Praca badawcza pt. „Pomiar ubóstwa na poziomie powiatów (LAU 1) – etap I”

Praca powstała w ramach Projektu ”Wsparcie systemu monitorowania polityki spójności w perspektywie finansowej 2007-2013 oraz programowania i monitorowania polityki spójności w perspektywie finansowej 2014–2020”

Projekt współfinansowany przez Unię Europejską ze środków Programu Operacyjnego Pomoc Techniczna 2007–2013

Nazwa jednostki opracowującej raport:

Centrum Badań i Edukacji Statystycznej GUS

Kierownik projektu:

Marcin Szymkowiak

Opracował zespół badawczy:

Maciej Beręsewicz, Anna Bieńkuńska, Piotr Jastrzębski,
Tomasz Józefowski, Tomasz Klimanek, Jacek Kowalewski, Mariusz Kraj,
Piotr Łysoń, Anna Małasiewicz, Andrzej Młodak, Tomasz Panek,
Tomasz Piasecki, Michał Pietrzak, Łukasz Wawrowski

Spis treści

Wprowadzenie

Poznanie poziomu i skali ubóstwa w celu eliminowania jego przyczyn oraz przeciwdziałanie jego negatywnym skutkom jest istotnym wyzwaniem polityki społecznej w każdym kraju. Dotyczy to również Polski, dla której problem ubóstwa jest szczególnie ważny z punktu widzenia prawidłowego realizowania polityki spójności, której jednym z głównych celów jest wspieranie włączania społecznego i walka z ubóstwem. Niezwykle ważne jest ponadto zagadnienie monitorowania tego zjawiska, co jest podkreślane przez wiele instytucji międzynarodowych, w tym Komisję Europejską, która m.in. rok 2010 ogłosiła Europejskim Rokiem Walki z Ubóstwem i Wykluczeniem Społecznym. W konsekwencji spowodowało to uruchomienie wielu inicjatyw związanych z analizą tego zjawiska, jego pomiarem oraz metod ograniczających jego ujemne skutki. Kompleksowe zbadanie ubóstwa, jego terytorialnego zasięgu i rozmieszczenia jest zatem niezwykle ważne z punktu widzenia prowadzenia właściwej polityki społecznej i podejmowania odpowiednich decyzji przez różne instytucje. Wymaga to jednak posiadania informacji na możliwie najniższym poziomie agregacji przestrzennej.

W Polsce podjęto już pierwsze tego typu działania o charakterze naukowo-badawczym. Przykładowo, w wyniku współpracy podjętej przez Departament Badań Społecznych i Warunków Życia Głównego Urzędu Statystycznego, Ośrodek Statystyki Małych Obszarów Urzędu Statystycznego w Poznaniu oraz Bank Światowy, stworzono mapę ubóstwa dla wszystkich podregionów w Polsce dla 2011 roku. Był to istotny krok naprzód, gdyż Główny Urząd Statystyczny publikuje jedynie dane na temat przestrzennego zróżnicowania stopy ubóstwa, z wykorzystaniem danych pochodzących z badania EU-SILC, na poziomie całego kraju, regionów oraz w układzie wojewódzkim. Uzyskanie oszacowań stopy ubóstwa na poziomie podregionów, ze względu

na małe liczebności próby w odpowiednich przekrojach w EU-SILC, możliwe było dzięki zastosowaniu metod jakie oferuje statystyka małych obszarów (model Faya-Herriota na poziomie obszaru). Skorzystanie z estymacji bezpośredniej (estymator Horvitza-Thompsona) obarczone byłoby zbyt dużym błędem i charakteryzowałoby się niską precyzją oszacowań.

Na całym świecie obserwowany jest od kilkudziesięciu lat intensywny rozwój metodologii statystyki małych obszarów. Stanowi ona remedium na pojawiające się z jednej strony rosnące zapotrzebowanie informacyjne na niskich poziomach agregacji przestrzennej, a z drugiej strony, umożliwia redukcję kosztów badań i obciążeń respondentów. Dzięki swoim własnościom estymatory wykorzystywane w statystyce małych obszarów umożliwiają uzyskiwanie wiarygodnych szacunków na niskich poziomach agregacji oraz dla bardziej szczegółowych domen, dla których klasyczne metody estymacji charakteryzują się zbyt dużą wariancją estymatorów.

Wraz z intensywnym rozwojem teorii w zakresie statystyki małych obszarów nastąpił rozwój jej zastosowań. Wystarczy wspomnieć o takich obszarach jak rynek pracy, rolnictwo, demografia czy statystyka przedsiębiorstw, w których istnieje szereg praktycznych rozwiązań i aplikacji. Dotyczy to w sposób szczególny również ubóstwa, gdzie różnego rodzaju wskaźniki (na przykład stopa ubóstwa) możliwe są do oszacowania na niskich szczeblach podziału terytorialnego jedynie z wykorzystaniem metodologii statystyki małych obszarów. Praktyczne zastosowania, akceptowalne przez odbiorców informacji statystycznej, wymagają jednak posiadania odpowiedniej jakości zmiennych pomocniczych. Zmienne takie mogą pochodzić z wielu źródeł statystycznych takich jak spisy, rejestry czy badania reprezentacyjne. „Pożyczanie mocy", które stanowi fundament statystyki małych obszarów w produkcji rzetelnych, wiarygodnych i akceptowalnych społecznie szacunków na poziomie lokalnym, stanowi zatem wyzwanie dla współczesnej statystyki publicznej.

Głównym celem raportu jest ocena metodologii statystyki małych obszarów w kontekście jej wykorzystania do zagadnienia pomiaru ubóstwa w Polsce na niskich poziomach agregacji przestrzennej. Docelowo ma to umożliwić estymację stopy ubóstwa na poziomie lokalnym (LAU 1) z akceptowalnym błędem szacunku. Realizacja głównego celu w szczególności obejmować będzie:

- ocenę podejść (estymatorów) stosowanych w zakresie statystyki małych obszarów i prezentowanych od strony teoretycznej w literaturze przedmiotu, które mogą być przydatne w estymacji stopy ubóstwa,
- przegląd praktycznych zastosowań i implementacji w różnych krajach

w zakresie oszacowania wybranych charakterystyk opisujących ubóstwo z wykorzystaniem statystyki małych obszarów,

- kwerendę potencjalnych źródeł danych w kontekście poszukiwania zmiennych pomocniczych do budowy odpowiednich modeli statystycznych, które będzie można wykorzystać na potrzeby estymacji stopy ubóstwa na poziomie LAU 1,

- weryfikację dostępnego oprogramowania, które może być wykorzystane na potrzeby szacowania stopy ubóstwa w Polsce na poziomie LAU 1.

Raport składa się z 7 zasadniczych rozdziałów. W rozdziale pierwszym opisano najważniejsze praktyczne zastosowania statystyki małych obszarów w kontekście ubóstwa na całym świecie. Wskazano na doświadczenia wybranych państw, prace prowadzone przez Bank Światowy oraz opisano w tym zakresie dorobek Polski.

Rozdział drugi zawiera opis aktualnego stanu wiedzy z zakresu wykorzystania metodologii statystyki małych obszarów, przede wszystkim w kontekście zagadnienia estymacji wybranych charakterystyk opisujących ubóstwo. Dokonano w nim syntetycznego streszczenia najważniejszych pozycji literaturowych, które mogą być przydatne w pogłębionych studiach z zakresu statystyki małych obszarów i jej zastosowań w obszarze ubóstwa.

Rozdział trzeci stanowi kwerendę najważniejszych źródeł danych statystycznych, które mogą być przydatne w konstrukcji odpowiednich modeli służących estymacji stopy ubóstwa na niskich poziomach agregacji przestrzennej. Wskazano w nim na rolę spisów, rejestrów administracyjnych i najważniejszych badań reprezentacyjnych jako potencjalnych źródeł zmiennych pomocniczych.

W rozdziale czwartym opisano najważniejsze zmienne pomocnicze, które mogą być przydatne w zagadnieniu szacowania stopy ubóstwa na niskich poziomach agregacji przestrzennej z wykorzystaniem statystyki małych obszarów. Wskazano przy tym na cztery różne poziomy poszukiwania potencjalnych zmiennych pomocniczych obejmujących poziom regionu, społeczności, gospodarstwa domowego i osoby.

Rozdział piąty poświęcony został syntetycznemu opisowi najważniejszych estymatorów statystyki małych obszarów, które mogą być przydatne do szacowania wybranych charakterystyk opisujących ubóstwo, w tym i stopę ubóstwa.

W rozdziale szóstym opisano potencjalne korzyści jakie można uzyskać z zastosowania wybranych metod wielowymiarowej analizy danych. Ze względu

na to, że ubóstwo można traktować jako zjawisko wielowymiarowe, które determinuje wiele różnych czynników, metody te mogą okazać się szczególnie przydatne i użyteczne.

Ostatni, siódmy rozdział raportu, poświęcony został opisowi najważniejszych narzędzi informatycznych i oprogramowaniu, które może być przydatne w zagadnieniu estymacji i mapowania ubóstwa na niskich poziomach agregacji przestrzennej. Jest to szczególnie istotne, zważywszy na fakt, że metody jakie oferuje statystyka małych obszarów wymagają zastosowania skomplikowanych od strony matematycznej i czasochłonnych algorytmów.

1

Zastosowanie metodologii statystyki małych obszarów do estymacji charakterystyk ubóstwa – przykłady zastosowań

Rozdział zawiera przegląd zastosowań technik statystyki małych obszarów do estymacji wybranych charakterystyk ubóstwa w krajach europejskich, Stanach Zjednoczonych oraz projektów Banku Światowego, który podejmował szereg inicjatyw współpracy z krajowymi urzędami i instytucjami statystycznymi.

Generalnie można wyodrębnić cztery grupy metod statystyki małych obszarów stosowanych w mapowaniu ubóstwa, są to:

- model Faya-Herriota (projekty prowadzone we współpracy z Bankiem Światowym),
- model ELL (projekty prowadzone we współpracy z Bankiem Światowym),
- hierarchiczne podejście bayesowskie (Hiszpania, USA),
- podejście typu M-quantile (Włochy).

1.1 Polska

Na podstawie listu intencyjnego podpisanego 26.06.2013 roku została podjęta współpraca Departamentu Badań Społecznych i Warunków Życia w Głównym Urzędzie Statystycznym i Ośrodka Statystyki Małych Obszarów w Urzędzie Statystycznym w Poznaniu z Bankiem Światowym, której celem było wykorzystanie technik z zakresu statystyki małych obszarów do stworzenia map ubóstwa na poziomie podregionów (NUTS 3). Powstałe w wyniku tej

współpracy opracowanie pt. **Mapy ubóstwa na poziomie podregionów w Polsce z wykorzystaniem estymacji pośredniej**, chociaż ma status pracy eksperymentalnej i nie stanowiło oficjalnych wyników GUS, pokazuje możliwości polskiej statystyki publicznej w tym obszarze badań.

Efektem współpracy było wykazanie, że uzyskanie szacunków dla podziału terytorialnego niższego niż regiony (NUTS 1) czy województwa (NUTS 2) jest możliwe poprzez zastosowanie metod estymacji pośredniej. Wykorzystują one informacje spoza badanej domeny, co zwykle przyczynia się do zwiększenia precyzji szacunków. Techniki te, między innymi ze względu na odmienny proces estymacji, który jest oparty na przyjętym modelu, stanowią wyzwanie dla statystyki publicznej w wielu krajach. Dlatego też w przypadku współpracy GUS, US Poznań i Banku Światowego podjęto próby oszacowania stopy ubóstwa na niższym poziomie agregacji przestrzennej niż prezentowane dotychczas, tj. dla wszystkich podregionów w Polsce. Wykorzystano w tym celu podejście wypracowane przez Bank Światowy we współpracy z wieloma krajowymi urzędami statystycznymi do szacowania poziomu ubóstwa w przekroju podregionów lub niższym.

W pracach nad zagadnieniem ubóstwa wykorzystano podejście modelowe. Ze względu na formę dostępnych danych, dobre własności empiryczne oraz prostotę wybrano model Faya–Herriota. Wybór zmiennych do modelu opierał się przede wszystkim na przesłankach merytorycznych. W tym celu posługiwano się zależnością regresyjną stopy ubóstwa od wybranych zmiennych objaśniających. Decydując, czy dana zmienna znajdzie się w modelu, kierowano się jego poprawnością merytoryczną. Po uwzględnieniu zmiennej w modelu przeprowadzano kompleksową analizę, mającą na celu stwierdzenie, czy istotność i znak współczynnika, stojącego przy danej zmiennej, znajduje swoje odzwierciedlenie w rzeczywistości. Do budowy modeli wykorzystano dane pochodzące z kilku źródeł statystycznych. Z badania EU-SILC wykorzystano jedynie stopę ubóstwa jako zmienną objaśnianą. Użycie innych zmiennych z tego badania w charakterze zmiennych objaśniających doprowadziłoby do zwiększenia błędu losowego i obciążonych oszacowań parametrów stojących przy tych zmiennych w modelu regresji. W związku z tym, w charakterze zmiennych objaśniających, rozważano dane pochodzące z Narodowego Spisu Powszechnego Ludności i Mieszkań 2011 (NSP 2011), Narodowego Spisu Powszechnego Ludności i Mieszkań 2002 (NSP 2002) oraz dane pochodzące z Banku Danych Lokalnych (BDL) z lat 2005–2011. Łącznie analizowano i brano pod uwagę ponad 200 potencjalnych zmiennych objaśniających, charakteryzujących różne aspekty: potencjał demograficzny, charakter miejski/wiejski, aktywność ekonomiczną, infrastrukturę mieszkaniową, wybrane

charakterystyki gospodarstw domowych, budżety jednostek terytorialnych, infrastrukturę drogową, ochronę środowiska, ochronę zdrowia i opiekę społeczną, migracje wewnętrzne.

W ostatecznym modelu uwzględniono 6 zmiennych objaśniających:

- odsetek liczby osób samotnych do 25 roku życia,
- liczba pokoi przypadająca na członka gospodarstwa domowego,
- odsetek gospodarstw domowych posiadających łazienkę,
- odsetek gospodarstw domowych z dwiema osobami powyżej 25 roku życia z wykształceniem co najwyżej zawodowym,
- gęstość zaludnienia,
- stosunek liczby osób wymeldowanych do liczby zameldowanych na pobyt stały w podregionie.

Otrzymane wyniki wskazały na silne przestrzenne zróżnicowanie stopy ubóstwa w Polsce w podregionach. Wyraźnie zaznaczył się podział kraju na grupy podregionów Polski centralnej, wschodniej i zachodniej. Ponadto charakterystyczne było to, że w podregionach okalających duże miasta występował znacznie niższy poziom ubóstwa niż w pozostałych podregionach województwa.

Podjęta próba zastosowania metodologii statystyki małych obszarów do mapowania ubóstwa była ogromnym wyzwaniem naukowym dla polskiej statystyki publicznej, ale na podkreślenie zasługuje ogromnie przychylne przyjęcie uzyskanych wyników i zachęty do podjęcia dalszych prac, których efektem mają być szacunki ubóstwa na niższych poziomach agregacji przestrzennej.

1.2 Hiszpania

Molina i in. [64] opisują próbę wykorzystania hierarchicznego podejścia bayesowskiego w statystyce małych obszarów do szacowania wskaźników ubóstwa na podstawie danych z hiszpańskiego badania EU-SILC z 2006 roku. Chociaż badanie EU-SILC jest zaprojektowane tak, aby zapewnić wiarygodne oszacowania dla całego kraju i dużych regionów kraju (Autonomous Communities), to jednak pojawiły się potrzeby informacyjne związane z dostarczeniem danych statystycznych dla niższych podziałów terytorialnych kraju. Zaproponowana metodologia hierarchicznego podejścia bayesowskiego umożliwiła uzyskanie wiarygodnych oszacowań zdefiniowanych w literaturze przedmiotu

wskaźników ubóstwa dla 52 prowincji hiszpańskich według płci, które to przekroje stanowiły tzw. małe obszary. Zmiennymi pomocniczymi były: wskaźnik pięciu pięcioletnich grup wieku osób narodowości hiszpańskiej, trzy poziomy charakterystyk wykształcenia oraz trzy poziomy statusu na rynku pracy (bezrobotny, pracujący, bierny zawodowo). Każda zmienna objaśniająca o liczbie kategorii n została przekształcona na $n-1$ zmiennych zero-jedynkowych. W przypadku zmiennych ciągłych hierarchiczne podejście bayesowskie wymaga posiadania pełnej informacji spisowej. Jednakże w tym badaniu wykorzystywano jedynie zmienne zero-jedynkowe, zatem potrzebne były jedynie liczebności osób posiadających określone kategorie zmiennych objaśniających. Kolejnym krokiem było wykorzystanie danych z badania aktywności ekonomicznej ludności (LFS – Spanish Labour Force Survey) do utworzenia przybliżonej postaci pełnej macierzy spisowej zmiennych pomocniczych. Dokonano tego poprzez multiplikację określonej obserwacji w badaniu aktywności ekonomicznej ludności tyle razy, ile wynosiła określona dla tej obserwacji waga z losowania. Wykorzystanie do tej procedury danych jednostkowych z badania LFS wynikało z faktu, że wielkość próby w badaniu LFS jest dużo większa niż wielkość próby w badaniu EU-SILC (155 333 wobec 34 389), co ma bezpośrednie przełożenie na jakość danych. Zmienną, która charakteryzuje sytuację dochodową każdej jednostki (osoby) w badaniu EU-SILC i jest wykorzystywana zarówno przez hiszpański urząd statystyczny (INE) jak i EUROSTAT do pomiaru ubóstwa, jest tzw. roczny dochód ekwiwalentny netto. Definiuje się go jako roczny dochód netto gospodarstwa domowego podzielony przez wskaźnik wielkości gospodarstwa domowego zgodnie z określoną skalą ekwiwalentności OECD. W ten sposób wielkość ta może być interpretowana jako dochód per capita i określać dochód pojedynczego członka gospodarstwa domowego.

Uzyskane wyniki wykazały, że zarówno w odniesieniu do skali, jak i intensywności największym ubóstwem charakteryzują się południowe i zachodnie prowincje Hiszpanii. Co więcej, wykorzystanie modelu jednostkowego, czyli przypisanie dochodu poszczególnym członkom gospodarstwa domowego, a nie całemu gospodarstwu domowemu, umożliwiło uwzględnienie kategorii płci w definiowaniu małego obszaru (domeny studiów). Okazało się, że sytuacja kobiet bez względu na lokalizację przestrzenną prowincji Hiszpanii jest gorsza w porównaniu z sytuacją mężczyzn.

Należy podkreślić, że w odróżnieniu od przypadków gdzie estymuje się parametry takie jak średnie lub wartości globalne dla małego obszaru, zaproponowana metoda dla oszacowania parametrów nieliniowych, wymaga posiadania pełnej informacji w zakresie zmiennych pomocniczych, a nie jedynie

wiedzy na temat wartości globalnych dla określonych domen.

1.3 Włochy

Pratesi i in. [29] przedstawili zastosowanie podejścia M-quantile w szacowaniu charakterystyk ubóstwa w ujęciu małych obszarów dla 29 włoskich prowincji zlokalizowanych w trzech regionach: Toskanii, Lombardii i Campanii. Zdaniem Autorów informacje dotyczące ubóstwa, nierówności oraz wskaźników jakości życia są obecnie jednymi z najbardziej oczekiwanymi danymi w skali Unii Europejskiej. Zatem oszacowanie przeciętnego dochodu ekwiwalentnego i odpowiednich kwantyli rozkładu powinno towarzyszyć oszacowaniu wskaźników ubóstwa, takich jak:

- stopa ubóstwa,
- indeks luki zagrożenia deprywacją. materialną

Autorzy zwracają uwagę na fakt, że chociaż średnie dla małych obszarów są często docelowymi oszacowaniami parametrów w innych aplikacjach, to poleganie jedynie na nich, może nie w pełni charakteryzować obraz poziomu rozwoju małego obszaru. Celem wyboru wspomnianych wyżej trzech regionów Włoch była charakterystyka podziału północ-południe, gdzie każdy z regionów reprezentował jedną z części kraju (północ, część centralną i południe). Dane pochodziły z włoskiego badania EU-SILC dla 2007 roku, a zmienną szacowaną był dochód ekwiwalentny gospodarstw domowych. Zmienne pomocnicze pochodziły z badania EU-SILC oraz spisu powszechnego i opisywały status własności, wiek głowy gospodarstwa domowego, liczbę lat edukacji głowy gospodarstwa domowego oraz wielkość gospodarstwa domowego.

Należy zwrócić uwagę, że do oszacowania błędu średniokwadratowego oszacowań typu M-quantile stosowane jest podejście bootstrap typu nieparametrycznego wprowadzone i opisane przez Tzavidisa [100].

1.4 USA

Od początku lat 90-tych XX wieku w USA dał się zaobserwować gwałtowny wzrost potrzeb informacyjnych w zakresie śledzenia zmian sytuacji ekonomicznej gospodarstw domowych w przekroju małych obszarów geograficznych. Stało się jasne, że informacje pozyskiwane co 10 lat w spisach powszechnych w ramach tzw. długiej formy (rodzaj kwestionariusza spisowego o szerszym spektrum pytań spisowych) są niewystarczające i nie umożliwiają

detekcji poważnych zmian sytuacji dochodowej społeczeństwa zachodzących w okresach międzyspisowych w kraju ani nie dają odpowiedzi na pytania dotyczące rozkładu dochodów czy ubóstwa dla małych społeczności (hrabstwa, miasta i inne obszary wyodrębniane w ramach poszczególnych stanów).

W 1993 roku udało się zbudować koalicję pięciu agencji federalnych (departamenty: rolnictwa, edukacji, zdrowia i usług dla ludności, mieszkalnictwa i rozwoju miast, rynku pracy), które, przy kluczowym współudziale Wydziału Statystyki Dochodów urzędu podatkowego USA (Statistics of Income Division of the Internal Revenue Service), przeznaczyły na rzecz amerykańskiego Biura Spisowego (U.S. Census Bureau) pewne fundusze umożliwiające rozpoczęcie badań na rzecz oszacowania charakterystyk ubóstwa i sytuacji dochodowej na podstawie najbardziej aktualnego spisu. Niestety w 1993 roku, pomimo przygotowania odpowiednich aktów prawnych wprowadzających metodologię produkcji danych pospisowych w zakresie ubóstwa w przekroju stanów, hrabstw i innych jednostek lokalnych oraz pozytywnego głosowania w Izbie Reprezentantów, amerykański Senat nie podjął debaty nad proponowanymi rozwiązaniami.

We wrześniu 1994 roku Kongres przegłosował akty prawne, które wskazywały, że rozdział funduszy federalnych pomiędzy okręgami szkolnymi (school district) powinien być oparty na najbardziej aktualnych, zadawalających danych udostępnionych przez Departament Handlu. Akty te zobowiązywały także sekretarza ds. edukacji (odpowiednik ministra ds. edukacji) do weryfikacji danych dotyczących ludności w okręgach szkolnych publikowanych przez Departament Handlu w zakresie ich aktualności i wiarygodności. Ponadto sekretarz ds. edukacji miał sfinansować odpowiednie ciało doradcze zwane panelem Narodowej Akademii Nauk, które miało opiniować oszacowania ubóstwa dla małych obszarów publikowane przez Biuro Spisowe, które miały być podstawą alokacji funduszy. Pierwsze raporty panelu zostały opublikowane w latach 1997–1999 [76], [77], [78].

We wrześniu 2000 roku panel Narodowej Akademii Nauk opublikował raport **Small Area Income and Poverty Estimates, Priorities for 2000 and Beyond**, którego celem była identyfikacja i sprawdzenie obszarów, dla których konieczne są pogłębione badania i dalsze doskonalenie modeli stosowanych w programie SAIPE do otrzymywania szacunków dla małych obszarów. W ramach programu SAIPE amerykańskie Biur Spisowe dostarcza corocznie następujących danych z zakresu dochodu i ubóstwa w przekroju okręgów szkolnych, hrabstw i stanów USA:

- liczba ludzi żyjących w ubóstwie,

- liczba dzieci do 5 lat żyjących w ubóstwie (tylko w przekroju stanów USA),
- liczba dzieci w wieku 5–17 lat w rodzinach żyjących w ubóstwie,
- liczba dzieci w wieku poniżej 18 lat żyjących w ubóstwie,
- mediana dochodu gospodarstw domowych.

Ponadto dla poziomu okręgów szkolnych publikowane są szacunki następujących zmiennych:

- ogólna liczba ludności,
- liczba dzieci w wieku 5-17 lat,
- liczba dzieci w wieku 5-17 lat w rodzinach żyjących w ubóstwie.

Powyższe oszacowania nie pochodzą ze statystyk wyliczanych bezpośrednio na podstawie rejestrów administracyjnych, ani nie są wartościami estymatora bezpośredniego na podstawie badania reprezentacyjnego. Dla hrabstw i stanów dane dochodowe i dotyczące ubóstwa są modelowane w taki sposób, że łączy się dane badania reprezentacyjnego z szacunkami liczby ludności i danymi rejestrów administracyjnych. Dla poziomu okręgów szkolnych w celu oszacowania charakterystyk ubóstwa wykorzystuje się oszacowania uzyskane z modelu dla hrabstw, dane z federalnych rejestrów podatkowych oraz dane pochodzące z wieloletnich badań reprezentacyjnych.

Od 2005 roku w procesie pozyskiwania szacunków dla zadanych przekrojów stosuje się dane z badania reprezentacyjnego amerykańskiego społeczeństwa (ACS – American Community Survey); we wcześniejszych latach stosowano dane z corocznych badań społeczno-ekonomicznych stanowiących uzupełnienie bieżących badań ludnościowych (Annual Social and Economic Supplements of the Current Population Survey). Bezpośrednie szacunki dla pojedynczych lat w oparciu o badanie ACS są dostępne corocznie dla hrabstw i innych obszarów, gdzie wielkość populacji wynosi co najmniej 65 tys. osób. Oszacowania dla trzyletnich okresów są co roku dostępne dla obszarów o wielkości 20 tys. i więcej. Oszacowania dla pięcioletnich okresów są dostępne co roku dla wszystkich hrabstw, okręgów szkolnych i innych małych obszarów (np. obwodów spisowych).

1.5 Bank Światowy

Bank Światowy może pochwalić się największą chyba skalą współpracy z wieloma krajami ze wszystkich kontynentów w zakresie tworzenia map ubóstwa: Albania, Bangladesz, Bułgaria, Boliwia, Brazylia, Chiny, Ekwador, Filipiny, Gruzja, Gwatemala, Honduras, Indonezja, Kambodża, Kenia, Madagaskar, Malawi, Maroko, Mozambik, Nepal, Nikaragua, Panama, Papua i Nowa Gwinea, Paragwaj, Polska, Republika Dominikany, Rumunia, Południowa Afryka, Sri Lanka, Uganda, Wietnam, Zachodni Brzeg i Gaza.

Eksperci Banku Światowego zwracają uwagę, że mapy ubóstwa - przestrzenny opis rozkładu ubóstwa w danym kraju mają olbrzymie znaczenie dla decydentów, polityków i analityków zajmujących się analizami społeczno-ekonomicznymi. Niestety praktycznie wszystkie badania gospodarstw domowych charakteryzują się na tyle małymi próbami, że poziom agregacji przestrzennej potrzebny wspomnianym wyżej grupom odbiorców docelowych nie może być uzyskany przy założeniu odpowiedniej jakości szacunków. Z kolei często dane pochodzące ze spisów nie zawierają informacji potrzebnych do wyznaczenia wskaźników ubóstwa. Z tych powodów w strukturach Banku Światowego powstał specjalny zespół ds. rozwoju badań w zakresie ubóstwa i nierówności (Development Research Group, Poverty and Inequality – DECRG-PI), który rozwinął metodologię opartą na podejściu ELL [26] w celu wsparcia krajowych urzędów i agencji statystycznych w pracach dotyczących oszacowania wskaźników bogactwa/ubóstwa. Zespół oprócz technicznego wsparcia, wypracowania szeregu narzędzi analitycznych, organizuje także warsztaty dla pracowników instytucji statystycznych i badaczy w ramach prowadzonej współpracy w krajach, dla których opracowywane są mapy ubóstwa – najczęściej były to przypadki krajów rozwijających się Azji, Afryki i Ameryki Południowej.

Metodologia ELL wypracowana przez zespół Banku Światowego obejmuje imputację, do zbiorów spisu ludności, które najczęściej nie zawierają charakterystyk dochodowych, czy mierników konsumpcji, danych pochodzących z badań reprezentacyjnych budżetów gospodarstw domowych. Z kolei te drugie zbiory charakteryzują się zbyt małą wielkością próby, aby z odpowiednią precyzją dostarczyć oszacowań dla małych obszarów. Pierwszy krok polega na konstrukcji modelu regresji wydatków lub spożycia, gdzie zmiennymi objaśniającymi są takie przykładowe zmienne (dostępne zarówno w spisie jaki i w badaniu reprezentacyjnym) jak: wielkość gospodarstwa domowego, wykształcenie, charakterystyki mieszkania i infrastruktury z nim związanej, ale także ważne zmienne demograficzne. Na drugim etapie wykorzystywane

są oszacowania parametrów skonstruowanych modeli regresji (współczynniki regresji i związane z nimi standardowe błędy szacunku) w celu oszacowania wydatków lub spożycia dla każdego gospodarstwa domowego w spisie. Ostatnim krokiem jest wykorzystanie danych jednostkowych dla gospodarstw domowych w celu oszacowania wskaźników ubóstwa dla małych obszarów.

Z reguły oszacowania pośrednie charakteryzują się pewnym stopniem niepewności, dlatego metodologia opracowana przez Bank Światowy kładzie duży nacisk na zapewnienie swoistego benchmarkingu w stosunku do zmiennych, które są dostępne na poziomie małych obszarów, ale jednocześnie umożliwiają ocenę obrazu wyłaniającego się z oszacowań uzyskanych za pomocą statystyki małych obszarów. Najczęściej są to mocno skorelowane zmienne pochodzące ze spisu lub dostępnych rejestrów administracyjnych. Oprócz krzyżowego porównania „sensowności" uzyskanych szacunków, takie porównania mogą ujawnić występowanie nieoczywistych związków z ubóstwem takich zmiennych jak: klimat, stan inwentarza hodowlanego per capita, odległości do najbliższych ośrodków związanych ze świadczeniem usług zdrowotnych, edukacyjnych, kulturalnych etc.

2

Ubóstwo w świetle literatury z zakresu statystyki małych obszarów

W rozdziale tym opisany zostanie stan wiedzy obejmujący teoretyczne aspekty z zakresu statystyki małych obszarów ze szczególnym zwróceniem uwagi na jej zastosowania w badaniach poświęconych ubóstwu. Zawarte zostały w nim streszczenia najważniejszych pozycji literaturowych obejmujących artykuły naukowe, prezentacje, monografie oraz opracowania, w których nacisk położony został zarówno na warstwę aplikacyjną jak i teoretyczną statystyki małych obszarów. Opisane w tym rozdziale pozycje mają przybliżyć aktualny stan wiedzy w tym zakresie. Ze względu na zróżnicowany charakter opisywanych źródeł, ich objętość oraz stopień skomplikowania, przy każdej z omówionych pozycji zawarto informację o skali jej trudności. Przyjęto przy tym trzy poziomy informujące, że dana pozycja nie wymaga znajomości zaawansowanych metod matematyczno-statystycznych [1], wymaga pewnego stopnia znajomości tych technik [2] oraz, że jest przeznaczona dla bardzo zaawansowanych odbiorców, którzy posiadają odpowiedni warsztat matematyczny [3]. W rozdziale tym opisane zostały ponadto najważniejsze projekty badawcze, które poświęcone były teoretycznym i praktycznym aspektom statystyki małych obszarów, również uwzględniające te, w których uwaga skupiona była na zagadnieniach estymacji ubóstwa na niskich poziomach agregacji przestrzennej.

2.1 Statystyka małych obszarów a problem ubóstwa – przegląd literatury

Bartosińska D. (2006), *Attempts at Applying Small Area Estimation Method in Agricultural Sample Surveys in Poland*, Statistics in Transition, Vol. 7,

No. 6, pp.1203—1218 http://pts.stat.gov.pl/cps/rde/xbcr/pts/PTS_sit_7_6.pdf

Streszczenie: Artykuł poświęcony jest zastosowaniu metod statystyki małych obszarów w reprezentacyjnych badaniach rolnych w Polsce na poziomie powiatów, przy użyciu Powszechnego Spisu Rolnego jako źródła danych pomocniczych. Autor przybliżył badania rolne prowadzone przez Główny Urząd Statystyczny w kontekście estymacji dla małych obszarów. Przedstawione zostały potencjalne źródła zmiennych pomocniczych oraz wykorzystane metody SMO, tj. estymator bezpośredni, model regresji na poziomie obszaru i na poziomie jednostki. Dla uzyskania bardziej precyzyjnych oszacowań wykorzystane zostało podejście bayesowskie. Autor przedstawił wyniki estymacji dla prostego podejścia (estymacja bezpośrednia, regresyjna) oraz dla EB i HB opartych na modelach budowanych na poziomie obszaru lub jednostki. W artykule rozważony został dodatkowo model regresji na poziomie jednostki i obszaru, stosowany w celu uniknięcia efektu występującego w sytuacji dopasowania modelu na poziomie jednostki, a następnie dla predykcji na podstawie danych na poziomie obszaru. W przypadku dopasowania modelu dla szacowania wybranych charakterystyk opisujących ubóstwo na poziomie jednostki i dalszego prognozowania przy użyciu danych na poziomie powiatu model taki może okazać się bardzo pomocny.

Skala trudności: [1]

Best N., Richardson S., Clarke P., Gomez-Rubio V. (2008), *A Comparison of Model-Based Methods for Small Area Estimation, Working paper BIAS: Research Programme* http://www.bias-project.org.uk/papers/ComparisonSAE.pdf

Streszczenie: W pracy autorzy dokonali przeglądu metod statystyki małych obszarów bazujących na modelu z wykorzystaniem różnych źródeł danych. Szczególna uwaga została zwrócona na to, jak efektywnie wykorzystywać różne informacje z badań statystycznych oraz jak radzić sobie w przypadku estymacji w domenach, w których nie ma dostępnych bezpośrednich danych jednostkowych. Rozważano metody oparte na schemacie losowania (estymator Horvitza–Thompsona), estymację regresyjną oraz estymatory EBLUP, w których parametry są szacowane w wykorzystaniem metody największej wiarygodności, jak również za pomocą podejścia bayesowskiego. Z rodziny ogólnych liniowych modeli mieszanych autorzy rozważali model z efektami mieszanymi, zawierającymi efekty stałe jak i losowe, budowanych na poziomie

obszaru oraz na poziomie jednostki. Dodatkowo rozważano sposób uwzględniania w modelach korelacji przestrzennych pomiędzy obszarami na przykładzie liniowych modeli mieszanych z korelacją przestrzenną typu SAR i CAR. W celu zbadania jakości rozważanych modeli przeanalizowano i opisano różne kryteria porównawcze i selekcji modelu do danych pochodzących ze Szwecji (zrównoważony dochód na gospodarstwo domowe) oraz Anglii i Walii (dochód na gospodarstwo domowe).

Skala trudności: [3]

Betti G., Lemmi A., Lewandowski P., Neri L., Salvati N., Zięba A. (2006), *Analiza wskaźnikowa wykluczenia i integracji społecznej na poziomie powiatów z wykorzystaniem statystyki małych obszarów w: Wykluczenie i integracja społeczna w Polsce. Ujęcie wskaźnikowe*, Ministerstwo Pracy i Polityki Społecznej, Warszawa, s. 157-201, rszarf.ips.uw.edu.pl/wykluczenie/raport_undp.pdf

Streszczenie: Autorzy wykorzystują Badanie Aktywności Ekonomicznej Ludności (BAEL) oraz Badanie Budżetów Gospodarstw Domowych (BBGD) do szacowania pewnych charakterystyk ekonomicznych na szczeblu powiatowym. Mała bądź zerowa próba w powiatach we wspomnianych badaniach uniemożliwia zastosowanie klasycznych metod estymacji, w związku z czym został wykorzystany model Faya-Herriota na poziomie obszaru z racji dostępności odpowiednich danych. Z wykorzystaniem tej metody oraz danych pochodzących z BBGD, Narodowego Spisu Powszechnego 2002 oraz Banku Danych Lokalnych został oszacowany przeciętny dochód ekwiwalentny, a także dochód per capita w powiatach Polski dla roku 2005. Na tej podstawie autorzy wyznaczyli trzy granice ubóstwa, na poziomie 50%, 60% i 70% mediany dochodów, które następnie wykorzystali do estymacji stopy ubóstwa w powiatach. Wynikiem tych prac są mapy ubóstwa w przekroju powiatów dla trzech granic ubóstwa oraz utworzonego dodatkowo wskaźnika kompozytowego. Ponadto opracowanie zostało wzbogacone o przestrzenną analizę otrzymanych wyników. W dalszej części pracy autorzy skupiają się na estymacji wskaźników wykluczenia z rynku pracy na podstawie danych z BAEL.

Skala trudności: [1]

Datta G.S., Ghosh M., Steorts R., Maples J. (2009), *Bayesian Benchmarking with Applications to Small Area Estimation*, Reaearch Report Series (Statistics 2009-01) https://www.census.gov/srd/papers/pdf/rrs2009-01.pdf

Streszczenie: Wykorzystując podejście modelowe, można otrzymać oszacowania, które różnią się od oszacowań bezpośrednich zwłaszcza dla obszarów z małą liczebnością próby. Może również wystąpić sytuacja, w której wyniki uzyskane za pomocą modelu w małych obszarach po agregacji na poziom większego obszaru różnią się od wyników wiarygodnych, uzyskanych dla tego poziomu. W takich przypadkach można zastosować technikę benchmarkingu lub rakingu. W artykule pokrótce opisany został benchmarking oraz dokonano przeglądu odnośnie literatury przedmiotu. Autorzy wyprowadzili estymator BBE (Benchmarked Bayes Estimator) dla modeli na poziomie obszaru za pomocą średniej ważonej oraz średniej ważonej i ważonej zmienności. Wynikowe estymatory stanowią bardziej ogólną klasę dla wielu estymatorów uwzględniających benchmarking proponowanych wcześniej w literaturze. Otrzymane estymatory zastosowane zostały na przykładzie programu SAIPE (The Small Area Income and Poverty Estimates) dla przypadku estymacji liczby ubogich dzieci w wieku szkolnym. Rozważanym modelem był, często stosowany w statystyce małych obszarów, klasyczny model Faya-–Herriota.

Skala trudności: [3]

Elbers C., Lanjouw J.O., Lanjouw P. (2003), *Micro-Level Estimation of Poverty and Inequality*, Econometrica, Vol. 71, No. 1, ss. 355-364. http://siteresources.worldbank.org/DEC/Resources/micestpovineq.pdf

Streszczenie: Artykuł przedstawia autorski model estymacji poziomu ubóstwa dla obszarów, w których występuje mała liczebność próby. Podejście to zostało określone mianem metody/modelu ELL od nazwisk jej twórców i jest wykorzystywane w pracach Banku Światowego. Celem artykułu jest empiryczna weryfikacja własności tego modelu na podstawie danych statystycznych przekazanych przez Narodowy Instytut Statystyczny i Spisowy Ekwadoru. Przeprowadzone symulacje wskazują na znaczną poprawę precyzji otrzymanych szacunków stopy ubóstwa nawet dla bardzo mało licznych prób w szczegółowo zdefiniowanych przekrojach. Ponadto autorzy powołują się na doświadczenia z innych krajów w stosowaniu tej metody wskazując jej uniwersalność.

