Wydział Informatyki i Gospodarki Elektronicznej

UNIWERSYTET EKONOMICZNY
W POZNANIU

Praca doktorska

STATYSTYCZNA INTEGRACJA DANYCH W BADANIACH SPOŁECZNO-EKONOMICZNYCH

mgr Wojciech Roszka

Promotor

dr hab. Elżbieta Gołata, prof. nadzw. UEP

Uniwersytet Ekonomiczny w Poznaniu

Wydział Informatyki i Gospodarki Elektronicznej

Katedra Statystyki

Poznań 2013

Spis treści

WSTĘP

Problem

Informacja w demokratycznym społeczeństwie pełni istotną rolę w szczególności jako podstawa podejmowania decyzji administracyjnych, społecznych a przede wszystkim biznesowych (np. kierowanie inwestycji w rejony najbardziej ich potrzebujące, czy kampanii marketingowych do odpowiednich segmentów rynkowych). Dlatego też podmioty zgłaszające popyt na informację oczekują by była ona rzetelna oraz aktualna. Spełnienie tych postulatów jest trudne, a często wręcz niemożliwe. Jest to zadanie niezmiernie kosztowne, czasochłonne, wymagające ogromnej wiedzy oraz zaangażowania specjalistów. Przeprowadzenie badania specjalnego trwającego wiele dni, a nawet tygodni oznacza ograniczoną aktualność informacji, co przekłada się na jej użyteczność. Podmioty zgłaszające popyt na informację oczekują również, aby była ona wyczerpująca pod względem merytorycznym oraz by dopuszczalne były szacunki w możliwie najbardziej szczegółowych przekrojach.

Często jednak, ograniczenia finansowe uniemożliwiają przeprowadzenie badań na wystarczająco licznej próbie, która pozwoliłaby na szacunki dla małych domen. Dodatkowo postulat pełności merytorycznej badania wymusza konstrukcję długiego kwestionariusza, co skutkuje zwiększeniem liczby odmów i braków odpowiedzi [Al, Bakker 2000]. Tak więc „klasyczne" podejście do zbierania informacji stosowane w badaniach statystycznych, przy postulatach zgłaszanych przez odbiorców danych, implikuje sprzeczność między interesami strony popytowej (instytucjami administracji samorządowej, przedsiębiorstwami publicznymi i prywatnymi), a możliwościami strony podażowej (organami statystyki publicznej i instytutami badawczymi, agencjami badań rynku).

Rozwiązaniem problemów związanych z kosztownością i czasochłonnością badań wydaje się być wykorzystanie rejestrów administracyjnych w systemie statystyki publicznej. Obejmują one swym zakresem dużą liczbę jednostek, jak również dostarczają informację bogatą merytorycznie. Organy statystyki publicznej mogą pozyskiwać rejestry od dysponujących nimi instytucji i urzędów. Dodatkowo, dynamicznie rozwijająca się informatyzacja i wzrost mocy obliczeniowej komputerów powodują, że pozyskiwanie i przetwarzanie rejestrów jest stosunkowo szybkie i tanie. Ponadto podmioty administracji publicznej, które gromadzą dane, zapewniają rzetelność i duże pokrycie badanej populacji.

W różnych krajach (m.in. Holandii i krajach skandynawskich) podjęto próbę połączenia wielu źródeł danych, zarówno administracyjnych, jak i pochodzących z badań reprezentacyjnych. Lata doświadczeń w tym zakresie, jak również wysoka jakość źródeł umożliwiły przeprowa-

dzenie tzw. spisu wirtualnego. Jest to forma spisu powszechnego, w którym wszystkie informacje gromadzi się z dostępnych, już istniejących źródeł. Repozytoria danych łączone są przy użyciu unikalnych kluczy połączeniowych, będących odpowiednikami numeru PESEL, czy numeru ubezpieczenia społecznego. Umożliwia to ograniczenie kosztów badania, zmniejszenie obciążenia respondentów, a także uzyskanie zintegrowanych repozytoriów o szerokim spektrum informacyjnym, przy zachowaniu wszelkich wymogów związanych z jakością danych i ich bezpieczeństwem.

Wieloźródłowość niesie jednak pewne trudności, takie jak brak zmiennych mogących służyć jako klucz połączeniowy, różne definicje zmiennych zawartych w poszczególnych rejestrach, odmienne sposoby wypełniania rejestrów, jak również istnienie zduplikowanych rekordów. Dostosowanie rejestrów do wymogów statystyki publicznej jest w związku z powyższym dużym wyzwaniem.

Rozwiązaniem problemu dostępności informacji spełniającej wymogi określone nie tylko przez niezależne organizacje międzynarodowe i instytuty statystyki publicznej, ale przede wszystkim formułowanych przez gospodarkę, wydaje się być zastosowanie metod statystycznej integracji danych. Polegają one na łączeniu dostępnych źródeł w zintegrowany zbiór zawierający zmienne z baz wejściowych w sytuacji, gdy klucz połączeniowy nie jest dostępny. Pozwala to nie tylko na oszczędność kosztów i czasu, ale umożliwia również powiększenie zasobów informacyjnych już istniejących zbiorów oraz weryfikację spójności zawartych w nich danych.

Uzasadnienie wyboru tematu

W odróżnieniu od deterministycznych metod integracji, metodologia statystycznej integracji danych polega na łączeniu dwóch (lub więcej) repozytoriów nie posiadających unikatowego klucza połączeniowego na podstawie zestawu tzw. zmiennych wspólnych. Są to cechy, które występują w obu zbiorach, charakteryzują się taką samą (lub bardzo zbliżoną) definicją oraz zgodnością pod względem wariantów cech. W zależności od podejścia metodologicznego, integrowane zbiory mogą zawierać informacje o tych samych jednostkach lub nie. Zestaw zmiennych wspólnych nie zawsze w pełni identyfikuje jednostki, jednak na podstawie specjalnie określonych kryteriów, np. podobieństwa par rekordów, można z dużym prawdopodobieństwem wskazać te same jednostki lub jednostki do siebie bardzo podobne.

W przypadku braku możliwości wykorzystania unikalnego klucza połączeniowego lub gdy źródła są rozłączne, wykorzystanie metod statystycznej integracji danych może umożliwić łączną obserwację zmiennych nieobserwowanych łącznie w żadnym z pojedynczych repozy-

toriów. Dodatkowo, wykorzystanie zbiorów danych o różnym pokryciu stwarza możliwość estymacji na podstawie liczniejszego z integrowanych źródeł, a nawet na podstawie zintegrowanego zbioru o liczebności będącej sumą baz wejściowych.

Łączna obserwacja cech opisujących różne zjawiska przy zwiększonej liczebności próby stwarza przesłanki do podjęcia próby konstrukcji kompleksowych zintegrowanych repozytoriów danych społeczno-gospodarczych. Tym samym zawarte w nich informacje będą w większym stopniu zaspokajać potrzeby odbiorców, jak również gospodarki kraju. Integracja informacji z różnych źródeł umożliwiać może również osiągnięcie efektu synergii informacyjnej.

Uzyskane w wyniku integracji źródła danych muszą jednocześnie być zgodne z ogólnie przyjętymi standardami[1], charakteryzować się następującymi własnościami:

- użytecznością – spełniać aktualne i potencjalne wymogi użytkowników;
- dokładnością – uzyskane estymatory powinny charakteryzować się możliwie niskim obciążeniem, zgodnością oraz możliwie wysoką efektywnością;
- terminowością – szacunki powinny dotyczyć możliwie nieodległego momentu czasowego;
- dostępnością – powinny być ogólnie dostępne oraz darmowe;
- porównywalnością – dane powinny być przetwarzane w taki sposób, by wyniki uzyskane na podstawie różnych źródeł były między sobą zgodne;
- spójnością – zastosowane definicje, warianty, a także populacje i jednostki powinny być zgodne niezależnie od źródła danych.

Ważne jest w szczególności spełnienie wymogów użyteczności, dokładności i spójności oraz zapewnienie wysokiej jakości szacunków. Nie mniej istotna jest także możność poprawy precyzji szacunków dla możliwie niskich poziomów agregacji przestrzennej i merytorycznej.

Cel pracy i hipotezy badawcze

Głównym celem pracy jest ocena możliwości i zasadności stosowania metod statystycznej integracji danych (nie posiadających unikatowego klucza połączeniowego lub nie zawierających informacji o tych samych jednostkach) dla rozszerzenia zakresu merytorycznego szacunków, a także weryfikacja precyzji estymacji na podstawie zintegrowanych danych. Cel ten zostanie zrealizowany poprzez następujące cele pomocnicze:

[1] Definicja jakości według Eurostatu [Working Group, Sixth Meeting: "Assessment of quality in statistics", Luxembourg, 2-3 October 2003]

1. empiryczną weryfikację wybranych metod statystycznej integracji danych i ewaluację jakości połączenia różnych źródeł,
2. badanie jakości zintegrowanych źródeł oraz sprawdzenie zgodności i precyzji estymacji przeprowadzonej na ich podstawie,
3. empiryczną ewaluację metod statystycznej integracji danych w kontekście zgodności rozkładów badanych zmiennych, ich wzajemnych relacji oraz spójności szacunków.

Stosownie do wyżej postawionych celów sformułowano następujące hipotezy badawcze:

1. Zastosowanie metod statystycznej integracji danych pozwala uzyskać informacje spełniające wymogi użyteczności, dokładności i spójności.
2. Statystyczna integracja danych stwarza możliwość łącznej obserwacji zmiennych z różnych badań, do tej pory wspólnie nieobserwowanych.
3. Istnieje możliwość wykorzystania zintegrowanych źródeł w celu estymacji na niskim poziomie agregacji przestrzennej.

Według Autora, ***novum* pracy** będzie adaptacja nowatorskich metod statystycznej integracji danych pochodzących z różnych źródeł dla potrzeb polskiej statystyki publicznej, a także próba estymacji oraz wnioskowania statystycznego na podstawie zintegrowanych zbiorów.

Metody badawcze i źródła danych

Przygotowana rozprawa ma charakter teoretyczno-empiryczny. Wśród wykorzystanych metod badawczych znalazły się studia literaturowe, metody statystycznej integracji danych, statystyki wielowymiarowej, wnioskowania statystycznego, w tym estymacji, oceny precyzji oraz badania empiryczne.

W literaturze zasadniczo wyróżnia się dwie metody łączenia danych: **deterministyczną** i **stochastyczną**. Podstawą łączenia deterministycznego (*deterministic record linkage*) jest identyczność wybranych pól w łączonych rekordach, tzw. klucz identyfikacyjny. Warunkiem zastosowania metody deterministycznej jest zgodność wartości zmiennych kluczowych w obydwu zbiorach. Błędy oraz braki danych występujące w zmiennych kluczowych uniemożliwiają zastosowanie integracji deterministycznej.

Różne pochodzenie zbiorów danych sprawia, że nie zawsze możliwe jest zdefiniowanie unikatowych kluczy. W takich przypadkach, dla integracji repozytoriów, konieczne jest zastosowanie metod stochastycznych, wśród których wyróżnia się dwa główne nurty: probabilistycz-

ne łączenie rekordów (*probabilistic record linkage*) oraz parowanie statystyczne[2] (*statistical matching, data fusion*). W niektórych opracowaniach wymienia się również metody geostatystyczne [Blum, Calvo 2001].

Bez względu na rodzaj metod stochastycznych, dane wejściowe stanowią dwa (lub więcej) zbiory zawierające informacje o jednostkach tej samej populacji generalnej. Ponadto wektory zmiennych w integrowanych źródłach zawierają część wspólną, tj. zmienne, które występują w każdym ze zbiorów. Zmienne wspólne muszą charakteryzować się taką samą (lub bardzo zbliżoną) definicją oraz identycznymi wariantami (kategoriami). Na podstawie wybranych zmiennych wspólnych, tzw. parujących, oblicza się pewne miary podobieństwa między poszczególnymi rekordami. Następnie łączy się rekordy „najbardziej" do siebie podobne. Celem tego postępowania jest utworzenie nowego zbioru danych zawierającego łączną informację ze wszystkich integrowanych repozytoriów.

W metodzie **probabilistycznego łączenia rekordów** zakłada się, że łączone repozytoria danych zawierają informacje o tych samych jednostkach. Ponieważ zbiory nie posiadają unikatowego klucza połączeniowego, wykorzystuje się informację zawartą w zmiennych parujących w celu obliczenia tak zwanej wagi połączeniowej. Jest to przekształcenie prawdopodobieństwa, że porównywana para rekordów należy do tej samej jednostki. Najczęściej, w literaturze przedstawia się probabilistyczne łączenie rekordów jako proces kilkustopniowy. W pierwszej kolejności należy doprowadzić zmienne parujące do porównywalności poprzez np. harmonizację ich definicji i wariantów. W przypadku zmiennych o charakterze tekstowym, które są najbardziej „narażone" na błędy (literówki, różne wielkości liter itp.), dokonuje się operacji standaryzacji poprzez zastosowanie takich narzędzi jak parsery (analizatory składniowe), czy komparatory łańcuchowe[3]. W celu optymalizacji algorytmu integracji, bardzo często dzieli się zbiory na rozłączne grupy i scalania dokonuje się wyłącznie w obrębie wydzielonych warstw. Taka operacja nazywana jest grupowaniem lub warstwowaniem. Kolejnym krokiem jest obliczenie wagi połączeniowej. Jej wysoka wartość sugeruje, że porównywane rekordy należą do tej samej jednostki, zaś niska – że nie należą. Ponieważ integrowane zbiory zawierają informację o tych samych jednostkach, połączony zbiór odnosi się do jednostek rzeczywistych. Stąd zachodzi potrzeba zachowania odpowiednich procedur bezpieczeństwa uniemożliwiających ujawnienie danych jednostkowych.

[2] *Statistical matching* – polskie tłumaczenie tego terminu jako „parowanie statystyczne" jest przedmiotem dyskusji. W niniejszej pracy wybrano to określenie ze względu na fakt, że w literaturze najczęściej wykorzystywane jest podejście łączenia w pary rekordów najbardziej do siebie podobnych (pod względem wybranych charakterystyk).

[3] Za pomocą komparatorów łańcuchowych oblicza się pewne miary podobieństwa wartości w porównywanych rekordach integrowanych baz.

Metoda probabilistycznego łączenia rekordów nadaje się do łączenia badań pełnych i rejestrów administracyjnych w sytuacji, gdy integrowane bazy pozbawione są informacji w pełni identyfikujących jednostkę (np. z powodu potrzeby zachowania tajemnicy statystycznej) lub gdy zmienne stanowiące unikatowy klucz połączeniowy zawierają błędy i braki danych.

Parowanie statystyczne jest metodą łączenia zbiorów danych, które nie zawierają informacji o tych samych jednostkach. Na podstawie wektora zmiennych parujących oblicza się miary podobieństwa porównywanych rekordów. Podobnie jak w nurcie probabilistycznego łączenia rekordów, integruje się jednostki „najbardziej" do siebie podobne, jednak z definicji są to tzw. jednostki nierzeczywiste (syntetyczne). W nurcie parowania statystycznego zakłada się, że jeżeli jednostki podobne są do siebie pod względem pewnych wyróżnionych cech (np. płci, wieku, wykształcenia itp.), będą również charakteryzować się wysoką zgodnością w kwestiach opisywanych przez dołączane zmienne. W literaturze jako miarę podobieństwa poszczególnych rekordów najczęściej wymienia się funkcję odległości lub tworzy się modele regresji i regresji stochastycznej w celu imputacji braków danych.

Metody parowania statystycznego, z racji założenia o niewystępowaniu w integrowanych zbiorach informacji o tych samych jednostkach, nadają się do łączenia plików z badań reprezentacyjnych. Ze względu na syntetyczny charakter jednostek w zintegrowanych zbiorach, nie występuje niebezpieczeństwo ujawnienia informacji o rzeczywistych jednostkach statystycznych.

W **metodach geostatystycznych** integracja dokonywana jest poprzez przyporządkowanie poszczególnym rekordom identyfikatorów przestrzennych (np. współrzędnych geograficznych, kodów pocztowych, itp.). Łączenie następuje w przypadku zgodności lub wysokiego podobieństwa rekordów identyfikujących położenie obiektów w przestrzeni. Dużą rolę odgrywają w tej metodzie narzędzia geograficznych systemów informacyjnych (*Geographic Information System, GIS)*, które służą do gromadzenia i przetwarzania danych przestrzennych. Nurt ten znajduje się obecnie w fazie rozwojowej. Po raz pierwszy na szerszą skalę zastosowano go w badaniach izraelskiego urzędu statystycznego [Blum, Calvo 2001]. Metody te nie będą przedmiotem rozważań w niniejszej pracy.

Punkt ciężkości rozważań w pracy skierowany zostanie na zastosowanie statystycznych metod integracji dla potrzeb skonstruowania określonego zintegrowanego repozytorium danych społeczno-ekonomicznych. W tym celu wykorzystane zostaną dwa jednostkowe zbiory danych wejściowych:

— Badanie Budżetów Gospodarstw Domowych,
— Badanie Dochodów i Jakości Życia EU-SILC.

W prowadzonym badaniu empirycznym zamiarem Autora będzie obserwacja powiązań między wydatkami gospodarstw domowych a charakterystykami związanymi z jakością życia, a także dochodami członków gospodarstw domowych. Związki takie nie są obserwowane w żadnym z pojedynczych źródeł. Jednocześnie podjęta zostanie próba utworzenia zintegrowanego zbioru o liczebności umożliwiającej tworzenie szacunków o zwiększonej, w porównaniu z którymkolwiek ze zbiorów wejściowych, precyzji.

Struktura pracy

Strukturę pracy podporządkowano realizacji celu głównego, celów szczegółowych oraz empirycznej weryfikacji sformułowanych hipotez. Na rozprawę składają się: wstęp, pięć rozdziałów, zakończenie, bibliografia oraz aneks tabelaryczny.

Rozdział pierwszy stanowi opis istoty systemu statystycznego opartego na zintegrowanych źródłach danych. Scharakteryzowano w nim ideę integracji, a także istotę systemu statystyki publicznej opartego na rejestrach administracyjnych. W kolejnych punktach wskazano możliwości łączenia repozytoriów danych pochodzących z badań reprezentacyjnych, problemy estymacji w zintegrowanych źródłach, a także ważny aspekt jakości danych statystycznych ze szczególnym uwzględnieniem rejestrów administracyjnych. Opisano także problem bezpieczeństwa danych w kontekście wymogów wynikających z zapisów ustawy o statystyce publicznej, w szczególności nakazu zachowania tajemnicy statystycznej, a także społecznych obaw związanych z możliwym ujawnieniem informacji wrażliwych.

W rozdziale drugim przedstawiono doświadczenia różnych krajów i podmiotów wynikające z utworzenia zintegrowanych systemów statystycznych. Szczególny nacisk położono na badania spisowe i reprezentacyjne oparte na zintegrowanych źródłach. Przytoczono krótko także praktyki prywatnych firm badawczych na przykładzie GfK oraz AC Nielsen. Opisano również doświadczenia polskiej statystyki publicznej na przykładzie badania „Przepływy ludności związane z zatrudnieniem". W ostatnim punkcie rozdziału przedstawiono prace Eurostatu i instytutów statystycznych krajów europejskich w dziedzinie integracji na podstawie ostatnich międzynarodowych projektów takich jak MEETS, CENEX oraz Data Integration.

W rozdziale trzecim scharakteryzowano znajdujące się w posiadaniu organów polskiej statystyki publicznej przykładowe źródła danych. Dokonano próby oceny ich zawartości informacyjnej i jakości. Opisano także źródła danych pozostające w gestii instytucji spoza systemu polskiej statystyki publicznej na przykładzie Polskiego Generalnego Sondażu Społeczne-

go i Diagnozy Społecznej. Zaproponowano ideę konstrukcji zintegrowanego systemu statystyki społecznej opartego na przedstawionych źródłach.

Rozdział czwarty posiada charakter metodologiczny. Przedstawiono w nim metody statystycznej integracji danych w kontekście możliwości ich wykorzystania w procesie konstrukcji zintegrowanego repozytorium danych społeczno-ekonomicznych. Przedmiotem dyskusji są metody harmonizacji zbiorów danych oraz oceny ich jakości. Szczegółowo opisano algorytmy statystycznej integracji danych, wskazano różne podejścia metodologiczne oraz nowatorskie rozwiązania w tej dziedzinie. Szczególny nacisk położono na zagadnienia oceny jakości zintegrowanych danych w kontekście możliwości ich wykorzystania przez statystykę publiczną.

Rozdział piąty jest rozdziałem empirycznym. Przedstawiono w nim koncepcję utworzenia modułu zintegrowanego repozytorium danych społeczno-ekonomicznych. Przeprowadzono kompleksowe badanie empiryczne, które pozwoliło zweryfikować postawione hipotezy badawcze. Dokonano także próby estymacji wybranych charakterystyk w warunkach zintegrowanych źródeł danych, jak również oceniono precyzję otrzymanych szacunków. Wskazano korzyści wynikające z integracji, a także problemy powstałe w toku badania.

W zakończeniu sformułowano wnioski i spostrzeżenia uzyskane w toku prowadzonych dociekań oraz przedstawiono perspektywy dalszych badań.

Chciałbym złożyć wyrazy podziękowania Promotorowi, Pani prof. dr hab. Elżbiecie Gołacie za życzliwość, pomoc oraz opiekę podczas pisania niniejszej rozprawy, a nade wszystko za daleko idącą cierpliwość i wiarę we mnie w chwilach kryzysu i zwątpienia.

ROZDZIAŁ I. INTEGRACJA DANYCH W BADANIACH STATYSTYCZNYCH

1.1. Rejestry administracyjne

Informacja w systemie statystyki publicznej pozyskiwana jest drogą badań statystycznych, zarówno próbkowych, jak i pełnych, których etapy przeprowadzania są ustalone i dobrze opisane w literaturze [m.in. Paradysz *et al.* 2004, Aczel 2000, Witkowski *et al.* 2009]. W „klasycznym" podejściu do badania zjawisk społeczno-ekonomicznych, punktem wyjścia są potrzeby informacyjne odbiorców. Dążąc do ich zaspokojenia, służby statystyczne projektują badania obejmujące różne zagadnienia, wśród których można wymienić aktywność ekonomiczną ludności, budżety gospodarstw domowych, czy działalność przedsiębiorstw. W kolejnym etapie sporządzany jest operat losowania[4] i na jego podstawie losuje się próbę. Następnie prowadzona jest obserwacja statystyczna. Zebrany materiał statystyczny po odpowiedniej kontroli formalnej i merytorycznej oraz przetworzeniu (imputacji braków danych, przeważeniu, edycji danych itp.) jest podstawą szacunków, których wyniki publikowane są w formie tabel i wykresów statystycznych. Publikacje te służą wielu różnym celom, wśród których wymienić można wspomaganie organów rządowych i samorządowych w formułowaniu strategii rozwoju oraz prowadzenia polityki społecznej i gospodarczej.

Mając na względzie obciążenie respondentów oraz koszty, organy statystyki publicznej przeprowadzają badania dotyczące określonych, 'pojedynczych[5]' tematów. Podejście takie utrudnia tworzenie wielowymiarowych szacunków obejmujących różne zagadnienia społeczno-gospodarcze. Dodatkowo, ograniczenia budżetowe[6] powodują, że liczebność próby w badaniu jest zwykle zbyt niska, by szacunki mogły być dokonywane dla małych jednostek terytorialnych. Może to powodować brak zaspokojenia potrzeb informacyjnych samorządów (np. powiatów) dotyczących szczegółowej informacji o kształtowaniu się zjawisk na ich terytorium. Jednocześnie niewielka próba powoduje trudności w wykryciu i badaniu zjawisk rzadkich w skali kraju (np. przestępczości wśród mniejszości etnicznych), które w skali małej domeny mogą być dużym problemem (np. w rejonach przygranicznych).

Uwzględnienie potrzeb wynikających z globalizacji gospodarki wymaga połączenia informacji z różnych dziedzin. Badania statystyczne obejmujące szeroki zakres merytoryczny anali-

[4] Do tworzenia operatów wykorzystywane są również rejestry administracyjne.

[5] Przez pojedynczy temat można rozumieć np. rynek pracy ujęty w Badaniu Aktywności Ekonomicznej Ludności. Należy być świadomym, że zagadnienia rynku pracy są same w sobie bardzo złożone i różnorodne. Jednak w badaniu tym nie ma powiązania rynku pracy np. z jakością życia.

[6] Badanie Aktywności Ekonomicznej Ludności w 2012 roku obejmujące około 54,7 tys. gospodarstw domowych kosztowało około 41 mln złotych, zaś Badanie Budżetów Gospodarstw Domowych w tym samym okresie aż 58 mln złotych [Program Badań Statystycznych Statystyki Publicznej na 2012 rok, 2011].

zowanych zagadnień są jednak bardzo kosztowne. Ich przeprowadzenie wiąże się jednocześnie z bardzo dużym obciążeniem respondentów [van der Laan 2000] i wynikającym z tego tytułu wzrostem liczby braków odpowiedzi i odmów wypełnienia kwestionariusza[7], nawet przy zastosowaniu nowoczesnych, w mniejszym stopniu obciążających respondentów, metod zbierania informacji (CATI[8], CAWI[9] itp.). Wysokie koszty zbierania informacji w badaniach z długimi (obszernymi) kwestionariuszami mogą również prowadzić do zmniejszenia próby dodatkowo utrudniając wnioskowanie dla małych domen.

Informacja na niskim poziomie agregacji przestrzennej dostępna jest z badań pełnych. W badaniu takim pomiarem objęte są wszystkie jednostki należące do populacji docelowej. Ze względu jednak na zasięg takiego badania, jest ono dużym wyzwaniem zarówno finansowym, jak i organizacyjnym. W przypadku spisu powszechnego, nawet bogate kraje nie mogą sobie pozwolić na powtarzanie go częściej niż raz na kilka lat. W okresach międzyspisowych powstaje luka informacyjna, której badania reprezentacyjne nie są w stanie wypełnić. Zapewnienie precyzyjnej informacji dla małych domen jest dużym wyzwaniem dla organów statystyki publicznej pod względem metodologicznym (np. związane z zastosowaniem metod statystyki małych obszarów).

Wykorzystanie rejestrów administracyjnych może dostarczyć informacji na niskim poziomie agregacji z dużą częstotliwością. W art. 13 ust. 1 ustawy z dnia 29 czerwca 1995 roku o statystyce publicznej, ustawodawca nakazuje organom administracji rządowej i samorządowej przekazywanie danych administracyjnych służbom statystyki publicznej w terminach i formie określonych w programie badań statystycznych. Zbiory te opisują pojedyncze zagadnienia, takie jak bezrobocie rejestrowane, ruch naturalny i wędrówkowy ludności, czy działalność podmiotów gospodarczych, bez możliwości dokonywania wielowymiarowych szacunków ukazujących relacje i zależności w funkcjonowaniu społeczeństwa, gospodarki i państwa jako całości. Dodatkowo definicje cech zawartych w rejestrach mogą się różnić od przyjętych w statystyce publicznej. Rejestry administracyjne, z definicji, tworzone są do wypełniania zadań publicznych [Ustawa z dnia 17 lutego 2005 r. o informatyzacji działalności podmiotów realizujących zadania publiczne, Dz.U. Nr 64, poz.565, z późn. zm.], nie zaś bezpośrednio do celów statystycznych. Z odrębności systemów statystycz-

[7] Przytoczyć tu można stale zmniejszający się poziom realizacji próby w badaniu Polski Generalny Sondaż Społeczny. W pierwszej edycji badania, w 1992 roku, wynosił 82,4%. W następnych latach ulegał stałemu spadkowi, by w 2008 roku wynosić już zaledwie 51,8% [Cichomski *et al.* 2009].

[8] *Computer-Assisted Telephone Interview* – wywiad wspomagany telefonicznie.

[9] *Computer-Assisted Web Interview* – wspomagany komputerowo wywiad przy pomocy strony www.

nych i administracyjnych (por. tabela 1.1) wynika, iż wykorzystanie rejestrów w statystyce publicznej nie może być automatyczne.

Tabela 1.1. System administracyjny a system statystyczny

Charakterystyka		System administracyjny	System statystyczny
1. Cel powstania		Podejmowanie decyzji administracyjnych	Dokonywanie szacunków i analiz
2. Zbiorowość		Wszystkie podmioty (jednostki) prawnie podległe danemu gestorowi	Wszystkie podmioty (jednostki) objęte badaniem statystycznym
3. Jednostka		Pojedynczy element zbiorowości Pobierane dane niezbędne są do podejmowania decyzji administracyjnych	Pojedynczy element zbiorowości Pozyskiwane informacje są podstawą szacunków dotyczących populacji lub jej podgrup
4. Cecha	**Definicja**	Wynika z aktów prawnych Mogą być odrębne dla różnych rejestrów	Wynika z ustaleń organizacji międzynarodowych Wymóg porównywalności
	Warianty	Nie muszą być zestandaryzowane	Zharmonizowane i porównywalne
5. Błędy		Błędy nielosowe Brak kontroli statystycznej	Błędy losowe i nielosowe Kontrola statystyczna
6. Użyteczność		Dobre źródło informacji dla małych obszarów	Jakość i możliwości szczegółowej analizy ograniczone wielkością próby
7. Terminowość i punktualność		Zróżnicowane w zależności od źródła Niektóre bardzo aktualne, inne mniej terminowe niż badania statystyczne	Zróżnicowane w zależności od badania Często mają charakter retrospektywny
8. Dostępność i przejrzystość		Wpływ uregulowań prawnych Możliwe bariery techniczne i instytucjonalne	Bezpośrednia kontrola urzędu statystycznego
9. Porównywalność	**W czasie**	Zależy od zmieniających się w czasie regulacji prawnych	Bezpośrednia kontrola urzędu statystycznego
	W przestrzeni	Porównywalność w skali kraju, Brak porównywalności w skali międzynarodowej	Bezpośrednia kontrola urzędu statystycznego

Źródło: opracowanie własne na podstawie [Penneck 2007]

Po pierwsze, informacje zawarte w rejestrach służą jako podstawa w podejmowaniu decyzji administracyjnych, które wpływają na funkcjonowanie podmiotów. Informacje w systemach statystycznych służą do analiz, na podstawie których formułowane są wnioski o całej populacji (por. tabela 1.1).

Zbiorowość w rejestrach administracyjnych jest określana za pomocą aktów prawnych. Jest to tak zwana zbiorowość *de iure*. W badaniach statystycznych zbiorowością określa się wszystkie jednostki określone z punktu widzenia celu badania (pod względem rzeczowym, czasowym i przestrzennym). Bardzo często zbiorowości administracyjne i statystyczne, nawet nazwane tak samo, nie pokrywają się w pełni. W celu zapewnienia porównywalności obu populacji może zaistnieć potrzeba ich harmonizacji, która najczęściej wiąże się z wyodrębnieniem

ich części wspólnej. Podejście takie powoduje usunięcie części jednostek, a to oznacza utratę informacji.

W systemach administracyjnych jednostka jest przedmiotem decyzji i działań poszczególnych organów wykonawczych, a baza danych służy pozyskaniu informacji na temat określonego podmiotu. W systemach statystycznych jednostka jest traktowana raczej jako część zbiorowości, dla której tworzone są pewne informacje agregatowe – jednostka nie leży więc w centrum zainteresowania (wyłączając badania monograficzne).

Definicja cechy w systemie administracyjnym wynika z aktów prawnych i może być różna dla różnych rejestrów. Warianty cechy nie muszą być spójne, ponieważ zwykle system administracyjny nie jest podstawą tworzenia zestawień statystycznych. W systemie statystycznym definicje cech, podobnie jak warianty, są spójne dla wszystkich badań (często wynika to z przyjętych ustaleń organizacji międzynarodowych).

W kontekście występowania błędów, w rejestrach wszystkie dane muszą być zgodne i nie mogą ich zawierać, jednak nie ma konieczności by zapis danej kategorii był taki sam w każdym rekordzie (np. kod pocztowy pisany z myślnikiem lub bez, pełny zapis nazw ulic lub skrócony itp.). Występujące nieścisłości (np. brak numeru PESEL, czy NIP) nie mają charakteru losowego i zazwyczaj wynikają z awarii systemów kontroli. W badaniach statystycznych poszczególne warianty cech muszą być ujednolicone, by możliwe było tworzenie spójnych komunikatów. Prowadzony również w badaniach statystycznych rachunek błędów powoduje, że nieścisłości w wynikach są kontrolowane i podejmowane są działania w celu ich redukcji. Różnice w podejściu do gromadzenia danych administracyjnych i statystycznych mogą powodować rozbieżności w publikowanych komunikatach.

Rejestry administracyjne obejmując z obowiązku prawnego całą populację mogą stanowić podstawę do tworzenia statystyk dla małych obszarów (również jako źródło pomocnicze dla estymacji pośredniej). W przypadku systemów statystycznych, możliwości publikacji szacunków na określonym poziomie agregacji są w dużej mierze ograniczone przez wielkość próby (pokrycie rzadko przekracza 1% populacji).

Również terminowość rejestrów może być większa niż w przypadku badań statystycznych (które mogą mieć charakter retrospektywny). Duża część zbiorów będących w posiadaniu organów publicznych jest uzupełniana na bieżąco i komunikaty tworzone na ich podstawie mogłyby być publikowane z częstotliwością miesięczną (a nawet i większą).

Dostępność i przejrzystość danych publikowanych przez organy statystyki publicznej są ściśle kontrolowane, co wynika w dużej mierze z uregulowań organizacji międzynarodowych. W przypadku systemów administracyjnych, decydujący wpływ mają regulacje prawne

danego państwa oraz związane z nimi zmiany definicji w czasie. Przystosowanie poszczególnych rejestrów do konkretnych celów publicznych może tworzyć również problemy techniczne i instytucjonalne. Architektura baz i hurtowni danych może być bardzo odmienna, a zapisy prawne, zwłaszcza dotyczące ochrony danych osobowych, mogą zniechęcać instytucje do udostępniania swoich repozytoriów.

Porównywalność danych w systemach statystycznych jest również ściśle kontrolowana. Bardzo wiele badań (np. BAEL, EU-SILC) zawierają komponent panelowy, w którym śledzi się zmiany charakterystyk wybranych jednostek w czasie, przez co dane muszą być spójne dla różnych okresów. W systemach administracyjnych porównywalność danych zależy od obowiązujących w czasie badania rozwiązań prawnych.

Należy zaznaczyć, że wielu gestorów rejestrów posiada własne służby statystyczne (np. Zakład Ubezpieczeń Społecznych, Narodowy Bank Polski) tworzące sprawozdania na podstawie danych administracyjnych. Publikowane na ich podstawie komunikaty są przeznaczone jednak na potrzeby tych instytucji i nie uwzględniają potrzeb innych odbiorców.

Idea systemu statystyki opartego na zintegrowanych źródłach wiąże się z łączeniem repozytoriów danych w taki sposób, by możliwe były:

- łączna estymacja zmiennych nie występujących łącznie w pojedynczych źródłach,
- łączna estymacja na podstawie integrowanych źródeł uwzględniająca szeroki zakres cech oraz łączną liczbę obserwacji,
- indywidualne generowanie zestawień,
- prezentacja danych z różnych źródeł ukazująca wzajemne relacje i zależności społeczno-ekonomiczne.

Pionierem w wykorzystaniu źródeł administracyjnych w systemie statystyki publicznej były Finlandia [*Statistics Finland* 2004] oraz Norwegia [Tonder 2008]. Już w spisie w 1970 wykorzystano w tych krajach rejestr ludności, stopniowo wprowadzając w następnych spisach dalsze rejestry. W 1981 roku do tych państw dołączyła Dania [Borchsenius 2000], która od razu włączyła do systemu spisowego wszystkie dostępne rejestry (Finlandia wprowadziła spis w pełni oparty o rejestry dopiero w 1990, a Norwegia w 2011 roku). Innymi krajami wykorzystującymi administracyjne źródła danych w systemie spisowym są m.in. Austria [*Statistics Austria* 2008], Australia [Ralphs, Tutton 2011], Holandia [Nordholt 2004], Izrael [Kamen 2005], Kanada [Ballano 2009], Nowa Zelandia [Bycroft 2011], Stany Zjednoczone [Prevost, Leggieri 1999], Szwajcaria [*Swiss Federal Statistical Office* 2008] oraz Szwecja [Bruhn 2001, Wallgren, Wallgren 2007].

1.2. Idea integracji różnych źródeł dla potrzeb statystycznych

Celem integracji jest stworzenie nowego, bogatszego, zbioru danych, opisującego pewną populację docelową. Ze względu na różne architektury systemów administracyjnych oraz ich niestatystyczne przeznaczenie, rejestry wymagają dostosowania do potrzeb organów statystyki publicznej. Prace dostosowujące rozpoczynają się od badania infrastruktury danych celem uzyskania potrzebnej wiedzy o zawartości informacyjnej zbiorów, częstości publikacji, zakresie terytorialnym. W kolejnym kroku łączy się zbiory w sposób deterministyczny za pomocą kluczy połączeniowych (np. numer PESEL). Otrzymany w ten sposób zbiór mikrodanych poddaje się procesowi harmonizacji (ujednolicania populacji definicji oraz korekty błędów). W kolejnym kroku zharmonizowane źródło poddaje się procesom analitycznym (imputacja, kalibracja, agregacja, estymacja) umożliwiającymi publikację końcowych komunikatów statystycznych.

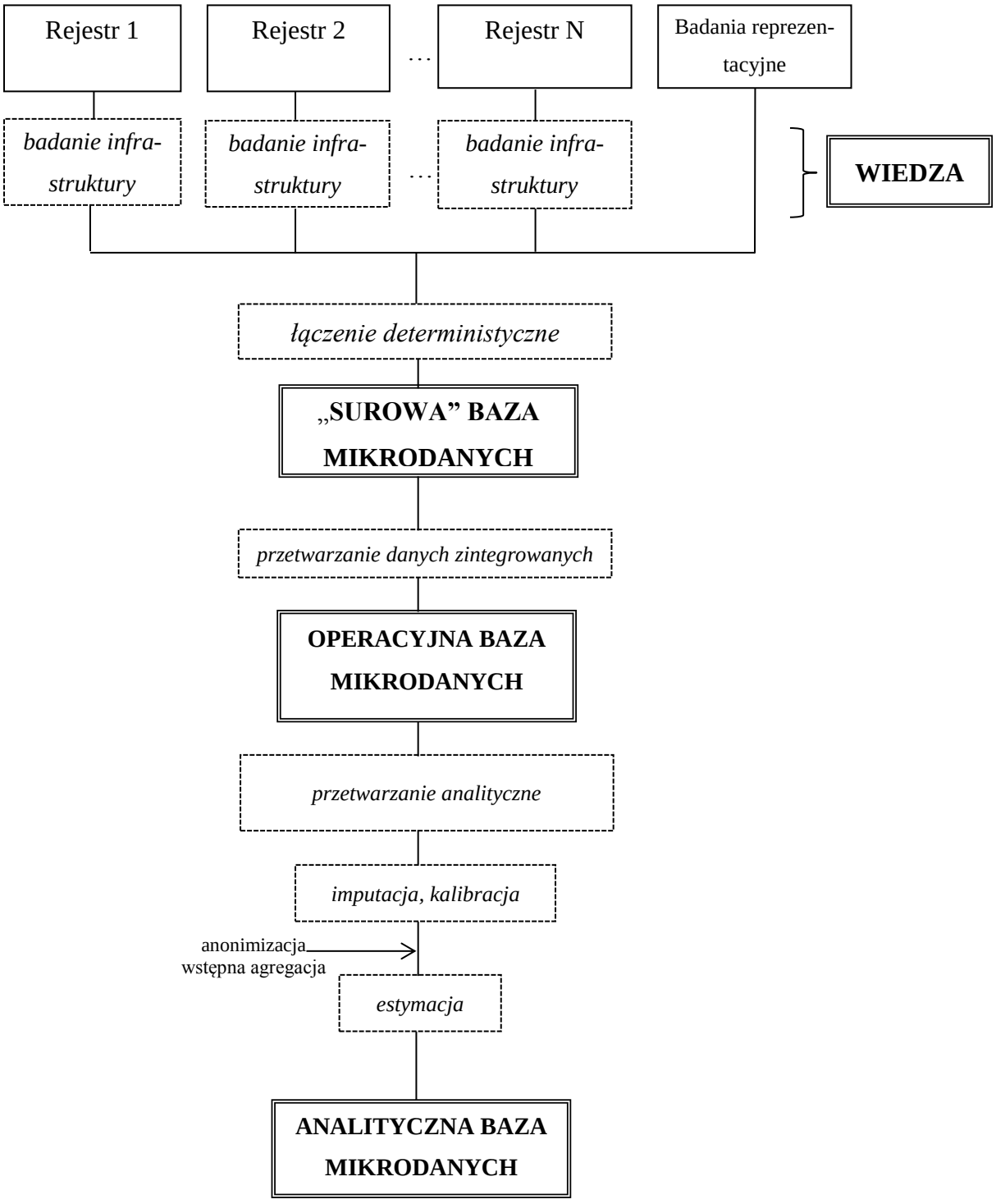
Procedurę integracji repozytoriów danych pochodzących z różnych źródeł dla potrzeb statystyki publicznej przedstawić można w postaci schematu.

Badanie infrastruktury

Pierwszym etapem statystycznej integracji jest badanie infrastruktury dostarczonych od gestorów zbiorów (por. schemat 1.1). Rejestry administracyjne różnią się przede wszystkim ze względu na cel, dla którego zostały utworzone i instytucję zarządzającą. Ponadto odmienność rejestrów wynika ze sposobu przechowywania danych (różne architektury baz danych [Dygaszewicz 2010]), niezgodności momentów referencyjnych (dnia, na który rejestr jest aktualny), czy różnic w definicjach zmiennych. Dodatkowo objęta rejestrem populacja oraz definicje i warianty uwzględnionych cech mogą odbiegać od przyjętych w systemie statystyki publicznej. Istnieje zatem potrzeba przetworzenia rejestrów administracyjnych w taki sposób, by odpowiadały potrzebom badań statystycznych i mogły zostać włączone do systemu badań statystyki publicznej. Proces ten określa się mianem badania infrastruktury statystycznej [Paradysz 2007] i przeprowadza w następujących etapach:

1. Zebranie informacji o dostępnych zbiorach (w tym określenie regulacji prawnych)
2. Identyfikacja populacji docelowej.
3. Sporządzenie formularza metadanych.

Schemat 1.1. Integracja repozytoriów danych pochodzących z różnych źródeł



Źródło: opracowanie własne

Nowozelandzki Urząd Statystyczny [*Data Integration Manual* 2006], korzystając z wieloletniego doświadczenia w integracji, zaproponował trzyetapową procedurę zbierania informacji o rejestrach przeznaczonych do integracji:

1. Zestawienie istniejącej wiedzy wewnętrznej

   W przypadku, gdy te same lub podobne zbiory były używane wcześniej, wiedza i doświadczenie pracowników biorących udział przy ich łączeniu może być bardzo pomocna. Zebrana jest ona zwykle w postaci dokumentów lub opisu stosowanych procedur. Wraz ze wzrostem liczby integrowanych zbiorów, wypracowanie wewnętrznego systemu gromadzenia wiedzy i dzielenia się nią prowadzi do zwiększenia efektywności pracy nad kolejnymi projektami.

2. Przegląd ogólnie dostępnych informacji

   Bardzo często potrzebne i użyteczne informacje na temat integrowanych zbiorów znajdują się na stronach internetowych poszczególnych gestorów. W wielu przypadkach również dostępne są bardziej szczegółowe informacje w postaci tzw. „często zadawanych pytań" (*Frequently Asked Questions, FAQ*) lub udostępnionych słowników danych[10].

3. Spotkanie z dostawcami danych

   Spotkanie z dostawcami danych umożliwia efektywny transfer wiedzy od osób, które na co dzień pracują z danymi repozytoriami. Podczas takich spotkań zadawane pytania oraz przekazywane dokumenty bardzo często w szybki sposób rozwiewają wątpliwości i problemy.

Ważnym zagadnieniem jest również określenie regulacji prawnych umożliwiających wykorzystanie rejestrów w statystyce publicznej. Art. 13 pkt. 4 ustawy z dnia 29 czerwca 1995r. o statystyce publicznej nakłada na gestorów rejestrów administracyjnych obowiązek nieodpłatnego przekazywania danych, w tym całych zbiorów, w zakresie, formie i terminach określanych w programie badań statystycznych GUS. Jednak dopiero ustawa o narodowym spisie powszechnym ludności i mieszkań w 2011 roku (art. 8) umożliwia utworzenie tzw. Bazy Danych NSP 2011 złożonej ze zintegrowanych administracyjnych źródeł informacji. Wykorzystanie połączonych rejestrów w okresach międzyspisowych wymaga dodatkowych aktów prawnych, w tym odpowiednich zapisów o ochronie danych osobowych w zintegrowanych bazach danych, a także każdorazową zgodę na integrację.

[10] Centralny element systemu zarządzania bazą danych, w którym przechowuje się m.in. opisy relacji i perspektyw, deklaracje kluczy głównych, grup użytkowników i uprawnień, informacje o indeksach, plikach i ich strukturach [Abramowicz *et al.* 2007].

Ważnym punktem zbierania informacji o danych źródłowych jest identyfikacja populacji docelowych w obu zbiorach. Istnieje bowiem możliwość różnego pokrycia na pozór zbieżnej w każdym ze źródeł populacji celu. Każde repozytorium danych odnosi się do pewnej populacji docelowej, teoretycznej, o której informacje powinno gromadzić

Zweryfikować należy również zgodność definicji jednostek w obu zbiorach (np. czy zbiory zawierają dane o obywatelach, gospodarstwach domowych, czy przedsiębiorstwach). Także warianty cech w bazach źródłowych mogą się różnić od wariantów w zintegrowanym zbiorze, np. zmiennej „stan cywilny" w jednej bazie może być kategoria „żonaty/zamężna", zaś w drugiej oba te warianty mogą występować oddzielnie. Przy dużej liczbie zmiennych, przekodowanie wszystkich tak, by zawierały dokładnie te same warianty może okazać się czasochłonnym, ale koniecznym procesem.

Ostatnią częścią etapu zbierania informacji o danych źródłowych jest sporządzenie formularza zawierającego „dane o danych" – tzw. metadanych („informacji o danych") [*Data Integration Manual* 2006]. Jest to ważny element, zapewniający odpowiednią jakość przyszłych wyników. Jakość danych rozumiana jest sześciowymiarowo, poprzez: przydatność, dokładność, terminowość, dostępność, interpretowalność oraz spójność. Wymienione wymiary jakości mogą być użyte jako kryterium oceny wszystkich szacunków uzyskanych na podstawie zintegrowanych źródeł danych, a odnoszą się zarówno do zbiorów zintegrowanych, jak i źródłowych.

Efektem etapu badania infrastruktury rejestrów jest wiedza umożliwiająca przeprowadzenie integracji, edycji danych i w końcowym etapie utworzenia spójnych i rzetelnych komunikatów statystycznych.

Gill [2001] oszacował, że czas potrzebny na przeprowadzenie całego procesu integracji można podzielić w następujący sposób:

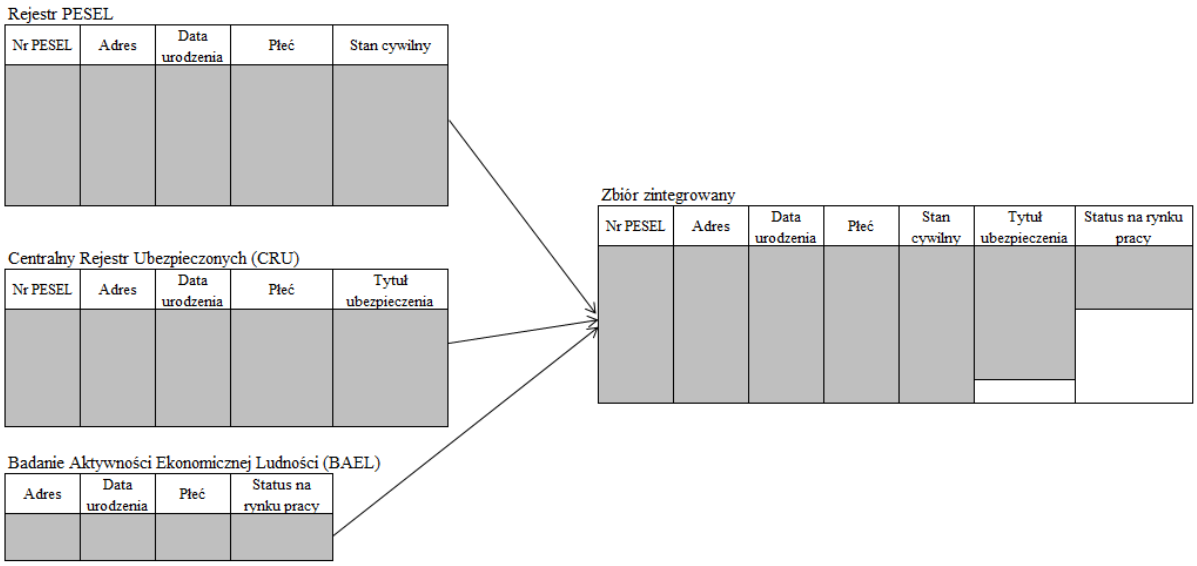
- 75% czasu pracy – przygotowanie zbiorów,
- 5% czasu pracy – przeprowadzenie łączenia rekordów,
- 20% czasu pracy – sprawdzanie poprawności wyników.

Przygotowanie zbiorów do integracji oraz weryfikacja poprawności połączenia to praca w dużej mierze manualna, wymagająca odpowiedniej wiedzy dotyczącej zarówno algorytmu integracji, jak również w dziedzinie której dotyczą integrowane zbiory (np. definicje zmiennych, specyfika populacji itp.). Przeprowadzanie łączenia zwykle wykonywane jest za pomocą odpowiedniego oprogramowania.

Łączenie deterministyczne

Następnym etapem jest łączenie rejestrów na podstawie unikatowych identyfikatorów jednostek (tzw. łączenie deterministyczne)[11]. W źródłach administracyjnych występują zwykle takie zmienne jak, np. numer PESEL dla osób, lub REGON dla podmiotów gospodarczych. W badaniach reprezentacyjnych jednostkę można zidentyfikować na podstawie zestawu takich zmiennych, jak wiek, płeć i adres zamieszkania[12]. Są to tzw. unikalne klucze połączeniowe. Na ich podstawie możliwe jest jednoznaczne (deterministyczne) wskazanie, które rekordy w łączonych zbiorach dotyczą tej samej jednostki.

Schemat 1.2. Integracja zbiorów danych pochodzących z różnych źródeł



Uwaga, kolor:
szary – informacje obserwowane,
biały – informacje nieobserwowane.
Źródło: opracowanie własne

Jak pokazano na schemacie 1.2, do rejestru PESEL, zawierającego m.in. dane demograficzne dodano, stosując numer PESEL jako unikalny klucz połączeniowy, informacje z Centralnego Rejestru Ubezpieczonych (CRU). W ten sposób uzyskano łączną informację na temat stanu cywilnego (zmienna obserwowana w zbiorze PESEL) i tytułu ubezpieczenia społecznego (zmienna obserwowana w CRU). Następnie, określając jako klucz kombinację wartości

[11] Problematyka deterministycznego łączenia rekordów pozostaje poza głównym tematem rozprawy. Omawiając statystyczną integrację danych, nie sposób jednak pominąć łączenia na podstawie unikatowych identyfikatorów.
[12] Adres zamieszkania nie jest przedmiotem pomiaru w badaniach próbkowych, jednak zazwyczaj jest on znany z operatu losowania.

zmiennych płeć, wiek oraz adres zamieszkania, dodano informację o statusie na rynku pracy (zmienna obserwowana w BAEL). Zbiory CRU i BAEL charakteryzują się mniejszym pokryciem niż PESEL[13], stąd łączna obserwacja cech (kolor szary) możliwa jest tylko dla tych jednostek, które występowały w każdym ze zbiorów. Dla pozostałych jednostek występują braki danych (kolor biały). Zintegrowany w ten sposób zbiór danych może być bogatym i tanim (nie ma potrzeby pomiaru statystycznego – dane już istnieją) źródłem informacji społeczno-gospodarczych.

Efektem zastosowania łączenia deterministycznego jest surowa baza danych zawierająca informacje o wszystkich zmiennych z łączonych baz. Są to zmienne niezharmonizowane, tj. o różnych definicjach, wariantach, momentach referencyjnych zgodnych z obowiązującymi w bazach wejściowych.

Przetwarzanie danych zintegrowanych

Połączenie zbiorów jest dopiero początkowym krokiem do pełnej integracji statystycznej. W celu uzyskania zbioru danych odpowiadającego celom statystycznym należy ujednolicić zawarte w nich informacje. Ten etap w literaturze nosi nazwę przetwarzania danych zintegrowanych (*micro-integration processing*). Wallgren, Wallgren [2007] oraz Linder [2004] wymieniają następujące etapy przetwarzania zintegrowanych rejestrów:

- kodowanie zmiennych – ujednolicanie wariantów cech,
- edycja braków danych – imputacja,
- wyrównywanie momentów lub okresów referencyjnych rejestrów – w celu zapewnienia synchronizacji w czasie[14],
- tworzenie jednostek pochodnych (np. gospodarstw domowych złożonych z osób mieszkających pod jednym adresem),
- tworzenie zmiennych pochodnych (np. utworzenie zmiennej „dochody całkowite" będącej sumą dochodów z różnych źródeł),
- porównywanie zmiennych z różnych źródeł w celu korekty błędów.

Etap ten zostanie szerzej omówiony w punkcie 1.5 „Rzetelność zintegrowanych danych".

[13] Np. do BAEL w 2011 roku wylosowano łącznie około 75 tys. gospodarstw domowych z około 13-milionowej populacji, co daje pokrycie rzędu ok. 0,6%.

[14] Przykładem rozwiązania tego problemu na potrzeby NSP jest praca Bijaka [2009].

Operacyjna baza mikrodanych

Etap przetwarzania danych zintegrowanych umożliwia utworzenie operacyjnej bazy mikrodanych[15], w której znajdują się cechy o ujednoliconym charakterze, zdefiniowane według norm przyjętych w statystyce publicznej oraz o zweryfikowanej jakości. Przetwarzanie danych zintegrowanych jest stałym procesem, który zapewnia spójność i wysoką jakość informacji statystycznej. Bakker [2010] wskazuje cztery podstawowe zalety przetworzonej, operacyjnej bazy danych:

- rzetelność i wiarygodność komunikatów statystycznych sporządzonych na podstawie zintegrowanych źródeł jest poprawiona (w porównaniu do bazy „surowej"),
- możliwa jest publikacja szacunków na niskim, niedostępnym dla badań reprezentacyjnych, poziomie agregacji przestrzennej i merytorycznej,
- zmienne z różnych źródeł są połączone i możliwa jest ich łączna obserwacja,
- możliwe jest przeprowadzenie badań panelowych.

Rejestry administracyjne zawierają informacje o bardzo dużej liczbie jednostek, całej zdefiniowanej populacji. Natomiast badania reprezentacyjne dotyczą tylko jej części określonej przez próbę losową. Stąd też zintegrowana, operacyjna baza danych będzie zawierać pełną informację wyłącznie dla jednostek, które wystąpiły w każdym z integrowanych źródeł. Rekordy, które wystąpiły tylko w pojedynczych źródłach, dla cech dołączonych z innych źródeł będą charakteryzować się brakami danych.

Przetwarzanie analityczne i analityczna baza mikrodanych

Ostatnim etapem integracji danych jest proces przetwarzania analitycznego[16] mikrodanych. Polega on na imputacji braków danych, kalibracji wag analitycznych oraz estymacji i publikacji finalnych komunikatów statystycznych. Celem procesu przetwarzania analitycznego jest zachowanie spójności numerycznej danych w sensie uzyskania takich samych wyników dla wszystkich oszacowań bez względu na źródło pochodzenia zmiennych w zintegrowanym zbiorze. Końcowym efektem integracji jest analityczna baza mikrodanych, na podstawie której opracowywane są finalne komunikaty statystyczne.

1.3. Zastosowanie metod statystycznych w integracji danych

Przedmiotem tej rozprawy są metody niedeterministryczne. Ich zastosowanie jest często konieczne w przypadkach, gdy klucz połączeniowy nie jest dostępny lub zbiory danych nie za-

[15] Noszącej również nazwę rejestru statystycznego [Wallgren, Wallgren 2007].
[16] Zostanie on szerzej opisany w sekcji „Rzetelność i spójność zintegrowanych danych".

wierają informacji o tych samych jednostkach. Łączenie deterministyczne nie jest w takich sytuacjach możliwe. Wówczas integracja danych przeprowadzona może być poprzez wykorzystanie metod stochastycznych. Sytuacja taka dotyczy np. łączenia zbiorów pochodzących z badań reprezentacyjnych. Zastosowanie metod stochastycznych nie jest jeszcze szeroko rozpowszechnione, jednak dla poprawy rzetelności i jakości integracji niezbędny jest ich dalszy rozwój .

Integracja administracyjnych zbiorów danych na potrzeby spisów powszechnych przebiega w sposób deterministyczny, tj. każda jednostka w każdym zbiorze jest identyfikowana przez unikalny klucz połączeniowy wspólny dla wszystkich źródeł. Podejście takie gwarantuje, że rekordy w zintegrowanym zbiorze dotyczą konkretnej jednostki rzeczywistej - osoby[17], którą opisuje wiele cech pozyskanych z różnych źródeł.

Pomimo wysokiej jakości rejestrów administracyjnych może zaistnieć sytuacja, gdy unikalny klucz połączeniowy nie będzie dostępny (np. usunięty ze względu na ochronę danych osobowych), zmienna kluczowa będzie zawierała braki danych lub będą w niej występować nieścisłości (np. błędnie wprowadzony numer PESEL). Takie sytuacje zdarzają się rzadko (jakość zmiennych kluczowych podlega szczególnej kontroli), jednak są możliwe[18]. Jeżeli frakcja rekordów z błędnym numerem identyfikacyjnym jest bardzo mała i błędy występują losowo, takie przypadki można pominąć bez większej szkody dla późniejszych szacunków. Błędy jednak mogą powstawać nielosowo, np. w jednej gminie, gdzie narzędzia kontroli okazały się nieskuteczne. W procesie integracji wiele rekordów dotyczących konkretnych jednostek może zostać niepołączonych. Wówczas szacunki tworzone dla jednostki terytorialnej, w której nie udało się zintegrować części rekordów mogą być obciążone.

Niemożność dołączenia jednostek do zintegrowanego zbiorów danych może wystąpić również w sytuacji, gdy rejestr administracyjny łączony jest z danymi badania reprezentacyjnego. Cechy takie jak numer PESEL, czy NIP, będące potencjalnymi zmiennymi kluczowymi, nie są przedmiotem pomiaru w badaniach częściowych. Utworzenie złożonego klucza połączeniowego, w którego skład wchodzą zmienne typu płeć, data urodzenia, czy adres (jak to miało miejsce w holenderskim spisie wirtualnym w 2001 roku) może być niewystarczające[19] (np. bliźnięta tej samej płci mogą dzielić wszystkie te charakterystyki).

[17] Jednostką taką może być także mieszkanie lub gospodarstwo domowe.

[18] W zbiorze administracyjnym Narodowego Funduszu Zdrowia blisko 67 tys. rekordów nie posiadało informacji o numerze PESEL, a 20,5 tys. osób miało go błędnie wpisany (informacja ta pochodzi z badań przeprowadzonych w Ośrodku Statystyki Małych Obszarów Urzędu Statystycznego w Poznaniu).

[19] W spisie wirtualnym w Holandii nie udało się połączyć około 3 procent jednostek zawartych w badaniu aktywności ekonomicznej (LFS).

Propozycją rozwiązań problemu braku niektórych (lub wszystkich) wartości unikalnego klucza połączeniowego w integrowanych repozytoriach są metody statystycznej integracji danych. Metodologia statystycznej integracji danych polega na łączeniu dwóch (lub więcej) źródeł danych (lub ich części) nie posiadających unikatowego klucza połączeniowego na podstawie zestawu tzw. zmiennych wspólnych. Są to cechy, które występują w obu zbiorach, charakteryzują się taką samą (lub bardzo zbliżoną) definicją oraz zgodnością pod względem wariantów cech.

W zależności od podejścia metodologicznego, integrowane zbiory mogą zawierać informacje o tych samych jednostkach lub nie. Zestaw zmiennych wspólnych nie zawsze w pełni identyfikuje jednostki, jednak na podstawie specjalnie określonych kryteriów, np. podobieństwa par rekordów, można z dużym prawdopodobieństwem wskazać te same jednostki lub jednostki do siebie bardzo podobne.

Metody statystycznej integracji danych rozwijane są od lat 60-tych XX wieku [Anderson 1957], jednak ich szersze zastosowanie nastąpiło dopiero na początku XXI wieku [Raessler 2002, D'Orazio *et al.* 2006]. Nagły rozwój metodologii statystycznej integracji danych wynika głównie z upowszechnienia się rozwiązań informatycznych takich jak bazy i hurtownie danych [Winkler 2005]. Duże znaczenie w tym względzie miała tzw. „rewolucja cyfrowa" objawiająca się intensywnym rozwojem mocy obliczeniowej komputerów [van der Putten *et al.* 2002] oraz umasowienie dostępu do Internetu. Dzięki tym czynnikom poprawiła się jakość danych i czas przetwarzania informacji [Barr, Turner 1981]. Dodatkowo w czasach tzw. „kryzysu gospodarczego" i nacisku na ograniczanie wydatków pojawiła się potrzeba wykorzystania już dostępnej informacji do poszerzania wiedzy o różnych zjawiskach społeczno-gospodarczych. „Rewolucja cyfrowa" spowodowała również zjawisko „zalewu informacji" o różnej, często wątpliwej jakości. Stąd też pojawiła się potrzeba wypracowania metod, które zapewniłyby rzetelność i spójność danych będących podstawą decyzji administracyjnych i biznesowych [Cohen 1991].

W literaturze zasadniczo wyróżnia się dwie metody integracji danych: deterministyczną i stochastyczną. Podstawą łączenia deterministycznego jest identyczność wybranych pól w łączonych rekordach, tzw. klucz identyfikacyjny[20]. W metodzie stochastycznej wyróżnia się dwa główne nurty:

1. Probabilistyczne łączenie rekordów (*probabilsitic record linkage*),

[20] Łączenie deterministyczne jest podstawową metodą integracji źródeł administracyjnych.

2. Parowanie statystyczne[21] (*statistical matching, data fusion, data merging, data matching, mass imputation, file concatenation*).

Probabilistyczne łączenie rekordów

W metodzie probabilistycznego łączenia rekordów zakłada się, że łączone repozytoria danych zawierają informacje o tych samych jednostkach[22]. Ponieważ żaden z łączonych zbiorów nie zawiera unikatowego klucza, należy znaleźć cechy (pola rekordu), które ten klucz mogą utworzyć. Przykładem tworzenia klucza jest odnajdywanie podobieństwa w ciągu znaków w rekordach zawierających imię, nazwisko, datę i miejsce urodzenia, wiek, płeć, itp.

Schemat 1.3. Wyszukiwanie rekordów dotyczących tej samej jednostki

ID	Nazwa	Adres	Nr telefonu
1432	Świnka sp. j.	ul. Mickiewicza 1a	+22 7456969
1433	Rowerek	Trzeciego Maja 15	591987321
1434	Bucik sp. z o.o.	al. Niepodległości 10	581596325

ID_rec	Nazwa	Adres	Nr telefonu
D1215	Rowerk	3 Maja 15	12 591-987-321
D1354	Świnka	Mickiewicza 1	745-69-69
D1236	Bucik	Niepodległości 10/1	14 581596325

Źródło: opracowanie własne na podstawie [Fortini *et al.* 2006]

W integrowanych zbiorach, ze względu na brak kompatybilności, często te same cechy inaczej się nazywają, a ich warianty czy wartości zapisywane są w niejednolity sposób (np. adresy, numery telefonów, nazwy własne itp., por. schemat 1.3). Fakt ten może wynikać zarówno z przyjętych przez gestorów odmiennych standardów zapisu czy obowiązujących regulacji administracyjnych lub po prostu z różnych błędów (np. ortograficznych, typograficznych, wynikających z niedoskonałości sprzętu i oprogramowania skanującego itp.). Metody probabilistycznego łączenia rekordów umożliwiają połączenie rekordów, które choć różnią się sposobem zapisu należą do tej samej jednostki rzeczywistej. Integracja odbywa się poprzez porównanie wartości zmiennych występujących w obu zbiorach.

[21] Polskie tłumaczenie – „parowanie statystyczne" jest przedmiotem dyskusji. W niniejszej pracy wybrano to określenie ze względu na fakt, że w literaturze najczęściej wykorzystywane jest podejście łączenia w pary rekordów najbardziej do siebie podobnych (pod względem wybranych charakterystyk).

[22] Jest to podstawowe założenie w integracji administracyjnych baz danych.

Najczęściej w literaturze przedstawia się probabilistyczne łączenie rekordów jako proces kilkustopniowy. Pierwszym krokiem jest zebranie informacji o danych źródłowych oraz wybór zmiennych, na podstawie których przeprowadzone zostanie łączenie (wybór tzw. zmiennych parujących). W kolejnym kroku przygotowuje się zbiory do procesu integracji poprzez usunięcie duplikatów oraz standaryzację wariantów cech parujących. Następnie dokonuje się operacji grupowania (nazywanej również blokowaniem) mającej na celu podział na podzbiory, w ramach których nastąpi łączenie rekordów (np. integracja mieszkańców jednego powiatu lub podmiotów jednej gałęzi przemysłu). Grupowania dokonuje się z w celu optymalizacji algorytmu integracji poprzez zredukowanie liczby porównań par rekordów. Następnie na podstawie różnych algorytmów[23] przeprowadza się integrację oraz sprawdzenie efektywności połączenia.

Głównym zadaniem metody probabilistycznego łączenia rekordów jest ustalenie, czy para rekordów należy do tego samej jednostki czy nie. Decyzję tę podejmuje się najczęściej na podstawie oceny prawdopodobieństwa (lub jego przekształceń), że dana para rekordów należy do tej samej jednostki [Blakely, Salmond 2002; Fellegi, Sunter 1969]. W rzeczywistości jednak nie jest możliwym dokładne wskazanie, które pary rekordów zawierają informacje o tym samym podmiocie, a które z całą pewnością nie zawierają. Zamiast tego możliwa jest obserwacja par zaklasyfikowanych jako prawdopodobne połączenie i niepołączenie za pomocą tzw. wag połączeniowych obliczanych na podstawie przekształcenia prawdopodobieństwa, że dana para rekordów należy do tej samej jednostki przy zgodności wartości wszystkich zmiennych wspólnych [Winkler 2005].

Parowanie statystyczne

Parowanie statystyczne to grupa metod służących do integracji dwóch (lub więcej) źródeł danych (zwykle pochodzących z badań próbkowych) odnoszących się do tej samej populacji generalnej. Ponieważ prawdopodobieństwo wylosowania tej samej jednostki do dwóch różnych badań reprezentacyjnych jest bardzo małe (zbliżone do zera), zakłada się, że integrowane zbiory są rozłączne, tzn. nie zawierają informacji o tych samych jednostkach rzeczywistych.

[23] Zostaną one szczegółowo opisane w rozdziale IV.

Schemat 1.4. Parowanie statystyczne, struktura integrowanych zbiorów

Zbiór A

$\boldsymbol{Y_1}$	…	$\boldsymbol{Y_Q}$	$\boldsymbol{X_1}$	…	$\boldsymbol{X_P}$
y_{11}^A	…	y_{1Q}^A	x_{11}^A	…	x_{1P}^A
…	…	…	…	…	…
y_{a1}^A	…	y_{aQ}^A	x_{a1}^A	…	x_{aP}^A
…	…	…	…	…	…
$y_{n_A1}^A$	…	$y_{n_AQ}^A$	$x_{n_A1}^A$	…	$x_{n_AP}^A$

Zbiór B

$\boldsymbol{X_1}$	…	$\boldsymbol{X_P}$	$\boldsymbol{Z_1}$	…	$\boldsymbol{Z_R}$
x_{11}^B	…	x_{1P}^B	z_{11}^B	…	z_{1R}^B
…	…	…	…	…	…
x_{b1}^B	…	x_{bP}^B	z_{b1}^B	…	z_{bR}^B
…	…	…	…	…	…
x_{n1}^B	…	$x_{n_BP}^B$	$z_{n_B1}^B$	…	$z_{n_BR}^B$

Uwaga:
zmienne $\boldsymbol{X}$ – zmienne wspólne
zmienne $\boldsymbol{Y}$ – zmienne obserwowane wyłącznie w zbiorze A
zmienne $\boldsymbol{Z}$ – zmienne obserwowane wyłącznie w zbiorze B
Źródło: opracowanie własne

W każdym zbiorze (oznaczonym jako A i B) znajduje się zwykle pewien wektor (o liczebności P) identycznych zmiennych (w badaniach społecznych mogą to być np. zmienne demograficzne) o tych samych lub zbliżonych definicjach i wariantach. Nazywa się je zmiennymi wspólnymi i oznacza jako $\boldsymbol{X}$. Zbiór A (o liczebności n_A) zawiera także wektor zmiennych $\boldsymbol{Y}$ (o liczebności 1 do Q), które są obserwowane wyłącznie w tym zbiorze. Podobnie w zbiorze B (o liczebności n_B) występuje analogiczny wektor – $\boldsymbol{Z}$ nieobserwowanych w zbiorze A (o liczebności R; por. schemat 1.4). Celem parowania statystycznego jest analiza związków pomiędzy zmiennymi $\boldsymbol{Y}$ i $\boldsymbol{Z}$ nieobserwowanymi łącznie w pojedynczym źródle.

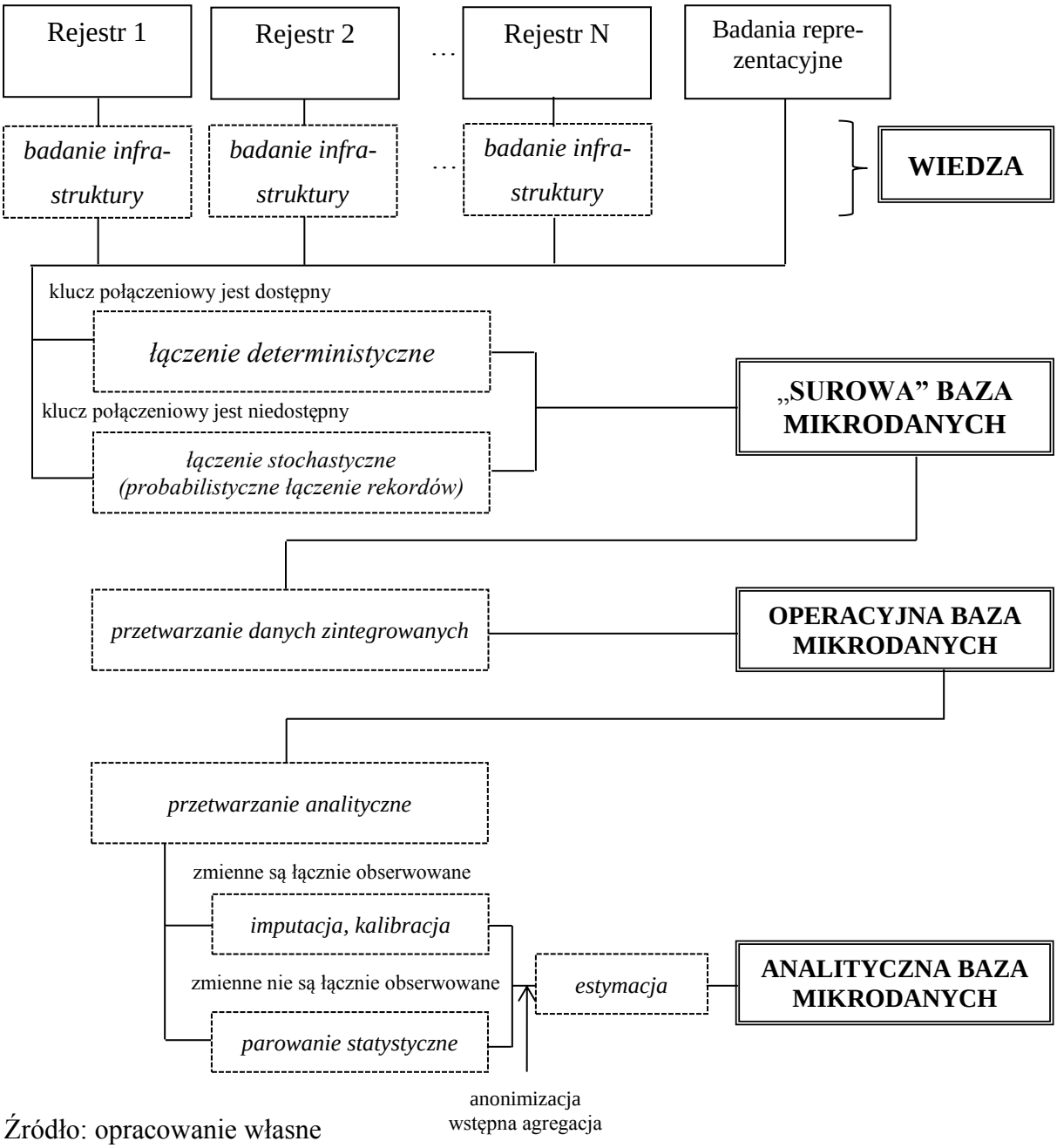
Produktem integracji danych metodą parowania statystycznego są jednostki syntetyczne. Oznacza to, że w zintegrowanym zbiorze obserwacji podlegają podmioty niewystępujące w rzeczywistości. U źródeł koncepcji tworzenia zbiorowości hipotetycznych, nierzeczywistych jest przypuszczenie, że jednostki, które są do siebie podobne pod względem określonych cech (np. demograficznych, jak wiek, płeć, miejsce zamieszkania, czy wykształcenie i ekonomicznych, jak aktywność ekonomiczna, źródło utrzymania, czy dochody) będą również podobne pod względem innych cech będących przedmiotem analiz.

Dzięki zastosowaniu metody parowania statystycznego możliwa jest łączna obserwacja cech nieobserwowanych wspólnie w żadnym ze źródeł. Umożliwia to dokonywanie analiz wielo-

wymiarowych, jak np. badanie współzależności (wyznaczenie współczynnika korelacji, czy też utworzenie tabeli kontyngencji).

Statystyczne metody integracji danych mogą pełnić funkcję wspomagającą w projektowaniu systemu statystyki publicznej opartej na zintegrowanych zbiorach danych. Dzięki tej metodologii możliwe jest łączenie rekordów nie posiadających unikalnego klucza połączeniowego (lub posiadających wartość błędną) – za pomocą metod probabilistycznego łączenia rekordów, jak również zapewnienie łącznej obserwacji cech nieobserwowanych wspólnie w pojedynczym źródle.

Schemat 1.5. Integracja danych z różnych źródeł z wykorzystaniem metod stochastycznych



Źródło: opracowanie własne

Powyższe metody uwzględniono na schemacie 1.1, przestawiając kompleksowo ideę statystycznej integracji danych (por. schemat 1.5). W przypadku niedostępności klucza połączeniowego, przy tworzeniu „surowej" bazy danych można zastosować metodę probabilistycznego łączenia rekordów (wymagającą harmonizacji zmiennych parujących). Przy przetwarzaniu analitycznym zintegrowanych i zharmonizowanych zbiorów, gdy integrowane zmienne nie są obserwowane łącznie w żadnym ze źródeł stosuje się parowanie statystyczne.

1.4. Spójność zintegrowanych danych

Estymacja jest ostatnim etapem tworzenia analitycznej bazy mikrodanych (por. schemat 1.5). W celu utworzenia zgodnych, rzetelnych i spójnych szacunków często trzeba się zmierzyć z szeregiem problemów.

Stopień pokrycia badanej populacji przez integrowane zbiory jest bardzo różny. Rejestry administracyjne zawierają informacje o bardzo dużej liczbie jednostek, natomiast badania reprezentacyjne charakteryzują się niewielkim pokryciem. Stąd też zintegrowana, operacyjna baza danych będzie zawierać pełną informację wyłącznie dla jednostek, które wystąpiły w każdym z integrowanych źródeł (kolor czarny, por. schemat 1.6). Rekordy, które wystąpiły tylko w jednym ze źródeł, dla cech dołączonych z innych zbiorów będą charakteryzować się brakami danych (kolor biały).

Schemat 1.6. Struktura operacyjnej bazy danych dotyczących zatrudnienia

Status na rynku pracy	Źródło danych		
	Ewidencja ludności	Rejestr zatrudnienia	Badanie aktywności ekonomicznej
pracujący	■	■	□
			■
			□
bezrobotni	■	□	□
			■
bierni zawodowo	■	□	□
			■
			□

Uwaga, kolor:
czarny – informacje obserwowane,
biały – informacje nieobserwowane.
W Polsce nie istnieje rejestr zatrudnienia. Przykład opiera się na doświadczeniach holenderskich.
Źródło: opracowanie własne na podstawie [Linder 2004]

Przykładowo, informacje z ewidencji ludności charakteryzują się pełnym pokryciem dla wszystkich osób, niezależnie od ich statusu na rynku pracy. Są to jednak zwykle dane demograficzne (płeć, wiek, miejsce zamieszkania). Stąd, by uzyskać informację o aktywności ekonomicznej ludności konieczne jest dołączenie rekordów z rejestru zatrudnienia i badania aktywności ekonomicznej ludności. Rejestr zatrudnienia zawiera pełną informację o ludności pracującej, natomiast osoby bezrobotne i bierne zawodowo (np. uczące się) nie są w nim ujęte. Informację o bezrobotnych i biernych zawodowo zawiera badanie aktywności ekonomicznej, obejmuje ono jednak zazwyczaj niewielką część badanej populacji (próbę losową).

Niezwykle istotnym zatem etapem jest odpowiednie dostosowanie mikrodanych z bazy operacyjnej, by możliwe było tworzenie dobrej jakości szacunków o populacji generalnej. Wyróżnia się trzy podstawowe wymagania, które muszą spełniać dane w bazie statystycznej [Kroese *et al.* 2000]:

- rzetelność – szacunki powinny być nieobciążone,
- spójność – dane nie mogą być sprzeczne, np. liczba ludności w ujęciu gmin musi być taka sama, bez względu na to, z której bazy pochodzą zmienne;
- zachowanie tajemnicy statystycznej – brak możliwości identyfikacji jednostek za pomocą publikowanych rezultatów.

Dwa pierwsze wymagania mogą zostać spełnione poprzez zastosowanie jednego z trzech podejść [Kroese *et al.* 2000]: kalibracji, masowej imputacji i mikrosymulacji lub iteracyjnego dopasowania proporcjonalnego.

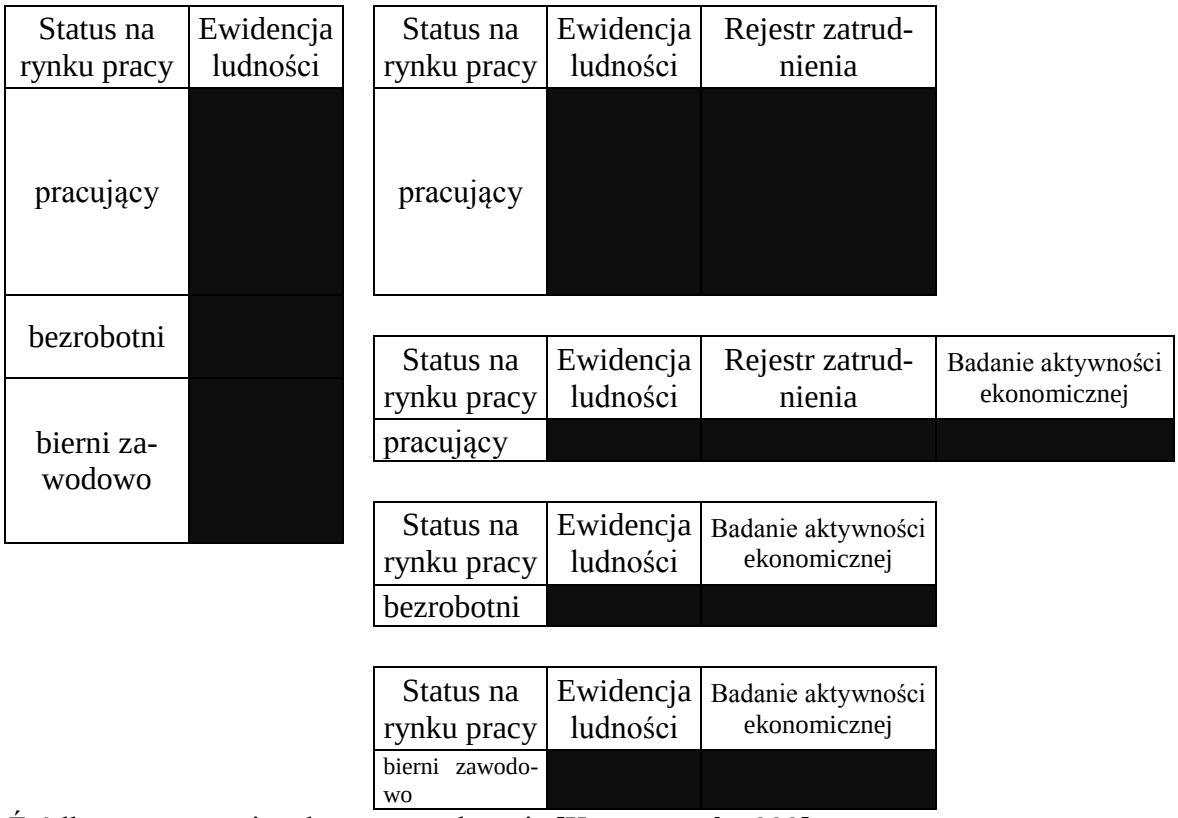
Kalibracja

Operacyjna baza mikrodanych dzielona jest na wszystkie możliwe podzbiory danych według przyjętego kryterium („bloki danych", por. schemat 1.7) $S_1 \cup S_2$. Przykładowo, informacje z ewidencji ludności (najczęściej demograficzne: wiek, płeć, adres zamieszkania itp.) dostępne są dla wszystkich jednostek w populacji i tworzą jeden podzbiór. Informacje z badania aktywności ekonomicznej ludności dla pracujących pokrywają się z informacjami z rejestru zatrudnienia i tworzą kolejny podzbiór danych, itd.

Poszczególnym rekordom w „blokach danych" przyporządkowane są wagi początkowe $d_i^* = \lambda\pi_{1i}^{-1}$ oraz $d_i^* = (1-\lambda)\pi_{2i}^{-1}$. Wagi początkowe następnie są kalibrowane w sposób zapewniający zgodność szacunków, minimalizujący błędy, a także niwelujący skutki braków odpowiedzi. Algorytm kalibracji wykorzystuje liniową kombinację wag pierwotnych:

$d_i^* = \lambda\pi_{1i}^{-1} + (1-\lambda)\pi_{2i}^{-1}$. Dla każdego „bloku danych" definiowany jest indywidualny zestaw wag.

Schemat 1.7. Podzbiory wyznaczone na podstawie zintegrowanej bazy danych



Źródło: opracowanie własne na podstawie [Kroese *et al.* 2000]

Kalibracja przeprowadzana dla każdego z bloków danych oddzielnie może prowadzić do uzyskiwania innych wartości dla tej samej cechy. Taka sytuacja stoi w sprzeczności z potrzebą zachowania spójności danych. W holenderskim spisie wirtualnym zastosowano więc nowatorską metodę wielokrotnego ważenia [Kroese *et al.* 2000, Gouweleeuw, Hartgens 2004]. Polega ona na dostosowaniu wag kalibracyjnych uzyskanych dla wszystkich podzbiorów danych w taki sposób, by uzyskane rezultaty były zgodne dla badanych cech bez względu na źródło. Odbywa się to poprzez rekalibrację wag w odniesieniu do rozkładów brzegowych zmiennych występujących w każdym „bloku". Zapewnia się w ten sposób spójność oszacowań przy jednoczesnym zachowaniu ich rzetelności rozumianej jako nieobciążoność [Zhang 2012].

Masowa imputacja i mikrosymulacja

Różny zakres pokrycia merytorycznego przez integrowane źródła powoduje, że wiele rekordów w połączonym zbiorze charakteryzuje się brakami danych dla niektórych cech (por. schemat 1.6). Zastosowanie metod imputacji w celu uzupełnia tych braków wartościami syntetycznymi zapewnia łączną obserwację wszystkich zmiennych w zintegrowanym zbiorze danych jednostkowych. Gwarantuje również numeryczną spójność szacunków (por. schemat 1.8).

Wadą takiego podejścia jest to, że imputowane wartości są w dużej mierze nierzeczywiste, nawet jeżeli wynikają z dobrze dobranego modelu uzupełniania braków. W wyniku imputacji powstają jednostki i charakterystyki syntetyczne (o nierzeczywistych wartościach cech), co z jednej strony powoduje, że maleje niebezpieczeństwo identyfikacji prawdziwych jednostek, ale z drugiej strony może prowadzić do obniżenia rzetelności danych. Dodatkowo, nawet jeżeli do imputacji stosuje się model skonstruowany w oparciu o wartości empiryczne, to często na podstawie stosunkowo niewielkiej liczby obserwacji szacuje się wielką liczbę wartości teoretycznych (np. imputując aktywność ekonomiczną z BAEL – na podstawie ok. 200 tys. rekordów - do rejestru PESEL – dla ok. 30 mln rekordów[24]).

Schemat 1.8. Zintegrowane repozytorium danych z zaimputowanymi wartościami

Status na rynku pracy	Źródło danych		
	Ewidencja ludności	Rejestr zatrudnienia	Badanie aktywności ekonomicznej
aktywni zawodowo			
bezrobotni			
bierni zawodowo			

Uwaga, kolor:
czarny – informacje obserwowane,
szary – informacje imputowane.

Źródło: opracowanie własne na podstawie [Kroese *et al.* 2000]

[24] Biorąc pod uwagę, że populację w BAEL stanowią osoby w wieku 15 lat i starsze.

Alternatywą dla masowej imputacji może być parowanie statystyczne. W metodach tych kładzie się nacisk, by, o ile to możliwe, dołączane wartości były wartościami empirycznymi[25], jak również dołącza się jedną zmienną jednocześnie dokładając starań, by model integracji zapewniał zgodność wielowymiarowych rozkładów w integrowanych zbiorach [Raessler 2002].

By uniknąć dołączania wielu „sztucznych" wartości, imputacja może być przeprowadzana tylko dla wybranej subpopulacji, dla której liczebność próby w badaniu reprezentacyjnym jest dostatecznie duża (np. dla miast liczących pow. 500 tys. mieszkańców).

Zintegrowany przy wykorzystaniu metod deterministycznych oraz statystycznych zbiór danych może służyć tworzeniu komunikatów statystycznych, jak również organom administracji publicznej dla realizacji różnych celów społecznych (np. kierowanie wyższych dotacji z funduszu rehabilitacji osób niepełnosprawnych do zakładów zatrudniających najwięcej takich ludzi, czy też dotowanie stanowisk pracy na obszarach o najniższych wynagrodzeniach). Zintegrowane zbiory danych, zapewniające wysokie pokrycie i łączną obserwację cech z różnych obszarów funkcjonowania społeczeństwa, gospodarki i państwa są również punktem wyjścia do tworzenia kompleksowych systemów statystycznych zapewniających wsparcie w tworzeniu symulacji i prognoz skuteczności działań organów administracji państwowej (polityki podatkowej, opieki społecznej etc.) i prywatnych (inwestowanie w określone rejony, lokowanie produktów, zdobywanie klientów poprzez precyzyjnie przeprowadzane kampanie marketingowe). Takie wykorzystanie zintegrowanych źródeł nosi nazwę mikrosymulacji (*microsimulation*) i jest wykorzystywane m.in. w Europie [Atkinson *et al.* 1999], Kanadzie [Morrison 1998] oraz Australii [Kelly 2003, Hardling *et al.* 2009].

Podstawą mikrosymulacji jest zbiór danych jednostkowych pochodzący z badania reprezentacyjnego. Może on jednak nie zawierać kompleksowej informacji o różnych charakterystykach bądź próba może okazać się niewystarczająca, co z kolei wpływa na jakość szacunków. Dlatego dąży się do utworzenia przy pomocy statystycznych metod integracji jednego zbioru zawierającego wszystkie zmienne będące przedmiotem badań w przekroju terytorialnym. W literaturze [Rahman 2008] wskazuje się dwa podstawowe sposoby tworzenia przestrzennego syntetycznego zbioru mikrodanych: syntetyczną rekonstrukcję (wykorzystującą parowanie statystyczne) oraz ponowne ważenie (*reweighting*), będące metodą analogiczną do

[25] Np. poszukiwanie tzw. „statystycznych bliźniąt" [Bacher 2002], czyli wyszukiwanie w jednym zbiorze rekordów najbardziej podobnych do tych w drugim zbiorze i ich łączenie [Raessler 2002, D'Orazio *et al.* 2006]. Inną metodą jest tzw. wielokrotna imputacja [Raessler 2002], gdzie dla jednej dołączanej wartości wyznacza się kilka wynikających z modelu lub najbardziej podobnych w innym zbiorze i albo dołącza się rekord wylosowany z tego zbioru „bliźniąt", albo np. imputuje się wartość średnią.

podwójnego ważenia. Syntetyczny zbiór danych jednostkowych zwierający informacje o zmiennych w przekroju małych jednostek terytorialnych (domen) umożliwia zastosowanie modelowania mikrosymulacyjnego (*microsimulation modelling*). Podejście to umożliwia, w szczególności prowadzanie badań symulacyjnych dotyczących różnych podejść w modelowaniu estymatorów SMO. Wyodrębnia się tutaj dwa etapów:

1. Utworzenie systemu wag dla jednostek w przekroju małych domen używając metody przeważania.
2. Zastosowanie nowoutworzonego systemu wag do oszacowania wybranych zmiennych wynikowych w przekroju małych domen.

Każda z powyższych faz składa się z kilku podetapów. Ballas *et al.* [2005] wyróżnia cztery główne kroki związane z modelowaniem symulacyjnym:

1. Utworzenie syntetycznego zbioru danych jednostkowych.
2. Wylosowanie próby z tego zbioru za pomocą metody Monte Carlo w celu utworzenia subpopulacji na poziomie jednostkowym dla badanej małej domeny (obszaru).
3. Przeprowadzenie badań symulacyjnych, których zadaniem jest określenie co się zdarzy w pewnych sytuacjach przy różnych możliwych scenariuszach (tzw. *What-if simulations*).
4. Zastosowanie modelowania dynamicznego (uwzględniającego czynnik czasu) w celu aktualizacji wyjściowego zbioru danych jednostkowych.

Mikrosymulacja może okazać się przydatna w ocenie rozkładu różnych cech społeczno - ekonomicznych, takich jak dochód, podatki i świadczenia socjalne, na niskim poziomie agregacji, podczas gdy dane spisowe (lub z innych źródeł) z tego samego okresu są niedostępne. Ponadto przestrzenne modelowanie mikrosymulacyjne posiada pewne szczególne cechy, wśród których można wyróżnić:

- większą swobodę w wyborze poziomu agregacji przestrzennej,
- utworzenie wieloźródłowej bazy danych jednostkowych zawierającej informacje dla małych obszarów,
- możliwość agregacji i dezagregacji przestrzennych zbiorów mikrodanych,
- możliwość aktualizacji oraz prognozowania.

Iteracyjne dopasowanie proporcjonalne

Celem metody iteracyjnego dopasowania proporcjonalnego (*iterative proportional fitting*, IPF) jest korekta szacunków w tabelach kontyngencji otrzymanych z badań częścio-

wych w stosunku do wyników badań pełnych. Metodę tę często zalicza się do metod statystyki małych obszarów [Simpson, Tranmer 2003; Zhang, Chambers 2004].

Algorytm metody IPF jest stosunkowo mało skomplikowany i polega na rozszacowaniu liczebności komórek w wierszach w oparciu o liczebności brzegowe z badania pełnego. Następnie zaś postępuje się w analogiczny sposób dla kolumn. Iteracyjny charakter metody IPF polega na powtarzaniu tych czynności do momentu, w którym zostanie osiągnięte kryterium konwergencji (zbieżności). Jest ono określane jako spójność sumy liczebności w komórkach tabeli kontyngencji z odpowiednimi liczebnościami brzegowymi z dodatkowego źródła informacji[26].

Algorytm metody iteracyjnego dopasowania proporcjonalnego rozpoczyna się od skonstruowania tabeli kontyngencji dla dwóch cech (A i B). W poszczególnych komórkach tej tabeli znajdują się liczebności cząstkowe pochodzące z bazy danych badania reprezentacyjnego $p_{i,j}$ (por. schemat 1.9), natomiast liczebności brzegowe dla wierszy i kolumn to wartości pochodzące z innego źródła danych (odpowiednio: R_i i C_j). W przypadku niezgodności wariantów cech, należy je ujednolicić , by kategorie badania częściowego, jak i dodatkowego źródła były identyczne.

Schemat 1.9. Dane wejściowe w metodzie IPF

Zmienna A	Zmienna B				
	Wariant 1	Wariant 2	…	Wariant n	SUMA
Wariant 1	$p_{1,1}$	$p_{1,2}$	…	$p_{1,n}$	$R_1 = \sum_j p_{1,j}$
Wariant 2	$p_{2,1}$	$p_{2,2}$	…	$p_{2,n}$	$R_2 = \sum_j p_{2,j}$
…	…	…	…	…	…
Wariant m	$p_{m,1}$	$p_{m,2}$	…	$p_{m,n}$	$R_m = \sum_j p_{m,j}$
SUMA	$C_1 = \sum_i p_{i,1}$	$C_2 = \sum_i p_{i,2}$	…	$C_n = \sum_i p_{i,n}$	$\sum_i R_i = \sum_j C_j$

Źródło: Bijak *et al.* [2007]

Algorytm dopasowania liczebności cząstkowych do liczebności brzegowych składa się z dwóch kroków [Norman 1999]:

1. $p_{i,j}^{2k} = \frac{p_{i,j}^{2k-1}}{\sum_j p_{i,j}^{2k-1}} \times R_i$ (1.1)

[26] Ponieważ algorytm polega na dopasowaniu liczebności do wartości brzegowych dla wierszy, a następnie dla kolumn, często metoda IPF w literaturze określana jest jako *raking* (zgrabianie).

2. $p_{i,j}^{2k+1} = \frac{p_{i,j}^{2k}}{\sum_j p_{i,j}^{2k}} \times C_j$ (1.2)

W kroku pierwszym, wyrażonym wzorem (1.1), wartości komórek wierszowych rozszacowuje się w oparciu o sumy wierszowe R_i. Analogicznie algorytm przebiega w drugim kroku (wzór (1.2)), gdzie wartości obliczone w pierwszym kroku rozszacowuje się dla kolumn w oparciu o sumy kolumnowe C_j. Kolejne iteracje wykonywane są tak długo, dopóki nie zostanie osiągnięte kryterium zbieżności (konwergencji). Zbieżność osiągnięta jest wtedy, gdy zachodzi równość (z założoną wcześniej precyzją, np. 0,1[27]) $\sum_j p_{i,j} = R_i$ oraz $\sum_i p_{i,j} = C_j$ dla wszystkich $p_{i,j}$. Konwergencja nie zostanie osiągnięta, jeżeli wśród liczebności brzegowych występują zera lub jeżeli $\sum_i R_i \neq \sum_j C_j$. Również wśród liczebności cząstkowych mogą wystąpić zera, wówczas komórki takie nie zostaną uwzględnione w procedurze. By uzyskać dla liczebności zerowych wartości wynikowe należy zamienić je na pewną bardzo małą wartość (w porównaniu z innymi), np. 0,001 [Hunsinger 2008]. Bishop *et al.* [1975] sugerują, by liczba zerowych liczebności nie przekraczała 30% wszystkich komórek, jeżeli „zera" rozłożone są w miarę równomiernie w tabeli, lub 10% jeżeli zerowe wartości są „zgrupowane". Należy również zwrócić uwagę, że metoda IPF dostarcza wyników z wartościami ułamkowymi. By otrzymać wyniki w wartościach całkowitych i jednocześnie zachować zbieżność należy oddzielnie przeprowadzić zaokrąglanie.

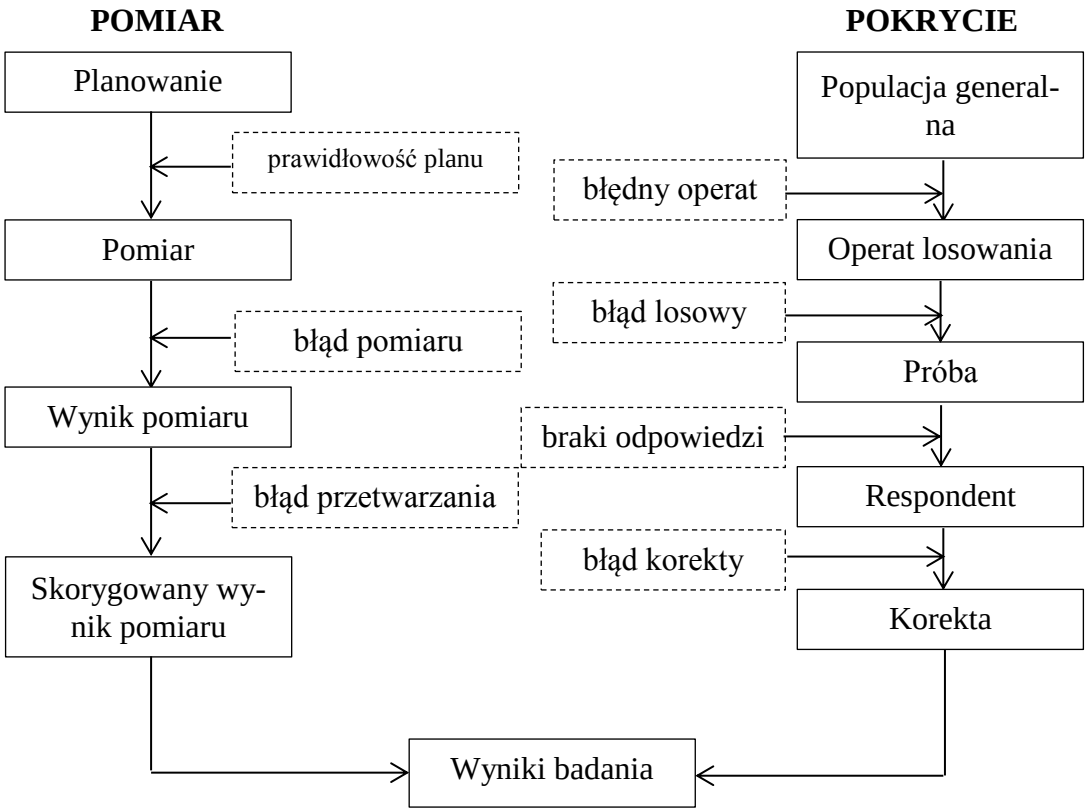
Metoda IPF było szeroko stosowana do szacowania wartości dla małych domen przed rozwinięciem metody wielokrotnego ważenia [Rahman 2008].

1.5. Rzetelność zintegrowanych danych

Jakość danych zintegrowanych wynika z jakości poszczególnych komponentów: badań reprezentacyjnych, rejestrów administracyjnych oraz wynikających z zastosowanych procedur. W zależności od źródła danych, występować mogą różne błędy, które wpływają na jakość zbioru zintegrowanego. Za każdym razem podejmować należy działania minimalizujące prawdopodobieństwo występowania błędów. W niniejszej sekcji opisane zostaną błędy, które mogą się pojawić w badaniach reprezentacyjnych i rejestrach administracyjnych.

[27] W praktyce rzadko się zdarza, by liczebności zsumowały się dokładnie do liczebności brzegowych.

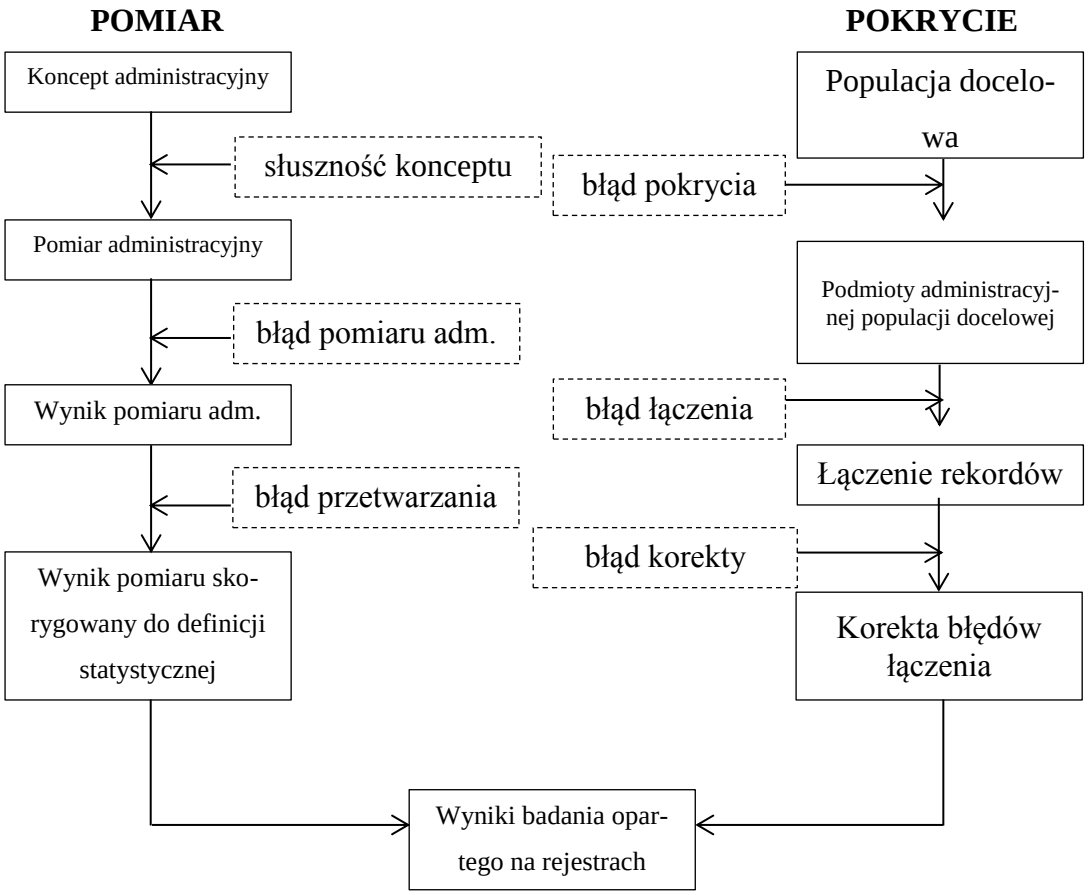
Schemat 1.10. Rachunek błędów w badaniach reprezentacyjnych



Źródło: na podstawie [Groves *et al.* 2004]

Rachunek błędów w badaniach reprezentacyjnych jest dobrze opisany i szeroko wykorzystywany [Kordos 1988]. Groves *et al.* (2004) wprowadza pojęcie całkowitego błędu badania (*total survey error*), który dezagregowany jest na dwa podstawowe komponenty: błąd pomiaru i błąd pokrycia. Błąd pomiaru wynika z zastosowanych definicji mierzonych cech, niedoskonałych metod pomiaru oraz metod przetwarzania danych. Błędy pokrycia są przede wszystkim skutkiem zastosowania błędnego (niepełnego) operatu losowania, błędów losowych, błędów wynikających z braków odpowiedzi i błędów powstałych w skutek imputacji i kalibracji danych (por. schemat 1.10).

Schemat 1.11. Błędy występujące w zintegrowanych repozytoriach danych



Źródło: na podstawie [Bakker 2010]

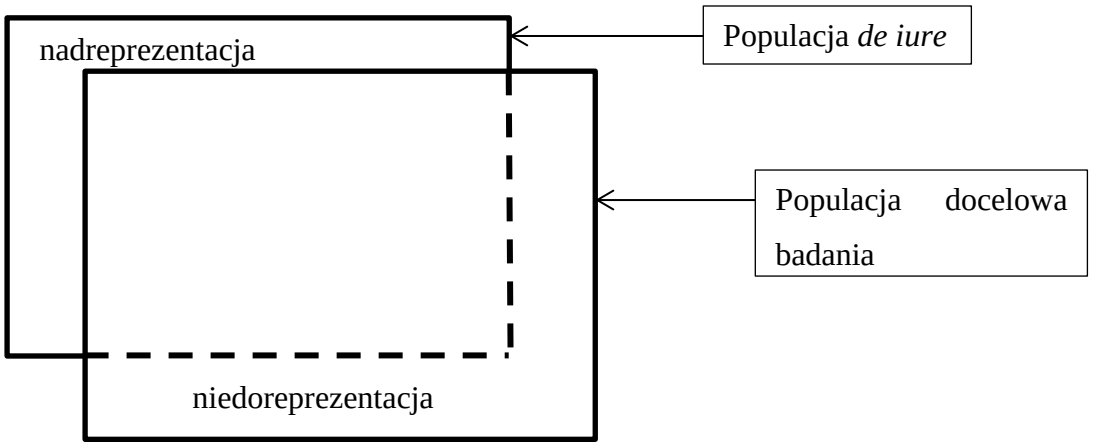
Pomiar w rejestrach administracyjnych rozpoczyna się od określenia teoretycznego przedmiotu i podmiotu badania (por. schemat 1.11). Nosi on nazwę konceptu administracyjnego (*administrative concept*, Bakker 2010). Etap pomiaru wynika z uwarunkowań prawnych i administracyjnych i odbywa się za pomocą ankiet i formularzy, które skonstruowane są w ściśle określony sposób umożliwiający precyzyjne pobranie informacji od podmiotów podlegających danej instytucji. Jeżeli podmiot lub badana cecha ma charakter złożony (np. gospodarstwo domowe lub rodzina w rejestrze pomocy społecznej, gdzie na określenie o podleganiu pomocy państwa składa się bardzo wiele, często niemierzalnych, czynników), prawdopodobieństwo pojawienia się błędów jest wyższe. Stąd też instytucje będące gestorami rejestrów formułują pytania w formularzach w sposób szczególnie precyzyjny[28]. Rozmiar błędów w zintegrowanych źródłach zależy również w dużej mierze od kontroli

[28] Często są to dokumenty wielostronicowe, a bardzo dużej liczbie rubryk . Choć coraz częściej papierowe formularze zastępowane są elektronicznymi, ich zawiłość może powodować błędy.

wprowadzanych informacji, która często wiąże się z wykorzystaniem odpowiedniej dokumentacji, zeznań podatkowych, dokumentów urzędowych, wizją lokalną itp.. Dane z formularzy wprowadzane są do systemów administracyjnych. Pomimo skomputeryzowania urzędów, zdarza się, że informacje wprowadzane są ręcznie[29]. Z tego względu również na tym etapie mogą pojawić się błędy i nieścisłości w systemie administracyjnym.

W kolejnym kroku informacje są przetwarzane oraz harmonizowane na potrzeby zastosowania w badaniach statystycznych. Czas potrzebny na przetworzenie informacji może spowodować opóźnienie w przekazaniu danych przez co informacje z rejestru mogą stracić na aktualności. Również dostosowanie danych rejestrowych do wymogów sprawozdawczości statystycznej może generować kolejne błędy. Wynikać mogą one z dostosowania definicji administracyjnych do statystycznych, a także imputacji braków danych, tworzenia zmiennych pochodnych[30] z kilku zmiennych lub kodowania zmiennych tekstowych.

Schemat 1.12. Błąd pokrycia



Źródło: opracowanie własne na podstawie [Bakker 2010]

Pokrycie populacji docelowej jest niejednolite w różnych rejestrach i wynika ściśle z uwarunkowań prawnych. Przykładowo, rejestr administracyjny PESEL zawiera informacje o osobach przebywających stale na terytorium Rzeczpospolitej Polskiej ponad 3 miesiące [Ustawa z dnia 10 kwietnia 1974 o ewidencji ludności i dowodach osobistych (Dz. U. z 2006 r. Nr 139, poz. 993 ze zm.)]. Nie obejmuje on jednak osób przebywających na terytorium Polski nielegalnie, pomimo, że należą one do populacji docelowej rejestru. Zdarzają się również

[29] Np. ze względu na konieczność złożenia „tradycyjnego" podpisu pod wprowadzonymi informacjami na formularzu papierowym.

[30] Na etapie przetwarzania danych w rejestrach administracyjnych tworzone są często zmienne pochodne (np. wiek na podstawie daty urodzenia, stan cywilny na podstawie daty zawarcia związku małżeńskiego, rozwodu bądź śmierci współmałżonka itp.).

opóźnienia związane z uzupełnianiem rejestru o osoby nowonarodzone oraz imigrantów. Może skutkować to niedoreprezentacją rejestru. Z kolei opóźnienia związane z rejestracją zgonów i emigracji mogą skutkować nadreprezentacją populacji w zbiorze.

Błędy pokrycia mogą wynikać przede wszystkim z różnic definicji populacji docelowej rejestru (*de iure*) i populacji statystycznej (por. schemat 1.12). Zwykle pokrywają się one tylko częściowo i konieczna jest harmonizacja, która zwykle sprowadza się do usunięcia jednostek nie spełniających określonych warunków. Prowadzić to może do utraty informacji tym większej, im mniejszą frakcję populacji generalnej stanowi część wspólna populacji *de iure* i docelowej.

Łączenie administracyjnych źródeł danych odbywa się najczęściej na podstawie unikalnego klucza połączeniowego (nr PESEL, NIP, REGON itp.), a w przypadku braku unikalnego klucza – na podstawie kombinacji zmiennych (tzw. klucza złożonego), np. płeć, data urodzenia oraz adres zamieszkania. Na etapie integracji mogą pojawić się dwa rodzaje błędów: brak połączenia oraz połączenie błędne [Fellegi, Sunter 1969; Arts *et al.* 2000]. Brak połączenia to sytuacja, w której nie można jednoznacznie połączyć rekordów, natomiast błędne połączenie występuje, gdy połączone są rekordy należące do różnych jednostek. Jeżeli brak połączeń nie ma charakteru losowego, może powodować obciążenie szacunków [Bakker 2010].

Przetwarzanie danych zintegrowanych

Rezultatem integracji rejestrów jest baza mikrodanych zawierająca informacje ze wszystkich łączonych źródeł. Definicje zmiennych nie są ujednolicone, ich jakość jest różna, różny jest stopień pokrycia populacji celu w poszczególnych źródłach. Etap przetwarzania danych zintegrowanych umożliwia utworzenie operacyjnej bazy mikrodanych zawierającej dane ujednolicone w wyniku harmonizacji. Przetwarzanie danych zintegrowanych jest stałym procesem, który zapewnia spójność i jakość informacji statystycznej m.in. przez zachowanie zalecanych przez Eurostat takich aspektów jakości danych, jak:

- przejrzystość,
- zasięg – pokrycie populacji docelowej,
- kompletność – analiza pokrycia informacyjnego zintegrowanych rejestrów odnośnie produktu statystycznego, który mają tworzyć,
- zakres czasowy – ujednolicenie momentów referencyjnych zbiorów, porównywalność danych w czasie,
- aktualność,

— kontrola błędów statystycznych – porównanie informacji z różnych źródeł w celu eksploracji i korekty błędów,
— poufność i bezpieczeństwo – konieczność publikowania komunikatów w sposób uniemożliwiający identyfikację jednostek.

Proces przetwarzania danych zintegrowanych może służyć do korekty błędów związanych z integracją. W szczególności wykorzystuje się tę metodologię do korekty błędów związanych z pokryciem oraz harmonizacją definicji cech i jednostek. Korekta niedoreprezentacji może odbywać się za pomocą kilku sposobów [Groves *et al.* 2004; Stoop 2005]:

— łączenie różnych źródeł celem utworzenia kompletnej listy jednostek populacji docelowej,
— nadanie wag analitycznych poszczególnym jednostkom w taki sposób, by odzwierciedlały liczebność całej populacji docelowej,
— imputacja wartości dla brakujących jednostek.

Brak dodatkowych źródeł, czy duża liczba brakujących jednostek mogą czasem uniemożliwiać korektę błędu niedoreprezentacji. Wówczas możliwe jest wykorzystanie informacji pochodzących z badań reprezentacyjnych dotyczących populacji celu. Dołączenie rekordów z takiego badania umożliwi oszacowanie rozmiaru niedoreprezentacji poprzez zastosowanie wag analitycznych przyporządkowanych rekordom z badania pomocniczego [Bakker 2010].

Błędy związane z nadreprezentacją najczęściej korygowane są poprzez usuwanie jednostek nienależących do populacji docelowej. Błędy te wynikają zwykle z różnych momentów referencyjnych zbiorów danych oraz opóźnień w uzupełnianiu rejestrów. Wykrywanie jednostek należących do badanej populacji generalnej odbywa się na ogół poprzez łączenie rejestrów z innymi źródłami danych (np. rejestru bezrobotnych z rejestrem zatrudnionych). Umożliwia to wykrycie podmiotów już nienależących do zbiorowości będącej obiektem zainteresowania.

Inne błędy w zintegrowanych źródłach danych polegają głównie na:

— różnym zdefiniowaniu zmiennych (np. płaca brutto w jednym zbiorze, podczas gdy w drugim podaje się płacę netto),
— błędnym pomiarze (np. w jednym rejestrze jednostka jest oznaczona jako mężczyzna, w drugim jako kobieta),

— błędnym przetworzeniu danych (np. roczne wynagrodzenie nie jest sumą wszystkich miesięcznych zarobków).

Harmonizacja integrowanych zbiorów w celu wyeliminowania niespójności wiąże się z wyborem dla każdej zmiennej najbardziej wiarygodnego źródła lub konstrukcji zmiennych pochodnych odpowiadających definicjom statystycznym.

