NR 9 (640)
WRZESIEŃ 2014

WIADOMOŚCI STATYSTYCZNE

CZASOPISMO GŁÓWNEGO URZĘDU STATYSTYCZNEGO
I POLSKIEGO TOWARZYSTWA STATYSTYCZNEGO

MIĘDZYNARODOWY ROK STATYSTYKI 2013
KONFERENCJA NAUKOWA
STATYSTYKA — WIEDZA — ROZWÓJ

Tomasz PIASECKI

Metody imputacji w badaniach gospodarstw domowych

Występowanie braków odpowiedzi w badaniach statystycznych jest istotnym problemem związanym z ich realizacją. Radzenie sobie z tym zjawiskiem i ograniczenie jego negatywnych konsekwencji dla uzyskiwanych wyników stanowią ważne wyzwanie metodologiczne.

W stosunku do pozycyjnych braków odpowiedzi, tj. takich, gdy jednostka statystyczna biorąca udział w badaniu nie udziela odpowiedzi na niektóre pytania, możliwe jest przyjęcie przez badacza różnych strategii. Przyjęcie restrykcyjnych wymogów co do kompletności zbieranych wywiadów (ograniczenie możliwości nieudzielenia odpowiedzi przez respondenta na poszczególne pytania) pozwala uzyskać spójne i kompletne zbiory danych wynikowych, jednak może powodować wzrost częstości występowania jednostkowych braków odpowiedzi (takich, gdy jednostka statystyczna w ogóle nie bierze udziału w badaniu). Dopuszczenie odmów odpowiedzi na pytania pozwala ograniczyć występowanie odmów udzielenia wywiadu w ogóle, pogarsza jednak kompletność i spójność wewnętrzną oraz może zmniejszać użyteczność uzyskanego zbioru danych. Pojawiają się wtedy w nim pozycyjne braki danych, wymagające przyjęcia określonej procedury postępowania. Rozwiązaniem, które przywraca w dużym stopniu zbiorowi danych niekompletnych użyteczność i funkcjonal-

ność, podobnie jak w przypadku danych kompletnych, może być imputacja (choć z istotnymi zastrzeżeniami dotyczącymi wnioskowania na podstawie takiego zbioru).

Autor omawia w artykule przykłady zastosowania procedur imputacyjnych w polskiej statystyce publicznej, dotyczące dwóch badań statystyki społecznej, tj. *Europejskiego badania dochodów i warunków życia* (EU-SILC) oraz *Badania spójności społecznej*.

Są to badania ważne i bardzo obszerne. EU-SILC jest regularnym badaniem statystyki publicznej, realizowanym co roku. Jego zakres tematyczny dotyczy szeroko rozumianych warunków życia oraz dochodów ludności. Badaniu podlega duża liczba cech reprezentujących różne typy dochodu, opisujących w sposób kompleksowy całość dochodów gospodarstwa domowego.

Badanie spójności społecznej ma charakter wieloaspektowy — łączy tematy dotyczące różnych dziedzin życia i zjawisk społecznych. Jego celem jest przedstawienie złożonego obrazu jakości i warunków życia. Badanie to zostało zrealizowane przez polską statystykę po raz pierwszy w 2011 r. Planowane są kolejne edycje w odstępie 4—5 lat.

Obydwa badania są badaniami reprezentacyjnymi, realizowanymi przez ankieterów statystycznych na próbie gospodarstw domowych. W badaniach tych imputacja stosowana jest w przypadku braków pozycyjnych dotyczących dochodów gospodarstw domowych oraz ich składowych (w przypadku EU-SILC mamy do czynienia z bardzo dużą liczbą składowych, zarówno dotyczących całego gospodarstwa jak i jego członków). Imputacja pozycyjna w tych badaniach dotyczy przede wszystkim dochodów, gdyż generalnie dla pozostałych zmiennych wymaga się kompletności zapisów już podczas zbierania danych. Jest to kompromis między obydwoma wspomnianymi wyżej podejściami do występowania braków odpowiedzi, który szerzej opisuje dalsza część artykułu.

PODSTAWOWE POJĘCIA

Imputacja polega na zastąpieniu braków danych w zbiorze wyników badania tzw. wartościami imputacyjnymi. Są to oszacowania brakujących wartości prawdziwych, tworzone w trakcie imputacji, najbardziej właściwe ze względu na przyjęte kryteria, założenia, zastosowane metody. Imputację można więc określić jako proces szacowania danych brakujących[1].

Imputacja jest wykonywana dla danych jednostkowych i prowadzi do uzyskania pozornie kompletnego zbioru jednostkowego. Należy jednak pamiętać, że stosowane metody imputacji pozwalają na poprawne wnioskowanie na podstawie zbioru imputowanego dopiero dla określonych agregatów. Nie jest natomiast właściwym traktowanie wartości imputacyjnych danych jednostkowych na równi z informacjami pochodzącymi z wywiadu i wnioskowanie na ich podsta-

[1] Balicki A. (2004).

wie o pojedynczych jednostkach statystycznych. Na imputację należy zatem patrzeć raczej jako na ekwiwalent lub uzupełnienie metod wnioskowania statystycznego i uogólniania wyników niż na ekwiwalent wywiadu.

Z punktu widzenia zakresu przedmiotowego imputacji, tj. tego, jaka część rekordu danych jest imputowana, **wyróżnić można dwa typy imputacji:**

- **pozycyjną** — w sytuacji, gdy imputowane są brakujące informacje o jednostkach statystycznych, które wzięły udział w badaniu, ale nie udzieliły odpowiedzi na pojedyncze pytania (imputacja dotyczy tylko niektórych pól w danym rekordzie);
- **brakujących rekordów** — gdy imputowana jest pełna informacja o jednostkach statystycznych, które w ogóle nie wzięły udziału w badaniu (pewne rekordy danych są imputowane w całości).

Obydwa typy imputacji odpowiadają dwóm podstawowym typom braków danych: imputacja pozycyjna dotyczy pozycyjnych braków danych (*item nonresponse*), zaś imputacja brakujących rekordów — jednostkowych braków danych (*unit nonresponse*).

W badaniach społecznych statystyki publicznej stosowana jest przede wszystkim imputacja pozycyjna. Pewne elementy imputacji brakujących rekordów występują w badaniu EU-SILC, gdzie imputuje się brakujące formularze indywidualne (osobowe). Dotyczy to wyłącznie pojedynczych osób należących do gospodarstw domowych, które (jako gospodarstwo) wzięły udział w badaniu. Nie jest to więc pełny brak jednostkowy — dla danego gospodarstwa domowego, dysponuje się bowiem uzyskanym od respondentów wywiadem zbiorczym gospodarstwa oraz częścią wywiadów indywidualnych.

W przypadku wystąpienia braków jednostkowych w omawianych badaniach (niemożności uzyskania wywiadu od całego gospodarstwa domowego), podstawową metodą niwelowania skutków takich braków jest odpowiednia korekta wag uogólniających (w tym kalibracja), a nie imputacja.

