięcznej oznacza, że w ciągu roku inne gospodarstwo domowe w każdym miesiącu bierze udział w badaniu i prowadzi zapisy w specjalnie przeznaczonych do tego celu książeczkach budżetowych.

3. W wylosowanym mieszkaniu badanie przeprowadzane jest w danym miesiącu w ciągu dwóch kolejnych lat (w 2002–2003 w podpróbce pierwszej i 2001–2002 w podpróbce drugiej).
4. W ramach każdego terenowego punktu badań losuje się rezerwową próbę mieszkań. Jest to konsekwencją nie przystępowania do badania gospodarstw domowych zamieszkujących wylosowane mieszkanie. Zastępowanie gospodarstw domowych, nie biorących udziału w badaniu, odbywa się w oparciu o gospodarstwa zamieszkujące mieszkania z próby rezerwowej według kolejności ich wylosowania.

Biorąc pod uwagę powyższe założenia, wylosowano w każdym terenowym punkcie badań po 24 mieszkania (co daje łączną liczbę 32 400 mieszkań) oraz próbę rezerwową liczącą 150 mieszkań.

Nie wszystkie gospodarstwa domowe, zamieszkujące 32 400 mieszkań wylosowanych do próby, wzięły jednak udział w badaniu. W związku z powyższym, zastosowano sekwencyjną procedurę zamiany gospodarstw domowych z próby rezerwowej. Losowy sposób zamiany gospodarstw, które nie podjęły badania polegał na dobieraniu kolejnych mieszkań z losowo uporządkowanego ciągu mieszkań tak długo, aż gospodarstwa domowe z dwóch mieszkań, w ramach danego terenowego punktu badań, wyraziły chęć uczestniczenia w badaniu. Losowa zamiana gospodarstw domowych oznaczała zatem, że gospodarstwo zastępujące gospodarstwo nieprzystępujące do badania, mogło charakteryzować się zupełnie innymi cechami społeczno-demograficznymi (skład osobowy, źródło utrzymania itd.).



Rysunek. 5.2. Liczba gospodarstw domowych w przekroju województw zbadanych w 2002r.

Źródło: Opracowanie własne

Ostatecznie w 2002 roku w badaniu udział wzięło 32 342 gospodarstwa domowe, z czego w województwie wielkopolskim ich liczba wyniosła 2 599, por. rysunek 5.2.

Istotnym zagadnieniem, z punktu widzenia jakości danych, jest skala i przyczyny niepodjęcia badania przez gospodarstwa domowe wylosowane do próby. W 2002 roku, nawet po uwzględnieniu efektu zastępowania mieszkań, nie uzyskano żadnej informacji (o źródle utrzymania, składzie osobowym) od wielu gospodarstw domowych. Ich udział w całkowitej liczbie gospodarstw domowych wylosowanych do badania budżetów wyniósł 17,4%. Najczęstszą przyczyną niepodjęcia przez wylosowane gospodarstwo domowe udziału w badaniu były odmowy, wiek bądź choroba.

Ze względu na dużą frakcję gospodarstw domowych nie przystępujących do badania, struktura próby zbadanej, ze względu na cechy społeczno-demograficzne, różni się w istotny sposób od struktury próby wylosowanej. Główny Urząd Statystyczny, celem redukcji obciążenia wynikającego z niepodejmowania badania, przeważa uzyskane wyniki przez wskaźniki odnoszące się do struktury gospodarstw domowych według liczby osób pochodzących z badania aktywności ekonomicznej ludności.

Istotnym problemem są również braki odpowiedzi dla wielu kategorii wydatków w odniesieniu do gospodarstw, które przystąpiły do badania. W wielu przypadkach frakcja braków odpowiedzi przekracza 50%. Innym źródłem błędów nielosowych, jak wykazuje praktyka badań reprezentacyjnych prowadzonych w Polsce i w innych krajach, a opartych na oświadczeniach osób badanych, są zaniżane celowo niektóre pozycje jak: wydatki na alkohol, tytoń, wyroby cukiernicze, żywienie w placówkach gastronomicznych itp. Stąd tak ważna rola omawianych w pracy metod, które zniwelować mogą ujemny wpływ błędów nielosowych na jakość uzyskanych wyników.

W kolejnym podrozdziale podjęta zostanie empiryczna ocena rozważanych w pracy estymatorów kalibracyjnych dla niektórych kategorii wydatków gospodarstw domowych.

5.2. Empiryczna ocena estymatorów kalibracyjnych średnich wydatków gospodarstw domowych

Głównym celem przeprowadzonych analiz, w ramach tego podrozdziału, było oszacowanie średnich wydatków gospodarstw domowych na niektóre kategorie towarów i usług konsumpcyjnych w przekroju powiatów województwa wielkopolskiego. Jako zmienne badane przyjęto wydatki na napoje alkoholowe, wyroby tytoniowe i narkotyki oraz wydatki na łączność. Ich wybór wynikał z faktu, że występujące tutaj braki odpowiedzi charakteryzowały się znaczną dyspersją. W przypadku wydatków na napoje alkoholowe, wyroby tytoniowe i narkotyki dodatkowo uwzględniono fakt, że gospodarstwa domowe celowo zaniżają tę pozycję, por. GUS (2003b). Jako zmienną pomocniczą, ze względu na silne skorelowanie, przyjęto dochód gospodarstw domowych ogółem.

Na potrzeby badania wykorzystano trzy typy estymatorów wartości globalnej: estymator bezpośredni (por. wzór 2.3), estymator kalibracyjny ze znanym wektorem $\mathbf{\check{X}}$ oszacowanych wartości globalnych zmiennych pomocniczych (por. wzór 2.34) oraz estymator syntetyczny ilorazowy (por. wzór 5.1). W pierwszej kolejności oszacowano wartości globalne analizowanych wydatków. Następnie, w oparciu o informacje pochodzące z NSP'2002 o liczbie gospodarstw domowych dla poszczególnych powiatów, dokonano oszacowania średnich wydatków przypadających na gospodarstwo domowe.

Informacje tę wykorzystano również celem zapewnienia sumowalności wag kalibracyjnych do liczby gospodarstw domowych w odpowiednich powiatach.

Wybór estymatora kalibracyjnego $\hat{Y}_{\tilde{\mathbf{x}}}$ podyktowany był brakiem informacji o wartości globalnej dochodów wszystkich gospodarstw domowych z danego powiatu. Wybór estymatora bezpośredniego i syntetyczno ilorazowego wynikał natomiast z konieczności porównania uzyskanych wyników dla analizowanych grup wydatków.

W podejściu zaprezentowanym w pracy, celem wyznaczenia ocen syntetycznego estymatora ilorazowego, połączono podobne domeny (powiaty województwa wielkopolskiego) w większe obszary (skupiska). W estymacji syntetycznej przyjmuje się bowiem założenie, że mały obszar (domena) jest podobny do większego obszaru – zazwyczaj zawierającego w sobie ten pierwszy. Takie założenie pozwala na podział oszacowanej wartości globalnej dla obszaru większego na części odpowiadające poszczególnym małym obszarom (domenom), por. G. Dehnel (2003), E. Gołata (2004). Ponieważ założenia estymacji syntetycznej są często w praktyce niemożliwe do zrealizowania, stosownym wydaje się łączenie podobnych powiatów w grupy, co może pozwolić uniknąć sztywnych założeń o identycznej relacji pomiędzy zmienną szacowaną Y, a zmienną pomocniczą X. Na potrzeby pracy, korzystając z hierarchicznej metody Warda, dokonano typizacji powiatów ze względu na sytuację panującą na rynku pracy, gdyż ma ona wpływ na strukturę wydatków i dochodów gospodarstw domowych. Jako zmienne diagnostyczne opisujące rynek pracy w powiatach w 2002 roku przyjęto, por. A. Witkowska, M. Witkowski (2006):

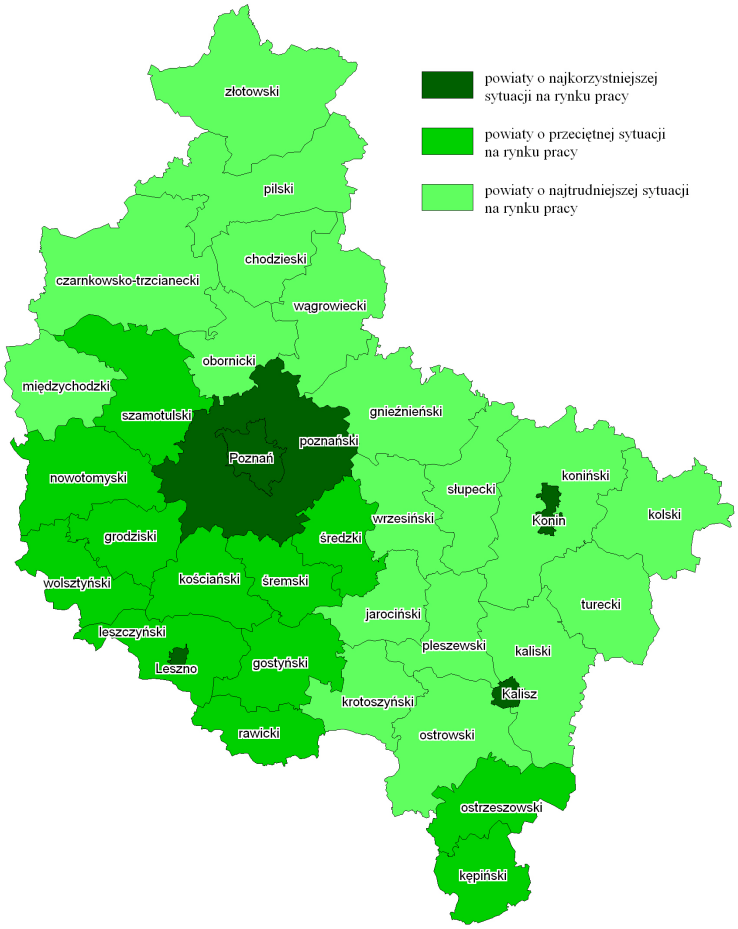
- udział procentowy zatrudnionych w sektorze prywatnym w liczbie zatrudnionych ogółem – (stymulanta),
- przeciętne miesięczne wynagrodzenie (brutto) – (stymulanta),
- udział procentowy bezrobotnych do 25 roku życia w ogólnej liczbie bezrobotnych – (destymulanta),
- udział procentowy długotrwale bezrobotnych w ogólnej liczbie bezrobotnych – (destymulanta),
- udział procentowy bezrobotnych bez stażu lub ze stażem do 1 roku w ogólnej liczbie bezrobotnych – (destymulanta),
- udział procentowy bezrobotnych z wykształceniem wyższym w ogólnej liczbie bezrobotnych – (destymulanta).

Ocenę estymatora syntetycznego ilorazowego wartości globalnej wyznaczono ze wzoru:

$$\hat{Y}_{syn,d} = \frac{\sum\limits_{i \in d} d_i x_i \sum\limits_{i \in Sk} d_i y_i}{\sum\limits_{i \in Sk} d_i x_i}, \tag{5.1}$$

gdzie x_i oznacza wartość zmiennej pomocniczej dla i – tej jednostki badania, d – domenę (mały obszar), a Sk skupisko obejmujące podobne domeny ze względu na analizowany zestaw cech.

W wyniku zastosowania hierarchicznej metody Warda, wyodrębniono trzy skupiska zawierające powiaty o podobnej sytuacji panującej na rynku pracy w województwie wielkopolskim w 2002 roku, por. rysunek 5.3.



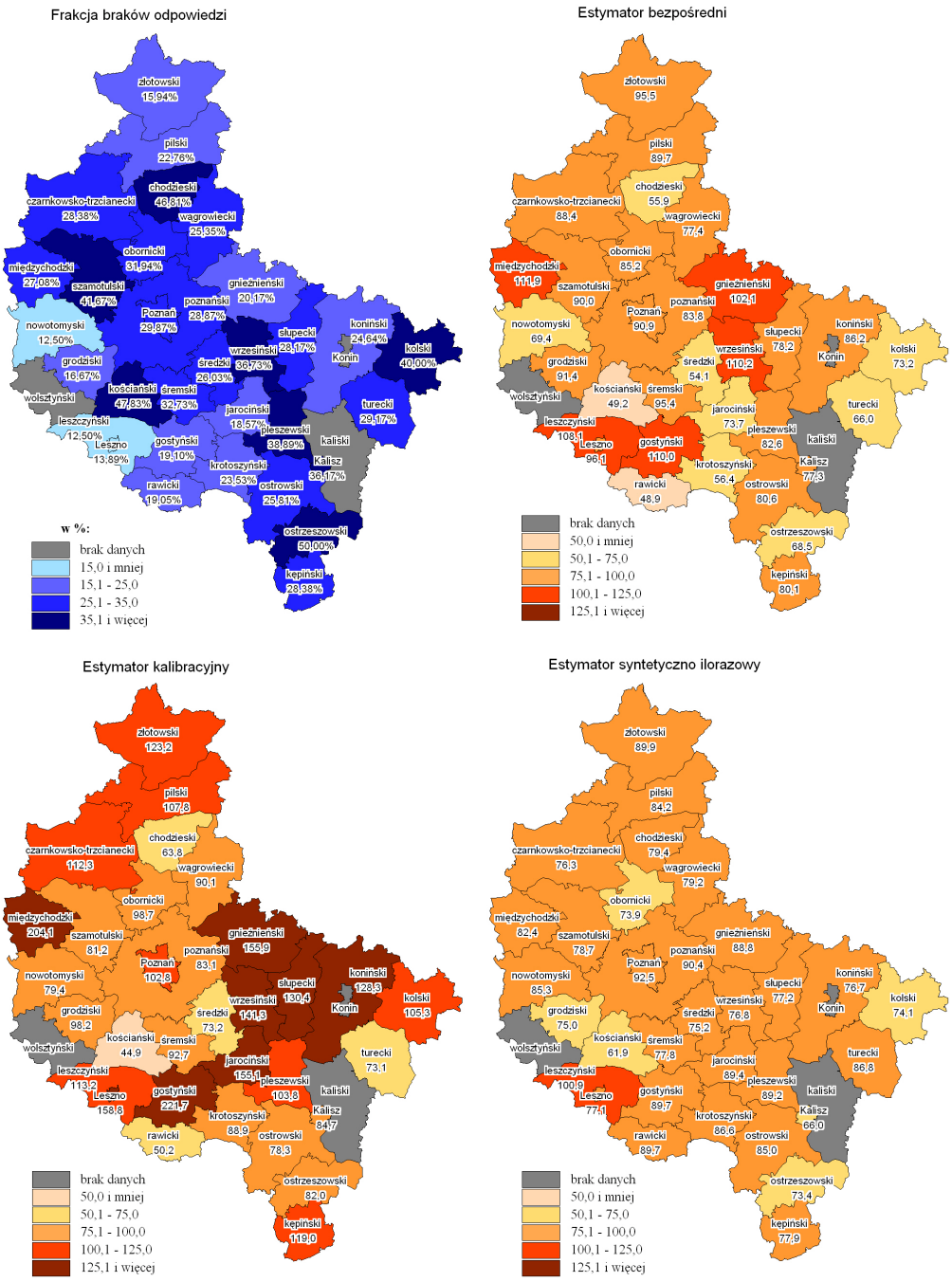
Rysunek. 5.3. Powiaty podobne pod względem sytuacji panującej na rynku pracy w województwie wielkopolskim w 2002r.

Źródło: Opracowanie własne

Pierwsze skupisko obejmuje powiaty, gdzie sytuacja na rynku pracy była najkorzystniejsza. Oprócz powiatu poznańskiego w jego skład weszły miasta na prawach powiatu tj. Kalisz, Leszno, Konin i Poznań. Drugie skupisko obejmowało w zasadzie powiaty leżące w południowej i południowo-zachodniej części województwa wielkopolskiego – w sumie 12 powiatów. Ostanie – trzecie skupisko – tworzyły powiaty o najtrudniejszej sytuacji na rynku pracy. Generalnie, swoim zasięgiem obejmowało ono powiaty wschodniej i północnej części województwa (łącznie 18 powiatów).

Na rysunkach 5.4 i 5.5 zestawiono wyniki estymacji wydatków gospodarstw domowych na napoje alkoholowe, wyroby tytoniowe i narkotyki w przekroju powiatów województwa wielkopolskiego w 2002 roku[42]. Występujące tutaj braki odpowiedzi, charakteryzowały się znaczną dyspersją: od 12,5% w nowotomyskim i leszczyńskim do

[42] W 2002 roku w badaniu budżetów gospodarstw domowych nie zostało wylosowane do próby żadne gospodarstwo z trzech powiatów: kaliskiego, wolsztyńskiego i Konina. Powiaty te zaznaczono na wykresach mapowych szarym kolorem. Powiaty na wykresach 5.5, 5.7, 5.10 oraz 5.13 uporządkowano według rosnących wartości frakcji braków odpowiedzi.



Rysunek. 5.4. Frakcja braków odpowiedzi i oszacowania średnich wydatków na napoje alkoholowe, wyroby tytoniowe i narkotyki (w PLN) gospodarstw domowych z wykorzystaniem estymatora bezpośredniego, kalibracyjnego i syntetycznego ilorazowego w przekroju powiatów województwa wielkopolskiego w 2002r.

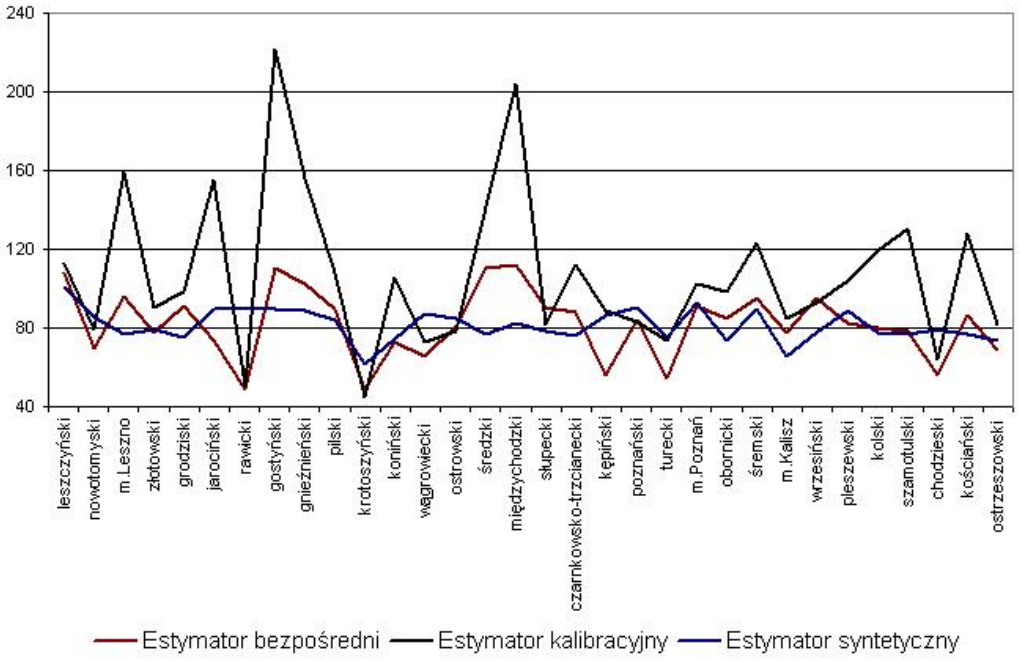
Źródło: Opracowanie własne

50% braków odpowiedzi w powiecie ostrzeszowskim. Pod względem braku odpowiedzi nie widać jednak większych różnic między poszczególnymi częściami Wielkopolski. Zasadniczo nie różnicuje powiatów ani wielkość powiatów, ani ich charakter, ani historia (zabór pruski, zabór rosyjski).

Oprócz frakcji braków odpowiedzi, rysunek 5.4 przedstawia natężenie wydatków na napoje alkoholowe, wyroby tytoniowe i narkotyki w PLN. Jeśli zaufać szacunkom na podstawie estymatora kalibracyjnego, to najwięcej się wydaje na ten rodzaj dóbr w powiecie gostyńskim i międzychodzkim – ponad 200 zł, najmniej w powiecie kościańskim, rawickim i chodzieskim – poniżej 65 zł.

Estymator kalibracyjny znacznie pogłębia różnice między powiatami, por. rysunek 5.4. Tylko w nielicznych powiatach estymator kalibracyjny pokazuje takie same lub zbliżone oceny do estymatora bezpośredniego i syntetycznego.

Na rysunku 5.5 rzuca się w oczy dość duża zbieżność ocen estymatora syntetycznego i bezpośredniego. Oceny estymatora kalibracyjnego znacznie od nich odbiegają. Dla wielu powiatów (gnieźnieńskiego, gostyńskiego, jarocińskiego, konińskiego, Leszna, międzychodzkiego, słupeckiego i wrzesińskiego) ocena estymatora kalibracyjnego wydatków na alkohol, wyroby tytoniowe i narkotyki, jest znacznie wyższa od szacunków uzyskanych z wykorzystaniem estymatora bezpośredniego i syntetycznego.



Rysunek. 5.5. Porównanie ocen estymatora bezpośredniego, kalibracyjnego i syntetycznego ilorazowego średnich wydatków na napoje alkoholowe, wyroby tytoniowe i narkotyki (w PLN) w przekroju powiatów województwa wielkopolskiego w 2002r.

Źródło: Opracowanie własne

Na podstawie tego, co wiadomo z analizy wyników badań symulacyjnych z rozdziału czwartego, większym zaufaniem bylibyśmy skłonić obdarzyć oceny estymatora kalibracyjnego. Wynika to również z faktu, że gospodarstwa domowe celowo zaniżają wydatki na używki, por. GUS (2003b). Przy tym różnice byłyby jeszcze większe,

gdyby można było uwzględnić w estymatorze kalibracyjnym różnorodność zachowań konsumpcyjnych respondentów i nierespondentów. Można bowiem przypuszczać, że wśród nierespondentów więcej wydaje się na napoje alkoholowe, wyroby tytoniowe i narkotyki.