Skala trudności: [3]

Estaban M.D., Morales D., Pérez A., Santamaría L. (2011), *Small Area Estimation of Poverty Proportions Under Area-Level Time Models*, Computational Statistics and Data Analysis, Vol. 56, No. 10, ss. 2840-2855

Streszczenie: Celem artykułu jest ocena możliwości wykorzystania modelu na poziomie obszaru z szeregiem czasowym. Autorzy proponują model wykorzystujący przeszłe dane kwartalne, który umożliwi estymację stopy ubóstwa w zadanych domenach. Takie podejście pozbawione jest ograniczeń związanych estymacją związków nieliniowych, ale jest trudne aplikacyjnie w przeciwieństwie do standardowych liniowych modeli statystyki małych obszarów. Kolejną niedogodność tworzy dostęp do danych, które muszą obejmować okres sprzed roku, który stanowi obiekt zainteresowania. W pracy dokonano estymacji stopy ubóstwa w podziale na płeć we wszystkich 52 prowincjach Hiszpanii. Szacunku dokonano na podstawie Hiszpańskiego Badania Dochodów i Warunków Życia (SILC) z lat 2004-2006. Wykazano, że prowincje położone w północnej i północno-zachodniej części kraju charakteryzują się dużo niższą stopą ubóstwa niż te położone w części południowej.

Skala trudności: [3]

Fabrizi E., Ferrante M.R., Pacei S. (2005), *Estimation of Poverty Indicator Sub-National Level Using Multivariate Small Area Models*, Statistics in Transition, Vol. 7, No. 3, ss. 587—608

Streszczenie: Głównym celem tej pracy jest oszacowanie kilku dochodowych wskaźników ubóstwa spośród tych zaproponowanych w Leaken na poziomie niższym niż krajowy. Autorzy chcąc wyjść naprzeciw rosnącemu zapotrzebowaniu na szczegółowe dane dotyczące ubóstwa w przekrojach regionalnych proponują podejście hierarchiczne bayesowskie oparte na wielowymiarowym modelu na poziomie obszaru. Bazując na danych pochodzących z Europejskiego Badania Panelowego Gospodarstw Domowych oraz Włoskiego Instytutu Statystycznego konstruowane są modele opisujące poziom ubóstwa w regionach Włoch. Ocena otrzymanych szacunków dokonana jest przy pomocy metody bootstrap oraz obliczonego współczynnika zmienności. Wynikiem przeprowadzonych prac jest otrzymanie oszacowań kilku wskaźników opisujących ubóstwo dla 21 regionów Włoch, które cechują się dużo niższym błędem szacunku, niż te otrzymane w sposób bezpośredni.

Skala trudności: [2]

Ferretti C., Molina I. (2012), *Fast EB Method for Estimating Complex Poverty Indicators in Large Population*, Journal of the Indian Society of Agricultural Statistics Vol. 66, No. 1, ss. 105-120 tp://isas.org.in/jisas/jsp/volume/vol66/09-Caterina%20Ferretti.pdf

Streszczenie: Celem artykułu jest estymacja stopy ubóstwa w regionach Toskanii, a także na poziomie prowincji oraz samorządu w oparciu o dane pochodzące z Włoskiego Badania Dochodów i Warunków Życia (SILC). Autorzy w definiowaniu wskaźnika ubóstwa nie wykorzystali klasycznej definicji ubogiego gospodarstwa domowego opartej wyłącznie na dochodach, a zdecydowali się na zastosowanie bardziej złożonego wskaźnika opartego na zbiorach rozmytych. Ponadto został wykorzystany estymator empiryczny bayesowski w szybszej wersji niż ta zaproponowana oryginalnie przez Molinę i Rao w 2010 roku. W ten sposób przy małym spadku efektywności estymatora znacznie zmniejszył się czas konieczny na dokonanie szacunków. Wykorzystanie tej metody pozwoliło na otrzymanie oszacowań stopy ubóstwa w 10 prowincjach oraz 54 jednostkach samorządowych Toskanii. Otrzymane wyniki wskazują, że najbardziej dotknięte ubóstwem są rejony znajdujące się w północnej oraz centralnej części tego regionu.

Skala trudności: [3]

Hastings D., Maine N., Brown G., Cruddas M. (2003), *Development of Improved Estimation Methods for Local Area Unemployment Levels and Rates*, Technical report, Labor Market Trends, Vol. 111, No. 1, pp. 37–43.

Streszczenie: W artykule opisane zostało podejście oparte na modelu służącym do estymacji bezrobocia według definicji ILO w małych obszarach oraz przedstawiono wyniki dla wybranych poziomów agregacji przestrzennej (unitary authorities/local authority districts, UA/LAD) w Wielkiej Brytanii w latach 1995/96-–1999/2000. Autorzy przytoczyli ogólne informacje o badaniu aktywności ekonomicznej - LFS (Labour Force Survey), na podstawie którego jedynie jedna czwarta wyników na docelowym poziomie UA/LAD może być publikowana. W związku z tym zaproponowane zostało podejście oparte na modelu oraz opisane zostały związki bezrobocia, dla którego dane uzyskano na podstawie danych z LFS z informacjami dodatkowymi, pochodzącymi ze spisów powszechnych lub źródeł administracyjnych. W artykule wyszczególnione zostały wytyczne dotyczące zastosowania estymatorów i ich ograniczeń. Autorzy przedstawili analizę obszarów, dla których mogą być publikowane wyniki oszacowań z LFS i oszacowań na podstawie modelu, a także zaprezentowali porównanie wiarygodności oszacowań przy pomocy mapy tematycznej. Z analizy autorów wynika, że w podejściu modelowym mogą być publikowane wyniki dla wszystkich docelowych obszarów. Jedynie w 3 do 5 (na 406) obszarów wyniki bezpośrednie były bardziej wiarygodne w rozważanych latach.

Skala trudności: [1]

Inglese F., Russo A., Russo M. (2008), *Diagnostics of Small Area Model-Based Estimators.*

http://old.sis-statistica.org/files/pdf/atti/rs08_spontanee_a_2_5.pdf

Streszczenie: Głównym celem krótkiego artykułu jest przegląd metod diagnostycznych dla estymacji dla małych obszarów opartych na modelu. Ocena taka jest przydatna podczas wyboru odpowiedniego modelu spośród rozważanych modeli jakie oferuje statystyka małych obszarów. Metody opierają się na kluczowym założeniu, że oszacowania bezpośrednie dla małych obszarów są nieobciążone, a związane z nimi przedziały ufności pokrywają odpowiednie oszacowania. Autorzy przedstawili diagnostykę obciążenia wyników, dopasowania modelu, pokrycia i kalibracji. Podane metody zostały zobrazowane dla badania aktywności ekonomicznej w 2001 roku we Włoszech (szacowanie liczby zatrudnionych) przyjmując sześć estymatorów: EBLUP i syntetyczny na poziomie jednostki lub obszaru oraz estymator przestrzenny proporcjonalny i nieproporcjonalny. Na podstawie analizy wyników można zauważyć, że najlepszym estymatorem na podstawie przeprowadzonej diagnostyki jest EBLUP dopasowany na poziomie jednostki. Artykuł może być pomocny w dokonywaniu diagnostyki dla budowanych modeli służących oszacowaniu wybranych charakterystyk związanych z ubóstwem.

Skala trudności: [1]

Krieg, S., Blaess, V., Smeets, M. (2012), *Small Area Estimation of Turnover of the Structural Business Survey*, Statistics Netherlands.

Streszczenie: W przypadku małych prób podejście wykorzystujące uogólniony estymator regresyjny (GREG) skutkuje wysoką wariancją oszacowań. W związku z tym alternatywnie rozważa się wykorzystanie podejścia modelowego (ang. model-based estimation). Autorzy w artykule przeprowadzili badanie symulacyjne mające na celu ocenę podejścia modelowego opartego na regresji wielopoziomowej (ang. multilevel regression) na przykładzie szacowania obrotów przedsiębiorstw badanych w ramach prac w Statistics Netherlands (badanie: the Structural Business Survey of Statistics Netherlands). Zastosowany estymator EBLUP w znaczącym stopniu zmniejszył błędy szacunku w porównaniu do estymatora GREG. Jednakże, w związku z tym, że obroty charakteryzują się silną asymetrią, estymator EBLUP jest obciążony. Autorzy zastosowali transformację zmiennej określającej obroty, co wpłynęło

na zmniejszenie obciążenia oraz asymetrii oszacowań estymatorem EBLUP, jednak kosztem wyższej wariancji. Zaproponowane podejście przez autorów może być jednym z podejść stosowanych do estymacji ubóstwa w małych obszarach z wykorzystaniem regresji wielopoziomowej oraz przy występowaniu asymetrycznych rozkładów badanych cech.

Skala trudności: [2]

Kubacki J., Jędrzejczak A., Piasecki T. (2012), *Wykorzystanie metod statystyki małych obszarów do opracowania wyników badań statystycznych*, Urząd Statystyczny w Łodzi, Raport z pracy metodologicznej 3.065.

Streszczenie: Głównym celem raportu jest próba przedstawienia najważniejszych metod jakie oferuje statystyka małych obszarów, które mogą być wykorzystane przez urzędy statystyczne w produkcji odpowiednich danych. Autorzy opracowania w części teoretycznej dokonują kompleksowego opisu najważniejszych podejść stosowanych w statystyce małych obszarów ze szczególnym zwróceniem uwagi na estymację bezpośrednią oraz wykorzystującą podejście modelowe. W części poświęconej estymacji bezpośredniej skupiają swoją uwagę na metodach replikacyjnych, wykorzystujących metodę bootstrap, metodę linearyzacji Taylora oraz uogólnionej funkcji wariacyjnej. W części poświęconej wykorzystaniu modeli w statystyce małych obszarów dokonują przeglądu i szczegółowego opisu estymatora EBPLUP oraz empirycznej i hierarchicznej estymacji bayesowskiej. W warstwie aplikacyjnej autorzy pokazują potencjalne korzyści płynące z wykorzystania opisanych technik statystyki małych obszarów w Badaniu Budżetów Gospodarstw Domowych na poziomie województwa i powiatu. Opisane w części teoretycznej metody są na tyle ogólne, że można je również zaaplikować w badaniu dotyczącym ubóstwa, korzystając z danych pochodzących z EU-SILC oraz źródeł alternatywnych, takich jak spisy czy rejestry.

Skala trudności: [3]

Lehtonen R., Veijanen A. (2012), *Small Area Estimation Poverty by Model Calibration*, Journal of The Indian Society of Agricultural Statistics, 66(1), 125–133.

Streszczenie: Głównym celem artykułu jest próba wykorzystania podejścia kalibracyjnego do szacowania stopy ubóstwa na niskich poziomach agregacji przestrzennej. W klasycznym podejściu kalibracyjnym wagi wynikające ze schematu losowania próby odtwarzają wartości globalne zmiennych pomocniczych. Autorzy w swoim artykule skupiają się na tzw. modelowym ujęciu

kalibracji i proponują dwa podstawowe estymatory tzw. semi-bezpośredni i semi-pośredni estymator kalibracyjny. Zaproponowane estymatory wykorzystują przestrzenne zależności między zmiennymi objaśniającymi. Autorzy, w drodze symulacji z wykorzystaniem danych pochodzących z fińskich rejestrów związanych z tematyką ubóstwa, dokonują porównania zaproponowanych estymatorów ze znanymi i szeroko wykorzystywanymi przez urzędy statystyczne estymatorami tj. Horvitza-Thompsona i GREG.

Skala trudności: [3]

Longhurst J., *Model-Based Estimation of Income: Measuring Change Over Time*, 7th Meeting of the National Statistics Methodology Advisory Committee.

Streszczenie: Autorzy publikacji wykazują, że istnieje powszechna i silna potrzeba posiadania lepszych informacji o dochodach na poziomie małego obszaru. ONS, w związku z decyzją rządu o nie zawarciu pytania o dochody w Narodowym Spisie Powszechnym w 2001 roku, podjął decyzję o zastosowaniu technik modelowania opracowanych w prowadzonym przez nich projekcie Small Area Estimation Project (SAEP) przy tworzeniu szacunków dochodów dla niskich poziomów agregacji (okręgi). Oszacowania oparte na modelu i wyznaczone przedziały ufności zostały uzyskane z wykorzystaniem średnich tygodniowych przychodów gospodarstw domowych dla wszystkich okręgów w Anglii oraz Walii, a także zostały opublikowane na stronie internetowej ONS jako dane eksperymentalne.

Skala trudności: [1]

Marhuenda Y., Molina I., Morales D. (2013), *Small Area Estimation with Spatio-Temporal Fay-Herriot Models*, Computational Statistics and Data Analysis, Vol. 58, ss. 308-325

Streszczenie: W artykule porównano oszacowania stopy ubóstwa otrzymane z wykorzystaniem standardowego modelu Faya-Herriota z modelem Faya-Herriota uwzględniającym korelacje przestrzenne oraz czynnik czasu pod kątem ich efektywności. Autorzy w badaniu symulacyjnym wykazują, że zmodyfikowany model Faya-Herriota cechuje się mniejszym obciążeniem oraz błędem szacunku niż model oryginalny. Oprócz badań symulacyjnych została także przeprowadzona aplikacja modelu z wykorzystaniem danych rzeczywistych. W tym celu wykorzystano Hiszpańskie Badanie Dochodów i Warunków Życia (SILC) z lat 2004–2008. Wyniki wskazują na znaczną przewagę modelu

Faya-Herriota z korelacją przestrzenną i czynnikiem czasu w zakresie wielkości błędu oszacowania. Zaproponowany przez autorów model charakteryzuje się największą efektywnością. W rezultacie prowadzonych prac otrzymano oszacowania stopy ubóstwa dla 52 prowincji Hiszpanii, które cechowały się dużo niższym błędem szacunku, niż te otrzymane w sposób bezpośredni. Artykuł zawiera także mapę tematyczną, która prezentuje przestrzenne zróżnicowanie ubóstwa w Hiszpanii, które jest zgodne z wcześniejszymi publikacjami i wiedzą na ten temat w badanym kraju.

Skala trudności: [1]

McCulloch Ch. E., (1998) *Model-Based Approaches to Small Area Estimation with Binary Data* http://www.amstat.org/sections/SRMS/Proceedings/papers/1998_008.pdf

Streszczenie: Według autora publikacji statystyka małych obszarów jest długoletnim problemem w badaniach reprezentacyjnych, który pojawia się w różnych kontekstach dotyczących estymacji dla obszarów o małym zasięgu przestrzennym. Jeżeli w badaniu przynajmniej niektóre warstwy mają próby o bardzo małej liczebności, jest potrzebne specjalne podejście do szacunków z wykorzystaniem statystyki małych obszarów. W takim przypadku bezpośrednie oszacowania mogą okazać się wysoce nieefektywne a korzystne mogą okazać się techniki, które „pożyczają moc". Autor postawił sobie za cel w tym artykule opisanie pewnych modeli mieszanych, odpowiednich do analizy binarnych danych z badania oraz porównanie i skonfrontowanie metod estymacji dla różnych modeli. Autor porusza również kwestie różnic pomiędzy wnioskowaniem opartym na modelu (ang. model-based) a opartym na schemacie losowania próby (ang. designed-based). Zagadnienia omawiane w publikacji mogą okazać się pożyteczne przy mapowaniu ubóstwa na etapie wyboru odpowiedniego modelu dla zmiennych binarnych.

Skala trudności: [2]

Molina I., Rao J.N.K. (2013) *A Review of Poverty Mapping Procedures*, 59th World Statistics Congress, Hong Kong, China (Session IPS080) http://www.statistics.gov.hk/wsc/IPS080-P3-S.pdf

Streszczenie: W artykule dokonano przeglądu procedur mapowania ubóstwa, korzystając z technik jakie oferuje statystyka małych obszarów. Wyjaśnione zostały metody estymacji obejmujące podejście bezpośrednie i pośrednie. Opisano również podstawowe modele, popularne w statystyce małych

obszarów – model Faya-Herriota dla poziomu obszaru oraz model Battesse'a, Hartera i Fullera (BHF) dla poziomu jednostki będący modelem regresji liniowej zawierającym losowy efekt obszaru. Autorzy zaznaczają, że modele z efektami losowymi należą do ogólnej klasy liniowych modeli mieszanych szczególnie często wykorzystywanych w naukach społecznych. Następnie autorzy wymieniają trzy popularne podejścia szacowania stopy ubóstwa występujące w literaturze – model Faya-Herriota, regularnie stosowany przez biuro spisowe US Census Bureau, ale także w ramach projektu SAIPE i EURAREA, metodę zaproponowaną przez Elbersa, Lanjouw i Lanjouw (ELL), używaną przez Bank Światowy, która jest szczególnie użyteczna w przypadku złożonych i nieliniowych wskaźników ubóstwa, podobną do modelu BHF oraz estymatory empiryczne bayesowskie (EB), opisane w ramach projektu SAMPLE. Celem rozszerzenia podstawowych modeli na poziomie jednostki, do szacowania stopy ubóstwa autorzy proponują szybką metodę EB, procedurę hierarchiczną bayesowską (HB) bazującą na klasycznym EB. Rozszerzeniem ELL mogą być modele zakładające mieszaninę rozkładów normalnych, a także procedury bazujące na M-kwantylach.

Skala trudności: [2]

Molina, I., Rao, J.N.K. (2010), *Small Area Estimation of Poverty Indicators, Canadian Journal of Statistics*, 38(3), 369–385.

Streszczenie: Autorzy proponują zastosowanie statystyki małych obszarów do estymacji stopy ubóstwa z wykorzystaniem metody Empirical Best oraz regresji z zagnieżdżonym błędem (ang. nested error model). Podkreślony jest aspekt nieliniowości sposobu szacowania ubóstwa oraz problemów z tego wynikających. Zastosowany został również parametryczny bootstrap do oszacowania błędu średniokwadratowego dla proponowanego modelu. W wyniku badania symulacyjnego, porównującego zaproponowane podejście oraz oszacowania bezpośrednie i metodę ELL, stwierdzono znaczną poprawę błędów szacunku przy wykorzystaniu modelu EB oraz regresji z zagnieżdżonym błędem. Autorzy zastosowali prezentowane podejście do estymacji stopy ubóstwa w Hiszpanii dla małych obszarów wskazując znaczną poprawę precyzji w porównaniu z podejściem bezpośrednim. Prezentowane w artykule podejście jest kluczowe dla estymacji stopy ubóstwa dla powiatów ponieważ może wskazać na znaczną poprawę estymacji w porównaniu do metodologii stosowanej przez Bank Światowy (metoda ELL).

Skala trudności: [3]

Molina I., Rao J.N.K., Nandram B. (2014), *Small Area Estimation of General Parameters with Application to Poverty Indicators: A Hierarchical Bayes Approach*, The Annals of Applied Statistics, Vol. 8, No. 2, 852–885 http://arxiv.org/pdf/1407.8384.pdf

Streszczenie: Głównym celem artykułu jest próba wykorzystania metod jakie oferuje statystyka małych obszarów, ze szczególnym uwzględnieniem tzw. podejścia bayesowskiego, do szacowania wybranych parametrów i wskaźników opisujących ubóstwo. Autorzy artykułu z wykorzystaniem metod bayesowskich dokonali oszacowania stopy ubóstwa na poziomie prowincji w Hiszpanii w przekroju płci. Ze względu na małe liczebności próby w Hiszpańskim Badaniu Dochodów i Warunków Życia (SILC) we wskazanych przekrojach nie jest możliwe wyznaczenie odpowiednich oszacowań stopy ubóstwa z wykorzystaniem estymatora bezpośredniego co jest konsekwencją niskiej precyzji. W związku z tym autorzy zaproponowali sposób uzyskania odpowiednich oszacowań i stworzenia mapy tematycznej ubóstwa dla Hiszpanii z wykorzystaniem podejścia hierarchicznego bayesowskiego, którego jedną z zalet jest możliwość szacowania złożonych parametrów na niskich poziomach agregacji przestrzennej. Autorzy na potrzeby dokonania szacunków skorzystali z odmiany podejścia hierarchicznego bayesowskiego w postaci tzw. modelu regresji z zagnieżdżonym błędem. Wykorzystując rzeczywiste dane w badaniu symulacyjnym dokonali oceny zaproponowanego podejścia, a także stworzyli mapę ubóstwa na poziomie prowincji w Hiszpanii w przekroju płci. Wskazali dzięki temu rejony, w których kobiety są bardziej narażone na zjawisko ubóstwa, a także wyznaczyli oceny punktowe stopy ubóstwa wraz z oceną ich precyzji . Zgodnie z uzyskanymi wynikami w większość prowincji Hiszpanii kobiety są w większym stopniu aniżeli mężczyźni narażeni na ubóstwo. Stworzone mapy pokazały również, że zjawisko ubóstwa w większym stopniu dotyczy prowincji zlokalizowanych w południowej jak i wschodniej części Hiszpanii.

Skala trudności: [3]

Myrskylä M. (2007), *Generalised Regression Estimation for Domain Class Frequencies, Research Reports*, 247, Statistics Finland https://helda.helsinki.fi/bitstream/handle/10138/23380/generali.pdf?sequence=1

Streszczenie: Głównym celem pracy jest analiza własności uogólnionego estymatora regresyjnego (GREG) dla przypadku estymacji liczebności i proporcji w domenach. Rodzina estymatorów typu GREG tworzy klasę estymato-

rów opartych na schemacie losowania wspomaganych modelem. Autor zwraca uwagę, że przypadku szacowania liczebności bardziej adekwatne będą modele typu logistycznego lub modele typu logistycznego z liniowymi efektami stałymi (L-GREG). W pracy porównana została teoretyczna i empiryczna dokładność rozważanych estymatorów, bazując na eksperymencie Monte Carlo i różnych schematach losowania. Badano również jakość standardowych oszacowań wariancji dla przyjętego schematu losowania oraz zaproponowano estymator wariancji biorący pod uwagę różnicę między dopasowaniem modelu w próbie i spisie powszechnym, będącym według autora dobrą alternatywą dla standardowego estymatora wariancji. Artykuł może być przydatny w badaniach polegających na szacowaniu liczby osób ubogich w szczegółowo zdefiniowanych domenach.

Skala trudności: [3]

National Statistics (2006), *Model-Based Estimates of ILO Unemployment for LAD/UAs in Great Britain*, Guide for Users

Streszczenie: Celem artykułu jest opis metodologii statystyki małych obszarów, która może być przydatna w kontekście szacowania stopy bezrobocia według definicji Międzynarodowej Organizacji Pracy (ang. International Labour Organisation, ILO) na niskich poziomach agregacji przestrzennej. W artykule zamieszczono mapy tematyczne, prezentujące oszacowania stopy bezrobocia na poziomie LAD/UA dla Wielkiej Brytanii. Omówione metody, jak również poddane szczegółowej analizie w artykule kwestie wątpliwe, mogą okazać się cenną wskazówką przy mapowaniu ubóstwa – na przykład w kontekście szacowania i mapowania stopy ubóstwa.

Skala trudności: [2]

Pratesi M., Pedreschi D., Giannotti F, Marchetti S., Salvati N., Maggino F. (2012), *Small Area Model-Based Estimators Using Big Data Sources*, ISTAT.

Streszczenie: Jak wskazują autorzy opracowania, bieżące monitorowanie wskaźników opisujących ubóstwo i nierówności społeczne wymagają dokładnych i rzetelnych danych, które umożliwiają podejmowanie właściwych decyzji dotyczących polityki społecznej. Zdaniem autorów jest to możliwe dzięki wykorzystaniu metod jakie oferuje statystyka małych obszarów, które w połączeniu ich z nowymi źródłami danych w postaci tzw. Big Data może przyczynić się do znacznego zaspokojenia informacyjnego w wielu dziedzinach życia, w tym z zakresu ubóstwa. Artykuł jest szczególnie interesujący w kon-

tekście wykorzystania nowych źródeł danych w statystyce społecznej, nad których przydatnością w ciągu ostatnich lat wywiązała się w obrębie urzędów statystycznych gorąca debata.

Skala trudności: [1]

Quintaes V., Hansen N., Silva, Pessoa D.D., Pedro S. (2011), *a Fay–Herriot Model for Estimating the Proportion of Households in Poverty in Brazilian Municipalities*, Int. Statistical Inst.: Proc. 58th World Statistical Congress, 2011, Dublin (Session CPS016) http://2011.isiproceedings.org/papers/950809.pdf

Streszczenie: W referacie autorzy skupili uwagę na wykorzystaniu klasycznego modelu Faya-Herriota do estymacji odsetka gospodarstw zagrożonych ubóstwem na niskich poziomach agregacji przestrzennej (gminy) w Brazylii. Jednostką odpowiedzialną za dostarczenie odpowiednich oszacowań w Brazylii związanych z ubóstwem jest Instytut Geografii i Statystyki. Jest on zobligowany, przy istniejących ograniczeniach finansowych, dostarczać rzetelnych i aktualnych informacji na temat wybranych wskaźników opisujących ubóstwo w Brazylii na możliwie najniższych poziomach agregacji przestrzennej. Przedstawione w artykule wyniki badań zawierają pierwsze oszacowania odsetka ubogich gospodarstw w Brazylii na poziomie gmin. W tym celu autorzy wykorzystali obszarowy model Faya-Herriota. Dane do budowy odpowiedniego modelu pochodziły z rejestrów administracyjnych, spisu oraz badania reprezentacyjnego wydatków rodzinnych. Model Faya-Herriota nie był do tej pory wykorzystywany w Brazylii na potrzeby mapowania ubóstwa. Przez długi okres czasu Instytut Geografii i Statystyki korzystał z metodologii zaproponowanej przez Bank Światowy. W artykule autorzy dokonali również próby porównania wyników uzyskanych z wykorzystaniem podejścia stosowanego przez Bank Światowy z tym, jaki oferuje model Faya-Herriota na poziomie obszaru. W artykule wskazano również rolę zmiennych pomocniczych w budowie właściwych modeli oraz dokonano ich kwerendy.

Skala trudności: [2]

Rahman A. (2008), *A Review of Small Area Estimation Problems and Methodological Developments*, Series NATSEM Discussion Papers, Issue 66 http://www.natsem.canberra.edu.au/storage/Azizur_paper%20in%20new%20template_Work_CX%20-%20final%20edit.pdf

Streszczenie: Głównym celem pracy jest przegląd metod używanych w sta-

tystyce małych obszarów. Autor opisał techniki jakie oferuje statystyka małych obszarów w podziale na dwie grupy - pierwsza dotyczyła podejścia statystycznego opartego na modelu, natomiast druga dotyczyła podejścia bazującego na tzw. mikrosymulacji geograficznej. W pracy dokonano porównania różnych modeli mikrosymulacji, a także samych technik statystycznych i podejścia geograficznego. Opisana została estymacja oparta na schemacie losowania (Horvitz-Thompson, GREG) oraz na różnych modelach statystycznych. Rozważono podejścia modelowe obejmujące estymatory syntetyczne, złożone i demograficzne. Autor przedstawił modele na poziomie obszaru (area level) i poziomie jednostki (unit level) takie jak BLUP, EBLUP, EB i HB. Dla tych estymatorów dokonano porównań obejmujących ich zastosowania, własności, założenia i sposoby estymacji wariancji, a także złożoności estymatorów i sposobu estymacji parametrów modelu. W celu porównania opisywanych metod statystycznych, analizie poddano technikę zwaną „mikrosymulacją przestrzenną", bazującą na teoriach ekonomicznych. Przedstawiono metody mikrosymulacji – statyczną, dynamiczną i przestrzenną oraz pokazano techniki tworzenia mikrodanych wykorzystujące informację przestrzenną. Wymienione w pracy estymatory mogą być wykorzystane w wielu dziedzinach, w tym do estymacji ubóstwa. Autor wskazuje jednak na trudności związane z założeniami i warunkami, które w pewnych przypadkach nie mają odzwierciedlenia w praktyce.

Skala trudności: [3]

Rao J.N.K. (2003), *Small Area Estimation*, Wiley Series in Survey Methodology.

Streszczenie: Monografia jednego z najwybitniejszych autorytetów z zakresu statystyki małych obszarów, która stanowi najważniejszą pozycję z zakresu tej dyscypliny wiedzy. Opisane w niej zostały wszystkie najważniejsze techniki i metody jakie oferuje statystyka małych obszarów i które stosowane są w praktycznych zastosowaniach przez urzędy statystyczne, w tym w zakresie ubóstwa. W monografii omówiono kompleksowo estymatory bezpośrednie oraz estymatory pośrednie wykorzystujące podejście modelowe. Prezentowane w monografii profesora Rao techniki estymacji mogą z powodzeniem być wykorzystywane w mapowaniu ubóstwa na różnych poziomach agregacji przestrzennej.

Skala trudności: [3]

Rao, J. N. K. (2012), *Small Area Estimation: Methods and Applications, Application of Small Area Estimation Techniques in the Social Sciences*, Mexico City.

Streszczenie: Prezentacja J.N.K. Rao wygłoszona podczas seminarium Application of Small Area Estimation Techniques in the Social Sciences w Meksyku, poświęcona jest przeglądowi dotychczasowych metod oraz zastosowań statystyki małych obszarów. Autor przedstawia główne podejścia proponowane w statystyce małych obszarów: podejścia wykorzystujące modele liniowe z efektami stałymi i losowymi, model Faya-Harriota dla obszaru, estymatory typu EBLUP, Empirical Best, estymację bayesowską (hierarchiczny estymator bayesowski), estymację odporną oraz opartą na M-kwantylach. Wskazuje problemy związane z estymacją błędów średniokwadratowych (MSE) oraz prezentuje podejścia estymacji MSE oparte na metodach Monte Carlo, jak i bootstrap (naiwny, parametryczny, wykorzystujący kalibrację). W ostatniej części przedstawione są projekty poświęcone estymacji ubóstwa z wykorzystaniem metodologii statystyki małych obszarów z podkreśleniem jej roli w projekcie SAIPE. Prezentacja zakończona jest 10 rekomendacjami związanymi z zastosowaniem statystyki małych obszarów do badań społecznych, w tym estymacji stopy ubóstwa.

Skala trudności: [3]

Rao, J. N. K., Molina I., *Small Area Estimation Methods, Applications and Practical Demonstration*. http://halweb.uc3m.es/esp/Personal/personas/imolina/MiDocencia/SmallAreaEstimation/SAEstimation_Part1.pdf.

Streszczenie: Prezentacja przygotowana przez Rao oraz Molinę zawiera przegląd estymatorów stosowanych w statystyce małych obszarów. Autorzy wprowadzają podstawowe oznaczania wychodząc od estymacji bezpośredniej (estymator Horvitza-Thompsona) oraz wskazania miar ubóstwa (m.in. stopa ubóstwa), następnie prezentując podejście oparte na estymacji wykorzystującej zmienne pomocnicze (estymator post-stratyfikacyjny, estymator regresyjny oraz kalibrację). W drugiej części wprowadzone zostały oznaczenia związane z problemem estymacji dla małych obszarów oraz rozważane są podejścia oparte na schemacie losowania – estymatory syntetyczne, BARE, SPREE, Jamesa-Steina oraz estymatory złożone. Prezentacja ma na celu przedstawić zarówno stronę teoretyczną, jak również praktyczną stosowania metodologii statystyki małych obszarów. W związku z tym prezentowane estymatory zilustrowane zostały przykładami realnych ich zastosowań w róż-

nych badaniach, w tym z zakresu ubóstwa.

Skala trudności: [3]

Saei A., Chambers R. (2003), *Small Area Estimation: A Review of Methods Based on the Application of Mixed Models*, Methodology Working Paper M03/16.

Streszczenie: W statystyce małych obszarów wyróżnia się dwa zasadnicze podejścia do zagadnienia estymacji nieznanych parametrów w szczegółowo zdefiniowanych domenach. Jedno z nich wykorzystuje wagi ze schematu losowania próby. Drugie bazuje na odpowiednio zbudowanych modelach, które łączą zależności pomiędzy zmiennymi pochodzącymi z różnych źródeł (badanie próbkowe, spisy, rejestry). W dokumencie autorzy skupili swoją uwagę na drugim podejściu i szczegółowo omówili idę tzw. modeli mieszanych, które w statystyce małych obszarów wykorzystywane są w coraz większym stopniu. Idea modelowania w statystyce ma bardzo długą historię jednak w statystyce małych obszarów wykorzystywana jest stosunkowo od niedawna. Autorzy w opracowaniu dokonali kompleksowego przeglądu modeli mieszanych i ich zastosowań w obszarze estymacji pośredniej. Swoją uwagę skupili na wykorzystaniu w modelowaniu tzw. efektów stałych, które mogą być przydatne w wyjaśnianiu zmienności między domenami w odniesieniu do zmiennej, dla której dokonuje się estymacji parametrów (wartość globalna, średnia itd.) oraz tzw. efektów losowych, które mogą być przydatne w wyjaśnianiu „niewyjaśnionej" zmienności danej zmiennej między domenami. Modele mieszane łączą w sobie wykorzystanie obydwu typów efektów i są wykorzystywane w statystyce małych obszarów na przestrzeni ostatnich 30 lat. Autorzy w swoim raporcie dokonali kompleksowego przeglądu zarówno liniowych jak i nieliniowych modeli mieszanych a także szczegółowo omówili techniki estymacji parametrów tych modeli.

Skala trudności: [3]

Salvati N., Chandra H., Chambers R., (2010), *Model Based Direct Estimation of Small Area Distributions*, Centre for Statistical & Survey Methodology, The University of Wollongong, Working Paper 20-10, Wollongong http://ro.uow.edu.au/cgi/viewcontent.cgi?article=1069&context=cssmwp

Streszczenie: Wiele pozycji literaturowych poświęconych statystyce małych obszarów skupia się na takich parametrach jak wartość globalna czy średnia. Użytkownicy danych z badania są jednak często zainteresowani rozkładem

zmiennej badanej w skończonej populacji, a także innymi miarami (jak mediana, kwartyle czy percentyle), które charakteryzują kształt tego rozkładu na poziomie małego obszaru. Autorzy artykułu zaprezentowali w nim estymator bezpośredni oparty na modelu (ang. model-based direct estimator, MBDE). Jest to estymator funkcji rozkładu na poziomie małego obszaru. Jest on definiowany jako ważona suma danych z próby pochodzących z danego obszaru, z wagami otrzymanymi z kalibracji oszacowań funkcją sklejaną funkcji rozkładu skończonej populacji. Jest to możliwe przy odpowiedniej specyfikacji modelu regresji z losowymi efektami dla obszaru. Autorzy poddają również dyskusji średni błąd kwadratowy oszacowań tego estymatora bezpośredniego. Symulacje Monte Carlo oparte są zarówno na symulowanych, jak również rzeczywistych zbiorach danych i pokazują, że omawiany estymator i jego średni błąd kwadratowy wypadają dobrze w porównaniu do alternatywnych estymatorów. Artykuł może być przydatny w kontekście szacowania miar pozycyjnych w małych obszarach a związanych ze strefą ubóstwa.

Skala trudności: [3]

Torelli N., Trevisani M. (2008), *Labour Force Estimates for Small Geographical Domains in Italy: Problems, Data and Models*, Working Paper n. 118, Rivista internazionale di scienze sociali, Vol. 01, Universita cattolica del Sacro Cuore.

Streszczenie: W publikacji dokonano przeglądu metod, problemów i możliwości, a także rodzaju danych, które mogą być zastosowane w estymacji dla małych obszarów w celu uzyskania wiarygodnych informacji na poziomie prowincji lub niższym w badaniu aktywności ekonomicznej (LFS) we Włoszech. Autorzy przedstawili metodologię LFS oraz przybliżyli popularne modele i estymatory, w tym: model Faya-Herriota (FH), który prowadzi do otrzymania estymatora typu EBLUP, estymator HB oraz model Battessego, Hartera i Fullera dla poziomu jednostki, podobny do modelu FH. Dodatkowo, wspomniane zostały modele szeregów czasowych, mające duże znaczenie dla LFS w kontekście szacowania np. wskaźnika bezrobocia. Po zwięzłym przeglądzie metod, omówiono zastosowanie statystyki małych obszarów we włoskim LFS, a także poruszono kwestię doboru zmiennych pomocniczych. Autorzy opisali także nowsze osiągnięcia statystyki małych obszarów, tj. podejście hierarchiczne bayesowskie w modelach dla poziomu obszaru z danymi typu count-data (liczebności). Przedstawione zostały problemy, które pojawiają się w trakcie modelowania, np. niewłaściwe rozmieszczenie małych obszarów w przypadku różnych źródeł badań (niepoprawnie wyznaczone granice

małych obszarów). Może do tego dochodzić m.in. w przypadku łączenia jednostkowych danych pochodzących z różnych źródeł danych statystycznych.

Skala trudności: [3]

Tzavidis, N. (2010), *What is Poverty Mapping?*, Prezentacja w ramach konferencji Methods at Manchester, Manchester, 20 maja 2010, http://www.methods.manchester.ac.uk/events/whatis/povertymapping.pdf

Streszczenie: Prezentacja Nikosa Tzavidisa wygłoszona podczas konferencji Methods at Manchester poświęcona jest metodologii mapowania ubóstwa. Autor zdefiniował podstawowe pojęcia wykorzystywane w badaniach nad ubóstwem – stopę ubóstwa, granicę ubóstwa oraz mapowanie ubóstwa. Omówione zostały trzy główne metody wykorzystywane do estymacji ubóstwa dla małych obszarów – Metoda ELL (Bank Światowy), Empirical Best Predictor (Molina i Rao) oraz podejście oparte na M–kwantylach (Tzavidis i Chambers). Autor skupił się na bezpośrednim porównaniu omawianych metod z punktu widzenia ich przydatności oraz różnic w estymacji stopy ubóstwa. Zaprezentowane zostały wyniki na podstawie badania EU-SILC dla Toskanii, Lombardii oraz Kampanii uzyskane w ramach projektu SAMPLE oraz AMELI. Prezentowane rezultaty wskazują na znaczną poprawę szacunków w porównaniu do stosowania estymatorów bezpośrednich na poziomie NUTS 4 i NUTS 3.

Skala trudności: [1]

Tzavidis, N., Chambers, R., Salvati, N., Chandra, H. (2012), *Small Area Estimation in Practice: An Application to Agricultural Business Survey Data*, Journal of the Indian Society of Agricultural Statistics, 66(1), 213–228.