1.6. Bezpieczeństwo informacji

W społeczeństwach o rozwiniętych strukturach obywatelskich cenione jest prawo do prywatności. Ingerencja państwa w tę sferę życia często spotyka się z gwałtownymi protestami. Społeczeństwa takie często zdają sobie sprawę z zagrożeń, jakie wynikają z integracji administracyjnych źródeł danych (zespolenia informacji o każdym obywatelu w jedną bazę) dla prywatności . Obawa przed wykorzystaniem danych zgromadzonych w rządowych repozytoriach przeciw obywatelom (np. w celach podatkowych) lub przed wyciekiem danych sprawia, że bardzo często nie dostrzega się lub pomija korzyści takiego podejścia do statystyki publicznej. Dlatego też badacze przykładają dużą wagę do kwestii bezpieczeństwa informacji.

Fellegi [1997] stwierdza, że na pojawienie się metodologii integracji danych wpływ miały cztery czynniki:

1. **Powojenny rozwój państwa dobrobytu** (*welfare state*) wraz z rozbudowanym systemem podatkowym dały początek dużym repozytoriom danych o obywatelach i przedsiębiorstwach.
2. **Rozwój informatyzacji** umożliwiający utrzymanie i zarządzanie tymi repozytoriami, praktycznie niczym nieograniczona możliwość dodawania nowych informacji i związany z tym dostęp do olbrzymiej ilości danych.
3. **Zwiększenie roli państwa** i związany z tym wzrost popytu na szczegółowe informacje, które mogą być dostarczane przy użyciu rejestrów administracyjnych.
4. **Obawy społeczne** dotyczące niebezpieczeństwa naruszenia prywatności jednostek oraz konieczność zapobieżenia temu niebezpieczeństwu.

O ile trzy pierwsze czynniki związane są z bodźcami wspierającymi rozwój, o tyle czynnik czwarty związany jest z ograniczeniami, które miały duży wpływ na kształtowanie się metod integracji, a także na ich zastosowanie. Duże repozytoria danych osobowych niemal od samego początku istnienia spotykały się z nieufnością, szczególnie w kontekście dyskusji o roli państwa w życiu społecznym i prawach jednostki. W obawie przed zbytnią ingerencją w życie prywatne obywateli, stosowanie integracji danych spotyka się ze społecznym oporem. Dlatego też urzędy statystyczne krajów wysoko rozwiniętych opracowały szereg założeń, jakie

musi spełniać proces integracji, by można go przeprowadzić. Przykładowo, Kanadyjski Urząd Statystyczny stosowanie metod integracji w badaniach statystycznych warunkuje:

- wyłącznie potrzebą masowych badań statystycznych,
- zgodnością publikacji wyników z zapisami prawnymi o zachowaniu tajemnicy statystycznej,
- dużymi społecznymi korzyściami zastosowania integracji, które służącymi interesowi społecznemu,
- sytuacją, w której zastosowanie metodologii jest jedyną możliwością pozyskania informacji lub, ze względu na koszt i inne ograniczenia, jest jedyną realną opcją,
- łączenie rekordów nie zakłóca innych badań w urzędzie statystycznym,
- proces łączenia poddany jest szczegółowej kontroli oraz ocenie.

Szeroki zakres informacyjny zintegrowanych źródeł danych łączy się z dużym ryzykiem ujawnienia informacji wrażliwych. Łączenie zbiorów przy pomocy unikalnego klucza połączeniowego, którym zwykle jest numer PESEL lub jego odpowiednik, powoduje, że dane osobowe muszą występować w całym procesie integracji. Są one przechowywane w dużych repozytoriach i hurtowniach danych i dopóki nie zostaną z nich usunięte[31] istnieje ryzyko ich ujawnienia. Dodatkowo, w przypadku występowania jednostek o rzadkich lub szczególnych charakterystykach, możliwa jest ich identyfikacja nawet na podstawie zagregowanych tabel wynikowych.

Zabezpieczenie przed ujawnieniem informacji poufnych i wrażliwych może zapewnić przestrzeganie czterech podstawowych reguł [Wallgren, Wallgren 2007]:

1. **Minimalizacja publikowania informacji na podstawie cech o charakterze tekstowym** – zawierają one zwykle nazwy, adresy i opisy poszczególnych jednostek; należy je zakodować w określone grupy umożliwiające publikowanie komunikatów statystycznych.
2. **Minimalizacja użycia numerów identyfikacyjnych jednostek** – numery takie jak PESEL, NIP, REGON, itp. powinny być używane z daleko posuniętą ostrożnością; o ile to możliwe, powinny zostać przekodowane w numery sztuczne, identyfikujące rekordy wyłącznie na potrzeby integracji [Nordholt 2004].
3. **Tworzenie tabel wynikowych w sposób minimalizujący możliwość odczytania informacji o jednostkach indywidualnych** – agregacja powinna być przeprowadzona tak, by informacje o jednostkach rzadkich lub szczególnych były niemożliwe do od-

[31] Co następuje zwykle dopiero jakiś czas po zakończeniu integracji i publikacji wyników, np. w NSP 2011 dane osobowe zostaną usunięte dopiero dwa lata po zakończeniu spisu [GUS 2009].

czytania (dotyczy to np. bardzo dużych przedsiębiorstw czy osób o rzadko spotykanym zawodzie występujących jako pojedyncza obserwacja w danej kategorii).

4. **Specjaliści** zajmujący się pracą nad jednostkowymi zbiorami danych powinni zostać **zaprzysiężeni,** a dostęp do danych powinien być ograniczony w taki sposób, by uniemożliwić skopiowanie informacji na zewnętrzne nośniki danych.

Zmienne o charakterze tekstowym można pogrupować w taką liczbę kategorii, która umożliwia publikowanie wyników w sposób bezpieczny. Liczba wariantów pogrupowanych cech musi być odpowiednio mała. Np. dane o charakterze adresowym (miejscowość, ulica, numer domu) można zagregować do poziomu samej miejscowości, gminy a nawet klasy miejscowości zamieszkania. W przypadku na przykład nazw zakładów pracy lub stanowisk pracy, informacje takie agregowane są zwykle według kodu klasyfikacji działalności gospodarczej lub kodu zawodu wg ISCO[32].

W celu minimalizacji wykorzystania unikalnych numerów identyfikacyjnych, są one często przekodowane w numery sztuczne. Zachowują wówczas swoją unikalność (jeden numer dla jednej jednostki), jednak na podstawie zmienionego numeru niemożliwe jest odszukanie jednostki w źródłach administracyjnych oraz ekstrakcja dodatkowych informacji (np. data urodzenia, czy płeć z numeru PESEL). Przykładem takiego sztucznego klucza są numery nadawane w holenderskim spisie wirtualnym. Każda jednostka w każdym z rejestrów była identyfikowana przez unikatowy klucz jakim był numer ubezpieczenia społecznego i podatkowego (*social security and fiscal numer, SoFi-number*). Jednak ze względu na ochronę danych osobowych, na potrzeby łączenia zbiorów klucz ten został przekodowany na tzw. numer identyfikujący rekord (*Record Identification Number, RIN-person*). Inne zmienne identyfikujące jednostkę jak data urodzenia oraz adres zamieszkania zostały przekształcone w zmienne wiek w momencie referencyjnym spisu oraz numer identyfikujący adres – *RIN-address*[33] [Nordholt 2004].

Istotne znaczenie ma ochrona przed ujawnieniem jednostek w raportach tabelarycznych. Szczególne ryzyko identyfikacji dotyczy jednostek rzadko występujących w badaniu lub posiadających np. rzadki na danym terytorium zawód [Hundepool, Willenborg 1997]. Podmiot taki powinno się objąć szczególną ochroną . Wyróżnia się trzy podstawowe metody zapobiegania ujawnienia informacji jednostkowych z tabel wynikowych:

— agregacja rzadkich wariantów;

[32] ISCO - *International Standard Classification of Occupations*; międzynarodowe standardy klasyfikacji zawodów wykonywanych.
[33] Holenderski spis wirtualny zostanie omówiony szerzej w rozdziale II.

— agregacja wariantów rzadkich tylko dla określonych subpopulacji lub usuwanie informacji o jednostkach szczególnych;
— tzw. „zasada dominacji".

Najprostszą i najczęściej stosowaną metodą jest łączenie wariantów opisywanych cech (np. zawód „statystyk lub matematyk" zamiast dwóch odrębnych kategorii). W praktyce jednak trudno jest rozróżnić czy cechy jednostki o rzadkich charakterystykach w danej subpopulacji są również unikalne w całej populacji. Usprawnieniem metody agregacji jest prezentowanie wartości zbiorczych o poziomie agregacji różnym dla różnych przekrojów [Herzog *et al.* 2007]. Np. jeżeli na jakimś terytorium występuje tylko jeden zakład pracy zatrudniający powyżej 500 pracowników, w tabeli wynikowej włącza się go w kategorię, o większej liczebności (np. 100 pracowników i więcej). Na terytorium, gdzie takich zakładów jest więcej, nie ma potrzeby zmniejszania liczby kategorii. Takie podejście nazywa się rekodowaniem globalnym (*global recoding*). Alternatywą jest usuwanie informacji o jednostkach szczególnych (*local suppression*) gdy np. ujawnienie zawodu jakiejś osoby może doprowadzić do jej zidentyfikowania. W takich przypadkach wartości danej zmiennej nie ujawnia się wcale. Obie metody prowadzą niestety do utraty informacji zarówno poprzez agregację, jak i usuwanie obserwacji. Dlatego też należy tak zoptymalizować działania, by strata informacji była jak najmniejsza.

Przykładem optymalizacji ochrony tajemnicy statystycznej poprzez odpowiednie konstruowanie tabel wynikowych mogą być doświadczenia holenderskiego urzędu statystycznego [Hundepool, Willenborg 1997]. Rozpoczęto tam prace nad ustanowieniem uniwersalnych zasad, wedle których bazy danych będą na tyle zabezpieczone przed ujawnieniem danych jednostkowych, że będą mogły być użyte przez badaczy. Efektem tych prac jest oprogramowanie μ-Argus oraz τ-Argus. Oprogramowanie μ-Argus umożliwia użytkownikowi określać globalne rekodowania interaktywnie. Gdy globalne rekodowania zostaną już ustalone, usuwanie jednostek szczególnie narażonych na identyfikację przeprowadzane jest automatycznie i optymalnie, tj. liczba usuniętych wartości jest minimalna.

„Zasadę dominacji" wykorzystuje oprogramowanie τ-Argus. Zasada ta mówi, że dana komórka w tabeli niesie niebezpieczeństwo ujawnienia informacji o jednostce, jeżeli mała liczba badanych odpowiada wysokiemu procentowi całości. Liczby, które uważa się za niebezpieczne to, odpowiednio: 3 i 70% (3 jednostki stanowią 70% analizowanej (pod)zbiorowości). Wartości takie w tabelach należy usunąć, jak również zapewnić brak możliwości ich określenia przy użyciu pozostałych informacji. Tajemnicę statystyczną w τ-Argus chroni się poprzez:

— przeprojektowanie tabeli w taki sposób, że poszczególne komórki zawierają bardziej uogólnione wartości agregatowe,

— zaokrąglenia uniemożliwiają dokładne obliczanie sum,
— usuwanie informacji z komórek pomocniczych zapobiega obliczeniu wartości dla komórek właściwych.

Alternatywą może okazać się integracja danych w sposób niedeterministyczny, a stochastyczny [Herzog *et al.* 2007]. Zamiast łączenia rekordów należących do tej samej jednostki, można łączyć rekordy podobne do siebie pod względem pewnych wybranych arbitralnie cech. Tworzone są wówczas jednostki syntetyczne, a zastosowanie odpowiednich narzędzi (parowania statystycznego, a w szczególności wielokrotnej imputacji [Raghunathan *et al.* 2003]) prowadzi do zgodności rozkładów analizowanych cech z rozkładem w populacji generalnej.

Nowozelandzki Urząd Statystyczny opracował procedurę *Oceny zagrożenia naruszenia prywatności* (*Privacy Impact Assessment, PIA*). Jest to procedura wdrażana w życie za każdym razem, gdy wprowadzana jest nowa technologia lub metodologia, udoskonalana jest dotychczasowa lub następuje jej użycie w nowych okolicznościach narażających prywatność. W przypadku integracji danych, umożliwiono publiczną krytykę stosowanych metod by poznać opinie osób obawiających się o swoje prawo do prywatności, nawet przy spełnieniu wymogów prawnych [*Statistics New Zealand* 2006].

Duże ograniczenia nałożone na stosowanie metodologii i publikację wyników powodują, że wszelkie działania podejmowane w celu łączenia repozytoriów danych są poddawane publicznej debacie. Kwestię tę poruszają szerzej Wallman i Coffey [1997] potwierdzając, że zachowanie tajemnicy statystycznej jest nie tylko wymogiem prawnym, ale przede wszystkim, ważnym punktem debaty społecznej na temat zastosowania administracyjnych źródeł danych w badaniach statystycznych. Autorzy wskazują, że zbyt wysokie oczekiwania społeczne dotyczące poufności spowodowały bardzo restrykcyjne przepisy utrudniające, a czasem wręcz uniemożliwiające wymianę informacji między gestorami rejestrów.

Przeprowadzone w Stanach Zjednoczonych w 1995 i 1996 roku badania pokazały niski poziom zaufania społecznego w stosunku do urzędów będących gestorami danych administracyjnych, a także wzrost przekonania, że dane nie są w należycie chronione [Singer *et al.* 1997]. Na podstawie tych samych badań okazało się również, że społeczeństwo nie jest wystarczająco dobrze poinformowane ani o tym, jakie dane podlegają wymianie, ani o kwestiach tajności danych. Zaistniała więc obawa, że opinia publiczna może kwestionować wykorzystanie danych administracyjnych w badaniach statystycznych. By zapobiegać utracie społecznego poparcia dla wykorzystania rejestrów administracyjnych i innych dostępnych źródeł do

badań statystycznych podjęto więc szereg działań, m.in. ukazujących społeczne i ekonomiczne korzyści wykorzystania informacji z rejestrów.

W Polsce, mimo przeprowadzonych na szeroką skalę konsultacji społecznych w związku z NSP 2011, w których brało udział wiele instytucji publicznych i prywatnych, stowarzyszeń oraz związków wyznaniowych, w prasie pojawiły się krytyczne oceny łączenia rejestrów administracyjnych i badań reprezentacyjnych[34]. Reakcje takie świadczą o nieufności również polskiej opinii publicznej do zbierania i przetwarzania danych jednostkowych. Przekonanie społeczeństwa do słuszności i korzyści takiego podejścia do statystyki publicznej będzie jedną z najbardziej kluczowych kwestii w nadchodzących latach.

Osobną kwestią pozostają repozytoria danych pozostające w rękach prywatnych. O ile łączenie rejestrów administracyjnych w sektorze państwowym obłożone jest wieloma warunkami i ograniczeniami, nad zachowaniem których czuwają odpowiednie służby, o tyle ograniczeń takich nie ma (lub są mniej rygorystyczne) w sektorze prywatnym [Fellegi 1997]. Wiele firm posiada duże bazy danych o swoich klientach, nad którymi tylko one sprawują kontrolę. Konsekwencje nieautoryzowanego wykorzystania tych danych nie trudno sobie wyobrazić, np.:

- dana osoba może otrzymywać konkretne reklamy w zależności od tego, jakie informacje zebrał o niej reklamodawca;
- obywatel może nie otrzymać kredytu nawet jeżeli informacje o jego finansach są błędne lub nieaktualne;
- przy dostępie do informacji niejawnych (np. informacji o przebytych chorobach), mogą pojawić się problemy z otrzymaniem ubezpieczenia,
- podczas rozpraw sądowych, jedna ze stron może uzyskać przewagę poprzez wejście w posiadanie informacji o adwersarzu.

Sytuacja taka wyraźnie wskazuje, że istnieje poważny problem informacji o obywatelach przechowywanych w prywatnych bazach danych. Odruchową reakcją byłoby wprowadzenie odpowiednich regulacji prawnych ograniczających możliwość wykorzystania takich zbiorów. Jednak rozwój technik informacyjnych spowodował, że takie pliki z łatwością można przenieść na serwer znajdujący się w innym kraju bez jakiejkolwiek straty dla ich integralności, czy możliwości użycia. Zastosowanie znaleźć więc muszą inne rodzaje nacisku na prywatnych gestorów danych osobowych. Jedną z nich może być wymaganie od podmiotu, który

[34] W szczególności duże kontrowersje wzbudził artykuł prasowy porównujący zbieranie danych o Polakach do „metod rodem z Orwella" [Leszczyńska 2009]. Wywołał on stanowczą reakcję Głównego Urzędu Statystycznego dementującego, jakoby zbieranie danych z rejestrów administracyjnych i badań reprezentacyjnych miało cel inny niż analizy statystyczne.

żąda danych osobowych od danego obywatela, by w zamian dostarczył informacji o polityce zarządzania danymi osobowymi, metodach zabezpieczania warunków na jakich dane są udostępniane innym podmiotom. Choć administracja publiczna może nie mieć bezpośredniej możliwości kontroli, czy dane zapisy są w praktyce stosowane, świadome swoich praw społeczeństwo może wywierać odpowiednią presję, a także może mieć możliwość składania formalnych zażaleń, jeżeli zapisy prawne nie będą respektowane. Wydaje się, że w takich okolicznościach firmy dysponujące bazami danych osobowych, w obawie przed utratą reputacji bądź przewagi konkurencyjnej, we własnym interesie będą przestrzegać tajemnicy tych informacji oraz nie będą udostępniać ich innym podmiotom.

1.7. Wnioski

Wykorzystanie rejestrów administracyjnych i ich integracja, także z badaniami reprezentacyjnymi, jest procesem nietrywialnym, wymagającym zaangażowania odpowiednich służb i środków. Problemy, jakie na drodze integracji można napotkać zmuszają do stałej poprawy algorytmów harmonizacji, łączenia i analizy danych zintegrowanych. Jest to również proces interdyscyplinarny, to znaczy angażujący ekspertów z wielu dziedzin nauki, m.in. informatyki, statystyki, socjologii, administracji, prawa i badania opinii społecznej. Jednak doświadczenia krajów, które stosowały integrację danych w badaniach statystycznych (przede wszystkim w spisach powszechnych) wskazują, że korzyści wypływające z tworzenia zintegrowanych źródeł danych przewyższają koszty ich powstania. Doświadczenia te zostaną opisane w rozdziale II.

ROZDZIAŁ II. DOTYCHCZASOWE DOŚWIADCZENIA W INTEGRACJI DANYCH

Metody integracji danych administracyjnych i pochodzących z badań reprezentacyjnych są coraz częściej stosowane przez organy statystyki publicznej. Wykorzystanie istniejących już informacji umożliwia odciążenie respondentów, a także zmniejszenie kosztów i czasu przygotowania badania. Korzyści wypływające z integracji dostrzegają również podmioty prywatne, np. agencje badań rynku, które nie tylko stosują, ale także rozwijają metodologię łączenia informacji z badań częściowych. W niniejszym rozdziale opisane zostaną doświadczenia wybranych instytucji w zakresie stosowania technik integracji danych, zarówno metod deterministycznych, jak i stochastycznych.

Jako punkt wyjścia do rozważań, zaprezentowane zostaną doświadczenia Holenderskiego Urzędu Statystycznego (*Centraal Bureau voor de Statistiek*, CBS), który jest jedną z pierwszych instytucji, jaka przeprowadziła tzw. spis wirtualny. Jest to forma spisu powszechnego, w której nie przeprowadza się bezpośredniego pomiaru u respondentów, a całość danych pobierana jest z administracyjnych systemów informacyjnych oraz wcześniej przeprowadzonych badań reprezentacyjnych. Przedstawione zostaną etapy i techniki przeprowadzania takiego badania, wykorzystane źródła danych, a także problemy i sposoby ich rozwiązania.

W latach 2010 i 2011 w Polsce przeprowadzono, odpowiednio, Powszechny Spis Rolny (PSR 2010) oraz Narodowy Spis Powszechny Ludności i Mieszkań (NSP 2011). Badania te zostały przeprowadzone tzw. metodą mieszaną, będącą formą przejściową między spisem „tradycyjnym" (opartym o papierowe kwestionariusze spisowe wypełniane przez rachmistrzów odwiedzających osobiście gospodarstwa domowe), a spisem wirtualnym. W PSR 2010 oraz NSP 2011 informacje częściowo zebrano z udostępnionych przez gestorów administracyjnych repozytoriów danych, a brakujące dane uzupełniono poprzez kwestionariusze wypełniane bezpośrednio przez respondentów[35]. W rozdziale przedstawione zostaną etapy i formy zbierania i integracji danych z różnych źródeł, a także procedury umożliwiające utworzenie wynikowych zbiorów danych. Opisane zostaną również problemy estymacji związane z dołączania do bazy spisowej repozytoria utworzonego na podstawie tzw. części reprezentacyjnej spisu.

W dalszej kolejności przedstawione zostaną wybrane badania społeczno-gospodarcze prowadzone na podstawie zintegrowanych źródeł danych bieżącej sprawozdawczości statystycznej. W szczególności opisane zostaną doświadczenia krajów zachodnich (Wielkiej Brytanii i Włoch) w kontekście łączenia (deterministycznego i stochastycznego) różnych źró-

[35] Należy jednak zaznaczyć, że respondenci mieli możliwość wypełnienia kwestionariuszy przez Internet (dokonać tzw. „samospisu"), co również pozwoliło na zaoszczędzenie czasu i kosztów pomiaru.

deł w celu utworzenia kompleksowych systemów sprawozdawczych. Wspomniane zostaną również polskie doświadczenia z wykorzystania źródeł administracyjnych w Badaniu przepływów ludności związanych z zatrudnieniem, gdzie dane systemu POLTAX posłużyły do szacowania dojazdów do pracy, a także projekt MEETS, w którym wykorzystano rejestry administracyjne na potrzeby statystyki przedsiębiorstw.

Na koniec rozdziału przedstawione zostanie znaczenie rozwoju i szerokiego wykorzystania metodologii statystycznej integracji dla instytucji Unii Europejskiej na przykładzie Eurostatu. Opisane zostaną dwa projekty badawcze: *CENEX-ISAD* oraz *ESSnet on Data Integration* poświęcone rozwojowi metodologii integracji danych w europejskich urzędach statystycznych. Omówiony zostanie również problem analiz *ad hoc* przy użyciu zintegrowanych repozytoriów danych (zarówno administracyjnych, jak i statystycznych) oraz wskazane zostaną doświadczenia prywatnych firm zajmujących się badaniami rynkowymi i marketingowymi w dziedzinie statystycznej integracji danych.

2.1. Spisy powszechne

2.1.1. Spis wirtualny w Holandii

Spis wirtualny jest pełnym badaniem statystycznym wykorzystującym istniejące i dostępne źródła danych, bez potrzeby przeprowadzania pomiaru obciążającego respondentów [van der Laan 2000]. Poza rejestrami administracyjnymi wykorzystuje się w nim również wcześniej przeprowadzone badania reprezentacyjne[36]. Przeprowadza się go w miejsce tradycyjnego spisu powszechnego w celu obniżenia kosztów badania, zmniejszenia obciążenia respondentów oraz poprawy jakości danych [Nordholt 2004].

Pierwszy spis wirtualny w Holandii przeprowadzono w 2001 roku (kolejny odbył się w roku 2011 [Nordholt 2005]). Wykorzystano takie administracyjne źródła danych jak ewidencja ludności (źródło podstawowe), rejestry zatrudnienia, skarbowy i pomocy społecznej, a także reprezentacyjne badanie aktywności ekonomicznej ludności (*Labour Force Survey, LFS*) [Linder 2004].

[36] Nie ma potrzeby ponownego zadawania tych samych pytań w spisie.

Tabela 2.1. Źródła danych wykorzystane w spisie wirtualnym w Holandii

Lp.	Nazwa	Rodzaj badania/ populacja objęta badaniem	Rekord/jednostka	Liczba rekordów	Zmienne wykorzystane w spisie
1	Ewidencja ludności	pełne	osoba	ok. 16 mln osób	płeć
					wiek
					kraj urodzenia
					obywatelstwo
					kraj zamieszkania
					region zamieszkania
					stan cywilny
					sytuacja rodzinna
			gospodarstwo domowe	ok. 6,9 mln gospodarstw domowych	skład gospodarstwa domowego
					status gospodarstwa domowego
					typ gospodarstwa domowego
					wielkość gospodarstwa domowego
					liczba dzieci
					inne osoby w gospodarstwie domowym
2	Rejestr Ubezpieczeń Pracowniczych	pełne	pracujący	ok. 6,5 mln	rodzaj zatrudnienia
					klasyfikacja działalności gospodarczej miejsca zatrudnienia
					wynagrodzenie brutto
3	Badanie Zatrudnienia i Wynagrodzeń	częściowe	pracujący	ok. 3 mln	rodzaj zatrudnienia
					wymiar czasu pracy
					miejsce zatrudnienia
4	Rejestr podatkowy samozatrudnionych	pełne	osoba samozatrudniona	ok. 800 tys.	charakter samozatrudnienia
					klasyfikacja działalności gospodarczej
5	Rejestr Ubezpieczeń Bezrobotnych	pełne	bezrobotny	ok. 440 tys.	źródło pomocnicze[1]
6	Rejestr Ubezpieczeń Niepełnosprawnych	pełne	osoba niepełnosprawna	ok. 1 mln	źródło pomocnicze[1]
7	Rejestr Świadczeń Społecznych	pełne	osoba objęte opieką społeczną	ok. 580 tys.	źródło pomocnicze[1]
8	Rejestr Podatkowy	pełne	świadczeniobiorca	ok. 7,2 mln miejsc zatrudnienia/ok. 2,7 mln emerytur i rent	zatrudnienie: tak/nie
					rodzaj świadczeń emerytalno-rentowych
9	Badanie aktywności ekonomicznej ludności	częściowe	osoba	ok. 120 tys.	poziom wykształcenia
					zawód
					aktywność ekonomiczna

Uwaga: [1]Zmienne ze źródeł pomocniczych nie zostały bezpośrednio wykorzystane w spisie, a służyły jedynie jako zmienne pomocnicze w procesie integracji.

Źródło: na podstawie [van der Laan 2000]

W spisie wykorzystano informacje pochodzące z 9 różnych źródeł (por. tabela 2.1). Charakteryzowało je różne pokrycie informacyjne, zarówno w odniesieniu do liczby jednostek, jak i zmiennych te jednostki opisujących.

Najważniejszym źródłem w spisie wirtualnym była Ewidencja Ludności (*Population Register*), zasilana w Holandii przez rejestry osobowe administracji samorządowej. Jako najbardziej rzetelne źródło danych o największym pokryciu populacji, Ewidencja Ludności stała się „kręgosłupem" spisu, do którego dołączano informacje z pozostałych źródeł danych [Nordholt 2004]. Również proces harmonizacji danych odbywał się w odniesieniu do kategorii i definicji stosowanych w tym rejestrze.

Informacje zawarte w Ewidencji Ludności odzwierciedlały stan prawny ludności Holandii. Oszacowanie stanu faktycznego wymagało korekty[37] przy pomocy modeli probabilistycznych skonstruowanych na podstawie badań reprezentacyjnych odzwierciedlających stan faktyczny [Linder 2004].

Rejestry pracownicze[38] oraz badania reprezentacyjne dotyczące pracowników[39] dostarczały informacji dotyczących m.in. wynagrodzeń, miejsc pracy oraz klasyfikacji działalności gospodarczej miejsca zatrudnienia (lub samozatrudnienia). Ze względu na administracyjny charakter rejestrów pracowniczych, często wykorzystywanych pośrednio w celach podatkowych, rzetelność danych była na wysokim poziomie (za podanie błędnych informacji groziła odpowiedzialność karna i skarbowa). Informacje niezawarte w rejestrach uzupełniane były danymi z badań reprezentacyjnych (np. wymiar czasu pracy). Szczególnie ważną rolę pełniło badanie aktywności ekonomicznej ludności. Zawierało ono szereg informacji niedostępnych w rejestrach, jak poziom wykształcenia, czy zawód wykonywany. Dodatkowo badanie to służyło jako podstawa identyfikacji statusu na rynku pracy (aktywny zawodowo, pracujący, bezrobotny, formalnie bierny zawodowo, ale pomagający członek rodziny).

Rejestr podatkowy (*FiBase-register, fiscal administration*) zawierał informacje o dochodach pochodzących z pracy zarobkowej oraz świadczeń społecznych, jak również o świadczeniach emerytalno-rentowych. Służył on jako ważne źródło informacji o aktywności ekonomicznej ludności (m.in. zawierał informację o biernych zawodowo). Wykorzystano go w procesie integracji jako główne źródło do estymacji zatrudnienia i różnych rodzajów dochodu.

Rejestr Ubezpieczeń Bezrobotnych, Rejestr Ubezpieczeń Niepełnosprawnych oraz Rejestr Świadczeń Społecznych pełniły w spisie rolę źródeł informacji pomocniczych w procesie

[37] Wynikało to m.in. z opóźnień w rejestracji narodzin i zgonów, a także trudności w ujęciu imigrantów.
[38] Rejestr Ubezpieczeń Pracowniczych, Rejestr podatkowy osób samozatrudnionych
[39] Badanie Zatrudnienia i Wynagrodzeń, Badanie Aktywności Ekonomicznej Ludności

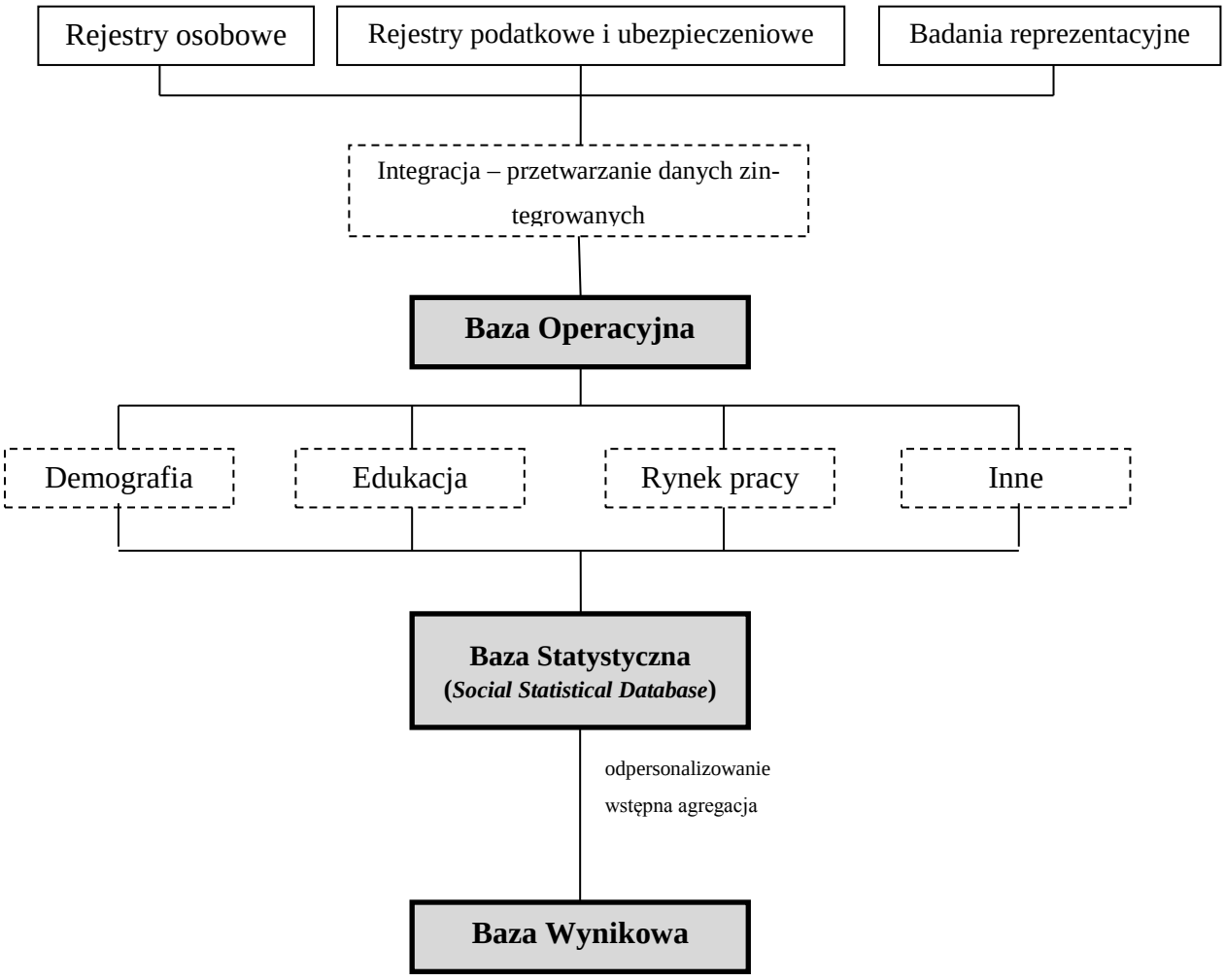
integracji. Repozytoria te zawierały dane o zatrudnieniu i świadczeniach społecznych, które wykorzystano w procesie harmonizacji.

Każda jednostka w każdym z rejestrów była identyfikowana przez unikatowy klucz jakim był numer ubezpieczenia społecznego i podatkowego (*social security and fiscal numer, SoFi-number*). Jednak ze względu na ochronę danych osobowych, na potrzeby łączenia zbiorów klucz ten został przekodowany na tzw. numer identyfikujący rekord (*Record Identification Number, RIN-person*). Inne zmienne identyfikujące jednostkę jak data urodzenia oraz adres zamieszkania zostały przekształcone w zmienną zawierającą wiek respondenta w momencie referencyjnym spisu oraz *RIN-address* (zakodowane informacje adresowe). Wysoka jakość holenderskich danych administracyjnych umożliwiła połączenie w sposób deterministyczny niemal 100 procent wszystkich rekordów [Nordholt 2004].

Do zintegrowanych w ten sposób rejestrów dołączono, również w sposób deterministyczny, informacje z Badania Aktywności Ekonomicznej Ludności oraz Badania Zatrudnienia i Wynagrodzeń. Ponieważ numer *SoFi* nie był przedmiotem pomiaru w badaniach reprezentacyjnych, na podstawie zmiennych płeć, data urodzenia oraz adres zamieszkania utworzono zmienną pochodną identyfikującą poszczególne osoby. Na podstawie tego klucza udało się dołączyć do rejestrów około 97% jednostek z badania częściowego.

Zintegrowana baza danych (por. schemat 2.1), zawierała informacje jednostkowe oraz łącznie obserwowane cechy ze wszystkich zbiorów. Nosiła nazwę Bazy Operacyjnej (*baseline*). Zawierała ona dużą liczbę pustych komórek, co wynika z obserwacji niektórych zmiennych tylko dla ograniczonej liczby jednostek (np. w badaniu reprezentacyjnym). Dodatkowo, ponieważ zintegrowane repozytorium danych jednostkowych nie spełniała wymagań zachowania tajemnicy statystycznej, dane zostały przeważone i zagregowane w tematy ujęte w planie publikacji. W ten sposób utworzono bazę statystyczną (*StatBase*), nazwaną Bazą Danych Społecznych (*Social Statistical Database, SSD*). Repozytorium to zawiera wiele milionów rekordów o osobach, gospodarstwach domowych, zatrudnieniu i świadczeniach społecznych opisanych za pomocą tysięcy zmiennych z różnych źródeł.

Schemat 2.1. Integracja danych pochodzących z różnych źródeł w Spisie Powszechnym w Holandii w 2001 roku



Źródło: opracowanie własne na podstawie [Everaers, van der Laan 2003]

Etap przeważania danych przeprowadzono dzieląc Bazę Operacyjną (por. schemat 2.1) na podzbiory danych („bloki danych") odnoszących się do różnych zagadnień opisywanych w spisie (demografia, edukacja, rynek pracy itp.) [Gouweleeuw, Hartgers 2004]. Dla każdego z bloków danych utworzono zestaw wag początkowych, które następnie poddano kalibracji. Umożliwiło to uzyskanie spójnych wyników („jedna liczba dla jednego zjawiska") dla wszystkich informacji zawartych w różnych zbiorach danych [Kroese *et al.* 2000]. W procesie estymacji zastosowano technikę wielokrotnego ważenia oprogramowaną w najnowszej wersji pakietu VRD (*Filling Reference Databse*) opracowanego przez holenderski urząd statystyczny. Główną funkcjonalnością VRD było oszacowanie warto-

ści w tabelach kontyngencji poprzez wielokrotne ważenie. Oprogramowanie umożliwiało również oszacowanie wariancji estymatorów.

Rzetelność oszacowań zapewniono poprzez uwzględnienie największej liczby rekordów. Tabele opisujące cechy pochodzące z rejestrów administracyjnych wyznaczano wyłącznie na podstawie zmiennych rejestrowych o wysokim pokryciu. Tabele zawierające przynajmniej jedną cechę pochodzącą z badania reprezentacyjnego szacowano na podstawie takiej kombinacji zbiorów pochodzących z rejestrów i badań, która zawierała największą liczbę obserwacji.

Połączenie informacji z rejestrów i badań reprezentacyjnych, po harmonizacji, przeważeniu i imputacji braków danych pozwoliło na utworzenie Bazy Statystycznej, nazwanej Bazą Danych Społecznych (*Social Statistical Database*). Zawierała ona jednostkowe informacje o populacji docelowej – ludności Holandii.

Ostatnim etapem spisu było utworzenie bazy wynikowej (*StatLine*). Jest to hurtownia danych zawierająca kostki danych[40] (pobieranych ze *StatBase*) z informacjami o wszystkich zagadnieniach, których opisanie zaplanowano w spisie. Agregację w Bazę Wynikową przeprowadzono ze względu na ochronę informacji wrażliwych, a także w celu zwiększenia wydajności spisowego systemu sprawozdawczości.

2.1.2. Narodowy Spis Powszechny Ludności i Mieszkań 2011

Wykorzystanie rejestrów administracyjnych w polskiej statystyce publicznej jest na etapie wstępnym. Przykładowymi badaniami, w których pełniły one rolę wspomagającą były „Badanie przepływów ludności związanych z zatrudnieniem", Narodowy Spis Powszechny 2011 (również Powszechny Spis Rolny 2010), a także statystyka przedsiębiorstw (m.in. projekt MEETS[41]). Wykorzystuje się w nich różne zbiory danych w celu zebrania niektórych informacji, aktualizacji operatu losowania oraz aktualizacji bazy budynków, mieszkań i adresów. Coraz szerszy dostęp do administracyjnych repozytoriów danych stwarza dla polskiej statystyki publicznej szansę na rozwój metod związanych z wykorzystaniem rejestrów w sprawozdawczości statystycznej, jak również unowocześnienie infrastruktury statystycznej [Paradysz 2008].

[40] Kostka danych (inaczej OLAP, *OnLine Analytical Processing*) to zestaw częściowo zagregowanych danych przyspieszający proces przetwarzania informacji [Hand *et al.* 2005].

[41] Badanie przepływów ludności związanych z zatrudnieniem, a także projekt MEETS zostaną szczegółowo opisane w dalszej części tego rozdziału

W Narodowym Spisie Powszechnym 2011 przyjęto tzw. model mieszany[42]. Część informacji zebrano ze źródeł administracyjnych i poza administracyjnych, natomiast pozostałe dane niezawarte w rejestrach zebrano bezpośrednio od respondentów poprzez rachmistrzów spisowych lub w formie samospisu. Pomiar podzielono na badanie pełne i badanie reprezentacyjne (20-procentowa próba). W badaniu pełnym za pomocą tzw. krótkiego kwestionariusza zbierano podstawowe informacje społeczne takie jak stan cywilny, obywatelstwo, czy narodowość. W badaniu reprezentacyjnym pomiaru dokonano tzw. długim kwestionariuszem. Badano za jego pomocą m.in. gospodarstwo domowe i osoby je zamieszkujące (aktywność ekonomiczną, źródła utrzymania, wykształcenie, dojazdy do pracy, niepełnosprawność, przynależność do wspólnot wyznaniowych itp.) oraz cechy mieszkania (metraż, wyposażenie, własność itp.). Kwestionariusz badania reprezentacyjnego zawierał również pytania z badania pełnego.

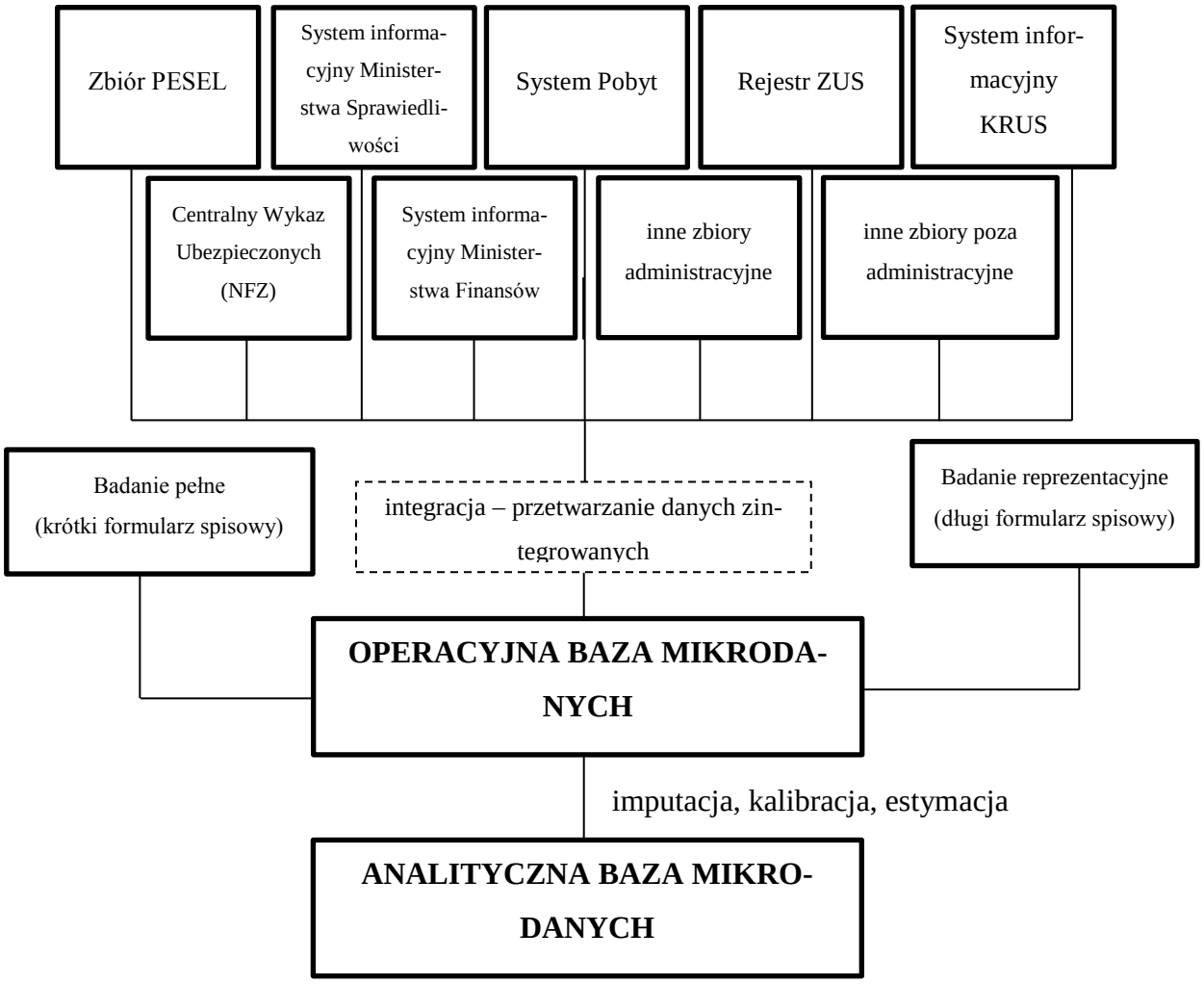
Dane administracyjne na potrzeby spisu pobrano m.in. ze zbiorów centralnych: Ministerstwa Finansów, Ministerstwa Spraw Wewnętrznych i Administracji, Ministerstwa Sprawiedliwości, Urzędu do Spraw Cudzoziemców, Zakładu Ubezpieczeń Społecznych, Kasy Rolniczego Ubezpieczenia Społecznego, Narodowego Funduszu Zdrowia, Głównego Urzędu Geodezji i Kartografii, Państwowego Funduszu Rehabilitacji Osób Niepełnosprawnych oraz licznych systemów rozproszonych (urzędów marszałkowskich, starostw powiatowych, powiatowych/miejskich zespołów ds. orzekania o niepełnosprawności, urzędów gmin/miast) oraz systemów poza administracyjnych (głównie zarządców i administratorów zasobów mieszkaniowych, przedsiębiorców wykonujących działalność gospodarczą w zakresie sprzedaży energii elektrycznej, dostawców publicznie dostępnych usług telekomunikacyjnych) [Dygaszewicz 2012]. Tak duża liczba wykorzystanych zbiorów administracyjnych wynikała z licznych celów spisu, a co się z tym ściśle wiąże, potrzebie zebrania dużej ilości informacji.

Dane, po procesie ich harmonizacji, zostały połączone zbiór danych jednostkowych nazwany Operacyjną Bazą Mikrodanych (por. schemat 2.2). System Operacyjnej Bazy Mikrodanych obejmuje infrastrukturę sprzętowo-systemowo-narzędziową (sprzęt komputerowy, oprogramowanie systemowe, oprogramowanie narzędziowe) oraz oprogramowanie aplikacyjne [Janczur-Knapek 2012]. Baza ta umożliwia połączenie danych przekazanych w formie elektro-

[42] Jako pośredni między spisem „klasycznym", opartym o papierowy kwestionariusz i rachmistrzów spisowych dokonujących pomiaru, a spisem opartym w pełni na źródłach administracyjnych, w żaden sposób nieangażującym respondentów (tzw. spisem wirtualnym).

nicznej przez gestorów rejestrów i pochodzących od rachmistrzów spisowych (z samospisu internetowego) oraz dalsze przetwarzanie dostarczonych informacji.

Schemat 2.2. Integracja danych w Narodowym Spisie Powszechnym 2011



Źródło: opracowanie własne na podstawie [Dygaszewicz 2011]

Analityczna Baza Mikrodanych przechowuje odpersonalizowane wartości zmiennych spisowych w ostatecznej wersji, zebrane podczas spisu, na których będą dokonywane wszelkie analizy statystyczne udostępniane publicznie (opracowania tabelaryczne, agregaty, analizy przestrzenne).

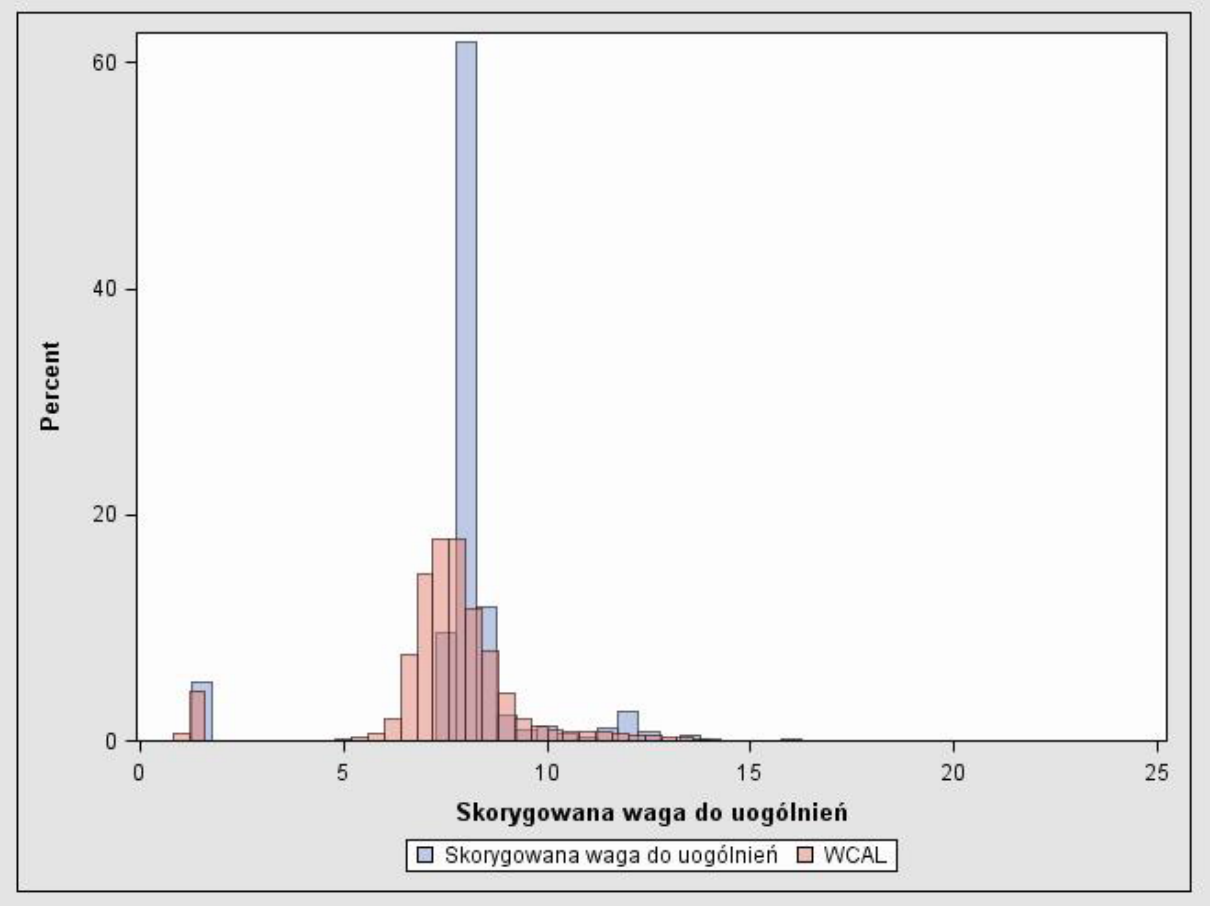
Wyniki pochodzące z badania reprezentacyjnego (długiego kwestionariusza) były obciążone ze względu na odmowy i braki odpowiedzi [Paradysz *et al.* 2012]. Zaistniała potrzeba modyfikacji wag analitycznych wynikających ze schematu doboru jednostek do próby uwzględniającej inną niż zaplanowana liczebność, a także dostosowania liczebności populacji generalnej do wynikającej z badania pełnego (38 966 806). Jako metodę korekty błędów wynikają-

cych z braków odpowiedzi wybrano podejście kalibracyjne. Zbiór pochodzący z badania reprezentacyjnego liczył 7 591 638 rekordów. Wagi kalibracyjne utworzono w taki sposób, by sumowały się do liczebności populacji jednocześnie w następujących domenach:

- płeć – 2 warianty,
- klasa miejscowości zamieszkania – 2 warianty (miasto, wieś),
- wiek:
  - dla powiatów: w dwu- lub pięcioletnich grupach wieku[43],
  - dla województw – w jednorocznych grupach wieku.

Do obliczeń wykorzystano środowisko SAS LAN przy użyciu makra CALMAR (***Cal****ibration on* ***Mar****gins*) napisanego na potrzeby Francuskiego Urzędu Statystycznego. Adaptację oprogramowania do potrzeb polskiego spisu przeprowadzono w Ośrodku Statystyki Małych Obszarów w Urzędzie Statystycznym w Poznaniu.

Wykres 2.1. Rozkład wag oryginalnych i kalibracyjnych w części reprezentacyjnej NSP 2011



Źródło: [Paradysz *et al.* 2012]

[43] W niektórych powiatach nie było informacji o ludności w każdym wieku wg płci i klas miejscowości zamieszkania. Dwuletnie grupy wieku utworzono dla części populacji, np. uczących się.

Wyznaczone wagi kalibracyjne („WCAL") były skupione wokół wag wynikowych (por. wykres 2.1). Różnice między wagami wyjściowymi a zmodyfikowanymi były stosunkowo niewielkie (przeciętnie 6%). Kolejnym etapem kalibracji była analiza tabel wynikowych uzyskanych w oparciu o zmienne ze spisu reprezentacyjnego, które nie zostały wykorzystane w procesie kalibracji jako zmienne pomocnicze. Celem tej analizy była ewaluacja spójności danych skalibrowanych.

Doświadczenia zebrane podczas Narodowego Spisu Ludności i Mieszkań 2011 oraz Powszechnego Spisu Rolnego 2010 pozwolą udoskonalić metodologię zbierania i integracji danych z różnych źródeł oraz sposób i szybkość udostępniania wyników w nadchodzących spisach, a także w okresach między spisowych.

2.2. Badania społeczne

2.2.1. System statystyki sąsiedztwa

Program statystyki sąsiedztwa zapoczątkowany w Wielkiej Brytanii w 2000 roku (przygotowania trwały od 1998 roku), ma za zadanie dostarczać informacji pozwalających wykrywać nieprawidłowości w funkcjonowaniu małych społeczności, walczyć z patologiami społecznymi, zapobiegać ubóstwu i bezrobociu, wspierać ochronę zdrowia, wspomagać samorząd lokalny i przedsiębiorców w lepszym inwestowaniu środków i ochronie środowiska. Program adresowany jest również do każdego obywatela kraju, dzięki czemu każdy może znaleźć odpowiedź na pytania „jakie jest moje sąsiedztwo?" oraz „jak moje sąsiedztwo przedstawia się na tle gminy i państwa jako całości?" [*Neigbourhood Statistics Programme Evaluation Report* 2006].

Na potrzeby Programu sporządzono odpowiednią dokumentację oraz uchwalono szereg ustaw umożliwiających wsparcie rządu centralnego. Do najważniejszych należą: Narodowa Strategia Odnowy Sąsiedztw (*National Strategy for Neighbourhood Renewal*) wyznaczająca cele oraz środki dla realizacji Programu oraz Zasady dostępu do danych i ich tajności (*Data Access and Confidentiality Policy*) wprowadzające ścisłe reguły postępowania mające na celu zapobieżenie ujawnieniu danych oraz ochronę przed rozpoznaniem pojedynczych respondentów.

Jednym z głównych problemów w statystyce lokalnej jest określenie dla jakiego obszaru mają być zbierane i publikowane dane. W Programie Statystyki Sąsiedztwa odniesiono się do terminu „sąsiedztwo" (*neighbourhood*), który jest bardzo intuicyjny i nie ma dokładnej definicji. Za sąsiedztwo uważa się ludzkie osiedla oddzielone od innych „naturalnymi" granicami takimi, jak rzeki, ulice, wygląd zabudowań (np. bloki i domki jednorodzinne), czy też prawo wła-

sności (np. własność spółdzielcza i prywatna). W Wielkiej Brytanii, na potrzeby spisu powszechnego w roku 2001, zastosowano podział na tzw. „obszary wynikowe”[44] (*Output Area, OA*). Dla celów statystyki sąsiedztwa utworzono także 34378 „specjalnych obszarów wynikowych” (*super output area,* SOA) składających się z 4 – 6 „obszarów wynikowych”. Liczba mieszkańców SOA liczy średnio 1500 osób. Źródłem danych dla potrzeb Statystyki Sąsiedztwa są rejestry administracyjne, wyniki spisów ludności oraz badania ankietowe (por. tabela 2.2). Na ich podstawie konstruowany jest ujednolicony na obszarze danego kraju historycznego (Anglii, Walii, Szkocji i Irlandii Północnej) wskaźnik wykluczenia społecznego (*indice of deprivation*)[45], opracowany w Centrum Badań nad Nierównościami Społecznymi (*Social Disadvantage Research Centre*) Uniwersytetu w Oxfordzie.

Tabela 2.2. Wybrane źródła danych w Programie Statystyki Sąsiedztwa w ujęciu dziedzin

Lp.	Dziedzina	Źródło danych	Nazwa angielska	Rodzaj źródła danych
1	Ludność	Spis Powszechny 2001	Census 2001	badanie pełne
		Rejestr Ministerstwa Pracy i Emerytur	Department for Work and Pensions (DWP)	rejestr administracyjny
		Rejestr Ministerstwa Finansów	Ministry of Finance	rejestr administracyjny
		Roczne Badanie Ludności	Annual Population Survey	badanie reprezentacyjne
		Badanie Czasu Pracy i Wynagrodzeń	Annual Survey of Hours and Earnings (ASHE)	badanie reprezentacyjne
		Rejestr Ministerstwa Edukacji	Department for Education and Skills (DfES)	rejestr administracyjny
		Rejestr Ministerstwa Spraw Wewnętrznych	Home Office (HO)	rejestr administracyjny
		Rejestr Kancelarii Premiera	Office for the Deputy Prime Minister	rejestr administracyjny
2	Rynek pracy	Spis Powszechny 2001	Census 2001	badanie pełne
		Badanie Zatrudnienia	Annual Employment survey (AES)	badanie reprezentacyjne
		Badanie Aktywności Ekonomicznej Ludności	Annual Local Labour Force Survey (ALALFS)	badanie reprezentacyjne
		Badanie Czasu Pracy i Wynagrodzeń	Annual Survey of Hours and Earnings (ASHE)	badanie reprezentacyjne
		Rejestr Ministerstwa Spraw Konstytucyjnych	Department for Constitutional Affairs (DCA)	rejestr administracyjny

[44] Utworzono je w celu porównywalności wyników między poszczególnymi obszarami, ponieważ hrabstwa (lub, w przypadku Irlandii Północnej, dystrykty) – jednostki administracyjne najniższego rzędu, często bardzo różnią się między sobą. Każdy „obszar wynikowy” składa się ze 125 gospodarstw domowych zamieszkałych przez co najmniej 300 mieszkańców, średnio przez pięciuset.

[45] Ze względu na różnicę w budowie obszarów wynikowych w każdym z krajów historycznych Wielkiej Brytanii, dla każdego z nich tworzy się osobne wskaźniki. Nie są one jednak do końca porównywalne ze sobą (choć trwają prace nad ich ujednoliceniem).

Lp.	Dziedzina	Źródło danych	Nazwa angielska	Rodzaj źródła danych
		Rejestr Ministerstwa Pracy i Emerytur	Department for Work and Pensions (DWP)	rejestr administracyjny
		Rejestr Ministerstwa Zdrowia Publicznego	Health & Safety Executive (HSE)	rejestr administracyjny
		Rejestr Podatkowy	Inland Revenue	rejestr administracyjny
		Rejestr Przedsiębiorstw	Business Registers Unit (BRU)	rejestr administracyjny
		Badanie Mikroprzedsiębiorstw	Small Business Service	badanie reprezentacyjne
3	Szkolnictwo i edukacja	Spis Powszechny 2001	Census 2001	badanie pełne
		rejestr Ministerstwa Edukacji	Department for Education and Skills (DfES)	rejestr administracyjny
		Badanie Aktywności Ekonomicznej Ludności	Annual Local Labour Force Survey (ALALFS)	badanie reprezentacyjne
		Rada Finansowania Szkolnictwa Wyższego w Anglii	Higher Education Funding Council for England	źródło pozaadministracyjne
		Bazy danych Uniwersytetu w Oxford	Oxford University, UCAS	źródło pozaadministracyjne
4	Ochrona zdrowia	Spis Powszechny 2001	Census 2001	badanie pełne
		Rejestr nosicieli chorób przenoszonych drogą płciową i HIV	CDSC (HIV/AIDS) KC60 (STDs)	rejestr administracyjny
		rejestr Ministerstwa Zdrowia Publicznego	Health & Safety Executive (HSE)	rejestr administracyjny
		Rejestr Ministerstwa Edukacji	Department for Education and Skills (DfES)	rejestr administracyjny
		Rejestr wypadków drogowych Ministerstwa Transportu	Department for Transport: Road Casualties	rejestr administracyjny
		Rejestr Ministerstwa Pracy i Emerytur	Department for Work and Pensions (DWP)	rejestr administracyjny
		Narodowy Rejestr Emisji do Atmosfery	National Atmospheric Emissions Inventory (NAEI)	rejestr administracyjny
		Badanie Stanu Mieszkalnictwa	English House Condition Survey (EHCS)	badanie reprezentacyjne
		Rejestr przestępstw Ministerstwa Spraw Wewnętrznych	Home Office notable offences	rejestr administracyjny
		Badanie Stanu Zdrowia	Health Survey for England	badanie reprezentacyjne
5	Mieszkalnictwo	Spis Powszechny 2001	Census 2001	badanie pełne
		Rejestr hipoteczny	Land Registry	rejestr administracyjny
		Badanie stanu mieszkań	The English House Condition Survey	badanie reprezentacyjne
		Rejestr Agencji Pustostanów	The Empty Homes Agency	źródło pozaadministracyjne
6	Przestępczość	Rejestr Ministerstwa Spraw Wewnętrznych	Home Office (HO)	rejestr administracyjny
		Rejestr przestępstw Ministerstwa Spraw Wewnętrznych	Home Office notable offences	rejestr administracyjny
7	Jakość życia	Spis Powszechny 2001	Census 2001	badanie pełne
		Badanie utylizacji odpadów	Municipal Waste Man-	badanie reprezenta-

Lp.	Dziedzina	Źródło danych	Nazwa angielska	Rodzaj źródła danych
		miejskich	agement Survey	cyjne
		Badanie turystyki	United Kingdom Tourism Survey	badanie reprezentacyjne
		Krajowy rejestr wykorzystania terenu	National Land Use Database (NLUD)	rejestr administracyjny
		Rejestr wypadków drogowych Ministerstwa Transportu	Department for Transport: Road Casualties	rejestr administracyjny
		Badanie dochodów i jakości życia	Survey on Income and Life Conditions	badanie reprezentacyjne
		Rejestr Ministerstwa Transportu	Department for Transport (DfT)	rejestr administracyjny

Źródło: na podstawie [*Data Provision for Neighbourhood Renewal*, 2005]

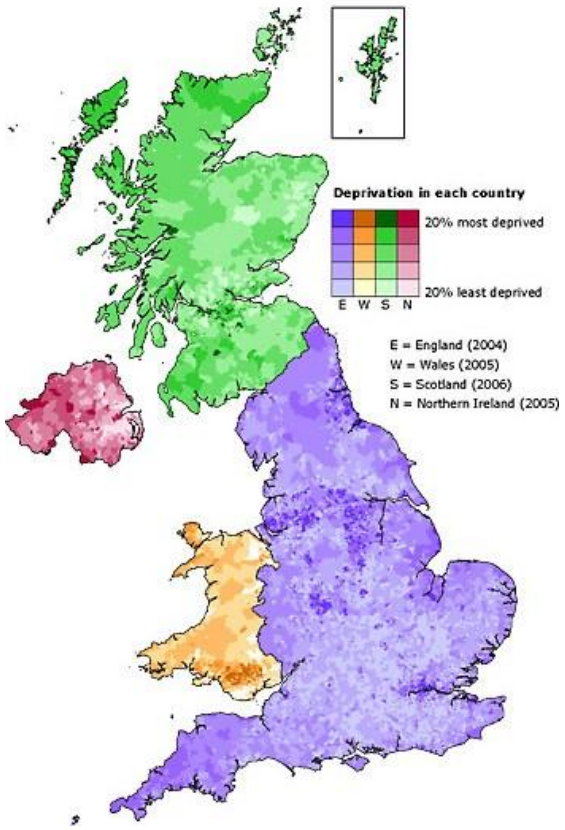
Dane do konstrukcji wskaźnika wykluczenia społecznego obejmują siedem dziedzin: ludność, rynek pracy, edukację, ochronę zdrowia, warunki mieszkaniowe, przestępczość i jakość życia (por. tabela 2.2). W każdym przypadku wykorzystano zbiory danych od wielu gestorów, pochodzące ze źródeł administracyjnych, poza administracyjnych oraz organów statystyki publicznej. Ich harmonizacja i integracja wymagała pracy dużego zespołu, wsparcia ustawodawczego Parlamentu, a także dużych środków finansowych. Integracja przebiegała w sposób deterministyczny.

W Programie Statystyki Sąsiedztwa bardzo ciekawie rozwiązano problem prezentacji wyników badań. Dane dostępne są z poziomu Internetu, gdzie po wpisaniu kodu pocztowego lub zaznaczeniu jednostki administracyjnej (parafii, obwodu pocztowego, hrabstwa itp.) prezentowane są przejrzyste zestawienia wyników. Przyjmują one formę tabelaryczną, wykresów oraz kartogramów (por. rys. 2.1 – 2.2).

Rysunek 2.1. przedstawia obszary wynikowe SOA pogrupowane w kwintylowe grupy. Dla każdego kraju wyniki przedstawianie są innym kolorem, im ciemniejszy odcień, tym na danym obszarze występuje większe natężenie wykluczenia.

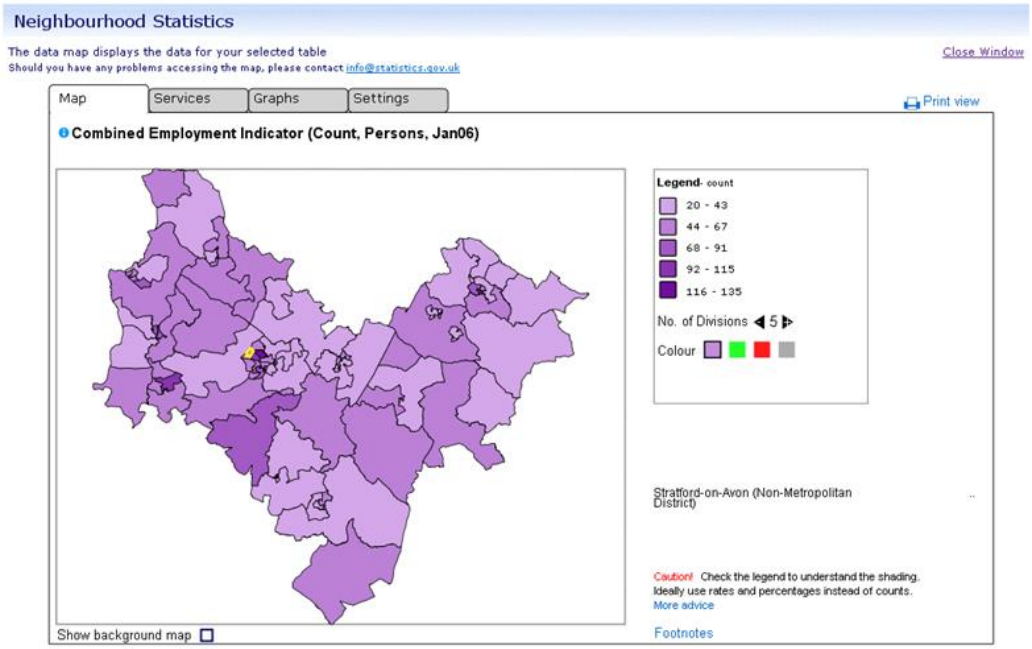
Kartograficzna forma prezentacji wyników, dostępna na stronie internetowej, umożliwia przeprowadzenie analizy porównawczej dla całego regionu (por. rysunek 2.2). Ponadto, dla wybranego obszaru, można uzyskać szczegółowe zestawienia wartości poszczególnych wskaźników. Wyboru obszaru dokonuje się poprzez wpisanie kodu pocztowego, co pozwala na uzyskanie informacji o wartości wskaźnika syntetycznego oraz szczegółowych informacji dla każdej z siedmiu dziedzin życia. Lokalizacja wybranego terytorium na mapie możliwa jest także poprzez wskazanie kursorem. Dodatkowo menu umożliwia dostęp do informacji o metodologii badań, stosowanych pojęciach, definicjach i klasyfikacjach.

Rysunek 2.1. Mapa ubóstwa Wielkiej Brytanii w 2001 roku



Źródło: http://neighbourhood.statistics.gov.uk

Rysunek 2.2. Kartogram zróżnicowania ubóstwa w gminie Stratford-upon-Avon



Źródło: http://neighbourhood.statistics.gov.uk

Wśród podstawowych korzyści Programu Statystyki Sąsiedztwa wymienia się przede wszystkim zwiększenie wiarygodności informacji lokalnych, ułatwienie dostępu do informacji, poprawę jakości danych, rozwój „małych ojczyzn", zwiększenie aktywności władz samorządowych oraz efektywności działań sektora publicznego i prywatnego [*Neigbourhood Statistics Programme Evaluation Report* 2006].

Wyniki otrzymane za pomocą metod zastosowanych w Programie Statystyki Sąsiedztwa są swoistą „bramą" do narodowego zbioru danych statystycznych. Wykorzystane w Programie informacje zostały pozyskane z urzędu statystycznego oraz badań specjalnych towarzyszących Programowi. Połączenie danych z różnych źródeł w jednym systemie spowodowało uproszczenie gromadzenia i przetwarzania informacji statystycznych.

Zastosowanie Programu Statystyki Sąsiedztwa przyniosło w Wielkiej Brytanii szereg korzyści. Udostępniona dokumentacja, jak również metodologia mogłaby znaleźć zastosowanie na terenie innych krajów, w tym oczywiście w Polsce. Do tej pory, w polskiej statystyce, badania małych obszarów (np. poniżej poziomu gminy) nie są rozpowszechnione. Część samorządów przeprowadza takie badania na własną rękę, jednak z powodu braku porównywalności, ich użyteczność jest ograniczona. Przeprowadzenie badań zakrojonych na szeroką skalę, pod patronatem rządu, ujednolicanie dostępnych rejestrów jako źródeł danych, jak również doprowadzenie do zgodności metodologii statystyki sąsiedztwa ze spisem powszechnym na pewno nie będzie procesem tanim ani szybkim. Jednak rozwój metod integracji danych oraz rosnące doświadczenie w ich wykorzystaniu może przyczynić się do budowy analogicznego systemu również w Polsce.

2.2.2. Macierz rachunków społecznych

Macierz rachunków społecznych (*Social Accounting Matrix*, SAM) jest systemem informacji statystycznej zawierającym różne dane społeczno-gospodarcze w zbiorze o strukturze macierzowej. Zawarte w nim są informacje m.in. o: dochodzie na osobę, wydatkach i dochodach podmiotów gospodarczych w podziale na różne kategorie, wzroście gospodarczym.

Utworzenie macierzy rachunków społecznych (*Social Accounting Matrix,* SAM) we Włoszech wpisuje się w system rachunków narodowych zalecanych przez Unię Europejską ESA95. Celem utworzenia macierzy rachunków społecznych było [D'Orazio *et al.* 2006]:

— uporządkowanie informacji o strukturze społeczno-gospodarczej kraju w pewnym (najczęściej rocznym) przekroju czasowym,

— utworzenie statystycznej bazy do tworzenia wiarygodnych modeli ekonomicznych umożliwiających wgląd w stan gospodarki narodowej, jak również symulacji wpływu różnych środków interwencyjnych administracji państwowej na gospodarkę.

Moduł SAM dotyczący gospodarstw domowych (inne moduły dotyczyły m.in. przedsiębiorstw) został podzielony na dwie części: dochody gospodarstw domowych oraz ich rozchody (w podziale na różne kategorie). Informacje na temat dochodów zbierane są w Badaniu Dochodów Gospodarstw Domowych (*Survey on Household Income and Wealth,* SHIW), które we Włoszech przeprowadzane jest przez Narodowy Bank Włoch. Dane o rozchodach natomiast pochodzą z Badania Budżetów Gospodarstw Domowych (*Household Budget Survey,* HBS) przeprowadzanego przez Włoski Urząd Statystyczny. Obie ankiety realizowane są niezależnie przez różne instytucje i służą innym celom. W żadnym z tych badań dochody i rozchody gospodarstw nie są obserwowane łącznie, natomiast wspólny jest szereg cech społeczno-gospodarczych gospodarstw domowych oraz ich członków. Dzięki takiej strukturze danych wejściowych możliwe było rozważenie zastosowania metod parowania statystycznego dla połączenia tych dwóch źródeł danych w celu łącznej analizy przychodów i rozchodów [Coli *et al.* 2005].

Integracja przebiegła w trzech etapach:

— harmonizacja definicji zmiennych i populacji,

— zdefiniowanie modelu integracji,

— aplikacja wybranych metod.

Populacje celu w obu badaniach charakteryzowały się pewnymi różnicami. Choć zbiorowość w obu przypadkach określono jako gospodarstwa domowe we Włoszech, jednostkę zdefiniowano różnorako. W badaniu HBS gospodarstwo domowe określono jako jedną osobę lub zespół osób połączonych więzami małżeństwa, pokrewieństwa, powinowactwa, adopcji, kuratelą lub innymi związkami, niekoniecznie wspólnie się utrzymujących. Natomiast w HBS gospodarstwo zdefiniowano jako jedną lub kilka osób wspólnie się utrzymujących, bez względu na stopień pokrewieństwa lub powinowactwa. Przyporządkowanie tak zdefiniowanych jednostek do ujednoliconej populacji okazało się zadaniem trudnym, jednak w toku badania rozkładów różnych cech uznano, że populacje te są w dużej mierze zbieżne. Przyjęto więc, że analizowane próby są wylosowane z populacji o zgodnej definicji zbiorowości [D'Orazio *et al.* 2006].

Ze względu na różne źródła pochodzenia zbiorów danych, konieczne było także ujednolicenie definicji zmiennych wspólnych. Części z nich nie można było zharmonizować i zostały one

odrzucone z analizy. Pozostałe zmienne zostały tak przekodowane, by ich warianty w obu zbiorach były zgodne, bądź utworzono zunifikowane zmienne pochodne. Szczególną uwagę poświęcono zmiennym opisującym różne charakterystyki głów gospodarstw domowych (płeć, wiek, wykształcenie itp.) gdyż cechuję się one istotną współzależnością z wieloma charakterystykami gospodarstw domowych jako całości.

Schemat 2.3. Model macierzy rachunków społecznych

		Wydatki gospodarstw domowych					Dochody gospodarstw domowych				
		C_1	…	C_u	…	C_U	M_1	…	M_v	…	M_V
Kategorie gospodarstw domowych	T_1	c_{11}	…	c_{1u}	…	c_{1U}	m_{11}	…	m_{1v}	…	m_{1V}
	…	…	…	…	…	…	…	…	…	…	…
	T_w	c_{w1}	…	c_{wu}	…	c_{wU}	m_{w1}	…	m_{wv}	…	m_{wV}
	…	…	…	…	…	…	…	…	…	…	…
	T_W	c_{W1}	…	c_{Wu}	…	c_{WU}	m_{W1}	…	m_{Wv}	…	m_{WV}

Źródło: na podstawie [D'Orazio *et al.* 2006]

Kolumny w macierzy rachunków społecznych reprezentują dwie grupy zmiennych (por. schemat 2.3):

— wydatki gospodarstw domowych: $\boldsymbol{C} = (C_1, \dots, C_u, \dots, C_U)$, gdzie subskrypty $1 \dots U$ reprezentują różne kategorie, np. wydatki na żywność, dobra trwałego użytku itp.,

— dochody gospodarstw domowych: $\boldsymbol{M} = (M_1, \dots, M_v, \dots, M_V)$, gdzie subskrypty $1 \dots V$ reprezentują różne kategorie dochodów, np. wynagrodzenia, dochody z dywident, itp.

Wiersze T_w, $w = 1, \dots, w, \dots, W$ reprezentują różne kategorie gospodarstw domowych, np. wykształcenie głowy gospodarstwa, główne źródła utrzymania, itp.

Zbiór danych HBS zawierała wektor zmiennych społeczno-demograficznych ($\boldsymbol{X}$), wspólnych z analogicznym wektorem w zbiorze SHIW. Dodatkowo dostępne były informacje o wydatkach gospodarstw domowych w ujęciu szczegółowych kategorii oraz informacje o zagregowanych miesięcznych dochodach gospodarstw domowych w ujęciu 14 kategorii. Zbiór SHIW zawierała wektor zmiennych społeczno-demograficznych $\boldsymbol{X}$, szczegółowe informacje o dochodach gospodarstw domowych oraz zagregowane dane opisujące wydatki. Szacunki łącznych rozkładów dochodów i wydatków gospodarstw domowych, próbowano przeprowadzić na podstawie jednego źródła (np. HBS). Okazało się jednak, że szacunki przeprowadzone oddzielnie na podstawie obu zbiorów generowały różne rezultaty dla tych sa-

mych grup zmiennych. Ponadto repozytorium HBS zawierało mało szczegółową informację o źródłach dochodów gospodarstw domowych, natomiast SHIW obejmowało jedynie mocno zagregowane cele wydatków.

Celem utworzenia macierzy rachunków społecznych była łączna obserwacja cech demograficznych$-\boldsymbol{X}$, charakteryzujących dochody - $\boldsymbol{M}$ i wydatki $\boldsymbol{C}$ dla każdej kategorii gospodarstw domowych - T_w. Utworzono model, w którym dokonano następującej dekompozycji:

$$P(\boldsymbol{X}, \boldsymbol{M}, \boldsymbol{C}|T_w) = P(\boldsymbol{C}|\boldsymbol{X}, \boldsymbol{M}, T_w)P(\boldsymbol{X}, \boldsymbol{M}|T_w) \quad (2.1)$$

Łączny rozkład zmiennych $\boldsymbol{X}$ oraz $\boldsymbol{M}$ może zostać oszacowany na podstawie informacji zawartych w SHIW. Określenie łącznego rozkładu $P(\boldsymbol{C}|\boldsymbol{X}, \boldsymbol{M})$ było bardziej problematyczne ze względu ma fakt, że w SHIW kategorie wydatków były mało szczegółowe. Podobny problem dotyczył dochodów w HBS. Postanowiono zatem dołączyć szczegółową informację o dochodach gospodarstw domowych z badania SHIW do HBS, zawierającego szczegółową informację o wydatkach.

Stosując techniki parowania statystycznego, dokonano integracji dwóch źródeł danych nie zawierających informacji o tych samych jednostkach przy wykorzystaniu różnych podejść metodologicznych[46]:

— przy założeniu warunkowej niezależności $\boldsymbol{M}$ i $\boldsymbol{C}$ przy danym $\boldsymbol{X}$, czyli przy założeniu, że wektor cech społeczno-demograficznych tłumaczy wszystkie zależności między dochodami a wydatkami gospodarstw domowych,

— przy wykorzystaniu dodatkowych informacji płynących ze zagregowanych kategorii wydatków w SHIW oraz zagregowanych kategorii dochodów zawartych w HBS.

Uzyskana metodami parowania statystycznego macierz rachunków społecznych umożliwiła szczegółowe szacunki łącznych rozkładów dochodów i wydatków gospodarstw domowych w ujęciu szczegółowych kategorii. Staranna analiza definicji zmiennych i ich wariantów przyczyniła się do wysokiej jakości zintegrowanych danych, zaś aplikacja wielu modeli integracji umożliwiła wszechstronną ocenę otrzymanych szacunków.

2.2.3. Badanie dojazdów do pracy

Badanie „Przepływy ludności związane z zatrudnieniem" przeprowadzono w Ośrodku Statystyki Miast w Urzędzie Statystycznym w Poznaniu przy wykorzystaniu informacji z systemu podatkowego POLTAX. Celem badania było oszacowanie natężenia oraz kierunków dojazdów do pracy osób zatrudnionych na podstawie stosunku pracy według stanu na dzień 31

[46] Podejścia te szczegółowo opisane są w rozdziale czwartym.

grudnia 2006 roku. Estymacja tych informacji umożliwia delimitację przestrzeni społeczno-gospodarczej kraju, a opublikowana na stronie internetowej Urzędu Statystycznego w Poznaniu tzw. macierz dojazdów do pracy może stanowić podstawę pogłębionych analiz dotyczących np. rynku pracy.

Bazę danych POLTAX udostępniło Ministerstwo Finansów. Repozytorium zawierało dane o formularzach podatkowych dotyczących roku 2006, które wpłynęły do urzędów skarbowych. Zasięg terytorialny informacji obejmował cały kraj. Zbiór danych zawierał informacje z następujących zeznań podatkowych [Kruszka 2010]:

- PIT-11/8B – Informacja o dochodach oraz pobranych zaliczkach na podatek dochodowy,
- PIT-40 – Roczne obliczenie podatku od dochodu uzyskanego przez podatnika w roku podatkowym.

W udostępnionych przez gestora zbiorach danych podatkowych (łącznie udostępniono 401 zbiorów z każdego urzędu skarbowego w Polsce) zawartych było 19 mln rekordów obejmujących 67 zmiennych, wśród których można wymienić informacje adresowe płatnika wraz z jego numerem NIP, informacje adresowe podatnika wraz z jego numerem NIP, dane dotyczące przychodu podatnika, kosztów uzyskania przychodów oraz osiągniętego dochodu w danym roku podatkowym.

Populacja generalna badania obejmowała pracowników najemnych[47]. W wyniku harmonizacji obejmującej usunięcie duplikatów oraz weryfikację formalną i logiczną[48], wyjściowy zbiór liczący 19 mln rekordów został zredukowany do około 9,5 mln rekordów. Szczegółowym badaniem objęto osoby będące pracownikami najemnymi dojeżdżającymi do pracy do gminy innej niż miejsce zamieszkania. Wyodrębnienie tej grupy osób możliwe było dzięki informacji w zeznaniu podatkowym o zwiększonych kosztach uzyskania przychodu z tytułu dojazdu do pracy. Liczebność tak zdefiniowanej zbiorowości wynosiła 2,3 mln osób.

Integracja danych z różnych zeznań podatkowych przeprowadzona została na podstawie informacji adresowych, numeru NIP oraz okresu złożenia oświadczenia przez płatnika i podatnika. Zweryfikowano również logiczne połączenia pomiędzy zapisami dotyczącymi przychodów, kosztów ich uzyskania oraz dochodami podatnika. W przypadku występowania jednego podatnika w zbiorze więcej niż jeden raz (wiele rekordów z tym samym numerem

[47] Zbiór zawierał informację również o osobach pracujących w ramach umów cywilno-prawnych (np. umowy zlecenia, umowy o dzieło itp.). Rekordy takie zostały usunięte ze zbioru.

[48] Harmonizacja dotyczyła również momentu referencyjnego badania, który ustalono na 31 grudnia 2006 roku. Z tego względu z udostępnionego zbioru usunięto wszystkie rekordy, w których okres rozliczeniowy nastąpił po tym terminie.

NIP) do badania wybierano rekord, w którym występowała największa wartość przychodu podatnika za 2006 rok.

Wykorzystanie jednostkowych zbiorów danych administracyjnych, których pokrycie równe jest liczebności populacji generalnej umożliwiło publikację rezultatów na niskim, niedostępnym dla „klasycznych" badań reprezentacyjnych poziomie agregacji[49]. Podstawową jednostką terytorialną w badaniu była gmina, przy czym gminy miejsko-wiejskie podzielono na część miejską i wiejską, i każdą z nich traktowano jako oddzielną jednostkę terytorialną (łącznie 3062 jednostek). Dla każdej gminy zebrano informacje o liczbie wyjeżdżających oraz przyjeżdżających do pracy.

W końcowym etapie zweryfikowano zbiór pod kątem poprawności danych adresowych z rejestrem podziału terytorialnego TERYT (głównie dla podatników) oraz Bazą Jednostek Statystycznych (dla płatników).

Schemat 2.4. Macierz przepływów związanych z zatrudnieniem

gmina miejsca zamieszkania i	gmina miejsca pracy j					W_i
	1	2	3	…	m	
1	-	a_{12}	a_{13}	…	a_{1m}	W_1
2	a_{21}	-	a_{23}	…	a_{2m}	W_2
3	a_{31}	a_{32}	-	…	a_{3m}	W_3
…	…	…	…	-	…	…
n	a_{n1}	a_{n2}	a_{n3}	…	-	W_n
P_j	P_1	P_2	P_3	…	P_m	$\sum_{j=1}^{m} P_j = \sum_{i=1}^{n} W_i$

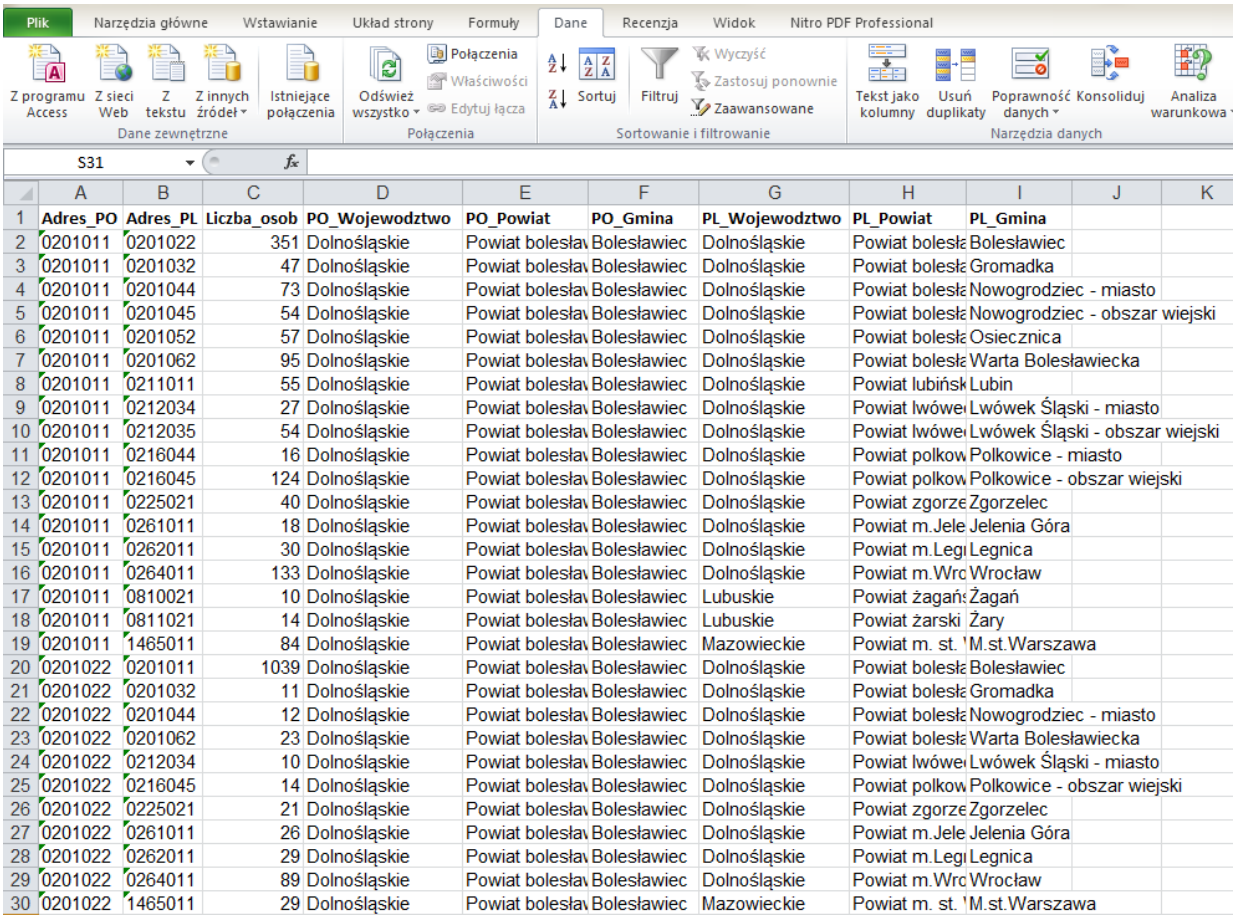
Uwaga:
$i = 1,2,3,\ldots,n$ – numer gminy miejsca zamieszkania,
$j = 1,2,3,\ldots,m$ – numer gminy miejsca pracy,
a_{ij} – liczba osób mieszkających w *i*-tej gminie pracujących w *j*-tej gminie,
$W_i = \sum_{i=1}^{m} a_{ij}$ – liczba osób wyjeżdżających do pracy z *i*-tej gminy,
$P_j = \sum_{j=1}^{n} a_{ij}$ - liczba osób przyjeżdżających do pracy do *j*-tej gminy.
Źródło: na podstawie [Kruszka 2010]

Zharmonizowane dane posłużyły do konstrukcji macierzy przepływów ludności w związku z zatrudnieniem. Macierz ta określa relacje między liczbą osób zamieszkałych w *i*-tej gminie dojeżdżających do pracy w gminie *j*-tej (por. tabela 2.4). Ze względu na uwzględnienie w badaniu wyłącznie ludności dojeżdżającej do pracy do innej gminy, na przekątnej macierzy znajdują się zera.

[49] Badanie przepływów ludności związanych z zatrudnieniem przeprowadzono również m.in. w 2010 roku na podstawie danych BAEL. Ze względu na częściowy charakter badania, wyniki opublikowano na poziomie województw.

Macierz przepływów ludności związanych z zatrudnieniem udostępniona jest na stronie internetowej Urzędu Statystycznego w Poznaniu, na podstronie Ośrodka Statystyki Miast[50] w formacie Microsoft Excel. Wygodna forma udostępnienia wyników badania (por. rysunek 2.3), połączona z rzetelnością i wysoką jakością zebranych informacji umożliwia szerokie wykorzystanie rezultatów prac badawczych do różnego rodzaju analiz.

Rysunek 2.3. Forma udostępnienia wyników „Badania przepływów ludności związanych z zatrudnieniem"



	A	B	C	D	E	F	G	H	I	J	K
1	Adres_PO	Adres_PL	Liczba_osob	PO_Wojewodztwo	PO_Powiat	PO_Gmina	PL_Wojewodztwo	PL_Powiat	PL_Gmina		
2	0201011	0201022	351	Dolnośląskie	Powiat bolesła	Bolesławiec	Dolnośląskie	Powiat bolesła	Bolesławiec		
3	0201011	0201032	47	Dolnośląskie	Powiat bolesła	Bolesławiec	Dolnośląskie	Powiat bolesła	Gromadka		
4	0201011	0201044	73	Dolnośląskie	Powiat bolesła	Bolesławiec	Dolnośląskie	Powiat bolesła	Nowogrodziec - miasto		
5	0201011	0201045	54	Dolnośląskie	Powiat bolesła	Bolesławiec	Dolnośląskie	Powiat bolesła	Nowogrodziec - obszar wiejski		
6	0201011	0201052	57	Dolnośląskie	Powiat bolesła	Bolesławiec	Dolnośląskie	Powiat bolesła	Osiecznica		
7	0201011	0201062	95	Dolnośląskie	Powiat bolesła	Bolesławiec	Dolnośląskie	Powiat bolesła	Warta Bolesławiecka		
8	0201011	0211011	55	Dolnośląskie	Powiat bolesła	Bolesławiec	Dolnośląskie	Powiat lubińsk	Lubin		
9	0201011	0212034	27	Dolnośląskie	Powiat bolesła	Bolesławiec	Dolnośląskie	Powiat lwówe	Lwówek Śląski - miasto		
10	0201011	0212035	54	Dolnośląskie	Powiat bolesła	Bolesławiec	Dolnośląskie	Powiat lwówe	Lwówek Śląski - obszar wiejski		
11	0201011	0216044	16	Dolnośląskie	Powiat bolesła	Bolesławiec	Dolnośląskie	Powiat polkow	Polkowice - miasto		
12	0201011	0216045	124	Dolnośląskie	Powiat bolesła	Bolesławiec	Dolnośląskie	Powiat polkow	Polkowice - obszar wiejski		
13	0201011	0225021	40	Dolnośląskie	Powiat bolesła	Bolesławiec	Dolnośląskie	Powiat zgorze	Zgorzelec		
14	0201011	0261011	18	Dolnośląskie	Powiat bolesła	Bolesławiec	Dolnośląskie	Powiat m.Jele	Jelenia Góra		
15	0201011	0262011	30	Dolnośląskie	Powiat bolesła	Bolesławiec	Dolnośląskie	Powiat m.Legi	Legnica		
16	0201011	0264011	133	Dolnośląskie	Powiat bolesła	Bolesławiec	Dolnośląskie	Powiat m.Wrc	Wrocław		
17	0201011	0810021	10	Dolnośląskie	Powiat bolesła	Bolesławiec	Lubuskie	Powiat żagańs	Żagań		
18	0201011	0811021	14	Dolnośląskie	Powiat bolesła	Bolesławiec	Lubuskie	Powiat żarski	Żary		
19	0201011	1465011	84	Dolnośląskie	Powiat bolesła	Bolesławiec	Mazowieckie	Powiat m. st. '	M.st.Warszawa		
20	0201022	0201011	1039	Dolnośląskie	Powiat bolesła	Bolesławiec	Dolnośląskie	Powiat bolesła	Bolesławiec		
21	0201022	0201032	11	Dolnośląskie	Powiat bolesła	Bolesławiec	Dolnośląskie	Powiat bolesła	Gromadka		
22	0201022	0201044	12	Dolnośląskie	Powiat bolesła	Bolesławiec	Dolnośląskie	Powiat bolesła	Nowogrodziec - miasto		
23	0201022	0201062	23	Dolnośląskie	Powiat bolesła	Bolesławiec	Dolnośląskie	Powiat bolesła	Warta Bolesławiecka		
24	0201022	0212034	10	Dolnośląskie	Powiat bolesła	Bolesławiec	Dolnośląskie	Powiat lwówe	Lwówek Śląski - miasto		
25	0201022	0216045	14	Dolnośląskie	Powiat bolesła	Bolesławiec	Dolnośląskie	Powiat polkow	Polkowice - obszar wiejski		
26	0201022	0225021	21	Dolnośląskie	Powiat bolesła	Bolesławiec	Dolnośląskie	Powiat zgorze	Zgorzelec		
27	0201022	0261011	26	Dolnośląskie	Powiat bolesła	Bolesławiec	Dolnośląskie	Powiat m.Jele	Jelenia Góra		
28	0201022	0262011	29	Dolnośląskie	Powiat bolesła	Bolesławiec	Dolnośląskie	Powiat m.Legi	Legnica		
29	0201022	0264011	89	Dolnośląskie	Powiat bolesła	Bolesławiec	Dolnośląskie	Powiat m.Wrc	Wrocław		
30	0201022	1465011	29	Dolnośląskie	Powiat bolesła	Bolesławiec	Mazowieckie	Powiat m. st. '	M.st.Warszawa		

Uwaga:
Adres_PO – kod TERYT gminy zamieszkania,
Adres_PL – kod TERYT gminy zatrudnienia,
Liczba_osob – liczba osób dojeżdżających z gminy i do gminy *j*,
PO_Wojewodztwo – nazwa województwa zamieszkania,
PO_Powiat – nazwa powiatu zamieszkania,
PO_Gmina – nazwa gminy zamieszkania,
PL_Wojewodztwo – nazwa województwa zatrudnienia,
PL_Powiat – nazwa powiatu zatrudnienia,
PL_Gmina – nazwa gminy zatrudnienia.
Źródło: Ośrodek Statystyki Miast Urzędu Statystycznego w Poznaniu

[50] http://www.stat.gov.pl/cps/rde/xbcr/poznan/ASSETS_Dojazdy_internet.xls (dostęp z dnia 31 grudnia 2012 roku)

Publikacja macierzy przepływów ludności umożliwiła badanie zależności przestrzennych natężenia dojazdów do pracy oraz cech lokalnych rynków pracy [Gołata *et al.* 2011; Gruchociak 2010, 2012; Filas-Przybył *et al.* 2012; Musiał, Roszka 2012]. Wykorzystując inne dostępne informacje na poziomie powiatów, jak bezrobocie rejestrowane i liczba ludności w wieku produkcyjnym, zdiagnozowano silne oddziaływanie dużych miast oraz wyodrębniono strefy ich wpływu. Natomiast analizy geostatystyczne bazujące na newtonowskiej koncepcji grawitacji umożliwiły określenie zależności pomiędzy gęstością sieci transportowej a stopniem grawitacji jednostek przestrzennych.

2.2.4. Inne badania społeczne

W krajach zachodnich integracja danych jest coraz bardziej popularną metodą wspomagania badań statystycznych. Stosunkowo niskie koszty, bogactwo informacyjne rejestrów, ich duża liczba i odpowiednie rozwiązania prawne – wszystko to sprzyja nowemu podejściu do statystyki publicznej. Coraz nowocześniejsze rozwiązania metodologiczne i informatyczne prowadzą do tworzenia zaawansowanych systemów wspomagających integrację danych.

Przykładem zastosowania statystycznej metody integracji danych było przeprowadzone w Stanach Zjednoczonych łączenie Krajowego Rejestru Dochodów i odpowiadającego mu rejestru płac z rejestru przedsiębiorstw [Steel, Konschnik 1997]. Proces łączenia probabilistycznego był procesem pomocniczym dla przeprowadzonego wcześniej łączenia deterministycznego wykorzystującego Numer Identyfikacyjny Pracodawcy (*Employer Identyfication Number*, EIN). Numer ten posiadało jednak zaledwie 30% jednostek biorących udział w badaniu. Pozostałe 70% należało połączyć w sposób stochastyczny. Ze względu na to, że integrowane zbiory zawierały informacje o tych samych jednostkach, zastosowano metodę probabilistycznego łączenia rekordów. Do procesu integracji wykorzystano następujące zmienne:

- Nazwa – zmienna zawierająca nazwisko,
- Adres – nazwa ulicy, miasto, stan, kod pocztowy,
- Klasyfikacja działalności – odpowiednik PKD,
- Roczna lista płac – wypłaty i koszty pracy.

Kryterium włączenia danego rekordu do analizy było posiadanie co najmniej sześciu liter nazwiska oraz pierwszej litery imienia. Kryterium połączenia pary rekordów była wartość sumy wag łączonych zmiennych. Zmienne wykorzystane w procesie łączenia podzielono na trzy grupy: nazwa, adres oraz branża. Niska wartość wagi dla nazw dyskwalifikowała rekordy

jako połączenie. W pozostałych grupach zmiennych, niska waga dla jednej cechy wymagała, by druga cecha posiadała wysoką wagę. Użycie parsera wykluczyło z analizy 0,16% jednostek. W procesie deduplikacji, pozostawiano rekord o najwyższej wadze połączeniowej. Pozostałe rekordy usuwano. Zintegrowane repozytorium służyło do określenia różnych zależności na rynku pracy.

Jednym z przeprowadzonych w Polsce badań, w których wykorzystano zintegrowane administracyjne zbiory danych było połączenie zbiorów Krajowego Rejestru Nowotworów (KRN) oraz bazy zgonów z powodu chorób nowotworowych z repozytoriów GUS [Borkowski, Mielniczuk 2003]. Ze względu na usunięcie danych osobowych (imienia, nazwiska oraz miesiąca w dacie urodzenia), jako zmienne parujące wybrano: dzień urodzenia, rok urodzenia, płeć, rodzaj nowotworu oraz województwo zamieszkania (49 województw wg podziału administracyjnego sprzed 1999 roku). Tak utworzony klucz nie był unikatowy zarówno w zbiorze KRN jak i GUS, co było przyczyną wprowadzenia modyfikacji do metody deterministycznej. Modyfikacja polegała na losowaniu rekordu w sytuacji powtórzeń tej samej wartości klucza. Otrzymano 68,1% zgodności zgonów oraz 58,7% zgodności daty. Jednak ze względu na brak unikatowości klucza rekordów, wyników nie rozpatrywano jednostkowo, a jedynie wnioskowano dla populacji.

Techniki parowania statystycznego są szeroko wykorzystywane także w badaniach marketingowych i rynkowych (w tej branży parowanie statystyczne nazywane jest *data fusion*), zwłaszcza przez duże grupy, jak GfK [Raessler 2002; Wildner, Scherübl 2006] i AC Nielsen [Soong, de Montigny 2001]. Według Raessler [2002] techniki parowania statystycznego mają wręcz swój początek w badaniach rynku – poprzez zainteresowanie badaczy marketingowych zachowaniem konsumentów w celu ulepszenia sposobów docierania różnego rodzaju mediów (głównie telewizji) do potencjalnych klientów (por. tabela 2.5).

Schemat 2.5. Integracja danych w badaniach marketingowych i rynkowych

ATRYBUT	Panel gospodarstwa domowego	Panel telewizyjny	Sparowany plik
nr jednostki	13	425	425
płeć	K	K	K
wiek	35 – 40	35 – 40	35 – 40
wykształcenie	wyższe	wyższe	wyższe
stan cywilny	zamężna	rozwiedziona	rozwiedziona
przychody netto	3500 – 4000	3000 – 3500	3000 – 3500
miejsce zamieszkania	dom wolnostojący	dom wolnostojący	dom wolnostojący
zwierzęta domowe	tak	tak	tak
ilość nabywanych płatków	1 kg/tydz.	**pytania niezadane**	1 kg/tydz.
ilość nabywanego wina	3 l/tydz.		3 l/tydz.
ilość nabywanego mięsa	2 kg/tydz.		2 kg/tydz.
czy wypożycza samochód	**pytania niezadane**	nie	nie
czy ogląda opery mydlane		nie	nie
czy ogląda wiadomości		regularnie	regularnie
czy przełącza kanał w czasie, gdy nadawane są reklamy		tak	tak

Źródło: na podstawie Raessler [2002]

Przeprowadzenie badań o szerokim zakresie tematycznym, zawierających wszystkie potrzebne informacje było problematyczne ze względu na koszty, czas i obciążenie respondentów. Wykorzystanie technik parowania statystycznego w panelowych badaniach rynku umożliwiło nie tylko obniżenie kosztów i obciążenia respondentów, ale również zwiększyło zasób informacji pochodzących z różnych badań tematycznych.

W GfK techniki statystycznej integracji danych stosuje się głównie w celu łączenia informacji z panelowych badań gospodarstw domowych i tzw. paneli telewizyjnych [Wildner, Scherübl 2006], tj. badań, których głównym celem jest pomiar skuteczności reklam telewizyjnych. Połączenie informacji o zachowaniach konsumenckich gospodarstw domowych oraz zwyczajach związanych z oglądaniem telewizji może umożliwić pomiar i śledzenie wpływu reklam na zmianę popytu na określoną grupę produktów i usług [Wildner 2000].

2.3. Badania przedsiębiorstw – projekt MEETS

Projekt MEETS (*Modernisation of European Enterprise and Trade Statistics*) był programem Komisji Europejskiej zapoczątkowanym decyzją nr 1297/2008/EC z dnia 16 grudnia 2008 roku. Przesłankami jego powołania były m.in.:

— dynamiczne przemiany zachodzące w gospodarce oraz potrzeba podążania sprawozdawczości statystycznej za tymi zmianami,
— potrzeba ujednolicenia międzynarodowej metodyki badań nad gospodarką i przedsiębiorczością,
— potrzeba wykorzystania danych administracyjnych w statystyce przedsiębiorstw.

Celem projektu było [*Wykorzystanie danych administracyjnych w statystyce przedsiębiorstw* 2011]:

— zbadanie użyteczności dostępnych źródeł administracyjnych dla krótkookresowej i rocznej statystyki przedsiębiorstw w odniesieniu do podmiotów gospodarczych o liczbie pracujących przekraczającej 9 osób,
— zmniejszenie wynikających ze sprawozdawczości statystycznej, obciążeń podmiotów gospodarczych o liczbie pracujących większej niż 9 osób,
— zwiększenie efektywności prowadzonych szacunków,
— zastosowanie kalibracji w przypadku braków danych,
— zastosowanie nowych metod estymacji pośredniej (statystyka małych obszarów (SMO)) w celu poprawy precyzji szacunku,
— rozszerzenie zakresu przekrojów, w których prezentowane są wyniki badania.

Prace nad projektem w Polsce prowadzone były w urzędach statystycznych w Poznaniu i Katowicach, przy wsparciu merytorycznym pracowników Uniwersytetu Ekonomicznego w Poznaniu.

W ramach projektu zrealizowano trzy obszary tematyczne:

1. Diagnoza i analiza dostępnych źródeł administracyjnych w celu wykorzystania ich w statystyce przedsiębiorstw, a w szczególności:
    a. System podatkowy
        i. podatek dochodowy od osób fizycznych, ryczałt od przychodów ewidencjonowanych, karta podatkowa – PIT,
        ii. podatek dochodowy od osób prawnych – CIT,
        iii. podatek od towarów i usług – VAT,
        iv. Krajowa Ewidencja Podatników – KEP.
    b. System ubezpieczeń społecznych
        i. Centralny Rejestr Płatników Składek (CRPS),
        ii. Centralny Rejestr Ubezpieczonych (CRU).

2. Zastosowanie metod statystyki małych obszarów celem zwiększenia precyzji szacunków rocznej i krótkookresowej statystyki przedsiębiorstw. Dodatkowo prace w tym obszarze tematycznym miały na celu umożliwienie przedstawiania szacunków w dotychczas niedostępnych przekrojach (np. informacje o przedsiębiorstwach wg klasyfikacji PKD w ujęciu przestrzennym).
3. Wykorzystanie informacji pochodzących z różnych źródeł w celu zastosowania metod kalibracyjnych do korekty błędów wynikających z braków danych w badaniach przedsiębiorstw.

W ramach pierwszego obszaru tematycznego dokonano przeglądu dostępnych statystycznych repozytoriów danych opisujących przedsiębiorczość w Polsce, takich jak:

- DG-1 „Meldunek o działalności gospodarczej",
- F-01/I-01 „Sprawozdanie o przychodach, kosztach i wyniku finansowym oraz o nakładach na środki trwałe",
- SP „Roczna ankieta przedsiębiorstwa",
- F-02 „Statystyczne sprawozdanie finansowe",
- Z-06 „Sprawozdanie o pracujących, wynagrodzeniach i czasie pracy",
- Statystyczny rejestr przedsiębiorstw,

a także rejestry administracyjne systemów podatkowego i ubezpieczeń społecznych. Zadaniem przeglądu była eksploracja oraz weryfikacja zawartości informacyjnej analizowanych źródeł, ich porównywalności, zgodności definicji i klasyfikacji między źródłami, momentów referencyjnych zbiorów, a docelowo ocena możliwości wykorzystanie zintegrowanych źródeł w sprawozdawczości statystycznej. Dodatkowo dokonano oceny systemów administracyjnych pod kątem zawartości informacyjnej o rynku pracy oraz przychodach i kosztach działalności oraz podatków.

W ramach działań w drugim obszarze tematycznym wybrano listę zmiennych ze wszystkich źródeł administracyjnych mogących służyć jako cechy pomocnicze w estymacji pośredniej (szczegółowo zagadnienie to opisuje [Gołata 2012]). W szczególności, w systemie podatkowym wyspecyfikowano zmienne dotyczące dochodów, przychodów, podatków oraz miejsca zamieszkania podatnika (siedziby podmiotu). W systemie ubezpieczeń społecznych uzyskano zmienne dotyczące rynku pracy. Opis wybranych zbiorów stanowiących system ubezpieczeń społecznych wraz z metodą pozyskiwania informacji o rynku pracy zostanie opisany w rozdziale trzecim.

Trzeci obszar tematyczny zrealizowano identyfikując populacje obejmujące poszczególne rejestry, ich liczebności, a także wyszczególniając zmienne mogące służyć jako podstawa tworzenia równań kalibracyjnych. Sformułowano również wytyczne dotyczące sposobu i terminów przekazywania odpowiednich zbiorów danych służbom statystyki publicznej w celu regularnego tworzenia odpowiednich komunikatów statystycznych.

Finalnym efektem prac w projekcie MEETS była zintegrowane repozytorium danych [Dehnel, Gołata 2012; Gołata, Dehnel 2012]. Na podstawie dostępnych zmiennych zawierających wartości numerów NIP oraz REGON połączono w sposób deterministyczny administracyjne zbiory danych także z repozytoriami pochodzącymi ze źródeł statystycznych. Dokonano deduplikacji rekordów zawierających ten sam numer identyfikacyjny. Rekordy, dla których nie był dostępny żaden z kluczy połączeniowych nie zostały połączone. Odsetek niepołączonych rekordów wahał się od 1,4% w województwach lubelskim i lubuskim do aż 13,5% w województwie mazowieckim. Dzięki harmonizacji źródeł danych utworzony został rejestr statystyczny o szerokim spektrum informacyjnym oraz dużym pokryciu, całej badanej populacji przedsiębiorstw.

W ostatnim kroku dokonano ewaluacji zintegrowanego zbioru poprzez ocenę rozkładu wybranych zmiennych oraz modelowanie relacji między nimi. Zauważono pewne rozbieżności w rozkładzie analizowanych zmiennych w ujęciu różnych źródeł. Również analizując relacje między tymi samymi zmiennymi z różnych rejestrów zaobserwowano pewne odchylenia. Na przykład, zaledwie około 40% jednostek wykazywało podobne wartości w każdym ze źródeł. Dodatkowo okazało się, że prawie wszystkie przedsiębiorstwa w sprawozdawczości statystycznej DG-1 wykazują niższe przychody niż w zeznaniach podatkowych.

Jako postulaty wynikające z badań i analiz związanych z projektem MEETS wyszczególniono przede wszystkim [Dehnel, Gołata 2012]:

- poprawę kompletności rejestrów zarówno pod względem merytorycznym, jak i pokrycia,
- ustalenie regularnych terminów oraz formy przekazywania rejestrów służbom statystyki publicznej,
- wyodrębnienie w ramach statystyki publicznej komórek, których zadaniem byłaby regularna współpraca z gestorami danych administracyjnych.

Wyróżniono również szereg możliwości związanych z wykorzystaniem rejestrów administracyjnych w statystyce przedsiębiorstw, wśród których wymienia się m.in.:

— regularne i rzetelne źródło informacji o liczbie i szczegółowej charakterystyce podmiotów gospodarczych,
— ocenę efektów gospodarczych,
— analizę terytorialnego zróżnicowania rozwoju gospodarczego.

W ramach projektu MEETS wykonano szereg badań symulacyjnych, których celem była weryfikacji możliwości wykorzystania metod statystyki małych obszarów, w tym kalibracji, przy wykorzystaniu zmiennych z rejestrów, do tworzenia szacunków na potrzeby statystyki publicznej [*Wykorzystanie danych administracyjnych w statystyce przedsiębiorstw* 2011]. Pokazały one dużą przydatność wspomagania dotychczasowej sprawozdawczości statystycznej danymi pochodzenia administracyjnego.

2.4. Projekty Eurostatu

2.4.1. CENEX-ISAD

W 2005 r. Eurostat wprowadził ideę utworzenia Europejskich Centrów Rozwoju (*European Centres and Networks of Excellence,* CENEX) w dziedzinie statystyki, jako sposób wzmocnienia współpracy między urzędami statystycznymi w Europie. Jedną z dziedzin projektu CENEX była integracja administracyjnych i ankietowych źródeł danych (*Integration of Survey and Administrative Data,* ISAD). Integracja różnych źródeł danych zyskuje coraz większą popularność w krajowych urzędach statystycznych [*Description of the action* 2006]. Wiele państw członkowskich ESS (*European Statistical System*) stosuje obecnie lub zamierza stosować metody integracji w różnych dziedzinach, jak przeprowadzanie wirtualnych spisów powszechnych, badanie demografii przedsiębiorstw czy konstruowanie macierzy rachunków społecznych. Pomimo rozwoju innych ważnych zagadnień metodologicznych (jak metody doboru próby, analiza i przetwarzanie braków danych, jakość danych), integracja różnych źródeł jest stosunkowo nowym problemem i istnieje potrzeba rozbudowy metodologicznych podstaw. Projekt CENEX-ISAD miał służyć państwo-członkom ESS jako platforma wymiany doświadczeń i wiedzy w dziedzinie integracji danych. Cel ten realizowano poprzez rozpowszechnianie dobrych praktyk, jak również projekty szkoleniowe i warsztaty dla pracowników naukowych i operacyjnych urzędów statystycznych.

Projekt trwał od grudnia 2006 do czerwca 2008 roku, a jego beneficjentami były urzędy statystyczne z Włoch (koordynator), Holandii, Hiszpanii, Austrii oraz Czech.

Integrację danych w projekcie rozumiano jako łączenie dwóch lub więcej źródeł w jedno repozytorium o zwiększonej, poprzez łączną obserwację cech ze wszystkich źródeł, zawartości informacyjnej. Podejścia metodologiczne rozpatrywano w dwóch grupach:

— łączenie rekordów (*record linkage*) – dla integracji repozytoriów zawierających informacje o tych samych jednostkach:
    — deterministyczne łączenie rekordów (*exact record linkage*) – w przypadku dostępności unikalnego klucza połączeniowego,
    — probabilistyczne łączenie rekordów (*probabilistic record linkage*) – w przypadku braku dostępności unikalnego klucza połączeniowego;
— parowanie statystyczne (*statistical matching*) – dla integracji repozytoriów nie zawierających informacji o tych samych jednostkach, ale odnoszących się do tych samym populacji generalnych.

Dodatkowo przedmiotem rozważań było zagadnienie przetwarzania danych zintegrowanych (*micro integration processing*) mające na celu zapewnienie wysokiej jakości danych statystycznych pochodzących z połączonych źródeł danych.

Projekt został podzielony na cztery główne pakiety robocze (*work package*, WP) oraz jeden roboczy pakiet administracyjny:

— WP0 – Koordynacja i zarządzanie projektem,
— WP1 – Obecny stan wiedzy w dziedzinie integracji danych,
— WP2 – Opracowanie dobrych praktyk w dziedzinie integracji danych administracyjnych i ankietowych,
— WP3 – Oprogramowanie służące integracji,
— WP4 – Rozwój narzędzi komunikacyjnych oraz podnoszenie świadomości ważności metod integracji w ESS.

W ramach pakietu WP1 dokonano przeglądu najnowszej literatury z zagadnień probabilistycznego łączenia rekordów, parowania statystycznego oraz przetwarzania danych zintegrowanych. Dodatkowo w ramach prac nad tym pakietem roboczym, opisano wybrane doświadczenia krajów biorących udział w projekcie w dziedzinie metod statystycznej integracji danych [*Report of WP1. State of the art on statistical methodologies for integration of surveys and administrative data* 2008]. Wśród uczestników projektu przeprowadzono również badanie ankietowe, którego celem była weryfikacja zakresu i formy wykorzystywania oraz rozwoju metod statystycznej integracji danych w praktyce urzędów statystycznych państw – beneficjentów projektu.

Efektem prac w ramach pakietu roboczego WP2 było wypracowanie dobrych praktyk w dziedzinie harmonizacji źródeł danych, analizowanych metod statystycznej integracji danych oraz przetwarzania danych zintegrowanych [*Report of WP2. Recommendations on the*

use of methodologies for the integration of surveys and administrative data 2008]. Dodatkowo, każdy z uczestniczących w projekcie urzędów opracował studia przypadków dla opisywanych w dokumencie zagadnień.