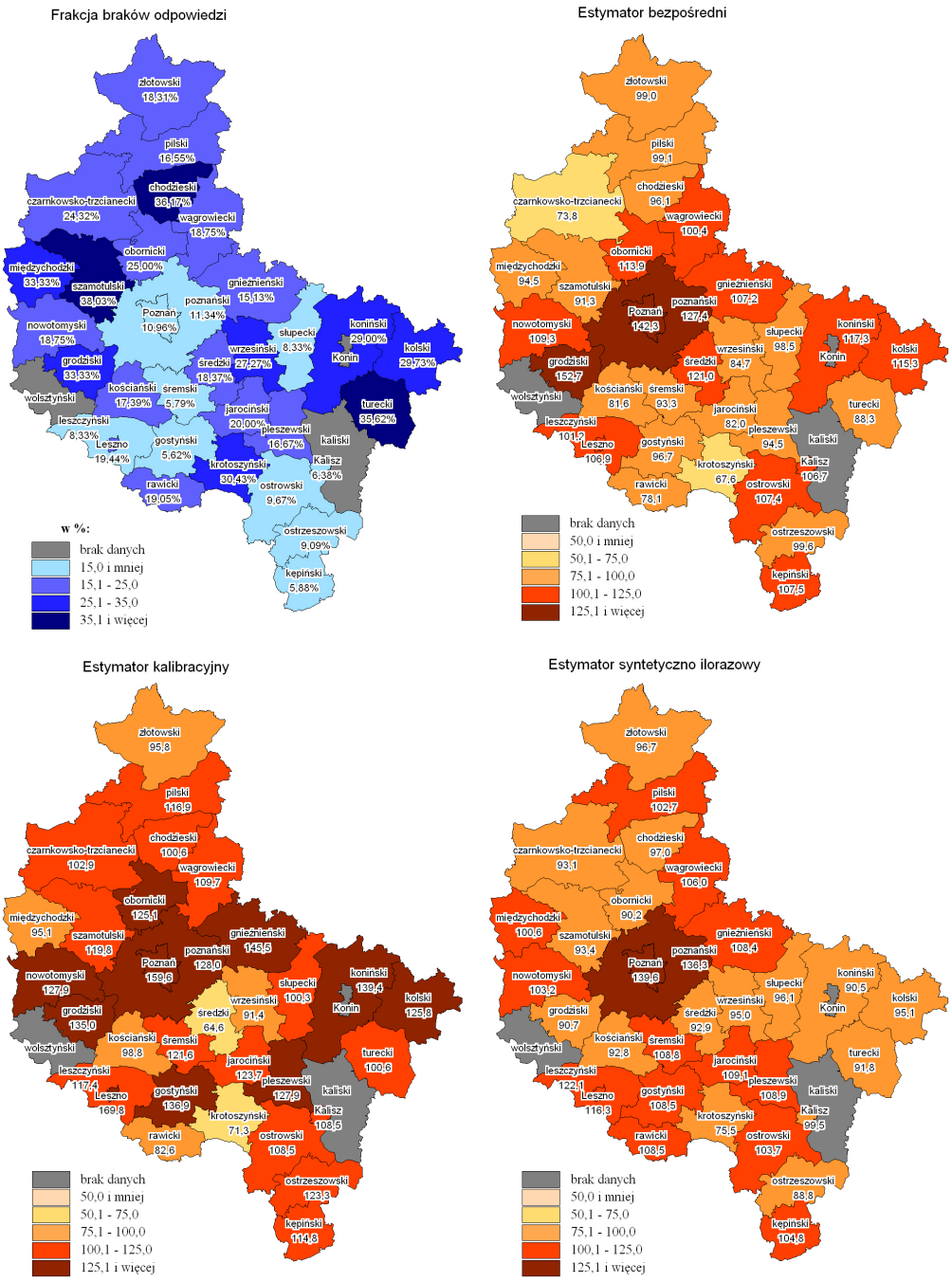
Drugą z analizowanych cech były wydatki na łączność, które obejmowały wydatki na usługi pocztowe, sprzęt i usługi telekomunikacyjne.

Występujące tutaj braki odpowiedzi, podobnie jak w przypadku wydatków na alkohol, wyroby tytoniowe i narkotyki, charakteryzowały się znaczną dyspersją: od 5,6% w powiecie gostyńskim do 38% w powiecie szamotulskim. Przestrzenna analiza zróżnicowania braków odpowiedzi prowadzi do wniosku, że nie występują większe różnice między poszczególnymi rejonami Wielkopolski, por. rysunek 5.6. Można jednak zauważyć, że mniejszy odsetek braków odpowiedzi wydatków na łączność jest dostrzegalny w miastach na prawach powiatu (Poznań, Kalisz). Podobnie jak dla wydatków na używki, bardzo dużą frakcję braków odpowiedzi zaobserwowano w powiecie szamotulskim i chodzieskim. Dla większości jednak powiatów frakcja braków odpowiedzi na łączność była na niższym poziomie, w porównaniu z odsetkiem braków na napoje alkoholowe, wyroby tytoniowe i narkotyki. Związane to było najprawdopodobniej z drażliwością pytania o wydatki na używki.

Rysunek 5.6 przedstawia – oprócz frakcji braków odpowiedzi – także natężenie średnich wydatków gospodarstw domowych na łączność (w PLN) w przekroju powiatów województwa wielkopolskiego. Według ocen jakie daje estymator kalibracyjny najwięcej wydaje się na łączność w Lesznie i Poznaniu oraz w powiatach: obornickim, poznańskim, gnieźnieńskim, nowotomyskim, grodziskim, gostyńskim, pleszewskim, konińskim oraz kolskim. Znaczące różnice między oceną estymatora kalibracyjnego a bezpośredniego i syntetycznego można przede wszystkim zaobserwować w powiatach, w których frakcja braków odpowiedzi była na stosunkowo wysokim poziomie. Dotyczy to zwłaszcza powiatów: szamotulskiego, kolskiego, Leszna, konińskiego oraz jarocińskiego.

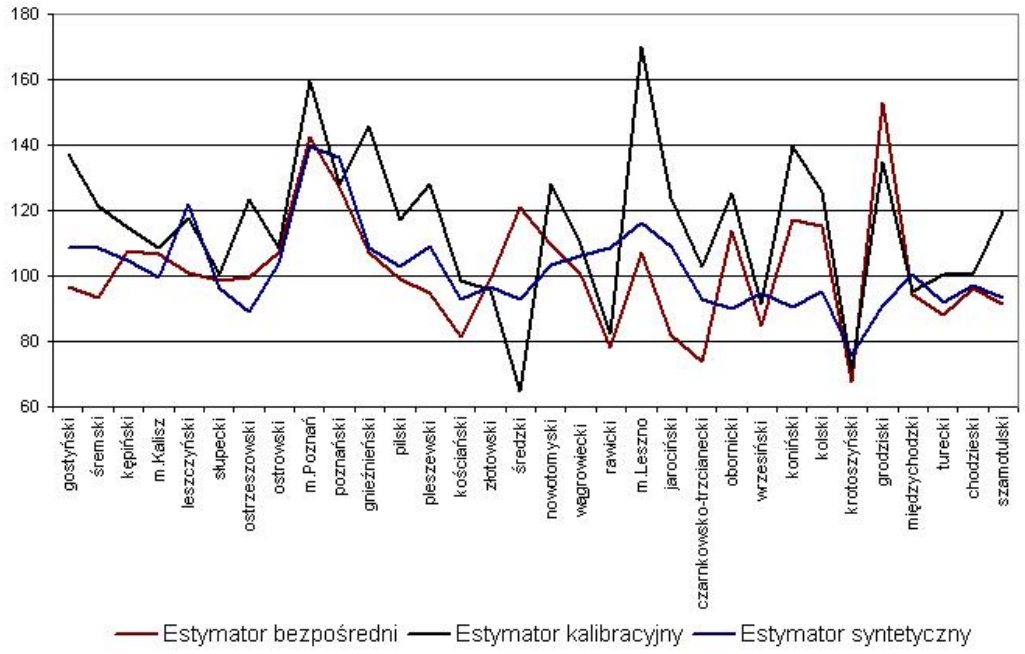
Estymator kalibracyjny pogłębia różnice między powiatami. Nie jest to jednak już tak wyraźne dla wydatków na łączność, jak w przypadku wydatków na alkohol, wyroby tytoniowe i narkotyki.

Na rysunku 5.7 można zauważyć, że występuje znaczna zbieżność ocen estymatora syntetycznego, bezpośredniego i kalibracyjnego dla wielu powiatów (Kalisz, słupecki, poznański, wągrowiecki, złotowski, wrzesiński, międzychodzki, ostrowski, chodzieski, krotoszyński oraz kępiński). Istnieją jednak powiaty, dla których ocena estymatora kalibracyjnego wydatków na łączność jest znacznie wyższa od szacunków uzyskanych z wykorzystaniem estymatora bezpośredniego i syntetycznego. Dotyczy to przede wszystkim powiatów: gnieźnieńskiego, konińskiego, pleszewskiego, gostyńskiego, ostrzeszowskiego, pilskiego, Leszna oraz Poznania. Wyjątek stanowi powiat średzki, dla którego ocena estymatora kalibracyjnego wydatków na łączność – jest znacznie niższa od ocen estymatora bezpośredniego i syntetycznego. Wynikać to może z faktu, że dla tego powiatu niektóre wagi kalibracyjne były bardzo małe w stosunku do wyjściowych wag wynikających ze schematu losowania próby.



Rysunek. 5.6. Frakcja braków odpowiedzi i oszacowania średnich wydatków na łączność (w PLN) gospodarstw domowych z wykorzystaniem estymatora bezpośredniego, kalibracyjnego i syntetycznego ilorazowego w przekroju powiatów województwa wielkopolskiego w 2002r.

Źródło: Opracowanie własne



Rysunek. 5.7. Porównanie ocen estymatora bezpośredniego, kalibracyjnego i syntetycznego ilorazowego średnich wydatków na łączność (w PLN) w przekroju powiatów województwa wielkopolskiego w 2002r.

Źródło: Opracowanie własne

5.3. Empiryczna ocena estymatorów kalibracyjnych mediany wydatków gospodarstw domowych

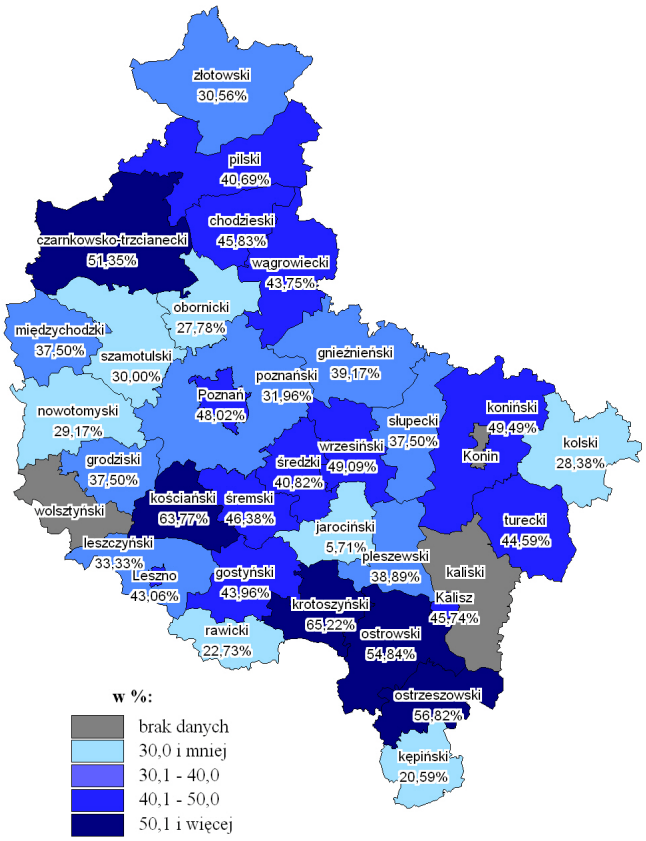
Głównym celem przeprowadzonych analiz, w ramach tego podrozdziału, było oszacowanie mediany wydatków gospodarstw domowych na niektóre kategorie towarów i usług konsumpcyjnych w przekroju powiatów województwa wielkopolskiego. Jako zmienne badane przyjęto wydatki gospodarstw domowych na energię elektryczną oraz wydatki na makaron. Podobnie jak w przypadku wydatków, dla których szacowano średnią, ich wybór wynikał z faktu, że braki odpowiedzi charakteryzowały się znaczną dyspersją i dla wielu powiatów występowały na dużą skalę. Jako zmienną pomocniczą, ze względu na silne skorelowanie, przyjęto także dochód gospodarstw domowych ogółem.

Na potrzeby badania wykorzystano dwa rodzaje estymatorów mediany: bezpośredni (por. wzór 3.83) oraz kalibracyjny ze znanym wektorem $\hat{Q}_{\boldsymbol{x},0.5}$ (por. wzór 3.85). Wybór estymatora kalibracyjnego mediany podyktowany był tym, że nieznane były mediany dochodów gospodarstw domowych z poszczególnych powiatów (na poziomie populacji). Dlatego dokonano ich oszacowania z wykorzystaniem estymatora bezpośredniego w oparciu o informacje pochodzące z badania budżetów gospodarstw domowych. W związku z tym, jedyną składową wektora $\hat{Q}_{\boldsymbol{x},0.5}$ była oszacowana na podstawie próby mediana dochodów gospodarstw domowych. Wybór estymatora bezpośredniego wynikał natomiast z potrzeby porównania uzyskanych wyników dla analizowanych grup wydatków.

W badaniu wykorzystano także informacje pochodzące z NSP'2002 o liczbie gospodarstw domowych w powiatach województwa wielkopolskiego. Dzięki temu spełnione

zostało równanie kalibracyjne zakładające, że wagi kalibracyjne sumują się do liczebności populacji, patrz rozdział 3, wzór 3.87.

Na rysunku 5.8 przedstawione zostały informacje na temat frakcji braków odpowiedzi na energię elektryczną w przekroju powiatów województwa wielkopolskiego w 2002 roku. Frakcja ta wahała się od 6% w przypadku powiatu jarocińskiego do 65% dla powiatu krotoszyńskiego. Z wyłączeniem jednak powiatu jarocińskiego, który stanowił wyjątek, frakcja ta przekraczała zazwyczaj 20%, a dla wielu powiatów nawet 50% (czarnkowsko-trzcianecki, kościański, krotoszyński, ostrowski i ostrzeszowski). Pod względem braków odpowiedzi, także dla wydatków na energię elektryczną, nie widać większych różnic między poszczególnymi częściami Wielkopolski.

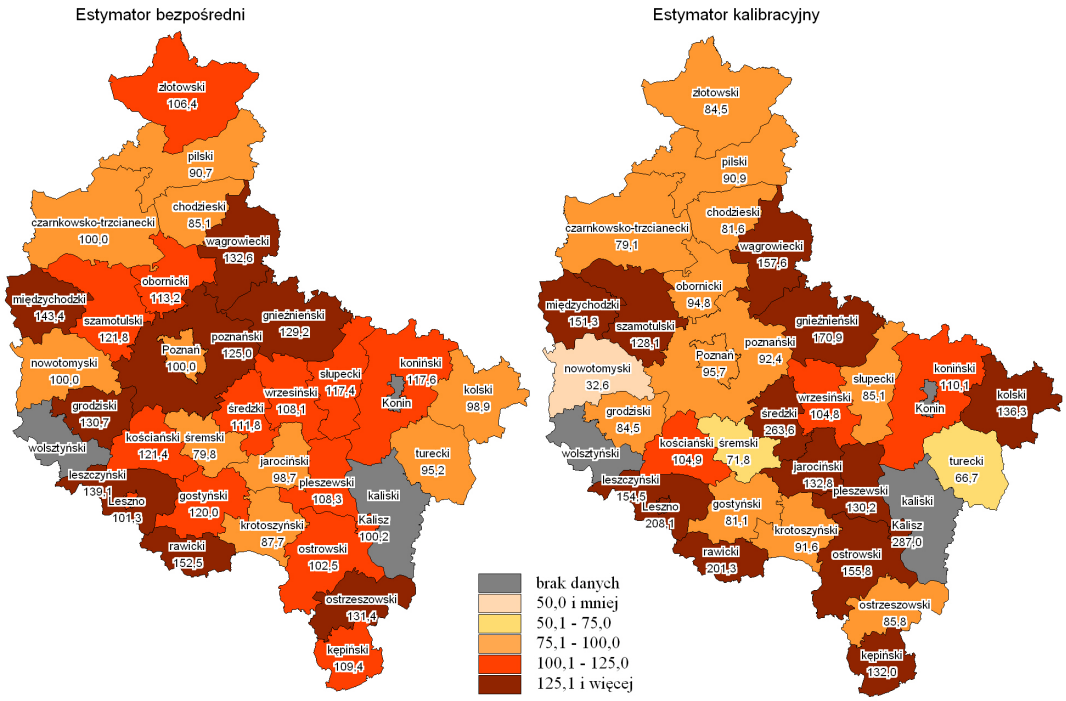


Rysunek. 5.8. Frakcja braków odpowiedzi wydatków na energię elektryczną w przekroju powiatów województwa wielkopolskiego w 2002r.

Źródło: Opracowanie własne

Rysunek 5.9 przedstawia ocenę estymatora kalibracyjnego i bezpośredniego mediany wydatków gospodarstw domowych na energię elektryczną (w PLN) w przekroju powiatów województwa wielkopolskiego. Gdyby za podstawę wyciąganych wniosków przyjąć medianę oszacowaną w oparciu o estymator kalibracyjny, to największą wartość przyjmuje ona w powiecie średzkim i Kaliszu. Oznacza to, że 50% gospodarstw

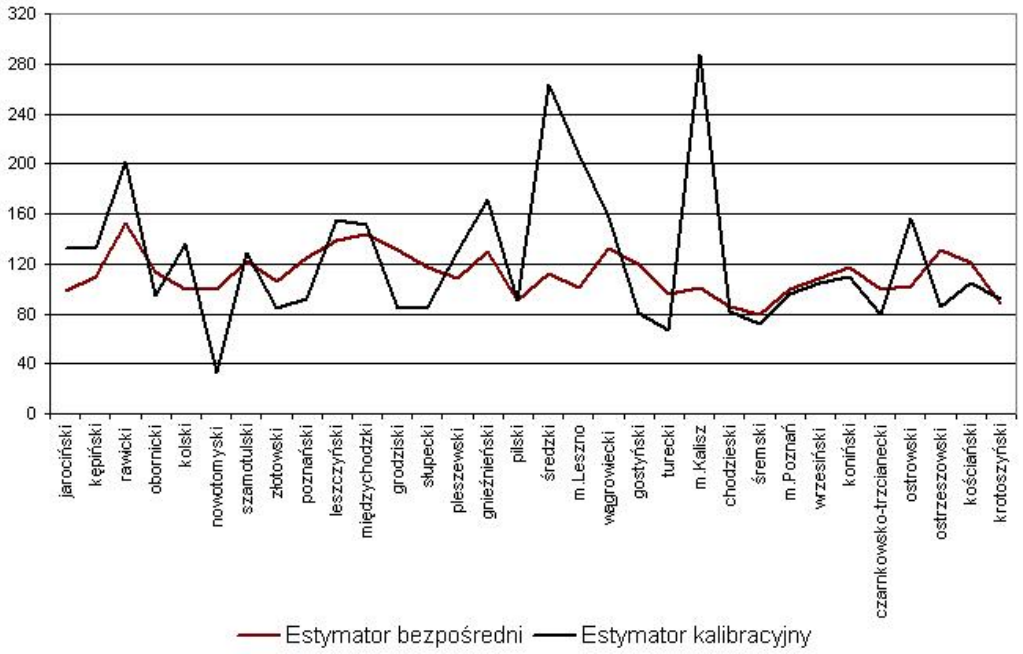
domowych w tych powiatach wydaje na energię elektryczną ponad 260 zł i 286 zł odpowiednio. Najmniejszą wartość mediana przyjmuje z kolei w powiecie nowotomyskim, z czego wynika, że 50% gospodarstw domowych ma wydatki na energię elektryczną nie większe niż około 33 zł.



Rysunek. 5.9. Mediana wydatków na energię elektryczną w przekroju powiatów województwa wielkopolskiego w 2002r.

Źródło: Opracowanie własne

Jak pokazuje rysunek 5.10 estymator kalibracyjny mediany pogłębia różnice między powiatami – choć dla kilku z nich daje zbliżone oceny do estymatora bezpośredniego (chodzieski, koniński, krotoszyński, międzychodzki, pilski, śremski, wrzesiński i Poznań). Estymator kalibracyjny w swych ocenach znacznie odbiega od ocen estymatora bezpośredniego przede wszystkim dla powiatów, w których frakcja braków odpowiedzi jest stosunkowo duża (średzki, Kalisz, ostrowski). Dla powiatów, gdzie frakcja braków odpowiedzi była stosunkowo niższa, oceny wydatków na energię elektryczną uzyskane w oparciu o estymator kalibracyjny i Horvitza-Thompsona – były zbliżone (z wyjątkiem nowotomyskiego). Różnica dla tego powiatu wynikała najprawdopodobniej z faktu, że niektóre z uzyskanych dla niego wag kalibracyjnych były bardzo małe w stosunku do wyjściowych wag wynikających ze schematu losowania próby.



Rysunek. 5.10. Porównanie ocen estymatora bezpośredniego i kalibracyjnego mediany wydatków na energię elektryczną w przekroju powiatów województwa wielkopolskiego w 2002r.

Źródło: Opracowanie własne

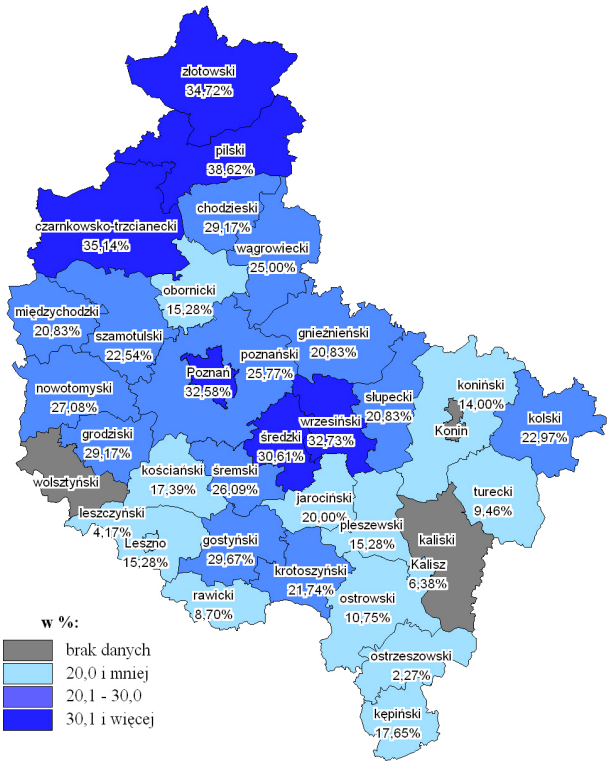
Drugą z analizowanych cech były wydatki na makaron stanowiące jedną z wielu pozycji wydatków gospodarstw domowych na żywność.

