PRACE NAUKOWE
Uniwersytetu Ekonomicznego we Wrocławiu

RESEARCH PAPERS
of Wrocław University of Economics

Nr 426

Taksonomia 26
Klasyfikacja i analiza danych – teoria i zastosowania

Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2016

Redaktor Wydawnictwa: Agnieszka Flasińska

Redaktor techniczny: Barbara Łopusiewicz

Korektor: Barbara Cibis

Łamanie: Beata Mazur

Projekt okładki: Beata Dębska

Tytuł dofinansowany ze środków Narodowego Banku Polskiego
oraz ze środków Sekcji Klasyfikacji i Analizy Danych PTS

Informacje o naborze artykułów i zasadach recenzowania
znajdują się na stronach internetowych
www.pracenaukowe.ue.wroc.pl
www.wydawnictwo.ue.wroc.pl

ISSN 1899-3192 (Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu)
e-ISSN 2392-0041
ISSN 1505-9332 (Taksonomia)

Wersja pierwotna: publikacja drukowana

Zamówienia na opublikowane prace należy składać na adres:
Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
ul. Komandorska 118/120, 53-345 Wrocław
tel./fax 71 36 80 602; e-mail:econbook@ue.wroc.pl
www.ksiegarnia.ue.wroc.pl

Druk i oprawa: TOTEM

Spis treści

Wstęp

W dniach 14–16 września 2015 r. w Hotelu Novotel Gdańsk Marina w Gdańsku odbyła się XXIV Konferencja Naukowa Sekcji Klasyfikacji i Analizy Danych PTS (XXIX Konferencja Taksonomiczna) „Klasyfikacja i analiza danych – teoria i zastosowania", zorganizowana przez Sekcję Klasyfikacji i Analizy Danych Polskiego Towarzystwa Statystycznego oraz Katedrę Statystyki Wydziału Zarządzania Uniwersytetu Gdańskiego. Przewodniczącymi Komitetu Organizacyjnego konferencji byli prof. dr hab. Mirosław Szreder oraz dr hab. Krzysztof Najman, prof. nadzw. UG, sekretarzami naukowymi dr hab. Kamila Migdał-Najman, prof. nadzw. UG oraz dr hab. Anna Zamojska, prof. nadzw. UG, a sekretarzem organizacyjnym Anna Nowicka z Fundacji Rozwoju Uniwersytetu Gdańskiego.

Konferencja Naukowa została dofinansowana ze środków Narodowego Banku Polskiego.

Zakres tematyczny konferencji obejmował takie zagadnienia, jak:

a) teoria (taksonomia, analiza dyskryminacyjna, metody porządkowania liniowego, metody statystycznej analizy wielowymiarowej, metody analizy zmiennych ciągłych, metody analizy zmiennych dyskretnych, metody analizy danych symbolicznych, metody graficzne),

b) zastosowania (analiza danych finansowych, analiza danych marketingowych, analiza danych przestrzennych, inne zastosowania analizy danych – medycyna, psychologia, archeologia, itd., aplikacje komputerowe metod statystycznych).

Zasadniczymi celami konferencji SKAD były prezentacja osiągnięć i wymiana doświadczeń z zakresu teoretycznych i aplikacyjnych zagadnień klasyfikacji i analizy danych. Konferencja stanowi coroczne forum służące podsumowaniu obecnego stanu wiedzy, przedstawieniu i promocji dokonań nowatorskich oraz wskazaniu kierunków dalszych prac i badań.

W konferencji wzięło udział 81 osób. Byli to pracownicy oraz doktoranci następujących uczelni i instytucji: AGH w Krakowie, Politechniki Łódzkiej, Politechniki Gdańskiej, Politechniki Opolskiej, Politechniki Wrocławskiej, Szkoły Głównej Gospodarstwa Wiejskiego w Warszawie, Szkoły Głównej Handlowej w Warszawie, Uniwersytetu im. Adama Mickiewicza w Poznaniu, Uniwersytetu Ekonomicznego w Katowicach, Uniwersytetu Ekonomicznego w Krakowie, Uniwersytetu Ekonomicznego w Poznaniu, Uniwersytetu Ekonomicznego we Wrocławiu, Uniwersytetu Gdańskiego, Uniwersytetu Jana Kochanowskiego w Kielcach, Uniwersytetu Łódzkiego, Uniwersytetu Mikołaja Kopernika w Toruniu, Uniwersytetu Przyrodniczego w Poznaniu, Uniwersytetu Szczecińskiego, Uniwer-

sytetu w Białymstoku, Wyższej Szkoły Bankowej w Toruniu, a także przedstawiciele NBP i PBS Sp. z o.o.

W trakcie dwóch sesji plenarnych oraz trzynastu sesji równoległych wygłoszono 58 referatów poświęconych aspektom teoretycznym i aplikacyjnym zagadnienia klasyfikacji i analizy danych. Odbyła się również sesja plakatowa, na której zaprezentowano 14 plakatów. Obradom w poszczególnych sesjach konferencji przewodniczyli profesorowie: Józef Pociecha, Eugeniusz Gatnar, Tadeusz Trzaskalik, Krzysztof Jajuga, Marek Walesiak, Barbara Pawełek, Feliks Wysocki, Ewa Roszkowska, Andrzej Sokołowski, Andrzej Bąk, Tadeusz Kufel, Mirosław Krzyśko, Krzysztof Najman, Małgorzata Rószkiewicz, Mirosław Szreder.