Wynikiem realizacji pakietu WP3 był przegląd dostępnych komercyjnych i niekomercyjnych programów komputerowych służących implementacji metod probabilistycznego łączenia rekordów i parowania statystycznego w praktyce [*Report of WP3. Software tools for integration methodologies* 2008]. Jednocześnie opisano najpopularniejsze narzędzia programistyczne służące poprawie jakości danych (SAS, Oracle, Netrics). Do użycia zarekomendowano, w przypadku probabilistycznego łączenia rekordów, program *fuzzDateStat 19* będący makropoleceniem napisanym w środowisku SAS. Dla integracji danych metodą parowania statystycznego zaproponowano program *SAMWIN* [Sacco 2008], który powstał w środowisku C++ dla potrzeb tworzenia macierzy rachunków społecznych (stąd nazwa: SAM – ***S**ocial **A**ccounting **M**atrix*, WIN – przeznaczony do pracy w środowisku Microsoft Windows).

Pakiet WP4 zrealizowano organizując szkolenie z metod statystycznej integracji danych, które odbyło się w Budapeszcie w dniach 14 – 16 listopada 2007 roku. Projekt zwieńczono konferencją podsumowującą prace, która odbyła się w Wiedniu w dniach 29 – 30 maja 2008 roku.

2.4.2. ESSnet on Data Integration

Doceniając rosnącą potrzebę łączenia repozytoriów danych z różnych źródeł, Eurostat przyznał kolejny grant poświęcony metodologii statystycznej integracji danych. Projekt *ESSnet on Data Integration* w założeniu miał być rozwinięciem prac w projekcie CENEX-ISAD. Celem projektu *Data Integration* były:

- opracowanie metod oceny jakości zintegrowanych źródeł,
- opracowanie i udoskonalenie metod integracji (probabilistycznego łączenia rekordów, parowania statystycznego, przetwarzania danych zintegrowanych) w celu ich stosowania w codziennej praktyce urzędów statystycznych.

W porównaniu do CENEX-ISAD, *Data Integration* był projektem bardziej rozbudowanym, trwającym dłużej, angażującym więcej państw-beneficjentów oraz obejmującym więcej działań.

Projekt trwał 24 miesiące – od stycznia 2010 do grudnia 2011 roku. Uczestniczyły w nim urzędy statystyczne z 6 państw: Włochy (koordynator), Holandia, Polska[51], Hiszpania, Szwajcaria oraz Norwegia. Składał się z sześciu pakietów roboczych:

- WP1 – Obecny stan wiedzy w dziedzinie integracji danych (uaktualnienie informacji zgromadzonych w WP1 CENEX-ISAD),
- WP2 – Rozwój metod statystycznej integracji danych,
- WP3 – Rozwój narzędzi informatycznych służących integracji,
- WP4 – Studia przypadków,
- WP5 – Rozpowszechnienie wiedzy z dziedziny integracji danych w państwach członkowskich ESS,
- WP6 – Zarządzanie projektem.

W ramach uaktualniania stanu wiedzy z dziedziny integracji danych dokonano przeglądu literatury z lat 2008[52] - 2011 w dziedzinach probabilistycznego łączenia rekordów, parowania statystycznego oraz przetwarzania danych zintegrowanych [*Report on WP1. State of the art on statistical methodologies for data integration* 2011]. Dodatkowo zwrócono uwagę na tzw. błąd ekologiczny[53] w parowaniu statystycznym oraz poddano szczególnej analizie problem zachowania tajemnicy statystycznej w zintegrowanych zbiorach.

W ramach realizacji pakietu WP2, opisano praktyczne zastosowanie metod integracji danych w pracy urzędów statystycznych [*Report on WP2. Methodological developments* 2011]. W szczególności zwrócono uwagę na podejście bayesowskie w probabilistycznym łączeniu rekordów, edycję błędów wynikających z niewłaściwych „dopasowań" rekordów w integracji, a także opisano sposoby radzenia sobie z różnymi definicjami jednostek, zmiennych i okresów referencyjnych w integrowanych zbiorach. Jednocześnie przedstawiono podejście bootstrapowe w szacunkach oraz modele i techniki w przetwarzaniu danych zintegrowanych.

W ramach rozwoju narzędzi informatycznych w integracji danych, opisano nowe programy umożliwiające automatyczną integrację danych różnymi technikami. Na potrzeby probabilistycznego łączenia rekordów zaprezentowano oprogramowanie RELAIS 2.3 (***Re**cord **L**inkage **A**t **IS**tat*). Jest to program opracowany we Włoskim rzędzie Statystycznym w środowisku

[51] W pracach brał udział Główny Urząd Statystyczny oraz Urząd Statystyczny w Poznaniu. Koordynatorem projektu z polskiej strony był dr Marcin Szymkowiak, konsultant w Ośrodku Statystyki Małych Obszarów w Urzędzie Statystycznym w Poznaniu oraz pracownik naukowo-dydaktyczny Uniwersytetu Ekonomicznego w Poznaniu.

[52] Przegląd obejmował lata po zakończeniu projektu CENEX-ISAD.

[53] Wnioskowanie o zależnościach na poziomie jednostkowym w oparciu o dane zagregowane. Problem został opisany w rozdziale 4.

programistycznym Java oraz R [Scannapieco *et al.* 2010]. Na potrzeby integracji metodą parowania statystycznego zaprezentowano pakiet programu statystycznego R o nazwie *StatMatch*. Umożliwia on wykorzystanie darmowego środowiska analitycznego R oraz zastosowanie szeregu technik parowania rekordów [D'Orazio 2011].

W części poświęconej studiom przypadku, opisano sześć doświadczeń urzędów statystycznych w dziedzinie integracji [*Report on WP4 Case studies* 2011]. Przedstawiono metody wykorzystania rejestrów administracyjnych do szacowania charakterystyk rynku pracy. Następnie opisano zagadnienie jakości danych w zintegrowanych źródłach na przykładzie szacunków dotyczących zatrudnienia. Kolejnym doświadczeniem była integracja rejestrów administracyjnych i badań reprezentacyjnych w celu poprawy szacunków dotyczących wykształcenia ludności. Omówiono problem błędów występujących w zintegrowanych źródłach, jak również przedstawiono polskie doświadczenia w dziedzinie parowania statystycznego. W tej sekcji opisano integrację zbiorów danych Mikrospisu 1995 z Badaniem Aktywności Ekonomicznej Ludności z tego samego roku [Roszka 2011]. Celem analizy było weryfikacja możliwości zastosowania technik statystycznej integracji danych na polskich źródłach.

Rozpowszechniane wiedzy dotyczącej metod i technik integracji danych zrealizowano poprzez cztery grupy działań: spotkania, szkolenia, kurs oraz warsztaty. W ramach realizacji WP5 odbyło się pięć spotkań. Ich celem była wymiana doświadczeń oraz wiedzy. Przeprowadzono również trzy szkolenia. Pod koniec trwania projektu przeprowadzono kurs obejmujący wszystkie poruszane w projekcie zagadnienia. Miał on na celu ukazanie problemów wynikających z rosnących potrzeb informacyjnych europejskich społeczeństw oraz sposobów ich rozwiązania metodami integracji. Projekt zwieńczyły warsztaty odbywające się w Madrycie w listopadzie 2011 roku. Zaproszeni prelegenci (Polskę reprezentowała m.in. Gołata [2011]) wygłosili referaty dotyczące doświadczeń i problemów wynikających ze stosowania metod integracji danych.

Zadaniem pakietu WP6 było zapewnienie właściwego zarządzania Projektem i osiągnięcia jego celów zgodnie z jej harmonogramem i budżetem.

2.5. Wnioski

Metody integracji są już dość szeroko wykorzystywane w praktyce, a systemy informacyjne konstruowane w oparciu o tę metodologię są rozbudowane i pełnią coraz ważniejszą rolę społeczną. Instytucje wykorzystujące metody integracji danych podkreślają zalety związane z ograniczeniem kosztów i czasu przeprowadzania badań, zmniejszenie obciążenia respondentów, zwiększenie jakości danych oraz korzyści związane z nowymi możliwościami esty-

macji (np. na niższym poziomie agregacji przestrzennej, publikowanie danych w wielu wymiarach jednocześnie bez straty na jakości estymatorów). Uwypuklają również problemy wynikające z wdrożenia systemu statystyki opartego na zintegrowanych źródłach, niepełnej zgodności danych z systemów administracyjnych z wymaganiami statystyki publicznej i związaną z tym konieczność harmonizacji. W zdecydowanej większości integracji dokonuje się w sposób deterministyczny. Wykorzystanie tego podejścia gwarantuje prawdziwość połączeń rekordów w wielu zbiorach. W przypadku konieczności wnioskowania na podstawie zmiennych ze źródeł o niskim pokryciu wykorzystuje się głównie znane i dobrze opisane w literaturze techniki kalibracji. Integracja stochastyczna nadal jest na etapie testowania własności i możliwości wykorzystania w komunikatach statystycznych. Jednak coraz większa liczba dostępnych źródeł danych, również z badań reprezentacyjnych sprawia, że metodologia statystycznej integracji danych zacznie odgrywać coraz większą rolę w pracach urzędów statystycznych i firm badawczych. Coraz bardziej ujednolicana (zwłaszcza przez organy statystyki publicznej) metodologia badań częściowych, coraz częstsze wykorzystanie zharmonizowanych źródeł administracyjnych, a także konieczność oszczędności czasu i kosztów mogą przyczynić się do popularyzacji i dalszego rozwoju metod statystycznej integracji danych.

W Polsce przeprowadzanych jest wiele badań reprezentacyjnych, zarówno przez Główny Urząd Statystyczny, jak i inne instytucje publiczne i prywatne. Opisują one różne dziedziny życia społeczno-gospodarczego, ale niejednokrotnie przedmiot ich badań częściowo się pokrywa. Żadne z tych źródeł oddzielnie nie zapewnia pełnego opisu zjawisk społecznych, a ograniczona liczebność próby uniemożliwia szacunek w ujęciu regionalnym czy lokalnym. Zasadnym wydaje się więc podjęcie próby ich integracji celem zwiększenia merytorycznego i przestrzennego zakresu szacunków oraz ich precyzji . Istnieje również wiele rejestrów administracyjnych, które mogłyby dostarczyć dodatkowych informacji w szczegółowym przekroju terytorialnym. Ich zawartość informacyjna i możliwości wykorzystania są przedmiotem badań Głównego Urzędu Statystycznego.

W rozdziale III opisane zostaną wybrane rejestry administracyjne udostępnione organom statystyki publicznej w celu rozpoznania ich przydatności. Opisane zostaną również wybrane badania reprezentacyjne, ich zawartość merytoryczna, techniki doboru próby oraz pomiaru. Celem tego opisu będzie identyfikacja zbiorów danych, których integracja byłaby możliwa. Jednocześnie zostanie dokonana próba oceny jakości danych oraz ich dostępności, by w dalszej kolejności podjąć próbę ich integracji za pomocą wybranych metod.

ROZDZIAŁ III. POTENCJALNE ŹRÓDŁA DANYCH DLA BADAŃ OPARTYCH NA INTEGRACJI

Dane gromadzone na potrzeby administracyjne i sprawozdawczości statystycznej w Polsce są analogiczne do zbieranych w innych krajach. Składają się na nie różnego rodzaju rejestry administracyjne, badania reprezentacyjne oraz badania przeprowadzane przez inne organy niż statystyka publiczna. Zawartość informacyjna badań reprezentacyjnych jest ogólnie znana. Zobowiązania międzynarodowe, duża liczba publikacji oraz udostępnianie jednostkowych (odpersonalizowanych) zbiorów danych m.in. ośrodkom naukowym przyczyniły się do rozpowszechnienia metodologii pomiaru, analizy i publikacji wyników.

Dostęp do rejestrów administracyjnych, nawet tych udostępnionych organom statystyki publicznej jest utrudniony. Wynika to głównie z konieczności zachowania tajemnicy statystycznej (rejestry często zawierają informacje osobowe). Utrudnienia w dostępie sprawiają, że rozpoznanie zawartości informacyjnej rejestrów, weryfikacja definicji i wariantów cech, momentów referencyjnych, populacji oraz spójności merytorycznej i statystycznej administracyjnych zbiorów danych nadal są przedmiotem badań.

W niniejszym rozdziale przedstawione zostaną wybrane źródła danych, które potencjalnie można wykorzystać w spisie wirtualnym, czy dyskutowanym w tej pracy systemie danych społeczno-ekonomicznych. Przedstawione zostaną cztery rejestry administracyjne: PESEL, ZUS, NFZ oraz POLTAX. Rejestr PESEL opisany zostanie w kontekście zastosowania jako „kręgosłup" spisu opartego na rejestrach czy badania społecznego (podobnie jak Ewidencja Ludności w holenderskim spisie wirtualnym). Dlatego nacisk położony zostanie głównie na listę zawartych w nim zmiennych, konstrukcję potencjalnej zmiennej kluczowej, jaką jest numer PESEL oraz jego funkcje i obecne wykorzystanie w różnych systemach administracyjnych. Przedstawiona zostanie również weryfikacja zgodności informacji w spisie z innymi źródłami danych oraz zgodności wybranych struktur ludności. Rejestry ZUS i NFZ zostaną opisane w kontekście możliwości wyznaczenia zmiennych pochodnych opisujących strukturę aktywności ekonomicznej ludności Polski na niskim poziomie agregacji przestrzennej. Dodatkowo zbadana zostanie zawartość merytoryczna rejestrów, jakość (m.in. pod kątem występowania braków danych w zmiennych kluczowych i duplikatów) oraz zgodność opracowanych na ich podstawie rezultatów z analogicznymi strukturami z badań reprezentacyjnych. Zaprezentowany zostanie również rejestr POLTAX jako potencjalne źródło danych w badaniach społecznych i gospodarczych.

W dalszej kolejności opisane zostaną wybrane badania reprezentacyjne: Badanie Aktywności Ekonomicznej Ludności, Badanie Budżetów Gospodarstw Domowych oraz Badania Dochodów i Warunków Życia EU-SILC. Należą one do najważniejszych badań reprezentacyjnych przeprowadzanych w Polsce[54]. Wskazany zostanie cel badań, poddane pomiarowi cechy, schematy doboru próby, definicje populacji, jednostek, stosowane klasyfikacje. Ze względu na dostępność zbiorów danych, a także możliwość ich porównania, zostaną opisane badania przeprowadzone w 2005 roku. Wyjątek stanowi badanie EU-SILC. Okres referencyjny głównych cech poddanych pomiarowi – dotyczących źródeł i wielkości dochodu – określony jest na 1 stycznia do 31 grudnia roku poprzedniego. Z tego względu zdecydowano się przedstawić edycję badania z 2006 roku.

Jako ostatnie zostaną opisane wybrane badania reprezentacyjne przeprowadzane przez instytucje spoza sektora statystyki publicznej: Polski Generalny Sondaż Społeczny (PGSS) oraz Diagnoza Społeczna (DS). O ile badania przeprowadzane przez GUS w przeważającej części dotyczą cech obiektywnych, łatwo poddających się pomiarowi (np. dochód, wydatki, aktywność na rynku pracy), o tyle w badaniach PGSS i DS podjęto próbę pomiaru i opisu cech o charakterze subiektywnym – opinii, poglądów, postaw, zachowań, stanu zdrowia, pragnień, planów na przyszłość itp. Badania te stanowią cenne uzupełnienie oceny sytuacji społeczno-ekonomicznej w Polsce. Ze względu na pełną dostępność jednostkowych zbiorów danych (dane ze wszystkich edycji badań dostępne są w Internecie), nie ograniczono się do opisu edycji badań tylko z 2005 roku.

We wnioskach zaprezentowana zostanie idea zintegrowanego systemu danych społecznych, wykorzystującego informacje pochodzące z przedstawionych źródeł. Konstrukcja nawiązuje do Bazy Danych Społecznych będącej produktem holenderskiego spisu wirtualnego. Jednocześnie zaproponowane zostanie wykorzystanie doświadczeń polskich, związanych z pracami nad Analityczną Bazą Mikrodanych NSP 2011. Wartością dodaną w proponowanym systemie będzie wykorzystanie metod statystycznej integracji danych w celu zapewnienia spójności numerycznej szacunków, możliwości łącznej obserwacji zmiennych ze wszystkich źródeł, a także umożliwienia szacowania charakterystyk pochodzących z badań reprezentacyjnych na niskim poziomie agregacji.

[54] Podobne jest znaczenie tych badań w poszczególnych krajach europejskich oraz UE i Eurostacie.

3.1. Wybrane rejestry administracyjne jako źródło informacji w statystyce publicznej

3.1.1. Rejestr Powszechnego Elektronicznego Systemu Ewidencji Ludności (PESEL)

Rejestr PESEL – Powszechny Elektroniczny System Ewidencji Ludności prowadzony jest od 1979 roku i zawiera dane osób przebywających stale na terytorium RP, zameldowanych na pobyt stały lub czasowy trwający ponad 3 miesiące[55], a także osób ubiegających się o wydanie dowodu osobistego lub paszportu oraz osób, dla których odrębne przepisy przewidują potrzebę posiadania numeru PESEL. Rejestr działa na mocy ustawy z 10 kwietnia 1974 r. o ewidencji ludności i dowodach osobistych (Dz. U. z 1974 r. Nr 14, poz. 85 ze zm.)[56].

W rejestrze PESEL przechowywane są następujące informacje:

1. numer PESEL,
2. nazwisko i imiona aktualne,
3. nazwisko rodowe,
4. nazwiska i imiona poprzednie,
5. imiona i nazwiska rodowe rodziców,
6. data i miejsce urodzenia,
7. płeć,
8. obywatelstwo,
9. numer aktu urodzenia i oznaczenie urzędu stanu cywilnego, który ten akt sporządził,
10. stan cywilny,
11. imię i nazwisko rodowe małżonka,
12. data zawarcia związku małżeńskiego, numer aktu małżeństwa i oznaczenie urzędu stanu cywilnego, który ten akt sporządził, data rozwiązania związku małżeńskiego, sygnatura akt i oznaczenie sądu, który rozwiązał małżeństwo, data zgonu małżonka, numer aktu zgonu i oznaczenie urzędu stanu cywilnego, który ten akt sporządził,
13. adres i data zameldowania na pobyt stały,
14. poprzednie adresy zameldowania na pobyt stały wraz z określeniem okresu zameldowania, tryb wymeldowania,

[55] Po 1 stycznia 2016 dane o zameldowaniu nie będą przechowywane w zbiorze PESEL.

[56] Ustawa ta zostanie uchylona 1 stycznia 2015 ustawą z 24 września 2010 r. o ewidencji ludności (Dz. U. z 2010 r. Nr 217, poz. 1427), która zachowa prawną ciągłość jego prowadzenia.

15. adres zameldowania na pobyt czasowy trwający ponad 3 miesiące wraz z określeniem okresu zameldowania,
16. stopień wojskowy, nazwa, seria i numer wojskowego dokumentu osobistego,
17. seria i numer aktualnego dowodu osobistego oraz serie i numery poprzednich dowodów osobistych oraz daty ich wydania, daty ważności, oznaczenie organów wydających,
18. data zgonu oraz numer aktu zgonu i oznaczenie urzędu stanu cywilnego, który akt sporządził.

Ponadto, dla cudzoziemców, gromadzie się następujące dane:

1. seria i numer karty pobytu wydanej w związku z udzieleniem zezwolenia na osiedlenie się, zezwolenia na pobyt rezydenta długoterminowego Wspólnot Europejskich, zgody na pobyt tolerowany lub nadaniem statusu uchodźcy w RP oraz data jej wydania, data ważności, oznaczenie organu, który ją wydał,
2. seria i numer karty pobytu obywatela Unii Europejskiej wydanej w związku z udzieleniem zezwolenia na pobyt na terytorium RP oraz data jej wydania, data ważności, oznaczenie organu, który ją wydał,
3. seria i numer dokumentu pobytu członka rodziny obywatela Unii Europejskiej wydanego w związku z udzieleniem zezwolenia na pobyt na terytorium RP oraz data jego wydania, data ważności, oznaczenie organu, który go wydał,
4. seria i numer karty pobytu wydanej w związku z udzieleniem zezwolenia na zamieszkanie na czas oznaczony lub zgody na pobyt tolerowany oraz data jej wydania, data ważności, oznaczenie organu, który ją wydał,
5. seria i numer karty pobytu obywatela Unii Europejskiej wydanej w związku z udzieleniem zezwolenia na pobyt czasowy na terytorium RP oraz data jej wydania, data ważności, oznaczenie organu, który ją wydał,
6. seria i numer dokumentu pobytu członka rodziny obywatela Unii Europejskiej wydanego w związku z udzieleniem zezwolenia na pobyt czasowy na terytorium RP oraz data jego wydania, data ważności, oznaczenie organu, który go wydał,
7. seria i numer tymczasowego zaświadczenia tożsamości cudzoziemca oraz data jego wydania, data ważności, oznaczenie organu, który je wydał.

Numer PESEL jest unikatowym, 11-cyfrowym numerem identyfikacyjnym. Poza unikalnością, w numerze PESEL zawarte są również informacje o płci i wieku. Pierwsze sześć cyfr

odnosi się do daty urodzenia osoby (w systemie zapisu RRMMDD[57]), natomiast informacja o płci zawarta jest na dziesiątej pozycji numeru – cyfry parzyste odnoszą się do kobiet, nieparzyste zaś do mężczyzn [Józefowski, Rynarzewska-Pietrzak 2010]. Cyfry siódma, ósma i dziewiąta to tzw. liczba porządkowa osoby (numer nadania). Jedenasta cyfra PESEL to liczba kontrolna, służąca do wychwytywania przekłamań numeru. Jest ona obliczana na podstawie pierwszych dziesięciu cyfr[58]. Numer PESEL służy bardzo wielu celom. W obecnie obowiązującym systemie prawa w Polsce istnieje ponad 200 aktów prawnych, które wskazują obowiązek podania lub wykorzystania omawianego numeru [Stawecki 2005]. Z numeru PESEL korzysta się w wielu ewidencjach, np.: gospodarstw rolnych i zwierząt gospodarskich, gruntów i budynków, kierowców, osób posiadających uprawnienia budowlane, podatników, wojskowej i zatrudnienia oraz bezrobotnych i poszukujących pracy. PESEL jest składnikiem wpisu w Krajowym Rejestrze Karnym i rejestrach prowadzonych przez ZUS, a także listy maklerów giełdowych. Wiele danych zgromadzonych w ewidencji ludności, nie tylko PESEL, jest wykorzystywanych do tworzenia rejestru wyborców. Rejestr ten składa się z wybranych danych[59], co sprawia, że mamy do czynienia z selektywnym wykorzystaniem istniejących źródeł danych określonego rejestru publicznego.

W wymienionych rejestrach numer PESEL służy dwom celom. Po pierwsze, jest używany jako klucz identyfikujący osobę fizyczną i pozwalający na odróżnienie jej od innych osób, które mogą nosić to samo imię i nazwisko. Dzięki temu ewidencja ludności pełni funkcję ochronną, np. ogranicza ryzyko przyznania renty osobie nieuprawnionej, bądź noszącej takie samo imię i nazwisko oraz mieszkającej w tej samej miejscowości. Po drugie, numer PESEL w niektórych instytucjach wykorzystywany jest jako klucz wyszukiwawczy, służy więc funkcji informacyjnej.

PESEL pełni również funkcję kontrolną, polegającą m.in. na badaniu, czy dokonana przez osobę fizyczną czynność zgłoszenia meldunkowego jest zgodna z prawem.

Inną funkcją rejestru PESEL jest funkcja fiskalna, która jako opłata za udostępnienie danych stanowi dochód budżetu państwa. Zgodnie z przepisem odpowiedniego aktu wykonawczego

[57] Zapis cyfry dotyczącej miesiąca urodzenia różni się w zależności od stulecia narodzenia danej osoby. Np. osobom urodzonym w latach 1900 – 1999 numery kolejnych miesięcy zapisywane są od 01 – 12, urodzonym w latach 2000 – 2099 od 21 – 32, itd.

[58] Aby sprawdzić czy dany PESEL jest prawidłowy należy wykonać następujące działania: pierwszą cyfrę mnożymy przez 1, drugą przez 3, trzecią przez 7, czwartą przez 9, piątą przez 1, szóstą przez 3, siódmą przez 7, ósmą przez 9, dziewiątą przez 1, dziesiątą przez 3. Tak uzyskane 10 iloczynów dodajemy do siebie, po czym do sumy dodajemy liczbę kontrolną. Jeżeli wynik tej operacji jest podzielny przez 10, to podany numer PESEL jest prawidłowy. Jeśli ostatnia cyfra wyniku jest różna od zera, numer PESEL jest błędny [Kobus, Smolka 2008].

[59] Mówi o tym art. 11 ust. 4 Ordynacji Wyborczej do Sejmu Rzeczypospolitej Polskiej i Senatu Rzeczypospolitej Polskiej oraz art. 2 ust. 2 rozporządzenia MSWiA z 16 sierpnia 2001 r. w sprawie rejestru wyborców (Dz. U. nr 88, poz 962)

wynosi 4% kwoty najniższego wynagrodzenia pracowników, określonego w odrębnych przepisach.

Liczne wykorzystanie rejestru PESEL ma także negatywne konsekwencje. Po pierwsze, podanie numeru PESEL jest obecnie wymagane przy wypełnianiu większości różnego rodzaju wniosków, co rodzi pewien problem, ponieważ wyklucza osoby, które nie posiadają swojego numeru PESEL (między innymi osoby bezdomne). Tak więc niezamierzoną funkcją rejestru PESEL jest funkcja selekcyjna. Posiadanie numeru identyfikacyjnego stanowi tym samym warunek dostępu do różnego rodzaju świadczeń lub uprawnień (świadczenia z funduszu alimentacyjnego, uprawnienia budowlane, rejestrowanie bezrobotnych, itp.).

Drugim powodem, dla którego uniwersalność wykorzystywania numeru PESEL jako klucza identyfikującego osobę fizyczną może budzić wątpliwości, jest potrzeba ochrony danych osobowych, a przez to wolności i prywatności obywateli i innych mieszkańców kraju. Większość centralnych baz danych i łączy telekomunikacyjnych, które stanowią techniczną infrastrukturę społeczeństwa administracyjnego, pozostaje pod bezpośrednią kontrolą MSWiA[60].

Rejestr PESEL, poza funkcjami administracyjnymi, zaczyna pełnić, bądź już pełni również funkcje związane ze wspomaganiem organów statystyki publicznej. Najczęściej wykorzystuje się go jako operat losowania, bądź jako źródło struktur demograficznych na niskich poziomach podziału administracyjnego kraju. W tym drugim charakterze, który nie narusza ustawy „o ochronie danych osobowych", zasoby PESEL na poziomie wojewódzkim były wykorzystywane, między innymi, w pracach badawczych na rzecz Urzędu Miasta Poznania także przez Centrum Statystyki Regionalnej [Paradysz 2005].

Szerokie prace analityczne dotyczące możliwości wykorzystania rejestru administracyjnego PESEL w NSP 2011 przeprowadzono w Ośrodku Statystyki Małych Obszarów w Urzędzie Statystycznym w Poznaniu. Dokonano ocen zgodności struktur ludności. Ze względów technicznych, populację rejestru ograniczono do ludności zameldowanej na pobyt stały w województwie wielkopolskim [Józefowski, Rynarzewska-Pietrzak 2010].

W ocenie struktur ludności, porównano dane dostarczane przez statystykę publiczną oraz otrzymane ze zbioru PESEL. Uzyskano wysoką zgodność struktur według pięcioletnich grupach wieku – różnica w analizowanych zbiorach nie przekraczała 0,2%. Wyjątek stanowiły rozbieżności zaobserwowane dla osób w wieku 0 lat oraz grupie wieku 20 – 24 lat (rzędu

[60] W statucie MSWiA nadanym zarządzeniem prezesa RM z 26 czerwca 2002 r. (M.P. nr 26, poz. 434) przewiduje się, że w ramach struktury organizacyjnej tego resortu działa m.in. Departament Rejestrów Państwowych.

6%). Różnica ta wynikała z opóźnień w rejestracji niemowląt oraz dużej mobilności ludności w wieku 20 – 24 lat[61], nie uwzględnionej w rejestrze.

Ocena zgodności czasowej wykazała około dwutygodniowe opóźnienia w ujmowaniu informacji o liczbie urodzeń w poszczególnych dniach miesiąca. Wykazano również, że dni świąteczne charakteryzują się niezgodnościami w zapisie do rejestru znacznie częściej niż dni powszednie.

Zgodność terytorialną zbadano w odniesieniu do różnic w przekroju powiatów. Wykazano nadwyżkę liczby ludności w dużych miastach, przy niedoszacowaniu w powiatach mniej zurbanizowanych.

Wśród wniosków z możliwości wykorzystania rejestru PESEL w NSP 2011 wymieniono przede wszystkim, takie zalety rejestru jak: możliwość otrzymania struktur ludności na niskich poziomach agregacji przestrzennej oraz stosunkowo niewielkie różnice w strukturach ludności w ujęciach merytorycznych, jak płeć, czy wiek. Wśród wad natomiast wyszczególniono przede wszystkim opóźnienia w rejestrowaniu ruchu naturalnego i wędrówkowego ludności.

Rejestr administracyjny PESEL, pomimo swoich niedoskonałości, jest dobrym źródłem danych do wykorzystania w statystyce publicznej. Jego bezpośrednie użycie w pracach organów statystycznych wymaga oczywiście dużo pracy związanej z harmonizacją czasową, przestrzenną i strukturalną zawartości informacyjnej zbioru, a także dostosowania definicji i wariantów cech. Jednak doświadczenie NSP 2011 wskazuje, że może on pełnić w polskim systemie statystycznym analogiczną rolę, jak zbiór Ewidencji Ludności w holenderskim spisie wirtualnym.

3.1.2. Rejestr Zakładu Ubezpieczeń Społecznych (ZUS)

W celu zbadania przydatności rejestrów administracyjnych do potrzeb statystyki publicznej podjęto próbę wykorzystania rejestru Zakładu Ubezpieczeń Społecznych do oszacowania aktywności ekonomicznej ludności na niskim poziomie agregacji przestrzennej[62]. Rejestr ZUS zawiera informacje o osobach ubezpieczonych w podziale na płatników – osoby fizyczne oraz osoby prawne. Każdy rekord opisano szeregiem zmiennych z których najważniejszą w kontekście analizy rynku pracy jest „kod ubezpieczenia". Zawiera ona informa-

[61] Np. przyjazdy i wyjazdy na studia.

[62] W Polsce, źródłami informacji o aktywności ekonomicznej są: Badanie Aktywności Ekonomicznej Ludności - badanie reprezentacyjne, któregom wyniki uogólniane są co najwyżej do poziomu województw [Zgierska 2010] oraz informacje pochodzące ze sprawozdawczości bieżącej powiatowych urzędów pracy. W tym drugim przypadku informacja jest pełna (w sensie pokrycia), jednak dotyczy wyłącznie bezrobocia.

cję o tytule ubezpieczenia płatnika, na podstawie której możliwe jest uzyskanie informacji o statusie na rynku pracy [Nowakowska 2008].

Analizie poddano repozytorium danych rejestru administracyjnego ZUS z momentem referencyjnym ustalonym przez gestora na 31 stycznia 2010 roku. Zbór zawierał 17 364 001 zanonimizowanych[63] rekordów dotyczących osób ubezpieczonych (symbol U) w podziale na:

- płatników – osoby fizyczne prowadzące pozarolniczą działalność lub płatnicy o statusie „nieustalony" oraz ubezpieczonych, dla których płatnicy złożyli poprawny komplet dokumentów rozliczeniowych za dany okres – w liczbie 4 324 364 rekordów;
- płatników - osoby prawne oraz jednostki organizacyjne nieposiadające osobowości prawnej oraz ubezpieczeni, dla których płatnicy złożyli poprawny komplet dokumentów rozliczeniowych za dany okres – w liczbie 13 039 637 rekordów.

Zakres przeprowadzonych prac obejmował:

- kontrolę momentu referencyjnego,
- kontrolę zbioru w zakresie braków danych,
- eksplorację kodów ubezpieczeń w celu uzyskania informacji o aktywności ekonomicznej ludności,
- usunięcie duplikatów,
- kontrolę momentu referencyjnego zbioru.

Moment referencyjny rejestru ZUS ustalono na podstawie przeprowadzonej analizy zmiennej „data wpisu do rejestru". Zaobserwowano, że dzienna liczba wpisów do rejestru dokonana po 10 stycznia 2010 roku była zdecydowanie niższa niż w przed tym dniem. Przypuszczalnie były to wpisy z błędną datą. Stąd ustalenie dokładnego dnia momentu referencyjnego rejestru było problematyczne (np. ze względu na wątpliwości dotyczące opóźnienia). W związku z tym ustalono arbitralnie, że moment referencyjny rejestru to 10 stycznia 2010 roku.

Braki danych zostały przeanalizowane przede wszystkim dla zmiennych „kod tytułu ubezpieczenia" – jako zmiennej kluczowej w analizie, „data urodzenia" – jako zmienna pierwotna do wyliczenia zmiennej pochodnej „wiek", „kod TERYT miejsca zamieszkania" – jako zmiennej służącej do agregowania wyników w przekroju terytorialnym oraz „płeć".

[63] Tzw. zmienne wrażliwe (imię, nazwisko, nr PESEL, adres zamieszkania, itp.) zostały zakodowane w sposób uniemożliwiający identyfikację poszczególnych osób.

Tabela 3.1. Struktura braków danych w analizowanych zmiennych rejestru ZUS

Zmienna	Odsetek braków danych
Data urodzenia	0,0003%
Kod TERYT miejsca zamieszkania	10,76%
Płeć	0,11%
Kod ubezpieczenia	0,00%

Źródło: opracowanie własne na podstawie rejestru ZUS

Największą frakcję braków danych zaobserwowano dla zmiennej „kod TERYT miejsca zamieszkania" (por. tabela 3.1). By zminimalizować liczbę braków danych zintegrowano w sposób deterministyczny (na podstawie zmiennej *id_pesel*[64]) rejestr ZUS z rejestrem administracyjnym PESEL, który również zawierał informację o miejscu zamieszkania. W miejsce braków danych w rejestrze ZUS dołączono odpowiednie informacje z rejestru PESEL, co pozwoliło na redukcję liczby braków do 23 655 obserwacji (zaledwie 0,14% wszystkich jednostek).

W celu przedstawienia aktywności ekonomicznej w przekroju terytorialnym oraz ze według płci i wieku wykorzystano sześcioznakową zmienną „kod tytułu ubezpieczenia". Ostatni znak kodu dotyczył stopnia niepełnosprawności danej osoby, przedostatni uprawnień emerytalno–rentowych, natomiast pierwsze cztery znaki odnosiły się do kodów tytułów ubezpieczeń. W opracowaniu pod redakcją G. Nowakowskiej [2008] dokonano podziału kodów tytułów ubezpieczeń na formy aktywności ekonomicznej [*ibidem*, załącznik 2].

[64] Zmienna *id_pesel* zawierała zakodowany numer PESEL osób. Numer ten został zakodowany w celu uniemożliwienia identyfikacji jednostek.

Tabela 3.2. Struktura zbioru ZUS ze względu na formę aktywności

Status (symbol)		Liczba
Pracownicy najemni (A)		**11 104 132**
w tym	osoby wykonujące umowę agencyjną, itp. – w siedzibie zleceniodawcy (A1)	766 537
	duchowni (A2)	28 154
	żołnierze służby zasadniczej (A3)	1 455
	osoby wykonujące odpłatnie pracę w czasie odbywania kary (A4)	17 873
Pracujący na własny rachunek (B)		**1 820 921**
Pracujący, pomagający członkowie rodzin (C)		**36 689**
Inni (nie zaliczani do pracujących) (D)		**204 314**
w tym	osoby na urlopie wychowawczym (D1)	197 534
Bezrobotni (E)		**2 063 296**
Nieustalony status na rynku pracy (X)		**982 753**

Źródło: opracowanie własne na podstawie rejestru ZUS i [Nowakowska 2008].

Najliczniejszą grupę na rynku pracy stanowili pracownicy najemni. Kod ubezpieczenia umożliwiał dodatkowo rozbicie tej grupy na podkategorie umożliwiające dodatkową analizę różnych form zatrudnienia (por. tabela 3.2). Możliwe było również wyodrębnienie innych form aktywności, w tym dwumilionową grupę osób bezrobotnych oraz około dwustutysięczną grupę osób na urlopie wychowawczym[65]. Prawie milion osób posiadało nieustalony status na rynku pracy (kody ubezpieczeń nie wskazywały jednoznacznie na status).

Występowanie duplikatów (powtarzających się jednostek) w zbiorze związane było z ubezpieczeniem jednej osoby z wielu tytułów (np. pracę na kilku etatach). W badanym zbiorze 2 368 279 osób było ubezpieczonych z więcej aniżeli jednego tytułu, co przekładało się zawyżenie liczby pracujących (por. tabela 3.3).

Tabela 3.3. Liczba osób ubezpieczonych z więcej niż jednego tytułu

Liczba tytułów ubezpieczeń	Liczba osób
1	13 846 177
2	2 040 657
3	263 619
4	46 133
5	11 002
6 i więcej	6 868

Źródło: opracowanie własne na podstawie rejestru ZUS

[65] Przedstawienie osób na urlopie wychowawczym w ujęciu płci umożliwi analizę wykorzystania świadczeń przez ojców.

Eliminacji duplikatów wynikała z celu określenia aktywności na rynku pracy dla populacji ludności w wieku 15 i więcej lat by umożliwić porównywalność w wynikami BAEL. Dlatego pozostawiono jeden rekord dotyczący głównego miejsca pracy danej osoby. W BAEL jest ono określone na podstawie deklaracji respondenta. W rejestrze ZUS nie ma takiej informacji, dlatego zdecydowano, że główne miejsce pracy określić wg następującego algorytmu[66]:

— jeżeli osoba ubezpieczona z wielu tytułów we wszystkich przypadkach posiada ten sam status zatrudnienia, w zbiorze pozostaje dowolny (losowy) rekord;

— jeżeli osoba ubezpieczona z wielu tytułów posiada różne statusy zatrudnienia, jako główne miejsce pracy uznaje się ważniejsze, przy czym kryteria ważności ustalono w sposób przedstawiony w tabeli 3.4.

Tabela 3.4. Kryterium ważności statusów zatrudnienia

Status (symbol)	**Ważność (1 – najważniejsze)**
Pracownicy najemni (A)	1
Pracujący na własny rachunek (B)	2
Pracujący, pomagający członkowie rodzin (C)	3
Inni (nie zaliczani do pracujących) (D)	4
Bezrobotni (E)	5
Nieustalony status na rynku pracy (X)	6

Źródło: opracowanie własne na podstawie rejestru ZUS

Po usunięciu duplikatów zbiór zawierał 16 214 456 rekordów – dotyczących osób aktywnych zawodowo.

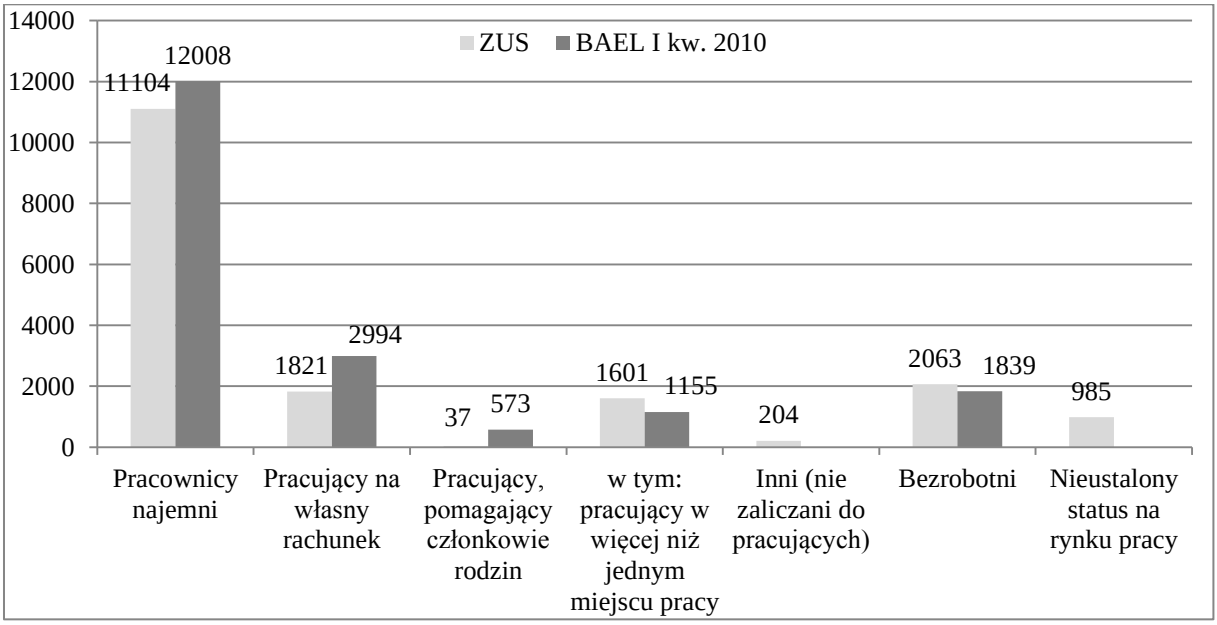
Jak określono wyżej, celem analizy było oszacowanie aktywności ekonomicznej na niskim poziomie agregacji przestrzennej (np. powiatów) z uwzględnieniem płci i wieku. Dane pozyskane z rejestru dotyczą źródła pełnego, zawierającego informacje o wszystkich osobach ubezpieczonych w Zakładzie Ubezpieczeń Społecznych. Należy jednak zwrócić uwagę, że rejestr ZUS nie uwzględnia ludności ubezpieczonej w Kasie Rolniczego Ubezpieczenia Społecznego, która dotyczy przede wszystkim rolników. Baza ta zawierała dane o 1 425 500 świadczeniobiorcach oraz 1 570 328 ubezpieczonych[67]. W związku z tym do otrzymanych szacunków należy podchodzić z dystansem mając świadomość, że uzyskane wyniki dotyczą ludności aktywnej zawodowo z wyłączeniem osób posiadających gospodarstwo rolne.

[66] Kryterium ‘ważności’ statusu zatrudnienia ustalono w sposób arbitralny.

[67] Źródło: portal internetowy www.krus.gov.pl; liczba świadczeniobiorców – stan na 31 grudnia 2009 roku, liczba ubezpieczonych – stan na 30 września 2009 roku.

W pierwszym kroku, w celu sprawdzenia ogólnej zgodności struktur aktywności zawodowej według rejestru ZUS i BAEL[68], zestawiono poszczególne warianty cechy „status na rynku pracy" z obu zbiorów.

Wykres 3.1. Struktura ludności aktywnej zawodowo w Polsce w I kwartale 2010 na podstawie ZUS i BAEL (w tys. osób)



Źródło: opracowanie własne na podstawie rejestru ZUS i BAEL

W rejestrze ZUS wydaje się występować niedoszacowanie liczby osób pracujących w porównaniu z analogicznymi strukturami wynikającymi z szacunków BAEL (por. wykres 3.1). Odstępstwem od tej reguły jest liczba pracujących w więcej niż jednym miejscu pracy. Rejestr ZUS zawiera informacje o ponad 450 tys. takich osób więcej niż BAEL. Podobnie w zbiorze danych ZUS znajduje się o ponad 200 tys. więcej rekordów osób bezrobotnych niż wynika to z szacunków BAEL. Może to wskazywać na błąd szacunku powyższych kategorii w BAEL lub różnice definicyjne, jednak potwierdzenie (bądź odrzucenie) tej hipotezy wymaga dalszych badań.

W celu syntetycznej oceny zgodności struktur aktywności ekonomicznej wynikającej z rejestru ZUS i BAEL obliczono wskaźniki podobieństwa dla pracujących, bezrobotnych oraz ogółem w przekroju województw w rejestrze ZUS i BAEL: $W_{p1} = \sum_{i=1}^{k}(min_{ZB}w_i)$ oraz $W_{p2} = \frac{\sum_{i=1}^{k}(min_{ZB}w_i)}{\sum_{i=1}^{k}(max_{ZB}w_i)}$. Współczynniki te informują w jakim stopniu podobne są struktury ana-

[68] W celach porównawczych wykorzystano szacunki BAEL z I kwartału 2010 roku.

lizowanych zmiennych w obu plikach. W przypadku osób pracujących współczynniki podobieństwa wynosiły: $W_{p1} = 94{,}5\%$, $W_{p2} = 89{,}6\%$, dla bezrobotnych: $W_{p1} = 94{,}3\%$, $W_{p2} = 89{,}1\%$, a dla pracujących i bezrobotnych ogółem: $W_{p1} = 94{,}2\%$, $W_{p2} = 89{,}0\%$. Wskazują one na wysokie podobieństwo struktur w obu zbiorach, co sugeruje, że rejestr ZUS może być alternatywnym źródłem informacji o aktywności ekonomicznej ludności.

Rejestr ZUS, ze względu na pełne pokrycie (w rozumieniu liczby osób ubezpieczonych w tej instytucji), umożliwia szacunki o dowolnym stopniu agregacji przestrzennej (ograniczonej dokładnością zmiennej „kod TERYT miejsca zamieszkania”[69]) i merytorycznej (ograniczonej liczbą zmiennych w zbiorze). Natomiast BAEL pozwala, z zadowalającą precyzją na szacowanie aktywności ekonomicznej w przekroju płci, pięcioletnich grup wieku oraz województw osobno. Szacunki na podstawie rejestru ZUS, o tak dużym stopniu szczegółowości mogą okazać się niezwykle pomocne np. dla samorządów chcących kierować środki pomocowe do najbardziej zubożonych lub zagrożonych ubóstwem regionów. Również w przypadku starania się o środki unijne, ważnym aspektem jest informacja na niskim stopniu agregacji umożliwiająca uargumentowanie potrzeb finansowych dla danego obszaru. Z tego punktu widzenia rejestr ZUS może okazać się bardzo ważnym źródłem informacyjnym.

Z drugiej strony należy jednak pamiętać, że około 2 milionów osób ubezpieczonych jest w KRUS, co oznacza, że wszelkie analizy i szacunki czynione na podstawie danych z ZUS nie będą analizami pełnymi. Dołączenie informacji o aktywności ekonomicznej z rejestru KRUS jest nadal obiektem badań Ośrodka Statystyki Małych Obszarów działającym przy Urzędzie Statystycznym w Poznaniu.

Należy również pamiętać, że rejestr ZUS nie zawiera informacji dla wielu cech, jakie oferuje Badanie Aktywności Ekonomicznej Ludności (np. wykształcenia, stopnia pokrewieństwa z głową gospodarstwa domowego, informacji o głównym miejscu pracy, itp.) co oznacza, że rejestr może pełnić jedynie funkcje pomocnicze w statystyce publicznej.

3.1.3. Rejestr Narodowego Funduszu Zdrowia (NFZ)

Prace nad rejestrem administracyjnym Narodowego Funduszu Zdrowia były kontynuacją prac przeprowadzonych na rejestrze ZUS. Rejestr NFZ zawierał informacje o 38 740 372 rekordach[70] opisanych przez 44 zmienne, wśród których wyróżnić można:

— płeć,

— datę urodzenia,

[69] W analizowanym zbiorze kody TERYT sięgały poziomu gminy.

[70] Rekordem w rejestrze było ubezpieczenie zdrowotne.

— miejsce zameldowania/zamieszkania (do poziomu gminy włącznie),
— tytuł uprawnienia do ubezpieczenia,
— informacje o numerze PESEL (odpersonalizowane).

Zakres podmiotowy rejestru obejmował:

— osoby posiadające obywatelstwo państwa członkowskiego Unii Europejskiej lub państwa członkowskiego Europejskiego Porozumienia o Wolnym Handlu (EFTA),
— osoby nieposiadające obywatelstwa państwa członkowskiego Unii Europejskiej lub państwa członkowskiego Europejskiego Porozumienia o Wolnym Handlu (EFTA) – na podstawie odrębnych przepisów szczegółowych.

Zakres przedmiotowy rejestru – obejmował dwie grupy ubezpieczonych:

— Centralny Rejestr Członków Rodziny Ubezpieczonych Uprawnionych do Ubezpieczenia Zdrowotnego – 8 572 781 rekordów,
— Centralny Wykaz Ubezpieczonych – 30 074 341 rekordów.

Podobnie jak w przypadku rejestru ZUS, celem badania była weryfikacja możliwości wykorzystania rejestru NFZ dla potrzeb sprawozdawczości rynku pracy. Zakres przeprowadzonych prac obejmował:

— kontrolę momentu referencyjnego,
— kontrolę zbioru w zakresie braków danych,
— eksplorację kodów ubezpieczeń w celu uzyskania informacji o aktywności ekonomicznej ludności,
— usunięcie duplikatów.

Kontrolę momentu referencyjnego przeprowadzono z wykorzystaniem zmiennej „data zapisu do rejestru". Dzienna liczba zapisów do rejestru utrzymywała się na podobnym poziomie do dnia 17 listopada 2009 roku. Po tym dniu liczba zapisów zdecydowanie spadła. Na tej podstawie arbitralnie ustalono, że rejestr jest aktualny na 17 listopada 2009 roku.

Tabela 3.5. Struktura braków danych w analizowanych zmiennych rejestru NFZ

Zmienna	Odsetek braków danych
płeć	0 %
data urodzenia	0,00023%
kod TERYT gminy zamieszkania	0,00003%
tytuł uprawnienia do ubezpieczenia	31,58%
id_PESEL	0,17%

źródło: opracowanie własne na podstawie rejestru NFZ

Zbiór danych NFZ charakteryzował się stosunkowo niską liczbą braków danych. Wyjątek stanowiła zmienna „tytuł uprawnienia do ubezpieczenia", w której niemal co trzecia jednostka nie posiadała ważnej wartości (por. tabela 3.5). Braków danych w tej zmiennej zanotowano 13 528 850, z czego wśród osób do 24 lat zanotowano aż 9 165 042 braków odpowiedzi. Ze względu na brak informacji od gestora o przyczynie tej sytuacji, wysunięto przypuszczenie, że są to osoby nie posiadające tytułu do ubezpieczenia, a więc bierne zawodowo, ubezpieczone jako członkowie rodzin płatników.

Tabela 3.6. Liczba powtórzeń numeru PESEL w zbiorze NFZ

Liczba powtórzeń	**Częstość**
1	38 641 521
2	11 086
3	210
4	8
5	5
6 i więcej	6

źródło: opracowanie własne na podstawie rejestru NFZ

Bardzo niewielki odsetek rekordów w zbiorze NFZ posiadał ten sam numer PESEL. Duplikaty wynikały ze zmiany statusu z 'członka rodziny płatnika' na 'płatnika' i odwrotnie. Ze względu na niedużą liczbę powtórzeń (por. tabela 3.6), duplikaty usunięto losowo. Należy jednak zwrócić uwagę, że 66 987 osób nie miało informacji o numerze PESEL, a 20 549 osób miało błędny nr PESEL (zakodowany jako 0). Dla tych osób nie sposób było sprawdzić występowanie duplikatów.

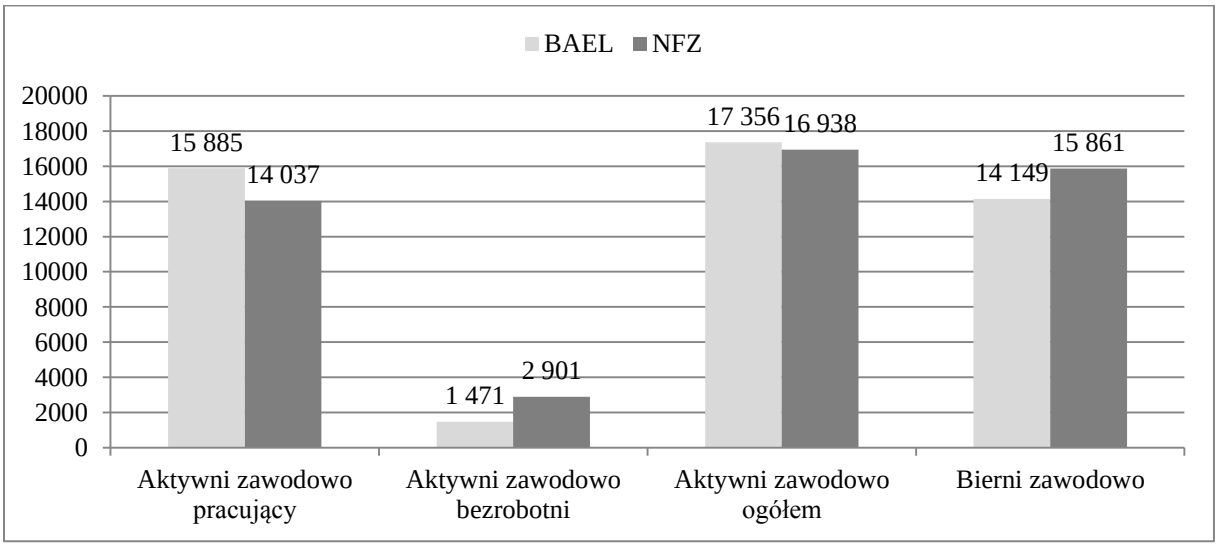
Wobec faktu, że zbiór danych NFZ, ze względu na swoją liczebność, zawierała informację o wszystkich mieszkańcach Polski, możliwym było, oprócz wyznaczenia liczby aktywnych zawodowo (jak w rejestrze ZUS), ustalenie liczby osób biernych zawodowo. Za osoby bierne zawodowo uznano osoby o nieustalonym statusie na rynku pracy (symbol X) oraz osoby którym nie przyporządkowano kodu ubezpieczenia (por. tabela 3.7).

Tabela 3.7. Struktura aktywności ekonomicznej ludności w wieku 15 lat i więcej wg rejestru NFZ

Status (symbol)		Liczba
Pracownicy najemni (A)		10 346 249
w tym:	osoby wykonujące umowę agencyjną, itp. – w siedzibie zleceniodawcy (A1)	1 351 190
	duchowni (A2)	18 339
	żołnierze służby zasadniczej (A3)	30 889
Pracujący na własny rachunek (B)		1 629 112
Pracujący, pomagający członkowie rodzin (C)		42 655
Inni (nie zaliczani do pracujących) (D)		113 113
w tym:	osoby na urlopie wychowawczym (D1)	26 766
Bezrobotni (E)		2 901 421
Nieustalony status na rynku pracy (X)		7 959 287
brak numeru ubezpieczenia		7 762 171

źródło: opracowanie własne na podstawie rejestru NFZ i [Nowakowska 2008]

Wykres 3.2. Struktura ludności aktywnej zawodowo w Polsce w IV kwartale 2009 na podstawie NFZ i BAEL (w tysiącach osób)



Źródło: opracowanie własne na podstawie rejestru NFZ i BAEL

Porównanie struktur rynku pracy otrzymanych na podstawie rejestru NFZ z analogicznymi strukturami wynikającymi z Badania Aktywności Ekonomicznej Ludności przeprowadzono dla III kwartału 2009 roku. Analiza porównawcza wykazała stosunkowo dużą zgodność między źródłami, jednak zauważyć można, że rejestr NFZ zdaje się niedoszacowywać pracujących (względem BAEL) oraz przeszacowywać liczbę osób bezrobotnych (por. wykres 3.2).

Tabela 3.8. Struktura ludności aktywnej zawodowo na podstawie NFZ i BAEL (w tysiącach osób), Polska, przekrój województw, IV kwartał 2009

Województwo	NFZ 17 listopada 2009		BAEL IV kw. 2009	
	aktywni zawodowo	bierni zawodowo	aktywni zawodowo	bierni zawodowo
dolnośląskie	1 354	1 141	1 343	1 097
kujawsko-pomorskie	927	850	949	772
lubelskie	863	999	1 109	860
lubuskie	478	388	495	397
łódzkie	1 140	1 060	1 414	1 124
małopolskie	1 385	1 368	1 362	1 181
mazowieckie	2 383	2 036	2 580	1 844
opolskie	436	424	424	356
podkarpackie	866	937	980	775
podlaskie	467	565	528	446
pomorskie	1 015	876	892	774
śląskie	2 066	1 934	1 887	1 708
świętokrzyskie	531	575	686	515
warmińsko-mazurskie	653	573	624	561
wielkopolskie	1 556	1 339	1 425	1 112
zachodniopomorskie	790	676	659	626
OGÓŁEM	**16 910**	**15 740**	**17 357**	**14 148**

Źródło: opracowanie własne na podstawie rejestru NFZ i BAEL

Zagregowane struktury rynku pracy porównano również w ujęciu województw (niższy poziom agregacji przestrzennej nie jest dostępny w BAEL). Współczynniki podobieństwa dla aktywnych i biernych zawodowo były wysokie i wynosiły odpowiednio 94,9% i 97,3%. Niedoszacowania aktywnych zawodowo w NFZ względem BAEL były szczególnie widoczne w województwach o stosunkowo niedużej liczbie mieszkańców (lubelskie, lubuskie, podkarpackie, podlaskie, świętokrzyskie), natomiast w województwach „dużych" struktury rynku pracy były bardziej zbliżone (por. tabela 3.8).

Dzięki bardzo dużemu (prawdopodobnie pełnemu) pokryciu rejestru NFZ, przy dużej zgodności struktur rynku pracy z BAEL, możliwym było wyznaczenie podstawowych charakterystyk rynku pracy na niskim poziomie agregacji przestrzennej oraz w ujęciu podstawowych charakterystyk demograficznych (płeć i wiek).

Rejestr NFZ, odpowiednio zharmonizowany, może być dobrym źródłem informacji dla statystyki publicznej. W zasadzie rejestr zawiera dane o aktywności ekonomicznej oraz o różnych charakterystykach ludności całego kraju, może być przydatny jako źródło szacunków w dowolnych przekrojach terytorialnych. Zasoby rejestru NFZ mogą mieć duże znaczenie dla władz samorządowych – zwłaszcza w kontekście członkostwa Polski w Unii Europejskiej. Duża liczebność rejestru może również predestynować go jako źródło informacji pomocniczych w estymacji pośredniej.

Rejestr nie zawiera jednak wielu informacji dodatkowych, jak wykształcenie mieszkańców, czy źródło utrzymania (dostępne w badaniach reprezentacyjnych).

3.1.4. Rejestr POLTAX

Rejestr POLTAX jest administracyjnym repozytorium ewidencjonującym i przechowującym dane o podatnikach. Wykorzystywany jest w urzędach skarbowych, zaś administrowany i rozwijany jest przez Ministerstwo Finansów. Jest przedsięwzięciem organizacyjno-technicznym, którego celem jest tworzenie i utrzymanie systemu informatycznego wspomagającego działalność administracji podatkowej.

Zręby systemu POLTAX sięgają roku 1989, w którym Ministerstwo Finansów podjęło decyzję o zakupie systemu informatycznego wspomagającego urzędy skarbowe w obsłudze wymiaru i poboru podatków planowanych do wprowadzenia w kolejnych latach. Gotowy system został oddany do użytku w 1995 roku. W 2003 r. Ministerstwo Finansów opracowało koncepcję rozbudowy i modyfikacji funkcjonujących rozwiązań informatycznych, uwzględniającą rosnące wymagania odnośnie wykorzystania systemów informacyjnych w pracy administracji podatkowej i zakładającą elektroniczną komunikację z otoczeniem. W wyniku tych prac powstała wizja systemu obejmującego wszystkie obszary działania administracji podatkowej i umożliwiającego obsługę kluczowych odbiorców usług [Ministerstwo Finansów 2010].

System początkowo wdrożono w 300 urzędach skarbowych oraz 49 izbach skarbowych. Po reformie administracyjnej z 1998 roku, system zaczął działać w 401 urzędach skarbowych powstałych w nowoutworzonych powiatach. Obejmuje podatników rozliczających się z urzędami skarbowymi m.in. za pomocą formularzy PIT oraz CIT.

W zakresie informacji o podatkach od osób fizycznych (PIT), w systemie przechowuje się informacje w 25 formularzy podatkowych, wśród których można wymienić:

— Deklaracja roczna o pobranych zaliczkach na podatek dochodowy - PIT-4R,

— Deklaracja do wymiaru zaliczek podatku dochodowego od dochodów z działów specjalnych produkcji rolnej - PIT-6,
— Deklaracja roczna o zryczałtowanym podatku dochodowym - PIT-8AR,
— Informacja o dochodach oraz pobranych zaliczkach na podatek dochodowy - PIT-11,
— Deklaracja o osiągniętych przychodach z odpłatnego zbycia nieruchomości lub praw majątkowych, objętych zryczałtowanym podatkiem dochodowym - PIT-23,
— Zeznanie o wysokości uzyskanego przychodu, wysokości dokonanych odliczeń i należnego ryczałtu od przychodów ewidencjonowanych za rok - PIT-28,
— Zeznanie o wysokości osiągniętego dochodu (poniesionej straty) w roku podatkowym - PIT-36, PIT-37, PIT-38, PIT-39,
— Roczne obliczenie podatku przez płatnika na wniosek podatnika - PIT-40.

W ramach formularzy podatkowych PIT zbierane są informacje m.in. o adresie zamieszkania podatnika, miejsca pracy, osiągniętym przychodzie, kosztach uzyskania przychodu oraz dochodzie.

W zakresie podatku dochodowego od osób prawnych (przedsiębiorstw i spółek, CIT) zbierane są informacje z 13 formularzy podatkowych, wśród których można wymienić:

— Informacja podatnika podatku dochodowego od osób prawnych o otrzymanych/przekazanych darowiznach - CIT-D,
— Deklaracja o wysokości podatku dochodowego od dochodów z tytułu udziału w zyskach osób prawnych - CIT-6AR,
— Deklaracja o wysokości pobranego podatku dochodowego od dochodów z dywidend oraz innych przychodów z tytułu udziału w zyskach osób prawnych - CIT-6R,
— Zeznanie o wysokości osiągniętego dochodu (poniesionej straty) przez podatnika podatku dochodowego od osób prawnych - CIT-8,
— Zeznanie o wysokości osiągniętego dochodu (poniesionej straty) przez podatkową grupę kapitałową - podatnika podatku dochodowego od osób prawnych - CIT-8A,
— Zeznanie o wysokości osiągniętego dochodu (poniesionej straty) przez podatkową grupę kapitałową - podatnika podatku dochodowego od osób prawnych - za rok podatkowy - CIT-8B,
— Deklaracja o wysokości przychodu za wywóz ładunków i pasażerów przyjętych do przewozu w porcie polskim, uzyskanego przez zagraniczne przedsiębiorstwo żeglugi handlowej od zagranicznych zleceniodawców - CIT-9R.

O osobach prawnych w rejestrze POLTAX gromadzone są informacje m.in. o numerze NIP i REGON oraz nazwie podatnika, adresie siedziby firmy, dochodzie i kwocie należnego podatku.

Obecnie rejestr POLTAX jest podstawą tworzonych nowoczesnych rozwiązań informatycznych: e-Podatki oraz jego składowej – e-Deklaracje. Głównym celem projektu e-Podatki jest uproszczenie systemu poboru podatków poprzez usprawnienie (zoptymalizowanie) wewnętrznych procesów biznesowych administracji podatkowej. Cel ten osiągnięty ma zostać poprzez budowę i wdrożenie centralnych systemów informatycznych, upowszechnienie dokumentu elektronicznego i jego obiegu wewnętrznego w administracji podatkowej oraz uproszczenie procedur w zakresie wymiaru i poboru podatków. Projekt e-Deklaracje jest składową projektu e-Podatki i stanowi platformę umożliwiającą osobom fizycznym i prawnym składanie zeznań podatkowych drogą elektroniczną.

Rejestr administracyjny POLTAX może być bogatym źródłem informacyjnym o dochodach ludności oraz finansach przedsiębiorstw. Poprzez przeprowadzenie na jego podstawie „Badania dojazdów do pracy" wykazano, że możliwe jest wykorzystanie informacji skarbowych w zasilaniu informacyjnym statystyki publicznej. Wysokie pokrycie rejestru oraz jego cenna zawartość informacyjna[71] stanowi o przydatności rejestru w badaniach społecznych i przedsiębiorstw.

3.2. Wybrane badania reprezentacyjne w systemie statystyki publicznej

3.2.1 Badanie Aktywności Ekonomicznej Ludności (BAEL)

Badanie Aktywności Ekonomicznej Ludności jest największym cyklicznie przeprowadzanym badaniem reprezentacyjnym w Polsce i należy do podstawowych źródeł informacji o aktywności ekonomicznej mieszkańców Polski. Przeprowadzane jest od maja 1992 roku zgodnie z metodologią wypracowaną na XIII Międzynarodowej Konferencji Statystyków Pracy w październiku 1982 r. i stale udoskonalaną zgodnie z wymogami Eurostatu. Badanie przeprowadzane jest w cyklu kwartalnym. Do IV kwartału 2009 roku badanie przeprowadzano na próbie około 25 tys. gospodarstw domowych, natomiast od I kwartału roku 2010 próba została zwiększona około dwukrotnie ze względu na wynikającą ze zobowiązań międzynaro-

[71] Informacje o dochodach ludności zbierane są m.in. w Badaniu Budżetów Gospodarstw Domowych oraz Badaniu Dochodów i Jakości Życia EU-SILC. Ze względu na fakt, iż informacje dochodowe należą do tzw. pytań drażliwych, mogą być one obarczone błędem. Należy również zaznaczyć, że w badaniach reprezentacyjnych pomiarowi podlegają również źródła dochodu, wśród których nie wszystkie mogą być ujęte w źródłach administracyjnych (np. dochody zwolnione z podatku, nie podlegające zgłoszeniu organom skarbowym).

dowych konieczność zmniejszenia błędu oszacowań (zwiększenie liczebności próby nie umożliwiło dokonywania szacunków w bardziej szczegółowych przekrojach).

Wyniki Badania Aktywności Ekonomicznej Ludności stanowią podstawę szacowania wielu charakterystyk rynku pracy, jak stopa bezrobocia, wskaźnik aktywności zawodowej, czy wymiar czasu pracy osób pracujących. W przypadku osób bezrobotnych pomiarowi podlegają również takie ważne czynniki jak przyczyna pozostawania bez pracy oraz czas i metody poszukiwania pracy. Szacunki na podstawie BAEL stanowią ważne źródło informacji dla kreowania oraz monitorowania efektów polityki społeczno-gospodarczej kraju. Stanowią również ważne źródło porównawcze z innymi krajami Europy pod względem organizacji i rozwoju rynku pracy – szczególnie w kontekście przynależności Polski do Unii Europejskiej.

Jednostką badania jest osoba w wieku 15 lat i więcej. Przedmiotem badania jest sytuacja w zakresie aktywności ekonomicznej ludności, tzn. fakt wykonywania pracy, pozostawania bezrobotnym lub biernym zawodowo w badanym tygodniu [Zgierska 2012]. Pomiar płci, wieku, województwa i klasy miejscowości zamieszkania umożliwia przedstawienie szacunków w odpowiednich przekrojach.

Tabela 3.9. Liczba osób i gospodarstw domowych poddanych pomiarowi w BAEL według kwartałów 2005 roku

Kwartał	Liczba gospodarstw domowych	Liczba osób
I	25598	43578
II	25470	45988
III	25566	45529
IV	25502	50053
Ogółem	102136	185148

Źródło: opracowanie własne na podstawie BAEL 2005

W 2005 roku w badaniu udział wzięło ponad 102 tys. gospodarstw domowych, w których pomiarowi poddano ponad 185 tys. osób w wieku 15 lat i więcej (por. tabela 3.9). Wśród przebadanych w 2005 roku osób, 73341 było badanych więcej niż jeden raz. Eliminacja obserwacji powtórzonych powoduje zatem, że unikatowych obserwacji było 111807. Usunięcie obserwacji powtórzonych powoduje również wyłączenie z szacunków przeprowadzanych na podstawie badania oceny sezonowości na rynku pracy.

Wśród najważniejszych cech poddanych pomiarowi w Badaniu Aktywności Ekonomicznej Ludności dla osób można wymienić:

— województwo zamieszkania,

— klasa miejscowości zamieszkania,
— płeć,
— wiek,
— stopień pokrewieństwa z głową gospodarstwa domowego,
— stan cywilny,
— poziom wykształcenia,
— symbol ,
— symbol PKD dodatkowego miejsca zatrudnienia,
— główne i dodatkowe źródło utrzymania,
— rodzaj aktywności ekonomicznej,
— rodzaj aktywności ekonomicznej w poprzednim roku.

Rodzaj aktywności ekonomicznej posiada trzy podstawowe warianty:

— **pracujący** – osoby, które w badanym tygodniu wykonywały jakąkolwiek pracę przynoszącą zarobek lub dochód; osoby pracujące dzieli się na dwie podkategorie:
  — pracujący w pełnym wymiarze czasu pracy – osoby pracujące, które w badanym tygodniu wykonywały pracę zarobkową przez co najmniej 40 godzin,
  — pracujący w niepełnym wymiarze czasu pracy – osoby pracujące, które w badanym tygodniu wykonywały pracę zarobkową przez mniej niż 40 godzin;
— **bezrobotni** – osoby w wieku 15 – 74 lat, które nie pracowały w badanym tygodniu i nie mają pracy, aktywnie poszukujące pracy i gotowe do jej podjęcia; lub osoby w wieku 15 – 74 lat, które nie pracowały w badanym tygodniu, nie mają pracy, nie poszukują pracy, ponieważ mają pracę załatwioną i oczekują na jej rozpoczęcie w ciągu trzech miesięcy i są gotowe ją podjąć.
— **bierni zawodowo** – to osoby, które:
  — nie pracowały w badanym tygodniu, nie miały pracy i nie poszukują pracy,
  — nie pracowały w badanym tygodniu, nie miały pracy, poszukują pracy, ale nie były zdolne (gotowe) do jej podjęcia,
  — mają 75 lat i więcej, nie pracowały w badanym tygodniu i nie mają pracy, aktywnie poszukują pracy i są gotowe do jej podjęcia,

— mają 75 lat i więcej, nie pracowały w badanym tygodniu, nie mają pracy, nie poszukują pracy, ponieważ mają pracę załatwioną i oczekują na jej rozpoczęcie w ciągu trzech miesięcy i są gotowe ją podjąć,
— nie pracowały w badanym tygodniu, nie miały pracy, poszukują pracy, ale nie podjęły aktywnych starań w ciągu ostatnich czterech tygodni, aby tę pracę znaleźć,
— nie pracowały w badanym tygodniu, nie miały pracy, nie poszukują pracy, ponieważ mają pracę załatwioną i oczekują na jej rozpoczęcie w okresie dłuższym niż 3 miesiące,
— nie pracowały w badanym tygodniu, nie miały pracy, nie poszukują pracy, ponieważ mają pracę załatwioną i oczekują na jej rozpoczęcie w ciągu 3 miesięcy, ale nie są gotowe do jej podjęcia.

Osoby pracujące i bezrobotne określa się mianem aktywnych zawodowo. Stopa bezrobocia według metodologii BAEL jest wyliczana jako stosunek liczby bezrobotnych do ogółu aktywnych zawodowo (pracujących i bezrobotnych – po uwzględnieniu wag analitycznych).

Badanie Aktywności Ekonomicznej Ludności jest przeprowadzane metodą reprezentacyjną, co oznacza, że istnieje możliwość uogólniania wyników pomiaru częściowego na całą populację generalną. Jednostki do badania losowane są metodą warstwową dwustopniową w tzw. cyklu rotacyjnym. Jednostkami losowania pierwszego stopnia są rejony statystyczne lub obwody spisowe (na wsiach). Jednostkami losowania drugiego stopnia są mieszkania. Jednostki pierwszego stopnia losowane są zgodnie ze schematem warstwowym, przy czym warstwy tworzą województwa. Na drugim stopniu losowania losuje się określoną liczbę mieszkań w ramach każdej jednostki terytorialnej. Pomiarowi poddaje się wszystkie osoby w wieku 15 lat i więcej zamieszkujące dane mieszkanie i tworzące gospodarstwo (lub gospodarstwa) domowe w danym lokalu.

Estymacja odbywa się poprzez określenie sposobu wyznaczania wag analitycznych dla każdej jednostki, przy czym wagi wyznaczane są w trzech etapach:

1. W pierwszej kolejności wyznacza się tzw. wagi pierwotne będące odwrotnością prawdopodobieństwa wyboru do próby poszczególnych mieszkań. Następnie wylicza się tzw. współczynniki realizacji $R = \frac{K-N}{K}$, gdzie K jest oszacowaniem (według wag pierwotnych) liczby mieszkań kwalifikujących się do badania, zaś N oszacowaniem liczby mieszkań kwalifikujących się do badania, lecz nie dających się zbadać z obojętnie jakich względów.

2. Następnie wylicza się tzw. wagi wtórne będące ilorazem wag pierwotnych i współczynnika realizacji R. Wagi wtórne są wagami finalnymi dla oszacowań dotyczących gospodarstw domowych.
3. W ostatnim etapie wylicza się wagi finalne dla osób. Wyliczenie odbywa się poprzez pomnożenie wag wtórnych przez odpowiednie modyfikatory. Modyfikatory konstruowane są w taki sposób, by dopasować oszacowania do aktualnych struktur demograficznych[72].

Rotacyjny charakter schematu doboru jednostek do próby oznacza, że dane mieszkanie losowane jest do próby kilkukrotnie, w systemie „dwa kwartały w badaniu, dwa kwartały przerwy, dwa kwartały w badaniu" [Zgierska 2010].

Rozmiar próby umożliwia uogólnienie wyników z „zadowalającą" precyzją w przekrojach pięcioletnich grup wieku, płci oraz klasy miejscowości zamieszkania w ujęciu województw.

3.2.2. Badanie Budżetów Gospodarstw Domowych (BBGD)

Badanie Budżetów Gospodarstw Domowych jest jednym z najdłużej przeprowadzanych badań w historii polskiej statystyki publicznej. Zostało zapoczątkowane już w okresie międzywojennym, jednak dopiero od lat 50-tych XX wieku przeprowadzane jest cyklicznie. Wyniki badania publikowane są z częstotliwością roczną, jednak w wyniku wielu zmian metodologicznych, pomiar przeprowadzany jest z częstością kwartalną.

BBGD stanowi podstawowe źródło informacji o warunkach bytowych poszczególnych grup ludności. Dostarcza danych o przychodach oraz rozchodach gospodarstw domowych w ujęciu szczegółowych kategorii oraz wyposażenia gospodarstw w dobra trwałego użytku. Umożliwia również przeprowadzanie szczegółowych analiz dotyczących różnic w poziomie dochodów i wydatków, spożycia oraz wyposażenia w różnych typach gospodarstw, a także umożliwia wskazanie przyczyn powstawania tych różnic.

Szacunki przeprowadzone na podstawie BBGD służą wielu celom, m.in. jako podstawa oceny bytowej mieszkańców kraju, jako narzędzie kreowania i kontroli polityki społeczno-ekonomicznej państwa i samorządów. Wykorzystywane są w sporządzaniu prognoz dotyczących spożycia indywidualnego, a także ustalania minimalnego wynagrodzenia. Równie ważnym aspektem badania jest szacowanie skali ubóstwa materialnego.

[72] Bilanse ludności sporządzane były w oparciu o struktury demograficzne wynikające z NSP 2002 (z odpowiednimi modyfikacjami). Struktury BAEL będą szacowane w oparciu o NSP 2011 wraz z zakończeniem opracowania jego wyników.

Jednostką badania jest gospodarstwo domowe, bez względu na liczbę stanowiących je osób. Gospodarstwo domowe w badaniu definiowane jest jako zespół osób (lub jedna osoba dla gospodarstw jednoosobowych) razem mieszkających i wspólnie się utrzymujących. Wielkość gospodarstwa określana jest liczbą osób wchodzących w skład gospodarstwa. Przedmiotem badania jest budżet gospodarstwa domowego rozumiany jako 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. Pomiarowi podlegają jednocześnie cechy demograficzne członków gospodarstwa domowego (w tym głowy), ich aktywności ekonomicznej, wyposażenia mieszkania, a także o subiektywnej oceny sytuacji materialnej gospodarstwa.

Badanie Budżetów Gospodarstw Domowych przeprowadzane jest metodą reprezentacyjną. Umożliwia to uogólnianie wyników na wszystkie gospodarstwa domowe w kraju. W 2005 roku zastosowano dwustopniowy, warstwowy schemat losowania z różnymi prawdopodobieństwami wyboru na pierwszym stopniu. Operat losowania jednostek pierwszego stopnia stanowił wykaz tzw. terenowych punktów badań (tpb) opracowany na potrzeby Narodowego Spisu Powszechnego 2002. Wykaz ten jest co roku aktualizowany w związku ze zmianami podziału administracyjnego państwa, powstawania nowych, jak i wyburzania starych budynków. Terenowy punkt badań w mieście liczy około 250 mieszkań, natomiast na wsi około 150. Na potrzeby BBGD w 2005 roku utworzono 29 tys. terenowych punktów badań [Marciniak 2006]. Operat losowania II stopnia stanowiły wykazy zamieszkanych mieszkań w poszczególnych terenowych punktach badań. Losowanie mieszkań przeprowadzono w oparciu o następujące założenia:

- w badaniu stosuje się model rotacji całkowitej z miesięcznym okresem wymiany próby[73],
- dla każdego miesiąca losuje się po dwa mieszkania w danym tpb, a w wylosowanych mieszkaniach badane są wszystkie gospodarstwa domowe,
- w wylosowanym mieszkaniu badanie przeprowadzane jest w danym miesiącu w dwóch kolejnych latach, tj. w 2004–2005 w podpróbce pierwszej w oraz 2005–2006 w podpróbce drugiej,
- w każdym tpb losuje się rezerwową próbę mieszkań ze względu na możliwość nieprzystąpienia do badania gospodarstw domowych zamieszkujących w wylosowanych mieszkaniach; mieszkania z próby rezerwowej wchodzą do bada-

[73] Rotacja całkowita oznacza, że wymianie podlegają wszystkie gospodarstwa domowe uczestniczące w badaniu w danym okresie.

nia w kolejności, w jakiej zostały wylosowane w miejsce mieszkań zamieszkanych przez gospodarstwa domowe nieprzystępujące do badania.

Wyniki badania przedstawiane są raz do roku w ujęciu makroregionów (NUTS1), klasy miejscowości zamieszkania, wielkości gospodarstw domowych (mierzoną liczbą członków) oraz typu gospodarstwa [Departament Badań Społecznych i Warunków Życia 2006].

3.2.3. Badanie Dochodów i Warunków Życia (EU-SILC)

Badanie Dochodów i Warunków Życia - EU-SILC (*Community Statistics on Income and Living Conditions* - Statystyka Dochodów i Warunków Życia Krajów Unii Europejskiej) jest międzynarodowym badaniem przeprowadzanym rocznie we wszystkich krajach Unii Europejskiej. Zostało one ustanowione rozporządzeniem Parlamentu Europejskiego (1177/2003 z modyfikacjami zawartymi w rozporządzeniu 553/2005) i zostało wdrożone w roku 2004 w większości krajów UE[74]. Przyczynkiem do wprowadzenia badania była konieczność stałej modyfikacji i dostosowywania do potrzeb odbiorców realizowanych badań statystycznych wywołana wzrostem zapotrzebowania użytkowników na różnego rodzaju informacje dotyczące szeroko rozumianych warunków życia ludności.

Celem badania EU-SILC jest pozyskanie podstawowego źródła porównywalnych na poziomie Unii Europejskiej danych z zakresu sytuacji dochodowej, ubóstwa i innych aspektów warunków życia ludności. W badaniu pozyskiwane są dane zarówno przekrojowe, jak i longitudinalne (uwzględniające zmiany w czasie).

Jednostkami badania w EU-SILC są prywatne gospodarstwa domowe oraz osoby w wieku 16 lat i więcej wchodzące w skład tych gospodarstw. Wielkość próby w każdym z krajów powinna zapewnić reprezentatywność wyników na poziomie narodowym zarówno w przypadku danych przekrojowych, jak i panelowych, przy czym okres obserwacji dla próby panelowej powinien wynosić przynajmniej 4 lata. Gospodarstwa domowe i ich członkowie poddawani są pomiarowi za pomocą odrębnych kwestionariuszy[75].

Badanie realizowane jest w okresie maj-czerwiec danego roku. W badaniu EU-SILC 2006 stosowane są różne okresy referencyjne. Okresem odniesienia dla zmiennych dochodowych jest ostatni pełny rok kalendarzowy (2005). Dla innych zmiennych prezentowa-

[74] Austria, Belgia, Grecja, Dania i Luksemburg rozpoczęły badanie w 2003 roku. Niemcy, Wielka Brytania, Holandia oraz nowe kraje członkowskie (z wyjątkiem Estonii) rozpoczęły badanie w roku 2005.

[75] W przypadku wywiadu indywidualnego dopuszcza się realizację tzw. wywiadu zastępczego przeprowadzonego z inną osobą z gospodarstwa domowego, która może udzielić wiarygodnych informacji o osobie objętej badaniem (dotyczy to osób zaliczonych w skład gospodarstwa domowego, a nieobecnych w miejscu zamieszkania w okresie trwania badania).

nych w tabelach okresem odniesienia jest sytuacja bieżąca[76]. Pytania dotyczące dochodów należą do kategorii pytań drażliwych i niektórzy respondenci odmawiają podania informacji o wysokości uzyskiwanych dochodów. Aby zmniejszyć wpływ braku odpowiedzi, w badaniu EU-SILC stosuje się imputację brakujących danych. W zależności od rodzaju brakujących informacji w badaniu stosuje się różne metody imputacji: metodę *hot-deck*, imputację regresyjną z symulowanymi resztami (imputację stochastyczną), imputację regresyjną deterministyczną oraz imputację dedukcyjną.

Pomiarowi w badaniu podlegają cechy z takich dziedzin, jak:

- podstawowe informacje dotyczące cech demograficznych respondentów,
- uczestnictwo w procesie edukacji,
- ocena stanu zdrowia,
- deprywacja materialna i niematerialna (wybrane aspekty)[77],
- warunki mieszkaniowe,
- aktywność ekonomiczna,
- poziom i źródła dochodów.

Dodatkowo, w każdym roku przeprowadzane są badania modułowe odpowiadające aktualnym potrzebom organów Unii Europejskiej. W 2006 roku przeprowadzono moduł poświęcony formom spędzania czasu wolnego (*Social participation*). Dokonano pomiaru takich cech, jak częstotliwość korzystania z różnych dóbr kultury (kina, teatru, koncertów, miejsc o szczególnych walorach kulturowych), uprawiania sportu, kontaktu z rodziną, przyjaciółmi lub sąsiadami, a także uczestnictwa w różnych organizacjach (m.in. politycznych, zawodowych, religijnych, sportowych itp.) [*European Commission* 2006].

Badanie EU-SILC jest przeprowadzane metodą reprezentacyjną, co umożliwia uogólnianie jego wyników na całą populację generalną. Stosowany schemat doboru próby jest dwustopniowy z różnymi prawdopodobieństwami wyboru na pierwszym stopniu. Jednostki pierwszego stopnia są przed losowaniem warstwowane, przy czym warstwami w 2005 roku były województwa (poziom NUTS2). Jako operat losowania wykorzystano Urzędowy Rejestr Podziału Terytorialnego Kraju TERYT. Jednostkami pierwszego stopnia (JPS) były obwody spisowe. Na drugim stopniu losowane były mieszkania. Wszystkie gospodarstwa domowe miesz-

[76] Ze względu na chęć zachowania zgodności czasowe opisywanych badań, zdecydowano, że opisane zostanie badanie przeprowadzone w 2006 roku. Powodem takiej decyzji było przypuszczenie, że cechy opisujące wyposażenie gospodarstw domowych, ich typ, cechy przestrzenne i sytuację społeczno-gospodarczą charakteryzują się mniejszą zmiennością, niż uzyskany dochód.

[77] W badaniu szacowany jest m.in. wskaźnik zagrożenia ubóstwem. Gospodarstwa zagrożone ubóstwem wg metodologii Eurostatu to gospodarstwa, których dochód ekwiwalentny jest mniejszy od 60% mediany dochodów ekwiwalentnych w kraju.

kające w wylosowanych mieszkaniach powinny wziąć udział w badaniu. W roku 2006 założono przebadanie 18162 adresów, jednak niedoskonałości operatu losowania (m.in. brak możliwości dotarcia do mieszkania lub brak mieszkania pod wskazanym adresem) nawiązano kontakt wyłącznie z 17224 mieszkaniami[78]. Ogółem przebadano 14914 gospodarstw domowych[79] i 34893 członków gospodarstw w wieku 16 lat i więcej.

Reprezentatywność wyników osiągnięto dzięki konstrukcji odpowiedniego systemu wag analitycznych. Wagi początkowe (pierwotne) wyznaczone zostały jako odwrotność prawdopodobieństwa wylosowania mieszkania do próby. Ze względu na odmowy odpowiedzi, wagi pierwotne zostały skorygowane o wskaźnik kompletności w poszczególnych klasach miejscowości oraz poddane kalibracji.

Wyniki badania przedstawiane są w podziale na makroregiony (NUTS1), klasę miejscowości zamieszkania, wielkość gospodarstw domowych (mierzoną liczbą członków) oraz jego typ.

3.3. Badania spoza systemu statystyki publicznej

3.3.1. Polski Generalny Sondaż Społeczny (PGSS)[80]

Polski Generalny Sondaż Społeczny od 1992 roku jest stałym programem badań statutowych Instytutu Studiów Społecznych Uniwersytetu Warszawskiego, finansowanym przez Komitet Badań Naukowych działający przy Ministerstwie Nauki i Szkolnictwa Wyższego. Badanie przeprowadzano cyklicznie realizując kolejne fale, wśród których część cech podlegała pomiarowi stale, a część mierzono w ramach specjalnych modułów tematycznych. Badanie przeprowadzono w latach 1992, 1993, 1994, 1995, 1997, 1999, 2002, 2005, 2008. W latach 1992-1995 badanie PGSS odbyło się w tym samym terminie (maj-czerwiec). W 1997 i 1999 roku badanie odbyło się w listopadzie i grudniu (około 20% wywiadów z badania PGSS 1997 zrealizowano w lutym i marcu), badanie w 2002 zrealizowano w kwietniu, zaś badanie 2005 i 2008 odpowiednio w styczniu i lutym.

Ogólnym celem badania jest systematyczny pomiar skutków zmian społecznych w Polsce. Wśród pozostałych celów badania wymienia się przede wszystkim:

[78] Wśród mieszkań, z którymi nawiązano kontakt, w 1851 spotkano się z odmową udziału w badaniu, w 240 mieszkaniach stwierdzono czasową nieobecność lokatorów, w 166 przypadkach brak było możliwości nawiązania kontaktu z gospodarstwem (m.in. z powodu chorób, podeszłego wieku mieszkańców lub alkoholizmu), natomiast w 48 zanotowano inne powody niewzięcia udziału w badaniu. Wśród gospodarstw, które zgodziły się wziąć udział w badaniu, w pięciu przypadkach nie dokonano pomiaru (nie podano przyczyn).

[79] W badaniu nie stosuje się próby rezerwowej.

[80] Opis sporządzono na podstawie dokumentacji badania [Cichomski *et al.* 2009].

— niekomercyjne udostępnianie zbiorów danych i dokumentacji metodologicznej PGSS społeczności badaczy i studentom nauk społecznych w Polsce oraz każdej zainteresowanej osobie i instytucji,
— dostarczanie badaczom w Polsce danych i wskaźników bezpośrednio porównywalnych z wynikami badań w innych krajach (m.in. poprzez dołączanie różnorodnych modułów problemowych z międzynarodowych badań *International Social Survey Programme* (ISSP), realizowanych w 45 krajach),
— umożliwienie systematycznych eksperymentów poświęconych testowaniu różnic w budowie kwestionariusza, w treści zadawanych pytań i w konstrukcji skali odpowiedzi,
— wydzielanie modułów tematycznych w kolejnych falach badania celem umożliwienia współpracy z różnymi ośrodkami naukowymi w Polsce.

We wszystkich falach badania pomiarowi podlegały takie grupy cech jak:

— charakterystyki społeczno-demograficzne,
— wskaźniki indywidualnych postaw, zogniskowanych na problemach życia rodzinnego, małżeństwa, sytuacji kobiety, aborcji, życia seksualnego oraz wartościach wychowawczych,
— położenie w systemie stratyfikacji i nierówności społeczne (m.in. aktywność ekonomiczna, pozycja zawodowa, poziom i charakter wykształcenia, sytuacja materialna – dochody indywidualne, dochody gospodarstwa domowego, sytuacja mieszkaniowa oraz wyposażenie mieszkania),
— nastroje ekonomiczne Polaków, a w tym skale poziomu zadowolenia z dochodów indywidualnych i z sytuacji materialnej rodziny oraz skale ocen bieżącego stanu i kierunków zmian w sytuacji gospodarczej kraju,
— subiektywne oceny pozycji społecznej i jej zmian w czasie,
— poglądy i opinie o strukturze społecznej, nierównościach społecznych i zmianach systemowych w Polsce,
— zachowania i preferencje wyborcze,
— subiektywne opinie dotyczące efektywności systemu politycznego,
— orientacje polityczne i ideologiczne,
— religia i religijność,
— postawy wobec etycznych dylematów współczesności (aborcji, rozwodów, eutanazji, kary śmierci) oraz przestrzegania prawa,

— satysfakcja z własnego życia,
— stan i ocena własnego zdrowia (w tym także palenie tytoniu i picie alkoholu).

Dodatkowo, w różnych falach badania przeprowadzono wiele modułów tematycznych, wśród których wskazać można:

— zagadnienia nierówności społecznych (1992 i 1999),
— ochrona środowiska (1993),
— rodzina i społeczne role kobiety (1994 i 2002),
— seksualizm (1994),
— tożsamość narodowa (1995 i 2005),
— postawy wobec pracy (1997),
— opinia o roli rządu i państwa (1997 i 2008),
— religia i religijność (1999),
— sieci społeczne (2002),
— społeczeństwo obywatelskie (2005),
— czas wolny i wypoczynek, aktywność fizyczna i uprawianie sportów (2008).

Jednostką badania była osoba dorosła (w wieku 18 lat i więcej), a badanie realizowano na terytorium całego kraju. Dane dla kolejnych edycji PGSS były gromadzone drogą indywidualnych wywiadów kwestionariuszowych, realizowanych na ogólnopolskich, reprezentatywnych próbach. Próba zastosowana w badaniach PGSS 1992-2002 była próbą adresową mieszkań, dobieraną z operatu Głównego Urzędu Statystycznego. W ramach wylosowanych gospodarstw domowych ankietowano wszystkich jego dorosłych członków.

W latach 2005 i 2008 operat losowania stanowił rejestr PESEL. Zmieniono również schemat doboru próby do badania. W 2005 roku zastosowano dwuetapowy warstwowy dobór próby.

W 2008 roku dobór próby był analogiczny, z pewnymi modyfikacjami dotyczącymi przede wszystkim rozdzielania niektórych obszarów na podobszary, w zależności od liczby ludności je zamieszkującej.

Tabela 3.10. Próba wylosowana i jej realizacja w poszczególnych edycjach PGSS

Edycja PGSS	Próba wylosowana	Poziom realizacji próby	
		N	%
1992	2000	1647	82,4
1993	2000	1649	82,5
1994	2000	1609	80,5
1995	2000	1603	80,2
1997	3200	2402	75,1
1999	3406	2282	67,0
2002	4008	2473	61,7
2005	2106	1277	60,6
2008	2495	1293	51,8
RAZEM	**23215**	**16235**	**69,9**

Źródło: [Cichomski *et al.* 2009]

Poziom realizacji próby w kolejnych falach badania PGSS systematycznie się zmniejszał (por. tabela 3.10). Długość kwestionariusza, a co się z tym wiąże, obciążenie respondenta sprawiało, że badanie charakteryzowało się bardzo dużym odsetkiem odmów odpowiedzi. Brak doboru tzw. próby rezerwowej (jak to ma miejsce w niektórych badaniach GUS, np. BBGD) sprawił, że ostateczna liczebność próby była dużo niższa niż pierwotnie zakładano. Odmowy odpowiedzi znalazły odzwierciedlenie w sposobie przeważania wyników. Korekta pierwotnych waga analitycznych, wynikających ze schematu doboru jednostek do próby, odbyła się za pomocą tzw. wag stratyfikacyjnych. Celem ich konstrukcji było dopasowanie struktury próby zrealizowanej do struktury badanej zbiorowości ze względu na następujące cechy społeczno-demograficzne: region zamieszkania, wielkość miejscowości, płeć i wiek.

Wyniki Polskich Generalnych Sondaży Społecznych udostępniane są nieodpłatnie, z pełną dokumentacją na stronie *http://pgss.iss.uw.edu.pl*. Forma udostępnienia to m.in. zbiór danych jednostkowych w formacie IBM SPSS wraz z pełnym opisem etykiet zmiennych i ich wariantów. Dodatkowo dostępne są wszystkie kwestionariusze oraz raporty z poszczególnych badań. Nie ma żadnych ograniczeń co do korzystania z danych.

3.3.2. Diagnoza Społeczna (DS)[81]

Diagnoza Społeczna jest badaniem ankietowym przeprowadzanym przez Radę Monitoringu Społecznego działającej przy niepublicznej Wyższej Szkole Finansów i Zarządzania w Warszawie. Celem badania jest pomiar warunków i jakości życia Polaków. W zamyśle autorów badanie jest „próbą uzupełnienia diagnozy opartej na wskaźnikach instytucjonalnych o kompleksowe dane na temat gospodarstw domowych oraz postaw, stanu ducha i zachowań osób tworzących te gospodarstwa".

Badaniu podlegają, za pomocą dwóch odrębnych kwestionariuszy, zarówno gospodarstwa domowe, jak i ich członkowie w wieku 16 lat i więcej. W badaniu uwzględnia się wszystkie ważne, zdaniem autorów, aspekty życia mieszkańców kraju, zarówno ekonomiczne, jak i społeczne. Diagnoza Społeczna ma charakter panelowy, tzn. w kolejnych falach badania część gospodarstw domowych podlega kolejnemu pomiarowi celem uchwycenia zmian oraz ich dynamiki w sytuacji gospodarstw i ich członków. Badanie rozpoczęło się w roku 2000, a kolejne edycje miały miejsce w latach 2003, 2005, 2007, 2009 oraz 2011. Na rok 2013 zaplanowana jest kolejna edycja. Badanie zawsze prowadzone jest w marcu, co służyć ma wytrąceniu efektu sezonowości. W 2009 i 2011 r. ze względu na wielkość próby pomiar przedłużył się do połowy kwietnia.

Projekt obejmuje wiele aspektów związanych z sytuacją gospodarstw domowych i poszczególnych obywateli. Uwzględnione w nim wskaźniki społeczne podzielić można na trzy ogólne klasy:

1. struktura demograficzno-społeczna gospodarstw domowych,
2. warunki życia gospodarstw domowych, związane z ich kondycją materialną, dostępem do świadczeń medycznych, do kultury i wypoczynku, edukacji i nowoczesnych technologii komunikacyjnych,
3. jakość, styl życia i cechy indywidualne obywateli.

Cechy demograficzno-społeczne gospodarstw nie są przedmiotem analizy, a służą jedynie do grupowania wyników oraz określenia różnic pomiędzy poszczególnymi grupami. Właściwym przedmiotem analizy są przede wszystkim warunki życia gospodarstw domowych i jakość życia obywateli. Strukturę cech poddanych pomiarowi można w pewnym uproszczeniu podzielić na obiektywne (np. dochody, warunki mieszkaniowe) i subiektywne (np. ocena jakości życia). Pomiar warunków życia gospodarstwa domowego obejmował:

— sytuację dochodową gospodarstwa domowego i sposób gospodarowania dochodami,

[81] Opis sporządzono na podstawie raportu z badania [Czapiński *et al.* 2005].

— wyżywienie,
— zasobność materialną gospodarstwa domowego, w tym wyposażenie w nowoczesne technologie komunikacyjne (telefon komórkowy, komputer, dostęp do Internetu),
— warunki mieszkaniowe,
— pomoc społeczną, z jakiej korzysta gospodarstwo domowe,
— kształcenie dzieci,
— uczestnictwo w kulturze i wypoczynek,
— korzystanie z usług systemu opieki zdrowotnej,
— sytuację gospodarstwa domowego na rynku pracy,
— korzystanie z pomocy społecznej,
— ubezpieczenia i zabezpieczenie emerytalne,
— ubóstwo, bezrobocie, niepełnosprawność i inne aspekty wykluczenia społecznego.

Wskaźniki jakości i stylu życia indywidualnych respondentów obejmowały:

— ogólny dobrostan psychiczny (w tym: wolę życia, poczucie szczęścia, zadowolenie z życia, symptomy depresji psychicznej),
— zadowolenie z poszczególnych dziedzin i aspektów życia,
— subiektywną ocenę materialnego poziomu życia,
— różne rodzaje stresu życiowego (w tym: stres administracyjny, tzw. „kafkoski", związany z kontaktami z administracją publiczną, stres zdrowotny, stres rodzicielski, stres finansowy, stres pracy, stres ekologiczny, stres małżeński, problemy związane z opieką nad osobami starszymi, stresowe wydarzenia losowe, jak napad, włamanie, aresztowanie),
— objawy somatyczne (miara dystresu),
— strategie radzenia sobie ze stresem,
— ocenę kontaktów z systemem opieki zdrowotnej,
— finanse osobiste (w tym: dochody osobiste, ubezpieczenia i zabezpieczenie emerytalne),
— system wartości, skłonność do ryzyka, styl życia oraz indywidualne zachowania i nawyki (m.in. palenie papierosów, nadużywanie alkoholu, używanie narkotyków, praktyki religijne),
— postawy i zachowania społeczne, w tym kapitał społeczny,
— wsparcie społeczne,
— ogólną ocenę procesu transformacji i jego wpływu na własne życie respondentów,

— korzystanie z nowoczesnych technologii komunikacyjnych – komputera, Internetu, telefonu komórkowego.

W badaniu zastosowano dwa rodzaje ankiet: osobną dla gospodarstw domowych (skierowaną do głów gospodarstw domowych) i osobną dla poszczególnych członków gospodarstwa domowego. Badanie realizowane jest przez zawodowych ankieterów Głównego Urzędu Statystycznego. Nadzór nad badaniem ankietowym sprawuje od strony organizacyjnej Biuro Badań i Analiz Polskiego Towarzystwa Statystycznego.

Badanie przeprowadzane jest metodą reprezentacyjną, umożliwiającą uogólnienie wyników na całą populację gospodarstw domowych w kraju oraz wszystkich Polaków w wieku 16 lat i więcej. Np. w 2005 roku badaniem objęto 3851 gospodarstw domowych w całym kraju oraz 8828 osób w wieku 16 lat i więcej. Dodatkowo zebrano podstawowe informacje (przede wszystkim demograficzne) o wszystkich członkach gospodarstw domowych, niezależnie od wieku, w liczbie ogółem 12887. Ze względu na panelowy charakter badania w próbie znalazła się część gospodarstw i ich członków, którzy podlegali pomiarowi we wcześniejszych falach badania. Schemat doboru jednostek do próby był warstwowy i dwustopniowy. Warstwy stanowiły klasy miejscowości zamieszkania[82] w ramach każdego z 16 województw. Jednostkami losowania pierwszego stopnia w warstwach miejskich w poszczególnych województwach były rejony statystyczne (obejmujące co najmniej 250 mieszkań), a w warstwach wiejskich obwody statystyczne. Na drugim stopniu losowano systematycznie po dwa mieszkania z uporządkowanej losowo listy mieszkań, niezależnie wewnątrz każdej z warstw utworzonych na pierwszym stopniu. W przypadku odmów uczestnictwa w badaniu włączano gospodarstwa z prób rezerwowych, należących do tego samego rejonu statystycznego. Uzyskane w badaniu wyniki, w celu zachowania reprezentatywności, tak dla badania w 2005 roku jak dla panelu 2000-2005, w skali kraju oraz dla poszczególnych województw i wyróżnionych klas miejscowości zamieszkania, podlegały odpowiedniemu ważeniu.

Badanie Diagnoza Społeczna, mimo iż finansowane głównie z pieniędzy prywatnych, ma charakter publiczny. Wyniki badania udostępnione są bezpłatnie pod adresem *http://diagnoza.com*. Forma udostępnienia to m.in. zbiór danych jednostkowych w formacie IBM SPSS wraz z pełnym opisem etykiet zmiennych i ich wariantów. Dodatkowo dostępne są wszystkie kwestionariusze oraz raporty z poszczególnych fal badania. Nie ma żadnych ograniczeń co do korzystania z danych.

[82] Miasta pow. 100 tys. mieszkańców, miasta poniżej 100 tys. mieszkańców oraz wieś.

3.4. Wnioski

Rejestry administracyjne stanowią cenne źródło danych społeczno-ekonomicznych. Stopień pokrycia populacji oraz podstawowe cechy demograficzne połączone z możliwością tworzenia zmiennych pochodnych rozszerzających zawartość merytoryczną czynią z nich wartościowe źródło informacji. Pomimo iż liczba cech poddanych pomiarowi nie jest tak duża jak w badaniach reprezentacyjnych, rejestry mogą służyć jako bogate źródło zmiennych pomocniczych. Zastosowanie zbiorów administracyjnych w integracji może w istotny sposób wzbogacić zasoby wiedzy społeczno-gospodarczej.

Prowadzane przez organy statystyki publicznej badania próbkowe charakteryzują się szerokim zakresem merytorycznym przy stosunkowo niewielkiej próbie. Zastosowanie metody reprezentacyjnej umożliwia uogólnianie wyników na całą populację generalną, jednak ograniczenia związane z liczebnością próby powodują, że zadowalającą jakość oszacowań można uzyskać tylko w nielicznych przekrojach (tak terytorialnych, jak i merytorycznych). Pomiarowi w opisywanych badaniach reprezentacyjnych często poddane są podobne grupy cech. Stosowane są również podobne metody doboru próby. Zobowiązania międzynarodowe często wymuszają także stosowanie ujednoliconych definicji. Spełnienie powyższych warunków stanowi dobrą podstawę do stosowania metod statystycznej integracji danych (w szczególności parowania statystycznego) celem zwiększenia zawartości merytorycznej badań, a także podjęcia próby zwiększenia liczebności wejściowych repozytoriów.

Instytucje znajdujące się poza systemem statystyki publicznej również prowadzą badania społeczne, często korzystają z bogatego doświadczenia Głównego Urzędu Statystycznego, stosują metodologię zbliżoną do badań statystyki publicznej. Celem tych badań jest prowadzenie analiz porównawczych oraz rozszerzenie zakresu informacji społecznych. Ze względu na bardzo szeroką zawartość merytoryczną badań (długi kwestionariusz), pomiar dokonywany jest na stosunkowo niedużej próbie. Jednak duża liczba cech, z których wiele jest wspólnych z badaniami statystyki publicznej, również stwarza możliwość integracji ze źródłami administracyjnymi i pochodzącymi z badań statystyki publicznej. Dodatkowo, niczym nieograniczona dostępność zbiorów jednostkowych badań PGSS i DS daje sposobność szerokiego wykorzystania tych repozytoriów.

Zintegrowanie istniejących zbiorów danych może stać się przyczynkiem do konstrukcji repozytorium informacji o społeczno-ekonomicznych aspektach życia ludności. Rejestr PESEL, zawierający dane o podstawowych charakterystykach demograficznych ludności mógłby zostać, podobnie jak Ewidencja Ludności w holenderskim spisie wirtualnym, „kręgosłupem" takiego repozytorium. Informacje o aktywności ekonomicznej, wymiarze czasu pracy, czy

stopniu niepełnosprawności zawarte w rejestrach ZUS i NFZ mogłyby zostać dołączone deterministycznie na podstawie numeru PESEL. Następnie dołączenie danych o dochodach ludności z rejestru POLTAX wraz z informacją o powiązaniu miejsca pracy i miejsca zamieszkania umożliwiłoby łączną, kompleksową informację o wzajemnych powiązaniach między aktywnością ekonomiczną, uzyskiwanych dochodach i charakterystykach demograficznych (por. tabela 3.11). W przypadku braku informacji o numerze PESEL, istnieje możliwość utworzenia klucza złożonego, np. z imienia, nazwiska, adresu i daty urodzenia.

Tabela 3.11. Źródła danych w konstrukcji zintegrowanego repozytorium danych społecznych

Lp.	Nazwa	Rodzaj badania/ populacja objęta badaniem	Rekord/ jednostka	Liczba rekordów	Wybrane zmienne/zagadnienia
1	PESEL	pełne	osoba	ok. 38,7 mln	płeć
					data urodzenia
					stan cywilny
					adres
2	ZUS	pełne	płatnik ubezpieczenia społecznego	16 214 456	płeć
					data urodzenia
					adres
					status na rynku pracy[1]
					wymiar czasu pracy
					status emerytalno-rentowy
					stopień niepełnosprawności
3	NFZ	pełne	osoba	38 647 138	płeć
					data urodzenia
					adres
					status na rynku pracy[1]
4	POLTAX[2]	pełne	pracujący, emeryci, renciści	ok. 19 mln	adres miejsca zamieszkania
					adres miejsca pracy
					przychód
					dochód
					wysokość podatku
					wysokość składek ubezp.
5	BAEL	częściowe	osoba w wieku 15 lat i więcej	111 807[3]	płeć
					wiek
					wykształcenie
					stan cywilny
					status na rynku pracy
					wymiar czasu pracy
					zawód wykonywany
			gospodarstwo domowe	59 994[3]	źródło utrzymania gosp. dom.
					typ gospodarstwa dom.
6	BBGD	częściowe	osoba	107 124[3]	płeć
					wiek
					stan cywilny
					status na rynku pracy
					wykształcenie
			gospodarstwo domowe	34 767[3]	wydatki gosp. dom. w przekroju szczegółowych kategorii
					dochody gosp. dom. w przekroju

Lp.	Nazwa	Rodzaj badania/ populacja objęta badaniem	Rekord/ jednostka	Liczba rekordów	Wybrane zmienne/zagadnienia
					szczegółowych kategorii
					charakterystyka lokalu zajmowanego przez gosp. dom.
					wyposażenie gosp. dom.
					wielkość spożycia towarów i usług w gosp. dom.
7	EU-SILC	częściowe	osoba w wieku 16 lat i więcej	36 590[4]	płeć
					wiek
					stan cywilny
					wykształcenie
					dostęp do różnych usług
					dochody osobiste w przekroju szczegółowych kategorii
					różne aspekty jakości życia
			gospodarstwo domowe	14 914[4]	dochody gosp. dom. w przekroju szczegółowych kategorii
					charakterystyka lokalu zajmowanego przez gosp. dom
					wyposażenie gosp. dom.
					subiektywna sytuacja materialna
					warunki mieszkaniowe
					wskaźnik ubóstwa materialnego
8	PGSS	częściowe	osoba w wieku 18 lat i więcej	1293[5]	różne charakterystyki społeczno-demograficzne
					nastroje społeczne
					preferencje wyborcze
					opinie dotyczące otoczenia społeczno-gospodarczego
					sposoby spędzania wolnego czasu
9	DS	częściowe	osoba w wieku 16 lat i więcej	26 420[6]	różne charakterystyki społeczno-demograficzne
					stan zdrowia
					systemy wartości i postawy społeczne
					opinie dotyczące otoczenia społeczno-gospodarczego
			gospodarstwo domowe	12 387[6]	sytuacja dochodowa
					warunki mieszkaniowe
					uczestnictwo w kulturze i wypoczynek
					różne aspekty wykluczenia społecznego

Uwaga:
[1] Zmienna pochodna zmiennej „kod ubezpieczenia".
[2] W zakresie osób fizycznych.
[3] Dotyczy 2005 roku. Dla liczebności BAEL wyeliminowano duplikaty wynikające z rotacyjnego doboru próby.
[4] Dotyczy 2006 roku.
[5] Dotyczy 2008 roku.
[6] Dotyczy 2011 roku.
Liczba rekordów (po deduplikacji) dla ZUS i NFZ została wyznaczona na podstawie analizy zawartości informacyjnej opisanej w rozdziale. Dla PESEL i POLTAX podano wartości orientacyjne ze względu na brak dostępu do zbiorów jednostkowych.
Źródło: opracowanie własne

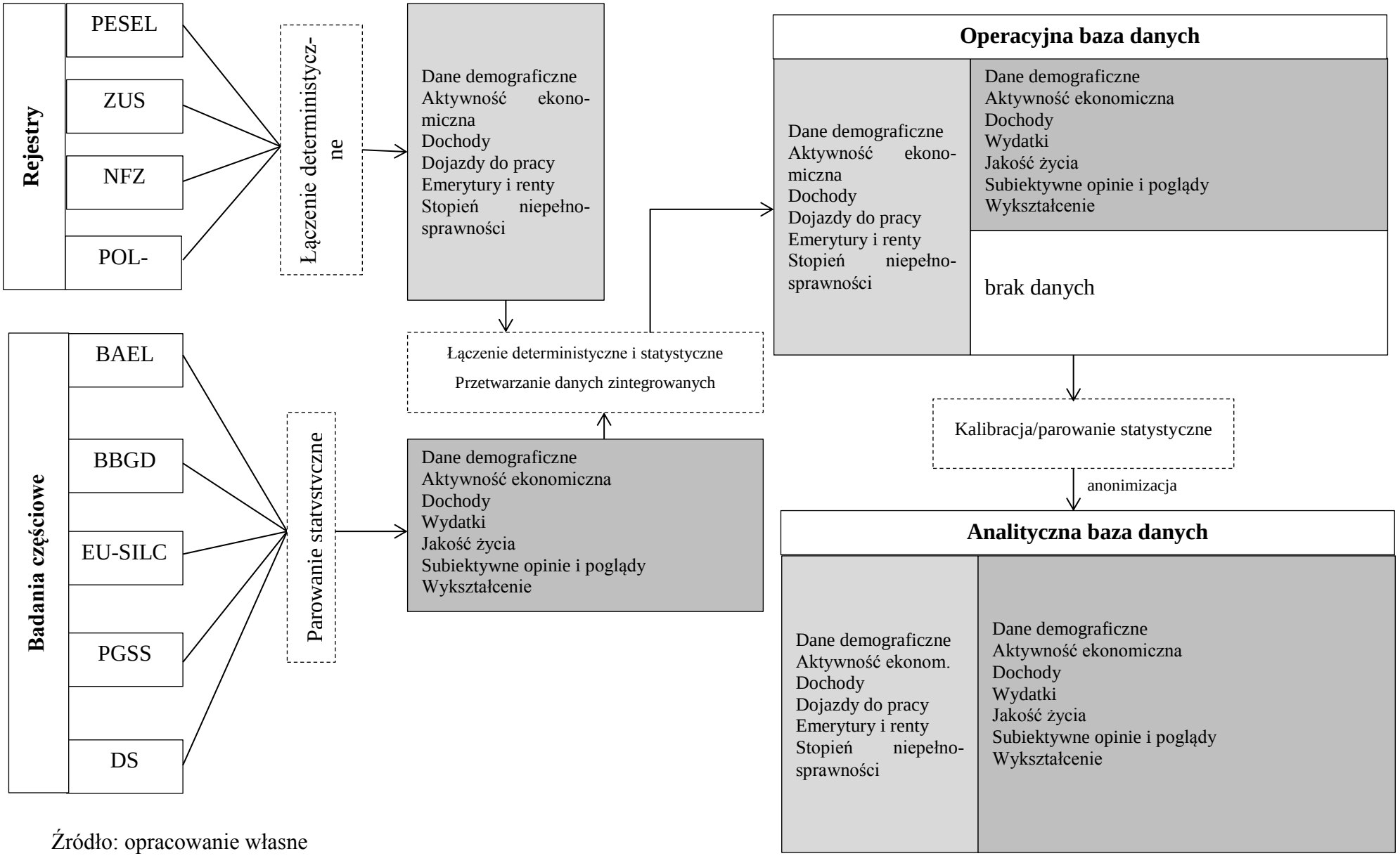
Zbiory danych pochodzące z badań reprezentacyjnych zawierają informację bogatą merytorycznie, jednak o niskim pokryciu. Ich integracja, metodami parowania statystycznego, zapewni łączną obserwację cech o bardzo szerokim spektrum – cech obiektywnych, jak dochody i wydatki, jak również subiektywnych, jak opinie dotyczące jakości życia, wyznawanych poglądów i sposobów spędzania wolnego czasu. Wykorzystanie technik umożliwiających konkatenację repozytoriów[83] pozwoliłoby na częściowe zniwelowanie problemów wynikających z niedostatecznie dużej liczebności próby dla wnioskowania dla szczegółowych przekrojów.

Zintegrowane repozytoria rejestrów i badań reprezentacyjnych można połączyć stosując metody zarówno deterministyczne, jak i statystyczne (por. schemat 3.1). Łączenie deterministyczne mogłoby odbyć się poprzez utworzenie klucza złożonego danych wych[84] i podstawowych charakterystyk demograficznych (np. wiek/data urodzenia, płeć). W przypadku, gdy klucz złożony nie identyfikowałby w sposób jednoznaczny jednostek, można zastosować metody statystyczne.

Tak zintegrowane repozytorium danych społeczno-ekonomicznych należy poddać procedurze przetwarzania danych zintegrowanych. Harmonizacja definicji, wariantów cech, jak również populacji zapewni rzetelność danych w zintegrowanym zbiorze (por. pkt. 1.5). Powstała w ten sposób operacyjna baza mikrodanych zawierałaby wiele braków danych, np. dla cech występujących wyłącznie w badaniach reprezentacyjnych. Spójność oszacowań może zostać zapewniona poprzez techniki kalibracji lub zmodyfikowane techniki masowej imputacji – przy wykorzystaniu parowania statystycznego (por. pkt. 1.4). Następnie zintegrowane repozytoria po anonimizacji rekordów utworzą analityczną bazę mikrodanych. Zastosowanie parowania statystycznego w celu dołączenia wartości zmiennych z badań reprezentacyjnych do rekordów pochodzących z rejestrów umożliwia przedstawienie szacunków na niskich poziomach agregacji. Może stać się to przyczynkiem do utworzenia kompleksowego systemu informacji społecznych, jak ma to miejsce w przypadku Programu Statystyki Sąsiedztwa w Wielkiej Brytanii.

[83] Opisanych w rozdziale IV i V.
[84] Znanych z operatu losowania.