Występujące tutaj braki odpowiedzi, podobnie jak w przypadku wydatków na energię elektryczną, charakteryzowały się znaczną dyspersją: od 2% w powiecie ostrzeszowskim do 39% w powiecie pilskim, por. rysunek 5.11. Dla większości jednak powiatów, frakcja braków odpowiedzi na makaron, była na niższym poziomie w porównaniu z odsetkiem braków na energię elektryczną. Jedynie w przypadku powiatów: złotowskiego, pilskiego, czarnkowsko-trzcianeckiego, wrzesińskiego, średzkiego oraz Poznania braki odpowiedzi przekraczały 30%. Dla znacznej liczby powiatów frakcja braków odpowiedzi wydatków na makaron nie przekraczała jednak 20%. Dotyczyło to w zasadzie powiatów zlokalizowanych w południowej części województwa wielkopolskiego (kępiński, ostrzeszowski, ostrowski, pleszewski, jarociński, rawicki, kościański, leszczyński i Leszno).

Przyjmując za podstawę wyciąganych wniosków medianę wydatków na makaron, oszacowaną w oparciu o estymator kalibracyjny, można zauważyć, że największą wartość przyjmuje ona w powiecie poznańskim, ostrzeszowskim i ostrowskim.

Najmniejszą wartość mediana przyjmuje z kolei w powiecie gnieźnieńskim i gostyńskim. Jak pokazuje rysunek 5.12 dla wielu powiatów występują znaczne różnice w ocenach uzyskanych za pomocą obydwu analizowanych estymatorów. Jest to zwłaszcza widoczne na rysunku 5.13.

Estymator kalibracyjny mediany pogłębia różnice między powiatami. Tylko dla kilku powiatów oceny estymatorów są zbieżne. Dotyczy to w zasadzie powiatów: grodziskiego, kościańskiego, krotoszyńskiego, leszczyńskiego, słupeckiego, śremskiego oraz wągrowieckiego. W przypadku powiatów: konińskiego, międzychodzkiego, nowotomy-

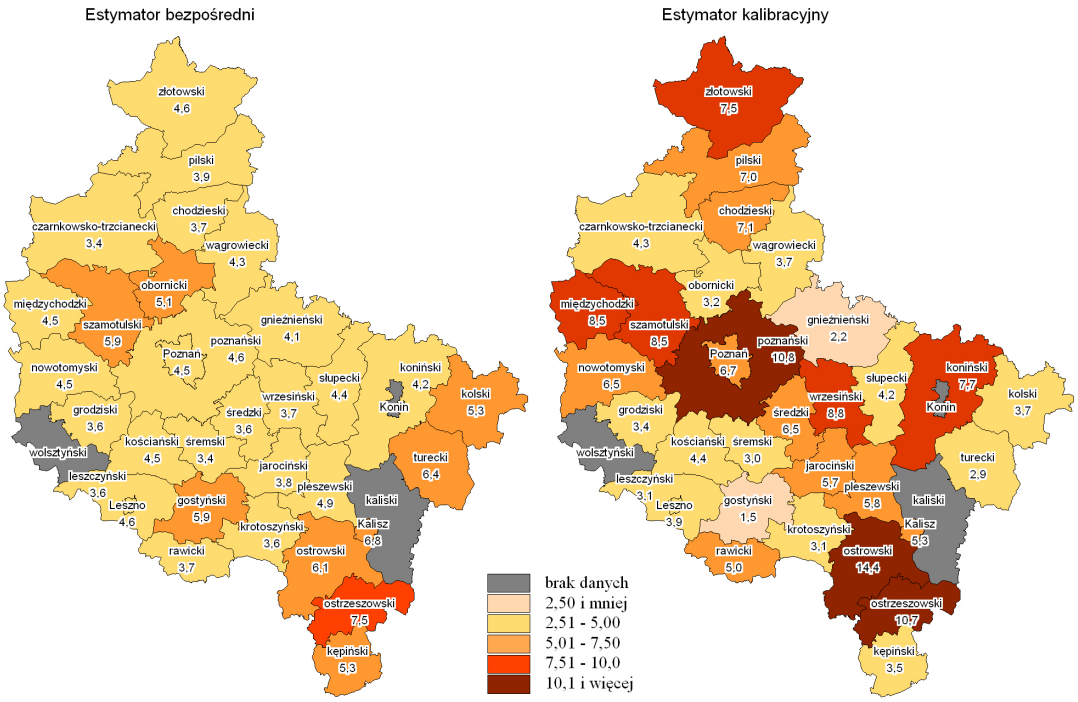


Rysunek. 5.11. Frakcja braków odpowiedzi wydatków na makaron w przekroju powiatów województwa wielkopolskiego w 2002r.

Źródło: Opracowanie własne

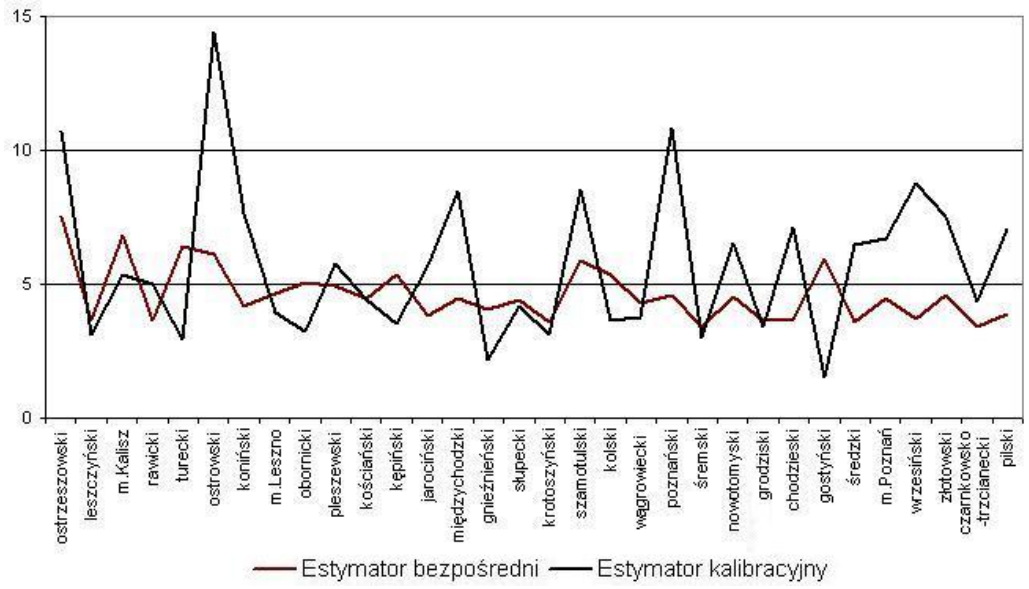
skiego, ostrowskiego, ostrzeszowskiego, pilskiego, poznańskiego, szamotulskiego, średzkiego, wrzesińskiego i złotowskiego oceny estymatora kalibracyjnego mediany wydatków na makaron znacznie przewyższają oszacowania, jakie daje estymator bezpośredni. Jest to zauważalne przede wszystkim dla powiatu ostrowskiego, gdzie różnica ocen obydwu estymatorów mediany wynosi ponad 8 zł. W niektórych powiatach można zaobserwować odwrotną sytuację. Odnosi się to do powiatów: gnieźnieńskiego, gostyńskiego, kępińskiego, kolskiego, obornickiego oraz tureckiego, dla których ocena estymatora bezpośredniego mediany wydatków na makaron jest wyższa od oceny estymatora kalibracyjnego. W odniesieniu do wydatków na makaron znaczne różnice można zaobserwować zarówno w sytuacji, gdy frakcja braków odpowiedzi była na niskim poziomie (powiat ostrowski – 11%) jak i wtedy, gdy była ona wysoka (powiat wrzesiński – 33%).

Opierając się na wynikach badań symulacyjnych rozdziału czwartego, większym zaufaniem bylibyśmy skłonni obdarzyć oceny estymatora kalibracyjnego – zarówno w przypadku wydatków na energię elektryczną jak i na makaron. Estymator kalibra-



Rysunek. 5.12. Mediana wydatków na makaron w przekroju powiatów województwa wielkopolskiego w 2002r.

Źródło: Opracowanie własne



Rysunek. 5.13. Porównanie ocen estymatora bezpośredniego i kalibracyjnego mediany wydatków na makaron w przekroju powiatów województwa wielkopolskiego w 2002r.

Źródło: Opracowanie własne

cyjny mediany charakteryzował się bowiem mniejszym obciążeniem od estymatora bezpośredniego we wszystkich rozważanych sytuacjach, tj. gdy frakcja braków odpowiedzi była mała jak i duża.

5.4. Wnioski

Wykonana empiryczna weryfikacja przydatności estymatorów kalibracyjnych w tym rozdziale, jest pierwszym ich zastosowaniem w badaniu budżetów gospodarstw domowych w przekroju powiatowym. Estymator kalibracyjny wydaje się przedstawiać bardziej przekonywująco różnice między powiatami województwa wielkopolskiego pod względem badanych cech. Tradycyjne estymatory klasy SMO zazwyczaj niwelują różnice między małymi obszarami. Estymator kalibracyjny te różnice uwypukla. Bardziej pogłębiona analiza poszczególnych powiatów, wydaje się wskazywać, że estymatory kalibracyjne lepiej opisują rzeczywistość. Przy ostatecznej ocenie naszych wyników należy mieć jednak na uwadze, że wyniki estymacji mogą ulec zmianie. Pogłębiając problem oceny jakości estymatorów kalibracyjnych, analiza taksonometryczna mogłaby z powodzeniem spełniać rolę kontrolną w stosunku do wyników estymacji pośredniej, a w szczególności z uwzględnieniem braków odpowiedzi.

Zakończenie

Praca nawiązuje do dynamicznie rozwijającej się na świecie metody kalibracji, poświęconej estymacji w badaniach statystycznych z brakami odpowiedzi i innymi błędami nielosowymi. Prezentowane podejście jest pierwszą w Polsce próbą kompleksowego przedstawienia estymatorów kalibracyjnych, które nie są u nas powszechnie znane i wykorzystywane.

Kalibracja wydaje się zadowalającym rozwiązaniem na braki odpowiedzi w badaniach społeczno-ekonomicznych. Szczególnie obiecująco wyglądają szacunki wydatków w budżetach gospodarstw domowych, obarczonych znacznymi błędami systematycznymi. To, co pokazują estymatory kalibracyjne, jest bardziej logiczne i lepiej odpowiada naszemu stanowi wiedzy niż obraz po estymacji bezpośredniej, por. na przykład rysunek 5.4. Merytoryczna ocena uzyskanych wyników wybranych kategorii wydatków z badania budżetów gospodarstw domowych, pozwoliła sformułować wniosek, że estymatory kalibracyjne dają bardziej przekonywujące i zróżnicowane oszacowania parametrów, w przeciwieństwie do estymatorów bezpośrednich, które te różnice niwelowały.

Dodatkowym argumentem przemawiającym na korzyść podejścia kalibracyjnego są wyniki badań symulacyjnych w sytuacjach, gdy braki odpowiedzi nie mają charakteru losowego. Jest to również zauważalne w badaniu budżetów gospodarstw domowych, w którym zazwyczaj uchylają się od udzielania odpowiedzi gospodarstwa o najniższych bądź najwyższych wartościach analizowanych cech (na przykład wydatki na alkohol i inne używki). Mechanizm powstawania braków odpowiedzi nie ma zatem charakteru losowego. W świetle uzyskanych wyników badań symulacyjnych, estymatory kalibracyjne są lepsze od estymatorów tradycyjnych, gdyż redukują obciążenie i poprawiają precyzję szacunków. Nawet w przypadku niewielkich braków odpowiedzi, obciążenie estymatorów kalibracyjnych jest od dwóch do czterech razy mniejsze niż estymatora bezpośredniego, por. tabela 4.2. Zawarte wyniki badań wskazują ponadto, że procedura estymacji ma wpływ na szacunek danego parametru. Symulacje, oparte na dwóch zbiorach danych: rzeczywistych i wygenerowanych z rozkładu logarytmiczno-normalnego, pozwoliły na sformułowanie wniosku, że na poprawę precyzji szacunku wpływa uwzględnienie wartości cech dodatkowych silnie skorelowanych z cechą badaną oraz zastosowanie technik bardziej złożonych, takich jak estymatory kalibracyjne regresyjne czy uogólnione estymatory kalibracyjne kwantyla rzędu α.

Wyniki przeprowadzonych badań i symulacji umożliwiły sformułowanie wniosków ogólniejszej natury, a mianowicie:

- W przypadku istnienia braków odpowiedzi, wszystkie estymatory są obciążone — w znacznie mniejszym jednak stopniu odnosi się to do estymatorów kalibracyjnych niż bezpośrednich. Dotyczyło to również sytuacji, gdy należało zastąpić wartość globalną bądź odpowiedni kwantyl zmiennych pomocniczych na poziomie całej populacji, oceną odpowiedniego estymatora z uwzględnieniem danych z próby.
- Estymatory bezpośrednie charakteryzują się większą wariancją i względnym błędem szacunku od kalibracyjnych.

Należy zatem przypuszczać, że omówione w rozprawie podejście kalibracyjne znajdzie w warunkach polskich szersze zastosowanie. Będzie to szczególnie ważne w kontekście prac prowadzonych przez Główny Urząd Statystyczny, a także ze względu na rosnący popyt na wiarygodną informację na niskich poziomach agregacji przestrzennej.

Ważną część rozprawy stanowiły również teoretyczne rozważania nad podejściem kalibracyjnym w estymacji statystycznej. Jest to pierwsza tego typu praca w Polsce, w której w kompleksowy sposób udowodniliśmy twierdzenia o wagach oraz podstawowe własności estymatorów kalibracyjnych. Zaproponowaliśmy własne estymatory kalibracyjne kwantyli dokonując uogólnienia na przypadek, gdy dla każdej zmiennej pomocniczej dysponujemy więcej aniżeli jednym kwantylem. Uwzględniliśmy przy tym, że w badaniu występują braki odpowiedzi. Jak pokazały wyniki symulacji, włączenie dodatkowych informacji może w znaczący sposób ograniczyć obciążenie i wariancję estymatorów. Zważywszy na fakt, że w niektórych badaniach szacuje się kwantyle jako podstawowe miary zróżnicowania, zaprezentowana autorska metodologia wyznaczania kwantyli w badaniach z brakami odpowiedzi, może stanowić interesującą propozycję ich późniejszego stosowania w praktyce.

Wśród ważnych zagadnień, wymagających uwagi w najbliższym czasie w kontekście rozważanego w pracy podejścia kalibracyjnego, wskazać można kilka istotnych problemów:

1. Konieczność włączenia w badaniach z brakami odpowiedzi podejścia modelowego. Rozważane w pracy estymatory kalibracyjne abstrahują od modelu opisującego zależność pomiędzy zmienną y, a zmiennymi pomocniczymi $x_1, \ldots, x_k$. Wydaje się, że uwzględnienie podejścia wspomaganego modelem, z jednoczesnym spełnieniem odpowiednich równań kalibracyjnych, może poprawić proces estymacji i prowadzić do uzyskania bardziej przekonywujących wyników, por. C. Cassel, P. Lundquist, J. Selén (2002).
2. Pogłębienie badań nad estymatorami kalibracyjnymi, w których konstrukcji wykorzystać można inne, aniżeli rozważane w pracy, funkcje odległości oraz próby porównania ich własności, zarówno od strony teoretycznej jak i empirycznej. Dotyczyć to może w szczególności estymatorów, dla których wagi kalibracyjne można wyznaczać w oparciu o funkcje G, których przegląd zamieszczono w rozdziale pierwszym rozprawy.
3. Pogłębienie badań nad odpornymi estymatorami kalibracyjnymi, które uwzględnieniać będą nie tylko fakt występowania w próbie braków odpowiedzi, ale również niwelować będą ujemny wpływ wartości odstających na jakość wyników. Do tej pory, w ramach podejścia kalibracyjnego, nie stosowano na szerszą skalę metod odpornych, które różnicowałyby dodatkowo wagi poprzez ich odpowiednie wykalibrowanie w odniesieniu do jednostek, dla których zaobserowowano wartości eks-

tremalne, por. P. Duchesne (1999). Odporne estymatory kalibracyjne stanowiłyby zatem swego rodzaju remedium na współwystępujący w wielu badaniach statystycznych problem braków odpowiedzi i wartości odstających.

4. Konieczność integracji różnych zbiorów danych, w kontekście wykorzystania zmiennych pomocniczych z alternatywnych względem badania budżetów gospodarstw domowych, źródeł informacji. W przypadku szacowania parametrów opisujących wydatki gospodarstw domowych na różne artykuły i dobra konsumpcyjne oprócz informacji zawartych w samym badaniu budżetów gospodarstw domowych, szczególnie ważnym źródłem zmiennych pomocniczych mógłby być, na przykład rejestr podatników, płatników i innych podmiotów oraz dokumentów „POLTAX", a także baza danych o podatnikach podatku dochodowego od osób fizycznych (PIT). Pełne wykorzystanie informacji pomocniczych, zawartych w tych źródłach, mogłoby dodatkowo poprawić proces estymacji oraz uwiarygodnić wyniki badań z budżetów gospodarstw domowych.
5. Zastosowanie estymatorów kalibracyjnych w innych, aniżeli badanie budżetów gospodarstw domowych, badaniach prowadzonych przez Główny Urząd Statystyczny. W zasadzie może to dotyczyć wszystkich badań prowadzonych metodą reprezentacyjną, gdyż braki odpowiedzi są permanentną częścią każdego z nich. Wśród badań, w odniesieniu do których w pierwszej kolejności warto byłoby zastosować podejście kalibracyjne, należy zaliczyć: badanie aktywności ekonomicznej ludności oraz badania towarzyszące spisom powszechnym, na przykład badanie dzietności.

Na koniec należy podkreślić, że praca ma charakter metodologiczno–poznawczy, przy czym większy akcent położono na metody konstrukcji estymatorów kalibracyjnych. Próba zastosowania kalibracji była egzemplifikacją omawianych metod niż zasadniczą kompleksową oceną wydatków na określone kategorie dóbr i usług w przekroju powiatów województwa wielkopolskiego. Wątek poznawczy został świadomie zredukowany do empirycznej weryfikacji wybranych metod z zakresu podejścia kalibracyjnego w badaniach statystycznych z brakami odpowiedzi.

Dokonując podsumowania prowadzonych w pracy rozważań można stwierdzić, że kalibracja jako metoda estymacji będzie odgrywać w polskiej praktyce badań statystycznych coraz większą rolę. Wynika to z dwóch zasadniczych powodów. Po pierwsze, wiele badań prowadzonych przez Główny Urząd Statystyczny jest obarczonych błędami nielosowymi w postaci braków odpowiedzi. Po drugie, rosnący popyt na informację na niskich poziomach agregacji przestrzennej, takich jak powiaty, gminy, a nawet ich części wymuszać będzie od instytucji przeprowadzających badanie stosowanie bardziej wyrafinowanych metod. Kalibracja jako metodologia estymacji w ujęciu proponowanym w pracy, ma szansę sprostać temu zapotrzebowaniu.

Literatura

[1] Alkaya A., Esin A. (2005), „*Calibration Estimator*", G. U. Journal of Science, 18(4), 591–601.

[2] Andersson C., Nordberg L. (1998), „*A User's Guide to CLAN97*", Statistics Sweden.

[3] Andersson G., Thorburn D. (2005), „*An Optimal Calibration Distance Leading to the Optimal Regression Estimator*", Survey Methodology, Vol. 31, No. 1, 95–99.

[4] Ash S. (2003), „*Simultaneous Calibration Estimators for Two-Phase Samples*", 2003 Join Statistical Meetings - Section on Survey Research Methods, 395–400.

[5] Bankier M.D., Houle A.M., Luc M. (1997), „*Calibration Estimation in the 1991 and 1996 Canadian Censuses*", Proceedings, Section on Survey Research Methods, American Statistical Association, 66–75.

[6] Beaumont J-F. (2005), „*Calibrated Imputation in Surveys under a Quasi Model Assisted Approach*", Journal of the Royal Statistical Society B, 67, 445–458.

[7] Bethlehem J. (1997), „*Bascula: Current Status and Future Developments*", Statistics Netherlands, 1–48.

[8] Bethlehem J. (2002), „*Weighting Nonresponse Adjustments Based on Auxiliary Information*", In: Groves R., Dillman D., Eltinge J., Little RJA. eds. Survey Nonresponse. New York: John Wiley & Sons, Inc.

[9] Brewer K.R.W. (1999), „*Cosmetic Calibration with Unequal Probability Sampling*", Survey Methodology, Vol. 25, No. 2, December 1999, 193–203.

[10] Campanelli P. (1997), „*Testing Survey Questions: New Directions in Cognitive Interviewing*", Bulletin de Methodologie Sociologique, 55, 5–17.

[11] Cannell C. F., Miller P. V., Oksenberg L. (1981), „*Research on Interviewing Techniques*", Sociological Methodology, San Francisco: Jossey-Bass.

[12] Cassel C., Lundquist P., Selén J. (2002), „*Model Based Calibration for Survey Estimation with an Example from Expenditure Analysis*", Research Methods Development, Statistics Sweden, Örebro, 1–22.

[13] Çelikbiçak M. B., Oruç Ö. E. (2006), „*Methods Used in Reduction of Bias Arising from Nonresponse*", Journal of Arts and Sciences, No. 6, 33–51.

[14] Chambers R. L., Skinner C. J., Wang S. (1999), „*Intelligent Calibration*", Bulletin of the International Statistical Institute: 52nd Session Proceedings. Finland, International Statistical Institute, 321–324.

[15] Chambers R. L., Dunstan R. (1986), „*Estimating Distribution Functions from Survey Data*",Biometrika, 73, 597–604.

[16] Chambers R. L., Dorfman A.H., Hall P. (1992), „*Properties of Estimators of Finite Population Distribution Functions*",Biometrika, 79, 577–582.

[17] Chang T., Kott P. S. (2005), „*Using Calibration Weighting to Adjust for Nonresponse*", National Agricultural Statistical Service, 1–27.

[18] Changbao Wu., Luan Y. (2003), „*Optimal Calibration Estimators Under Two-Phase Sampling*", Journal of Official Statistics, Vol. 19, No. 2, 119–131.

[19] Chen G., Chen J. (1996), „*A Transformation Method for Finite Population Sampling Calibrated with Empirical Likelihood*", Survey Methodology, Vol. 22, No. 2, December 1996, 139–146.

[20] Chen J., Wu C. (2002), „*Estimation of Distribution Function and Quantiles Using the Model-Calibrated Pseudo Empirical Likelihood Method*", Statistica Sinica, 12, 1223–1239.

[21] Chen P., Penne M.A., Singh A.C. (2001), „*Experience with the Generalized Exponential Model for Weight Calibration for the National Household Survey on Drug Abuse*", Research Triangle Institute, 604–609.

[22] Crouse C., Kott P.S. (2004), „*Evaluating Alternative Calibration Schemes for an Economic Survey with Large Nonresponse*", ASA Proceedings of the Survey Research Methods Section, 1509–1515.