Streszczenie: Celem artykułu jest przedstawienie zastosowania statystyki małych obszarów do estymacji charakterystyk przedsiębiorstw rolniczych z wykorzystaniem badania przedsiębiorstw rolniczych w Australii (ang. Agricultural Business Survey). Autorzy rozważają znane w literaturze podejścia (m.in. EBLUP), jak również nowo zaproponowane oparte na M-kwantylach czy uwzględniające podejście odporne (Robust EBLUP). Artykuł przedstawia krok po kroku podejście stosowane w statystyce małych obszarów – od weryfikacji istnienia efektów między obszarami oraz wewnątrz obszarów (m.in. zastosowanie modeli mieszanych) po zastosowanie estymacji z wykorzystaniem estymatora EBLUP, odpornego EBLUP oraz M-kwantyli. Wyniki prezentowane przez autorów wskazują, że rozważane metody zmniejszają błędy

szacunku w porównaniu do estymacji bezpośredniej. Rozważana jest również charakterystyka tego typu badań, w której występują wartości odstające. Podejście zaproponowane przez autorów wskazuje na potencjalne możliwości wykorzystania tych metod do szacowania dochodów/wydatków będących podstawą estymacji stopy ubóstwa.

Skala trudności: [3]

Tzavidis N., Salvati N., Pratesi M., Chambers R. (2007), *M-Quantile Models with Application to Poverty Mapping*, Statistical Methods and Applications, Vol. 17, No. 3, ss. 393-411.

Streszczenie: W artykule zaproponowano nowatorskie podejście do estymacji poziomu ubóstwa w jednostkach przestrzennych cechujących się małą liczebnością próby. Klasyczne podejście modelowe bazuje na silnych założeniach dotyczących rozkładu i jest wrażliwe na wartości odstające. z kolei autorzy pracy opisują możliwości estymacji z wykorzystaniem regresji kwantylowej, która nie zależy tak silnie od rozkładu zmiennej i automatycznie łagodzi wpływ obserwacji odstających. Opisana metodologia została zilustrowana praktycznym zastosowaniem na danych pochodzących z Living Standards Measurement Survey przeprowadzonego w Albanii w 2002 roku. Model M-kwantylowy został wykorzystany do estymacji stopy ubóstwa w dystryktach Albanii. Otrzymane wyniki wskazują na najbiedniejsze rejony w części górskiej położonej w północnej i północno-wschodniej części kraju. W artykule zaprezentowano także mapy tematyczne dla estymowanych kwantyli rozkładu.

Skala trudności: [3]

US w Poznaniu, GUS (2013), Mapy ubóstwa na poziomie podregionów w Polsce z wykorzystaniem estymacji pośredniej, Warszawa. http://stat.gov.pl/z-prac-studialnych/mapy-ubostwa-na-poziomie-podregionow-w-polsce-z-wykorzystaniem-estymacji-posredniej,4,1.html

Streszczenie: W raporcie przygotowanym wspólnie przez pracowników Ośrodka Statystyki Małych Obszarów Urzędu Statystycznego w Poznaniu, Głównego Urzędu Statystycznego w Warszawie i Banku Światowego skupiono uwagę na wykorzystaniu modelu Faya-Herriota do szacowania stopy ubóstwa na poziomie podregionów w Polsce. Głównym celem raportu było stworzenie odpowiedniej mapy tematycznej ukazującej zasięg ubóstwa w Polsce na niepublikowanym do tej pory poziomie agregacji przestrzennej. Dotychczas

publikowany przez Główny Urząd Statystyczny w Warszawie wskaźnik obrazujący stopę ubóstwa był przedstawiany jedynie na poziomie całego kraju i w przekroju województw. Wykorzystanie podejścia modelowego z zakresu statystyki małych obszarów (model Faya-Herriota) oraz zmiennych pochodzących z różnych źródeł (EU-SILC), NSP 2011 i BDL umożliwiło estymację stopy ubóstwa na niższym poziomie z akceptowalną precyzją szacunku. Było to jedno z pierwszych praktycznych zastosowań metod jakie oferuje statystyka małych obszarów w mapowaniu ubóstwa w Polsce.

Skala trudności: [2]

van den Brakel J., Bethlehem J. (2008), *Model-Based Estimation for Official Statistics*, Statistics Netherlands, http://www.cbs.nl/NR/rdonlyres/B0CF3D94-85C1-437A-824A-7ECC30A05028/0/200802x10pub.pdf

Streszczenie: Głównym celem artykułu było wskazanie wad i zalet dwóch podejść wykorzystywanych w statystyce małych obszarów (bazującego na schemacie losowania próby i wykorzystującego odpowiednio zdefiniowany model) na potrzeby estymacji parametrów, które wykorzystywane są przez urzędy statystyczne różnych państw w bieżącej produkcji danych statystycznych. Autorzy opracowania wskazali sytuacje, w których wykorzystanie podejścia modelowego może przynieść wymierne korzyści w publikacji oficjalnych danych. Wskazali, że krajowe urzędy statystyczne w większym stopniu wykorzystują techniki bazujące na schemacie losowania próby a podejście modelowe wykorzystywane jest przez statystykę publiczną w ograniczonym zakresie. Poprzez dogłębną analizę wad i zalet modeli, które w statystyce małych obszarów zaczynają odgrywać coraz większą rolę oraz pokazując możliwe obszary ich zastosowań wykazują, że modelowanie może stanowić bardzo ważną technikę estymacji parametrów dla szczegółowo zdefiniowanych domen. Dotyczyć to może również ubóstwa, w którym wykorzystanie odpowiednio zbudowanych modeli stanowić może jedyną drogę szacowania parametrów (np. stopa ubóstwa).

Skala trudności: [1]

2.2 Projekty badawcze

2.2.1 SAIPE

SAIPE (**Small Area Income and Poverty Estimates**) jest programem prowadzonym przez amerykańskie biuro spisowe, którego celem jest dostar-

czanie corocznych szacunków dla poniższych zmiennych:

- liczba osób dotkniętych ubóstwem na poziomie hrabstw oraz stanów,
- liczba dzieci poniżej 5 roku życia dotkniętych ubóstwem na poziomie stanów,
- liczba dzieci na utrzymaniu w wieku od 5 do 17 lat w rodzinach dotkniętych ubóstwem na poziomie hrabstw i stanów,
- liczba dzieci poniżej 18 roku życia dotkniętych ubóstwem na poziomie hrabstw i stanów,
- mediana dochodów gospodarstw domowych na poziomie hrabstw i stanów,
- liczba ludności na poziomie okręgów szkolnych,
- liczba dzieci w wieku od 5 do 17 lat na poziomie okręgów szkolnych,
- liczba dzieci na utrzymaniu w wieku od 5 do 17 lat w rodzinach dotkniętych ubóstwem na poziomie okręgów szkolnych.

Uzyskiwane informacje są szczególności istotne dla władz lokalnych i wykorzystywane są, między innymi, do zarządzania federalnymi programami pomocowymi oraz alokacją federalnych funduszy. Informacje te nie są jednak dostarczane w oparciu o estymator bezpośredni z badania reprezentacyjnego, ani też bezpośrednio ze spisów lub rejestrów administracyjnych. Reprezentacyjne badanie **American Community Survey** prowadzone corocznie umożliwia przy wykorzystaniu estymatora bezpośredniego podawanie powyższych informacji dla obszarów powyżej 65 tysięcy mieszkańców. Raz na trzy lata badanie jest tak skonstruowane, aby dostarczać potrzebnych danych dla obszarów o liczbie mieszkańców 20 tysięcy i więcej. Natomiast raz na pięć lat na podstawie tego badania można uzyskać te informacje dla wszystkich analizowanych przekrojów. W celu uzyskania każdego roku takich szacunków konstruowane są specjalne modele, które łączą dane pochodzące z badań reprezentacyjnych, spisów powszechnych czy tez rejestrów administracyjnych. Jako zmienne pomocnicze wykorzystywane są, między innymi, szacunki zmiennych z okresów wcześniejszych, stany ludności w okresach między spisami oraz informacje pochodzące z rejestru podatkowego. Z rodziny estymatorów statystyki małych obszarów stosowany jest, między innymi, model Faya-Ferriota na poziomie obszaru. Ostatecznie uzyskiwane oszacowania,

mimo iż są oparte na modelach, zachowają zgodność obszarową tzn. szacunki dla niższych poziomów agregacji przestrzennej sumują się do odpowiednich szacunków dla obszarów z wyższego poziomu. Stosowana metodologia w ramach programu SAIPE z roku na rok w niewielkim stopniu jest korygowana i ulepszana w celu uzyskiwania możliwie najbardziej wiarygodnych oszacowań.

2.2.2 SAMPLE

Projekt SAMPLE (**Small Area Methods for Poverty and Living Condition Estimates** - Estymacja ubóstwa i warunków życia za pomocą metod estymacji małych obszarów) jest programem badawczym finansowanym przez Komisję Europejską. Został on przeprowadzony w okresie od marca 2008 roku do marca 2011 roku w ramach Siódmego Ramowego (7PR) programu Unii Europejskiej. Projekt był realizowany przez międzynarodowe konsorcjum, którego głównym koordynatorem była prof. Monica Pratesi, Universita di Pisa – Dipartimento di Statistica e Matematica Applicata all'Economia. Pozostałe instytucje, których pracownicy brali udział w projekcie to:

- Universita di Siena – Centro Interdipartimentale di Ricerca sulla Distribuzione del Reddito (Włochy),
- Cathie Marsh Centre for Census and Survey Research, University of Manchester (Wielka Brytania),
- Departamento de Estadistica, Universided Carlos III de Madrid (Hiszpania),
- Centro de Investigacion Operativa, Universidad Miguel Hernandez de Elche (Hiszpania),
- Provincia di Pisa – U.O. e Ricerche – Osservatorio per le Politiche Sociali – Ufficio Politiche Comunitarie (Włochy),
- Simurg Ricerche (Włochy),
- Centrum Edukacji Statystycznej, Główny Urząd Statystyczny (Polska),
- Szkoła Główna Handlowa (Polska).

Celem projektu SAMPLE była identyfikacja i opracowanie nowych wskaźników, które pomogą w zrozumieniu zjawisk z zakresu nierówności i wykluczenia społecznego oraz ubóstwa. Podjęte w ramach projektu inicjatywy przyczyniły się do skonstruowania stosownych modeli oraz wdrożenia procedur

do szacowania analizowanych wskaźników wraz z ich odpowiednią precyzją. Podstawą do prowadzenia szacunków było Europejskie Badanie Dochodów i Warunków Życia (EU-SILC). Dodatkowo prowincja Pizza miała zwiększoną próbę gospodarstw domowych podlegających badaniu. Alternatywnymi źródłami informacji o wybranych aspektach warunków życia ludności były rejestry administracyjne i lokalne bazy danych. W ramach projektu określono, między innymi, następujące cele szczegółowe:

- identyfikacja nowych wskaźników ubóstwa i wykluczenia społecznego na poziomie NUTS 3 oraz NUTS 4;
- opracowanie nowych modeli szacowania alternatywnych wskaźników na poziomie lokalnym, poprzez integrację danych pochodzących EU-SILC i dostępnych lokalnych administracyjnych baz danych;
- rozwój nowych metod estymacji wskaźników na poziomie lokalnym poprzez efektywne wykorzystanie wielu dostępnych źródeł danych;
- określenie i wdrożenie rozwiązań, które będą angażować zainteresowane strony w proces produkcji i interpretacji alternatywnych wskaźników ubóstwa i wykluczenia społecznego;
- rozwój narzędzi (oprogramowania, ankiet, formularzy, przewodników z obszaru najlepszych praktyk), które wspomogą cały proces produkcji rozważanych wskaźników i tym samym przyczynią się do lepszego zrozumienia zjawiska wykluczenia społecznego i ubóstwa na poziomie lokalnym.

Projekt składał się z sześciu pakietów roboczych. Odpowiadają one sześciu głównym obszarom prowadzonych prac. Każdy pakiet składał się z grupy zadań, które były realizowane przez zespoły partnerów projektu:

1. Nowe wskaźniki i modele badania nierówności i ubóstwa ze szczególnym uwzględnieniem wykluczenia społecznego, stopnia podatności na zagrożenie ubóstwem i deprywacji;
2. Estymacja wskaźników ubóstwa i nierówności społecznych za pomocą statystyki małych obszarów;
3. Integracja danych z badania EU-SILC z danymi pochodzącymi z rejestrów administracyjnych;

4. Standaryzacja i rozwój zastosowań oprogramowania do prowadzenia szacunków z dziedziny warunków życia;

5. Zarządzanie;

6. Informacja i upowszechnianie wyników.

Z punktu widzenia zastosowania metodologii estymacji pośredniej istotne są rezultaty prac w ramach drugiego pakietu roboczego w obszarze:

- metodologii szacowania dystrybuanty dochodu w małym obszarze,
- zastosowania modeli przestrzennych do szacowania ubóstwa,
- zastosowania modeli czasowych do szacowania ubóstwa,
- zastosowania modeli przestrzenno-czasowych do szacowania ubóstwa.

2.2.3 AMELI

Jedną z przesłanek uruchomienia projektu AMELI (**Advanced Methodology Laeken Indicators**) stanowiła próba zaspokojenia oczekiwań na wiarygodne i wysokiej jakości oszacowania wskaźników statystycznych z obszaru spójności społecznej. Projekt rozpoczął się w kwietniu 2008 roku i trwał do marca 2011 roku. Głównym instytucjonalnym koordynatorem projektu był University of Trier (Niemcy). Natomiast pozostali partnerzy to Federal Statistical Office of Germany, University of Applied Sciences Northwestern Switzerland, Swiss Federal Statistical Office, Statistics Austria, Statistics Finland, University of Helsinki, Vienna University of Technology (Austria), Statistical Office of the Republic of Slovenia oraz Statistics Estonia. W ramach prowadzonych działań dokonano przeglądu obecnie funkcjonujących wskaźników monitorujących w sposób wielowymiarowy zjawisko ubóstwa i wykluczenia społecznego. Analizowane były wskaźniki lejkenowskie (Laeken Indicators) w kontekście spójności społecznej. Szczególny nacisk położono na aspekty metodologiczne związane z zagadnieniami jakości wskaźników oraz ich własnościom statystycznym i matematycznym w kontekście oczekiwań użytkowników informacji statystycznej. Prace obejmowały, między innymi, metody pomiaru jakości uzyskiwanych oszacowań oraz podejść stosowanych do radzenia sobie z brakami odpowiedzi oraz wartościami odstającymi. Część prowadzonych działań była poświęcona implementacji metodologii statystyki małych obszarów w celu uzyskiwania oszacowań estymowanych wskaźników dla

niższych poziomów agregacji przestrzennej, dla których ze względu na niewielką próbę nie można zastosować estymacji bezpośredniej. Projekt składał się z dziesięciu pakietów roboczych:

1. Wskaźniki lejkenowskie,
2. Estymacja,
3. Estymacja wariancji,
4. Odporność,
5. Jakość danych,
6. Symulacje,
7. Analizy,
8. Wizualizacja,
9. Wsparcie dla polityki,
10. Raport końcowy.

W ramach drugiego pakietu roboczego - Estymacja zastosowano, między innymi, cztery estymatory statystyki małych obszarów do estymacji wskaźnika ubóstwa tj. estymator typu GREG, DRE, SYN, EBLUB.

2.2.4 ESSnet on Small Area Estimation

Głównym celem projektu **ESSnet on Small Area Estimation** było zrewidowanie metod i doświadczeń z zakresu statystyki małych obszarów poprzez kwerendę obecnego stanu wiedzy oraz przegląd metod oceny jakości estymatorów pośrednich. W szczególności projekt nakierowany był na dostarczenie użytecznych narzędzi i wytycznych, które ułatwią w sposób możliwie przystępny na dokonywanie estymacji oraz rozpowszechnianie wiedzy i zdobytych doświadczeń wśród narodowych instytutów statystycznych na temat statystyki małych obszarów. Działania podjęte w ramach projektu miały zachęcić do korzystania z metod statystyki małych obszarów do produkcji danych statystycznych przez narodowe urzędy statystyczne.

W projekcie zaangażowane były urzędy statystyczne ośmiu krajów Unii Europejskiej (Narodowy Instytut Statystyczny Włoch — ISTAT – pełnił rolę koordynatora całego projektu):

- Francji - Institut National de la Statistique et des Etudes Economiques, INSEE,
- Hiszpanii - Instituto Nacional de Estadistica de Espana, INE,
- Holandii - Centraal Bureau voor de Statistiek, CBS,
- Niemiec - Statistisches Bundesamt, DESTATIS,
- Norwegii - Statistisk Sentralbyra, SSB,
- Polski - Główny Urząd Statystyczny, GUS,
- Wielkiej Brytanii - Office for national Statistics, ONS,
- Włoch – Istituto Nazionale di Statistica, ISTAT,
- a także Szwajcarii posiadającej status partnera w tym projekcie.

Projekt realizowany był w latach 2010–2012. Podzielono go na 7 pakietów roboczych (Work Packages):

- **Zarządzanie projektem** - realizowane w nim prace związane były z administrowaniem całością projektu ze strony włoskiego koordynatora,
- **Analiza obecnego stanu wiedzy w zakresie metodologii statystyki małych obszarów** - przegląd dorobku w zakresie metodologii statystyki małych obszarów od momentu zakończenia projektu EURAREA, zaktualizowanie przeglądu literatury w zakresie statystyki małych obszarów, opis obecnych zastosowań statystyki małych obszarów w krajach europejskich i pozaeuropejskich, utworzenie bazy wiedzy na temat wykorzystywanych metod w zakresie statystyki małych obszarów.
- **Ocena jakości estymacji** - badania nad metodologią oceny jakości szacunków przy coraz mniejszych domenach (top-down assessment), przegląd metod oceny porównawczej estymatorów, przegląd metod oceny jakości estymacji oraz diagnostyki zastosowanych modeli, przegląd doświadczeń różnych krajów w zakresie wykorzystania różnych podejść do oceny jakości estymacji pośredniej.

- **Oprogramowanie** - przegląd dostępnego oprogramowania umożliwiającego estymację pośrednią ze szczególnym naciskiem na programy nie wymagające płatnych licencji (R), dostarczenie ujednoliconych procedur postepowania, rekomendacji oraz potwierdzenie użyteczności testowanych programów.

- **Analiza studium przypadku** - przeprowadzenie studium przypadku w dziedzinach zaproponowanych przez uczestników projektu;

- **Wskazówki i wytyczne** - podsumowanie wyników poprzednich pakietów roboczych, szczególnie WP3, WP4 i dostarczenie praktycznych wskazówek i wytycznych w kontekście implementacji metodologii estymacji pośredniej do produkcji danych statystycznych.

- **Transfer wiedzy i know-how** - rozpowszechnianie wiedzy z zakresu statystyki małych obszarów do krajów, które nie uczestniczyły w projekcie poprzez prowadzenie kursów i szkoleń w tym zakresie; utworzenie i prowadzenie strony internetowej związanej z tematyką estymacji pośredniej.

Szczególnie istotny był pakiet WP2, który zawierał przegląd praktycznych zastosowań metodologii statystyki małych obszarów oraz dostępnej literatury przedmiotu, w której opisano również zastosowania estymacji pośredniej do pomiaru ubóstwa. Pakiet WP6, przedstawiał z kolei zestandaryzowany ogólny opis procesu postępowania jaki należałoby przeprowadzić, aby dokonać przejścia od szacunków opartych na estymatorze bezpośrednim poprzez estymatory syntetyczne, do coraz bardziej złożonych estymatorów statystyki małych obszarów opartych na modelach. Wskazówki te można również wykorzystać w procesie produkcji danych statystycznych z obszaru ubóstwa przy wykorzystaniu metodologii statystyki małych obszarów.

2.2.5 EURAREA

Projekt EUAREA był finansowany przez Eurostat i został przeprowadzony w okresie od stycznia 2001 roku do czerwca 2004 roku w ramach Siódmego Ramowego (7PR) programu Unii Europejskiej w celu zweryfikowania podejść stosowanych w metodologii statystyki małych obszarów i jej praktycznych zastosowań. W skład konsorcjum EURAREA wchodziły urzędy statystyczne pięciu państw europejskich: Wielkiej Brytanii, Hiszpanii, Finlandii, Szwecji, Włoch oraz Polski, która była reprezentowana przez naukowców z Akademii Ekonomicznej w Poznaniu. Głównym koordynatorem projektu był Urząd

Statystyczny Wielkiej Brytanii, a ekspertami byli specjaliści w zakresie statystyki małych obszarów z całego świata. Celem konsorcjum EURAREA było wypracowanie technik estymacji oraz opracowanie odpowiedniego oprogramowania wspierającego Europejski System Statystyczny w zakresie szacowania podstawowych charakterystyk rynku pracy oraz poziomu życia ludności w skali lokalnej. Zadania projektu realizowane były w czterech grupach tematycznych:

- **Temat 1:** Pożyczanie mocy w czasie - próba odpowiedzi na pytanie, w jaki sposób najlepiej wykorzystać dane z badań z lat wcześniejszych, aby zwiększyć precyzję ocen dla bieżącego roku.
- **Temat 2:** Pożyczanie mocy w przestrzeni - zwiększenie efektywności estymacji poprzez wykorzystanie informacji z innych obszarów. Temat ten miał dwa aspekty: wykorzystanie przestrzennej korelacji reszt oraz problem estymacji dla jednostek przestrzennych różnej wielkości (tzw. korelacja ekologiczna).
- **Temat 3:** Adaptowanie metod standardowych dla złożonych schematów losowania próby - przystosowanie estymatorów stosowanych w statystyce małych obszarów do wykorzystania w przypadku stosowania złożonych schematów losowania. Wyróżniono tutaj dwa zagadnienia: wybór estymatorów i ocenę ich właściwości oraz optymalny dla statystyki małych obszarów schemat doboru próby.
- **Temat 4:** Estymacja parametrów dla małego obszaru oraz dla grup wyróżnionych w jego ramach - określenie pośredniej metody szacunku dla małego obszaru między techniką estymacji dla tylko jednej zmiennej, a techniką typu SPREE zachowującą strukturę badanej cechy.

Szczegółowo opisane estymatory i wypracowane w ramach tego projektu oprogramowanie w języku 4GL systemu SAS, ze względu na swoją uniwersalność, mogą stanowić cenne źródło wiedzy na temat wykorzystania metodologii statystyki małych obszarów, również w zakresie ubóstwa.

3

Możliwe źródła danych w estymacji poziomu ubóstwa

Obecny paradygmat statystyki małych obszarów skupia się na wykorzystaniu metod bazujących na modelu lub nim wspomaganych. W związku z tym istnieje potrzeba budowy związków regresyjnych pomiędzy wskaźnikiem ubóstwa lub dochodem, a innymi zmiennymi objaśniającymi. Rozdział ten zawiera opis potencjalnych źródeł danych które mogą być wykorzystane w estymacji poziomu ubóstwa na poziomie powiatów. Przedstawiono zarówno źródło dla zmiennej objaśnianej, jak i potencjalne zbiory danych dla zmiennych objaśniających.

3.1 Europejskie Badanie Dochodów i Warunków Życia

Europejskie Badanie Dochodów i Warunków Życia (EU-SILC) jest przeprowadzane w Polsce od 2005 roku, kiedy to zostało wdrożone przez GUS na podstawie rozporządzenia Unii Europejskiej. Celem badania jest „dostarczenie porównywalnych dla krajów Unii Europejskiej danych dotyczących warunków życia ludności" [36]. Zakresem badania objętych jest wiele cech, m.in. demograficznych, dotyczących edukacji, stanu zdrowia, warunków mieszkaniowych, aktywności ekonomicznej oraz dochodów.

EU-SILC jest badaniem gospodarstw domowych dobieranych zgodnie z teorią metody reprezentacyjnej. Co roku badanych jest około 24 000 mieszkań metodą bezpośredniego wywiadu. Zebrane w ten sposób dane pozwalają na publikowanie rezultatów w takich przekrojach terytorialnych jak kraj oraz regiony (NUTS 1), a także w wybranych przekrojach przedmiotowych, m.in. grupy wieku czy grupy społeczno-ekonomiczne.

Europejskie Badanie Dochodów i Warunków Życia stanowi cenne źródło informacji o skali ubóstwa w Polsce. Na podstawie deklarowanych dochodów po uwzględnieniu transferów społecznych oblicza się ekwiwalentny dochód gospodarstwa domowego[1]. Następnie wyznacza się medianę tych dochodów i określa tzw. granicę ubóstwa czyli próg dochodów poniżej których gospodarstwo domowe uznawane jest za ubogie. W badaniu EU-SILC granicę ubóstwa stanowi 60% mediany dochodów ekwiwalentnych.

W pracach nad estymacją stopy ubóstwa na poziomie powiatów, ekwiwalenty dochód gospodarstw domowych obliczony w ramach badania EU-SILC będzie pełnić rolę zmiennej objaśnianej. Aby uniknąć błędów związanych z doborem próby zmienne objaśniające zostaną wybrane z innych badań: spisów powszechnych oraz rejestrów administracyjnych. Źródła potencjalnych zmiennych zostaną omówione w kolejnych punktach.

3.2 Spisy powszechne

Najpopularniejszym źródłem zmiennych objaśniających są dane pochodzące ze spisów powszechnych. Badanie to, pokrywające całą populację, charakteryzuje się brakiem błędu losowego. Z kolei część spisu powszechnego, będąca badaniem reprezentacyjnym umożliwia dokładną estymację w bardziej szczegółowych przekrojach niż w większości przeprowadzanych badań statystyki publicznej takich jak EU-SILC czy Badanie Budżetów Gospodarstw Domowych.

3.2.1 Narodowy Spis Powszechny Ludności i Mieszkań 2002 i Powszechny Spis Rolny 2002

Narodowy Spis Powszechny Ludności i Mieszkań w 2002 r. (zwany dalej skrótowo NSP 2002) został przeprowadzony na terenie całego kraju w dniach od 21 maja do 8 czerwca 2002 roku razem z powszechnym spisem rolnym – według stanu w dniu 20 maja 2002 r. o godz. 24:00. Spisowi temu, mającemu – z wyjątkiem badania dzietności kobiet, o których niżej – charakter pełny podlegały:

- osoby stale zamieszkałe i czasowo przebywające w mieszkaniach, budynkach, obiektach i pomieszczeniach,
- mieszkania i budynki, w których znajdują się mieszkania zamieszkane

[1] uwzględniający wielkość gospodarstwa domowego oraz wiek ich członków

lub niezamieszkane oraz zamieszkane obiekty zbiorowego zakwaterowania i inne zamieszkane pomieszczenia nie będące mieszkaniami,

- osoby niemające miejsca zamieszkania.

Spis ten nie obejmował:

- szefów i cudzoziemskiego personelu przedstawicielstw dyplomatycznych i urzędów konsularnych państw obcych, członków rodzin tych osób oraz innych osób korzystających z przywilejów i immunitetów na mocy umów, ustaw lub powszechnie ustalonych zwyczajów międzynarodowych (pozostali cudzoziemcy przebywający w Polsce byli spisywani na ogólnych zasadach),
- osób ubiegających się o azyl,
- mieszkań, budynków, obiektów i pomieszczeń będących własnością przedstawicielstw dyplomatycznych i urzędów konsularnych państw obcych.

Osoby objęte spisami obowiązane były do udzielenia rachmistrzom spisowym ścisłych, wyczerpujących i zgodnych z prawdą odpowiedzi na pytania zawarte w wymienionych formularzach. Osoby prawne i inne jednostki organizacyjne wypełniły we własnym zakresie stosowne formularze i przekazywały je do właściwego terytorialnie urzędu statystycznego. Od strony technicznej w pracach spisowych (także w PSR 2002) formularze były wydrukowane na papierze i wypełniane przez rachmistrzów w sposób tradycyjny. Zgromadzone w ten sposób dane digitalizowano następnie za pomocą technologii OCR (ang. *Optical Character Recognition* — optyczne rozpoznawanie znaków) i przetwarzano już w postaci cyfrowej.

Zakres tematyczny NSP 2002 został ukształtowany w oparciu o ówczesne realia społeczno – ekonomiczne. Postępująca integracja międzynarodowa znalazła wówczas swoje odbicie m.in. w procesie dostosowywania krajowych systemów statystycznych, w tym wyników spisów, do wymogów międzynarodowych. Biuro Statystyczne ONZ, Europejska Komisja Gospodarcza oraz Unia Europejska wspólnie przygotowały propozycje tematów rekomendowanych do uwzględnienia w spisach powszechnych przeprowadzanych około 2000 r. Propozycje te brały pod uwagę znaczenie konkretnych aspektów sytuacji ludnościowej oraz potrzeby informacyjne poszczególnych krajów, a także potrzeby organizacji międzynarodowych związane z monitorowaniem i rozwijaniem polityki społecznej i regionalnej. Oczywiście każdy kraj ma także pewne tematy ściśle powiązane ze swoją specyfiką, które muszą być bezwzględnie

badane w kolejnych spisach i z tego względu powinny stanowić stały element tematyki spisowej.

Biorąc to wszystko pod uwagę NSP 2002 obejmował następujące tematy:

- Geograficzne rozmieszczenie ludności według miejsca zamieszkania oraz przebywania;
- Migracje wewnętrzne i zagraniczne ludności;
- Demograficzna charakterystyka osób: płeć, wiek, stan cywilny (formalno-prawny i faktyczny);
- Charakterystyka demograficzna gospodarstw domowych i rodzin: pozycja osób w gospo-darstwie domowym i rodzinie, wielkość i skład gospodarstwa domowego i rodziny;
- Charakterystyka społeczna osób: poziom wykształcenia oraz uczęszczanie do szkoły, kraj urodzenia, obywatelstwo, deklarowana narodowość i język używany w rozmowach w domu;
- Niepełnosprawność prawna i biologiczna;
- Aktywność ekonomiczna ludności: pracujący, bezrobotni, bierni zawodowo, pracujący w indywidualnych gospodarstwach rolnych, zawód; rodzaj działalności zakładu pracy;
- Główne i dodatkowe źródła utrzymania osób oraz pobieranie świadczeń społecznych;
- Źródła utrzymania gospodarstwa domowego; samodzielność gospodarowania;
- Gospodarstwa zbiorowe i rodziny w tych gospodarstwach;
- Mieszkania zamieszkane i niezamieszkane: stan zasobów mieszkaniowych;
- Wielkość mieszkań i ich wyposażenie;
- Samodzielność zamieszkiwania;
- Charakterystyka budynków.

Tematyka ludnościowa została wzbogacona dzięki przeprowadzeniu dwóch badań towarzyszących spisowi, a mianowicie: badania dzietności oraz badania długookresowych migracji ludności, jakie miały miejsce w latach 1989–2002.

Badanie dzietności kobiet zostało przeprowadzone metodą reprezentacyjną w sposób ankietowy na próbie prawie 350 tys. kobiet, będących w wieku powyżej 16 lat (bez względu na ich stan cywilny), czyli ok. 2% całej populacji kobiet w wieku 16 i więcej lat oraz 3,5% kobiet w wieku rozrodczym (15–49 lat). Z uwagi na wrażliwość poruszanych zagadnień, ankietowe badania dzietności są z założenia badaniami, w których uczestnictwo jest dobrowolne.

Badaniem długookresowych migracji ludności w latach 1989–2002 zostało objętych prawie 4 mln osób, które zmieniały w latach 1989–2002 miejsce zamieszkania na pobyt stały lub na okres co najmniej 12 miesięcy. Badanie to objęło migracje wewnętrzne oraz migracje zagraniczne ludności. Uzyskane w spisie informacje o migracjach pozwoliły określić faktyczne rozmiary przemieszczeń, ich zasięg przestrzenny i główne kierunki mobilności ludności w latach 90-tych. Ustalenie rozmiarów takiej kategorii migracji i liczby osób migrujących w bieżących badaniach jest niezwykle trudne z uwagi na stosunkowo skromny zakres dostępnych informacji.

Powszechnym Spisem Rolnym w 2002 r. (zwanym dalej PSR 2002) objęto:

- gospodarstwa indywidualne o powierzchni użytków rolnych powyżej 1 ha,
- gospodarstwa indywidualne o powierzchni użytków rolnych od 0,1 do 1 ha włącznie,
- osoby fizyczne będące właścicielami zwierząt gospodarskich, nieposiadające użytków rolnych lub posiadające użytki rolne o powierzchni mniejszej niż 0,1 ha,
- pozostałe gospodarstwa rolne będące w użytkowaniu osób prawnych i jednostek organizacyjnych niemających osobowości prawnej.

Wykaz gospodarstw indywidualnych powstał w oparciu o statystyczny rejestr podatkowy (system SPGC), który aktualizowany był przez rachmistrzów w czasie obchodu przedspisowego oraz uzupełniany o wykaz właścicieli zwierząt gospodarskich nieposiadających użytków rolnych lub posiadających użytki rolne o powierzchni mniejszej niż 0,1 ha. Kartotekę gospodarstw rolnych będących w użytkowaniu osób prawnych i jednostek organizacyjnych niemających osobowości prawnej utworzono na podstawie Bazy Jednostek Statystycznych GUS.

Dla gospodarstw rolnych, których użytkownicy odmówili udziału w Powszechnym Spisie Rolnym 2002 r. oraz w przypadkach, gdy kontakt z użytkownikami gospodarstw był niemożliwy, dokonano imputacji danych o gospodarstwie.

Tematyka Powszechnego Spisu Rolnego 2002 r. została ujęta na formularzach spisowych w następujących działach:

- Powierzchnia gospodarstwa;
- Struktura własnościowa użytków rolnych gospodarstwa;
- Struktura dochodów;
- Działalność gospodarcza;
- Pracujący w gospodarstwie rolnym;
- Powierzchnia zasiewów;
- Powierzchnia inna;
- Pogłowie zwierząt gospodarskich;
- Rozdysponowanie produkcji rolniczej;
- Infrastruktura gospodarstwa;
- Budynki i budowle;
- Magazynowanie w gospodarstwie;
- Nawozy i pestycydy w gospodarstwie;
- Maszyny i urządzenia rolnicze;
- Wybrane wydatki w gospodarstwie.

W ramach spisu w gospodarstwach indywidualnych o powierzchni użytków rolnych powyżej 1 ha zebrano następujące informacje:

1) dane o osobach będących użytkownikami gospodarstw rolnych:
   a) nazwisko i imiona użytkownika gospodarstwa rolnego,
   b) poziom wykształcenia rolniczego osoby kierującej,
   c) wkład pracy w gospodarstwo rolne w okresie 12 miesięcy poprzedzających badanie — liczba godzin przepracowanych w gospodarstwie,

2) o liczbie pracowników najemnych stałych i pracowników dorywczych zatrudnionych w gospodarstwie rolnym,

3) o użytkowaniu gruntów, a w szczególności:

   a) o powierzchni gruntów ogółem, w tym użytków rolnych (gruntów ornych, sadów, łąk i pastwisk trwałych), lasów i gruntów leśnych oraz pozostałych gruntów,

   b) o powierzchni zasiewów głównych upraw,

4) o pogłowiu zwierząt gospodarskich według gatunków i grup produkcyjno-użytkowych oraz liczbie pni pszczelich,

5) o rozdysponowaniu produkcji rolniczej,

6) o budynkach, infrastrukturze i wyposażeniu technicznym gospodarstw, w tym:

   a) rodzaje budynków, budowli i ich powierzchnia,

   b) liczba ciągników rolniczych i innych środków transportowych oraz maszyn rolniczych,

   c) źródła zaopatrzenia w wodę, sposoby odprowadzania ścieków i usuwania śmieci,

   d) wyposażenie w sieć elektryczną i telefon,

   e) rodzaje urządzeń melioracyjnych,

7) o stosowaniu nawozów i pestycydów w gospodarstwie w okresie 12 miesięcy poprzedzających badanie,

8) o zadłużeniu gospodarstw rolnych,

9) o działalności gospodarczej (rolniczej i pozarolniczej) prowadzonej przez użytkownika gospodarstwa rolnego lub osobę dorosłą pozostającą z użytkownikiem we wspólnym gospodarstwie domowym,

10) o ważniejszych wydatkach poniesionych w okresie 12 miesięcy poprzedzających badanie, w tym na:

    a) zakup gruntów,

    b) budowę lub modernizację budynków,

    c) powiększenie stada podstawowego,

d) zakup ciągników rolniczych i innych środków transportowych oraz maszyn rolniczych.

Informacje o wieku i płci użytkownika, a także osobach pracujących wyłącznie lub głównie w gospodarstwach rolnych pozyskano z Narodowego Spisu Powszechnego Ludności i Mieszkań, który, jak wspomniano, odbył się jednocześnie z Powszechnym Spisem Rolnym 2002 r.

Dla osób fizycznych użytkujących gospodarstwa indywidualne o powierzchni użytków rolnych od 0,1 do 1 ha włącznie oraz dla właścicieli zwierząt gospodarskich zebrano w/w informacje, z wyłączeniem tych, o których mowa w pkt. 7, 6e, 8 i 10, a w ramach punktów 3b i 6b zebrano dane w ograniczonym zakresie tematycznym. W ramach spisu u osób prawnych i w jednostkach organizacyjnych niemających osobowości prawnej zebrano informacje, o których mowa w punktach 2–10.

Podsumowując, należy stwierdzić, że dzięki swemu pełnemu charakterowi oba spisy powszechne stanowią dobre źródło potencjalnych zmiennych pomocniczych do estymacji rozpatrywanego wskaźnika ubóstwa. Należy jednak wziąć przy tym pod uwagę fakt, że uwagi na znaczniejszy upływ czasu od przeprowadzania tychże spisów, zakres wykorzystania tych danych jest ograniczony. Oznacza to, że takie informacje przydatne byłyby przede wszystkim do odzwierciedlenia czynników retrospektywnych lub uwzględnienia roli opóźnień określonych regresorów w modelach ekonometrycznych (niektóre zjawiska mogą oddziaływać na inne z dużym 'poślizgiem' czasowym).

3.2.2 Narodowy Spis Powszechny Ludności i Mieszkań 2011

Narodowy Spis Powszechny Ludności i Mieszkań 2011 (zwany dalej skrótowo NSP 2011) został przeprowadzony metodą mieszaną, tzn. dane dla spisu 2011 były pozyskiwane ze źródeł administracyjnych — rejestrów i systemów informacyjnych (przede wszystkim wykorzystane do przygotowania i aktualizacji wykazu adresowo-mieszkaniowego oraz do utworzenia operatu adresowo-mieszkaniowego) oraz zbierane bezpośrednio od ludności w ramach badania reprezentacyjnego oraz tzw. badania pełnego. Oprócz tego przeprowadzone zostały dwa pełne badania, obejmujące osoby przebywające w obiektach zbiorowego zakwaterowania oraz osoby bezdomne. Zastosowane rozwiązania miały przede wszystkim zmniejszyć koszty spisu oraz obciążenie osób objętych spisem, przy jednoczesnym zachowaniu dobrej jakości wyników spisu.

W ustawie o NSP 2011 przyjęte zostało założenie jak najszerszego wy-

korzystania systemów informacyjnych administracji publicznej, jako źródeł danych dla potrzeb spisu, co w konsekwencji oznaczało, że informacje przewidziane do zebrania w trakcie spisu pobrane zostały przede wszystkim z dostępnych źródeł administracyjnych, a następnie wykorzystane do przygotowania i aktualizacji wykazu adresowo-mieszkaniowego oraz do utworzenia operatu adresowo-mieszkaniowego do losowania próby do badania reprezentacyjnego, a także jako bezpośrednie źródło danych spisowych.

W badaniu pełnym realizowanym drogą internetową oraz poprzez wygenerowanie informacji dostępnych w źródłach administracyjnych pozyskano podstawowe dane demograficzno-społeczne i adresowe osób, które nie zostały objęte badaniem reprezentacyjnym lub spisem w obiektach zbiorowego zakwaterowania lub badaniem bezdomnych.

