Urząd Statystyczny
w Łodzi

Prace nad metodologią badania dochodów i warunków życia

związane z realizacja projektu
„Zmiany i ulepszenia metod imputacji i ważenia oraz wdrożenie nowych flag w polskim badaniu EU-SILC”

Ośrodek Statystki Matematycznej

Tomasz Piasecki
Dawid Formella

Konferencja naukowa MET2023, Warszawa, 3-5.07.2023

Zmiany i ulepszenia metod imputacji i ważenia oraz wdrożenie nowych flag w polskim badaniu EU-SILC (Cel 5)

Projekt współfinansowany ze środków Unii Europejskiej na podstawie umowy z Komisją Europejską o udzielenie dotacji nr **101052514-2021-PL-ILC-SILC** w ramach programu ***Single Market Programme (SMP ESS)***

Okres realizacji: 27.09.2021 – 26.01.2023

Jednostka wiodąca: Urząd Statystyczny w Łodzi

Pozostałe **jednostki uczestniczące w realizacji**:

Departament Badań Społecznych GUS

Departament Programowania i Koordynacji GUS

Cele projektu

Zapewnienie zgodności polskiego badania EU-SILC z wymogami nowego rozporządzenia ramowego dot. statystyki społecznej (*Rozporządzenie Parlamentu Europejskiego i Rady UE nr 2019/1700 ustanawiające wspólne ramy statystyk europejskich dotyczących osób i gospodarstw domowych, opartych na danych na poziomie indywidualnym zbieranych metodą doboru prób*) oraz wprowadzenie w badaniu ulepszeń metodologicznych służących poprawie zaspokojenia potrzeb użytkowników (poprzez wzrost jakości uzyskiwanych danych w zakresie m.in. dokładności i użyteczności) oraz ograniczeniu obciążenia respondentów

Główne obszary tematyczne/zadania

- Opracowanie nowej metodyki estymacji wykorzystującej dane administracyjne dotyczące podatku dochodowego od osób fizycznych (PIT) do kalibracji wag uogólniających
- Wdrożenie nowych standardów oznaczania tzw. flag dla zmiennych dochodowych (metainformacja opisującą metodologię pozyskania konkretnej wartości w zbiorze danych jednostkowych)
- Opracowanie ulepszeń metodologicznych stosowanego algorytmu imputacji danych brakujących dot. dochodów, obejmujących m.in.:
  - wprowadzenie (poszerzenie zakresu stosowania) imputacji dedukcyjnej (jako alternatywy dla imputacji statystycznej w przypadku braków odpowiedzi lub w celu ograniczenia wywiadu)
  - zmianę sposobu konwersji dochodów pomiędzy wartością netto i wartością brutto (w przypadku braku jednej z wartości lub w celu ograniczenia wywiadu)

Opracowanie nowej alternatywnej metodyki estymacji (Obszar 1)

- **Kalibracja wag w EU-SILC – stan obecny**
  - kalibracja zintegrowana: te same wagi dla gospodarstw domowych i osób, uwzględniają zarówno warunki kalibracyjne dotyczące gospodarstw jak i osób
  - stosowane warunki kalibracyjne uwzględniają następujące informacje spoza próby:
    - struktura gospodarstw domowych wg wielkości
    - struktura demograficzna populacji wg płci i wieku
    - informacje terytorialne: NUTS-2 oraz podział na miasto i wieś – w powiazaniu z powyższymi klasyfikacjami
- **Idea alternatywnej metodyki estymacji**: zastosowanie dodatkowych warunków kalibracyjnych odzwierciedlających rozkłady dochodów obserwowane w danych administracyjnych zbieranych przez ministerstwo Finansów w związku poborem podatku PIT

Opracowanie nowej alternatywnej metodyki estymacji

- **Wykorzystywane dane**: agregaty utworzone na podstawie danych zanonimizowanych pochodzących z Ministerstwa Finansów (nie rozważano parowania indywidualnych danych jednostkowych) dotyczące:
  - liczby/odsetka osób (podatników) w określonych przedziałach dochodu (określonych jako kwantyle rozkładu)
  - sumy dochodów w tak wyznaczonych przedziałach
- **Potencjalne efekty**
  - zapewnienie zgodności rozkładów uzyskiwanych z badania z rozkładami ze źródła administracyjnego (może być również wadą/zagrożeniem, jeśli w badaniu mamy możliwość uzyskania informacji nie objętych rejestracją administracyjną)
  - zapobieżenie zniekształceniom rozkładu dochodu wynikającym z (nielosowego) występowania braków odpowiedzi (redukcja potencjalnego obciążenia/błędu nielosowego)
  - zmniejszenie błędu losowego poprzez nałożenie na ostateczne dane wynikowe dodatkowych nielosowych warunków