[23] Dehnel G. (2003), „*Statystyka małych obszarów jako narzędzie oceny rozwoju ekonomicznego regionów*", Wydawnictwo Akademii Ekonomicznej w Poznaniu, Poznań.

[24] Dehnel G. (2009), „*Testowanie estymacji odpornej dla małych obszarów w oparciu o dane z NSP'2002*", Prace przygotowawcze do spisów PSR 2010 i NSP 2011, Prace Podgrupy ds. metod statystyczno-matematycznych na rzecz spisów zrealizowane w 2008r., Część I, Główny Urząd Statystyczny, Warszawa.

[25] Deville J-C., Särndal C-E. (1992), „*Calibration Estimators in Survey Sampling*", Journal of the American Statistical Association, Vol. 87, 376–382.

[26] Deville J-C., Särndal C-E, Sautory O. (1993), „*Generalized Raking Procedures in Survey Sampling*", Journal of the American Statistical Association, Vol. 88, 1013–1020.

[27] Deville J-C. (2000), „*Generalized Calibration and Application to Weighting for Nonresponse*", COMPSTAT – Proceedings in Computational Statistics, 65–76.

[28] Dillman D.A., Eltinge J.J., Groves R.M., Little R.J.A. (2002), „*Survey Nonresponse in Design, Data Collection and Analysis*", In Survey Nonresponse, Groves R.M., Dillman D.A., Eltinge J.L., Little R.J.A (eds)., New York: Wiley, 3–26.

[29] Domański Cz., Pruska K. (2001), „*Metody statystyki małych obszarów*", Wydawnictwo Uniwersytetu Łódzkiego, Łódź.

[30] Dorfman A.H. (1993), „*A Comparision of Design-based and Model-based Estimators of the Finite Population Distribution Function*", Australian Journal of Statistics, 35, 29–41.

[31] Duchesne P. (1999), „*Robust Calibration Estimators*", Survey Methodology, Vol. 25, Number 1, June 1999, 43–56.

[32] Dupont F. (1994), „*Calibration Used as a Nonresponse Adjustment*", New Approaches in Classification and Data Analysis, Springer Verlag, 539–548.

[33] Dupont F. (1995), „*Alternative Adjustments when There Are Several Levels of Auxiliary Information*", Survey Methodology, 21, 125–136.

[34] Dygaszewicz J. (2007), „*Narodowy Spis Powszechny Ludności i Mieszkań 2011. Założenia metodyczne*", Materiał na posiedzenie Rady Programowej narodowego spisu powszechnego ludności i mieszkań 2011 r., Warszawa 2007, 1–61.

[35] Estevao V. M. (1994), „*Calculation of G-Weights under Calibration and Bound Constraints*", Report, Statistics Canada.

[36] Estevao V. M., Särndal C-E. (1999), „*The Use of Auxiliary Information in De-*

sign-Based Estimation for Domains", Survey Methodology, Vol.25, No. 2, December 1999, 213–221.

[37] Estevao V., Särndal C-E. (2000), „*A Functional Form Approach to Calibration*", Journal of Official Statistics, Vol. 16, No. 4, 379–399.

[38] Estevao V. M., Särndal C-E. (2002), „*The Ten Cases of Auxiliary Information for Calibration in Two-Phase Sampling*", Journal of Official Statistics, Vol. 18, 233–255.

[39] Estevao V. M., Särndal C-E. (2003), „*A New Perspective on Calibration Estimators*", 2003 Join Statistical Meetings - Section on Survey Research Methods, 1346–1356.

[40] Estevao V. M., Särndal C-E. (2006), „*Survey Estimates by Calibration on Complex Auxiliary Information*", International Statistical Review, Vol. 74, Number 2 (2006), 127–147.

[41] Éltető Ö., László M. (2002), „*Household Surveys in Hungary*", Statistics in Transition, Vol. 5, No. 4, 521–540.

[42] Folsom R.E., Singh A.C. (2000), „*The Generalized Exponential Model for Sampling Weight Calibration for Extreme Values, Nonresponse and Poststratification*", Proceedings of the Section on Survey Research Mthods, American Statistical Association, Washington DC, 598–603.

[43] Gambino J. (1999), „*Discussion of «Issues in Weighting Household and Business Surveys»*", Statistics Canada, Household Survey Methods Division, 1–2.

[44] Gołata E. (2004), „*Estymacja pośrednia bezrobocia na lokalnym rynku pracy*", Prace Habilitacyjne 11, Wydawnictwo Akademii Ekonomicznej w Poznaniu.

[45] Gołata E. (2009), „*Testowanie estymatorów klasy SMO (statystyki małych obszarów) w oparciu o dane z NSP'2002 oraz z bieżących badań reprezentacyjnych BAEL'2002*", Prace przygotowawcze do spisów PSR 2010 i NSP 2011, Prace Podgrupy ds. metod statystyczno-matematycznych na rzecz spisów zrealizowane w 2008r., Część I, Główny Urząd Statystyczny, Warszawa.

[46] Groves R. M. (2006), „*Nonresponse Rates and Nonresponse Bias in Household Surveys*", Public Opinion Quarterly, Vol. 70, No. 5, Special Issue 2006, 646-675.

[47] GUS (2003a), „*Użytkowanie gruntów i ich jakość 2002*", Główny Urząd Statystyczny, Departament Statystyki Rolnictwa i Środowiska, Warszawa 2003.

[48] GUS (2003b), „*Budżety gospodarstw domowych w 2002 r.*", Główny Urząd Statystyczny, Informacje i opracowania statystyczne, Departament Statystyki Społecznej, Warszawa 2003.

[49] GUS (2007), „*Warunki życia ludności Polski w latach 2004–2005*", Główny Urząd Statystyczny, Departament Statystyki Społecznej, Warszawa 2007.

[50] GUS (2008), „*Dochody i warunki życia ludności (raport z badania EU-SILC 2006r.)*", Główny Urząd Statystyczny, Informacje i opracowania statystyczne, Departament Pracy i Warunków Życia, Warszawa 2008.

[51] Hansen M.H., Hurwitz W.N. (1943), „*On the Theory of Sampling from Finite Populations*", Annals of Mathematical Statistics,14, 333–362.

[52] Harms T. (2003), „*Extensions of the Calibration Approach: Calibration of Distribution Functions and its Link to Small Area Estimators*", Chintex working paper#13, Federal Statistical Office, Germany.

[53] Harms T., Duchesne P. (2006), „*On Calibration Estimation for Quantiles*", Survey Methodology, Vol. 32, June 2006, 37–52.

[54] Hidiroglou M.A., Särndal C-E. (1998), „*Use of Auxiliary Information for Two-phase Sampling*", Survey Methodology, Vol. 24, No. 1, June 1998, 11–20.

[55] Jayasuriya B., Valliant R. (1995), „*An Application of Regression and Calibration Es-*

timation to Post-Stratification in a Household Survey", U.S. Department of Labor, Bureau of Labor Statistics, Office of Survey Methods Research, 902–907.

[56] Kalton G., Kasprzyk D. (1986), „*The Treatment of Missing Data*", Survey Methodology, 12, 1–16.

[57] Kalton G., Maligalig D.S. (1991), „*A Comparision of Weighting Adjustment for Nonresponse*", Proceedings of the Bureau of the Census Annual Research Conference, 409–428.

[58] Kalton G., Flores-Cervantes I. (2003), „*Weighting Methods*", Journal of Official Statistics, Vol. 19, No. 2, Statistics Sweden, 81–97.

[59] Kim J-M., Sungur E. A., Heo T-Y. (2007), „*Calibration approach estimators in stratified sampling*", Statistics & Probability Letters 77(2007), 99–103.

[60] Klimanek T. (2009), „*Testowanie estymatorów klasy SMO (statystyki małych obszarów) w powszechnym spisie rolnym (PSR) na podstawie PSR'2002 oraz badań struktury rolnej 2005 i 2007*", Prace przygotowawcze do spisów PSR 2010 i NSP 2011, Prace Podgrupy ds. metod statystyczno-matematycznych na rzecz spisów zrealizowane w 2008r., Część I, Główny Urząd Statystyczny, Warszawa.

[61] Knäuper B., Belli R.F., Hill D.H., Herzog A.R. (1997), „*Question Difficulty and Respondents' Cognitive Ability: The Effect on Data Quality*", Journal of Official Statistics, 13, 181–199.

[62] Kordos J. (1988), „*Jakość danych statystycznych*", Państwowe Wydawnictwo Ekonomiczne, Warszawa 1988.

[63] Kordos J. (1997), „*40 lat badań budżetów gospodarstw domowych w Polsce*", „Wiadomości Statystyczne", nr 3, 27–42.

[64] Kordos J. (2004), „*Podstawowe badania społeczne statystyki publicznej w Polsce*", [w:] Statystyka społeczna – wybrane zagadnienia, Szkoła Główna Handlowa w Warszawie, 29–57.

[65] Kott P.S. (1996), „*Calibration Estimators Based on Several Separate Stratifications*", Proceedings of the Survey Research Methods Section, American Statistical Association, 819–823.

[66] Kott P.S. (2003a), „*An Overview of Calibration Weighting*", 2003 Joint Statistical Meetings - Section on Survey Research Methods, 2241–2252.

[67] Kott P.S. (2003b), „*A Practical Use for Instrumantal Variable Calibration*", Journal of Official Statistics, Vol. 19, No. 3, 265–272.

[68] Kott P.S. (2006), „*Using Calibration Weighting to Adjust for Nonresponse and Coverage Errors*", Survey Methodology, Vol. 32, No. 2, 133–142.

[69] Kovačević M.S. (1997), „*Calibration Estimation of Cumulative Distribution and Quantile Functions from Survey Data*", Proceedings of the Survey Methods Section, Statistical Society of Canada, 139–144.

[70] Kroese B., Renssen R.H., Trijssenaar M. (2000), „*Weighting or Imputation: Constructing a Consistent Set of Estimates Based on Data from Different Sources*", [in:] Integrating Statistics Administrative Registers and Household Surveys, Vol. 15, Netherlands Official Statistics, 23–31.

[71] Kuk A.Y.C. (1988), „*Estimation of Distribution Functions and Medians under Sampling with Unequal Probabilities*", Biometrika, 75, 97–103.

[72] Kuk A.Y.C., Mak T.K. (1989), „*Median Estimation in the Presence of Auxiliary Information*", Journal of the Royal Statistical Society, Series B (Methodological), 51, 261–269.

[73] Leeuw E.D., Hox J., Huisman M. (2003), „*Prevention and Treatment of Item Nonresponse*", Journal of Official Statistics, Vol. 19, No. 2, 153–176.
[74] Little R., Rubin D. (1987), „*Statistical Analysis with Missing Data*", New York: Wiley.
[75] Longford N.T. (2005), „*Missing Data and Small-Area Estimation*", Springer.
[76] Luan Y. (2001), „*Model-Calibration Approach under Two-Phase Sampling*", Unpublished master's essay, Department of Statistics and Actuarial Science, University of Waterloo, Canada.
[77] Lundström S. (2005), „*Estimation in the Presence of Nonresponse and Other Survey Imperfections - A Handbook at Statistics Sweden*", Klostergatan 23, Örebro, Sweden, 1–2.
[78] Meeden G. (1995), „*Median Estimation Using Auxiliary Information*", Survey Methodology, 21, 71–77.
[79] Montanari G. E., Ranalli M. G. (2003), ,*Neural Networks for Calibration Estimation of Finite Population Parameters*",Proceedings of the Conference of the Italian Statistical Society SIS 2003, Napoli, Italy, 1–4.
[80] Mukhopadhyay P. (2005), „*Cosmetic and Calibration Estimators*", Pakistan Journal of Statistics, Vol. 21(1), 15–26.
[81] Nieuwenbrook N.J., Boonstra H.J. (2002), „*Bascula 4.0 for Weighting Sample Survey Data with Estimation of Variances*", Survey Statistician, Software Reviews.
[82] Nordberg L. (2000), „*CLAN - Statistics Sweden's Software for Computation of Point and Standard Error Estimates in Sample Surveys*", Statistics Sweden, Örebro, 187–193.
[83] Paradysz J., Szymkowiak M. (2007), „*Imputacja i kalibracja jako remedium na braki odpowiedzi w badaniu budżetów gospodarstw domowych* ", [w:] Taksonomia 14. Klasyfikacja i analiza danych - teoria i zastosowania. Wydawnictwo Akademii Ekonomicznej we Wrocławiu, s. 74-80.
[84] Paradysz J., Szymkowiak M. (2007), „*Źródła danych ludnościowych*", [w:] Metodologia Badań Demograficznych, Komitet Nauk Demograficznych PAN, Warszawa, 7–26.
[85] Paradysz J., Szymkowiak M. (2008), „*Taksonometryczne podstawy kalibracji w statystyce małych obszarów* ", [w:] Taksonomia 15. Klasyfikacja i analiza danych - teoria i zastosowania. Wydawnictwo Akademii Ekonomicznej we Wrocławiu, Nr 7(1207), s. 215–222.
[86] Paveone A. (2006), „*Nonresponse Bias Reduction Integrating Calibration Constraints with Variance and Covariance Structure of Auxiliary Variables*", ISTAT, 407–410.
[87] Plikuskas A. (2006), „*Nonlinear Calibration*", Statistics Lithuania, Institute of Mathematics and Informatics, 1–6.
[88] Plikuskas A. (2003), „*Calibrated Weights for the Estimators of the Ratio*", Lithuanian Mathematical Journal, 43, 543–547.
[89] Plikuskas A. (2007), „*Calibrated Estimators of the Population Covariance* ", Acta Applicandae Mathematicae: An International Survey Journal on Applying Mathematics and Mathematical Applications, Springer Netherlands, Volume 97, Numbers 1–3, July, 177–187.
[90] Pratesi M., Rocco E. (2006), „*Two Steps Calibration in the Presence of Nonresponse*", European Conference on Quality in Survey Statistics, 1–8.
[91] Ranalli Giovanna M. (2008), „*Recent Developments in Calibration Estimation*", Proccedings of the XLIV Meeting of the Italian Statistical Society, June 25–27, 2008, Invited papers CLEUP, 355–362.

[92] Rancourt E. (2000), „*Edit and Imputation: From Suspicious to Scientific Techniques*", Statistics Canada, Household Survey Methods Division Tunneys Pasture, Ottawa, 1–4.
[93] Rao J.N.K., Kovar J.G., Mantel H.J. (1990), „*On Estimating Distribution Functions and Quantiles from Survey Data Using Auxiliary Information*", Biometrika, 77, 365–375.
[94] Rao J.N.K., Singh A.C. (1997), „*A Ridge-shrinkage Method for Range-Restricted Weight Calibration in Survey Sampling*", Proceedings of the Section on Survey Research Methods, American Statistical Association.
[95] Rao J.N.K. (2003), „*Small Area Estimation*", John Wiley & Sons, Inc.
[96] Ren R. (2002), „*Estimation de la fonction de répartition et des fractiles d'une population finie*", Actes des journées de méthodologie statistique, INSEE Méthodes, Tome 1, 100, 263–289.
[97] Rendtel U., Harms T. (2006), „*Weighting and Calibration for Household Panels*", Methodology of Longitudinal Surveys, (ed. P. Lynn), John Wiley & Sons, Ltd., 1–24.
[98] Renssen R.H. (1998), „*Use of Statistical Matching Techniques in Calibration Estimation*", Survey Methodology, Vol. 24, No. 2, December 1998, 171–183.
[99] Rubin D.B. (1976), „*Inference and Missing Data*", Biometrika, 63, 581–590.
[100] Rueda M.M., Arcos A., Martnez M.D. (2003), „*Difference Estimators of Quantiles in Finite Populations*",Test, 12, 481–496.
[101] Rueda M., Martnez S., Martnez H., Arcos A. (2006), „*Mean Estimation with Calibration Techniques in Presence of Missing Data*", Computational Statistics & Data Analysis 50 (2006), 3263–3277.
[102] Rueda M., Martnez S., Martnez H., Arcos A. (2007), „*Estimation of the Distribution Function with Calibration Methods*", Journal of Statistical Planning and Inference 137(2007), 435–448.
[103] Sautory O. (2003), „*Calmar 2: A New Version of the Calmar Calibration Adjustment Program*", Proceedings of Statistics Canada's Symposium 2003, Challenges in Survey Taking for the Next Decade, Statistics Canada - Catalogue no. 11-522-XIE.
[104] Särndal C-E., Swensson B., Wretman J.H. (1992), „*Model Assisted Survey Sampling*, New York, Springer-Verlag.
[105] Särndal C-E., Lundström S. (1999), „*Calibration as a Standard Method for Treatment of Nonresponse*", Journal of Official Statistics, Vol. 15, No. 2, 305–327.
[106] Särndal C-E., Lundström S. (2001), „*Estimation in the Presence of Nonresponse and Frame Imperfections*", Statistics Sweden, Örebro.
[107] Särndal C-E., Lundström S. (2002), „*A Handbook on Estimation in the Presence of Nonresponse and Frame Imperfections*", The International Conference on Improving Surveys, ICIS 2002, 1–8.
[108] Särndal C-E., Lundström S. (2005), „*Estimation in Surveys with Nonresponse*", John Wiley & Sons, Ltd.
[109] Särndal C-E. (2007), „*The Calibration Approach in Survey Theory and Practice*", Survey Methodology, Vol. 33, No. 2, 99–119.
[110] Särndal C-E., Lundström S. (2007), „*Assessing Auxiliary Vectors for Control of Nonresponse Bias in the Calibration Estimator*", Methodology Reports from Statistics Sweden, Research and Development Department, 1–59.
[111] Särndal C-E., Lundström S. (2008), „*Assessing Auxiliary Vectors for Control of Nonresponse Bias in the Calibration Estimator*", Journal of Official Statistics, Vol. 24, No. 2, 167–191.

[112] Sikkel D., Hox J., Leeuw E. (2004), „*Using Auxiliary Data for Adjustment in Longitudinal Research*", 1–11.
[113] Singh A.C., Wu S., Boyer R. (1995), „*Longitudinal Survey Nonresponse Adjustment by Weight Calibration for Estimation of Gross Flows*", Proceedings of the Survey Research Methods Section, American Statistical Association, 396–401.
[114] Singh A.C., Mohl C.A. (1996), „*Understanding Calibration in Survey Sampling*", Survey Methodology, Vol. 22, No. 2, December 1996, 107–115.
[115] Singh A.C., Mohl C.A. (1997), „*Calibration Estimators with application to FAMEX Survey and Computer Program Documentation*", Methodology Branch Working Paper, Statistics Canada.
[116] Singh S., Horn S., Yu F. (1998), „*Estimation of Variance of General Regression Estimator: Higher Level Calibration Approach*", Survey Methodology, Vol. 24, No. 1, June 1998, 99–119.
[117] Singh S., Horn S., Chowdhury S., Yu F. (1999), „*Calibration of the Estimators of Variance*", Australian&New Zealand Journal of Statistics, Volume 41, Number 2, June, 199–212.
[118] Singh S., Horn S., Tracy D.S. (2001), „*Hybrid of Calibration and Imputation: Estimation of Mean in Survey Sampling*", Statistica 61 (1), 27–41.
[119] Skinner Ch. (1998), „*Calibration Weighting and Non-Sampling Errors*", Proceedings of the Seminar on New Techniques and Technologies for Statistics, Sorrento, 33–43.
[120] Son Ch. K., Jung Y. M. (2004), „*The Calibration of Horvitz–Thompson Variance Estimator under the Unit Nonresponse*", Proceedings of the Spring Conference, Korean Statistical Society, 45–50.
[121] Stukel D.M., Hidiroglou M.A., Särndal C-E. (1996), „*Variance Estimation for Calibration Estimators: A Comparision of Jackknifing versus Taylor Linearization*", Survey Methodology, Vol. 22, No. 2, December 1996, 117–125.
[122] Stukel D.M., Boyer R. (1992), „*Calibration Estimation: An Application to the Canadian Labour Force Survey*", Methodology Branch Working Paper, SSMD 92-009E, Statistics Canada.
[123] Sugden R.A., Smith T.M. Fred. (2007), „*Design-Based Properties of Linear Calibrated Estimators of a Finite Population Total*", International Statistical Review, 75, 218–223.
[124] Sverchkov M., Dorfman A. H., Ernst L. R., Moerhle T. G., Paben S. T., Ponikowski Ch. H. (2005)„„*On Non-Response Adjustment via Calibration*", Bureau of Labor Statistics, 1–6.
[125] Szymkowiak M. (2004), „*Wyznaczanie optymalnej alokacji próby pomiędzy małe obszary dla estymatora złożonego*", [w:] Taksonomia 11. Klasyfikacja i analiza danych - teoria i zastosowania, Wydawnictwo Akademii Ekonomicznej we Wrocławiu, s. 413–419.
[126] Szymkowiak M. (2007), „*Przyczynek do kalibracji w badaniach statystycznych z brakami odpowiedzi*", [w:] Kapitał Ludzki i Wiedza w Gospodarce. Wyzwania XXI Wieku, Zeszyty Naukowe Wydziału Informatyki i Gospodarki Elektronicznej AE w Poznaniu, 194–204.
[127] Szymkowiak M. (2007), „*Optymalna alokacja próby w statystyce małych obszarów*", [w:] Statystyka Regionalna w jednoczącej się Europie, Internetowa Oficyna Wydawnicza Centrum Statystyki Regionalnej, Poznań, s.184–189.
[128] Szymkowiak M. (2008a), „*Imputacja i kalibracja – nowe możliwości estymacji w ba-*

daniach statystycznych z brakami odpowiedzi", Zeszyty Naukowe AE w Poznaniu, w druku.

[129] Szymkowiak M. (2009), „*Testowanie estymatorów kalibracyjnych dla małych obszarów w oparciu o dane z NSP 2002 oraz z bieżących badań reprezentacyjnych budżetów gospodarstw domowych 2002*", Prace przygotowawcze do spisów PSR 2010 i NSP 2011, Prace Podgrupy ds. metod statystyczno-matematycznych na rzecz spisów zrealizowane w 2008r., Część I, Główny Urząd Statystyczny, Warszawa.

[130] Théberge A. (1999), „*Extensions of Calibration Estimators in Survey Sampling*", Journal of the American Statistical Association, 94, 635–644.

[131] Théberge A. (2000), „*Calibration and Restricted Weights*", Survey Methodology, Vol. 26, Number 1, June 2000, 99–107.

[132] Thibaudeau Y., Shao J., Mulrow J. (2006), „*A Study of Basic Calibration Estimators and their Variance Estimators in Presence of Nonresponse*", U.S. Census Bureau, 1–7.

[133] Tillé Y. (2002), „*Unbiased Estimation by Calibration on Distribution in Simple Sampling Designs Without Replacement*", Survey Methodology, Vol. 28, No. 1, June 2002, 77–85.