Przeprowadzone w ramach NSP 2011 badanie reprezentacyjne dostarczyło danych, których nie można było pozyskać z rejestrów i systemów informacyjnych. Badanie to zostało przeprowadzone na próbie losowej ok. 20% mieszkań w skali kraju. Jednostką losowania było mieszkanie, a dokładniej — jego adres. Zbiór mieszkań, który stanowił podstawę do losowania próby został przygotowany w postaci odpowiedniego operatu losowania z „głębokim" warstwowaniem. Z uwagi na fakt, że przyjęta została zasada jednostopniowego losowania mieszkań, zastosowany schemat losowania oraz alokację próby w poszczególnych powiatach (we wszystkich poprzednich spisach w badaniach reprezentacyjnych towarzyszących tym spisom stosowano losowanie dwustopniowe) —- wspomniany operat wymagał szczególnego przygotowania.

W efekcie do badania reprezentacyjnego wylosowano ponad 2 744 tys. mieszkań spośród prawie 13,5 mln mieszkań znajdujących się w operacie losowania. Mieszkania były losowane z każdej z prawie 70,5 tys. warstw, zaś wielkość próby w poszczególnych warstwach wahała się od niemal 6% do ponad 49%.

Zakres tematyczny badania reprezentacyjnego w NSP 2011 uwzględniał sześć dużych obszarów tematycznych:

- ludność i jej charakterystyka demograficzno-społeczna,
- aktywność ekonomiczna,
- migracje wewnętrzne i zagraniczne ludności,
- narodowość i wyznanie,
- gospodarstwa domowe i rodziny
- budynki i mieszkania.

W ramach tych obszarów można wyróżnić 15 tematów badawczych. Formularz długi, o szerokim zakresie tematycznym, zawierał ponad 120 pytań. Respondenci odpowiadali przeciętnie na 70–80 pytań, w zależności od płci, wieku respondenta, jego mobilności i aktywności zawodowej.

Badanie reprezentacyjne w zdecydowanej większości zostało przeprowadzone metodą bezpośredniego wywiadu rachmistrzów z mieszkańcami wylosowanego mieszkania (metoda CAPI), ale respondenci mogli także spisać się sami przez Internet — z takiej możliwości skorzystało ok. 2% osób.

Do badania reprezentacyjnego stosowany był formularz, o szerokim zakresie tematycznym z dużą liczbą pytań (ponad 120 pytań), przygotowany w wersji aplikacji na urządzenia przenośne typu handheld oraz aplikacji internetowej w trybie on-line.

Na podstawie danych z badania reprezentacyjnego można dokonywać oszacowań stosownych wielkości dla całej populacji wykorzystując wagi kalibracyjne przygotowane specjalnie do tego celu. Oczywiście, najlepsze w kontekście wsparcia estymacji dla małych obszarów wydają się dane ze źródeł administracyjnych, gdyż zebrano je dla wszystkich osób. Jednakże występują tutaj dwa problemy. Po pierwsze, duża część bardzo istotnych z punktu widzenia celów naszej analizy danych pochodzi z badania reprezentacyjnego (a więc ich dostępność ograniczona jest do relatywnie niezbyt dużej części populacji) co może generować dodatkowe obciążenie wyników. Po drugie, w przypadku zmiennych, dla których dane pozyskiwano z administracyjnych źródeł danych weryfikację jakościową (tzw. „czyszczenie") przeprowadzono tylko dla rekordów wylosowanych do badania reprezentacyjnego.

3.2.3 Powszechny Spis Rolny 2010

Powszechny Spis Rolny przeprowadzono w 2010 roku przy zastosowaniu nowoczesnej technologii gromadzenia danych statystycznych, wychodzącej naprzeciw współczesnym oczekiwaniom ich użytkowników. W odróżnieniu od rozwiązań spisowych stosowanych w latach wcześniejszych, podstawowym źródłem informacyjnym stały się tutaj zasoby administracyjne. Informacje zawarte w tychże bazach posłużyły do identyfikacji oraz określenia kluczowych cech jednostek objętych spisem. Pozyskiwaniu danych z rejestrów i ewidencji towarzyszyły badania uzupełniające. Ze względu na konieczność ograniczenia kosztów Powszechny Spis Rolny w 2010 r. (oznaczany dalej skrótem PSR 2010) został przeprowadzony jako:

1. badanie pełne w gospodarstwach rolnych:

(a) osób fizycznych o powierzchni użytków rolnych wynoszącej:

i. co najmniej 1 ha,

ii. poniżej 1 ha spełniających następujące progi fizyczne: 0,5 ha dla plantacji drzew owocowych, 0,5 ha dla plantacji krzewów owocowych, 0,5 ha razem dla warzyw i truskawek gruntowych, 0,5 ha dla chmielu, 0,3 ha dla szkółek sadowniczych i ozdobnych, 0,1 ha dla truskawek gruntowych, 0,1 ha Dla warzyw i truskawek pod osłonami, 0,1 ha dla tytoniu, 0,1 ha dla kwiatów i roślin ozdobnych pod osłonami, 10 sztuk dla bydła ogółem, 5 sztuk dla krów ogółem, 50 sztuk dla trzody chlewnej ogółem, 10 sztuk dla loch, 20 sztuk dla owiec ogółem, 20 sztuk dla kóz ogółem, 100 sztuk dla drobiu ogółem oraz 5 sztuk dla koni ogółem.

(b) osób prawnych i jednostek organizacyjnych niemających osobowości prawnej.

2. badanie reprezentacyjne w gospodarstwach rolnych osób fizycznych o powierzchni poniżej 1 ha użytków rolnych (innych niż określone w pkt. 1.a.ii).

Wykaz gospodarstw rolnych do spisu powstał na bazie danych administracyjnych, które przed spisem zostały zweryfikowane przez urzędy gmin oraz rachmistrzów spisowych, w trakcie obchodu przedspisowego. Badanie to przeprowadzane przez rachmistrzów spisowych w samych gospodarstwach rolnych dotyczyło w pierwszym rzędzie informacji, które nie są dostępne w powyższych rejestrach oraz uzupełnienia braków znajdujących się tam danych. Informacje zebrane w ten sposób (określane także jako dane „z natury") posłużyło również do oceny jakości baz, których gestorami są organy administracji państwowej i samorządowej jak również ułatwiło wychwycenie występujących w nich niespójności, utrudniających posługiwanie się nimi przez służby statystyki publicznej, które mają niebagatelny wpływ także na jakość statystyki małych obszarów.

Jak podają Autorzy opracowania GUS (2011), losowanie gospodarstw rolnych przeprowadzone zostało oddzielnie w każdej gminie przy wykorzystaniu schematu losowania warstwowego. Próba gospodarstw o liczebności 150 tys. (ok. 30% ogółu) rozdzielona została pomiędzy gminy w taki sposób, aby błąd względny oszacowania łącznej powierzchni użytków rolnych (tj. z uwzględnieniem gospodarstw rolnych o powierzchni powyżej 1 ha, które ujęto w badaniu pełnym) był jednakowy dla wszystkich gmin, tj. $v(\hat{y}) \leqslant 0,01497$.

W celu alokacji próby pomiędzy gminy wykorzystany został specjalny program optymalizacyjny, który jednocześnie:

- rozdzielał próbę gospodarstw pomiędzy gminy,
- w każdej gminie populację gospodarstw o powierzchni użytków rolnych poniżej 1 ha, dzielił w sposób optymalny na trzy warstwy. W efekcie ustalone zostały granice warstw, różne w zależności od gminy,
- wydzielał w danej gminie tzw. warstwę górną zawierającą gospodarstwa największe w klasie gospodarstw poniżej 1 ha. Gospodarstwa te badane były w 100%,
- dokonywał optymalnego rozdziału próby pomiędzy pozostałe warstwy.

W celu wykonania losowania jako operat wykorzystano wykaz, który obejmował wszystkie gospodarstwa (nie tylko poniżej 1 ha), w którym dla każdego gospodarstwa zapisano m.in. symbol gminy oraz powierzchnię użytków rolnych.

Dane uzyskane z Powszechnego Spisu Rolnego mogą zatem stanowić istotne źródło pomocnicze estymacji dla małych obszarów. Jednak specyfika twego spisu powoduje, ze zasób informacji, które można byłoby ewentualnie wykorzystać jako zmienne pomocnicze w szacowaniu wskaźników ubóstwa jest stosunkowo wąski i dotyczy właściwie tylko zagadnień związanych z pracującymi w rolnictwie, ewentualnie niektórych społeczno-demograficznych cech użytkowników gospodarstw rolnych.

3.3 Źródła administracyjne

Rejestry administracyjne wykorzystywane są przede wszystkim w systemie informacji administracyjnych. Do celów statystycznych stosowane są rejestry ogólnokrajowe. Rozbudowa systemu rejestrów administracyjnych może w przyszłości prowadzić do ograniczenia ilości pracochłonnych, czasochłonnych i kosztownych badań statystycznych. Należy jednak zaznaczyć, że rejestry administracyjne nie mogą w pełni zastąpić bezpośredniego gromadzenia danych z badań - są to metody uzupełniające się. Dane z rejestrów można pobierać w dowolnym momencie, są one zatem bardziej aktualne niż dane pochodzące ze spisów przeprowadzanych w około 10-letnich interwałach czasowych [91].

Należy podkreślić, że zbiory danych utworzone z rejestrów administracyjnych opisują pojedyncze zagadnienia, na przykład bezrobocie rejestrowane,

czy działalność podmiotów gospodarczych, nie dając możliwości dokonania wielowymiarowych szacunków obejmujących różnorodne relacje i zależności w funkcjonowaniu społeczeństwa, gospodarki i państwa jako całości. Dodatkowo często definicje cech zawartych w rejestrach mogą różnić się od przyjętych w systemie statystyki publicznej. Wszystkie dane w rejestrach muszą być zgodne i nie mogą zawierać błędów jednak zdarzają się w nich nieścisłości. Powodują one, że często jakość wyników na podstawie rejestrów jest gorsza niż w badaniach tradycyjnych, na przykład przez różne zapisy kodów pocztowych, czy brak numerów PESEL czy NIP w rekordach. W przeciwieństwie do badań statystycznych nie jest prowadzony rachunek błędów.

Integracja danych administracyjnych i badań reprezentacyjnych, przy zachowaniu odpowiednich procedur i metod może prowadzić do utworzenia zbioru o szerokim spektrum informacyjnym i wysokim pokryciu. Koszty pozyskania informacji z takich zbiorów są dużo niższe niż w przypadku klasycznych badań statystycznych, nie występuje dodatkowe obciążenie respondentów, a zbiory mogą charakteryzować się wysoką jakością [85].

3.3.1 Powszechny Elektroniczny System Ewidencji Ludności

Gestorem systemu jest Ministerstwo Spraw Wewnętrznych. Informacje z Powszechnego Elektronicznego Systemu Ewidencji Ludności, w skrócie PESEL, w latach pomiędzy spisami powszechnymi mogą przybliżyć stan oraz migracje ludności w danym obszarze [34].

Podczas estymacji ubóstwa na poziomie powiatów z systemu PESEL mogą być przydatne informacje dotyczące wieku i płci osób (w szczególności liczby ludności w określonych grupach wiekowych), liczbie ludności według płci, miejsca zameldowania na pobyt stały lub czasowy, wymeldowania z miejsca pobytu stałego, kraju pochodzenia (dla cudzoziemców).

W Banku Danych Lokalnych znajdują się potencjalnie użyteczne informacje na poziomie gmin dla roku 2009 opracowane na podstawie PESEL: ludność w grupach wieku oraz ludność według płci.

3.3.2 POLTAX

Międzynarodowym skrótem POLTAX określa się zasoby baz danych dotyczących podatników będących obywatelami polskimi i płacących podatki od dochodów pochodzących z różnych źródeł przewidziane odpowiednimi przepisami prawa jako dochód budżetu państwa polskiego i – pośrednio – w części

także budżetów jednostek polskiego samorządu terytorialnego oraz realizacji przez nich obowiązku podatkowego. Gestorem tych baz jest Ministerstwo Finansów Rzeczypospolitej Polskiej. Z punktu widzenia badania ubóstwa podstawowym i najważniejszym elementem tego systemu jest Centralny Rejestr Podmiotów - Krajowa Ewidencja Podatników (CRP KEP) stworzona w formie elektronicznej dla usprawnienia i spełnienia obowiązków wynikających z przepisów prawa oraz wykonywania przez administrację podatkową określonych prawem zadań realizowanych dla dobra publicznego. CRP KEP funkcjonuje w oparciu o obowiązek ewidencyjny nałożony stosownymi przepisami prawa na osoby fizyczne, osoby prawne oraz jednostki organizacyjne niemające osobowości prawnej, które na podstawie odrębnych ustaw są podatnikami, inne podmioty, które na podstawie odrębnych ustaw są podatnikami, płatników podatków oraz podmioty będące, na podstawie odrębnych ustaw, płatnikami składek ubezpieczeniowych.

Dokument [34] precyzuje, że zakres informacyjny bazy obejmuje osoby fizyczne nieprowadzące działalności gospodarczej lub niebędące zarejestrowanymi podatnikami podatku od towarów i usług objęte rejestrem PESEL, dla których identyfikatorem podatkowym jest numer PESEL - nie dokonują zgłoszenia identyfikacyjnego. Dane przekazywane są z rejestru PESEL. Identyfikatorem podatkowym NIP objęte są pozostałe podmioty, które podlegają obowiązkowi ewidencyjnemu, tj. osoby fizyczne, osoby prawne oraz jednostki organizacyjne niemające osobowości prawnej, które na podstawie odrębnych ustaw są podatnikami, płatnikami podatków, płatnikami składek ubezpieczeniowych i zdrowotnych: W przypadku spółek cywilnych, osobowych spółek handlowych i podmiotów podlegających wpisowi do rejestru przedsiębiorców na zasadach określonych dla spółek osobowych są to dane dotyczące wspólników, w tym również identyfikator podatkowy poszczególnych wspólników.

Z punktu widzenia analizy ubóstwa szczególnie interesujące są dane z bazy obejmującej podatników podatku dochodowego od osób fizycznych (PIT). Zaliczki na podatek dochodu pobierane przez płatników wpływają w sposób ciągły, ale końcowe rozliczenia podatkowe przekazywane są przez płatników i podatników jednorazowo w ciągu roku. Płatnik przekazuje informacje od odprowadzonych zaliczkach na podatek dochodowy do końca lutego roku następującego po roku podatkowym, zaś podatnik jest zobowiązany złożyć, zeznanie podatkowe do końca kwietnia owego roku następnego w stosunku do roku podatkowego. Tak więc dane możliwe do uzyskania mają charakter roczny i obejmują następujące cechy:

1. Stan i charakterystyka demograficzna ludności

- charakter przebywania
- czas przebywania
- data urodzenia/wiek
- płeć

2. Aktywność ekonomiczna

   - Pełna nazwa i adres siedziby lub miejsca prowadzenia działalności gospodarczej lub pracodawcy(-ów)
   - stan zatrudnienia we własne firmie lub u pracodawcy(-ów) (wspólnicy, pracownicy, członkowie rodzin)

3. Charakterystyka społeczno -ekonomiczna ludności

   - rodzaj źródła utrzymania (praca, emerytura, renta, działalność gospodarcza, najmu, inne źródła
   - dochody podatnika z różnych źródeł
   - wysokość kosztów uzyskania przychodu (w tym fakt korzystania z ich zwiększenia z tytułu zamieszkiwania poza miejscowością, w której zlokalizowane jest jego miejsce pracy)
   - wysokość wpłaconego podatku.

Właśnie informacje o dochodach ludności mają w przypadku badania ubóstwa —- i nie tylko — kolosalne znaczenie. Uwidoczniło się ono już w przypadku prac w ramach paneuropejskiego programu statystycznego monitoringu miast URBAN AUDIT, gdzie zasoby POLTAX stanowiły podstawę estymacji kierunków i natężenia dojazdów do pracy, które to dane były podstawą wyznaczania szerszych stref miejskich (ang. *Large Urban Zones* —- LUZ), odzwierciedlających zasięg funkcjonalnego oddziaływania miast. Szczególne znaczenie miały wówczas:

- informacja o dochodach oraz pobranych zaliczkach na podatek dochodowy (PIT-11/8B) zawierająca dane na temat osób zatrudnionych, które przekazywały zaliczki za pośrednictwem swoich pracodawców, lecz finalnego rozliczenia rocznego dokonywały samodzielnie (np. ze względu na korzystanie z prawa do ulg podatkowych);

- roczne obliczenie podatku od dochodu uzyskanego przez podatnika w roku dochodowym (PIT-40) przygotowanego dla osób zatrudnionych, które wszystkich operacji skarbowych dokonywały za pośrednictwem

swoich pracodawców oraz dla podatników będących emerytami bądź rencistami.

Otrzymane dane pozwoliły na przeprowadzenie w Ośrodku Statystyki Miast Urzędu Statystycznego w Poznaniu prac studialnych, które zaowocowały utworzeniem unikatowego zbioru podatników, dla których źródło przychodu w 2006 r. stanowiła praca najemna. Z tej bazy wyodrębniono podzbiór osób, dla których gmina zamieszkania różniła się od gminy lokalizacji miejsca pracy i równocześnie informacja o zwiększonych kosztach uzyskania przychodu oraz ich wysokość odpowiadały odpowiednim stawkom przysługującym osobie dojeżdżającej do pracy w 2006 r. (w myśl stosownych przepisów Ministerstwa Finansów)Weryfikacja otrzymanej bazy danych (m.in. w drodze eliminacji rekordów z powtarzającymi się numerami NIP podatnika, łączenia zbiorów, analizy dokumentów z punktu widzenia charakteru pracy oraz jej odległości od miejsca zamieszkania) i konieczne obliczenia przetestowane (zob. [72], [47], [31]). Podobne doświadczenie miało miejsce w czasie NSP 2011.

Zasoby POLTAX miały także inne zastosowanie w programie URBAN AUDIT. Wymagane tam były bowiem również informacje ilustrujące zróżnicowanie dochodów gospodarstw domowych (mediana i kwintyle dochodów, liczba gospodarstw osiągających dochód poniżej połowy średniej krajowej itp.). i tutaj przeprowadzono stosowne oszacowania oparte na wysokości zadeklarowanego w zeznaniach PIT dochodów podatników.

W kontekście badania ubóstwa warto byłoby rozważyć szersze wykorzystanie rejestru podatkowego POLTAX w połączeniu z innymi bazami. Dobry przykład stanowi tutaj szeroka analiza, jaka była możliwa np. w Holandii. Tam bowiem (zob. [94]) podatnicy mieszkający pod tym samym adresem byli określani mianem „jądrowych" (ang. *nucleus*) członków gospodarstwa domowego. Do nich dołączano następnie inne osoby również zamieszkałe pod tym samym adresem, ale niebędące podatnikami (na podstawie ewidencji ludności lub wyników spisu powszechnego). Tym sposobem można było poczynić wiele obserwacji statystycznych w zakresie sytuacji gospodarstw domowych (mediana dochodów, przepływy kapitału ludzkiego, źródła utrzymania itp.), kluczowe nie tylko dla statystyki miast, ale i (a może przede wszystkim - dla statystyki ubóstwa. Jak wiadomo, koszt przeprowadzania spisów powszechnych jest bardzo duży, zaś ich częstotliwość mała (średnio odbywają się raz na 10 lat). Tak więc regularne wykorzystanie tak cennego dla analizy ubóstwa źródła administracyjnego jakim jest POLTAX w wielu przypadkach wydaje się być szczególnie efektywne i bezpieczne (informacje byłyby przekazywane

w sposób ściśle chroniony bezpośrednio do służb statystycznych). Pozytywne doświadczenia wyniesione ze spisów powszechnych w tym zakresie ([97]) są tutaj szczególne istotnym argumentem przemawiającym za stałością kompleksowego i pełnego korzystania z tych danych przez służby statystyki publicznej. A takie informacje jak powyżej przedstawione mogą być naprawdę kluczowymi zmiennymi pomocniczymi w estymacji zagrożenia ubóstwem dla małych obszarów [34].

3.3.3 SyriuszStd

Jak podano w publikacji *Ubóstwo w Polsce w świetle badań GUS* [35], czynnikiem decydującym o statusie społecznym, w tym o sytuacji materialnej jednostki i jej rodziny jest miejsce zajmowane na rynku pracy. Ubóstwem zagrożone są przede wszystkim osoby bezrobotne i rodziny osób bezrobotnych.

System SyriuszStd prowadzony jest przez Ministerstwo Pracy i Polityki Społecznej. W systemie zbierane są informacje dotyczące klientów RP zgłaszających się do jednostki PUP, m.in. dane, obywatelstwo, adresy (zameldowania, zamieszkania, korespondencji), okresy mające wpływ na przyznanie prawa do zasiłku (okresy zatrudnienia, urlopy wychowawcze, służba wojskowa), wykształcenie, zawody, uprawnienia i umiejętności oraz stopień niepełnosprawności. Osoba rejestrowana podaje również informacje dotyczące członków swojej rodziny, m.in. stopień pokrewieństwa, status członka rodziny na rynku pracy, stopień niepełnosprawności oraz czy jest zgłaszany do ubezpieczenia zdrowotnego.

Dane generowane są z systemu w formie sprawozdań MPiPS-01 wraz z załącznikami, MPiPS-02, MPiPS-06 oraz MPiPS-07 przez powiatowe i wojewódzkie urzędy pracy, a następnie zgodnie z Programem badań statystycznych statystyki publicznej na dany rok przekazywane do Urzędu Statystycznego w Bydgoszczy. Na ich podstawie publikowane są dane dot. osób bezrobotnych i poszukujących pracy zarejestrowanych w urzędach pracy. Bezrobocie jest niezwykle ważnym czynnikiem zróżnicowania ubóstwa, a bezrobocie rejestrowane może być jego przybliżeniem [34].

W Banku Danych Lokalnych można znaleźć wiele informacji dotyczących bezrobocia rejestrowanego. Udostępniane są dane na poziomie powiatów o osobach bezrobotnych zarejestrowanych według płci (dostępne za lata 2003–2013, dane na poziomie gmin) lub wieku ogółem i w podziale na płeć (2000–2013) oraz według poziomu wykształcenia ogółem (2000–2013) i w podziale na płeć (2004–2013). Potencjalnie użytecznymi danymi w BDL jest liczba bezrobotnych rejestrowanych pozostających bez pracy dłużej niż rok:

ogółem (dostępne za lata 2006–2013) lub w pewnych udziałach w liczbie bezrobotnych (2003–2013), czy też w ludności aktywnej zawodowo (2006–2013). Kolejną pomocną podgrupą dostępną w BDL na poziomie powiatów są bezrobotni zarejestrowani wg czasu pozostawania bez pracy ogółem (dostępne za lata 2003–2013) i w podziale na płeć (2004–2013). Przydatne mogą się okazać także dane informujące o liczbie bezrobotnych według płci i typu, m.in. nowo zarejestrowani, wyrejestrowani, z prawem do zasiłku, dotychczas niepracujący, poprzednio pracujące (1999–2013), według stażu pracy ogółem (2003–2013) i w podziale na płeć (2004–2013).

3.3.4 POMOST

Rejestr POMOST prowadzony jest przez Ministerstwo Pracy i Polityki Społecznej. System ten obejmuje wszystkie osoby korzystające z pomocy społecznej oraz jednostki pomocy społecznej. Od 2002 roku rozwijany jest jako część szerszego systemu informatycznego SYRIUSZ, zmierzającego do stworzenia zintegrowanej bazy wiedzy o publicznym systemie usług społecznych.

W rejestrze znajdują się szczegółowe dane opisujące świadczenia, świadczeniobiorców i ich rodziny. Można znaleźć m.in. informacje dotyczące składu rodziny, źródła utrzymania, stanu cywilnego, czy pozycji na rynku pracy. Przydatnymi w kontekście modelowania ubóstwa mogą być zatem na przykład dane o świadczeniobiorcach według grup wieku lub składu rodzinnego/liczby osób w rodzinie, w szczególności osób samotnych i samotnych z dziećmi, osób o niskim wykształceniu pobierających świadczenia, osób bezrobotnych, pobierających świadczenia a utrzymujących się z emerytury lub renty lub innych niezarobkowych źródeł utrzymania [42].

3.3.5 System Informacji Oświatowej

Dysponentem Systemu Informacji Oświatowej, w skrócie SIO, jest Ministerstwo Edukacji Narodowej. Gromadzone są w nim dane dotyczące bazy materialnej placówek systemu oświaty, kosztów ich prowadzenia, zbiorcze dane o uczniach (m.in. według płci, wieku, miejsca zamieszkania, klas, rodzajów kształcenia, rodzajów zajęć w których uczestniczą), a także indywidualne dane o nauczycielach (m.in. według wieku, płci, poziomu wykształcenia, przygotowania pedagogicznego, stopnia awansu zawodowego, otrzymywanego wynagrodzenia)[34].

W Banku Danych Lokalnych można znaleźć informacje dotyczące liczebności danych typów szkół i liczbie nauczycieli w danym typie szkół (podsta-

wowe, gimnazjalne, ogólnokształcące, policealne, zasadnicze zawodowe, ponadgimnazjalne, itd.) — w większości na poziomie gmin — które mogłyby posłużyć do stworzenia zmiennych dotyczących np. ilości szkół lub nauczycieli przypadających na osobę mieszkającą w danym obszarze. Użytecznymi mogą być również wskaźniki komputeryzacji według typu szkoły (dostępne na poziomie gmin w większości za lata 2008–2013), uczniowie przypadający na 1 komputer z dostępem do Internetu przeznaczony do użytku uczniów ogółem (2008–2013) lub w podziale na miasto/wieś (2002–2013), współczynniki skolaryzacji brutto i netto dostępne na poziomie gmin dla wykształcenia podstawowego i gimnazjalnego (2003–2013) oraz ponadpodstawowego i ponadgimnazjalnego (2002–2013) .

3.3.6 Budżety Jednostek Samorządu Terytorialnego

Budżety jednostek samorządu terytorialnego, w skrócie BESTI@, to system, którego gestorem jest Ministerstwo Finansów. Obejmuje on poszczególne jednostki samorządu terytorialnego. System zawiera informacje z wykonania planu dochodów i wydatków budżetowych, nadwyżce/deficycie oraz przychodów i rozchodów jednostek samorządu terytorialnego [34].

Użytecznymi z punktu widzenia estymacji ubóstwa mogą być dostępne w Banku Danych Lokalnych dane na poziomie powiatów otrzymane na podstawie omawianego systemu dotyczące dochodów lub wydatków na jednego mieszkańca (w latach 2002–2013), a także wydatki według poszczególnych działów (m.in. turystykę, oświatę i wychowanie, kultury i dziedzictwa narodowego, ochrony zdrowia, transportu i łączności), dostępne na poziomie powiatów w latach 2001–2013. Przydatne mogą być również wskaźniki dotyczące dochodów i wydatków budżetów jednostek samorządów terytorialnych, w tym udziały wydatków na drogi publiczne w wydatkach ogółem (lata 2008–2013) oraz udział wydatków inwestycyjnych gmin i powiatów w wydatkach ogółem (2005–2013).

3.3.7 Kompleksowy System Informatyczny Zakładu Ubezpieczeń Społecznych

Źródło utrzymania jest związane z ubóstwem, a bardziej niż przeciętnie podatne na ubóstwo są gospodarstwa domowe utrzymujące się z renty, gdy główni żywiciele nie żyją lub są niezdolni do pracy w wyniku choroby lub podeszłego wieku. Jedną z grup o najniższych dochodach często są emeryci. Dodatkowo obecność osoby niepełnosprawnej w gospodarstwie domo-

wym znacznie zwiększa ryzyko zagrożenia ubóstwem [35].

Kompleksowy System Informatyczny Zakładu Ubezpieczeń Społecznych (KSI ZUS) prowadzony jest przez Zakład Ubezpieczeń Społecznych. System obejmuje swoją funkcjonalnością procesy zachodzące w ZUS oraz procesy mające miejsce w relacji ZUS z podmiotami zewnętrznymi.

System między innymi przechowuje informacje o składkach na Fundusz Ubezpieczeń społecznych, indywidualnych kontach ubezpieczonych, świadczeniach - emeryturach, rentach i zasiłkach ubezpieczenia społecznego. Znajdują się tu również informacje o nie opłacanych składkach i nienależnie pobranych świadczeniach, zwolnieniach lekarskich i procesach postępowania orzeczniczego. KSI ZUS obsługuje także dane z dokumentów ubezpieczeniowych płatników składek i dokumentów dotyczących zwolnień z tytułu choroby. KSI ZUS korzysta z rejestrów urzędowych PESEL, KE i RECON, wspomaga przekazywanie składek do OFE, NFZ i Krajowego Urzędu Pracy.

System poza udostępnianiem zagregowanych danych dotyczących statutowej działalności ZUS, pozwala na generowanie raportów i analiz danych z zakresu ewidencji płatników składek i ubezpieczonych, rozliczeń z płatnikami składek, dochodzenia należności z tytułu składek na ubezpieczenie społeczne oraz nienależnie pobranych świadczeń, kontroli płatników składek, absencji chorobowej, świadczeń emerytalno-rentowych i zasiłków, orzecznictwa i prewencji rentowej [34].

Bank Danych Lokalnych oferuje w swoich zasobach informacje na poziomie województw za lata 1999–2013 dotyczące przeciętnej liczby emerytów i rencistów ogółem oraz osób pobierających świadczenia z pozarolniczego systemu ubezpieczeń społecznych (razem/emerytury/renty z tytułu niezdolności do pracy/renty rodzinne) i osób pobierających emerytury i renty, rolników indywidualnych. Ponadto również na poziomie województw za lata 1999–2013 dostępne są przeciętne miesięczne emerytury i renty brutto dla świadczeń społecznych z pozarolniczego systemu ubezpieczeń społecznych (razem/emerytury/renty z tytułu niezdolności do pracy/renty rodzinne), świadczeń rolników indywidualnych oraz świadczeń z pozarolniczego systemu ubezpieczeń społecznych - emerytura brutto w relacji do przeciętnego miesięcznego wynagrodzenia brutto. Nie są dostępne żadne dane na poziomie powiatów.

3.3.8 Krajowy System Monitoringu Świadczeń Rodzinnych

W sferze ubóstwa warunków życia znajduje się ponad połowa gospodarstw domowych, w których głównym źródłem utrzymania były świadczenia spo-

łeczne [35], czyli między innymi świadczenia dotyczące rodziny i renty rodzinne.

Dysponentem Krajowego Systemu Monitoringu Świadczeń Rodzinnych - ZC (KSMSR-zc) oraz Krajowego Systemu Monitoringu Świadczeń Rodzinnych - SPR (KSMSR-spr) jest Ministerstwo Pracy i Polityki Społecznej. Celem KSMSR-zc jest prowadzenie monitoringu realizacji przyznawania i wypłaty świadczeń rodzinnych, a zatem przechowywane są informacje dotyczące zbiorowości osób otrzymujących świadczenia rodzinne i składów rodzin (w tym informacje o rodzajach i wysokości dochodów) oraz realizatorów świadczeń rodzinnych.

KSMSR-spr ma na celu monitoring realizacji zadań przez organy właściwe w zakresie przyznawania i wypłaty świadczeń rodzinnych, w tym także osobom uprawnionym do świadczeń rodzinnych w ramach koordynacji systemów zabezpieczenia społecznego. W związku z tym gromadzone są informacje dotyczące zbiorowości osób otrzymujących świadczenia rodzinne i składów rodzin oraz jednostki organizacyjne realizujące zadania związane z przyznaniem świadczeń rodzinnych [34].

Użytecznymi danymi otrzymanymi na podstawie rozważanych systemów mogłyby być m.in. informacje o wieku, stanie cywilnym, dochodach osób pobierających i rodzaju świadczenia rodzinnego. Znacznie szerszy zakres przydatnych z modelowaniu ubóstwa informacji dostarcza drugi z rozważanych systemów. Na jego podstawie można otrzymać dane zagregowane o liczbie rodzin pełnych/niepełnych z lub bez dziecka niepełnosprawnego pobierających świadczenia rodzinne według skategoryzowanych dochodów na osobę i liczbie rodzin pobierających świadczenia rodzinne na daną liczbę dzieci. Dodatkowo dostępne są informacje o jednostkach realizujących świadczenia m.in. wydatki i ich liczba na świadczenia z dotacji z budżetu państwa lub ze środków własnych, czy też liczba wydanych decyzji. W systemie gromadzone są również dane o świadczeniach z podziałem na rodzaje (m.in. zasiłki rodzinne w grupach wiekowych, z tytułu urodzenia dziecka, opieki w czasie urlopu wychowawczego, samotnego wychowania dziecka, rozpoczęcia roku szkolnego, zasiłki pielęgnacyjne, jednorazowe zapomogi), liczbie tych świadczeń i kwoty.

Bank Danych Lokalnych w swoich zasobach posiada informacje na poziomie gmin w latach 2008–2013, dotyczące rodzin otrzymujących zasiłki rodzinne na dzieci, dzieciach, na które rodzice otrzymują zasiłek rodzinny — ogółem i w wieku do lat 17, a także udziale dzieci w wieku do lat 17, na które rodzice otrzymują zasiłek rodzinny w ogólnej liczbie dzieci w tym wieku. Na poziomie gmin, także za lata 2008–2013, można uzyskać dane

o kwotach świadczeń rodzinnych, zasiłków rodzinnych (wraz z dodatkami) oraz zasiłków pielęgnacyjnych. Brak jest danych za rok 2005.

3.3.9 Krajowy System Monitoringu Pomocy Społecznej

W przypadku ubóstwa dochodowego wskaźnik zagrożenia ubóstwem warunków życia osiąga wysokie wartości dla rodzin wielodzietnych — w sferze ubóstwa znalazło się co piąte małżeństwo z co najmniej 3 dzieci. Do kategorii gospodarstw domowych dotkniętych ubóstwem warunków życia należą także rodziny niepełne [35]. Ze wspomnianej publikacji można wnioskować również, że zdecydowana większość zwracających się o pomoc finansową oraz korzystających z tej pomocy to gospodarstwa ubogie pod względem dochodowym i warunków życia.

Dysponentem systemu jest Ministerstwo Pracy i Polityki Społecznej. Krajowy system monitoringu i pomocy społecznej, w skrócie KSMPS, zawiera informacje o rodzinach korzystających z pomocy społecznej, kwotach wydanych na świadczenia, powodów przyznanej pomocy, typów rodzin otrzymujących świadczenia, liczbie i kwocie świadczeń oraz liczbie korzystających i formie pomocy wykorzystywanych przy realizacji programów w zakresie dożywiania i przeciwdziałania przemocy w rodzinie.

W Banku Danych Lokalnych znajdują się potencjalnie przydatne informacje na poziomie gmin opracowane przez Główny Urząd Statystyczny na podstawie KSMPS dotyczące: liczby gospodarstw i liczbie osób korzystających ze środowiskowej pomocy społecznej (dostępne za lata 2008–2012), udziału osób w gospodarstwach domowych korzystających ze środowiskowej pomocy społecznej w ludności ogółem (dostępne za lata 2008–2012), udziałów procentowych osób według grup wieku korzystających ze środowiskowej pomocy społecznej w ogólnej liczbie osób w tym wieku (dostępne za lata 2010–2013). Brak dostępnych danych dla roku 2005.

Na poziomie województw znajdują się przydatne dane dostępne za lata 2009–2013 dotyczące rzeczywistej liczby osób ogółem korzystających ze świadczeń, liczbie osób otrzymujących pomoc pieniężną razem oraz w podziale na zasiłki stałe, okresowe i celowe, a także liczbie osób otrzymujących pomoc niepieniężną razem oraz w podziale na schronienie, posiłek i ubranie. Brak dostępnych danych dla lat 2005 i 2008, a także danych na pożądanym poziomie powiatu.

Innymi potencjalnie przydatnymi zmiennymi mogłyby być dane zawierające informacje m.in. o kwotach wydanych na świadczenia, typach rodzin

objętych pomocą społeczną (według liczby osób, liczby osób i liczby dzieci), liczbie osób i liczbie rodzin, którym przyznano świadczenia według powodów przyznanej pomocy (ubóstwa, sieroctwa, bezdomności, niepełnosprawności, długotrwałej choroby, itd.), liczbie rodzin zastępczych, którym przyznano świadczenia i kwocie tych świadczeń, liczbie uczniów objętych programem dożywiania, szacunkowej liczbie dzieci i osób wymagających dożywiania, liczbie punktów prowadzących dożywianie [101].

3.3.10 Elektroniczny Krajowy System Monitoringu Orzekania o Niepełnosprawności

Jak wspomniano wcześniej, obecność osoby niepełnosprawnej w gospodarstwie domowym znacznie zwiększa ryzyko zagrożenia ubóstwem. Osoby niepełnosprawne często ponoszą większe wydatki, mając jednocześnie niższe dochody z powodu drugorzędnej pozycji na rynku pracy, zazwyczaj są bezrobotni.

System EKSMOON, czyli Elektroniczny Krajowy System Monitoringu Orzekania o Niepełnosprawności, jest prowadzony przez Ministerstwo Pracy i Polityki Społecznej. Składa się z trzech modułów: powiatowego, wojewódzkiego i centralnego. Poziom powiatowy związany jest m.in. z przyjmowaniem i rejestracją wniosków o wydanie orzeczenia o niepełnosprawności/stopniu niepełnosprawności/wskazaniach do ulg i uprawnień, rejestracją odwołań od orzeczeń, rejestracją posiedzeń składów orzekających, rejestracją legitymacji osób niepełnosprawnych i skierowań na badania diagnostyczne. Poziom wojewódzki dotyczy m.in. rejestracji i rozpatrywania odwołań od decyzji powiatowych, rejestracją informacji o posiedzeniach składów orzekających, rejestracją skierowań na badania i wyników badań specjalistycznych, wykonywaniem sprawozdawczości i informacją o przeprowadzonych kontrolach. Moduł centralny wykorzystywany jest przez Biuro Pełnomocnika Rządu do spraw Osób Niepełnosprawnych w celu nadzoru nad poprawnością orzekania, kontroli poprawności doboru składów orzekających, monitoringu w zakresie terminowości załatwiania wniosków i odwołań w zespołach powiatowych, przygotowania zestawień centralnych powiatowych w zakresie liczby i celu wniosków, sposobu załatwienia sprawy, liczby wydanych orzeczeń i legitymacji, przygotowania zestawień centralnych wojewódzkich oraz generowania zestawień centralnych dotyczących osób niepełnosprawnych.

Dane gromadzone w systemie obejmują informacje dotyczące osób (podstawowe informacje, wykształcenie i zawód, forma zatrudnienia i czasu pracy, daty i rodzaje wydanych orzeczeń, symbolu przyczyny, daty powstania oraz

stopnia niepełnosprawności), wniosków w sprawie orzeczenia o niepełnosprawności (m.in. wydane orzeczenia, złożone odwołania, skierowania na badania, składy orzekające, przeprowadzone kontrole) [34].