Schemat 3.1. Koncepcja utworzenia zintegrowanego repozytorium danych społecznych



Źródło: opracowanie własne

W kolejnym rozdziale przedstawione zostaną metody statystycznej integracji zbiorów danych. Szczególna uwaga poświęcona zostanie metodologii integracji zbiorów, w których występują te same jednostki (np. rejestrów administracyjnych, spisów itp.), jak również integracji zbiorów danych pochodzących z badań próbkowych. Przedstawione zostaną metody harmonizacji definicji zmiennych, ich wariantów, a także zaprezentowane zostaną formalne metody poszukiwania rekordów odnoszących się do tych samych jednostek lub rekordów podobnych. Szczegółowo opisane zostaną metody oceny jakości zintegrowanych w sposób statystyczny zbiorów danych, jak również zilustrowane zostaną problemy wynikające ze stosowanych metod.

ROZDZIAŁ IV. STATYSTYCZNE METODY INTEGRACJI DANYCH

4.1. Klasyfikacja metod

W literaturze przedmiotu zasadniczo wyróżnia się dwie grupy metod integracji: deterministyczną i stochastyczną [Raessler 2002; Lenz, Zwick 2009; D'Orazio *et al.* 2006; Herzog *et al.* 2007; i in.]. Integracja deterministyczna polega w głównej mierze na dołączaniu zmiennych z jednego zbioru do drugiego na podstawie unikatowego klucza połączeniowego. Kluczem takim w źródłach administracyjnych może być np. numer PESEL (dla osób), numer REGON (dla przedsiębiorstw) czy też numer rejestracyjny samochodu. Integracja danych w takim przypadku odbywa się na podstawie wartości zmiennej. Unikalny klucz połączeniowy może również zostać utworzony na podstawie kilku zmiennych. W takiej sytuacji ich wartości muszą identyfikować każdą jednostkę w łączonych zbiorach. Ponadto oba łączone zbiory muszą zawierać wszystkie zmienne składające się na klucz, a ich warianty powinny być w pełni zharmonizowane. Braki danych, czy błędne zapisy zmiennych tworzących klucz powodują, że integracja deterministyczna może okazać się niemożliwa. W wielu repozytoriach danych klucz nie występuje lub jest celowo usuwany (np. ze względu na ochronę danych osobowych). W takich przypadkach korzystanie z metody deterministycznej jest niemożliwe. Jest ona nieprzydatna także w sytuacji łączenia zbiorów z badań próbkowych. Prawdopodobieństwo wylosowania do próby w różnych badaniach tej samej jednostki jest na tyle małe, że można założyć, że zbiory takie są rozłączne. Do integracji baz danych stosuje się wówczas metody stochastyczne.

Stochastyczne metody integracji danych umożliwiają połączenie zbiorów poprzez:

- obliczenie na podstawie wartości zmiennych parujących tak zwanej „wagi połączeniowej" wyrażającej iloraz prawdopodobieństwa, że badana para rekordów należy do tej samej jednostki oraz prawdopodobieństwa, że do niej nie należy (zbiory zawierające informacje o tych samych jednostkach)
- łączenie rekordów podobnych co do wartości tzw. zmiennych parujących (zbiory rozłączne).

Integracja metodami stochastycznymi oznacza łączenie dwóch (lub więcej) zbiorów danych w celu uzyskania jednego zbioru, w którym zmienne ze wszystkich integrowanych zbiorów są obserwowane łącznie. Wśród metod stochastycznych w literaturze przedmiotu wymienia się dwa główne nurty:

1. probabilistyczne łączenie rekordów (*probabilistic record linkage*),

2. parowanie statystyczne (*statistical matching, data fusion, data merging, data matching, mass imputation, microsimulation modeling, file concatenation*)

Pierwszy nurt wykorzystywany jest w przypadku, gdy integrowane zbiory zawierają informacje o tych samych rzeczywistych jednostkach pochodzących z jednej populacji. W pierwszej kolejności dokonuje się, w każdym ze zbiorów, identyfikacji rekordów, które z wysokim prawdopodobieństwem zawierają informacje o tej samej jednostce (np. osobie, gospodarstwie domowym, czy też przedsiębiorstwie). Następnie rekordy te łączy się uzyskując zbiór danych zawierający łączną obserwację cech z integrowanych źródeł. Identyfikacja rekordów należących do tej samej jednostki odbywa się na podstawie analizy wartości występujących w tak zwanych zmiennych parujących wybieranych z zestawu zmiennych występujących w obu integrowanych zbiorach (np. adres zamieszkania, płeć, data urodzenia, stan cywilny etc.). Jeżeli zmienne te są zgodne (lub prawie zgodne) dla danych rekordów w integrowanych zbiorach, uznaje się je za należące do tej samej jednostki.

Drugą grupę metod określanych parowaniem statystycznym stosuje się, gdy niemożliwa jest identyfikacja tych samych jednostek w analizowanych zbiorach. Sytuacja taka zachodzi w przypadku łączenia zbiorów danych pochodzących z badań próbkowych, gdy prawdopodobieństwo wylosowania tej samej jednostki do dwóch prób jest bardzo niskie. W takiej sytuacji nie identyfikuje się rekordów należących do tej samej jednostki (co wobec braku ich występowania jest niemożliwe), a wyszukuje się rekordy podobne pod względem wartości pewnych zmiennych parujących. Efektem jest jednostkowy, syntetyczny zbiór danych[85] zawierający łączną informację o integrowanych zmiennych. Jednostki występujące w nowo utworzonym zbiorze są jednostkami nierzeczywistymi (syntetycznymi), a więc nie posiadają swoich odpowiedników w rzeczywistości, jednak są „reprezentatywne" dla badanej populacji pod względem rozkładu dołączanych cech.

Należy zwrócić uwagę, że obie metody, choć przeznaczone do różnych zadań, posiadają pewne wspólne charakterystyki:

— działanie na jednostkowych zbiorach danych;

— działanie na zmiennych występujących w obu zbiorach (tzw. zmiennych wspólnych).

Charakterystykami, które odróżniają te dwie metody są:

— identyfikacja rekordów należących do tej samej jednostki (probabilistyczne łączenie rekordów),

[85] Efektem mogą być również charakterystyki łącznego rozkładu zmiennych, jak współczynnik korelacji, czy tabela kontyngencji.

— identyfikacja rekordów z integrowanych zbiorów będących reprezentatywnymi dla populacji docelowej (parowanie statystyczne).

Pierwsza ze wspólnych charakterystyk prowadzi do identyfikacji populacji, które są opisywane przez integrowane źródła danych i ich harmonizacji. Druga wspólna charakterystyka wymaga harmonizacji definicji, klasyfikacji, okresów referencyjnych oraz innych własności zmiennych wspólnych w celu zapewnienia ich pełnej homogeniczności.

Pierwsza z cech odróżniających oba nurty dotyczy wyodrębnienia zmiennych wspólnych (najbardziej odpowiednich dla identyfikacji rekordów należących do tej samej jednostki) oraz zmiennych grupujących (dzielących zbiór na rozłączne klasy połączeniowe). Zdefiniowana musi zostać postać funkcji porównującej rekordy, jak również mechanizm powstawania błędów połączenia. Natomiast w przypadku parowania statystycznego, zdefiniowany musi zostać pewien statystyczny model łączący integrowane zbiory oraz należy wybrać „najbardziej odpowiednie" zmienne wspólne, za pomocą których wyznaczone zostaną reprezentatywne rekordy populacji docelowej.

Wykorzystanie poszczególnych metod podyktowane jest zarówno charakterem dostępnych zbiorów, ich jakością oraz zgodnością badanych zbiorowości. Zbiory danych często opisują odmienne (przynajmniej częściowo) populacje, charakteryzują się różnymi definicjami zmiennych, okresami referencyjnymi, a także brakami danych i błędami w zapisie wariantów poszczególnych cech. Kwestie te są bardzo ważne w kontekście integracji zbiorów danych pochodzących z różnych źródeł, a ich właściwa diagnoza i harmonizacja to istotne składowe omawianego podejścia.

4.2. Harmonizacja źródeł danych przed integracją

Harmonizacja jest pierwszym, bardzo ważnym, etapem integracji danych, bez względu na stosowaną metodę. Umożliwia ona m.in. porównanie rozkładów zmiennych z różnych źródeł oraz późniejszą ocenę rezultatów integracji. Harmonizacja jest procesem czaso- i pracochłonnym, jednak niezbędnym. Według van der Laana [2000] wyróżnić można 8 następujących etapów (por. również [Scanu 2008]):

1. harmonizacja definicji jednostek;
2. harmonizacja okresów referencyjnych;
3. badanie kompletności populacji;
4. harmonizacja zmiennych;
5. harmonizacja wariantów cech;
6. korekta błędów pomiaru;

7. korekta związana z brakami danych (imputacja);
8. tworzenie zmiennych pochodnych.

W celu dokonania pewnego uproszczenia można podzielić powyższe etapy w trzy grupy: zgodność populacji i jednostek (1, 2, 3), harmonizację zmiennych (4, 5, 8) oraz inne aspekty operacyjne oraz związana z nimi operacja blokowania (6, 7). Powyższe etapy harmonizacji danych można odnieść zarówno do probabilistycznego łączenia rekordów, jak i parowania statystycznego. Przeprowadzając je należy jednak pamiętać o odmienności problemów rozwiązaniu których służą poszczególne metody. W poprzedniej sekcji wspomniano o różnicach w zastosowaniu obu nurtów metodologicznych:

— probabilistyczne łączenie rekordów stosuje się w przypadkach, gdy celem badania jest identyfikacja rekordów (w dwóch zbiorach) należących do tej samej jednostki;
— celem parowania statystycznego jest łączna analiza (przynajmniej) dwóch zmiennych, które nie są łącznie obserwowane w jednym zbiorze danych; dokonuje się tego poprzez zbadanie dwóch różnych zbiorów danych, gdzie jedna z analizowanych zmiennych znajduje się w pierwszym zbiorze, a druga w drugim.

Harmonizacja zmiennych wspólnych w integrowanych zbiorach, populacji i definicji jednostek jest niezwykle ważnym etapem integracji. Źle przeprowadzona harmonizacja lub zupełne pominięcie tego etapu może skutkować nieakceptowalną jakością zintegrowanych zbiorów [Gill 2001].

Zgodność populacji i jednostek

Integracja dwóch źródeł danych jest uzasadniona gdy:

— okresy referencyjne obu integrowanych zbiorów są takie same;
— zbory odnoszą się do dwóch takich samych lub różnych, lecz częściowo pokrywających się populacji.

W pierwszym przypadku zastosować można zarówno metodę probabilistycznego łączenia rekordów, jak i parowania statystycznego, w zależności od charakteru integrowanych źródeł. Drugi przypadek występuje znacznie częściej i wymaga decyzji dotyczące techniki integracji. Za pomocą metody probabilistycznego łączenia rekordów można podjąć próbę wykrycia rekordów należących do tej samej jednostki, a dla pozostałych rekordów dokonać integracji metodą parowania statystycznego. Nie są to jednak procesy automatyczne. W pierwszej kolejności w integrowanych zbiorach, oznaczonych jako A i B, należy wyodrębnić zbiory $A1$ oraz $B1$, których zawartość odnosi się do wspólnej części populacji. Należy zweryfikować,

czy uzyskane w ten sposób podpróbki są reprezentatywne dla badanej zbiorowości. Jeżeli weryfikacja przebiegnie pomyślnie, zastosowanie metody parowania statystycznego na zbiorach $A1$ i $B1$ może być zasadne. Alternatywnym podejściem do zastosowania techniki parowania statystycznego na zbiorach A i B jest przyjęcie założenia, że te dwie badane populacje są próbą losową pochodzącą z tego samego procesu generowania danych. Innymi słowy można przyjąć, że dwa zbiory jednostek, które nie mają żadnego „połączenia" nie zmienią rozkładu analizowanych zmiennych. W takim przypadku nie ma potrzeby redukcji zbiorów A i B do podzbiorów $A1$ i $B1$ [Scanu 2008].

Typowym przypadkiem braku homogeniczności integrowanych zbiorów jest ich prze- sunięcie w czasie. Jeżeli część rekordów należy do tej samej jednostki (w dwóch różnych okresach czasowych), można rozważyć zastosowanie probabilistycznego łączenia rekordów w celu utworzenie, na przykład, badania panelowego.

Gdy dwa zbiory odnoszą się do dwóch różnych (rozłącznych) populacji, żadna z metod integracji nie będzie właściwa [Scanu 2008].

Harmonizacja zmiennych

Zmienne występujące w obu zbiorach (tzw. zmienne wspólne), potencjalne zmienne parujące[86], muszą charakteryzować się pełną homogenicznością. Oznacza to, że zarówno rozkłady, jak i definicje tych zmiennych muszą cechować się wysoką zgodnością. W zbiorach danych pochodzących z różnych źródeł spełnienie obu tych warunków w pełni może okazać się trudne, dlatego przed przystąpieniem do integracji należy przeprowadzić harmonizację zmiennych wspólnych. Najczęściej spotykanymi problemami występującymi na tym etapie są:

- różne definicje zmiennych i występowanie różnych wariantów cech,
- występowanie braków danych,
- różnice w rozkładach.

W przypadku niezgodności definicji i klasyfikacji można wyróżnić trzy rodzaje zmiennych wspólnych:

1. Zmienne, dla których nie ma możliwości przeprowadzenia harmonizacji.

Zmienne takie nie powinny być uznawane za 'wspólne', a więc nie powinno się rozważać możliwości zastosowania ich jako zmiennych parujących. Sytuacja taka zdarza się stosun-

[86] Zmienne biorące udział w procesie integracji.

kowo często, zwłaszcza, gdy do integracji przeznaczone są zbiory pochodzące z różnych instytucji [Scanu 2008].

2. Zmienne, które można zharmonizować modyfikując ich warianty.

Cechy jakościowe zawierają często wiele wariantów. Ich harmonizacja zwykle odbywa się poprzez agregację w taki sposób, by utworzone warianty pochodne były zgodne w odpowiadających sobie zmiennym wspólnym w obu integrowanych zbiorach (np. kategorie „miasto do 10 tys. mieszkańców" i „miasto od 10 tys. do 20 tys. mieszkańców" w zmiennej „Klasa miejscowości zamieszkania" można połączyć w nową kategorię „miasto do 20 tys. mieszkańców").

3. Nowe zmienne wspólne, będące zmiennymi pochodnymi.

W przypadku braku odpowiednich zmiennych wspólnych lub ich niewystarczającej liczby, istnieje możliwość utworzenia nowych zmiennych poprzez przekształcenie innych cech zawartych w integrowanych zbiorach. Wówczas takie zmienne pochodne, jeżeli spełniają określone kryteria (jakościowe i definicyjne), mogą zostać użyte jako zmienne parujące.

Zmienne wspólne powinny odznaczać się również odpowiednią jakością. Oznacza to m.in., że nie powinny zawierać braków danych. W przypadku, gdy w zmiennych wspólnych występują całkowite braki danych[87], jednostki takie należy usunąć ze zbioru i kontynuować integrację na uzyskanym w ten sposób zbiorze. W sytuacji występowania częściowych braków odpowiedzi w cechach wspólnych, do problemu można podejść dwojako: użyć tylko tych zmiennych, które braków nie zawierają lub rozważyć również użycie zmiennych dotkniętych problemem braków odpowiedzi. W drugim przypadku należy zastosować metody imputacji w celu zastąpienia braków danych odpowiednimi wartościami [Scanu 2008].

Trzecia kwestia dotyczy zgodności rozkładów zmiennych wspólnych.. Jest to wynikiem założenia, że integrowane zbiory dotyczą tej samej populacji. W sytuacji, gdy rozkłady zmiennych wspólnych bardzo się różnią, może zachodzić podejrzenie, że próby nie zostały wylosowane z tej samej populacji lub ich momenty referencyjne mocno się różnią. Częstszą sytuacją są różnice w rozkładach zmiennych wspólnych, które wynikają ze zmienności próby.

Różnice w rozkładach odpowiadających sobie zmiennych wspólnych można zbadać m.in. powszechnie wykorzystywanymi testami statystycznymi:

[87] Brak danych dla jednostki dla całego wektora zmiennych wspólnych.

— testem równości frakcji (najczęściej wykorzystywanym dla zmiennych jakościowych mierzonych w skali nominalnej),
— testem zgodności χ^2 (najczęściej wykorzystywanym dla zmiennych jakościowych mierzonych w skali co najmniej porządkowej),
— testem Kołmogorowa - Smirnowa (najczęściej wykorzystywanym dla zmiennych ciągłych).

Testy te są znane i dobrze opisane w literaturze, jednak dla dużych prób[88] wykazują one tendencję do odrzucenia hipotezy o równości rozkładów lub frakcji już przy bardzo niewielkich różnicach. Większość „klasycznych" testów statystycznych zostało skonstruowanych do weryfikacji hipotez dla prób losowanych schematem prostym. Podczas gdy integrowane zbiory pochodzą często z badań o złożonym schemacie losowania, przez co wyniki testów mogą okazać się niemiarodajne.

M. Scanu [2008] zaproponował ocenę zgodności rozkładów cech wspólnych za pomocą tzw. „podejścia empirycznego". Jego istotą jest porównanie rozkładów odpowiednich cech metodami wizualnymi oraz zastosowanie pewnych prostych miar:

— dla zmiennych ciągłych – porównanie histogramów,
— dla zmiennych jakościowych – porównanie różnic frakcji poszczególnych wariantów:
  - dla „dużych" frakcji – akceptowalne są różnice mniejsze niż 5%;
  - dla „małych" frakcji – akceptowalne są różnice mniejsze niż 2%,
— dla zmiennych ilościowych oraz jakościowych – obliczenie tzw. „całkowitego zakresu zmienności" (*total variation distance*[89]):

$$\Delta(w_A, w_B) = \frac{1}{2}\sum_{i=1}^{k}\left|w_{A,i} - w_{B,i}\right| \qquad (4.1)$$

gdzie: $w_{A,i}$ i $w_{B,i}$ to frakcje poszczególnych, i-tych kategorii zmiennych wspólnych w poszczególnych zbiorach. $\Delta \leq 6\%$ oznacza, że rozkłady są „akceptowalnie" zgodne.

— dla zmiennych ilościowych możliwe jest również porównanie parametrów rozkładów zmiennych wspólnych, na przykład: $\frac{\bar{x}_A}{\bar{x}_B}$, $\frac{s_A}{s_B}$, $\frac{\rho_A}{\rho_B}$ itp.

Gołata [2009] zaproponowała by w procesie harmonizacji stosować dodatkowe kryterium oceny zgodności rozkładów za pomocą współczynników podobieństwa:

[88] Integracja danych statystycznych dotyczy zbiorów liczących zazwyczaj co najmniej kilka tysięcy obserwacji.
[89] Miara ta wywodzi się z teorii prawdopodobieństwa, gdzie służy do szacowania odległości między dwiema zmiennymi losowymi [Janson *et al.* 2001].

$$W_{p_1} = \sum_{i=1}^{k} min_{AB}(w_i), \quad (4.2)$$

$$W_{p_2} = \frac{\sum_{i=1}^{k} min_{AB}(w_i)}{\sum_{i=1}^{k} max_{AB}(w_i)}. \quad (4.3)$$

Zwykle w badaniach empirycznych $W_{p1} < W_{p2}$, dlatego też kryterium „akceptowalnej" zgodności rozkładów to $W_{p1} \geq 97\%$ oraz $W_{p2} \geq 95\%$.

Jeżeli zmienne wspólne w integrowanych zbiorach spełniają poszczególne kryteria podobieństwa, mogą być wykorzystane jako zmienne parujące.

Zbiory danych, zwłaszcza pochodzące ze źródeł administracyjnych, zawierają zmienne o charakterze tekstowym. Zastosowanie takich cech w analizach statystycznych niesie za sobą wiele problemów. Wszelkiego rodzaju błędy typograficzne (np. „literówki"), a nawet pisownia wielką lub małą literą sprawiają, że wartości odnoszące się do jednego wariantu mogą zostać zaklasyfikowane (w standardowym oprogramowaniu statystycznym, np. SAS, SPSS, R) jako odmienne warianty (np. „Anna" i „anna", mimo iż oznaczają to samo imię, zostaną potraktowane jako odmienne warianty). W przypadku porównywania wartości poszczególnych zmiennych w procesie integracji, należy wszystkie wpisy w zmiennych tekstowych zharmonizować. Manualna harmonizacja wartości zawartych w zmiennych tekstowych, zwłaszcza w wielkich zbiorach danych, może być czasochłonna i nie eliminuje wszystkich błędów. Zastosowanie automatycznych metod porównujących wartości tekstowe może ten proces przyspieszyć, jednocześnie zachowując jego wyższą niż manualna skuteczność. Wśród metod harmonizacji zmiennych tekstowych wymienia się:

— komparatory łańcuchowe,

— edycję danych.

Komparatory łańcuchowe

Komparator to funkcja porównująca wartości tekstowe (zmienne typu *string*). Jaro [1989] zaproponował komparator służący do korekty takich błędów typograficznych jak wstawienie dodatkowego znaku, usunięcie znaku czy transpozycję (zamianę miejsc) znaków:

$$\Phi_{s_1,s_2} = \frac{1}{3}\left(\frac{\mathrm{N_C}}{\mathrm{len_{s_1}}} + \frac{\mathrm{N_C}}{\mathrm{len_{s_2}}} + \frac{\frac{1}{2}\mathrm{N_t}}{\mathrm{N_C}}\right), \quad (4.4)$$

gdzie:

Φ_{s_1,s_2} - funkcja zgodności dwóch wartości tekstowych s_1 i s_2,

$\mathrm{N_C}$ - liczba wspólnych znaków w s_1 i s_2,

N_t - liczba transpozycji,

len_{s_i} – liczba znaków (długość) i-tej wartości tekstowej.

W kolejnych latach zaproponowano szereg poprawek powyższej funkcji:

- Poprawka McLaughlina [1993] – przyporządkowuje wartość 0,3 do każdego podobnego znaku. Podobne, ale niezgodne znaki mogą się pojawić w wyniku błędów kopiowania, np.: 1 (jeden) a l (mała litera L) lub V a B (znajdujące się blisko siebie na klawiaturze). Każda zgadzająca się para znaków otrzymuje wartość 1. Znaki zgodne są znajdowane w pierwszej kolejności, a następnie znaki podobne. Liczba wspólnych znaków N_C rośnie o 0,3 dla każdego podobnego znaku.
- Poprawka Winklera [1990] – wprowadza dodatkowe wartości w sytuacji zgodności znaków z początku wyrazu. Przeprowadzone badania empiryczne Pollocka i Zamory (1984) wykazały, że najmniej błędów zawierają pierwsze człony wyrazów, oraz że liczba błędów narasta monotonicznie wraz z przesuwaniem się znaków w prawo. Ppoprawka ta koryguje wartość komparatora tekstowego o stałą, jeżeli pierwsze cztery znaki tekstu są zgodne oraz o wartości odpowiednio niższe, jeżeli zgadzają się pierwsze trzy, dwa lub jeden znak.
- Poprawka Lyncha i Winklera [1994] – zwiększa wartość komparatora tekstowego, jeżeli wyraz składa się z więcej aniżeli 6 liter i więcej niż połowa znaków za pierwszymi czterema się zgadza.

Wariacją komparatora tekstowego jest tzw. metoda bigramów. Polega ona na porównywaniu kolejnych dwuliterowych części wyrazu. Np. ze słowa „bigram" porównuje się następujące pary: „bi", „ig", „gr", „ra", „am". Bigram przyjmuje dwie wartości: 0 i 1. Wartością funkcji jest iloraz liczby zgodnych bigramów w odniesieniu do wszystkich badanych.

W tabeli 4.1 zilustrowane zostało porównanie wartości poszczególnych komparatorów tekstowych (oparte o badania Portera i Winklera [2007]). By wartość bigramu uczynić bardziej porównywalną do innych komparatorów, dokonuje się następującej korekty: jeżeli *x* jest wartością funkcji bigramu, używa się przekształcenia $f(x)=x^{0,2435}$ jeżeli *x* jest większe od 0,8 lub 0 w przeciwnym wypadku [Porter, Winkler 2007]. Jeżeli któryś z porównywanych wyrazów zawiera mniej niż 4 znaki, komparatory Jaro i Winklera przyjmują wartość 0. W każdym przypadku ustalany jest arbitralny próg, powyżej którego dane ciągi znaków uznawane są za takie same, zaś poniżej – za różne.

Tabela 4.1. Porównanie komparatorów tekstu

Ciągi znaków		Jaro	Winkler	McLaughlin	Lynch	Bigram
SHACKLEFORD	SHACKELFORD	0,970	0,982	0,982	0,989	0,925
DUNNINGHAM	CUNNIGHAM	0,896	0,896	0,896	0,931	0,917
NICHLESON	NICHULSON	0,926	0,956	0,969	0,977	0,906
JONES	JOHNSON	0,790	0,832	0,860	0,874	0
MASSEY	MASSIE	0,889	0,933	0,953	0,953	0,845
ABROMS	ABRAMS	0,889	0,922	0,946	0,952	0,906
HARDIN	MARTINEZ	0	0	0	0	0
ITMAN	SMITH	0	0	0	0	0
JERALDINE	GERALDINE	0,926	0,926	0,948	0,966	0,972
MARHTA	MARTHA	0,944	0,961	0,961	0,971	0,845
MICHELLE	MICHAEL	0,869	0,921	0,938	0,944	0,845
JULIES	JULIUS	0,889	0,933	0,953	0,953	0,906
TANYA	TONYA	0,867	0,880	0,916	0,933	0,883
DWAYNE	DUANE	0,822	0,840	0,873	0,896	0
SEAN	SUSAN	0,783	0,805	0,845	0,845	0,800
JON	JOHN	0,917	0,933	0,933	0,933	0,847
JON	JAN	0	0	0,860	0,860	0
BROOKHAVEN	BRROKHAVEN	0,933	0,947	0,947	0,964	0,975
BROOK HALLOW	BROOK HLLW	0,944	0,967	0,967	0,977	0,906
DECATUR	DECATIR	0,905	0,943	0,960	0,965	0,921
FITZRUREITER	FITZENREITER	0,856	0,913	0,923	0,945	0,932
HIGBEE	HIGHEE	0,889	0,922	0,922	0,932	0,906
HIGBEE	HIGVEE	0,889	0,922	0,946	0,952	0,906
LACURA	LOCURA	0,889	0,900	0,930	0,947	0,845
IOWA	IONA	0,833	0,867	0,867	0,867	0,906
1ST	IST	0	0	0,844	0,844	0,947

Źródło: [Porter, Winkler 2007]

Porter i Winkler [2007] wskazują, że najlepsze rezultaty otrzymuje się zwykle za pomocą komparatora łańcuchowego z poprawką Lyncha i metody bigramów.

Edycja danych

Edycja danych jest procesem wykrywania i poprawy błędnych danych lub takich, co do których istnieje podejrzenie, że zawierają błędy. Do błędów, które można usunąć w procesie edycji należą m.in.: dane typu tekstowego w zmiennych numerycznych, wartości wykraczające poza dopuszczalny przedział (np. „155 lat" zamiast „15 lat", daty urodzenia z przyszłości). Najczęściej zamienia się takie wartości na systemowe braki danych.

Standaryzacja nazw i adresów polega na zastępowaniu różnie zapisanych słów o tym samym znaczeniu jednakowymi, np.: „ul." na „ulica". Za pomocą odpowiedniego oprogramowania komputerowego można wyszukiwać podobnie brzmiące nazwy w rekordach zmiennych

składających się na klucz połączeniowy, oddzielić fragmenty tekstu takie jak całe nazwy lub adresy w oddzielne słowa używając dowolnego znaku (np. spacji) jako separatora (delimitera). Każde słowo poddane takiej obróbce jest następnie porównywane ze słownikiem (tabelą zawierającą zestandaryzowane nazwy), by nadać mu odpowiednią pisownię. Po zakończeniu procesu standaryzacji, tekst nazwy jest parsowany (poddany działaniu analizatora składniowego) na porównywalne komponenty [Winkler 2005]. Schemat 4.1 przedstawia przykładowe działanie parsera (analizatora składniowego) następujących zestandaryzowanych nazw:

1.DR John J Smith MD

2. Smith DRY FRM

3. Smith & Son ENTP

Schemat 4.1. Przykład nazw poddanych procesowi parsowania

	PRE	FIRST	MID	LAST	POST1	POST2	BUS1	BUS2
1	DR	John	J	Smith	MD			
2				Smith			DRY	FRM
3				Smith			Son	Entp

Źródło: Winkler [2005]

Zastosowanie analizatora składniowego zwiększa efektywność łączenia rekordów, a zestandaryzowane nazwy mogą być od siebie odróżnione nawet jeżeli taka sama lub podobna nazwa odnosi się do różnych obiektów. Zwiększa to prawdopodobieństwo prawidłowego połączenia rekordów odnoszących się do tej samej jednostki.

Gdy wartości rekordów zostaną zestandaryzowane, można przystąpić do wykrywania i usuwania duplikatów. Występują one stosunkowo często w administracyjnych repozytoriach danych. Zwykle tworzone są przez przypadek, w procesie wypełniania formularzy (więcej niż jednego) lub poprzez wielokrotne wprowadzanie danych do rejestru (jeden formularz dwa lub więcej razy). Do wykrywania duplikatów używa się procedury zwanej deduplikacją. Procedura ta może przyjąć postać podobną do integracji dwóch plików, przy czym w tym przypadku łączy się plik z samym sobą szukając rekordów odnoszących się do tej samej jednostki. Wykrywanie zduplikowanych rekordów może również odbywać się metodą „ręczną" (np. filtrowanie zbioru w celu wykrycia tych samych wartości w rekordach). Jednak przy dużej ilości danych, zwłaszcza w zbiorach bez zmiennych o unikalnych wartościach metoda manualna może zająć dużo czasu, a tzw. „czynnik

ludzki" może doprowadzić do powstania kolejnych błędów (np. usunięcia niepowtarzającej się obserwacji, przeoczenia zdublowanych rekordów). Bhattacharya i Getoor [2004] zaproponowali metodę iteracyjnej deduplikacji zbioru danych będącą metodą probabilistyczną. Polega ona na obliczaniu funkcji odległości między poszczególnymi obserwacjami i traktowaniu za zdublowane te jednostki, między którymi odległość jest „mała".
Innym podejściem jest zignorowanie duplikatów, jednak pożądane jest wskazanie ich liczby. Umożliwi to oszacowanie wpływu duplikatów na zintegrowany zbiór.

Blokowanie i inne aspekty operacyjne

Integracja danych wymaga, by każdy rekord z jednej bazy został porównany (pod względem wartości zmiennych parujących) z każdym rekordem z drugiej. Jeżeli jeden z tych zbiorów (lub oba) zawierają informację o bardzo dużej liczbie jednostek, liczba koniecznych porównań znacznie wzrasta, a co się z tym wiąże – wzrasta czas potrzebnaydo wykonania algorytmu. Dodatkowo, tylko niewielka część rekordów zostanie połączona. Przykładowo, łączenie dwóch zbiorów, z których każdy zawiera 1000 rekordów oznacza, że należy sprawdzić aż milion możliwych połączeń (1000×1000), przy czym połączonych ze sobą może zostać maksymalnie 1000, zaś 999000 będzie niepołączonych.
W celu zredukowania liczby sprawdzanych możliwych połączeń, wybierana jest zmienna (lub zmienne) wspólne, której warianty dzielą zbiory wejściowe na podzbiory. Zmienna taka nazywa się zmienną blokującą (warstwującą, grupującą). Przykładowo dla cechy „płeć" i wyboru wariantu „mężczyzna", sprawdzamy połączenia jedynie tych rekordów, którym odpowiada właśnie ta wartość cechy. Podejście takie zapewnia ograniczenie liczby sprawdzeń o około połowę. Analogicznie sytuacja wygląda dla innych zmiennych, a nawet całych zestawów zmiennych blokujących. Na przykład zastosowanie zmiennych blokujących „płeć" i „miesiąc urodzenia" zmniejsza liczbę połączeń już do około 1/24 pierwotnej liczby [por. *Data Integration Manual* 2006].Niezmiernie ważna jest decyzja, czy do integracji użyć danych niezharmonizowanych (surowych), czy zharmonizowanych. W sytuacji, gdy użyte zostaną dane zharmonizowane, zmienne nie są cechami faktycznie obserwowanymi, lecz w pewnym stopniu syntetycznie zrekonstruowanymi. Dane surowe z kolei mogą być obciążone błędami logicznymi. Mogą wystąpić również inne błędy, których nie da się w prosty sposób „naprawić". Takie mogą prowadzić do różnic w rozkładach zmiennych wspólnych.
Nie ma określonych wytycz6nych, których danych należy użyć w integracji. Zdarza się, że zbiory przeznaczone do łączenia zostają przekazane organom statystyki publicznej już po

przeprowadzonym procesie harmonizacji. W takiej sytuacji należy dokładnie przeanalizować poszczególne etapy edycji zbiorów i jeżeli na przykład imputowane wartości stanowią znaczny odsetek, nie należy przeprowadzać integracji [Di Zio 2007].

4.3. Probabilistyczne łączenie rekordów

Procedura *record linkage*[90] odnosi się do wyszukiwania w dwóch zbiorach rekordów zawierających informacje o tej samej jednostce. Procedurę tę często określa się również jako metodę wyszukiwania informacji (*information retrieval*) w celu tworzenia, deduplikacji[91] oraz zarządzania listą nazwisk i adresów [Winkler 2005]. Ze względu na charakter łączenia oraz zastosowanie zmiennych zawierających klucz połączeniowy, procedurę *record linkage* dzieli się na deterministyczną (*deterministic record linkage, exact matching*) oraz probabilistyczną (*probabilistic record linkage*). W łączeniu deterministycznym poszukuje się dokładnej zgodności (lub niezgodności) wartości zmiennych będących tzw. zmiennymi kluczowymi. W najprostszym, i najczęściej spotykanym przypadku, zwykle jest to jedna zmienna zawierająca unikatową wartość dla każdej jednostki. Każdy rekord w integrowanych zbiorach musi zawierać odpowiednią, pozbawioną błędów (np. typograficznych) wartość dla tej zmiennej. Klucz połączeniowy może stanowić również zestaw zmiennych wybranych w taki sposób, że poszczególne warianty cech kluczowych łącznie identyfikują konkretną jednostkę w zbiorach.

W przypadku, gdy zmienne kluczowe o unikalnych wartościach nie są dostępne lub zawierają wspomniane wyżej błędy, niemożliwe jest wykorzystanie metod deterministycznych. W takich przypadkach możliwe jest zastosowanie metody probabilistycznego łączenia rekordów (*probabilistic record linkage*). Idea tej metody sprowadza się do wyboru kilku zmiennych (nazywanych zmiennymi parującymi), które zawarte są w obu zbiorach i oszacowania na ich podstawie prawdopodobieństwa, że poszczególne rekordy należą do tej samej jednostki.

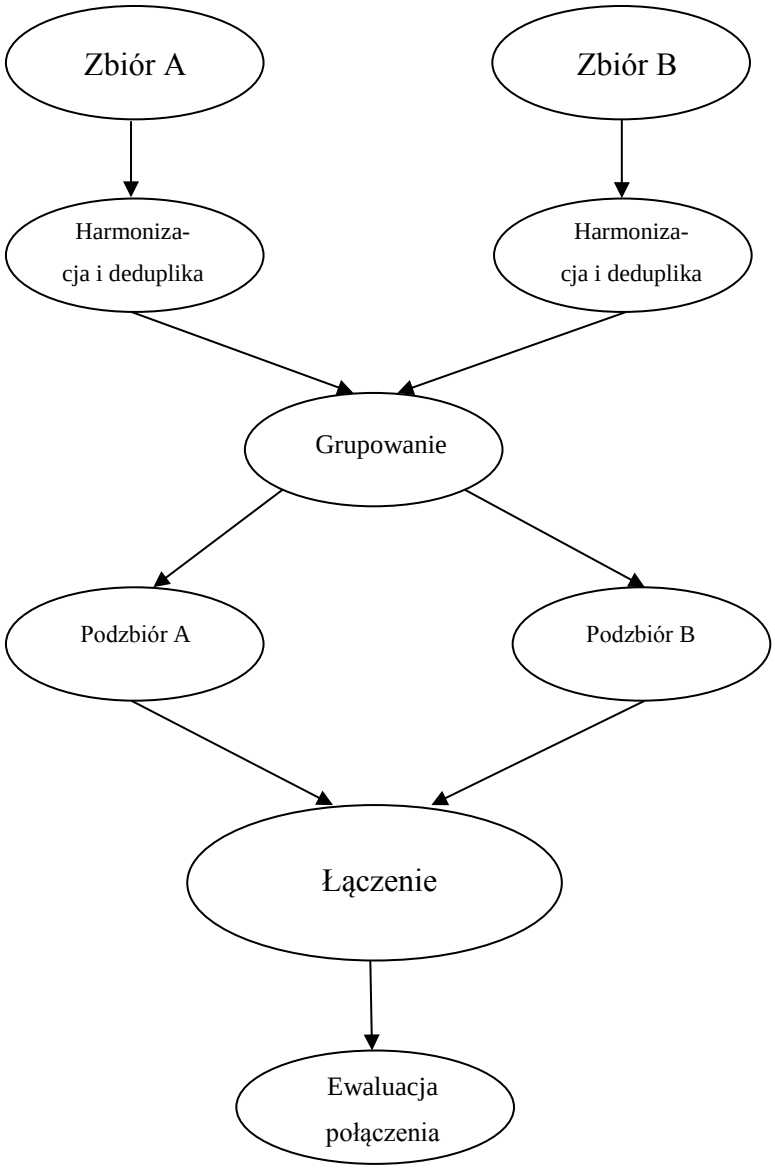
W integrowanych repozytoriach często te same wartości zapisywane są w niejednolity sposób (np. adresy, numery telefonów, nazwy własne itp.). Może to wynikać zarówno z przyjętych przez gestorów różnych standardów zapisu lub różnorodnego rodzaju błędów zapisu (np. ortograficznych, typograficznych, wynikających z niedoskonałości sprzętu i oprogramowania skanującego itp.). Za pomocą metodyki probabilistycznego łączenia

[90] Procedura ta ma szerszy kontekst niż probabilistyczne łączenie rekordów. Obejmuje ona wszystkie metody, deterministyczne i probabilistyczne.
[91] Wyszukiwanie rekordów odnoszących się do tej samej jednostki w tej samej bazie, tzw. duplikatów.

rekordów można również zintegrować takie jednostki porównując pewne wartości w zmiennych występujących w obu zbiorów, które choć różnią się sposobem zapisu należą do tej samej jednostki. Dokonuje się tego za pomocą komparatorów tekstowych i parserów.

Najczęściej w literaturze przedstawia się probabilistyczne łączenie rekordów jako proces kilkustopniowy. Pierwszym krokiem jest zebranie informacji o danych źródłowych oraz wybór zmiennych, na podstawie których przeprowadzone zostanie łączenie (zmienne parujące). W kolejnym kroku przygotowuje się zbiory do procesu integracji poprzez usunięcie duplikatów oraz standaryzację wariantów cech parujących. Następnie przeprowadza się operację grupowania (nazywaną również blokowaniem) mającą na celu podział integrowanych repozytoriów na podzbiory, w których znajdą się jednostki w jakiś sposób do siebie podobne (są to np. mieszkańcy jednego powiatu lub przedstawiciele jednej gałęzi przemysłu). Grupowania dokonuje się z w celu optymalizacji algorytmu poprzez zredukowanie liczby połączeń. Następnie na podstawie odpowiednich algorytmów łączy się bazy oraz sprawdza efektywność (por. schemat 4.2).

Schemat 4.2. Algorytm integracji danych metodą probabilistycznego łączenia rekordów



Źródło: na podstawie [*Data Integration Manual* 2006]

4.3.1. Proces łączenia

Głównym zadaniem metody probabilistycznego łączenia rekordów jest ustalenie, czy para rekordów należy do tego samej jednostki czy nie. Decyzję tę podejmuje się najczęściej na podstawie prawdopodobieństwa (lub jego przekształceń), że dana para rekordów należy (lub nie) do tej samej jednostki [Blakely, Salmond 2002; Fellegi, Sunter 1969].

Niech *m* oznacza empiryczne prawdopodobieństwo zgodności wartości zmiennych parujących przy założeniu, że porównywana para jest dokładnym połączeniem (rekordy należą do tej samej jednostki). Natomiast niech u oznacza empiryczne prawdopodobieństwo niezgodności wartości zmiennych parujących przy założeniu, że porównywana para jest niepołącze-

niem (rekordy nie należą do tej samej jednostki). Wskaźniki zdefiniowane we wzorach (4.5) i (4.6) wykorzystują te wartości w celu ustalenia przynależności porównywanych rekordów do tej samej jednostki. Prawdopodobieństwa *m* i u służą do obliczenia wag zgodności i niezgodności. Wagi zgodności (w_z) i niezgodności (w_n) wyrażają się wzorami [Blakely, Salmond 2002]:

$$w_z = \frac{ln\left(\frac{m}{u}\right)}{ln2}, \tag{4.5}$$

$$w_n = \frac{ln\left(\frac{1-m}{1-u}\right)}{ln2}. \tag{4.6}$$

Przykładem może być zmienna „miesiąc urodzenia". Prawdopodobieństwo, że rekordy posiadające tę samą jej wartość nie należą do tej samej jednostki (niepołączenie) wynosi około $^1/_{12} = 0{,}083$. Wartość ta będzie zatem prawdopodobieństwem u (por. tabela 4.2). Ponieważ prawie we wszystkich zmiennych występują błędy, prawdopodobieństwo *m* (a więc w przypadku zgodności wartości zmiennych parujących) nigdy nie osiąga jedności. Jego wartość wyznaczana jest podczas ustalania strategii łączenia na podstawie informacji z poprzednich badań (w których zastosowano metodologię *record linkage*). W przykładzie z miesiącem urodzenia jako zmienną parującą, przyjmijmy założenie, że prawdopodobieństwo *m* wynosi 0,95 [Blakely, Salmond 2002].

Porównywanej parze, dla której występuje zgodność pod względem miesiąca urodzenia, przyporządkowana zostanie waga zgodności równa 3,51. Natomiast parze, która nie zgadzała się co do wartości dla tej zmiennej przyporządkowana zostanie wartość -4,20 (waga niezgodności, por. tabela 4.2). Algorytm ten powtarzany jest dla wszystkich zmiennych parujących w obrębie pary rekordów, a suma wag połączeniowych nazywana jest wagą łączną. Waga łączna dla danej porównywanej pary jest sumą wszystkich wag zgodności i niezgodności dla zmiennych parujących. Będzie ona dużą liczbą dodatnią, jeżeli wszystkie lub większość zmiennych parujących zgadza się co do wartości i dużą liczbą ujemną, jeżeli wszystkie lub większość zmiennych parujących się nie zgadza. Za prawdopodobne połączenie uznane zostaną te pary rekordów, dla których wartość wagi połączeniowej jest największa.

Tabela 4.2. Przykład obliczania wag zgodności i niezgodności

Wynik porównania	Prawdopodobne połączenie	Prawdopodobne niepołączenie	Waga
Zgodność	0,95 (m)	0,083 (u)	$w_z = \frac{ln\left(\frac{m}{u}\right)}{ln2} = 3{,}51$
Niezgodność	0,05 (1-m)	0,917 (1-u)	$w_n = \frac{ln\left(\frac{1-m}{1-u}\right)}{ln2} = -4{,}20$

Źródło: na podstawie [Blakely, Salmond 2002]

Celem metody probabilistycznego łączenia rekordów jest znalezienie dokładnych połączeń[92]. W rzeczywistości jednak nie jest możliwym dokładne wskazanie, które pary rekordów są dokładnymi połączeniami, a które z całą pewnością są niepołączeniami[93]. Zamiast tego możliwa jest obserwacja par zaklasyfikowanych jako dokładne połączenie i niepołączenie za pomocą wag łącznych. Zadaniem jest wyznaczenie progu wagowego (dla wagi łącznej), powyżej którego pary uznawane są jako prawdopodobne połączenie[94], zaś poniżej jako prawdopodobne niepołączenie[95]. W najlepszym przypadku, znakomita większość prawdopodobnych połączeń jest dokładnym połączeniem i, analogicznie, większość prawdopodobnych niepołączeń jest niepołączeniami [Bakely, Salmond 2002].

Fellegi i Sunter [1969] zaproponowali metodę weryfikacji prawidłowości połączenia rekordów w dwóch zbiorach danych za pomocą funkcji podobieństwa dwóch połączonych jednostek statystycznych. Model stosuje się w przypadku, gdy oba zbiory są już połączone. Ideą metody jest zaklasyfikowanie par w przestrzeni $\boldsymbol{A} \times \boldsymbol{B}$ utworzonej z połączonych zbiorów A i B w zbiór M dokładnych połączeń oraz zbiór U niepołączeń. Klasyfikacja odbywa się za pomocą wzoru będącego ilorazem prawdopodobieństw:

$$R = \frac{P(\gamma \in \Gamma|M)}{P(\gamma \in \Gamma|U)}, \tag{4.7}$$

gdzie:

γ - arbitralny wzór zgodności (np. waga zgodności lub niezgodności) w przestrzeni porównawczej Γ.

Wskaźnik R lub jego inna monotonicznie rosnąca funkcja (np. logarytm naturalny) jest wagą łączną.

[92] Dokładne połączenie - porównywana para rekordów, która w rzeczywistości należy samej jednostki.
[93] Niepołączenie – porównywana para rekordów, która w rzeczywistości nie należy samej jednostki.
[94] Prawdopodobne połączenie – porównywana para rekordów, co do której istnieje wysokie prawdopodobieństwo, że należy do tej samej jednostki.
[95] Prawdopodobne niepołączenie – porównywana para rekordów, co do której istnieje wysokie prawdopodobieństwo, że nie należy do tej samej jednostki.

Wartość R jest następnie porównywalna z wartościami progowymi T_μ i T_λ, które są określone granicznymi błędami a *priori* na, odpowiednio, fałszywe połączenie i fałszywe niepołączenie. Jeżeli spełniony będzie warunek:

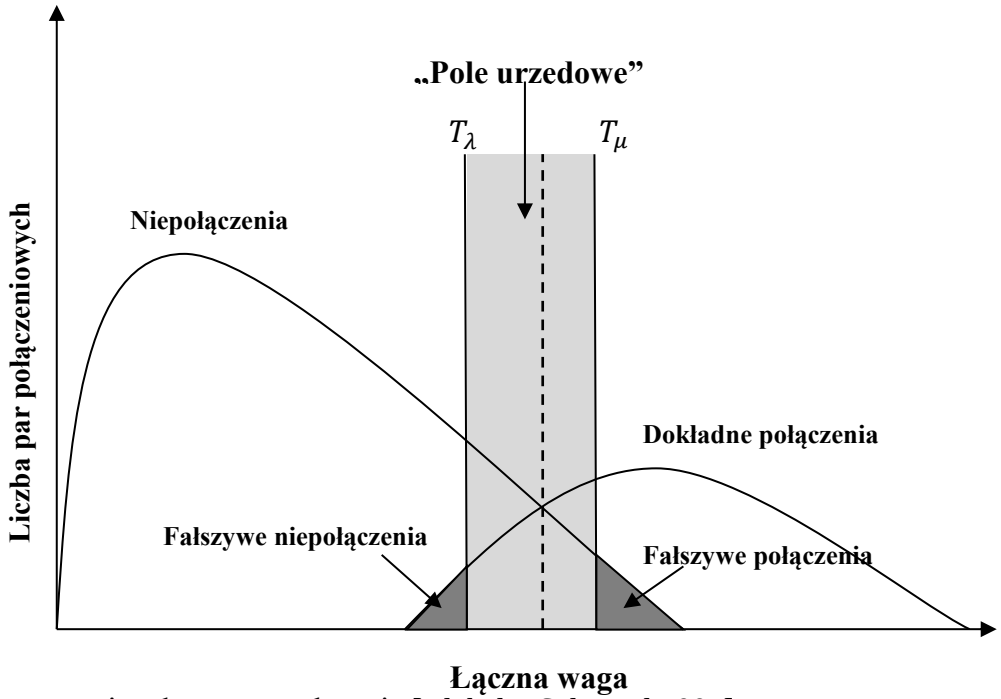
— $R > T_\mu$ – to para jest uważana za dokładne połączenie,

— $T_\lambda \leq R \leq T_\mu$ – to połączenie jest możliwe; przedział ten nazywany jest „polem niedecyzyjnym" lub „polem urzędowym"[96],

— $R < T_\lambda$ – to para uznawana jest za niepołączenie.

Procedura opracowana przez Fellegi i Suntera jest zgodna z intuicją. Jeżeli $\gamma \in \Gamma$ składa się głównie z połączeń, to stosunek R będzie duży. Analogicznie, jeżeli $\gamma \in \Gamma$ składa się głównie z niepołączeń, stosunek R będzie mały [Winkler 2005]. Wadą tej metody jest stosunkowo częste klasyfikowanie połączeń do „pola urzędowego" (por. schemat 4.3).

Schemat 4.3. Liczba par połączeniowych dla dokładnych połączeń i niepołączeń w odniesieniu do wartości wagi łącznej



Źródło: opracowanie własne na podstawie [Blakely, Salmond 2002]

Par zaklasyfikowanych jako dokładne niepołączenie jest zwykle znacznie więcej niż tych zaklasyfikowanych jako dokładne połączenie. Dodatkowo, wśród par klasyfikowanych jako dokładne połączenie (lub niepołączenie) występują fałszywe niepołączenia (lub fałszywe połączenia), czyli rekordy źle zaklasyfikowane. Na etapie weryfikacji jakości integracji licz-

[96] Ponieważ bardzo często zgodność takich par można sprawdzić w źródłach urzędowych.

ba takich przypadków jest szacowana (ponieważ nie można empirycznie zweryfikować, które pary zostały źle zaklasyfikowane) i obliczany jest ich wpływ na ogólny rozkład dołączanych cech. Zaznaczony na wykresie przedział (T_μ, T_λ) przedstawia „pole urzędowe". Oznacza to, że poprawność zaklasyfikowania rekordów, dla których wartość wagi połączeniowej znajdzie się w tym przedziale, należy zweryfikować w źródłach urzędowych. Natomiast przerywaną linią zaznaczono pojedynczą wartość progową [Blakely, Salmond 2002].

Zaletą zastosowania pojedynczej wartości jest brak konieczności manualnej („urzędowej") weryfikacji połączenia, co pozwala na zaoszczędzenie czasu. Wadą natomiast jest większa liczba par rekordów zaklasyfikowana jako fałszywe połączenie lub fałszywe niepołączenie. W praktyce częściej stosuje się podejście Fellegi i Suntera (z „polem urzędowym") [Bernier 1997].

4.3.2. Ocena jakości połączenia

Procedura probabilistycznego łączenia rekordów narażona jest na błędy analogiczne jak w przypadku wnioskowania statystycznego: zaklasyfikowanie jako niepołączenie pary rekordów w rzeczywistości odnoszące się do tej samej jednostki (błąd I rodzaju) oraz, przeciwnie, zaklasyfikowanie jako połączenie pary rekordów nie odnoszącej się do tej samej jednostki (błąd II rodzaju). W ujęciu statystycznym, dokładność połączenia jest rozpatrywana w ramach ilorazów fałszywych połączeń oraz fałszywych niepołączeń. Rozpatrywane będą więc takie miary jak dodatnia wartość predykcyjna i czułość, będące algebraicznymi przekształceniami ilorazów fałszywych połączeń i fałszywych niepołączeń [Cibella, Scanu, Tuoto 2008].

By zdefiniować wymienione wyżej wskaźniki, należy założyć, że znane są następujące wartości:

- liczba par rekordów połączonych prawidłowo (prawdziwie pozytywne) – n_m;
- liczba par rekordów połączonych nieprawidłowo (fałszywie pozytywne, błąd I rodzaju, α) - n_{fp};
- liczba par rekordów niepołączonych prawidłowo (prawdziwie negatywne) - n_u;
- liczba par rekordów niepołączonych nieprawidłowo (fałszywie negatywne, błąd II rodzaju, β) - n_{fn};
- ogólna liczba par rekordów odnoszących się do tych samych jednostek - N_m;
- ogólna liczba par rekordów nie odnoszących się do tej samej jednostki - N_u.

Iloraz fałszywych połączeń definiowany (*false match rate, fmr*) jest jako:

$$fmr = \frac{n_{fp}}{(n_m + n_{fp})}, \tag{4.8}$$

czyli np. iloraz błędnie połączonych par rekordów i ogólnej liczby wszystkich połączonych par. Iloraz fałszywych połączeń odpowiada poziomowi ufności $1 - \alpha$ dla jednostronnego testu statystycznego. Dodatnia wartość predykcyjna (*positive predicted value, ppv*), a więc prawdopodobieństwo, że pary zaklasyfikowane jako dokładne połączenia są nimi w rzeczywistości[97], może zostać oszacowana z ilorazu fałszywych połączeń:

$$ppv = \frac{n_m}{(n_m + n_{fp})} = 1 - fmr, \tag{4.9}$$

i odpowiada ilorazowi liczby prawidłowo połączonych par rekordów i ogólnej liczbie wszystkich połączonych par. Z drugiej strony, iloraz fałszywych niepołączeń (*false non-match rate*, *fnmr*) definiowany jest jako:

$$fnmr = \frac{n_{fn}}{N_m}. \tag{4.10}$$

Iloraz fałszywych niepołączeń odpowiada poziomowi błędu II rodzaju β. Analogicznie do dodatniej wartości predykcyjnej, czułość może zostać zdefiniowana jako:

$$cz = \frac{n_m}{N_m} = 1 - fnmr. \tag{4.11}$$

Można również obliczyć iloraz połączeń (*ip*) zdefiniowany jako:

$$ip = \frac{(n_m + n_{fp})}{N_m}. \tag{4.12}$$

Inną miarą dokładności połączenia jest swoistość definiowana jako:

$$sw = \frac{n_u}{N_u}. \tag{4.13}$$

Różnica między czułością a swoistością polega na tym, że czułość mierzy odsetek prawidłowo zaklasyfikowanych par rekordów, natomiast swoistość – odsetek prawidłowo zaklasyfikowanych niepołączeń.

Zależności te przedstawia tabela 4.3.

[97] Analogicznie, ujemna wartość predykcyjna będzie prawdopodobieństwem, że pary zaklasyfikowane jako dokładne niepołączenie są nimi w rzeczywistości.

Tabela 4.3. Czułość i swoistość oraz dodatnia i ujemna wartość predykcyjna

	Dokładne połączenie	**Niepołączenie**	**Wartość predykcyjna**
Prawdopodobne połączenie	Prawdziwie pozytywne - n_m	Fałszywie pozytywne (błąd I rodzaju) - n_{fp}	**Dodatnia** $\frac{n_m}{(n_m+n_{fp})}$
Prawdopodobne niepołączenie	Fałszywie negatywne (błąd II rodzaju) - n_{fn}	Prawdziwie negatywne - n_u	**Ujemna** $\frac{n_u}{(n_u+n_{fn})}$
Suma	N_m	N_u	
	Czułość $\frac{n_m}{N_m}$	**Swoistość** $\frac{n_u}{N_u}$	

Źródło: opracowanie własne na podstawie na podstawie [Blakely, Salmond 2002]

Charakterystyki te będą się różnić w zależności od wag progowych. Obniżenie wartości progowej spowoduje wzrost czułości, ale zwiększy również liczbę fałszywie pozytywnych prawdopodobnych połączeń. Podwyższenie wartości progowej natomiast zmniejszy czułość, ale zmniejszy też liczbę fałszywie pozytywnych prawdopodobnych połączeń.

Z powyższych powodów potrzebny jest kompromis, w którym należy utrzymać stosunkowo wysoką swoistość (niską liczbę fałszywie pozytywnych prawdopodobnych połączeń) kosztem czułości. W strategii tej ryzyko popełnienia błędu I rodzaju jest stosunkowo niskie, jednak cierpi na tym moc statystyczna.

Wyżej opisane mierniki jakości są używane w empirycznej ocenie jakości połączenia. W ocenie jakości połączenia zastosować można jedną z następujących metod.

— **Próbkowanie i „urzędowa" weryfikacja**

Metoda oceny jakości połączenia poprzez próbkowanie polega na wylosowaniu (lub wyborze celowym) rekordów z obu zbiorów spośród połączeń (M) i niepołączeń (U). Wylosowane pary poddawane są następnie procedurze pogłębionej analizy weryfikacyjnej poprawności połączenia [Hogan, Wolter 1998]. Procedura oceny łączenia wylosowanych rekordów jest dokładniejsza niż ocena całości i wykonywana przez wykwalifikowany personel. Dlatego uznaje się ją za pozbawioną błędów[98].. Obciążenie procedury probabilistycznego łączenia rekordów jest szacowane na podstawie rozbieżności, między wynikami oryginalnego połączenia, a rezultatami empirycznej weryfikacji na podstawie próby.

[98] Czasem wręcz dokonuje się tej procedury manualnie, by mieć pewność, że jest wykonana perfekcyjnie

— **Oszacowania ilorazu fałszywych połączeń**

Belin i Rubin [1995] proponują konstrukcję modelu szacowania ilorazu fałszywych połączeń dla każdej możliwej wartości progowej. W modelu tym rozkład obserwowanych wag połączeniowych jest interpretowany jako złączenie wag dla prawdziwych i fałszywych połączeń. Podejście to opiera się na oszacowaniu prawdopodobieństwa połączenia pary rekordów razem z błędem standardowym jako funkcją wagi połączeniowej.

— **Szacowania częstości błędów**

Torelli i Pagiaro [1999] zaproponowali metodę, która umożliwia oszacowanie częstości występowania błędów na podstawie prawdopodobieństwa, że każda para rekordów jest dokładnym połączeniem. Estymatory konstruowane są metodą największej wiarygodności przy zastosowaniu tzw. „algorytmu maksymalizacji oczekiwań"[99] oraz jego modyfikacji. W metodzie tej oblicza się odsetek fałszywych niepołączeń jako sumy prawdopodobieństw połączenia par rekordów poniżej wartości progowej (np. T_λ). Analogicznie oblicza się odsetek fałszywych połączeń. W metodzie tej nie jest wymagane stosowanie próby testowej.

4.4. Parowanie statystyczne

Parowanie statystyczne to grupa metod służących do integracji dwóch (lub więcej) źródeł danych (zwykle pochodzących z badań próbkowych) odnoszących się do tej samej populacji generalnej. Celem integracji jest jednoczesna obserwacja zmiennych nie obserwowanych łącznie w żadnym ze źródeł oraz wnioskowanie o ich łącznym rozkładzie [Gilula *et al.* 2006, Rodgers 1984, Raessler 2002, D'Orazio et al. 2006, Moriarity 2009 i in.].

Ponieważ prawdopodobieństwo wylosowania tej samej jednostki do dwóch różnych badań reprezentacyjnych jest zbliżone do zera, zakłada się, że integrowane zbiory są rozłączne. W każdym ze zbiorów (oznaczonych jako A i B) znajduje się pewien wspólny wektor zmiennych o tych samych lub zbliżonych definicjach i wariantach (oznaczony jako $\boldsymbol{X}$). Zbiór A zawiera wektor zmiennych obserwowanych wyłącznie w nim, oznaczony jako $\boldsymbol{Y}$, natomiast zbiór B zawiera analogiczny wektor – $\boldsymbol{Z}$ (por. schemat 4.4).

Zmienne $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$ są zmiennymi losowymi o funkcji gęstości $f(x, y, z)$, gdzie $x \in \boldsymbol{X}, y \in \boldsymbol{Y}, z \in \boldsymbol{Z}$. Zakłada się, że $\boldsymbol{X} = (X_1, \dots, X_P)'$, $\boldsymbol{Y} = (Y_1, \dots, Y_Q)$, $\boldsymbol{Z} = (Z_1, \dots, Z_R)$ to wektory zmiennych losowych o wymiarach odpowiednio: P, Q oraz R. Zakłada się ponadto, że A i B

[99] Algorytm maksymalizacji oczekiwań (*Expectation Maximalization Algorithm*) jest metodą konstrukcji estymatorów największej wiarygodności dla zbiorów obserwowanych częściowo [Dempster, Laird, Rubin 1977].

to dwie niezależne próby złożone z n_A i n_B niezależnie wylosowanych obserwacji. Wektor $\boldsymbol{Z}$ nie jest obserwowany w zbiorze A, a wektor $\boldsymbol{Y}$ nie jest obserwowany w zbiorze B. Wektor $(x_a^A, y_a^A) = \left(x_{a1}^A, \dots, x_{aP}^A; y_a^A, \dots, y_{aQ}^A\right)$, gdzie $a = 1, \dots, n_a$ to wektor złożony z obserwowanych wartości zmiennych dla jednostek w zbiorze A. Analogicznie, wektor $(x_b^B, z_b^B) = (x_{b1}^B, \dots, x_{bP}^B; z_b^B, \dots, y_{bR}^B)$, gdzie $b = 1, \dots, n_b$ to wektor złożony z obserwowanych wartości zmiennych dla jednostek w zbiorze B [Di Zio 2007].

Schemat 4.4. Dane wejściowe w parowaniu statystycznym

Zbiór A

$\boldsymbol{Y_1}$	...	$\boldsymbol{Y_Q}$	$\boldsymbol{X_1}$	...	$\boldsymbol{X_P}$
y_{11}^A	...	y_{1Q}^A	x_{11}^A	...	x_{1P}^A
...	...	...	...	...	...
y_{a1}^A	...	y_{aQ}^A	x_{a1}^A	...	x_{aP}^A
...	...	...	...	...	...
$y_{n_A1}^A$	...	$y_{n_AQ}^A$	$x_{n_A1}^A$	...	$x_{n_AP}^A$

Zbiór B

$\boldsymbol{X_1}$	...	$\boldsymbol{X_P}$	$\boldsymbol{Z_1}$	...	$\boldsymbol{Z_R}$
x_{11}^B	...	x_{1P}^B	z_{11}^B	...	z_{1R}^B
...	...	...	...	...	...
x_{b1}^B	...	x_{bP}^B	z_{b1}^B	...	z_{bR}^B
...	...	...	...	...	...
$x_{n_B1}^B$	...	$x_{n_BP}^B$	$z_{n_B1}^B$	...	$z_{n_BR}^B$

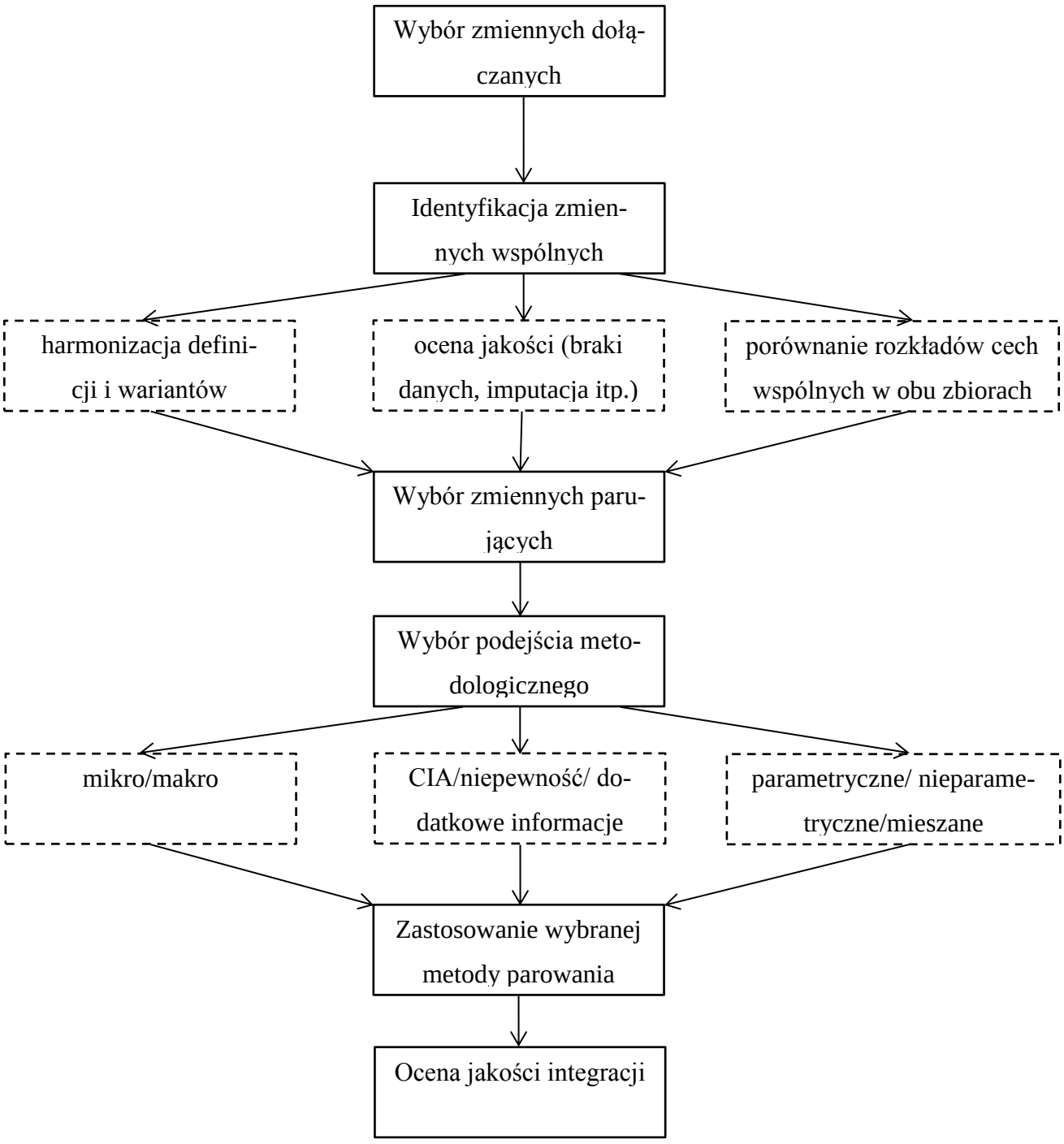
Źródło: opracowanie własne

Zbiory A i B powinny dotyczyć tej samej populacji generalnej, a więc opisują ten sam rodzaj jednostek statystycznych (osoby, gospodarstwa domowe, przedsiębiorstwa itp.), a także charakteryzować się podobnym okresem referencyjnym. W przypadku niespełnienia któregoś z powyższych warunków, zbiory należy zharmonizować (ujednolicić populację i/lub wyrównać okresy referencyjne [Bijak 2009]). W przypadku braku możliwości zharmonizowania zbiorów (np. gdy zbiory dotyczą rozłącznych populacji) nie można przeprowadzać integracji danych.

Algorytm statystycznej integracji danych metodą parowania statystycznego inicjowany jest poprzez wybór tzw. zmiennych dołączanych (por. schemat 4.5). Są to wybrane zmienne z wektora $\boldsymbol{Y}$ ze zbioru A, które mają być dołączone do zbioru B oraz, analogicznie, wybrane zmienne z wektora $\boldsymbol{Z}$ ze zbioru B, które mają zostać dołączone do zbioru A. Zbiór, do którego w danym kroku algorytmu integracji dołączane są zmienne nazywa się zbiorem biorcy (*recipient*), natomiast zbiór, z którego pobierane są informacje o zmiennych dołącza-

nych nosi miano zbioru dawcy (*donor*). Wybór zmiennych dołączanych jest podyktowany zwykle potrzebami informacyjnymi, a w zależności od charakteru tych zmiennych stosowane są określone metody i reguły integracji.

Schemat 4.5. Algorytm parowania statystycznego



Źródło: na podstawie [D'Orazio 2012]

W kolejnym kroku identyfikuje się wektor zmiennych wspólnych $\boldsymbol{X}$. Są to zmienne występujące w obu zbiorach i charakteryzujące się takimi samymi lub podobnymi definicjami. Jeżeli definicje nie są w pełni spójne, należy je zharmonizować.

Dalszym etapem algorytmu jest wybór zmiennych parujących. Integrowane zbiory zwykle współdzielą wiele zmiennych. Przeprowadzane badania społeczno-ekonomiczne, pomimo odrębnych celów, prawie zawsze posiadają wektor zmiennych o podobnych definicjach. Najczęściej są to cechy demograficzno-społeczne (np. płeć, wiek, wykształcenie, miejsce zamieszkania, aktywność ekonomiczna, itp.). Metody wyboru zmiennych parujących zostaną opisane szczegółowo w dalszej części rozdziału.

Wybierając metodę integracji danych, należy rozważyć cel integracji, stawiane założenia, charakter dołączanych zmiennych, dostępność informacji dodatkowych oraz możliwość wykorzystania informacji płynącej ze schematu losowania próbek. W parowaniu statystycznym zasadniczo wyróżnia się dwa główne podejścia metodologiczne [D'Orazio *et al.* 2006]:

— podejście makro – oszacowanie określonych związków (np. korelacji, współczynników regresji, tabeli kontyngencji) między wektorami zmiennych $\boldsymbol{Y}$ i $\boldsymbol{Z}$ bez tworzenia syntetycznego, pełnego zbioru danych (zawierającego łączną obserwację $\boldsymbol{X}$, $\boldsymbol{Y}$ i $\boldsymbol{Z}$).

— podejście mikro – utworzenie syntetycznego, jednostkowego zbioru danych zawierającego łączną obserwację $\boldsymbol{X}$, $\boldsymbol{Y}$ i $\boldsymbol{Z}$.

Ponieważ zmienne $\boldsymbol{Y}$ oraz $\boldsymbol{Z}$ nie są łącznie obserwowane w żadnym ze źródeł, w procesie estymacji związków pomiędzy tymi cechami zwykle przyjmuje się założenie, że zmienne $\boldsymbol{Y}$ i $\boldsymbol{Z}$ są warunkowo niezależne przy danym $\boldsymbol{X}$ [Raessler 2002, D'Orazio et al. 2006, Moriarity 2009]. Nazywa się to założeniem o warunkowej niezależności (*conditional independence assumption,* CIA). Oznacza to, że funkcja gęstości dla ($\boldsymbol{X}$, $\boldsymbol{Y}$, $\boldsymbol{Z}$) posiada następującą własność:

$$f(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}) = f_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x}) f_{\boldsymbol{Z}|\boldsymbol{X}}(\boldsymbol{z}|\boldsymbol{x}) f_{\boldsymbol{X}}(\boldsymbol{x}), \forall\, \boldsymbol{x} \in \mathcal{X}, \boldsymbol{y} \in \mathcal{Y}, \boldsymbol{z} \in \mathcal{Z} \qquad (4.14)$$

gdzie $f_{\boldsymbol{Y}|\boldsymbol{X}}$ to warunkowa funkcja gęstości dla $\boldsymbol{Y}$ przy danym $\boldsymbol{X}$, $f_{\boldsymbol{Z}|\boldsymbol{X}}$ to warunkowa funkcja gęstości dla $\boldsymbol{Z}$ przy danym $\boldsymbol{X}$, a $f_{\boldsymbol{X}}$ to gęstość brzegowa $\boldsymbol{X}$. Przy prawdziwości założenia o warunkowej niezależności, do oszacowania (4.14) wystarczą informacje o brzegowym rozkładzie $\boldsymbol{X}$ *a* także o związkach pomiędzy $\boldsymbol{X}$ i $\boldsymbol{Y}$ oraz $\boldsymbol{X}$ i $\boldsymbol{Z}$. Informacje te dostępne są w zbiorach A i B.

Należy jednak podkreślić, że prawdziwość założenia o warunkowej niezależności nie może zostać przetestowana przy wykorzystaniu informacji z $A \cup B$. Przyjęcie fałszywego założe-

nia może prowadzić do błędnych wniosków wynikających z integracji przy pomocy parowania statystycznego. Wyróżnia się tutaj trzy możliwości postepowania:

— wykorzystanie dodatkowych źródeł informacji (np. wcześniejszych doświadczeń lub za przeprowadzenia pomocniczego badania), które potwierdzą prawdziwość CIA,

— użycie dodatkowych źródeł informacji w toku integracji,

— w przypadku braku dodatkowych informacji o związkach między ($\boldsymbol{X}$, $\boldsymbol{Y}$, $\boldsymbol{Z}$) – rozważenie tzw. niepewności (*uncertainty*) dla właściwości modelu integracji.

W przypadku posiadania dodatkowych informacji, możliwe jest wyznaczenie punktowych estymatorów dla parametrów modelu integracji. Gdy dodatkowe informacje nie są dostępne, w procesie parowania wykorzystywana jest estymacja przedziałowa dla nieznanych charakterystyk, jak np. macierz korelacji ($\boldsymbol{Y}$, $\boldsymbol{Z}$). Im węższe są szacowane przedziały, tym lepszą jakością charakteryzują się zintegrowane zbiory danych. Produktem zastosowania metod parowania statystycznego przy wykorzystaniu estymacji przedziałowej są:

— dla podejścia makro – przedziały wiarygodnych wartości szacowanych parametrów (np. wariancji, kowariancji, współczynników korelacji);

— dla podejścia mikro – rodzina syntetycznych zbiorów danych utworzonych przy wykorzystaniu różnych wiarygodnych parametrów stosowanego modelu integracji.

Syntetyczny zbiór danych w podejściu mikro może być utworzony dwojako:

— poprzez dołączenie, na podstawie podobieństwa, wartości zmiennych wspólnych, rekordów zbioru *dawcy* do zbioru *biorcy (*– por. schemat 4.6); liczebność zintegrowanego zbioru jest równa liczebności zbioru biorcy;

— poprzez konkatenację zbiorów danych - dołączenie Z ze zbioru B do A oraz Y ze zbioru A do B (por. schemat 4.7); wymaga zastosowania określonych metod imputacji; liczebność nowo utworzonego syntetycznego zbioru jest równa sumie liczebności A i B ($S = A \cup B$).

Schemat 4.6. Zintegrowany zbiór danych

	$\boldsymbol{Y_1}$	…	$\boldsymbol{Y_Q}$	$\boldsymbol{X_1}$	…	$\boldsymbol{X_P}$	$\boldsymbol{Z_1}$	…	$\boldsymbol{Z_R}$
Zbiór A+B	y_{11}^A	…	y_{1Q}^A	x_{11}^A	…	x_{1P}^A	$\tilde{z}_{11}^B$	…	$\tilde{z}_{1R}^B$
	…	…	…	…	…	…	…	…	…
	y_{a1}^A	…	y_{aQ}^A	x_{a1}^A	…	x_{aP}^A	$\tilde{z}_{b1}^B$	…	$\tilde{z}_{bR}^B$
	…	…	…	…	…	…	…	…	…
	$y_{n_A1}^A$	…	$y_{n_AQ}^A$	$x_{n_A1}^A$	…	$x_{n_AP}^A$	$\tilde{z}_{n_A1}^B$	…	$\tilde{z}_{n_AR}^B$

Źródło: opracowanie własne

Schemat 4.7. Schemat danych wejściowych w podejściu konkatenacji baz

Zbiory	$\boldsymbol{Y_1}$	...	$\boldsymbol{Y_Q}$	$\boldsymbol{X_1}$	**...**	$\boldsymbol{X_P}$	$\boldsymbol{Z_1}$	**...**	$\boldsymbol{Z_R}$
A	y_{11}^A	...	y_{1Q}^A	x_{11}^A	...	x_{1P}^A	brak danych – do imputacji		
	...	...	...	...	...	...			
	y_{a1}^A	...	y_{aQ}^A	x_{a1}^A	...	x_{aP}^A			
	...	...	...	...	...	...			
	$y_{n_A1}^A$	...	$y_{n_AQ}^A$	$x_{n_A1}^A$	...	$x_{n_AP}^A$			
B	brak danych – do imputacji			x_{11}^B	...	x_{1P}^B	z_{11}^B	...	z_{1R}^B
				...	...	...	...	...	...
				x_{b1}^B	...	x_{bP}^B	z_{b1}^B	...	z_{bR}^B
				...	...	...	...	...	...
				$x_{n_B1}^B$	...	$x_{n_BP}^B$	$z_{n_B1}^B$	...	$z_{n_BR}^B$

Źródło: na podstawie Di Zio [2007]

W pracach m.in. Kadane [1978], Paas [1986], Cohen [1991] oraz Singh *et al.* [1993] wykazano, że integracja przy założeniu o warunkowej niezależności zwykle prowadzi do oszacowań o zadowalającej jakości.

W celu otrzymania punktowego estymatora gęstości $f(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})$ należy odnieść się do zewnętrznych źródeł informacji. Singh *et al.* [1993] określił dwa rodzaje zewnętrznych źródeł informacyjnych:

— trzeci zbiór danych C, w którym ($\boldsymbol{X}$, $\boldsymbol{Y}$, $\boldsymbol{Z}$) lub ($\boldsymbol{Y}$, $\boldsymbol{Z}$) są łącznie obserwowane;

— wiarygodne wartości nieznanych relacji ($\boldsymbol{Y}$, $\boldsymbol{Z}|\boldsymbol{X}$) lub ($\boldsymbol{Y}$, $\boldsymbol{Z}$).

W praktyce wystąpić może jednak wiele problemów. Zbiór C może pochodzić z innej populacji generalnej, być nieaktualny lub niespójny ze zbiorami A i B (w sensie zastosowanych definicji zmiennych, np. w przypadku wykorzystania niezharmonizowanego rejestru administracyjnego). Przeprowadzenie dodatkowego badania w celu uzyskania łącznej informacji dla ($\boldsymbol{X}$, $\boldsymbol{Y}$, $\boldsymbol{Z}$) lub ($\boldsymbol{Y}$, $\boldsymbol{Z}$) rodzi z kolei problemy natury ekonomicznej (koszty i czas przeprowadzenia badania) oraz statystycznej (nowy zbiór danych może charakteryzować się brakami danych oraz błędami losowymi i nielosowymi obciążającymi estymatory).

Podsumowując, parowanie statystyczne, zarówno dla podejścia makro, jak i mikro przeprowadzić można przy:

— założeniu o warunkowej niezależności,

— wykorzystaniu zewnętrznych źródeł informacji,

— przeprowadzeniu analizy niepewności.

Założenie o warunkowej niezależności stosowane jest najczęściej, ze względu na łatwość aplikacji, a także, jak wskazuje praktyka, na dobrą jakość integracji.

Techniki wykorzystywane w procesie integracji za pomocą parowania statystycznego można dodatkowo podzielić na parametryczne, nieparametryczne i mieszane. W metodach parametrycznych zakłada się postać rozkładu zmiennych w integrowanych zbiorach i używa metod wnioskowania statystycznego w procesie integracji. Wykorzystuje się je głównie przy dołączaniu zmiennych ciągłych. W podejściu makro metody parametryczne zwykle opierają się na kalibracji [Sarndal *et al.* 1992], natomiast w podejściu mikro na imputacji brakujących wartości [Rubin 1986, Raessler 2002, D'Orazio *et al.* 2006]. Przy wykorzystaniu metod nieparametrycznych nie wnioskuje się o rozkładzie cech lub postaci modelu, a wykorzystuje głównie metody typu *hot deck* (np. losową, najbliższego sąsiedztwa, rangową – w przypadku podejścia mikro) lub estymator jądrowy (w podejściu makro).

Metody mieszane zwykle przebiegają dwustopniowo. W pierwszym kroku wykorzystywane są metody parametryczne w celu oszacowania imputowanych wartości (np. modele regresji), zaś w drugim wykorzystuje się techniki nieparametryczne do utworzenia syntetycznego zbioru danych z łączną obserwacją ($\boldsymbol{X}$, $\boldsymbol{Y}$, $\boldsymbol{Z}$). Literatura [m.in. D'Orazio et al. 2006, Moriarity 2009, Moriarity i Scheuren 2001 oraz 2003] wskazuje na podejście mieszane jako najlepsze w sensie zapobiegania nierzetelnym szacunkom w przypadku źle dobranego modelu, a także pozwala na dołączanie wartości „żywych" (zaobserwowanych w rzeczywistości) zamiast tych wynikających z modelu (teoretycznych). Modele mieszane wykorzystuje się wyłącznie w podejściu mikro. Rozwinięcie tych podejść zostanie przedstawione w dalszej części rozdziału.

Ponieważ za pomocą parowania statystycznego integruje się głównie zbiory danych pochodzących z badań reprezentacyjnych, ostatnim elementem różnicującym stosowane podejścia jest wykorzystanie (lub nie) schematu doboru jednostek do próby i wynikających z niego prawdopodobieństw inkluzji oraz wag analitycznych. W tej dziedzinie wyróżnia się dwa podstawowe podejścia:

- klasyczne – zakłada, że zbiory A i B są próbami wylosowanymi niezależne i o jednakowym rozkładzie (*independent and indetically distributed, i.i.d.*) z nieskończonej populacji – w procesie wnioskowania nie wykorzystuje się informacji wynikających ze schematu doboru próby;
- oparte na schemacie losowania - zbiory A i B są próbami wylosowanymi zgodnie ze złożonym schematem losowania z tej samej, skończonej populacji – we wnioskowaniu wykorzystuje się schemat doboru jednostek do próby.

4.4.1. Wybór zmiennych parujących

Duża liczba cech parujących nie zawsze pozwala na konstrukcję najlepszego modelu w sensie predykcji dołączanej zmiennej [D'Orazio 2012]. Wybór zmiennych parujących ze zbioru wektora zmiennych wspólnych X może zostać dokonany dwojako:

— w sposób ekspercki;

— za pomocą metod statystycznych.

Jednym z najprostszych podejść jest wyłączenie ze zbioru zmiennych wspólnych tych cech, które nie wyjaśniają zmienności zmiennych dołączanych w sposób istotny, a pozostawienie tych cech, których moc predykcyjna dla zmiennych dołączanych jest dostatecznie wysoka. W praktyce przeprowadza się analizę współzależności między cechami $\boldsymbol{X}$ a $\boldsymbol{Y}$ w zbiorze A oraz $\boldsymbol{X}$ i $\boldsymbol{Z}$ w zbiorze B. Jako zmienne parujące wyznacza się podzbiór cech $\boldsymbol{X}$ istotnie korelujący zarówno z cechami $\boldsymbol{Y}$, jak i $\boldsymbol{Z}$ [Singh *et al.* 1988; Cohen 1991].

Współzależność między cechami wspólnymi i dołączanymi można rozpatrywać jedno-, jak i wielowymiarowo. Najprostszą metodą jest analiza współzależności między parami zmiennych $\boldsymbol{X}$ i $\boldsymbol{Y}$ lub $\boldsymbol{Z}$. W zależności od poziomu pomiaru badanych par zmiennych, zastosować należy różne miary współzależności [Agresti 1990].

W przypadku, gdy każda z analizowanych zmiennych mierzona jest na poziomie nominalnym, najczęściej stosowaną miarą współzależności jest test niezależności χ^2. Statystyka testowa ma postać $\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(n_{ij}-\hat{n}_{ij})^2}{\hat{n}_{ij}}$, gdzie n_{ij} oraz $\hat{n}_{ij}$ oznaczają, odpowiednio, liczebności empiryczne i teoretyczne w tabeli kontyngencji cech X i Y (Z) [Witkowski 2009]. Siłę zależności mierzy się najczęściej za pomocą współczynnika kontyngencji C-Pearsona wyrażonego wzorem $C = \sqrt{\frac{\chi^2}{\chi^2+n}}$. Współczynnik ten jest unormowany w przedziale $\langle 0,1\rangle$, gdzie 0 oznacza brak zależności, zaś 1 zależność doskonałą [Paradysz 2004].

W przypadku, gdy obie analizowane cechy mierzone są na poziomie porządkowym, zastosować można miary współzależności przeznaczone dla tego rodzaju cech. Do najczęściej używanych zaliczyć można współczynnik γ Goodmana i Kruskala, d-Somersa (asymetryczne), d-Somersa (symetryczne), τ_b – Kendalla, τ_c – Kendalla oraz współczynnik korelacji rang Spearmana [Górniak, Wachnicki 2010]. Wszystkie te miary oparte są na porównaniu rang dla wartości poszczególnych obserwacji. We wzorach wykorzystuje się różne rodzaje par:

— pary zgodnie (o tej samej wartości rangi dla X i Y) - N_c;

— pary niezgodne (o różnej wartości rangi dla X i Y) - N_d;

— pary wiązane (o tej samej wartości rangi) dla cechy X, ale o różnej wartości dla Y - T_x;

— pary wiązane (o tej samej wartości rangi) dla cechy Y, ale o różnej wartości dla X - T_y;

— pary wiązane ze względu na obie zmienne - T_{xy}.

Miary współzależności dla cech porządkowych wyrażają się następującymi wzorami:

— γ Goodmana i Kruskala: $\gamma = \frac{N_c - N_d}{N_c + N_d}$ – ignoruje rangi powiązane i przyjmuje wartości z przedziału $\langle -1,1 \rangle$. Wartość 1 osiąga, gdy wszystkie przypadki skoncentrowane są na przekątnej tabeli kontyngencji. Wartość 0 nie oznacza niezależności cech.

— d-Somersa :

asymetryczne, dla Y jako zmiennej zależnej: $d_{yx} = \frac{N_c - N_d}{N_c + N_d + T_x}$;symetryczne: $d_S = \frac{N_c - N_d}{N_c + N_d + \frac{1}{2}(T_x + T_y)}$;

Obie miary uwzględniają wiązania, a ich interpretacja jest analogiczna jak w przypadku γ Goodmana i Kruskala.

— τ_b Kendalla: $\tau_b = \frac{N_c - N_d}{\sqrt{(N_c + N_d + T_x)(N_c + N_d + T_y)}}$ – przybiera wartości zbliżone do współczynnika korelacji liniowej Pearsona (zwłaszcza, gdy liczba kategorii każdej z analizowanych zmiennych jest większa od 5); jest unormowany w przedziale $\langle -1,1 \rangle$.

— τ_c Kendalla: $\tau_c = \frac{N_c - N_d}{\frac{1}{2}N^2\left(\frac{m-1}{m}\right)}$, gdzie N oznacza liczbę jednostek, zaś m mniejszą z liczby wierszy lub kolumn w tabeli kontyngencji; jest unormowany w przedziale $\langle -1,1 \rangle$. Jest trudno interpretowalny ze względu na silną zależność wartości od wielkości analizowanej tabeli.

— Współczynnik korelacji rang Spearmana jest wariantem współczynnika korelacji liniowej Pearsona. W swojej najprostszej formie jest to współczynnik korelacji Pearsona obliczony dla rang obserwacji. Unormowany jest w przedziale $\langle -1,1 \rangle$, gdzie 0 oznacza brak zależności.

W przypadku, gdy jedna z cech jest mierzona na poziomie porządkowym (np. X), a druga co najmniej interwałowym (np. Y), jako miarę ich współzależności wykorzystuje się najczęściej współczynnik η^2 Pearsona: $\eta^2 = \frac{\sum_{i=1}^{I}(\bar{y}_i - \bar{y})^2 n_i}{s_Y^2 n}$, gdzie n to liczebność próby, n_i to licz-

ba jednostek, dla których $X = i$, $i = 1, \dots, I$, $\bar{y}$ to średnia dla cechy Y, a $\bar{y}_i$ to średnia warunkowa dla cechy Y dla $X = i$. Miarę tę interpretuje się jako część wariancji zmiennej ilościowej wyjaśnionej przez przynależność do danej kategorii zmiennej porządkowej. W przypadku, gdy zmienną zależną jest zmienna jakościowa, wartość współczynnika η^2 pozwala ocenić skuteczność przewidywania przynależności obserwacji do kategorii tej zmiennej na podstawie wartości jakie przyjmuje zmienna ilościowa. Jest unormowana w przedziale $\langle 0,1 \rangle$. Stosowana jest także dla dwóch zmiennych ilościowych, jeżeli charakter związku między nimi nie jest liniowy.

Jeżeli obie analizowane cechy mierzone są na poziomie co najmniej interwałowym, a związek pomiędzy nimi jest liniowy[100], do analizy współzależności zastosować można współczynnik korelacji liniowej Pearsona. Wyrażony jest on wzorem $r = \frac{\frac{1}{n}\sum_{i=1}^{N} x_i y_i - \bar{x}\bar{y}}{S_x S_y}$ i unormowany jest w przedziale $\langle -1,1 \rangle$.

Tabela 4.5. Wybrane miary związku między dwiema zmiennymi

Współczynnik	Poziom pomiaru	Zakres	Uwagi
Chi-kwadrat χ^2	N,N	0 do ∞	Stosowany do testowania hipotezy o niezależności zmiennych
Phi-Yule'a ϕ	N,N	0 do 1	Unormowany tylko dla tabel 2x2
C-Pearsona	N,N	0 do 1	
V-Kramera	N,N	0 do 1	
γ Goodmana i Kruskala	N,N	0 do 1	0 nie wyklucza zależności
d – Somersa	P,P	-1 do 1	Uwzględnia pary wiązane
τ_b – Kendalla	P,P	-1 do 1	Może osiągnąć -1 lub 1 tylko w tabelach kwadratowych
τ_c – Kendalla	P,P	-1 do 1	
r – Pearsona	I,I	-1 do 1	
ρ – Spearmana	P,P	-1 do 1	
η^2 Pearsona	I,N	0 do 1	

Uwaga:
N – zmienna mierzona na skali nominalnej, P – na skali porządkowej, I – na skali co najmniej interwałowej
Źródło: na podstawie [Górniak, Wachnicki 2010]

Zależność między cechami wspólnymi a dołączanymi zwykle nie jest jednowymiarowa. Wartości zmiennych dołączanych zależą zwykle od łącznego wpływu różnych zmiennych wspólnych. Bardziej złożone metody wyboru zmiennych parujących wymagają więc zastosowania metod wielowymiarowej analizy statystycznej. Służą one do redukcji liczby zmien-

[100] Hipotezę o liniowości związku między analizowanymi zmiennymi zweryfikować można np. za pomocą testu na liniowość funkcji regresji dwóch zmiennych [Witkowski 2009].

nych parujących bez utraty informacji. Do najczęściej wykorzystywanych metod wielowymiarowych w wyborze zmiennych parujących należą regresja krokowa, metody hierarchicznej analizy skupień, drzewa regresyjne i klasyfikacyjne [D'Orazio *et al.* 2006].

W przypadku, gdy zmienne Y i Z są mierzone na poziomie co najmniej interwałowym, do wyboru zmiennych parujących można zastosować metody regresji krokowej ze zmiennymi X jako zmiennymi niezależnymi. Tworzy się w takim przypadku modele regresji $\hat{Y}(X)$ oraz $\hat{Z}(X)$ stosując dowolne metody redukcji liczby zmiennych niezależnych (np. metoda krokowa wprzód i wstecz, metoda Hellwiga itp.).

Z zagadnieniem regresji związane jest zjawisko współliniowości cech. Wynika ono z silnej korelacji między zmiennymi niezależnymi i niesie za sobą wiele niekorzystnych konsekwencji [Gatnar, Walesiak 2009]:

- niemożliwy staje właściwy pomiar siły oddziaływania zmiennych objaśniających na zmienną objaśnianą,
- oceny wariancji estymatorów są zawyżone,
- wartości statystyk testowych t testu istotności parametrów równania regresji dla zmiennych skorelowanych są małe, podczas gdy statystyka F testu istotności całego wektora parametrów regresji wskazuje na „istotność" modelu jako całości,
- oszacowania są wrażliwe na niewielkie zmiany liczby obserwacji.

Skutki występowania współliniowości[101] mogą powodować błędne wyniki analizy regresji krokowej. W takich przypadkach najczęściej stosuje się metody hierarchicznej analizy skupień wykorzystującej miary odległości między zmiennymi [D'Orazio *et al.* 2006].

Jeżeli zmienna dołączana jest cechą jakościową lub istnieje nieliniowy związek między cechą dołączaną, a zmiennymi wspólnymi, dobrym narzędziem wyboru zmiennych parujących są drzewa klasyfikacyjne i regresyjne (*Classification And Regression Trees, CART,* por. Gatnar, Walesiak [2004] oraz [2009]; Rószkiewicz [2002] oraz [2012]). Metody te służą do podziału próby na klasy obserwacji o homogenicznych wartościach zmiennej objaśnianej. Wynik końcowy przedstawiony jest w formie drzewa składającego się z „korzenia" (przedstawianego u góry wykresu) oraz „gałęzi" prowadzących do kolejnych węzłów. Im wyżej na wykresie pojawia się dana zmienna, tym większy ma ona wpływa na zmienną objaśnianą.

[101] Do badania występowania współliniowości najczęściej wykorzystuje się współczynnik VIF (*variance inflation factor,* współczynnik inflacji wariancji) wyrażony wzorem $VIF = \frac{1}{1-R_j^2}$, gdzie R_j^2 to współczynnik determinacji liniowej w modelu, w którym zmienną objaśnianą jest X_j, a zmiennymi objaśniającymi wszystkie pozostałe $m-1$ zmiennych X [Gruszczyński *et al.* 2012].

Dodatkową zaletą metod CART jest ich nieparametryczność – nie istnieją żadne założenia co do rozkładu analizowanych cech.

Wielość sytuacji, z którymi spotykają się badacze, różne rodzaje rozważanych zmiennych oraz celów dokonywania integracji danych powoduje, że wybór odpowiedniej metody powinien zostać przeprowadzany bardzo starannie, z uwzględnieniem ich zalet i ograniczeń. W dalszej części rozdziału opisane zostaną szczegółowo poszczególne podejścia metodologiczne.

4.4.2. Podejście makro

W podejściu makro wnioskuje o łącznym rozkładzie cech $\boldsymbol{Y}$ i $\boldsymbol{Z}$ nie konstruując syntetycznego zbioru danych jednostkowych. Przy założeniu, że zmienne dołączane mają charakter ciągły (oraz rozkład normalny) szacowanymi parametrami są wektor wartości oczekiwanych $\boldsymbol{\mu}$ oraz macierz wariancji i kowariancji $\boldsymbol{\Sigma}$.

<u>Założenie o warunkowej niezależności</u>

A. Podejście parametryczne

Założenie o warunkowej niezależności stosuje się w przypadku braku jakiejkolwiek dodatkowej informacji o szacowanych parametrach łącznego rozkładu zmiennych $\boldsymbol{Y}$ i $\boldsymbol{Z}$. Załóżmy, że zbiory A i B są próbami wylosowanymi niezależne i o jednakowym rozkładzie (*i.i.d.*) liczącymi odpowiednio n_A i n_B obserwacji. Rozkład łączny zmiennych ($\boldsymbol{X}$, $\boldsymbol{Y}$, $\boldsymbol{Z}$) jest rozkładem normalnym o parametrach:

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_X \\ \mu_Y \\ \mu_Z \end{pmatrix} \text{ oraz } \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_X^2 & \sigma_{XY} & \sigma_{XZ} \\ \sigma_{XY} & \sigma_Y^2 & \sigma_{YZ} \\ \sigma_{XZ} & \sigma_{YZ} & \sigma_Z^2 \end{pmatrix}.$$

Dla zmiennych ciągłych estymowanym parametrami są:

$$\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \left[\begin{pmatrix} \mu_X \\ \mu_Y \\ \mu_Z \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & \sigma_{XY} & \sigma_{XZ} \\ \sigma_{XY} & \sigma_Y^2 & \sigma_{YZ} \\ \sigma_{XZ} & \sigma_{YZ} & \sigma_Z^2 \end{pmatrix} \right] \tag{4.15}$$

W przypadku zmiennych jakościowych, zbiór A zawiera zmienne $X = [1, \dots, I]$ oraz $Y = [1, \dots, J]$. Zbiór B zawiera zmienne $X = [1, \dots, I]$ oraz $Z = [1, \dots, K]$, gdzie i, j, k to warianty zmiennych, odpowiednio, X, Y i Z. Szacowanym parametrem jest wówczas:

$$\theta_{ijk} = P(X = i, Y = j, Z = k), \tag{4.16}$$

gdzie $0 \leq \theta_{ijk} \leq 1$ oraz $\sum_{i,j,k} \theta_{ijk} = 1$.

Uogólniając powyższe rozważania, łączny rozkład ($\boldsymbol{X}$, $\boldsymbol{Y}$, $\boldsymbol{Z}$) można zapisać wzorem:

$$f(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z};\boldsymbol{\theta}) = f_{\boldsymbol{X}}(\boldsymbol{x};\boldsymbol{\theta}_{\boldsymbol{x}})f_{\boldsymbol{Y}|\boldsymbol{X}}\left(\boldsymbol{y}|\boldsymbol{x};\ \boldsymbol{\theta}_{\boldsymbol{Y}|\boldsymbol{X}}\right)f_{\boldsymbol{Z}|\boldsymbol{X}}\left(\boldsymbol{z}|\boldsymbol{x};\ \boldsymbol{\theta}_{\boldsymbol{Z}|\boldsymbol{X}}\right). \quad (4.17)$$

Parametry μ_X oraz σ_X^2 mogą zostać oszacowane na podstawie informacji zarówno ze zbioru A, zbioru B, jak i zbioru $A \cup B$. Parametry μ_Y oraz σ_{XY} mogą zostać oszacowane na podstawie informacji ze zbioru A, natomiast parametry μ_Z oraz σ_{XZ} na podstawie informacji ze zbioru B. Na podstawie informacji z integrowanych zbiorów A i B nie można oszacować parametru σ_{YZ}. Przy założeniu o warunkowej niezależności (CIA) między zmiennymi $\boldsymbol{Y}$ i $\boldsymbol{Z}$ przy danym $\boldsymbol{X}$ zachodzi równość:

$$\sigma_{YZ} = \frac{\sigma_{XY}\sigma_{XZ}}{\sigma_X^2}. \quad (4.18)$$

Oznacza to również, że współczynnik korelacji:

$$\rho_{YZ} = \rho_{XY}\rho_{XY} \quad (4.19)$$

przy czym $\rho_{YZ|X} = 0$.

Szacunek wartości powyższych parametrów przeprowadzić można wykorzystując wszystkie informacje pochodzące z próby [Kadane 1978, Moriarity i Scheuren 2001]:

— dla parametrów $\boldsymbol{X}$

$$\hat{\mu}_X = \bar{x}_{A\cup B} = \frac{n_A\bar{x}_A + n_B\bar{x}_B}{n_A+n_B}, \quad (4.20)$$

$$\hat{\sigma}_X^2 = S_{X,A\cup B}^2 = \frac{(n_A-1)S_{X,A}^2 + (n_B-1)S_{X,B}^2}{n_A+n_B-1}, \quad (4.21)$$

— dla parametrów $\boldsymbol{Y}$

$$\hat{\mu}_Y = \bar{y}_A = \frac{\sum_{a=1}^{n_A} y_a}{n_A}, \quad (4.22)$$

$$\hat{\sigma}_Y^2 = S_{Y,A}^2 = \frac{\sum_{a=1}^{n_A}(y_a-\bar{y}_A)^2}{n_A-1}, \quad (4.23)$$

— dla parametrów $\boldsymbol{Z}$

$$\hat{\mu}_Z = \bar{z}_B = \frac{\sum_{b=1}^{n_B} z_b}{n_B}, \quad (4.24)$$

$$\hat{\sigma}_Z^2 = S_{Z,B}^2 = \frac{\sum_{b=1}^{n_B}(z_b-\bar{z}_B)^2}{n_B-1}, \quad (4.25)$$

— kowariancja $\boldsymbol{XY}$ (szacowana na podstawie informacji ze zbioru A)

$$\hat{\sigma}_{XY} = S_{XY,A} = \frac{\sum_{a=1}^{n_A}(x_a-\bar{x}_A)(y_a-\bar{y}_A)}{n_A-1}, \quad (4.26)$$

— kowariancja $\boldsymbol{XZ}$ (szacowana na podstawie informacji ze zbioru B)

$$\hat{\sigma}_{XZ} = S_{XZ,B} = \frac{\sum_{b=1}^{n_B}(x_b-\bar{x}_B)(z_b-\bar{z}_B)}{n_B-1}. \quad (4.27)$$

Zatem, przy założeniu o warunkowej niezależności, kowariancję $\boldsymbol{YZ}$ można wyznaczyć z następującej równości:

$$\hat{\sigma}_{YZ} = \frac{S_{XY,A}S_{XZ,B}}{S^2_{X,A\cup B}}. \tag{4.28}$$

W efekcie szacowana macierz wariancji i kowariancji $\boldsymbol{XYZ}$ będzie miała postać:

$$\widehat{\boldsymbol{\Sigma}} = \begin{bmatrix} S^2_{X,A\cup B} & S_{XY,A} & S_{XZ,B} \\ S_{XY,A} & S^2_{Y,A} & \hat{\sigma}_{YZ} \\ S_{XZ,B} & \hat{\sigma}_{YZ} & S^2_{Z,B} \end{bmatrix}. \tag{4.29}$$

W świetle CIA powyższy sposób rozumowania jest dobry, jednak może prowadzić do pewnych problemów przy szacowaniu współczynnika regresji β_{YX}, ponieważ podmacierz

$$\widehat{\boldsymbol{\Sigma}} = \begin{bmatrix} S^2_{Y,A} & \hat{\sigma}_{YZ} \\ \hat{\sigma}_{YZ} & S^2_{Z,B} \end{bmatrix} \tag{4.30}$$

może nie okazać się dodatnio półokreślona.

Anderson [1984] zaproponował procedurę, opartą o estymator największej wiarygodności (*maximum likelihood, ML*), prowadzącą do prawidłowego, w sensie zachowania dodatniej półokreśloności, oszacowania macierzy $\widehat{\boldsymbol{\Sigma}}$. W pierwszym kroku algorytmu szacuje się wartość oczekiwaną $\hat{\mu}_X = \bar{x}_{A\cup B}$ oraz wariancję $\hat{\sigma}^2_X = S^2_{X,A\cup B}$. Parametry wykorzystujące informacje z $\boldsymbol{Y}$ szacuje się na podstawie równania regresji:

$$Y = \alpha_Y + \beta_{YX}X + \varepsilon_{Y|X} \tag{4.31}$$

gdzie:

$$\hat{\beta}_{YX} = \frac{S_{YX,B}}{S_{X,B}}, \tag{4.32}$$

$$\alpha_Y = \bar{y}_A - \hat{\beta}_{YX}\bar{x}_A. \tag{4.33}$$

Wynika z tego, że:

$$\hat{\mu}_Y = \hat{\alpha}_Y + \hat{\beta}_{YX}\hat{\mu}_X, \tag{4.34}$$

$$\hat{\sigma}^2_Y = S^2_{Y,A} + \hat{\beta}_{YX}\left(S^2_{X,A\cup B} - S^2_{X,A}\right), \tag{4.35}$$

$$\hat{\sigma}_{XY} = \hat{\beta}_{YX}S^2_{X,A\cup B}. \tag{4.36}$$

Analogicznie wyznacza się parametry dla $\boldsymbol{Z}$: $\hat{\mu}_Z$, $\hat{\sigma}^2_Z$ oraz $\hat{\sigma}_{XZ}$. Następnie wyznaczone wariancje i kowariancje podstawia się do wzoru (4.28) otrzymując estymator $\hat{\sigma}_{YZ}$.

Dla zmiennych jakościowych szukanym parametrem jest częstość (4.16). Przy założeniu CIA można wyznaczyć [D'Orazio *et al.* 2006]:

$$P(X = i, Y = j, Z = k) = P(Y = j|X = i)P(Z = k|X = i)P(X = i), \tag{4.37}$$

$$\theta_{ijk} = \theta_{j|i}\theta_{k|i}\theta_{i..} = \frac{\theta_{ij.}}{\theta_{i..}}\frac{\theta_{i.k}}{\theta_{ij.}}\theta_{i..} = \frac{\theta_{ij.}\theta_{i.k}}{\theta_{i..}}. \tag{4.38}$$

Wartości brzegowe tabeli $Y \times Z$ uzyskiwane są z:

$$\sum_i \theta_{ijk} = \sum_{i=1}^{I} \frac{\theta_{ij.}\theta_{i.k}}{\theta_{i..}}. \quad (4.39)$$

Niech $n_{A,ij.}$ będą liczebnościami w tabeli $X \times Y$ uzyskanej ze zbioru A, $n_{B,i.k}$ – liczebnościami w tabeli $X \times Z$ uzyskanej ze zbioru B. Wykorzystując estymator największej wiarygodności[102] [Anderson 1957] otrzymuje się:

$$\hat{\theta}_{i..} = \frac{n_{A,i..}+n_{B,i..}}{n_A+n_B}, \quad (4.40)$$

$$\hat{\theta}_{j|i} = \frac{n_{A,ij.}}{n_{A,i..}}, \quad (4.41)$$

$$\hat{\theta}_{k|i} = \frac{n_{B,i.k}}{n_{B,i..}}. \quad (4.42)$$

Wykorzystanie metod parametrycznych w podejściu makro wymaga ustalenia postaci modelu integracji. Pewne problemy mogą się pojawić przy dużej liczbie zmiennych mierzonych na różnych skalach. Ułatwieniem może być wówczas transformacja:

- zamiana zmiennych jakościowych na dychotomiczne i traktowanie ich jak zmiennych ciągłych;
- kategoryzacja zmiennych ciągłych (problem utraty informacji).

B. Podejście nieparametryczne

Podejście nieparametryczne opiera się na szacowaniu estymatorów jądrowych [D'Orazio *et al.* 2006]. Zostały one szczegółowo opisane w [Wand, Jones 1995], [Silverman 1986] oraz [Eubank 1988]. Ze względu na jego rzadkie wykorzystywanie, podejście to nie będzie przedmiotem rozważań w niniejszej rozprawie.

Wykorzystanie informacji dodatkowych

W podejściu makro najczęściej stosowane są dwie metody wykorzystujące informację dodatkową:

- metoda Kadane'a,
- podejście Renssena.

W metodzie Kadane'a wykorzystuje się znajomość kowariancji $\boldsymbol{Y}$ i $\boldsymbol{Z}$ dostępną z innych źródeł, natomiast w podejściu Renssena wykorzystuje się informacje o łącznym rozkładzie

[102] Raessler [2002] zaproponowała szacowanie KMNK.

cech $\boldsymbol{Y}$ i $\boldsymbol{Z}$ oszacowaną z dodatkowego źródła danych C. Dodatkowo znane są wagi analityczne wynikające ze schematu losowania jednostek do prób.

Metoda zaproponowana przez Kadane'a [1978] wykorzystuje macierz wariancji i kowariancji zmiennych $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$. Przy pewnych założeniach dotyczących kowariancji σ_{YZ} (znanej z innych źródeł) oraz wykorzystaniu określonych równań regresji, wyznacza się łączny rozkład $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$[103].

Metodą, w której oprócz dodatkowych informacji wykorzystuje się informacje pochodzące ze schematu losowania próby jest podejście kalibracyjne Renssena [1998]. Oparte jest ono na algorytmie kalibracji wag analitycznych wynikających ze schematu losowania, oddzielnie dla A i B. Algorytm składa się z dwóch faz:

- harmonizacja wag analitycznych w obu zbiorach do liczebności ogólnej $\boldsymbol{X}$,
- wykorzystanie dodatkowego źródła danych C, w którym $\boldsymbol{X}$, $\boldsymbol{Y}$ i $\boldsymbol{Z}$ są łącznie obserwowane w celu oszacowania związku między $\boldsymbol{Y}$ i $\boldsymbol{Z}$.

Wynikiem procedury Renssena jest tabela kontyngencji $Y \times Z$.

Niech d_k oznacza wagi początkowe, a w_k finalne wagi kalibracyjne. Wagi finalne uzyskiwane są jako rozwiązanie zagadnienia optymalizacji $\min[\sum_{k\in r} D(d_k, w_k)]$, gdzie $D(d, w)$ to miara odległości, z zastrzeżeniem, że $\sum_{k=1}^{m} w_k x_k = \sum_{k=1}^{N} x_k$ oraz $\sum_{k=1}^{m} w_k = N$.

Pierwsza faza polega na harmonizacji wag w integrowanych zbiorach. Wybiera się podzbiór zmiennych $\boldsymbol{X_1} \subseteq \boldsymbol{X}$, dla których znane są liczebności w populacji generalnej:

- wagi w_a w zbiorze A są kalibrowane w taki sposób, by wagi kalibracyjne $w_a^{(1)}$ spełniały warunek $\sum_{a\in A} w_a^{(1)} \boldsymbol{x}_{1a} = \boldsymbol{t}_1$, gdzie $\boldsymbol{t}_1$ oznacza wektor wartości globalnych w populacji,
- wagi w_b w zbiorze B są kalibrowane w taki sposób, by wagi kalibracyjne $w_b^{(1)}$ spełniały warunek $\sum_{b\in B} w_b^{(1)} \boldsymbol{x}_{1b} = \boldsymbol{t}_1$.

Jeżeli istnieją jakieś zmienne $\boldsymbol{X_2} \subseteq \boldsymbol{X}$, dla których liczebności w populacji nie są znane, w kolejnym kroku wyznaczany jest łączny estymator (*pooled estimate*) [104]:

$$\hat{\boldsymbol{t}}_2 = \lambda \sum_{a\in A} w_a^{(1)} \boldsymbol{x}_{2a} + (1-\lambda) \sum_{b\in B} w_b^{(1)} \boldsymbol{x}_{2b}, \qquad (4.43)$$

gdzie $0 \leq \lambda \leq 1$. Następnie wagi $w_a^{(1)}$ i $w_b^{(1)}$ są rekalibrowane w taki sposób, że:

[103] Szerzej o metodzie piszą Moriarty, Scheuren [2001].

[104] Wartość λ może być wyznaczona w sposób ekspercki. Jeżeli nie istnieją przesłanki do wyznaczenia konkretnej wartości, zwykle przyjmuje się, że $\lambda = \frac{n_A}{n_A + n_B}$.

— w zbiorze A powstają wagi $w_a^{(2)}$ spełniające warunek $\sum_{a\in A} w_a^{(2)} \boldsymbol{x}_{1a} = \boldsymbol{t}_1$ oraz $\sum_{a\in A} w_a^{(2)} \boldsymbol{x}_{2a} = \hat{\boldsymbol{t}}_2$,

— w zbiorze B powstają wagi $w_b^{(2)}$ spełniające warunek $\sum_{b\in B} w_b^{(2)} \boldsymbol{x}_{1b} = \boldsymbol{t}_1$ oraz $\sum_{b\in B} w_b^{(2)} \boldsymbol{x}_{2b} = \hat{\boldsymbol{t}}_2$.

Wagi kalibracyjne $w_a^{(2)}$ i $w_b^{(2)}$ mogą zostać użyte do wyznaczenia estymatorów łącznych rozkładów w A i B. Dla zmiennych jakościowych, przy CIA, łączny rozkład $\boldsymbol{Y}$ i $\boldsymbol{Z}$ może zostać oszacowany za pomocą (4.38).

Posiadając informacje pomocnicze w postaci dodatkowego zbioru C, w którym $\boldsymbol{X}$, $\boldsymbol{Y}$ i $\boldsymbol{Z}$ są łącznie obserwowane istnieją dwa alternatywne sposoby oszacowania łącznego rozkładu $\boldsymbol{Y}$ i $\boldsymbol{Z}$:

— niekompletna dwukierunkowa stratyfikacja (*incomplete two-way stratification*),

— syntetyczna dwukierunkowa stratyfikacja (*synthetic two-way stratification*).

Niekompletna dwukierunkowa stratyfikacja jest prostszą procedurą. Polega ona na kalibracji wag w_c w zbiorze C poprzez ograniczenie ich do liczebności populacji zmiennej Y ze zbioru A oraz zmiennej Z ze zbioru B.

Druga z procedur wymaga oszacowania łącznego rozkładu $\boldsymbol{Y}$ i $\boldsymbol{Z}$ przy założeniu o warunkowej niezależności, a następnie dokonania korekty przy użyciu wartości resztowych obliczonych ze zbioru C (pomiędzy liczebnościami empirycznymi i teoretycznymi dla $\boldsymbol{Y}$ i $\boldsymbol{Z}$).

W praktyce zdarzają się sytuacje, w których procedura kalibracji w podejściu Renssena jest nieskuteczna (tzn. algortym nie osiąga zbieżności, pojawiają się ujemne wagi itp.). Ma to miejsce zwłaszcza w przypadku, gdy wektor $\boldsymbol{X}$ zawiera zmienne mierzone na różnej skali lub (i) gdy zmienne jakościowe charakteryzują się dużą liczbą wariantów [Szymkowiak 2007]. W takich przypadkach należy grupować warianty cech jakościowych lub (i) kategoryzować zmiennej ilościowe.

Problem błędu ekologicznego

Wnioskowanie ekologiczne polega na wykorzystywaniu informacji zagregowanych w celu wnioskowania na poziomie jednostkowym [Hudson *et al.* 2010]. Szacując relację między zmiennymi, w sytuacji braku danych jednostkowych, wnioskujemy na poziomie indywidualnym, na podstawie informacji danych w postaci tabeli kontyngencji. Podstawowym problemem jest fakt, że wiele różnych zależności na poziomie indywidualnych osób nie można opisać poprzez wielkości uzyskane w wyniku agregacji np. na poziomie województw (np.

podczas uśredniania cech dla danej domeny). Może to prowadzić do błędów we wnioskowaniu. Błąd taki określa się mianem ekologicznego. W parowaniu statystycznym jest on ściśle związany z zagadnieniem niepewności [D'Orazio 2006]. Problem błędu ekologicznego wymaga szczególnie skrupulatnego rozważenia ponieważ jest związany z szacowaniem postaci łącznego rozkładu cech $\boldsymbol{Y}$ i $\boldsymbol{Z}$ przy znajomości wyłącznie rozkładów brzegowych. Wśród metod wnioskowania ekologicznego w literaturze wymienia się, między innymi, regresję ekologiczną [Goodman 1953, Chambers, Steel 2001]. Istotę i znaczenie wnioskowania ekologicznego dobrze ilustruje poniższy przykład zaczerpnięty z pracy Di Zio [2012].
Rozważania dotyczyły szacowania współzależności między płcią a skłonnością do głosowania na podstawie zagregowanych informacji zawartych w tabeli kontyngencji (por. tab.4.4). Celem było określenie relacji dla każdego i-tego okręgu, np. poprzez oszacowanie frakcji głosujących kobiet β_k oraz frakcji głosujących mężczyzn β_m dla całej populacji.