[134] Tíngdahl S. (2004), „*Nonresponse Bias for Some Common Estimators and its Change Over Time in the Data Collection Process*", Working Paper Series No. 13, Statistics Sweden, Örebro University, 1–35.

[135] Toompere K. (2006), „*Strength of Auxiliary Information for Compensating Nonresponse*", Workshop on Survey Sampling Theory and Methodology, August 24–28, 2006, Ventspils, 153–159.

[136] Tracy D. S., Singh S., Arnab R. (2003), „*Note on Calibration in Stratified and Double Sampling*", Survey Methodology, Vol. 29, No. 1, June 2003, 99–104.

[137] Welsh A.H., Ronchetti E. (1998), „*Bias-Calibrated Estimation From Sample Surveys Containing Outliers*", Journal of the Royal Statistical Society, Series B, 60, 413–428.

[138] Williams M. (2006), „*Handbook on Methodological Aspects Related to Sampling Designs and Weights Estimations*", Task Force on the Implementation of NACE Rev. 2, Version 1.0, ONS.

[139] Witkowska A., Witkowski M. (2006), „*Taksonometryczna analiza rynku pracy w województwie wielkopolskim w latach 2000–2003*", Wydawnictwo Akademii Ekonomicznej we Wrocławiu.

[140] Wu. C., Sitter R.R. (2001a), „*Variance Estimation for the Finite Population Distribution Function with Complete Auxiliary Information*", The Canadian Journal of Statistics, 29, 289–308.

[141] Wu. C., Sitter R.R. (2001b), „*A Model-Calibration Approach to Using Complete Auxiliary Information from Survey Data*", Journal of the American Statistical Association, 96, 185–193.

[142] Wu. C. (2003), „*Optimal Calibration Estimators in Survey Sampling*", Biometrika, 90(4), 937–951.

[143] Wu. C. (2007), „*Optimal Calibration and Empirical Likelihood Methods in Survey Sampling*", Proceedings of Statistics Canada's Symposium 2002, Modelling Survey Data for Social and Economic Research.

[144] Vaish A.K., Gordek H., Singh A.C. (2000), „*Variance Estimation Adjusted for Weight Calibration via the Generalized Exponential Model with Application to the National Household Survey on Drug Abuse*,Research Triangle Institute, 616–621.

[145] Vanderhoeft C., Waeytens E., Museux J.M. (2000), „*Generalised Calibration with*

SPSS 9.0 for Windows", Statistics Belgium, Department for Methodology and Coordination, 1–14.

[146] Vanderhoeft C. (2001), „*New Strategies in Calibration at Statistics Belgium*", Statistics Belgium, Department for Methodology and Coordination, 1–2.

[147] Vanderhoeft C. (2002), „*g-CALIB 1.0: SPSS Based Software for Generalised Calibration*", Statistics Belgium, Department for Methodology and Coordination, 1–7.

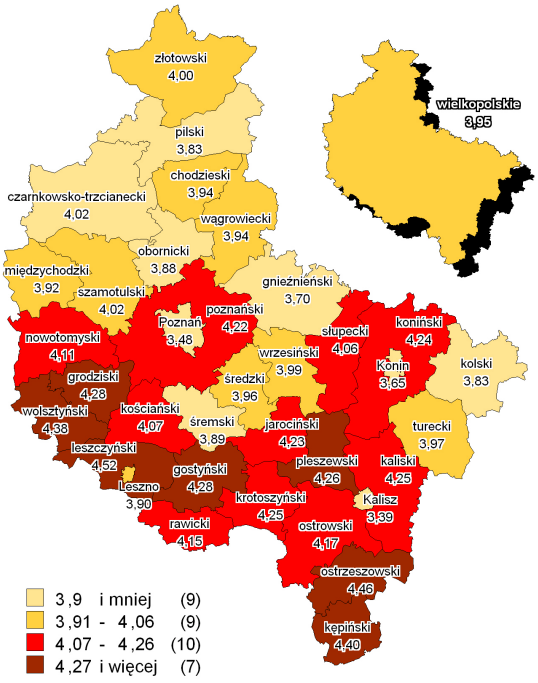
[148] Vaughn B. J. (2006), „*An Empirical Comparison of Weighting Class Adjustment and Propensity Modeling Adjustment for Nonresponse*", 1–40.

[149] Yucel R.M., Yulei H., Zaslavsky A.M. (2008), „*Using Calibration to Improve Rounding in Imputation*", The American Statistician, A Publication of the American Statistical Association, Vol. 62, No. 2, 125–129.

[150] Zanutto E. L., Zaslavsky A. M. (2006), „*A Model for Estimating and Imputing Nonrespondent Census Households under Sampling for Nonresponse Follow-up*", Survey Methodology, Vol. 32, No. 1, 65–76.

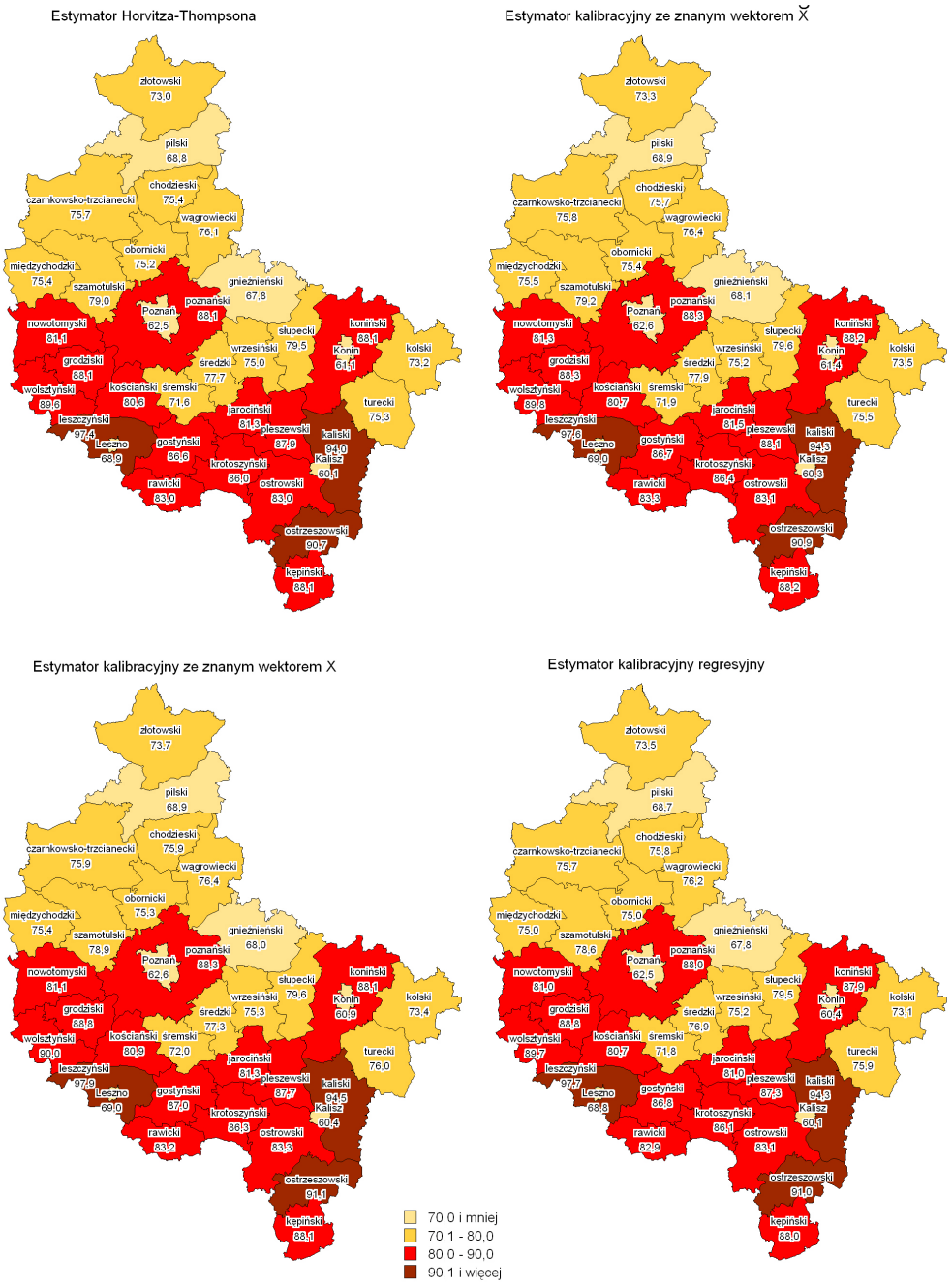
Dodatek A

Wyniki badań symulacyjnych



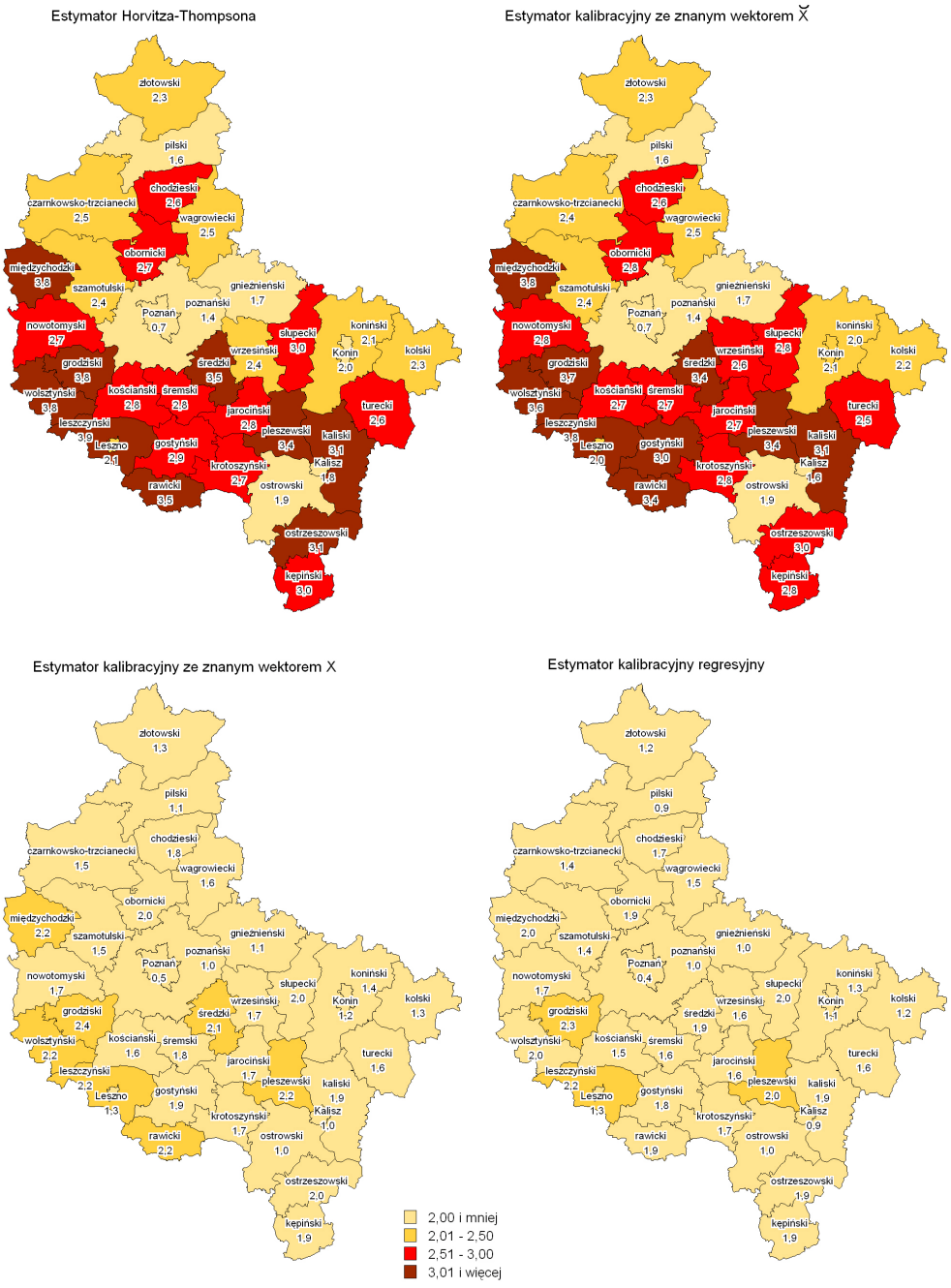
Rysunek. A.1. Średnia liczba izb w mieszkaniach w przekroju powiatów województwa wielkopolskiego w 2002r.

Źródło: Opracowanie własne na podstawie danych z NSP'2002



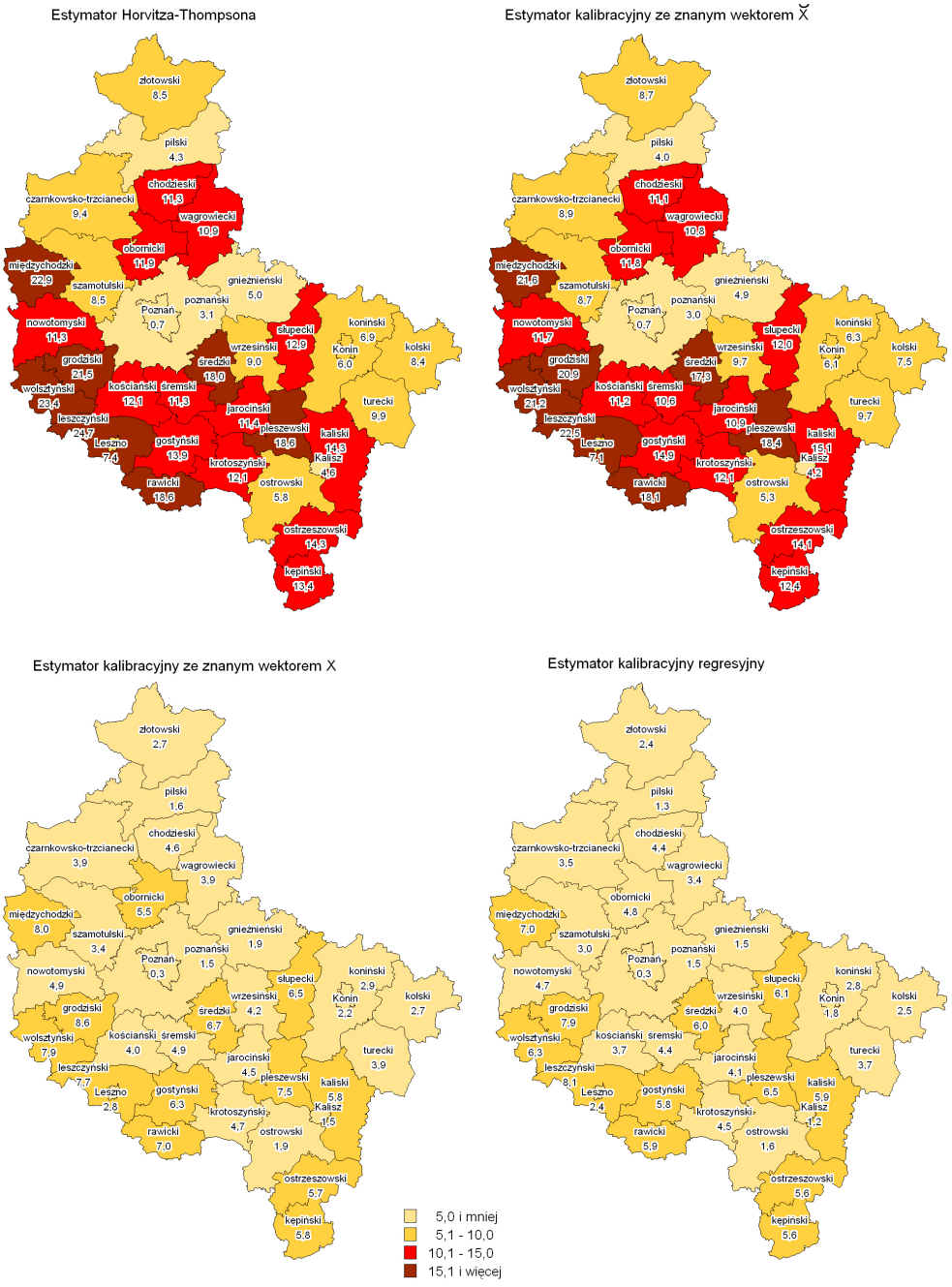
Rysunek. A.2. Wartość oczekiwana estymatorów średniej powierzchni mieszkań (w m^2) w przekroju powiatów województwa wielkopolskiego – wariant 1, próba 1%, frakcja braków danych 10%

Źródło: Opracowanie własne na podstawie wyników badań symulacyjnych



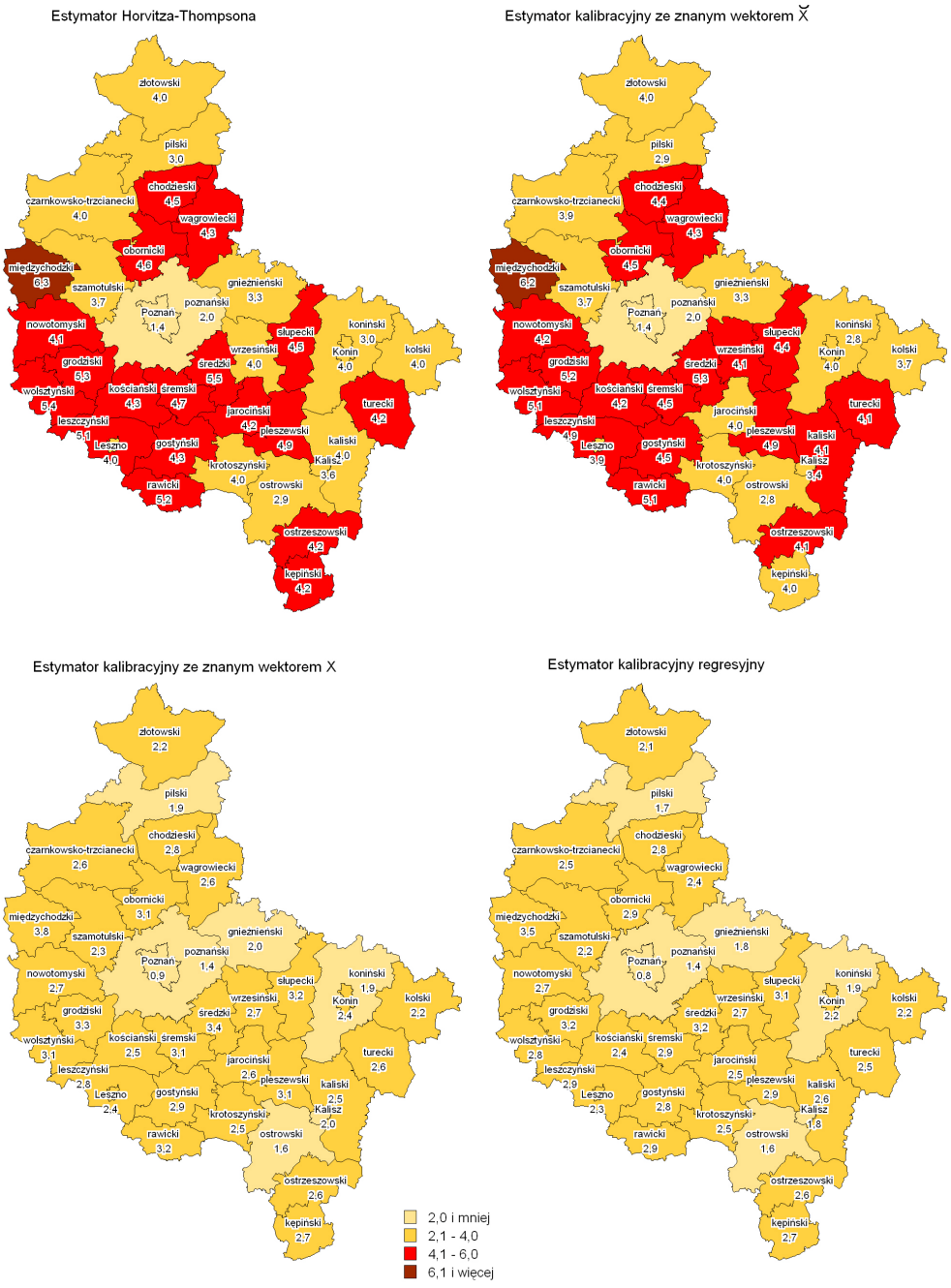
Rysunek. A.3. Wartość oczekiwana obciążenia estymatorów średniej powierzchni mieszkań (w m^2) w przekroju powiatów województwa wielkopolskiego – wariant 1, próba 1%, frakcja braków danych 10%

Źródło: Opracowanie własne na podstawie wyników badań symulacyjnych



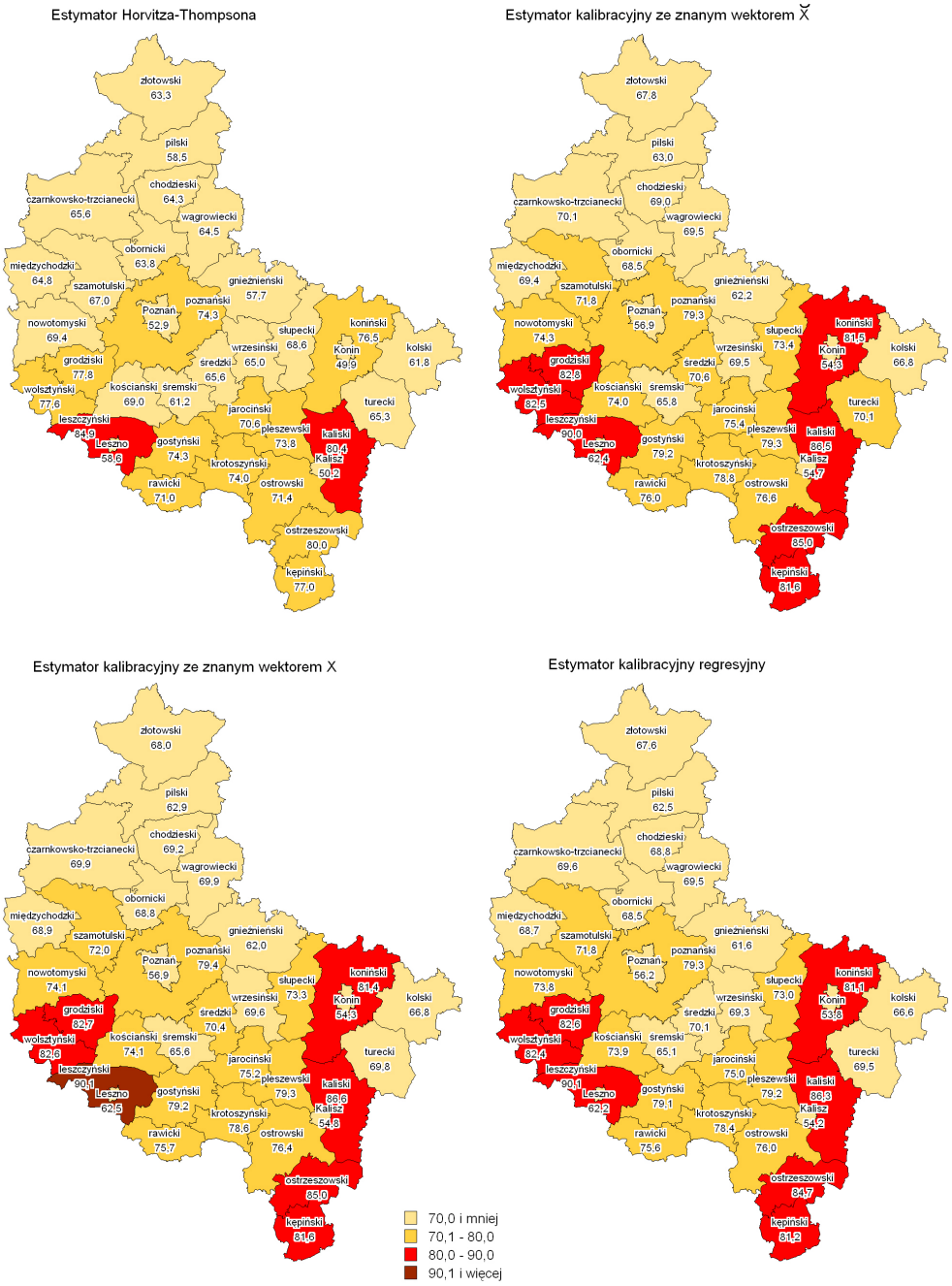
Rysunek. A.4. Wariancja estymatorów średniej powierzchni mieszkań w przekroju powiatów województwa wielkopolskiego – wariant 1, próba 1%, frakcja braków danych 10%