Bank Danych Lokalnych w swoich zasobach posiada głównie informacje o niepełnosprawności na podstawie Narodowych Spisów Powszechnych na poziomie powiatów. Pozostałe informacje o niepełnosprawnych w BDL dotyczą bezrobocia rejestrowanego i ofert pracy dla osób bezrobotnych (poziom powiatów, lata 2012–2013), a także edukacji dzieci i młodzieży ze specjalnymi potrzebami (gminy, lata 1999–2013) oraz o uczniach szkół dla dzieci i młodzieży niepełnosprawnej (gminy, 2012–2013). Na poziomie gmin za lata 2008–2012 dostępne są dane dotyczące mieszkańców niepełnosprawnych intelektualnie placówek stacjonarnej pomocy społecznej.

3.3.11 System Bank Danych Drogowych

Zagrożenie ubóstwem w znacznie większym stopniu dotyczy mieszkańców wsi niż ośrodków miejskich [35]. W ośrodkach miejskich znajduje się znacznie więcej dróg o nawierzchni twardej oraz znacznie większe natężenie ruchu niż ma to miejsce na terenach zamiejskich. Szczególnie na terenach wiejskich w których to przeważają drogi o powierzchni gruntowej.

System Bank Danych Drogowych, w skrócie BDD, prowadzony jest przez Generalną Dyrekcję Dróg Krajowych i Autostrad. Jest to system służący do ewidencjonowania, gromadzenia i zarządzania danymi dotyczącymi sieci drogowej. BDD umożliwia udostępnianie informacji o długości dróg jedno jezdniowych, dwu jezdniowych, liczbie i długości autostrad i dróg ekspresowych, liczbie pasów ruchu na danym odcinku drogi, szerokości poboczy, rodzaju nawierzchni dróg i poboczy itd. Możliwe jest również uzyskanie informacji o stanie technicznym dróg i natężeniu ruchu [34].

W Banku Danych Lokalnych na poziomie powiatów zamieszczono dane o drogach publicznych powiatowych oraz gminnych według typu nawierzchni (twarda lub gruntowa) w formie wskaźników na 100 km^2 lub 10 tys. ludności za lata 2005–2012, a także długość dróg według typu nawierzchni (twarda/twarda ulepszona/gruntowa) w kilometrach za lata 2001–2012. Niestety brak jest informacji o natężeniu ruchu czy też długości dróg jedno i dwu jezdniowych lub danych dotyczących poboczy.

3.3.12 Rejestr Cen i Wartości Nieruchomości

Nieruchomości w ośrodkach wielkomiejskich oraz podmiejskich osiągają zazwyczaj Wyższe ceny transakcyjne. Zagrożenie ubóstwem na tych obszarach jest niższe niż na terenach wiejskich, gdzie ceny nieruchomości są zwykle niższe.

Gestorem powyższego Rejestru (RCiWN) jest Główny Urząd Geodezji i Kartografii. Prowadzona baza zawiera ceny transakcyjne nieruchomości oraz rejestru wartości nieruchomości określanych przez rzeczoznawców majątkowych. Źródłem danych są akty notarialne przekazywane przez notariuszy, wyciągi z operatów szacunkowych od rzeczoznawców majątkowych, ewidencja gruntów i budynków [34].

Przydatnymi informacjami pochodzącymi z RCiWN mogłyby być na przykład średnie wartości cen transakcyjnych nieruchomości na rynku pierwotnym i wtórnym. Średnie ceny transakcyjne są wyższe w ośrodkach miejskich i podmiejskich, gdzie zagrożenie ubóstwem bywa niższe niż na pozostałych terenach. W Banku Danych Lokalnych dane takie nie są umieszczane. Kwartalną informację o cenach mieszkań i sytuacji na rynku nieruchomości od roku 2010 zamieszcza na swojej stronie Narodowy Bank Polski- opracowanie stanowi opis zmian na rynku nieruchomości. Dodatkowo dostępne są raporty roczne, podsumowujące dane lata na rynku nieruchomości oraz opisujące zmiany w największych miastach Polski.

3.3.13 Krajowy Rejestr Urzędowy Podmiotów Gospodarki Narodowej

Krajowy Rejestr Urzędowy Podmiotów Gospodarki Narodowej, czyli REGON, prowadzony jest przez Prezesa Głównego Urzędu Statystycznego. Rejestr REGON jest bieżąco aktualizowanym zbiorem informacji o podmiotach gospodarki narodowej prowadzonym w systemie informatycznym w postaci centralnej bazy danych oraz terenowych baz danych w 16 urzędach statystycznym. Wpisowi do rejestru podlegają osoby prawne, jednostki organizacyjne niemające osobowości prawnej oraz osoby fizyczne prowadzące działalność gospodarczą, w tym prowadzące indywidualne gospodarstwa rolne.

Rejestr zawiera informacje między innymi o danych adresowych firmy lub danych osobowych prowadzących działalność, numerze identyfikacji podatkowej, formie prawnej i formie własności, rodzajach prowadzonej działalności, dat związanych z działalnością, a także przewidywanej liczbie pracujących i zatrudnionych. W rejestrze REGON rozróżnia się jednostki lokalne (tj. zor-

ganizowana całość) nierolnicze i gospodarstwa rolne [32].

Użytecznymi informacjami podczas estymacji ubóstwa związanymi z rejestrem REGON mogą być zamieszczone w Banku Danych Lokalnych dane na poziomie gmin o podmiotach wg klas miejscowości (lata 2002–2013), PKD i rodzajów działalności (2009–2013), PKD i sektorów własnościowych (2009–2013). Dodatkowo udostępnione są wskaźniki do poziomu gminy za lata 2002–2013 dotyczące podmiotów, m.in. podmiotów na 1000 mieszkańców w wieku produkcyjnym, podmiotów wpisanych do rejestru REGON na 10 tys. ludności, jednostek nowo zarejestrowanych na 10 tys. ludności, podmiotów nowo zarejestrowanych na 10 tys. ludności w wieku produkcyjnym, czy też osób fizycznych prowadzących działalność na 100 tys. osób w wieku produkcyjnym. Innymi wskaźnikami są również informacje o fundacjach, stowarzyszeniach i organizacjach społecznych na 1000 lub 10 tys. mieszkańców. W BDL znajdują się także podmioty według klas wielkości na 10 tys. mieszkańców w wieku produkcyjnym.

4

Przegląd potencjalnych zmiennych pomocniczych w estymacji stopy ubóstwa

Analiza ubóstwa bez zrozumienia jego przyczyn jest niemożliwa. Te z kolei pełnią bardzo istotną rolę w statystyce małych obszarów. Charakterystyki związane ze zjawiskiem będącym przedmiotem analiz służą jako zmienne pomocnicze przy estymacji poziomu ubóstwa na docelowym poziomie terytorialnym.

Przed przystąpieniem do badania należy zdać sobie sprawę z pewnych ograniczeń. Pierwszą trudnością jest konieczność oddzielenia przyczyny od skutku. Powszechnie wiadomo, że osoby ubogie cechują się niższym wykształceniem. Niemniej powstaje pytanie czy przyczyną znalezienia się w strefie ubóstwa było niskie wykształcenie czy też ubóstwo nie pozwoliło na zdobycie odpowiedniego wykształcenia. Kierowanie się wyłącznie miarami statystycznymi przy określaniu zależności jest niewystarczające — konieczna jest dogłębna analiza potencjalnej przyczyny ubóstwa.

Po drugie, przedstawione determinanty będą wyłącznie przybliżonymi przyczynami ubóstwa. Nie sposób określić co jest rzeczywistym powodem niskiego wykształcenia członków ubogiego gospodarstwa domowego. Czy przeszkodą w edukacji były zbyt wysokie opłaty, brak szkoły w pobliżu miejsca zamieszkania czy też jakość samej edukacji. Pomiar tych bezpośrednich przyczyn jest niestety na chwilę obecną niemożliwy.

Należy także pamiętać, że nie ma podstaw, aby twierdzić, że przyczyny ubóstwa są zamkniętym zbiorem cech. Jest to zjawisko silnie zorientowane terytorialnie, więc kluczowa jest analiza przyczyn w kontekście danego kraju [37].

Determinanty ubóstwa można obserwować na czterech poziomach:

- poziom regionu,

- poziom społeczności,
- poziom gospodarstwa domowego,
- poziom osoby.

4.1 Poziom regionu

Wiele charakterystyk na poziomie regionu można powiązać ze zjawiskiem ubóstwa. W tej grupie znajdą się cechy opisujące geograficzne odosobnienie, brak zasobów naturalnych oraz niekorzystny klimat. Ponadto wskaźniki związane z jakością władz lokalnych, ekonomią, gospodarką oraz wymiarem sprawiedliwości także są obserwowane na tym poziomie.

Stopa bezrobocia — Schiller [86] wskazuje, że najbardziej przekonującą przyczyną ubóstwa jest pozostawanie bez pracy. Cecha ta jest obserwowana na poziomie powiatów w oparciu o dane pochodzące z Narodowych Spisów Powszechnych. Ponadto można wykorzystać także stopę bezrobocia rejestrowanego pochodzącą z Powiatowych Urzędów Pracy — dostępne są wówczas dane roczne (2003–2013) w dodatkowych przekrojach przedmiotowych.

Przeciętne wynagrodzenia — poziom wynagrodzeń ma duży wpływ na kształtowanie się ubóstwa w regionie. Zależą one od wielu czynników: liczby zakładów pracy, dostępności wykwalifikowanej siły roboczej itp. Dane dotyczące wartości przeciętnych wynagrodzeń pochodzą ze sprawozdawczości przedsiębiorstw i dostępne są w Banku Danych Lokalnych na poziomie powiatów dla lat 2002–2013.

Przedsiębiorczość — Ignaciuk i Kątowski [43] wskazują, że wspieranie przedsiębiorczości powinno przynieść pozytywne skutki w postaci redukcji liczby osób dotkniętych zjawiskiem ubóstwa. Rozwój przedsiębiorczości można obserwować za pomocą liczby podmiotów gospodarczych według rejestru REGON. Dane te są dostępne w Banku Danych Lokalnych za lata 2005–2013 do poziomu gmin.

Miejsce zamieszkania — liczne badania wskazują, że tereny wiejskie są dużo bardziej narażone na występowanie gospodarstw ubogich aniżeli osoby żyjące w miastach [81], [35]. Informacje o liczbie ludności według miejsca zamieszkania dostępne są w Banku Danych Lokalnych na poziomie powiatów za lata 1995–2013.

Dochody własne budżetów — zamożność jednostki samorządu terytorialnego może istotnie wpływać na poziom życia mieszkańców. Dane na temat budżetów powiatów dostępne są w Banku Danych Lokalnych dla lat 1999–2013 w dodatkowych, szczegółowych przekrojach przedmiotowych.

Dojazdy do pracy — badania przeprowadzone w USA wykazały, że osoby ubogie wydawały większą część swoich dochodów na dojazdy do pracy [83]. Oznacza to, że osoby te nie zrażały się dużą odległością do miejsca pracy i ważniejsze było dla nich stałe zatrudnienie. Dane na temat dojazdów do pracy dostępne są w Polsce za lata 2006 oraz 2011 na poziomie gmin. Zagadnienie to jest szerzej opisane w podrozdziale poświęconemu rejestrowi podatkowemu POLTAX, który był podstawą przeprowadzenia owych analiz.

4.2 Poziom społeczności

Główną determinantą ubóstwa na poziomie społeczności jest dostęp do infrastruktury. Wśród głównych charakterystyk można wymienić utwardzone drogi, dostęp do elektryczności, wodociągów, kanalizacji. Ponadto kluczowa jest odległość do ośrodków handlowych, szkół czy lokalnych ośrodków administracyjnych (urzędów).

Mieszkania wyposażone w instalacje — informacje na temat wyposażenia mieszkań w instalacje takie jak wodociąg, kanalizację, gaz z sieci, łazienkę, ustęp oraz centralne ogrzewanie. Ponadto kategorie wodociągu i ustępu dostępne są w bardziej szczegółowym podziale. Zostały zebrane w trakcie Narodowego Spisu Ludności i Mieszkań 2002 oraz 2011. Dane dostępne w Banku Danych Lokalnych na poziomie powiatów.

Drogi publiczne — wskaźniki zbudowane na podstawie długości dróg publicznych umożliwią ocenę dostępu mieszkańców lokalnej społeczności do transportu drogowego. Dane dostępne są w Banku Danych Lokalnych na poziomie powiatów za lata 2001–2013.

Szkolnictwo — dostęp do edukacji ma również istotny udział w kształtowaniu poziomu ubóstwa. Dostępność instytucji pozadomowej opieki nad dziećmi umożliwia rodzicom bilansowanie życia zawodowego i rodzinnego i wpływa na polepszenie warunków życia [109].

Władza lokalna — władze lokalne powinny skupiać swoje działania na umożliwieniu osobom ubogim dostępu do usług społecznych, gospodarczych i kulturalnych. Prowadzona w ten sposób polityka społeczna to opracowanie i wdrażanie rozwiązań administracyjnych mających na celu pomoc osobom znajdującym się w sferze ubóstwa [51]. Niemniej jest to cecha bardzo trudno mierzalna i w tym miejscu jest jedynie sygnalizowana.

Pomoc społeczna — gospodarstwa, których dochód na osobę nie przekracza progu ubóstwa socjalnego mogą skorzystać ze świadczeń pomocy społecznej. Są one udzielane w celu zastąpienia utraconego dochodu oraz w przypadku szczególnych potrzeb rodziny. W efekcie świadczenia z pomocy społecznej powinny mieć wpływ na ograniczenie ubóstwa. Niemniej analizy wskazują, że wpływ ten jest znikomy [110]. Zagregowane do poziomu gmin dane za lata 2008–2012 dotyczące beneficjentów środowiskowej pomocy społecznej oraz placówek pomocy społecznej dostępne są na stronie Banku Danych Lokalnych.

Ochrona zdrowia — ograniczony dostęp do ochrony zdrowia jest przedstawiany jako jeden z czynników ubóstwa. Nierównomierny dostęp do opieki zdrowotnej może przyczynić się do kulminacji niekorzystnych warunków życiowych [20]. Dane dotyczące ochrony zdrowia dostępne są w Banku Danych Lokalnych w następujących kategoriach: liczba łóżek w szpitalu (powiaty, 2005–2013), przychodnie (gminy, 1995–2013), kadra medyczna (powiaty, 1999–2013).

4.3 Poziom gospodarstwa domowego i osoby

Bardzo ważne determinanty ubóstwa obserwuje się na poziomie gospodarstwa domowego oraz osób je stanowiące. Są to między innymi struktura gospodarstwa domowego, płeć głowy gospodarstwa domowego, wykształcenie oraz zaangażowanie w rynek pracy członków gospodarstwa domowego.

Wielkość gospodarstwa domowego — wykazano, że gospodarstwa domowe o dużej liczbie członków są bardziej narażone na znalezienie się w strefie ubóstwa niż gospodarstwa 4-osobowe i mniejsze [33]. Związane jest to z występowaniem mniejszego dochodu na osobę w przypadku licznych gospodarstw domowych. Dane na temat liczebności gospodarstw domowych zostały zebrane podczas NSP 2002 i NSP 2011.

Płeć — wskazuje się także na nierówności w dostępie kobiet i mężczyzn do rynku pracy oraz usług społecznych. Dochody kobiet są przeciętnie o 20% niższe od zarobków mężczyzn, także wysokość wypłacanych emerytur jest diametralnie różna. Ponadto kobiety zmagają się z problemem powrotu na rynek pracy po urlopach macierzyńskich i wychowawczych — wszystkie te zjawiska wpływają na zwiększony poziom zagrożenia ubóstwem w grupie kobiet [8]. Informacje płci dostępne są na podstawie NSP 2002 oraz NSP 2011.

Wiek — według badań GUS, w Polsce ubóstwem zagrożeni są ludzie młodzi, do 17 roku życia. Szczególną grupę stanowią także osoby starsze, które ze względu na wiek i stan zdrowia mają ograniczone możliwości zarobkowe [35]. Dane na temat wieku zostały zebrane podczas NSP 2002 oraz NSP 2011.

Wykształcenie — bardzo istotną determinantą jest wykształcenie członków oraz głowy gospodarstwa domowego. Ukończony poziom edukacji bezpośrednio determinuje pozycję zajmowaną na rynku pracy [96]. Szczegółowe dane dostępne są na podstawie NSP 2002 oraz NSP 2011.

Aktywność ekonomiczna — posiadanie pracy nie jest dostatecznym gwarantem wyjścia z biedy. Ubóstwo osób pracujących staje się coraz powszechniejszym zjawiskiem i wpływa na nie niska płaca oraz praca w niepełnym wymiarze czasu [12]. Narodowe Spisy Powszechne 2002 i 2011 dostarczyły wielu danych dotyczących aktywności ekonomicznej Polaków.

Pozostawanie na utrzymaniu — Golinowska [30] w swojej pracy jako jedną z korelant ubóstwa uznaje relację liczby osób z dochodami do tych, które są na ich utrzymaniu. Szczególnie wysoki wskaźnik w tym zakresie obserwowany będzie w rodzinach wielodzietnych.

Źródła utrzymania — znalezienie się gospodarstwa w sferze ubóstwa związane jest z charakterem głównego źródła utrzymania. Gospodarstwa utrzymujące się z niezarobkowych źródeł, jak i gospodarstwa rencistów oraz rolników z dużo większym prawdopodobieństwem mogą się w tej grupie znaleźć [35]. Dane zgromadzone podczas NSP 2002 oraz NSP 2011.

Mniejszość etniczna — przynależność do mniejszości etnicznej wiąże się z narażeniem na dyskryminację, a w skrajnym przypadku – rasizmem.

Takim osobom trudniej jest dostać pracę, często też mieszkają w gorszych warunkach [25]. Przynależność etniczna była jednym z obszarów badanym podczas NSP 2002 oraz NSP 2011.

Niepełnosprawność — obecność osoby niepełnosprawnej w gospodarstwie domowym związana jest ze zwiększonymi wydatkami poświęconymi dla tej osoby. Często konieczne jest także zapewnienie permanentnej opieki, która dodatkowo zmniejsza możliwości zarobkowe takiego gospodarstwa [44]. Podczas Narodowego Spisu Powszechnego w 2002 oraz 2011 roku zostały zebrane informacje na temat niepełnosprawności poprzez wywiad bezpośredni (NSP 2002 i NSP 2011), a także z wykorzystaniem rejestru osób niepełnosprawnych (NSP 2011).

Warunki mieszkaniowe — mieszkania rodzin ubogich zwykle cechują się małym metrażem, na którym zwykle zmuszonych jest żyć kilka osób [98]. Dane na temat wielkości mieszkań zostały zebrane podczas NSP 2002 oraz NSP 2011.

5

Przegląd najważniejszych estymatorów wykorzystywanych w estymacji ubóstwa

W rozdziale tym opisane zostaną w syntetyczny sposób najważniejsze estymatory statystyki małych obszarów, które mogą być wykorzystywane w zagadnieniu estymacji wybranych charakterystyk ubóstwa na różnych poziomach agregacji przestrzennej.

5.1 Estymator bezpośredni

Niech dana będzie N – elementowa populacja $U = \{1, \ldots, N\}$. Z populacji tej losujemy zgodnie z określonym schematem losowania n – elementową próbę $s \subseteq U$. Niech π_i oznacza prawdopodobieństwo inkluzji i – tej jednostki do próby, tzn. $\pi_i = P(i \in s)$ dla $i = 1, \ldots, N$ – jest to tzw. prawdopodobieństwo inkluzji pierwszego rzędu. Niech π_{ij} oznacza prawdopodobieństwo inkluzji i – tej i j – tej jednostki do próby, tzn. $\pi_{ij} = P(i, j \in s)$ dla $i, j = 1, \ldots, N$ oraz $i \neq j$ – jest to tzw. prawdopodobieństwo inkluzji drugiego rzędu. Definiujemy wagę odpowiadającą jednostce i jako $d_i = \frac{1}{\pi_i}$ a jednostkom i oraz j jako $d_{ij} = \frac{1}{\pi_{ij}}$.

Załóżmy, że celem badania jest oszacowanie wartości globalnej pewnej zmiennej y, określonej wzorem:

$$Y = \sum_{i=1}^{N} y_i, \tag{5.1}$$

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

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

wzorem:

$$\hat{Y}_{HT} = \sum_{s} d_i y_i = \sum_{i=1}^{n} d_i y_i. \tag{5.2}$$

Jest to jeden z najczęściej wykorzystywanych przez urzędy statystyczne estymator w różnego rodzaju badaniach reprezentacyjnych. Estymator Horvitza-Thompsona może być wykorzystany dla dowolnie zdefiniowanego schematu losowania próby. W badaniach wykorzystywany jest często jako punkt odniesienia tzn. odgrywa rolę estymatora referencyjnego. Charakteryzuje się on dużą wariancją, zwłaszcza w sytuacji małej liczebności próby bądź dużych frakcji braków odpowiedzi w odniesieniu do zmiennej y. Jest to estymator nieobciążony. Oznacza to, że:

$$E\left(\hat{Y}_{HT}\right) = Y. \tag{5.3}$$

Estymator ten nie wykorzystuje żadnych dodatkowych informacji spoza próby poza wartościami zmiennej y. $\hat{Y}_{HT}$ jest nieobciążonym estymatorem Y, a jego wariancja wyraża się wzorem:

$$V\left(\hat{Y}_{HT}\right) = \sum\sum_{U}\left(\frac{d_i d_j}{d_{ij}} - 1\right) y_i y_j. \tag{5.4}$$

Statystyka postaci:

$$\hat{V}\left(\hat{Y}_{HT}\right) = \sum\sum_{s}\left(d_i d_j - d_{ij}\right) y_i y_j, \tag{5.5}$$

jest nieobciążonym estymatorem wariancji $V\left(\hat{Y}_{HT}\right)$.

Estymator ten jest wykorzystywany przez Główny Urząd Statystyczny do szacowania stopy ubóstwa w Polsce na poziomie całego kraju, regionów oraz województwa. Ze względu na małe liczebności próby dla niższych poziomów agregacji przestrzennej (podregiony, powiaty) i jego dużą wariancję nie jest możliwe wykorzystanie go dla tak zdefiniowanych domen.

5.2 Uogólniony estymator regresyjny – GREG

Jest to estymator z klasy estymatorów wykorzystujących koncepcję regresji uogólnionej. Na szeroką skalę estymatory te wykorzystywane są w badaniach statystycznych realizowanych przez urzędy statystyczne w różnego rodzaju badaniach. W chwili obecnej estymatory typu GREG, które budowane są w oparciu o podejście modelowe, stanowią bardzo szeroką klasę estymatorów wykorzystujących modele liniowe i nieliniowe. W estymatorach tego typu bardzo ważną rolę pełnią zmienne pomocnicze, w oparciu o które dokonuje się korekty wag wynikających ze schematu losowania próby.

Niech $\mathbf{X}$ oznacza wektor utworzony z wartości globalnych każdej zmiennej pomocniczej:

$$\mathbf{X} = \left(\sum_{i=1}^{N} x_{i1}, \sum_{i=1}^{N} x_{i2}, \ldots, \sum_{i=1}^{N} x_{ik} \right)^T. \tag{5.6}$$

Poszczególne składowe wektora (5.6) odnoszą się do wartości globalnych kolejnych zmiennych pomocniczych. Informacje na temat takich zmiennych można pozyskać ze spisów bądź z odpowiednich rejestrów administracyjnych.

Z kolei niech $\hat{\mathbf{X}}$ będzie wektorem złożonym z oszacowanych wartości globalnych wszystkich zmiennych pomocniczych z wykorzystaniem estymatora Horvitza-Thompsona:

$$\hat{\mathbf{X}} = \left(\sum_{i=1}^{n} d_i x_{i1}, \sum_{i=1}^{n} d_i x_{i2}, \ldots, \sum_{i=1}^{n} d_i x_{ik} \right)^T, \tag{5.7}$$

Załóżmy, że celem badania jest oszacowanie wartości globalnej zmiennej Y określonej wzorem (5.1).

Uogólnionym estymatorem regresyjnym (GREG) wartości globalnej (5.1) nazywamy statystykę postaci:

$$\hat{Y}_{GREG} = \hat{Y}_{HT} + \left(\mathbf{X} - \hat{\mathbf{X}} \right)^T \mathbf{B}_s \tag{5.8}$$

gdzie

$$\mathbf{B}_s = \left(\sum_{i=1}^{n} d_i c_i \mathbf{x}_i \mathbf{x}_i^T \right)^{-1} \left(\sum_{i=1}^{n} d_i c_i \mathbf{x}_i y_i \right), \tag{5.9}$$

jest wektorem złożonym ze współczynników regresji uzyskanych z modelu regresji pomiędzy zmienną objaśnianą y a zmiennymi objaśniającymi $x_{i1}, \ldots, x_{ik}$ z wykorzystaniem ważonej metody najmniejszych kwadratów, natomiast:

$$\mathbf{x}_i = (x_{i1}, \ldots, x_{ik})^T, \tag{5.10}$$

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

Wygodnie jest przedstawić uogólniony estymator regresyjny (5.8) jako ważoną sumę zaobserwowanych w próbie wartości y_k

$$\hat{Y}_{GREG} = \sum_{i=1}^{n} d_i g_i y_i, \tag{5.11}$$

gdzie

$$g_i = 1 + \lambda_s^T c_i \mathbf{x}_i, \tag{5.12}$$

a

$$\lambda_s^T = \left(\mathbf{X} - \hat{\mathbf{X}}\right)^T \left(\sum_{i=1}^{n} d_i c_i \mathbf{x}_i \mathbf{x}_i^T\right)^{-1}. \tag{5.13}$$

Waga g_i jest bliska jedności dla większości elementów z próby. W przypadku małych prób mogą wystąpić przypadki, w których wagi g_i są ujemne. W przypadku prób większych jest to bardzo rzadki przypadek podobnie jak i sytuacja, w której wagi g_i są większe od 4.

Estymator Horvitza-Thompsona jest szczególnym przypadkiem estymatora typu GREG. Otrzymany jest w sytuacji gdy $\mathbf{x}_i = c_i = 1$ dla $k \in s$ a w schemacie losowania spełniony jest warunek $\sum_s d_i = N$.

Wariancja estymatora GREG może być bardzo duża w przypadku małych prób oraz występowania wartości odstających. Można dokonać redukcji wariancji estymatora poprzez zwiększenie liczebności próby oraz odpowiedni dobór zmiennych pomocniczych, które powinny być silnie skorelowane ze zmienną Y. Estymator GREG nie jest nieobciążony ale obciążenie wraz ze wzrostem liczebności próby wykazuje tendencję malejącą.

W sytaucji gdy $\mathbf{x}_i = x_i$ i $c_i = \frac{1}{x_i}$ uzyskujemy tzw. estymator ilorazowy postaci:

$$\hat{Y}_{GREG} = \sum_{i=1}^{N} x_i \frac{\sum_{i=1}^{n} d_i y_i}{\sum_{i=1}^{n} d_i x_i}, \tag{5.14}$$

W sytaucji gdy $\mathbf{x}_i = (1, x_i)^T$ i $c_i = 1$ uzyskujemy tzw. estymator regresyjny.

Wariancja estymatora typu GREG wartości globalnej wyraża się wzorem:

$$V\left(\hat{Y}_{GREG}\right) = \sum\sum_{U} \left(\frac{d_i d_j}{d_{ij}} - 1\right) e_i e_j, \tag{5.15}$$

gdzie

$$e_i = y_i - \mathbf{x}_i^T \mathbf{B}_U, \tag{5.16}$$

a

$$\mathbf{B}_U = \left(\sum_{i=1}^{N} c_i \mathbf{x}_i \mathbf{x}_i^T\right)^{-1} \left(\sum_{i=1}^{N} c_i \mathbf{x}_i y_i\right). \tag{5.17}$$

Estymatorem wariancji jest statystyka postaci:

$$\hat{V}\left(\hat{Y}_{GREG}\right) = \sum\sum_{s} \left(d_i d_j - d_{ij}\right) \hat{e}_i \hat{e}_j, \tag{5.18}$$

gdzie

$$\hat{e}_i = y_i - \mathbf{x}_i^T \mathbf{B}_s, \tag{5.19}$$

a

$$\mathbf{B}_s = \left(\sum_{i=1}^{n} d_i c_i \mathbf{x}_i \mathbf{x}_i^T \right)^{-1} \left(\sum_{i=1}^{n} d_i c_i \mathbf{x}_i y_i \right). \tag{5.20}$$

Estymator typu GREG jest wykorzystywany przez niektóre krajowe urzędy statystyczne w badaniach poświęconych ubóstwu. Podobnie jednak jak estymator Horvitza-Thompsona jego zastosowanie ogranicza się zazwyczaj do poziomu kraju. Dla niższych poziomów agregacji przestrzennej jego zastosowanie jest ograniczone ze względu na zbyt dużą wariancję.

5.3 Estymator kalibracyjny

Kalibracja to metoda polegająca na skorygowaniu wyjściowych wag wynikających ze schematu losowania próby z wykorzystaniem informacji dodatkowych, tak aby spełnione były odpowiednie równania kalibracyjne. W rezultacie uzyskuje się równowagę rozumianą w ten sposób, że po zastosowaniu kalibracji próba jest „wyglądem" zbliżona do całej populacji (wagi sumują się na przykład do liczby wszystkich jednostek w populacji, sumy wag w odpowienio zdefiniowanych przekrojach również mogą sumować się do liczby wszystkich jednostek w tych przekrojach w populacji generalnej).

Załóżmy, podobnie jak w przypadku estymatora typu GREG, że celem badania jest oszacowanie wartości globalnej zmiennej Y określonej wzorem (5.1).

Niech ponadto $x_1, \ldots, x_k$ oznaczają zmienne pomocnicze, a $\mathbf{X}_j$ oznacza wartość globalną zmiennej x_j, $j = 1, \ldots, k$, tj.

$$\mathbf{X}_j = \sum_{i=1}^{N} x_{ij}, \tag{5.21}$$

gdzie x_{ij} oznacza wartość j – tej zmiennej pomocniczej dla i – tej jednostki badania. Do oszacowania wartości globalnej zmiennej y wykorzystujemy estymator Horvitza-Thompsona. W praktyce bardzo często zdarza się, że:

$$\sum_{s} d_i x_{ij} \neq \mathbf{X}_j, \tag{5.22}$$

co oznacza, że pewna korekta wag (kalibracja) jest pożądana.

Niech $\mathbf{d} = (d_1, \ldots, d_n)^T$ będzie wektorem wag wynikających ze schematu losowania próby, a $\mathbf{w} = (w_1, \ldots, w_n)^T$ poszukiwanym wektorem wag kalibracyjnych, gdzie n oznacza liczebność próby. Niech G będzie dowolną funkcją spełniającą następujące warunki:

- $G(\cdot)$ jest dwukrotnie różniczkowalna,
- $G(\cdot) \geqslant 0$,
- $G(1) = 0$,
- $G'(1) = 0$,
- $G''(1) = 1$.

Nowo wyznaczone wagi powinny nieznacznie się różnić od wag d_i oraz powinny spełniać warunek:

$$\sum_s w_i x_{ij} = \mathbf{X}_j. \tag{5.23}$$

Problem poszukiwania wag kalibracyjnych można opisać w następujący sposób:

- Minimalizacja funkcji odległości:

$$D(\mathbf{w}, \mathbf{d}) = \sum_{i=1}^{n} d_i G\left(\frac{w_i}{d_i}\right) \longrightarrow \min, \tag{5.24}$$

- Równania kalibracyjne:

$$\sum_{i=1}^{n} w_i x_{ij} = \mathbf{X}_j, \quad j = 1, \ldots, k, \tag{5.25}$$

- Warunki ograniczające:

$$L \leqslant \frac{w_i}{d_i} \leqslant U, \quad \text{gdzie: } L < 1 \text{ i } U > 1,\ i = 1, \ldots, n. \tag{5.26}$$

Istnieje pewna dowolność przy wyborze funkcji $G(\cdot)$. Najczęściej rozważa się w literaturze następujące jej postacie:

$$G_1(x) = \frac{1}{2}(x-1)^2, \tag{5.27}$$

$$G_2(x) = \frac{(x-1)^2}{x}, \tag{5.28}$$

$$G_3(x) = x(\log x - 1) + 1, \tag{5.29}$$

$$G_4(x) = 2x - 4\sqrt{x} + 2, \tag{5.30}$$

$$G_5(x) = \frac{1}{2\alpha}\int_1^x \sinh\left[\alpha\left(t - \frac{1}{t}\right)\right] dt, \tag{5.31}$$

gdzie α jest dodatnim parametrem, pozwalającym sterować stopniem rozrzutu wag kalibracyjnych w stosunku do wag wynikających ze schematu losowania próby (domyślnie parametr przyjmuje wartość 1), a sinh jest funkcją sinusa hiperbolicznego zdefiniowanego jako $\sinh(x) = \frac{e^x - e^{-x}}{2}$.

W praktycznych zastosowaniach najczęściej wykorzystuje się funkcję G w postaci $G_1(x) = \frac{1}{2}(x-1)^2$. w tym przypadku mamy bowiem:

$$D(\mathbf{w}, \mathbf{d}) = \sum_{i=1}^{n} d_i G\left(\frac{w_i}{d_i}\right) = \sum_{i=1}^{n} d_i \frac{1}{2}\left(\frac{w_i}{d_i} - 1\right)^2 = \frac{1}{2}\sum_{i=1}^{n} \frac{(w_i - d_i)^2}{d_i}. \quad (5.32)$$

Estymatorem kalibracyjnym wartości globalnej zmiennej Y jest:

$$\hat{Y}_{cal} = \sum_{i=1}^{n} w_i y_i, \quad (5.33)$$

gdzie wektor wag kalibracyjnych $\mathbf{w} = (w_1, w_2, \ldots, w_n)^T$ jest rozwiązaniem zadania minimalizacji:

$$\mathbf{w} = \text{argmin}_v D(\mathbf{v}, \mathbf{d}), \quad (5.34)$$

$$\mathbf{X} = \tilde{\mathbf{X}}, \quad (5.35)$$

przy czym

$$D(\mathbf{v}, \mathbf{d}) = \frac{1}{2}\sum_{i=1}^{n} \frac{(v_i - d_i)^2}{d_i}, \quad (5.36)$$

$$\tilde{\mathbf{X}} = \left(\sum_{i=1}^{n} w_i x_{i1}, \sum_{i=1}^{n} w_i x_{i2}, \ldots, \sum_{i=1}^{n} w_i x_{ik}\right)^T, \quad \mathbf{X} = \left(\sum_{i=1}^{N} x_{i1}, \sum_{i=1}^{N} x_{i2}, \ldots, \sum_{i=1}^{N} x_{ik}\right)^T. \quad (5.37)$$

Rozwiązaniem powyższego zadania minimalizacji jest wektor wag kalibracyjnych $\mathbf{w} = (w_1, w_2, \ldots, w_n)^T$, którego składowe spełniają równanie

$$w_i = d_i + d_i\left(\mathbf{X} - \hat{\mathbf{X}}\right)^T \left(\sum_{i=1}^{n} d_i \underline{\mathbf{x}}_i \underline{\mathbf{x}}_i^T\right)^{-1} \underline{\mathbf{x}}_i, \quad (5.38)$$

przy czym:

$$\hat{\mathbf{X}} = \left(\sum_{i=1}^{n} d_i x_{i1}, \sum_{i=1}^{n} d_i x_{i2}, \ldots, \sum_{i=1}^{n} d_i x_{ik}\right)^T, \quad (5.39)$$

$$\underline{\mathbf{x}}_i = (x_{i1}, x_{i2}, \ldots, x_{ik})^T. \quad (5.40)$$

Podobnie jak estymatory typu GREG, estymatory kalibracyjne wykorzystywane są w praktycznych zastosowaniach przez różne urzędy statystyczne

na całym świecie. Dotyczy to również zagadnienia estymacji ubóstwa, zazwyczaj jednak na wyższych poziomach agregacji przestrzennej takich jak kraj czy województwo w Polsce.

5.4 ELL

Metoda ELL została zaproponowana przez Elbersa, Lanjouw i Lanjouw (2003) i jest wykorzystywana głównie w pracach prowadzonych przez Bank Światowy. Jej podstawy teoretyczne stanowi model regresji z zagnieżdżonym błędem, gdzie w roli zmiennej objaśnianej występuje zlogarytmowany dochód gospodarstw domowych. Rozpatrujemy jednostki administracyjne oznaczone $d = 1, ..., D$ oraz gospodarstwa domowe w tych obszarach $j = 1, ..., N_d$. Postać tego modelu w oryginalnej formie opisana przez [4] jest następująca:

$$Y_{dj} = X_{dj}\beta + u_d + e_{dj}, \tag{5.41}$$

gdzie: Y_{dj} — zmienna objaśniana, X_{dj} — macierz zmiennych objaśniających, u_d — losowy efekt obszaru, $u_d \overset{iid}{\sim} N(0, \sigma_u^2)$, e_{dj} — błąd losowy, $e_{dj} \overset{iid}{\sim} N(0, \sigma_e^2)$.

W podejściu ELL losowy efekt obszaru (u_d) został zastąpiony losowym efektem grupy. Grupy te tworzone są przez podobne gospodarstwa domowe i mogą znacznie różnić się od określonych jednostek administracyjnych.

Kolejnym etapem estymacji jest wygenerowanie A prób bootstrapowych na podstawie modelu (5.41) oraz dostępnych danych z badania pełnego. Otrzymujemy zatem wektory oszacowań $\{Y_{dj}^{*(a)}; j = 1, ..., N_d; d = 1, ..., D\}$ dla $a = 1, ..., A$ prób. Na podstawie oszacowanych w próbach bootstrapowych wartości zmiennej objaśnianej oblicza się oceny stopy ubóstwa $\{F_d^{*(a)}; d = 1, ..., D\}$ dla $a = 1, ..., A$, a następnie uśrednia otrzymując końcowy szacunek zgodnie z formułą:

$$\hat{F}_d^{ELL} = \frac{1}{A}\sum_{a=1}^{A} F_d^{*(a)} =: F_d^{*(.)}, \tag{5.42}$$

gdzie: $\hat{F}_d^{ELL}$ — oszacowanie stopy ubóstwa w d-tej jednostce administracyjnej, $F_d^{*(a)}$ — oszacowanie stopy ubóstwa w d-tej jednostce administracyjnej na podstawie a-tej próby bootstrapowej.

W przypadku, gdy dany obszar d nie jest reprezentowany, wartości zmiennej objaśnianej $Y_{dj}^{(l)}$ dla $j = 1, ..., N_d$ są generowane metodą bootstrap według wzoru (5.41). Wówczas estymator syntetyczny przyjmuje postać:

$$\hat{F}_d^{S-ELL} = \frac{1}{N_d}\sum_{j=1}^{N_d} F_{dj}^{S-ELL}, \tag{5.43}$$

gdzie: $\hat{F}_d^{S-ELL}$ — syntetyczne oszacowanie stopy ubóstwa w d-tej jednostce administracyjnej, F_{dj}^{S-ELL} — funkcja przynależności j-tego gospodarstwa domowego do strefy ubóstwa.