WYMOGI W ZAKRESIE KOMPLETNOŚCI DANYCH I JAKOŚĆ WYNIKÓW

Jak wspomniano na wstępie, możliwe są różne podejścia (strategie) dotyczące wymogów służących zapewnieniu kompletności danych podczas ich zbierania (wywiadów). Wymogi te mogą być stosowane w sposób bardziej lub mniej ostry, co może skutkować brakiem występowania pozycyjnych braków danych (w przypadku ostrych wymogów) lub ich występowaniem w różnym zakresie, zależnie od przyjętej strategii. Wyróżnić można zatem:

- strategię restrykcyjną, czyli wymaganie udzielenia kompletnych odpowiedzi na wszystkie pytania wywiadu (wymóg pełnej kompletności), co stanowi warunek akceptacji wywiadu i przekazania danych do dalszego przetwarzania — podejście takie eliminuje występowanie pozycyjnych braków danych;
- strategię mniej restrykcyjną (łagodną), czyli dopuszczenie odmowy odpowiedzi na niektóre pytania wywiadu, nieskutkującą dyskwalifikacją całego wywiadu — oznacza to dopuszczenie wystąpienia braków pozycyjnych.

Obydwa te podejścia mają zalety i wady (zestawienie). W przypadku podejścia mniej restrykcyjnego negatywne skutki wiążą się głównie z powstaniem pozycyjnych braków danych, w przypadku podejścia restrykcyjnego są przede wszystkim wynikiem utraty części wywiadów (odrzucanych ze względu na niespełnienie kryteriów kompletności) oraz możliwości uzyskania odpowiedzi mniej rzetelnych, gdy zostają one „wymuszone". Wady podejścia zakładającego wymóg pełnej kompletności danych ujawniają się przede wszystkim w przypadku występowania w badaniu pytań problematycznych, drażliwych, wzbudzających szczególną niechęć do udzielania odpowiedzi wśród respondentów.

ZESTAWIENIE. PORÓWNANIE ROZMAITEGO PODEJŚCIA DO WYMOGU KOMPLETNOŚCI DANYCH

Wyszczególnienie	Wymóg pełnej kompletności	Dopuszczenie braków pozycyjnych
Zalety	• spójny i kompletny zbiór wynikowy • lepsza jakość danych wynikowych, jeżeli uda się uzyskać rzetelne odpowiedzi na problematyczne pytania	• zmniejszenie częstości występowania braków jednostkowych • większe prawdopodobieństwo, że uzyskane odpowiedzi na pytania problematyczne są rzeczywiście rzetelne
Wady	• możliwy wzrost częstości występowania braków jednostkowych • możliwe pogorszenie jakości danych wynikowych, jeżeli wymaganie pełnej kompletności skłoni respondentów do nierzetelnych odpowiedzi	• niekompletność zbioru danych wynikowych • trudności analityczne z przetwarzaniem uzyskanych danych • mniejsza elastyczność wykorzystania uzyskanych danych
Zastosowanie imputacji	• brak imputacji pozycyjnej	• wskazana imputacja pozycyjna

Ź r ó d ł o: opracowanie własne.

W badaniach statystyki publicznej (w tym dotyczących gospodarstw domowych) zazwyczaj przyjmuje się, co do zasady, założenie kompletności wywiadu, czyli wymaga się udzielenia odpowiedzi na wszystkie pytania. Nie zawsze eliminuje to całkowicie występowanie pozycyjnych braków danych (gdyż czasem przesłanki obiektywne rzeczywiście uniemożliwiają uzyskanie odpowiedzi, a brak nie jest na tyle istotny by dyskwalifikować wywiad), ale marginalizuje problem. Od opisanej reguły czyni się jednak wyjątki w przypadku pytań szczególnie drażliwych, na które respondenci odpowiadają z większą niechęcią lub wręcz trudno byłoby wymagać od nich, by udzielili odpowiedzi. Bezwzględne wymaganie kompletności wywiadu w zakresie dotyczącym takich pytań prowadziłoby do dyskwalifikacji dużej liczby wywiadów i znaczącego zwiększenia odsetka odmów, dlatego zwykle dopuszcza się odmowę odpowiedzi na dane pytanie bez dyskwalifikacji całego wywiadu. Zasada ta stosowana jest bardzo często do pytań dotyczących dochodów.

Z taką właśnie sytuacją mamy do czynienia w EU-SILC oraz *Badaniu spójności społecznej*. Zasady badania dopuszczają odmowę odpowiedzi przez respon-

denta na pytania dotyczące dochodów, co nie skutkuje dyskwalifikacją wywiadu. Na oznaczenie odmowy odpowiedzi stosowany jest specjalny kod. W przypadku pozostałych pytań ankiety generalnie stosuje się wymóg uzyskania odpowiedzi od respondenta, dlatego projektując algorytm imputacji można założyć kompletność (lub stan bliski kompletności) dla zmiennych niedotyczących dochodów.

Dopuszczenie braków pozycyjnych dla pytań dotyczących dochodów w żadnym wypadku nie oznacza zaniechania starań o uzyskanie odpowiedzi na dane pytanie przez ankietera — oznaczenie braku danych stosowane jest dopiero wtedy, gdy zabiegi te nie przynoszą rezultatu, a odmowa jest ostateczna.

METODY IMPUTACJI DANYCH BRAKUJĄCYCH

Istnieje wiele metod imputacji danych oraz różne kryteria ich klasyfikacji[2]. Z punktu widzenia poruszanej tematyki ważne jest rozróżnienie imputacji statystycznej od imputacji dedukcyjnej, zaś w zakresie imputacji statystycznej podział metod na deterministyczne i stochastyczne.

Omawiane metody są metodami elementarnymi, dotyczą tzw. imputacji pojedynczej (jednokrotnej). Alternatywę dla imputacji jednokrotnej stanowi tzw. imputacja wielokrotna[3]. Metoda ta nie jest stosowana w rozważanych tu badaniach, dlatego nie będzie szczegółowo omawiana. Zastosowano podejście oparte na użyciu elementarnych metod imputacji jednokrotnej, z wykorzystaniem modeli odrębnie dobieranych do poszczególnych zmiennych i weryfikowanych merytorycznie. Wybór taki motywowany jest dążeniem do zachowania jak największej kontroli nad imputacją oraz poprawnością merytoryczną i spójnością logiczną otrzymywanych wyników. Z tego powodu nie zdecydowano się na użycie metod zakładających większą automatyzację doboru modeli imputacyjnych, traktując je jako zbyt ryzykowne. Mimo tego, imputacja wielokrotna stanowi wciąż braną pod uwagę alternatywę w kontekście dalszych prac nad algorytmami imputacji ze względu na możliwość dostarczenia precyzyjnej informacji o wielkości błędu losowego związanego z niepewnością wartości imputacyjnych, a tym samym poprawy jakości uzyskiwanych oszacowań błędów dla uogólnień tworzonych na podstawie zbioru będącego wynikiem imputacji.

---

[2] Opis i klasyfikację metod imputacji znaleźć można m.in. w następujących opracowaniach: Longford N. T. (2005); Kalton G., Kasprzyk D. (1982).

[3] Podejście zaproponowane przez D. Rubina. W bardzo uproszczony sposób można je przedstawić jako polegające na kilkukrotnym zastosowaniu stochastycznych metod imputacji jednokrotnej (z uwzględnieniem w części stochastycznej wszystkich źródeł niepewności modeli generujących wartości imputacyjne) do całego zbioru (wszystkich zmiennych podlegających imputacji). W efekcie otrzymujemy kilka wariantów zbioru zaimputowanego. Posiadanie kilku wariantów zbioru pozwala oszacować niepewność wartości imputacyjnych, a tym samym nie tylko dokonać uogólnień na podstawie zbioru, ale także prawidłowo oszacować błędy tych uogólnień, uwzględniając niepewność imputacji. Opis metody znaleźć można m.in. w: Rubin D. B. (1987), Longford N. T. (2005).