Źródło: Opracowanie własne na podstawie wyników badań symulacyjnych



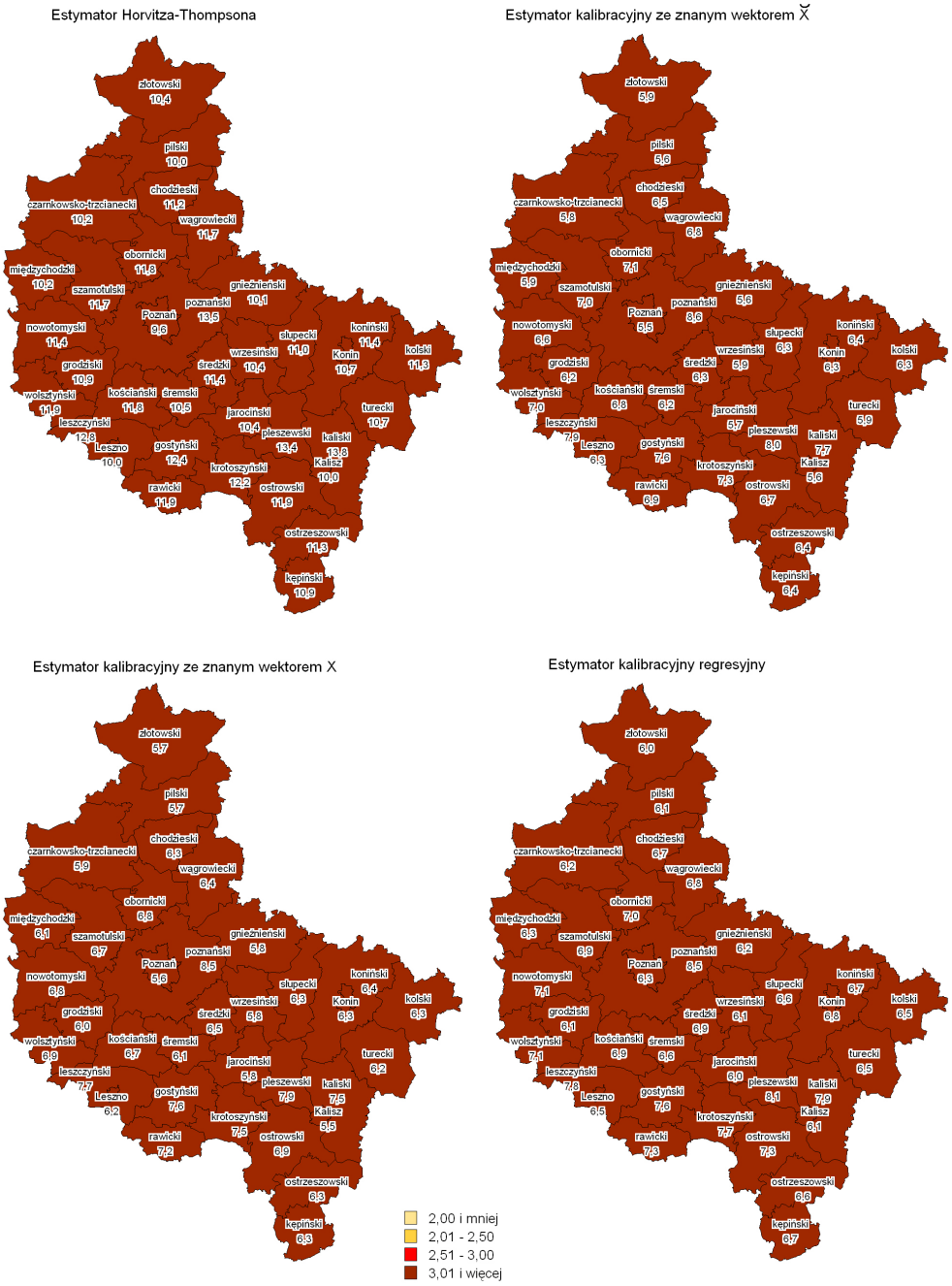
Rysunek. A.5. Względny błąd szacunku estymatorów średniej powierzchni mieszkań (w %) w przekroju powiatów województwa wielkopolskiego – wariant 1, próba 1%, frakcja braków danych 10%

Źródło: Opracowanie własne na podstawie wyników badań symulacyjnych



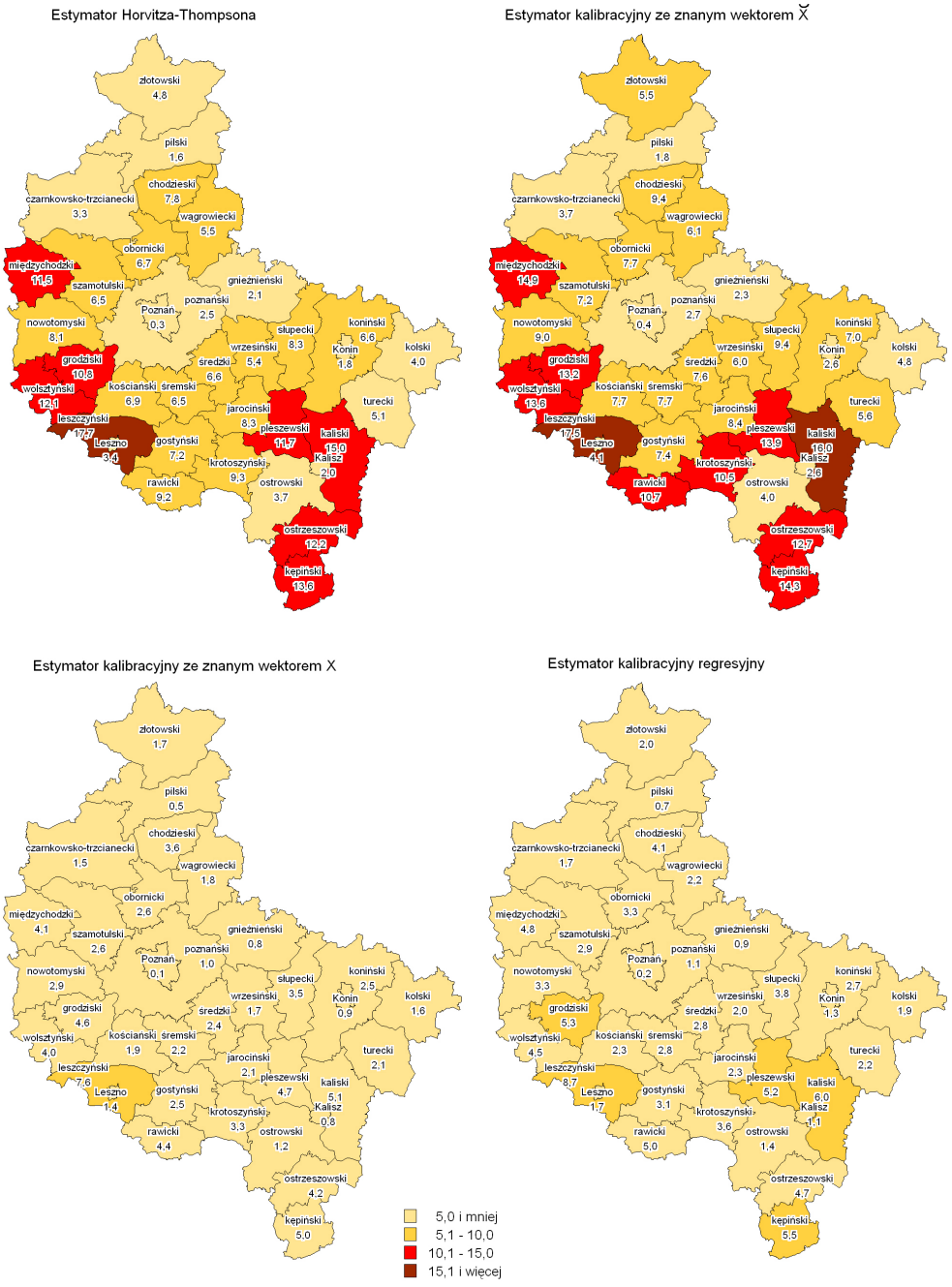
Rysunek. A.6. Wartość oczekiwana estymatorów średniej powierzchni mieszkań (w m^2) w przekroju powiatów województwa wielkopolskiego – wariant 3, próba 1%, frakcja braków danych 10%

Źródło: Opracowanie własne na podstawie wyników badań symulacyjnych



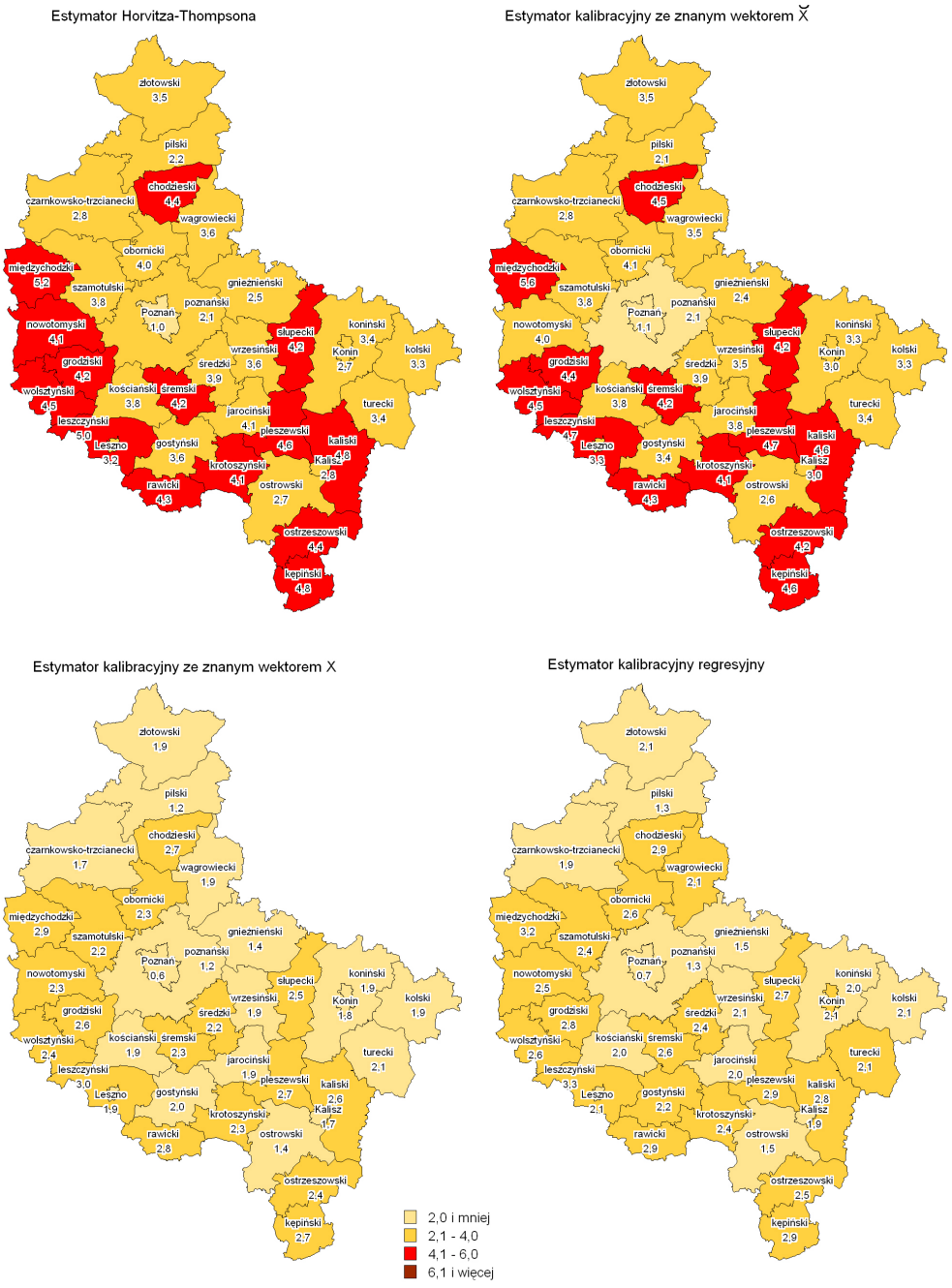
Rysunek. A.7. Wartość oczekiwana obciążenia estymatorów średniej powierzchni mieszkań (w m^2) w przekroju powiatów województwa wielkopolskiego – wariant 3, próba 1%, frakcja braków danych 10%

Źródło: Opracowanie własne na podstawie wyników badań symulacyjnych



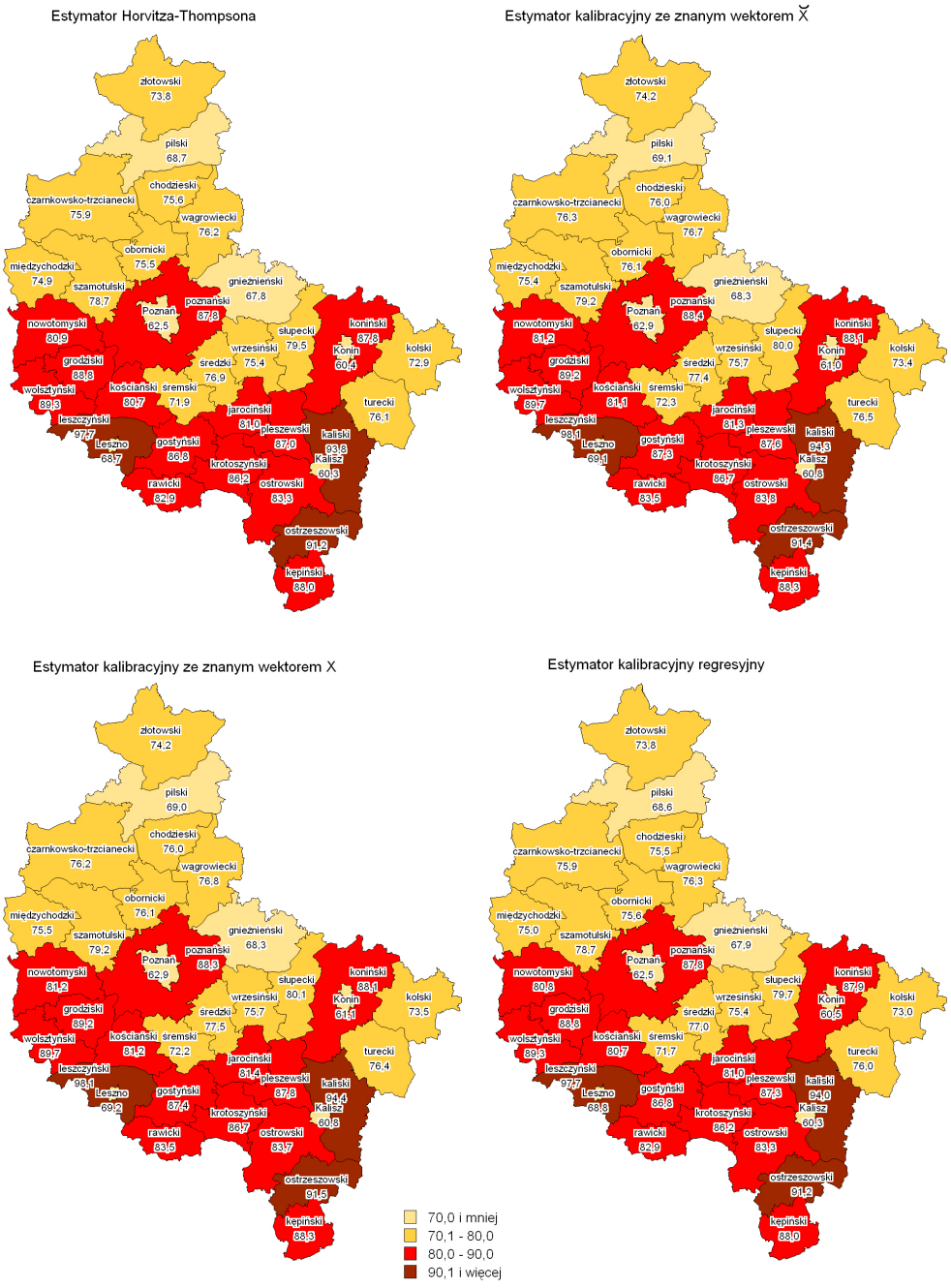
Rysunek. A.8. Wariancja estymatorów średniej powierzchni mieszkań w przekroju powiatów województwa wielkopolskiego – wariant 3, próba 1%, frakcja braków danych 10%

Źródło: Opracowanie własne na podstawie wyników badań symulacyjnych



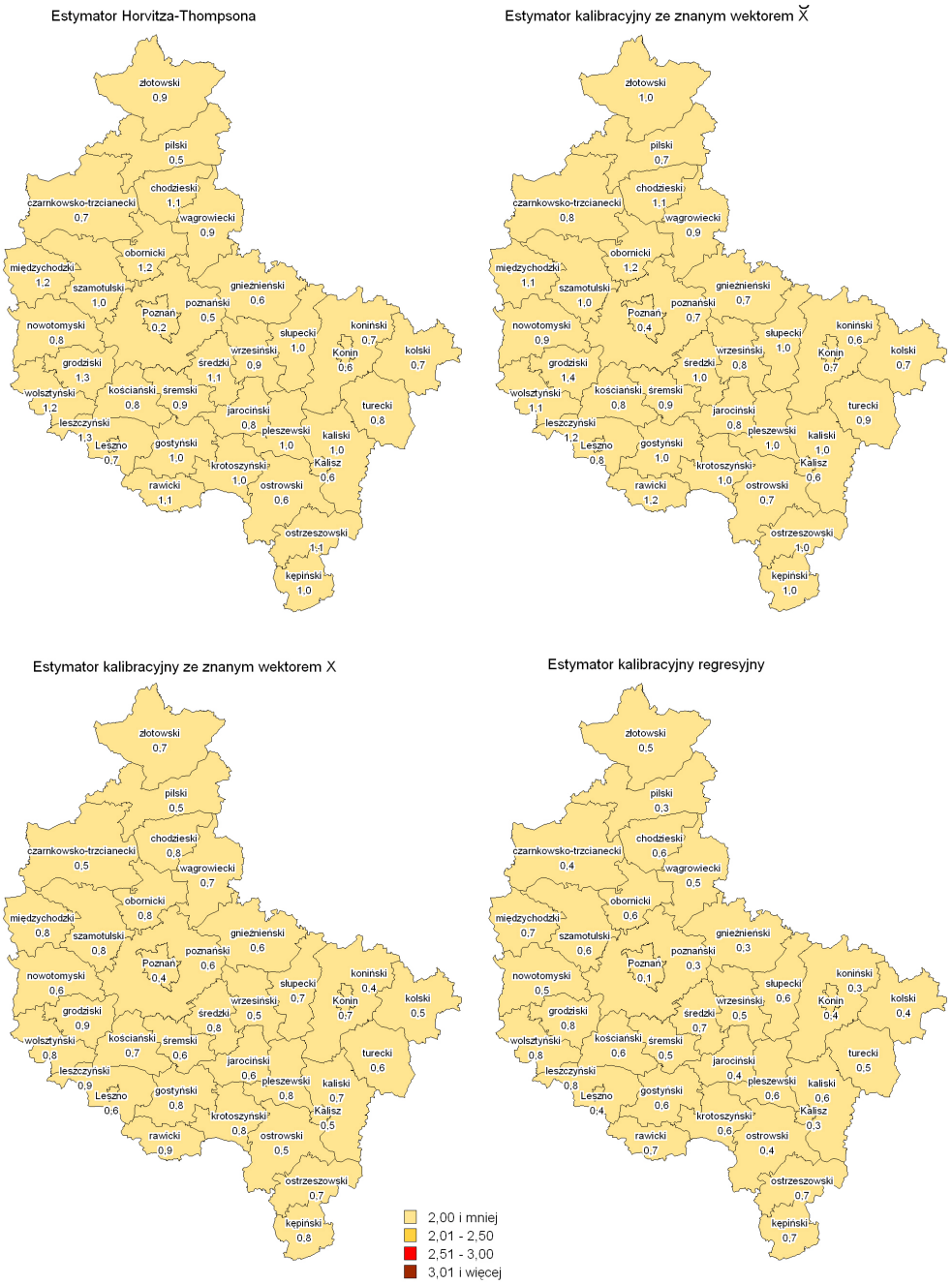
Rysunek. A.9. Względny błąd szacunku estymatorów średniej powierzchni mieszkań (w %) w przekroju powiatów województwa wielkopolskiego – wariant 3, próba 1%, frakcja braków danych 10%