Jakość estymacji oceniana jest na podstawie wielkości błędu średniokwadratowego (ang. mean square error, MSE). Jego wartość można oszacować na podstawie wygenerowanych już prób bootstrapowych według następującej formuły:

$$MSE(\hat{F}_d^{ELL}) = \frac{1}{A}\sum_{a=1}^{A}(F_d^{*(a)} - F_d^{*(.)})^2, \tag{5.44}$$

gdzie: $MSE(\hat{F}_d^{ELL})$ — błąd średniokwadratowy oszacowania stopy ubóstwa w d-tej jednostce administracyjnej, $\hat{F}_d^{ELL}$ — oszacowanie stopy ubóstwa w d-tej jednostce administracyjnej, $F_d^{*(a)}$ — oszacowanie stopy ubóstwa w d-tej jednostce administracyjnej na podstawie a-tej próby bootstrapowej.

Przy adekwatnym modelu otrzymane oszacowania charakteryzują się mniejszym obciążeniem oraz błędem średniokwadratowym aniżeli szacunki bezpośrednie wykorzystujące wyłącznie informacje pochodzące z próby. Metoda ELL jest szeroko stosowana przez Bank Światowy ze względu na jej średni stopień złożoności oraz dedykowane oprogramowanie PovMap [112]. Niemniej podejście to nie jest pozbawione wad. Główny zarzut to niewykorzystywanie danych pochodzących z badania próbkowego, a jedynie prognozownie dochodu w badaniu pełnym.

5.5 Estymator EBLUP na poziomie jednostki

Estymator EBLUP na poziomie jednostki jest kombinacją liniową estymatora bezpośredniego oraz syntetycznego regresyjnego dla jednostek, które nie trafiły do próby. Model składa się z dwóch elementów - efektu stałego oraz efektu losowego. Efekt stały związany jest ze zmniennymi dodatkowymi dla większego obszaru, a efekt losowy uwzględnia zróżnicowanie jednostek wewnątrz małego obszaru.

Model mieszany na poziomie jednostki może być użyty w przypadku gdy dostępne są zmienne dodatkowe dla badanych jednostek w ramach danego małego obszaru. Włączenie efektu losowego dla obszaru ma na celu uwzględenienie zmienności między obszarami poprzez badanie korelacji między jednostkami w danych obszarach. Podstawowy model mieszany dla jednostki jest inaczej nazywany modelem z zagnieżdżonym błędem i może być wyrażony następująco:

$$y_{di} = x_{di}^T \beta + u_d + e_{di}, \tag{5.45}$$

gdzie:

$$u_d \sim N(0, \sigma_u^2),$$

$$e_{di} \sim N(0, \sigma_e^2),$$

$$\forall i = 1, \ldots, N_d,$$

$$d = 1, \ldots, D.$$

Zmiennna y_{di} jest badaną zmienną dla i-tej jednostki w domenie d. Przy założeniu nieinformatywnego schematu losowania takiego, jak losowanie proste, powyższy model, przy założeniu poprawności dla populacji, może być zastosowany dla próby. W związku z tym, przy użyciu notacji macierzowej, model może być sformułowany następująco:

$$\mathbf{y}_s = \mathbf{x}_s \boldsymbol{\beta} + \mathbf{z}_s \mathbf{u} + \mathbf{e}_s, \tag{5.46}$$

gdzie $\mathbf{y}_s$ jest n-wymiarowym wektorem zmiennych obserwowanych dla zmiennej y, $\mathbf{x}_s$ jest $(n \times p)$-wymiarową macierzą zmiennych objaśniających dla jednostek ujętych w badaniu, $\mathbf{e}_s$ jest n-wymiarowym wektorem błędów, $\mathbf{z}_s$ jest $(n \times D)$-wymiarową macierzą określającą przynależność jednostek do obszarów, a $\mathbf{u}$ jest D-wymiarowym wektorem efektów losowych.

W celu oszacowania parametrów modelu możliwe jest wykorzystanie podejścia predykcyjnego (ang. *predictive approach*) czy bayesowskiego (ang. *bayesian approach*).

Wykorzystując podejście predykcyjne najlepszy liniowy nieobciążony predyktor (BLUP) otrzymujemy poprzez minimalizcję funkcji kadratowej w klasie wszystkich liniowych nieobciążonych estymatorów.

Estymator BLUP jest zależny od dwóch komponentów wariacyjnych (σ_s oraz σ_e), które są nieznane i muszą być estymowane z wykorzystaniem m.in. metody największej wiarygodności (ML) lub ograniczonej metody największej wiarygodności (REML).

W wyniku oszacowania otrzymujemy estymator EBLUP, który jest estymatorem złożonym o poniższej postaci:

$$\hat{\boldsymbol{\theta}}_{\boldsymbol{d}}^{\boldsymbol{EBLUP\_UNIT}} = \boldsymbol{\gamma}_d[\bar{\boldsymbol{y}}_d + (\bar{\mathbf{X}}_{\boldsymbol{d}}^{\boldsymbol{T}} \hat{\boldsymbol{\beta}} - \bar{\mathbf{x}}_d \hat{\boldsymbol{\beta}})] + (\mathbf{1} - \boldsymbol{\gamma}_d) \bar{\mathbf{X}}_{\boldsymbol{d}}^{\boldsymbol{T}} \hat{\boldsymbol{\beta}}, \tag{5.47}$$

gdzie

$$\gamma_d = \frac{\hat{\sigma}_d^2}{\hat{\sigma}_d^2 + \hat{\sigma}_e^2/n_d}, \tag{5.48}$$

oraz $\bar{\mathbf{X}}_d$ jest wektorem znanych wartości średnich zmiennych objaśniających dla domeny d oraz $\bar{\mathbf{x}}_d$ jest odpowiadającym wektorem średnich otrzymanych z próby. Efekty stałe w modelu są estymowane na podstawie wszystkich dostępnych informacji z obszaru większego, a w przypadku gdy wariancja między obszarami jest mała, oszacowania estymatora EBLUP są zbliżone do estymatora syntetycznego, w przeciwnym wypadku większa waga przypisywana jest dla estymatora bezpośredniego.

W literaturze statystyki małych obszarów powyższy model został znacznie rozwinięty. Zwróćmy uwagę, że podstawowy model na poziomie jednostki nie uwzględnia doboru jednostek z wykorzystaniem złożonych schematów losowania. W pracy [95] autorzy zaproponowali dwustopniowy model z zagnieżdżonym błędem dla badań, w których zastosowany został dwustopniowy, warstwowy schemat losowania. Z kolei w pracy [79] zaproponowano estymator pseudo-EBLUP, który uwzględnia problem różnych prawdopodobieństw inkluzji oraz ich wpływu na model. Kolejne rozszerzenie, które zostało zaproponowane uwzględnia możliwość szacowania wielu parametrów w małych obszarach jednocześnie. Pozostałe rozszerzenia dotyczyły m.in. uwzględnienia faktu, że zmienna objaśniana nie musi być ciągła co wpłynęło na użycie uogólnionego modelu mieszanego.

W literaturze pojawiają się następujące rekomendcje za i przeciw stosowaniu powyższego modelu dla jednostki:

- model może być zastosowany w przypadku dostępności zmiennych pomocniczych dla jednostek, wartości średnie powinny być znane na poziomie obszaru,
- model jest stosowany dla zmiennych ciągłych o rozkładach normalnych, w związku z tym należy dokonać pewnych transformacji w celu doprowadzenia do rozkładu normalnego,
- model poprawia oszacowania estymatora bezpośredniego gdy występuje silny związek przyczynowo-skutkowy między zmiennymi objaśniającymi, a zmienną objaśnianą,
- jeżeli zmienna objaśniana nie ma rozkładu normalnego można stosować uogólnione modele liniowe,
- zmienne objaśniające mogą być zarówno na poziomie jednostki, jak i obszaru,

- jeżeli model jest źle określony, oszacowania mogą być obarczone błędem,

- podstawowy model nie bierze pod uwagę nieprostych schematów losowania

- pojawiają się problemy w zgodności estymacji otrzymanych dla małych obszarów w odniesieniu do dużego obszaru (np. inne sumy). Problem ten jest rozważany w literaturze poświęconej benchmarkingowi, który ma na celu przeciwdziałać niezgodności wyników,

- zakłada się, że zmienna objaśniana ma rozkład normalny lub sprowadzony do rozkładu normalnego, co w przypadku rozkładów skośnych może być trudne do osiągnięcia. W związku z tym należy rozważyć stosowanie metod odpornych lub opartych na M-kwantylach,

- podstawowy model zakłada, że reszty są nieskorelowane w czasie oraz przestrzeni.

5.6 Estymator EBLUP na poziomie obszaru (model Faya-Herriota)

Estymator EBLUP dla obszaru stanowi liniowe połączenie bezpośredniego estymatora dla obszaru (domeny) oraz komponentu prognozy opartej na mieszanym modelu liniowym. Model ten wyraża związek między parametrem badanym a znanymi zmiennymi pomocniczymi dla poszczególnych domen, na które dzieli się cała populacja. Model uwzględnia efekt jednorodności wewnątrz obszaru (domeny).

Estymator EBLUP bazuje na mieszanym modelu liniowym, który wyraża związek między parametrem badanym a informacją pomocniczą na poziomie obszaru.

Niech θ_d będzie szacowanym parametrem dla każdej domeny d. Zakłada się istnienie liniowej relacji pomiędzy θ_d a zbiorem zmiennych towarzyszących o znanych wartościach w każdej badanej domenie. Relację tę wyraża równanie:

$$\theta_d = \mathbf{X}_d^T \beta + u_d, \tag{5.49}$$

gdzie $\mathbf{X}_d$ jest wektorem zmiennych towarzyszących dla domeny d natomiast u_d s (d=1,...,D) są efektami w domenie, które zgodnie z założeniem mają

rozkład z zerową średnią i wariancją σ_u^2. Efekty losowe wyjaśniają dodatkową zmienność, która nie jest wyjaśniona przez zmienne pomocnicze ujęte w modelu.

Poza modelem dla parametrów, należy określić schemat losowania. Zakłada się, że dostępny jest wynikający ze schematu losowania bezpośredni nieobciążony estymator $\hat{\theta}_d$ (ale niekoniecznie dla wszystkich domen), taki że:

$$\hat{\theta}_d = \theta_d + e_d, \tag{5.50}$$

gdzie e_d są to błędy losowe związane z estymatorami bezpośrednimi, dla których $E(e_d|\theta_d) = 0$, tzn. zgodnie z założeniem estymator bezpośredni jest nieobciążony, a $V(e_d|\theta_d) = \varphi_d$, gdzie wariancje φ_d są znane.

Łącząc równania (5.49) oraz (5.50) uzyskuje się liniowy model mieszany. Model ma następującą postać:

$$\hat{\theta}_d = \mathbf{X}_d^T \beta + u_d + e_d. \tag{5.51}$$

Zwykle do estymacji błędu średniokwadratowego (MSE) przyjmuje się, że e oraz u mają rozkład normalny, ale założenie to nie jest konieczne do szacowania parametru. Na podstawie modelu (5.51) uzyskuje się najlepszy empiryczny nieobciążony predyktor liniowy (EBLUP) postaci:

$$\hat{\theta}_d^{\text{EBLUP\_AREA}} = \gamma_d \hat{\theta}_d + (1 - \gamma_d)\, \mathbf{X}_{.d}^T \hat{\beta}, \tag{5.52}$$

gdzie

$$\gamma_d = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \varphi_d},$$

to waga estymatora bezpośredniego a $\hat{\beta}$ to estymator oparty na ważonej metodzie najmniejszych kwadratów, który szacuje wektor współczynnika regresji β, gdzie wagi do oszacowania β wynikają z macierzy diagonalnej, której dowolny składnik ma postać $\hat{\sigma}_u^2 + \varphi_d$. Estymację parametrów σ_u^2 oraz β trzeba przeprowadzać metodą rekurencyjną. Ponadto, jak wspomniano wcześniej, aby uniknąć problemów identyfikacji w stosunku do składników wariancji, zakłada się, że wariancje losowe $V(e_d|\theta_d) = \varphi_d$ (d=1, ..., D) są znane. Tym niemniej, jeśli dostępna jest informacja na poziomie jednostek, wtedy zakładając homeskedastyczność błędów losowych, wariancję φ_d można oszacować z modelu jednostkowego lub z uogólnionej funkcji wariancji. Tak czy inaczej, takie rozwiązanie miałoby wpływ na wielkość błędu średniokwadratowego prognozowanych wartości w domenie.

Więcej szczegółów na temat specyfikacji modelu, metod estymacji $\hat{\sigma}_u^2$ można znaleźć w Rao ([82] str. 115–120). Szczegóły na temat estymacji błędu średniokwadratowego są dostępne w Rao ([82] str. 103 oraz 128–130).

Poniżej sformułowane zostały najwazniejsze zalecenia dotyczące stosowania tej metody.

- Metoda może być stosowania do estymacji gdy liczebność próby dla jednej lub więcej badanych domen (obszarów) jest niewielka lub zerowa;
- Metoda może być stosowana na makrodanych odnoszących się do poziomu domeny;
- Metoda jest użyteczna do poprawy szacunków bezpośrednich jeśli dostępny jest zbiór zmiennych pomocniczych silnie powiązanych ze zmienną badaną;
- Wariancje szacunków dla małych obszarów muszą być znane. Zwykle stosuje się model wygładzony do szacowania wariancji, przy założeniu, że wariancje są znane. Takie podejście wpływa na wielkość błędu średniokwadratowego;
- Zmienne pomocnicze potrzebne są tylko na poziomie domeny

Poniżej sformułowane zostały ewentualne wady rozpatrywanej metody estymacji.

- Jeśli model nie jest odpowiednio wyspecyfikowany, estymator może wykazywać obciążenie;
- Sumując szacunki z małych obszarów w ramach większej domeny nie ma pewności uzyskania wartości równych szacunkom bezpośrednim na wyższym poziomie. Prostym sposobem zapewnienia zgodności w tym zakresie jest ilorazowa korekta estymatora EBLUP dla obszaru. Innym sposobem jest wprowadzenie ograniczenia za pomocą kalibracji (benchmarking) podczas obliczania szacunków dla małych obszarów.
- Wymagana jest symetria rozkładu, podczas gdy w badaniach często występuje skośność rozkładów (na przykład dochodów czy wydatków). Jeśli przekształcenie zmiennych nie wystarczy do redukcji skośności, może pojawić się konieczność zastosowania bardziej zaawansowanych metod.
- Założenie normalności rozkładu oraz znanej wartości wariancji może być nie do utrzymania przy małej wielkości próby.
- Szacowana wartość wariancji modelu σ_u^2 może wynosić 0. Jest to wynik niepożądany. W takim przypadku dobrą alternatywą są hierarchiczne metody Bayesowskie, które zawsze dają dodatnie wartości wariancji.

5.7 Estymacja dla danych panelowych

Określenie 'dane panelowe' oznacza powszechnie informacje statystyczne gromadzone poprzez powtarzalne obserwacje jednej lub więcej zmiennych dla jednej lub więcej jednostek obserwacji. W rezultacie dla każdej jednostki pozyskuje się wiele danych na temat tej samej zmiennej w różnych okresach czasu. Jednostką obserwacji może być tutaj osoba, gospodarstwo domowe czy obszar przestrzenny. Specyficzną cechą takich danych jest fakt, że mają one zarówno formę przekrojową jak też czasową, a zatem mogą być analizowane jednocześnie jako dane strukturalne i szeregi czasowe. Pochodzą one zazwyczaj z badań panelowych lub przekrojowych (zawierających pytania retrospektywne). Dobrym przykładem w tym kontekście jest panelowy charakter Badania Aktywności Ekonomicznej Ludności (BAEL), opartego na rotacyjnym schemacie doboru próby: próba kwartalna składa się z czterech prób elementarnych. W takim zestawie co kwartał wymieniana jest jedna z nich. Tak więc w każdym panelu dwie podpróby elementarne były badane w kwartale poprzednim, jedna została nowo wprowadzona do badania, natomiast jedną badano rok wcześniej. Niektórzy badacze (np. J. Brüderl (citeBR05)) utrzymują, że dane panelowe rejestrują jedynie "migawki" (ang. *snapshots*) z ciągłą lub dyskretną zmienną zależną) i stąd posiadają one mniejszą wartość informacyjną niż dane historyczne. Jednakże, jeśli dane panelowe są pozyskiwane regularnie a ich jakość pozostaje pod ścisłą kontrolą, stanowią one także wartościowy materiał statystyczny do estymacji dla małych obszarów. Estymacja także daje możliwość wyzyskania wielu logicznych lub stochastycznych współzależności między jednostkami i okresami czasu, co w sposób istotny ulepsza jakość oszacowań wartości ogółem i innych wyników z tym związanych.

Wielu analityków (np. P. D. Allison ([3]), C. Dougherty ([23]) lub nawet J. Brüderl ([11]) dostrzega liczne korzyści płynące z danych panelowych. W odróżnieniu od klasycznych danych przekrojowych, informacja panelowa może bowiem przyczynić się do redukcji problemu obciążenia estymacji wynikłej z nieobserwowanej heterogeniczności danych (pojawiającej się szczególnie w badaniach nieeksperymentalnych, a takimi są na ogół badania statystyczne). Dzięki zmniejszeniu współliniowości oraz zwiększeniu liczby stopni swobody uzyskane oszacowania są zazwyczaj bardziej efektywne. Z drugiej strony, dane panelowe umożliwiają wykrycie indywidualnych tendencji w zakresie dynamiki (np. efekty wynikłe z zastosowania kohort osób według wieku czy gospodarstw domowych według liczby członków), co w przypadku tradycyjnych danych przekrojowych jest trudne lub nawet czasami niemożliwe.

Dane panelowe mają częstokroć dużą liczbę obserwacji. Zaleta ta daje sposobność wybrania z nich efektywnych informacji o zdarzeniach zachodzących w kolejnych okresach czasu (lub momentach czasowych). Jeśli badanie panelowe ma pełne pokrycie (tzn. jeśli dla każdej jednostki w każdym okresie czasu mamy stosowną obserwację), to nazywa się ono *zbilansowanym* (ang. *balanced*). W przeciwnym razie określa się je jako *niezbilansowane* (ang. *unbalanced*).

Niech n będzie liczbą jednostek, m — liczbą obserwowanych zmiennych, p — liczbą zmiennych nieobserwowalnych zaś τ — liczb a okresów czasu, dla których zgromadzono te dane. Ogólny ekonometryczny model estymacji dla danych panelowych jest dany wzorem (za C. Doughertym ([23])):

$$y_{it} = \beta_1 + \sum_{j=2}^{m} \beta_j x_{ijt} + \sum_{k=2}^{p} \gamma_k z_{ik} + \delta t + \epsilon_{it}, \tag{5.53}$$

gdzie $Y = [y_{it}]$ jest zmienną zależną będącą celem estymacji, $X_j = [x_{ijt}]$ to interesujące nas zmienne objaśniające, $Z_k = [z_i k]$ są nieobserwowalnymi zmiennymi objaśniającymi, δ oznacza współczynnik trendu zaś ϵ_{it} to czynnik zakłócający (spełniający klasyczne założenia stosowane w modelach regresji), $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, m$, $t = 1, 2, \ldots, \tau$, $k = 1, 2, \ldots, p$.

Największy problem dotyczy zmiennych Z_k, które są nieobserwowalne, ale mają zazwyczaj istotny wpływ na ukrytą heterogeniczność danych. Jak można łatwo zauważyć na podstawie (5.53), przyjmujemy, że ta heterogeniczność jest niezmienna w czasie. W praktyce ma to miejsce bardzo często. Jednakowoż, to założenie nie ułatwia nam rozwiązania problemu nieobserwowalnych zmiennych – w dalszym ciągu brak informacji na temat całego nieobserwowalnego komponentu w (5.53). Dlatego też wpływ zmiennych Z_k, $k = 1, 2, \ldots, p$, jest reprezentowany przez łączny efekt nieobserwowalny, tzn.

$$\alpha_i \stackrel{df}{=} \sum_{k=2}^{p} \gamma_k z_{ik},$$

$i = 1, 2, \ldots, n$, wobec czego wzór (5.53) może być zapisany jako

$$y_{it} = \beta_1 + \sum_{j=2}^{m} \beta_j x_{ijt} + \alpha_i + \delta t + \epsilon_{it}, \tag{5.54}$$

dla $i = 1, 2, \ldots, n$ i $t = 1, 2, \ldots, \tau$.

Jeśli badanie panelowe jest zbilansowane, tzn. jeśli zbiór zmiennych obserwowalnych X_j, $j = 1, 2, \ldots, m$, i dane gromadzone dla nich obejmują wszystkie aspekty analizowanych zjawisk społeczno – ekonomicznych, wówczas nie ma nieobserwowalnych charakterystyk i czynniki α_i, $i = 1, 2, \ldots, n$, mogą

zostać pominięte. W takim przypadku model (5.54) jest tożsamy z funkcją regresji wielorakiej zawierającą czynnik czasowy i stąd dla estymacji parametrów można zastosować rozległą klasyczną metodę najmniejszych kwadratów (ang. *pooled ordinary least squares — pooled OLS*).

Regresja tego typu jest regresją względem wszystkich danych wykorzystującą klasyczną metodę najmniejszych kwadratów po umieszczeniu wszystkich danych razem bez rozróżnienia pomiędzy komponentami: przekrojowym i czasowym. Jak ujmuje to J. M. Wooldridge ([108]) ideą tej metody jest fakt, że każdego roku losuje się z odpowiedniej populacji nową próbkę (nowe próbki). Oczywiście, z uwagi na możliwe zmiany rozkładów zmiennych w czasie obserwacje takie mogą być wprawdzie niezależne, ale niekoniecznie jednakowo rozłożone. Stąd ważną rzeczą jest wyzyskanie niezależności poszczególnych przekrojach dla korekcji tego problemu. Jednakże taka korekcja przekrojów może doprowadzić do braku powtarzalności w czasie. Celem redukcji tych problemów J. M. Wooldridge ([107]) sugeruje stosowanie w tym kontekście wielu analitycznych metod korekcji heteroskedastyczności (nierówności wariancji), testów, zmiennych instrumentalnych, itp.

J. Hecht i E. M. Haye ([38]) zaobserwowali, że nawet jeśli model czasowy i przekrojowy przy zastosowaniu różnych metod, takich jak: rozległa OLS, uogólniona metoda najmniejszych kwadratów (ang. *Generalized Least Squares — GLS*), regresji pozornie niepowiązanej (ang. *Seemingly Unrelated Regression (SUR)*), techniki korekcji autokorelacji, heteroskedastyczności i korelacji reszt międzyrównaniowych /panelowych zapewniały identyczne współczynniki, to generowane przez nie oszacowania były różne. C. Heij i in. ([39]) zauważyli, że estymator OLS odpowiada estymacji parametrów dla każdego równania z osobna przy użyciu OLS i że ta estymacja jest zgodna, ale nieefektywna jeśli zakłócenia dla różnych jednostek ujawniają jednoczesną korelację, a zbiór regresorów (zmiennych objaśniających) różni się pomiędzy poszczególnymi równaniami. J. M. Wooldridge ([108]) zauważa także, że o ile stosowanie metody OLS dla każdego równania odrębnie pozwala na łatwe testowanie hipotez dotyczących współczynników w obrębie danego równania, to nie zapewnia ona wygodnej drogi dla testowania efektywności ograniczeń międzyrównaniowych. Stąd lepszym wyjściem w tym zakresie wydaje się być rozwiązywanie układu równań w sposób kompleksowy, wielowymiarowy i stosowanie na przykład estymatorów opartych na wykonalnej metodzie najmniejszych kwadratów (ang. *the feasible General Least Squares — FGLS*), która szacuje strukturę heteroskedastyczności uzyskanej z OLS (zob. [52]). Zaletą tego podejścia jest brak kowariancji błędów w czasie. Dla typowo rotacyjnego badania jakim jest BAEL można też zastosować bardziej

specyficzne, acz efektywne, współczynniki autokorelacji błędów oszacowań ([105]).

Oprócz wyżej wymienionych technik najmniejszych kwadratów, znane są dwa główne podejścia analityczne uwzględniające charakter błędów. Pierwsze z nich to regresja stałych efektów (ang. *fixed effects regression*). Według tej opcji efekty nieobserwowalne są albo eliminowane albo zastępowane przez zmienne nieme (tj. zerojedynkowe przyjmujące wartość 1 gdy dana obserwacja dotyczy danej jednostki i 0 w przeciwnym razie). Można więc albo wycentrować w obrębie poszczególnych grup jednostek (np. małych obszarów, do których należą) wszystkie zmienne i błędy w czasie — czyli poprzez odjęcie średniej spowodować, by miały one średnią zero (takie podejście nazywa się modelem stałych efektów wewnątrzgrupowych — ang. *within—group fixed effects model*) bądź też zastosować model pierwszych różnic (ang. *first differences regression*), polegający na eliminacji nieobserwowalnych efektów poprzez obliczenia różnicy pomiędzy dwoma kolejno analizowanymi okresami czasu bądź wykorzystać wspomniane już zmienne nieme (ang. *least squares dummy variable (LSDV) regression*) zastępujące zmienne Z_k.

Drugą zasadniczą opcją jest regresja efektów losowych (ang. *random effects regression*). Stosuje się ją głównie wtedy, gdy zmienne objaśniające są stałe dla każdej jednostki. W przeciwieństwie do opcji z efektami stałymi zmienne Z_k $(k = 1, 2, \ldots, p)$ są tutaj traktowane jako wylosowane z danego rozkładu prawdopodobieństwa. Szczegóły na temat tych opcji można znaleźć w książce C. Dougherty'ego ([23]).

5.8 Model m–kwantylowy

Ten rodzaj opartej na modelu estymacji dla małych obszarów wykorzystuje regresję kwantylową jako podstawę oszacowań. Jest to specyficzny typ analizy regresji, zaproponowany przez Koenkera i Bassetta ([49]). Polega on na tym, że dla każdej wartości $q \in (0, 1)$, konstruowany jest odrębny model regresyjny ukazujący zależność q–tego kwantyla[1] warunkowego rozkładu zmiennej objaśnianej Y pod warunkiem $\boldsymbol{X}$ ($\boldsymbol{X}$ to zmienna lub zmienne objaśniająca(-e)), który ulega zmianie wraz ze zmianą $\boldsymbol{X}$. Na przykład, gdy $q = 0,5$ prosta regresji kwantylowej ukazuje jak mediana warunkowego rozkładu Y zmienia się w zależności od $\boldsymbol{X}$. W przypadku $q = 0,25$ mamy w podobny sposób

[1]Kwantyl rzędu $q \in (0, 1)$ zmiennej losowej X to taka wielkość wynikająca z jej rozkładu, że co najmniej $q * 100\%$ obserwacji jest od niej nie większych i co najwyżej $(1 - q) * 100\%$ jest od niej nie mniejszych.

do czynienia z pierwszym kwartylem, a więc odpowiednia prosta regresji oddziela górne 75% warunkowego rozkładu zmiennej objaśnianej od dolnego 25%. Ogólny model regresyjny jest postaci:

$$Q_q(Y|\boldsymbol{X}) = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}_q, \tag{5.55}$$

gdzie współczynniki $\boldsymbol{\beta}_q$ szacuje się poprzez minimalizację funkcji

$$\sum_{i=1}^{n} \left| y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} \right| \Big((1-q) I\left(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} \leqslant 0 \right) + qI\left(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} > 0 \right) \Big),$$

względem $\boldsymbol{\beta}$, zaś $I(a)$ oznacza indykator warunku a, czyli $I(a) = 1$ gdy a zachodzi i $I(a) = 0$ w przeciwnym razie. Celem rozwiązania tego problemu stosuje się zazwyczaj metody programowania liniowego (Koenker i D'Orey, [50]). W bardziej ogólnym ujęciu zagadnienie przyjmuje formę:

$$\int \boldsymbol{\psi}_q(Y - Q) f(Y|\boldsymbol{X}) dY = 0,$$

gdzie $\boldsymbol{\psi}$ oznacza ustaloną funkcję wpływu. Wtedy model (5.55) przyjmuje postać

$$Q_q(Y|\boldsymbol{X}; \boldsymbol{\psi}) = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}_{\boldsymbol{\psi}}(q),$$

A zatem dopuszczamy odmienny zestaw parametrów dla każdej wartości q. Modele wielowymiarowe przyjmują, że zmienność odpowiadająca warunkowemu rozkładowi Y pod warunkiem $\boldsymbol{X}$ może być przynajmniej częściowo wyjaśniona przez daną strukturę hierarchiczną. Tej idei pomiaru efektów hierarchicznych metodą m—kwantylową poświęcono wiele prac z zakresu estymacji dla małych obszarów (Chambers i Tzavidis [14]; Tzavidis i in. [100], Tzavidis i Brown ([99])).

W praktyce, dla jednostki i w skupieniu j (np. gminie) z wartościami Y_{ij} i $\boldsymbol{X}_{ij}$ i współczynnikiem mkwantylowym charakteryzującym warunkową zmienność w badanej populacji p_{ij} takim, że $Q_{p_{ij}}(\boldsymbol{X}_{ij}; \boldsymbol{\psi}) = Y_{ij}$. Zauważmy, że te współczynniki są wyznaczone na poziomie populacji. Jeśli hierarchiczna struktura nie wyjaśnia części zmienności danych z populacji, oczekujemy, że jednostki wewnątrz skupień zdefiniowanych przez tę hierarchię będą miały podobne współczynniki m–kwantylowe. Skupienie jest bowiem charakteryzowane przez położenie rozkładu współczynników m–kwantylowych należących do niego jednostek. W praktyce do estymacji tego rodzaju modeli najczęściej wykorzystuje się metody bootstrapowe (zob. [99]).

6

Metody wielowymiarowej analizy danych w estymacji i mapowaniu przestrzennego rozkładu ubóstwa

W tej części raportu przedstawimy możliwości wykorzystania najistotniejszych narzędzi wielowymiarowej analizy danych w statystyce małych obszarów. Szczególną uwagę poświęcimy metodom użytecznym w badaniu przestrzennego zróżnicowania ubóstwa. Prezentacja dotyczyć będzie metod klasyfikacji i porządkowania obiektów wielocechowych, roli podejścia opartego na zbiorach rozmytych w estymacji i analizie przestrzennego zróżnicownaia ubóstwa oraz znaczenie informacji na temat rozmaicie rozumianego sąsiedztwa obiektów w tym kontekście.

6.1 Klasyfikacja i porządkowanie obszarów przestrzennych

Jest to kluczowa gałąź taksonomii, która odgrywa istotną rolę w dwóch aspektach statystki małych obszarów. Po pierwsze, umożliwia efektywny podział jednostek na warstwy, co wydatnie poprawia jakość estymacji. Z drugiej strony, dzięki tym sposobom można dokonywać efektywnej analizy przestrzennego zróżnicowania zjawisk społeczno – ekonomicznych, takich jak właśnie ubóstwo, co pozwala na skuteczne kształtowanie polityki regionalnej. Istotę tego rodzaju analizy stanowi fakt, że badamy tutaj złożone zjawisko społeczno—gospodarcze, opisane przez wiele zmiennych statystycznych o różnorodnym charakterze. Każda jednostka (zwana tutaj *obiektem wielocechowym*) opisana jest więc określonym zestawem danych. Wyróżnia się dwa zasadnicze kierunki analizy taksonomicznej (por. [67]):

- **Klasyfikacja obiektów**, polegająca na grupowaniu obiektów w celu utworzenia ich jednorodnych klas ze względu na wewnętrzne zróżnicowanie badanych cech, a tym samym również ze względu na badane zjawisko złożone,

- **Porządkowanie obiektów**, sprowadzające się do konstrukcji syntetycznego miernika rozwoju obrazującego zróżnicowanie obiektów w zakresie przedmiotowego zjawiska złożonego oraz dokonywanego na tej podstawie rankingu obiektów, co również prowadzi do uzyskania określonych grup obiektów podobnych. Wyróżnia się tutaj podział na metody bezwzorcowe i wzorcowe. Te pierwsze polegają na konstrukcji globalnych mierników agregatowych i dokonywaniu odpowiednich grupowań na podstawie tylko znormalizowanych wartości cech oraz pomiaru ich odległości. W drugim przypadku konstruujemy tzw. taksonomiczny wzorzec rozwoju, czyli sztuczny obiekt odniesienia, dla którego wartości cech stanowią pewnego rodzaju optimum, a następnie mierzymy odległość poszczególnych obiektów od tegoż wzorca, po czym na bazie tychże odległości konstruujemy miernik kompleksowy lub grupujemy obiekty.

Podstawę analizy stanowi zestaw zmiennych zgromadzonych według siedmiu zasad sformułowanych m.in. przez T. Śmiłowską ([114]):

- istotność z punktu widzenia analizowanych zjawisk,
- jednoznaczność i precyzyjność zdefiniowania,
- wyczerpanie zakresu zjawiska,
- logiczność wzajemnych powiązań,
- zachowana proporcjonalność reprezentacji zjawisk cząstkowych,
- mierzalność – w sensie możliwości liczbowego wyrażenia poziomu cechy,
- dostępność i kompletność informacji statystycznych (dla wszystkich badanych obiektów).

Ponadto preferuje się wybierać do analizy taksonomicznej zmienne o charakterze wskaźnikowym (np. liczbę pracujących w przeliczeniu na 1000 ludności, odsetek gospodarstw domowych z szerokopasmowym dostępem do Internetu, wskaźnik zagrożenia ubóstwem, itp.). Pozostawanie przy wartościach w ujęciu bezwzględnym może bowiem prowadzić do zafałszowania wyników

— pewne (bywa, że nieliczne) obiekty ze swej natury lub określonych specyficznych uwarunkowań z nimi związanych mogą charakteryzować się wielkościami znacznie wyższymi od innych.

Zanim przystąpimy do zasadniczej analizy, zgromadzony zestaw cech należy poddać stosownemu przeglądowi. Jego pierwszy etap stanowi weryfikacja zmiennościowa zestawu cech. Dokonuje się tutaj eliminacji tych spośród nich, które wykazują zbyt niskie zróżnicowanie — z taksonomicznego punktu widzenia stanowią one bowiem niewielką wartość analityczną (mają znikomą zdolność różnicującą, a przecież celem analizy taksonomicznej jest odzwierciedlenie zróżnicowania). W tym celu ustalamy, dla których spośród badanych cech współczynnik zmienności przyjmuje wartości poniżej arbitralnie ustalonego progu. Zazwyczaj przyjmuje się go na poziomie 10%. Wieloletnie doświadczenia w tym zakresie skłaniają nas do stwierdzenia, że specyfika określonych dużych agregacji przestrzennych powoduje, iż mimo identycznej wyjściowej kolekcji zmiennych, dla każdej z nich zbiór charakterystyk finalnie poddawanych analizie może być odmienny.

Drugi etap kontroli polega na poddaniu ukształtowanego w etapie pierwszym zestawu zmiennych weryfikacji korelacyjnej. Jej cel stanowi ustalenie, które cechy wykazują nadmierne skorelowanie z innymi (i stanowią tym samym nośnik podobnej informacji). Ważną rzeczą jest także spełnienie postulatu, aby owa ocena uwzględniała wszystkie właściwości modelu, czyli reprezentowała ujęcie kompleksowe. Można tego dokonać stosując np. metodę parametryczną opartą na maksymalizacji sum elementów wierszach/kolumnach macierzy korelacji (zob. [114]). Jednakże lepsze podejście w tym zakresie, pozwalające uniknąć lub zniwelować wiele niedogodności metody parametrycznej (np. wrażliwość na symetrię czy preferowanie brzegowości rozkładów) stanowi metoda odwróconej macierzy korelacji (p. [62], [58], [67]). Dla danego zestwu cech wyznaczamy mianowicie macierz współczynników korelacji Pearsona, a następnie macierz do niej odwrotną. Diagonalne elementy tejże macierzy należą do przedziału $[1, \infty)$. Jeśli wartość któregoś z nich przekracza arbitralnie ustalony próg (najczęściej 10), to oznacza to, iż macierz jest wadliwie uwarunkowana numerycznie. Innymi słowy, odpowiadająca jej cecha okazuje się nadmiernie skorelowana z innymi. Można więc ją usunąć. Jeśli takich niepokojących wartości diagonalnych występuje więcej, to należy dokładnie przeanalizować korelację pomiędzy poszczególnymi „podejrzanymi" cechami, a następnie dokonać takiej eliminacji, aby była ona jak najskromniejsza, a jednocześnie zapewniła odpowiednie nieskorelowanie pozostałych cech. Weryfikacja korelacyjna wymusza więc niejako zastosowanie pewnej uznaniowości w podejmowaniu decyzji dyskryminacyjnej.

Powyższe czynności prowadzą do uzyskania *zestawu cech diagnostycznych*, który stanowi podstawę dalszych analiz mających teraz na celu ujednorodnienie wszystkich zmiennych. W pierwszym rzędzie należy określić kierunek wpływu poszczególnych charakterystyk na rozwój agregatowy. Wyróżniamy zatem trzy kategorie cech:

- ***stymulanty*** – zmienne, których wyższe wartości decydują o lepszym poziomie rozpatrywanego zjawiska w badanym obiekcie; przykładem może tutaj być wpływ długości dróg publicznych o twardej nawierzchni na 100km^2 powierzchni miasta na rozwój infrastruktury komunalnej,

- ***destymulanty*** – zmienne wykazujące działanie odwrotne do stymulant, tzn. wzrost ich wartości prowadzi do pogorszenia się sytuacji obiektu pod omawianym względem; taką cechą jest dla przykładu stopa bezrobocia w kontekście pozytywnie postrzeganego rozwoju rynku pracy,

- ***nominanty*** – cechy charakteryzujące się najkorzystniejszą z punktu widzenia oceny obiektów wartością, tzw. optymalnym poziomem nasycenia lub wartością nominalną. Innymi słowy, jeśli λ_j jest optymalnym poziomem nasycenia cechy – nominanty X_j, $j \in 1, 2, \ldots, m$, to oznacza to, że albo wartości cechy X_j niższe od λ_j są niekorzystne, zaś wyższe – korzystne z punktu widzenia wpływu na rozwój agregatowy, bądź też zachodzi prawidłowość odwrotna – jeśli $x_{ij} > \lambda_j$ wówczas wzrost wartości cechy oznacza słabszą pozycję obiektu w świetle rozpatrywanej dziedziny, zaś, gdy $x_{ij} < \lambda_j$ – wyższą. Zatem nominanty są funkcjami monotonicznymi z ekstremum lokalnym dla wartości nominalnej — przyrost wartości do optymalnego poziomu nasycenia wywiera pozytywny wpływ na ocenę, podczas, gdy dalszy jej wzrost, po przekroczeniu tej granicy generuje wpływ negatywny lub odwrotnie. Przykładami takich cech mogą być: wydatki konsumpcyjne, akumulacja kapitału czy udział inwestycji w Produkcie Krajowym Brutto, które w pewnych regionach zazwyczaj okazują się stymulantami, w innych zaś – destymulantami.

Określanie charakteru zmiennych powinno się dokonywać w oparciu o kryteria merytoryczne lub analizę korelacyjną, bądź też poprzez testowanie zgodności rozkładów teoretycznych. Można także przeprowadzić ex post weryfikację określenia charakteru zmiennych, wyznaczając wartość współczynnika jej korelacji ze zmienną syntetyczną. Stymulanty powinny być wówczas z taką metacechą skorelowane dodatnio, destymulanty – ujemnie, zaś nominanty

nie powinny wykazywać wyraźnych korelacji z pozostałymi zmiennymi. Nadmieńmy jeszcze, że w analizie taksonomicznej nie uwzględnia się cech neutralnych, czyli takich, które nie wywierają wpływu na rozwój agregatowy w badanej dziedzinie. Dla ujednolicenia charakteru cech destymulanty i nominanty zamieniane są na stymulanty, na przykład poprzez odwrócenie wartości destymulujących lub zmianę ich znaku (tzn. w miejsce wartości destymulant lub destymulujących części nominant przyjmuje się ich odwrotności lub wartości do nich przeciwne).