Tabela 4.4. Skłonność do głosowania w ujęciu płci dla *i*-tego okręgu wyborczego

Płeć głosującego	Skłonność do głosowania		
	Głosował(a)	Nie głosował(a)	Ogółem
Kobieta	β_{ki}	$1-\beta_{ki}$	p_i
Mężczyzna	β_{mi}	$1-\beta_{mi}$	$1-p_i$
Ogółem	q_i	$1-q_i$	1

Uwaga:
p_i – odsetek kobiet,
q_i – odsetek głosujących,
β_{ki} – odsetek głosujących kobiet,
β_{mi} – odsetek głosujących mężczyzn.
Źródło: na podstawie [Di Zio 2012]

Niech zmienna $\boldsymbol{X}$ będzie 41 wariantową cechą zawierającą informacje o okręgach wyborczych, $\boldsymbol{Y}$ dychotomiczną zmienną określającą płeć osoby, a $\boldsymbol{Z}$ dychotomiczną zmienną określającą, czy wyborca głosował, czy też nie. Zamiast szacowania łącznego rozkładu $(\boldsymbol{Y},\boldsymbol{Z})$ przy danym $\boldsymbol{X}$, w tym przypadku ma miejsce oszacowanie warunkowego rozkładu $\boldsymbol{Z}$ przy danym $(\boldsymbol{X},\boldsymbol{Y})$.
Regresja ekologiczna służy do wnioskowania o współzależności w sytuacji posiadania jedynie częściowej informacji. Model regresji ekologicznej Goodmana[105] [1953] wykorzystuje wszystkie dostępne informacje w taki sposób, że:

$$q_i = \beta_{ki}p_i + \beta_{mi}(1-p_i) \qquad (4.44)$$

[105] Model jest skonstruowany dla dychotomicznych zmiennych Y i Z.

Płaszczyznę regresji (4.44) nazywa się linią tomograficzną. W modelu zakłada się, że $\beta_{ki} = \beta_k$ oraz $\beta_{mi} = \beta_m$, a więc, że skłonność do głosowania jest taka sama we wszystkich okręgach. Jest to więc założenie analogiczne do założenia o warunkowej niezależności.
Chambers i Steel [2001] zaproponowali rozwinięcie modelu poprzez utworzenie wszystkich możliwych modeli wykorzystujących dostępne dane, np.:

$$\beta_{mi} = \gamma q_i. \quad (4.45)$$

Dolne i górne granice dla parametru γ mogą zostać wyznaczone przez tzw. granice Frecheta. Przedział przez nie wyznaczony odpowiada zagadnieniu niepewności. Zagadnienie to wraz z granicami Frecheta zostanie bardziej szczegółowo przedstawione w sekcji 4.4.3.

4.4.3. Podejście mikro

Podejście mikro sprowadza się do utworzenia syntetycznego zbioru danych przedstawiającego łączną informację o cechach $\boldsymbol{XYZ}$. Zbiór ten powstaje w wyniku imputacji brakujących wartości w pliku A i B (por. schematy 4.6 i 4.7). Syntetyczność nowo utworzonego, zintegrowanego zbioru polega na tym, że jednostki w nim obserwowane nie są jednostkami rzeczywistymi, gdyż przyłączone wartości zmiennych $\boldsymbol{Y}$ oraz $\boldsymbol{Z}$ nie są wartościami empirycznie obserwowanymi dla konkretnych jednostek, osób czy gospodarstw domowych. W pierwszej kolejności przedstawione zostaną parametryczne metody integracji danych stosowane w podejściu mikro. Następnie omówione zostaną również metody nieparametryczne oraz podejście mieszane wykorzystujące elementy obu poprzednich.

<u>Założenie o warunkowej niezależności</u>

A. Metody parametryczne

Jeżeli zakłada się warunkową niezależność między ciągłymi cechami $\boldsymbol{Y}$ i $\boldsymbol{Z}$ przy danym $\boldsymbol{X}$ oraz korzysta się z podejścia parametrycznego, zastosować można zasadniczo dwie techniki integracji [D'Orazio *et al.* 2006, Raessler 2002]:

— imputację regresyjną,

— stochastyczną imputację regresyjną.

Imputacja regresyjna w parowaniu statystycznym polega na konstrukcji modeli regresji $Z(X)$ oraz $Y(X)$, a następnie imputacji wartości teoretycznych wynikających z modeli, odpowiednio do zbiorów A oraz B. Proces ten składa się z trzech etapów:

a) Do zbioru A imputowane są wartości teoretyczne wynikające z modelu:

$$\hat{z}_a^{(A)} = \hat{\alpha}_Z + \hat{\beta}_{ZX} x_a, a = 1,2, \dots, n_A. \quad (4.46)$$

b) Do zbioru B imputowane są wartości teoretyczne wynikające z modelu:

$$\hat{y}_b^{(B)} = \hat{\alpha}_Y + \hat{\beta}_{YX} x_b,\ b = 1,2,\dots,n_B. \tag{4.47}$$

c) Konkatenacja zbiorów A i B : $S = A \cup B;\ n_S = n_A + n_B$.

Estymację punktową i przedziałową parametrów modeli (4.46) i (4.47) przeprowadzić można klasyczną metodą najmniejszych kwadratów (KMNK).

Podejście to charakteryzuje się dużą prostotą. Wartości teoretyczne uzyskuje się poprzez podstawienie do równań (4.46) i (4.47) wartości, odpowiednio x_a i x_b. Skonstruowane modele regresji mogą w dobry sposób przybliżać prawdziwe, nieznane wartości. Jako wadę podejścia imputacji regresyjnej postrzegać można ć możliwość przyjmowania przez wartości teoretyczne wielkości spoza empirycznego przedziału zmienności dołączanych zmiennych (np. ujemne wynagrodzenia). Niewątpliwym minusem zastosowania modeli regresji jest również fakt, że imputowane wartości leżą wyłącznie na prostej lub płaszczyźnie (hiperpłaszczyźnie) regresji. Niweluje to zmienność próby, co może prowadzić do obciążeń estymatora wariancji dołączanych cech.

Little i Rubin [2002] zaproponowali alternatywę dla metody imputacji regresyjnej. Przy założeniu, że braki danych generowane są w sposób losowy[106], można uzyskać lepsze rezultaty niż w przypadku imputacji regresyjnej, jeżeli do wartości teoretycznych wynikających z modeli regresji dolosowane są wartości z określonego rozkładu. Podejście to nazwane zostało stochastyczną imputacją regresyjną i polega na dolosowaniu do wartości teoretycznych wynikających z równań (4.46) i (4.47) wartości składnika losowego, w taki sposób, że:

$$\tilde{z}_a^{(A)} = \hat{z}_a^{(A)} + e_a = \hat{\alpha}_Z + \hat{\beta}_{ZX} x_a + e_a \tag{4.48}$$

gdzie $e_a \sim N\left(0, \hat{\sigma}_{Z|X}\right)$, oraz

$$\tilde{y}_b^{(B)} = \hat{y}_b^{(B)} + e_b = \hat{\alpha}_Y + \hat{\beta}_{YX} x_b + e_b \tag{4.49}$$

gdzie $e_b \sim N\left(0, \hat{\sigma}_{Y|X}\right)$.

Rozwinięciem metody stochastycznej imputacji regresyjnej jest zaproponowana przez Raessler [2002] metoda wielokrotnej imputacji. Jest ona rozwinięciem podejścia Rubina [1986] do integracji zbiorów danych pochodzących ze złożonych schematów losowania. W podejściu tym uwzględnia się dodatkowo schemat losowania każdego z badań – dokonuje się przekształcenia prawdopodobieństwa inkluzji poszczególnych jednostek w każdym ze zbiorów w taki sposób, by syntetyczny zbiór odzwierciedlał liczebność populacji generalnej. Prawdopodobieństwo inkluzji każdej i-tej jednostki w zintegrowanym zbiorze jest sumą

[106]Mechanizm generowania braków danych MAR (*Missing At Random*).

prawdopodobieństw wylosowania do próby w badaniach A i B pomniejszoną o prawdopodobieństwo wylosowania tej jednostki do obu badań jednocześnie:

$$\pi_{A\cup B,i} = \pi_{A,i} + \pi_{B,i} - \pi_{A\cap B,i} \tag{4.50}$$

Liczebność badań częściowych stanowi zwykle bardzo niewielki odsetek liczebności całej populacji. Również instytucje przeprowadzające pomiar dbają by respondenci nie byli nadmiernie obciążeni obowiązkami wynikającymi z udziału w badaniu i starają się nie uwzględniać tej samej jednostki w kilku badaniach jednocześnie. Dlatego, wzór (4.54) można uprościć do postaci:

$$\pi_{A\cup B,i} \cong \pi_{A,i} + \pi_{B,i}. \tag{4.51}$$

Wynikająca ze schematu losowania waga analityczna jest odwrotnością prawdopodobieństwa wylosowania jednostki do próby. W zintegrowanym zbiorze będzie więc przyjmować postać:

$$w_{i_{A\cup B}} = \frac{1}{\pi_{A\cup B,i}}. \tag{4.52}$$

W praktyce jednak zwykle prawdopodobieństwa inkluzji nie są dostępne, natomiast zbiory danych zawierają określone wagi analityczne. Aby syntetyczny zbiór danych odzwierciedlał liczebność populacji generalnej, dokonuje się przekształcenia wag wg następującej formuły:

$$w'_{i_{A\cup B}} = \frac{w_{i_{A\cup B}}}{\sum_{i=1}^{s} w_{i_{A\cup B}}} N, \tag{4.53}$$

gdzie:

$w'_{i_{A\cup B}}$ - zharmonizowana waga analityczna dla i-tej jednostki w zintegrowanym zbiorze,

$w_{i_{A\cup B}}$ - oryginalna waga analityczna,

N – liczebność populacji generalnej.

Schemat 4.8. Dane wejściowe w podejściu Rubina

<table>
<tr><td>y_1</td><td>x_{11}</td><td>…</td><td>$\mathrm{x}_{\mathrm{p}1}$</td><td rowspan="4">brak danych</td><td>w_{A1}</td></tr>
<tr><td>y_2</td><td>x_{12}</td><td>…</td><td>x_{p2}</td><td>w_{A2}</td></tr>
<tr><td>…</td><td>…</td><td>…</td><td>…</td><td>…</td></tr>
<tr><td>y_{n_A}</td><td>x_{1n_A}</td><td>…</td><td>x_{pn_A}</td><td>w_{An_A}</td></tr>
<tr><td rowspan="4">brak danych</td><td>x_{11}</td><td>…</td><td>$\mathrm{x}_{\mathrm{p}1}$</td><td>z_1</td><td>w_{B1}</td></tr>
<tr><td>x_{12}</td><td>…</td><td>x_{p2}</td><td>z_2</td><td>w_{B2}</td></tr>
<tr><td>…</td><td>…</td><td>…</td><td>…</td><td>…</td></tr>
<tr><td>x_{1n_B}</td><td>…</td><td>x_{pn_B}</td><td>z_{n_B}</td><td>w_{Bn_B}</td></tr>
</table>

Źródło: Rubin [1986]

W kolejnym kroku wykorzystuje się metodę wielokrotnej imputacji w celu oszacowania wartości braków danych z wykorzystaniem przekształconych wag analitycznych (por. schemat 4.8).

Każdy brak danych jest imputowany za pomocą pewnej liczby[107] (m) wartości. Te m wartości są uporządkowane w takim sensie, że pierwszy zestaw wartości tworzy pierwszy zbiór danych itd. Oznacza to, że tworzonych jest m kompletnych zbiorów danych. Każdy z tych zbiorów jest analizowany za pomocą standardowych procedur wykorzystujących informację pełną w taki sposób, jakby wartości imputowane były prawdziwe.

Niech $M: A \rightarrow B$ będzie dowolną metodą statystyczną wymagającą kompletnych danych, w której dane wejściowe to A, a dane wyjściowe to B. Wektor zmiennych X zawiera braki danych.

1. Estymuje się parametry wielowymiarowego rozkładu R danych X.
2. Wykonuje się w pętli dużą liczbę powtórzeń $i = 1,2,\cdots,N$ następujących czynności:
    a) Uzupełnia się braki danych w X wartościami wylosowanymi z rozkładu R, otrzymując X_i,
    b) Stosuje się metodę M obliczając $Y_i = M(X_i)$;
3. Uśrednia (łączy) się wyniki $Y_1, Y_2, \cdots, Y_N$ aby otrzymać Y [Davey *et al.* 2001].

Na potrzeby wielokrotnej imputacji tworzy się m modeli, gdzie do wartości teoretycznych wynikających z modeli imputacji regresyjnej dolosowane są różne wartości resztowe. Odzwierciedla to zmienność próby, a także umożliwia przeprowadzenie estymacji punktowej i przedziałowej dla nieznanych wartości braków danych (jest to również rozwiązanie problemu niepewności, opisanego w sekcji 4.4.3).

Estymatorem dla każdego z t ($t = 1,2,\ldots,m$) podstawień jest $\hat{\theta}^{(t)} = \hat{\theta}(U_{obs}, U_{mis}^{(t)})$, gdzie U_{obs} to wartości obserwowane dla danej cechy, zaś $U_{mis}^{(t)}$ to zaimputowane braki danych [Raessler 2004]. Wariancję tego estymatora można wyrazić jako $\widehat{var}\left(\hat{\theta}^{(t)}\right) = \widehat{var}(\hat{\theta}(U_{obs}, U_{mis}^{(t)}))$. Estymatorem punktowym wielokrotnej imputacji jest średnia arytmetyczna z m podstawień:

$$\hat{\theta}_{MI} = \frac{1}{m}\sum_{t=1}^{m} \hat{\theta}^{(t)}. \qquad (4.54)$$

[107] Literatura [Raessler 2002, Rubin 1987] wskazuje, że liczba imputacji nie musi być duża. Mówi się wręcz o 3 – 5. Wynika to z faktu, że Rubin [1987] wykazał, że efektywność określonej liczby podstawień w porównaniu do sytuacji, gdyby była ich nieskończona liczba można wyrazić wzorem $1 + \frac{\lambda}{m}$ gdzie λ to frakcja braków danych. Np. dla frakcji braków rzędu 0,6 dla 20 podstawień, efektywność wynosi $1 + \frac{0,6}{20} = 1,03$ i oznacza, że oszacowany estymator charakteryzuje się błędem standardowym o 3% większym niż ten oszacowany na podstawie dążącej do nieskończoności liczby imputacji.

Wariancja estymatora wielokrotnej imputacji dzieli się na wariancję wewnątrzgrupową i wariancję międzygrupową. Wariancja międzygrupowa wyraża się wzorem:

$$B = \frac{1}{m-1}\sum_{t=1}^{m}(\hat{\theta}^{(t)} - \hat{\theta}_{MI})^2, \tag{4.55}$$

zaś wariancję wewnątrzgrupową można zapisać jako wyrażenie:

$$W = \frac{1}{m}\sum_{t=1}^{m}\widehat{var}\left(\hat{\theta}^{(t)}\right). \tag{4.56}$$

Wariancja ogólna jest sumą wariancji wewnątrz- i międzygrupowej zmodyfikowanym o składnik $\frac{m+1}{m}$ zwiększający dyspersję estymatora, co ma odzwierciedlać niepewność co do prawdziwych wartości imputowanych braków danych:

$$T = W + \frac{m+1}{m}B. \tag{4.57}$$

Estymacji przedziałowej w wielokrotnej imputacji dokonuje się szacując przedział ufności:

$$\hat{\theta}_{MI} - t_{v,\frac{\alpha}{2}}\sqrt{T} < \theta < \hat{\theta}_{MI} + t_{v,\frac{\alpha}{2}}\sqrt{T}\,, \tag{4.58}$$

gdzie liczba stopni swobody $v = (m-1)(1 + \frac{W}{(1+\frac{1}{m})B})^2$.

Główną zaletą podejścia parametrycznego jest „oszczędność" modelu – niewielka liczba predyktorów wyjaśnia dużą część zmienności dołączanych wartości. Wśród wad należy wymienić przede wszystkim konieczność specyfikacji modelu. Źle skonstruowany model imputacji może generować rezultaty o słabej jakości[108]. Dodatkowo, imputowane wartości są sztuczne, tj. wynikające wyłącznie z modelu, nie mające swoich odpowiedników w rzeczywistości (imputowane wartości nie są wartościami empirycznymi, zaobserwowanymi w rzeczywistości). Problem ten rozwiązuje się zwykle przy wykorzystaniu podejścia mieszanego.

B. Metody nieparametryczne

W przypadku podejścia nieparametrycznego, stosuje się metody często wykorzystywane w przypadku imputacji w zbiorach danych pochodzących z badań próbkowych [D'Orazio 2012]. W takiej sytuacji nie stosuje się podejścia konkatenacji zbiorów. Zintegrowany, syntetyczny zbiór danych jest zbiorem biorcy po dołączeniu wartości ze zbioru dawcy (por. schemat 4.6). Problem wyboru zbioru biorcy i dawcy jest przedmiotem sporu w literaturze. Ponieważ zwykle jeden ze zbiorów jest bardziej liczny od drugiego, D'Orazio *et al.* [2006] sugerują, by biorcą był zbiór mniejszy, ponieważ w przeciwnym wypadku niektóre rekordy będą dołączone więcej niż jeden raz, co może prowadzić do zniekształcenia rozkładu impu-

[108] Metody oceny jakości integracji zostaną przedstawione w sekcji 4.5.

towanych wartości (rozkład będzie „sztuczny"). Z kolei Raessler [2002] postuluje, by biorcą był zbiór większy. Podejście to argumentuje koniecznością wykorzystania wszystkich dostępnych informacji.

Na potrzeby dalszych rozważań, niech A będzie zbiorem biorcy zawierającym zmienne $\boldsymbol{X}$ i $\boldsymbol{Y}$, zaś B zbiorem dawcy zawierającym zmienne $\boldsymbol{X}$ i $\boldsymbol{Z}$. Syntetyczny, zintegrowany zbiór S jest tworzony poprzez imputację zmiennych $\boldsymbol{Z}$ w A. Imputowane wartości $\boldsymbol{Z}$ są obserwowane w B – są więc wartościami rzeczywistymi („żywymi" - *live values*). Najczęściej wykorzystywanymi metodami nieparametrycznymi są zaproponowane przez Singh *et al.* [1993] metody typu *hot deck*[109]:

— losowa,

— najbliższego sąsiada (najmniejszej odległości),

— rangowa.

Metoda losowa polega na losowym doborze zmiennej $\boldsymbol{Z}$ ze zbioru dawcy do zbioru biorcy. By zachować jak największą zgodność dołączanych wartości, zbiory A i B dzielone są na jak największą liczbę homogenicznych grup (na podstawie wartości wybranych zmiennych, najlepiej jakościowych, $\boldsymbol{X}$) - $\boldsymbol{X_G}$. Grup takich powinno być możliwie dużo. Losowe dołączanie przebiega wtedy w obrębie wyznaczonych grup.

Metoda najbliższego sąsiada polega na wybraniu dla każdego rekordu ze zbioru A najbardziej podobnego rekordu ze zbioru B. „Podobieństwo" to mierzone jest odległością między wartościami zmiennych parujących wybranych z wektora zmiennych wspólnych $\boldsymbol{X}$ $\boldsymbol{(X_M \subseteq X)}$:

$$d_{ab} = (x_{M.a}, x_{M.b}) = min,\ b = 1,2,\dots,n_b. \qquad (4.59)$$

Wartość $\boldsymbol{Z}$ jest następnie imputowana w A. W przypadku, gdy kilka rekordów dawcy charakteryzuje się taką samą odległością do danego rekordu biorcy, dołączany rekord wybiera się losowo.

Wariacją metody najbliższego sąsiada jest metoda k najbliższych sąsiadów. W metodzie tej dla każdego rekordu biorcy wybiera się k najbliższych sąsiadów (rekordów o najmniejszej odległości), a następnie spośród nich losowo dobiera się dołączany rekord. Aluja-Banet *et al.* [2007] zastosował podejście, w którym k najbliższym sąsiadom przyporządkowuje się „wagi" odwrotnie proporcjonalne do odległości w taki sposób, że rekordy dawcy o mniej-

[109] W imputacji metody typu *hot deck* polegają na zastępowaniu braków porównywalnymi wartościami z tego samego zbioru danych.

szym dystansie do rekordu biorcy charakteryzują się większym prawdopodobieństwem przyłączenia.

Do obliczenia odległości między rekordami w zbiorach A i B można użyć dowolnej funkcji odległości, która spełnia następujące założenia [D'Orazio et al. 2006]:

— jest symetryczna: $d_{ab} = d_{ba}$,

— jest nieujemna: $d_{ab} \geq 0$,

— jest tożsama: $d_{aa} = 0$,

— ma własności metryki:

- spełnia zasadę identyczności przedmiotów nierozróżnialnych: $d_{ab} = 0 \Leftrightarrow a = b$,
- spełnia nierówność trójkąta: $d_{ab} \leq d_{ac} + d_{cb}$.

Powyższe założenia spełnia klasa funkcji odległości opisanych na podstawie miary Minkowskiego [Gatnar, Walesiak 2009]:

$$d_{ab} = \sqrt[\lambda]{\sum_{p=1}^{P} \left| x_{ap} - x_{bp} \right|^{\lambda}}, \tag{4.60}$$

gdzie P oznacza liczbę zmiennych parujących (rekord jest P-wymiarowy), a $\lambda \geq 1$.

Najczęściej stosowanymi metrykami opartymi na mierze Minkowskiego są:

— metryka miejska (Manhattan, $\lambda = 1$)

$$d_{ab} = \sum_{p=1}^{P} \left| x_{ap} - x_{bp} \right|, \tag{4.61}$$

— metryka euklidesowa ($\lambda = 2$)

$$d_{ab} = \sqrt{\sum_{p=1}^{P} \left(x_{ap} - x_{bp} \right)^2}, \tag{4.62}$$

— metryka Czebyszewa ($\lambda \to \infty$)

$$d_{ab} = \max_p \left| x_{ap} - x_{bp} \right|. \tag{4.63}$$

Wśród miar odległości, które również można wykorzystać w procesie integracji można wymienić

— odległość Mahalanobisa

$$d_{ab} = (\boldsymbol{x_a} - \boldsymbol{x_b})^T \Sigma_{\boldsymbol{XX}}^{-1} (\boldsymbol{x_a} - \boldsymbol{x_b}), \tag{4.64}$$

gdzie $\Sigma_{\boldsymbol{XX}}$ to macierz wariancji i kowariancji $\boldsymbol{X}$,

— uogólnioną miarą odległości GDM1 i GDM2 [Gatnar, Walesiak 2009].

W przypadku, gdy wśród zmiennych parujących $\boldsymbol{X_M}$ występują cechy o różnym poziomie pomiaru, zastosować można jedno z następujących podejść [D'Orazio et al. 2006]:

— zamiana zmiennych jakościowych na ilościowe (np. poprzez rangowanie[110]) i zastosowanie którejkolwiek miary (4.60) – (4.64),

— zastosowanie miary odległości uwzględniającej różny charakter zmiennych w wektorze zmiennych parujących.

Wśród miar odległości uwzględniających różny charakter zmiennych często wymienia się współczynnik niepodobieństwa Gowera (*Gower's dissimilarity coefficient*). Współczynnik Gowera wyznacza się poprzez obliczenie uśrednionej odległości dla wszystkich zmiennych:

$$d_{ab} = \frac{1}{P}\sum_{p=1}^{P} c_p d_{abp}, \qquad (4.65)$$

gdzie $c_p = 1$ dla zmiennych binarnych (zdychotomizowanych zmiennych jakościowych) oraz $c_p = \frac{1}{R_p}$ dla zmiennych ilościowych i jakościowych porządkowych, gdzie R_p to rozstęp. Odległością d_{abp} może być każda metryka, choć ze względu na występowanie zdychotomizowanych zmiennych jakościowych najczęściej oblicza się odległość miejską ($d_{abp} = 0$ jeżeli warianty się zgadzają i $d_{abp} = 1$ w przeciwnym przypadku).

By zoptymalizować algorytm integracji[111] bardzo często, podobnie jak w przypadku podejścia losowego, zbiory dzieli się na rozłączne podzbiory. Dołączanie rekordów odbywa się wtedy w podzbiorach wyznaczonych przez zmienne $\boldsymbol{X_G} \subseteq \boldsymbol{X}$ (np. łączone są osobno rekordy dla mężczyzn i osobno dla kobiet w przypadku, gdy zmienną grupującą jest płeć).

Bacher [2002] zaproponował algorytm umożliwiający parowanie zbiorów metodą kwadratowej odległości euklidesowej stosując wagi oparte na odchyleniu standardowym. Dla zmiennych ilościowych wagi mają postać $w_p = \frac{1}{s_p}$, gdzie s_p oznacza odchylenie standardowe *p*-tej zmiennej ilościowej. Dla zmiennych jakościowych wagi wyrażają się wzorem $w_{p'k} = \frac{1}{\sqrt{2}\times s_{p'k}}$, gdzie $s_{p'k}$ jest odchyleniem standardowym dla p'-tej zmiennej jakościowej oraz jej *k*-tego wariantu (zmiennej zero- jedynkowej utworzonej ze zmiennej jakościowej). Kwadratowa odległość euklidesowa zastosowana w algorytmie wyraża się wzorem:

$$d_{ab}^2 = \sum_{p'=1}^{P'}\sum_{k=1}^{K_i} w_{p'k}^2 \left(x_{ap'k} - x_{bp'k}\right)^2 + \sum_{p=1}^{P} w_p^2 \left(x_{ap} - x_{bp}\right). \qquad (4.66)$$

Kryterium połączenia to minimalizacja powyższej funkcji odległości. Alternatywnie można zastosować drugie kryterium połączenia, a mianowicie subiektywny próg $d_{ab}^2 < c$, powyżej wartości którego minimalna wartość kwadratowej odległości euklidesowej nie jest trakto-

[110] Zmienne mierzone na poziomie nominalnym można wykorzystać np. do blokowania.

[111] Przy bardzo dużej liczbie rekordów porównywanie każdego rekordu z każdym może okazać się bardzo czasochłonnym procesem.

wana jako połączenie. Próg ten, w przypadku algorytmu dla metody najmniejszej odległości jest nieunormowany, a jego wielkość zależy od liczby zmiennych.

W metodzie najbliższego sąsiada prawie na pewno (w przypadku, gdy zbiór biorcy jest liczniejszy od zbioru dawcy – na pewno) zaistnieje sytuacja, w której jeden rekord dawcy będzie przyporządkowany więcej niż jeden raz[112]. Sytuacja taka może prowadzić do zniekształcenia rozkładu dołączanych wartości $\boldsymbol{Z}$ zwłaszcza, gdy grupa rekordów, czy nawet jeden rekord będzie dołączany szczególnie często. By temu zapobiec, Di Zio *et al.* [2006] oraz Raessler [2002] zaproponowali tzw. podejście ograniczone (*constrained*). Każdy rekord ze zbioru dawcy jest dołączany do rekordu biorcy tylko raz (przy założeniu, że $n_A \leq n_B$). Odległość obliczona między poszczególnymi rekordami jest ważona w taki sposób, by zminimalizować sumaryczną odległość między wszystkimi połączonymi rekordami. Rozwiązywane jest więc zadanie optymalizacyjne, takie, że [Kadane 1978]:

$$\sum_{a=1}^{n_A}\sum_{b=1}^{n_B}(d_{ab}w_{ab}) \rightarrow min \tag{4.67}$$

przy ograniczeniach:

$$\sum_{b=1}^{n_B} w_{ab} = 1\,,\ a = 1,2,\dots,n_a, \tag{4.68}$$

$$\sum_{a=1}^{n_A} w_{ab} \leq 1, b = 1,2,\dots,n_b, \tag{4.69}$$

gdzie $w_{ab} \in \{0,1\}$, $w_{ab} = 1$ jeżeli rekordy są połączone oraz $w_{ab} = 0$ w przeciwnym przypadku[113].

Główną zaletą podejścia ograniczonego jest lepsze odwzorowanie rozkładu dołączanej zmiennej (jest idealne, jeżeli $n_A = n_B$) niż w przypadku podejścia nieograniczonego. Wśród wad natomiast można wymienić większą średnią odległość niż w podejściu nieograniczonym oraz skomplikowanie obliczeniowe, które może znacznie wydłużyć proces integracji.

Metodę rangową wykorzystuje się w głównej mierze w sytuacji, gdy dostępna jest tylko jedna zmienną parująca X mierzona na skali co najmniej porządkowej [Singh *et al.* 1990]. Do integracji wykorzystywane jest uporządkowanie wartości w zmiennej – tzw. rangi (rangowanie jest użyteczne zwłaszcza wtedy, gdy rozkłady cechy X są różne ze względu na błędy pomiaru). Jednostki w obu zbiorach rangowane są oddzielnie. W kolejnym kroku obliczana jest wartość dystrybuanty empirycznej rozkładu zmiennej w zbiorze biorcy:

$$\hat{F}_X^A(x) = \frac{1}{n_A}\sum_{a=1}^{n_A} I(x_a \leq x),\ \ x \in \mathcal{X}, \tag{4.70}$$

oraz zbiorze dawcy:

[112] Takie łączenie nosi nazwę nieograniczonego (*unconstrained*).

[113] W przypadku, gdy $n_A = n_B \Rightarrow \sum_{a=1}^{n_A} w_{ab} = 1$

$$\hat{F}_X^B(x) = \frac{1}{n_B}\sum_{b=1}^{n_B} I(x_b \leq x), \;\; x \in \mathcal{X}. \tag{4.71}$$

Następnie dla każdego rekordu biorcy (a) przyporządkowywany jest rekord dawcy (b^*), taki, że:

$$\left|\hat{F}_X^A(x_a^A) - \hat{F}_X^B\left(x_{b^*}^B\right)\right| = \min_{1\leq b\leq n_B}\left|\hat{F}_X^A(x_a^A) - \hat{F}_X^B(x_b^B)\right|. \tag{4.72}$$

C. Metody mieszane

Trzecim typem głównych technik w parowaniu statystycznym jest podejście mieszane. Wykorzystuje się w nim metody zarówno parametryczne, jak i nieparametryczne. Parowanie statystyczne w podejściu mieszanym przeprowadzane jest zazwyczaj w dwóch krokach [Rubin 1986,1987]:

1. konstruowany jest model parametryczny oraz szacowane są jego parametry,
2. zintegrowany, syntetyczny zbiór danych jest tworzony przy wykorzystaniu technik nieparametrycznych.

Podejście mieszane ma dwie ważne zalety:

— w przypadku braku możliwości konstrukcji modelu o zadowalającej jakości, wykorzystanie technik nieparametrycznych może zniwelować błąd losowy,

— imputowane wartości są obserwowane w rzeczywistości – nie są teoretyczne.

Dla zmiennych ciągłych podejście mieszane można rozpisać jako:

1. Konstrukcję modelu regresji[114] na podstawie informacji ze zbioru dawcy (na potrzeby rozważań można przyjąć B) $Z = g(x;\theta)$. Estymacji parametrów θ. Na podstawie oszacowanego modelu obliczane są wartości teoretyczne $\tilde{z}_a$ w zbiorze A[115].
2. Dla każdego rekordu w zbiorze biorcy wyszukiwany jest „najbliższy sąsiad" w zbiorze dawcy na podstawie odległości między wartościami teoretycznymi w A i empirycznymi w B: $d_{ab}(\tilde{z}_a, z_b) = min.$

Dla zmiennych jakościowych w pierwszym kroku wykorzystywane są modele logliniowe (por. Di Zio *et al.* [2006], Singh *et al.* [1988, 1993]).

[114] D'Orazio [2011] zaproponował stosowanie w miejsce modelu regresji drzew regresyjnych i klasyfikacyjnych.

[115] W zależności od stosowanego modelu regresji, mogą to być wartości teoretyczne lub wartości teoretyczne skorygowane o składnik losowy. W przypadku metody wielokrotnej imputacji algorytm metody mieszanej wykonuje się oddzielnie dla każdego podstawienia.

Raessler [2002] zaproponowała również tzw. metodę współczynników skłonności (*propensity scores matching*[116]). W metodzie tej zbiór biorcy poszerzany jest o zmienną S, taką, że $S_i = 1$ dla wszystkich jednostek zbioru biorcy. Zmienna ta dołączana jest również do zbioru dawcy tak, że $S_j = 0$ dla wszystkich jednostek. Konkatenacja zbiorów A i B umożliwia zastosowanie modelu logitowego lub probitowego w celu oszacowania współczynników skłonności[117]. Współczynnik skłonności $e(x_i) = P(S = 1|X = x_i) = g(x_i'\beta)$ jest definiowany jako warunkowe prawdopodobieństwo, że jednostka i, $i = 1,2,\dots,n;\ n = n_A + n_B$ należy do pewnej (eksperymentalnej) grupy dla $X = x$. Wartości współczynników skłonności (dla modelu logitowego) oblicza się ze wzoru:

$$\hat{e}(x_i) = g(x_i'\hat{\beta}) = \frac{1}{1+e^{-x_i'\hat{\beta}}}. \qquad (4.73)$$

Następnie dołącza się rekordy dawcy do rekordów biorcy, dla których różnica miedzy oszacowanymi współczynnikami skłonności jest najmniejsza.

Wykorzystanie informacji dodatkowych

W podejściu mikro, dodatkowe informacje pobierane są z pomocniczego źródła C zawierającego łączną obserwację wszystkich zmiennych (por. schemat 4.9). Zbiór ten zwykle jest stosunkowo niewielką próbą, na podstawie której szacunki łącznego rozkładu $(\boldsymbol{Y}, \boldsymbol{Z})$ nie charakteryzują się zadowalającą jakością. Wykorzystanie jednak informacji z tego zbioru może przyczynić się do oszacowań bliższych rzeczywistości niż przy założeniu o warunkowej niezależności [D'Orazio *et al.* 2006].

[116] O technice *propensity scores matching* szerzej pisze Trzciński [2009].

[117] W modelach wykorzystywanych w tej technice S traktuje się jako zmienną zależną, a **X** to wektor zmiennych niezależnych. Tworzy się model z wyrazem wolnym. Zmienne **Y** ani **Z** nie są używane w procedurze.

Schemat 4.9. Dane wejściowe w sytuacji posiadania pomocniczych informacji

Zbiór A

$\boldsymbol{Y_1}$	...	$\boldsymbol{Y_Q}$	$\boldsymbol{X_1}$	...	$\boldsymbol{X_P}$
y_{11}^A	...	y_{1Q}^A	x_{11}^A	...	x_{1P}^A
...	...	...	...	...	...
y_{a1}^A	...	y_{aQ}^A	x_{a1}^A	...	x_{aP}^A
...	...	...	...	...	...
$y_{n_A1}^A$	...	$y_{n_AQ}^A$	$x_{n_A1}^A$	...	$x_{n_AP}^A$

Zbiór C

$\boldsymbol{Y_1}$	...	$\boldsymbol{Y_Q}$	$\boldsymbol{X_1}$	...	$\boldsymbol{X_P}$	$\boldsymbol{Z_1}$	...	$\boldsymbol{Z_R}$
y_{11}^C	...	y_{1Q}^C	x_{11}^C	...	x_{1P}^C	z_{11}^C	...	z_{1R}^C
...	...	...	...	...	...	...	...	...
$y_{n_C1}^C$	...	$y_{n_CQ}^C$	$x_{n_C1}^C$	...	$x_{n_CP}^C$	$z_{n_C1}^C$	...	$z_{n_CR}^C$

Zbiór B

$\boldsymbol{X_1}$	...	$\boldsymbol{X_P}$	$\boldsymbol{Z_1}$	...	$\boldsymbol{Z_R}$
x_{11}^B	...	x_{1P}^B	z_{11}^B	...	z_{1R}^B
...	...	...	...	...	...
x_{b1}^B	...	x_{bP}^B	z_{b1}^B	...	z_{bR}^B
...	...	...	...	...	...
$x_{n_B1}^B$	...	$x_{n_BP}^B$	$z_{n_B1}^B$	...	$z_{n_BR}^B$

Źródło: opracowanie własne

W **metodach parametrycznych** zbiory są poddawane procesowi konkatenacji, w taki sposób, że $S = A \cup B \cup C$. Następnie stosowana jest imputacja regresyjna lub stochastyczna imputacja regresyjna, w których modele tworzone są z wykorzystaniem informacji ze zbioru C [D'Orazio *et al.* 2006].

Korzystając z **metod nieparametrycznych**, najczęściej wykorzystuje się metodę najbliższego sąsiada [Singh *et al.* 1993]. Jeżeli dodatkowa próba C zawiera informacje o wysokiej rzetelności (np. jest całkowicie zharmonizowana pod względem populacji, definicji zmiennych i czasu z A i B), imputuje się $\boldsymbol{Z}$ przy wykorzystaniu zbioru C jako dawcy używając odległości:

— $d_{ac}\big((x_a, y_a), (x_c, y_c)\big)$, jeżeli C zawiera $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$,

— $d_{ac}(y_a, y_c)$, jeżeli C zawiera $(\boldsymbol{Y}, \boldsymbol{Z})$.

W przypadku, gdy próba C zawiera informacje o wątpliwej lub niskiej rzetelności, wtedy procedura przebiega dwustopniowo:

1. imputuje się $\boldsymbol{Z}$ do zbioru A używając zbioru C jako dawcy oraz odległości:

— $d_{ac}\big((x_a, y_a), (x_c, y_c)\big)$, jeżeli C zawiera $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$,

— $d_{ac}(y_a, y_c)$, jeżeli C zawiera $(\boldsymbol{Y}, \boldsymbol{Z})$.

2. imputuje się $\boldsymbol{Z}$ do zbioru A używając zbioru B jako dawcy oraz odległości $d_{ab}\big((x_a, \tilde{z}_a), (x_b, z_b)\big)$, gdzie $\tilde{z}_a$ to wartości imputowane w kroku 1.

W **metodach mieszanych**, przy dostępności informacji dodatkowych używa się technik analogicznych do tych dostępnych dla CIA. Są one szczegółowo opisane w [D'Orazio *et al.* 2006].

Analiza niepewności

Jeżeli założenie o warunkowej niezależności (CIA) jest nieprawdziwe i nie występują dodatkowe informacje, których można by użyć w toku integracji, należy przeanalizować tzw. „przestrzeń niepewności". Jest to zbiór wszystkich możliwych rozkładów zmiennych losowych $(\boldsymbol{Y}, \boldsymbol{Z}|\boldsymbol{X})$ zgodnych z dostępną informacją, tj. obserwowanym brzegowym rozkładem $(\boldsymbol{Y}, \boldsymbol{X})$ oraz $(\boldsymbol{Z}, \boldsymbol{X})$ [D'Orazio 2012]. W zależności od podejścia metodologicznego, w przypadku analizy niepewności rezultatem parowania statystycznego jest:

— dla podejścia makro: zbiór tak samo prawdopodobnych szacunków parametrów,

— dla podejścia mikro: rodzina jednostkowych zbiorów danych utworzonych na podstawie tak samo prawdopodobnych szacunków parametrów modelu integracji.

Analizę niepewności dla podejścia mikro przeprowadza się najczęściej przy wykorzystaniu metody wielokrotnej imputacji. Przedziały ufności dla parametrów modelu integracji są powiększane o czynnik $\frac{m+1}{m}$ (por. równanie 4.57).

Dla podejścia makro, analizę niepewności przeprowadza się tworząc przedziały dla szacowanych parametrów:

— dla zmiennych ciągłych: współczynnika korelacji ρ_{YZ},

— dla zmiennych jakościowych: dla liczebności komórek tabeli kontyngencji θ_{jk} (tzw. granice Frecheta).

Dla zmiennych ciągłych, głównym problemem w parowaniu statystycznym jest oszacowanie współczynnika korelacji ρ_{YZ} dla nieobserwowanych łącznie zmiennych. Macierz korelacji zmiennych (X, Y, Z) ma postać:

$$\rho = \begin{pmatrix} 1 & \rho_{XY} & \rho_{XZ} \\ \rho_{YX} & 1 & \rho_{YZ} \\ \rho_{ZX} & \rho_{ZY} & 1 \end{pmatrix}, \tag{4.74}$$

gdzie na podstawie informacji z $A \cup B$ nie można wyznaczyć jedynie $\hat{\rho}_{YZ}$. Jeżeli założenie o warunkowej niezależności jest prawdziwe, to:

$$\rho_{YZ} = \rho_{XY}\rho_{XZ} \quad (4.75)$$

Przy braku dodatkowej informacji o wartości ρ_{YZ} lub $\rho_{YZ|X}$[118] i przy braku założenia o warunkowej niezależności, jedyną dostępną informacją jest [Kadane 1978, Rubin 1986, Moriarity i Scheuren 2001, 2003]:

$$\rho_{XY}\rho_{XZ} - \sqrt{[(1-\rho_{XY}^2)(1-\rho_{XZ}^2)]} \leq \rho_{YZ} \leq \rho_{XY}\rho_{XZ} + \sqrt{[(1-\rho_{XY}^2)(1-\rho_{XZ}^2)]} \quad (4.76)$$

ze względu na fakt, że macierz korelacji musi być dodatnio półokreślona ($\det\rho \geq 0$). Szacunek $\rho_{YZ} = \rho_{XY}\rho_{XZ}$ jest centralnym punktem przedziału. Wartość optymalną ρ_{YZ} wyznacza się ze wzoru [D'Orazio 2012]:

$$\rho_{YZ}^{*} = \rho_{YZ}^{G} - \rho_{YZ}^{CIA} = \rho_{YZ}^{CIA} - \rho_{YZ}^{D} \quad (4.77)$$

gdzie ρ_{YZ}^{G} to górna granica przedziału niepewności, ρ_{YZ}^{D} to dolna granica, a ρ_{YZ}^{CIA} to wartość ρ_{YZ} przy założeniu warunkowej niezależności.

Dla przypadku z wieloma zmiennymi macierz korelacji przyjmuje postać:

$$\boldsymbol{\Sigma} = \begin{pmatrix} \Sigma_{XX} & \Sigma_{YX} & \Sigma_{ZX} \\ \Sigma_{XY} & \Sigma_{YY} & \Sigma_{ZY} \\ \Sigma_{XZ} & \Sigma_{YZ} & \Sigma_{ZZ} \end{pmatrix}. \quad (4.78)$$

Wartość wektora współczynników korelacji $\boldsymbol{YZ}$ wyznacza się ze wzoru [Kiesl, Raessler 2006]:

$$\hat{\Sigma}_{YZ} = \Sigma_{ZX}\Sigma_{XX}^{-1}\Sigma_{XY}, \quad (4.79)$$

natomiast przedziały niepewności dla (4.79) wyznacza się w dwóch etapach [Kiesl, Raessler 2006]:

1. Wyznaczenie wektorów własnych macierzy:

$$\tilde{C} = (I - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY})^{-1}(\Sigma_{ZZ} - \Sigma_{ZX}\Sigma_{XX}^{-1}\Sigma_{XZ})^{-1}. \quad (4.80)$$

2. Wyznaczenie długości półosi elipsoidy prawdopodobnych korelacji $\boldsymbol{YZ}$: $\frac{1}{\sqrt{\lambda_i}}$, gdzie λ_i to *i*-ta wartość własna.

Przedział niepewności dla (4.79) przyjmuje więc postać:

$$\hat{\Sigma}_{YZ} - \frac{1}{\sqrt{\lambda_i}} \leq \Sigma_{YZ} \leq \hat{\Sigma}_{YZ} + \frac{1}{\sqrt{\lambda_i}} \quad (4.81)$$

Im węższe jest przedziały (4.76) lub (4.81), tym mniejsza jest niepewność.

Dla zmiennych jakościowych szacowanymi parametrami są liczebności tabeli kontyngencji $(\boldsymbol{Y}, \boldsymbol{Z})$**:**

$$\theta_{ijk} = P(X = i, Y = j, Z = k), i = 1, .., I; j = 1, \dots, J; k = 1, \dots, K. \quad (4.82)$$

Przy braku dostępnej informacji, niepewność parametru (4.16) jest opisana przez przedział:

[118] Jeżeli założenie o warunkowej niezależności jest prawdziwe, to $\rho_{YZ|X} = 0$.

$$0 \leq \theta_{ijk} \leq 1, \quad \sum_{i,j,k} \theta_{ijk} = 1. \tag{4.83}$$

Ze zbioru A i B można wynaczyć:

$$\hat{\theta}_{ij.} = \hat{P}(X = i, Y = j) = \hat{\theta}_{j|i}\hat{\theta}_{i..} = \frac{n_{A,ij}}{n_{A,i.}} \frac{n_{A,i.}+n_{B,i.}}{n_A+n_B}, \tag{4.84}$$

$$\hat{\theta}_{i.k} = \hat{P}(X = i, Z = k) = \hat{\theta}_{k|i}\hat{\theta}_{i..} = \frac{n_{B,ik}}{n_{B,i.}} \frac{n_{A,i.}+n_{B,i.}}{n_A+n_B}. \tag{4.85}$$

Informacja ta zawęża przedział niepewności do wszystkich rozkładów spełniających ograniczenia:

$$0 \leq \theta_{ijk} \leq 1, \tag{4.86}$$

$$\sum_{i,j,k} \theta_{ijk} = 1, \tag{4.87}$$

$$\sum_{k} \theta_{ijk} = \hat{\theta}_{ij.}, \tag{4.88}$$

$$\sum_{k} \theta_{ijk} = \hat{\theta}_{i.k}. \tag{4.89}$$

Przedział dla prawdopodobnych wartości rozkładu brzegowego $\boldsymbol{YZ}$ ($H(y,z)$) opisują tzw. granice Frecheta [Fréchet 1951]:

$$\max\{0; F(Y) + G(Z) - 1\} \leq H(y,z) \leq \min\{F(Y), G(Z)\}. \tag{4.90}$$

Dla zmiennych jakościowych empiryczne granice Frecheta można zapisać następująco:

$$\max\{0;\ \theta_{j.} + \theta_{.k} - 1\} \leq \theta_{jk} \leq \min\{\theta_{j.}; \theta_{.k}\}. \tag{4.91}$$

Wykorzystując informacje zawarte w wektorze $\boldsymbol{X}$ można wyznaczyć granice:

$$\sum_i \theta_{i..} \max\{0;\ \theta_{j|i} + \theta_{k|i} - 1\} \leq \theta_{.jk} \leq \sum_i \theta_{i..} \min\{\theta_{j|i}; \theta_{k|i}\}. \tag{4.92}$$

Parametr przy założeniu o warunkowej niezależności:

$$\theta_{ijk} = \theta_{j|i}\theta_{k|i}\theta_{i..} = \frac{\theta_{ij.}\theta_{i.k}}{\theta_{i..}} \tag{4.93}$$

oraz jego szacunek

$$\hat{\theta}_{ijk} = \frac{n^A_{ij.}}{n^A_{i..}} \frac{n^B_{i.k}}{n^B_{i..}} \frac{n^A_{i..}+n^B_{i..}}{n} \tag{4.94}$$

nie jest środkowym punktem przedziału niepewności, jest jednak w nim zawarty. Im węższy jest przedział (4.92), tym mniejsza jest niepewność.

Analiza niepewności winna być zastosowana na początku procesu integracji. W przypadku braku podstaw do wysunięcia przypuszczenia o warunkowej niezależności i braku dodatkowej informacji zewnętrznej, należy utworzyć przedziały niepewności dla szacowanych parametrów. Jeżeli są „akceptowalnie" wąskie, istnieją przesłanki do przeprowadzenia integracji metodą parowania statystycznego. W przeciwnym wypadku nie ma statystycznych podstaw do zastosowania integracji.

4.4.4. Ocena jakości integracji przy zastosowaniu parowania statystycznego

Ocena jakości integracji stanowi najważniejszy punkt analizy i nie jest zadaniem łatwym. Z formalnego punktu widzenia ewaluacja jakości połączenia powinna odbyć się poprzez oszacowanie błędu średniokwadratowego [D'Orazio *et al.* 2006]:

$$MSE(\hat{\theta}) = E\left((\hat{\theta} - \theta)^2\right), \tag{4.95}$$

Równanie (4.95) spełnia tożsamość:

$$MSE(\hat{\theta}) = var(\hat{\theta}) + \left(b(\hat{\theta})\right)^2, \tag{4.96}$$

gdzie $b(\hat{\theta}) = E(\hat{\theta}) - \theta$ to obciążenie estymatora.

W przypadku technik parowania statystycznego jakość i precyzja rezultatów zależą od dwóch składowych: jakości zbiorów danych wejściowych A i B (np. błędów losowych i nielosowych) oraz dokładności zastosowanej metody integracji. Przy założeniu, że jakość zbiorów wejściowych jest wysoka (np. dokonana została korekta błędów nielosowych – imputacja, kalibracja itp. oraz błąd losowy jest niewielki i kontrolowany), precyzja wyników integracji będzie zależeć głównie od „zdolności" zastosowanej techniki integracji do odtwarzania prawdziwego, nieznanego łącznego rozkładu cech $\boldsymbol{Y}$ i $\boldsymbol{Z}$ (σ_{YZ} dla metody makro lub jednostkowego zbioru danych mogącego być uznanym za próbę wylosowaną z prawdziwej populacji w przypadku metody mikro). W literaturze zasaniczo wyróżnia się cztery metody oceny jakości integracji określane następująco:

- metoda prosta (*simple measures*),
- szum integracyjny (*matching noise*),
- składana baza danych (*folded database*),
- „ważność" integracji (*validity evaluation*).

Barr i Turner [1981, 1990] oraz Rodgers [1984] zaproponowali prostą metodę oceny jakości integracji poprzez porównanie różnych charakterystyk rozkładu (np. średnich, odchyleń standardowych itp.) dołączanych cech w zbiorze zintegrowanym i zbiorach wejściowych. Dla nieparametrycznych metod mikro (typu *hot deck*) zaproponowano również porównanie relacji (np. współzależności) między $\boldsymbol{X}$ i $\boldsymbol{Z}$, jak również charakterystyk rozkładu $\boldsymbol{Z}$ w zbiorze zintegrowanym i zbiorze dawcy.

Innym sposobem oceny jakości, użytecznym zwłaszcza dla nieparametrycznych metod mikro typu *hot deck* przy założeniu o warunkowej niezależności jest tzw. szum integracyjny (*matching noise*, Paas [1985], D'Orazio *et al.* [2006]). Jest to „odległość" między prawdziwym nieznanym łącznym rozkładem $\boldsymbol{Z}$ przy danym $\boldsymbol{X}$ a rozkładem między imputowaną war-

tością $\widetilde{\boldsymbol{Z}}$ przy danym $\boldsymbol{X}$. Jeżeli oba te rozkłady są „podobne", imputowany zbiór biorcy jest reprezentatywny dla łącznego rozkładu $(\boldsymbol{X}, \boldsymbol{Z})$, przy założeniu warunkowej niezależności $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$. Szum integracyjny można zbadać posiadając dodatkowe źródło informacji o prawdziwym łącznym rozkładzie $(\boldsymbol{X}, \boldsymbol{Z})$ lub poprzez badania symulacyjne.

W pracy Marella *et al.* [2008] wykazano, że w metodzie *k* najbliższych sąsiadów szum integracyjny jest mniejszy niż w innych metodach *hot deck* i maleje wraz ze wzrostem liczebności próby dawcy.

Paas [1986] zaproponował metodę „składanej bazy danych" (*folded database*). Należy ona do metod symulacyjnych i polega na losowym podziale jednego z wejściowych zbiorów danych (zwykle bardziej licznego) na trzy podzbiory danych $\boldsymbol{G}'$, $\boldsymbol{G}''$ i $\boldsymbol{G}'''$ w taki sposób, że każdy z podzbiorów zawiera pewną liczbę zmiennych z wejściowego zbioru. Podzbiory dzielone są na dwie podpróbki A i B. Następnie z podpróbki A usuwany jest blok zmiennych $\boldsymbol{G}'''$, a z podpróbki B - $\boldsymbol{G}''$. Otrzymując w ten sposób sytuację analogiczną do tej w parowaniu statystycznym, podpróbki integrowane są zgodnie z okreslonym algorytmem. Otrzymane szacunki porównywane są ze zintegrowanym źródłem $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$ lub $(\boldsymbol{X}, \boldsymbol{Z})$. Należy jednak założyć, że zmienne $\boldsymbol{G}'$, $\boldsymbol{G}''$, $\boldsymbol{G}'''$ generowane są w sposób podobny do $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}$.

Raessler [2002] zaproponowała metodę oceny „ważności" otrzymanych wyników poprzez weryfikację czterech poziomów poprawności integracji, gdzie poziom pierwszy jest najtrudniejszy do weryfikacji, a poziom 4 najłatwiejszy:

— **Poziom 1**: **Reprodukcja nieznanych wartości $\boldsymbol{Z}$ w pliku biorcy** – prawdziwe, nieznane wartości wektora zmiennych Z w pliku biorcy są reprodukowane. Jeżeli w efekcie otrzymujemy prawdziwą wartość, sytuacje taką określa się „trafieniem" (*hit* - dla każdej jednostki zbioru biorcy). Można obliczyć „współczynnik trafień" (*hit ratio*).

Poziom ten jest najbardziej wymagający ze wszystkich. Ponieważ reprodukowane wartości są nieznane, współczynnik trafień może zostać obliczony wyłącznie za pomocą badań symulacyjnych. W ogólnym rozumieniu, dokładna reprodukcja wartości możliwa jest wtedy i tylko wtedy, gdy zmienne $\boldsymbol{X}$ w sposób deterministyczny wyjaśniają zmienność zmiennych $\boldsymbol{Z}$. W takim przypadku imputowana wartość z^B jest prawdziwa dla każdego $X = x$. Zwykle jednak taka sytuacja nie ma miejsca, zwłaszcza, gdy zmienne $\boldsymbol{Z}$ mają charakter ciągły oraz posiadają wielowymiarową strukturę. W rozkładzie ciągłym prawdopodobieństwo wylosowania określonej wartości wynosi zero, więc obliczanie liczby „trafień" jest bezcelowe. W przypadku rozkładu dyskretnego lub gdy zmienne $\boldsymbol{Z}$ mają charakter jakościowy można

obliczyć współczynnik trafień będący stosunkiem liczby prawidłowo imputowanych wartości do liczby imputacji ogółem. Należy jednak zwrócić uwagę, że współczynnik ten nie informuje, czy łączny rozkład został zachowany.

— **Poziom 2**: **Zachowanie łącznego rozkładu** – prawdziwy łączny rozkład zmiennych $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$ jest odzwierciedlony w zintegrowanym zbiorze.

Przy założeniu, że jednostki z obu zbiorów zostały wylosowane niezależnie, sparowany plik może zostać uznany jako próba losowa o łącznym rozkładzie $\tilde{f}_{\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z}}$. Najważniejszym celem parowania statystycznego jest wygenerowanie próby, która może zostać uznana jako prawdziwa próba wylosowana z rozkładu $f_{\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z}}$. Umożliwiałoby to przeprowadzanie analiz statystycznych na zintegrowanym („sparowanym") pliku. Jest to możliwe tylko wtedy, gdy zmienne dołączane $\boldsymbol{Y}$ oraz $\boldsymbol{Z}$ są warunkowo niezależne przy danym $\boldsymbol{X}$.

— **Poziom 3: Struktura korelacji** zmiennych jest zachowana w zintegrowanym pliku: $\widetilde{cov}(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}) = cov(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$. Poprawnie odwzorowane również są rozkłady brzegowe: $\tilde{f}_{\boldsymbol{XY}} = f_{\boldsymbol{X}}$ oraz $\tilde{f}_{\boldsymbol{XZ}} = f_{\boldsymbol{XZ}}$.

Czasami analityka interesują bardziej szczegółowe kwestie związane np. z zależnościami analizowanych zmiennych wyrażonymi poprzez macierz korelacji. W takim przypadku zintegrowany zbiór musi być rozpatrywany jako zbiór wygenerowany losowo ze sztucznej populacji, która charakteryzuje się, co najmniej, tymi samymi wartościami momentów oraz strukturą korelacji co populacja będąca obiektem badań. Zależność pomiędzy Y oraz Z wygenerowana przez proces parowania, może być mierzona jako kowariancja $cov(Y, Z) = E\big(cov(Y, Z|X)\big) + cov\big(E(Y|Z), E(Z|X)\big)$, a $\widetilde{cov}(Y, Z) = cov\big(E(Y|Z), E(Z|X)\big)$, jednak tylko $E(Y|X = x)$ i $E(Z|X = x)$ mogą być otrzymane z integrowanych zbiorów. Z tego wynika, że łączna kowariancja $\widetilde{cov}(Y, Z)$ równa jest prawdziwej kowariancji $cov(Y, Z)$ wtedy i tylko wtedy, gdy Y oraz Z są warunkowo nieskorelowane przy danym $X = x$, np. $E\big(cov(Y, Z|X)\big) = 0$. Należy również zwrócić uwagę, że zmienne, które są warunkowo niezależne są również warunkowo nieskorelowane, jednak nie zawsze występuje zależność odwrotna.

Wszystkie trzy powyższe poziomy mogą być sprawdzone wyłącznie poprzez przeprowadzenie badań symulacyjnych. Empiryczne przetestowanie ich nie jest możliwe.

— **Poziom 4**: Po przeprowadzeniu parowania statystycznego **brzegowy i łączny rozkład zmiennych** w pliku dawcy powinien zostać zachowany w zintegrowanym zbio-

rze. Wtedy należy się spodziewać, że spełnione zostaną równości $\tilde{f}_Z = f_Z$ oraz $\tilde{f}_{XZ} = f_{XZ}$ jeżeli $\boldsymbol{Z}$ jest imputowane do próby $(\boldsymbol{X}, \boldsymbol{Y})$.

Minimalnym wymaganiem dla procedury parowania statystycznego powinno być zachowanie rozkładu, który jest już obserwowany w integrowanych plikach. W praktyce rozkłady empiryczne zmiennych wspólnych $\boldsymbol{X}$, jak również zmiennych dołączanych $\boldsymbol{Z}$ w sparowanym pliku mogą być porównane z ich empirycznymi rozkładami w pliku dawcy w celu sprawdzenia ich zgodności. Rozkłady empiryczne $\hat{\tilde{f}}_X$ oraz $\hat{\tilde{f}}_{X,Z}$ nie powinny się różnić od $\hat{f}_X$ oraz $\hat{f}_{X,Z}$ dla więcej niż dwóch prób losowych wylosowanych z tej samej populacji. Można również zastosować do integrowanych plików wagi wynikające ze schematu ich losowania.

Zgodność rozkładu zmiennych Z oraz rozkładów łącznych X-Z może zostać w prosty sposób obliczona za pomocą np. testu zgodności χ^2 lub testu Kołmogorowa-Smirnowa. Dla bardzo dużych prób, kiedy nawet niewielkie różnice prowadzą do odrzucenia hipotezy zerowej, warto stosować inne miary podobieństwa, np. (4.1), (4.2), (4.3) itp.

Niemieckie Stowarzyszenie Analiz Medialnych[119] wystosowało postulaty dotyczące kontroli jakości zintegrowanych repozytoriów [Raessler 2002]:

- najpierw porównywane są empiryczne rozkłady zmiennych wspólnych X w pliku dawcy i biorcy w celu ocenyzgodności,
- następnie porównywany jest rozkład empiryczny dołączonych zmiennych Z w pliku biorcy i dawcy,
- w ostatnim etapie porównuje się łączny rozkład $f_{X,Z}$ obserwowany w pliku dawcy z rozkładem łącznym $\tilde{f}_{X,Z}$ obserwowanym w pliku zintegrowanym.

4.5. Wnioski

Wielość podejść metodologicznych do statystycznej integracji danych wymaga starannej analizy integrowanych zbiorów pod kątem zawartości informacyjnej, pokrycia, definicji populacji i jednostek, a także wykorzystanych skal pomiarowych i rozkładów zmiennych – zarówno jedno-, jak i wielowymiarowych. Na każdym etapie statystycznej integracji danych uwzględnia się jakość danych wejściowych w celu zapewnienia jak najwyższej jakości zbioru zintegrowanego.

W następnym rozdziale przedstawione zostanie badanie empiryczne będące koncepcją utworzenia modułu zintegrowanego repozytorium danych społeczno-ekonomicznych. Wykorzy-

[119] *Arbeitsgemeinschaft Media Analyse (AG.MA)*

stane zostaną wybrane metody integracji. Wyniki posłużą weryfikacji postawionych hipotez badawczych.

Jako literackie podsumowanie metod statystycznej integracji danych, przed przejściem do ich empirycznego zastosowania, można przytoczyć cytat Ivana Fellegi'ego[120] (1977):
„*W sytuacji, gdy badacze społeczni tak chciwie pragną bogatych w informacje zbiorów danych, parowanie statystyczne może wydawać się ogromnie atrakcyjną procedurą tworzenia zbiorów zawierających logiczne powiązania zmiennych znajdujących się w oddzielnych źródłach […]. Chciałbym najpierw zobaczyć rzetelną ocenę takich łącznych rozkładów zanim zdjąłbym z procedury tabliczkę: „**UWAGA! NIEBEZPIECZEŃSTWO! STOSOWAĆ Z ZACHOWANIEM OSTROŻNOŚCI!***".

[120] Ivan Peter Fellegi (ur. 1935) – kanadyjski statystyk węgierskiego pochodzenia. Razem z Alanem Sunterem w 1969 opracował model probabilistycznego łączenia rekordów, który został opublikowany w artykule "A Theory for Record Linkage" [1969].

ROZDZIAŁ V. KONSTRUKCJA ZINTEGROWANEGO REPOZYTORIUM DANYCH SPOŁECZNYCH

5.1. Koncepcja badania empirycznego

Dla weryfikacji przydatności metod statystycznej integracji danych podjęto próbę konstrukcji zintegrowanego zbioru danych społecznych dotyczącego charakterystyk gospodarstw domowych. Zasadniczym celem było wykazanie, że metody parowania statystycznego umożliwiają dobre jakościowo odwzorowanie rozkładów brzegowych i łącznych zmiennych dołączanych w zintegrowanym zbiorze danych. Utworzony zbiór może stanowić podstawę dalszych badań właściwości metod integracji, a także możliwości ich wykorzystania w praktyce urzędów statystycznych. Jako cele szczegółowe badania empirycznego określono:

- wykazanie, że przedstawione w rozdziale IV metody parowania statystycznego są lepsze od integracji losowej,
- weryfikację jakości integracji w zależności od relacji między wielkością (liczebnością) zbiorów dawcy i biorcy,
- udowodnienie, że szacunki na podstawie zintegrowanego zbioru są lepsze w sensie precyzji i dokładności, niż w zbiorach wejściowych,
- ocenę możliwości otrzymania na podstawie zintegrowanego repozytorium danych szacunków o zwiększonej precyzji oraz szerszym zakresie merytorycznym.

Ze względu na dostępność danych, jak również zawartość merytoryczną, badanie empiryczne przeprowadzone zostało przy wykorzystaniu zbiorów Badania Budżetów Gospodarstw Domowych (BBGD) z 2005 roku[121] oraz Badania Dochodów i Warunków Życia (EU-SILC) z 2006 roku[122]. Integracja obejmuje informacje dotyczące gospodarstw domowych.

Badanie Budżetów Gospodarstw Domowych i Badanie Dochodów i Jakości Życia w dużej mierze charakteryzują się podobnym zakresem merytorycznym i metodą przeprowadzenia.

[121] Zbiór danych został zakupiony przez Katedrę Statystyki na Uniwersytecie Ekonomicznym w Poznaniu na potrzeby badań naukowych.

[122] Wybrano zbiór z roku 2006 ze względu na fakt, że okres referencyjny głównych cech dotyczących dochodów gospodarstw domowych w badaniu EU-SILC ustalony został na rok poprzedzający badanie. Założono, że pozostałe cechy, dotyczące wyposażenia gospodarstwa, warunków życia i cech demograficzno-społecznych charakteryzują się mniejszą zmiennością niż kategorie finansowe. W ten sposób starano się zachować zgodność cech wspólnych EU-SILC z BBGD.

Zbiór EU-SILC został udostępniony Autorowi w ramach prac w projekcie *WORKABLE - Making Capabilities Work (2009-2012), 7th Framework Programme, large-scale collaborative project, European Commission* realizowanym przez Centrum Studiów nad Polityką Społeczną działającym przy Uniwersytecie im. Adama Mickiewicza w Poznaniu.

Pomiarowi w obu badaniach poddane są gospodarstwa domowe w Polsce. Ich definicje w obu przypadkach są zbieżne, a zbliżona liczebność populacji generalnej potwierdza tę tezę (por. tabela 5.1). Oba badania mają charakter częściowy i przeprowadzone zostały metodą reprezentacyjną o podobnym schemacie doboru jednostek do próby. Brak próby rezerwowej w EU-SILC spowodował jednak, że liczba wypełnionych kwestionariuszy była niższa od zakładanej[123].

Tabela 5.1. Podstawowa charakterystyka BBGD i EU-SILC

Charakterystyka	**Badanie Budżetów Gospodarstw Domowych**	**Badanie Dochodów i Warunków Życia**
Czas realizacji	cały rok 2005	2 maja – 19 czerwca 2006
Zbiorowość badania	gospodarstwa domowe w Polsce	gospodarstwa domowe w Polsce
Definicja jednostki	Zespół osób mieszkających razem i wspólnie utrzymujących się (gospodarstwo domowe wieloosobowe) lub osoba utrzymującą się samodzielnie, tzn. niełącząca swoich dochodów z innymi osobami, bez względu na to, czy mieszka sama, czy też z innymi osobami (gospodarstwo jednoosobowe).	Zespół osób spokrewnionych ze sobą lub niespokrewnionych, mieszkających razem i wspólnie utrzymujących się (gospodarstwo domowe wieloosobowe) lub osoba utrzymująca się samodzielnie, bez względu na to, czy mieszka sama, czy też z innymi osobami (gospodarstwo domowe jednoosobowe).
Metoda doboru próby	reprezentacyjna	reprezentacyjna
Schemat losowania	dwustopniowy, warstwowy	dwustopniowy, warstwowy
Próba rezerwowa	TAK	NIE
Przedmiot badania	— budżet gospodarstwa domowego (zestawienie różnych źródeł dochodów i wydatków) — wyposażenie gospodarstwa domowego — wielkość spożycia produktów i usług	— sytuacja dochodowa — wyposażenie gospodarstwa domowego — ubóstwo — różne aspekty warunków życia ludności
Zakładana liczebność populacji generalnej (suma wag analitycznych)	13 332 605	13 300 839
Wielkość próby (roczna)	34 767	14 914 (zakładana próba 18 494)

Źródło: opracowanie własne

[123] W BBGD w przypadku odmowy odpowiedzi dolosowane zostały gospodarstwa z próby rezerwowej.

W obu badaniach pomiarowi poddano wiele takich samych lub zbliżonych charakterystyk. Wymienić tutaj można: typ gospodarstwa domowego, wyposażenie gospodarstw domowych w niektóre dobra trwałego użytku czy niektóre charakterystyki zajmowanych mieszkań oraz dochody gospodarstw domowych w ujęciu szczegółowych kategorii.

Jednak każde z badań posiada własną specyfikę. Badanie Budżetów ukazuje przede wszystkim materialne warunki bytu gospodarstw domowych w ujęciu obiektywnym. Natomiast EU-SILC w większym stopniu uwzględnia wymiar subiektywny np. ocenę stanu zdrowia, sytuacji finansowej czy poczucia bezpieczeństwa. Tak więc połączenie zbiorów danych obu badań umożliwia pełny opis warunków życia gospodarstw domowych uwzględniający zarówno ocenę obiektywną jak i subiektywną. Poniżej przedstawiona zostanie charakterystyka zakresu merytorycznego obu badań, w wyniku której wskazane zostaną zmienne integrowane w eksperymentalnej konstrukcji modułu macierzy rachunków społecznych dotyczących gospodarstw domowych.

Zbiór danych jednostkowych BBGD pochodził z roku 2005. Repozytorium danych składało się z siedmiu tabel zapisanych w formacie .dbf. Dwie z nich zawierały informacje na temat członków gospodarstw domowych, natomiast pozostałe charakteryzowały gospodarstwa domowe (por. tabela 5.2).

Tabela 5.2. Zawartość tabel zbioru danych BBGD 2005

Nazwa tabeli	Liczba rekordów	Rekord / jednostka badania
BR2005_01a_1	34 767	gospodarstwo domowe
BR2005_01a_4	107 124	osoba
BR2005_01a_5	86 371	osoba
BR2005_04	34 767	gospodarstwo domowe
BR2005_Przychody	142 066	kategoria przychodu (skumulowane dane dla gospodarstwa)
BR2005_Rozchody	2 446 005	kategoria rozchodu (skumulowane dane dla gospodarstwa)
BR2005_Kategorie	34 767	gospodarstwo domowe

Źródło: opracowanie własne na podstawie opisu zmiennych zawartym w pliku *BR2005-Struktury*

Tabela *BR2005_01a_1*, sporządzona na podstawie informacji z kwestionariusza BR01a (*karta statystyczna gospodarstwa domowego*) zawierała informacje na temat warunków mieszkaniowych, wyposażenia w podstawowe urządzenia AGD. W tabeli *BR2005_04*, utworzonej na podstawie informacji z kwestionariusza BR04 (*informacje uzupełniają-*

ce o gospodarstwie domowym), znalazły się cechy opisujące wyposażenie gospodarstwa domowego w urządzenia RTV, AGD, dobra luksusowe oraz inną własność gospodarstwa (m.in. garaż, dom letniskowy, działka rekreacyjna).

Tabela *BR2005_Przychody* zawiera informacje o przychodach gospodarstw domowych podzielonych na 70 szczegółowych kategorii ujętych w 11 grupach:

- dochody z pracy najemnej,
- dochody z pracy na własny rachunek poza gospodarstwem rolnym w użytkowaniu indywidualnym,
- dochody z własności,
- dochody z wynajmu nieruchomości,
- świadczenia z ubezpieczeń społecznych,
- świadczenia z pomocy społecznej,
- inne dochody,
- sprzedaż użytkowanych artykułów konsumpcyjnych,
- sprzedaż majątku rzeczowego (niezwiązanego z działalnością gospodarczą),
- przychody finansowe,
- przychody z tytułu prowadzenia gospodarstwa rolnego.

Rekordy w tabeli *BR2005_Przychody* dotyczą poszczególnych kategorii przychodu dla każdego gospodarstwa (w formie pieniężnej i niepieniężnej). Z tego powodu jedno gospodarstwo opisane jest za pomocą wielu rekordów.

Tabela danych *BR2005_Rozchody* opisuje rozchody (wydatki) gospodarstw domowych podzielonych na 391 szczegółowych kategorii, zagregowanych w 60 podkategorii i 18 następujących kategorii głównych:

- żywność i napoje bezalkoholowe,
- napoje alkoholowe, wyroby tytoniowe i narkotyki,
- odzież i obuwie,
- użytkowanie mieszkania lub domu i nośniki energii,
- wyposażenie mieszkania i prowadzenie gospodarstwa domowego,
- zdrowie,
- transport,
- łączność,
- rekreacja i kultura,
- edukacja,

— restauracje i hotele,
— pozostałe wydatki na towary i usługi,
— pozostałe wydatki,
— podatki i inne opłaty,
— rozchody kapitałowe (rzeczowe),
— rozchody finansowe,
— rozchody bieżące związane z gospodarstwem rolnym,
— rozchody inwestycyjne na gospodarstwo rolne.

Podobnie jak w przypadku przychodów, rekordy w tabeli danych odnoszą się do każdej szczegółowej kategorii rozchodów dla każdego gospodarstwa.

Z pięciu tabel opisujących gospodarstwa domowe utworzono kompleksowy, jednostkowy zbiór danych łącząc je na podstawie unikalnego identyfikatora gospodarstwa domowego. Dla kategorii dochodów i rozchodów dokonano agregacji poszczególnych rekordów w taki sposób, by jeden rekord zawierał informacje dla jednego gospodarstwa. Zbiór danych EU-SILC 2006 składała się z czterech tabel danych (por. tabela 5.3). Dwie zawierają informacje dla osób (zbiór UDB_c06R i UDB_c06P), a dwie dla gospodarstw domowych (UDB_c06D i UDB_c06H).

Tabela 5.3. Zawartość tabel zbioru danych EU-SILC 2006

Nazwa tabeli	Liczba rekordów	Rekord / jednostka badania
UDB_c06D	14 914	gospodarstwo domowe
UDB_c06H	14 914	gospodarstwo domowe
UDB_c06R	45 122	członek gospodarstwa domowego/osoba (bez względu na wiek)
UDB_c06P	36 589	członek gospodarstwa domowego/osoba w wieku 16 lat i więcej

Źródło: opracowanie własne

Tabele osób podzielone zostały na rejestry osobowe (R) oraz zbiór danych o osobach (P, por. tabela 5.3). W pliku rejestru zawarte zostały podstawowe informacje o wszystkich członkach gospodarstwa domowego (bez względu na wiek), takie jak rok urodzenia, wiek, płeć, ID rodziców[124], rodzeństwa, małżonka lub partnera, status zamieszkania, aktywności ekonomicznej oraz edukacyjny. Natomiast zbiór danych o osobach dotyczy członków gospodarstwa domowego w wieku 16 lat i więcej. Zawiera on takie cechy jak: stan cywilny,

[124] Jeżeli w gospodarstwie zamieszkiwali rodzice respondenta.

obywatelstwo, poziom wykształcenia, subiektywne opinie o stanie zdrowia, przebytych chorobach, historię aktywności zawodowej w roku poprzedzającym badanie oraz wykonywany zawód (zgodnie z klasyfikacją ISCO-88[125]), informacje o barierach w dostępie do zawodu, ochrony zdrowia, a także o dochodach członków gospodarstw domowych w rozbiciu na 12 następujących kategorii zbiorczych (brutto i netto):

- pieniężne dochody z pracy najemnej,
- niepieniężne dochody z pracy najemnej,
- dochody z pracy na własny rachunek,
- wartość dóbr wytworzonych przez członka gospodarstwa na potrzeby własne,
- dochody z indywidualnych planów oszczędnościowych,
- świadczenia dla bezrobotnych,
- świadczenia związane z wiekiem,
- renty rodzinne,
- świadczenia chorobowe (w tym odszkodowania z tytułu uszczerbku na zdrowiu),
- świadczenia dla niepełnosprawnych,
- stypendia.

Tabele dla gospodarstw domowych również podzielono na rejestr i zbiór danych. Rejestr dla gospodarstw domowych zawierał informacje o lokalizacji gospodarstwa[126] oraz schemacie doboru jednostek do próby (waga pierwszego stopnia losowania, waga drugiego stopnia oraz waga finalna). Zbiór danych gospodarstw domowych zawierał szczegółowe charakterystyki gospodarstw domowych dotyczące warunków zamieszkania (m.in. wyposażenie gospodarstwa w niektóre dobra codziennego użytku), warunków życia (subiektywna ocena sytuacji finansowej, ocena sąsiedztwa gospodarstwa) oraz dochody nieujęte w zbiorze danych o członkach gospodarstwa (osiągane przez gospodarstwo jako całość), takie jak: całkowity dochód brutto, całkowity dochód rozporządzalny brutto oraz ekwiwalentny dochód do dyspozycji, a także poszczególne kategorie przychodów gospodarstwa ujęte w 12 następujących kategorii (brutto i netto):

[125] ISCO (*International Standard Classification of Occupations*) - klasyfikacja zawodów i specjalności dla potrzeb rynku pracy. Klasyfikacja została opracowana na podstawie Międzynarodowego Standardu Klasyfikacji Zawodów ISCO-88, przyjętego na XIV Międzynarodowej Konferencji Statystyków Pracy w Genewie w 1987 r. oraz jej nowej edycji z 1994 r., tzw. ISCO-88 (COM), dostosowanej do potrzeb Unii Europejskiej.

[126] Informacje te zostały mocno ograniczone. Dostępne są jedynie informacje o makroregionie (poziom NUTS 1), w którym znajduje się gospodarstwo oraz gęstości zaludnienia terytorium. Informacje o takich podstawowych charakterystykach gospodarstwa jak klasa miejscowości zamieszkania zostały usunięte z udostępnionego zbioru.

— czynsz przypisany,
— dochód z wynajmu własności lub gruntu,
— świadczenia dotyczące rodziny (w tym zasiłki rodzinne z dodatkami, zasiłki macierzyńskie),
— świadczenia dotyczące wykluczenia społecznego (w tym świadczenia z pomocy społecznej),
— dodatki mieszkaniowe,
— regularne transfery otrzymywane od osób spoza gospodarstwa domowego,
— dochody kapitałowe (z własności finansowej),
— dochody związane z posiadaniem hipoteki,
— dochody dzieci do lat 16,
— stałe podatki majątkowe,
— regularne transfery otrzymywane od osób z gospodarstwa domowego,
— zwroty podatku dochodowego.

Dochody członków gospodarstw domowych oraz gospodarstw mierzone są rocznie dla okresu referencyjnego od 1 stycznia do 31 grudnia roku poprzedzającego badanie. Ze względu na zachowanie porównywalności międzynarodowej, dochody mierzone są w Euro, jednak istnieje możliwość ich przeliczenia na złote[127]. Zbiór danych zawierał również zmienne zamieszczone w dołączonym do badania module – uczestnictwa w życiu społecznym (uczęszczanie do kina i teatru, częstotliwość kontaktów i spotkań z rodziną i przyjaciółmi, możliwość otrzymania pomocy od rodziny i przyjaciół, uczestnictwo w wolontariacie, partiach politycznych, związkach zawodowych, organizacjach zawodowych, organizacjach religijnych itp.).

Tabele danych (rejestru i zbioru) dla gospodarstw domowych zintegrowano w sposób deterministyczny na podstawie unikalnego klucza połączeniowego (innego niż w przypadku BBGD). Dodatkowo z tabeli osób dołączono informacje o zagregowanych dla gospodarstw domowych dochodach ich członków .

Przeprowadzone badanie infrastruktury statystycznej obu integrowanych zbiorów umożliwiło sformułowanie następujących hipotez, których weryfikacja pozwoli zrealizować określony na wstępie cel badania empirycznego.

[127] Przeliczenia można dokonać poprzez przemnożenie dochodów w Euro przez zmienną zawierającą średnioroczny kurs złotówki do Euro - zmienna HX010. W 2006 roku przeliczono wartość Euro na 4,023 zł.

1. Precyzja szacunków na poziomie NUTS 1, w zintegrowanym zbiorze jest większa dla zmiennych wspólnych i dołączanych.
2. W zintegrowanym zbiorze akceptowalna będzie precyzja szacunków na poziomie NUTS 2.
3. Merytoryczny zakres szacunków na podstawie zbioru zintegrowanego można rozszerzyć między innymi o:
    a. oszacowanie nieznanego współczynnika korelacji między wydatkami gospodarstw domowych i dochodami głów gospodarstw domowych,
    b. konstrukcję tabeli kontyngencji ukazującej możliwości sfinansowania tygodniowego urlopu poza miejscem zamieszkania w przekroju województw,
    c. oszacowanie wydatków gospodarstw domowych według możliwości sfinansowania tygodniowego urlopu poza miejscem zamieszkania w przekroju województw,
    d. oszacowanie dochodów głów gospodarstw domowych w przekroju województw.

Precyzja oszacowania zmierzona zostanie błędem standardowym oszacowania przed i po integracji dla najlepszej jakościowo metody integracji.

5.2. Wybór zmiennych dołączanych

Zakres tematyczny Badania Budżetów Gospodarstw Domowych i Badania Dochodów i Warunków życia wydaje się podobny. Zarówno w jednym, jak i drugim badaniu pomiarowi podlegają takie grupy tematyczne cech jak:

- wyposażenie gospodarstwa domowego,
- subiektywna ocena stanu finansowego gospodarstwa,
- informacje o rodzinie,
- informacje o rodzaju i wielkości zajmowanego lokalu,
- dochody.

Wśród ważniejszych zakresów tematycznych omawianych oddzielnie w każdym z badań wymienić można:

- BBGD:
    - wydatki gospodarstw domowych,
    - niektóre rodzaje wyposażenia gospodarstwa,
    - grupa i podgrupa społeczno-ekonomiczna gospodarstwa,

— główne i dodatkowe źródło utrzymania.

— EU-SILC

— warunki życia,

— dochody poszczególnych członków gospodarstwa domowego.

Na potrzeby badania empirycznego postanowiono dołączyć:

1. wariant 1 – zmienne ilościowe
    a. do zbioru EU-SILC – wydatki ogółem gospodarstwa domowego,
    b. do zbioru BBGD – dochody głowy gospodarstwa domowego[128].
2. wariant 2 – cechy jakościowe
    a. do zbioru EU-SILC – identyfikator województwa,
    b. do zbioru BBGD – czy stać na tygodniowy urlop rocznie poza domem.

Dołączenie do zbioru EU-SILC wydatków gospodarstw umożliwi ich łączną obserwację z dochodami głów gospodarstw domowych. Dołączenie identyfikatora województwa umożliwi z kolei oszacowanie jakości estymatorów dla zmiennej dochody głowy gospodarstwa domowego na poziomie NUTS1. Zmienna „czy stać na tygodniowy urlop rocznie poza domem" zostanie dołączona jako przykład cechy opisującej warunki życia ludności.

W przypadku zmiennych dołączanych wydatki ogółem gospodarstwa domowego i identyfikator województwa, zbiorem dawcy będzie BBGD, a EU-SILC będzie biorcą. Dla zmiennych dołączanych „dochody głowy gospodarstwa domowego" i „czy gospodarstwo stać na tygodniowy urlop rocznie poza domem" dawcą będzie EU-SILC, zaś BBGD biorcą. W tej sytuacji dołączane będą rekordy ze zbioru mniejszego do większego. W literaturze taka okoliczność uznawana jest za niepożądaną ze względu na możliwość dołączania wariantów cech dawcy wielokrotnie do zbioru biorcy. W niniejszym badaniu zweryfikowana zostanie możliwość rozwiązania tego problemu.

5.3. Detekcja zmiennych wspólnych i parujących

Wśród cech znajdujących się w obu zbiorach danych, jako zmienne wspólne (charakteryzujące się podobnymi definicjami i wariantami) zidentyfikowano następujące:

— rodzaj nieruchomości,

— tytuł prawny do zajmowanego lokalu,

— liczba pokoi do dyspozycji gospodarstwa domowego,

[128] Ponieważ w zbiorze EU-SILC głowa gospodarstwa domowego nie jest wyszczególniona, za głowę uznano osobę o najwyższym dochodzie. Zmienna zostanie dołączona jako przykład dochodów indywidualnych członków gospodarstw domowych (nie są one poddane pomiarowi w BBGD).

— wanna lub prysznic w mieszkaniu,
— ustęp spłukiwany dla wyłącznego użytku domowego,
— możliwość wiązania końca z końcem,
— czy w gospodarstwie jest telewizor kolorowy,
— czy w gospodarstwie jest komputer,
— czy w gospodarstwie jest telefon,
— czy w gospodarstwie jest pralka automatyczna,
— czy w gospodarstwie jest pralka automatyczna,
— czy gospodarstwo posiada samochód,
— typ gospodarstwa domowego,
— wielkość gospodarstwa domowego,
— ekwiwalentna wielkość gospodarstwa domowego,
— całkowity rozporządzalny dochód gospodarstwa domowego,
— ekwiwalentny dochód do dyspozycji,
— region (NUTS 1).

Braki danych w zmiennych wspólnych występowały sporadycznie jedynie w zbiorze EU-SILC. Zmienne dotknięte problemem braków danych to:

— HH010 – typ budynku (n=45; 0,3%),
— HH030 – liczba pokoi do użytkowania gospodarstwa (n=23; 0,2%),

Ze względu na ich niewielką liczbę, brakujące wartości zaimputowano w sposób losowy. Warianty zmiennych wspólnych zharmonizowano poprzez ich agregację (por. załącznik 1). W kolejnym kroku sprawdzono zgodność rozkładów cech wspólnych w obu zbiorach za pomocą współczynnika zgodności Δ (wzór 4.1, por. tabela 5.4 i 5.5).

Tabela 5.4. Porównanie rozkładów jakościowych zmiennych wspólnych po harmonizacji

Zmienna	**Wariant**	**Zbiór danych**		Δ
		BBGD	**EU-SILC**	
Region	region centralny	21,43	21,82	0,83
	region południowy	21,27	21,11	
	region wschodni	16,77	16,46	
	region północno-zachodni	15,22	15,38	
	region południowo-zachodni	10,97	10,60	
	region północny	14,34	14,62	
Rodzaj budynku	budynek wielorodzinny	59,21	55,96	4,03
	dom jednorodzinny	35,11	39,14	
	dom jednorodzinny w zabudowie szeregowej	5,36	4,61	

Zmienna	Wariant	Zbiór danych		Δ
		BBGD	EU-SILC	
	inne	0,32	0,29	
Tytuł prawny do zajmowanego mieszkania	własność	53,73	55,70	3,08
	najem wg cen rynkowych	2,68	3,79	
	najem poniżej cen rynkowych	1,33	1,27	
	inne	42,25	39,23	
Typ biologiczny gospodarstwa domowego	małżeństwo bez dzieci	16,89	21,22	5,82
	małżeństwo z 1 dzieckiem na utrzymaniu	11,27	11,33	
	małżeństwo z 2 dzieci na utrzymaniu	11,63	10,41	
	małżeństwo z 3 więcej dzieci na utrzymaniu	5,50	3,80	
	samotny rodzic z co najmniej jedną osobą na utrzymaniu	5,62	2,97	
	gospodarstwa jednoosobowe	24,80	24,70	
	inne gospodarstwa z osobami na utrzymaniu	9,44	10,87	
	pozostałe gospodarstwa	14,85	14,70	
Liczba pokoi	1	15,93	13,41	4,03
	2	36,56	35,34	
	3	29,43	29,14	
	4	9,86	11,07	
	5	4,98	6,19	
	6 i więcej	3,24	4,85	
Czy jest ustęp spłukiwany?	tak	90,15	91,20	1,05
	nie	9,85	8,80	
Czy jest łazienka?	tak	90,04	89,98	0,06
	nie	9,96	10,02	
Czy dochody pozwalają na wiązanie końca z końcem?	z wielką trudnością	18,44	21,01	6,85
	z trudnością	22,94	25,71	
	z pewną trudnością	39,00	32,34	
	raczej łatwo	15,06	14,86	
	łatwo	3,76	4,92	
	bardzo łatwo	0,80	1,16	
Czy gospodarstwo posiada TV?	tak	98,16	96,87	1,29
	nie	1,84	3,13	
Czy gospodarstwo posiada komputer?	tak	38,57	44,37	5,80
	nie	61,43	55,63	
Czy gospodarstwo posiada telefon?	tak	65,18	92,39	27,21
	nie	34,82	7,61	
Czy gospodarstwo posiada pralkę?	tak	96,52	96,23	0,29
	nie	3,48	3,77	
Czy gospodarstwo posiada samochód?	tak	47,35	50,66	3,32
	nie	52,65	49,34	

Źródło: opracowanie własne

Wśród jakościowych zmiennych wspólnych nieakceptowalnie wysokimi różnicami frakcji ($\Delta \leq 6\%$) w analizowanych zbiorach charakteryzowały się: (i) czy dochody pozwalają na

związanie końca z końcem, (ii) czy gospodarstwo posiada telefon (por. tabela 5.4). Pierwsza z nich, po dalszej agregacji[129] umożliwiającej zwiększenie zgodności rozkładu, została użyte jako zmienna grupująca w metodach nieparametrycznych. Drugą zmienną, ze względu na różnicę definicyjną (w BBGD uwzględniono wyłącznie telefony komórkowe, w EU-SILC zarówno stacjonarne, jak i komórkowe), pominięto w procesie integracji.

Tabela 5.5. Porównanie rozkładów ilościowych zmiennych wspólnych po harmonizacji

Zmienna	**Statystyka**	**Zbiór danych**		**Wskaźnik**
		BBGD	EU-SILC	
Liczba osób w gospodarstwie domowym	Średnia	2,83	2,84	0,998
	Mediana	3,00	3,00	--
	Wariancja	2,561	2,597	0,986
	Odchylenie standardowe	1,600	1,612	0,993
	Minimum	1	1	--
	Maksimum	14	14	--
	Rozstęp międzykwar.	2	2	--
	Skośność	,914	,953	--
Ekwiwalentna wielkość gospodarstwa domowego	Średnia	1,825	1,833	0,996
	Mediana	1,800	1,800	--
	Wariancja	,492	,514	0,957
	Odchylenie standardowe	,7013	,7169	0,978
	Minimum	1,0	1,0	--
	Maksimum	6,7	6,9	--
	Rozstęp międzykwar.	1,0	1,0	--
	Skośność	,833	,918	--
Dochód rozporządzalny gospodarstwa	Średnia	2155,6894	2286,2650	0,943
	Mediana	1800,0000	1833,7505	--
	Wariancja	3451099,871	3266838,180	1,056
	Odchylenie standardowe	1857,71361	1807,43968	1,028
	Minimum	-83648,13	0,00	--
	Maksimum	85344,40	23458,33	--
	Rozstęp międzykwar.	1514,50	1846,00	--
	Skośność	2,720	2,532	--
Ekwiwalentny dochód gospodarstwa domowego	Średnia	1222,7919	1288,6312	0,949
	Mediana	1033,3333	1073,2917	--
	Wariancja	991775,549	828935,213	1,196
	Odchylenie standardowe	995,87928	910,45879	1,094
	Minimum	-46471,18	-238,06	--
	Maksimum	42100,00	13411,11	--
	Rozstęp międzykwar.	718,65	815,28	--
	Skośność	4,670	3,043	--

Źródło: opracowanie własne

[129] Agregacja umożliwi zwiększenie zgodności rozkładów zmiennych, ale spowoduje znaczną utratę informacji. Stąd decyzja o wykorzystaniu ich w grupowaniu, nie zaś w integracji.

Wśród ilościowych zmiennych wspólnych wszystkie charakteryzowały się zgodnością na poziomie co najmniej 94% (zarówno dla średniej, jak i miar dyspersji; por. tabela 5.5). Uznano więc, że wszystkie zmienne mogą zostać wykorzystane w procesie integracji.
Ostatecznie lista zmiennych wspólnych w procesie integracji jest następująca:

- region,
- rodzaj budynku,
- tytuł prawny do zajmowanego mieszkania,
- typ biologiczny gospodarstwa domowego,
- liczba pokoi,
- czy jest ustęp spłukiwany,
- czy jest łazienka,
- czy gospodarstwo posiada tv,
- czy gospodarstwo posiada komputer,
- czy gospodarstwo posiada pralkę,
- czy gospodarstwo posiada samochód,
- liczba osób w gospodarstwie domowym,
- ekwiwalentna wielkość gospodarstwa domowego,
- dochód rozporządzalny gospodarstwa,
- ekwiwalentny dochód gospodarstwa domowego.

Określony zestaw zmiennych wspólnych stanowił podstawę wyboru zmiennych parujących dla każdej z dołączanych cech. W celu wyboru zmiennych parujących wykorzystano metodę drzewa klasyfikacyjnego i regresyjnego CART (pkt. 4.4.1). Rolę zmiennej objaśnianej każdorazowo przypisano zmiennej dołączanej, natomiast zmienne wspólne stanowiły potencjalny zestaw zmiennych parujących. W wyniku zastosowania powyższej procedury określono następujące zmiennych parujących:

1. Dla cechy - wydatki gospodarstw domowych:
   - czy jest łazienka,
   - czy jest ustęp spłukiwany,
   - czy gospodarstwo posiada samochód,
   - liczba pokoi,
   - rodzaj budynku,
   - ekwiwalentny dochód do dyspozycji ,

— dochód do dyspozycji ,
— wielkość gospodarstwa domowego;

2. Dla cechy - dochody głów gospodarstw domowych:
   — czy jest ustęp spłukiwany,
   — czy gospodarstwo posiada pralkę,
   — czy gospodarstwo posiada samochód,
   — czy gospodarstwo posiada tv,
   — liczba pokoi,
   — rodzaj budynku,
   — tytuł prawny do zajmowanego mieszkania,
   — ekwiwalentny dochód do dyspozycji,
   — dochód do dyspozycji,
   — wielkość gospodarstwa domowego;
3. Dla cechy - identyfikator województwa:
   — Liczba pokoi
   — Rodzaj budynku
   — Tytuł prawny do zajmowanego mieszkania
   — ekwiwalentny dochód do dyspozycji ,
   — dochód do dyspozycji ,
   — wielkość gospodarstwa domowego,
   — ekwiwalentna wielkość gospodarstwa domowego;
4. Dla cechy - czy stać na tygodniowy urlop rocznie poza domem:
   — czy jest łazienka,
   — czy gospodarstwo posiada samochód,
   — liczba pokoi,
   — rodzaj budynku,
   — tytuł prawny do zajmowanego mieszkania,
   — ekwiwalentny dochód do dyspozycji,
   — dochód do dyspozycji,
   — ekwiwalentna wielkość gospodarstwa domowego.

Wybrane zmienne parujące, zgodnie z kryterium przyjętym w metodzie CART, były najsilniej skorelowane z odpowiednimi zmiennymi dołączanymi.

5.4. Metoda integracji

Ze względu na brak dodatkowych informacji o łącznych charakterystykach integrowanych cech, w wyniku analizy niepewności, zostanie przyjęte założenie o warunkowej niezależności (CIA). W kolejnym kroku przeprowadzona zostanie integracja mikro. Wykorzystane zostaną następujące metody nieparametryczne, parametryczne i mieszane:

— **nieparametryczne:**

1. losowa - losowy dobór wartości zmiennej Z ze zbioru B (dawcy) do zbioru A (biorcy) oraz losowy dobór zmiennej Y ze zbioru A (dawcy) do zbioru B (biorcy),
2. najbliższego sąsiedztwa - dobór zmiennej Z ze zbioru B (dawcy) do zbioru A (biorcy) oraz dobór zmiennej Y ze zbioru A (dawcy) do zbioru B (biorcy);

— **parametryczne – stochastyczna imputacja regresyjna (podejście Rubina):**

1. harmonizacja wag: $w'_{i_{A\cup B}} = \frac{w_{i_{A\cup B}}}{\sum_{i=1}^{s} w_{i_{A\cup B}}} N$, (wzór 4.53),
2. do zbioru A imputowane są wartości teoretyczne wynikające z modelu: $\tilde{z}_a^{(A)} = \hat{z}_a^{(A)} + e_a = \hat{\alpha}_Z + \hat{\beta}_{ZX} x_a + e_a$ (wzór 4.48),
3. do zbioru B imputowane są wartości teoretyczne wynikające z modelu: $\tilde{y}_b^{(B)} = \hat{y}_b^{(B)} + e_b = \hat{\alpha}_Y + \hat{\beta}_{YX} x_b + e_b$ (wzór 4.49);

— **mieszane:**

1. stochastyczna imputacja regresyjna (podejście Rubina),
2. dla każdego rekordu w zbiorze biorcy wyszukiwany jest „najbliższy sąsiad” w zbiorze dawcy na podstawie odległości między wartościami teoretycznymi w A a empirycznymi w B: $d_{ab}(\tilde{z}_a, z_b) = min$.

Integrację metodą losową przeprowadzono w celu weryfikacji hipotezy, że zastosowanie statystycznych metod integracji poprawia jakość zintegrowanych zbiorów .

5.4.1. Integracja losowa

Integrację losową zalicza się do metod naiwnych. Jest to metoda nieparametryczna. Badanie empiryczne miało charakter symulacji składającej się ze stu iteracji[130] łączenia losowego. W każdej iteracji do każdego rekordu zbioru biorcy dołączono losowo wybrany rekord zbioru dawcy. Następnie przeprowadzono syntetyczną ocenę jakości integracji według podejścia Raessler (czwarty poziom)[2002]:

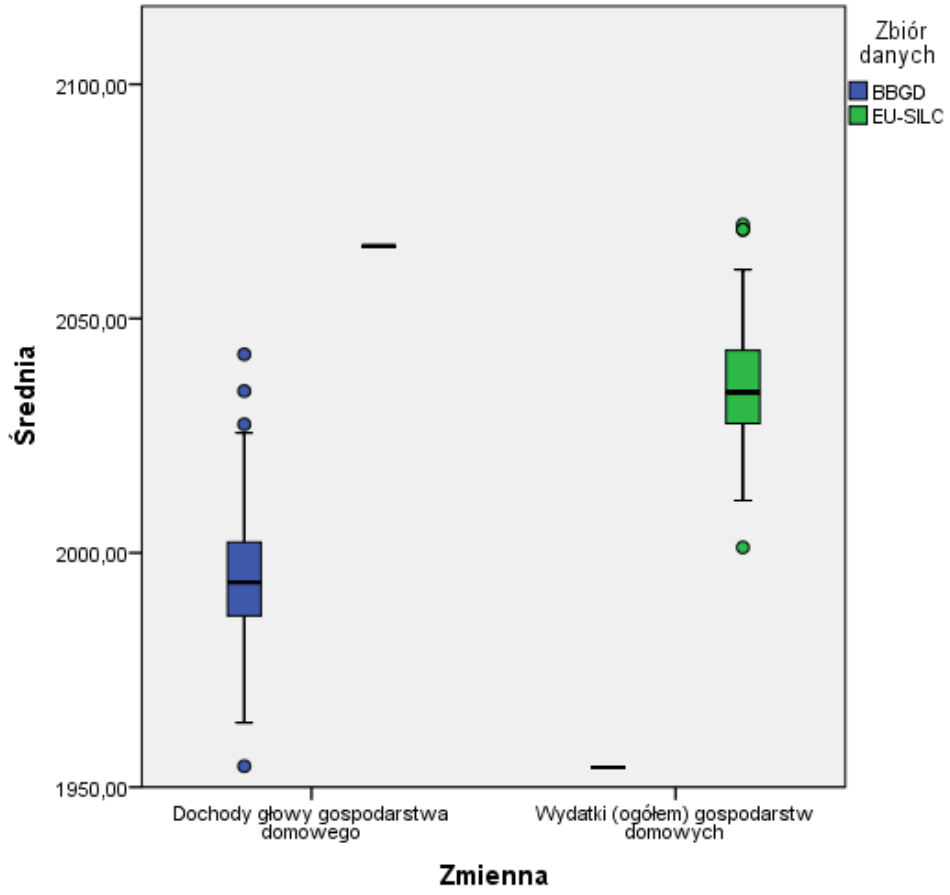
— ocena zgodności rozkładów zmiennych dołączanych w zbiorze dawcy i biorcy,

[130] Utworzono 100 zintegrowanych zbiorów danych.

— ocena zgodności rozkładów łącznych zmiennych dołączanych w zbiorze biorcy z wybranymi zmiennymi wspólnymi w porównaniu z analogicznymi rozkładami empirycznymi w zbiorze dawcy.

Analizę zgodności rozkładów zmiennych ilościowych w zbiorze biorcy z wartościami empirycznymi (w zbiorze dawcy) przeprowadzono wykorzystując wskaźniki zgodności dla średnich arytmetycznych $\frac{\bar{x}_A}{\bar{x}_B}$ oraz odchylenia standardowego $\frac{\hat{s}_A}{\hat{s}_B}$. Im wartości tych wskaźników są bliższe jedności, tym większa zgodność analizowanych rozkładów.

Wykres 5.1. Rozkład średnich arytmetycznych dołączanych zmiennych ilościowych, 100 iteracji integracji losowej



Uwaga:
Poziome kreski obok wykresu pudełkowego dla danej cechy oznaczają wartość empiryczną.
Źródło: opracowanie własne

Tabela 5.6. Charakterystyki rozkładu wskaźników zgodności dla średnich arytmetycznych dołączanych zmiennych ilościowych, 100 iteracji integracji losowej

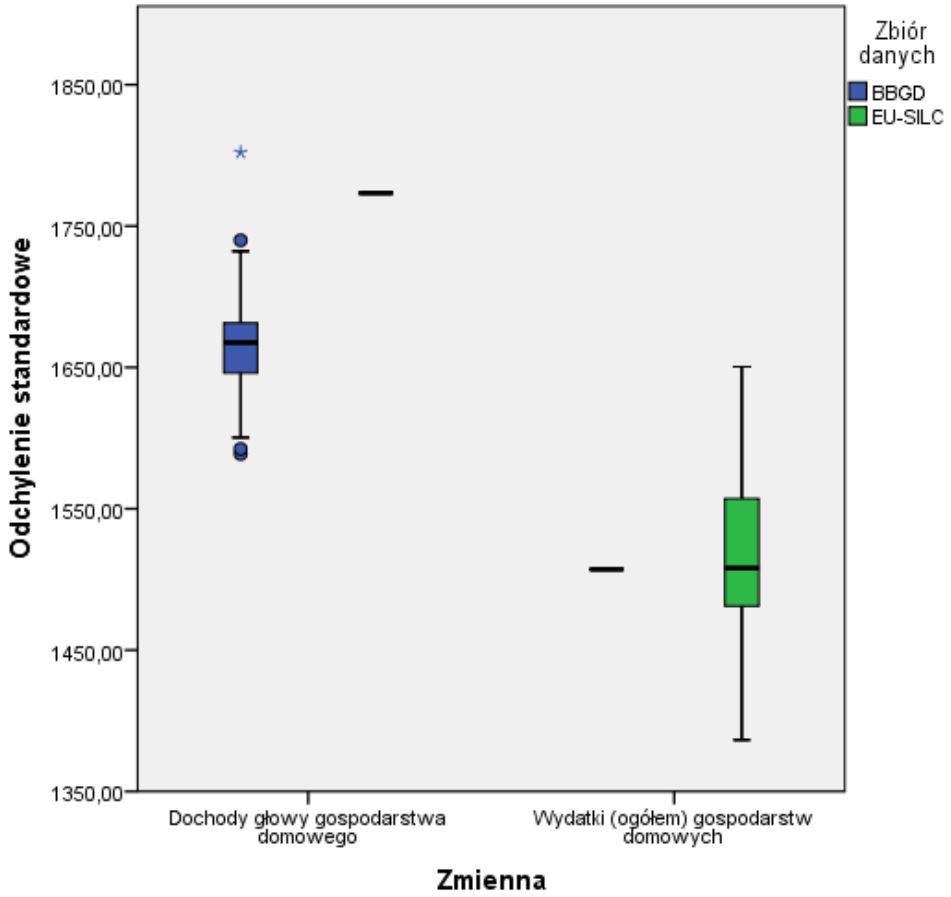
Statystyka	Dochody głowy gospodarstwa domowego	Wydatki (ogółem) gospodarstwa domowego
Średnia	0,96563	1,04172
Mediana	0,96526	1,04098
Odchylenie standardowe	0,00689	0,00679
Wariancja	0,00005	0,00005
Skośność	0,32432	0,27759
Rozstęp	0,04255	0,03526
Minimum	0,94626	1,02402
Maksimum	0,98881	1,05928

Źródło: opracowanie własne

Średnia arytmetyczna zmiennej „dochody głowy gospodarstwa domowego" (dołączanej ze zbioru EU-SILC do BBGD) w każdej iteracji została niedoszacowana (por. wykres 5.1). Niedoszacowanie jednak wynosiło przeciętnie niecałe 3,5% (por. tabela 5.6) i można je uznać za niewielkie (najmniejsza różnica wynosiła 1,1%, zaś największa 5,4%).
Z kolei dla średniej arytmetycznej zmiennej „wydatki (ogółem) gospodarstwa domowego" (dołączanej z BBGD do EU-SILC) w każdej iteracji nastąpiło przeszacowanie (por. wykres 5.1). Podobnie jednak jak w przypadku zmiennej „dochody", różnice nie były duże i przeszacowanie względem wartości empirycznej wynosiło średnio 4% (por. tabela 5.6).
Zarówno dla zmiennej „dochody głowy", jak i „wydatki", wartość średniej arytmetycznej nie różniła się znacząco w kolejnych iteracji. Algorytm losowego dopasowania informuje o stosunkowo małym zróżnicowaniu. Jednak średnie empiryczne w każdym przypadku były poza zakresem zmienności losowej integracji.

Dyspersja wydatków gospodarstwa domowego została odzwierciedlona znacznie lepiej. Przeciętnie wartości imputowane różniły się od empirycznych zaledwie o 0,7%. Niemniej należy zauważyć wysoką zmienność odchyleń standardowych dla każdej iteracji. Rozstęp współczynników zgodności wartości tego parametru rozkładu wynosi aż 17,5%, tj. od 8-procentowego niedoszacowania do prawie 10-procentowego przeszacowania (por. tabela 5.7).

Wykres 5.2. Rozkłady odchyleń standardowych dołączanych zmiennych ilościowych, 100 iteracji integracji losowej



Uwaga:
Poziome kreski obok wykresu pudełkowego dla danej cechy oznaczają wartość empiryczną.

Źródło: opracowanie własne

Tabela 5.7. Charakterystyki rozkładu wskaźników zgodności odchylenia standardowego dołączanych zmiennych ilościowych, 100 iteracji integracji losowej

Statystyka	Dochody głowy gospodarstwa domowego	Wydatki (ogółem) gospodarstwa domowego
Średnia	0,93965	1,00689
Mediana	0,94041	1,00061
Odchylenie standardowe	0,01893	0,03604
Wariancja	0,00036	0,00130
Skośność	0,51319	0,30939
Rozstęp	0,12033	0,17525
Minimum	0,89591	0,91984
Maksimum	1,01624	1,09509

Źródło: opracowanie własne

Odchylenie standardowe dochodów głowy gospodarstwa domowego również zostało niedoszacowane (por. wykres 5.2). Przeciętna różnica między wartościami imputowanymi, a empirycznymi wynosi 6%. Największe niedoszacowanie odchylenia standardowego wyniosło 10,5%, zaś w jednym przypadku zaobserwowano przeszacowanie wartości o 1,6%.

Dla dołączanych cech jakościowych (województwo oraz czy gospodarstwo stać na tygodniowy urlop rocznie), ocenę zgodności rozkładów brzegowych przeprowadzono wykorzystując wskaźnik podobieństwa Δ (wzór 4.1) określony dla frakcji poszczególnych kategorii. Rozkład brzegowy dołączonej do zbioru EU-SILC zmiennej „województwo" w 100 iteracjach różnił się średnio od empirycznego (w zbiorze BBGD) o 2,01% (por. tabela 5.8). W 50 iteracjach różnica między rozkładami była nie większa niż 2,02%, zaś w pozostałych 50 iteracji nie mniejsza. W najbardziej 'zgodnej; iteracji przeciętna różnica wynosiła zaledwie 1,28%, zaś w iteracji najmniej 'zgodnej' – 3,01%. Rozkład współczynników podobieństwa Δ charakteryzował się niedużą zmiennością, z rozstępem o wartości 1,72% i zaledwie jedną obserwacją odstającą (por. wykres 5.3).

Tabela 5.8. Charakterystyki rozkładu współczynników podobieństwa Δ dla frakcji dołączanych cech jakościowych,100 iteracji integracji losowej (w %)

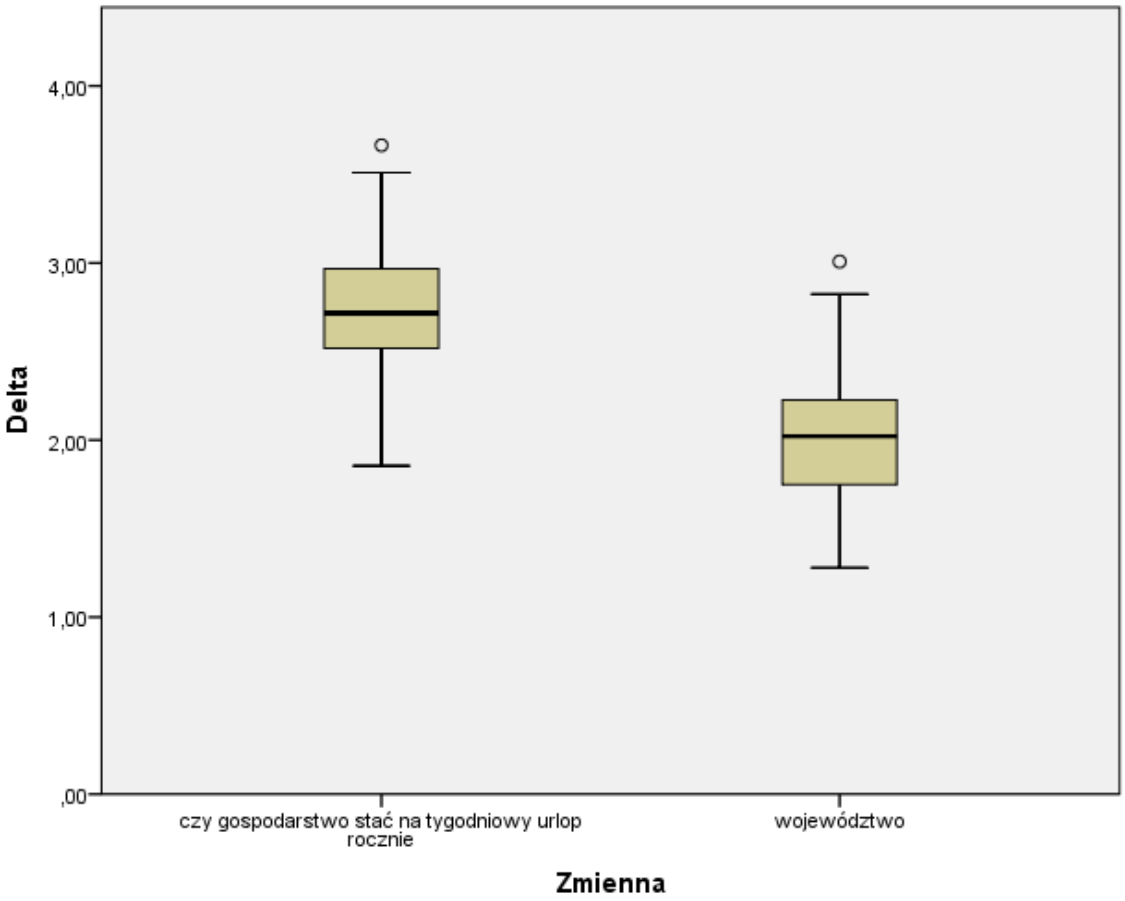
Statystyka	**Cecha: Województwo**	**Cecha: Czy gospodarstwo stać na tygodniowy urlop rocznie?**
Średnia	2,0096	2,7368
Mediana	2,0213	2,7170
Odchylenie standardowe	0,3415	0,3603
Skośność	0,2546	-0,0046
Rozstęp	1,7290	1,8110
Minimum	1,2790	1,8540
Maksimum	3,0080	3,6650

Źródło: opracowanie własne

Zmienna „czy gospodarstwo stać na tygodniowy urlop rocznie" dołączana ze zbioru EU-SILC do zbioru BBGD charakteryzowała się nieco mniejszą zgodnością rozkładów brzegowych niż zmienna „województwo". Jednak różnice w porównaniu z rozkładem empirycznym były stosunkowo nieduże. Średnio rozkłady brzegowe zmiennej dołączanej w 100 iteracjach różniły się o 2,74%. W 50 iteracjach różnice nie przekraczały

2,72%, a w pozostałych 50 były większe (por. tabela 5.8). Iteracja o najbardziej zgodnym rozkładzie brzegowych charakteryzowała się wartością wskaźnika na poziomie 1,85%, zaś iteracja o najmniejszej zgodności – 3,67%. Podobnie jak w przypadku cechy „województwo", rozkład charakteryzował się stosunkowo niewielką dyspersją – rozstęp wynosił 1,81%, przy jednej obserwacji odstającej (por. wykres 5.3).

Wykres 5.3. Rozkład współczynników podobieństwa Δ dla frakcji cech jakościowych, 100 iteracji integracji losowej (w %)



Źródło: opracowanie własne

Zgodność rozkładów brzegowych wartości imputowanych z wartościami empirycznymi dołączanych cech jakościowych można uznać za zadowalającą (Δ w każdym przypadku była mniejsza od 6%).

Należy zauważyć, że współczynniki zgodności rozkładów dołączanych zmiennych ilościowych (zarówno średniej, jak i odchylenia standardowego) są znacznie lepsze w przypadku wydatków, niż dochodów głów. W głównej mierze wynika to z faktu, że zmienna „wydatki"

była dołączana ze zbioru o większej liczebności do zbioru mniejszego. Prawdopodobieństwo wylosowania dokładnie tego samego rekordu biorcy jest niższe, niż w przypadku dołączania wartości ze zbioru mniejszego do większego (jak to miało miejsce w przypadku zmiennej „dochody głów"; por. tabela 5.9). Dodatkowo warto zwrócić uwagę, że część rekordów nie została w ogóle wykorzystana w procesie integracji. Decyzja, który ze zbiorów ma być dawcą, a który biorcą znajduje swoje odzwierciedlenie w jakości integracji. Choć doświadczenie pokazuje, że niemal zawsze część rekordów zostanie nieprzyłączona, bez względu na wariant statystycznej integracji danych [Roszka 2011b].

Tabela 5.9. Rozkład sparowań rekordów dawcy, 100 iteracji integracji losowej

Liczba sparowań	Kierunek integracji	
	BBGD → EU-SILC	EU-SILC → BBGD
0	67,81%	21,36%
1	23,93%	22,95%
2	6,35%	18,51%
3	1,49%	13,28%
4	0,33%	8,94%
5	0,07%	5,77%
6 i więcej	0,02%	9,18%

Źródło: opracowanie własne

Rozkłady łączne dla cech jakościowych z wybranymi zmiennymi wspólnymi zbadano za pomocą współczynnika kontyngencji Pearsona. Integracja losowa nie zachowała rozkładów łącznych dla cech jakościowych. Współczynniki kontyngencji dla cechy „województwo" w pliku biorcy (EU-SILC) są w dużej mierze niedoszacowane, a dodatkowo w każdym przypadku były zbliżone do zera (por. tabela 5.10).

