Konferencja naukowa
„STATYSTYKA – WIEDZA - ROZWÓJ”
Łódź 17 – 18 października 2013 r.

Imputacja dochodów w badaniach statystyki publicznej dotyczących gospodarstw domowych

Tomasz Piasecki

Urząd Statystyczny w Łodzi
Ośrodek Statystyki Matematycznej

Występowanie braków odpowiedzi w badaniach statystycznych jest istotnym problemem związanym z ich realizacją, a sposoby radzenia sobie z tym zjawiskiem i ograniczania jego negatywnych konsekwencji dla wyników badań stanowią ważne wyzwanie metodologiczne.

W stosunku do pozycyjnych braków odpowiedzi, tj. takich, gdy jednostka statystyczna bierze udział w badaniu, ale z jakichś powodów nie udziela odpowiedzi na niektóre pytania, możliwe jest przyjęcie przez badacza różnych strategii. Przyjęcie restrykcyjnych wymogów co do kompletności zbieranych wywiadów (ograniczenie możliwości nie udzielenia odpowiedzi przez respondenta na poszczególne pytania) pozwala uzyskać spójne i kompletne zbiory danych wynikowych, jednak może powodować wzrost częstości występowania jednostkowych braków odpowiedzi, tj. takich, gdy jednostka statystyczna w ogóle nie bierze udziału w badaniu. Z drugiej strony, dopuszczenie odmów odpowiedzi na poszczególne pytania pozwala ograniczyć występowanie odmów udzielenia wywiadu w ogóle, pogarsza jednak kompletność i spójność wewnętrzną oraz może zmniejszać użyteczność uzyskanego zbioru danych wynikowych. W takim przypadku w zbiorze danych wynikowych pojawiają się pozycyjne braki danych, wymagające przyjęcia jakiejś procedury postępowania z nimi. Rozwiązaniem, przywracającym w dużym stopniu zbiorowi danych niekompletnych użyteczność i funkcjonalność zbliżoną jak w w przypadku danych kompletnych (choć z zastrzeżeniami co do wnioskowania, o których należy pamiętać) może być imputacja.

Referat omawia dwa przykłady zastosowania procedur imputacyjnych w badaniach statystyki publicznej – w odniesieniu do *Europejskiego Badania Dochodów i Warunków Życia* (*EU-SILC*) oraz *Badania Spójności Społecznej*. We wspomnianych badaniach imputacja stosowana jest w przypadku braków pozycyjnych dotyczących dochodów i ich składowych (w przypadku EU-SILC w bardzo rozbudowanym podziale).

Podstawowe pojęcia

Imputacja polega na zastąpieniu braków danych występujących w zbiorze danych wynikowych badania sztucznie utworzonymi wartościami imputacyjnymi.

Wartości imputacyjne są oszacowaniami brakujących wartości prawdziwych, najbardziej właściwymi ze względu na przyjęte kryteria, założenia, zastosowane metody.

Imputacja:

- jest prowadzona na poziomie danych jednostkowych, prowadzi do uzyskania pozornie kompletnego zbioru jednostkowego,
- umożliwia poprawne wnioskowanie na poziomie agregatów, nie gwarantuje możliwości rzetelnego wnioskowania o jednostkach objętych imputacją w zakresie objętych nią cech.

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

- **Imputacja pozycyjna**
  Imputowane są brakujące informacje o jednostkach statystycznych, które wzięły udział w badaniu, ale nie udzieliły odpowiedzi na pojedyncze pytania (imputowane pojedyncze pola w danym rekordzie).
- **Imputacja brakujących rekordów**
  Imputacja pełnej informacji o jednostkach statystycznych, które w ogóle nie wzięły udziału w badaniu (pewne rekordy danych są imputowane w całości).

Obywa 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*).

Tam, gdzie w badaniach społecznych statystyki publicznej stosuje sie imputację (a w szczególności w przypadku omawianych dwóch badań), jest to przede wszystkim **imputacja pozycyjna**. Pewne elementy imputacji brakujących rekordów występują w badaniu EU-SILC, gdzie imputuje się brakujące formularze indywidualne (osobowe) w przypadku pojedynczych osób należących do gospodarstw domowych, które wzięły udział w badaniu (nie jest to więc pełny brak jednostkowy – mamy wywiad gospodarstwa i część wywiadów indywidualnych).