Generalnie wyróżnić można **dwa podstawowe typy imputacji — dedukcyjną i statystyczną. Imputacja dedukcyjna** opiera się na zależnościach między zmiennymi i na regułach; można ją taktować jako część procesu redagowania danych. Wartość imputacyjna wyznaczana jest bezpośrednio na podstawie tych zależności. Imputacja dedukcyjna ma charakter deterministyczny, tzn. wykorzystując przyjęte reguły można wartość imputacyjną wyznaczyć w sposób jednoznaczny.

W przypadku **imputacji statystycznej** do imputacji danych brakujących wykorzystywana jest pozostała część zbioru danych, czyli dane uzyskane od innych respondentów („niebrakujące") dotyczące zmiennej imputowanej. Imputacja statystyczna opiera się na określonym modelu (wyspecyfikowanym jawnie lub określonym niejawnie). Wykorzystuje ona zależności wykryte metodami statystycznymi w dostępnym zbiorze danych. Pozwala to na nieobciążoną estymację parametrów populacji, gdy modele (założenia) są prawidłowe.

Metody imputacji statystycznej podzielić można na:

- **deterministyczne** — gdy w tworzeniu wartości imputacyjnej nie występuje element losowy. Wartość imputacyjną określa jednoznacznie metoda i wejściowy zbiór danych. Dla danego zbioru danych, powtarzając imputację, otrzymamy zawsze te same wartości imputacyjne (ten sam zbiór wyjściowy);
- **stochastyczne** — gdy tworzenie wartości imputacyjnej zawiera element losowy. Dla danego wejściowego zbioru danych możemy uzyskać różne zbiory wartości zaimputowanych (zbiory wyjściowe). Zbiór wyjściowy (pozycje w tym zbiorze podlegające imputacji) nie jest więc określony w sposób jednoznaczny.

Metody deterministyczne zapewniają zwykle większą precyzję uogólnień dokonywanych na podstawie zbioru zaimputowanego niż metody stochastyczne, gdyż nie wprowadzają dodatkowego źródła błędu losowego. Z tego punktu widzenia mogą wydawać się optymalne. Ich wadą jest jednak to, że całkowicie pomijają rzeczywistą niepewność związaną z oszacowaniem wartości imputacyjnej, wskutek czego zniekształcają rozkłady imputowanych zmiennych (w tym zaniżają miary rozrzutu i błędu). Rzeczywisty błąd uogólnień jest zwykle nieco mniejszy niż w przypadku imputacji stochastycznej, ale jego oszacowanie staje się obciążone. Obciążone zostają też wartości wszystkich miar, których wartość zależy od kształtu rozkładu.

Metody stochastyczne generują dodatkowy błąd wynikający z imputacji. Jest on związany z elementem losowym w nich zawartym, co w pewnym (zwykle niewielkim) stopniu zwiększa błąd uogólnień. Zaletą metod stochastycznych (wielokrotnie przesądzającą o przewadze nad metodami deterministycznymi) jest jednak to, że lepiej zachowują rozkłady zmiennych, nie zniekształcając ich znacząco. W mniejszym stopniu niż metody deterministyczne zniekształcają one zależności między zmiennymi, natomiast nie zniekształcają kształtu rozkładów jednowymiarowych, nie powodują obciążenia miar rozrzutu, od których z kolei zależą miary błędu.

Ponadto wyróżnić można metody tworzące sztuczne wartości zmiennych (np. średnie wartości teoretyczne z modelu) oraz tzw. metody „oparte na dawcach", czyli takie, w których wartość imputacyjna przenoszona jest z innego rekordu, a nie generowana sztucznie.

W omawianych badaniach[4] stosowane są następujące metody imputacji:
spośród metod **deterministycznych:**

- **imputacja średnia** — za wartość imputacyjną przyjmowana jest średnia z obserwacji prawidłowych („niebrakujących"). Zwykle stosowana jest imputacja średnia w klasach (grupach) imputacyjnych, utworzonych ze względu na określone kryteria. Mamy więc wtedy do czynienia z imputacją średnią warunkową. Dobór zmiennych grupujących stanowi określenie „modelu" takiej imputacji. Kryteria grupowania mogą mieć postać hierarchiczną;
- **deterministyczna imputacja regresyjna** — realizowana na podstawie modelu regresyjnego objaśniającego zmienną imputowaną za pomocą zmiennych kompletnych (lub kompletnych w takim zakresie, jaki jest wystarczający do dopasowania modelu i dokonania imputacji dla konkretnego rekordu). Za wartość imputacyjną przyjmowana jest wartość teoretyczna z modelu;

spośród metod **stochastycznych:**

- **hot-deck**[5] — imputacja danymi innego rekordu (tzw. dawcy), wylosowanego spośród rekordów kompletnych (przynajmniej w zakresie imputowanej zmiennej). Zwykle wybór dawcy ograniczany jest do rekordów należących do tej samej klasy (grupy) imputacyjnej, a więc rekordów spełniających określone kryteria podobieństwa. Ich dobór stanowi określenie „modelu" takiej imputacji. Podobnie jak w przypadku imputacji średnią, kryteria mogą mieć charakter hierarchiczny, co jest stosowane w EU-SILC;
- **stochastyczna imputacja regresyjna** — podobnie jak w wariancie deterministycznym, opiera się na modelu regresyjnym objaśniającym zmienną imputowaną. Oprócz części deterministycznej modelu uwzględnia składnik losowy, którego „realizacje" (reszty losowe) tworzone są (pseudo) losowo przy użyciu odpowiedniego generatora. Wartość imputacyjną stanowi wartość teoretyczna z modelu uzupełniona o resztę losową. Możliwe są różne generatory, jak również dodatkowe warunki i reguły stosowane przy generowaniu reszt. Specyficzne elementy z tym związane występują w obydwu omawianych badaniach i wymagają odrębnego opisu.

Wybór procedury imputacji w przypadku każdej imputowanej zmiennej wymaga określenia metody, modelu, doboru zmiennych grupujących, a także innych specyficznych elementów związanych z poszczególnymi metodami. Zagadnienia te ze względu na ich specyfikę przedstawiono odrębnie dla każdego z omawianych badań.

---

[4] Wszystkie wymienione metody używane są w EU-SILC. W *Badaniu spójności społecznej* zastosowana została stochastyczna imputacja regresyjna.

[5] Metoda ta bywa też określana w literaturze jako imputacja losowa (w klasach). Z kolei określenie *hot-deck* bywa również rozumiane nieco inaczej niż tu opisano (Kalton G., Kasprzyk D., 1982). W artykule jest ono stosowane zgodnie z przedstawionym opisem.

EUROPEJSKIE BADANIE DOCHODÓW I WARUNKÓW ŻYCIA (EU-SILC)

Istotny element zakresu tematycznego badania EU-SILC stanowią dochody ludności. Informacje zbierane są przez ankieterów statystyki publicznej w wywiadzie osobistym, na wylosowanej próbie gospodarstw domowych (adresów), przy użyciu kwestionariusza gospodarstwa domowego oraz kwestionariusza indywidualnego. Uzyskiwane dane obejmują kilkaset zmiennych dotyczących osoby i gospodarstwa domowego. Informacja o dochodach występuje na obydwu poziomach pomiaru. Ma bardzo rozbudowany charakter, obejmuje dużą liczbę zmiennych reprezentujących różne kategorie (typy) dochodów, np. z pracy najemnej i na rachunek własny (z rozbiciem na pracę w rolnictwie i poza rolnictwem) czy wiele różnych typów świadczeń. Wszystkie te typy dochodów stanowią odrębne zmienne podlegające imputacji[6].