Źródło: Opracowanie własne na podstawie wyników badań symulacyjnych



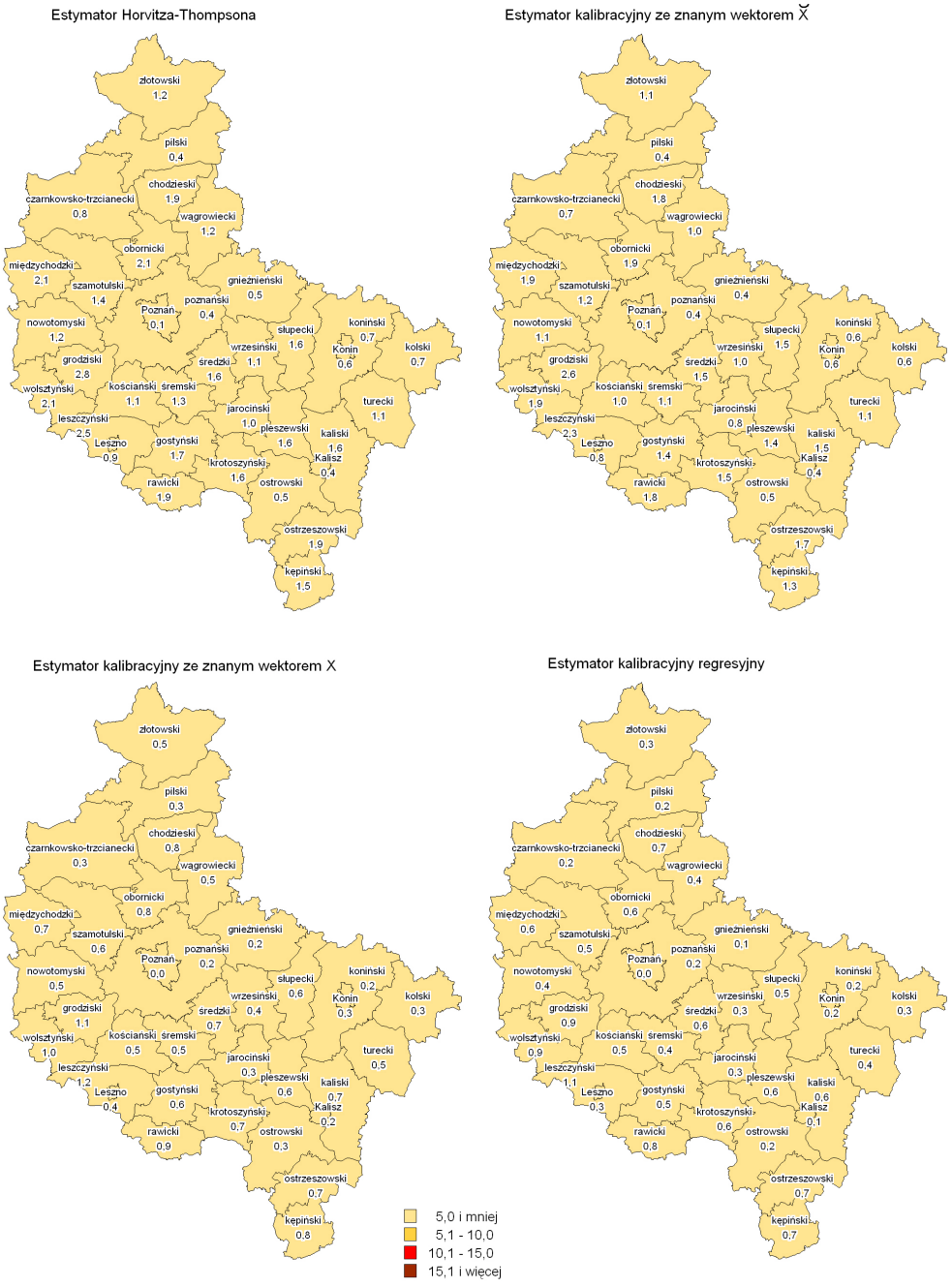
Rysunek. A.10. Wartość oczekiwana estymatorów średniej powierzchni mieszkań (w m^2) w przekroju powiatów województwa wielkopolskiego – wariant 1, próba 10%, frakcja braków danych 20%

Źródło: Opracowanie własne na podstawie wyników badań symulacyjnych



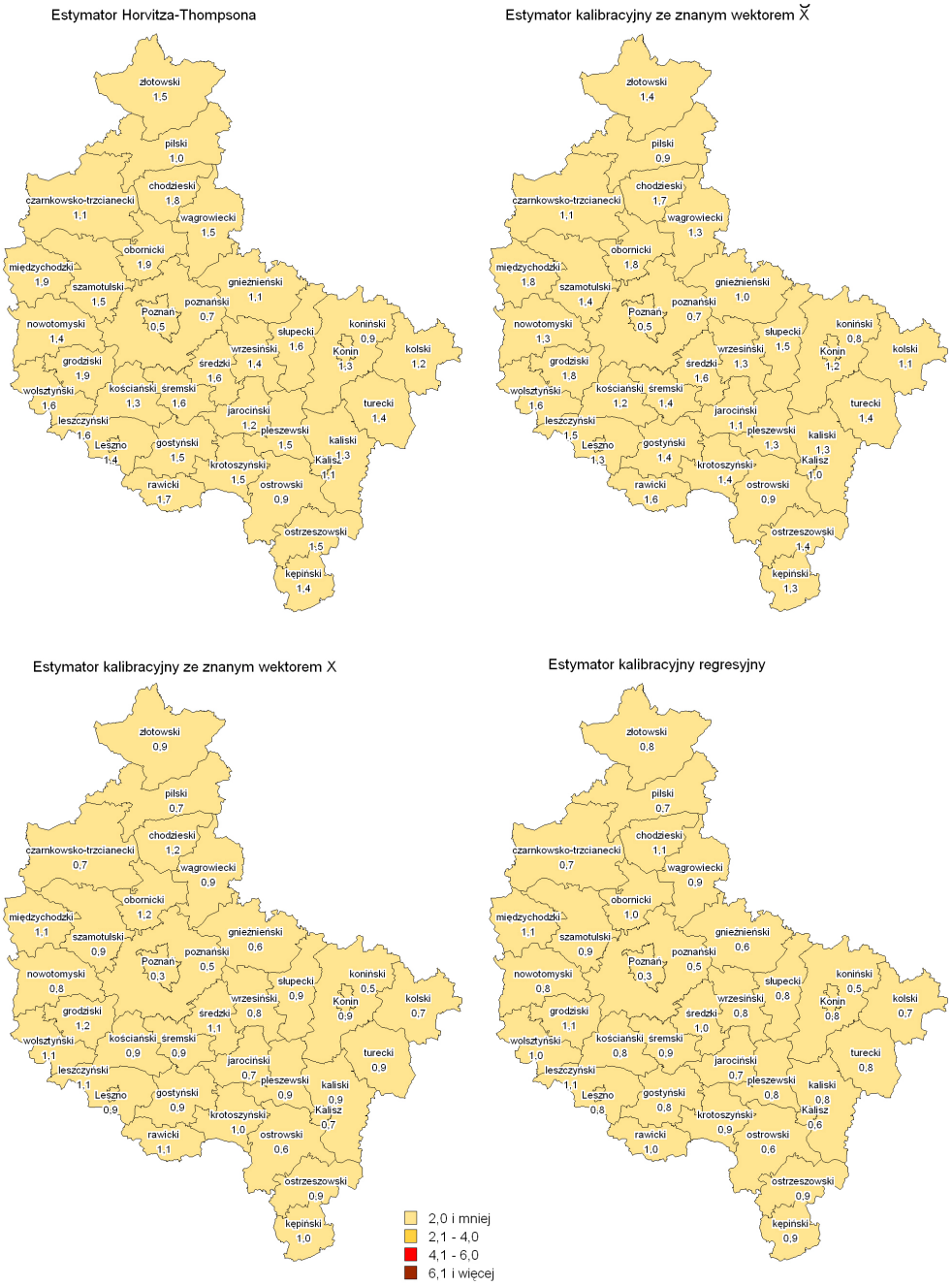
Rysunek. A.11. Wartość oczekiwana obciążenia estymatorów średniej powierzchni mieszkań (w m^2) w przekroju powiatów województwa wielkopolskiego – wariant 1, próba 10%, frakcja braków danych 20%

Źródło: Opracowanie własne na podstawie wyników badań symulacyjnych



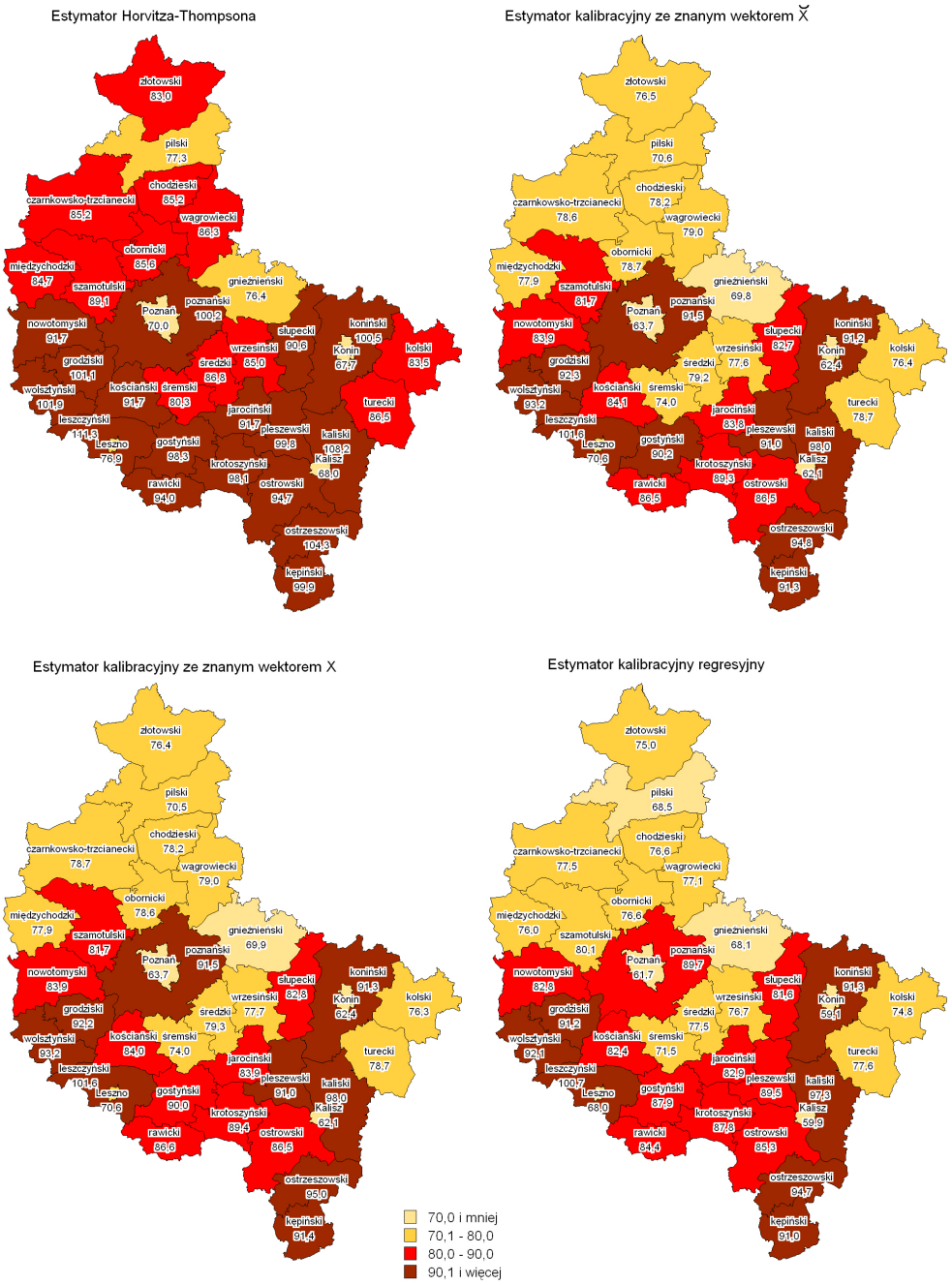
Rysunek. A.12. Wariancja estymatorów średniej powierzchni mieszkań w przekroju powiatów województwa wielkopolskiego – wariant 1, próba 10%, frakcja braków danych 20%

Źródło: Opracowanie własne na podstawie wyników badań symulacyjnych



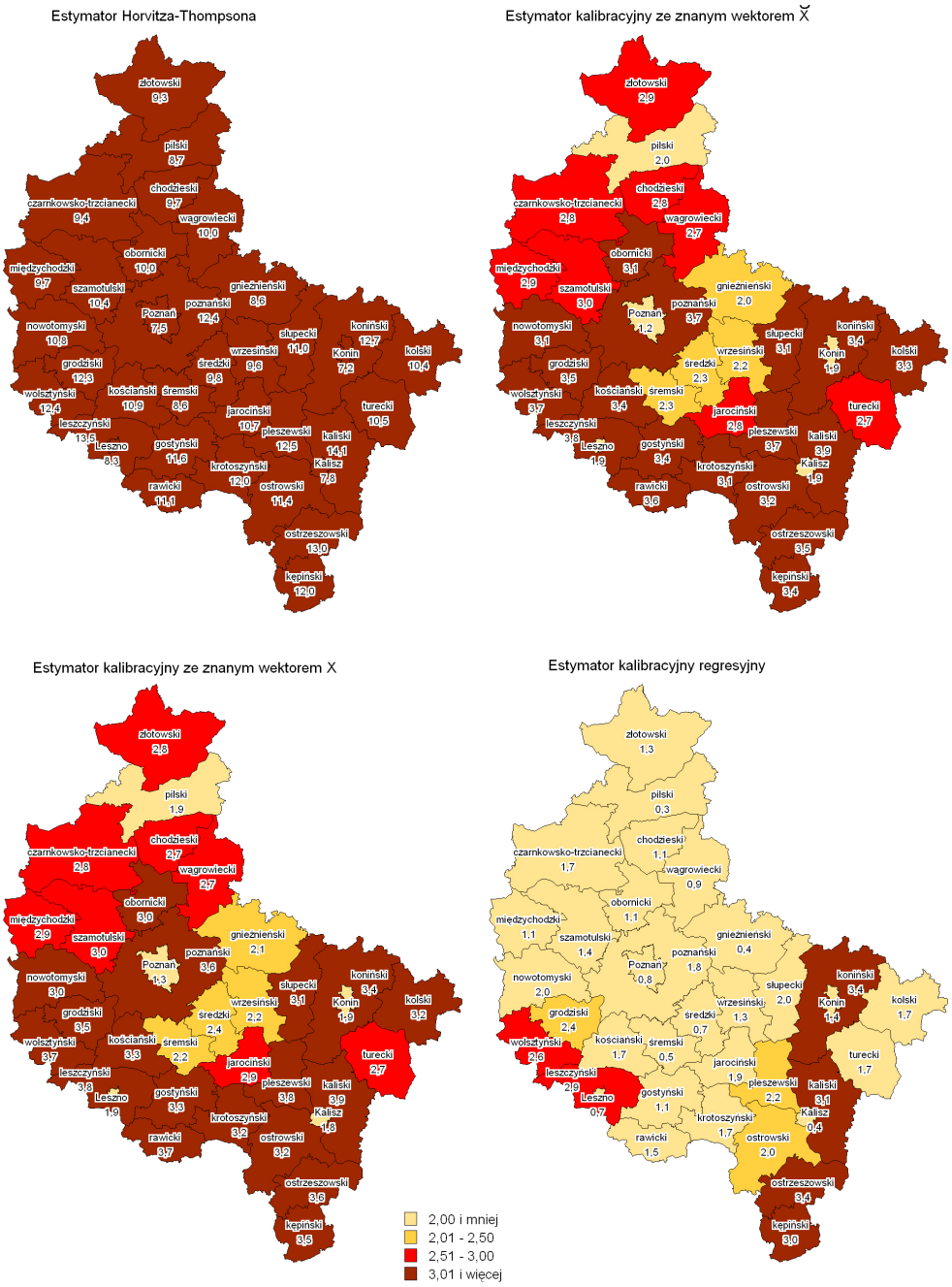
Rysunek. A.13. Względny błąd szacunku estymatorów średniej powierzchni mieszkań (w %) w przekroju powiatów województwa wielkopolskiego – wariant 1, próba 10%, frakcja braków danych 20%

Źródło: Opracowanie własne na podstawie wyników badań symulacyjnych



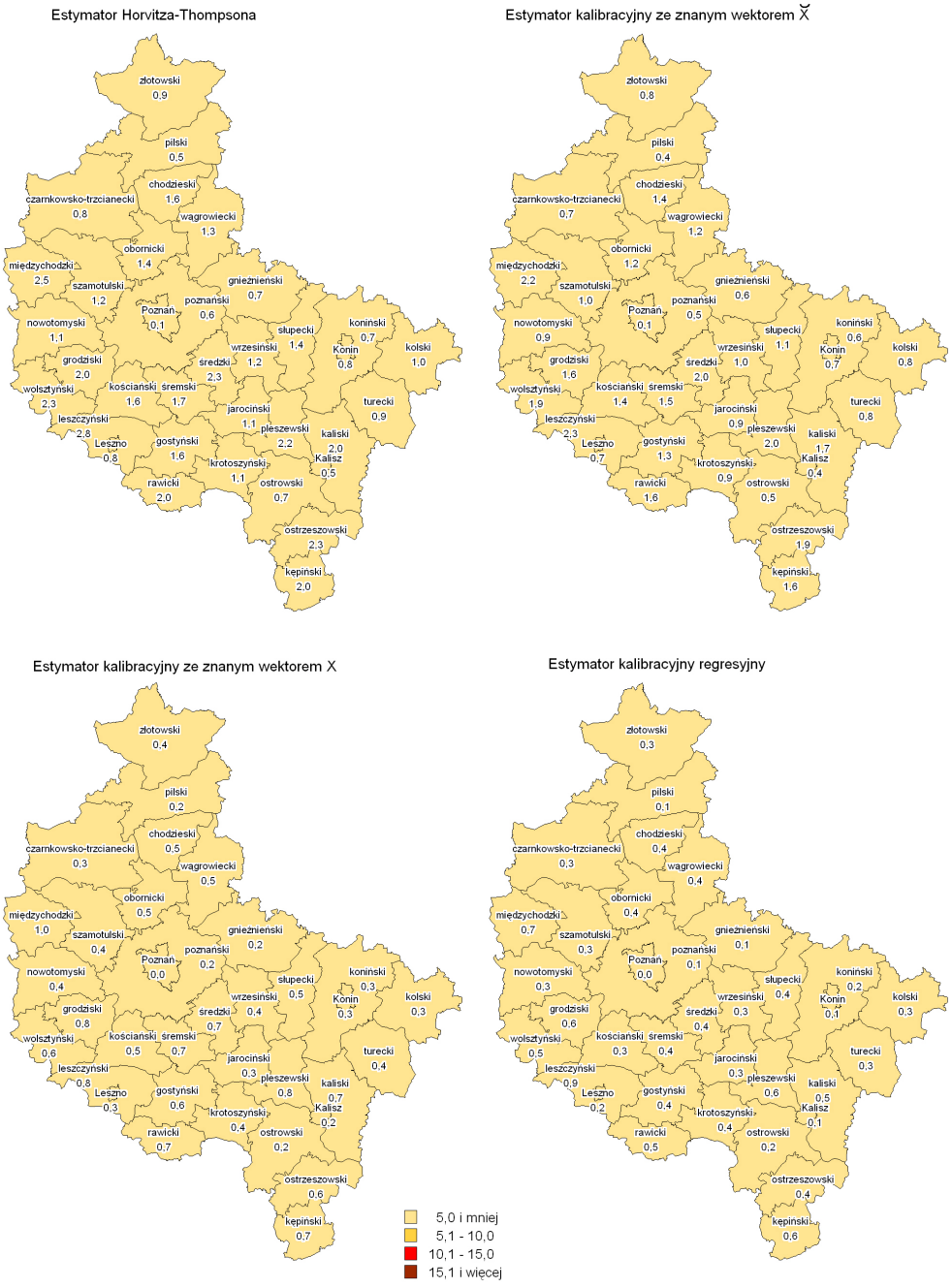
Rysunek. A.14. Wartość oczekiwana estymatorów średniej powierzchni mieszkań (w m^2) w przekroju powiatów województwa wielkopolskiego – wariant 2, próba 10%, frakcja braków danych 20%

Źródło: Opracowanie własne na podstawie wyników badań symulacyjnych



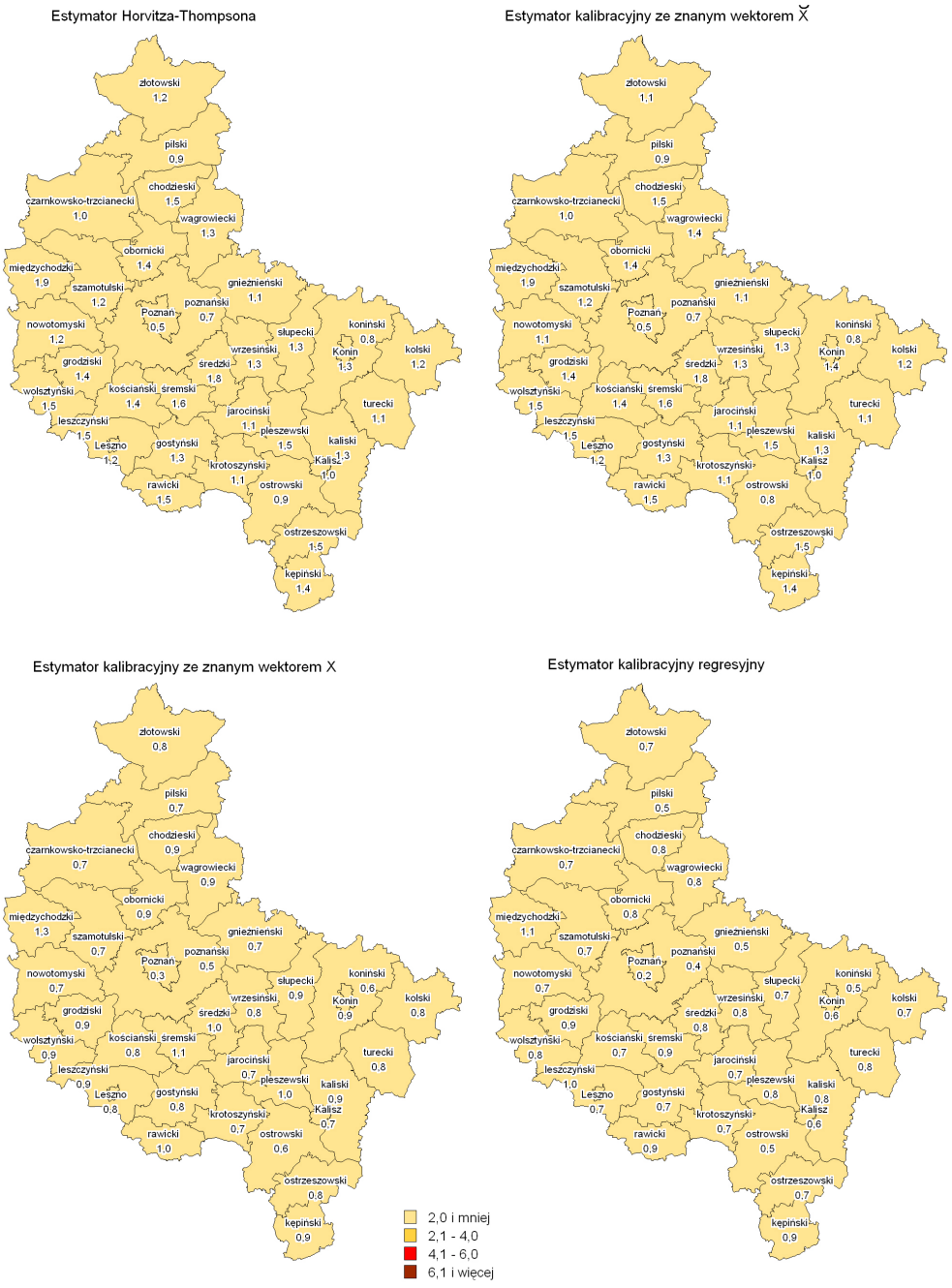
Rysunek. A.15. Wartość oczekiwana obciążenia estymatorów średniej powierzchni mieszkań (w m^2) w przekroju powiatów województwa wielkopolskiego – wariant 2, próba 10%, frakcja braków danych 20%