Podstawową metodą **niwelowania skutków występowania braków jednostkowych** w omawianych badaniach jest odpowiednia **korekta wag** (w tym kalibracja).

Strategie dotyczące wymogów co do kompletności danych w badaniu

Można wyróżnić dwie podstawowe strategie, możliwe do zastosowania w badaniu:

- **Strategia restrykcyjna**: wymaganie udzielenia kompletnych odpowiedzi na wszystkie pytania wywiadu (wymóg pełnej kompletności), jako warunek akceptacji wywiadu i przekazania danych do dalszego przetwarzania (eliminuje występowanie pozycyjnych braków danych),
- **Strategia mniej restrykcyjna**: dopuszczenie odmowy odpowiedzi na niektóre pytania wywiadu, nie skutkującej dyskwalifikacją całego wywiadu, tzn. dopuszczenie wystąpienia braków pozycyjnych.

Porównanie strategii

Strategia	Wymóg pełnej kompletności	Dopuszczenie braków pozycyjnych
Zalety	• spójny i kompletny zbiór wynikowy • lepsza jakość danych wynikowych, jeżeli uda sie uzyskać rzetelne odpowiedzi na problematyczne pytania	• zmniejszenie częstości występowania braków jednostkowych • większe prawdopodobieństwo, iż uzyskane odpowiedzi na pytania problematyczne są rzeczywiście rzetelne
Wady	• możliwy wzrost częstości występowania braków jednostkowych • gorsza jakość danych wynikowych, jeżeli wymaganie pełnej kompletności wymusi nierzetelne odpowiedzi	• niekompletność zbioru danych wynikowych • trudności analityczne z przetwarzaniem uzyskanych danych • Mniejsza elastyczność wykorzystania uzyskanych danych
Konsekwencje dot. imputacji	• brak imputacji pozycyjnej	• wskazana imputacja pozycyjna

W **badaniach statystyki publicznej** przyjmuje sie zazwyczaj:

- **co do zasady** – **strategię** bardziej **restrykcyjną**, zakładającą wymóg kompletności wywiadu,
- dla niektórych badań, **w szczególnych przypadkach** – pytań szczególnie drażliwych, trudnych, na które respondenci odpowiadają z dużą niechęcią, lub wręcz trudno byłoby naciskać na nich by udzielili odpowiedzi – **strategię mniej restrykcyjną**, dopuszczającą odmowę odpowiedzi na takie pytanie.

Strategia dopuszczająca brak odpowiedzi, bez dyskwalifikacji całego wywiadu, stosowana jest często w przypadku pytań dotyczących dochodu. Z taką sytuacja mamy takze do czynienia w omawianych badaniach.

Dopuszczenie braków pozycyjnych nie oznacza zaniechania starań o uzyskanie odpowiedzi na dane pytanie przez ankietera, oznaczenie braku danych traktowane jest jako „ostateczność".

Metody imputacji danych brakujących

W omawianych badaniach do imputacji danych brakujących stosowana jest imputacja pojedyncza (jednokrotna). Nie są stosowane procedury imputacji wielokrotnej. Przedstawiona dalej klasyfikacja i krótki opis dotyczą metod, które są stosowane do imputacji danych w przypadku omawianych badań.

Możemy wyróżnić dwa podstawowe **rodzaje imputacji**:

- **Imputacja dedukcyjna**
  Bazuje na zależnościach między zmiennymi i regułach (redagowanie danych). Ma charakter deterministyczny. Wartość imputacyjna wyznaczana jest bezpośrednio w oparciu o te zależności.
- **Imputacja statystyczna**
  Wykorzystuje do imputacji danych brakujących pozostałą część zbioru danych, tj. dane kompletne („nie brakujące"), przekazane przez respondentów. Opiera się na określonym modelu. Pozwala na nieobciążoną estymację parametrów populacji gdy modele (założenia) są prawidłowe.

Metody **imputacji statystycznej** możemy podzielić na:

- **Metody deterministyczne**

  Dla danego zbioru danych, powtarzając proces imputacji, otrzymamy zawsze te same wartości imputacyjne. Metody te:

  - charakteryzują się większą precyzją (nie wprowadzają dodatkowego źródła błędu losowego), jednakże:
  - zniekształcają rozkłady zmiennych (w tym zaniżają miary rozrzutu i błędu).