Teksty 25 recenzowanych artykułów naukowych stanowią zawartość prezentowanej publikacji z serii „Taksonomia" nr 26. Pozostałe recenzowane artykuły znajdują się w „Taksonomii" nr 27.

W pierwszym dniu konferencji odbyło się posiedzenie członków Sekcji Klasyfikacji i Analizy Danych Polskiego Towarzystwa Statystycznego, któremu przewodniczył prof. dr hab. Józef Pociecha. Ustalono plan przebiegu zebrania obejmujący następujące punkty:

A. Sprawozdanie z działalności Sekcji Klasyfikacji i Analizy Danych PTS.
B. Informacje dotyczące planowanych konferencji krajowych i zagranicznych.
C. Organizacja konferencji SKAD PTS w latach 2016 i 2017.
D. Wybór przedstawiciela Rady Sekcji SKAD PTS do IFCS.
E. Dyskusja nad kierunkami rozwoju działalności Sekcji.

Prof. dr hab. Józef Pociecha otworzył posiedzenie Sekcji SKAD PTS. Sprawozdanie z działalności Sekcji Klasyfikacji i Analizy Danych PTS przedstawiła sekretarz naukowy Sekcji dr hab. Barbara Pawełek, prof. nadzw. UEK. Poinformowała, że obecnie Sekcja liczy 231 członków. Przypomniała, że na stronie internetowej Sekcji znajdują się regulamin, a także deklaracja członkowska. Poinformowała, że zostały opublikowane zeszyty z serii „Taksonomia" nr 24 i 25 (PN UE we Wrocławiu nr 384 i 385). W „Przeglądzie Statystycznym" (zeszyt 4/2014) ukazało się sprawozdanie z ubiegłorocznej konferencji SKAD, która odbyła się w Międzyzdrojach, w dniach 8–10 września 2014 r. Prof. Barbara Pawełek przedstawiła także informacje dotyczące działalności międzynarodowej oraz udziału w ważnych konferencjach członków i sympatyków SKAD.

W konferencji Międzynarodowego Stowarzyszenia Towarzystw Klasyfikacyjnych (IFCS – International Federation of Classification Societies) w dniach 6–8 lipca 2015 r. w Bolonii, zorganizowanej przez Università di Bologna, udział wzięło 19 osób z Polski (w tym 17 członków Sekcji), które wygłosiły 15 referatów (wkład członków SKAD – 79,0%). Ponadto prof. Józef Pociecha był członkiem Komitetu Naukowego Konferencji z ramienia SKAD, członkiem Międzynarodowego Komitetu Nagród IFCS oraz organizatorem i przewodniczącym sesji nt. „Classification models for forecasting of economic processes".

W konferencji „European Conference on Data Analysis" (Colchester, 2–4 września 2015 r.) zorganizowanej przez The German Classification Society (GfKl) we współpracy z The British Classification Society (BCS) i Sekcją Klasyfikacji i Analizy Danych PTS (SKAD) udział wzięło 18 osób z Polski (w tym 14 członków Sekcji), które wygłosiły 15 referatów (wkład członków SKAD – 66,0%). Ponadto profesorowie Krzysztof Jajuga oraz Józef Pociecha byli członkami Komitetu Naukowego konferencji, prof. Andrzej Dudek został poproszony przez organizatorów o przygotowanie referatu i wygłoszenie na Sesji Plenarnej „Cluster analysis in XXI century, new methods and tendencies", prof. Krzysztof Jajuga był przewodniczącym sesji plenarnej, przewodniczącym sesji nt. „Finance and economics II" oraz organizatorem i przewodniczącym sesji nt. „Data analysis in finance", prof. Józef Pociecha był organizatorem i przewodniczącym sesji nt. „Outliers in classification procedures – theory and practice", prof. Andrzej Dudek był przewodniczącym sesji nt. „Machine learning and knowledge discovery II".

Kolejny punkt posiedzenia Sekcji obejmował zapowiedzi najbliższych konferencji krajowych i zagranicznych, których tematyka jest zgodna z profilem Sekcji. Prof. dr hab. Józef Pociecha poinformował o dwóch wybranych konferencjach krajowych (były to XXXIV Konferencja Naukowa „Multivariate Statistical Analysis MSA 2015", Łódź, 16–18 listopada 2015 r. i X Międzynarodowa Konferencja Naukowa im. Profesora Aleksandra Zeliasia nt. „Modelowanie i prognozowanie zjawisk społeczno-gospodarczych", Zakopane, 10–13 maja 2016 r.) oraz o trzech wybranych konferencjach zagranicznych. Konferencja „European Conference on Data Analysis" odbędzie się na Uniwersytecie Ekonomicznym we Wrocławiu w dniach 26–28 września 2017 r. W przeddzień tej konferencji, tj. 25.09.2017 r., odbędzie się Niemiecko-Polskie Sympozjum nt. „Analizy danych i jej zastosowań GPSDAA 2017". Następna konferencja Międzynarodowego Stowarzyszenia Towarzystw Klasyfikacyjnych (IFCS) odbędzie się w 2017 r. w Tokio. W 2019 r. Niemiecko-Polskie Sympozjum nt. „Analizy danych i jej zastosowań GPSDAA 2019" organizuje prof. Andreas Geyer-Schultz w Karlsruhe.