Badanie służy nie tylko publikacji uogólnionych wyników na użytek krajowy. Jego efektem jest także zbiór danych jednostkowych przekazywany do Eurostatu spełniający specjalne wymogi, co do użyteczności analitycznej oraz musi mieć ściśle określoną strukturę. Wymogi te są istotne z punktu widzenia imputacji i rzutują na postać zaprojektowanego algorytmu. Ważne jest na przykład, że ostateczny zbiór danych jednostkowych otrzymany w wyniku badania musi m.in. umożliwiać wyznaczanie nieliniowych wskaźników nierównomierności dochodów. Skutkuje to zaleceniem, by przy imputacji stosować metody, które nie deformują rozkładów zmiennych.

Projektując badanie przyjęto za dopuszczalne nieudzielenie przez respondentów podczas wywiadu odpowiedzi na pytania dotyczące wszystkich składowych dochodu gospodarstwa domowego (zarówno dotyczące osoby, jak i gospodarstwa). Powstałe pozycyjne braki danych są odpowiednio oznaczane i poddawane imputacji. Podlegają jej także dochody osób, które w ogóle nie udzieliły wywiadu indywidualnego, ale należą do gospodarstw, w których wywiadu udzielono (imputacja brakujących formularzy indywidualnych w części dotyczącej dochodów).

Ważnym argumentem za przyjęciem takiego podejścia w przypadku EU-SILC jest jego panelowy charakter. Każde gospodarstwo biorące udział w badaniu obserwowane jest czterokrotnie w kolejnych latach. Obserwacje te tworzą panel. Z tego powodu konsekwencje odmowy odpowiedzi dotyczącej całego wywiadu bądź dyskwalifikacji wywiadu ze względu na braki odpowiedzi byłyby poważniejsze niż gdybyśmy mieli do czynienia z jednorazową obserwacją. Utrata wywiadu rzutowałaby na jakość danych panelowych — tracilibyśmy nie tylko bieżący wywiad, ale także informacje dotyczące danego gospodarstwa z wywiadów wcześniejszych (które traciłyby przez to na wartości) oraz przyszłych (bo nie zostałyby w ogóle przeprowadzone).

[6] Opis badania i jego metodologii, w tym podstawowe informacje na temat imputacji (*Dochody*..., 2013).

Przed imputacją danych w badaniu stawia się następujące cele:

- uzyskanie kompletnego (pozbawionego braków) zbioru danych, zawierającego tzw. zmienne obowiązkowe (porównywalne dla całej Unii Europejskiej);
- zapewnienie, na potrzeby statystyki krajowej, możliwości uogólnień bardziej szczegółowych niż poziom zmiennych obowiązkowych, których wartości będą spójne z wartościami uogólnień dokonywanych na podstawie zbioru zmiennych obowiązkowych;
- zachowanie rozkładu zmiennych imputowanych, czyli uniknięcie deformacji rozkładu tych zmiennych w imputacji;
- zapewnienie możliwości obliczania wskaźników (w tym wskaźników zmienności i nierównomierności dochodów) na podstawie zbioru po imputacji.

Algorytm imputacji danych dla badania został opracowany w Ośrodku Statystyki Matematycznej Urzędu Statystycznego w Łodzi[7]. Placówka ta odpowiada też za realizację imputacji oraz rozwijanie i aktualizowanie algorytmu.

Szczegółowa postać algorytmu oraz główne jego założenia i zasady są pochodną celów, jakie postawiono przed imputacją. Zasady te tworzą jednocześnie schemat procedury imputacyjnej. Można je przedstawić następująco:

- preferowane są metody stochastyczne. Imputacja najważniejszej składowej każdej zmiennej dochodowej (zwykle dochód netto) wykonywana jest zawsze metodą stochastyczną. Imputacja deterministyczna może dotyczyć podatków, składek obciążających dochód itp., które mają stosunkowo niewielki udział w wartości globalnej poszczególnych komponentów dochodu (zmiennych finalnych). Preferencja dla metod stochastycznych wynika z potrzeby zachowania (uniknięcia znaczących zniekształceń) rozkładów i charakterystyk i zmiennych finalnych;
- poszczególne zmienne imputowane są oddzielnie, z wyjątkiem zmiennych ściśle powiązanych ze sobą zależnościami merytorycznymi oraz imputacji brakujących wywiadów indywidualnych. To, że ten proces realizowany jest oddzielnie dla każdej ze zmiennych, nie wyklucza istnienia statystycznych zależności między wartościami imputacyjnymi, wynikających albo z zastosowania wspólnych zmiennych objaśniających, albo z faktu, że zmienna zaimputowana wcześniej może być zmienną objaśniającą w modelu imputacyjnym;
- z zasady imputowane są dochody miesięczne;
- dopuszcza się i stosuje imputację jednej zmiennej za pomocą kilku alternatywnych metod (modeli) dla różnych podzbiorów rekordów ze względu na:
  - różną dostępność informacji o zmiennych pomocniczych (objaśniających) dla poszczególnych rekordów,
  - dostępność informacji o dochodzie tego samego typu z poprzedniego roku lub jej brak dla danego rekordu (możliwość wykorzystania danych panelowych).

[7] Informacje o algorytmie można znaleźć w publikacji GUS *Dochody*... (2013), s. 31—36, także w raportach z wcześniejszych edycji badania. Głównymi autorami algorytmu są T. Piasecki i D. Cybart.

Zasady te nie określają szczegółowo konkretnych metod imputacji. Są one dobierane na każdym kroku algorytmu (odpowiadającym zmiennej lub zmiennej i podzbiorowi rekordów), zgodnie z przedstawionymi zasadami, spośród zestawu metod. Stosowane są:

- metoda *hot-deck* (zwykle w klasach imputacyjnych),
- stochastyczna imputacja regresyjna,
- deterministyczna imputacja regresyjna,
- imputacja dedukcyjna.

Podstawowe znaczenie mają metoda *hot-deck* oraz stochastyczna imputacja regresyjna stosowane do imputacji najważniejszych zmiennych. Wybór pomiędzy nimi jest uzależniony od możliwości dopasowania modelu regresyjnego dobrze objaśniającego zmienną imputowaną, liczby obserwacji, na podstawie których taki model można dopasować, dostępności potencjalnych zmiennych objaśniających i ich liczby.

W przypadku metody *hot-deck* stosowane są hierarchiczne kryteria wyodrębniania klas (grup) imputacyjnych. Zmienne pomocnicze (zmienne grupujące, stanowiące kryterium wyodrębnienia klas) dla poszczególnych zmiennych imputowanych uporządkowano od najważniejszych do najmniej ważnych. W przypadku, gdy nie można znaleźć dawcy o odpowiadających wartościach wszystkich zmiennych pomocniczych lub gdy powstająca w ten sposób klasa imputacyjna nie jest wystarczająco liczna, sekwencyjnie pomija się kolejne kryteria, poczynając od najmniej ważnych. Taki hierarchiczny model grupowania ustalany jest dla każdego kroku algorytmu imputacji, w którym stosowana jest metoda *hot-deck*.