Jeden z naczelnych postulatów analizy taksonomicznej to porównywalność zmiennych. Tymczasem cechy są zazwyczaj wyrażone przy pomocy różnorodnych jednostek pomiarowych, z rozmaitą dokładnością, a ich wartości mogą charakteryzować się też odmiennym przedziałem zmienności. Konieczne zatem staje się ujednolicenie charakterystyk oraz ustalenie sztywniejszych ram ich rozpiętości. Można tego dokonać na drodze normalizacji. Termin ten pochodzi od łacińskiego słowa *normalis* – uregulowany i oznacza niwelację, ujednolicanie, wyrównywanie cech przy zachowaniu określonych kryteriów. Musi to być takie prze-kształcenie, aby otrzymane w jego wyniku „poprawione" zmienne charakteryzowały się identycznymi wielkościami pewnych miar tendencji centralnej bądź przyjmowały wartości z tego samego, ściśle określonego, spójnego, zwartego i domkniętego podzbioru zbioru liczb rzeczywistych $\mathbb{R}$. Wyróżniamy trzy zasadnicze rodzaje przekształceń normalizacyjnych:

- ***standaryzacja*** — ma ona na celu uzyskanie zmiennych o wariancji lub medianowym odchyleniu bezwzględnym równym 1 (w tym drugim przypadku mówimy o standaryzacji pozycyjnej). Osiąga się to np. poprzez odjęcie średniej arytmetycznej i podzielenie tej różnicy przez odchylenie standardowe lub wykorzystanie pozycyjnych odpowiedników tych statystyk opisowych (tj. mediany i medianowego odchylenia bezwzględnego odpowiednio),

- ***unitaryzacja*** – jej główną ideę stanowi transformacja cech diagnostycznych do takiej postaci, aby przedział ich zmienności miał stałą długość 1. W tego rodzaju przekształceniach dzieli się na ogół wartość cechy przez jej rozstęp, czyli różnicę pomiędzy jej maksymalną a minimalną wartością, maksymalną wartość bezwzględną odchylenia od średniej lub mediany, itp.,

- ***przekształcenia ilorazowe*** – ich punktem odniesienia (tzn. wielkością kompleksową przez którą dzielimy wyjściową wartość cechy) bywa

średnia arytmetyczna, wartość minimalna, maksymalna czy mediana cechy, bądź też suma jej wartości, suma kwadratów wartości cechy, mediana jej składników albo pierwiastek z tejże sumy.

Szczegółowy opis formuł normalizacyjnych mozna znaleźć np. w pracach [111] i [67]. W ostatnich latach preferowane jest podejście wykorzystujące medianę Webera (będącą wielowymiarowym uogólnieniem klasycznej mediany), także w wersji uciętej (zob. np. [58], [67] czy [70]). Pozwala ono na wyzyskanie kompleksowych informacji o danym zjawisku złożonym, traktując zestaw cech jako integralną całość i biorąc tym samym pod uwagę również powiązania między cechami diagnostycznymi niezależnie od ich formalnej korelacji. Można też tworzyć formuły normalizacyjne z wykorzystaniem innych punktów osobliwych w przestrzeni wielowymiarowej (zob. [68]). W zakresie wyboru metody klasyfikacyjnej, warto oprzeć się na typologii P.H.A. Sneatha i R.R. Sokala ([89]), uściślonej przez M. Sobczyka ([90]) składającej się z ośmiu kryteriów, którymi należy się kierować podczas wyboru metody klasyfikacyjnej.

Podstawą sukcesu, czyli efektywnego uporządkowania badanych obiektów stanowi właściwe *rozpoczęcie procesu grupowania*. Na tym etapie wyróżniamy sposoby: aglomeracyjny (łączeniowy) oraz podziałowy (dedukcyjny). Pierwszy z nich zakłada, że każdy obiekt należący do klasyfikowanego zbioru stanowi odrębną grupę; w kolejnych krokach liczba grup (poprzez łączenie podobnych grup w jedną) ulega redukcji, aż pozostanie jedna grupa, zawierająca wszystkie klasyfikowane obiekty. W odwrotnym kierunku działają podejścia dedukcyjne. Tam grupa pełna jest dzielona sukcesywnie na części będące jej podzbiorami, aż do uzyskania grup będących wyłącznie jednoelementowymi podzbiorami zbioru obiektów.

Ustalenie kierunku przebiegu grupowania pociąga za sobą konieczność określenia *hierarchii grupowania*. Wyróżniamy tutaj metody hierarchiczne i niehierarchiczne. Charakterystyczną cechą metod hierarchicznych (np. taksonomia wrocławska ([28]) czy metoda Warda ([103]) są szczeble łączenia lub rozpadania się grup. Generowane grupy niższego rzędu zawierają w sobie rozłączne grupy poziomów niższych. Porządek tworzenia grup nie odgrywa tutaj znaczenia. Procedura niehierarchiczna prowadzi natomiast do uzyskiwania rozmaitych konfiguracji grup, które mogą zarówno zachodzić na siebie, jak również teoriomnogościowych skupień elementów niższego rzędu. Ze względu na ścisłość, wysoką efektywność oraz praktyczne znaczenie postępowania hierarchicznego, w naszych rozważaniach będziemy stosować tylko takie podejście.

W wyniku wyboru stosowanej inicjacji i hierarchizacji, kształtuje się odpowiednia *rozłączność podzbiorów*. Możemy więc uzyskać grupy rozłączne (nie zachodzące na siebie) oraz grupy nierozłączne (posiadające wspólne elementy). Ważna jest tutaj – jak i w innych tego typu sytuacjach – konsekwencja: jeśli stosujemy metodę grupowania rozłącznego, to na każdym jej etapie otrzymujemy cząstkowe grupowania w postaci podzbiorów rozłącznych. Inaczej może zdarzyć się w przypadku klasyfikacji nierozłącznej – tutaj najczęściej zachodzi konieczność ustalenia poziomu przynależności danego obiektu do konkretnego zbioru, a więc realizacja podstawowego założenia teorii zbiorów rozmytych.

Sposób przeprowadzania grupowania dzieli metody taksonomiczne na sekwencyjne (w których występują pewne powtarzające się ciągi operacji, np. podziału lub łączenia podzbiorów obiektów) oraz równoczesne (w których tego rodzaju powtórzeń nie ma).

Kryterium grupowania obiektów może być lokalne — wówczas optymalizacja zachodzi na każdym poziomie grupowania — lub też globalne — kryterium optymalizacji jest uniwersalne dla całego procesu taksonomicznego. Owa optymalizacja obejmuje zazwyczaj liczbę grup, podobieństwo wewnątrzgrupowe i zróżnicowanie międzygrupowe. Stosując dwa ostatnie mierniki zakładamy -– rzecz jasna — ustaloną liczbę grup.

Technika realizacji algorytmu pozwala wyróżnić procedury bezpośrednie i iteracyjne. Pierwsze poprzez jeden ciąg operacji prowadzą wprost do osiągnięcia określonego rodzaju klasyfikacji optymalnej w świetle przyjętych założeń, drugie — stosują kolejne przybliżenia optimum klasyfikacyjnego na drodze zmiany przynależności grupowej obiektów lub porządkowania skupisk.

Pod względem *wagi otrzymywanych podzbiorów* wyróżniamy procedury ważone i nieważone. W procedurach ważonych różnicujemy znaczenie oraz istotność grup wyznaczonych w danej klasyfikacji. W ten sposób wyróżniamy cechy bądź obiekty bardziej lub mniej ważne z punktu widzenia celów badania. Ważenie nie jest jednak zbyt zalecane przez statystyków.

Algorytm może zawierać procedury *uczenia się* i wówczas nazywamy go adaptacyjnym. W takim przypadku zakładamy możliwość zmiany algorytmu, w zależności od aktualnie uzyskiwanych rezultatów. Trzeba wówczas brać pod uwagę możliwość przeliczania struktury wag, tak dla cech, jak i obiektów. Najczęściej jednak spotykamy się z metodami nieadaptacyjnymi, pozwalającymi na bezpośrednie rozwiązywanie badanych problemów.

We wszystkich tych opcjach wykorzystuje się rozmaite formuły odległości obiektów (zob. np. [67]). W ostatnich latach pojawiło się sporo uogólnień

w tym zakresie – na przykład wykorzystanie metryki Minkowskiego w metodzie Warda ([54]). Istotną sprawą staje się także kwestia doboru wskaźników jakości grupowania. Istnieje cały ich szereg opartych na zgodności lub niezgodności klasyfikacji par obiektów (np. indeksy Randa czy Peirce'ego); prowadzone są badania i dyskusje na temat ich użyteczności (zob. np. [1]).

Nowatorskim kierunkiem badań w tym zakresie jest analiza danych symbolicznych (w tym głównie przedziałowych) w tym kontekście. Wprowadza się nowe metody grupowania, oparte na specyficznych odległościach takich danych czy specyficznych statystykach pozycyjnych (zob. np. [24], [71], [21], [22]). W statystyce ubóstwa może mieć to istotne znaczenie albowiem często właśnie zamiast liczbami posługujemy się tam przedziałami (przedziały ufności dla oszacowań, przedziały wartości cech diagnostycznych itp.).

Celem porządkowania obiektów jest natomiast budowa miernika kompleksowego, zwanego też *metacechą*, czyli uzyskanie jednostkowej zmiennej $\mu = (\mu_1, \mu_2, \ldots, \mu_n))$ posiadającej określone własności, która łączyłaby w sobie specyfikę poszczególnych cech diagnostycznych oraz jak najrzetelniej odzwierciedlałaby badane zjawisko widziane z całościowego postrzegania istoty rzeczy. Wyróżnia się tutaj metody bezwzorcowe i wzorcowe.

Z metodologicznego punktu widzenia, konstrukcje *mierników bezwzorcowych* opierają się na określonych funkcjach wielowymiarowych, których argumentami są wartości znormalizowanych cech diagnostycznych. Charakterystyczne bywa też podejście normalizacyjne — z określonych powodów konstrukcja danego typu miernika może preferować pewien ustalony rodzaj transformacji normalizacyjnej. Bywają też metody, w których wybór owego przekształcenia jest dalece swobodniejszy. Do najważniejszych rodzajów takich konstrukcji należą Jeden z najprostszych modeli konstrukcji metacechy stanowi metoda unitaryzacji zerowanej (np. K. Kukuła ([53], M. Wierzbiński i M. Sobolewski ([104])). Metoda średnich rang (np. [61]) polega na nadaniu rang każdej zestymulowanej zmiennej z osobna, po czym obliczeniu dla każdego obiektu średniej arytmetycznej tychże rang.

Jak wynika z samego określenia zagadnienia, podstawę koncepcji konstrukcji miernika wzorcowego stanowi pojęcie taksonomicznego wzorca rozwojowego. Jest to pewien sztucznie skonstruowany, de facto idealistycznie ukształtowany obiekt, charakteryzujący się pewnymi optymalnymi własnościami wyrażonymi odpowiednio określonymi funkcjami wartości poszczególnych cech diagnostycznych. Do stanu tego obiektu – wzorca w zakresie badanego zjawiska przyrównuje się sytuację w rzeczywistych analizowanych obiektach, ustalając jak bardzo każdy z nich odległy jest od tego poziomu, który w założeniu powinny one osiągać.

Najpowszechniejszą metodą konstrukcji taksonomicznego wzorca rozwoju jest metoda maksimum. Polega ona na zdefiniowaniu współrzędnych rzeczonego wzorca jako maksymalnych wartości poszczególnych jednorodnie zestymulowanych i znormalizowanych cech diagnostycznych. Następnie obliczamy odległości poszczególnych obiektów od wzorca, a potem na ich podstawie konstruujemy metacechę. Najpopularniejsze w tym zakresie jest podejście Z. Hellwiga ([40]), który przyjął jako miernik kompleksowy dopełnienie ilorazu odległości danego obiektu od wzorca i sumy średniej oraz dwukrotności odchylenia standardowego takich odległości do jedności. Jednak w ostatnich latach rozwinięto szereg metod alternatywnych, bazujących na różnych typach odległości i wiążących się z nią modyfikacją formuły Z. Hellwiga, np. wykorzystujące medianę Webera ([58], [68]). Bada się również możliwości zastosowania tej idei w analizie danych przedziałowych ([75]). Wzorzec taki może być w praktyce określony także w sposób egzogeniczny — poprzez zewnętrznie ustalone standardy, np. normy czy progi UE, co ma szczególne znaczenie właśnie w statystyce poziomu życia i ubóstwa.

Na podstawie wartości miernika kompleksowego obiekty zostają pogrupowane na cztery klasy. Stosowana tutaj najczęściej metoda grupowania nosi nazwę *metody trzech średnich*. Granicami klas są wówczas średnia arytmetyczna wartości miernika oraz średnie wartości miernika dla zbiorów, dla których wartość miernika jest mniejsza (odpowiednio większa) od średniej globalnej. Pozycyjnym odpowiednikiem tego podejścia jest *metoda trzech median*, w której średnia zastąpiona jest przez medianę właśnie. Inna metoda grupowania konstrukcję wartości progowych opiera na określaniu korzystnych i niekorzystnych progach wartości metacechy. Są one najczęściej określone poprzez sumę średniej arytmetycznej i odchylenia standardowego pomnożonego przez odpowiednią stałą (zazwyczaj -2, 0 i 2). W pozycyjnej odmianie tej metody zamiast średniej arytmetycznej stosujemy medianę, a zamiast odchylenia standardowego — medianowe odchylenie bezwzględne. Stałe, przez które mnożymy medianowe odchylenie bezwzględne to wówczas: 2,5, 0 i 2,5 (zob. [67]).

6.2 Podejście rozmyte (Fuzzy Approach)

To podejście oparte na teorii zbiorów rozmytych stanowi bardzo efektywne narzędzie do estymacji i badania przestrzennego zróżnicowania ubóstwa. Za prekursorów tego podejścia w tym kontekście uważa się Cerioli i Zani ([13]), którzy zastosowali je do analizy ubóstwa we włoskiej prowincji Parma. Póź-

niejszy wkład wnieśli tutaj też: Dagum i in. ([19]), Cheli i in. ([15]), Martinetti ([63]), Cheli i Lemmi ([17]), Betti and Verma ([7]), Betti i in. ([6]), Lemmi i Betti ([55]).

Teoria zbiorów rozmytych umożliwia pomiar relatywnego ubóstwa czy deprywacji każdego gospodarstwa domowego, szacowanie przeciętnej wartości wskaźnika ubóstwa dla całej populacji rozpatrywanych gospodarstw domowych oraz pomiar relatywnego udziału każdego komponentu w poziom ubóstwa ogółem. Dzięki temu możliwe staje się określenie najważniejszych cech lub wymiarów ubóstwa, co ma istotne znaczenie dla kształtowania polityki regionalnej w zakresie redukcji tego niekorzystnego zjawiska społeczno—ekonomicznego poprzez precyzyjnie – dzięki takiemu wskaźnikowi – celowane zmiany instytucjonalne, strukturalne, technologiczne, itp.

Teoria ta zakłada definiowanie zbioru ubogich (czyli prezentujących pewien stopień ubóstwa w zakresie przynajmniej jednej zmiennej objaśniającej) gospodarstw domowych, poziomu przynależności danego gospodarstwa domowego do tego zbioru (wyrażonego stosownym prawdopodobieństwem) oraz poziomów ubóstwa dla tego gospodarstwa domowego jak i dla całej populacji.

Przynależność i–tego gospodarstwa domowego a_i do zbioru gospodarstw ubogich ($i = 1, 2, \ldots, n$) ze względu na j–ty atrybut ($j = 1, 2, \ldots, m$) dana jest wzorem

$$\mu_j(a_i) = x_{ij}, 0 \leqslant x_{ij} \leqslant 1.$$

Wskaźnik ubóstwa i–tego gospodarstwa domowego, czyli poziom przynależności do zbioru gospodarstw ubogich definiowany jest jako średnia ważona prawdopodobieństw atrybutowych:

$$\mu(a_i) = \frac{\sum_{j=1}^{m} x_{ij} w_j}{\sum_{j=1}^{m} w_j},$$

gdzie w_j to waga przypisana cesze j. Dzięki takiej formule możemy mierzyć względną deprywację, poziom wykluczenia społecznego i inne wskaźniki społeczne. Waga może być definiowana rozmaicie (np. metodą delficką w oparciu o opinie eksperckie). Obrazuje ona intensywność deprywacji ze względu na daną cechę. Najczęściej jest to funkcja odwrotna do poziomu deprywacji pod danym względem w populacji — im mniejsza liczba gospodarstw, u których obserwuje się deprywację z uwagi na tę cechę, tym większa jest waga w_j. Podejście to reprezentuje propozycja Cerioli i Zani ([13]):

$$w_j = \log\left[\frac{n}{\sum_{i=1}^{n} x_{ij} n_i}\right] \geqslant 0,$$

gdzie $\sum_{i=1}^{n} x_{ij} n_j > 0$ i gdzie n_i jest wagą przyporządkowaną i–tej obserwacji z wylosowanej próbki, gdy dane pochodzą z badania reprezentacyjnego.

Indeks dla całej populacji ma postać:

$$\mu = \frac{\sum_{i=1}^{n} \mu(a_i) n_i}{\sum_{i=1}^{n} n_i}.$$

Cheli i Betti ([15]) ukazują jak można takie podejście wykorzystać do konstrukcji rozmytego wskaźnika ubóstwa monetarnego opartego na ekwiwalentnym dochodzie gospodarstw domowych (należącego do klasy uogólnionych wskaźników nierówności Giniego). Do estymacji rzeczonego wskaźnika stosują oni podejście m–kwantylowe.

Warto wspomnieć także o podejściu rozmytym w grupowaniu, które reprezentuje algorytm Ben–Israela i Iyiguna oparty na optymalizacji prawdopodobieństwa przynależności do skupień oraz środków ciężkości tychże skupień poprzez optymalizację funkcji celu wykorzystującej odległości obiektów od skupień i stosowny algorytm iteracyjny ([5], [46]).

6.3 Wykorzystanie macierzy sąsiedztwa

W badaniu przestrzennego zróżnicowania złożonych zjawisk społeczno – ekonomicznych, takich jak ubóstwo, istotną rolę odgrywa sąsiedztwo obszarów przestrzennych. Pojęcie „sąsiedztwo" zazwyczaj kojarzy się z ich bliskim położeniem w wymiarze fizycznym, tzn. w kontekście istniejących wspólnych granic wytyczonych na drodze realizacji decyzji administracyjnych lub uzgodnień politycznych. Założenia te stanowią także fundament funkcjonowania systemu badań, analiz i udostępniania danych statystycznych.

Oprócz sąsiedztwa rozumianego w sensie fizycznym, ważne są inne jego rodzaje, np. sąsiedztwo społeczno — gospodarcze. Bliskość obiektów ocenia się tutaj pod względem określonych zjawisk demograficznych, socjalnych, ekonomicznych i innych. Wspomnijmy w tym miejscu chociażby, że odległości pomiędzy miejscowościami coraz częściej mierzy się czasem, jaki jest potrzebny na pokonanie dzielącego je dystansu (np. w kontekście organizacji ważnych imprez i spotkań międzynarodowych; dla przykładu podajmy, że podczas przygotowań organizacyjnych do finałów Mistrzostw Europy w piłce nożnej, które w 2012 r. odbyły się w Polsce i na Ukrainie wymagano, aby bazy pobytowo — treningowe reprezentacji biorących udział w turnieju były zlokalizowane tak, by do najbliższego lotniska międzynarodowego dało się dojechać w czasie nie dłuższym niż jedna godzina). Tak więc dla kształtowania polityki rozwoju regionalnego sąsiedztwo społeczno — gospodarcze

jest równie ważne jak fizyczne (a czasem nawet istotniejsze). W niektórych państwach statystyka publiczna wychodzi naprzeciw temu zapotrzebowaniu. Na przykład w Wielkiej Brytanii od kilku lat istnieje system statystyki sąsiedztw, obejmujący bazę danych statystycznych dla różnorodnych obszarów przestrzennych stworzonych przez grupowanie podstawowych jednostek terytorialnych (np. rejonów administracyjnych, statystycznych, okręgów ochrony zdrowia, itp.) pod względem rozmaitych cech ekonomicznych i społecznych (zob. http://www.neighbourhood.statistics.gov.uk). Warto w tym kontekście wspomnieć także o szerszych strefach miejskich (ang. *Larger Urban Zone* — *LUZ*) obrazujących obszar funkcjonalnego oddziaływania miasta a wyznaczanych poprzez grupowanie gmin pod względem odsetka osób dojeżdżających z nich do pracy w danym mieście (zob. np. D. Rogalińska ([84]), A. Młodak ([69]), Józefowski i Młodak ([47])). To też jeden z ważniejszych czynników wpływających na zamożność ludności.

Konsekwencją tych tendencji stało się zainteresowanie badaczy metodologią tworzenia i analizy strukturalnej sąsiedztw. W ostatnich latach niektórzy specjaliści z zakresu wielowymiarowej analizy danych (jak np. zmarły w 2010 r. prof. dr hab. Wiesław Wagner — zob. [102]) sporo uwagi poświęcali macierzy sąsiedztwa, zwanej także macierzą bliskości (ang. *contiguity matrix*) obrazującej układ geograficznego sąsiedztwa obszarów przestrzennych wchodzących w skład określonego szerszego terytorium (np. gmin tworzących powiat, czy powiatów tworzących województwo), a także identyfikującej jednostki wewnętrzne i brzegowe takiego układu oraz ścieżki sąsiedztwa, tj. sposoby „przejścia" od jednej jednostki do drugiej przez kolejne jednostki sąsiadujące i ich wspólne granice bez wykraczania poza dany obszar nadrzędny. Tak skonstruowaną macierz wykorzystywano pomocniczo, np. w klasyfikacji obiektów pod kątem danego zjawiska społeczno – gospodarczego dokonywanej przy pomocy analizy skupień (zob. np. B. H. Wipperman ([106])). Szczegółową konceptualizację obu typów sąsiedztwa ze wskazaniem jego szczególnie ciekawych przypadków zawiera praca ([73]).

Ogólnie konstrukcję macierzy sąsiedztwa $\mathbf{W} = [w_{ij}]$ rozmiaru $n \times n$ (gdzie $n \in \mathbb{N}$ jest liczbą obiektów [1]), to każdy jej elementy dane są wzorem:

$$w_{ij} \stackrel{df}{=} \begin{cases} \eta_{ij} & \text{jeśli obiekty } i \text{ i } j \text{ sąsiadują ze sobą,} \\ 0 & \text{jeśli obiekty } i \text{ i } j \text{ nie są sąsiadami,} \end{cases} \tag{6.1}$$

przy czym $\eta_{ij} \in (0, 1]$ jest wielkością wyrażającą siłę sąsiedztwa, zwaną także *przestrzenną wagą sąsiedztwa*. Standardowo przyjmuje się tę wartość jako 1

[1] $\mathbb{N}$ oznacza zbiór liczb naturalnych

dla każdej pary sąsiadujących ze sobą obiektów. W innym przypadku wyższa wartość η_{ij} w (6.1) oznacza większą siłę sąsiedztwa obiektów i i j w rozpatrywanym sensie, $i, j = 1, 2, \ldots, n$.

Z praktycznego punktu widzenia w niektórych analizach stosuje się normalizację macierzy sąsiedztwa. W literaturze najbardziej znane są dwa podejścia w tym zakresie. Pierwsze z nich ([48]), opiera sie na maksymlanej weartosci własnej macierzy $\mathbf{W}$. Oznaczmy tę wartość przez τ. Wobec tego wszystkie elementy macierzy $\mathbf{W}$ są dzielone przez τ i w efekcie znormalizowana macierz sąsiedztwa będzie składała się z elementów równych 0 lub $1/\tau$. Inne podejście – sugerowane przez LeSage and Pace ([56]) i zwane standaryzacją wierszową -– oparte jest na standaryzacji wierszy macierzy sąsiedztwa. Oznacza to, że normalizujemy macierz w ten sposób, żeby suma niezerowych elementów w każdym wierszu była równa 1, czyli dzielimy każdy element przez sumę elementów w wierszu, do którego ona należy. Taka normalizacja staje się konieczna na w przypadku analizy regresji opartej na macierzy sąsiedztwa celem zapewnienia przestrzeni parametrycznej dla parametrów autoregresji przestrzennej (zob. [73]).

Oba typy sąsiedztwa były w ostatnich latach analizowane z różnych punktów widzenia. Kelejian and Prucha ([48]) badali przekrojowy autoregresyjny model przestrzenny (ang. *cross–sectional autoregressive spatial model – CAS*) będący dwurównaniowym modelem ekonometrycznym obejmującym zestaw zmiennych objaśniających i dwa różne typy sąsiedztwa wyrażone dwiema odmiennymi macierzami sąsiedztwa:

$$\boldsymbol{y} = \mathbf{W}\boldsymbol{\beta} + \lambda\mathbf{W}\boldsymbol{y} + \boldsymbol{u}, \tag{6.2}$$

$$\boldsymbol{u} = \gamma\mathbf{W}^{*}\boldsymbol{u} + \epsilon, \tag{6.3}$$

gdzie $\mathbf{W}$ i $\mathbf{W}^{*}$ oznaczają dwie różne macierze sąsiedztwa. Autorzy pracy ([48]) badali użyteczność modelu danego wzorami (6.2) i (6.3) w analizie współzależności zmiennych statystycznych proponując trzystopniową procedurę estymacji jego parametrów. Wagner i Mantaj ([102]) rozpatrywali z kolei matematyczne i praktyczne własności macierzy sąsiedztwa wymiarze fizycznym dochodząc do ciekawych rezultatów dotyczących cech kwadratu macierzy sąsiedztwa i jej kształtu dla ciągów sukcesywnych sąsiadów, a także jej zastosowanie w taksonomii. LeSage and Pace ([56]) zajmowali się użytecznością przestrzennego ekonometrycznego modelu Bayesowskiego, który uwzględnia sąsiedztwo obszarów. Ci sami badacze (LeSage and Pace ([57])) weryfikowali hipotezę o wrażliwości przestrzennego modelu regresyjnego (ang. *Spatial Regression Model – SAR*):

$$\boldsymbol{y} = \alpha\mathbf{1}(n) + \gamma\mathbf{W}\boldsymbol{y} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

gdzie $\mathbf{1}(n)$ oznacza wektor jedynek długości n, na poszczególne formy specyfikacji i badali jak można interpretować odpowiednie jego parametry. Młodak ([73]) koncentrował swą uwagę na praktycznych własnościach obu typów sąsiedztwa, zaproponował efektywną metodę ich porównywania oraz badał możliwości estymacji wariancji zakłóceń i wektorów opóźnień przestrzennych a także korelacji efektów i predykcji w modelach CAS i SAR. Z kolei studia przeprowadzone w pracy ([74]) dotyczyły możliwości wykorzystania zasobów informacyjnych jakie niosą macierze sąsiedztwa w grupowaniu obiektów przy użyciu podejść opartych na odległości.

7

Narzędzia informatyczne w estymacji ubóstwa

Narzędzia informatyczne są istotnym elementem umożliwiającym szacowanie modeli statystyki małych obszarów oraz wskaźników ubóstwa. Liczba dostępnych pakietów statystycznych jest szeroka, zawierają się w niej programy zarówno płatne jak i bezpłatne. Na potrzeby poniższego raportu wyselekcjonowano cztery pakiety statystyczne poświęcone bezpośrednio lub pośrednio tematyce ubóstwa, jak również statystyce małych obszarów. Większość prezentowanych narzędzi jest darmowa i wykorzystywana tak przez instytucje publiczne jak i prywatne. Do szczegółowego opisu zostały wybrane następujące narzędzia – PovMap , R, SAS oraz Stan. Pierwszy jest stworzony przez Bank Światowy i przeznaczony wyłącznie do tworzenia map ubóstwa w ramach projektu Mapowanie Ubóstwa z wykorzystaniem statystyki małych obszarów. Drugi pakiet - R – to darmowe środowisko do analiz statystycznych, w którym zaimplementowane zostały zarówno estymatory SMO, jak również oparte na nich miary ubóstwa (stopa ubóstwa, granica ubóstwa). Pakiet ten był wykorzystywany m.in. w projekcie AMELI oraz SAMPLE. Trzecim narzędziem prezentowanym w poniższym opracowaniu jest SAS tworzony od 1966 roku przez Jamesa Goodnighta i do tej pory rozwijany przez SAS Institute. SAS wykorzystywany jest przez wiele urzędów statystycznych na świecie, m.in. przez Główny Urząd Statystyczny czy Statistics New Zeland. W ramach projektu EURAREA SAS był wykorzystywany do tworzenia programów poświęconych statystyce małych obszarów (modele dla jednostki oraz obszaru). Ostatnim, a zarazem najmłodszym omawianym oprogramowaniem jest Stan tworzony przez zespół prof. Andrew Gelmana. Pakiet implementuje podejście bayesowskie do estymacji, które jest wykorzystane w statystyce małych obszarów (m.in. hierarchiczna estymacja bayesowska).

7.1 PovMap

PovMap jest jedynym oprogramowaniem dedykowanym mapowaniu ubóstwa, czyli nowo opracowanej metodzie szacowania poziomu dobrobytu i stopnia nierówności na niskich poziomach agregacji przestrzennej. Bazuje ono na procedurze zaproponowanej przez [26]. Metoda ta wykorzystuje dane z badania reprezentacyjnego gospodarstw domowych do oszacowania modelu dochodów/wydatków wykorzystując metody statystyki małych obszarów, a następnie stosuje go do danych pochodzących ze spisu lub rejestru gospodarstw domowych, które nie posiadają informacji o wysokości dochodów/wydatków. Oszacowane charakterystyki są następnie wykorzystywane do wyznaczania miar ubóstwa w małych domenach. W PovMap dostępne są następujące miary ubóstwa:

- Miary z rodziny FGT (stopa ubóstwa, głębokość ubóstwa, dotkliwość ubóstwa),
- Uogólniona entropia (ang. *generalised entropy*),
- Miary Atkinsona (ang. *Atkinson class of measures*),
- Miary Giniego (ang. *Gini Index*).

Oprogramowanie stworzone przez Bank Światowy implementuje podejście, w którym wydzielone są następujące etapy modelowania. W pierwszym kroku oprogramowanie umożliwia badanie zgodności wariantów oraz rozkładów zmiennych ze spisu/rejestru oraz z badania reprezentacyjnego. Krok ten jest ważny ponieważ model budowany na podstawie badania częściowego jest następnie aplikowany do spisu/rejestru. W tym zakresie PovMap ofertuje zarówno tabelaryczną, jak i wizualną ocenę rozkładów wraz z testowaniem ich zgodności. Oferuje również podstawowe operacje na danych, które umożliwią przetwarzanie zbiorów. Zmienne, które mają takie same definicje jak również rozkłady, są wykorzystywane w kolejnym etapie. W kroku drugim przeprowadza się modelowanie wydatków/dochodów z wykorzystaniem regresji liniowej z efektami stałymi i losowymi. Możliwe jest oszacowanie efektów losowych związanych z gospodarstwem domowym, jak również z wybranymi domenami. PovMap umożliwia w tym kroku diagnozowanie otrzymanego modelu oraz wizualizację otrzymanych wyników. Oszacowany model następnie w kolejnym, ostatnim kroku, jest aplikowany do danych pochodzących ze spisu powszechnego/rejestru administracyjnego. Otrzymane wyniki są następnie agregowane dla wybranych domen. W tym kroku przeprowadzane

jest także szacowanie wariancji estymatorów z wykorzystaniem metody bootstrap.

Aktualnie dostępna wersja pakietu PovMap 2.0 dostarcza dodatkowo podstawowe możliwości przetwarzania danych, porównywania rozkładów zmiennych, analizy wariancji oraz korelacji [112]. Dostępna jest również eksperymentalna wersja PovMap 2.1, która umożliwia dodatkowo mapowanie żywienia[1]. Należy zwrócić uwagę, że pakiet ogranicza się jedynie do zastosowania metodologii proponowanej przez [26] co w znacznym stopniu ogranicza możliwości porównania różnych metod do szacowania dochodów / wydatków gospodarstw domowych. Dodatkowo, pakiet wykorzystuje podejście dla jednostki (ang. unit-level, estymator EB), co wskazuje, że oprogramowanie można wykorzystać jedynie w przypadku posiadania dostępu do danych ze spisu powszechnego/rejestru, który można połączyć z badaniem reprezentacyjnym. Warto zaznaczyć, że oprogramowanie działa jedynie na systemie operacyjnym Windows, a dane pochodzące ze spisu/rejestru muszą zostać wyeksportowane do pliku (np. txt, sas7dbat) aby móc zaimportować lokalnie do program PovMap .

Do niewątpliwych zalet tego narzędzia należy zaliczyć szybkość działania programu, który został napisany w języku C++ oraz implementuje, stworzony specjalnie na potrzeby PovMap , sposób przetrzymywania danych. PovMap umożliwia przeprowadzanie obliczeń na dużych zbiorach danych, które są podstawą szacowania modelu dla jednostki. Jednocześnie pomimo "okienkowego" charakteru PovMap użytkownik otrzymuje raport aktywności wraz ze skryptami, które umożliwiają odtworzenie oraz powtarzalność prowadzonych analiz.

7.2 R

R bazuje na języku S, który powstał w 1976 roku, a jednym z jego twórców był John Chambers. W 1995 roku Ross Ihaka oraz Robert Gentlemen napisali pierwszą wersję R, która miała zostać przeznaczona do celów edukacyjnych. Od 1997 roku istnieje R Core Team, który kieruje rozwojem oraz bierze aktywny udział w rozwijaniu systemu. W tej grupie znajduje się również John Chambers, który tworzył język S. Oprócz R Core team istnieje wiele osób, które tworzą pakiety/biblioteki rozszerzające możliwości systemu (m.in. Hadley Wickham, Duncan Temple Lang, Dirk Eddelbuettel). W skrócie R jest środowiskiem, w którym są zaimplementowane metody statystyczne

[1] http://www.iresearch.worldbank.org/PovMap/

oraz metody analizy i wizualizacji danych, co podkreśla uniwersalność systemu. Obecnie dostępnych jest ponad 4000 pakietów, które umożliwiają rozszerzają podstawowe możliwości R.

Należy zauważyć, że R jest darmowym systemem udostępnianym na licencji GNU GPL-2, która pozwala na używanie języka R do dowolnych celów, w tym komercyjnych. Oznacza to, że możliwe jest wykorzystanie tego pakietu na potrzeby urzędów statystycznych, ministerstw czy ogólnie dla celów statystyki publicznej. Pakiety rozszerzające możliwości R pogrupowane są w grupy (tzw. task view), które zarządzane są przez specjalistów z danego zakresu tematycznego. W przypadku pakietów poświęconych estymacji z wykorzystaniem badań częściowych, statystyki małych obszarów lub innych tematów z zakresu statystyki publicznej stworzony został dział Official Statistics zarządzany przez prof. Matthiasa Templ'a z Politechniki Wiedeńskiej (Vienna University of Technology).

System R oferuje kilka pakietów poświęconych bezpośrednio problematyce statystyki małych obszarów oraz estymacji wskaźników ubóstwa – można tu wymienić m.in. **laeken, sae, rsae, hbsae** czy **mme**. Oprócz gotowych pakietów znaleźć można również programy poświęcone wybranym aspektom statystyki małych obszarów. Jednym z takich przykładów są programy przygotowane w ramach projektu SAMPLE na które składa się 12 programów implementujących rozwiązania zaproponowane w ramach tego projektu bezpośrednio związane z estymacją wskaźników ubóstwa dla małych domen.

Pierwszy pakiet -– **laeken** -– został stworzony w ramach projektu AMELI i zawiera implementację różnych wskaźników wykluczenia oraz ubóstwa z wykorzystaniem badania EU-SILC. Funkcja *arpr* umożliwia oszacowanie stopy ubóstwa wykorzystując estymację bezpośrednią, *arpt* natomiast pozwala określić wartość graniczna dla stopy ubóstwa. Pakiet umożliwia również szacowanie wariancji estymatora bezpośredniego metodą bootstrap (funkcja *variance* i *bootVar*) – podejście naiwne polegające na wielokrotnym pobieraniu próby przy określonym schemacie losowania oraz podejście wykorzystujące kalibrację (ang. *calibrated bootstrap*). Niestety, pakiet nie zawiera implementacji podejść znanych ze statystyki małych obszarów (model dla obszaru czy jednostki), jednak może zostać wykorzystany do oszacowania estymatorów bezpośrednich, które będą podstawą do dalszej pracy z pakietami poświęconymi statystyce małych obszarów. W kolejnych akapitach omówione zostaną pakiety poświęcone stricte podejściom znanym ze statystyki małych obszarów.

Pakiet **sae**, stworzony przez Isabel Molinę (Universidad Carlos III de Madrid) oraz Yolanda Marhuenda (Universidad Miguel Hernández de El-

che) umożliwia szacowanie charakterystyk o rozkładach ciągłych (np. dochodów/wydatków) z wykorzystaniem podejścia modelowego dla obszaru oraz jednostki. Zaimplementowane są następujące estymatory:

- Estymator bezpośredni – funkcja: *direct*,
- Empirical Best – funkcja *ebBHF* (model dla jednostki),
- EBLUP z wykorzystaniem modelu liniowego z zagnieżdżonym błędem – funkcja *eblupBHF*,
- EBLUP oparty na modelu Faya-Harriota – funkcja *eblupFH*,
- EBLUP oparty na modelu Faya-Harriota uwzględniający czynnik przestrzenny (macierz odległości, autokorelacji przestrzennej) – funkcja *eblupSFH*,
- EBLUP oparty na modelu Faya-Harriota uwzględniający czynnik przestrzenny oraz czasu – funkcja *eblupSTFH*.

Do estymacji modeli wykorzystywana jest metoda ML albo REML. Natomiast w przypadku estymacji błędu średniokwadratowego (MSE) wykorzystywane są dwa podejścia bootstrapowe – parametryczne (funkcje *pbmseBHF*, *pbmseebBHF*, *pbmseSFH*, *pbmseSTFH*) oraz nieparametryczne (funkcja *npbmseSFH*), a także algebraicznie (funkcje *mseFH*, *mseSFH*). Pakiet umożliwia szacowanie modelu dla jednostki metodą EB oraz obszaru (oparte na modelu Faya Harriota). Funkcja *ebBHF* umożliwia wskazanie szacowanego wskaźnika jako argumentu w postaci funkcji napisanej w języku R. Podstawą pakietu są funkcje zawarte w innych pakietach – nlme oraz MASS, które umożliwiają szacowanie modeli mieszanych.