W następnym punkcie posiedzenia podjęto kwestię organizacji kolejnych konferencji SKAD. SKAD 2016 zorganizuje Katedra Metod Statystycznych Wydziału Ekonomiczno-Socjologicznego Uniwersytetu Łódzkiego.

W kolejnej części zebrania dokonano wyboru przedstawiciela Rady Sekcji SKAD PTS do IFCS na kadencję 2016–2019. Powołano Komisję Skrutacyjną, której przewodniczącym został prof. Tadeusz Kufel, a członkami dr hab. Iwona Konarzewska i dr Dominik Rozkrut. Profesor Józef Pociecha poprosił zebranych o proponowanie kandydatur zgłaszając jednocześnie prof. Andrzeja Sokołowskiego. Wobec braku następnych kandydatur listę zamknięto. Komisja Skrutacyjna przeprowadziła głosowanie tajne. W głosowaniu uczestniczyło 41 członków Sekcji. Profesor Andrzej Sokołowski został przedstawicielem Rady Sekcji SKAD PTS do

IFCS na kadencję 2016–2019, uzyskując następujący wynik: 39 głosów na „tak", 1 głos na „nie", 1 głos był nieważny.

W ostatnim punkcie zebrania dyskutowano nad kierunkami rozwoju działalności Sekcji obejmującymi następujące problemy: udział w międzynarodowym ruchu naukowym (wspólne granty, publikacje), umiędzynarodowienie konferencji SKAD (uczestnicy zagraniczni, dwujęzyczność konferencji), wydawanie własnego czasopisma.

Profesor Józef Pociecha zamknął posiedzenie Sekcji SKAD.

Krzysztof Jajuga, Marek Walesiak

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS nr 426 • 2016

Taksonomia 26 ISSN 1899-3192
Klasyfikacja i analiza danych – teoria i zastosowania e-ISSN 2392-0041

Wojciech Roszka
Uniwersytet Ekonomiczny w Poznaniu
e-mail: wojciech.roszka@gmail.com

SYNTETYCZNE ŹRÓDŁA DANYCH W ANALIZIE PRZESTRZENNEGO ZRÓŻNICOWANIA UBÓSTWA

SYNTHETIC DATA SOURCES IN SPATIAL POVERTY ANALYSIS

DOI: 10.15611/pn.2016.426.16

Streszczenie: Celem artykułu jest wykorzystanie metody wielokrotnej imputacji w tworzeniu syntetycznych zbiorów danych o pełnym pokryciu w analizie przestrzennego zróżnicowania ubóstwa. Podejście to umożliwia tworzenie szacunków o zwiększonej precyzji na niskich poziomach agregacji przestrzennej, niemożliwych do uzyskania z wykorzystaniem estymacji bezpośredniej. Uzyskane rezultaty porównano z pracami studialnymi przeprowadzonymi metodą Faya-Herriota w Ośrodku Statystyki Małych Obszarów w Urzędzie Statystycznym w Poznaniu i otrzymano zbieżne rezultaty.

Słowa kluczowe: statystyka małych obszarów, wielokrotna imputacja, integracja danych, mapowanie ubóstwa.

Summary: The aim of of this article is employing the method of multiple imputation in creating full coverage synthetic data sets in the analysis of spatial poverty differentiation. This approach allows for the creation of estimates with improved accuracy at low levels of spatial aggregation, impossible to obtain with the use of the direct estimation. The results were compared with the work carried out by the Center for Small Areas Statistics at the Statistical Office in Poznań with the use of Fay-Herriot model. The point estimates were consistent.

Keywords: small area estimation, multiple imputation, data integration, poverty mapping.

1. Wstęp

Dostarczanie rzetelnej, aktualnej, wielowymiarowej informacji dla odbiorców administracyjnych jest jednym z głównych zadań statystyki publicznej. W szczególności ważne jest wspomaganie państwa w walce z różnymi niepożądanymi zjawiskami społecznymi, jakim jest m.in. ubóstwo. Ważna jest nie tylko informacja o jego wielkości, lecz także, a nawet przede wszystkim, o jego przestrzennym zróżnicowaniu. Dostarczenie szczegółowej informacji o terytorialnym zróżnicowaniu wskaźników

jakości życia może przyczynić się do lepszego dysponowania ograniczonymi środkami pomocowymi, ale też wskazywać miejsca, gdzie konieczne są różnego rodzaju inwestycje, nie tylko infrastrukturalne, ale również społeczne.