- **Metody stochastyczne**

  Proces tworzenia wartości imputacyjnej zawiera element losowy – dla danego zbioru danych możemy uzyskać różne zbiory wartości zaimputowanych. Metody te:

  - generują pewien dodatkowy błąd wynikający z procesu imputacji (mniejsza precyzja), jednakże:
  - lepiej zachowują rozkłady zmiennych

Ponadto, wyróżnić można metody „oparte na dawcach" (zarówno deterministyczne, jak i stochastyczne), tzn. takie, w których wartość imputacyjna przenoszona jest z innego rekordu, jako metody, które nie tworzą sztucznych wartości zmiennych.

Stosowane **metody deterministyczne** imputacji danych brakujących:

- **Imputacja średnią**
  Za wartość imputacyjną przyjmowana jest średnia z obserwacji prawidłowych („niebrakujących"). Zwykle stosowana jest imputacja średnią w klasach (grupach) imputacyjnych, wyróżnionych ze względu na określone kryteria. Mamy więc wtedy do czynienia z imputacją średnią warunkową. Dobór zmiennych grupujących stanowi określenie „modelu" takiej imputacji.
  Kryteria mogą mieć postać hierarchiczna.

- **Imputacja regresyjna** (deterministyczna)
  Na podstawie modelu regresyjnego objaśniającego zmienną, dla której występuje brak danych za pomocą zmiennych kompletnych (lub kompletnych w takim zakresie, jaki jest wystarczający dla dopasowania modelu i dokonania imputacji dla konkretnego rekordu). Za wartość imputacyjną przyjmowana jest wartość teoretyczna z modelu.

Stosowane **metody stochastyczne** imputacji danych brakujących:

- **Metoda hot-deck**
  Imputacja danymi innego rekordu (tzw. dawcy), wylosowanego spośród rekordów kompletnych. Zwykle wybór dawcy dokonywany jest spośród rekordów należących do tej samej klasy (grupy) imputacyjnej, tj spośród rekordów spełniających określone kryteria podobieństwa. Dobór tych kryteriów stanowi określenie „modelu" takiej imputacji.
  Podobnie jak w przypadku imputacji średnią, kryteria mogą mieć charakter hierarchiczny, co jest stosowane w badaniu 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 przybliżenia.

Określenie procedury imputacji dla każdej imputowanej zmiennej wymaga określenia metody, modelu, doboru zmiennych grupujących, a także określenia innych specyficznych elementów związanych z poszczególnymi metodami. Zagadnienia te przedstawione są odrębnie dla każdego z omawianych badań, ze względu na powiązanie z ich specyfiką.

Europejskie Badanie Dochodów i Warunków Życia (EU-SILC)

Podstawowe informacje o badaniu

- Badanie regularne, prowadzone z **częstotliwością raz do roku**.
- Dotyczy warunków życia ludności oraz dochodów.
- Badanie obejmuje kilkaset zmiennych na poziomie osoby i gospodarstwa domowego.
- Bardzo rozbudowana informacja o dochodach, obejmuje dużą liczbę zmiennych reprezentujących różne kategorie / typy dochodów (np. dochód z pracy najemnej, dochód z pracy na rachunek własny poza rolnictwem / w rolnictwie, wiele różnych typów świadczeń).
- **Efektem badania** jest **zbiór danych jednostkowych przekazywany do Eurostatu**, spełniający określone wymogi co do jego użyteczności analitycznej, posiadający ściśle określoną strukturę
- Zbiór danych jednostkowych musi m. in. **umożliwiać wyznaczanie nieliniowych wskaźników nierównomierności dochodów**.
- Pozycyjne braki danych podlegające imputacji dotyczą poszczególnych składowych dochodu gospodarstwa domowego, występujących na poziomie osobowym i poziomie gospodarstwa. Dla zmiennych tych dopuszczalne jest nieudzielenie odpowiedzi przez respondenta, a powstałe braki danych są odpowiednio oznaczane.
- Imputacji podlegają także dochody dotyczące osób, które nie udzieliły wywiadu indywidualnego, należących do gospodarstw, które wywiadu udzieliły (imputacja brakujących formularzy indywidualnych w części dotyczącej dochodów)
- Imputowane zmienne dochodowe podlegają agregacjom na kilku poziomach szczegółowości.