Podstawową metodą grupowania jest grupowanie według zmiennych jakościowych, których poszczególne kategorie służą bezpośrednio do wyznaczania klas. Jednak stosuje się również grupowanie według zmiennych ilościowych — w takim przypadku klasy imputacyjne tworzone są na podstawie grup kwantylowych.

W przypadku stochastycznej imputacji regresyjnej możliwe są różne sposoby generowania reszt losowych. W EU-SILC zastosowano podejście, w którym reszta dla danej pozycji imputowanej otrzymywana jest poprzez losowy wybór ze zbioru rzeczywistych reszt modelu. Resztę losuje się nie ze zbioru wszystkich reszt, lecz z ograniczonego, odpowiednio wyodrębnionego podzbioru. Stanowią go reszty dotyczące rekordów, dla których wartość teoretyczna z modelu jest względnie bliska wartości teoretycznej dla rekordu, którego dotyczy imputacja. Takie postępowanie ma na celu dodatkowe zabezpieczenie przed skutkami ewentualnego niedopasowania modelu lub heteroskedastyczności reszt. Pozwala przez to poszerzyć zakres stosowania stochatycznej imputacji regresyjnej.

Omówienie konkretnych modeli stosowanych przy imputacji poszczególnych zmiennych, a więc zestawu zmiennych grupujących w przypadku imputacji metodą *hot-deck* oraz zmiennych objaśniających modelu w przypadku imputacji regresyjnej, wykraczałoby poza zakres artykułu. Przedstawione zostaną jedynie ogólne zasady doboru tych zmiennych (określanych wspólnie mianem zmiennych pomocniczych).

Zmienne pomocnicze dobrano tak, by odzwierciedlały zależności, jakie zgodnie z logiką i wiedzą merytoryczną o badanych zjawiskach powinny występować w zbiorze danych, uwzględniając dostępność potencjalnych zmiennych objaśniających na formularzu. Zależności te przetestowano na zbiorze danych zarejestrowanych (uzyskanych od respondentów, nieimputowanych) i w większości przypadków okazały się istotne.

Badanie powtarzane jest przez 4 lata w tych samych gospodarstwach domowych z wymianą jednej czwartej próby co roku. Wskutek tego dla części obserwacji dokonywanych w danym roku dostępne są dane z lat poprzednich (panelowe). Dla części badanych jednostek (gospodarstw i osób) takich danych nie posiadamy — dotyczy to wymienianej w danym roku części próby, osób nowych w gospodarstwie oraz tych, których w poprzednim roku nie udało się zbadać. Rozróżnienie na osoby badane pierwszy raz i badane po raz kolejny jest ważne z punktu widzenia stosowanego modelu, a więc zestawu zmiennych pomocniczych.

W przypadku jednostek, dla których dostępna jest informacja z poprzedniego roku i dotyczy ona tego samego typu dochodu, ta właśnie informacja (wartość dochodu) traktowana jest jako najważniejsza zmienna pomocnicza; informacje bieżące (dotyczące danego roku) mogą mieć charakter uzupełniający (dodatkowe zmienne pomocnicze). Natomiast w sytuacji gdy nie dysponujemy taką informacją z poprzedniego roku, stosowany jest model wykorzystujący informacje bieżące, w możliwie najlepszy sposób, mogące objaśnić tworzenie danego typu dochodu lub mające wpływ na jego wartość (np. w przypadku dochodu z pracy najemnej: staż pracy, zawód, rodzaj działalności miejsca pracy, także lokalizacja terytorialna — województwo czy klasa miejscowości).

Dla zobrazowania skali imputacji pozycyjnej w badaniu i jej znaczenia dla otrzymywanych wyników na wykr. 1 pokazano statystykę imputacji dotyczącą kilku kluczowych zmiennych otrzymywanych z badania (w 2012 r.). Dochód całkowity oraz do dyspozycji i wartość świadczeń rodziny wyznaczane były dla gospodarstwa domowego, pozostałe uwzględnione zmienne dotyczyły osób.

Na wykresie pokazano odsetek rekordów, których dotyczy imputacja dla każdej ze zmiennych, z rozróżnieniem na imputację całkowitą i częściową. Imputacja całkowita ma miejsce, gdy cała wartość danej zmiennej (dla osoby lub gospodarstwa) pochodziła z imputacji. Poszczególne zmienne dochodowe powstają często jako suma wartości uzyskanych w odpowiedzi na kilka pytań. Możliwa jest zatem sytuacja, gdy część wartości zmiennej pochodzi z odpowiedzi respondenta, a część z imputacji — i jest to przypadek imputacji częściowej. Imputacja częściowa występuje bardzo często dla zmiennych „wynikowych", które składają się z wielu „cząstkowych" komponentów. Przykładem takiej zmiennej są dochód całkowity (będący sumą dochodów w gospodarstwie oraz dochodów wszystkich jego członków) oraz do dyspozycji. Odsetki mają za podstawę liczbę rekordów, których dotyczy dany typ dochodu, a więc np. w przypadku dochodów z pracy najemnej — liczbę pracujących najemnie.

Ze względu na występowanie imputacji częściowych trudne jest określenie skali imputacji tylko na podstawie odsetka rekordów, których dotyczy imputacja. Imputa-

cja częściowa może bowiem oznaczać zarówno to, że z imputacji pochodzi niewielka część wartości danej zmiennej, jak i to, że jest ona imputowana prawie w całości. Dlatego lepszą miarą faktycznej skali imputacji jest udział wartości imputowanych w uogólnieniu[8] dla danej zmiennej, który również został pokazany na wykr. 1.

Wykr. 2 prezentuje dane mogące zobrazować wpływ imputacji na wyniki uogólnień. Wpływ ten pokazano na przykładzie wyliczonej wartości przeciętnego dochodu do dyspozycji w gospodarstwie domowym w przeliczeniu na osobę, w podziale według głównego źródła utrzymania gospodarstwa. Na wykr. 2 pokazano wyniki uogólnienia, które można by uzyskać w trzech wariantach:

- z pominięciem wartości stanowiących braki pozycyjne przy użyciu wag oryginalnych,

[8] W przypadku gdy — w rekordzie danych — część komponentów zmiennej „wynikowej" pochodzi z imputacji, część została zarejestrowana (informacje otrzymane od respondenta), możemy policzyć oddzielnie sumę komponentów imputowanych i zarejestrowanych wchodzących w skład zmiennej. Wartość zmiennej jest oczywiście sumą tych sum. Odnosząc sumę komponentów imputowanych do wartości zmiennej otrzymujemy „udział", który obrazuje skalę imputacji danej zmiennej dla danego rekordu. Jeżeli sumę komponentów imputowanych uogólnimy na całą populację i odniesiemy do uogólnienia wartości globalnej zmiennej „wynikowej", otrzymamy udział wartości imputowanych w uogólnieniu będący miarą skali imputacji zmiennej „wynikowej" dla całej populacji. Oczywiście miarę tę możemy obliczyć także dla zmiennych złożonych z pojedynczych komponentów. W przypadku takich zmiennych udział dla pojedynczego rekordu przyjmuje zawsze wartość 0 lub 1, natomiast dla populacji — wartość z przedziału od 0 do 1, zależną od odsetka rekordów imputowanych oraz zróżnicowania wag uogólniających przypisanych rekordom imputowanym i nieimputowanym.