By wypełnić swoje zobowiązania, organy statystyczne przeprowadzają wiele badań specjalnych o tematyce społeczno-ekonomicznej. Jednym z badań, w którym dokonuje się pomiaru wskaźników jakości życia, w tym tych związanych z różnymi wymiarami ubóstwa, jest Badanie Dochodów i Warunków Życia (*European Union Statistics on Income and Living Conditions*, EU-SILC). Wielkość próby w badaniu EU-SILC umożliwia jednak agregację rezultatów co najwyżej na poziomie makroregionów (grup województw, poziom NUTS1), ponieważ szacunki na niższych poziomach agregacji przestrzennej charakteryzują się nieakceptowalnie dużym błędem losowym.

By zwiększyć użyteczność, w kontekście uzyskania szacunków dla małych domen[1], informacji pochodzących z badań reprezentacyjnych, często stosuje się metody statystyki małych obszarów (estymacja pośrednia, SMO). Stosowane w SMO estymatory zwykle poprawiają efektywność szacunków dla małych domen [Rao 2003] i w Polsce przeprowadzane są eksperymentalne prace nad wykorzystaniem estymacji pośredniej w mapowaniu ubóstwa, tj. jego przestrzennego zróżnicowania [Wawrowski 2014; Szymkowiak i in. 2013].

Alternatywną dla badania terytorialnego zróżnicowania różnych zjawisk społeczno-ekonomicznych jest konstrukcja syntetycznych źródeł danych. Podejście to polega na tworzeniu źródeł o pełnym pokryciu z wykorzystaniem istniejących baz pochodzących z badań reprezentacyjnych oraz wykorzystaniu informacji dodatkowych, najczęściej pochodzących ze spisu powszechnego. W kontekście ubóstwa, Eurostat podjął już pierwsze prace na wykorzystaniem badania EU-SILC do konstrukcji tego typu zbiorów [Alfons i in. 2011].

Celem niniejszego opracowania jest próba oszacowania zróżnicowania ubóstwa na poziomie NUTS 3 na podstawie syntetycznego repozytorium danych jednostkowych o pełnym pokryciu skonstruowanego w oparciu o zbiór danych EU-SILC z 2011 r. oraz publikacje spisowe. Zostaną wykorzystane techniki iteracyjnego dopasowania proporcjonalnego (*Iterative Proportional Fitting*, IPF) oraz wielokrotnej imputacji (*Multiple Imputation*, MI). Uzyskane rezultaty zostaną sprawdzone pod kątem jakości, a także porównane z rezultatami uzyskanymi innymi metodami.

2. Badanie Dochodów i Warunków Życia

Badanie Dochodów i Warunków Życia – EU-SILC – jest międzynarodowym badaniem przeprowadzanym rocznie we wszystkich krajach Unii Europejskiej. Zostało ono ustanowione rozporządzeniem Parlamentu Europejskiego (1177/2003 z mody-

[1] Przy tym „małą domeną" nazywa się poziom agregacji, który przy zastosowaniu „klasycznej" estymacji bezpośredniej charakteryzuje się dużym błędem losowym, uniemożliwiającym publikację.

fikacjami zawartymi w rozporządzeniu 553/2005) i zostało wdrożone w 2004 r. w większości krajów UE. Przyczynkiem do wprowadzenia badania była konieczność stałej modyfikacji i dostosowywania do potrzeb odbiorców realizowanych badań statystycznych wywołana wzrostem zapotrzebowania użytkowników na różnego rodzaju informacje dotyczące szeroko rozumianych warunków życia ludności.

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

Badanie realizowane jest w okresie maj-czerwiec danego roku. Okresem odniesienia dla danych finansowych jest rok poprzedzający badanie, natomiast dla pozostałych charakterystyk momentem referencyjnym jest dzień badania [Łysoń (red.) 2012].

Jednym z celów badania jest oszacowanie wielkości tzw. ubóstwa materialnego. Określane jest ono poprzez oszacowanie frakcji gospodarstw domowych znajdujących się poniżej progu ubóstwa jako tzw. wskaźnik zagrożenia ubóstwem po uwzględnieniu w dochodach transferów społecznych. Definiowany jest on jako odsetek osób z ekwiwalentnym dochodem do dyspozycji poniżej progu zagrożenia ubóstwem, który wynosi 60% krajowej mediany ekwiwalentnych dochodów do dyspozycji po transferach społecznych [Łysoń (red.) 2012].

W 2011 r. efektywna liczebność próby wynosiła 12 871 gospodarstw domowych, co stanowiło ok. 65% próby zakładanej. Stosunkowo wysoka frakcja odmów odpowiedzi wymusiła korektę wag wynikających z prawdopodobieństwa inkluzji (tzw. wag początkowych) o wskaźnik kompletności obliczony w ujęciu klas miejscowości zamieszkania.

Tworząc wagi finalne, zastosowano m.in. metody kalibracji wykorzystując dane demograficzne [Łysoń (red.) 2012].

Informacje o tym, czy gospodarstwo domowe znajduje się poniżej progu ubóstwa przechowywane są w zmiennej HX080 (zmienna zero-jedynkowa, gdzie 1 oznacza, że gospodarstwo znajduje się poniżej progu ubóstwa), która jest pochodną zmiennej HX090, gdzie przechowywane są informacje o ekwiwalentnym dochodzie do dyspozycji.