Przed imputacją danych w badaniu stawia się następujące **cele**:

- **Uzyskanie kompletnego** (pozbawionego braków) **zbioru danych** zawierającego tzw. zmienne obowiązkowe (porównywalne na poziomie UE; zbiór przekazywany do Eurostatu)
- **Zachowanie rozkładu zmiennych** imputowanych
- **Zapewnienie możliwości obliczania** wskaźników, w tym **wskaźników zmienności i nierównomierności dochodów**, na podstawie zbioru po imputacji
- Zapewnienie, na potrzeby statystyki krajowej, możliwości uogólnień bardziej szczegółowych, niż poziom zmiennych obowiązkowych, zgodnych z uogólnieniami na podstawie zbioru UE

Algorytm imputacji danych w badaniu został opracowany przez pracowników Ośrodka Statystyki Matematycznej Urzędu Statystycznego w Łodzi; Ośrodek odpowiada też za coroczna realizację procesu imputacji oraz doskonalenie, rozwijanie i aktualizowanie algorytmu.

Szczegółowa postać algorytmu oraz główne jego założenia są wynikiem celów, jakie postawiono przed procesem imputacji. Konstruując **algorytm imputacji danych w badaniu EU-SILC** przyjęto następujące **główne zasady**:

- Preferowane są metody stochastyczne.
  Imputacja najważniejszej składowej każdej zmiennej dochodowej (zwykle dochód netto) realizowana 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 zmiennych finalnych.

- Poszczególne zmienne imputowane oddzielnie, za wyjątkiem zmiennych ściśle powiązanych ze sobą oraz imputacji brakujących wywiadów indywidualnych.
- Z zasady imputowane są dochody miesięczne.
- Możliwe stosowanie kilku metod (modeli) dla jednej zmiennej dla różnych podzbiorów rekordów, ze względu na:
    - różną dostępność informacji o zmiennych pomocniczych dla poszczególnych rekordów,
    - dostępność informacji o dochodzie tego samego typu z poprzedniego roku lub jej brak dla danego rekordu (możliwość wykorzystania danych panelowych).

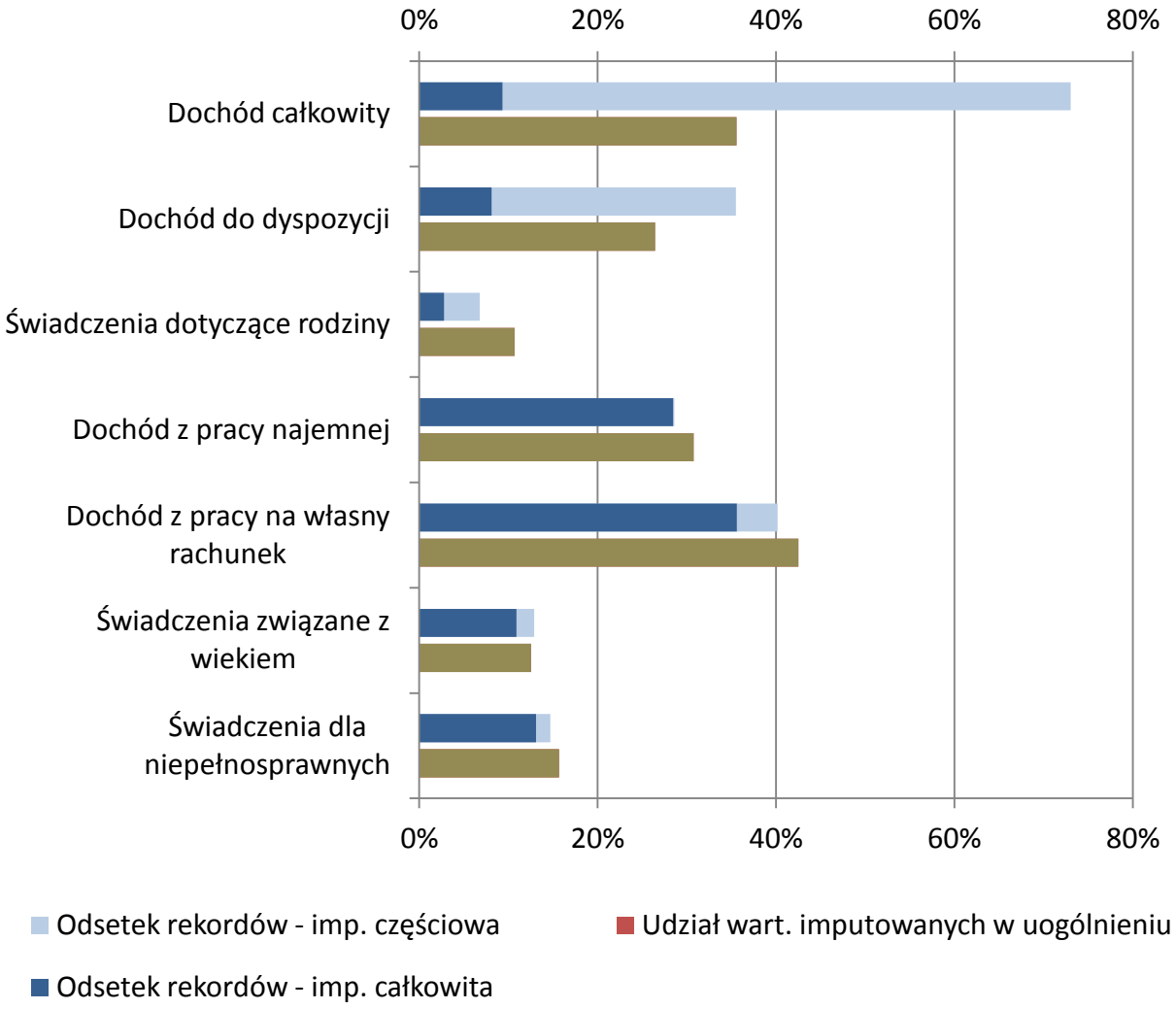
Stosowane metody imputacji:

- 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, jako metody stochastyczne stosowane do imputacji najważniejszych zmiennych. Warto zwrócić uwagę na **specyficzne** dla algorytmu imputacji w tym badaniu **rozwiązania szczegółowe**, przyjęte w przypadku tych metod:

- W przypadku **metody hot-deck**:
    - Stosowane są hierarchiczne kryteria wyodrębniania klas (grup) imputacyjnych. Zmienne pomocnicze (grupujące, kryterialne) dla poszczególnych zmiennych imputowanych zostały uporządkowane od najważniejszych do najmniej ważnych. W przypadku, gdy nie można znaleźć dawcy o odpowiadających wartościach wszystkich zmiennych pomocniczych, sekwencyjnie pomija się kolejne kryteria, poczynając od najmniej ważnych
    - Jako kryteria grupowania stosowane są zarówno zmienne jakościowe, jak i ilościowe. W przypadku zmiennych ilościowych grupowanie odbywa się według grup kwantylowych.
- W przypadku **stochastycznej imputacji regresyjnej**:
    - Imputowane reszty otrzymywane są poprzez losowy wybór ze zbioru rzeczywistych reszt modelu. Są one jednak losowane nie zbioru wszystkich reszt, lecz z ograniczonego, odpowiednio wyodrębnionego podzbioru. Ograniczony podzbiór stanowią reszty dotyczące rekordów, dla których wartość teoretyczna z modelu jest względnie bliska wartości teoretycznej dla rekordu, którego dotyczy imputacja.
      Takie postępowanie stanowi dodatkowe zabezpieczenie przed skutkami ewentualnego niedopasowania modelu lub heteroskedastyczności reszt.

Skala i częstość stosowania imputacji
dla wybranych zmiennych obowiązkowych
w badaniu EU-SILC (edycja 2012)



Odsetki mają za podstawę liczbę wszystkich rekordów w zbiorze wynikowym (badanych gospodarstw / osób), których dotyczy dany typ dochodu. Pokazują, jaka część rekordów, których dotyczy dany typ dochodu, została poddana imputacji całkowitej lub częściowej.

Imputacja całkowita oznacza sytuację, gdy cała wartość danej zmiennej (dla danego rekordu) pochodzi z imputacji. Z imputacją częściową mamy do czynienie, gdy część wartości zmiennej pochodzi z wywiadu od respondenta, część z imputacji. Może to mieć miejsce, gdy na zmienną składa sie kilka komponentów, a brak danych wystąpił tylko dla części z nich imputacja częściowa występuje przede wszystkim w przypadku zmiennych będących sumą wielu komponentów.