- z pominięciem wartości stanowiących braki pozycyjne oraz z korektą wag uogólniających ze względu na te braki,
- na podstawie wszystkich obserwacji w zbiorze, w tym obserwacji stanowiących braki pozycyjne, po zaimputowaniu tych braków.

Oczywiście wyniki w pierwszym wariancie są obciążone, gdyż pominięta zostaje część wartości globalnej zmiennej — ta, która odpowiada uogólnieniu na podstawie obserwacji stanowiących braki. Drugi wariant można natomiast potraktować jako swego rodzaju eksperyment pokazujący podejście alternatywne w stosunku do imputacji — gdyby zastosować korektę wag zamiast imputacji braków pozycyjnych. Wagi zostały skorygowane informacją o brakach odpowiedzi w zakresie rozważanej zmiennej w warstwach według województw i klas miejscowości. Tak skorygowane wagi nie mają zastosowania w przypadku uogólnień dla danych zaimputowanych, tj. w trzecim wariancie, gdyż wtedy uogólnienie odbywa się na podstawie całego zbioru, a „brakująca" część wartości globalnej uzupełniona jest przez imputację — stosowane są zatem wagi oryginalne. Trzeci wariant odpowiada faktycznie przyjętemu sposobowi uogólniania przy przeprowadzeniu imputacji braków pozycyjnych i uogólnieniu na podstawie zbioru zaimputowanego.

Wykr. 1 pokazuje znaczącą skalę imputacji w badaniu. Dotyczy ona dużej części zbioru danych, co pokazują odsetki rekordów imputowanych dla zmiennych będących sumą wielu komponentów, przede wszystkim dochodu całkowitego gospodarstwa. W przypadku ponad 70% gospodarstw imputacji podlegał przynajmniej jeden z komponentów dochodu.

Skala imputacji nie jest jednak tak duża, jakby wskazywał przywołany odsetek, gdyż większość przypadków imputacji dla zmiennych łączących wiele komponentów to imputacje częściowe, będące zwykle skutkiem braków danych dotyczących pojedynczych składników. Potwierdzają to liczby opisujące udział wartości imputowanych w tworzeniu wartości globalnej. Tak mierzona skala występowania imputacji w przypadku zmiennych „wynikowych" nie różni się znacząco w stosunku do zmiennych będących pojedynczymi komponentami.

Spośród pojedynczych komponentów dochodu imputacja odgrywa największą rolę w przypadku dochodu z pracy na rachunek własny, szczególnie poza rolnictwem. Jest to składnik dochodu, o którym uzyskanie informacji od respondentów jest najtrudniejsze. Konsekwencją tego jest fakt, że w przypadku gospodarstw utrzymujących się z tego źródła najsilniejszy jest również wpływ imputacji na wartości uogólnień.

Analiza wpływu zastosowania imputacji na wartości uogólnień (wykr. 2) potwierdza konieczność przeprowadzenia imputacji braków pozycyjnych w przypadku zastosowanej metody uogólnienia (tzn. uogólnienia za pomocą wag wspólnych dla całego zbioru, bez wykluczenia rekordów dotkniętych brakami pozycyjnymi).

Porównanie uogólnień uzyskanych z użyciem imputacji oraz korekty wag nie wskazuje na występowanie znaczących różnic. Nieco bardziej widoczny efekt dotyczy jedynie — jak wcześniej wspomniano — dochodu gospodarstw utrzymujących się z pracy na własny rachunek. Zbieżność uogólnień pozwala traktować obydwie metody jako w dużym stopniu równoważne ze względu na efekty estymacji, stanowiące dla siebie alternatywę. Należy jednak zauważyć, że aby korekta wag była równie skuteczna, musiałaby być wykonywana odrębnie dla każdej zmiennej, a nawet dla konkretnych zestawów zmiennych w zakresie analiz wielowymiarowych. Jednak to znacząco utrudniałoby analizę i zmniejszało praktyczną użyteczność zbioru.

BADANIE SPÓJNOŚCI SPOŁECZNEJ

Badanie to odnosi się do jakości i warunków życia pojmowanych znacznie szerzej niż warunki materialne, a w odniesieniu do warunków materialnych szerzej niż tylko zagadnienia związane z dochodem. Sposób badania dochodu jest znacznie mniej szczegółowy niż w EU-SILC, natomiast bardzo wiele uwagi poświęca się pozadochodowym i pozamaterialnym aspektom jakości życia. Badane są one wielowymiarowo, również z punktu widzenia zależności i powiązań z dochodem i sytuacją materialną. Chociażby z tego powodu informacja o dochodzie gospodarstwa jest także tutaj informacją bardzo ważną[9].

[9] Informacje na temat badania i jego metodologii znajdują się w publikacji *Jakość…* (2013). Informacje na temat imputacji dochodów tamże, s. 296—298.

W formularzu występuje kilka pytań dotyczących dochodu, jednakże ich celem jest wyprowadzenie jednej zmiennej reprezentującej przeciętny miesięczny dochód ogółem gospodarstwa domowego za rok poprzedzający badanie. Większa liczba pytań służy uzyskaniu informacji pomocniczych, które mogą być wykorzystane do imputacji głównej zmiennej dochodowej w przypadku braku odpowiedzi. Fakt, że wyprowadzana jest jedna główna zmienna dochodowa, nie wyklucza stosowania w dalszych analizach informacji, np. o dochodzie na osobę czy dochodzie ekwiwalentnym, ale wszystkie one wyprowadzane są z tej zmiennej.

Podobnie jak w przypadku EU-SILC, dla pytań dotyczących dochodu dopuszcza się nieudzielenie odpowiedzi przez respondenta, co nie skutkuje dyskwalifikacją wywiadu. Powstałe w ten sposób braki pozycyjne są imputowane.

Celem imputacji jest uzyskanie kompletnej informacji dotyczącej głównej zmiennej dochodowej, czyli średniomiesięcznego dochodu gospodarstwa domowego za rok poprzedzający badanie. Zmienna ta powinna być imputowana w taki sposób, aby umożliwiać:

- elastyczną, wielowymiarową analizę, w której informacja o dochodzie jednostkowym jest łączona z innymi informacjami z badania;
- wyliczenia mierników, których konstrukcja wymaga informacji o różnych parametrach rozkładu dochodów, w tym również miar pozycyjnych i nieliniowych.

W badaniu występują dwa główne pytania dotyczące łącznego dochodu netto gospodarstwa domowego. Są to pytania o:

- dochód za poprzedni rok (jako pytanie o dochód roczny bądź średni dochód miesięczny — do wyboru respondenta) — będący bezpośrednim źródłem zainteresowania;
- aktualny dochód miesięczny — którego nie można przeliczyć na interesującą nas wartość, ale będący bardzo dobrym źródłem informacji pomocniczej dla jej imputacji.

Przy każdym z pytań o dochód respondent dostaje możliwość:

- udzielenia odpowiedzi wprost, podając wartość dochodu;
- wskazania przedziału, w którym mieści się jego dochód, na przygotowanej liście przedziałów dochodowych, jeżeli nie zgadza się podać wartości dochodu.