Opracowanie nowej alternatywnej metodyki estymacji

- **Realizacja prac**
  - Rozpoznanie danych administracyjnych
  - Określenie typów dochodów, dla których rozwiązanie może mieć zastosowanie (dochody z pracy najemnej oraz dochody tytułu emerytur i rent, po rozpoznaniu źródeł nie uwzględniono dochodów z pracy na rachunek własny)
  - Wybór metody kalibracji i opracowanie potencjalnych formuł (zastawów warunków kalibracyjnych)
  - Wykonanie obliczeń na danych historycznych dla rozważanych formuł, analiza uzyskanych wyników (m.in. ocena uwarunkowania numerycznego, porównanie z dotychczasowymi oszacowaniami, ocena precyzji estymacji)
  - Wybór optymalnej formuły spośród rozważanych na podstawie przeprowadzonych analiz oraz ocena uzyskanych efektów z punktu widzenia możliwej rekomendacji zastosowania

Flagi zmiennych dochodowych (Obszar 2)

- W EU-SILC każdej zmiennej dochodowej (zawierającej kwotę składowej dochodu osoby/gospodarstwa lub dochodu całkowitego) towarzyszą (w nowym standardzie) dwie **zmienne zawierające metainformacje dot. metodologii pozyskania lub wyprowadzenia informacji** zapisanych w konkretnych rekordach (dla konkretnych osób/gospodarstw domowych)
- **Nowy standard kodowania flag**:
  - zmienna _F (**flaga**) – 2-cyfrowy kod zawierający informacje nt. źródła danych i sposobu wyprowadzenia zmiennej (1.cyfra – np. wywiad, dane administracyjne, imputacja dedukcyjna, imputacja bazująca na modelu, konwersja, ...) oraz tego, czy informacja źródłowa, na podstawie której jest wyprowadzana, zawiera kwotę brutto czy netto (2. cyfra)
  - zmienna _IF (***imputation factor***) – określa udział wartości pochodzącej z innego źródła niż imputacja (wywiad, źródło administracyjne) w ostatecznie wyprowadzonej wartości (zmienne są często naliczane za pomocą złożonych formuł, część składników może być imputowana, część nie):
    - IF = 0 : wartość w całości imputowana
    - IF = 1 : nie występuje imputacja
    - IF $\in$ (0, 1) : częściowa imputacja

Flagi zmiennych dochodowych

- **Wdrożenie nowego standardu**
  - do 2020 r. stosowana była jedna zmienna dla flagi, od 2021 informacja rozdzielona na 2 zmienne (_F i _IF)
  - od 2021 r. rozszerzone raportowanie imputacji (np. metoda, wcześniej tylko *IF* zawarty we fladze) oraz źródła pochodzenia informacji
  - modyfikacja algorytmów imputacji i programów imputacyjnych w celu zapewnienia rozszerzonego raportowania zawierającego m.in. informację o sposobie imputacji
  - opracowanie i oprogramowanie algorytmów agregacji flag - konieczność określenia „głównego źródła” symbolizowanego we fladze _F w przypadku zmiennych będących sumą (wynikiem transformacji) wielu składowych

Imputacja dedukcyjna (Obszar 3)

- **Stan obecny**
  - stosowana jest przede wszystkim imputacja statystyczna – na podstawie (odpowiednio modelowanych) zależności empirycznych obserwowanych w zbiorze danych uzyskanych z wywiadu (nie wymagających imputacji)
  - imputacja statystyczna stosowana w odniesieniu do wszystkich kategorii dochodów (z pracy, z tytułu świadczeń, z innych tytułów)
  - nieliczne przypadki zastosowania imputacji dedukcyjnej – np. podatek od dochodów kapitałowych, świadczenie 500+

Imputacja dedukcyjna

- **Rozwiązania rozważane w projekcie:**
  - zastosowanie imputacji dedukcyjnej dla szerokiej grupy świadczeń społecznych (testy dotyczyły 29 rodzajów świadczeń)
  - imputacja dedukcyjna na podstawie reguł wynikających z przepisów prawnych regulujących przyznawanie oraz wysokość kwot świadczeń, bez wykorzystania imputacji statystycznej
  - wariant „minimalny”: zastosowanie w przypadku braków odpowiedzi
  - wariant „maksymalny”: zastosowanie zamiast pytania o kwotę świadczenia w wywiadzie (ograniczenie wywiadu)
- **Potencjalne efekty**
  - poprawa jakości uzyskiwanych danych, zwiększenie adekwatności wartości imputacyjnych na poziomie danych jednostkowych
  - redukcja obciążenia respondenta poprzez rezygnacją z niektórych pytań i ograniczenie długości wywiadu
  - redukcja obciążenia ankietera oraz kosztów badania