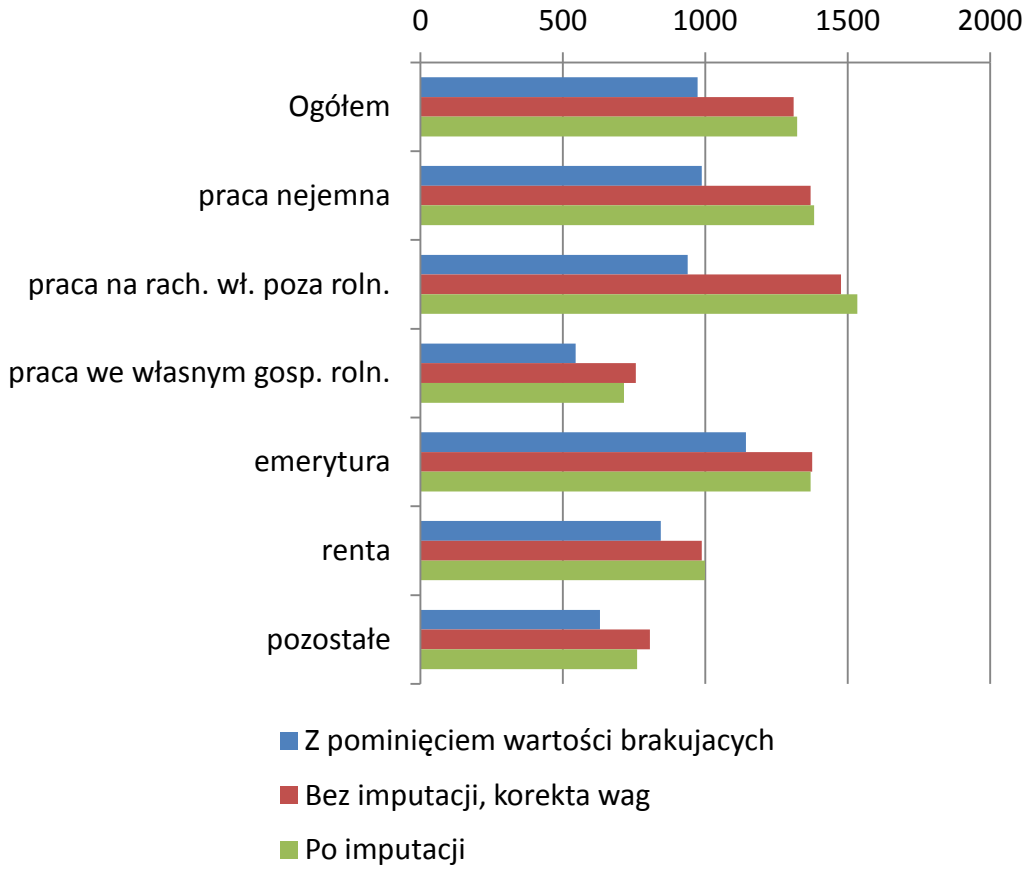
Udział wartości imputowanych w uogólnieniu oznacza udział wartości pochodzących z imputacji w tworzeniu wartości globalnej danej zmiennej. Ze względu na występowanie imputacji częściowych jest to miara najlepiej pokazująca skalę imputacji.

Wpływ imputacji na wartości uogólnień

Dochód do dyspozycji na osobę

według deklarowanego głównego źródła utrzymania gospodarstwa

EU-SILC – edycja 2012



Źródło: opracowanie własne.

Obliczenia wykonane eksperymentalnie z użyciem wstępnych wag celem prezentacji skutków zastosowania imputacji, nie stanowią oficjalnych danych statystyki publicznej.