Tabela 5.10. Charakterystyki rozkładu współczynników kontyngencji zmiennej „województwo" z wybranymi zmiennymi wspólnymi, 100 iteracji integracji losowej

Zmienna	**Statystyka**	**Współczynnik kontyngencji**	
		EU-SILC	**Empiryczna**
Czy jest łazienka ?	**Średnia**	**0,033**	0,132
	Odchylenie standardowe	0,007	
	Mediana	0,033	
	Minimum	0,020	
	Maksimum	0,046	
Czy jest ustęp spłukiwany ?	**Średnia**	**0,033**	0,136
	Odchylenie standardowe	0,006	
	Mediana	0,033	
	Minimum	0,018	
	Maksimum	0,049	

Zmienna	Statystyka	Współczynnik kontyngencji	
		EU-SILC	Empiryczna
Typ biologiczny gospodarstwa domowego	**Średnia**	**0,089**	0,125
	Odchylenie standardowe	0,006	
	Mediana	0,089	
	Minimum	0,074	
	Maksimum	0,103	
Czy gospodarstwo posiada komputer?	**Średnia**	**0,036**	0,071
	Odchylenie standardowe	0,007	
	Mediana	0,036	
	Minimum	0,021	
	Maksimum	0,058	
Czy gospodarstwo posiada pralkę?	**Średnia**	**0,038**	0,053
	Odchylenie standardowe	0,007	
	Mediana	0,038	
	Minimum	0,021	
	Maksimum	0,057	
Czy gospodarstwo posiada samochód?	**Średnia**	**0,035**	0,081
	Odchylenie standardowe	0,006	
	Mediana	0,035	
	Minimum	0,022	
	Maksimum	0,049	
Czy gospodarstwo posiada TV?	**Średnia**	**0,035**	0,047
	Odchylenie standardowe	0,007	
	Mediana	0,035	
	Minimum	0,016	
	Maksimum	0,054	
Liczba pokoi	**Średnia**	**0,079**	0,152
	Odchylenie standardowe	0,006	
	Mediana	0,078	
	Minimum	0,057	
	Maksimum	0,094	
Rodzaj budynku	**Średnia**	**0,059**	0,271
	Odchylenie standardowe	0,007	
	Mediana	0,058	
	Minimum	0,048	
	Maksimum	0,085	
Tytuł prawny do zajmowanego mieszkania	**Średnia**	**0,065**	0,184
	Odchylenie standardowe	0,007	
	Mediana	0,064	
	Minimum	0,049	
	Maksimum	0,084	

Uwaga:
Wartości empiryczne współczynnika korelacji zostały oszacowane na podstawie zbioru BBGD

Źródło: opracowanie własne

Analogiczną sytuację zaobserwowano w badaniu relacji cechy „czy gospodarstwo stać na tygodniowy urlop rocznie" z wybranymi zmiennymi wspólnymi. Współczynniki kontyngencji w pliku biorcy (BBGD) są zbliżone do zera i nie są zbieżne do wartości empirycznych (por. tabela 5.11).

Tabela 5.11. Charakterystyki rozkładu współczynników kontyngencji zmiennej „czy gospodarstwo stać na tygodniowy urlop rocznie" z wybranymi zmiennymi wspólnymi, 100 iteracji integracji losowej

Zmienna	**Statystyka**	**Współczynnik kontyngencji**	
		BBGD	**Empiryczna**
Czy jest łazienka ?	**Średnia**	**0,004**	0,185
	Odchylenie standardowe	0,003	
	Mediana	0,003	
	Minimum	0	
	Maksimum	0,013	
Czy jest ustęp spłukiwany ?	**Średnia**	**0,004**	0,176
	Odchylenie standardowe	0,003	
	Mediana	0,004	
	Minimum	0	
	Maksimum	0,013	
Typ biologiczny gospodarstwa domowego	**Średnia**	**0,014**	0,176
	Odchylenie standardowe	0,004	
	Mediana	0,013	
	Minimum	0,007	
	Maksimum	0,027	
Czy gospodarstwo posiada komputer?	**Średnia**	**0,004**	0,179
	Odchylenie standardowe	0,003	
	Mediana	0,004	
	Minimum	0	
	Maksimum	0,013	
Czy gospodarstwo posiada pralkę?	**Średnia**	**0,005**	0,298
	Odchylenie standardowe	0,004	
	Mediana	0,005	
	Minimum	0	
	Maksimum	0,018	
Czy gospodarstwo posiada samochód?	**Średnia**	**0,004**	0,043
	Odchylenie standardowe	0,003	
	Mediana	0,003	
	Minimum	0	
	Maksimum	0,019	
Czy gospodarstwo posiada	**Średnia**	**0,005**	0,275

Zmienna	Statystyka	Współczynnik kontyngencji	
		BBGD	Empiryczna
TV?	Odchylenie standardowe	0,004	
	Mediana	0,005	
	Minimum	0	
	Maksimum	0,017	
Liczba pokoi	**Średnia**	**0,011**	0,144
	Odchylenie standardowe	0,003	
	Mediana	0,011	
	Minimum	0,003	
	Maksimum	0,023	
Rodzaj budynku	**Średnia**	**0,009**	0,121
	Odchylenie standardowe	0,004	
	Mediana	0,009	
	Minimum	0,002	
	Maksimum	0,018	
Tytuł prawny do zajmowanego mieszkania	**Średnia**	**0,01**	0,038
	Odchylenie standardowe	0,004	
	Mediana	0,009	
	Minimum	0,002	
	Maksimum	0,023	

Uwaga:
Wartości empiryczne współczynnika korelacji zostały oszacowane na podstawie zbioru EU-SILC

Źródło: opracowanie własne

Ważnym kryterium oceny zgodności jest analiza łącznych rozkładów zmiennych dołączanych ze zmiennymi wspólnymi. W celu przetestowania zgodności rozkładów łącznych zmiennych ilościowych wykorzystano współczynnik korelacji liniowej Pearsona.
Dla zmiennej „dochody głowy gospodarstwa domowego", integracja losowa w żaden sposób nie zachowała łącznego rozkładu. Dla wybranych zmiennych wspólnych, przeciętnie współczynnik korelacji wartości imputowanych był bliski zera (por. tabela 5.12). Siła zależności między analizowaną zmienną dołączaną, a zmiennymi charakteryzującymi dochody gospodarstwa domowego (dochód rozporządzalny i ekwiwalentny) była bardzo silna, zaś wśród wartości imputowanych najwyższa wartość tego współczynnika była niewiele większa od 0,01. Analogiczna sytuacja miała miejsce dla korelacji między wielkością i ekwiwalentną wielkością gospodarstwa domowego.

W przypadku zmiennej „wydatki (ogółem) gospodarstwa domowego" zaistniała analogiczna sytuacja. Przeciętna korelacja wśród wartości imputowanych była bliska zerowej (por. tabela

5.13). Zdarzyły się również przypadki ujemnego kierunku zależności, sprzecznego ze znakiem empirycznej wartości współczynnika korelacji.

Tabela 5.12. Charakterystyki rozkładu współczynników korelacji zmiennej „dochody głowy gospodarstwa domowego" z wybranymi zmiennymi wspólnymi, 100 iteracji integracji losowej

Zmienna	**Statystyka**	**Współczynnik korelacji**	
		BBGD	**Empiryczna**
Ekwiwalentny dochód gospodarstwa domowego	Średnia	0,0006	0,8537
	Odchylenie standardowe	0,0053	
	Mediana	0,0008	
	Minimum	-0,0135	
	Maksimum	0,0139	
Dochód rozporządzalny gospodarstwa	Średnia	0,0005	0,8865
	Odchylenie standardowe	0,0051	
	Mediana	0,0007	
	Minimum	-0,0101	
	Maksimum	0,0151	
Liczba osób w gospodarstwie domowym	Średnia	0,0005	0,1541
	Odchylenie standardowe	0,0058	
	Mediana	-0,0001	
	Minimum	-0,0124	
	Maksimum	0,0168	
Ekwiwalentna wielkość gospodarstwa domowego	Średnia	0,0005	0,1536
	Odchylenie standardowe	0,0059	
	Mediana	0,0004	
	Minimum	-0,0122	
	Maksimum	0,0165	

Uwaga:

Wartości empiryczne współczynnika korelacji zostały oszacowane na podstawie zbioru EU-SILC

Źródło: opracowanie własne

Tabela 5.13. Charakterystyki rozkładu współczynników korelacji zmiennej „wydatki (ogółem) gospodarstwa domowego" z wybranymi zmiennymi wspólnymi, 100 iteracji integracji losowej

Zmienna	Statystyka	Wartość	
		EU-SILC	Empiryczna
Ekwiwalentny dochód gospodarstwa domowego	Średnia	-0,0019	0,4840
	Odchylenie standardowe	0,0096	
	Mediana	-0,0027	
	Minimum	-0,0295	
	Maksimum	0,0230	
Dochód rozporządzalny gospodarstwa	Średnia	-0,0019	0,5880
	Odchylenie standardowe	0,0099	
	Mediana	-0,0022	
	Minimum	-0,0325	
	Maksimum	0,0289	
Liczba osób w gospodarstwie domowym	Średnia	0,0005	0,2685
	Odchylenie standardowe	0,0084	
	Mediana	-0,0001	
	Minimum	-0,0257	
	Maksimum	0,0284	
Ekwiwalentna wielkość gospodarstwa domowego	Średnia	0,0003	0,2790
	Odchylenie standardowe	0,0087	
	Mediana	0,0000	
	Minimum	-0,0287	
	Maksimum	0,0290	

Uwaga:
Wartości empiryczne współczynnika korelacji zostały oszacowane na podstawie zbioru BBGD
Źródło: opracowanie własne

Integracja metodą losową w sposób zadowalający odzwierciedla podstawowe charakterystyki rozkładów zmiennych dołączanych. Nie odzwierciedla jednak rozkładów łącznych, co uniemożliwia jej empiryczne wykorzystane w pracach organów statystyki publicznej lub firm badawczych. Wyniki integracji uzyskane tą metodą przyjęto jako podstawę porównań dla bardziej zaawansowanych metod.

5.4.2. Statystyczna integracja danych społecznych

Odrzucając metodę losową jako technikę umożliwiającą odzwierciedlenie empirycznych rozkładów cech dołączanych i łącznych w pliku zintegrowanym, podjęto próbę parowania statystycznego wymienionymi metodami bardziej złożonymi, tj.:

— najbliższego sąsiedztwa (*nearest neighbour distance, NND*),

— stochastyczną imputację regresyjną,

— mieszaną (*predictive mean matching, PMM*).

Wykorzystano oprogramowanie IBM SPSS 20. Dla metody najbliższego sąsiada wykorzystano autorski kod w edytorze poleceń SYNTAX (por. Załącznik) oparty o algorytm Bachera [2002]. Przy zastosowaniu metody imputacji stochastycznej oraz metody mieszanej skorzystano z funkcjonalności IBM SPSS – Wielokrotne podstawienia.

Metoda NND

Ze względu na fakt, że metoda najbliższego sąsiada wymaga porównania każdego rekordu w zbiorze dawcy z każdym rekordem w zbiorze biorcy i związanej z tym czasochłonności obliczeń[131], zbiory dawcy i biorcy podzielono na 12 warstw (grup) wyznaczonych przez regiony oraz „czy dochody pozwalają na wiązanie końca z końcem?" (por. tabela 5.14). Zmienną tą zagregowano w dwie kategorie: „z trudnością"[132] oraz „łatwo"[133]. Utworzenie warstw pozwoliło na zredukowanie liczby połączeń ponad ośmiokrotnie. Cecha „czy dochody pozwalają na wiązanie końca z końcem?" charakteryzowała się stosunkowo dużą różnicą rozkładu dla niezagregowanych wariantów, jednak po agregacji współczynnik podobieństwa wynosił $\Delta = 1{,}34\%$. U podstaw decyzji o utworzeniu warstw leżało również założenie, że w wyróżnionych grupach jednostki charakteryzują się większą zgodnością co do wartości integrowanych cech.

Tabela 5.14. Liczebności integrowanych zbiorów w ujęciu wyznaczonych warstw

Warstwa		Zbiór		Liczba porównań rekordów
Nr	**Warianty cech grupujących**	**BBGD**	**EU-SILC**	
1	region centralny; z trudnością	5761	2377	13 693 897
2	region centralny; łatwo	1495	624	932 880
3	region południowy; z trudnością	5774	2479	14 313 746
4	region południowy; łatwo	1486	635	943 610
5	region wschodni; z trudnością	4980	2368	11 792 640
6	region wschodni; łatwo	1096	494	541 424
7	region północno-zachodni; z trudnością	4312	1802	7 770 224
8	region północno-zachodni; łatwo	1019	423	431 037
9	region południowo-zachodni; z trudnością	3029	1230	3 725 670
10	region południowo-zachodni; łatwo	781	333	260 073
11	region północny; z trudnością	4109	1755	7 211 295
12	region północny; łatwo	925	394	364 450
OGÓŁEM		**34 767**	**14 914**	**61 980 946**

Źródło: opracowanie własne

[131] Porównanie każdego rekordu z każdym wymagałoby obliczenia odległości dla $34767 \times 14914 = 518\,515\,038$ par rekordów.
[132] Złożoną z wariantów: „z pewną trudnością", „z trudnością", „z wielką trudnością".
[133] Złożoną z wariantów: „raczej łatwo", „łatwo", „bardzo łatwo".

Jako miarę odległości między parami rekordów wykorzystano kwadratową odległość euklidesową. W celu wyłączenia wpływu jednostki pomiaru, ilościowe zmienne parujące poddano standaryzacji. Jakościowe zmienne parujące zdychotomizowano w $k-1$ kategorii. W celu minimalizacji utraty informacji, usunięto kategorię w każdym przypadku najmniej liczną. W sytuacji rekordów o tej samej minimalnej odległości, rekordy parowano losowo.

W celu weryfikacji przypuszczenia, że wybór zmiennych parujących poprawia jakość połączenia, zastosowano dwa podejścia:

1. Wykorzystanie zmiennych parujących wybranych metodą CART.
2. Przeprowadzenie integracji przy wykorzystaniu wszystkich dostępnych zmiennych wspólnych jako parujących.

Metoda stochastycznej imputacji regresyjnej i mieszana

W metodzie imputacji stochastycznej oraz mieszanej dokonano korekty wyboru zmiennych parujących w oparciu o współczynniki standaryzowane Beta[134], będące współczynnikiem korelacji wielorakiej danej zmiennej niezależnej ze zmienną zależną przy wyłączeniu wpływu pozostałych zmiennych niezależnych. Wybierano cechy o współczynnikach „wyraźnie" większych od pozostałych (w sposób ekspercki). Do budowy modeli liniowych wykorzystano funkcjonalność programu IBM SPSS – Automatyczne modelowanie liniowe. Utworzono dwa rodzaje modeli dla każdej zmiennej dołączanej (zależnej): dla Polski ogółem oraz 6 modeli dla każdego regionu z osobna.

Ze względu na stosunkowo niskie dopasowanie wartości teoretycznych do danych empirycznych, postanowiono zlogarytmować zmienną „wydatki" (por. tabela 5.15). Logarytmowanie zmiennej „dochód głowy" przynosiło pogorszenie jakości modelu, wykorzystano więc wartości oryginalne.

[134] Ze względu na dużą liczbę obserwacji w zbiorach oraz zastosowanie wag analitycznych „klasyczne" metody doboru zmiennych do modelu regresji (np. krokowa wprzód i wstecz) nie wykluczały żadnych cech.

Tabela 5.15. Wartości współczynników determinacji R^2 dla utworzonych modeli

Model	Zmienna zależna			
	wyd	$ln(wyd)$	$doch\_glowy$	$ln(doch\_glowy)$
region1	0,478	0,656	0,791	0,483
region2	0,479	0,653	0,700	0,456
region3	0,460	0,621	0,696	0,488
region4	0,458	0,666	0,735	0,422
region5	0,463	0,651	0,724	0,436
region6	0,476	0,630	0,724	0,433
Polska	0,459	0,644	0,711	0,445

Uwaga:
wyd – wydatki ogółem gospodarstwa domowego
$doch\_glowy$ – dochody głowy gospodarstwa domowego
Źródło: opracowanie własne

W następnym kroku dokonano wielokrotnej imputacji, korzystając z funkcjonalności wielokrotne podstawienia programu IBM SPSS 20. Utworzono dwa modele:

1. na podstawie zmiennych wspólnych, jako zmiennych niezależnych i empirycznych wartości zmiennej rozchody netto gospodarstwa domowego zaimputowano wartości do jednostek pochodzących z EU-SILC;
2. podstawie zmiennych wspólnych, jako zmiennych niezależnych i empirycznych wartości zmiennej dochody głowy gospodarstwa domowego zaimputowano wartości do jednostek pochodzących z BBGD.

Dokonano 100 imputacji na podstawie modeli regresji liniowej. Wartości resztowe zostały wylosowane za pomocą metody Monte Carlo opartej o łańcuchy Markova (*Markov Chain Monte Carlo*, MCMC – por. [IBM SPSS Missing Values 20 2011], [Grzenda 2012]). W celu harmonizacji wag analitycznych zastosowano podejście Rubina.

Ostatecznie utworzono 100 pełnych zbiorów danych dla dwóch podejść metodologicznych: wielokrotnej imputacji stochastycznej oraz modelu mieszanego (w programie IBM SPSS nazywany *Predictive Mean Matching*, PMM).

5.5. Ocena jakości integracji

Oceny jakości integracji, podobnie jak w przypadku podejścia losowego dokonano w oparciu o czwarty poziom podejścia Raessler [2002] (por. pkt. 4.4.4.). Dodatkowo dokonano porównania metod wewnątrz grup metodologicznych w oparciu o ich charakterystyki i założenia.

Ponieważ efektem parowania statystycznego jest jeden zintegrowany zbiór danych, dla każdej zmiennej dołączanej w wyniku analizy porównawczej wybrana zostanie najlep-

sza, w kontekście zaimplementowanej metody oceny jakości, metoda. Następnie utworzony zostanie zintegrowany zbiór danych, na podstawie którego oszacowane i ocenione zostaną łączne charakterystyki zmiennych dołączanych.

5.5.1. Ocena algorytmów połączenia

W pierwszym kroku oceny jakości integracji z zastosowaniem wybranych metod dokonano porównania poszczególnych grup metod między sobą. W tym celu wykorzystano własności metod:

- najbliższego sąsiada:
  - rozkład funkcji najbliższej odległości,
  - liczba przyłączeń poszczególnych rekordów dawcy do biorcy;
- imputacji stochastycznej i mieszanej:
  - przedziały niepewności dla średniej arytmetycznej (wzory 4.58 – 4.62),
  - przedziały niepewności dla nieznanej wartości współczynnika korelacji między zmiennymi dołączanymi.

Poza analizą rozkładów cech dołączanych, w metodzie najbliższego sąsiada ocenie merytorycznej poddane zostaną również charakterystyki związane z rozkładem funkcji minimalnej odległości (na podstawie które parowane są rekordy ze zbioru dawcy i biorcy), a także liczba, jaką jeden rekord ze zbioru dawcy został dołączony do zbioru biorcy. Pierwsza z wymienionych charakterystyk ma związek z efektywnością algorytmu, natomiast druga z optymalnym wykorzystaniem informacji zawartych w zbiorze dawcy.

Tabela 5.16. Rozkład liczby sparowań rekordów w pliku dawcy według zmiennej dołączanej i metody doboru zmiennych parujących

Liczba sparowań	Zmienna dołączana, metoda doboru zmiennych					
	Wydatki (ogółem) gospodarstwa domowego, CART	**Dochody głowy gospodarstwa domowego, CART**	**Województwo, CART**	**Czy gospodarstwo stać na tygodniowy urlop rocznie, CART**	**Wydatki, województwo, wszystkie zmienne wspólne**	**Dochody głowy, czy stać na urlop, wszystkie zmienne wspólne**
0	69,39	31,66	68,37	33,21	71,70	33,75
1	21,86	18,60	23,28	17,80	19,36	19,31
2	6,23	14,68	6,21	13,88	5,83	14,01
3	1,77	10,68	1,59	10,61	1,83	9,55
4	0,55	7,60	0,38	7,67	0,69	6,83
5	0,13	5,06	0,11	5,26	0,29	4,87
6	0,05	3,71	0,05	3,65	0,12	3,09
7 i więcej	0,02	8,00	0,01	7,91	0,17	8,59

Uwaga, kolor:
szary – integracja ze zbioru większego do mniejszego (BBGD → EU-SILC),
biały – integracja za zbioru mniejszego do większego (EU-SILC → BBGD).
Źródło: opracowanie własne

Bardzo duża część rekordów (bez względu na zastosowaną metodę) nie została wykorzystana w procesie integracji. Rekordy te nie spełniły kryterium podobieństwa wyznaczonego przez zastosowany algorytm. Należy zauważyć, że w przypadku, gdy dawca jest zbiorem liczniejszym (w tabeli zaznaczono szarym kolorem), zdecydowana większość rekordów nie jest przyłączana w ogóle (por. tabela 5.16). W idealnym przypadku, liczba nieprzyłączonych rekordów powinna być różnicą między liczebnościami zbioru dawcy i biorcy (w analizowanym przypadku winno to być $34767 - 14914 = 19853$; 57,1%). Sytuacja taka jest bardzo rzadko spotykana. Zaobserwowano jednak, że algorytm w przypadku dołączania zbioru większego do mniejszego charakteryzuje się większą efektywnością względem analizowanego kryterium. Jednocześnie można zauważyć, że w przypadku rezygnacji z doboru zmiennych parujących i wykorzystania wszystkich zmiennych wspólnych, algorytm jest nieco mniej efektywny niż przy uwzględnieniu tylko zmiennych istotnie skorelowanych ze zmienną dołączaną. Algorytm wykorzystujący wszystkie zmienne wspólne jako parujące przeciętnie częściej nie wykorzystuje większej liczby rekordów dawcy, a także charakteryzuje się większą liczbą rekordów przyłączanych wiele razy (szczególnie 7 i więcej razy).

Rozkład funkcji najmniejszej odległości[135] charakteryzuje się przede wszystkim skrajną asymetrią prawostronną (por. tabela 5.17). Jest to cecha bardzo pożądana, ponieważ zadaniem algorytmu jest wyszukanie rekordów jak najbardziej do siebie podobnych (a więc o możliwie zbliżonej do zera funkcji odległości). Im bardziej asymetryczny rozkład funkcji odległości tym bardziej efektywny algorytm znajdowania rekordów podobnych (tzw. „statystycznych bliźniąt").

Tabela 5.17. Charakterystyka rozkładu standaryzowanej funkcji minimalnej odległości

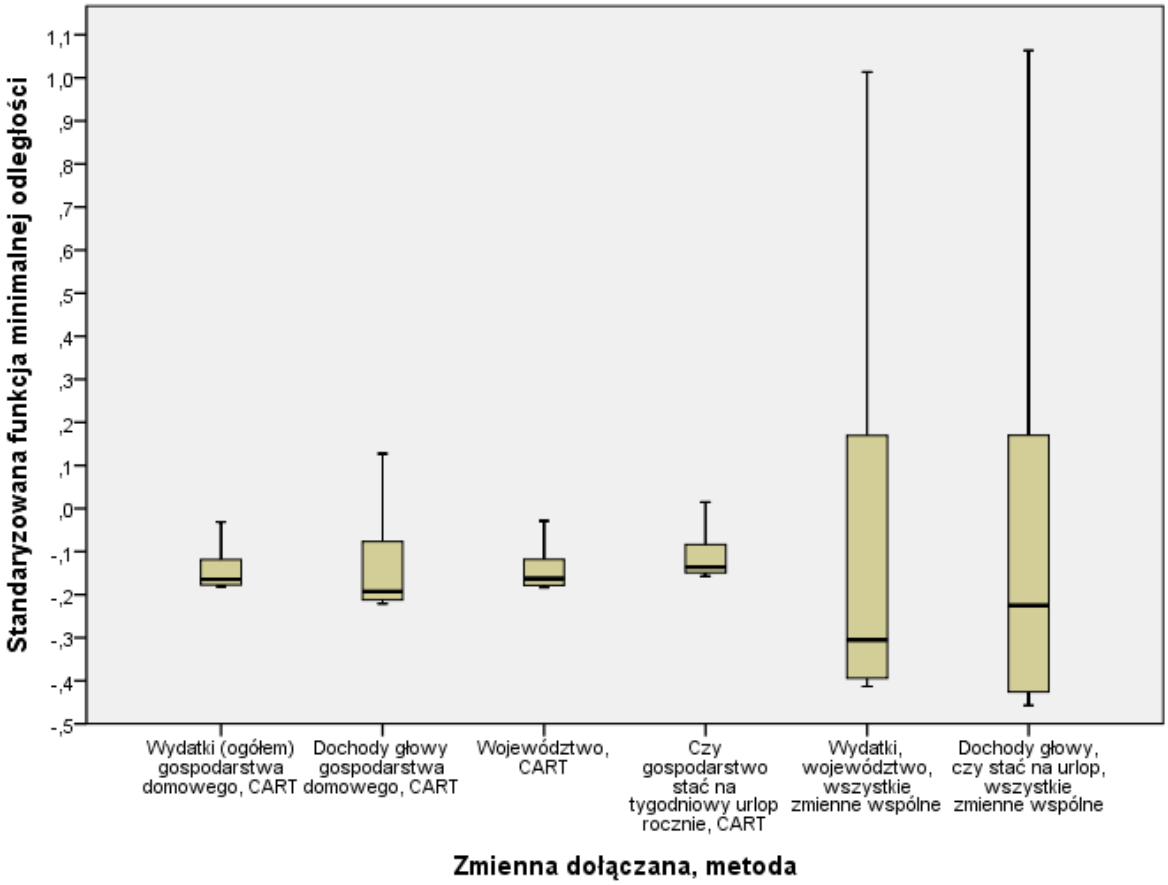
Statystyka	Zmienna dołączana, metoda doboru zmiennych					
	Wydatki (ogółem) gospodarstwa domowego, CART	**Dochody głowy gospodarstwa domowego, CART**	**Województwo, CART**	**Czy gospodarstwo stać na tygodniowy urlop rocznie, CART**	**Wydatki, województwo, wszystkie zmienne wspólne**	**Dochody głowy, czy stać na urlop, wszystkie zmienne wspólne**
Średnia	0	0	0	0	0	0
Odchylenie standardowe	1	1	1	1	1	1
Minimum	-0,181	-0,221	-0,183	-0,157	-0,413	-0,457
Maksimum	36,452	50,189	35,322	54,013	30,189	44,482
Rozstęp	36,632	50,409	35,505	54,170	30,602	44,939
Rozstęp międzykwartylowy	0,058	0,136	0,060	0,066	0,564	0,596
Skośność	13,987	26,471	14,015	30,610	10,185	18,529

Uwaga, kolor:
szary – integracja ze zbioru większego do mniejszego (BBGD → EU-SILC),
biały – integracja za zbioru mniejszego do większego (EU-SILC → BBGD).
Źródło: opracowanie własne

Analogicznie jak w przypadku liczby sparowań, algorytm wykorzystujący wszystkie zmienne wspólne jako parujące wydaje się mniej efektywny pod kątem rozkładu funkcji minimalnej odległości. Charakteryzuje się on mniejszą asymetrią prawostronną (choć nadal skrajną), jak również dużo mniejszą dyspersją. Również rozstęp międzykwartylowy funkcji, który można uznać za przedział zawierający typowe odległości, jest większy niż w pozostałych przypadkach (por. wykres 5.4).

[135] W analizowanym przykładzie funkcja odległości ze względu na różną skalę, wynikającą z różnej liczby wykorzystanych zmiennych parujących, została poddana standaryzacji.

Wykres 5.4. Charakterystyka rozkładu standaryzowanej funkcji minimalnej odległości



Uwaga:
Ze względu na bardzo dużą liczbę obserwacji odstających i skrajnych, ograniczono zakres funkcji minimalnej odległości do 1,1 oraz przedstawiono wyłącznie rozkład obserwacji typowych.
Źródło: opracowanie własne

Podsumowując porównanie różnych wariantów metody najbliższego sąsiada można wysnuć dwa podstawowe wnioski:

— algorytm jest bardziej efektywny przy dołączaniu wartości ze zbioru większego do mniejszego,

— wybór zmiennych parujących z wektora zmiennych wspólnych poprawia efektywność algorytmu.

Tabela 5.18. Ocena estymatorów średniej arytmetycznej dołączanych zmiennych w zintegrowanym zbiorze, wybrane metody imputacji stochastycznej i mieszane

Zmienna dołączana	Statystyka	Imputacja stochastyczna, Polska ogółem	PMM, Polska ogółem	Imputacja stochastyczna, modele dla regionów	PMM, modele dla regionów
Wydatki (ogółem) gospodarstwa domowego	B	8,14	32,25	12,66	30,05
	W	8,80	10,06	8,52	10,05
	T	17,03	42,64	21,32	40,40
	$\sqrt{T}$	4,13	6,53	4,62	6,36
	v	513 208,26	10 288 061,84	1 152 446,90	8 907 036,88
	$t_{v,\frac{\alpha}{2}}$	2,2414093	2,2414031	2,2414057	2,2414031
	$\hat{\theta}_{MI} - t_{v,\frac{\alpha}{2}}\sqrt{T}$	1 950,86	2 005,29	1 953,06	1 998,92
	$\hat{\theta}_{MI} + t_{v,\frac{\alpha}{2}}\sqrt{T}$	1 969,36	2 034,56	1 973,75	2 027,41
	szerokość przedziału	18,50	29,27	20,70	28,49
Dochody głowy gospodarstwa domowego	B	35,20	5,23	23,47	5,29
	W	14,68	11,88	14,58	11,71
	T	50,23	17,17	38,28	17,05
	$\sqrt{T}$	7,09	4,14	6,19	4,13
	v	26 004 633,27	387 341,75	11 428 640,55	384 094,70
	$t_{v,\frac{\alpha}{2}}$	2,2414029	2,2414114	2,2414030	2,2414115
	$\hat{\theta}_{MI} - t_{v,\frac{\alpha}{2}}\sqrt{T}$	2 006,58	2 004,91	2 007,75	2 003,73
	$\hat{\theta}_{MI} + t_{v,\frac{\alpha}{2}}\sqrt{T}$	2 038,35	2 023,48	2 035,48	2 022,24
	szerokość przedziału	31,77	18,57	27,74	18,51

Uwaga, w tabelach 5.18 – 5.20 przyjęto oznaczenia:

$B = \frac{1}{m-1}\sum_{t=1}^{m}(\hat{\theta}^{(t)} - \hat{\theta}_{MI})^2$ – wariancja międzygrupowa (wzór 4.55),

$W = \frac{1}{m}\sum_{t=1}^{m}\widehat{var}\left(\hat{\theta}^{(t)}\right)$ – wariancja wewnątrzgrupowa (wzór 4.56),

$T = W + \frac{m+1}{m}B$ – wariancja ogólna (wzór 4.57),

$v = (m-1)(1 + \frac{W}{(1+\frac{1}{m})B})^2$ – liczba stopni swobody

$t_{v,\frac{\alpha}{2}}$ – kwantyl rozkładu t-Studenta

$\hat{\theta}_{MI} - t_{v,\frac{\alpha}{2}}\sqrt{T} < \theta < \hat{\theta}_{MI} + t_{v,\frac{\alpha}{2}}\sqrt{T}$ – przedział niepewności dla parametru θ (wzór 4.58).

Źródło: opracowanie własne

Tabela 5.19. Ocena estymatorów odchylenia standardowego dołączanych zmiennych w zintegrowanym zbiorze, wybrane metody imputacji stochastycznej i mieszane

Zmienna dołączana	Statystyka	Imputacja stochastyczna, Polska ogółem	PMM, Polska ogółem	Imputacja stochastyczna, modele dla regionów	PMM, modele dla regionów
Wydatki (ogółem) gospodarstwa domowego	B	593,23	1 209,56	191,76	936,66
	W	0,0004	0,0004	0,0003	0,0004
	T	599,16	1 221,65	193,68	946,03
	$\sqrt{T}$	24,48	34,95	13,92	30,76
	v	144,51	218,36	112,32	187,22
	$t_{v,\frac{\alpha}{2}}$	2,2650814	2,2569901	2,2719350	2,2595942
	$\hat{\theta}_{MI} - t_{v,\frac{\alpha}{2}}\sqrt{T}$	1 422,43	1 495,65	1 426,69	1 516,20
	$\hat{\theta}_{MI} + t_{v,\frac{\alpha}{2}}\sqrt{T}$	1 533,31	1 653,42	1 489,93	1 655,20
	szerokość przedziału	110,89	157,77	63,24	139,00
Dochody głowy gospodarstwa domowego	B	36,01	18,11	29,85	16,45
	W	0,0006	0,0005	0,0006	0,0005
	T	36,37	18,30	30,15	16,62
	$\sqrt{T}$	6,03	4,28	5,49	4,08
	v	103,22	100,71	102,46	100,53
	$t_{v,\frac{\alpha}{2}}$	2,2746407	2,2756524	2,2749712	2,2756524
	$\hat{\theta}_{MI} - t_{v,\frac{\alpha}{2}}\sqrt{T}$	1 895,95	1 707,50	1 890,58	1 696,98
	$\hat{\theta}_{MI} + t_{v,\frac{\alpha}{2}}\sqrt{T}$	1 923,39	1 726,97	1 915,57	1 715,54
	szerokość przedziału	27,44	19,47	24,98	18,55

Źródło: opracowanie własne

Warianty imputacji stochastycznej, a także metody mieszanej porównano wykorzystując własności estymatora wielokrotnej imputacji. Porównanie wariancji międzygrupowej (B), wewnątrzgrupowej (W) oraz ogólnej (T) można przedstawić rozpiętość przedziałów niepewności.

W przypadku modeli dla zmiennej „wydatki", najlepsze wyniki dla średniej arytmetycznej uzyskano stosując metodę imputacji stochastycznej wykorzystując model utworzony dla całego kraju ogółem (por. tabela 5.20). Dla zmiennej „dochody" najwęższy przedział niepewności oszacowano za pomocą metody mieszanej, dla modeli oszacowanych dla regionów. Bardzo zbliżoną szerokość przedziału niepewności zaobserwowano dla metody mieszanej z modelem oszacowanym dla całego kraju ogółem.

Tabela 5.20. Ocena estymatorów współczynnika korelacji dołączanych zmiennych w zintegrowanym zbiorze, wybrane metody imputacji stochastycznej i mieszane

Zmienna dołączana	Statystyka	Imputacja stochastyczna, Polska ogółem	PMM, Polska ogółem	Imputacja stochastyczna, modele dla regionów	PMM, modele dla regionów
Korelacja $z(\hat{\rho}^{(t)})$	B	0,00006	0,00013	0,00003	0,00013
	W	2E-15	2E-15	2E-15	2E-15
	T	0,00006	0,00013	0,00003	0,00013
	$\sqrt{T}$	0,01	0,01	0,01	0,01
	v	99,00	99,00	99,00	99,00
	$t_{v,\frac{\alpha}{2}}$	2,2760035	2,2760035	2,2760035	2,2760035
	$\hat{\theta}_{MI} - t_{v,\frac{\alpha}{2}}\sqrt{T}$	0,5611	0,5534	0,5661	0,5463
	$\hat{\theta}_{MI} + t_{v,\frac{\alpha}{2}}\sqrt{T}$	0,5849	0,5884	0,5824	0,5817
	szerokość przedziału	0,0238	0,0350	0,0164	0,0354

Uwaga:
Kolor szary oznacza przedział niepewności dla współczynnika korelacji ρ.
$z(\hat{\rho}^{(t)})$ to z-transformowany szacunek dla ρ: $z(\hat{\rho}^{(t)}) = \frac{1}{2} ln \frac{1+\rho_{YZ}^{(t)}}{1-\rho_{YZ}^{(t)}}$; $z(\hat{\rho}^{(t)})$ posiada rozkład normalny o stałej wariancji $\frac{1}{n-3}$. Przedziały ufności podano dla ρ (oznaczone kolorem szarym).
Źródło: opracowanie własne

Analogiczne wyniki zaobserwowano dla estymatora odchylenia standardowego, choć przedziały niepewności dla tego parametru rozkładu są szersze niż w przypadku średniej arytmetycznej. Dodatkowo zaobserwowano większe różnice przeciętnych wartości odchylenia w ujęciu różnych wariantów niż w przypadku średniej arytmetycznej (por. tabela 5.20). Przedziały niepewności oszacowane dla nieznanego współczynnika korelacji między zmiennymi „wydatki" i „dochody głów" są dla każdej metody bardzo podobne. Jednak najlepsze rezultaty osiągnięto dla imputacji stochastycznej, zarówno dla kraju ogółem, jak i regionów. Wśród ogólnych wniosków na temat metod wielokrotnej imputacji można wymienić:

- dolosowanie składnika resztowego do wartości empirycznych modelu regresji umożliwia wyznaczenie estymatorów o dobrych właściwościach,
- różne warianty metodologiczne zwracają podobne wartości średnie,
- odchylenie standardowe jest odzwierciedlane gorzej niż średnia,
- można przypuszczać, że jakość wyników uzależniona jest od zdolności predykcyjnej modelu regresji,
- nieznana korelacja jest odzwierciedlana z podobną jakością przez każdą z metod.

Należy mieć na uwadze, że na przedstawionej powyżej analizie nie można oprzeć oceny jakości integracji, gdyż ta sprowadza się do sprawdzenia zbieżności rozkładów wartości dołączanych (lub/i w zbiorze zintegrowanym) do wartości empirycznych.

5.5.2. Ocena charakterystyk rozkładów cech dołączanych

Zgodność rozkładów cech jakościowych w zbiorze dawcy i zintegrowanym zbadano za pomocą współczynnika zgodności Δ. Dokonano również graficznego porównania liczebności poszczególnych wariantów tych cech w zbiorze dawcy oraz zbiorze biorcy (po integracji).
Dla zmiennej „województwo", szacunki dokonane oboma zastosowanymi wariantami metody NND okazały się zbieżne do wartości empirycznych. W przypadku zastosowania metody najbliższego sąsiada z wykorzystaniem wybranych metodą CART zmiennych parujących, przeciętna różnica w rozkładzie wynosiła zaledwie 0,975% (por. tabela 5.21). Metoda NND przy wykorzystaniu wszystkich zmiennych parujących w procesie integracji zwróciła jeszcze bardziej zbieżne wyniki, o różnicy rzędu 0,635%. Tak niewielkie różnice mogą sugerować, że, przy założeniu zachowania rozkładów łącznych ze zmiennymi wspólnymi $\boldsymbol{X}$, szacunki w ujęciu województw utworzone na podstawie zintegrowanego zbioru będą charakteryzować się co najmniej taką samą jakością jak te utworzone na podstawie zbioru dawcy.

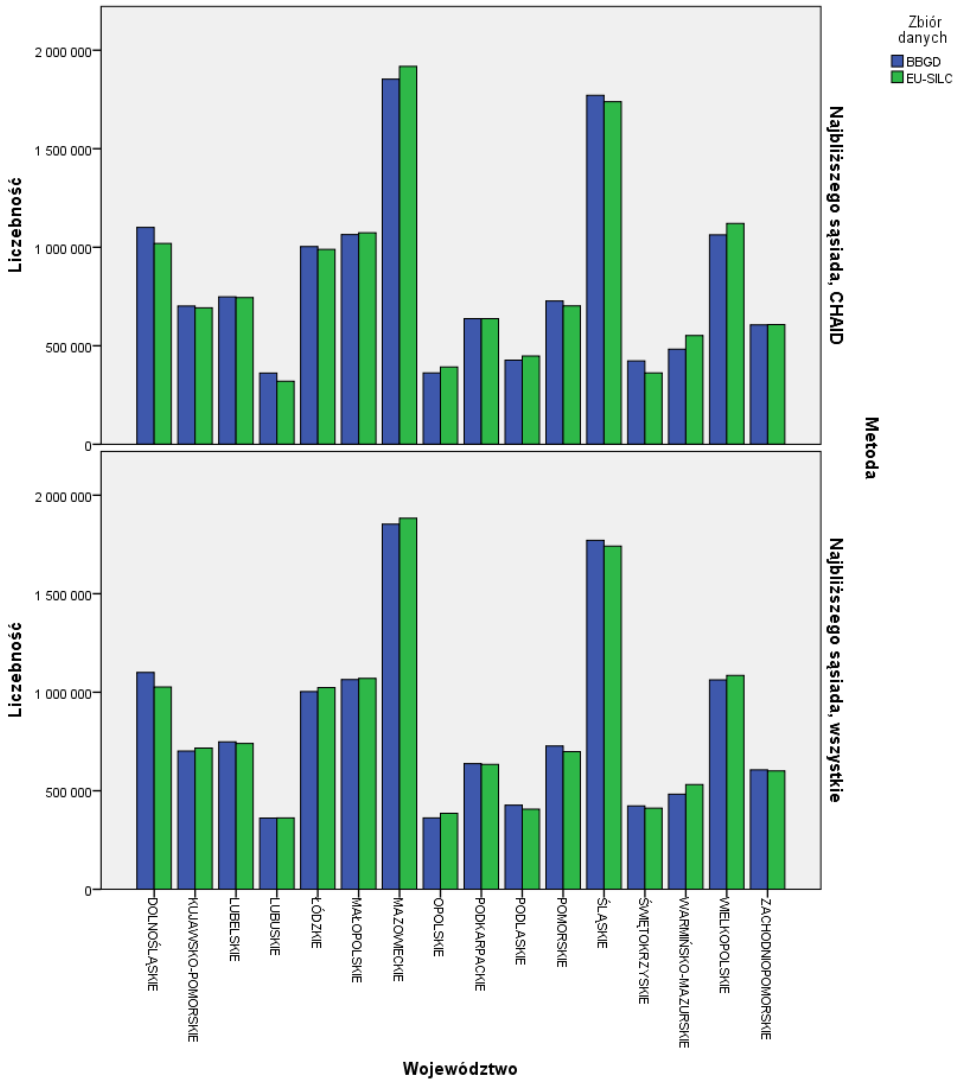
Tabela 5.21. Rozkłady brzegowe dla cechy „województwo” przed i po integracji

Metoda	Województwo	Zbiór danych						Δ
		Dawcy		Biorcy		Zintegrowany		
		BBGD		EU-SILC				
		N	%	N	%	N	%	
Najbliższego sąsiada, CART	dolnośląskie	1 100 231	8,25	1 019 344	7,65	1 059 075	7,95	0,975
	kujawsko-pomorskie	701 980	5,27	692 738	5,20	696 890	5,23	
	lubelskie	748 245	5,61	744 740	5,59	745 990	5,60	
	lubuskie	361 182	2,71	319 797	2,40	340 260	2,56	
	łódzkie	1 003 619	7,53	988 879	7,42	995 579	7,48	
	małopolskie	1 064 343	7,98	1 073 269	8,06	1 068 087	8,02	
	mazowieckie	1 853 326	13,90	1 917 421	14,40	1 884 106	14,15	
	opolskie	362 370	2,72	392 615	2,95	377 239	2,83	
	podkarpackie	637 782	4,78	637 572	4,79	637 248	4,79	
	podlaskie	427 299	3,20	447 898	3,36	437 304	3,28	
	pomorskie	727 256	5,45	702 413	5,27	714 354	5,36	
	śląskie	1 770 934	13,28	1 738 826	13,06	1 753 700	13,17	
	świętokrzyskie	423 011	3,17	362 402	2,72	392 442	2,95	
	warmińsko-mazurskie	482 582	3,62	552 185	4,15	517 036	3,88	
	wielkopolskie	1 062 550	7,97	1 120 946	8,42	1 091 014	8,19	
	zachodniopomorskie	605 897	4,54	607 715	4,56	606 398	4,55	
Najbliższego sąsiada, wszystkie	dolnośląskie	1 100 231	8,25	1 026 656	7,71	1 062 729	7,98	0,635
	kujawsko-pomorskie	701 980	5,27	716 884	5,38	708 955	5,32	
	lubelskie	748 245	5,61	740 248	5,56	743 746	5,59	
	lubuskie	361 182	2,71	362 108	2,72	361 402	2,71	
	łódzkie	1 003 619	7,53	1 023 524	7,68	1 012 890	7,61	
	małopolskie	1 064 343	7,98	1 070 153	8,03	1 066 530	8,01	
	mazowieckie	1 853 326	13,90	1 882 776	14,14	1 866 795	14,02	
	opolskie	362 370	2,72	385 302	2,89	373 585	2,81	
	podkarpackie	637 782	4,78	634 232	4,76	635 579	4,77	
	podlaskie	427 299	3,20	406 448	3,05	416 593	3,13	
	pomorskie	727 256	5,45	698 682	5,25	712 489	5,35	
	śląskie	1 770 934	13,28	1 741 943	13,08	1 755 257	13,18	
	świętokrzyskie	423 011	3,17	411 684	3,09	417 067	3,13	
	warmińsko-mazurskie	482 582	3,62	531 770	3,99	506 835	3,81	
	wielkopolskie	1 062 550	7,97	1 085 404	8,15	1 073 254	8,06	
	zachodniopomorskie	605 897	4,54	600 946	4,51	603 016	4,53	

Źródło: opracowanie własne

Graficzna analiza rozkładu cechy w zbiorze dawcy i biorcy (po integracji) wskazuje, że obie metody w sposób podobny odzwierciedlają rozkład dołączanej cechy (por. wykres 5.5).

Wykres 5.5. Rozkład brzegowy cechy „województwo”, wartości empiryczne i dołączone



Uwaga, kolor:
Niebieski – wartości empiryczne,
Zielony – wartości dołączone.
Źródło: opracowanie własne

Tabela 5.22. Rozkłady brzegowe cechy „czy gospodarstwo stać na tygodniowy urlop rocznie” przed i po integracji

Metoda	Czy gospodarstwo stać na tygodniowy urlop rocznie	Zbiór danych						Δ
		Dawca		Biorca		Zintegrowany		
		EU-SILC		BBGD				
		N	%	N	%	N	%	
Najbliższego sąsiada, CART	tak	4 502 121	33,80	4 042 930	30,32	4 269 653	32,06	1,74
	nie	8 816 639	66,20	9 289 675	69,68	9 047 069	67,94	
Najbliższego sąsiada, wszystkie	tak	4 502 121	33,80	4 114 229	30,86	4 305 278	32,33	1,47
	nie	8 816 639	66,20	9 218 376	69,14	9 011 444	67,67	

Źródło: opracowanie własne

Podobnie jak w przypadku zmiennej „województwo”, cecha dołączana „czy gospodarstwo stać na tygodniowy urlop rocznie” ze zbioru EU-SILC do zbioru BBGD zachowała dużą zbieżność z wartościami empirycznymi (por. tabela 5.22). Lepszy rezultat osiągnięto stosując metodę NND wykorzystującą wszystkie zmienne wspólne jako parujące. Warto również zauważyć, że zmienna była dołączana ze zbioru mniej licznego do bardziej licznego i osiągnięto dobre rezultaty. Wydaje się to wynikać z faktu, że cecha jakościowa posiada mało wariantów (w analizowanym przypadku dwa) i sztuczność rozkładu wynikająca z przyłączania wielokrotnie tych samych wartości zdaje się nie mieć wpływu na postać rozkładu cechy. W przypadku cech ilościowych (w szczególności ciągłych) wielokrotne dołączanie tej samej wartości zaburza własność ciągłości, to jest prawie zerowe prawdopodobieństwo wystąpienia dokładnie tej samej wartości.

Wnioski wyciągnięte na podstawie analizy zgodności rozkładu zmiennej „czy gospodarstwo stać na tygodniowy urlop rocznie”, potwierdza analiza graficzna. Rozkład wartości imputowanych (kolor niebieski) jest w dużej mierze zbliżony do rozkładu wartości empirycznych (por wykres 5.6).

Oceniając zgodność charakterystyk rozkładów zmiennych dołączanych w zbiorze zintegrowanym z wartościami empirycznymi, dla cech ilościowych posłużono się analizą struktury. Jako podstawę oceny zastosowano wskaźniki zgodności:

— dla średniej arytmetycznej: $\frac{\bar{x}_{int}}{\bar{x}_{emp}}$,

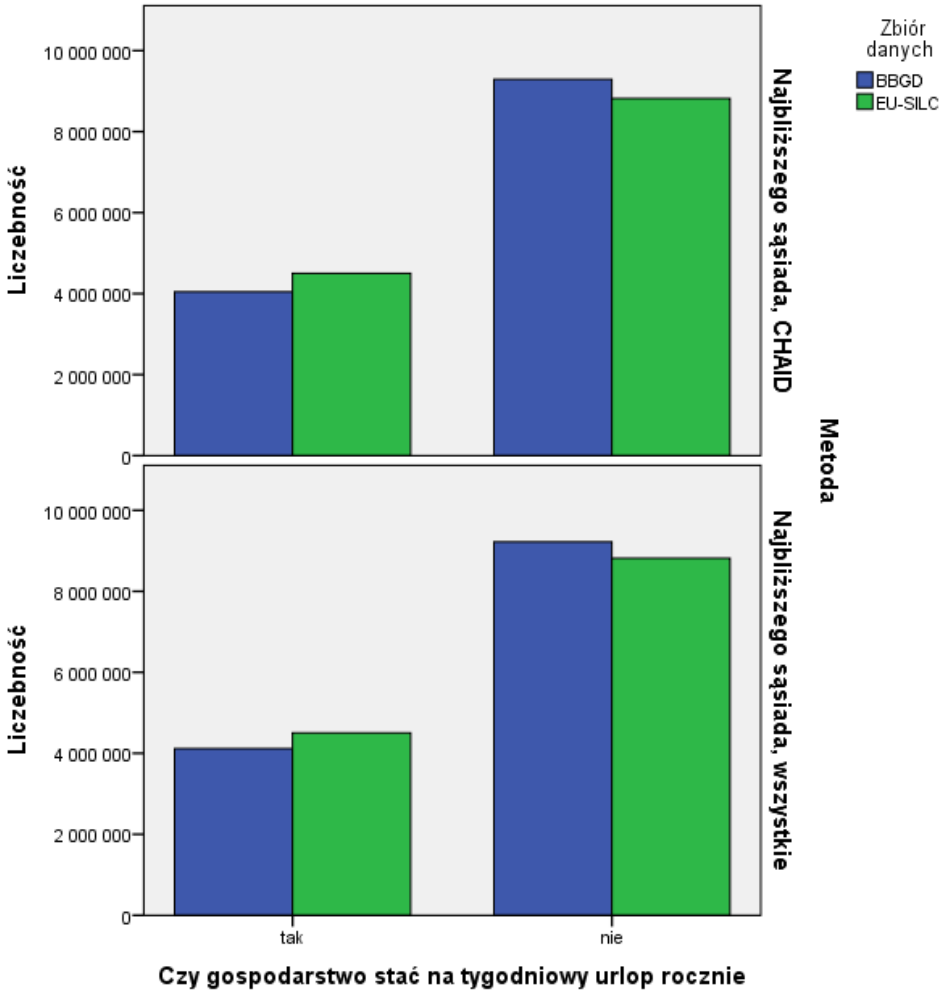
— dla wariancji: $\frac{\hat{s}^2_{int}}{\hat{s}^2_{emp}}$,

— dla odchylenia standardowego: $\frac{\hat{s}_{int}}{\hat{s}_{emp}}$,

gdzie subskrypt int oznacza dany parametr w zbiorze zintegrowanym, a emp w zbiorze dawcy (wartość empiryczna). Dokonano również porównania kształtu histogramów w zbiorze zintegrowanym i dawcy.

Dokonując integracji zmiennej „wydatki (ogółem) gospodarstwa domowego”, na podstawie każdej metody wartość średnia w zbiorze zintegrowanym była bardzo zbliżona do wartości empirycznej (por. tabela 5.23). Największą bezwzględną różnicę (kolumna $\left|1 - \frac{\bar{x}_{int}}{\bar{x}_{emp}}\right|$) zaobserwowano dla metody mieszanej PMM dla modelu dla całej Polski ogółem – 3,4%. Najniższą różnicę zwróciła otrzymano metodą imputacji stochastycznej dla modelu oszacowanego dla całej Polski – 0,3%.

Wykres 5.6. Rozkłady brzegowe cechy „czy gospodarstwo stać na tygodniowy urlop rocznie” przed i po integracji



Uwaga, kolor:
Niebieski – wartości dołączane,
Zielony – wartości empiryczne.
Źródło: opracowanie własne

Tabela 5.23. Charakterystyki rozkładów cechy „wydatki ogółem gospodarstw domowych” przed i po integracji

Metoda	Statystyka	Zbiór			$\frac{\bar{x}_{int}}{\bar{x}_{emp}}$	$\left\|1-\frac{\bar{x}_{int}}{\bar{x}_{emp}}\right\|$
		Dawca	Biorca	Zintegrowany		
		BBGD	EU-SILC			
Najbliższego sąsiada, CART	Średnia	1 954,20	1 982,21	1 968,20	1,007	0,007
	Mediana	1 602,67	1 629,39	1 615,31	--	
	Wariancja	2 271 514,07	2 086 521,57	2 179 261,95	0,959	0,041
	Odchylenie standardowe	1 507,15	1 444,48	1 476,23	0,979	0,021
	Minimum	102,20	102,20	102,20	--	
	Maksimum	45 500,96	26 642,00	45 500,96	--	

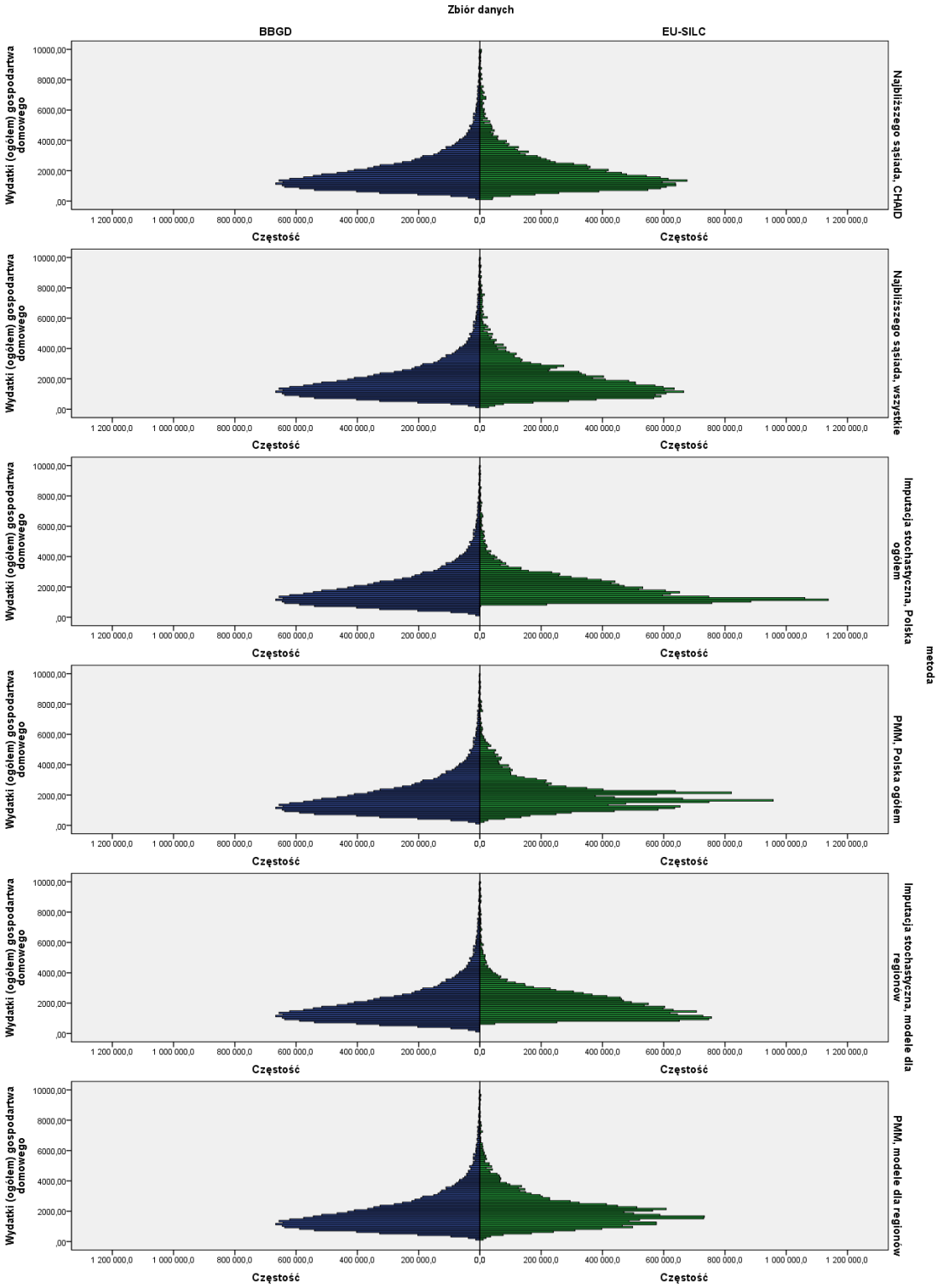
Metoda	Statystyka	Zbiór			$\frac{\overline{x}_{int}}{\overline{x}_{emp}}$	$\left\|1-\frac{\overline{x}_{int}}{\overline{x}_{emp}}\right\|$
		Dawca	Biorca	Zintegrowany		
		BBGD	EU-SILC			
	Rozstęp	45 398,76	26 539,80	45 398,76	--	
	Rozstęp międzykwartylowy	1 305,71	1 352,01	1 328,65	--	
	Skośność	5,22	3,83	4,57	--	
	Kurtoza	66,62	35,64	52,50	--	
Najbliższego sąsiada, wszystkie	Średnia	1 954,20	1 990,52	1 972,35	1,009	0,009
	Mediana	1 602,67	1 651,07	1 625,52	--	
	Wariancja	2 271 514,07	2 124 660,43	2 198 455,17	0,968	0,032
	Odchylenie standardowe	1 507,15	1 457,62	1 482,72	0,984	0,016
	Minimum	102,20	102,20	102,20	--	
	Maksimum	45 500,96	26 940,72	45 500,96	--	
	Rozstęp	45 398,76	26 838,52	45 398,76	--	
	Rozstęp międzykwartylowy	1 305,71	1 372,92	1 345,63	--	
	Skośność	5,22	4,32	4,79	--	
	Kurtoza	66,62	44,12	56,14	--	
Imputacja stochastyczna, Polska ogółem	Średnia	1 954,20	1 966,02	1 960,11	1,003	0,003
	Mediana	1 602,67	1 697,63	1 653,46	--	
	Wariancja	2 271 514,07	1 645 799,21	1 958 854,07	0,862	0,138
	Odchylenie standardowe	1 507,15	1 282,89	1 399,59	0,929	0,071
	Minimum	102,20	790,54	102,20	--	
	Maksimum	45 500,96	47 627,05	47 627,05	--	
	Rozstęp	45 398,76	46 836,52	47 524,85	--	
	Rozstęp międzykwartylowy	1 305,71	1 141,76	1 201,66	--	
	Skośność	5,22	10,33	7,23	--	
	Kurtoza	66,62	251,78	133,73	--	
PMM, Polska ogółem	Średnia	1 954,20	2 085,71	2 019,92	1,034	0,034
	Mediana	1 602,67	1 825,76	1 720,77	--	
	Wariancja	2 271 514,07	1 350 657,88	1 815 649,00	0,799	0,201
	Odchylenie standardowe	1 507,15	1 162,18	1 347,46	0,894	0,106
	Minimum	102,20	190,02	102,20	--	
	Maksimum	45 500,96	20 626,73	45 500,96	--	
	Rozstęp	45 398,76	20 436,70	45 398,76	--	
	Rozstęp międzykwartylowy	1 305,71	1 144,39	1 229,40	--	
	Skośność	5,22	2,45	4,40	--	
	Kurtoza	66,62	14,37	55,77	--	
Imputacja stochastyczna, modele dla regionów	Średnia	1 954,20	1 972,62	1 963,40	1,005	0,005
	Mediana	1 602,67	1 745,99	1 679,27	--	
	Wariancja	2 271 514,07	1 570 538,83	1 921 293,29	0,846	0,154
	Odchylenie standardowe	1 507,15	1 253,21	1 386,11	0,920	0,080

Metoda	Statystyka	Zbiór			$\frac{\overline{x}_{int}}{\overline{x}_{emp}}$	$\left\|1-\frac{\overline{x}_{int}}{\overline{x}_{emp}}\right\|$
		Dawca BBGD	Biorca EU-SILC	Zintegrowany		
	Minimum	102,20	607,30	102,20	--	
	Maksimum	45 500,96	38 068,56	45 500,96	--	
	Rozstęp	45 398,76	37 461,26	45 398,76	--	
	Rozstęp międzykwartylowy	1 305,71	1 185,75	1 244,76	--	
	Skośność	5,22	7,27	6,04	--	
	Kurtoza	66,62	124,72	88,31	--	
PMM, modele dla regionów	Średnia	1 954,20	2 072,19	2 013,17	1,030	0,030
	Mediana	1 602,67	1 845,29	1 724,49	--	
	Wariancja	2 271 514,07	1 404 695,19	1 841 810,27	0,811	0,189
	Odchylenie standardowe	1 507,15	1 185,20	1 357,13	0,900	0,100
	Minimum	102,20	147,72	102,20	--	
	Maksimum	45 500,96	17 389,38	45 500,96	--	
	Rozstęp	45 398,76	17 241,66	45 398,76	--	
	Rozstęp międzykwartylowy	1 305,71	1 225,45	1 277,83	--	
	Skośność	5,22	2,71	4,45	--	
	Kurtoza	66,62	17,83	55,58	--	

Uwaga:
Kolorem oznaczono najbardziej zbieżne współczynniki podobieństwa.
Źródło: opracowanie własne

Większe różnice zaobserwowano analizując zgodność miar dyspersji (por. tabela 5.23). Rozstęp wskaźnika zgodności dla wariancji wahał się od 3,2% (NND, wszystkie zmienne wspólne) do aż 20,1% (metoda mieszana, model dla Polski ogółem). Podobnie dużą zmienność zaobserwowano dla odchylenia standardowego – od 1,6% (NND, wszystkie zmienne) do 10,6% (PMM, Polska ogółem).

Wykres 5.7. Rozkłady zmiennej „wydatki (ogółem) gospodarstw domowych”, wartości empiryczne i dołączone



Uwaga, kolor:
Niebieski – wartości empiryczne,
Zielony – wartości dołączone.
Źródło: opracowanie własne

W kolejnym kroku porównano postaci histogramów rozkładu wartości empirycznych i wartości imputowanych. Celem tej analizy jest zweryfikowanie wniosków wypływających z analizy zgodności podstawowych parametrów rozkładu wartości empirycznych i w zbiorze zintegrowanym. Wykres 5.7 prezentuje histogramy wartości empirycznych (kolor niebieski) oraz imputowanych (kolor zielony). Kolejność porównywanych metod jest taka sama jak w tabeli 5.23. Dla lepszej ilustracji, wartości zmiennej na histogramach zostały ograniczone do 10000.

Postaci najbardziej zbieżne zaobserwowano dla metody najbliższego sąsiada. Metoda imputacji stochastycznej minimalne przeszacowały wydatki. Dodatkowo, model dla Polski ogółem charakteryzował się wyższą wysmukłością rozkładu. Metoda mieszana charakteryzowała się „postrzępieniem" rozkładu – część wartości była bardziej niż proporcjonalnie częściej imputowana.

Postaci histogramów rozkładu wskazują, że największą zgodnością z rozkładem empirycznym charakteryzują się wartości dołączane metodą najbliższego sąsiada (również inne charakterystyki rozkładu są bardzo zbliżone dla tych metod – por. tabela 5.23). Pomimo faktu, że metoda imputacji stochastycznej za pomocą modelu dla Polski ogółem charakteryzowała się bardzo zbliżoną średnią (0,3% różnicy) to różnice dla miar dyspersji oraz nieproporcjonalnie duża wysmukłość wydają się wykluczać tę metodę jako dobrze odzwierciedlającą rozkład empiryczny.

Tabela 5.24. Charakterystyki rozkładów cechy „dochody głowy gospodarstwa domowego" przed i po integracji

Metoda	**Statystyka**	**Zbiór**			$\frac{\overline{x}_{int}}{\overline{x}_{emp}}$	$\left\|1-\frac{\overline{x}_{int}}{\overline{x}_{emp}}\right\|$
		Dawca	**Biorca**	**Zintegrowany**		
		EU-SILC	**BBGD**			
Najbliższego sąsiada, CART	Średnia	2 065,48	1 868,64	1 967,01	0,952	0,048
	5% średnia obcięta	1 854,20	1 702,87	1 775,78	--	
	Mediana	1 566,00	1 483,82	1 525,90	--	
	Wariancja	3 144 682,13	2 027 856,49	2 595 665,26	0,825	0,175
	Odchylenie standardowe	1 773,33	1 424,03	1 611,11	0,909	0,091
	Minimum	0,00	0,00	0,00	--	
	Maksimum	23 775,96	23 775,96	23 775,96	--	
	Rozstęp	23 775,96	23 775,96	23 775,96	--	
	Rozstęp międzykwartylowy	1 516,64	1 264,19	1 380,58	--	
	Skośność	3,13	3,30	3,27	--	
	Kurtoza	16,86	20,56	19,01	--	

Metoda	Statystyka	Zbiór			$\frac{\overline{x}_{int}}{\overline{x}_{emp}}$	$\left\|1-\frac{\overline{x}_{int}}{\overline{x}_{emp}}\right\|$
		Dawca	Biorca	Zintegrowany		
		EU-SILC	BBGD			
Najbliższego sąsiada, wszystkie	Średnia	2 065,48	1 896,15	1 980,77	0,959	0,041
	5% średnia obcięta	1 854,20	1 733,72	1 791,54	--	
	Mediana	1 566,00	1 510,25	1 533,33	--	
	Wariancja	3 144 682,13	2 056 350,99	2 607 401,34	0,829	0,171
	Odchylenie standardowe	1 773,33	1 434,00	1 614,74	0,911	0,089
	Minimum	0,00	0,00	0,00	--	
	Maksimum	23 775,96	23 775,96	23 775,96	--	
	Rozstęp	23 775,96	23 775,96	23 775,96	--	
	Rozstęp międzykwartylowy	1 516,64	1 321,34	1 411,02	--	
	Skośność	3,13	3,24	3,24	--	
	Kurtoza	16,86	20,82	19,06	--	
Imputacja stochastyczna, Polska ogółem	Średnia	2 065,48	1 979,49	2 022,46	0,979	0,021
	5% średnia obcięta	1 854,20	1 821,36	1 835,18	--	
	Mediana	1 566,00	1 644,79	1 613,19	--	
	Wariancja	3 144 682,13	3 081 214,02	3 114 779,86	0,990	0,010
	Odchylenie standardowe	1 773,33	1 755,34	1 764,87	0,995	0,005
	Minimum	0,00	-82 271,15	-82 271,15	--	
	Maksimum	23 775,96	79 232,18	79 232,18	--	
	Rozstęp	23 775,96	161 503,32	161 503,32	--	
	Rozstęp międzykwartylowy	1 516,64	1 330,50	1 419,44	--	
	Skośność	3,13	3,03	3,08	--	
	Kurtoza	16,86	370,75	190,09	--	
PMM, Polska ogółem	Średnia	2 065,48	1 962,96	2 014,19	0,975	0,025
	5% średnia obcięta	1 854,20	1 773,40	1 811,63	--	
	Mediana	1 566,00	1 558,36	1 560,80	--	
	Wariancja	3 144 682,13	2 414 743,52	2 782 150,42	0,885	0,115
	Odchylenie standardowe	1 773,33	1 553,94	1 667,98	0,941	0,059
	Minimum	0,00	0,00	0,00	--	
	Maksimum	23 775,96	23 775,96	23 775,96	--	
	Rozstęp	23 775,96	23 775,96	23 775,96	--	
	Rozstęp międzykwartylowy	1 516,64	1 218,78	1 364,58	--	
	Skośność	3,13	4,55	3,74	--	
	Kurtoza	16,86	36,35	24,52	--	
Imputacja stochastyczna, modele dla regionów	Średnia	2 065,48	1 977,80	2 021,62	0,979	0,021
	5% średnia obcięta	1 854,20	1 820,25	1 834,47	--	
	Mediana	1 566,00	1 648,43	1 614,93	--	
	Wariancja	3 144 682,13	3 040 985,27	3 094 728,58	0,984	0,016
	Odchylenie standardowe	1 773,33	1 743,84	1 759,18	0,992	0,008

Metoda	Statystyka	Zbiór			$\frac{\overline{x}_{int}}{\overline{x}_{emp}}$	$\left\|1-\frac{\overline{x}_{int}}{\overline{x}_{emp}}\right\|$
		Dawca	Biorca	Zintegrowany		
		EU-SILC	BBGD			
	Minimum	0,00	-82 578,63	-82 578,63	--	
	Maksimum	23 775,96	73 786,61	73 786,61	--	
	Rozstęp	23 775,96	156 365,24	156 365,24	--	
	Rozstęp międzykwartylowy	1 516,64	1 320,12	1 413,81	--	
	Skośność	3,13	2,52	2,84	--	
	Kurtoza	16,86	353,28	179,38	--	
PMM, modele dla regionów	Średnia	2 065,48	1 960,55	2 012,99	0,975	0,025
	5% średnia obcięta	1 854,20	1 774,80	1 812,40	--	
	Mediana	1 566,00	1 577,40	1 572,95	--	
	Wariancja	3 144 682,13	2 339 993,75	2 744 881,14	0,873	0,127
	Odchylenie standardowe	1 773,33	1 529,70	1 656,77	0,934	0,066
	Minimum	0,00	0,00	0,00	--	
	Maksimum	23 775,96	23 775,96	23 775,96	--	
	Rozstęp	23 775,96	23 775,96	23 775,96	--	
	Rozstęp międzykwartylowy	1 516,64	1 221,56	1 351,38	--	
	Skośność	3,13	4,47	3,70	--	
	Kurtoza	16,86	36,56	24,44	--	

Uwaga:

Kolorem oznaczono najbardziej zbieżne współczynniki podobieństwa.

Źródło: opracowanie własne

Zmienna „dochód głowy gospodarstwa domowego" były dołączana ze zbioru mniej licznego (EU-SILC) do bardziej licznego (BBGD). W metodzie najbliższego sąsiada, niecałe 15000 rekordów ze zbioru dawcy zostały rozdystrybuowane w niemal 35000 rekordach w zbiorze biorcy. Uwidoczniły się podkreślane w literaturze (m.in. D'Orazio *et. al.* [2006]) wady – metoda nie odzwierciedla dobrze rozkładu empirycznego (m.in. ze względu na wielokrotne dołączanie tych samych wartości i wynikającą z tego faktu sztuczność rozkładu). O ile wartości średnie są w dużej mierze zbieżne z wartościami empirycznymi (niedoszacowanie rzędu 4 – 5 procent), o tyle odchylenie standardowe jest w każdym przypadku niedoszacowana aż o ok. 9% (a wariancja aż o ok. 17%). Nie należy się również spodziewać, że metody mieszane w sposób dobry odzwierciedlą rozkład empiryczny, ze względu na to, że za ich pomocą również dobiera się wartość empiryczną ze zbioru dawcy.

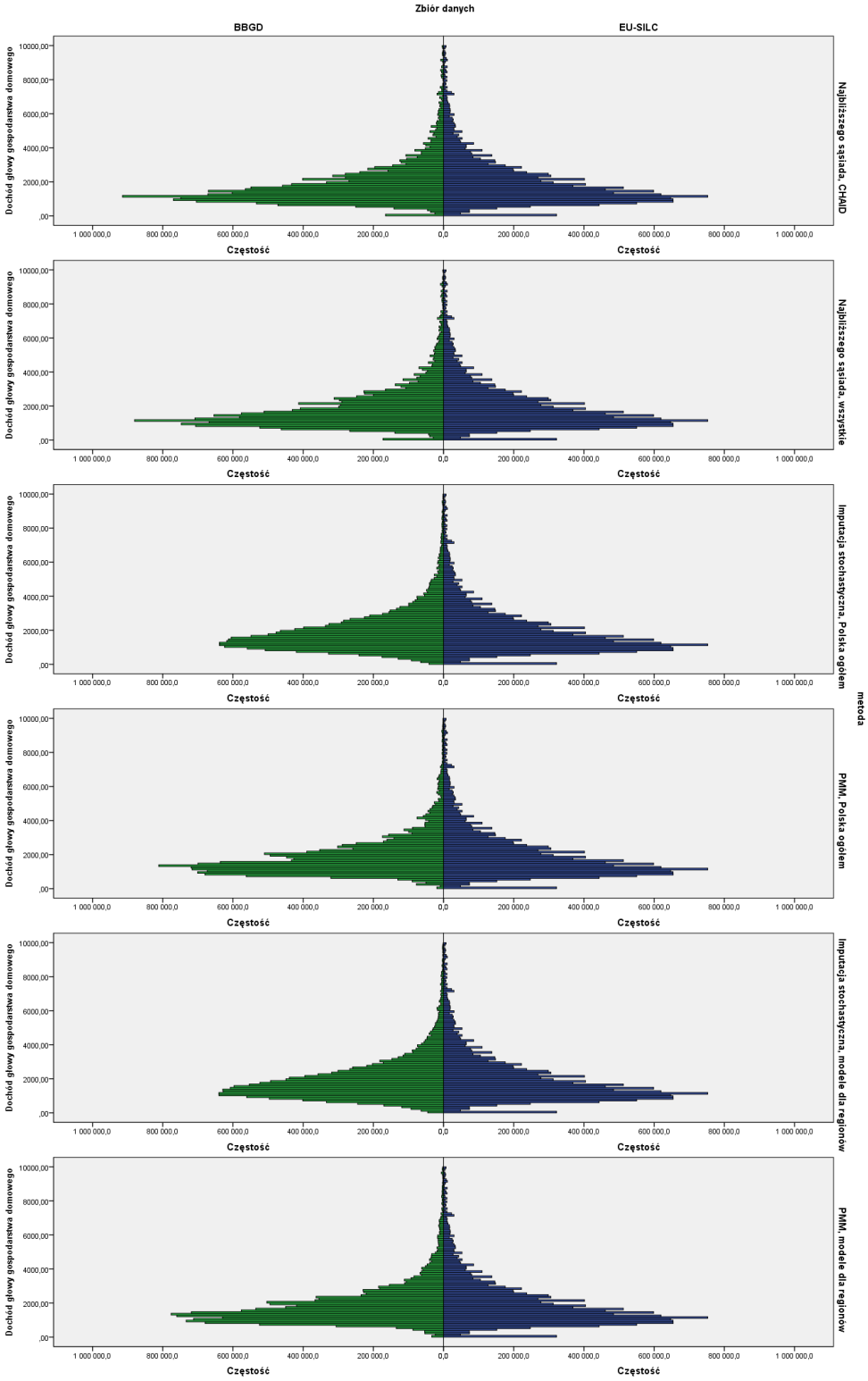
Wydaje się, że problem sztuczności rozkładu wartości dołączanych można rozwiązać stosując metodę imputacji stochastycznej. Wartości imputowane za jej pomocą są wartościami teoretycznymi (wynikającymi z modelu), co niweluje wpływ omawianej kwestii na postać

rozkładu wartości dołączanych. Pomimo faktu, że model[136] imputuje wartości spoza empirycznego obszaru zmienności cechy (wartości poniżej minimalnego i powyżej maksymalnego poziomu dochodu – por. tabela 5.24), w najlepszy wśród omawianych metod sposób odzwierciedla empiryczną wartość średniej arytmetycznej (niedoszacowanie rzędu zaledwie 2,1%) oraz miary dyspersji (niedoszacowanie wariancji o 1% oraz odchylenia standardowego o 0,5%).

Analiza postaci histogramów (por. wykres 5.8) wskazuje, że imputacja stochastyczna „wygładza" rozkład. Wielomodalność rozkładu empirycznego prawdopodobnie wynika ze skłonności respondentów do podawania „okrągłych" kwot dochodów. Choć metody NND i PMM lepiej odzwierciedlają tę skłonność, analiza rozkładów brzegowych przemawia za zastosowaniem wartości wynikających z imputacji stochastycznej w zintegrowanym zbiorze.

[136] Omawiany jest model dla Polski ogółem – charakteryzuje się większą zgodnością niż model dla regionów (por. tabela 5.24).

Wykres 5.8. Rozkłady zmiennej „dochody głowy gospodarstwa domowego”, wartości empiryczne i dołączone



Uwaga, kolor:
zielony – wartości dołączane,
niebieski – wartości empiryczne.
Źródło: opracowanie własne

Podsumowując porównanie rozkładów, najlepszą metodą integracji dla poszczególnych cech dołączanych są:

— wydatki (ogółem) gospodarstw domowych – metoda najbliższego sąsiada przy wykorzystaniu wszystkich zmiennych wspólnych jako parujących,

— dochody głowy gospodarstwa domowego – imputacja stochastyczna, model dla Polski ogółem,

— województwo - metoda najbliższego sąsiada przy wykorzystaniu wszystkich zmiennych wspólnych jako parujących,

— czy gospodarstwo stać na tygodniowy urlop rocznie – metoda najbliższego sąsiada przy wykorzystaniu wszystkich zmiennych wspólnych jako parujących.

Tabela 5.25. Miary zgodności dołączanych cech ilościowych, wybrane metody integracji

Zmienna dołączana	Metoda	Statystyka	$\frac{\bar{x}_{int}}{\bar{x}_{emp}}$	$\frac{\hat{s}_{int}}{\hat{s}_{emp}}$
wydatki	Losowa	**Średnia**	**1,042**	**1,007**
		Minimum	1,024	0,920
		Maksimum	1,059	1,095
	NND		**1,009**	**0,984**
dochody głowy	Losowa	**Średnia**	**0,966**	**0,940**
		Minimum	0,946	0,900
		Maksimum	0,989	1,016
	imputacja stochastyczna		**0,979**	**0,995**

Uwaga:
Dla metody losowej przedstawiono podstawowe charakterystyki dla 100 iteracji.
Źródło: opracowanie własne

Tabela 5.26. Miary zgodności Δ dołączanych cech jakościowych, wybrane metody integracji

Metoda	Statystyka	województwo	czy gospodarstwo stać na tygodniowy urlop rocznie
Losowa	**Średnia**	**2,010**	**2,737**
	Minimum	1,279	1,854
	Maksimum	3,008	3,665
NND		**0,635**	**1,470**

Uwaga:
Dla metody losowej przedstawiono podstawowe charakterystyki miary dla 100 iteracji.
Źródło: opracowanie własne

Miary zbieżności rozkładów wartości imputowanych przy wykorzystaniu „najlepszych" metod wskazują również, że osiągnięto generalnie lepsze rezultaty niż w przypadku integra-

cji losowej. Zarówno dla ilościowych zmiennych dołączanych (por. tabela 5.25), jak i jakościowych (por. tabela 5.26).

Kolejnym etapem jest porównanie rozkładów łącznych zmiennych dołączanych ze zmiennymi empirycznymi.

5.5.3. Ocena rozkładów łącznych

Dla zmiennych jakościowych przeprowadzono test niezależności χ^2. Na podstawie statystyki testowej testu, obliczono współczynnik kontyngencji C-Pearsona w zbiorze dawcy, zbiorze biorcy po integracji, a także w zbiorze zintegrowanym, zarówno dla zmiennej dołączanej „województwo" (por. tabela 5.27.) oraz „czy gospodarstwo stać na tygodniowy urlop rocznie" (por. tabela 5.28). Następnie obliczono współczynnik zgodności $\frac{C_{int}}{C_{emp}}$ dla współczynnika korelacji z każdą zmienną. Zarówno dla zmiennej dołączanej „województwo" (por. tabela 5.29), jak i „czy gospodarstwo stać na tygodniowy urlop rocznie" (por. tabela 5.30) najlepsze rezultaty osiągnięto za pomocą metody najbliższego sąsiada przy wykorzystaniu wszystkich zmiennych wspólnych jako parujących.

Tabela 5.27. Współczynnik kontyngencji C-Pearsona zmiennej dołączanej „województwo" z wybranymi zmiennymi wspólnymi

Zmienne wspólne	Wartości	Metoda				
		Wartość empiryczna	EU-SILC		Zintegrowany	
			Najbliższego sąsiada, CART	Najbliższego sąsiada, wszystkie	Najbliższego sąsiada, CART	Najbliższego sąsiada, wszystkie
Tytuł prawny do zajmowanego mieszkania	χ^2	464 624,41	478 643,61	505 511,83	446 053,68	459 281,11
	df	45	45	45	45	45
	C-Pearsona	0,18351	0,18625	0,19122	0,18003	0,18259
Czy gospodarstwo posiada komputer?	χ^2	68 034,55	55 292,55	48 373,25	52 346,96	53 168,91
	df	15	15	15	15	15
	C-Pearsona	0,07125	0,06430	0,06016	0,06257	0,06306
Czy gospodarstwo posiada pralkę?	χ^2	37 726,23	30 196,77	37 745,21	26 479,11	32 902,48
	df	15	15	15	15	15
	C-Pearsona	0,05312	0,04756	0,05316	0,04455	0,04965
Czy gospodarstwo posiada samochód?	χ^2	87 607,63	66 393,38	80 845,40	59 743,66	76 420,60
	df	15	15	15	15	15
	C-Pearsona	0,08080	0,07043	0,07768	0,06683	0,07554
Czy gospodarstwo posiada TV?	χ^2	29 846,98	69 030,09	89 602,88	36 922,06	50 001,76
	df	15	15	15	15	15
	C-Pearsona	0,04726	0,07181	0,08175	0,05258	0,06116
Czy jest łazienka ?	χ^2	236 191,05	237 469,04	284 838,74	225 709,68	252 363,20

Zmienne wspólne	Wartości	Metoda				
		Wartość empiryczna	EU-SILC		Zintegrowany	
			Najbliższego sąsiada, CART	Najbliższego sąsiada, wszystkie	Najbliższego sąsiada, CART	Najbliższego sąsiada, wszystkie
	df	15	15	15	15	15
	C-Pearsona	0,13194	0,13235	0,14470	0,12910	0,13638
Czy jest ustęp spłukiwany ?	χ^2	249 441,04	308 052,50	343 224,19	266 873,55	284 721,21
	df	15	15	15	15	15
	C-Pearsona	0,13552	0,15035	0,15850	0,14017	0,14468
Liczba pokoi	χ^2	314 699,64	423 600,49	391 045,33	343 375,31	323 366,48
	df	75	75	75	75	75
	C-Pearsona	0,15185	0,17566	0,16898	0,15859	0,15401
Rodzaj budynku	χ^2	1 058 910,91	1 012 355,82	1 067 378,60	1 005 850,95	1 043 927,62
	df	45	45	45	45	45
	C-Pearsona	0,27125	0,26578	0,27239	0,26501	0,26962
Typ biologiczny gospodarstwa domowego	χ^2	212 776,93	317 078,44	320 873,27	189 511,00	207 965,86
	df	105	105	105	105	105
	C-Pearsona	0,12533	0,15249	0,15338	0,11845	0,12400

Źródło: opracowanie własne

Tabela 5.28. Współczynnik kontyngencji C-Pearsona zmiennej dołączanej „czy gospodarstwo stać na tygodniowy urlop rocznie" z wybranymi zmiennymi wspólnymi

Zmienne wspólne	Wartości	Metoda				
		Wartość empiryczna	BBGD		Zintegrowany	
			Najbliższego sąsiada, CART	Najbliższego sąsiada, wszystkie	Najbliższego sąsiada, CART	Najbliższego sąsiada, wszystkie
Tytuł prawny do zajmowanego mieszkania	χ^2	19 214,55	13 594,22	27 456,92	14 313,45	20 363,91
	df	3	3	3	3	3
	C-Pearsona	0,03796	0,03192	0,04533	0,03277	0,03908
Czy gospodarstwo posiada komputer?	χ^2	1 302 004,36	675 053,90	1 094 601,39	981 088,00	1 207 003,78
	df	1	1	1	1	1
	C-Pearsona	0,29842	0,21953	0,27545	0,26195	0,28828
Czy gospodarstwo posiada pralkę?	χ^2	25 121,16	8 723,60	11 627,51	15 835,92	17 671,59
	df	1	1	1	1	1
	C-Pearsona	0,04339	0,02557	0,02952	0,03446	0,03640
Czy gospodarstwo posiada samochód?	χ^2	1 086 563,69	1 108 657,35	999 200,45	1 103 215,60	1 047 744,87
	df	1	1	1	1	1
	C-Pearsona	0,27464	0,27707	0,26404	0,27660	0,27007
Czy gospodarstwo posiada TV?	χ^2	51 002,59	372,77	130,07	17 038,90	13 939,93
	df	1	1	1	1	1
	C-Pearsona	0,06176	0,00529	0,00312	0,03575	0,03234
Czy jest łazienka ?	χ^2	472 457,16	405 172,83	384 678,87	437 539,35	427 108,20
	df	1	1	1	1	1

Zmienne wspólne	Wartości	Metoda				
		Wartość empiryczna	BBGD		Zintegrowany	
			Najbliższego sąsiada, CART	Najbliższego sąsiada, wszystkie	Najbliższego sąsiada, CART	Najbliższego sąsiada, wszystkie
	C-Pearsona	0,18509	0,17174	0,16746	0,17836	0,17628
Czy jest ustęp spłukiwany ?	χ^2	427 635,73	319 376,22	324 357,36	372 451,47	374 784,97
	df	1	1	1	1	1
	C-Pearsona	0,17638	0,15295	0,15411	0,16495	0,16545
Liczba pokoi	χ^2	283 533,00	309 393,31	321 433,26	299 273,61	306 104,93
	df	5	5	5	5	5
	C-Pearsona	0,14446	0,15060	0,15343	0,14830	0,14994
Rodzaj budynku	χ^2	196 346,59	152 036,14	153 706,12	167 627,36	170 174,47
	df	3	3	3	3	3
	C-Pearsona	0,12053	0,10618	0,10676	0,11150	0,11233
Typ biologiczny gospodarstwa domowego	χ^2	439 981,57	330 650,60	422 454,16	361 393,99	419 990,12
	df	7	7	7	7	7
	C-Pearsona	0,17882	0,15556	0,17525	0,16255	0,17486

Źródło: opracowanie własne

Jako miarę oceny zgodności rozkładów łącznych przyjęto wskaźniki zgodności:

— dla cech ilościowych: $\frac{\rho_{int}}{\rho_{emp}}$,

— dla cech jakościowych: $\frac{C_{int}}{C_{emp}}$,

gdzie ρ to współczynnik korelacji liniowej Pearsona, a C to współczynnik kontyngencji C-Pearsona.

Tabela 5.29. Współczynniki podobieństwa dla zmiennej „województwo” i wybranych zmiennych wspólnych

Zmienne wspólne	**Najbliższego sąsiada, CART**		**Najbliższego sąsiada, wszystkie**	
	$\frac{C_{int}}{C_{emp}}$	$\left\|1-\frac{C_{int}}{C_{emp}}\right\|$	$\frac{C_{int}}{C_{emp}}$	$\left\|1-\frac{C_{int}}{C_{emp}}\right\|$
Tytuł prawny do zajmowanego mieszkania	0,9810	1,8963%	0,9950	0,5001%
Czy gospodarstwo posiada komputer?	0,8782	12,1800%	0,8850	11,4960%
Czy gospodarstwo posiada pralkę?	0,8386	16,1369%	0,9346	6,5393%
Czy gospodarstwo posiada samochód?	0,8271	17,2851%	0,9349	6,5084%
Czy gospodarstwo posiada TV?	1,1126	11,2592%	1,2941	29,4116%
Czy jest łazienka ?	0,9785	2,1489%	1,0337	3,3658%
Czy jest ustęp spłukiwany ?	1,0343	3,4294%	1,0676	6,7619%
Liczba pokoi	1,0444	4,4363%	1,0142	1,4222%
Rodzaj budynku	0,9770	2,3033%	0,9940	0,6033%
Typ biologiczny gospodarstwa domowego	0,9451	5,4887%	0,9894	1,0613%
ŚREDNIE ODCHYLENIE	--	**7,6564**%	--	**6,7670**%

Źródło: opracowanie własne

Tabela 5.30. Współczynniki podobieństwa dla zmiennej „czy gospodarstwo stać na tygodniowy urlop rocznie” i wybranych zmiennych wspólnych

Zmienne wspólne	**Najbliższego sąsiada, CART**		**Najbliższego sąsiada, wszystkie**	
	$\frac{C_{int}}{C_{emp}}$	$\left\|1-\frac{C_{int}}{C_{emp}}\right\|$	$\frac{C_{int}}{C_{emp}}$	$\left\|1-\frac{C_{int}}{C_{emp}}\right\|$
Tytuł prawny do zajmowanego mieszkania	0,8633	13,6683%	1,0295	2,9509%
Czy gospodarstwo posiada komputer?	0,8778	12,2195%	0,9660	3,3962%
Czy gospodarstwo posiada pralkę?	0,7943	20,5697%	0,8390	16,0980%
Czy gospodarstwo posiada samochód?	1,0071	0,7123%	0,9834	1,6630%
Czy gospodarstwo posiada TV?	0,5788	42,1224%	0,5236	47,6436%
Czy jest łazienka ?	0,9636	3,6371%	0,9524	4,7565%
Czy jest ustęp spłukiwany ?	0,9352	6,4802%	0,9380	6,1957%
Liczba pokoi	1,0266	2,6584%	1,0380	3,7974%
Rodzaj budynku	0,9250	7,4971%	0,9319	6,8057%
Typ biologiczny gospodarstwa domowego	0,9090	9,1030%	0,9778	2,2199%
ŚREDNIE ODCHYLENIE	--	**11,8668**%	--	**9,5527**%

Źródło: opracowanie własne

Dla zmiennych ilościowych oszacowano współczynnik korelacji liniowej Pearsona w zbiorze dawcy, zbiorze biorcy po integracji, a także w zbiorze zintegrowanym. Następnie obliczono współczynnik zgodności $\frac{\rho_{int}}{\rho_{emp}}$ dla współczynnika korelacji z każdą zmienną.

Ponieważ w zbiorze występowały cztery ilościowe zmienne wspólne, obliczono średnie odchylenie współczynnika korelacji w zbiorze zintegrowanym od wartości empirycznych.

Tabela 5.31. Współczynniki korelacji liniowej Pearsona zmiennej „wydatki ogółem gospodarstw domowych" z wybranymi zmiennymi wspólnymi, wybrane metody integracji

Metoda	Zmienna	Zbiór danych			$\frac{\rho_{int}}{\rho_{emp}}$	$\left\|1-\frac{\rho_{int}}{\rho_{emp}}\right\|$	Średnie odchylenie
		BBGD (dawca)	EU-SILC (biorca)	Zintegrowany			
Najbliższego sąsiada, CART	Liczba osób	0,2685	0,2865	0,2773	1,0326	3,26%	5,38%
	Ekwiwalentna wielkość	0,2790	0,2920	0,2853	1,0228	2,28%	
	Dochód rozporządzalny	0,5880	0,6693	0,6271	1,0666	6,66%	
	Dochód ekwiwalentny	0,4840	0,5812	0,5292	1,0934	9,34%	
Najbliższego sąsiada, wszystkie	Liczba osób	0,2685	0,2820	0,2751	1,0246	2,46%	3,74%
	Ekwiwalentna wielkość	0,2790	0,2880	0,2834	1,0158	1,58%	
	Dochód rozporządzalny	0,5880	0,6409	0,6136	1,0436	4,36%	
	Dochód ekwiwalentny	0,4840	0,5522	0,5159	1,0658	6,58%	
Imputacja stochastyczna, Polska ogółem	Liczba osób	0,2685	0,3214	0,2919	1,0871	8,71%	13,71%
	Ekwiwalentna wielkość	0,2790	0,3332	0,3029	1,0860	8,60%	
	Dochód rozporządzalny	0,5880	0,8551	0,7071	1,2025	20,25%	
	Dochód ekwiwalentny	0,4840	0,6765	0,5677	1,1728	17,28%	
PMM, Polska ogółem	Liczba osób	0,2685	0,4024	0,3239	1,2062	20,62%	19,26%
	Ekwiwalentna wielkość	0,2790	0,4159	0,3359	1,2040	20,40%	
	Dochód rozporządzalny	0,5880	0,8723	0,7055	1,1999	19,99%	
	Dochód ekwiwalentny	0,4840	0,6750	0,5617	1,1604	16,04%	
Imputacja stochastyczna, modele dla regionów	Liczba osób	0,2685	0,3610	0,3092	1,1517	15,17%	17,40%
	Ekwiwalentna wielkość	0,2790	0,3693	0,3188	1,1427	14,27%	
	Dochód rozporządzalny	0,5880	0,8263	0,6922	1,1772	17,72%	
	Dochód ekwiwalentny	0,4840	0,7376	0,5927	1,2245	22,45%	
PMM, modele dla regionów	Liczba osób	0,2685	0,4396	0,3413	1,2709	27,09%	21,83%
	Ekwiwalentna wielkość	0,2790	0,4501	0,3521	1,2622	26,22%	
	Dochód rozporządzalny	0,5880	0,8271	0,6882	1,1704	17,04%	
	Dochód ekwiwalentny	0,4840	0,6826	0,5661	1,1696	16,96%	

Źródło: opracowanie własne

W przypadku zmiennej „wydatki (ogółem) gospodarstwa domowego" najmniejsze odchylenia rozkładów zaobserwowano w zbiorze zintegrowanym metodą najbliższego sąsiada przy wykorzystaniu wszystkich zmiennych wspólnych (por tabela 5.31). Dla metod imputacji stochastycznej i metody mieszanych stwierdzono „nieakceptowalnie" duże różnice w rozkładzie łącznym. Analiza ta potwierdza słuszność wyboru metody NND ze wszystkimi zmiennymi wspólnymi jako najlepszą metodę integracji dla zmiennej „wydatki".

Tabela 5.32. Współczynniki korelacji liniowej Pearsona zmiennej „dochody głowy gospodarstwa domowego" z wybranymi zmiennymi wspólnymi, wybrane metody integracji

Metoda	**Zmienna**	**Zbiór danych**			$\frac{\rho_{int}}{\rho_{emp}}$	$\left\|1-\frac{\rho_{int}}{\rho_{emp}}\right\|$	**Średnie odchylenie**
		EU-SILC	**BBGD**	**Zintegrowany**			
Najbliższego sąsiada, CART	Liczba osób	0,1541	0,1701	0,1601	1,0392	3,92%	6,37%
	Ekwiwalentna wielkość	0,1536	0,1738	0,1618	1,0530	5,30%	
	Dochód rozporządzalny	0,8865	0,7609	0,8236	0,9290	7,10%	
	Dochód ekwiwalentny	0,8537	0,7060	0,7756	0,9084	9,16%	
Najbliższego sąsiada, wszystkie	Liczba osób	0,1541	0,1674	0,1591	1,0326	3,26%	6,72%
	Ekwiwalentna wielkość	0,1536	0,1676	0,1592	1,0362	3,62%	
	Dochód rozporządzalny	0,8865	0,7241	0,8072	0,9105	8,95%	
	Dochód ekwiwalentny	0,8537	0,6705	0,7594	0,8895	11,05%	
Imputacja stochastyczna, Polska ogółem	Liczba osób	0,1541	0,1371	0,1456	0,9454	5,46%	5,37%
	Ekwiwalentna wielkość	0,1536	0,1328	0,1435	0,9340	6,60%	
	Dochód rozporządzalny	0,8865	0,9654	0,9262	1,0447	4,47%	
	Dochód ekwiwalentny	0,8537	0,9369	0,8959	1,0494	4,94%	
PMM, Polska ogółem	Liczba osób	0,1541	0,1521	0,1528	0,9920	0,80%	1,36%
	Ekwiwalentna wielkość	0,1536	0,1477	0,1508	0,9813	1,87%	
	Dochód rozporządzalny	0,8865	0,8729	0,8774	0,9897	1,03%	
	Dochód ekwiwalentny	0,8537	0,8336	0,8389	0,9826	1,74%	
Imputacja stochastyczna, modele dla regionów	Liczba osób	0,1541	0,1276	0,1410	0,9152	8,48%	6,80%
	Ekwiwalentna wielkość	0,1536	0,1234	0,1389	0,9040	9,60%	
	Dochód rozporządzalny	0,8865	0,9614	0,9240	1,0422	4,22%	
	Dochód ekwiwalentny	0,8537	0,9365	0,8954	1,0488	4,88%	
PMM, modele dla regionów	Liczba osób	0,1541	0,1431	0,1486	0,9649	3,51%	2,99%
	Ekwiwalentna wielkość	0,1536	0,1388	0,1466	0,9545	4,55%	
	Dochód rozporządzalny	0,8865	0,8647	0,8731	0,9848	1,52%	
	Dochód ekwiwalentny	0,8537	0,8236	0,8334	0,9762	2,38%	

Źródło: opracowanie własne

Analiza wybranych rozkładów łącznych wskazuje, że najbardziej zbieżne rezultaty osiągnięto przy zastosowaniu metody PMM (por. tabela 5.32). Dołączanie wartości empirycznych

do zbioru biorcy wydaje się zachowywać w lepszy sposób rozkłady łączne. Jednak ze względu na wspominany wcześniej fakt „sztuczności rozkładu" przy dołączaniu ilościowych wartości empirycznych ze zbioru mniejszego do większego, bardziej słusznym wydaje się zastosowanie metod imputujących wartości teoretyczne. Przeciętna różnica we wskaźnikach zbieżności dla imputacji stochastycznej (model dla całego kraju) wynosiła 5,37% odchylenie to można uznać za akceptowalne[137].

Analiza rozkładów łącznych wykazała, że metody uznane za najlepsze w procesie integracji przy ocenie charakterystyk rozkładów w zintegrowanym zbiorze zachowały również najbardziej zbieżne rozkłady łączne. Wyjątkiem jest metoda imputacji stochastycznej, która w sposób nieco gorszy odtwarza łączne rozkłady dla cechy „dochody głów" niż metoda PMM.

5.6. Ocena realizacji celów badania empirycznego i hipotez badawczych

Zbiory danych BBGD i EU-SILC z dołączonymi wartościami zmiennych poddano konkatenacji w taki sposób, że zintegrowany zbiór charakteryzował się łączną liczebnością będącą sumą liczebności obu zbiorów wejściowych, tj. 49681 jednostek (por. tabela 5.33). Był on liczniejszy od zbioru BBGG o 30%, natomiast od zbioru EU-SILC aż o 70%. Wagi analityczne w zintegrowanym zbiorze utworzono na podstawie wag każdego ze zbiorów poprzez operację harmonizacji zaproponowaną przez Rubina (wzór 4.54). Liczebność populacji generalnej ustalono na poziomie 13 316 722 gospodarstw domowych (średnia arytmetyczna liczebności wynikającej z sumy wag analitycznych w wejściowych zbiorach, por. tabela 5.1).

Tabela 5.33. Liczebność zbiorów wejściowych i zintegrowanego

Zbiór	Liczebność
BBGD	34767
EU-SILC	14914
Zintegrowany	49681

Źródło: opracowanie własne

Zintegrowany wybranymi metodami parowania statystycznego zbiór charakteryzował się zgodnymi rozkładami (brzegowymi i łącznymi) zarówno zmiennych wspólnych, jak i dołączanych. Zwiększona, dzięki konkatencji, próba rodzi podstawy do przypuszczeń,

[137] W literaturze nie istnieją wytyczne dotyczące „wystarczającego" stopnia zbieżności.

że nowoutworzony, zintegrowany zbiór danych, oprócz łącznej obserwacji cech dołączanych, stwarza możliwość pozyskania szacunków o lepszej jakości, aniżeli niezależnie w którymkolwiek ze zbiorów wejściowych. Dodatkowo, dołączenie zmiennej „województwo" do zbioru EU-SILC umożliwia szacunki średniego dochodu głowy gospodarstwa domowego oraz frakcji gospodarstw, które stać na tygodniowy urlop poza domem w ujęciu NUTS 2, co w zbiorze wejściowym nie było możliwe.

Ocena precyzja szacunków na poziomie NUTS 1 w zintegrowanym zbiorze

Dokonano porównania jakości estymatorów dla cech wspólnych i dołączanych w zbiorach wejściowych i zintegrowanym obliczając:

- dla zmiennych ilościowych (wspólnych i dołączanych) – błąd standardowy szacunku średniej arytmetycznej w ujęciu regionów (NUTS 1),
- dla zmiennych jakościowych (wspólnych i dołączanych) – błąd standardowy frakcji dla każdego wariantu w ujęciu regionów (NUTS 1).

Jako miarę poprawy (lub pogorszenia) jakości szacunków dla zmiennych ilościowych zaproponowano współczynniki $1-\frac{s_{int}(\bar{x})}{s_{BBGD}(\bar{x})}$ oraz $1-\frac{s_{int}(\bar{x})}{s_{EU-SILC}(\bar{x})}$, gdzie $s_{int}(\bar{x})$ oznacza błąd standardowy szacunku średniej arytmetycznej w zbiorze zintegrowanym, a $s_{BBGD}(\bar{x})$ oraz $s_{EU-SILC}(\bar{x})$ oznaczają błąd standardowy szacunku średniej arytmetycznej w zbiorze, odpowiednio, BBGD i EU-SILC. Dla zmiennych jakościowych zastosowano współczynniki $1-\frac{s_{int}(p)}{s_{BBGD}(p)}$ oraz $1-\frac{s_{int}(p)}{s_{EU-SILC}(p)}$, gdzie $s_{int}(p)$ oznacza błąd standardowy szacunku frakcji w zbiorze zintegrowanym, a $s_{BBGD}(p)$ i $s_{EU-SILC}(p)$ oznaczają błąd standardowy szacunku frakcji w zbiorze, odpowiednio, BBGD i EU-SILC. Wartości współczynników większe od 0 oznaczają zysk na jakości estymatora w zbiorze zintegrowanym.

Tabela 5.34. Bezwzględne i względne błędy szacunków wartości przeciętnych ilościowych zmiennych dołączanych w ujęciu regionów

Zmienna	Region	BBGD			EU-SILC			zintegrowany			$1-\frac{s_{int}(\bar{x})}{s_{emp}(\bar{x})}$
		$\bar{x}$	$S(\bar{x})$	CV	$\bar{x}$	$S(\bar{x})$	CV	$\bar{x}$	$S(\bar{x})$	CV	
Wydatki (ogółem) gospodarstwa domowego	centralny	1916	16,40	0,86	2130	29,48	1,38	2024	14,95	0,74	0,088
	południowy	2008	16,44	0,82	2072	27,59	1,33	2040	14,45	0,71	0,121
	wschodni	1970	19,27	0,98	2040	29,12	1,43	2005	16,19	0,81	0,160
	północno-zachodni	1973	18,20	0,92	2147	37,72	1,76	2061	18,11	0,88	0,005
	południowo-zachodni	2138	27,10	1,27	2121	44,00	2,07	2130	23,27	1,09	0,141
	północny	2058	21,28	1,03	1994	27,64	1,39	2026	16,51	0,82	0,224
Dochód głowy gospodarstwa domowego	centralny	1831	18,40	1,00	2270	34,05	1,50	2052	17,16	0,84	0,168
	południowy	2005	21,04	1,05	2097	42,49	2,03	2051	20,63	1,01	0,041
	wschodni	1928	18,61	0,97	2017	26,84	1,33	1972	15,27	0,77	0,198
	północno-zachodni	2000	20,86	1,04	2092	38,55	1,84	2046	19,31	0,94	0,095
	południowo-zachodni	2111	34,24	1,62	2134	37,40	1,75	2122	24,96	1,18	0,275
	północny	2108	25,16	1,19	1990	30,51	1,53	2049	18,99	0,93	0,223

Uwaga, kolory:

szary – szacunki na podstawie wartości dołączanych,

zielony – zysk na jakości szacunków w zbiorze zintegrowanym względem danego zbioru wejściowego.

Miara zysku definiowana jest jako $1-\frac{s_{int}(\bar{x})}{s_{BBGD}(\bar{x})}$ lub $1-\frac{s_{int}(\bar{x})}{s_{EU-SILC}(\bar{x})}$.

Źródło: opracowanie własne

Tabela 5.35. Frakcja gospodarstw domowych, które „stać na tygodniowy urlop rocznie” według regionów wraz z oceną precyzji szacunku

Region	Wariant cechy	BBGD			EU-SILC			Zintegrowany			$1 - \frac{s_{int}(p)}{s_{EU-SILC}(p)}$
		p	$S(p)$	CV	p	$S(p)$	CV	p	$S(p)$	CV	
centralny	tak	0,3355	0,0095	2,8377	0,3763	0,0155	4,1081	0,3561	0,0082	2,2923	0,472
	nie	0,6645	0,0068	1,0260	0,6237	0,0108	1,7287	0,6439	0,0058	0,9008	0,462
południowy	tak	0,3511	0,0095	2,6959	0,3808	0,0144	3,7688	0,3659	0,0079	2,1679	0,447
	nie	0,6489	0,0069	1,0710	0,6192	0,0109	1,7671	0,6341	0,0059	0,9289	0,462
wschodni	tak	0,2490	0,0112	4,4813	0,2580	0,0167	6,4756	0,2534	0,0093	3,6693	0,443
	nie	0,7510	0,0064	0,8514	0,7420	0,0094	1,2639	0,7466	0,0053	0,7091	0,436
północno-zachodni	tak	0,3266	0,0112	3,4229	0,3302	0,0177	5,3562	0,3284	0,0095	2,8791	0,465
	nie	0,6734	0,0078	1,1653	0,6698	0,0121	1,8022	0,6716	0,0066	0,9802	0,455
południowo-zachodni	tak	0,3073	0,0136	4,4219	0,3767	0,0209	5,5561	0,3414	0,0115	3,3798	0,449
	nie	0,6927	0,0090	1,2921	0,6233	0,0151	2,4259	0,6586	0,0078	1,1861	0,483
północny	tak	0,2569	0,0120	4,6904	0,2895	0,0188	6,5105	0,2733	0,0102	3,7465	0,457
	nie	0,7431	0,0072	0,9641	0,7105	0,0114	1,6110	0,7267	0,0061	0,8433	0,465

Uwaga, kolory:

szary – szacunki na podstawie wartości dołączanych,

zielony – zysk na jakości szacunków w zbiorze zintegrowanym względem danego zbioru wejściowego.

Miara zysku definiowana jest jako $1 - \frac{s_{int}(p)}{s_{EU-SILC}(p)}$.

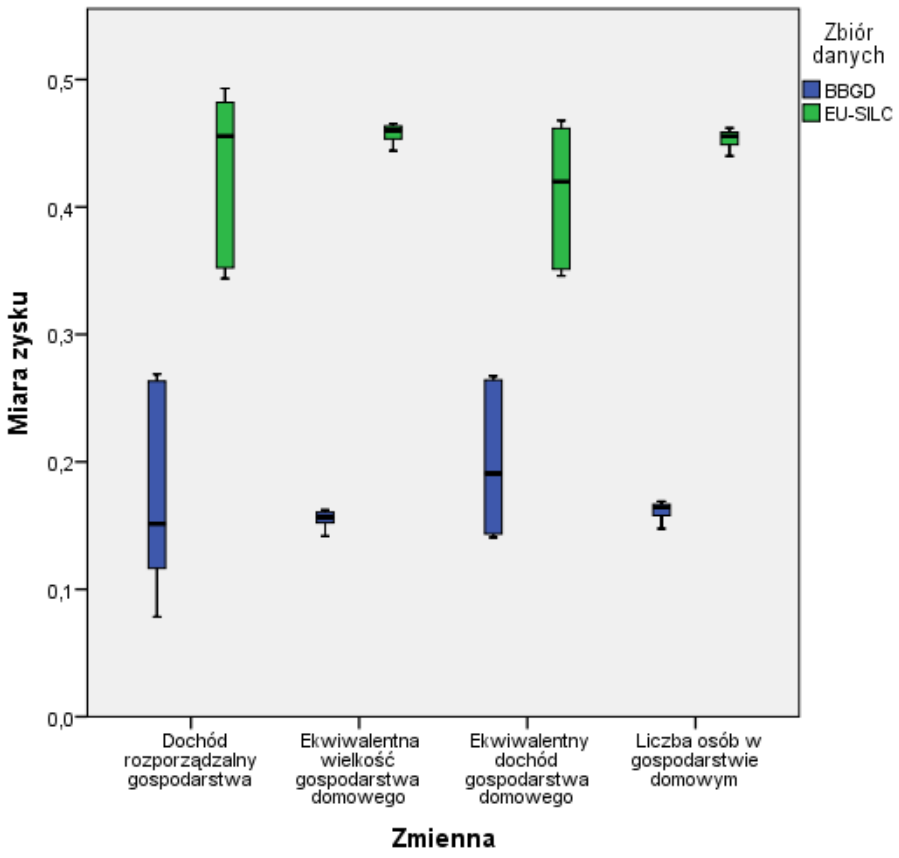
Źródło: opracowanie własne

Analizując rezultaty w przekroju regionów (NUTS 1) dla dołączanych zmiennych ilościowych, uzyskano poprawę jakości estymatora średniej arytmetycznej w dla każdej zmiennej w każdym regionie. W zbiorze zintegrowanym dla wydatków gospodarstw domowych największy zysk wyniósł 0,224 (region północny), zaś najmniejszy 0,005 (region północno-zachodni) (por. tabela 5.34).

Dla zmiennej jakościowej „czy gospodarstwo stać na tygodniowy urlop rocznie" w przekroju regionów w zintegrowanym zbiorze również zaobserwowano wzrost jakości w porównaniu z wartościami empirycznymi (EU-SILC). Zysk na jakości dla każdego regionu i wariantu był wysoki i wynosił od 0,436 do 0,483 (por. tabela 5.35).

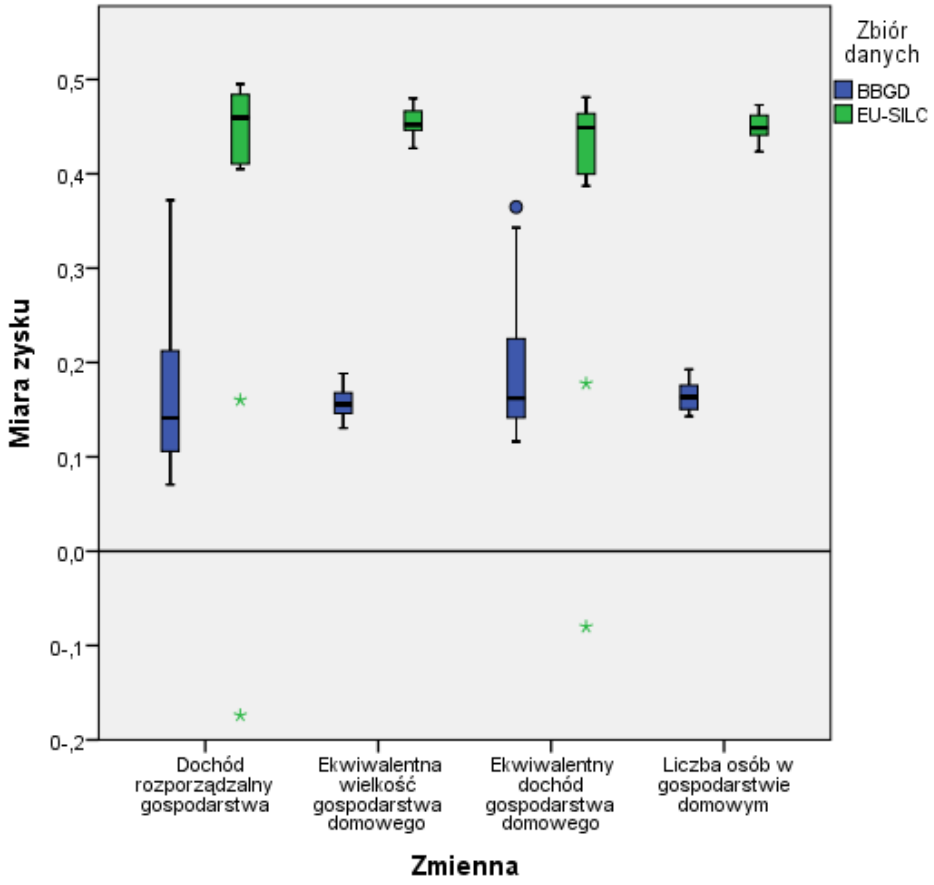
Jednocześnie pozytywnie zweryfikowano hipotezę, że możliwe jest utworzenie tabeli kontyngencji możliwości gospodarstwa na sfinansowanie tygodniowego urlopu poza miejscem zamieszkania w ujęciu województw (por. tabela 5.37).

Wykres 5.9. Rozkład miary zysku dla wspólnych cech ilościowych w ujęciu zbiorów wejściowych i regionów



Uwaga:
Dane do wykresu pobrano z tabeli A2 znajdującej się w aneksie.
Źródło: opracowanie własne

Wykres 5.10. Rozkład miary zysku dla wspólnych cech ilościowych w ujęciu zbiorów wejściowych i województw



Uwaga:
Dane do wykresu pobrano z tabeli A3 znajdującej się w aneksie.
Źródło: opracowanie własne

Dla ilościowych zmiennych wspólnych zaobserwowano podobną sytuację jak w przypadku cech dołączanych. W ujęciu regionów dla każdej zmiennej zaobserwowano zysk na jakości szacunku w zintegrowanym zbiorze (por. wykres 5.9), przy czym był on dużo wyższy w odniesieniu do zbioru EU-SILC (co wynika z jego mniejszej liczebności niż BBGD). W ujęciu województw w zdecydowanej większości przypadków również stwierdzono zysk. Tylko w dwóch przypadkach zanotowano stratę na jakości szacunku (por. wykres 5.10), co było spowodowane zwiększoną dyspersją cechy w zbiorze zintegrowanym.

Ocena precyzji szacunku na poziomie NUTS 2 w zintegrowanym zbiorze

W ujęciu województw (NUTS 2) zaobserwowano podobne przeciętne zyski na jakości estymatora średniej, zarówno dla zmiennych wspólnych, jak i dołączanych, w każdym ujęciu terytorialnym (por. tabela 5.35). W zaledwie dwóch przypadkach, dla zmiennej

„wydatki", zaobserwowano spadek jakości (spowodowany większą dyspersją zmiennej w zbiorze zintegrowanym).

Dla cechy jakościowej „czy gospodarstwo stać na tygodniowy urlop rocznie" możliwa była obserwacja w ujęciu województw w zintegrowanym zbiorze. Dodatkowo, w odniesieniu do zbioru wejściowego (EU-SILC), nastąpiła istotna poprawa jakości oszacowania. Został osiągnięty zysk na poziomie między 0,43 a 0,49 (por .tabela 5.36).

Niskie względne błędy szacunków – w zdecydowanej większości poniżej 2%, stwarzają przesłanki do próby szacowania charakterystyk wydatków gospodarstw domowych oraz dochodów ich głów w przekroju województw. W roku badania (2005) szacunki na tak niskim poziomie agregacji przestrzennej nie były publikowane.