3. Metodyka badania

Idea konstrukcji syntetycznego zbioru danych jest stosunkowo prosta i w dużej mierze opiera się na idei metody reprezentacyjnej – tj. wykorzystania schematu losowania i replikacji na podstawie wartości wag finalnych rekordów w zbiorze [Haslett i in. 2010].

3.1. Iteracyjne dopasowanie proporcjonalne

W celu zwiększenia jakości oszacowań oraz zapewnienia zgodności rozkładów brzegowych analizowanych cech z ograniczeniami spisowymi, w połączonym zbiorze danych dokonano przekształcenia wag analitycznych z pomocą metody iteracyjnego dopasowania proporcjonalnego (*Iterational Proportional Fitting*, IPF; [Peck 2011]). Liczebności cząstkowe zostały rozszacowane z wykorzystaniem modelu logliniowego [Peck 2011]:

$$N_{ij} = a_i b_i n_{ij} \tag{1}$$

zapisanego jako prawdopodobieństwa:

$$\pi_{ij} = a_i b_i p_{ij}, \tag{2}$$

gdzie π_{ij} i p_{ij} to, odpowiednio, prawdopodobieństwa oszacowane z próby i populacji (spisu):

$$\log\left(\frac{\pi_{ij}}{p_{ij}}\right) = \log(a_i) + \log(b_i) + \epsilon_{ij}. \tag{3}$$

Zakłada się, że liczebności empiryczne są zmiennymi niezależnymi o rozkładzie Poissona. Dopasowanie modelu przeprowadzane jest metodą największej wiarygodności przy użyciu algorytmu Newtona–Raphsona.

Na podstawie wag finalnych zmodyfikowanych poprzez algorytm IPF dokonano replikacji rekordów. W efekcie utworzono jednostkowy zbiór danych zawierający 13 568 068 jednostek (gospodarstw domowych). Celem badania było oszacowanie frakcji gospodarstw domowych poniżej progu ubóstwa, dlatego wartości zmiennej HX080 zostały usunięte dla rekordów zreplikowanych, a pozostawione wyłącznie dla rekordów oryginalnych. W tak przygotowanym zbiorze wykorzystano metodę wielokrotnej imputacji.

3.2. Wielokrotna imputacja

Na potrzeby wielokrotnej imputacji tworzy się *m* modeli, gdzie do wartości teoretycznych wynikających z modeli imputacji regresyjnej dolosowane są różne wartości resztowe:

$$\tilde{y}_i = \hat{y}_i + e_i = \hat{\alpha}_i + \hat{\beta} x_i + e_i, \tag{4}$$

gdzie $e_i \sim N\left(0, \hat{\sigma}_{Y|X}\right)$.

Odzwierciedla to zmienność próby, a także umożliwia przeprowadzenie estymacji punktowej i przedziałowej dla nieznanych wartości braków danych.

W wielokrotnej imputacji każdy brak danych jest imputowany za pomocą pewnej liczby (*m*) wartości. Te *m* wartości są uporządkowane w takim sensie, że pierwszy zestaw wartości tworzy pierwszy zbiór danych itd. Oznacza to, że tworzonych

jest *m* kompletnych zbiorów danych. Każdy z tych zbiorów jest analizowany za pomocą standardowych procedur wykorzystujących informację pełną w taki sposób, jakby wartości imputowane były prawdziwe.

Estymatorem dla każdego z *t* $(t = 1,2,\dots,m)$ podstawień jest $\hat{\theta}^{(t)} = \hat{\theta}(U_{obs}, U_{mis}^{(t)})$, gdzie U_{obs} to wartości obserwowane dla danej cechy, zaś $U_{mis}^{(t)}$ to zaimputowane braki danych [Raessler 2004]. Wariancję tego estymatora można wyrazić jako $\widehat{var}\left(\hat{\theta}^{(t)}\right) = \widehat{var}(\hat{\theta}(U_{obs}, U_{mis}^{(t)}))$. Estymatorem punktowym wielokrotnej imputacji jest średnia arytmetyczna z m podstawień:

$$\hat{\theta}_{MI} = \frac{1}{m}\sum_{t=1}^{m}\hat{\theta}^{(t)}. \tag{5}$$

Wariancja estymatora wielokrotnej imputacji dzieli się na wariancję wewnątrzgrupową i wariancję międzygrupową. Wariancja międzygrupowa wyraża się wzorem:

$$B = \frac{1}{m-1}\sum_{t=1}^{m}(\hat{\theta}^{(t)} - \hat{\theta}_{MI})^2, \tag{6}$$

wariancję wewnątrzgrupową zaś można zapisać jako wyrażenie:

$$W = \frac{1}{m}\sum_{t=1}^{m}\widehat{var}\left(\hat{\theta}^{(t)}\right). \tag{7}$$

Wariancja ogólna jest sumą wariancji wewnątrz- i międzygrupowej zmodyfikowany o składnik $\frac{m+1}{m}$ zwiększający dyspersję estymatora, co ma odzwierciedlać niepewność co do prawdziwych wartości imputowanych braków danych:

$$T = W + \frac{m+1}{m}B. \tag{8}$$

Estymacji przedziałowej w wielokrotnej imputacji dokonuje się, szacując przedział ufności:

$$\hat{\theta}_{MI} - t_{v,\frac{\alpha}{2}}\sqrt{T} < \theta < \hat{\theta}_{MI} + t_{v,\frac{\alpha}{2}}\sqrt{T}\,, \tag{9}$$

gdzie liczba stopni swobody $v = (m-1)(1 + \frac{W}{(1+\frac{1}{m})B})^2$.