Wnioski

- Imputacja dotyczy dużej części zbioru danych, przede wszystkim w przypadku zmiennych będących sumą wielu komponentów. Jest to związane głównie z występowaniem imputacji częściowej, co wynika zwykle z braków danych dotyczących pojedynczych komponentów.
- W przypadku tych zmiennych udział wartości imputowanych w tworzeniu wartości globalnej wskazuje na mniejszą skalę znaczenia zjawiska, niż by to wynikało z odsetka imputowanych rekordów. Skala ta 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. Wniosek ten potwierdza największy wpływ imputacji na wyniki uogólnienia w przypadku gospodarstw utrzymujących się z tego źródła.
- Prezentowane uogólnienia nie różnią się znacząco w wariantach z imputacja danych brakujących oraz z korektą wag. Należy jednak zauważyć, że korekta wag, by była równie skuteczna, musiałaby być wykonywana odrębnie dla każdej zmiennej.

Badanie Spójności Społecznej

Podstawowe informacje o badaniu

- Badanie zostało zrealizowane w 2011 roku, nie został ustalony stały kalendarz cyklicznej realizacji. Planuje się przeprowadzenie kolejnej edycji w odstępie 4 – 5 lat.
- Badanie wieloaspektowe. Dotyczy zagadnień związanych ze spójnością społeczną, ubóstwem, wykluczeniem, warunkami życia i jakością życia.
- Dochód badany jest wyłącznie jako dochód gospodarstwa domowego ogółem, bez rozbijania na komponenty. W formularzu występuje kilka pytań dotyczących dochodu, jednakże mają one charakter pomocniczy. Ich celem jest uzyskanie informacji o dochodzie ogółem, bądź też dostarczenie jak najlepszej informacji pomocniczej dla jego imputacji.
- Dla pytań dotyczących dochodu dopuszcza sie nieudzielenie odpowiedzi przez respondenta, co nie skutkuje dyskwalifikacją wywiadu. Powstałe w ten sposób braki pozycyjne są imputowane.

Celem imputacji jest uzyskanie kompletnej informacji dotyczącej **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 na poziomie jednostkowym jest łączona z innymi informacjami z badania,
- wyliczenia mierników, których konstrukcja wymaga informacji o różnych parametrach rozkładu dochodów, w tym również pozycyjnych i nieliniowych.

Warunki imputacji danych w badaniu

- Występują pytania o dwie kategorie dochodu ogółem:
    - dochód za poprzedni rok (jako pytanie o dochód roczny bądź średni dochód miesięczny – do wyboru respondenta) – będący bezpośrednim źródłem zainteresowania,
    - aktualny dochód miesięczny – nie dający się jednoznacznie 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,
    - jeżeli nie zgadza się podać wartości dochodu, wskazania przedziału, w którym mieści się jego dochód

  Dopiero w przypadku odmowy wskazania przedziału możemy mówić bez zastrzeżeń o braku odpowiedzi i mamy do czynienia z informacją tworzoną w całości za pomocą imputacji statystycznej.

Wobec takich uwarunkowań, dokonując imputacji mamy do dyspozycji stosunkowo precyzyjne pomocnicze **informacje dotyczące dochodu, które należy w pierwszej kolejności wykorzystać**:

- informację o przedziale dochodowym – jeśli została podana, imputacja ulega zawężeniu do granic przedziału,
- informację o dochodzie bieżącym – zmienną pomocniczą silnie skorelowana ze zmienną badaną.

Model opisujący zależność między dochodem a charakterystykami społeczno-ekonomicznymi gospodarstwa i jego głowy staje się podstawowym źródłem wiedzy, na podstawie której tworzona jest wartość imputacyjna, dopiero wtedy, gdy nie dysponujemy żadną z powyższych informacji.

Metoda imputacji

- Stosowana jest **stochastyczna imputacja regresyjna**.
- Imputowane **reszty** generowane są **z generatora liczb losowych** o wariancji odpowiadającej oszacowaniu wariancji składnika losowego modelu (z rozkładu normalnego dla modelu w postaci zlogarytmowanej).
- W przypadku, **gdy znamy przedział dochodowy**, **reszta generowana** jest z rozkładu **uciętego**, tak, by zagwarantować, że uzyskana wartość imputacyjna będzie należała do przedziału.