Tabela 5.36. Estymacja wartości przeciętnych dołączanych zmiennych ilościowych według województw wraz z oceną precyzji szacunku

Zmienna	Województwo	BBGD			EU-SILC			zintegrowany			$1-\frac{s_{int}(\bar{x})}{s_{emp}(\bar{x})}$
		$\bar{x}$	$S(\bar{x})$	CV	$\bar{x}$	$S(\bar{x})$	CV	$\bar{x}$	$S(\bar{x})$	CV	
Wydatki (ogółem) gospodarstwa domowego	dolnośląskie	2112	28,65	1,36	2123	50,67	2,39	2117	25,61	1,21	0,106
	kujawsko-pomorskie	2100	36,62	1,74	1946	44,69	2,30	2022	27,75	1,37	0,242
	lubelskie	1952	35,47	1,82	2097	56,94	2,72	2024	30,88	1,53	0,130
	lubuskie	1999	42,23	2,11	2113	78,22	3,70	2056	39,16	1,90	0,073
	łódzkie	1925	26,84	1,39	2058	44,50	2,16	1992	23,36	1,17	0,130
	małopolskie	1982	25,96	1,31	2186	48,23	2,21	2085	24,22	1,16	0,067
	mazowieckie	1912	20,68	1,08	2169	38,52	1,78	2042	19,27	0,94	0,068
	opolskie	2218	65,54	2,95	2116	87,94	4,16	2166	51,70	2,39	0,211
	podkarpackie	2006	34,96	1,74	1977	40,41	2,04	1991	26,04	1,31	0,255
	podlaskie	1994	47,54	2,38	2057	54,86	2,67	2024	35,41	1,75	0,255
	pomorskie	2026	34,40	1,70	2055	47,17	2,29	2040	27,25	1,34	0,208
	śląskie	2024	21,21	1,05	2002	33,16	1,66	2013	17,96	0,89	0,153
	świętokrzyskie	1924	37,10	1,93	2019	81,87	4,05	1971	38,97	1,98	-0,050
	warmińsko-mazurskie	2046	39,76	1,94	1978	52,57	2,66	2011	31,18	1,55	0,216
	wielkopolskie	1959	24,51	1,25	2107	45,23	2,15	2034	22,72	1,12	0,073
	zachodniopomorskie	1983	35,24	1,78	2241	88,03	3,93	2111	39,72	1,88	-0,127
Dochód głowy gospodarstwa domowego	dolnośląskie	2086	31,35	1,50	2090	41,00	1,96	2088	24,38	1,17	0,405
	kujawsko-pomorskie	2181	41,34	1,90	1851	39,79	2,15	2014	29,02	1,44	0,271
	lubelskie	1930	32,61	1,69	2093	51,76	2,47	2011	28,23	1,40	0,455
	lubuskie	1957	47,73	2,44	2168	99,12	4,57	2063	47,55	2,31	0,520
	łódzkie	1842	31,69	1,72	2109	47,85	2,27	1977	26,36	1,33	0,449
	małopolskie	1967	35,27	1,79	2252	55,58	2,47	2110	30,05	1,42	0,459
	mazowieckie	1825	22,57	1,24	2358	45,57	1,93	2094	22,27	1,06	0,511

Zmienna	Województwo	BBGD			EU-SILC			zintegrowany			$1-\frac{s_{int}(\bar{x})}{s_{emp}(\bar{x})}$
		$\bar{x}$	$S(\bar{x})$	CV	$\bar{x}$	$S(\bar{x})$	CV	$\bar{x}$	$S(\bar{x})$	CV	
	opolskie	2186	98,49	4,51	2252	81,96	3,64	2220	65,72	2,96	0,198
Dochód głowy gospodarstwa domowego	podkarpackie	1942	34,01	1,75	1927	43,65	2,27	1935	26,46	1,37	0,394
	podlaskie	1957	45,84	2,34	2123	61,91	2,92	2038	36,41	1,79	0,412
	pomorskie	2041	37,23	1,82	2169	61,63	2,84	2104	32,22	1,53	0,477
	śląskie	2028	26,16	1,29	2002	59,38	2,97	2015	27,70	1,37	0,533
	świętokrzyskie	1875	39,95	2,13	1913	58,43	3,05	1894	32,92	1,74	0,437
	warmińsko-mazurskie	2104	56,08	2,67	1943	54,41	2,80	2020	38,95	1,93	0,284
	wielkopolskie	1981	27,52	1,39	1962	45,11	2,30	1971	23,83	1,21	0,472
	zachodniopomorskie	2060	41,97	2,04	2281	84,39	3,70	2170	40,85	1,88	0,516

Uwaga, kolory:

szary – szacunki na podstawie wartości dołączanych,

czerwony – strata na jakości szacunków w zbiorze zintegrowanym względem danego zbioru wejściowego,

zielony – zysk na jakości szacunków w zbiorze zintegrowanym względem danego zbioru wejściowego.

Miara zysku definiowana jest jako $1-\frac{s_{int}(\bar{x})}{s_{BBGD}(\bar{x})}$ lub $1-\frac{s_{int}(\bar{x})}{s_{EU-SILC}(\bar{x})}$.

Źródło: opracowanie własne

Tabela 5.37. Frakcja gospodarstw domowych, które „stać na tygodniowy urlop rocznie” według województw i oceny precyzji szacunku

Województwo	Wariant cechy	BBGD			EU-SILC			Zintegrowany			$1-\frac{s_{int}(p)}{s_{EU-SILC}(p)}$
		p	$S(p)$	CV	p	$S(p)$	CV	p	$S(p)$	CV	
dolnośląskie	tak	0,3066	0,0158	5,1525	0,3788	0,0248	6,5430	0,3415	0,0135	3,9518	0,456
	nie	0,6934	0,0104	1,4988	0,6212	0,0179	2,8786	0,6585	0,0091	1,3847	0,490
kujawsko-pomorskie	tak	0,2530	0,0201	7,9344	0,2643	0,0314	11,8867	0,2587	0,0170	6,5592	0,460
	nie	0,7470	0,0118	1,5752	0,7357	0,0181	2,4552	0,7413	0,0099	1,3341	0,453
lubelskie	tak	0,2594	0,0192	7,4019	0,2657	0,0280	10,5364	0,2625	0,0159	6,0400	0,433
	nie	0,7406	0,0112	1,5157	0,7343	0,0162	2,2110	0,7375	0,0092	1,2537	0,431
lubuskie	tak	0,3176	0,0273	8,6078	0,2962	0,0435	14,6966	0,3069	0,0231	7,5141	0,470
	nie	0,6824	0,0189	2,7644	0,7038	0,0280	3,9778	0,6931	0,0156	2,2495	0,443
łódzkie	tak	0,2904	0,0165	5,6854	0,3048	0,0273	8,9458	0,2977	0,0142	4,7608	0,480
	nie	0,7096	0,0107	1,5011	0,6952	0,0165	2,3755	0,7023	0,0090	1,2785	0,456
małopolskie	tak	0,3351	0,0157	4,6874	0,3535	0,0240	6,7781	0,3444	0,0132	3,8256	0,450
	nie	0,6649	0,0109	1,6382	0,6465	0,0170	2,6361	0,6556	0,0092	1,4038	0,460
mazowieckie	tak	0,3600	0,0116	3,2309	0,4152	0,0187	4,4953	0,3878	0,0099	2,5646	0,467
	nie	0,6400	0,0088	1,3737	0,5848	0,0140	2,3909	0,6122	0,0075	1,2248	0,464
opolskie	tak	0,3094	0,0266	8,6113	0,3710	0,0391	10,5267	0,3411	0,0223	6,5222	0,430
	nie	0,6906	0,0176	2,5500	0,6290	0,0283	4,5021	0,6589	0,0151	2,2986	0,465
podkarpackie	tak	0,2306	0,0209	9,0759	0,2488	0,0310	12,4745	0,2397	0,0174	7,2769	0,438
	nie	0,7694	0,0114	1,4803	0,7512	0,0169	2,2454	0,7603	0,0095	1,2477	0,438
podlaskie	tak	0,2928	0,0251	8,5809	0,2995	0,0384	12,8329	0,2961	0,0211	7,1121	0,452
	nie	0,7072	0,0164	2,3129	0,7005	0,0244	3,4904	0,7039	0,0136	1,9337	0,443
pomorskie	tak	0,2749	0,0194	7,0678	0,3128	0,0312	9,9712	0,2935	0,0166	5,6696	0,466
	nie	0,7251	0,0120	1,6575	0,6872	0,0199	2,9004	0,7065	0,0104	1,4704	0,479
śląskie	tak	0,3607	0,0119	3,2875	0,3975	0,0179	4,5047	0,3789	0,0099	2,6204	0,445
	nie	0,6393	0,0090	1,4097	0,6025	0,0142	2,3625	0,6211	0,0077	1,2318	0,463

Województwo	Wariant cechy	BBGD			EU-SILC			Zintegrowany			$1 - \frac{s_{int}(p)}{s_{EU-SILC}(p)}$
		p	$S(p)$	CV	p	$S(p)$	CV	p	$S(p)$	CV	
świętokrzyskie	tak	0,2139	0,0260	12,1738	0,2171	0,0410	18,8938	0,2155	0,0220	10,2135	0,464
	nie	0,7861	0,0136	1,7320	0,7829	0,0199	2,5456	0,7845	0,0113	1,4344	0,435
warmińsko-mazurskie	tak	0,2353	0,0238	10,1086	0,2929	0,0359	12,2447	0,2655	0,0202	7,5993	0,437
	nie	0,7647	0,0136	1,7775	0,7071	0,0219	3,0931	0,7345	0,0118	1,6029	0,462
wielkopolskie	tak	0,3264	0,0152	4,6466	0,3317	0,0240	7,2349	0,3290	0,0128	3,8995	0,465
	nie	0,6736	0,0106	1,5774	0,6683	0,0163	2,4422	0,6710	0,0089	1,3284	0,454
zachodniopomorskie	tak	0,3323	0,0208	6,2522	0,3481	0,0327	9,3984	0,3402	0,0176	5,1688	0,463
	nie	0,6677	0,0148	2,2144	0,6519	0,0232	3,5658	0,6598	0,0125	1,8955	0,462

Uwaga, kolory:

szary – szacunki na podstawie wartości dołączanych,

zielony – zysk na jakości szacunków w zbiorze zintegrowanym względem danego zbioru wejściowego.

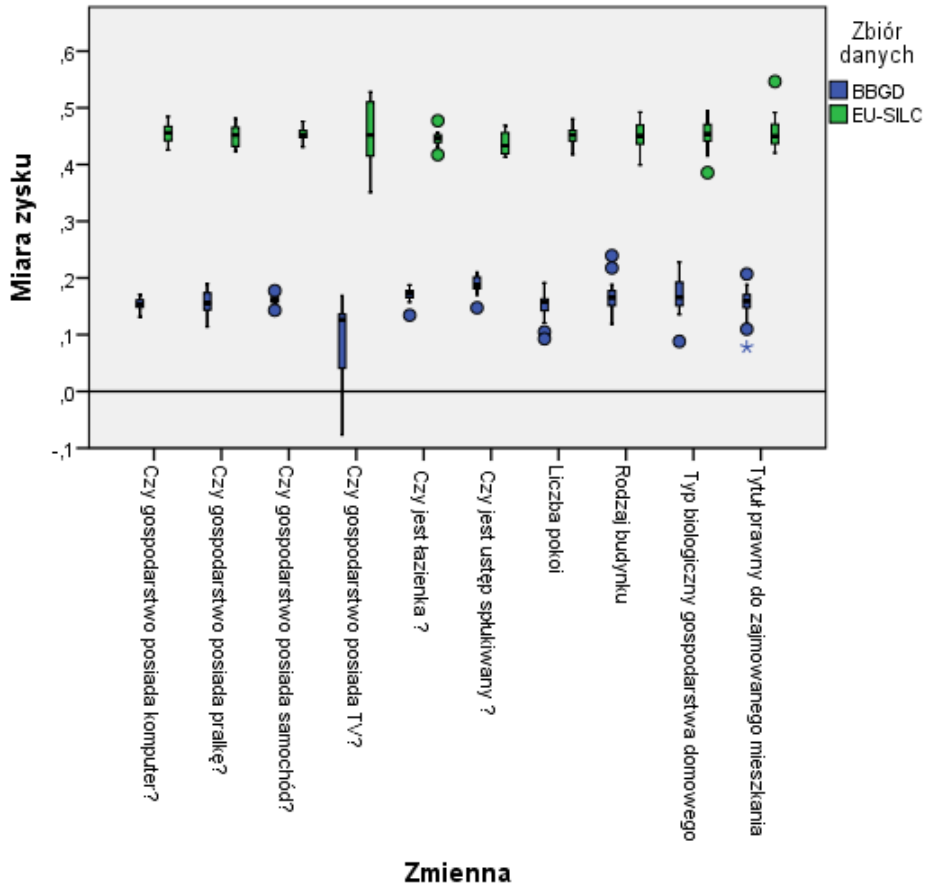
Miara zysku definiowana jest jako $1 - \frac{s_{int}(p)}{s_{EU-SILC}(p)}$.

Źródło: opracowanie własne

Jednocześnie **pozytywnie zweryfikowano hipotezy**, że w zbiorze zintegrowanym możliwe jest:

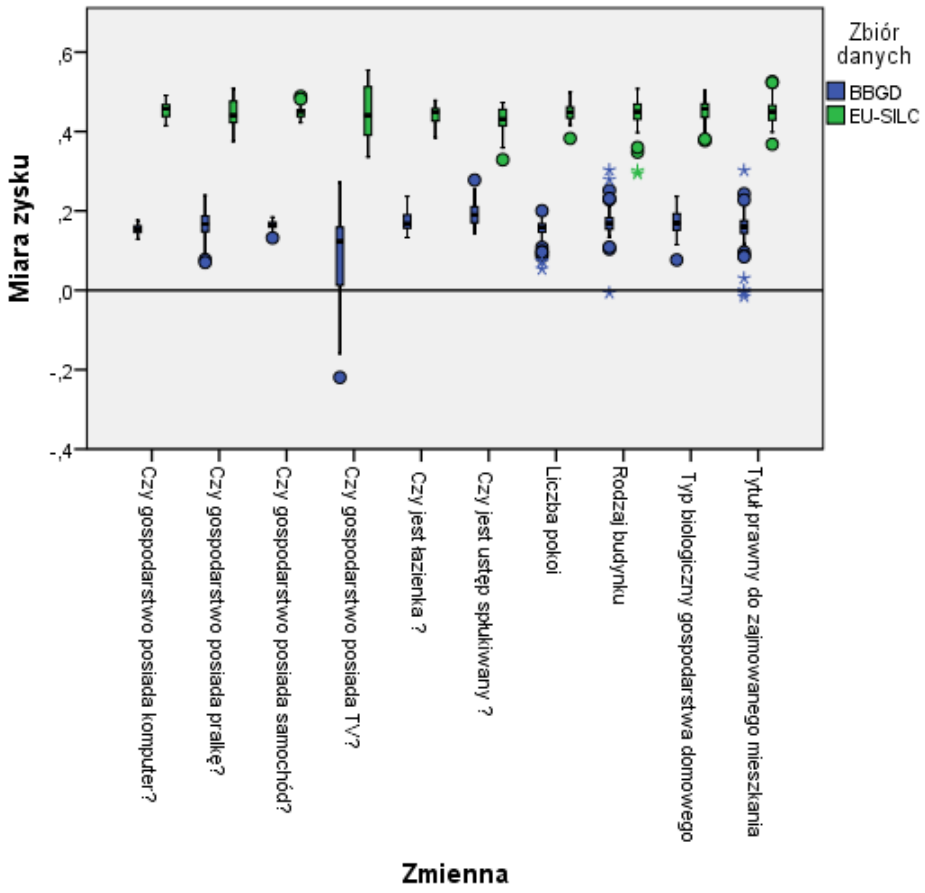
1. oszacowanie dochodów głów gospodarstw domowych w ujęciu województw (por. tabela 5.35),
2. utworzenie tabeli kontyngencji możliwości gospodarstwa na sfinansowanie tygodniowego urlopu poza miejscem zamieszkania w ujęciu województw (por. tabela 5.36).

Wykres 5.11. Rozkład miary zysku dla wspólnych cech jakościowych w ujęciu zbiorów wejściowych i regionów



Uwaga:
Warianty cech potraktowano łącznie.
Dane do wykresu pobrano z tabeli A4 znajdującej się w aneksie.
Źródło: opracowanie własne

Wykres 5.12. Rozkład miary zysku dla wspólnych cech jakościowych w ujęciu zbiorów wejściowych i województw



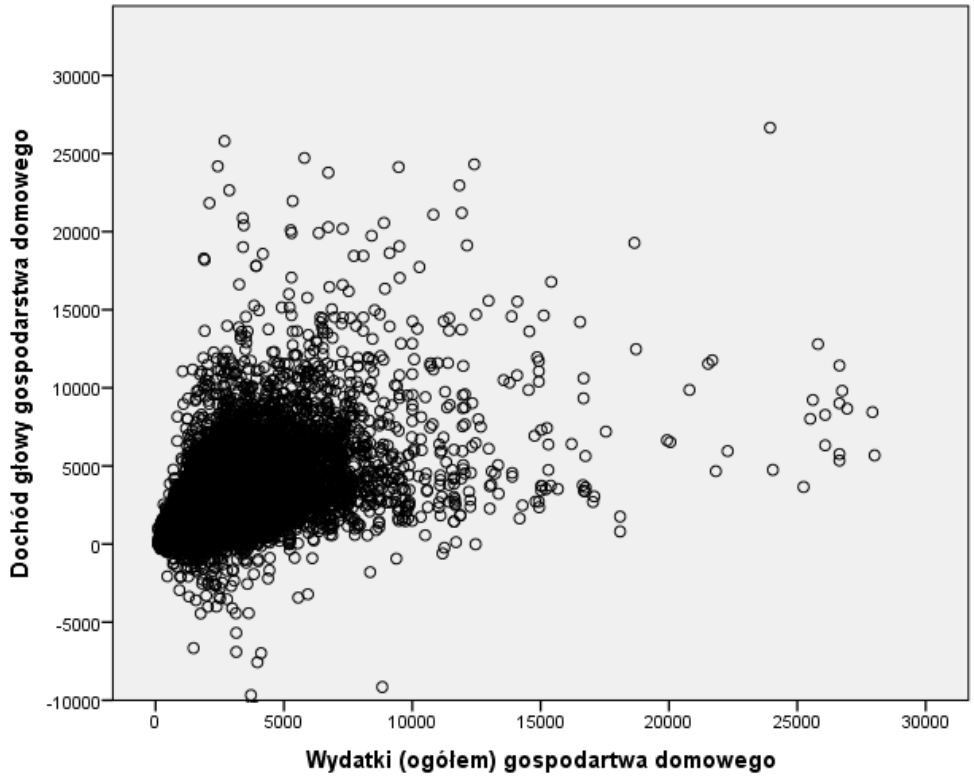
Uwaga:
Warianty cech potraktowano łącznie.
Dane do wykresu pobrano z tabeli A5 znajdującej się w aneksie.
Źródło: opracowanie własne

W odniesieniu do jakościowych cech wspólnych w ujęciu regionów w zintegrowanym zbiorze w znakomitej większości przypadków zaobserwowano zysk na jakości estymatora, znacznie wyższy w odniesieniu do zbioru EU-SILC, niż BBGD (por. wykres 5.11). Dla województw liczba przypadków spadku jakości była większa, jednak również zysk był zauważalny, zwłaszcza w odniesieniu do zbioru EU-SILC (por. wykres 5.12).

Oszacowanie współczynnika korelacji między ilościowymi zmiennymi dołączanymi

Oszacowany na podstawie zintegrowanego zbioru współczynnik korelacji liniowej Pearsona między zmiennymi „wydatki ogółem gospodarstw domowych", a „dochody głowy gospodarstwa domowego" wynosił **0,5678** (por. wykres 5.13). Błąd standardowy oszacowania współczynnika korelacji wynosił **0,003**. Analiza tej relacji nie była możliwa na podstawie żadnego ze zbiorów wejściowych.

Wykres 5.13. Korelogram ilościowych cech dołączanych w zintegrowanym zbiorze



Uwaga:
Dla lepszej ilustracji zmniejszono zakres zmiennych na wykresie.
Źródło: opracowanie własne

Należy zauważyć, że wartość tego współczynnika znajduje się w przedziale niepewności dla ρ oszacowanego dla każdej z metod imputacji stochastycznej i mieszanych (por. tabela 5.20).

Oszacowanie wydatków gospodarstw domowych według możliwości gospodarstwa na sfinansowanie tygodniowego urlopu poza miejscem zamieszkania w ujęciu województw

Zmienna „czy gospodarstwo stać na tygodniowy urlop rocznie" nie była obserwowana łącznie ani ze zmienną „województwo", ani wydatkami gospodarstw domowych. Integracja umożliwiła łączną obserwację tych cech. W zintegrowanym zbiorze możliwa stała się estymacja wydatków w ujęciu województw i możliwości sfinansowania urlopu (por. tabela 5.37).

Tabela 5.38. Przeciętne wydatki gospodarstw domowych według województw i możliwości sfinansowania urlopu wraz z oceną precyzji szacunku

Województwo	Czy gospodarstwo stać na tygodniowy urlop rocznie					
	tak			nie		
	$\bar{x}$	$S(\bar{x})$	CV	$\bar{x}$	$S(\bar{x})$	CV
dolnośląskie	2205	52,88	2,40	2072	28,05	1,35
kujawsko-pomorskie	2091	55,05	2,63	1998	32,12	1,61
lubelskie	2025	61,02	3,01	2024	35,80	1,77
lubuskie	2024	71,81	3,55	2070	46,69	2,26
łódzkie	1886	34,58	1,83	2037	29,72	1,46
małopolskie	2149	39,75	1,85	2051	30,42	1,48
mazowieckie	2025	30,25	1,49	2052	24,95	1,22
opolskie	2128	90,69	4,26	2185	62,95	2,88
podkarpackie	2046	60,34	2,95	1974	28,56	1,45
podlaskie	2022	66,78	3,30	2025	41,74	2,06
pomorskie	2010	47,99	2,39	2053	32,98	1,61
śląskie	2004	29,47	1,47	2018	22,66	1,12
świętokrzyskie	1945	73,01	3,75	1978	45,37	2,29
warmińsko-mazurskie	1936	58,54	3,02	2038	36,75	1,80
wielkopolskie	1993	37,26	1,87	2053	28,49	1,39
zachodniopomorskie	2127	76,82	3,61	2104	45,39	2,16

Źródło: opracowanie własne

Jakość szacunków wydaje się dobra, choć dla województw o stosunkowo niedużej liczebności próby (np. lubuskie, opolskie, świętokrzyskie) względne błędy szacunku były stosunkowo wysokie. Wzrost liczebności próby umożliwia dobre oszacowania w „pojedynczych" przekrojach. Dodanie kolejnych przekrojów danych powoduje wzrost błędów ponad akceptowalny poziom (ustalonym na 2%).

5.7. Wnioski

W badaniu empirycznym **osiągnięto cel główny** – wybrane jako najlepsze metody integracji w sposób dobry jakościowo odzwierciedliły rozkłady brzegowe i łączne w zintegrowanym zbiorze.

Zweryfikowano również cele szczegółowe:

1. Wykazano, że **metody losowe**, choć dobrze odzwierciedlają rozkłady brzegowe, **nie zachowują rozkładów łącznych**. Wydaje się, że można je stosować, gdy zbiory odnoszą się do tej samej populacji generalnej, ale nie zawierają żadnych zmiennych wspólnych lub dołączanych wartości jest bardzo mało. Jednak nawet

wtedy do analiz wielowymiarowych przeprowadzonych na zbiorze zintegrowanym metodą losową należy podchodzić bardzo ostrożnie.

2. Wybór zbioru dawcy i biorcy zwykle podyktowany jest celem integracji. Dla dołączanych zmiennych jakościowych wydaje się, że dołączanie rekordów ze zbioru mniejszego do większego nie zmienia w istotny sposób rozkładu cechy. Wynika to z tego, że cechy takie posiadają stosunkowo niewiele wariantów i wielokrotne dołączanie tych samych wartości nie zakłóca rozkładu. W przypadku cech ciągłych dołączanie wielokrotnie tych samych wariantów zakłócałoby ciągłość cechy (prawie zerowe prawdopodobieństwo pojawienia się tej samej wartości). W takich przypadkach właściwym wydaje się stosowanie metod imputujących wartości teoretyczne (wynikające z modelu).
3. Integracja danych metodą parowania statystycznego umożliwiła, dzięki dołączeniu zmiennej „województwo", niewystępującej w zbiorze EU-SILC, estymację na niedostępnym przed integracją poziomie agregacji przestrzennej. Otrzymane estymatory charakteryzowały się również dobrą jakością, przeciętnie dużo wyższą niż w jakimkolwiek zbiorze wejściowym.
4. Łączna obserwacja zmiennych z obu zbiorów zwiększyła zasób informacyjny w zintegrowanym zbiorze, m.in. poprzez możliwość szacowania łącznych charakterystyk cech niewystępujących łącznie w którymkolwiek ze zbiorów wejściowych. Zwiększona liczebność próby w zbiorze zintegrowanym przeciętnie zwiększyła również precyzję szacunków zarówno dla zmiennych dołączanych, jak i wspólnych.

Z przeprowadzonej analizy metod integracji można również wysnuć następujące wnioski:

— metoda najbliższego sąsiada, prosta i obarczone mniejszą liczbą założeń, zwraca lepsze rezultaty niż metody parametryczne – zwłaszcza w przypadku cech ciągłych, gdy dawcą jest zbiór liczniejszy, a biorcą zbiór mniej liczny,
— metody parametryczne, np. wielokrotna imputacja stochastyczna, dołączające wartości teoretyczne (wynikające z zastosowanego modelu) nadają się do parowania cech ciągłych w przypadku, gdy dawcą jest zbiór mniej liczny, a biorcą zbiór bardziej liczny,
— integracja zmiennych jakościowych, o stosunkowo niewielkiej liczbie wariantów, nie wymaga stosowania skomplikowanych metod parametrycznych (np.

imputacji za pomocą stochastycznej regresji logistycznej) – metoda najbliższego sąsiada w bardzo dobry sposób odzwierciedla rozkłady brzegowe i łączne,

— metody mieszane, choć za ich pomocą imputuje się wartości empiryczne, nie gwarantują rezultatów lepszych niż w przypadku metod nieparametrycznych (np. NND),

— za pomocą metod wielokrotnej imputacji można oszacować przedziały niepewności dla nieznanych szacowanych wartości, w tym współczynników korelacji między nieobserwowanymi łącznie cechami (metody nieparametryczne tego nie umożliwiają).

W przypadku stosowania metod nieparametrycznych, wybór zmiennych parujących z wektora zmiennych wspólnych nie poprawia jakości dołączanych zmiennych w zintegrowanym zbiorze. Wykorzystanie wszystkich zmiennych wspólnych zwraca rozkłady brzegowe i łączne bardziej zbliżone do empirycznych. Można to tłumaczyć przypuszczeniem, że większa liczba zastosowanych zmiennych w większym stopniu wyjaśnia zmienność zmiennej dołączanej, co ma odzwierciedlenie w wynikach. Jednocześnie jednak należy zauważyć, że algorytm metody najbliższego sąsiada jest bardziej efektywny w przypadku wyboru najbardziej skorelowanych predyktorów – mniej rekordów jest dołączanych wielokrotnie, a funkcja minimalnej odległości charakteryzuje się większą asymetrią prawostronną.

Bardzo ważnym aspektem jest harmonizacja zbiorów danych przed integracją. W obu repozytoriach występowały zmienne o takich samych lub zbliżonych definicjach. Ich warianty jednak często były rozbieżne i ich harmonizacja wymagała agregacji poszczególnych kategorii w warianty o takiej samej definicji. Agregacja w mniejszą liczbę zgodnych wariantów spowodowała utratę informacji. Oba badania są przeprowadzane przez tę samą instytucję, przyświeca im podobny cel, a pomiarowi podlegają bardzo podobne dziedziny życia społeczno-gospodarczego. Wydaje się zatem, że definicje cech, ich warianty, a także rozkłady powinny być zbieżne, nie tylko na potrzeby integracji danych, ale przede wszystkim w celach porównawczych. Wydaje się, że występujące w obu badaniach rozbieżności wynikają ze specyficznych zobowiązań międzynarodowych i konieczności porównań z innymi analogicznymi badaniami przeprowadzanymi w innych krajach europejskich. Jednak na paradoks zakrawa fakt, że wiele cech o tej samej nazwie i wydawałoby się definicji, posiada duże rozbieżności w rozkładach (np. czy gospodarstwo posiada telefon oraz niektóre zmienne docho-

dowe, nieopisane w niniejszej pracy), co *de facto* wyklucza je jako cechy parujące w procesie integracji, a nawet uniemożliwia proste analizy porównawcze.

Jako, że integracja danych wydaje się stawać nie tylko nowinką metodologiczną, ale przede wszystkim koniecznością (ze względu na koszty i czas przeprowadzania badań), projektowanie badań o już zharmonizowanych populacjach, definicjach i wariantach cech wspólnych zdaje się być krokiem w stronę efektywnego i optymalnego wykorzystania dostępnej informacji w statystyce publicznej. Wykorzystywanie technik parowania statystycznego w sektorze prywatnym (m.in. GfK), gdzie celowo projektuje się badania o mniejszym zakresie informacyjnym (w celu minimalizacji obciążeń respondentów) celem ich późniejszej integracji, wskazuje, że metody statystycznej integracji danych można i trzeba wykorzystywać w praktyce.

Rezultaty przeprowadzonego badania empirycznego wykazały, że utworzenie dobrego jakościowo zintegrowanego zbioru danych wymaga rozwiązania wielu problemów. Wśród nich można wymienić:

- utratę informacji przy harmonizacji zmiennych,
- wybór optymalnego wektora zmiennych parujących,
- wybór „dobrego" modelu integracji,
- konieczność szczegółowej analizy dla każdej zmiennej dołączanej,
- bardzo często konieczność pracy przy założeniu o warunkowej niezależności – nietestowalnego w warunkach dostępnych zbiorów,
- harmonizację i wykorzystanie wag analitycznych nie będących pierwotnymi wagami finalnymi, a wagami zmodyfikowanymi poprzez np. kalibrację,
- utrudnienia związane z usuwaniem przez gestorów ważnych, z punktu widzenia procesu integracji i późniejszych analiz, zmiennych (np. klasy miejscowości zamieszkania) ze względu na ochronę danych osobowych lub inne zobowiązania.

Doświadczenia zbierane przez statystyków w dziedzinie statystycznej integracji danych, przykładanie coraz większej wagi do łączenia zbiorów danych pochodzących z różnych źródeł przez organy statystyki publicznej oraz rozwój metodologii (m.in. poprzez projekty europejskiej, jak *ESSnet on Data Integration*) pozwalają przypuszczać, że metody parowania statystycznego będą w najbliższym czasie wykorzystywane w codziennej praktyce urzędów statystycznych. Większość zauważonych w toku wstępnych badań empirycznych i symulacyjnych problemów uda się wtedy wyeliminować lub przynajm-

niej zminimalizować. Można również uznać rozwiązanie opisywanych problemów i wątpliwości jako dalsze kierunki badań.

Możliwość oszczędności kosztów badań, poprzez np. losowanie mniejszej próby lub skrócenie kwestionariusza celem późniejszej integracji z innymi repozytoriami, czasu przeprowadzania pomiaru, optymalnego wykorzystania dostępnej informacji z pewnością nie zostaną przeoczone przez instytucje państwowe i prywatne. Wróży to dalszy rozwój metod statystycznej integracji danych, wzrost społecznej świadomości ich przydatności oraz współpracy między różnymi instytucjami w celu wypracowania tzw. „dobrych praktyk" i najlepszych metod łączenia repozytoriów.

ZAKOŃCZENIE

Przedstawiona dysertacja jest pierwszą w Polsce próbą kompleksowego opisu metod statystycznej integracji danych, ich zastosowań, a w szczególności:

- oceny efektywności algorytmów integracji,
- oceny jakości integracji dla cech jakościowych i ilościowych,
- oceny możliwości wnioskowania na podstawie zbiorów zintegrowanych.

O ile deterministyczne łączenie repozytoriów pochodzących z różnych źródeł było w Polsce praktykowane (m.in. PSR 2010, NSP 2011, projekt MEETS), o tyle podejście stochastyczne, mimo zaangażowania GUS w projekt *ESSnet on Data Integration*, a także prac podgrupy roboczej ds. metod statystyczno-matematycznych na rzecz spisu powszechnego [Gołata 2009, Roszka 2009], nie było dotąd przedmiotem empirycznego zastosowania.

Studia literaturowe wykazały szerokie spektrum zastosowań zintegrowanych zbiorów danych, głównie za pomocą metod deterministycznych. Wykorzystanie metod stochastycznych znajduje się obecnie w fazie początkowej, jednak dostrzegana jest potrzeba ich stosowania w codziennej praktyce urzędów statystycznych. Świadczy o tym choćby finansowanie przez Komisję Europejską projektów związanych z rozwojem metod statystycznej integracji danych (CENEX, *Data Integration*).

Przeprowadzane badanie empiryczne wykazało, że utworzenie zintegrowanego zbioru danych umożliwia rozszerzenie zakresu merytorycznego szacunków w porównaniu do zbiorów wejściowych. Jednocześnie zweryfikowano precyzję estymacji na podstawie zintegrowanego zbioru danych społeczno-ekonomicznych. Oceniono jakość zastosowanych metod statystycznej integracji, jak również zgodność rozkładów brzegowych i łącznych cech dołączanych w połączonym zbiorze.

W pracy wykazano też, że utworzone repozytorium charakteryzuje się użytecznością, dokładnością i spójnością. Dzięki zintegrowaniu BBGD oraz EU-SILC dowiedziono, że istnieje możliwość zwiększenia zasobów informacji społeczno-ekonomicznych w zakresie zarówno merytorycznym, jak i terytorialnym.

Zintegrowane repozytorium danych zawierało łączną informację o charakterystykach finansowych gospodarstw domowych i jakości życia. Oszacowano współczynnik korelacji między wydatkami gospodarstw domowych a dochodami ich głów, jak również sporządzono tabelę kontyngencji możliwości sfinansowania tygodniowego urlopu poza

miejscem zamieszkania w ujęciu województw. W zbiorach wejściowych nie było możliwości przeprowadzenia takich szacunków.

Dołączenie identyfikatorów województw do zbioru EU-SILC połączone z zastosowaniem konkatenacji zbiorów danych zwiększyło liczebność zintegrowanego zbioru. Repozytorium o większym pokryciu umożliwiło opracowanie szacunków o zwiększonej precyzji w ujęciu zarówno regionów, jak i województw.

Jednocześnie wyniki badania empirycznego umożliwiły sformułowanie wniosków ogólniejszej natury:

- harmonizacja definicji i wariantów zwykle sprowadza się do tworzenia zmiennych pochodnych z agregowanymi kategoriami, co może spowodować utratę informacji,
- przy założeniu o warunkowej niezależności (CIA), bez dostępu do informacji dodatkowej, możliwe jest skonstruowanie dobrych jakościowo estymatorów w zintegrowanym zbiorze,
- każda dołączana zmienna z wektora Y i Z powinna zostać przeanalizowana osobno poprzez m.in. wybór odpowiednich zmiennych parujących,
- metody nieparametryczne zwykle w dobry sposób odzwierciedlają charakterystyki rozkładu dołączanych cech, jak również rozkładów łącznych; wyjątkiem jest sytuacja, gdy dołącza się zmienną ciągłą ze zbioru mniejszego do większego – wtedy należy rozważyć zastosowanie metod parametrycznych,
- w kontekście przekazywania zbiorów danych przez gestorów występują utrudnienia związane z usuwaniem ważnych, z punktu widzenia procesu integracji i późniejszych analiz, zmiennych (np. klasy miejscowości zamieszkania ze względu na ochronę danych osobowych lub inne zobowiązania).

Wśród ważniejszych i pilnych zagadnień wymagających uwagi w kontekście rozważanego w pracy problemu statystycznej integracji danych można wymienić:

1. Potrzebę rozważenia podejścia z wykorzystaniem informacji dodatkowych, wraz z oceną ich jakości i zgodności z integrowanymi źródłami. W literaturze postuluje się, że wykorzystanie dodatkowych źródeł danych może zwiększyć jakość informacji w zbiorze zintegrowanym.
2. Pogłębienie badań nad możliwością wykorzystania technik parowania statystycznego i zintegrowanych w sposób statystyczny źródeł danych w statystyce małych obszarów, a w szczególności jako źródła informacji pomocni-

czych w estymacji pośredniej. Równocześnie należy rozważyć integrację więcej niż dwóch źródeł jednocześnie. Takie podejście umożliwi nie tylko łączną obserwację większej liczby zmiennych, zastosowanie podejścia konkatenacji plików może również umożliwić znaczne zwiększenie liczebności próby w zbiorze zintegrowanym.

3. Podjęcie próby integracji informacji z badań reprezentacyjnych z repozytoriami danych pochodzących z badań pełnych (np. spisu powszechnego) i rejestrów. W podejściu takim, po pierwsze, zniwelowany zostanie problem związany z wykorzystaniem zmodyfikowanych (za pomocą np. kalibracji) wag analitycznych, a po drugie, umożliwi otrzymanie informacji kompleksowej i pełnej dla analizowanych zjawisk społeczno-gospodarczych.
4. Zasugerowanie gestorom danych, a także organom przeprowadzającym badania reprezentacyjne ujednolicenie definicji zmiennych i wariantów cech wspólnych już na etapie projektowania badań w celu umożliwienia statystycznej integracji danych z jak najmniejszą utratą informacji na etapie harmonizacji danych przed łączeniem.
5. Przeprowadzenie kompleksowych badań empirycznych i symulacyjnych metodą probabilistycznego łączenia rekordów w celu niwelowania skutków braku klucza połączeniowego w integracji źródeł pełnych (rejestrów i spisów).

Podsumowując całość rozważań zawartych w pracy należy przypuszczać, że metody statystycznej integracji danych będą coraz częściej stosowane w badaniach statystycznych. Wynika to z dwóch zasadniczych powodów. Po pierwsze, rosnące koszty przeprowadzania badań, wzrost obciążeń respondentów i związanych z tym odmów odpowiedzi mogą wymusić na podmiotach przeprowadzających badania skrócenie kwestionariuszy i zakresu merytorycznego pojedynczych badań celem ich późniejszej integracji. Po drugie, wzrost popytu na informację szczegółową, na niskim poziomie agregacji spowoduje konieczność łączenia informacji z różnych źródeł celem zwiększenia efektywnej liczebności próby. Statystyczna integracja danych ma szansę w dużym stopniu wzbogacić warsztat metodologiczny organów statystycznych i sprostać oczekiwaniom odbiorców danych.

LITERATURA

Abramowicz W., Kaczmarek T., Flejter D. 2007,*Architectures for Deep Web Data Extraction and Integration* [w:] *Information Systems Architecture and Technology*, Oficyna Wydawnicza Politechniki Wrocławskiej, Wrocław

Aczel A. D. 2000, *Statystyka w zarządzaniu. Pełny wykład*, Wydawnictwo Naukowe PWN, Warszawa

Agresti A. 1990, *Categorical Data Analysis*, Wiley, New York

Al P., Bakker B. 2000, *Re-engineering social statistics by micro-integration of different sources; an introduction* [w:] *Integrating administrative registers and household surveys*, vol. 15, Netherlands Official Statistics, Voorburg/Heerlen

Aluja-Banet T., Daunis-i-Estadella J., Pellicer D. 2007, *GRAFT, a complete system for data fusion*, Journal of Computational Statistics and Data Analysis, 52, 635–649.

Anderson T.W. 1957 *Maximum likelihood estimates for a multivariate normal distribution when some observations are missing*, Journal of the American Statistical Association 52

Anderson T.W. 1957, *An Introduction to Multivariate Statistical Analysis*, Wiley, New York

Arts K., Bakker B.F.M., van Lith E., 2000, Lin*king administrative registers and household surveys*, [w:] *Reengineering Social Statistics by micro-integration of different sources*, Themanummer Netherlands Official Statistics, jrg. 15

Atkinson A. B., Bourguignon F., O'Donoghue C., Sutherland H., Utili F. 1999, *Microsimulation and the formulation of policy: a case study of targeting in the European Union*, EUROMOD, Working Papers Series, Working Paper No. EM2/99

Bacher J. 2002, *Statistisches Matching - Anwendungsmöglichkeiten, Verfahren und ihre praktische Umsetzung in SPSS*, ZA-Informationen, 51. Jg.

Bakker B. 2010, *Micro-Integration: State of the art* [w:] *Draft Report of WP1. State of the art on statistical methodologies for data integration*, ESSnet on Data Integration, WP1/D1.32/2010JUN

Ballano C. 2009, A *Census of Population Based on an Administrative Register*, Proceedings of Statistics Canada Symposium 2008, Data Collection: Challenges, Achievements and New Directions

Ballas D., Rossiter D., Thomas B., Clarke G.P., Dorling, D. 2005, *Geography Matters: Simulating the Local Impacts of National Social Policies*, York, Joseph Rowntree Foundation, UK

Barr R.S., Turner J.S. 1981, *Microdata file merging through large-scale network technolog*y, Mathematical Programming Study, Volume 15

Barr R.S., Turner J.S. 1990, *Quality issues and evidence in statistical file merging* [w:] *Data Quality Control: Theory and Pragmatics*, Marcel Dekker, New York

Belin, T.R., and Rubin, D.B., 1995, A *method for calibrating false-match rates in record linkage*, Journal of the American Statistical Association vol. 90.

Bernier, J., 1997, *Quantitative Evaluation of the Linkage Operations of the 1996 Census Reverse Record Check*, [w:] *Record Linkage Techniques*, Federal Committee on Statistical Methodology, Office of Management and Budget, Washington, DC

Bhattacharya I., Getoor L., 2004, *Iterative Record Linkage for Cleaning and Integration*, Department of Computer Science, University of Maryland, College Park, USA

Bijak J. 2009, *Wyrównanie wskaźników przetrwania oraz ich dekompozycja na poszczególne dni roku z uwzględnieniem sezonowości zgonów dla 66 podregionów Polski w latach 1999–2007*, podgrupa robocza do spraw metod statystycznych i matematycznych w NSP 2011, GUS

Bijak J., Kicinger A., Kupiszewski M., Śleszyński P. 2007, *Studium metodologiczne oszacowania rzeczywistej liczby ludności Warszawy*, Środkowoeuropejskie Forum Badań Migracyjnych i Ludnościowych, CEFMR Working Paper, 2/2007

Bishop Y., Fienberg S., Holland P. 1975, *Discrete Multivariate Analysis: Theory and Practice*, MIT Press, Cambridge, Massachussets

Blakely, T., Salmond, C., 2002, *Probabilistic Record linkage and a method to calculate the positive predictive value*, International Journal of Edidemiology 2002;31, Great Britain

Blum O., Calvo R. 2001, *Geospatial Data Collection and Analysis as Crucial Process in an Integrated Census*, FCSM - Federal Committee on Statistical Methodology, 2001 FCSM Conference, USA;

Borchsenius L. 2000, *From a conventional to a register-based census of population*, INSEE-Eurostat seminar on the censuses after 2001

Borkowski, W., Mielniczuk, H., 2003, *Łączenie rekordów na potrzeby analiz epidemiologicznych*, Przegląd Epidemiologiczny 2003, 57, Warszawa.

Bruhn A. 2001, *The next Population and Housing Census in Sweden is planned for 2005 – it will be totally register-based*, Symposium on Global Review of 200 Round of Population and Housing Censuses: Mid-Decade Assessment and Future Prospects, Statistic Division, Department of Economic and Social Affairs, United Nations, New York

Bycroft Ch. 2011, A *register-based census: what is the potential for New Zealand*, Statistics New Zealand, Tatauranga Aotearoa, Wellington, New Zealand

Chambers R. L., Steel D. G. 2001, *Simple methods for ecological inference in 2x2 tables*, Journal of the Royal Statistical Society, A, Volume 164

Cibella, N., Scanu, M., Tuoto, T., 2008, *Quality assessments* [w:] *ESSnet Statistical Methodology Project on Integration of Survey and Administrative Data . Report of WP1. State of the art on statistical methodologies for integration of surveys and administrative data.*

Cichomski B. (kierownik programu), Jerzyński T., Zieliński M. 2009, *Polskie Generalne Sondaże Społeczne: struktura skumulowanych wyników badań 1992-2008*, Instytut Studiów Społecznych, Uniwersytet Warszawski, Warszawa

Cohen M.L. 1991, *Statistical matching and microsimulation models* [w:] *Improving Information for Social Policy Decisions: The Uses of Microsimulation Modeling*, Vol. II: Technical Papers. Washington, DC: National Academy

Coli A., Tartamella F., Sacco G., Faiella I., Scanu M., D'Orazio M., Di Zio M., Siciliani I., Colombini S., Masi A. 2005, *La costruzione di un archivio di microdati sulle famiglie italiane ottenuto integrando l'indagine ISTAT sui consumi delle famiglie italiane e l'indagine Banca d'Italia sui bilanci delle famiglie italiane*, Technical report, *Documenti 12/2006*, Instituto Nazionale di Statistica, Rome

Czapiński J. (red.), Panek T.(red.), Baranowska A., Batorski D., Grabowska I., Grzelak J., Kotowska I., Łagodziński W., Muras M., Strzelecki P., Szumlicz T., Tymowska K. 2005, *Diagnoza Społeczna 2005. Warunki i jakość życia Polaków*, Rada Monitoringu Społecznego, Wyższa Szkoła Finansów i Zarządzania w Warszawie, Warszawa

D'Orazio M. 2012, *Statistical Matching and Imputation of Survey Data with the Package StatMatch for the R Environment*, Italian National Institute of Statistics (Istat), Rome, Italy

D'Orazio M., Di Zio M., Scanu M. 2006, *Statistical Matching. Theory and Practice*, John Wiley & Sons Ltd., England

Data Integration Manual, 2006, praca zbiorowa Statistics New Zealand, Wellington

Data Provision for Neighbourhood Renewal, Version 1, saved 23rd December 2005, www.data4nr.net

Davey A., Shanahan M. J., Schafer J. L. 2001, *Correcting for selective nonresponse in the National Longitudinal Survey of Youth using multiple imputation*. [w:] Hum Resources. 2001;36(3)

Dehnel G., Gołata E. 2012, *Wykorzystanie rejestrów administracyjnych w statystyce przedsiębiorstw* [w:] *Analiza wielowymiarowa w badaniach społeczno-ekonomicznych*, Zeszyty naukowe 227, Wydawnictwo Uniwersytetu Ekonomicznego w Poznaniu

Dempster, A., Laird, N., Rubin, D., 1977, *Maximum Likelihood from Incomplete Data via the EM Algorithm*, Journal of the Royal Statistical Society. Series B Methodological, Vol. 39, No. 1. (1977*).*

Departament Badań Społecznych i Warunków Życia 2006, *Uwagi metodyczne do Badania Budżetów Gospodarstw Domowych przeprowadzonego w 2005 r.*, GUS

Description of the action 2006, CENEX Statistical Methodology Area "Combination of survey and administrative data"

Di Zio M. 2007, *What is statistical matching*, Course on Methods for Integration of Surveys and Administrative Data, Budapest, Hungary

Di Zio M. 2012, *Connections between ecological inference and statistical matching* [w:] *Report on WP1. State of the art on statistical methodologies for data integration*, ESSnet on Data Integration

D'Orazio M. 2011, *Statistical Matching and Imputation of Survey Data with the Package StatMatch for the R Environment*, Italian National Institute of Statistics (Istat), Rome, Italy

Dygaszewicz J. 2010, *Integracja rejestrów publicznych*, Główny Urząd Statystyczny, Warszawa

Dygaszewicz J. 2011, *Narodowy Spis Powszechny Ludności i Mieszkań 2011. Podsumowanie spisu próbnego*, Centralne Biuro Spisowe

Dygaszewicz J. 2012, *Spisy powszechne jako źródło danych do analiz geoprzestrzennych*, Archiwum Fotogrametrii, Kartografii i Teledetekcji, Vol. 23

Eubank R. 1988, *Spline Smoothing and Nonparametric Regression*, Marcel Dekker, New York

European Commission 2006, *Description of SILC user database secondary target variables: Module 2006. Social participation*, Eurostat

Everaers P.C.J., van der Laan P. 2003, *The Dutch System of Social Statistics: Micro-Integration of Different Sources*, Expert Group Meeting on Setting the Scope of Social Statistics, United Nations Statistics Division in collaboration with the Siena Group on Social Statistics New York, 6-9 May 2003, ESA/STAT/AC.88/06

Fellegi, I., 1997, *Record linkage and public Policy – a dynamic resolution*, [w:] *Record Linkage Techniques*, Federal Committee on Statistical Methodology, Office of Management and Budget, Washington, DC

Fellegi, I., Sunter, A., 1969, A *theory for record linkage*, Journal of the American Statistical Association, American Statistical Association, vol. 64, no. 328, Washington DC

Filas-Przybył S., Klimanek T., Kowalewski J. 2012, *Analiza dojazdów do pracy za pomocą modelu grawitacji*, Taksonomia 19. Klasyfikacja i analiza danych – teoria i zastosowania, Prace naukowe Uniwersytetu Ekonomicznego we Wrocławiu 242

Fortini, M., Scannapieco, M., Tosco, L., and Tuoto, T. 2006, *Towards an Open Source Toolkit for Building Record Linkage Workflows*, Proceedings SIGMOD 2006 Workshop on Information Quality in Information Systems IQIS'06, Chicago, USA, 2006

Fréchet, M. 1951, *Sur les tableaux de corrélation dont les marges sont données*, Annales de l'Université de Lyon. Section A: Sciences mathématiques et astronomie 9

Gatnar E. (red.), Walesiak M. 2009, *Statystyczna analiza danych z wykorzystaniem programu R*, Wydawnictwo Naukowe PWN, Warszawa

Gatnar E., Walesiak M. (red.) 2004, *Metody statystycznej analizy wielowymiarowej w badaniach marketingowych*, Wydawnictwo Akademii Ekonomicznej we Wrocławiu

Gill, L., 2001, *Methods for Automatic Record Matching and Linkage and their use in National Statistics*, National Statistics Methodological Series No 25, National Statistics, United Kingdom

Gilula Z, McCulloch R.E., Rossi P.E. 2006, A *direct approach to data fusion*, Journal of Marketing Research, 43

Gołata E. 2011, *Data integration and SDE in Poland – experiences and problems*, ESSnet ESSnet Data Integration Workshop, Madrid 24-25 November 2011

Gołata E. 2012, *Estymacja charakterystyk przedsiębiorstw wspomagana zasobami rejestrów administracyjnych* [w:] *Analiza wielowymiarowa w badaniach społeczno-ekonomicznych*, Zeszyty naukowe 227, Wydawnictwo Uniwersytetu Ekonomicznego w Poznaniu

Gołata E., Dehnel G. 2012, *Rejestry administracyjne w analizie przedsiębiorczości*, Taksonomia 19. Klasyfikacja i analiza danych – teoria i zastosowania, Prace naukowe Uniwersytetu Ekonomicznego we Wrocławiu nr 242

Gołata E., Dehnel G., Gruchociak H. 2011, *Analiza przestrzenna w badaniu dojazdów do pracy w Polsce*, Taksonomia 18. Klasyfikacja i analiza danych – teoria i zastosowania, Prace naukowe Uniwersytetu Ekonomicznego we Wrocławiu nr 176

Gołata, E. 2009, *Raport: Opracowanie dla wybranych metod integracji danych reguł, procedur integracji danych z różnych źródeł . . .* ,materiały wewnętrzne GUS, Poznań

Goodman L. 1953, *Ecological regression and behavior of individuals*, American Sociological Review, Volume 18

Gouweleeuw J., Hartgers M. 2004, *The method of repeated weighting in the 2001 Census* [w:] *The Dutch Virtual Census of 2001. Analysis and Methodology*, Statistics Netherlands, Voorburg/Heerlen

Górniak J., Wachnicki J. 2010, *Pierwsze kroki w analizie danych*, SPSS Polska, Kraków

Groves R.M., Fowler F.J. jr., Couper M.P., J.M. Lepkowski, Singer E., Tourangeau R. 2004, *Survey Methodology,* New York: Wiley Interscience.

Gruchociak H. 2010, *Dojazdy do pracy w województwie wielkopolskim,* Wiadomości Statystyczne nr 9 (592), wrzesień 2010

Gruszczyński M. (red.), Bazyl M., Książek M., Owczarczuk M., Szulc A., Wiśniowski A., Witkowski B. 2012, *Mikroekonometria. Modele i metody analizy danych indywidualnych*, Wolters Kluwer Polska

Grzenda W. 2012, *Wybrane zagadnienia estymacji bayesowskiej* [w:] *Zaawansowane metody analiz statystycznych*, Oficyna Wydawnicza, Szkoła Główna Handlowa

Hand D., Mannila H., Smyth P. 2005, *Eksploracja danych*, Wydawnictwa Naukowo-Techniczne, Warszawa

Hardling A., Kelly S., Percival R., Keegan M. 2009, *Population Ageing and Government Age Pension Outlays*, ESRI International Collaboration Project, NATSEM, University of Canberras

Herzog T., Scheuren F., Winkler W. 2007, *Data Quality and Record Linkage Techniques*, Springer, New York, USA

Hogan, H. and Wolter, K., 1998, *Measuring accuracy in a post-enumeration survey*, Survey Methodology, vol. 14

Hudson I.L., Moore L., Beh E.J., Steel D.G. 2010, *Ecological inference techniques: an empirical evaluation using data describing gender and voter turnout at New Zealand elections, 1893–1919*, Journal of the Royal Statistical Society, A, Volume 173

Hundepool, A., Willenborg, L., 1997, *μ-Argus and τ -ARGUS: Software for Statistical Disclosure Control*, [w:] *Record Linkage Techniques*, Federal Committee on Statistical Methodology, Office of Management and Budget, Washington, DC

Hunsinger E. 2008, *Iterative Proportional Fitting For A Two – Dimensional Table*, Alaska Department of Labor and Workforce Development

i opracowania statystyczne, GUS, Warszawa

IBM SPSS Missing Values 20 2011, IBM White Papers

Janczur-Knapek M. 2012, *Spisy powszechne PSR 2010 i NSP 2011 oraz systemy informacji geograficznej w statystyce publicznej*, referat wygłoszony na Kongresie Statystyki Polskiej, Poznań

Janson S, Łuczak T., Ruciński A., *Random graphs*, Wiley, 2001.

Jaro, M. A., 1989, A*dvances In Record Linkage Methodology as Applied to Matchning the 1985 Census of Tampa*, Florida, Journal of the American Statistical Association

Józefowski T., Rynarzewska-Pietrzak B. 2010, O*cena możliwości wykorzystania rejestru PESEL w spisie ludności* [w:] *Pomiar i informacja w gospodarce*, Zeszyt Naukowy 149 pod red. E. Gołaty. Wydawnictwo Uniwersytetu Ekonomicznego w Poznaniu, Poznań

Kadane J.B. 1978, *Some statistical problems in merging data files* [w:] *Department of Treasury, Compendium of Tax Research*, Washington, DC: US Government Printing Office

Kamen C. S. 2005, *The 2008 Israel Integrated Census of Population and Housing, Basic conception and procedure*, State of Israel, Central Bureau of Statistics

Kelly S. 2003, *Australia's Microsimulation Model – Dynamod*, National Centre for Social and Economic Modelling, University of Canberra

Kiesl H., Raessler S. 2006, *How Valid Can Data Fusion Be?*, IAB Discussion Paper 15/2006, Nürnberg, Deutschland

Kordos J. 1988, *Jakość danych statystycznych*, Państwowe Wydawnictwo Ekonomiczne

Kroese B., Renssen R.H., Trijssenaar M. 2000, *Weighting or imputation: constructing a consistent set of estimates based on data from different sources*, „Netherlands Official Statistics", vol. 15, Summer 2000, Special issue: *Integrating administrative registers and household surveys*, Statistics Netherlands, Voorburg/Heerlen

Kruszka K. (red.) 2010, *Dojazdy do pracy w Polsce. Terytorialna identyfikacja przepływów ludności związanych z zatrudnieniem*, Główny Urząd Statystyczny, Urząd Statystyczny w Poznaniu

Lenz R., Zwick M. 2009, *Methodological aspects assuring remote access to German business microdata*. Bulletin of the 60th International Statistical Institute (ISI). Durban

Leszczyńska I. 2009, *Poufne dane o Polakach trafią do superbazy*, dziennik.pl, wydanie z 2009-03-09

Linder F. 2004, *The use of administrative registers and sample surveys in the Dutch Census* of 2001 [w:] *The Dutch Virtual Census of 2001. Analysis and Methodology*, Statistics Netherlands, Voorburg/Heerlen

Little R. J. A., Rubin D. B. 2002, *Statistical Analysis with Missing Data*, Wiley

Lynch, M., Winkler, W., 1994, *Improved String Comparator, Technical Report, Statistical Research Division*, Washington, DC: U.S. Bureau of the Census.

Marciniak G. (red.) 2006, *Budżety gospodarstw domowych w 2005 roku*, Informacje i opracowania statystyczne, Główny Urząd Statystyczny, Warszawa

Marella D., Scanu M., Conti P.L. 2008, *On the matching noise of some nonparametric imputation procedures*, Statistics and Probability Letters, 78

McLaughlin, G., 1993, *Private Communication of C-String-Comparison Routine*.

Ministerstwo Finansów 2010, *Studium wykonalności Projektu e-Podatki*, http://www.epodatki.mf.gov.pl/images/files/ePodatki_Studium_Wykonalnosci_SW_PU_6_3_v_3_1.pdf

Moriarity C. 2009, *Statistical Properties of Statistical Matching. Data Fusion Algorithm*, VDM Verlag Dr. Mueller, Saarbrucken, Deutschland

Moriarity C., Scheuren F. 2001, *Statistical matching: a paradigm for assessing the uncertainty in the procedure*, Journal of Official Statistics 17

Moriarity C., Scheuren F. 2003, A *note on Rubin's statistical matching using file concatenation with adjusted weights and multiple imputation*, Journal of Business and Economic Statistics 21

Morrison R. 1998, *Overview of DYNACAN: a full-fledged Canadian actuarial stochastic model designed for the fiscal and policy analysis of social security schemes*, http://www.actuaries.org/CTTEES_SOCSEC/Documents/dynacan.pdf

Musiał M., Roszka W. 2012, *Przestrzenne zróżnicowanie dojazdów do pracy*, referat wygłoszony na konferencji InterEconoMIX 2012 „Główne problemy współczesnej ekonomii z perspektywy młodych naukowców"

Neigbourhood Statistics Programme Evaluation Report, 2006, Office for National Statistics, Hampshire, UK

Neighbourhood Statistics Home Page, National Statistics, http://neighbourhood.statistics.gov.uk

Neighbourhood Statistics Service. Annual Report to Ministries 2005/06, 2006, Department for Communities and local Government, Norwich, UK

Nordholdt E.S. 2004, *Introduction to the Dutch Virtual Census of 2001* [w:] *The Dutch Virtual Census of 2001. Analysis and Methodology*, Statistics Netherlands, Voorburg/Heerlen

Nordholdt E.S. 2005, *Brief description of the methodology plan for the 2011 Census of population and housing in the Netherlands*, Statistics Netherlands, Voorburg

Norman P. 1999, *Putting Iterative Proportional Fitting on the Researcher's Desk*, School of Geography, University of Leeds, UK

Norman P. 1999, *Putting Iterative Proportional Fitting on the Researcher's Desk*, School of Geography, University of Leeds, UK

Nowakowska G. (red.) 2008, *Raport z prac na zbiorach kompleksowego systemu informatycznego ZUS*, Departament Pracy i Warunków Życia Ludności, GUS, Warszawa

Paas G. 1985, *Statistical record linkage methodology: state of art and future progress* [w:] *Bulletin of the International Statistical Institute, Proceedings of the 45th Session*, Vol. LI, Book 2, Voorburg, Netherlands ISI

Paas G. 1986, *Statistical match: evaluation of existing procedures and improvements by using additional information* [w:] *Microanalytic Simulation Models to Support Social and Financial Policy*, Amsterdam, Elsevier Science

Paradysz J. (red.), Dehnel G., Gołata E., Klimanek T., Szymkowiak M., Witkowska A., Witkowski M. 2004, *Statystyka*, Wydawnictwo Akademii Ekonomicznej w Poznaniu

Paradysz J. 2005, *Audyt miejski* [w:] *Przestrzenno-czasowe modelowanie i prognozowanie zjawisk gospodarczych.* Pod red. A. Zeliasia. Wydawnictwo Akademii Ekonomicznej w Krakowie, Kraków

Paradysz J. 2007, *Rejestry administracyjne jako źródło zasilania w statystyce regionalnej,* [w:] *Statystyka regionalna w jednoczącej się Europie*, Internetowa Oficyna Wydawnicza Centrum Statystyki Regionalnej, Poznań

Paradysz J. 2008, *Kryteria dobroci estymacji dla małych obszarów*, konferencja naukowa z okazji jubileuszu 90-lecia GUS: Statystyka społeczna: dokonania – szanse – perspektywy, Kraków, 28-30 stycznia

Paradysz J., Szymkowiak M., Wawrowski Ł., Ambroziak A., Meller E., Szkop A. 2012, *Raport z opisem wyników z zakresu możliwości wykorzystania kalibracji na potrzeby korygowania wag w złotym rekordzie*, materiały wewnętrzne GUS, Poznań

Penneck S. 2007, *Using administrative data for statistical purposes*, Economic & Labour Market Review

Porter, E., Winkler, W., 1997, *Approximate String Comparison and its Effect on an Advanced Record Linkage System* [w:] *Record Linkage Techniques*, Federal Committee on Statistical Methodology, Office of Management and Budget, Washington, DC

Prevost R., Leggieri Ch. 1999, *Expansion of Administrative Records Uses at the Census Bureau: A Long-Range Research Plan*, U.S. Bureau of the Census, Washington D.C.

Program badań statystycznych statystyki publicznej na 2012 rok, załącznik do rozporządzenia Rady Ministrów z dnia 22.07.2011 r. w sprawie programu badań statystycznych statystyki publicznej na rok 2012 (Dz. U. Nr 173, poz. 1030)

Raessler S. 2002, *Statistical Matching. A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches*, Springer, New York, USA

Raessler S. 2004, *Data fusion: identification problems, validity, and multiple imputation*, Austrian Journal of Statistics 33(1–2)

Raessler S., Kiesl H. 2009, *How useful are uncertainty bounds? Some recent theory with an application to Rubin's causal model*. 57th Session of the International Statistical Institute, Durban (South Africa), 16-22 August 2009

Raghunathan T. E., Reiter J. P., Rubin D. B. 2003, *Multiple imputation for statistical disclosure limitation*, Journal of Official Statistics 19

Rahman A. 2008, A *Review of Small Area Estimation Problems and Methodological Developments*, Discussion paper 66, NATSEM, University of Canberra

Ralphs M., Tutton P. 2011, *Beyond 2011: International models for census taking: current processes and future developments*, Beyond 2011 Project, Office for National Statistics

Renssen R.H. 1998, *Use of statistical matching techniques in calibration estimation*, Survey Methodology 24

Report of WP1. State of the art on statistical methodologies for integration of surveys and administrative data 2008, ESSnet Statistical Methodology Project on Integration of Survey and Administrative Data, Rome

Report of WP2. Recommendations on the use of methodologies for the integration of surveys and administrative data 2008, ESSnet Statistical Methodology Project on Integration of Survey and Administrative Data, Rome

Report of WP3. Software tools for integration methodologies 2008, ESSnet Statistical Methodology Project on Integration of Survey and Administrative Data, Rome

Report on WP1. State of the art on statistical methodologies for data integration 2011, ESSnet on Data Integration, Rome

Report on WP2. Methodological developments 2011, ESSnet on Data Integration, Rome

Report on WP4 Case studies 2011, ESSnet on Data Integration, Rome

Rodgers W.L. 1984, *An evaluation of statistical matching*, Journal of Business and Economic Statistics 2

Roszka W. 2009, *Raport: Opracowanie dla wybranych metod integracji danych reguł, procedur integracji danych z różnych źródeł*, materiały wewnętrzne GUS, Poznań

Roszka W. 2011a, *Statistical matching. A Polish case study* [w:] *Report on WP4. Case Studies*, ESSNet on Data Integration, Rome

Roszka W., 2011b, *Powiększanie zasobów informacyjnych ankietowych baz danych*, Zeszyty Naukowe Kolegium Analiz Ekonomicznych SGH, 23/2011, Warszawa, 2011

Rószkiewicz M. 2002, *Metody ilościowe w badaniach marketingowych*, Wydawnictwo Naukowe PWN, Warszawa

Rószkiewicz M. 2011, *Analiza klienta*, Predictive Solutions, Kraków

Rubin D.B. 1986, S*tatistical matching using file concatenation with adjusted weights and multiple imputations*, Journal of Business and Economic Statistics 4

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

Sarndal C.E., Swensson B., Wretman J. 1992, *Model-Assisted Survey Sampling*, Springer-Verlag, New York

Scannapieco M., Tosco L., Valentino L., Cibella N., Tuoto T., Fortini M. 2010, *RELAIS User's Guide*, Istat

Scanu M. 2008, *Some preliminary common aspects for record linkage and matching* [w:] *Report of WP2. Recommendations on the use of methodologies for the integration of surveys and administrative data*, CENEX-ISAD

Silverman B. W. 1986, *Density Estimation for Statistics and Data Analysis*, Chapman & Hall, London

Simpson L., Tranmer M. 2003, *Combining sample and census data in small area estimates: Iterative Proportional Fitting with standard software*, Cathie Marsh Centre for Census and Survey Research, University of Manchester, UK

Singer, E., van Hoewyk, J. 1997, *Public Attitudes Toward Data Sharing by Federal Agencies* [w:] *Record Linkage Techniques*, Federal Committee on Statistical Methodology, Office of Management and Budget, Washington, DC

Singh A. C., Armstrong J. B., Lemaitre G. E. 1988, *Statistical matching using log linear imputation* [w:] *Proceedings of the Section on Survey Research Methods*, American Statistical Association

Singh A. C., Mantel H., Kinack M., Rowe G. 1990, *On methods of statistical matching with and without auxiliary information*, Technical Report, DDMD-90-016, Statistics Canada

Singh A. C., Mantel H., Kinack M., Rowe G. 1993, *Statistical matching: Use of auxiliary information as an alternative to the conditional independence assumption*, Survey Methodology 19

Soong R., de Montigny M. 2001, *The anatomy of data fusion*, 2001 Worldwide Readership Research Symposium, Venice

Statistics Austria 2008, *Register-based census 2010 and census test 2006*, Joint UNECE/Eurostat meeting on population and housing censuses, Geneva

Statistics Finland 2004, *Use of Registers and Administrative Data Sources for Statistical Purposes. Best Practices of Statistics Finland*, Tilastokeskus, Statistikcentralen, Statistics Finland, Helsinki

Stawecki T. 2005, *Rejestry publiczne. Funkcje instytucji*, LexisNexis Polska

Steel P., Konschnik C., 1997, *Post-Matching Administrative Record Linkage Between Sole Proprietorship Tax Returns and the Standard Statistical Establishment List*,[w:] *Record Linkage Techniques*, Federal Committee on Statistical Methodology, Office of Management and Budget, Washington, DC

Stoop I.A.L. 2005, *The hunt for the last respondent. Nonresponse in sample* surveys, Den Haag: SCP

Swiss Federal Statistical Office, 2008, *The Swiss Census 2010: Moving towards a comprehensive system of household and person statistics*, Federal Statistical Office.

Szymkowiak M. 2007, *Przyczynek do kalibracji w badaniach statystycznych z brakami odpowiedzi* [w:] *Kapitał ludzki i wiedza w gospodarce wyzwania XXI wieku*, Wydawnictwo AE w Poznaniu, Poznań

Tonder J-K 2008, *The register-based statistical system – preconditions and processes*, IAOS Conference, Shanghai

Torelli, N., Paggiaro, A. 1999, *Una procedura per l'abbinamento di record nella rilevazione trimestrale delle forze di lavoro*.

Trzciński R. 2009, *Wykorzystanie techniki propensity scores matching w badaniach ewaluacyjnych*, Polska Agencja Rozwoju Przedsiębiorczości, Warszawa

Ustawa z dnia 10 kwietnia 1974 o ewidencji ludności i dowodach osobistych (Dz. U. z 2006 r. Nr 139, poz. 993 ze zm.)

Ustawa z dnia 17 lutego 2005 r. o informatyzacji działalności podmiotów realizujących zadania publiczne (Dz.U. Nr 64, poz.565, z późn. zm.)

Ustawa z dnia 29 czerwca 1995 roku o statystyce publicznej, tekst jednolity (Dz. U. 2012.591)

Ustawa z dnia 4 marca 2010 roku o narodowym spisie powszechnym ludności i mieszkań w 2011 r. (Dz.U. z 2010 nr 47 poz. 277)

van der Laan P. 2000, *Integrating administrative registers and household surveys*, „Netherlands Official Statistics", vol. 15, Summer 2000, Special issue*: Integrating administrative registers and household surveys*, Statistics Netherlands, Voorburg/Heerlen

van der Putten P., Kok J. N., Gupta A 2002, *Data Fusion through Statistical Matching*, Center for eBusiness, MIT, USA

Wallgren A., Wallgren B. 2007, *Register-based Statistics. Administrative Data for Statistical Purposes*, John Wiley and Sons Ltd.

Wallman, K., Coffey, J. 1997, *Sharing Statistical Information for Statistical Purposes*,[w:] *Record Linkage Techniques*, Federal Committee on Statistical Methodology, Office of Management and Budget, Washington, DC

Wand M., Jones C. 1995, *Kernel Smoothing*, Chapman & Hall, London

Wildner R. 2000, *Measuring Advertising Effectiveness in a Sample Matched from a Consumer Panel and a Television Panel*, Proceedings of the Fifth International Conference on Logic and Methodology, Cologne

Wildner R., Scherübl B. 2006, *Model-assisted analysis, simulation and forecasting with consumer panel data*, Yearbook of Marketing and Consumer Research, Vol.4 (2006), GfK

Winkler W. 1990, *String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage,* Proceedings of the Section on Survey Research Methods, American Statistical Association.

Winkler W. 2005, *Overview of Record Linkage and Current Research Directions*, U.S. Bureau of the Census, Washington

Witkowski J. 2010, *Rola statystyki publicznej we współczesnym świecie*, Wiadomości Statystyczne, 2, 2010

Witkowski J., Berger J., Walczak T. 2012, *Statystyka publiczna - rozwój historyczny i aktualne wyzwania*, Główny Urząd Statystyczny, Warszawa

Witkowski M. (red.), Szymkowiak M., Witkowska A. 2009, *Statystyka matematyczna w zarządzaniu*, Wydawnictwo Uniwersytetu Ekonomicznego w Poznaniu

Witkowski M., Klimanek T. 2006, *Prognozowanie gospodarcze i symulacje w przykładach i zadaniach*, Wydawnictwo Akademii Ekonomicznej w Poznaniu

Working Group (2003), *Assessment of quality in statistics. Item 6: Quality assessment of administrative data for statistical purposes*, Eurostat, Luxembourg, Doc. Eurostat/A4/Quality/03/item6

Wykorzystanie danych administracyjnych w statystyce przedsiębiorstw, 2011, Raport końcowy z wdrażania akcji przeprowadzonej w terminie od 01.11.2009 do 28.02.2011 w ramach umowy o dotację nr 30121.2009.004-2009.807, Projekt ME-ETS, Poznań-Katowice

Zawadzki J. 1997, *Wykorzystanie predyktorów z nielosowymi parametrami w prognozowaniu mikroekonomicznym*, Zeszyty NaukowePrace Katedry Ekonometrii i Statystyki, Uniwerstytet Szczeciński, Wydawnictwo Naukowe US

Zawadzki J. 1999, *Wykorzystanie modeli przyczynowo-opisowych w prognozowaniu brakujących danych*, Rozprawy i Studia Uniwersytetu Szczecińskiego, Wydawnictwo Naukowe US

Zgierska A. (red.) 2010, *Aktywność ekonomiczna ludności polski. I kwartał 2010,* Informacje

Zhang L., Chambers R. 2004, *Small area estimates for cross-classifications*, Journal of the Royal Statistical Society, Vol. 66, No. 2

Zhang L-C. (2012), *Micro calibration for data integration*, referat wygłoszony na Kongresie Statystyki Polskiej, Poznań

SPIS TABEL I RYSUNKÓW

Wykaz tabel w pracy

Wykaz schematów w pracy

Wykaz wykresów w pracy

Wykaz rysunków w pracy

ANEKS TABELARYCZNY

Tabela A1. Harmonizacja wariantów cech jakościowych

Nazwa	Warianty w BBGD	Warianty w EU-SILC	Warianty w zbiorze zharmonizowanym
Rodzaj budynku	**d6_1**	**HH010**	**rodz_bud**
	1 budynek wielorodzinny	1 dom wolnostojący	1 budynek wielorodzinny
	2 dom jednorodzinny w zabudowie szeregowej (również bliźniak)	2 bliźniak lub dom szeregowy	2 dom jednorodzinny
	3 dom jednorodzinny wolnostojący	3 apartament lub mieszkanie w budynku z mniej niż 10 mieszkań	3 dom jednorodzinny w zabudowie szeregowej
	4 inny	4 apartament lub mieszkanie w budynku z 10 lub więcej mieszkań	4 inne
Tytuł prawny do zajmowanego mieszkania	**d6_8**	**HH020**	**tytuł_prawny**
	1 własność, obciążona pożyczką lub kredytem hipotecznym	1 właściciel	1 własność
	2 własność, nieobciążona pożyczką lub kredytem hipotecznym	2 najemca lub sublokator płacący czynsz w wysokości rynkowej	2 najem wg cen rynkowych
	3 spółdzielcze prawo do zamieszkania (własnościowe lub lokatorskie)	3 wynajem po cenie poniżej rynkowej	3 najem poniżej cen rynkowych
	4 najem lub podnajem, opłata odstępnego według cen rynkowych	4 zakwaterowanie jest darmowe	4 inne
	5 najem lub podnajem, opłata odstępnego poniżej cen rynkowych		
	6 najem lub podnajem, bez odstępnego		
	7 inny		
Liczba pokoi	**d6_14**	**HH030**	**l_pok**
	1	1	1
	2	2	2
	3	3	3
	4	4	4
	5	5	5
	6	6 6 i więcej pokoi	6 6 i więcej pokoi
	7		
	8		
	9		

Nazwa	Warianty w BBGD	Warianty w EU-SILC	Warianty w zbiorze zharmonizowanym
	10		
	11		
	12		
Czy jest łazienka	**d6_22**	**HH080**	**łazienka**
	1 tak	1 tak	1 tak
	2 nie	2 nie	2 nie
Czy jest ustęp spłukiwany?	**d6_23**	**HH090**	**ustęp**
	1 tak	1 tak	1 tak
	2 nie	2 nie	2 nie
Czy dochody pozwalają na związanie końca z końcem?	**d7_3**	**HS120**	**dochwyst**
	1 z wielką trudnością	1 z wielką trudnością	1 z wielką trudnością
	2 z trudnością	2 z trudnością	2 z trudnością
	3 z pewną trudnością	3 z pewną trudnością	3 z pewną trudnością
	4 raczej łatwo	4 raczej łatwo	4 raczej łatwo
	5 łatwo	5 łatwo	5 łatwo
	6 bardzo łatwo	6 bardzo łatwo	6 bardzo łatwo
Czy gospodarstwo posiada odbiornik TV	**d1_01**[1]	**HS080**	**tv**
	0	1 tak	1 tak
	1	2 nie, nie stać mnie	2 nie
	2	3 nie, inne powody	
	3		
	4		
	5		
	6		
Czy gospodarstwo posiada komputer	**d1_12 d1_13**[2]	**HS090**	**comp**
	0	1 tak	1 tak
	1	2 nie, nie stać mnie	2 nie
	2	3 nie, inne powody	
	3		
	4		
Czy gospodarstwo posiada telefon	**d1_15 d1_16**[3]	**HS070**	**tel**
	0	1 tak	1 tak
	1	2 nie, nie stać mnie	2 nie
	2	3 nie, inne powody	
	3		
	4		
	5		
	6		
	7		

Nazwa	Warianty w BBGD	Warianty w EU-SILC	Warianty w zbiorze zharmonizowanym
	8		
Czy gospodarstwo posiada pralkę	**d1_17 d1_18[4]**	**HS100**	**pralka**
	0 1 2 3 4	1 tak 2 nie, nie stać mnie 3 nie, inne powody	1 tak 2 nie
Czy gospodarstwo posiada samochód?	**d1_28 d1_29[5]**	**HS110**	**samochód**
	0 1 2 3 4 5	1 tak 2 nie, nie stać mnie 3 nie, inne powody	1 tak 2 nie
Typ biologiczny gospodarstwa domowego	**typr**	**HX060**	**typr**
	1 małżeństwo bez dzieci 2 małżeństwo z 1 dzieckiem na utrzymaniu 3 małżeństwo z 2 dzieci na utrzymaniu 4 małżeństwo z 3 dzieci na utrzymaniu 5 małżeństwo z 4(i więcej) dzieci na utrzymaniu 6 matka z dziećmi na utrzymaniu 7 ojciec z dziećmi na utrzymaniu 8 małżeństwo z dziećmi na utrzymaniu i innymi osobami 9 matka z dziećmi na utrzymaniu i innymi osobami 10 ojciec z dziećmi na utrzymaniu i innymi osobami	5 gospodarstwo jednoosobowe 6 2 dorosłe osoby, bez dzieci i osób starszych na utrzymaniu 7 2 dorosłe osoby, bez dzieci, ale z co najmniej jedną osobą starszą na utrzymaniu 8 inne gospodarstwo bez dzieci na utrzymaniu 9 samotny rodzic z dziećmi na utrzymaniu 10 2 osoby dorosłem z jednym dzieckiem na utrzymaniu 11 2 osoby dorosłe z dwoma dziećmi na utrzymaniu 12 2 osoby dorosłem z trzema i więcej dziećmi na utrzymaniu 13 inne gospodarstwo z dziećmi na utrzymaniu 16 inne gospodarstwo	1 małżeństwo bez dzieci 2 małżeństwo z 1 dzieckiem na utrzymaniu 3 małżeństwo z 2 dzieci na utrzymaniu 4 małżeństwo z 3 więcej dzieci na utrzymaniu 5 samotny rodzic z co najmniej jedną osobą na utrzymaniu 6 gospodarstwa jednoosobowe 7 inne osoby z dziećmi na utrzymaniu 8 inne gospodarstwa z osobami na utrzymaniu 9 pozostałe gospodarstwa .

Nazwa	Warianty w BBGD	Warianty w EU-SILC	Warianty w zbiorze zharmonizowanym
	11 inne osoby z dziećmi na utrzymaniu 12 gospodarstwa jednoosobowe 13 pozostałe		

Uwaga:
W nagłówkach dla każdej zmiennej podano nazwę zmiennej w zbiorze danych.
[1] Podano liczbę odbiorników w gospodarstwie.
[2] d1_12 - Komputer osobisty z dostępem do Internetu; d1_13 - Komputer osobisty bez dostępu do Internetu; podano liczbę komputerów w gospodarstwie.
[3] d1_15 - Telefon komórkowy prywatny; d1_16 - Telefon komórkowy służbowy; podano liczbę telefonów w gospodarstwie.
[4] d1_17 - Pralka i wirówka elektryczna; d1_18 - Pralka automatyczna; podano liczbę urządzeń w gospodarstwie.
[5] d1_28 - Samochód osobowy prywatny; d1_29 - Samochód osobowy służbowy; podano liczbę pojazdów w gospodarstwie.
Źródło: opracowanie własne

Tabela A.2. Odchylenia standardowe estymatora średniej arytmetycznej wspólnych zmiennych ilościowych w zbiorach wejściowych i zintegrowanym, przekroj regionów (NUTS 1)

Zmienna	Region	BBGD			EU-SILC			Zintegrowany			$1-\frac{s_{int}(\bar{x})}{s_{BBGD}(\bar{x})}$	$1-\frac{s_{int}(\bar{x})}{s_{EU-SILC}(\bar{x})}$
		$\bar{x}$	$S(\bar{x})$	CV	$\bar{x}$	$S(\bar{x})$	CV	$\bar{x}$	$S(\bar{x})$	CV		
Liczba osób w gospodarstwie domowym	region centralny	2,7	0,02	0,66	2,6	0,03	1,04	2,7	0,01	0,56	0,1654	0,4546
	region południowy	2,8	0,02	0,65	2,8	0,03	1,02	2,8	0,02	0,55	0,1475	0,4620
	region wschodni	3,0	0,02	0,74	3,1	0,03	1,08	3,0	0,02	0,61	0,1667	0,4400
	region północno-zachodni	2,9	0,02	0,76	3,0	0,03	1,18	2,9	0,02	0,64	0,1578	0,4587
	region południowo-zachodni	2,7	0,02	0,93	2,7	0,04	1,40	2,7	0,02	0,77	0,1634	0,4567
	region północny	2,9	0,02	0,78	2,9	0,03	1,19	2,9	0,02	0,65	0,1690	0,4490
Ekwiwalentna wielkość gospodarstwa domowego	region centralny	1,8	0,01	0,44	1,7	0,01	0,70	1,8	0,01	0,38	0,1604	0,4579
	region południowy	1,8	0,01	0,44	1,8	0,01	0,70	1,8	0,01	0,38	0,1418	0,4653
	region wschodni	1,9	0,01	0,51	1,9	0,01	0,76	1,9	0,01	0,43	0,1601	0,4441
	region północno-zachodni	1,9	0,01	0,53	1,9	0,02	0,82	1,9	0,01	0,44	0,1522	0,4622
	region południowo-zachodni	1,8	0,01	0,62	1,8	0,02	0,96	1,8	0,01	0,52	0,1530	0,4635
	region północny	1,9	0,01	0,53	1,9	0,02	0,82	1,9	0,01	0,45	0,1624	0,4533
Dochód rozporządzalny gospodarstwa	region centralny	2377,7	24,32	1,02	2487,5	41,49	1,67	2433,0	21,48	0,88	0,1164	0,4821
	region południowy	2138,2	16,96	0,79	2357,4	30,84	1,31	2247,4	15,63	0,70	0,0783	0,4931
	region wschodni	1973,2	19,17	0,97	2076,9	28,99	1,40	2024,6	16,12	0,80	0,1594	0,4442
	region północno-zachodni	2155,6	29,07	1,35	2221,2	33,06	1,49	2188,6	21,41	0,98	0,2634	0,3523
	region południowo-zachodni	2096,7	29,50	1,41	2367,9	47,41	2,00	2229,9	25,27	1,13	0,1433	0,4670
	region północny	2108,6	29,93	1,42	2128,2	33,36	1,57	2118,5	21,89	1,03	0,2687	0,3439
Ekwiwalentny dochód gospodarstwa	region centralny	1380,4	13,26	0,96	1452,2	21,35	1,47	1416,6	11,36	0,80	0,1435	0,4679
	region południowy	1225,8	9,38	0,76	1333,8	14,90	1,12	1279,6	8,02	0,63	0,1445	0,4616

Zmienna	Region	BBGD			EU-SILC			Zintegrowany			$1-\frac{s_{int}(\bar{x})}{s_{BBGD}(\bar{x})}$	$1-\frac{s_{int}(\bar{x})}{s_{EU-SILC}(\bar{x})}$
		$\bar{x}$	$S(\bar{x})$	CV	$\bar{x}$	$S(\bar{x})$	CV	$\bar{x}$	$S(\bar{x})$	CV		
domowego	region wschodni	1067,4	8,94	0,84	1106,2	14,10	1,28	1086,6	7,68	0,71	0,1407	0,4554
	region północno-zachodni	1198,6	14,83	1,24	1235,9	16,62	1,34	1217,3	10,87	0,89	0,2672	0,3459
	region południowo-zachodni	1234,8	19,33	1,57	1344,9	23,94	1,78	1288,9	14,74	1,14	0,2374	0,3844
	region północny	1181,0	14,73	1,25	1199,3	16,71	1,39	1190,2	10,84	0,91	0,2642	0,3512

Źródło: opracowanie własne

Tabela A.3. Odchylenia standardowe estymatora średniej arytmetycznej wspólnych zmiennych ilościowych w zbiorach wejściowych i zintegrowanym w ujęciu województw (NUTS 2)

Zmienna	Województwo	BBGD			EU-SILC			Zintegrowany			$1-\frac{s_{int}(\bar{x})}{s_{BBGD}(\bar{x})}$	$1-\frac{s_{int}(\bar{x})}{s_{EU-SILC}(\bar{x})}$
		$\bar{x}$	$S(\bar{x})$	CV	$\bar{x}$	$S(\bar{x})$	CV	$\bar{x}$	$S(\bar{x})$	CV		
Liczba osób w gospodarstwie domowym	dolnośląskie	2,7	0,03	1,08	2,7	0,05	1,66	2,7	0,02	0,90	0,1588	0,4634
	kujawsko-pomorskie	2,9	0,04	1,27	3,0	0,05	1,85	2,9	0,03	1,04	0,1800	0,4400
	lubelskie	3,0	0,04	1,22	3,1	0,06	1,79	3,1	0,03	1,02	0,1525	0,4453
	lubuskie	2,8	0,05	1,91	2,8	0,08	2,78	2,8	0,04	1,56	0,1758	0,4460
	łódzkie	2,6	0,03	1,11	2,6	0,05	1,77	2,6	0,02	0,94	0,1552	0,4615
	małopolskie	3,0	0,03	1,09	3,0	0,05	1,72	3,0	0,03	0,93	0,1481	0,4623
	mazowieckie	2,7	0,02	0,83	2,7	0,03	1,29	2,7	0,02	0,70	0,1702	0,4512
	opolskie	2,7	0,05	1,78	2,8	0,07	2,57	2,7	0,04	1,45	0,1759	0,4390
	podkarpackie	3,2	0,04	1,39	3,2	0,06	1,98	3,2	0,04	1,13	0,1787	0,4300
	podlaskie	2,8	0,05	1,81	2,8	0,07	2,71	2,8	0,04	1,51	0,1703	0,4430
	pomorskie	2,9	0,04	1,28	3,0	0,06	2,03	2,9	0,03	1,08	0,1429	0,4729
	śląskie	2,7	0,02	0,79	2,7	0,03	1,25	2,7	0,02	0,67	0,1478	0,4615
	świętokrzyskie	3,0	0,05	1,67	3,0	0,07	2,48	3,0	0,04	1,39	0,1678	0,4422
	warmińsko-mazurskie	2,9	0,05	1,59	2,7	0,06	2,38	2,8	0,04	1,33	0,1927	0,4235
	wielkopolskie	3,1	0,03	1,00	3,1	0,05	1,57	3,1	0,03	0,85	0,1564	0,4591
	zachodniopomorskie	2,7	0,04	1,46	2,8	0,06	2,29	2,7	0,03	1,23	0,1469	0,4665
Ekwiwalentna wielkość gospodarstwa domowego	dolnośląskie	1,8	0,01	0,72	1,8	0,02	1,14	1,8	0,01	0,61	0,1490	0,4699
	kujawsko-pomorskie	1,9	0,02	0,87	1,9	0,02	1,28	1,9	0,01	0,72	0,1770	0,4420
	lubelskie	1,9	0,02	0,84	2,0	0,02	1,27	1,9	0,01	0,71	0,1460	0,4490
	lubuskie	1,8	0,02	1,29	1,8	0,03	1,90	1,8	0,02	1,06	0,1706	0,4496
	łódzkie	1,7	0,01	0,73	1,7	0,02	1,18	1,7	0,01	0,63	0,1495	0,4650
	małopolskie	1,9	0,01	0,76	1,9	0,02	1,21	1,9	0,01	0,65	0,1378	0,4681

Zmienna	Województwo	BBGD			EU-SILC			Zintegrowany			$1-\frac{s_{int}(\bar{x})}{s_{BBGD}(\bar{x})}$	$1-\frac{s_{int}(\bar{x})}{s_{EU-SILC}(\bar{x})}$
		$\bar{x}$	$S(\bar{x})$	CV	$\bar{x}$	$S(\bar{x})$	CV	$\bar{x}$	$S(\bar{x})$	CV		
	mazowieckie	1,8	0,01	0,56	1,8	0,02	0,87	1,8	0,01	0,47	0,1654	0,4544
	opolskie	1,8	0,02	1,19	1,8	0,03	1,76	1,8	0,02	0,98	0,1642	0,4462
	podkarpackie	2,0	0,02	0,97	2,0	0,03	1,40	2,0	0,02	0,80	0,1738	0,4332
	podlaskie	1,8	0,02	1,23	1,8	0,03	1,88	1,8	0,02	1,04	0,1608	0,4494
	pomorskie	1,8	0,02	0,86	1,9	0,03	1,41	1,9	0,01	0,74	0,1304	0,4800
	śląskie	1,8	0,01	0,53	1,8	0,01	0,84	1,8	0,01	0,46	0,1463	0,4625
	świętokrzyskie	1,9	0,02	1,15	1,9	0,03	1,73	1,9	0,02	0,96	0,1614	0,4465
	warmińsko-mazurskie	1,9	0,02	1,08	1,8	0,03	1,60	1,8	0,02	0,90	0,1883	0,4270
	wielkopolskie	1,9	0,01	0,70	2,0	0,02	1,11	2,0	0,01	0,60	0,1507	0,4625
	zachodniopomorskie	1,8	0,02	0,97	1,8	0,03	1,55	1,8	0,01	0,83	0,1411	0,4698
Dochód rozporzą-dzalny gospodarstwa	dolnośląskie	2074,1	35,07	1,69	2338,2	56,97	2,44	2201,6	30,08	1,37	0,1423	0,4719
	kujawsko-pomorskie	1950,8	34,67	1,78	1954,8	45,71	2,34	1952,8	27,09	1,39	0,2186	0,4074
	lubelskie	2010,1	36,47	1,81	2133,7	54,08	2,53	2071,6	30,43	1,47	0,1656	0,4373
	lubuskie	1906,7	121,40	6,37	2015,1	64,95	3,22	1961,0	76,25	3,89	0,3719	-0,1740
	łódzkie	2012,5	29,37	1,46	2031,2	53,04	2,61	2022,0	26,78	1,32	0,0880	0,4950
	małopolskie	2161,8	29,45	1,36	2328,9	52,64	2,26	2245,6	26,83	1,19	0,0888	0,4902
	mazowieckie	2575,4	33,64	1,31	2735,5	56,43	2,06	2656,1	29,46	1,11	0,1243	0,4780
	opolskie	2165,2	53,58	2,47	2446,8	84,78	3,47	2310,3	46,06	1,99	0,1404	0,4568
	podkarpackie	1981,2	30,82	1,56	2147,1	50,42	2,35	2063,9	27,12	1,31	0,1201	0,4621
	podlaskie	2005,7	49,68	2,48	2015,6	67,21	3,33	2010,5	39,43	1,96	0,2063	0,4134
	pomorskie	2311,0	51,04	2,21	2375,8	66,56	2,80	2342,8	39,61	1,69	0,2238	0,4048
	śląskie	2124,0	20,56	0,97	2374,9	37,84	1,59	2248,4	19,11	0,85	0,0706	0,4950
	świętokrzyskie	1863,1	38,25	2,05	1927,4	62,14	3,22	1894,8	33,27	1,76	0,1300	0,4645
	warmińsko-mazurskie	2033,0	73,95	3,64	2036,6	57,44	2,82	2034,9	48,24	2,37	0,3477	0,1602
	wielkopolskie	2295,3	32,51	1,42	2303,8	47,68	2,07	2299,6	26,60	1,16	0,1817	0,4420

Zmienna	Województwo	BBGD			EU-SILC			Zintegrowany			$1-\frac{s_{int}(\bar{x})}{s_{BBGD}(\bar{x})}$	$1-\frac{s_{int}(\bar{x})}{s_{EU-SILC}(\bar{x})}$
		$\bar{x}$	$S(\bar{x})$	CV	$\bar{x}$	$S(\bar{x})$	CV	$\bar{x}$	$S(\bar{x})$	CV		
	zachodniopomorskie	2059,0	33,65	1,63	2196,2	60,32	2,75	2127,3	30,57	1,44	0,0917	0,4933
Ekwiwalentny dochód gospodarstwa domowego	dolnośląskie	1227,6	23,73	1,93	1340,3	29,63	2,21	1282,0	18,16	1,42	0,2347	0,3872
	kujawsko-pomorskie	1082,6	17,01	1,57	1092,7	22,38	2,05	1087,7	13,28	1,22	0,2192	0,4067
	lubelskie	1085,6	16,97	1,56	1125,6	28,01	2,49	1105,5	14,99	1,36	0,1163	0,4647
	lubuskie	1106,2	63,70	5,76	1185,8	37,46	3,16	1146,1	40,46	3,53	0,3648	-0,0801
	łódzkie	1190,0	15,45	1,30	1200,5	25,11	2,09	1195,3	13,29	1,11	0,1397	0,4706
	małopolskie	1199,0	16,11	1,34	1275,4	25,21	1,98	1237,3	13,65	1,10	0,1526	0,4586
	mazowieckie	1483,6	18,51	1,25	1589,1	29,55	1,86	1536,7	15,79	1,03	0,1466	0,4656
	opolskie	1256,7	30,54	2,43	1357,2	38,71	2,85	1308,5	23,52	1,80	0,2301	0,3925
	podkarpackie	1044,2	14,91	1,43	1099,4	21,94	2,00	1071,7	12,39	1,16	0,1693	0,4353
	podlaskie	1129,7	22,44	1,99	1158,7	34,66	2,99	1143,9	18,96	1,66	0,1551	0,4530
	pomorskie	1306,7	25,73	1,97	1316,8	33,95	2,58	1311,7	20,06	1,53	0,2202	0,4091
	śląskie	1241,9	11,45	0,92	1369,7	18,35	1,34	1305,3	9,86	0,76	0,1395	0,4630
	świętokrzyskie	1007,2	17,69	1,76	1030,1	27,93	2,71	1018,5	15,15	1,49	0,1432	0,4574
	warmińsko-mazurskie	1134,7	35,35	3,12	1188,7	28,24	2,38	1163,0	23,23	2,00	0,3430	0,1776
	wielkopolskie	1225,8	15,84	1,29	1230,7	23,46	1,91	1228,2	13,02	1,06	0,1778	0,4449
	zachodniopomorskie	1206,0	17,57	1,46	1275,4	29,87	2,34	1240,5	15,50	1,25	0,1181	0,4811

Źródło: opracowanie własne

Tabela A.4. Odchylenia standardowe estymatora frakcji wspólnych zmiennych jakościowych w zbiorach wejściowych i zintegrowanym, przekrój regionów (NUTS 1)

Region	Zmienna	Wariant	BBGD			EU-SILC			Zintegrowany			$1-\frac{s_{int}(p)}{s_{BBGD}(p)}$	$1-\frac{s_{int}(p)}{s_{EU-SILC}(p)}$
			p	$S(p)$	CV	p	$S(p)$	CV	p	$S(p)$	CV		
region centralny	Rodzaj budynku	budynek wielorodzinny	0,5881	0,0078	1,33	0,5659	0,0132	2,33	0,5769	0,0067	1,17	0,1382	0,4889
		dom jednorodzinny	0,3588	0,0089	2,48	0,3914	0,0128	3,27	0,3753	0,0073	1,95	0,1761	0,4268
		dom jednorodzinny w zabudowie szeregowej	0,0494	0,0112	22,75	0,0393	0,0174	44,43	0,0443	0,0092	20,86	0,1777	0,4704
		inne	0,0037	0,0119	321,02	0,0034	0,0174	519,95	0,0035	0,0098	276,08	0,1828	0,4402
	Tytuł prawny do zajmowanego mieszkania	własność	0,5233	0,0079	1,50	0,5153	0,0120	2,32	0,5192	0,0066	1,26	0,1641	0,4509
		najem wg cen rynkowych	0,0265	0,0124	46,78	0,0304	0,0198	65,18	0,0285	0,0107	37,47	0,1387	0,4621
		najem poniżej cen rynkowych	0,0115	0,0120	104,24	0,0153	0,0202	131,85	0,0134	0,0107	79,58	0,1097	0,4706
		inne	0,4387	0,0091	2,08	0,4390	0,0147	3,34	0,4389	0,0077	1,76	0,1503	0,4727
	Typ biologiczny gospodarstwa domowego	małżeństwo bez dzieci	0,1718	0,0100	5,84	0,2237	0,0158	7,05	0,1980	0,0087	4,38	0,1357	0,4502
		małżeństwo z 1 dzieckiem na utrzymaniu	0,1152	0,0105	9,14	0,1112	0,0177	15,88	0,1132	0,0090	7,96	0,1443	0,4896
		małżeństwo z 2 dzieci na utrzymaniu	0,1138	0,0104	9,11	0,1037	0,0167	16,14	0,1087	0,0087	8,03	0,1575	0,4782
		małżeństwo z 3 więcej dzieci na utrzymaniu	0,0481	0,0105	21,82	0,0349	0,0159	45,58	0,0414	0,0085	20,53	0,1889	0,4657
		samotny rodzic z co najmniej jedną osobą na utrzymaniu	0,0518	0,0109	21,04	0,0308	0,0174	56,38	0,0412	0,0088	21,30	0,1942	0,4947
		gospodarstwa jednoosobowe	0,2692	0,0126	4,68	0,2825	0,0181	6,40	0,2759	0,0104	3,75	0,1772	0,4268
		inne gospodarstwa z osobami na utrzymaniu	0,0819	0,0104	12,72	0,1067	0,0161	15,08	0,0944	0,0090	9,51	0,1383	0,4422
		pozostałe gospodarstwa	0,1483	0,0102	6,86	0,1065	0,0148	13,92	0,1272	0,0082	6,44	0,1946	0,4471
	Liczba pokoi	1	0,2018	0,0112	5,53	0,1739	0,0180	10,37	0,1878	0,0094	4,99	0,1599	0,4802

Region	Zmienna	Wariant	BBGD			EU-SILC			Zintegrowany			$1-\frac{s_{int}(p)}{s_{BBGD}(p)}$	$1-\frac{s_{int}(p)}{s_{EU-SILC}(p)}$
			p	$S(p)$	CV	p	$S(p)$	CV	p	$S(p)$	CV		
		2	0,3624	0,0095	2,62	0,3519	0,0152	4,31	0,3571	0,0080	2,25	0,1537	0,4699
		3	0,2717	0,0098	3,62	0,2832	0,0154	5,42	0,2775	0,0083	2,99	0,1552	0,4597
		4	0,0882	0,0106	12,00	0,0989	0,0161	16,30	0,0935	0,0089	9,56	0,1551	0,4453
		5	0,0459	0,0107	23,42	0,0520	0,0159	30,50	0,0490	0,0090	18,38	0,1628	0,4325
		6 i więcej	0,0300	0,0108	35,87	0,0401	0,0162	40,34	0,0351	0,0092	26,28	0,1437	0,4300
	Czy jest ustęp spłukiwany ?	tak	0,8845	0,0040	0,45	0,8789	0,0064	0,73	0,8816	0,0034	0,39	0,1474	0,4686
		nie	0,1155	0,0111	9,59	0,1211	0,0163	13,43	0,1184	0,0092	7,77	0,1701	0,4351
	Czy jest łazienka ?	tak	0,8757	0,0041	0,47	0,8614	0,0068	0,79	0,8684	0,0036	0,41	0,1342	0,4773
		nie	0,1243	0,0112	8,97	0,1386	0,0163	11,75	0,1316	0,0093	7,06	0,1676	0,4300
	Czy gospodarstwo posiada TV?	tak	0,9813	0,0016	0,16	0,9686	0,0032	0,33	0,9749	0,0016	0,16	0,0265	0,5171
		nie	0,0187	0,0131	70,12	0,0314	0,0191	61,00	0,0251	0,0113	45,25	0,1330	0,4072
	Czy gospodarstwo posiada komputer?	tak	0,3957	0,0089	2,24	0,4373	0,0138	3,17	0,4167	0,0075	1,80	0,1545	0,4586
		nie	0,6043	0,0075	1,25	0,5627	0,0120	2,13	0,5833	0,0064	1,10	0,1502	0,4652
	Czy gospodarstwo posiada pralkę?	tak	0,9583	0,0024	0,25	0,9512	0,0040	0,42	0,9547	0,0021	0,22	0,1242	0,4793
		nie	0,0417	0,0129	30,94	0,0488	0,0198	40,63	0,0453	0,0110	24,26	0,1482	0,4460
	Czy gospodarstwo posiada samochód?	tak	0,4773	0,0081	1,69	0,5027	0,0126	2,50	0,4901	0,0068	1,39	0,1585	0,4588
		nie	0,5227	0,0085	1,63	0,4973	0,0133	2,67	0,5099	0,0072	1,41	0,1579	0,4596
region południowy	Rodzaj budynku	budynek wielorodzinny	0,5995	0,0077	1,28	0,5523	0,0125	2,26	0,5760	0,0066	1,14	0,1448	0,4732
		dom jednorodzinny	0,3690	0,0089	2,41	0,4250	0,0130	3,06	0,3969	0,0074	1,86	0,1696	0,4321
		dom jednorodzinny w zabudowie szeregowej	0,0297	0,0113	38,19	0,0219	0,0174	79,37	0,0258	0,0092	35,78	0,1862	0,4682
		inne	0,0018	0,0128	706,64	0,0009	0,0169	1974,31	0,0013	0,0098	730,20	0,2393	0,4209
	Tytuł praw-	własność	0,4917	0,0081	1,65	0,5154	0,0121	2,35	0,5035	0,0067	1,34	0,1691	0,4433

Region	Zmienna	Wariant	BBGD			EU-SILC			Zintegrowany			$1-\frac{s_{int}(p)}{s_{BBGD}(p)}$	$1-\frac{s_{int}(p)}{s_{EU-SILC}(p)}$
			p	$S(p)$	CV	p	$S(p)$	CV	p	$S(p)$	CV		
	ny do zajmowanego mieszkania	najem wg cen rynkowych	0,0248	0,0123	49,58	0,0404	0,0188	46,45	0,0326	0,0108	33,16	0,1211	0,4247
		najem poniżej cen rynkowych	0,0181	0,0119	65,79	0,0149	0,0196	132,07	0,0165	0,0100	60,45	0,1640	0,4915
		inne	0,4653	0,0088	1,90	0,4293	0,0139	3,24	0,4474	0,0075	1,67	0,1567	0,4643
	Typ biologiczny gospodarstwa domowego	małżeństwo bez dzieci	0,1658	0,0101	6,10	0,2147	0,0151	7,05	0,1902	0,0086	4,52	0,1512	0,4323
		małżeństwo z 1 dzieckiem na utrzymaniu	0,1221	0,0105	8,58	0,1163	0,0167	14,39	0,1192	0,0088	7,42	0,1563	0,4716
		małżeństwo z 2 dzieci na utrzymaniu	0,1296	0,0104	8,00	0,1069	0,0162	15,15	0,1183	0,0086	7,26	0,1714	0,4697
		małżeństwo z 3 więcej dzieci na utrzymaniu	0,0525	0,0106	20,20	0,0355	0,0158	44,55	0,0440	0,0085	19,36	0,1965	0,4604
		samotny rodzic z co najmniej jedną osobą na utrzymaniu	0,0527	0,0109	20,69	0,0231	0,0160	69,32	0,0380	0,0085	22,33	0,2223	0,4705
		gospodarstwa jednoosobowe	0,2539	0,0128	5,02	0,2539	0,0182	7,17	0,2539	0,0104	4,11	0,1810	0,4262
		inne gospodarstwa z osobami na utrzymaniu	0,0843	0,0104	12,33	0,1052	0,0164	15,54	0,0947	0,0090	9,46	0,1371	0,4519
		pozostałe gospodarstwa	0,1390	0,0102	7,36	0,1444	0,0158	10,91	0,1417	0,0086	6,08	0,1588	0,4533
	Liczba pokoi	1	0,1535	0,0118	7,71	0,1249	0,0180	14,44	0,1393	0,0097	6,99	0,1771	0,4600
		2	0,3548	0,0096	2,70	0,3543	0,0148	4,17	0,3545	0,0080	2,27	0,1612	0,4557
		3	0,3101	0,0096	3,08	0,2929	0,0148	5,05	0,3015	0,0080	2,65	0,1630	0,4585
		4	0,0997	0,0106	10,60	0,1126	0,0162	14,38	0,1061	0,0090	8,43	0,1535	0,4472
		5	0,0450	0,0107	23,69	0,0620	0,0165	26,58	0,0535	0,0092	17,29	0,1330	0,4391
		6 i więcej	0,0368	0,0106	28,77	0,0534	0,0165	30,96	0,0451	0,0093	20,56	0,1252	0,4392
	Czy jest ustęp spłukiwany ?	tak	0,9319	0,0031	0,33	0,9471	0,0041	0,43	0,9395	0,0024	0,26	0,2097	0,4134
		nie	0,0681	0,0117	17,25	0,0529	0,0180	33,98	0,0605	0,0096	15,88	0,1812	0,4652
	Czy jest	tak	0,9390	0,0029	0,31	0,9372	0,0045	0,48	0,9381	0,0024	0,26	0,1574	0,4563

Region	Zmienna	Wariant	BBGD			EU-SILC			Zintegrowany			$1-\frac{s_{int}(p)}{s_{BBGD}(p)}$	$1-\frac{s_{int}(p)}{s_{EU-SILC}(p)}$
			p	$S(p)$	CV	p	$S(p)$	CV	p	$S(p)$	CV		
	łazienka ?	nie	0,0610	0,0119	19,60	0,0628	0,0181	28,88	0,0619	0,0100	16,17	0,1628	0,4482
	Czy gospodarstwo posiada TV?	tak	0,9786	0,0017	0,17	0,9641	0,0034	0,35	0,9714	0,0017	0,17	0,0340	0,5113
		nie	0,0214	0,0129	60,28	0,0359	0,0192	53,47	0,0286	0,0112	39,30	0,1279	0,4142
	Czy gospodarstwo posiada komputer?	tak	0,4206	0,0087	2,06	0,4667	0,0130	2,78	0,4436	0,0072	1,63	0,1659	0,4429
		nie	0,5794	0,0078	1,34	0,5333	0,0123	2,31	0,5564	0,0066	1,19	0,1517	0,4642
	Czy gospodarstwo posiada pralkę?	tak	0,9691	0,0021	0,21	0,9614	0,0035	0,37	0,9653	0,0018	0,19	0,1138	0,4808
		nie	0,0309	0,0128	41,52	0,0386	0,0188	48,73	0,0347	0,0108	31,13	0,1575	0,4250
	Czy gospodarstwo posiada samochód?	tak	0,4807	0,0081	1,68	0,5006	0,0123	2,45	0,4906	0,0068	1,38	0,1641	0,4505
		nie	0,5193	0,0085	1,64	0,4994	0,0131	2,62	0,5094	0,0071	1,40	0,1617	0,4541
region wschodni	Rodzaj budynku	budynek wielorodzinny	0,4068	0,0104	2,55	0,3724	0,0157	4,21	0,3898	0,0086	2,21	0,1682	0,4492
		dom jednorodzinny	0,5504	0,0083	1,51	0,5878	0,0116	1,97	0,5689	0,0068	1,19	0,1882	0,4177
		dom jednorodzinny w zabudowie szeregowej	0,0388	0,0127	32,81	0,0363	0,0188	51,80	0,0376	0,0105	27,91	0,1770	0,4422
		inne	0,0040	0,0126	315,35	0,0036	0,0211	591,42	0,0038	0,0107	282,38	0,1538	0,4924
	Tytuł prawny do zajmowanego mieszkania	własność	0,6444	0,0075	1,16	0,6721	0,0105	1,56	0,6581	0,0061	0,92	0,1883	0,4205
		najem wg cen rynkowych	0,0185	0,0131	70,72	0,0250	0,0195	78,02	0,0217	0,0112	51,45	0,1458	0,4272
		najem poniżej cen rynkowych	0,0082	0,0136	166,11	0,0079	0,0198	250,77	0,0080	0,0112	138,93	0,1779	0,4360
		inne	0,3289	0,0110	3,33	0,2950	0,0164	5,56	0,3121	0,0091	2,90	0,1727	0,4469
	Typ biologiczny gospodarstwa domowego	małżeństwo bez dzieci	0,1621	0,0112	6,89	0,1958	0,0161	8,23	0,1788	0,0093	5,21	0,1668	0,4224
		małżeństwo z 1 dzieckiem na utrzymaniu	0,0940	0,0118	12,50	0,0990	0,0176	17,74	0,0965	0,0098	10,17	0,1652	0,4416
		małżeństwo z 2 dzieci na utrzymaniu	0,1082	0,0115	10,67	0,1013	0,0173	17,11	0,1048	0,0096	9,12	0,1722	0,4486

Region	Zmienna	Wariant	BBGD			EU-SILC			Zintegrowany			$1-\frac{s_{int}(p)}{s_{BBGD}(p)}$	$1-\frac{s_{int}(p)}{s_{EU-SILC}(p)}$
			p	$S(p)$	CV	p	$S(p)$	CV	p	$S(p)$	CV		
		małżeństwo z 3 więcej dzieci na utrzymaniu	0,0667	0,0115	17,20	0,0479	0,0166	34,61	0,0574	0,0092	16,03	0,1980	0,4448
		samotny rodzic z co najmniej jedną osobą na utrzymaniu	0,0537	0,0120	22,32	0,0298	0,0176	59,15	0,0419	0,0095	22,63	0,2092	0,4629
		gospodarstwa jednoosobowe	0,2313	0,0145	6,26	0,2240	0,0186	8,28	0,2277	0,0114	5,01	0,2125	0,3857
		inne gospodarstwa z osobami na utrzymaniu	0,1219	0,0110	9,06	0,1162	0,0168	14,48	0,1191	0,0092	7,72	0,1676	0,4531
		pozostałe gospodarstwa	0,1620	0,0111	6,88	0,1861	0,0168	9,02	0,1739	0,0094	5,40	0,1575	0,4411
	Liczba pokoi	1	0,1711	0,0126	7,37	0,1375	0,0181	13,14	0,1545	0,0102	6,61	0,1911	0,4354
		2	0,3432	0,0105	3,07	0,3187	0,0157	4,92	0,3310	0,0087	2,63	0,1724	0,4443
		3	0,2816	0,0107	3,79	0,2820	0,0157	5,57	0,2818	0,0088	3,13	0,1731	0,4375
		4	0,1095	0,0116	10,57	0,1277	0,0171	13,37	0,1185	0,0097	8,19	0,1617	0,4319
		5	0,0603	0,0118	19,52	0,0755	0,0173	22,87	0,0678	0,0099	14,62	0,1576	0,4258
		6 i więcej	0,0343	0,0116	33,99	0,0585	0,0175	29,97	0,0463	0,0102	22,07	0,1228	0,4179
	Czy jest ustęp spłukiwany ?	tak	0,8374	0,0052	0,62	0,8463	0,0074	0,87	0,8418	0,0042	0,50	0,1839	0,4286
		nie	0,1626	0,0120	7,37	0,1537	0,0168	10,93	0,1582	0,0097	6,15	0,1888	0,4212
	Czy jest łazienka ?	tak	0,8468	0,0050	0,59	0,8478	0,0073	0,86	0,8473	0,0041	0,49	0,1747	0,4362
		nie	0,1532	0,0122	7,96	0,1522	0,0170	11,16	0,1527	0,0099	6,48	0,1878	0,4173
	Czy gospodarstwo posiada TV?	tak	0,9795	0,0018	0,19	0,9500	0,0042	0,44	0,9649	0,0020	0,20	-0,0766	0,5276
		nie	0,0205	0,0145	70,92	0,0500	0,0186	37,24	0,0351	0,0121	34,42	0,1688	0,3511
	Czy gospodarstwo posiada komputer?	tak	0,3341	0,0101	3,04	0,3979	0,0147	3,68	0,3657	0,0084	2,30	0,1708	0,4259
		nie	0,6659	0,0075	1,13	0,6021	0,0117	1,95	0,6343	0,0064	1,01	0,1509	0,4535
	Czy gospo-	tak	0,9560	0,0027	0,28	0,9592	0,0038	0,39	0,9576	0,0022	0,23	0,1897	0,4236

Region	Zmienna	Wariant	BBGD			EU-SILC			Zintegrowany			$1-\frac{s_{int}(p)}{s_{BBGD}(p)}$	$1-\frac{s_{int}(p)}{s_{EU-SILC}(p)}$
			p	$S(p)$	CV	p	$S(p)$	CV	p	$S(p)$	CV		
	darstwo posiada pralkę?	nie	0,0440	0,0141	32,08	0,0408	0,0199	48,71	0,0424	0,0114	26,98	0,1892	0,4244
	Czy gospodarstwo posiada samochód?	tak	0,4941	0,0087	1,76	0,5188	0,0127	2,46	0,5063	0,0072	1,42	0,1739	0,4362
		nie	0,5059	0,0095	1,88	0,4812	0,0137	2,85	0,4937	0,0078	1,58	0,1774	0,4310
region północno-zachodni	Rodzaj budynku	budynek wielorodzinny	0,6221	0,0087	1,39	0,6046	0,0137	2,27	0,6133	0,0073	1,20	0,1524	0,4656
		dom jednorodzinny	0,2797	0,0111	3,96	0,3124	0,0167	5,36	0,2961	0,0093	3,14	0,1601	0,4444
		dom jednorodzinny w zabudowie szeregowej	0,0961	0,0126	13,12	0,0802	0,0198	24,70	0,0881	0,0105	11,87	0,1702	0,4722
		inne	0,0021	0,0133	627,12	0,0028	0,0216	769,62	0,0025	0,0117	474,46	0,1191	0,4591
	Tytuł prawny do zajmowanego mieszkania	własność	0,5357	0,0091	1,70	0,5469	0,0138	2,52	0,5414	0,0076	1,40	0,1659	0,4490
		najem wg cen rynkowych	0,0358	0,0139	38,78	0,0521	0,0220	42,23	0,0440	0,0122	27,80	0,1191	0,4440
		najem poniżej cen rynkowych	0,0126	0,0143	113,32	0,0129	0,0210	162,33	0,0128	0,0118	92,71	0,1717	0,4357
		inne	0,4158	0,0108	2,60	0,3880	0,0174	4,48	0,4019	0,0092	2,28	0,1524	0,4734
	Typ biologiczny gospodarstwa domowego	małżeństwo bez dzieci	0,1649	0,0119	7,22	0,2029	0,0178	8,78	0,1840	0,0101	5,47	0,1542	0,4350
		małżeństwo z 1 dzieckiem na utrzymaniu	0,1183	0,0123	10,42	0,1200	0,0206	17,12	0,1192	0,0106	8,89	0,1412	0,4848
		małżeństwo z 2 dzieci na utrzymaniu	0,1152	0,0123	10,66	0,1147	0,0200	17,43	0,1149	0,0105	9,10	0,1483	0,4769
		małżeństwo z 3 więcej dzieci na utrzymaniu	0,0538	0,0124	22,94	0,0329	0,0186	56,49	0,0433	0,0099	22,76	0,2012	0,4698
		samotny rodzic z co najmniej jedną osobą na utrzymaniu	0,0604	0,0127	20,95	0,0330	0,0192	58,03	0,0467	0,0100	21,52	0,2070	0,4756
		gospodarstwa jednoosobowe	0,2295	0,0153	6,67	0,2223	0,0212	9,53	0,2259	0,0124	5,49	0,1909	0,4153
		inne gospodarstwa z osobami na utrzymaniu	0,1059	0,0120	11,36	0,1052	0,0198	18,79	0,1055	0,0103	9,73	0,1464	0,4804