Pakiet **hbsae** został stworzony przez Harma Jana Boonstrę ze Statistics Netherlands i umożliwia szacowanie charakterystyk dla małych obszarów z wykorzystaniem podejścia bayesowskiego. Pakiet implementuje dwa modele – dla obszaru (ang. *area-level model*) oraz jednostki (ang. *unit-level model*). Zaproponowane podejście wykorzystuje hierarchiczny estymator bayesowski oraz estymację z wykorzystaniem metody REML. Funkcja *fSAE* umożliwia deklarowanie modelu oraz jego typu, natomiast funkcje *fSAE.Area* i *fSAE.Unit* wykorzystywane są do oszacowania poszczególnych modeli, w pierwszym przypadku modelu Faya–Herriota, w drugim Battese-Hartera-Fullera. Funkcja *aggr* służy do agregacji wyników modelu dla jednostki lub obszaru

dla wskazanej, wyższej, domeny. Pakiet ten wykorzystywany jest przez Statistics Netherlands do estymacji dla małych obszarów dla badań rynku pracy (Dutch LFS survey).

Kolejny pakiet - **rsae** - poświęcony statystyce małych obszarów został stworzony przez Tobiasa Schocha (University of Applied Sciences Northwestern, Szwajcaria) i implementuje podejście odporne do estymacji modeli dla jednostki oraz obszaru [87]. Podstawą estymacji jest uwzględnienie klasy M-estymatorów, które w skrócie umożliwiają nadanie wag dla poszczególnych obserwacji, tak aby niwelować wpływ wartości wpływowych na oszacowania parametrów, jak również wariacji estymatorów. Pakiet implementuje podejście znane z M-estymatorów Hubera dając możliwość estymacji parametrów metodą LTS albo S-estymatorów (funkcja *fitsaemodel*). Natomiast aby oszacować poziom badanych zmiennych w określonych domenach wykorzystuje się w tym celu funkcję *robpredict*. Podobnie jak pakiet hbsae, rsae nie posiada zaimplementowanej funkcji do estymacji stopy ubóstwa jednak można wykorzystać oszacowania otrzymane z modelu do budowy granicy ubóstwa.

Ostatni pakiet w R poświęcony statystyce małych obszarów to **mme** autorstwa Lopez-Vizcaino M.E. (Instituto Galego de Estatística), Lombardio M.J. (Universidade da Coruña) oraz Moralesa D. (Universidad Miguel Hernández de Elche). Pakiet implementuje podejście uogólnionego modelu liniowego z efektami stałymi i mieszanymi w przypadku gdy zmienna objaśniana ma rozkład wielomianowy (ang. *multinomial mixed model*). Mme zawiera trzy funkcje (*modefit1*, *modelfit2* oraz *modelfit3*), które wykorzystywane są do szacowania modelu z dwoma efektami losowymi – dla grupy osób zatrudnionych oraz bezrobotnych. Model uwzględnia również autokorelację stopnia 1 (AR(1)) w celu uwzględnienia efektu czasu. Pakiet implementuje również dwie metody szacowania błędu średniokwadratowego wykorzystując podejście analityczne (definicja g1, g2, g3) oraz parametrycznego bootstrapu (założenie rozkładu wielomianowego dla zmiennej objaśnianej oraz normalnego dla efektów losowych). Szczegółowy opis pakietu można znaleźć w [60] oraz [65]

Ostatnim elementem opisu pakietu statystycznego R oraz jego użyteczności w szacowaniu stopy ubóstwa jest zbiór funkcji opracowanych w ramach projektu SAMPLE, który był realizowany w okresie od marca 2008 do marca 2001, w ramach 7 Programu Ramowego Badań i Rozwoju Technologicznego UE. W projekcie zaproponowany został szereg funkcji zawierających istniejące oraz nowe estymatory statystyki małych obszarów przeznaczone do szacowania stopy ubóstwa na niskich poziomach agregacji przestrzennej.

- Estymator Faya-Harriota dla obszaru uwzględniający efekt obszaru, czasu oraz efektu i czasu (funkcje *fitFH, fitSpatialFH, fitSpatioTemporalFH*)
- Estymator Faya-Harriota dla jednostki uwzględniający efekt czasu i autokorelację (funkcja *REML.individual.indep, BETA.U.individual.indep*, REML.autocorr)
- Estymatory oparte na M-kwantylach, które dodatkowo uwzględniają efekt przestrzenny przez Regresję Ważoną Geograficznie (funkcje *mq.sae, npmq.sae, mqgwr.sae*)
- Estymatory nieparametryczne oparte na M-kwantylach umożliwiające estymację rozkładu badanej cechy (funkcje *mq.sae.quant, npmq.sae.quant*)

W projekcie znajdują się również funkcje przeznaczone bezpośrednio do estymacji stopy ubóstwa.

- oparte na estymacji z wykorzystaniem M-kwantyli - *mq.sae.poverty*,
- oparte na metodzie EB – *FGTpovertyEB, PBMSE.EB, FGTpovertyEBsample*,
- oparte na modyfikacji algorytmu EB - *FastEB*.

7.3 SAS

W podstawowej wersji pakietu SAS nie ma procedur bezpośrednio związanych z estymacją wskaźników ubóstwa czy wykluczenia społecznego, jak również estymatorów statystyki małych obszarów. Niemniej jednak znajdują się w nim procedury bezpośrednio związane z estymacją modeli liniowych z efektami stałymi i losowymi, jak również specjalne makroprogramy napisane w ramach projektu EURAREA.

W SAS znajdują się następujące procedury służące do estymacji modeli wykorzystywanych w statystyce małych obszarów. PROC MIXED, która służy do szacowania modeli liniowych z efektami stałymi oraz losowymi, PROC GLIMMIX przeznaczony do estymacji uogólnionych modeli liniowych z efektami stałymi oraz losowymi, PROC NLMIXED stworzona do estymacji modeli nieliniowych z efektami stałymi oraz losowymi, PROC HPMIXED – nowa procedura przeznaczona do estymacji modeli liniowych z efektami stałymi i losowymi, zoptymalizowana pod względem przetwarzania dużych zbiorów danych oraz PROC MCMC, która umożliwia estymację

modeli bayesowskich z wykorzystaniem algorytmu MCMC. Pierwsza z wymienionych procedur (PROC MIXED) wykorzystywana jest w praktyce do estymacji modeli statystyki małych obszarów, których przykłady można znaleźć w [27], [66], a także projekcie EURAREA. Procedura MIXED, jak i pozostałe omawiane procedury, umożliwiają szacowanie modeli, które uwzględniają:

- efekty losowe dla obszaru (area-level model) oraz jednostki (unit-level model),
- efekty przestrzenne (autokorelację przestrzenną, geograficznie ważoną regresję) przez wykorzystanie m.in. macierzy sąsiedztwa,
- efekty dla badań powtarzalnych (badania panelowe, autokorelacja).

Procedura MIXED posiada bardzo rozbudowane możliwości deklaracji macierzy efektów losowych, co pozwala zdecydowanie rozszerzyć gamę szacowanych modeli uwzględniając różne efekty i budując modele hierarchiczne. Dodatkowym atutem SAS jest integracja z bazami danych oraz możliwość estymacji nawet w przypadku bardzo dużych zbiorów danych, które mogą nie mieścić się do pamięci komputera.

Program SAS był wykorzystywany również w projekcie EURAREA poświęconym estymatorom statystyki małych obszarów. Na potrzeby projektu oprogramowano następujące estymatory z wykorzystaniem procedur PROC MIXED oraz PROC IML. Ta druga jest specjalną procedurą przeznaczoną do tworzenia programów w języku macierzowym.

- EBLUPGREG – implementuje model dla jednostki uwzględniający efekty stałe oraz losowe (dla obszaru, przestrzenne oraz czasu). W zależności od zadeklarowanego modelu w wyniku zwracane są estymatory: bezpośredni, GREG, syntetyczny, EBLUP, EBLUP uwzględniający korelację przestrzenną, EBLUP uwzględniający autokorelacje w czasie oraz jego rozszerzenia,
- EBLUP dla szeregów czasowych – makra: EBLUP_TS uwzględniające zmianę efektu dla obszaru w czasie, MEBLUP_TS rozszerza EBLUP_TS o efekt losowy dla każdej próby,
- EBLUP dla uwzględniający autokorelację przestrzenną – makra: FISHERSCORMIX oraz FISHERSCORMIX2 do szacowania efektów losowych wynikających z deklaracji macierzy korelacji przestrzennej,
- SPREE, GLSM oraz GLSMM – estymacja modeli log-liniowych.

7.4 Stan

Stan jest pakietem statystycznym napisanym w języku C++ umożliwiającym estymację z wykorzystaniem metod bayesowskich. Pakiet został stworzony przez m.in. Andrew Gelmana (Columbia University), Boba Carpentera (Columbia University) oraz Matta Hoffmana (Adobe Creative Technologies Lab) finansowany z grantów U. S. Department of Energy, U. S. National Science Foundation oraz U. S. Department of Education Institute of Education Sciences (szczegóły na stronie http://mc-stan.org/team.html). Pakiet możliwy jest do uruchomienia m.in. w środowisku R (pakiet RStan [92]), Python (PyStan) czy Mathlab (MathlabStan). Działa na systemach Windows, Linux czy Mac OS oraz udostępniony jest na licencji BSD oraz GPLv3, co umożliwia jego wykorzystanie w dowolnym zakresie [93].

Pakiet Stan jest alternatywą dla programów typu BUGS czy JAGS, które są wykorzystywane do estymacji modeli z wykorzystaniem podejścia bayesowskiego. Pakiet opiera się na metodzie próbkowania MCMC (a dokładnie metodzie Monte Carlo Hamiltona), która sprawuje się zdecydowanie lepiej niż próbkowanie Gibbsa. Dodatkowo wykorzystuje nowo opracowaną metodę NUTS (No-U-Turn sampler), której szczegóły można znaleźć w [41]. Stan umożliwia estymację modeli statystyki małych obszarów wykorzystując hierarchiczną estymację bayesowską dla zmiennych dyskretnych oraz ciągłych. W jasny sposób rozszerza możliwości pakietu R poprzez integrację przez pakiet RStan, następnie przekazuje odpowiednie dane do funkcji stan, która dokonuje estymacji zadanego modelu.

```
deklarowany_model <- '
  data {
      // określenie zmiennych wejściowych
  }
  parameters {
    // określenie parametrów modelu
  }
  transformed parameters {
    // przekształcenie parametrów
  }
  model {
    // deklaracja modelu
  }
'
```

Podobnie jak SAS, pakiet Stan nie posiada funkcji bezpośrednio powiązanych ze statystyką małych obszarów czy estymacją stopy ubóstwa, jednak zawiera gamę modeli szacowanych metodami bayesowskimi, które są wykorzystywane w estymacji dla małych obszarów.

Podsumowanie

Zamieszczone w prezentowanym opracowaniu informacje pozwalają na podkreślenie istotnego miejsca w statystyce, jakie zajmują kwestie związane z analizą poziomu życia ludności oraz sfomułowanie kilku istotnych spostrzeżeń dotyczących możliwości badania ubóstwa i doskonalenia narzędzi służących temu celowi.

Po pierwsze, tematyka ta ma swe poczesne miejsce w szeregu projektów badawczych realizowanych przez różne instytucje i organizacje zajmujące się statystyką małych obszarów, w tym Eurostat. Ważne jest więc korzystanie z doświadczeń płynących z takich projektów oraz aktywne w nich uczestnictwo. Sporo istotnych z rozpatrywanego punktu widzenia zmiennych pozyskiwanych jest poza naszym krajem ze źródeł międzynarodowych (co uwidocznił np. paneuropejski projekt statystycznego monitoringu miast URBAN AUDIT), zatem bieżące korzystanie z tego rodzaju danych byłoby bardzo istotne. Inne projekty – takie jak EURAREA, SAMPLE czy AMELI – pozwoliły na dostarczenie nowoczesnych narzędzi estymacyjnych służących doskonaleniu jakości szacunków ubóstwa na niższych poziomach agregacji przestrzennej. Badania nad ich użytecznością w naszych warunkach aktualnie się toczą. Należałoby w tym miejscu nadmienić, że także po zakończeniu realizowanego w ramach ESSnet projektu MeMoBuSt (akronim od ang. *Methodology for Modern Business Statistics* – metodologia nowoczesnej statystyki działalności gospodarczej) rozważa się możliwość podjęcia podobnej inicjatywy dla statystyki społecznej (w tym ubóstwa), albowiem podręcznik metodologiczny, który powstał w wyniku tego przedsięwzięcia, zawiera na ogół uniwersalne zagadnienia, które mogą być łatwo zastosowane i rozwijane również w tym przypadku. Warto by zatem także doskonalić doświadczenia polskiej statystyki publicznej, w takich dziedzinach jak obciążenia odpowie-

dzi, kontrola ujawniania danych, imputacja, przetwarzanie odpowiedzi i im podobnych w badaniach społecznych typu Badanie Aktywności Ekonomicznej Ludności, Badanie Budżetów Gospodarstw Domowych, itp.

Po drugie, w naszym kraju istnieje bardzo dużo źródeł danych statystycznych – różnorodnych i zasobnych w rozmaite dane mogące być pomocne w szacowaniu ubóstwa. Są to zarówno spisy powszechne, jak i bazy informacji pochodzących z bieżących badań prowadzonych przez statystykę publiczną, a także rejestry administracyjne. Kluczową sprawą w tym kontekście byłoby, aby możliwie w jak największym stopniu wykorzystać potencjał informacyjny, który w nich się znajduje, co obecnie nie zawsze jeszcze jest możliwe. Dotyczy to w pierwszym rzędzie rejestru podatkowego POLTAX (jako 'zagłębia' wiedzy o dochodach gospodarstw domowych), ale także niezcyfryzowanych zasobów informacyjnych Agencji Restrukturyzacji i Modernizacji Rolnictwa czy wykorzystywanego przez urzędy gmin systemu LPIS (System Identyfikacji Działek Rolnych — ang. *Land Parcel Identification System*) – co mogłoby być szczególnie użyteczne w przypadku ocen poziomu ubóstwa na obszarach wiejskich.

Staytystyka dysponuje aktualnie różnorodnymi narzędziami teoretycznymi, służącymi do estymacji i badania przestrzennego zróżnicowania takich zjawisk jak ubóstwo. Są to przede wszystkim rozmaitej klasy estymatory: bezpośrednie, złożone i oparte na podejściu Faya– Herriota, regresyjne, itp. Nie sposób także pominąć w tym kontekście istotnej roli narzędzi wielowymiarowej analizy danych, które należałoby z dużą intensywnością w tym celu rozwijać. Klasyfikacja i porządkowanie obiektów wielocechowych czy wykorzystanie informacji o – rozmaicie rozumianym – sąsiedztwie poszczególnych obszarów to bardzo istotne komponenty jakości obrazu ubóstwa. Dlatego ich stosowanie i doskonalenie winno wejść do stałej praktyki statystyki publicznej.

Nie należy też zapominać o tym, że efektywnych rezultatów w zakresie badania ubóstwa nie byłoby, gdyby nie szerokie zaplecze profesjonalnych narzędzi informatycznych, które ma do dyspozycji współczesny statystyk. Są to przede wszystkim pakiety i procedury funkcjonujące w specjalistycznych narzędziach typu SAS i R. Sporo prac badawczych wykorzystuje także w tym celu potencjał pakietu STATISTICA. Jednak dwa pierwsze środowiska dają większe możliwości kreatywnego tworzenia nowych rozwiązań informatycznych w tym zakresie.

Wykonana przez autorów tego opracowania – reprezentujących różne obszary zainteresowań badawczych w dziedzinie statystyki – praca przeglądowa umożliwiła zatem wskazanie źródeł dalszego zgłębiania statystyki ubóstwa

w zakresie naukowym i praktycznym oraz zasygnalizowanie zagadnień problemowych, które wymagałyby ewentualnej dalszej pracy badawczej i intensyfikacji stosownych działań w tym zakresie – szczególnie w dostępie do danych i tworzenia narzędzi analitycznych. Wydaje się, że może wnieść to istotny wkład w rozwój polskiej statystyki publicznej.

Bibliografia

[1] Albatineh, A. (2010). Means and Variances for a Family of Similarity Indices Used in Cluster Analysis. *Journal of Statistical Planning and Inference*, 140:2828—2038.

[2] Alfons, A. and Templ, M. (2013). Estimation of Social Exclusion Indicators from Complex Surveys: The R Package laeken. *Journal of Statistical Software*, 54(15):1–25.

[3] Allison, P. D. (1994). Using Panel Data to Estimate the Effects of Events. *Sociological Methods & Research*, 23:174—199.

[4] Battese, G., Harter, R., and Fuller, W. (1988). An Error-Components Model for Prediction of County Crop Areas Using Survey and Satellite Data. *Journal of the American Statistical Association*, 83:28–36.

[5] Ben—Israel, A. and Iyigun, C. (2008). Probabilistic d–Clustering. *Journal of Classification*, 25:5–26.

[6] Betti, G., Cheli, B., and Gambini, R. (2004). A Statistical Model for the Dynamics Between Two Fuzzy States: Theory and an Application to Poverty Analysis. *Metron*, 62:391–411.

[7] Betti G., Verma, V. (1999). Measuring the Degree of Poverty in a Dynamic and Comparative Context: a Multidimensional Approach Using Fuzzy Set Theory. *Proceedings ICCS–VI, Lahore, Pakistan*, 11:289–301.

[8] Blicharz, J., Klat-Wertelecka, L., and Rutkowska-Tomaszewska, E. (2014). *Ubóstwo w Polsce*. E-Wydawnictwo, Prawnicza i Ekonomiczna Biblioteka Cyfrowa, Wydział Prawa, Administracji i Ekonomii Uniwersytetu

Wrocławskiego, Wrocław. http://www.bibliotekacyfrowa.pl/dlibra/publication?id=50393.

[9] Boonstra, H. J. (2012). *hbsae: Hierarchical Bayesian Small Area Estimation*. R package version 1.0.

[10] Breidenbach, J. (2013). *JoSAE: Functions for Some Unit-Level Small Area Estimators and Their Variances*. R package version 0.2.2.

[11] Brüderl, J. (2005). *Panel Data Analysis*. University of Mannheim, Mannheim, Germany. http://www2.sowi.uni-mannheim.de/lsssm/veranst/Panelanalyse.pdf.

[12] Bukowski, M. and Magda, I. (2013). *Zatrudnienie w Polsce 2011. Ubóstwo a praca*. Ministerstwo Pracy i Polityki Społecznej, Warszawa. http://www.mpips.gov.pl/download/gfx/mpips/pl/defaultaktualnosci/5543/6190/1/ubostwo_a_praca.pdf.

[13] Cerioli, A. and Zani, S. (1990). A Fuzzy Approach to the Measurement of Poverty. *[in:] C. Dagum, M. Zenga, (eds.), Income and Wealth Distribution, Inequality and Poverty, Springer Verlag, Berlin*, page 272—284.

[14] Chambers, R. and Tzavidis, N. (2006). M-Quantile Models for Small Area Estimation. *Biometrika*, 93:255–268.

[15] Cheli, B. and Betti, G. (1999). A Totally Fuzzy and Relative Measures of Poverty in Dynamics Context. 57:83–104.

[16] Cheli, B., Ghellini, G., Lemmi, A., and Pannuzi, N. (1994). Measuring Poverty in the Countries in Transition via TFR Method: the Case of Poland in 1990–1991. *Statistics in Transition*, 1:585–636.

[17] Cheli, B. and Lemmi, A. (1999). A Totally Fuzzy and Relative Approach to the Multidimensional Analysis of Poverty. *Economic Notes*, 24:115–134.

[18] Costa, M. and De Angelis, L. (2008). The Multidimensional Measurement of Poverty: a Fuzzy Set Approach. *Statistica*, LXVIII(3–4):303–319.

[19] Dagum, C., R. Gambassi, R., and A. Lemmi, A. (1992). New Approaches to the Measurement of Poverty. *Poverty Measurement for Economies in Transition in Eastern European Countries, Polish Statistical Association and Central Statistical Office, Warsaw*, pages 201–225.

[20] Daniels, N., Kennedy, B., and Kawachi, I. (2006). Health and Inequality, or, Why Justice is Good for Our Health. In *Public Health, Ethics, and Equity*. Oxford University Press.

[21] De Carvalho, F. A. T., Brito, P., and Bock, H.-H. (2006). Dynamic Clustering for Interval Data Based on L2—Distance. *Computational Statistics*, 21:231—250.

[22] De Souza, R. and De Carvalho, F. (2004). Clustering of Interval Data Based on City-Block Distances. *Pattern Recognition Letters*, 25:353—365.

[23] Dougherty, C. (2007). *An Introduction to Econometrics*. Oxford University Press, Oxford, UK., 3rd edition.

[24] Dudek, A. (2013). *Metody analizy danych symbolicznych w badaniach ekonomicznych*. Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu, Wrocław.

[25] EAPN (2007). *Ubóstwo i nierówności w Unii Europejskiej*. EAPN, Warszawa.

[26] Elbers, C., Lanjouw, J., and Lanjouw, P. (2003). Micro–Level Estimation of Poverty and Inequality. *Econometrica*, 71(1):355–364.

[27] Fabrizi, E., Ferrante, M. R., and Pacei, S. (2007). Small Area Estimation of Average Household Income Based on Unit-Level Models for Panel Data. *Survey Methodology*, 33(2):187–198.

[28] Florek, K., Łukaszewicz, J., Perkal, J., Steinhaus, H., and Zubrzycki, S. (1951). Taksonomia wrocławska. *Przegląd Antropologiczny*, XVII:193—211.

[29] Giusti, C., Pratesi, M., Tzavidis, N., and Salvati, N. (2011). Recent Advances in Small Area Methodologies for Poverty Indicators. *Journal of Statistical Planning and Inference*.

[30] Golinowska, S., Morecka, Z., Styrc, M., Cukrowska, E., and Cukrowski, J. (2007). *Od ubóstwa do wykluczenia społecznego. Badania. Koncepcja. Wyniki. Propozycje*. IPiSS, Warszawa.

[31] GUS (2009). *Informacja o wynikach badania przepływów ludności związanych z zatrudnieniem w Polsce*. Materiał na konferencję prasową w dniu 23 października 2009 r., Główny Urząd Statystyczny, Departament Pracy i Warunków Życia oraz Urząd Statystyczny w Po-

znaniu. http://www.stat.gov.pl/cps/rde/xbcr/gus/POZ_Infor_wyn_bad_przeplyw_ludno_zwi_zatrudnieniem_Polsce.pdf.

[32] GUS (2011). *SYSTEM REGON Rejestracja wniosków CEUDG-1 Założenia*. Główny Urząd Statystyczny, Warszawa, kwiecień 2011 r. http://stat.gov.pl/cps/rde/xbcr/cois/CEIDG_zalacznik_do_SIWZ_CIS_3_2011.pdf.

[33] GUS (2013a). *Jakość życia, kapitał społeczny, ubóstwo i wykluczenie społeczne w Polsce*. Zakład Wydawnictw Statystycznych, Warszawa. http://stat.gov.pl/download/cps/rde/xbcr/gus/WZ_jakosc_zycia_2013.pdf.

[34] GUS (2013b). *Systemy informacyjne administracji publicznej. Źródła danych dla badań statystyki publicznej*. Główny Urząd Statystyczny, Warszawa, maj 2013 r.

[35] GUS (2013c). *Ubóstwo w świetle badań GUS*. Zakład Wydawnictw Statystycznych, Warszawa. http://stat.gov.pl/download/gfx/portalinformacyjny/pl/defaultaktualnosci/5487/1/5/7/wz_ubostwo_w_polsce_2013.pdf.

[36] GUS (2014). *Dochody i Warunki Życia Ludności Polski (raport z badania EU-SILC 2012)*. Zakład Wydawnictw Statystycznych, Warszawa. http://stat.gov.pl/download/gfx/portalinformacyjny/pl/defaultaktualnosci/5486/6/5/3/wz_dochody_warunki_zycia_raport_2012.pdf.

[37] Haughton, J. and Khandker, S. R. (2009). *Handbook of Poverty and Inequality*. The World Bank, Waszyngton.

[38] Hecht, J. and Haye, E. M. (2009). Pooling vs. Panel Models of Leverage for American, Asian, and European Firms. *European Journal of Economics*, 15:94—105.

[39] Heij, C., de Boer, P., Franses, P. H., Kloek, T., and van Dijk, H. K. (2004). *Econometric Methods with Applications in Business and Economics*. Oxford University Press, Oxford, UK.

[40] Hellwig, Z. (1968). Zastosowanie metody taksonomicznej do typologicznego podziału krajów ze względu na poziom ich rozwoju oraz zasoby i strukturę wykwalifikowanych kadr. *Przegląd Statystyczny*, XV(4):307—327.

[41] Hoffman, M. D. and Gelman, A. (2011). The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *arXiv preprint arXiv:1111.4246*.

[42] Hryniewicka, A. and Herbst, J. (2010). *Analiza danych zebranych w systemie informatycznym POMOST za 2008 rok*. Instytut Rozwoju Służb Społecznych, Warszawa. http://irss.pl/wp-content/uploads/2011/06/analizadanychSIPOMOST2008.pdf.

[43] Ignaciuk, E. and Kątowski, T. (2006). Zachowania decyzyjne podmiotów gospodarczych. In *Rozwój mikroprzedsiębiorstw jako sposób wychodzenia z ubóstwa*, pages 107–116. Katedra Mikroekonomii, Uniwersytet Szczeciński, Szczecin. mikroekonomia.net/system/publication_files/884/original/10.pdf.

[44] Inclusion Europe (2009). *Ubóstwo i Niepełnosprawność Intelektualna w Europie*. Inclusion Europe, Bruksela.

[45] Isabel and Marhuenda, Y. (2013). *sae: Small Area Estimation*. R package version 1.0-2.

[46] Iyigun, C. and Ben—Israel, A. (2013). The Multi–Facility Location Problem: a Probabilistic Decomposition Method. *Computational Optimization and Applications*. www.optimizationonline.org/DB_FILE/2012/08/3555.pdf.

[47] Józefowski, T. and Młodak, A. (2009). Observation of Flows of Population in Polish Statistics – Problems and Challenges. *[in:] Elsner E., Michel H. (eds.), Assistance for the Younger Generation. Statistics and Planning, in Big Agglomerations. Institut für Angewandte Demographie IFAD, Berlin, Germany*, pages 61—-76.

[48] Keleijan, H. and Prucha, I. (1998). A Generalized Spatial Two-Stage Least Squares Procedure for Estimating a Spatial Autoregressive Model with Autoregressive Disturbances. *Journal of Real Estate Finance and Economics*, 17:99–121.

[49] Koenker, R. and Bassett, G. (1978). Regression Quantiles. *Econometrica*, 46:33–50.

[50] Koenker, R. and D'orey, V. (1987). Computing Regression Quantiles. *Journal of the Royal Statistical Society, Series C: Applied Statistics*, 36:383–389.

[51] Komitet Regionów (2011). *Polityka społeczna i prawa socjalne: walka z ubóstwem i wykluczeniem społecznym*. Unia Europejska, Bruksela.

[52] Kuan, C.-M. (2004). *Introduction to Econometric Theory*. Institute of Economics, Academia Sinica, China.

[53] Kukuła, K. (2000). *Metoda unitaryzacji zerowanej*. Wydawnictwo Naukowe PWN, Warszawa.

[54] Lee, A. and Willcox, B. (2014). Minkowski Generalizations of Ward's Method in Hierarchical Clustering. *Journal of Classification*, 31:194–218.

[55] Lemmi, A. and Betti, G. (2006). *Fuzzy Set Approach to Multidimensional Poverty Measurement*. Springer Verlag, Berlin.

[56] LeSage, J. and Pace, R. (2009). *Introduction to Spatial Econometrics*. Taylor & Francis Group, Boca Raton.

[57] LeSage, J. and Pace, R. (2010). The Biggest Myth in Spatial Econometrics. *Working Paper, Available at SSRN*. http://ssrn.com/abstract=1725503.

[58] Lira, J., Wagner, W., and Wysocki, F. (2002). Mediana w zagadnieniach porządkowania obiektów wielocechowych. *J. Paradysz (red.), Statystyka regionalna w służbie samorządu lokalnego i biznesu, Internetowa Oficyna Wydawnicza Centrum Statystyki Regionalnej, Akademia Ekonomiczna w Poznaniu, Poznań*, pages 97—99.

[59] Lopez-Vizcaino, E., Lombardia, M., and Morales, D. (2014). *mme: Multinomial Mixed Effects Models*. R package version 0.1-5.

[60] López-Vizcaíno, E., Lombardía, M. J., and Morales, D. (2013). Multinomial-Based Small Area Estimation of Labour Force Indicators. *Statistical Modelling*, 13(2):153–178.

[61] Malina, A. and Wanat, S. (1995). Przestrzenna analiza rozwoju Polski. *Wiadomości Statystyczne*, XL(5):20—25.

[62] Malina, A. and Zeliaś, A. (1998). On Building Taxonometric Measures on Living Conditions. *Statistics in Transition*, 3:523–544.

[63] Martinetti, E. (1994). A New Approach to the Evaluation of Well-Being and Poverty by Fuzzy Set Theory. *Giornale Degli Economisti e Annali di Economia*, 53:367–388.

[64] Molina, I., Nandram, B., and Rao, J. (2014). Small Area Estimation of General Parameters with Application to Poverty Indicators: a Hierarchical Bayes Approach. *The Annals of Applied Statistics*, 8(2):852–885.

[65] Molina, I., Saei, A., and Jose Lombardia, M. (2007). Small Area Estimates of Labour Force Participation Under a Multinomial Logit Mixed Model. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170(4):975–1000.

[66] Mukhopadhyay, P. K. and McDowell, A. (2011). *Small Area Estimation for Survey Data Analysis Using SAS Software*.

[67] Młodak, A. (2006a). *Analiza taksonomiczna w statystyce regionalnej*. Centrum Doradztwa i Informacji DIFIN, Warszawa.

[68] Młodak, A. (2006b). Multilateral Normalisations of Diagnostic Features. *Statistics in Transition*, 7:1125–1139.

[69] Młodak, A. (2008a). Polish Experiences and Possibilities in Realisation of the URBAN AUDIT programme. *[w:] J. Dziechciarz (red.) "Globalization Impact on Regional and Urban Statistics", Proceedings from the* 25^{th} *SCORUS Conference on Regional and Urban Statistics and Research, Publishing House of the Wrocław University of Economics, Wrocław.*

[70] Młodak, A. (2008b). Zróżnicowanie kapitału ludzkiego na rynku pracy – analiza taksonomiczna. *Wiadomości Statystyczne*, LIV(11):53—69.

[71] Młodak, A. (2011). Classification of Multivariate Objects Using Interval Quantile Classes. *Journal of Classification*, 28:327—362.

[72] Młodak, A. (2012). Statystyka metropolii polskich – problemy i perspektywy. *Studia Regionalne i Lokalne*, 48(2):20–38. www.studreg.uw.edu.pl/pdf/2012_2_mlodak.pdf.

[73] Młodak, A. (2013). Neighbourhood of Spatial Areas in the Physical and Socio-Economical Context. *Computational Statistics*, 28:2379 —- 2414.

[74] Młodak, A. (2014a). *Distance—Based Clustering Using a Neighbourhood Matrix*. unpublished manuscript.

[75] Młodak, A. (2014b). On the Construction of an Aggregated Measure of the Development of Interval Data. *Computational Statistics*, 29:895—929.

[76] on Estimates of Poverty for Small Geographic Areas, P. and Council, N. R. (1997). Small-Area Estimates of School-Age Children in Pinterim Report 1: Evaluation of 1993 County Estimates for Title I Allocations.

[77] on National Statistics, C. and Council, N. R. (1998). Small-Area Estimates of School-Age Children in Poverty, Interim Report 2: Evaluation of Revised 1993 County Estimates for Title I Allocations.

[78] on National Statistics, C., Council, N. R., Citro, C., and Kalton, G. E. (1999). Small-Area Estimates of School-Age Children in Poverty, Interim Report 3: Evaluation of 1995 County and School District Estimates for Title I Allocations.

[79] Prasad, N. and Rao, J. (1999). On Robust Small Area Estimation Using a Simple Random Effects Model. *Survey Methodology*, 25:67–72.

[80] R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[81] Raczkowska, M. (2012). Ekskluzja społeczna na obszarach wiejskich w Polsce. *Roczniki Ekonomii Rolnictwa i Rozwoju Obszarów Wiejskich*, 99(4):49–55. www.wne.sggw.pl/czasopisma/pdf/RNR_2012_T99_z4_s49.pdf.

[82] Rao, J. (2003). *Small Area Estimation*. Wiley Series in Survey Methodology.

[83] Roberto, E. (2008). *Commuting to Opportunity: The Working Poor and Commuting in the United States*. Brookings, Waszyngton. http://web.stanford.edu/group/scspi/_media/pdf/key_issues/transportation_policy.pdf.

[84] Rogalińska, D. (2007). Wykorzystanie danych ze źródeł administracyjnych w statystyce miast. *Wiadomości Statystyczne*, LII(2):42–67.

[85] Roszka, W. (2012). System statystyki publicznej oparty na zintegrowanych źródłach danych. *Wiadomości Statystyczne*, (numer specjalny 2):205–221. http://keii.ue.wroc.pl/przeglad/Rok%202012/Zeszyt%20Specjalny%202/2012_spec_2_205-221.pdf.

[86] Schiller, B. R. (1998). *The Economics of Poverty and Discrimination*. Upper Saddle River: Prentice Hall.

[87] Schoch, T. (2012). Robust Unit-Level Small Area Estimation: a Fast Algorithm for Large Datasets. *Austrian Journal of Statistics*, 41(4):243–265.

[88] Schoch, T. (2014). *rsae: Robust Small Area Estimation.* R package version 0.1-5.

[89] Sneath, P. and Sokal, R. (1973). *Numerical Taxonomy.* ed. by W. H. Freeman and Company, San Francisco, USA.

[90] Sobczyk, M. (1995). Wybrane zagadnienia taksonomii numerycznej. *[w:] Rozwój metodologii badań statystycznych w Polsce, Seria: Biblioteka Wiadomości Statystycznych, Główny Urząd Statystyczny i Polskie Towarzystwo Statystyczne, Warszawa*, 44:80—100.

[91] Sobieszak, A. and Szałtys, D. (2007). Wykorzystanie rejestrów administracyjnych do Spisów Ludności i Mieszkań w niektórych krajach. *Wiadomości Statystyczne*, (5):67–80.

[92] Stan Development Team (2014a). RStan: the R interface to Stan, version 2.4.

[93] Stan Development Team (2014b). *Stan Modeling Language Users Guide and Reference Manual, Version 2.4.*

[94] Statistics Netherlands (2003). *URBAN AUDIT II. The Implementation in Netherlands.* Division of Social and Spatial Statistics, Department of Statistical Analysis, Voorburg, The Hague, The Netherlands.

[95] Stukel, D. M. and Rao, J. (1999). On Small-Area Estimation Under Two-Fold Nested Error Regression Models. *Journal of Statistical Planning and Inference*, 78(1):131–147.

[96] Szarfenberg, R. (2012). Ubóstwo i wykluczenie społeczne w Polsce — pomiar, wyjaśnianie, strategie przeciwdziałania. rszarf.ips.uw.edu.pl/pdf/uiws2012a.pdf [dostęp: 14.10.2014].

[97] Szałtys, D. and Stępień, R. (2011). Narodowy Spis Powszechny Ludności i Mieszkań w 2011 r. *Wiadomości Statystyczne*, (11):11–25.

[98] Tarkowska, E. (2013). *Ubóstwo dzieci w Polsce.* EAPN, Warszawa.

[99] Tzavidis, N. and Brown, J. J. (2010). Using M-Quantile Models as an Alternative to Random Effects to Model the Contextual Value-Added of Schools in London. *Department of Quantitative Social Science, Institute of Education, University of London, UK, DoQSS Working Paper No. 10-11.*

[100] Tzavidis, N., Marchetti, S., and Chambers, R. (2010). Robust Estimation of Small Area Means and Quantiles. *Australian and New Zealand Journal of Statistics*, 52:167–186.

[101] Urząd Statystyczny w Warszawie (2010). Krajowy System Monitoringu Pomocy Społecznej. http://www.trendyrozwojowemazowsza.pl/zrodlo/sac.

[102] Wagner, W. and Mantaj, A. (2010). Contiguity Matrix of Spatial Units and its Properties on Example of Land Districts of Podkarpackie Voivodship. *Statistics in Transition – New Series*, 11:187–203.

[103] Ward, J. (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, 58:236—244.

[104] Wierzbiński, M. and M., S. (2002). Klasyfikacja powiatów województwa podkarpackiego ze względu na poziom życia ludności. *[w:] K. Jajuga i M. Walesiak (red.) Taksonomia 9. Klasyfikacja i analiza danych – teoria i zastosowania. Seria: Prace Naukowe Akademii Ekonomicznej im. Oskara Langego we Wrocławiu, Wrocław*, 942:84—96.

[105] Wilak, K. (2014). Autokorelacja błędów oszacowań w Badaniu Aktywności Ekonomicznej Ludności. *Wiadomości Statystyczne*. Maszynopis złożony do publikacji.

[106] Wipperman, B. (2004). *Hierarchical Agglomerative Cluster Analysis with a Contiguity Constraint*. Simon Fraser University, Burnaby – Surrey – Vancouver, Canada. A project submitted in partial fulfilment of the requirements for the degree of Master of Science in the Department of Statistics and Actuarial Science, Simon Fraser University, Canada.

[107] Wooldridge, J. M. (2002). *Econometric Analysis of Cross–Section and Panel Data*. MIT Press, Cambridge, Massachusetts, USA.

[108] Wooldridge, J. M. (2009). *Introductory Econometrics: A Modern Approach*. South – Western Cengage Learning Inc., Mason, U. S. A.

[109] Wóycicka, I. (2007). Walka z ubóstwem wśród dzieci oraz promocja ich integracji społecznej. *Insytut Badań nad Gospodarką Rynkową, Warszawa.* ec.europa.eu/social/BlobServlet?docId=5164&langId=pl.

[110] Wóycicka, I. (2011). Czy zasiłki z pomocy społecznej skutecznie redukują ubóstwo dochodowe. In *Różne wymiary skuteczności w pomocy społecznej*. Insytut Rozwoju Służb Społecznych, Warszawa. http://irss.pl/wp-content/uploads/2011/06/R%C3%B3%C5%BCne-wymiary-suteczno%C5%9Bci-w-pomocy-spo%C5%82ecznej.pdf.

[111] Zeliaś, A. (2002). Some Notes on the Selection of Normalization of Diagnostic Variables. *Statistics in Transition*, 5:787—802.

[112] Zhao, Q. and Lanjouw, P. (2014). *Using PovMap2, A User's Guide.*

[113] Zimmer, S., J., K., and Nusser, A. (2012). A Hierarchical Clustering Algorithm for Multivariate Stratification in Stratified Sampling. *Proceedings of the Survey Research Methods Section of the American Statistical Association.* https://www.amstat.org/sections/srms/proceedings/y2012/files/306416_76476.pdf.

[114] Śmiłowska T. (1997). Statystyczna analiza poziomu życia ludności Polski w ujęciu przestrzennym. *Studia i Prace. Z Prac Zakładu Badań Statystyczno – Ekonomicznych Głównego Urzędu Statystycznego i Polskiej Akademii Nauk, Zeszyt 247, Warszawa.*