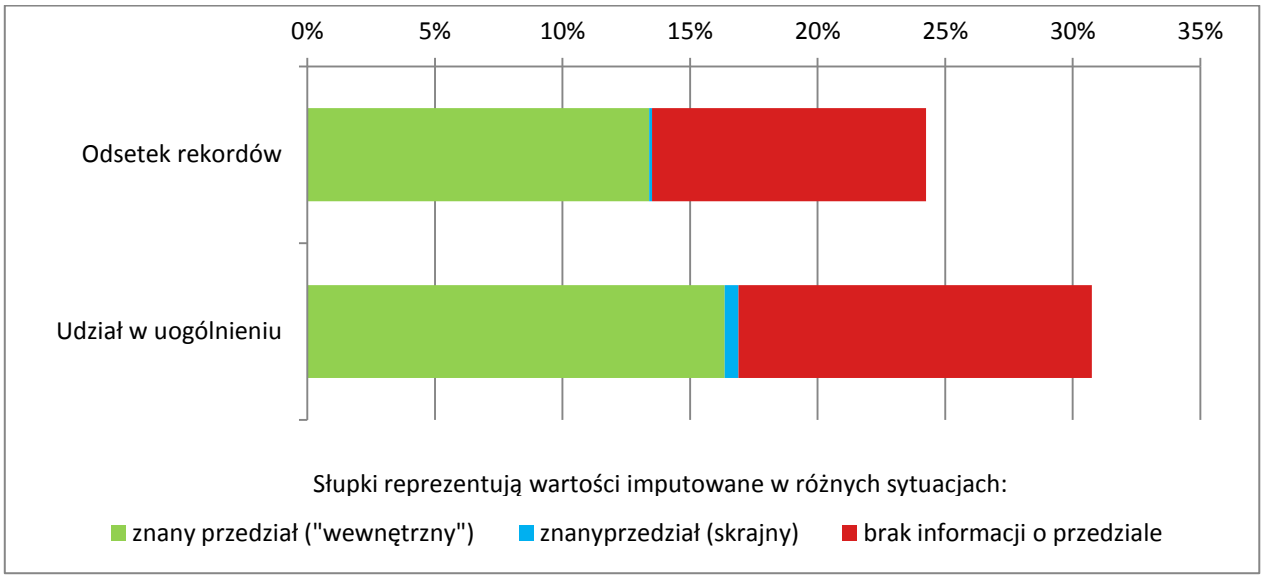
Stosowane są dwa **modele regresyjne**:

- **Model A** (preferowany) – objaśniający imputowaną zmienną (średni miesięczny dochód za rok poprzedni) za pomocą dochodu bieżącego.
- **Model B** – objaśniający zmienną imputowaną za pomocą charakterystyk społeczno ekonomicznych gospodarstwa domowego i jego głowy

Model A jest stosowany zawsze, gdy tylko może być użyty, co jest uzależnione przede wszystkim od dostępność informacji o dochodzie bieżącym. **Wykorzystuje informację o dochodzie** przekazaną przez respondenta i **daje dużo lepsze objaśnienie** niż model B.

Model B jest stosowany w przypadku braku możliwości zastosowania modelu A.

Skala i częstość stosowania imputacji
w badaniu spójności społecznej 2011
z uwzględnieniem różnej dostępności informacji o przedziale dochodowym



Liczba imputowanych rekordów
według dostępności informacji o przedziale dochodowym
i zastosowanego modelu

Dostępność informacji o przedziale dochodowym	Zastosowany model		Ogółem
	Model A (bardzo precyzyjny)	Model B (mniej precyzyjny)	
przedział „wewnętrzny”	368	1808	2176
przedział skrajny	5	16	21
brak informacji	293	**1481**	1774
Ogółem	666	3305	3971

Wnioski i uwagi

- Zastosowanie w kwestionariuszu badania przedziałów dochodowych (a – w mniejszym stopniu – także pytanie o dochód bieżący) znacząco podnosi jakość imputacji i otrzymywanych wyników.
- Udział rekordów imputowanych w tworzeniu wartości globalnej dochodu jest nieco większy, niż odsetek rekordów poddanych imputacji (wyższe przeciętne dochody).
- W przypadku imputowanych dochodów należących do przedziałów skrajnych udział w wartości globalnej jest znacznie wyższy niż w liczbie gospodarstw, jednak i tak na tyle mały, iż nie odgrywa większego znaczenia.
- Formalnie liczba rekordów imputowanych wynosi 3 971, przy próbie liczącej 14 884 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 wynosi jedynie 1481, co stanowi 10% zrealizowanej próby.