4. Badanie empiryczne

Zmienna HX080 jest zmienną dychotomiczną, dlatego utworzono model regresji logistycznej, w którym zmiennymi objaśniającymi były zmienne pochodzące z próby EU-SILC. Ze względu na złożoność gospodarstwa domowego, do modelu dobrano zmienne z trzech grup[2]: (1) charakterystyki głów gospodarstw domowych (GD): płeć

[2] Przy tym zmienne dobrano w taki sposób, aby zapewnić brak współliniowości wektora zmiennych objaśniających. Wykorzystano do tego m.in. wartości statystyki VIF.

głowy GD, czy głowa gospodarstwa domowego się uczy, poziom wykształcenia głowy GD, stan cywilny głowy GD, stan zdrowia głowy GD, wiek głowy GD; (2) charakterystyki GD: czy gospodarstwo stać na tygodniowy urlop poza miejscem zamieszkania, czy gospodarstwo potrafi związać koniec z końcem, klasa miejscowości zamieszkania, dochód do dyspozycji GD, województwo; (3) charakterystyki składu GD: liczba niepełnoletnich osób zamieszkałych w GD, liczba bezrobotnych w GD, liczba nieaktywnych zawodowo w GD, liczba niepełnosprawnych w GD.

Model oszacowany na podstawie danych z próby charakteryzował się wartością R^2 Nagelkerkego równą 0,782, a odsetek prawidłowo zaklasyfikowanych wartości wynosił 0,944.

Ze względu na wielkość syntetycznego zbioru danych, dokonano 10 imputacji[3] ($m = 10$), przy czym utworzono model zarówno bez interakcji między zmiennymi jakościowymi, jak i model z interakcjami drugiego stopnia[4].

Ideą przyświecającą zastosowaniu wielokrotnej imputacji jest fakt dolosowania składnika losowego do wartości teoretycznych wynikających z modelu. Oznacza to, że części gospodarstw domowych, choć powielonych w syntetycznym zbiorze, przyporządkowane zostaną różne wartości wskaźnika zagrożenia ubóstwem, co odzwierciedli zmienność zjawiska przy jednoczesnym zwiększeniu liczebności próby.

Otrzymane wyniki, dotyczące wskaźnika zagrożenia ubóstwem, w pierwszej kolejności porównano z opublikowanymi wynikami z badania EU-SILC 2011 (zob. [Łysoń (red.) 2012]). Ogólnie wyniki były zbliżone, choć te uzyskane za pomocą syntetycznego zbioru przy wykorzystaniu modelu bez interakcji wykazywały niższe natężenie ubóstwa, natomiast z wykorzystaniem modelu z interakcjami były bliższe wynikom oszacowanym za pomocą estymacji bezpośredniej (tab. 1).

Analizując rezultaty w ujęciu podregionów (NUTS 3[5]), porównano nie tylko estymatory punktowe oszacowane za pomocą różnych metod, ale również przedziały ufności. Co do zasady, estymacja punktowa na podstawie wielokrotnej imputacji, dla modelu bez, jak i z interakcjami, mieściły się w przedziale ufności dla oszacowań estymacją bezpośrednią (zob. rys. 1; dla przejrzystości przedstawiono jedynie model z interakcjami). Jednocześnie zwiększenie liczebności zbioru spowodowało zmniejszenie błędu standardowego.

---

[3] Literatura [Raessler 2002; Rubin 1987] wskazuje, że liczba imputacji nie musi być duża. Mówi się wręcz o 3 – 5. Wynika to z faktu, że D.B. Rubin [1987] wykazał, że efektywność określonej liczby podstawień w porównaniu do sytuacji, gdyby była ich nieskończona liczba można wyrazić wzorem $1 + \frac{\lambda}{m}$, gdzie λ to frakcja braków danych. Np. dla frakcji braków rzędu 0,6 dla 20 podstawień, efektywność wynosi $1 + \frac{0{,}6}{20} = 1{,}03$ i oznacza, że oszacowany estymator charakteryzuje się błędem standardowym o 3% większym niż ten oszacowany na podstawie dążącej do nieskończoności liczby imputacji.