Imputacja dedukcyjna

- **Realizacja prac:**
  - Rozpoznanie krajowych przepisów prawnych regulujących przyznawanie oraz wysokość kwot poszczególnych świadczeń oraz pozyskanie informacji w tym zakresie (obejmujące m.in bezpośrednie kontakty z instytucjami zaangażowanych w wypłacanie świadczeń krajowych)
  - Wytypowanie świadczeń, które spełniają warunki umożliwiające zastosowanie imputacji dedukcyjnej;
  - Opracowanie reguł imputacji dedukcyjnej w oparciu o rozpoznane przepisy prawne i informacje oraz opracowanie nowego algorytmu imputacji danych dochodowych stanowiącego implementację tych reguł;
  - Testowanie i weryfikacja nowych algorytmów imputacji na danych historycznych
  - Ocena możliwości redukcji zakresu danych uzyskiwanych z wywiadu dzięki imputacji dedukcyjnej niektórych komponentów dochodu
  - Opracowanie rekomendacji dotyczących wdrożenia
- **Przygotowanie wdrożenia**
  - Modyfikacja formularza (dla edycji 2023 i następnych) w celu pozyskania dodatkowych informacji jakościowych (zamiast ilościowych) potrzebnych do imputacji dedukcyjnej oraz zmiany sposobu grupowania świadczeń
  - Modyfikacja aplikacji CAPI, za pomocą której realizowany jest wywiad

Konwersja dochodu

- Wyliczenie wartości brutto na podstawie wartości netto lub wartości netto na podstawie wartości brutto w sytuacji, gdy znana jest tylko jedna z nich
- **Stan obecny** – konwersja statystyczna: empiryczna funkcja konwersji oszacowana na podstawie danych pochodzących od respondentów, dla których uzyskana informacje o obydwu kwotach (regresja)
- **Rozwiązania rozważane w projekcie**:
  - konwersja brutto-netto i netto-brutto na podstawie reguł wynikających z przepisów prawnych regulujących wysokość podatków oraz składek na ubezpieczenia społeczne dla różnych kategorii podatników / ubezpieczonych / typów dochodów
  - zastosowanie w sytuacji, gdy z wywiadu znana jest tylko jedna kwota (zwykle netto, respondent nie podaje / nie potrafi podać kwoty obciążeń, tj. podatków i składek)
  - rozważenie ograniczenia wywiadu do pytania o jedną z kwot (netto)

Konwersja dochodu

- **Realizacja prac**
  - Opracowanie reguł (funkcji) konwersji **brutto->netto** na podstawie przepisów prawnych w odniesieniu do różnych typów dochodów i kategorii osób (funkcja konwersji ma postać monotonicznej funkcji przedziałami liniowej)
  - Określenie zmiennych z badania, które pozwalają wyodrębnić kategorie osób podlegające różnym funkcjom konwersji (określić parametry konwersji)
  - Odwrócenie funkcji konwersji brutto->netto w celu otrzymania funkcji konwersji **netto->brutto** (w większości przypadków łatwiejsza do uzyskania z wywiadu jest wartość netto, którą należy konwertować na wartość brutto)
  - Wykonanie obliczeń na danych historycznych i analiza uzyskanych wyników m.in. przez porównanie z zależnościami obserwowanymi empirycznie w przypadku respondentów podających obie kwoty (analiza dotyczyła dochodów z pracy najemnej oraz emerytur i rent)
  - Ocena uzyskanych efektów z punktu widzenia rekomendacji dot. wdrożenia
  - Analiza dotyczyła stanu obecnie historycznego, nie obejmowała roku 2022 („Polski Ład”) – ewentualne wdrożenie będzie wymagało opracowania reguł na podstawie nowego ustabilizowanego stanu prawnego

Osiągnięcia i efekty

- Wdrożenie nowego standardu symbolizacji flag zmiennych dochodowych począwszy od edycji 2021 – dostosowanie polskiego badania EU-SILC do wymogów wynikających z nowej regulacji ramowej i wytycznych Eurostatu oraz zwiększenie użyteczności analitycznej zbioru danych jednostkowych
- Wdrożenie poszerzonej imputacji dedukcyjnej dotyczącej świadczeń społecznych począwszy od edycji 2023
  - wdrożenie dotyczy 15 świadczeń (spośród 29 rozważanych), w tym 11 w wariancie „maksymalnym” (całkowita rezygnacja z zadawania pytania o kwotę)
  - zmniejszenie obciążenia respondentów spowodowane ograniczeniem pytań o kwoty dochodów
  - spodziewane efekty dotyczące poprawy jakości pozyskiwanych informacji oraz ograniczenia braków odpowiedzi (respondenci chętniej odpowiadają na pytania jakościowe niż pytania o kwoty dochodów)
- Efekty prac dotyczących kalibracji wag oraz konwersji dochodów stanowią produkty o charakterze analitycznym oraz narzędzia, które mogą być podstawą dalszych prac zmierzających do wdrożenia rozwiązań