Tak więc dochód może być określony jednoznacznie przez respondenta co do wartości lub też jedynie co do przedziału, w którym mieści się jego wartość. Istnieje również trzecia ewentualność, kiedy respondent odmawia nawet wskazania przedziału dochodowego. Tylko w takim przypadku możemy mówić bez zastrzeżeń o braku odpowiedzi i mamy do czynienia z informacją tworzoną w całości za pomocą imputacji statystycznej.

Przyjęta metoda badania, dopuszczająca wskazanie przedziału dochodowego, ma istotny wpływ na ostateczny kształt przyjętej procedury i metody imputacji danych. Wobec opisanych uwarunkowań, dokonując imputacji, dla dużej części

rekordów mamy do dyspozycji stosunkowo precyzyjne pomocnicze informacje dotyczące dochodu, które należy w pierwszej kolejności wykorzystać. Są to:

- informacja o przedziale dochodowym — jeśli została podana przez respondentów, imputacja ulega zawężeniu do granic przedziału;
- informacja o dochodzie bieżącym — jest to zmienna pomocnicza silnie skorelowana ze zmienną badaną, którą — jeżeli została podana przez respondenta — należy w pierwszej kolejności wykorzystać do objaśnienia i imputacji głównej zmiennej dochodowej.

Wspomniana zależność między obydwiema zmiennymi dochodowymi stanowi podstawę głównego modelu imputacyjnego. Alternatywny model, opisujący zależność między dochodem a cechami społeczno-ekonomicznymi gospodarstwa i jego głowy, staje się źródłem, na podstawie którego tworzona jest wartość imputacyjna, dopiero wtedy, gdy nie dysponujemy żadną informacją o dochodzie.

Jako metodę imputacji zastosowano stochastyczną imputację regresyjną. Imputowana reszta jest generowana z rozkładu teoretycznego o odpowiednich parametrach. Metoda ma charakter stochastyczny, dzięki czemu imputacja w niewielkim stopniu zaburza rozkład imputowanej zmiennej. Uzyskane rozkłady są „podobne" do naturalnych; nie są tworzone sztucznie zbiory identycznych wartości imputacyjnych. Modele stosowane do imputacji mają postać potęgowo-wykładniczą, co oznacza, że stają się modelami liniowymi po transformacji logarytmicznej.

Zastosowana metoda pozwala objąć jednolitym podejściem metodologicznym sytuacje, gdy respondent podał granice przedziału dochodowego oraz gdy ich nie podał. Jest to ważna własność, która w kontekście rozważanego badania w dużym stopniu przesądziła o wyborze tej metody.

Imputowane reszty tworzone są za pomocą generatora liczb pseudolosowych. Stosowany jest rozkład normalny (dla modelu w postaci zlogarytmowanej) o wariancji odpowiadającej oszacowaniu wariancji składnika losowego modelu.

W sytuacji gdy nie określono przedziału, imputowane reszty generowane są z rozkładu bezwarunkowego. Jeśli natomiast znamy granice przedziału dochodowego, imputowana reszta jest generowana z rozkładu uciętego, tzn. przyjmuje wyłącznie wartości z założonego zakresu. Granica „ucięcia" rozkładu dla reszty losowej (zakres) ustalana jest tak, by zagwarantować, że uzyskana wartość imputacyjna będzie się mieściła w granicach zadeklarowanego przedziału dochodowego.

Ze względu na różną dostępność informacji dotyczącej zmiennych objaśniających (dostępność lub brak informacji o dochodzie bieżącym), dla każdego rekordu wymagającego imputacji stosowany jest jeden z dwóch alternatywnych modeli:

- jeżeli mamy informację na temat dochodu miesięcznego z miesiąca poprzedzającego badanie (dochodu bieżącego), używamy modelu, w którym zmiennymi objaśniającymi są dochód bieżący oraz opisowa informacja o zmianie dochodu w stosunku do poprzedniego roku (wzrost/spadek/w przybliżeniu bez zmian) — model A;

- jeżeli nie mamy informacji na temat dochodu bieżącego (brak danych), stosujemy model objaśniający dochód za pomocą obiektywnych czynników determinujących wysokość dochodu i możliwości gospodarstwa domowego w zakresie jego osiągania, takich jak: źródło utrzymania, rodzaj pracy, zawód i wykształcenie głowy gospodarstwa — model B.

Model A jest modelem o znacznie lepszym objaśnieniu i preferowanym w użyciu. Wykorzystuje przy tym informację o dochodzie przekazaną przez respondenta. Dlatego jest stosowany zawsze, gdy tylko jest to możliwe, biorąc pod uwagę posiadane informacje. Model B stosuje się tylko w przypadku braku możliwości zastosowania modelu A.

Wykr. 3 przedstawia statystykę dotyczącą imputacji w badaniu. Podobnie jak w przypadku EU-SILC, pokazano odsetek rekordów imputowanych oraz udział w uogólnieniu. Nie mamy tu przypadków imputacji częściowej, natomiast istotne jest rozróżnienie poszczególnych sytuacji ze względu na dostępność informacji o przedziale dochodowym. Wzięto pod uwagę nie tylko to, czy przedział został podany przez respondenta, ale również, czy zaznaczony przedział jest przedziałem „wewnętrznym" czy też skrajnym, tzn. pierwszym (najniższe dochody) lub ostatnim (najwyższe dochody). Z najmniejszym potencjalnym błędem imputacji mamy do czynienia w przypadku przedziałów „wewnętrznych", gdzie określona jest zarówno dolna, jak i górna granica możliwego dochodu. W przypadku przedziałów skrajnych jedna z tych granic pozostaje nieokreślona, natomiast jeśli respondent odmawia wskazania przedziału, nie znamy żadnej z nich, co oznacza największy zakres możliwego błędu.

W tablicy przedstawiono statystykę rekordów imputowanych, z tym że oprócz dostępności informacji o przedziale dochodowym zestawienie uwzględnia również rozróżnienie ze względu na zastosowany do imputacji model. Przed-

stawione liczby przypadków imputacji należy odnieść do liczby wszystkich gospodarstw zbadanych w badaniu (14884).

TABLICA. LICZBA IMPUTOWANYCH REKORDÓW WEDŁUG DOSTĘPNOŚCI INFORMACJI O PRZEDZIALE DOCHODOWYM ORAZ ZASTOSOWANEGO MODELU

Dostępność informacji o przedziale dochodowym	Ogółem	Zastosowany model	
		A — na podstawie informacji o dochodzie bieżącym	B — brak jakiejkolwiek informacji o dochodzie
Przedział „wewnętrzny”	2176	368	1808
Przedział skrajny	21	5	16
Brak informacji	1774	293	1481
Ogółem	3971	666	3305

Ź r ó d ł o: opracowanie własne.

Zastosowanie w kwestionariuszu przedziałów dochodowych znacząco podnosi jakość imputacji i otrzymywanych wyników, ponieważ zmniejsza i — co najważniejsze — precyzyjnie ogranicza zakres niepewności wartości imputacyjnej. Fakt, że większość przypadków imputacji jest realizowana w obrębie znanego przedziału dochodowego, minimalizuje negatywne konsekwencje występowania pozycyjnych braków danych. Czyni to uogólnienia dokonywane na podstawie zbioru zaimputowanego bardziej wiarygodnymi.