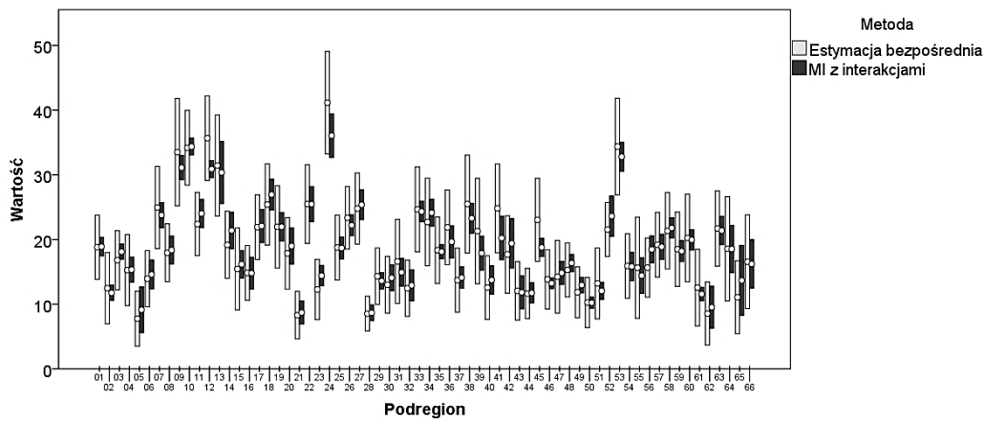
[4] Wykorzystano oprogramowanie IBM SPSS 22 wraz z modułem „Wielokrotne podstawienia". Czas obliczeń dla modelu bez interakcji wynosił 20 godzin, natomiast dla modelu z interakcjami – 6,5 doby.

[5] Estymatory bezpośrednie oszacowano przy wykorzystaniu dostępnego zbioru danych jednostkowych.

Tabela 1. Porównanie wielokrotnej imputacji z estymacją bezpośrednią dla oszacowań wskaźnika ubóstwa

Jednostka terytorialna	Estymacja bezpośrednia	*MI*	*MI* z interakcjami
Polska	17,7	17,3	17,6
Region centralny	15,6	15,7	16,2
Region południowy	16,1	15,7	15,6
Region wschodni	24,5	24,3	24,4
Region północno-zachodni	18,4	17,8	18,2
Region południowo-zachodni	13,0	14,3	14,4
Region północny	17,1	16,0	16,8

Źródło: opracowanie własne.



Rys. 1. Porównanie oszacowań punktowych i przedziałowych dla analizowanych metod estymacji
Źródło: opracowanie własne.

W kolejnym kroku porównano otrzymane rezultaty z wynikami badania przeprowadzonego na zamówienie Głównego Urzędu Statystycznego przez Ośrodek Statystyki Małych Obszarów w Urzędzie Statystycznym w Poznaniu [Szymkowiak i in. 2013]. Oszacowań dokonano tam za pomocą modelu Faya-Herriota (należącego do estymatorów klasy EBLUP), jednego z najczęściej wykorzystywanych modeli SMO.

W ujęciu przestrzennym zaobserwowano zbieżność oszacowań (tab. 2). Analizując zróżnicowanie ubóstwa z wykorzystaniem różnych rodzajów estymacji, można zaobserwować, że zastosowane metody SMO „zwiększają" oszacowania wskaźnika ubóstwa w podregionach o stosunkowo niskim jego natężeniu (np. podregiony 5, 23, 50, 62), a „zmniejszają" w podregionach o wysokim natężeniu analizowanego zjawiska (np. 9, 13, 24, 41, 45). Jednocześnie można również zaobserwować, że metody wielokrotnej imputacji są bardziej zbliżone do oszacowań bezpośrednich niż oszacowania metodą Faya-Herriota (co można uznać za zaletę, jako że estymatory bezpośrednie z definicji są nieobciążone).

Tabela 2. Porównanie przestrzennego zróżnicowania ubóstwa w ujęciu NUTS 3 z wykorzystaniem wybranych metod

Podr.	Est. bezp.	*FH*	*MI*	*MI* Int.	Podr.	Est. bezp.	*FH*	*MI*	*MI* Int.	Podr.	Est. bezp.	*FH*	*MI*	*MI* Int.
1	15,7	17,1	18,6	18,9	23	12,0	14,3	16,1	14,4	45	24,1	13,9	19,0	18,8
2	14,4	14,5	11,8	11,8	24	40,9	24,6	34,3	36,1	46	15,2	14,6	12,6	13,2
3	15,3	20,5	17,9	18,1	25	18,2	21,3	19,5	18,7	47	13,4	14,1	15,3	14,8
4	11,3	12,6	16,0	15,4	26	21,1	25,7	21,5	22,2	48	13,6	14,6	16,1	16,4
5	6,2	7,5	8,8	9,1	27	23,5	24,5	25,2	25,4	49	10,1	10,4	12,7	13,0
6	11,5	12,1	13,9	14,6	28	6,2	6,3	7,6	8,7	50	9,5	10,2	10,0	10,2
7	26,1	22,9	22,2	23,8	29	12,8	14,4	13,0	13,6	51	10,3	9,9	12,9	12,1
8	18,3	22,6	17,8	18,4	30	10,8	10,3	13,9	14,1	52	22,2	21,3	23,4	23,6
9	35,2	29,4	31,7	31,1	31	12,2	16,5	15,2	14,9	53	34,0	29,8	33,2	32,8
10	34,7	30,2	34,3	34,4	32	14,2	11,5	12,5	12,9	54	17,6	20,7	14,7	15,8
11	24,0	18,5	24,1	24,0	33	25,9	24,1	23,1	24,3	55	17,5	20,8	13,9	14,4
12	35,4	29,5	30,7	30,9	34	28,6	26,1	24,2	24,1	56	14,8	17,2	17,3	18,5
13	31,0	16,4	25,9	30,4	35	14,7	18,0	19,1	18,1	57	17,5	16,7	20,5	18,9
14	21,7	17,7	19,7	21,4	36	19,7	20,9	19,5	19,7	58	21,3	19,4	21,1	21,8
15	14,1	15,1	15,7	16,2	37	12,0	13,4	13,9	14,1	59	18,0	17,0	18,4	18,2
16	13,9	14,2	15,0	14,8	38	21,4	24,6	23,7	23,3	60	21,6	19,8	20,0	20,0
17	23,6	21,6	21,8	22,1	39	18,5	22,2	15,5	17,9	61	13,4	11,0	12,5	11,6
18	21,5	24,4	26,4	27,0	40	11,0	11,9	13,8	13,7	62	7,7	8,5	8,5	9,5
19	21,5	23,4	20,7	22,0	41	29,7	20,8	19,4	20,2	63	21,9	16,6	22,2	21,4
20	17,7	17,4	18,9	19,0	42	17,3	22,0	18,4	19,4	64	17,3	18,7	20,7	18,5
21	8,4	8,7	10,0	8,7	43	13,3	7,4	11,7	11,8	65	11,6	9,6	11,0	13,7
22	28,8	23,2	24,8	25,5	44	10,5	11,1	11,9	11,8	66	16,5	12,1	16,1	16,2