Źródło: Opracowanie własne na podstawie wyników badań symulacyjnych



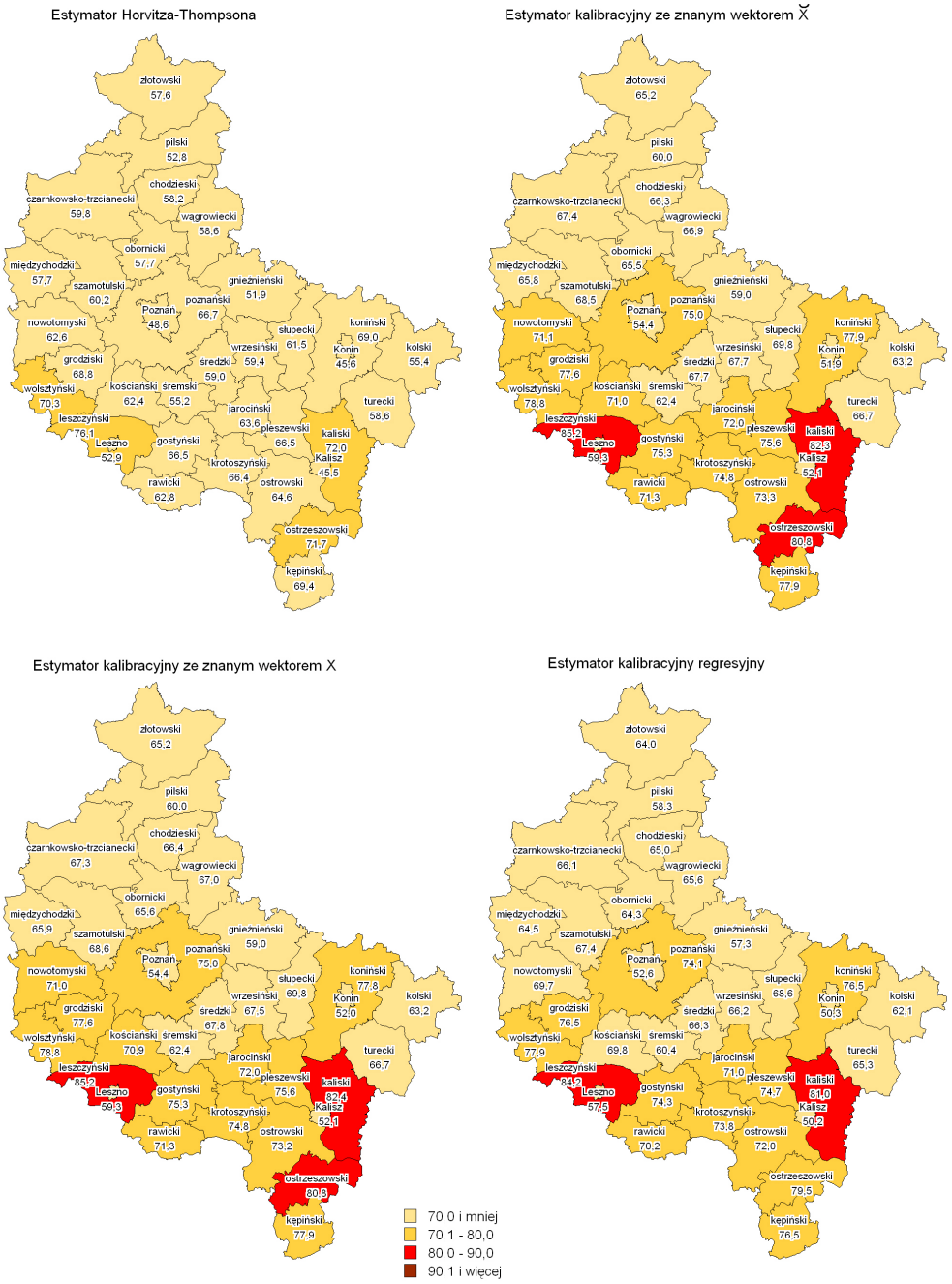
Rysunek. A.16. Wariancja estymatorów średniej powierzchni mieszkań w przekroju powiatów województwa wielkopolskiego – wariant 2, próba 10%, frakcja braków danych 20%

Źródło: Opracowanie własne na podstawie wyników badań symulacyjnych



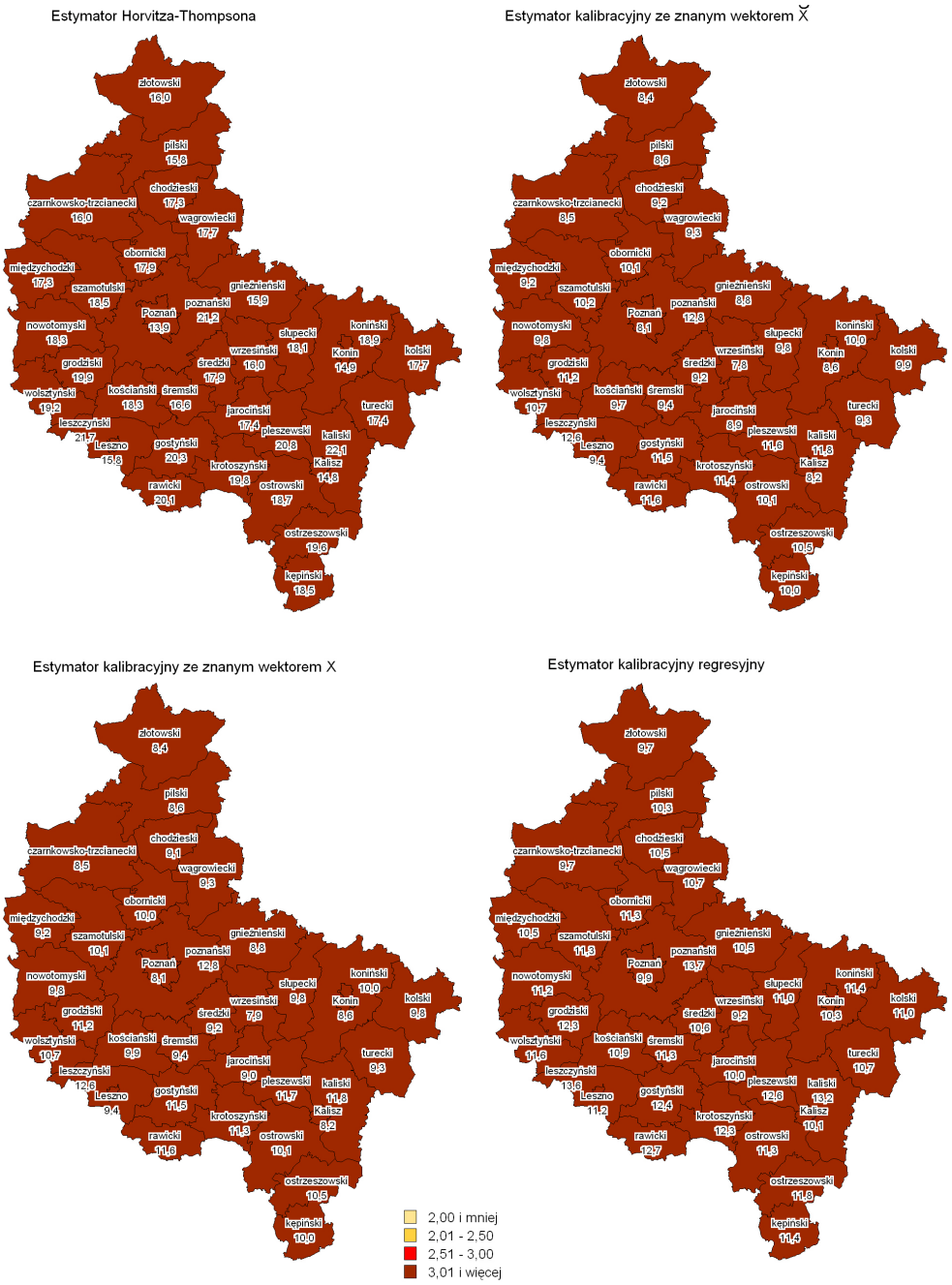
Rysunek. A.17. Względny błąd szacunku estymatorów średniej powierzchni mieszkań (w %) w przekroju powiatów województwa wielkopolskiego – wariant 2, próba 10%, frakcja braków danych 20%

Źródło: Opracowanie własne na podstawie wyników badań symulacyjnych



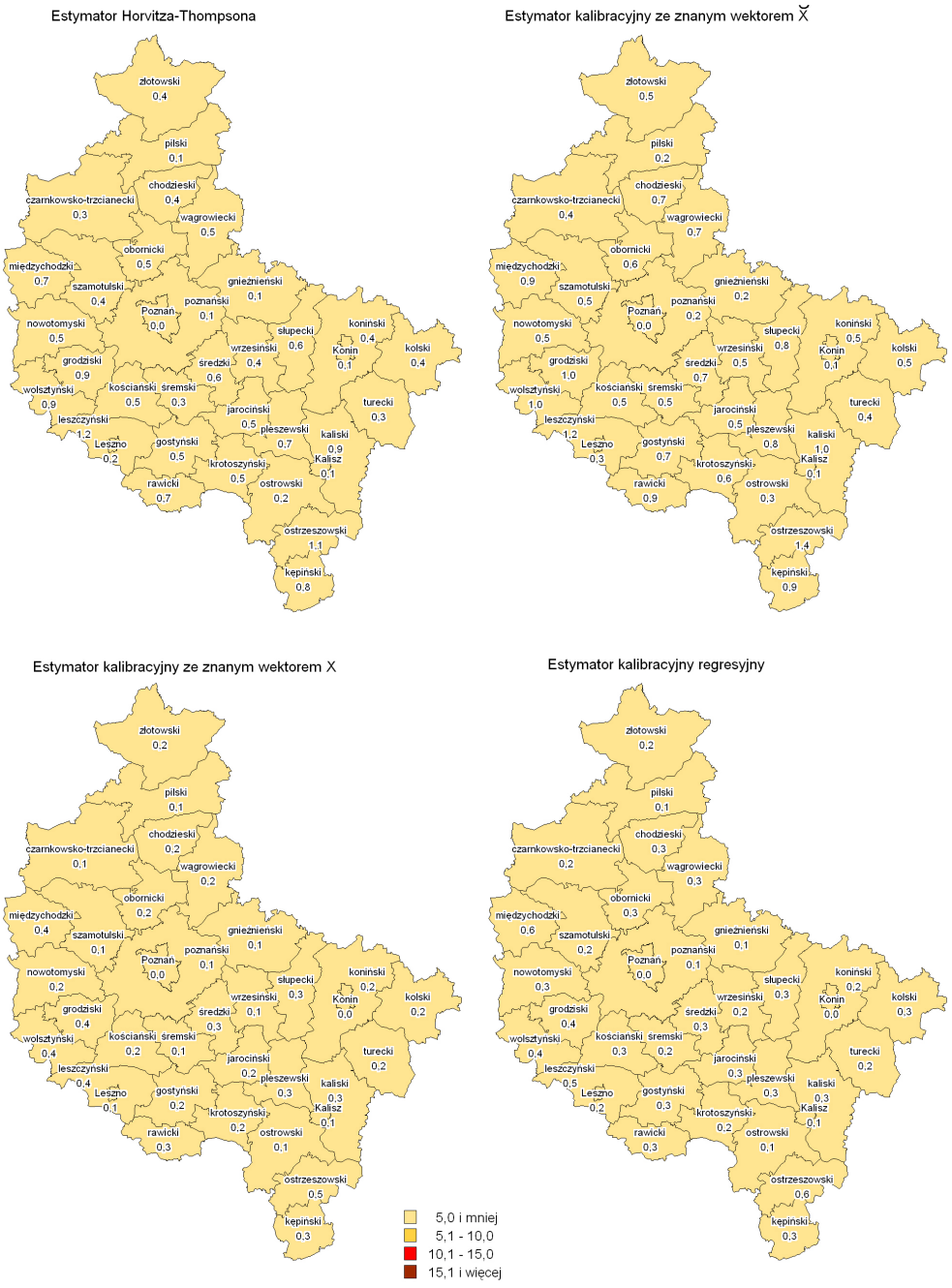
Rysunek. A.18. Wartość oczekiwana estymatorów średniej powierzchni mieszkań (w m^2) w przekroju powiatów województwa wielkopolskiego – wariant 3, próba 10%, frakcja braków danych 20%

Źródło: Opracowanie własne na podstawie wyników badań symulacyjnych



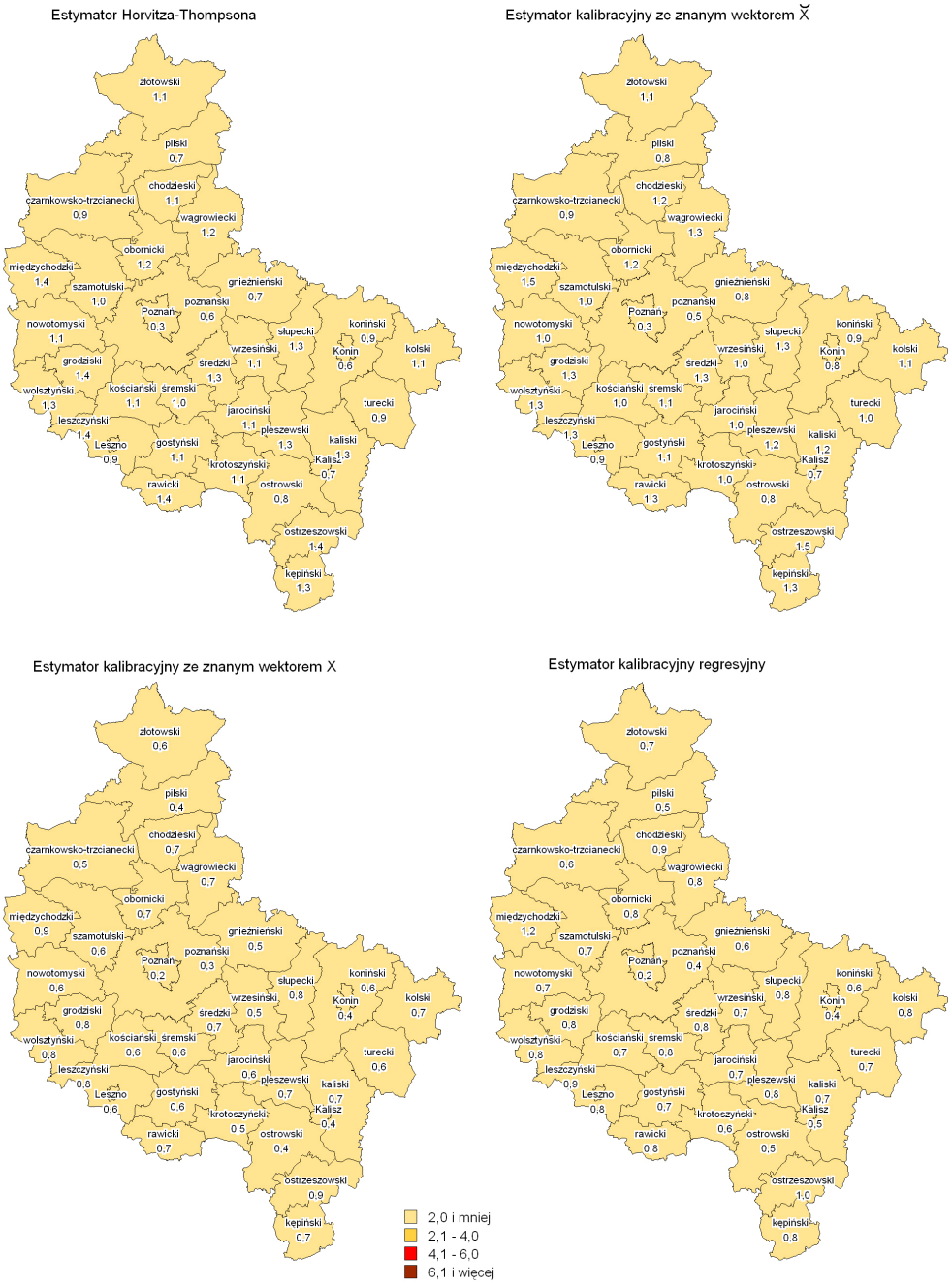
Rysunek. A.19. Wartość oczekiwana obciążenia estymatorów średniej powierzchni mieszkań (w m^2) w przekroju powiatów województwa wielkopolskiego – wariant 3, próba 10%, frakcja braków danych 20%

Źródło: Opracowanie własne na podstawie wyników badań symulacyjnych



Rysunek. A.20. Wariancja estymatorów średniej powierzchni mieszkań w przekroju powiatów województwa wielkopolskiego – wariant 3, próba 10%, frakcja braków danych 20%

Źródło: Opracowanie własne na podstawie wyników badań symulacyjnych



Rysunek. A.21. Względny błąd szacunku estymatorów średniej powierzchni mieszkań (w %) w przekroju powiatów województwa wielkopolskiego – wariant 3, próba 10%, frakcja braków danych 20%

Źródło: Opracowanie własne na podstawie wyników badań symulacyjnych

Spis rysunków

Spis tablic

Skorowidz

S

T

U

W

Z

Wykaz ważniejszych symboli i oznaczeń matematycznych

Oznaczenie	Nazwa
α	rząd kwantyla: $\alpha \in (0,1)$
k	liczba zmiennych pomocniczych
m	liczebność zbioru respondentów
n	liczebność próby
N	liczebność populacji
r	zbiór respondentów
ρ_{XY}	współczynnik korelacji liniowej Pearsona w populacji
$\mathbb{R}$	zbiór liczb rzeczywistych
s	próba
U	populacja
d_i	waga odpowiadająca i – tej jednostce wylosowanej do próby
d_{pi}	waga początkowa odpowiadająca i – tej jednostce wylosowanej do próby
y_i	wartość zmiennej y dla i – tej jednostki badania
π_i	prawdopodobieństwo inkluzji i – tej jednostki do próby
π_{ij}	prawdopodobieństwo inkluzji i – tej i j– tej jednostki do próby
$\mathbf{d}$	wektor wag wynikających ze schematu losowania próby
$\mathbf{d}_p$	wektor wag początkowych w podejściu kalibracyjnym
$\mathbf{w}$	wektor wag kalibracyjnych
$\boldsymbol{x}_i$	wektor złożony z wartości zmiennych pomocniczych dla i – tego respondenta
$\boldsymbol{z}_i$	wektor zmiennych instrumentalnych odpowiadających i – temu respondentowi
$Q_{y,\alpha}$	kwantyl rzędu α zmiennej y
$\hat{Q}_{y,cal,\alpha}$	estymator kalibracyjny kwantyla rzędu α zmiennej y
$\mathbf{X}$	wektor wartości globalnych zmiennych pomocniczych
$\hat{\mathbf{X}}$	wektor oszacowanych wartości globalnych zmiennych pomocniczych na zbiorze respondentów r
$\breve{\mathbf{X}}$	wektor oszacowanych wartości globalnych zmiennych pomocniczych na podstawie próby s

Oznaczenie	**Nazwa**
$\tilde{\mathbf{X}}$	wektor oszacowanych wartości globalnych zmiennych pomocniczych na zbiorze respondentów r z wykorzystaniem wag kalibracyjnych
$D(\mathbf{v}, \mathbf{d})$	funkcja odległości między wektorami $\mathbf{v}$ i $\mathbf{d}$
$F_y(t)$	wartość dystrybuanty zmiennej y w punkcie t
$\hat{F}_{y,cal}(t)$	wartość dystrybuanty interpolacyjnej zmiennej y w punkcie t
Y	wartość globalna zmiennej y
$\hat{Y}_{cal}$	uogólniony estymator kalibracyjny wartości globalnej zmiennej y
$\hat{Y}_{Fcal}$	estymator kalibracyjny wartości globalnej zmiennej y skonstruowany w oparciu o podejście funkcyjne
$\hat{Y}_{HT}$	estymator Horvitza-Thompsona wartości globalnej zmiennej y
$\hat{Y}_{RA}$	estymator kalibracyjny ilorazowy wartości globalnej zmiennej y
$\hat{Y}_{REG}$	estymator kalibracyjny regresyjny wartości globalnej zmiennej y
$\hat{Y}_{\mathbf{X}}$	estymator kalibracyjny wartości globalnej zmiennej y ze znanym wektorem $\mathbf{X}$ wartości globalnych zmiennych pomocniczych
$\hat{Y}_{\check{\mathbf{X}}}$	estymator kalibracyjny wartości globalnej zmiennej y ze znanym wektorem $\check{\mathbf{X}}$ oszacowanych wartości globalnych zmiennych pomocniczych
L	funkcja Lagrange'a
$\mathbb{R}^m$	przestrzeń wektorowa m–wymiarowa
$\frac{\partial L}{\partial w_i}$	pochodna cząstkowa funkcji Lagrange'a $L : \mathbb{R}^m \to \mathbb{R}$, gdzie $(i = 1, \ldots, m)$
$\boldsymbol{\lambda}$	wektor czynników nieoznaczonych Lagrange'a
$\operatorname{argmin}_{\mathbf{v}} D$	argument, dla którego funkcja D osiąga wartość minimalną na zbiorze wszystkich wektorów $\mathbf{v} \in \mathbb{R}^m$