Udział rekordów imputowanych w tworzeniu wartości globalnej dochodu jest większy niż odsetek rekordów poddanych imputacji. Oznacza to, że wyniki wskazują, iż przeciętne dochody gospodarstw odmawiających odpowiedzi są wyższe niż w przypadku gospodarstw, które odpowiedzi udzieliły. Innymi słowy, przeciętna wartość imputacyjna jest wyższa od przeciętnej wartości dochodu podanej przez respondenta. Wiarygodność tego wniosku podnosi fakt, że zaobserwowana prawidłowość w dużym stopniu wynika z danych uzyskanych dla gospodarstw, które zadeklarowały przedział dochodowy, a więc poznajemy ją nie tylko na podstawie modelowania statystycznego, ale także rzeczywistych deklaracji respondenta. Można zatem stwierdzić, że zastosowana imputacja skutecznie koryguje obciążenie związane z mniejszą skłonnością do udzielania odpowiedzi wśród osób o wyższych dochodach. Rezygnacja z imputacji mogłaby prowadzić do obciążenia wyników (gdyby nie użyto innej, równie skutecznej, metody korekty).

W przypadku imputowanych dochodów należących do przedziałów skrajnych, ich udział w uogólnieniu wartości globalnej jest znacznie wyższy niż w liczbie gospodarstw, co jest zjawiskiem naturalnym ze względu na ostatni przedział obejmujący najwyższe dochody. Imputacja w tym przedziale ze względu na brak górnego ograniczenia wiąże się z nieco większym ryzykiem niż w przypadku pozostałych przedziałów, jednak i tak jest to zawsze sytuacja lepsza, niż gdyby w przypadku osób o najwyższych dochodach nie został podany nawet przedział. Poza tym grupa ta, mimo opisanego efektu, ma stosunkowo niewielki wpływ na całkowitą wartość uogólnienia ze względu na niewielką liczebność.

Formalnie liczba rekordów imputowanych wynosiła 3971 przy próbie liczącej 14884 rekordy. Biorąc jednak pod uwagę, że dla części z nich znany był przedział dochodowy, dla części dochód bieżący, liczba rekordów imputowanych bez żadnej informacji o wartości dochodu wynosiła jedynie 1481 (10% próby). Tylko dla tych 10% gospodarstw należy przyjąć, że przypisana im wartość dochodu w całości pochodzi z imputacji statystycznej.

Podsumowanie

Stosowanie statystycznej imputacji pozycyjnych braków danych nie jest jak dotąd standardem obowiązującym w większości regularnych badań statystycznych. Jednakże coraz częściej potrzeba takiego postępowania pojawia się w poszczególnych badaniach i stopniowo wymusza rozszerzanie zakresu stosowania opisywanych rozwiązań.

Niezależnie od starań ankieterów o uzyskanie odpowiedzi od respondenta, problem braku odpowiedzi dotyczący niektórych badań oraz pewnych typów pytań czasem występujących w wielu badaniach pojawia się nieuchronnie, nawet wtedy, gdy taka możliwość nie jest z góry przewidywana przy projektowaniu badania. Przykładem takiego pytania, w przypadku którego obawy o wystąpienie braków odpowiedzi zawsze stają się zasadne, jest właśnie pytanie o dochód.

Opisane procedury imputacji dochodu w omówionych badaniach mają zarówno cechy wspólne, jak i charakterystyczne elementy wynikające ze specyfiki danego badania. Dzięki temu możliwe było z jednej strony spójne omówienie problemu i objęcie tematyki artykułu jednym wstępem teoretycznym, z drugiej zaś przedstawienie na przykładzie tych dwóch badań szerokiego spektrum zagadnień związanych z imputacją.

Niewątpliwie ciekawą i wartą powielania w innych badaniach praktyką związaną z badaniem spójności społecznej jest stosowanie pytania o przedział dochodowy. Patrząc szerzej, można tę praktykę określić jako zadawanie pytań o zjawiska bardzo ściśle powiązane ze zmienną imputowaną, czasami po prostu opisującą ją w sposób mniej precyzyjny lub bardziej jakościowy niż ilościowy. Realizacja badania pokazała, że w bardzo dużej liczbie przypadków respondenci są gotowi odpowiadać na takie pytania, nawet wtedy, gdy precyzyjnej wartości opisującej zjawisko (tu dochodu) nie są skłonni podać. Oczywiście podejście takie ma swoje ograniczenia związane przede wszystkim z obciążeniem respondenta dodatkowymi pytaniami — zastosowanie go np. w przypadku tak obszernego (z punktu widzenia informacji o dochodzie) badania jak EU-SILC byłoby trudne lub wręcz niemożliwe. Stawia to także dodatkowe wyzwania przed autorami badań, dotyczące wyznaczenia właściwych przedziałów (określenie ich liczby i granic), co ma potem istotny wpływ na ostateczne wyniki. Zastosowanie klasyfikacji zbyt szczegółowej czyni odpowiedź trudniejszą dla respondenta i zwiększa ryzyko niepodania przedziału w ogóle. Mniej szczegółowa klasyfikacja oznacza jednak mniejszą wartość informacji uzyskanej z deklaracji przedziału i większy zakres niepewności co do prawdziwej wartości.

Alternatywę dla imputacji w przypadku badań reprezentacyjnych stanowi korekta wag uogólniających ze względu na brak danych. W przypadku braku odpowiedzi ze strony całych jednostek statystycznych jest zwykle lepsza niż imputacja. Również w przypadku braków pozycyjnych, za pomocą odpowiedniej korekty wag, osiągnąć można efekty zbliżone, jak przy użyciu imputacji, co pokazały nawet bardzo uproszczone wyliczenia dotyczące danych z EU-SILC. Byłoby to jednak rozwiązanie bardzo niepraktyczne, gdyż wymagałoby stosowania wag tworzonych odrębnie dla poszczególnych zmiennych oraz ich kombinacji. Z kolei zastosowanie wag wspólnych dla całego zbioru może prowadzić do obciążenia oszacowań. Taki obraz sytuacji dostarcza argumentów przemawiających na korzyść imputacji pozycyjnej i uzasadnia jej racjonalność.

mgr Tomasz Piasecki — *Urząd Statystyczny w Łodzi*

LITERATURA

Balicki A. (2004), *Metody imputacji brakujących danych w badaniach statystycznych*, „Wiadomości Statystyczne", nr 9

Dochody i warunki życia ludności Polski (2013), GUS

Jakość życia, kapitał społeczny, ubóstwo i wykluczenie społeczne w Polsce (2013), GUS, Urząd Statystyczny w Łodzi

Kalton G., Kasprzyk D. (1982), *Imputing for Missing Survey Responses*, Proceedings of the Survey Research Methods Section, American Statistical Association

Longford N. T. (2005), *Missing Data and Small Area Estimation*, Springer Science + Business Media, Inc.

Rubin D. B. (1987), *Multiple Imputation for Nonresponse in Surveys*, Wiley & Sons, New York

SUMMARY

The study presents methodological solutions used in the European surveys of income and living conditions as well as social cohesion. The paper contains the results and conclusions of their implementation as an example of a different approach to estimating household income instead of current practice in statistical surveys.

РЕЗЮМЕ

Статья содержит методологические решения использованные в Европейском обследовании доходов и условий жизни, а также в Обследовании социальной сплоченности. Статья представляет также результаты и выводы связанные с их проведением в качестве примера другого подхода к оценке доходов домашних хозяйств по сравнению с практикой обследований существующей до сих пор в официальной статистике.