Region	Zmienna	Wariant	BBGD			EU-SILC			Zintegrowany			$1-\frac{s_{int}(p)}{s_{BBGD}(p)}$	$1-\frac{s_{int}(p)}{s_{EU-SILC}(p)}$
			p	$S(p)$	CV	p	$S(p)$	CV	p	$S(p)$	CV		
		pozostałe gospodarstwa	0,1519	0,0120	7,87	0,1689	0,0186	11,00	0,1605	0,0101	6,32	0,1512	0,4538
	Liczba pokoi	1	0,1272	0,0138	10,82	0,1048	0,0218	20,82	0,1159	0,0114	9,87	0,1687	0,4757
		2	0,3642	0,0111	3,05	0,3541	0,0172	4,87	0,3592	0,0093	2,60	0,1605	0,4589
		3	0,3126	0,0112	3,57	0,3121	0,0174	5,58	0,3123	0,0094	3,01	0,1583	0,4603
		4	0,1073	0,0124	11,59	0,1160	0,0191	16,44	0,1117	0,0105	9,39	0,1565	0,4497
		5	0,0551	0,0126	22,93	0,0648	0,0199	30,72	0,0600	0,0108	18,09	0,1416	0,4549
		6 i więcej	0,0335	0,0126	37,70	0,0482	0,0200	41,44	0,0409	0,0111	27,16	0,1207	0,4440
	Czy jest ustęp spłukiwany ?	tak	0,9307	0,0036	0,39	0,9422	0,0051	0,54	0,9365	0,0029	0,31	0,1941	0,4314
		nie	0,0693	0,0133	19,13	0,0578	0,0199	34,48	0,0635	0,0109	17,10	0,1803	0,4552
	Czy jest łazienka ?	tak	0,9227	0,0038	0,41	0,9296	0,0056	0,61	0,9262	0,0031	0,34	0,1779	0,4447
		nie	0,0773	0,0134	17,34	0,0704	0,0202	28,73	0,0738	0,0111	15,01	0,1734	0,4521
	Czy gospodarstwo posiada TV?	tak	0,9870	0,0016	0,16	0,9796	0,0030	0,31	0,9833	0,0015	0,15	0,0475	0,5094
		nie	0,0130	0,0153	117,53	0,0204	0,0227	110,84	0,0167	0,0132	79,06	0,1335	0,4161
	Czy gospodarstwo posiada komputer?	tak	0,3927	0,0104	2,65	0,4441	0,0159	3,59	0,4186	0,0088	2,09	0,1587	0,4504
		nie	0,6073	0,0087	1,44	0,5559	0,0140	2,53	0,5814	0,0075	1,28	0,1463	0,4692
	Czy gospodarstwo posiada pralkę?	tak	0,9756	0,0021	0,22	0,9744	0,0034	0,35	0,9750	0,0018	0,19	0,1501	0,4635
		nie	0,0244	0,0151	61,72	0,0256	0,0235	92,03	0,0250	0,0127	51,01	0,1535	0,4584
	Czy gospodarstwo posiada samochód?	tak	0,4944	0,0094	1,90	0,5463	0,0143	2,61	0,5205	0,0078	1,51	0,1640	0,4504
		nie	0,5056	0,0100	1,98	0,4537	0,0157	3,46	0,4795	0,0085	1,76	0,1569	0,4614
region południowo-zachodni	Rodzaj budynku	budynek wielorodzinny	0,7291	0,0086	1,18	0,6680	0,0149	2,24	0,6991	0,0076	1,09	0,1199	0,4913
		dom jednorodzinny	0,2210	0,0134	6,04	0,2813	0,0204	7,24	0,2506	0,0114	4,54	0,1488	0,4414
		dom jednorodzinny w zabudowie szeregowej	0,0469	0,0155	32,96	0,0472	0,0243	51,55	0,0470	0,0131	27,75	0,1557	0,4632

Region	Zmienna	Wariant	BBGD			EU-SILC			Zintegrowany			$1-\frac{s_{int}(p)}{s_{BBGD}(p)}$	$1-\frac{s_{int}(p)}{s_{EU-SILC}(p)}$
			p	$S(p)$	CV	p	$S(p)$	CV	p	$S(p)$	CV		
		inne	0,0029	0,0163	554,88	0,0036	0,0243	682,23	0,0033	0,0138	424,71	0,1549	0,4329
	Tytuł prawny do zajmowanego mieszkania	własność	0,5076	0,0111	2,18	0,5549	0,0165	2,98	0,5308	0,0092	1,74	0,1692	0,4433
		najem wg cen rynkowych	0,0321	0,0178	55,50	0,0392	0,0262	66,76	0,0356	0,0150	42,10	0,1586	0,4278
		najem poniżej cen rynkowych	0,0123	0,0166	135,01	0,0051	0,0291	570,38	0,0088	0,0132	150,35	0,2070	0,5464
		inne	0,4480	0,0123	2,74	0,4008	0,0200	4,99	0,4248	0,0105	2,46	0,1498	0,4774
	Typ biologiczny gospodarstwa domowego	małżeństwo bez dzieci	0,1815	0,0138	7,58	0,2177	0,0213	9,80	0,1993	0,0117	5,89	0,1470	0,4502
		małżeństwo z 1 dzieckiem na utrzymaniu	0,1110	0,0145	13,05	0,1116	0,0228	20,42	0,1113	0,0122	10,99	0,1559	0,4630
		małżeństwo z 2 dzieci na utrzymaniu	0,0978	0,0144	14,73	0,0956	0,0231	24,16	0,0967	0,0122	12,61	0,1533	0,4719
		małżeństwo z 3 więcej dzieci na utrzymaniu	0,0441	0,0146	33,19	0,0253	0,0196	77,62	0,0348	0,0113	32,58	0,2238	0,4216
		samotny rodzic z co najmniej jedną osobą na utrzymaniu	0,0653	0,0148	22,69	0,0369	0,0216	58,59	0,0514	0,0117	22,84	0,2083	0,4575
		gospodarstwa jednoosobowe	0,2733	0,0173	6,34	0,2557	0,0252	9,85	0,2646	0,0142	5,38	0,1790	0,4350
		inne gospodarstwa z osobami na utrzymaniu	0,0786	0,0143	18,17	0,1198	0,0249	20,79	0,0988	0,0130	13,18	0,0880	0,4770
		pozostałe gospodarstwa	0,1484	0,0141	9,48	0,1374	0,0229	16,67	0,1430	0,0119	8,32	0,1536	0,4802
	Liczba pokoi	1	0,1622	0,0164	10,12	0,1291	0,0253	19,64	0,1459	0,0135	9,28	0,1749	0,4655
		2	0,3708	0,0131	3,52	0,3582	0,0205	5,73	0,3646	0,0110	3,02	0,1572	0,4643
		3	0,2759	0,0134	4,87	0,2561	0,0210	8,21	0,2662	0,0113	4,23	0,1610	0,4641
		4	0,1032	0,0146	14,10	0,1179	0,0234	19,84	0,1105	0,0125	11,33	0,1399	0,4650
		5	0,0510	0,0147	28,81	0,0747	0,0245	32,82	0,0626	0,0132	21,01	0,1048	0,4631
		6 i więcej	0,0369	0,0145	39,44	0,0640	0,0238	37,15	0,0502	0,0132	26,28	0,0928	0,4451
	Czy jest	tak	0,8985	0,0051	0,57	0,9246	0,0070	0,76	0,9113	0,0041	0,45	0,2087	0,4168

Region	Zmienna	Wariant	BBGD			EU-SILC			Zintegrowany			$1-\frac{s_{int}(p)}{s_{BBGD}(p)}$	$1-\frac{s_{int}(p)}{s_{EU-SILC}(p)}$
			p	$S(p)$	CV	p	$S(p)$	CV	p	$S(p)$	CV		
	ustęp spłukiwany ?	nie	0,1015	0,0157	15,43	0,0754	0,0227	30,14	0,0887	0,0126	14,24	0,1936	0,4445
	Czy jest łazienka ?	tak	0,9017	0,0051	0,56	0,9123	0,0075	0,82	0,9069	0,0042	0,46	0,1780	0,4463
		nie	0,0983	0,0159	16,17	0,0877	0,0233	26,61	0,0931	0,0130	13,99	0,1807	0,4419
	Czy gospodarstwo posiada TV?	tak	0,9739	0,0026	0,27	0,9696	0,0044	0,45	0,9718	0,0023	0,23	0,1248	0,4810
		nie	0,0261	0,0184	70,52	0,0304	0,0265	87,18	0,0282	0,0153	54,27	0,1688	0,4220
	Czy gospodarstwo posiada komputer?	tak	0,3816	0,0124	3,24	0,4666	0,0186	3,98	0,4234	0,0104	2,45	0,1602	0,4412
		nie	0,6184	0,0102	1,65	0,5334	0,0172	3,23	0,5766	0,0089	1,54	0,1315	0,4846
	Czy gospodarstwo posiada pralkę?	tak	0,9620	0,0031	0,33	0,9655	0,0047	0,49	0,9637	0,0026	0,27	0,1766	0,4472
		nie	0,0380	0,0174	45,94	0,0345	0,0266	77,17	0,0363	0,0145	39,89	0,1709	0,4565
	Czy gospodarstwo posiada samochód?	tak	0,4325	0,0116	2,68	0,5031	0,0176	3,50	0,4672	0,0097	2,08	0,1615	0,4474
		nie	0,5675	0,0111	1,96	0,4969	0,0182	3,66	0,5328	0,0095	1,79	0,1429	0,4760
region północny	Rodzaj budynku	budynek wielorodzinny	0,6672	0,0084	1,25	0,6458	0,0133	2,06	0,6564	0,0071	1,08	0,1499	0,4666
		dom jednorodzinny	0,2554	0,0115	4,50	0,2847	0,0172	6,03	0,2702	0,0096	3,56	0,1630	0,4397
		dom jednorodzinny w zabudowie szeregowej	0,0724	0,0131	18,10	0,0655	0,0197	30,15	0,0689	0,0108	15,70	0,1744	0,4519
		inne	0,0050	0,0147	295,22	0,0040	0,0191	474,39	0,0045	0,0115	255,38	0,2176	0,3993
	Tytuł prawny do zajmowanego mieszkania	własność	0,5250	0,0095	1,80	0,5618	0,0139	2,47	0,5436	0,0078	1,44	0,1744	0,4367
		najem wg cen rynkowych	0,0264	0,0145	54,71	0,0442	0,0247	55,99	0,0354	0,0133	37,67	0,0782	0,4610
		najem poniżej cen rynkowych	0,0166	0,0146	88,34	0,0167	0,0226	135,77	0,0166	0,0123	74,01	0,1600	0,4564
		inne	0,4319	0,0109	2,53	0,3773	0,0175	4,64	0,4044	0,0092	2,29	0,1549	0,4715
	Typ biologiczny go-	małżeństwo bez dzieci	0,1718	0,0122	7,11	0,2156	0,0180	8,36	0,1939	0,0103	5,31	0,1571	0,4286
		małżeństwo z 1 dzieckiem	0,1120	0,0127	11,34	0,1223	0,0205	16,74	0,1172	0,0109	9,29	0,1427	0,4681

Region	Zmienna	Wariant	BBGD			EU-SILC			Zintegrowany			$1-\frac{s_{int}(p)}{s_{BBGD}(p)}$	$1-\frac{s_{int}(p)}{s_{EU-SILC}(p)}$
			p	$S(p)$	CV	p	$S(p)$	CV	p	$S(p)$	CV		
	spodarstwa domowego	na utrzymaniu											
		małżeństwo z 2 dzieci na utrzymaniu	0,1250	0,0126	10,12	0,0991	0,0197	19,84	0,1119	0,0104	9,31	0,1757	0,4699
		małżeństwo z 3 więcej dzieci na utrzymaniu	0,0647	0,0126	19,50	0,0494	0,0194	39,25	0,0570	0,0103	18,10	0,1828	0,4677
		samotny rodzic z co najmniej jedną osobą na utrzymaniu	0,0598	0,0130	21,73	0,0285	0,0181	63,30	0,0440	0,0100	22,79	0,2278	0,4443
		gospodarstwa jednoosobowe	0,2273	0,0157	6,92	0,2299	0,0227	9,88	0,2286	0,0129	5,66	0,1774	0,4301
		inne gospodarstwa z osobami na utrzymaniu	0,0957	0,0125	13,04	0,1036	0,0202	19,52	0,0997	0,0107	10,74	0,1417	0,4708
		pozostałe gospodarstwa	0,1437	0,0124	8,62	0,1515	0,0189	12,46	0,1476	0,0104	7,05	0,1602	0,4488
	Liczba pokoi	1	0,1225	0,0142	11,62	0,1187	0,0219	18,41	0,1206	0,0119	9,86	0,1642	0,4560
		2	0,4100	0,0110	2,67	0,3890	0,0171	4,38	0,3994	0,0092	2,31	0,1603	0,4600
		3	0,3139	0,0114	3,65	0,3160	0,0174	5,52	0,3149	0,0096	3,04	0,1635	0,4514
		4	0,0869	0,0131	15,05	0,0956	0,0202	21,18	0,0913	0,0111	12,14	0,1523	0,4523
		5	0,0441	0,0130	29,38	0,0490	0,0198	40,39	0,0466	0,0110	23,52	0,1548	0,4463
		6 i więcej	0,0226	0,0129	57,25	0,0317	0,0205	64,70	0,0272	0,0114	41,78	0,1219	0,4460
	Czy jest ustęp spłukiwany ?	tak	0,9283	0,0038	0,41	0,9440	0,0051	0,54	0,9362	0,0030	0,32	0,2087	0,4159
		nie	0,0717	0,0133	18,55	0,0560	0,0199	35,59	0,0638	0,0108	16,98	0,1860	0,4567
	Czy jest łazienka ?	tak	0,9179	0,0040	0,44	0,9212	0,0061	0,66	0,9196	0,0033	0,36	0,1709	0,4475
		nie	0,0821	0,0135	16,45	0,0788	0,0205	26,07	0,0804	0,0112	13,98	0,1676	0,4527
	Czy gospodarstwo posiada TV?	tak	0,9891	0,0015	0,15	0,9843	0,0027	0,27	0,9866	0,0014	0,14	0,0754	0,4961
		nie	0,0109	0,0161	146,70	0,0157	0,0240	152,20	0,0134	0,0138	103,45	0,1389	0,4229
	Czy gospo-	tak	0,3749	0,0109	2,90	0,4542	0,0162	3,58	0,4149	0,0091	2,20	0,1618	0,4386

Region	Zmienna	Wariant	BBGD			EU-SILC			Zintegrowany			$1-\frac{s_{int}(p)}{s_{BBGD}(p)}$	$1-\frac{s_{int}(p)}{s_{EU-SILC}(p)}$
			p	$S(p)$	CV	p	$S(p)$	CV	p	$S(p)$	CV		
	darstwo posiada komputer?	nie	0,6251	0,0088	1,40	0,5458	0,0143	2,62	0,5851	0,0075	1,29	0,1386	0,4729
	Czy gospodarstwo posiada pralkę?	tak	0,9716	0,0024	0,24	0,9683	0,0038	0,40	0,9700	0,0020	0,21	0,1387	0,4677
		nie	0,0284	0,0158	55,55	0,0317	0,0236	74,53	0,0300	0,0132	44,10	0,1591	0,4391
	Czy gospodarstwo posiada samochód?	tak	0,4418	0,0101	2,29	0,4684	0,0153	3,26	0,4552	0,0084	1,86	0,1648	0,4476
		nie	0,5582	0,0097	1,74	0,5316	0,0152	2,85	0,5448	0,0082	1,50	0,1565	0,4603

Źródło: opracowanie własne

Tabela A.5. Odchylenia standardowe estymatora frakcji wspólnych zmiennych jakościowych w zbiorach wejściowych i zintegrowanym, przekrój województw (NUTS 2)

Województwo	Zmienna	Wariant	BBGD			EU-SILC			Zintegrowany			$1-\frac{s_{int}(p)}{s_{BBGD}(p)}$	$1-\frac{s_{int}(p)}{s_{EU-SILC}(p)}$
			p	$S(p)$	CV	p	$S(p)$	CV	p	$S(p)$	CV		
dolnośląskie	Rodzaj budynku	budynek wielorodzinny	0,77	0,01	1,18	0,71	0,02	2,34	0,74	0,01	1,11	0,1040	0,5033
		dom jednorodzinny	0,17	0,02	9,31	0,23	0,02	10,74	0,20	0,01	6,86	0,1354	0,4479
		dom jednorodzinny w zabudowie szeregowej	0,05	0,02	33,36	0,06	0,03	50,83	0,05	0,02	27,73	0,1459	0,4696
		inne	0,00	0,02	868,31	0,00	0,03	581,35	0,00	0,02	486,42	0,1086	0,4016
	Tytuł prawny do zajmowanego mieszkania	własność	0,49	0,01	2,66	0,53	0,02	3,83	0,51	0,01	2,16	0,1615	0,4546
		najem wg cen rynkowych	0,03	0,02	60,89	0,04	0,03	74,04	0,04	0,02	46,43	0,1585	0,4302
		najem poniżej cen rynkowych	0,01	0,02	154,97	0,01	0,03	485,91	0,01	0,02	161,18	0,1880	0,5254
		inne	0,46	0,01	3,06	0,42	0,02	5,45	0,44	0,01	2,71	0,1465	0,4811
	Typ biologiczny gospodarstwa domowego	małżeństwo bez dzieci	0,18	0,02	8,94	0,21	0,03	11,90	0,20	0,01	7,01	0,1394	0,4615
		małżeństwo z 1 dzieckiem na utrzymaniu	0,11	0,02	15,49	0,11	0,03	24,70	0,11	0,01	13,13	0,1547	0,4667
		małżeństwo z 2 dzieci na utrzymaniu	0,10	0,02	16,95	0,10	0,03	27,16	0,10	0,01	14,36	0,1483	0,4741
		małżeństwo z 3 więcej dzieci na utrzymaniu	0,04	0,02	39,84	0,02	0,02	94,21	0,03	0,01	39,12	0,2196	0,4260
		samotny rodzic z co najmniej jedną osobą na utrzymaniu	0,07	0,02	23,67	0,04	0,02	62,38	0,06	0,01	23,93	0,2085	0,4542
		gospodarstwa jednoosobowe	0,28	0,02	7,18	0,26	0,03	11,28	0,27	0,02	6,11	0,1746	0,4409
		inne gospodarstwa z osobami na utrzymaniu	0,07	0,02	22,55	0,12	0,03	24,89	0,10	0,02	16,06	0,0765	0,4815
		pozostałe gospodarstwa	0,15	0,02	11,17	0,13	0,03	20,55	0,14	0,01	9,95	0,1514	0,4882
	Liczba pokoi	1	0,19	0,02	10,03	0,15	0,03	19,95	0,17	0,02	9,26	0,1694	0,4730

Województwo	Zmienna	Wariant	BBGD			EU-SILC			Zintegrowany			$1-\frac{s_{int}(p)}{s_{BBGD}(p)}$	$1-\frac{s_{int}(p)}{s_{EU-SILC}(p)}$
			p	$S(p)$	CV	p	$S(p)$	CV	p	$S(p)$	CV		
		2	0,38	0,02	4,01	0,36	0,02	6,66	0,37	0,01	3,46	0,1541	0,4697
		3	0,28	0,02	5,61	0,26	0,02	9,61	0,27	0,01	4,89	0,1545	0,4735
		4	0,10	0,02	17,61	0,11	0,03	25,39	0,10	0,01	14,29	0,1378	0,4706
		5	0,04	0,02	42,76	0,07	0,03	42,51	0,05	0,02	29,22	0,0852	0,4593
		6 i więcej	0,03	0,02	68,37	0,06	0,03	53,16	0,04	0,02	41,09	0,0536	0,4443
	Czy jest ustęp spłukiwany ?	tak	0,89	0,01	0,71	0,92	0,01	0,94	0,90	0,00	0,55	0,2059	0,4202
		nie	0,11	0,02	16,13	0,08	0,03	32,89	0,10	0,01	15,05	0,1868	0,4563
	Czy jest łazienka ?	tak	0,89	0,01	0,71	0,90	0,01	1,04	0,90	0,01	0,58	0,1769	0,4496
		nie	0,11	0,02	16,33	0,10	0,03	28,13	0,10	0,02	14,34	0,1751	0,4526
	Czy gospodarstwo posiada TV?	tak	0,97	0,00	0,34	0,97	0,01	0,55	0,97	0,00	0,29	0,1410	0,4758
		nie	0,03	0,02	69,66	0,03	0,03	98,48	0,03	0,02	56,58	0,1631	0,4431
	Czy gospodarstwo posiada komputer?	tak	0,39	0,01	3,71	0,47	0,02	4,66	0,43	0,01	2,84	0,1579	0,4464
		nie	0,61	0,01	1,95	0,53	0,02	3,84	0,57	0,01	1,81	0,1288	0,4912
	Czy gospodarstwo posiada pralkę?	tak	0,96	0,00	0,41	0,96	0,01	0,59	0,96	0,00	0,33	0,1892	0,4387
		nie	0,04	0,02	45,30	0,04	0,03	87,03	0,04	0,02	41,10	0,1684	0,4766
	Czy gospodarstwo posiada samochód?	tak	0,41	0,01	3,34	0,50	0,02	4,22	0,45	0,01	2,56	0,1581	0,4483
		nie	0,59	0,01	2,13	0,50	0,02	4,23	0,55	0,01	1,99	0,1322	0,4886
kujawsko-pomorskie	Rodzaj budynku	budynek wielorodzinny	0,64	0,01	2,29	0,60	0,02	3,86	0,62	0,01	2,00	0,1490	0,4662
		dom jednorodzinny	0,32	0,02	5,64	0,35	0,03	7,75	0,33	0,02	4,51	0,1660	0,4403
		dom jednorodzinny w zabudowie szeregowej	0,04	0,02	54,23	0,05	0,04	76,44	0,05	0,02	43,66	0,1409	0,4621
		inne	0,00	0,02	2324,9	0,00	0,03	1214,5	0,00	0,02	1118,4	0,1690	0,3480

Województwo	Zmienna	Wariant	BBGD			EU-SILC			Zintegrowany			$1-\frac{s_{int}(p)}{s_{BBGD}(p)}$	$1-\frac{s_{int}(p)}{s_{EU-SILC}(p)}$
			p	$S(p)$	CV	p	$S(p)$	CV	p	$S(p)$	CV		
					2			3			0		
	Tytuł prawny do zajmowanego mieszkania	własność	0,49	0,02	3,27	0,52	0,02	4,57	0,50	0,01	2,63	0,1730	0,4374
		najem wg cen rynkowych	0,03	0,03	81,81	0,05	0,04	87,46	0,04	0,02	57,77	0,0963	0,4560
		najem poniżej cen rynkowych	0,02	0,02	117,40	0,02	0,04	221,99	0,02	0,02	106,69	0,1849	0,4584
		inne	0,46	0,02	3,90	0,42	0,03	6,72	0,44	0,02	3,43	0,1557	0,4667
	Typ biologiczny gospodarstwa domowego	małżeństwo bez dzieci	0,17	0,02	12,19	0,18	0,03	16,31	0,18	0,02	9,67	0,1663	0,4342
		małżeństwo z 1 dzieckiem na utrzymaniu	0,12	0,02	18,13	0,12	0,03	28,46	0,12	0,02	15,30	0,1574	0,4619
		małżeństwo z 2 dzieci na utrzymaniu	0,12	0,02	16,85	0,12	0,03	26,40	0,12	0,02	14,22	0,1597	0,4592
		małżeństwo z 3 więcej dzieci na utrzymaniu	0,06	0,02	35,75	0,06	0,03	54,45	0,06	0,02	29,87	0,1623	0,4529
		samotny rodzic z co najmniej jedną osobą na utrzymaniu	0,06	0,02	34,29	0,03	0,03	99,51	0,05	0,02	36,00	0,2315	0,4361
		gospodarstwa jednoosobowe	0,21	0,03	12,53	0,22	0,04	17,10	0,22	0,02	10,06	0,1793	0,4238
		inne gospodarstwa z osobami na utrzymaniu	0,10	0,02	20,10	0,12	0,03	27,59	0,11	0,02	15,95	0,1337	0,4660
		pozostałe gospodarstwa	0,16	0,02	13,10	0,15	0,03	21,40	0,15	0,02	11,27	0,1684	0,4548
	Liczba pokoi	1	0,15	0,02	15,92	0,15	0,04	23,83	0,15	0,02	13,18	0,1560	0,4573
		2	0,38	0,02	4,86	0,34	0,03	8,47	0,36	0,02	4,31	0,1664	0,4560
		3	0,31	0,02	6,02	0,34	0,03	8,44	0,33	0,02	4,85	0,1621	0,4463
		4	0,09	0,02	24,57	0,09	0,03	37,55	0,09	0,02	20,54	0,1607	0,4551
		5	0,05	0,02	40,52	0,05	0,03	63,20	0,05	0,02	34,21	0,1664	0,4520
		6 i więcej	0,02	0,02	126,33	0,03	0,03	104,00	0,02	0,02	78,12	0,0729	0,4336
	Czy jest	tak	0,90	0,01	0,80	0,91	0,01	1,18	0,91	0,01	0,66	0,1750	0,4425

Województwo	Zmienna	Wariant	BBGD			EU-SILC			Zintegrowany			$1-\frac{s_{int}(p)}{s_{BBGD}(p)}$	$1-\frac{s_{int}(p)}{s_{EU-SILC}(p)}$
			p	$S(p)$	CV	p	$S(p)$	CV	p	$S(p)$	CV		
	ustęp spłukiwany ?	nie	0,10	0,02	22,61	0,09	0,03	36,83	0,09	0,02	19,39	0,1641	0,4598
	Czy jest łazienka ?	tak	0,89	0,01	0,86	0,88	0,01	1,38	0,89	0,01	0,74	0,1471	0,4620
		nie	0,11	0,02	20,42	0,12	0,03	28,57	0,11	0,02	16,47	0,1576	0,4469
	Czy gospodarstwo posiada TV?	tak	0,99	0,00	0,22	0,97	0,01	0,62	0,98	0,00	0,28	-0,2194	0,5546
		nie	0,01	0,03	302,97	0,03	0,04	143,75	0,02	0,03	134,77	0,0466	0,3842
	Czy gospodarstwo posiada komputer?	tak	0,37	0,02	4,88	0,45	0,03	6,01	0,41	0,02	3,70	0,1634	0,4364
		nie	0,63	0,01	2,29	0,55	0,02	4,23	0,59	0,01	2,10	0,1399	0,4709
	Czy gospodarstwo posiada pralkę?	tak	0,97	0,00	0,41	0,96	0,01	0,77	0,96	0,00	0,38	0,0773	0,4944
		nie	0,03	0,03	87,03	0,04	0,04	89,20	0,04	0,02	61,02	0,1422	0,4196
	Czy gospodarstwo posiada samochód?	tak	0,46	0,02	3,61	0,48	0,02	5,18	0,47	0,01	2,93	0,1654	0,4475
		nie	0,54	0,02	3,02	0,52	0,03	4,87	0,53	0,01	2,60	0,1605	0,4549
lubelskie	Rodzaj budynku	budynek wielorodzinny	0,41	0,02	4,39	0,37	0,03	7,25	0,39	0,01	3,82	0,1719	0,4443
		dom jednorodzinny	0,56	0,01	2,55	0,60	0,02	3,26	0,58	0,01	1,98	0,1920	0,4129
		dom jednorodzinny w zabudowie szeregowej	0,03	0,02	75,45	0,03	0,03	101,62	0,03	0,02	60,58	0,1974	0,4036
		inne	0,01	0,02	379,16	0,00	0,03	964,44	0,00	0,02	376,83	0,1942	0,4887
	Tytuł prawny do zajmowanego mieszkania	własność	0,65	0,01	1,96	0,67	0,02	2,65	0,66	0,01	1,56	0,1894	0,4182
		najem wg cen rynkowych	0,02	0,02	93,38	0,03	0,03	102,05	0,03	0,02	67,83	0,1589	0,4159
		najem poniżej cen rynkowych	0,01	0,03	326,60	0,01	0,03	350,24	0,01	0,02	237,65	0,2152	0,3680
		inne	0,32	0,02	6,00	0,29	0,03	9,85	0,30	0,02	5,20	0,1755	0,4427
	Typ biolo-	małżeństwo bez dzieci	0,17	0,02	11,28	0,19	0,03	14,27	0,18	0,02	8,75	0,1692	0,4261

Województwo	Zmienna	Wariant	BBGD			EU-SILC			Zintegrowany			$1-\frac{s_{int}(p)}{s_{BBGD}(p)}$	$1-\frac{s_{int}(p)}{s_{EU-SILC}(p)}$
			p	$S(p)$	CV	p	$S(p)$	CV	p	$S(p)$	CV		
	giczny gospodarstwa domowego	małżeństwo z 1 dzieckiem na utrzymaniu	0,11	0,02	19,24	0,14	0,03	20,99	0,12	0,02	13,93	0,1579	0,4185
		małżeństwo z 2 dzieci na utrzymaniu	0,11	0,02	17,62	0,10	0,03	28,59	0,11	0,02	15,15	0,1748	0,4470
		małżeństwo z 3 więcej dzieci na utrzymaniu	0,07	0,02	29,86	0,05	0,03	58,34	0,06	0,02	27,66	0,2085	0,4273
		samotny rodzic z co najmniej jedną osobą na utrzymaniu	0,05	0,02	38,51	0,03	0,03	95,69	0,04	0,02	38,32	0,2042	0,4639
		gospodarstwa jednoosobowe	0,21	0,03	11,81	0,21	0,03	15,33	0,21	0,02	9,38	0,2197	0,3768
		inne gospodarstwa z osobami na utrzymaniu	0,11	0,02	18,16	0,09	0,03	31,40	0,10	0,02	15,95	0,1708	0,4596
		pozostałe gospodarstwa	0,17	0,02	11,20	0,18	0,03	15,50	0,18	0,02	9,02	0,1692	0,4354
	Liczba pokoi	1	0,20	0,02	10,49	0,14	0,03	21,26	0,17	0,02	9,87	0,1969	0,4376
		2	0,35	0,02	5,10	0,33	0,03	7,84	0,34	0,01	4,31	0,1785	0,4337
		3	0,28	0,02	6,69	0,31	0,03	8,80	0,29	0,02	5,27	0,1725	0,4288
		4	0,09	0,02	21,69	0,11	0,03	27,85	0,10	0,02	16,89	0,1593	0,4353
		5	0,06	0,02	36,83	0,05	0,03	62,18	0,05	0,02	32,34	0,2003	0,4229
		6 i więcej	0,02	0,02	87,11	0,07	0,03	43,61	0,05	0,02	40,48	0,0734	0,3828
	Czy jest ustęp spłukiwany ?	tak	0,80	0,01	1,22	0,82	0,01	1,65	0,81	0,01	0,97	0,1935	0,4177
		nie	0,20	0,02	10,18	0,18	0,03	15,63	0,19	0,02	8,62	0,1910	0,4214
	Czy jest łazienka ?	tak	0,82	0,01	1,15	0,83	0,01	1,62	0,82	0,01	0,93	0,1849	0,4255
		nie	0,18	0,02	11,15	0,17	0,03	16,20	0,18	0,02	9,23	0,1904	0,4175
	Czy gospodarstwo posiada TV?	tak	0,98	0,00	0,34	0,95	0,01	0,76	0,96	0,00	0,36	-0,0523	0,5191
		nie	0,02	0,03	118,80	0,05	0,03	61,95	0,04	0,02	57,53	0,2016	0,3356
	Czy gospo-	tak	0,33	0,02	5,24	0,41	0,02	5,95	0,37	0,01	3,86	0,1761	0,4148

Województwo	Zmienna	Wariant	BBGD			EU-SILC			Zintegrowany			$1-\frac{s_{int}(p)}{s_{BBGD}(p)}$	$1-\frac{s_{int}(p)}{s_{EU-SILC}(p)}$
			p	$S(p)$	CV	p	$S(p)$	CV	p	$S(p)$	CV		
	darstwo posiada komputer?	nie	0,67	0,01	1,95	0,59	0,02	3,47	0,63	0,01	1,77	0,1470	0,4542
	Czy gospodarstwo posiada pralkę?	tak	0,95	0,00	0,50	0,96	0,01	0,69	0,96	0,00	0,40	0,1932	0,4185
		nie	0,05	0,02	52,70	0,04	0,03	79,70	0,04	0,02	44,27	0,1903	0,4229
	Czy gospodarstwo posiada samochód?	tak	0,50	0,01	2,97	0,53	0,02	4,08	0,52	0,01	2,38	0,1770	0,4319
		nie	0,50	0,02	3,31	0,47	0,02	4,97	0,48	0,01	2,78	0,1824	0,4242
lubuskie	Rodzaj budynku	budynek wielorodzinny	0,63	0,02	3,32	0,61	0,03	5,39	0,62	0,02	2,85	0,1534	0,4644
		dom jednorodzinny	0,29	0,03	9,38	0,31	0,04	13,24	0,30	0,02	7,59	0,1625	0,4461
		dom jednorodzinny w zabudowie szeregowej	0,08	0,03	37,83	0,07	0,05	65,60	0,08	0,03	33,14	0,1620	0,4706
		inne	0,00	0,03	1569,37	0,00	0,06	1551,79	0,00	0,03	1038,35	-0,0065	0,5013
	Tytuł prawny do zajmowanego mieszkania	własność	0,64	0,02	3,12	0,60	0,03	5,30	0,62	0,02	2,75	0,1491	0,4647
		najem wg cen rynkowych	0,02	0,03	196,32	0,06	0,06	93,78	0,04	0,03	87,31	0,0000	0,4009
		najem poniżej cen rynkowych	0,01	0,04	302,32	0,02	0,05	231,58	0,02	0,03	182,26	0,1266	0,3988
		inne	0,33	0,03	8,29	0,32	0,04	13,91	0,33	0,02	7,17	0,1515	0,4747
	Typ biologiczny gospodarstwa domowego	małżeństwo bez dzieci	0,18	0,03	15,76	0,22	0,04	19,56	0,20	0,02	12,02	0,1469	0,4446
		małżeństwo z 1 dzieckiem na utrzymaniu	0,12	0,03	25,41	0,13	0,05	37,54	0,12	0,03	20,67	0,1246	0,4855
		małżeństwo z 2 dzieci na utrzymaniu	0,10	0,03	31,36	0,09	0,05	54,45	0,09	0,03	27,51	0,1628	0,4696
		małżeństwo z 3 więcej dzieci na utrzymaniu	0,05	0,03	65,01	0,04	0,05	134,87	0,04	0,02	60,69	0,1752	0,4820
		samotny rodzic z co najmniej jedną osobą na utrzy-	0,06	0,03	52,74	0,02	0,05	257,41	0,04	0,02	61,59	0,2315	0,5036

Województwo	Zmienna	Wariant	BBGD			EU-SILC			Zintegrowany			$1-\frac{s_{int}(p)}{s_{BBGD}(p)}$	$1-\frac{s_{int}(p)}{s_{EU-SILC}(p)}$
			p	$S(p)$	CV	p	$S(p)$	CV	p	$S(p)$	CV		
		maniu											
		gospodarstwa jednoosobowe	0,27	0,04	13,23	0,24	0,05	21,53	0,25	0,03	11,48	0,1885	0,4273
		inne gospodar-stwa z osobami na utrzyma-niu	0,10	0,03	29,63	0,09	0,04	51,18	0,09	0,02	26,04	0,1729	0,4575
		pozostałe gospodarstwa	0,13	0,03	21,70	0,18	0,05	25,44	0,16	0,03	16,09	0,1325	0,4479
	Liczba pokoi	1	0,13	0,03	26,61	0,12	0,05	46,17	0,12	0,03	23,35	0,1633	0,4684
		2	0,35	0,03	7,73	0,35	0,04	11,80	0,35	0,02	6,48	0,1665	0,4483
		3	0,31	0,03	8,80	0,30	0,04	14,44	0,30	0,02	7,57	0,1581	0,4649
		4	0,12	0,03	25,35	0,11	0,05	41,38	0,12	0,03	21,72	0,1590	0,4653
		5	0,05	0,03	56,02	0,07	0,05	72,77	0,06	0,03	43,50	0,1315	0,4593
		6 i więcej	0,04	0,03	83,83	0,06	0,05	88,02	0,05	0,03	58,95	0,1084	0,4469
	Czy jest ustęp spłu-kiwany ?	tak	0,91	0,01	1,11	0,93	0,01	1,46	0,92	0,01	0,86	0,2102	0,4148
		nie	0,09	0,03	36,10	0,07	0,05	73,15	0,08	0,03	33,60	0,1817	0,4676
	Czy jest łazienka ?	tak	0,91	0,01	1,09	0,92	0,01	1,54	0,92	0,01	0,87	0,1894	0,4352
		nie	0,09	0,03	38,01	0,08	0,05	65,99	0,08	0,03	33,56	0,1819	0,4478
	Czy gospo-darstwo posiada TV?	tak	0,98	0,00	0,49	0,97	0,01	0,90	0,98	0,00	0,45	0,0799	0,4980
		nie	0,02	0,04	183,74	0,03	0,06	192,50	0,02	0,03	130,43	0,1462	0,4200
	Czy gospo-darstwo posiada komputer?	tak	0,40	0,02	6,21	0,46	0,04	8,30	0,43	0,02	4,87	0,1565	0,4527
		nie	0,60	0,02	3,61	0,54	0,03	6,44	0,57	0,02	3,25	0,1471	0,4669
	Czy gospo-darstwo posiada pral-kę?	tak	0,97	0,01	0,61	0,97	0,01	0,89	0,97	0,00	0,50	0,1836	0,4400
		nie	0,03	0,04	114,59	0,03	0,06	206,38	0,03	0,03	101,64	0,1620	0,4770
	Czy gospo-	tak	0,47	0,02	4,99	0,56	0,03	6,15	0,51	0,02	3,78	0,1713	0,4346

Województwo	Zmienna	Wariant	BBGD			EU-SILC			Zintegrowany			$1-\frac{s_{int}(p)}{s_{BBGD}(p)}$	$1-\frac{s_{int}(p)}{s_{EU-SILC}(p)}$
			p	$S(p)$	CV	p	$S(p)$	CV	p	$S(p)$	CV		
	darstwo posiada samochód?	nie	0,53	0,02	4,46	0,44	0,04	8,72	0,49	0,02	4,16	0,1448	0,4759
łódzkie	Rodzaj budynku	budynek wielorodzinny	0,63	0,01	1,96	0,59	0,02	3,54	0,61	0,01	1,75	0,1334	0,4906
		dom jednorodzinny	0,33	0,02	4,64	0,38	0,02	5,77	0,36	0,01	3,56	0,1722	0,4257
		dom jednorodzinny w zabudowie szeregowej	0,04	0,02	53,25	0,02	0,03	127,36	0,03	0,02	52,39	0,2030	0,4657
		inne	0,00	0,02	594,29	0,00	0,03	1214,34	0,00	0,02	560,55	0,2285	0,4096
	Tytuł prawny do zajmowanego mieszkania	własność	0,45	0,01	3,12	0,45	0,02	4,81	0,45	0,01	2,62	0,1625	0,4537
		najem wg cen rynkowych	0,01	0,02	148,24	0,02	0,03	187,03	0,02	0,02	113,90	0,1339	0,4517
		najem poniżej cen rynkowych	0,01	0,02	386,97	0,02	0,03	206,90	0,01	0,02	184,04	-0,0066	0,4137
		inne	0,53	0,01	2,64	0,52	0,02	4,32	0,52	0,01	2,26	0,1519	0,4695
	Typ biologiczny gospodarstwa domowego	małżeństwo bez dzieci	0,17	0,02	9,84	0,20	0,03	13,22	0,19	0,01	7,73	0,1364	0,4630
		małżeństwo z 1 dzieckiem na utrzymaniu	0,12	0,02	15,14	0,13	0,03	23,04	0,12	0,02	12,53	0,1376	0,4773
		małżeństwo z 2 dzieci na utrzymaniu	0,10	0,02	17,15	0,10	0,03	27,86	0,10	0,01	14,64	0,1548	0,4693
		małżeństwo z 3 więcej dzieci na utrzymaniu	0,04	0,02	42,61	0,03	0,03	85,91	0,04	0,01	39,48	0,1786	0,4745
		samotny rodzic z co najmniej jedną osobą na utrzymaniu	0,05	0,02	34,69	0,03	0,03	98,64	0,04	0,01	35,74	0,1991	0,4961
		gospodarstwa jednoosobowe	0,28	0,02	7,39	0,31	0,03	9,59	0,30	0,02	5,80	0,1808	0,4188
		inne gospodarstwa z osobami na utrzymaniu	0,08	0,02	21,83	0,09	0,03	27,98	0,09	0,01	16,94	0,1504	0,4419
		pozostałe gospodarstwa	0,15	0,02	10,96	0,11	0,03	23,63	0,13	0,01	10,50	0,1933	0,4553
	Liczba pokoi	1	0,22	0,02	8,27	0,21	0,03	14,17	0,21	0,02	7,23	0,1593	0,4696

Województwo	Zmienna	Wariant	BBGD			EU-SILC			Zintegrowany			$1-\frac{s_{int}(p)}{s_{BBGD}(p)}$	$1-\frac{s_{int}(p)}{s_{EU-SILC}(p)}$
			p	$S(p)$	CV	p	$S(p)$	CV	p	$S(p)$	CV		
		2	0,39	0,02	3,94	0,36	0,03	7,08	0,37	0,01	3,51	0,1520	0,4773
		3	0,26	0,02	6,38	0,29	0,03	8,98	0,27	0,01	5,14	0,1542	0,4538
		4	0,07	0,02	24,90	0,08	0,03	34,94	0,08	0,02	20,03	0,1454	0,4580
		5	0,04	0,02	51,49	0,05	0,03	55,85	0,04	0,02	37,18	0,1531	0,4186
		6 i więcej	0,02	0,02	99,03	0,03	0,03	104,67	0,02	0,02	70,34	0,1335	0,4292
	Czy jest ustęp spłukiwany ?	tak	0,85	0,01	0,90	0,84	0,01	1,48	0,84	0,01	0,78	0,1436	0,4724
		nie	0,15	0,02	12,23	0,16	0,03	16,74	0,16	0,02	9,81	0,1719	0,4311
	Czy jest łazienka ?	tak	0,83	0,01	0,97	0,81	0,01	1,64	0,82	0,01	0,84	0,1357	0,4783
		nie	0,17	0,02	11,04	0,19	0,03	14,47	0,18	0,02	8,69	0,1713	0,4272
	Czy gospodarstwo posiada TV?	tak	0,98	0,00	0,30	0,96	0,01	0,67	0,97	0,00	0,31	-0,0290	0,5320
		nie	0,02	0,02	98,57	0,04	0,03	72,18	0,03	0,02	57,80	0,1186	0,3972
	Czy gospodarstwo posiada komputer?	tak	0,35	0,02	4,36	0,39	0,02	6,15	0,37	0,01	3,49	0,1510	0,4607
		nie	0,65	0,01	1,84	0,61	0,02	3,14	0,63	0,01	1,62	0,1466	0,4675
	Czy gospodarstwo posiada pralkę?	tak	0,95	0,00	0,44	0,94	0,01	0,76	0,95	0,00	0,39	0,1146	0,4844
		nie	0,05	0,02	46,92	0,06	0,03	58,87	0,05	0,02	36,07	0,1464	0,4417
	Czy gospodarstwo posiada samochód?	tak	0,47	0,01	2,89	0,47	0,02	4,58	0,47	0,01	2,44	0,1533	0,4679
		nie	0,53	0,01	2,70	0,53	0,02	4,11	0,53	0,01	2,26	0,1644	0,4506
małopolskie	Rodzaj budynku	budynek wielorodzinny	0,48	0,01	2,98	0,41	0,02	5,79	0,45	0,01	2,76	0,1469	0,4801
		dom jednorodzinny	0,48	0,01	2,74	0,56	0,02	3,33	0,52	0,01	2,07	0,1795	0,4242
		dom jednorodzinny w zabudowie szeregowej	0,04	0,02	47,95	0,03	0,03	86,63	0,03	0,01	42,75	0,1732	0,4647
		inne	0,00	0,02	960,06	0,00	0,03	3660,5	0,00	0,02	1070,0	0,2518	0,4302

Województwo	Zmienna	Wariant	BBGD			EU-SILC			Zintegrowany			$1-\frac{s_{int}(p)}{s_{BBGD}(p)}$	$1-\frac{s_{int}(p)}{s_{EU-SILC}(p)}$
			p	$S(p)$	CV	p	$S(p)$	CV	p	$S(p)$	CV		
								2			3		
	Tytuł prawny do zajmowanego mieszkania	własność	0,62	0,01	1,81	0,65	0,02	2,58	0,64	0,01	1,47	0,1723	0,4432
		najem wg cen rynkowych	0,04	0,02	50,29	0,06	0,03	54,28	0,05	0,02	36,13	0,1421	0,4268
		najem poniżej cen rynkowych	0,03	0,02	71,35	0,01	0,03	227,26	0,02	0,02	75,11	0,1953	0,5235
		inne	0,31	0,02	5,39	0,28	0,03	9,18	0,29	0,01	4,72	0,1647	0,4600
	Typ biologiczny gospodarstwa domowego	małżeństwo bez dzieci	0,15	0,02	10,93	0,17	0,02	14,61	0,16	0,01	8,66	0,1622	0,4380
		małżeństwo z 1 dzieckiem na utrzymaniu	0,11	0,02	16,31	0,12	0,03	24,12	0,11	0,01	13,38	0,1428	0,4683
		małżeństwo z 2 dzieci na utrzymaniu	0,12	0,02	13,53	0,11	0,03	24,62	0,12	0,01	12,09	0,1662	0,4714
		małżeństwo z 3 więcej dzieci na utrzymaniu	0,06	0,02	29,50	0,04	0,03	61,78	0,05	0,01	27,69	0,1810	0,4754
		samotny rodzic z co najmniej jedną osobą na utrzymaniu	0,05	0,02	38,46	0,02	0,03	119,84	0,03	0,01	40,77	0,2179	0,4737
		gospodarstwa jednoosobowe	0,24	0,02	8,62	0,25	0,03	11,49	0,25	0,02	6,87	0,1871	0,4139
		inne gospodarstwa z osobami na utrzymaniu	0,12	0,02	14,06	0,12	0,03	22,33	0,12	0,01	11,89	0,1510	0,4700
		pozostałe gospodarstwa	0,15	0,02	10,68	0,17	0,03	14,69	0,16	0,01	8,54	0,1556	0,4477
	Liczba pokoi	1	0,15	0,02	12,58	0,12	0,03	25,04	0,13	0,02	11,70	0,1857	0,4562
		2	0,34	0,02	4,62	0,35	0,02	6,91	0,34	0,01	3,82	0,1599	0,4547
		3	0,30	0,02	5,10	0,27	0,02	9,15	0,29	0,01	4,55	0,1623	0,4677
		4	0,11	0,02	14,74	0,13	0,03	20,25	0,12	0,01	11,79	0,1551	0,4473
		5	0,06	0,02	30,70	0,09	0,03	29,80	0,07	0,02	20,85	0,1201	0,4298
		6 i więcej	0,03	0,02	52,91	0,05	0,03	54,17	0,04	0,02	36,67	0,0976	0,4501
	Czy jest	tak	0,93	0,01	0,54	0,94	0,01	0,72	0,94	0,00	0,43	0,2090	0,4154

Województwo	Zmienna	Wariant	BBGD			EU-SILC			Zintegrowany			$1-\frac{s_{int}(p)}{s_{BBGD}(p)}$	$1-\frac{s_{int}(p)}{s_{EU-SILC}(p)}$
			p	$S(p)$	CV	p	$S(p)$	CV	p	$S(p)$	CV		
	ustęp spłukiwany ?	nie	0,07	0,02	25,93	0,06	0,03	49,75	0,06	0,02	23,73	0,1872	0,4546
	Czy jest łazienka ?	tak	0,94	0,00	0,49	0,94	0,01	0,78	0,94	0,00	0,42	0,1482	0,4631
		nie	0,06	0,02	31,85	0,06	0,03	44,32	0,06	0,02	25,70	0,1650	0,4389
	Czy gospodarstwo posiada TV?	tak	0,97	0,00	0,33	0,95	0,01	0,67	0,96	0,00	0,32	0,0293	0,5132
		nie	0,03	0,02	74,14	0,05	0,03	64,65	0,04	0,02	47,87	0,1281	0,4115
	Czy gospodarstwo posiada komputer?	tak	0,43	0,01	3,30	0,46	0,02	4,62	0,45	0,01	2,65	0,1620	0,4499
		nie	0,57	0,01	2,19	0,54	0,02	3,67	0,55	0,01	1,91	0,1556	0,4596
	Czy gospodarstwo posiada pralkę?	tak	0,97	0,00	0,35	0,97	0,01	0,55	0,97	0,00	0,30	0,1510	0,4611
		nie	0,03	0,02	64,42	0,03	0,03	88,63	0,03	0,02	51,87	0,1728	0,4299
	Czy gospodarstwo posiada samochód?	tak	0,49	0,01	2,60	0,52	0,02	3,76	0,51	0,01	2,12	0,1636	0,4516
		nie	0,51	0,01	2,77	0,48	0,02	4,51	0,49	0,01	2,39	0,1616	0,4546
mazowieckie	Rodzaj budynku	budynek wielorodzinny	0,56	0,01	1,78	0,55	0,02	3,07	0,56	0,01	1,55	0,1397	0,4884
		dom jednorodzinny	0,37	0,01	2,92	0,40	0,02	3,96	0,39	0,01	2,33	0,1775	0,4275
		dom jednorodzinny w zabudowie szeregowej	0,06	0,01	24,50	0,05	0,02	44,99	0,05	0,01	21,95	0,1697	0,4708
		inne	0,00	0,01	380,52	0,00	0,02	561,62	0,00	0,01	314,02	0,1611	0,4497
	Tytuł prawny do zajmowanego mieszkania	własność	0,56	0,01	1,67	0,55	0,01	2,59	0,56	0,01	1,41	0,1642	0,4500
		najem wg cen rynkowych	0,03	0,02	46,25	0,04	0,02	66,02	0,04	0,01	37,41	0,1396	0,4649
		najem poniżej cen rynkowych	0,01	0,01	99,80	0,01	0,03	170,66	0,01	0,01	86,35	0,1428	0,4893
		inne	0,39	0,01	3,04	0,40	0,02	4,87	0,39	0,01	2,57	0,1473	0,4762
	Typ biolo-	małżeństwo bez dzieci	0,17	0,01	7,26	0,23	0,02	8,29	0,20	0,01	5,31	0,1357	0,4435

Województwo	Zmienna	Wariant	BBGD			EU-SILC			Zintegrowany			$1-\frac{s_{int}(p)}{s_{BBGD}(p)}$	$1-\frac{s_{int}(p)}{s_{EU-SILC}(p)}$
			p	$S(p)$	CV	p	$S(p)$	CV	p	$S(p)$	CV		
	giczny gospodarstwa domowego	małżeństwo z 1 dzieckiem na utrzymaniu	0,11	0,01	11,47	0,10	0,02	21,60	0,11	0,01	10,29	0,1484	0,4973
		małżeństwo z 2 dzieci na utrzymaniu	0,12	0,01	10,70	0,11	0,02	19,79	0,11	0,01	9,58	0,1584	0,4830
		małżeństwo z 3 więcej dzieci na utrzymaniu	0,05	0,01	25,23	0,04	0,02	53,55	0,04	0,01	23,89	0,1933	0,4616
		samotny rodzic z co najmniej jedną osobą na utrzymaniu	0,05	0,01	26,45	0,03	0,02	68,68	0,04	0,01	26,53	0,1916	0,4938
		gospodarstwa jednoosobowe	0,26	0,02	6,02	0,27	0,02	8,50	0,27	0,01	4,90	0,1752	0,4317
		inne gospodarstwa z osobami na utrzymaniu	0,08	0,01	15,65	0,11	0,02	17,85	0,10	0,01	11,48	0,1325	0,4419
		pozostałe gospodarstwa	0,15	0,01	8,78	0,11	0,02	17,23	0,13	0,01	8,15	0,1954	0,4424
	Liczba pokoi	1	0,19	0,01	7,36	0,16	0,02	14,69	0,17	0,01	6,79	0,1604	0,4878
		2	0,35	0,01	3,47	0,35	0,02	5,42	0,35	0,01	2,92	0,1540	0,4656
		3	0,28	0,01	4,39	0,28	0,02	6,81	0,28	0,01	3,68	0,1555	0,4629
		4	0,10	0,01	13,52	0,11	0,02	18,11	0,10	0,01	10,71	0,1590	0,4403
		5	0,05	0,01	25,80	0,05	0,02	36,32	0,05	0,01	20,93	0,1661	0,4391
		6 i więcej	0,04	0,01	36,72	0,05	0,02	42,01	0,04	0,01	27,13	0,1464	0,4303
	Czy jest ustęp spłukiwany ?	tak	0,90	0,00	0,50	0,90	0,01	0,80	0,90	0,00	0,43	0,1505	0,4659
		nie	0,10	0,01	14,33	0,10	0,02	20,47	0,10	0,01	11,70	0,1692	0,4381
	Czy jest łazienka ?	tak	0,90	0,00	0,51	0,89	0,01	0,87	0,89	0,00	0,45	0,1327	0,4772
		nie	0,10	0,01	13,81	0,11	0,02	18,05	0,11	0,01	10,85	0,1648	0,4320
	Czy gospodarstwo posiada TV?	tak	0,98	0,00	0,19	0,98	0,00	0,36	0,98	0,00	0,18	0,0677	0,5034
		nie	0,02	0,02	97,63	0,02	0,02	98,23	0,02	0,01	67,92	0,1437	0,4164
	Czy gospo-	tak	0,42	0,01	2,59	0,46	0,02	3,65	0,44	0,01	2,08	0,1563	0,4575

Województwo	Zmienna	Wariant	BBGD			EU-SILC			Zintegrowany			$1-\frac{s_{int}(p)}{s_{BBGD}(p)}$	$1-\frac{s_{int}(p)}{s_{EU-SILC}(p)}$
			p	$S(p)$	CV	p	$S(p)$	CV	p	$S(p)$	CV		
	darstwo posiada komputer?	nie	0,58	0,01	1,66	0,54	0,02	2,84	0,56	0,01	1,46	0,1521	0,4640
	Czy gospodarstwo posiada pralkę?	tak	0,96	0,00	0,30	0,95	0,00	0,50	0,96	0,00	0,26	0,1303	0,4759
		nie	0,04	0,02	40,91	0,05	0,02	55,36	0,04	0,01	32,49	0,1493	0,4491
	Czy gospodarstwo posiada samochód?	tak	0,48	0,01	2,09	0,52	0,02	2,97	0,50	0,01	1,68	0,1617	0,4535
		nie	0,52	0,01	2,05	0,48	0,02	3,49	0,50	0,01	1,80	0,1548	0,4643
opolskie	Rodzaj budynku	budynek wielorodzinny	0,60	0,02	3,51	0,57	0,03	5,76	0,58	0,02	3,04	0,1560	0,4588
		dom jednorodzinny	0,37	0,02	6,45	0,41	0,03	8,42	0,39	0,02	5,05	0,1742	0,4286
		dom jednorodzinny w zabudowie szeregowej	0,03	0,03	113,67	0,02	0,05	195,96	0,03	0,03	100,44	0,1939	0,4365
		inne	0,01	0,03	619,44	0,00	--	--	0,00	0,02	890,97	0,3029	--
	Tytuł prawny do zajmowanego mieszkania	własność	0,55	0,02	3,75	0,63	0,03	4,52	0,59	0,02	2,82	0,1927	0,4129
		najem wg cen rynkowych	0,03	0,03	132,14	0,03	0,05	150,76	0,03	0,03	97,52	0,1516	0,4237
		najem poniżej cen rynkowych	0,01	0,03	277,53	0,00	--	--	0,01	0,02	399,76	0,3019	--
		inne	0,41	0,03	6,13	0,34	0,04	11,72	0,37	0,02	5,67	0,1629	0,4638
	Typ biologiczny gospodarstwa domowego	małżeństwo bez dzieci	0,19	0,03	14,26	0,23	0,04	17,21	0,21	0,02	10,80	0,1676	0,4217
		małżeństwo z 1 dzieckiem na utrzymaniu	0,12	0,03	24,14	0,12	0,04	36,13	0,12	0,02	20,01	0,1603	0,4527
		małżeństwo z 2 dzieci na utrzymaniu	0,09	0,03	29,81	0,08	0,04	52,22	0,09	0,02	26,30	0,1679	0,4663
		małżeństwo z 3 więcej dzieci na utrzymaniu	0,05	0,03	59,53	0,03	0,04	136,29	0,04	0,02	58,61	0,2364	0,4092
		samotny rodzic z co najmniej jedną osobą na utrzymaniu	0,04	0,03	66,52	0,03	0,05	154,02	0,04	0,02	64,79	0,1951	0,4757

Województwo	Zmienna	Wariant	BBGD			EU-SILC			Zintegrowany			$1-\frac{s_{int}(p)}{s_{BBGD}(p)}$	$1-\frac{s_{int}(p)}{s_{EU-SILC}(p)}$
			p	$S(p)$	CV	p	$S(p)$	CV	p	$S(p)$	CV		
		gospodarstwa jednoosobowe	0,26	0,03	13,47	0,24	0,05	20,13	0,25	0,03	11,30	0,1918	0,4180
		inne gospodarstwa z osobami na utrzymaniu	0,09	0,03	30,02	0,12	0,05	37,80	0,11	0,02	22,84	0,1191	0,4643
		pozostałe gospodarstwa	0,16	0,03	17,83	0,15	0,04	28,17	0,15	0,02	15,10	0,1608	0,4591
	Liczba pokoi	1	0,09	0,03	38,32	0,08	0,05	62,10	0,08	0,03	33,03	0,1837	0,4387
		2	0,36	0,03	7,36	0,35	0,04	11,26	0,36	0,02	6,18	0,1663	0,4487
		3	0,28	0,03	9,81	0,25	0,04	15,82	0,26	0,02	8,46	0,1803	0,4375
		4	0,12	0,03	22,91	0,14	0,04	30,97	0,13	0,02	18,14	0,1481	0,4508
		5	0,08	0,03	34,20	0,09	0,05	50,08	0,09	0,02	27,96	0,1456	0,4633
		6 i więcej	0,07	0,03	38,83	0,09	0,04	48,06	0,08	0,02	29,71	0,1554	0,4320
	Czy jest ustęp spłukiwany ?	tak	0,93	0,01	0,91	0,94	0,01	1,23	0,94	0,01	0,72	0,2048	0,4179
		nie	0,07	0,03	44,68	0,06	0,04	71,86	0,06	0,03	38,72	0,2131	0,4046
	Czy jest łazienka ?	tak	0,94	0,01	0,81	0,94	0,01	1,26	0,94	0,01	0,69	0,1568	0,4528
		nie	0,06	0,03	55,65	0,06	0,04	71,34	0,06	0,03	43,70	0,1939	0,4019
	Czy gospodarstwo posiada TV?	tak	0,99	0,00	0,37	0,97	0,01	0,78	0,98	0,00	0,38	-0,0042	0,5127
		nie	0,01	0,04	285,48	0,03	0,05	186,28	0,02	0,03	157,84	0,1912	0,3470
	Czy gospodarstwo posiada komputer?	tak	0,36	0,02	6,64	0,46	0,04	7,64	0,41	0,02	4,87	0,1658	0,4272
		nie	0,64	0,02	3,10	0,54	0,03	5,92	0,59	0,02	2,90	0,1378	0,4673
	Czy gospodarstwo posiada pralkę?	tak	0,98	0,00	0,46	0,97	0,01	0,84	0,98	0,00	0,43	0,0706	0,4909
		nie	0,02	0,04	187,88	0,03	0,05	165,41	0,02	0,03	122,87	0,1808	0,3756
	Czy gospodarstwo posiada sa-	tak	0,50	0,02	4,32	0,52	0,03	6,21	0,51	0,02	3,51	0,1669	0,4467
		nie	0,50	0,02	4,71	0,48	0,04	7,28	0,49	0,02	3,99	0,1719	0,4391

Województwo	Zmienna	Wariant	BBGD			EU-SILC			Zintegrowany			$1-\frac{s_{int}(p)}{s_{BBGD}(p)}$	$1-\frac{s_{int}(p)}{s_{EU-SILC}(p)}$
			p	$S(p)$	CV	p	$S(p)$	CV	p	$S(p)$	CV		
	mochód?												
podkarpackie	Rodzaj budynku	budynek wielorodzinny	0,38	0,02	5,17	0,31	0,03	9,85	0,34	0,02	4,76	0,1666	0,4605
		dom jednorodzinny	0,57	0,02	2,61	0,64	0,02	3,10	0,61	0,01	1,97	0,2038	0,3978
		dom jednorodzinny w zabudowie szeregowej	0,04	0,02	58,46	0,05	0,04	79,60	0,04	0,02	46,63	0,1540	0,4462
		inne	0,01	0,02	459,24	0,01	0,04	793,98	0,01	0,02	397,35	0,1332	0,5004
	Tytuł prawny do zajmowanego mieszkania	własność	0,67	0,01	1,97	0,71	0,02	2,52	0,69	0,01	1,54	0,1996	0,4076
		najem wg cen rynkowych	0,01	0,02	188,12	0,01	0,04	313,61	0,01	0,02	161,52	0,1458	0,4823
		najem poniżej cen rynkowych	0,00	0,02	624,72	0,01	0,04	350,69	0,01	0,02	306,34	-0,0157	0,4253
		inne	0,31	0,02	6,66	0,26	0,03	11,77	0,29	0,02	5,94	0,1742	0,4501
	Typ biologiczny gospodarstwa domowego	małżeństwo bez dzieci	0,15	0,02	14,41	0,18	0,03	16,49	0,16	0,02	10,68	0,1688	0,4156
		małżeństwo z 1 dzieckiem na utrzymaniu	0,09	0,02	25,82	0,07	0,03	46,20	0,08	0,02	22,95	0,1716	0,4637
		małżeństwo z 2 dzieci na utrzymaniu	0,11	0,02	20,43	0,10	0,03	31,25	0,10	0,02	17,16	0,1735	0,4417
		małżeństwo z 3 więcej dzieci na utrzymaniu	0,08	0,02	27,39	0,06	0,03	50,21	0,07	0,02	24,70	0,1894	0,4453
		samotny rodzic z co najmniej jedną osobą na utrzymaniu	0,05	0,02	42,89	0,03	0,03	126,47	0,04	0,02	44,90	0,2204	0,4580
		gospodarstwa jednoosobowe	0,22	0,03	12,18	0,21	0,03	16,91	0,21	0,02	9,95	0,2131	0,3881
		inne gospodarstwa z osobami na utrzymaniu	0,16	0,02	12,80	0,14	0,03	20,60	0,15	0,02	11,00	0,1808	0,4386
		pozostałe gospodarstwa	0,16	0,02	13,37	0,21	0,03	14,78	0,18	0,02	9,71	0,1511	0,4270
	Liczba pokoi	1	0,15	0,02	16,39	0,13	0,03	27,49	0,14	0,02	14,35	0,1907	0,4313
		2	0,31	0,02	6,39	0,28	0,03	10,54	0,30	0,02	5,55	0,1762	0,4423
		3	0,28	0,02	7,11	0,27	0,03	10,43	0,28	0,02	5,88	0,1778	0,4321

Województwo	Zmienna	Wariant	BBGD			EU-SILC			Zintegrowany			$1-\frac{s_{int}(p)}{s_{BBGD}(p)}$	$1-\frac{s_{int}(p)}{s_{EU-SILC}(p)}$
			p	$S(p)$	CV	p	$S(p)$	CV	p	$S(p)$	CV		
		4	0,13	0,02	16,28	0,14	0,03	21,96	0,14	0,02	12,98	0,1689	0,4323
		5	0,08	0,02	26,46	0,11	0,03	28,55	0,10	0,02	19,07	0,1578	0,4166
		6 i więcej	0,05	0,02	45,47	0,07	0,03	47,88	0,06	0,02	32,28	0,1389	0,4270
	Czy jest ustęp spłukiwany ?	tak	0,88	0,01	0,94	0,86	0,01	1,46	0,87	0,01	0,80	0,1542	0,4488
		nie	0,12	0,02	19,26	0,14	0,03	23,47	0,13	0,02	14,78	0,1891	0,4028
	Czy jest łazienka ?	tak	0,89	0,01	0,87	0,87	0,01	1,43	0,88	0,01	0,76	0,1333	0,4608
		nie	0,11	0,02	22,17	0,13	0,03	24,47	0,12	0,02	16,24	0,1844	0,3977
	Czy gospodarstwo posiada TV?	tak	0,98	0,00	0,35	0,95	0,01	0,79	0,97	0,00	0,37	-0,0534	0,5194
		nie	0,02	0,03	123,13	0,05	0,03	71,33	0,03	0,02	63,35	0,1512	0,3638
	Czy gospodarstwo posiada komputer?	tak	0,37	0,02	4,87	0,41	0,03	6,46	0,39	0,02	3,84	0,1735	0,4308
		nie	0,63	0,01	2,37	0,59	0,02	3,71	0,61	0,01	2,02	0,1688	0,4374
	Czy gospodarstwo posiada pralkę?	tak	0,97	0,00	0,43	0,97	0,01	0,66	0,97	0,00	0,36	0,1504	0,4485
		nie	0,03	0,03	86,09	0,03	0,04	105,61	0,03	0,02	66,15	0,1788	0,4117
	Czy gospodarstwo posiada samochód?	tak	0,53	0,02	2,98	0,54	0,02	4,24	0,53	0,01	2,42	0,1748	0,4356
		nie	0,47	0,02	3,83	0,46	0,03	5,58	0,47	0,01	3,17	0,1838	0,4226
podlaskie	Rodzaj budynku	budynek wielorodzinny	0,48	0,02	4,72	0,48	0,03	7,07	0,48	0,02	3,92	0,1669	0,4467
		dom jednorodzinny	0,47	0,02	4,42	0,47	0,03	6,63	0,47	0,02	3,69	0,1712	0,4406
		dom jednorodzinny w zabudowie szeregowej	0,05	0,03	64,86	0,05	0,05	102,46	0,05	0,02	54,79	0,1545	0,4657
		inne	0,00	--	--	0,00	0,05	2081,38	0,00	0,03	2982,79	--	0,3014
	Tytuł prawny	własność	0,58	0,02	3,30	0,59	0,03	4,74	0,59	0,02	2,69	0,1737	0,4381

Województwo	Zmienna	Wariant	BBGD			EU-SILC			Zintegrowany			$1-\frac{s_{int}(p)}{s_{BBGD}(p)}$	$1-\frac{s_{int}(p)}{s_{EU-SILC}(p)}$
			p	$S(p)$	CV	p	$S(p)$	CV	p	$S(p)$	CV		
	do zajmowanego mieszkania	najem wg cen rynkowych	0,02	0,03	157,69	0,04	0,05	128,19	0,03	0,03	98,26	0,1159	0,4150
		najem poniżej cen rynkowych	0,01	0,03	306,56	0,00	0,04	1155,94	0,01	0,02	338,67	0,2431	0,4331
		inne	0,39	0,02	6,21	0,37	0,04	10,10	0,38	0,02	5,34	0,1652	0,4531
	Typ biologiczny gospodarstwa domowego	małżeństwo bez dzieci	0,16	0,03	16,27	0,22	0,04	17,35	0,19	0,02	11,64	0,1542	0,4223
		małżeństwo z 1 dzieckiem na utrzymaniu	0,09	0,03	31,15	0,09	0,04	47,95	0,09	0,02	26,06	0,1538	0,4631
		małżeństwo z 2 dzieci na utrzymaniu	0,10	0,03	27,33	0,10	0,04	41,46	0,10	0,02	22,80	0,1618	0,4530
		małżeństwo z 3 więcej dzieci na utrzymaniu	0,06	0,03	43,41	0,04	0,04	102,13	0,05	0,02	42,52	0,2079	0,4459
		samotny rodzic z co najmniej jedną osobą na utrzymaniu	0,05	0,03	56,11	0,02	0,04	162,04	0,04	0,02	58,17	0,2159	0,4573
		gospodarstwa jednoosobowe	0,29	0,03	10,95	0,28	0,04	15,16	0,29	0,03	8,91	0,2018	0,4000
		inne gospodarstwa z osobami na utrzymaniu	0,10	0,03	27,56	0,08	0,04	54,42	0,09	0,02	25,25	0,1655	0,4826
		pozostałe gospodarstwa	0,16	0,03	16,26	0,17	0,04	23,71	0,17	0,02	13,27	0,1448	0,4662
	Liczba pokoi	1	0,11	0,03	29,33	0,10	0,04	42,46	0,11	0,03	24,20	0,1878	0,4203
		2	0,34	0,03	7,50	0,30	0,04	13,33	0,32	0,02	6,66	0,1596	0,4682
		3	0,34	0,02	7,02	0,34	0,04	10,85	0,34	0,02	5,91	0,1652	0,4508
		4	0,12	0,03	23,20	0,15	0,04	26,70	0,13	0,02	17,26	0,1640	0,4209
		5	0,06	0,03	47,18	0,08	0,04	52,41	0,07	0,02	34,41	0,1435	0,4323
		6 i więcej	0,04	0,03	68,99	0,03	0,04	114,94	0,04	0,02	59,82	0,1769	0,4485
	Czy jest ustęp spłukiwany ?	tak	0,85	0,01	1,37	0,90	0,01	1,61	0,87	0,01	1,02	0,2305	0,3811
		nie	0,15	0,03	18,88	0,10	0,04	40,33	0,13	0,02	17,99	0,1951	0,4471
	Czy jest	tak	0,85	0,01	1,36	0,90	0,01	1,62	0,87	0,01	1,02	0,2278	0,3849

Województwo	Zmienna	Wariant	BBGD			EU-SILC			Zintegrowany			$1-\frac{s_{int}(p)}{s_{BBGD}(p)}$	$1-\frac{s_{int}(p)}{s_{EU-SILC}(p)}$
			p	$S(p)$	CV	p	$S(p)$	CV	p	$S(p)$	CV		
	łazienka ?	nie	0,15	0,03	19,28	0,10	0,04	40,15	0,13	0,02	18,22	0,1963	0,4430
	Czy gospodarstwo posiada TV?	tak	0,98	0,00	0,48	0,93	0,01	1,31	0,95	0,01	0,57	-0,1590	0,5512
		nie	0,02	0,03	137,99	0,07	0,04	59,68	0,05	0,03	59,14	0,1609	0,3457
	Czy gospodarstwo posiada komputer?	tak	0,31	0,02	7,95	0,37	0,04	9,63	0,34	0,02	6,01	0,1625	0,4335
		nie	0,69	0,02	2,44	0,63	0,03	4,37	0,66	0,01	2,20	0,1381	0,4679
	Czy gospodarstwo posiada pralkę?	tak	0,93	0,01	0,82	0,94	0,01	1,19	0,94	0,01	0,67	0,1810	0,4352
		nie	0,07	0,03	48,53	0,06	0,05	73,68	0,06	0,03	40,78	0,1867	0,4265
	Czy gospodarstwo posiada samochód?	tak	0,44	0,02	4,83	0,43	0,03	7,49	0,44	0,02	4,07	0,1638	0,4521
		nie	0,56	0,02	3,76	0,57	0,03	5,43	0,56	0,02	3,08	0,1750	0,4354
pomorskie	Rodzaj budynku	budynek wielorodzinny	0,69	0,01	1,90	0,68	0,02	3,16	0,68	0,01	1,64	0,1450	0,4767
		dom jednorodzinny	0,22	0,02	8,69	0,24	0,03	12,25	0,23	0,02	7,01	0,1552	0,4528
		dom jednorodzinny w zabudowie szeregowej	0,09	0,02	24,65	0,08	0,03	41,86	0,08	0,02	21,48	0,1688	0,4594
		inne	0,01	0,02	383,73	0,01	0,04	562,29	0,01	0,02	316,27	0,1659	0,4444
	Tytuł prawny do zajmowanego mieszkania	własność	0,55	0,01	2,70	0,58	0,02	4,00	0,57	0,01	2,22	0,1615	0,4568
		najem wg cen rynkowych	0,02	0,02	91,20	0,05	0,04	82,57	0,04	0,02	58,64	0,0303	0,4742
		najem poniżej cen rynkowych	0,01	0,02	168,31	0,01	0,03	293,25	0,01	0,02	148,84	0,1913	0,4378
		inne	0,41	0,02	4,46	0,36	0,03	8,13	0,38	0,02	4,00	0,1521	0,4767
	Typ biologiczny gospodarstwa domowego	małżeństwo bez dzieci	0,18	0,02	11,24	0,22	0,03	13,58	0,20	0,02	8,49	0,1494	0,4402
		małżeństwo z 1 dzieckiem na utrzymaniu	0,11	0,02	17,99	0,14	0,03	25,40	0,13	0,02	14,40	0,1256	0,4790
		małżeństwo z 2 dzieci na utrzymaniu	0,12	0,02	16,99	0,08	0,03	39,06	0,10	0,02	16,41	0,1762	0,4876

Województwo	Zmienna	Wariant	BBGD			EU-SILC			Zintegrowany			$1-\frac{s_{int}(p)}{s_{BBGD}(p)}$	$1-\frac{s_{int}(p)}{s_{EU-SILC}(p)}$
			p	$S(p)$	CV	p	$S(p)$	CV	p	$S(p)$	CV		
		małżeństwo z 3 więcej dzieci na utrzymaniu	0,07	0,02	30,44	0,04	0,03	83,21	0,05	0,02	30,97	0,1963	0,4853
		samotny rodzic z co najmniej jedną osobą na utrzymaniu	0,06	0,02	35,94	0,03	0,03	110,03	0,04	0,02	37,97	0,2251	0,4450
		gospodarstwa jednoosobowe	0,23	0,03	10,93	0,21	0,04	18,00	0,22	0,02	9,44	0,1691	0,4528
		inne gospodarstwa z osobami na utrzymaniu	0,09	0,02	22,26	0,11	0,03	29,81	0,10	0,02	17,43	0,1214	0,4745
		pozostałe gospodarstwa	0,14	0,02	14,48	0,16	0,03	19,40	0,15	0,02	11,41	0,1409	0,4589
	Liczba pokoi	1	0,11	0,02	21,46	0,09	0,04	42,88	0,10	0,02	19,81	0,1724	0,4750
		2	0,45	0,02	3,84	0,47	0,03	5,70	0,46	0,01	3,16	0,1559	0,4617
		3	0,28	0,02	6,76	0,26	0,03	11,79	0,27	0,02	5,94	0,1565	0,4729
		4	0,09	0,02	23,85	0,09	0,03	35,22	0,09	0,02	19,62	0,1517	0,4598
		5	0,04	0,02	52,12	0,05	0,03	73,05	0,04	0,02	41,99	0,1481	0,4559
		6 i więcej	0,03	0,02	66,60	0,04	0,03	76,92	0,04	0,02	49,32	0,1340	0,4429
	Czy jest ustęp spłukiwany ?	tak	0,96	0,00	0,47	0,97	0,01	0,62	0,97	0,00	0,37	0,2142	0,4144
		nie	0,04	0,02	55,61	0,03	0,03	115,86	0,03	0,02	52,20	0,1915	0,4587
	Czy jest łazienka ?	tak	0,94	0,01	0,57	0,95	0,01	0,89	0,95	0,00	0,48	0,1619	0,4603
		nie	0,06	0,02	39,01	0,05	0,04	65,39	0,05	0,02	33,64	0,1522	0,4762
	Czy gospodarstwo posiada TV?	tak	0,99	0,00	0,27	0,99	0,00	0,36	0,99	0,00	0,21	0,2187	0,4105
		nie	0,01	0,03	184,37	0,01	0,04	382,52	0,01	0,02	173,06	0,1994	0,4486
	Czy gospodarstwo posiada komputer?	tak	0,39	0,02	4,43	0,47	0,03	5,77	0,43	0,01	3,43	0,1560	0,4515
		nie	0,61	0,01	2,41	0,53	0,02	4,57	0,57	0,01	2,21	0,1350	0,4838
	Czy gospo-	tak	0,97	0,00	0,41	0,97	0,01	0,62	0,97	0,00	0,34	0,1684	0,4561

Województwo	Zmienna	Wariant	BBGD			EU-SILC			Zintegrowany			$1-\frac{s_{int}(p)}{s_{BBGD}(p)}$	$1-\frac{s_{int}(p)}{s_{EU-SILC}(p)}$
			p	$S(p)$	CV	p	$S(p)$	CV	p	$S(p)$	CV		
	darstwo posiada pralkę?	nie	0,03	0,03	84,76	0,03	0,04	142,93	0,03	0,02	73,47	0,1618	0,4671
	Czy gospodarstwo posiada samochód?	tak	0,42	0,02	3,98	0,47	0,03	5,53	0,44	0,01	3,18	0,1599	0,4523
		nie	0,58	0,02	2,65	0,53	0,03	4,74	0,56	0,01	2,36	0,1411	0,4817
śląskie	Rodzaj budynku	budynek wielorodzinny	0,67	0,01	1,31	0,64	0,01	2,20	0,66	0,01	1,14	0,1472	0,4687
		dom jednorodzinny	0,30	0,01	3,93	0,34	0,02	5,11	0,32	0,01	3,07	0,1684	0,4330
		dom jednorodzinny w zabudowie szeregowej	0,02	0,01	58,89	0,02	0,02	143,17	0,02	0,01	58,00	0,1994	0,4712
		inne	0,00	0,02	1017,20	0,00	0,02	2328,67	0,00	0,01	992,43	0,2310	0,4136
	Tytuł prawny do zajmowanego mieszkania	własność	0,41	0,01	2,70	0,43	0,02	3,83	0,42	0,01	2,19	0,1688	0,4415
		najem wg cen rynkowych	0,01	0,01	98,52	0,03	0,02	77,16	0,02	0,01	59,70	0,0872	0,4234
		najem poniżej cen rynkowych	0,01	0,01	119,17	0,01	0,02	162,19	0,01	0,01	94,32	0,1276	0,4686
		inne	0,56	0,01	1,81	0,52	0,02	3,09	0,54	0,01	1,59	0,1509	0,4677
	Typ biologiczny gospodarstwa domowego	małżeństwo bez dzieci	0,17	0,01	7,33	0,24	0,02	7,85	0,21	0,01	5,23	0,1468	0,4282
		małżeństwo z 1 dzieckiem na utrzymaniu	0,13	0,01	10,00	0,12	0,02	17,93	0,12	0,01	8,87	0,1627	0,4729
		małżeństwo z 2 dzieci na utrzymaniu	0,13	0,01	9,91	0,11	0,02	19,22	0,12	0,01	9,08	0,1742	0,4687
		małżeństwo z 3 więcej dzieci na utrzymaniu	0,05	0,01	27,51	0,03	0,02	63,16	0,04	0,01	26,78	0,2077	0,4479
		samotny rodzic z co najmniej jedną osobą na utrzymaniu	0,06	0,01	24,38	0,02	0,02	84,83	0,04	0,01	26,55	0,2241	0,4689
		gospodarstwa jednoosobowe	0,26	0,02	6,17	0,25	0,02	9,17	0,26	0,01	5,13	0,1775	0,4335
		inne gospodarstwa z osobami na utrzyma-	0,06	0,01	20,84	0,10	0,02	21,41	0,08	0,01	14,57	0,1252	0,4349

Województwo	Zmienna	Wariant	BBGD			EU-SILC			Zintegrowany			$1-\frac{s_{int}(p)}{s_{BBGD}(p)}$	$1-\frac{s_{int}(p)}{s_{EU-SILC}(p)}$
			p	$S(p)$	CV	p	$S(p)$	CV	p	$S(p)$	CV		
		niu											
		pozostałe gospodarstwa	0,13	0,01	10,07	0,13	0,02	15,78	0,13	0,01	8,50	0,1619	0,4570
	Liczba pokoi	1	0,15	0,01	9,75	0,13	0,02	17,64	0,14	0,01	8,72	0,1721	0,4620
		2	0,36	0,01	3,32	0,36	0,02	5,22	0,36	0,01	2,81	0,1617	0,4563
		3	0,31	0,01	3,87	0,31	0,02	6,01	0,31	0,01	3,26	0,1638	0,4530
		4	0,09	0,01	14,96	0,10	0,02	20,12	0,10	0,01	11,86	0,1527	0,4468
		5	0,04	0,01	35,45	0,05	0,02	46,13	0,04	0,01	27,70	0,1471	0,4466
		6 i więcej	0,04	0,01	34,09	0,05	0,02	37,65	0,05	0,01	24,75	0,1386	0,4332
	Czy jest ustęp spłukiwany ?	tak	0,93	0,00	0,41	0,95	0,01	0,54	0,94	0,00	0,32	0,2103	0,4121
		nie	0,07	0,02	23,02	0,05	0,02	46,29	0,06	0,01	21,30	0,1770	0,4732
	Czy jest łazienka ?	tak	0,94	0,00	0,39	0,94	0,01	0,60	0,94	0,00	0,33	0,1628	0,4520
		nie	0,06	0,02	24,88	0,06	0,02	38,05	0,06	0,01	20,81	0,1609	0,4549
	Czy gospodarstwo posiada TV?	tak	0,98	0,00	0,20	0,97	0,00	0,39	0,98	0,00	0,19	0,0401	0,5091
		nie	0,02	0,02	94,82	0,03	0,02	85,69	0,02	0,01	62,45	0,1295	0,4157
	Czy gospodarstwo posiada komputer?	tak	0,42	0,01	2,64	0,47	0,02	3,48	0,44	0,01	2,07	0,1681	0,4387
		nie	0,58	0,01	1,70	0,53	0,02	2,98	0,56	0,01	1,51	0,1492	0,4672
	Czy gospodarstwo posiada pralkę?	tak	0,97	0,00	0,27	0,96	0,00	0,48	0,96	0,00	0,24	0,0903	0,4911
		nie	0,03	0,02	54,23	0,04	0,02	58,02	0,04	0,01	38,91	0,1485	0,4218
	Czy gospodarstwo posiada samochód?	tak	0,47	0,01	2,19	0,49	0,02	3,23	0,48	0,01	1,80	0,1646	0,4497
		nie	0,53	0,01	2,04	0,51	0,02	3,22	0,52	0,01	1,73	0,1620	0,4537
świętokrzyskie	Rodzaj bu-	budynek wielorodzinny	0,38	0,02	6,57	0,37	0,04	9,73	0,37	0,02	5,45	0,1723	0,4391

Województwo	Zmienna	Wariant	BBGD			EU-SILC			Zintegrowany			$1-\frac{s_{int}(p)}{s_{BBGD}(p)}$	$1-\frac{s_{int}(p)}{s_{EU-SILC}(p)}$
			p	$S(p)$	CV	p	$S(p)$	CV	p	$S(p)$	CV		
	dynku	dom jednorodzinny	0,58	0,02	3,22	0,60	0,03	4,46	0,59	0,02	2,59	0,1793	0,4306
		dom jednorodzinny w zabudowie szeregowej	0,05	0,03	62,48	0,03	0,04	166,14	0,04	0,02	63,40	0,2120	0,4584
		inne	0,00	0,03	898,99	0,00	0,05	2145,36	0,00	0,02	870,54	0,1735	0,5073
	Tytuł prawny do zajmowanego mieszkania	własność	0,65	0,02	2,59	0,68	0,02	3,52	0,67	0,01	2,07	0,1845	0,4271
		najem wg cen rynkowych	0,02	0,03	185,98	0,02	0,04	210,27	0,02	0,02	137,36	0,1614	0,4180
		najem poniżej cen rynkowych	0,01	0,03	222,00	0,00	0,05	1051,51	0,01	0,02	254,73	0,2276	0,5166
		inne	0,32	0,03	8,05	0,29	0,04	13,01	0,30	0,02	6,93	0,1750	0,4420
	Typ biologiczny gospodarstwa domowego	małżeństwo bez dzieci	0,18	0,03	14,51	0,20	0,04	18,54	0,19	0,02	11,30	0,1712	0,4256
		małżeństwo z 1 dzieckiem na utrzymaniu	0,09	0,03	28,45	0,07	0,04	52,04	0,08	0,02	25,66	0,1916	0,4407
		małżeństwo z 2 dzieci na utrzymaniu	0,11	0,03	23,15	0,10	0,04	41,24	0,11	0,02	20,59	0,1755	0,4570
		małżeństwo z 3 więcej dzieci na utrzymaniu	0,06	0,03	47,12	0,04	0,04	107,74	0,05	0,02	45,47	0,1865	0,4783
		samotny rodzic z co najmniej jedną osobą na utrzymaniu	0,06	0,03	45,51	0,04	0,04	110,35	0,05	0,02	44,86	0,1979	0,4684
		gospodarstwa jednoosobowe	0,22	0,03	15,76	0,23	0,04	18,92	0,22	0,03	12,07	0,2118	0,3806
		inne gospodarstwa z osobami na utrzymaniu	0,12	0,03	20,46	0,15	0,04	25,76	0,14	0,02	15,72	0,1429	0,4486
		pozostałe gospodarstwa	0,16	0,03	16,11	0,17	0,04	23,00	0,16	0,02	13,10	0,1578	0,4503
	Liczba pokoi	1	0,22	0,03	12,83	0,18	0,04	22,50	0,20	0,02	11,41	0,1829	0,4420
		2	0,38	0,02	6,13	0,37	0,04	9,37	0,38	0,02	5,15	0,1685	0,4457
		3	0,23	0,03	10,72	0,20	0,04	18,61	0,22	0,02	9,49	0,1793	0,4452
		4	0,10	0,03	26,48	0,12	0,04	32,32	0,11	0,02	20,13	0,1498	0,4382

Województwo	Zmienna	Wariant	BBGD			EU-SILC			Zintegrowany			$1-\frac{s_{int}(p)}{s_{BBGD}(p)}$	$1-\frac{s_{int}(p)}{s_{EU-SILC}(p)}$
			p	$S(p)$	CV	p	$S(p)$	CV	p	$S(p)$	CV		
		5	0,04	0,03	70,23	0,07	0,04	57,75	0,05	0,02	43,69	0,0902	0,4291
		6 i więcej	0,03	0,03	91,65	0,05	0,04	84,72	0,04	0,02	60,93	0,1199	0,4253
	Czy jest ustęp spłukiwany ?	tak	0,82	0,01	1,50	0,81	0,02	2,33	0,82	0,01	1,27	0,1595	0,4506
		nie	0,18	0,03	15,25	0,19	0,04	20,51	0,18	0,02	12,18	0,1796	0,4221
	Czy jest łazienka ?	tak	0,83	0,01	1,48	0,81	0,02	2,36	0,82	0,01	1,27	0,1518	0,4562
		nie	0,17	0,03	15,76	0,19	0,04	20,23	0,18	0,02	12,34	0,1800	0,4174
	Czy gospodarstwo posiada TV?	tak	0,99	0,00	0,34	0,97	0,01	0,75	0,98	0,00	0,36	-0,0473	0,5248
		nie	0,01	0,03	253,64	0,03	0,04	156,22	0,02	0,03	135,54	0,1431	0,3743
	Czy gospodarstwo posiada komputer?	tak	0,30	0,02	7,86	0,37	0,03	9,39	0,34	0,02	5,90	0,1624	0,4317
		nie	0,70	0,02	2,35	0,63	0,03	4,19	0,66	0,01	2,13	0,1399	0,4628
	Czy gospodarstwo posiada pralkę?	tak	0,96	0,01	0,59	0,97	0,01	0,71	0,97	0,00	0,45	0,2389	0,3760
		nie	0,04	0,03	86,30	0,03	0,05	185,31	0,03	0,03	82,02	0,2001	0,4512
	Czy gospodarstwo posiada samochód?	tak	0,48	0,02	4,13	0,55	0,03	5,24	0,52	0,02	3,19	0,1780	0,4281
		nie	0,52	0,02	4,21	0,45	0,03	7,28	0,48	0,02	3,74	0,1643	0,4484
warmińsko-mazurskie	Rodzaj budynku	budynek wielorodzinny	0,68	0,02	2,37	0,67	0,02	3,70	0,67	0,01	2,01	0,1609	0,4512
		dom jednorodzinny	0,22	0,02	10,94	0,26	0,03	13,04	0,24	0,02	8,23	0,1691	0,4192
		dom jednorodzinny w zabudowie szeregowej	0,10	0,03	26,05	0,07	0,04	47,90	0,08	0,02	23,71	0,2039	0,4311
		inne	0,01	0,03	330,47	0,00	0,03	983,40	0,01	0,02	353,51	0,2788	0,3598
	Tytuł prawny do zajmowanego mieszkania	własność	0,53	0,02	3,53	0,60	0,03	4,18	0,57	0,01	2,64	0,1964	0,4059
		najem wg cen rynkowych	0,02	0,03	122,86	0,03	0,05	142,24	0,03	0,03	90,32	0,1184	0,4482
		najem poniżej cen rynko-	0,01	0,03	201,28	0,02	0,05	203,75	0,02	0,03	137,77	0,0846	0,4498

Województwo	Zmienna	Wariant	BBGD			EU-SILC			Zintegrowany			$1-\frac{s_{int}(p)}{s_{BBGD}(p)}$	$1-\frac{s_{int}(p)}{s_{EU-SILC}(p)}$
			p	$S(p)$	CV	p	$S(p)$	CV	p	$S(p)$	CV		
		wych											
		inne	0,43	0,02	4,96	0,34	0,03	10,00	0,39	0,02	4,69	0,1590	0,4701
	Typ biologiczny gospodarstwa domowego	małżeństwo bez dzieci	0,17	0,02	13,90	0,25	0,03	13,51	0,21	0,02	9,46	0,1576	0,4043
		małżeństwo z 1 dzieckiem na utrzymaniu	0,10	0,02	24,57	0,11	0,04	35,20	0,11	0,02	19,90	0,1452	0,4604
		małżeństwo z 2 dzieci na utrzymaniu	0,13	0,02	18,90	0,09	0,04	42,89	0,11	0,02	18,50	0,2015	0,4576
		małżeństwo z 3 więcej dzieci na utrzymaniu	0,07	0,02	35,68	0,05	0,04	73,03	0,06	0,02	33,52	0,1932	0,4605
		samotny rodzic z co najmniej jedną osobą na utrzymaniu	0,06	0,03	44,90	0,03	0,04	122,90	0,04	0,02	46,53	0,2256	0,4539
		gospodarstwa jednoosobowe	0,24	0,03	12,64	0,26	0,04	16,03	0,25	0,02	9,84	0,1876	0,4084
		inne gospodarstwa z osobami na utrzymaniu	0,09	0,02	26,64	0,07	0,04	56,36	0,08	0,02	25,30	0,1895	0,4684
		pozostałe gospodarstwa	0,13	0,02	18,47	0,14	0,04	24,82	0,14	0,02	14,72	0,1748	0,4252
	Liczba pokoi	1	0,11	0,03	25,89	0,12	0,04	34,65	0,11	0,02	20,56	0,1696	0,4293
		2	0,39	0,02	5,60	0,35	0,03	9,65	0,37	0,02	4,96	0,1659	0,4563
		3	0,36	0,02	6,03	0,36	0,03	8,79	0,36	0,02	4,98	0,1768	0,4326
		4	0,08	0,03	31,56	0,11	0,04	37,19	0,10	0,02	23,51	0,1433	0,4347
		5	0,04	0,02	69,20	0,05	0,04	75,10	0,04	0,02	49,67	0,1396	0,4250
		6 i więcej	0,02	0,03	146,48	0,02	0,04	282,03	0,02	0,02	132,45	0,1533	0,4995
	Czy jest ustęp spłukiwany ?	tak	0,91	0,01	0,89	0,96	0,01	0,92	0,94	0,01	0,63	0,2777	0,3290
		nie	0,09	0,03	30,82	0,04	0,04	79,07	0,06	0,02	31,51	0,2309	0,4313
	Czy jest łazienka ?	tak	0,92	0,01	0,87	0,94	0,01	1,05	0,93	0,01	0,66	0,2369	0,3846
		nie	0,08	0,03	32,40	0,06	0,04	63,90	0,07	0,02	30,29	0,2124	0,4292
	Czy gospo-	tak	0,99	0,00	0,27	0,99	0,00	0,30	0,99	0,00	0,20	0,2716	0,3453

Województwo	Zmienna	Wariant	BBGD			EU-SILC			Zintegrowany			$1-\frac{s_{int}(p)}{s_{BBGD}(p)}$	$1-\frac{s_{int}(p)}{s_{EU-SILC}(p)}$
			p	$S(p)$	CV	p	$S(p)$	CV	p	$S(p)$	CV		
	darstwo posiada TV?	nie	0,01	0,03	341,43	0,01	0,04	681,61	0,01	0,02	323,48	0,2677	0,3529
	Czy gospodarstwo posiada komputer?	tak	0,35	0,02	6,18	0,45	0,03	6,98	0,40	0,02	4,49	0,1672	0,4228
		nie	0,65	0,02	2,59	0,55	0,03	4,89	0,60	0,01	2,42	0,1393	0,4620
	Czy gospodarstwo posiada pralkę?	tak	0,98	0,00	0,44	0,98	0,01	0,61	0,98	0,00	0,35	0,1945	0,4239
		nie	0,02	0,03	133,48	0,02	0,05	213,95	0,02	0,03	114,63	0,1874	0,4351
	Czy gospodarstwo posiada samochód?	tak	0,45	0,02	4,40	0,45	0,03	6,45	0,45	0,02	3,63	0,1703	0,4413
		nie	0,55	0,02	3,52	0,55	0,03	5,22	0,55	0,02	2,92	0,1733	0,4369
wielkopolskie	Rodzaj budynku	budynek wielorodzinny	0,56	0,01	2,29	0,51	0,02	4,06	0,53	0,01	2,04	0,1461	0,4746
		dom jednorodzinny	0,36	0,01	3,93	0,41	0,02	5,12	0,39	0,01	3,07	0,1652	0,4383
		dom jednorodzinny w zabudowie szeregowej	0,08	0,02	22,21	0,07	0,03	36,95	0,08	0,01	19,15	0,1584	0,4690
		inne	0,00	0,02	545,43	0,00	0,03	1153,81	0,00	0,01	512,57	0,1811	0,4807
	Tytuł prawny do zajmowanego mieszkania	własność	0,54	0,01	2,24	0,58	0,02	3,13	0,56	0,01	1,80	0,1720	0,4410
		najem wg cen rynkowych	0,05	0,02	39,29	0,06	0,03	48,93	0,05	0,02	30,00	0,1352	0,4498
		najem poniżej cen rynkowych	0,01	0,02	165,09	0,01	0,03	235,11	0,01	0,02	134,88	0,1752	0,4315
		inne	0,40	0,01	3,78	0,35	0,02	6,92	0,37	0,01	3,40	0,1524	0,4779
	Typ biologiczny gospodarstwa domowego	małżeństwo bez dzieci	0,15	0,02	10,65	0,20	0,02	12,45	0,17	0,01	7,93	0,1541	0,4298
		małżeństwo z 1 dzieckiem na utrzymaniu	0,11	0,02	14,93	0,11	0,03	26,61	0,11	0,01	13,16	0,1503	0,4870
		małżeństwo z 2 dzieci na utrzymaniu	0,12	0,02	13,55	0,11	0,03	23,56	0,12	0,01	11,87	0,1574	0,4756
		małżeństwo z 3 więcej dzie-	0,06	0,02	28,47	0,04	0,03	73,68	0,05	0,01	28,63	0,1976	0,4833

Województwo	Zmienna	Wariant	BBGD			EU-SILC			Zintegrowany			$1-\frac{s_{int}(p)}{s_{BBGD}(p)}$	$1-\frac{s_{int}(p)}{s_{EU-SILC}(p)}$
			p	$S(p)$	CV	p	$S(p)$	CV	p	$S(p)$	CV		
		ci na utrzymaniu											
		samotny rodzic z co najmniej jedną osobą na utrzymaniu	0,05	0,02	31,81	0,04	0,03	72,59	0,05	0,01	30,84	0,1968	0,4669
		gospodarstwa jednoosobowe	0,20	0,02	10,84	0,19	0,03	14,98	0,20	0,02	8,79	0,1925	0,4110
		inne gospodarstwa z osobami na utrzymaniu	0,13	0,02	12,49	0,12	0,03	22,66	0,12	0,01	11,08	0,1504	0,4892
		pozostałe gospodarstwa	0,17	0,02	9,57	0,20	0,02	12,35	0,18	0,01	7,43	0,1523	0,4435
	Liczba pokoi	1	0,13	0,02	14,41	0,11	0,03	27,51	0,12	0,02	13,12	0,1697	0,4737
		2	0,33	0,02	4,65	0,32	0,02	7,60	0,32	0,01	3,99	0,1589	0,4634
		3	0,31	0,02	4,83	0,30	0,02	7,80	0,31	0,01	4,13	0,1598	0,4613
		4	0,11	0,02	14,99	0,13	0,03	20,17	0,12	0,01	11,89	0,1567	0,4430
		5	0,07	0,02	24,70	0,09	0,03	30,19	0,08	0,01	18,71	0,1368	0,4467
		6 i więcej	0,04	0,02	40,31	0,06	0,03	47,41	0,05	0,01	30,01	0,1321	0,4444
	Czy jest ustęp spłukiwany ?	tak	0,93	0,01	0,55	0,93	0,01	0,83	0,93	0,00	0,45	0,1678	0,4515
		nie	0,07	0,02	23,95	0,07	0,03	38,43	0,07	0,01	20,41	0,1628	0,4594
	Czy jest łazienka ?	tak	0,92	0,01	0,58	0,92	0,01	0,90	0,92	0,00	0,49	0,1566	0,4593
		nie	0,08	0,02	22,05	0,08	0,03	33,17	0,08	0,02	18,32	0,1610	0,4526
	Czy gospodarstwo posiada TV?	tak	0,99	0,00	0,20	0,98	0,00	0,38	0,99	0,00	0,19	0,0654	0,5023
		nie	0,01	0,02	176,05	0,02	0,03	167,90	0,01	0,02	119,64	0,1571	0,4011
	Czy gospodarstwo posiada komputer?	tak	0,40	0,01	3,51	0,45	0,02	4,75	0,43	0,01	2,78	0,1622	0,4465
		nie	0,60	0,01	1,99	0,55	0,02	3,48	0,57	0,01	1,77	0,1470	0,4696
	Czy gospodarstwo posiada pral-	tak	0,98	0,00	0,27	0,97	0,00	0,49	0,98	0,00	0,25	0,0909	0,4926
		nie	0,02	0,02	104,13	0,03	0,03	114,18	0,02	0,02	75,50	0,1484	0,4228

Województwo	Zmienna	Wariant	BBGD			EU-SILC			Zintegrowany			$1-\frac{s_{int}(p)}{s_{BBGD}(p)}$	$1-\frac{s_{int}(p)}{s_{EU-SILC}(p)}$
			p	$S(p)$	CV	p	$S(p)$	CV	p	$S(p)$	CV		
	kę?												
	Czy gospodarstwo posiada samochód?	tak	0,56	0,01	2,13	0,59	0,02	3,07	0,58	0,01	1,73	0,1653	0,4522
		nie	0,44	0,01	3,29	0,41	0,02	5,46	0,42	0,01	2,86	0,1634	0,4551
zachodniopomorskie	Rodzaj budynku	budynek wielorodzinny	0,73	0,01	1,84	0,77	0,02	2,53	0,75	0,01	1,46	0,1838	0,4358
		dom jednorodzinny	0,13	0,02	17,33	0,13	0,03	26,06	0,13	0,02	14,41	0,1623	0,4515
		dom jednorodzinny w zabudowie szeregowej	0,14	0,02	16,64	0,09	0,04	37,63	0,12	0,02	16,08	0,1825	0,4767
		inne	0,00	--	--	0,00	0,04	1395,29	0,00	0,03	1978,57	--	0,2939
	Tytuł prawny do zajmowanego mieszkania	własność	0,46	0,02	3,95	0,46	0,03	6,04	0,46	0,02	3,30	0,1610	0,4556
		najem wg cen rynkowych	0,02	0,03	103,22	0,03	0,04	124,57	0,03	0,02	78,09	0,1449	0,4384
		najem poniżej cen rynkowych	0,01	0,03	179,60	0,01	0,04	433,73	0,01	0,02	177,06	0,2087	0,4572
		inne	0,50	0,02	3,73	0,50	0,03	5,93	0,50	0,02	3,16	0,1554	0,4645
	Typ biologiczny gospodarstwa domowego	małżeństwo bez dzieci	0,18	0,02	12,46	0,20	0,03	15,98	0,19	0,02	9,68	0,1579	0,4382
		małżeństwo z 1 dzieckiem na utrzymaniu	0,13	0,02	17,68	0,14	0,04	27,21	0,13	0,02	14,68	0,1371	0,4799
		małżeństwo z 2 dzieci na utrzymaniu	0,11	0,02	20,22	0,13	0,04	28,85	0,12	0,02	16,24	0,1251	0,4800
		małżeństwo z 3 więcej dzieci na utrzymaniu	0,05	0,02	46,66	0,03	0,03	115,98	0,04	0,02	46,73	0,2249	0,4289
		samotny rodzic z co najmniej jedną osobą na utrzymaniu	0,07	0,02	32,15	0,04	0,04	97,02	0,05	0,02	33,77	0,2085	0,4806
		gospodarstwa jednoosobowe	0,26	0,03	10,55	0,26	0,04	14,60	0,26	0,02	8,55	0,1884	0,4155
		inne gospodarstwa z osobami na utrzymaniu	0,07	0,02	33,67	0,09	0,04	41,91	0,08	0,02	25,53	0,1145	0,4678

Województwo	Zmienna	Wariant	BBGD			EU-SILC			Zintegrowany			$1-\frac{s_{int}(p)}{s_{BBGD}(p)}$	$1-\frac{s_{int}(p)}{s_{EU-SILC}(p)}$
			p	$S(p)$	CV	p	$S(p)$	CV	p	$S(p)$	CV		
		pozostałe gospodarstwa	0,13	0,02	16,94	0,11	0,04	33,63	0,12	0,02	15,60	0,1673	0,4809
	Liczba pokoi	1	0,12	0,03	20,75	0,09	0,04	43,44	0,11	0,02	19,46	0,1711	0,4837
		2	0,43	0,02	4,59	0,42	0,03	7,19	0,43	0,02	3,87	0,1604	0,4573
		3	0,31	0,02	6,66	0,34	0,03	9,54	0,32	0,02	5,40	0,1563	0,4549
		4	0,09	0,02	25,53	0,10	0,04	36,99	0,09	0,02	20,86	0,1540	0,4549
		5	0,03	0,02	79,17	0,02	0,04	186,25	0,03	0,02	77,02	0,1893	0,4819
		6 i więcej	0,02	0,02	151,70	0,03	0,04	133,91	0,02	0,02	97,96	0,0969	0,4317
	Czy jest ustęp spłukiwany ?	tak	0,95	0,01	0,59	0,97	0,01	0,66	0,96	0,00	0,43	0,2561	0,3588
		nie	0,05	0,02	52,35	0,03	0,03	124,52	0,04	0,02	51,85	0,2316	0,4124
	Czy jest łazienka ?	tak	0,94	0,01	0,68	0,96	0,01	0,87	0,95	0,00	0,53	0,2219	0,4041
		nie	0,06	0,02	39,71	0,04	0,04	82,64	0,05	0,02	37,43	0,1985	0,4492
	Czy gospodarstwo posiada TV?	tak	0,99	0,00	0,26	0,98	0,01	0,58	0,98	0,00	0,27	-0,0238	0,5309
		nie	0,01	0,03	259,95	0,02	0,04	217,66	0,02	0,03	162,44	0,0713	0,4388
	Czy gospodarstwo posiada komputer?	tak	0,37	0,02	5,27	0,42	0,03	7,21	0,40	0,02	4,17	0,1535	0,4565
		nie	0,63	0,02	2,54	0,58	0,03	4,45	0,60	0,01	2,26	0,1448	0,4697
	Czy gospodarstwo posiada pralkę?	tak	0,97	0,00	0,43	0,98	0,01	0,57	0,98	0,00	0,34	0,2116	0,4158
		nie	0,03	0,03	98,12	0,02	0,04	220,92	0,02	0,02	93,34	0,1650	0,5086
	Czy gospodarstwo posiada samochód?	tak	0,40	0,02	4,81	0,46	0,03	6,54	0,43	0,02	3,79	0,1523	0,4588
		nie	0,60	0,02	2,75	0,54	0,03	4,87	0,57	0,01	2,46	0,1474	0,4664

Źródło: opracowanie własne

ZAŁĄCZNIK. KOD SPSS SYNTAX

DLA INTEGRACJI METODĄ NAJBLIŻSZEGO SĄSIADA

Uwaga:
Plik:

- dawca.sav – plik dawcy,
- biorca.sav – plik biorcy,
- baza.sav – połączone zharmonizowane pliki dawcy i biorcy (konkatenacja – dla zmiennych dołączanych występują braki danych).

Gwiazdką rozpoczęto linię polecenia będącą komentarzem do poleceń poniżej.
Kursywą zapisano część kodu, którą należy zmodyfikować w zależności od ścieżki dostępu do plików oraz nazw zmiennych parujących.

```
*definicja makropolecenia.
define !NND(!pos=!tokens(1)/!pos=!tokens(1)).
* wczytanie sharmonizowanego zbioru danych; S=1 – zbiór biorcy, S=0 zbiór dawcy.
get file="ścieżka dostępu\ baza.sav".
compute pre_1=1.
compute id=$casenum.
exe.
val lab S
0       "       Dawca "
1       "       Biorca ".

**generowanie liczb pseudolosowych – umożliwia dołączenie losowego rekordu w przypadku takiej samej wartości funkcji odległości.
compute h1=rv.uniform(0,100000000).
compute h2=rv.uniform(0,100000000).
compute h3=rv.uniform(0,100000000).
compute h4=rv.uniform(0,100000000).
compute h5=rv.uniform(0,100000000).
exe.

***sortowanie wg zmiennej rozdzielającej i zmiennej losowej zapewniające losowy dobór "bliźniaka" w przypadku większej liczby rekordów o tej samej minimalnej odległości.
set workspace=4097151.
sort cases by S(d) h4(a).
exe.

compute pom=0.
exe.
weight off.
sav out='ścieżka dostępu\baza.sav'.
* pętla po każdej warstwie.
!do !i=1 !to !1.
```

```
* zapisanie liczebności zbioru dawcy i biorcy.
get file="ścieżka dostępu\baza.sav".
select if !2=!i.
exe.
fre s.
if s=1 s1=1.
filter by s1.
AGGREGATE
 /OUTFILE=* MODE=ADDVARIABLES
 /n_don=N.
filter off.
if missing(n_don) n_don=0.
AGGREGATE
 /OUTFILE=* MODE=ADDVARIABLES
 /n_all=N.

sort cases by s1(d) s(a) h3(a).

* wejście w tryb macierzowy umożliwiające obliczenie odległości między każdą parą
rekordów.
SET MXLOOP=3000000.
MATRIX.
GET M /VARIABLES= S id pre_1 pom zmienna_wspólna_1 to zmienna_wspólna_n
n_don n_all.

compute n1=m(1,19).
compute i2=n1+1.
compute n2=m(1,20).
compute cc=99.

* petla licząca odległość (ddd) między wszystkimi parami rekordów i zapisująca odle-
głość minimalną
** Uwaga! Poniższą pętlę stosuję się w sytuacji, gdy dawcą jest zbiór więk-
szy, a biorcą mniejszy!.
loop #i=1 to n1.
+ compute ii=0.
+ compute dd=9999.
+  loop #j=i2 to n2.
+  compute ddd=0.
+   do if (m(#j,4) = 0).
+     loop #k=5 to 18.
+       compute ddd=ddd+(m(#i,#k)-m(#j,#k))*(m(#i,#k)-m(#j,#k)).
+     end loop.
+   do if (ddd < dd).
+     compute ii=#j.
+     compute dd=ddd.
+   end if.
+   end if.
+  end loop.
+  do if (dd < cc).
```

```
+   compute m(#i,4)=m(ii,2).
+   compute m(#i,3)=dd.
+   compute m(ii,3)=dd.
+  end if.
end loop.
```

* **Uwaga! Poniższą pętlę stosuję się w sytuacji, gdy dawcą jest zbiór mniejszy, a biorcą większy!.**

```
loop #j=i2 to n2.
+ compute ii=0.
+ compute dd=999.
+  loop #i=1 to n1.
+  compute ddd=0.
+   do if (m(#j,4) = 0).
+    loop #k=5 to 22.
+      compute ddd=ddd+(m(#i,#k)-m(#j,#k))*(m(#i,#k)-m(#j,#k)).
+    end loop.
+   do if (ddd < dd).
+    compute ii=#i.
+    compute dd=ddd.
+   end if.
+   end if.
+  end loop.
+  do if (dd < cc).
+   compute m(#j,4)=m(ii,2).
+   compute m(#j,3)=dd.
+   compute m(ii,3)=dd.
+  end if.
end loop.
```

*** Koniec pętli.**

```
save M/outfile=*.
END MATRIX.
compute s=col1.
compute warstwa=!i.
exe.

* Zapisanie pliku warstwy.
sav out=!concat("'ścieżka dostępu\ Warstwy\',!2,!i,'.sav'").

get file="ścieżka dostępu\ baza.sav".
!doend.

* Łączenie plików warstw w jeden.
get file=!concat("'ścieżka dostępu\Warstwy\',!2,'1.sav'").
!do !j=2 !to !1.
ADD FILES /FILE=* /FILE=!concat("'ścieżkadostępu\Warstwy\',!2,!j,'.sav'").
```

```
!doend.
sort cases by s(a) !2(a).
sav out=!concat('''ścieżka dostępu\Warstwy\results_all.sav''').

* Tworzenie plików tymczasowych.
get file=!concat('''ścieżka dostępu\Warstwy\results_all.sav''').
select if s=1.
exe.
compute key=col4.
compute ID=col2.
exe.
sort cases by id(a).
exe.
sav out=!concat('''ścieżka dostępu\Warstwy\temp_rec.sav''').

* Tworzenie pliku z informacją o liczbie dołączeń rekordów dawcy do zbioru biorcy.
get file=!concat('''ścieżka dostępu\Warstwy\temp_rec.sav''').
select if s=1.
exe.
AGGREGATE
 /OUTFILE=!concat('''ścieżka dostępu\Ewaluacja\n_match.sav''')
 /BREAK=key
 /N_BREAK=N.
get file="ścieżka dostępu\dawca.sav".
match files /file=* /keep=
ID.
sav out='ścieżka dostępu\Ewaluacja\ID_don.sav'.
get file=!concat('''ścieżka dostępu\Ewaluacja\n_match.sav''').
ren var key=ID.
sav out=!concat('''ścieżka dostępu\Ewaluacja\n_match.sav''').
get file="ścieżka dostępu\Ewaluacja\ID_don.sav".
match files /file=* /table=!concat('''ścieżka dostępu\Ewaluacja\n_match.sav''') /by id.
exe.
recode n_break (sysmis=0).
exe.
sav out=!concat('''ścieżka dostępu\Ewaluacja\n_match.sav''').


* Pozyskanie ID dawcy i dołączenie do biorcy.
get file="ścieżka dostępu\dawca.sav".
compute key=ID.
EXECUTE.
sort cases by key(a).
sav out="ścieżka dostępu\dawca.sav".
get file=!concat('''ścieżka dostępu\Warstwy\temp_rec.sav''').
sort cases by id(a).
match files /file=* /keep=ID key.
exe.
match files /file=* /table="ścieżka dostępu\biorca.sav" /by id.
exe.
```

```
sav out=!concat("'ścieżka dostępu\Ewaluacja\rec_common.sav'").
get file="ścieżka dostępu\dawca.sav".
compute key=ID.
exe.

sort cases by key(a).
sav out=!concat("'ścieżka dostępu\Ewaluacja\don_common.sav'").
get file=!concat("'ścieżka dostępu\Ewaluacja\rec_common.sav'").
add files file=* /file=!concat("'ścieżka dostępu\Ewaluacja\don_common.sav'").
exe.
sav out=!concat("'ścieżka dostępu\Ewaluacja\rec_don_eval.sav'").
get file="ścieżka dostępu\dawca.sav".
compute key=ID.
exe.
sort cases by key(a).
sav out=!concat("'ścieżka dostępu\Ewaluacja\temp_donor.sav'").
get file="ścieżka dostępu\biorca.sav".
sort cases by id(a).
MATCH FILES /FILE=*
 /FILE=!concat("'ścieżka dostępu\Warstwy\temp_rec.sav'")
 /DROP=col1 to col10
 /by id.
EXECUTE.
sav out=!concat("'ścieżka dostępu\Warstwy\rec_key.sav'").

*łączenie pliku biorcy i dawcy w jeden zbiór.
get file=!concat("'ścieżka dostępu\Warstwy\rec_key.sav'").
sort cases by key(A).
match files /file=* /table=!concat("'ścieżka dostępu\Ewaluacja\temp_donor.sav'") /by
key.
exe.
sav out=!concat("'ścieżka dostępu\Warstwy\INTEGRATED.sav'").
!doend.
!enddefine.
* wywołanie makropolecenia (!nnd – nazwa makropolecenia, 12 – liczba warstw, war-
stwa – nazwa zmiennej z numerem warstwy).
!nnd 12 warstwa.
```