FH – model Faya-Harriota; *MI* – wielokrotna imputacja bez interakcji; *MI* Int. – wielokrotna imputacja z interakcjami.

Źródło: opracowanie własne.

Zbieżność rezultatów uzyskanych za pomocą opisywanych metod potwierdza analiza korelacji[6] wartości oszacowań punktowych w ujęciu podregionów (tab. 3). Metody z użyciem wielokrotnej imputacji charakteryzują się wynikami najbardziej zbliżonymi do estymacji bezpośredniej, natomiast estymacja Faya-Herriota charakteryzuje się również dużą zbieżnością z pozostałymi metodami, jednak siła związku z nimi jest mniejsza.

Tabela 3. Macierz korelacji dla oszacowań wskaźnika ubóstwa za pomocą wybranych metod

Estymacja	Bezp.	EBLUP	*MI*	*MI* int
Bezp.	1	0,81	0,92	0,94
EBLUP	0,81	1	0,84	0,84
MI	0,92	0,84	1	0,99
MI int.	0,94	0,84	0,99	1

Źródło: opracowanie własne.

[6] Wszystkie współczynniki korelacji były istotnie różne od zera na poziomie istotności $\alpha = 0{,}01$.

5. Zakończenie

W artykule pokazano sposób konstrukcji syntetycznych zbiorów danych o pełnym pokryciu na podstawie publikacji spisowych i zbiorów danych pochodzących z badań reprezentacyjnych. Wykazano, że syntetyczne zbiory danych mogą służyć jako baza estymacji dla małych domen.

Jednocześnie należy wskazać na pewne problemy wynikające z zastosowanego podejścia. Przede wszystkim dołączanie informacji do wielkich zbiorów danych jest czasochłonne i trudne obliczeniowo. Jednocześnie jakość dołączanej informacji z próby do syntetycznego zbioru danych w dużej mierze zależy od specyfikacji modelu imputacji oraz od jakości zbioru wejściowego.

Wśród zalet opisywanej metody można wymienić przede wszystkim fakt, że rzetelnie skonstruowany syntetyczny, jednostkowy zbiór danych, przy odpowiednim, dobrym jakościowo, dołączeniu informacji o zmiennej celu umożliwia tworzenie wielowymiarowych zestawień z pozostałymi zmiennymi tworzącymi zbiór.

Jako dalsze kierunki badań można wskazać utworzenie zbiorów z większą liczbą zmiennych, w tym zmiennych mierzalnych, jak również utworzenie syntetycznych zbiorów dla okresów międzyspisowych, m.in. z wykorzystaniem danych rejestrowych i informacji pochodzących ze sprawozdawczości bieżącej.

Syntetyczne zbiory danych o pełnym pokryciu mogą być alternatywą dla modelowego podejścia do statystyki małych obszarów [Rao 2003], jak również mogą być źródłem informacji dodatkowej dla podejścia modelowego.

Literatura

Alfons A., Filzmoser P., Hulliger B., Kolb J-P., Kraft S., Munnich R., Templ M., 2011, *Synthetic Data Generation of SILC Data*, European Commision, Community Research, AMELI Project, Trier.

Haslett S., Jones G., Noble A., Ballas D., 2010, *More or Less? Comparing Small Area Estimation, Spatial Microsimulation, and Mass Imputation*, Section on Survey Research Methods – JSM, American Statistical Association, Alexandria–Vancouver.

Łysoń P. (red.), 2012, *Dochody i warunki życia ludności Polski (raport z badania EU-SILC 2011)*, Informacje i Opracowania Statystyczne, GUS, Departament Badań Społecznych i Warunków Życia, Warszawa.

Peck J., 2011, *Extension Commands and Rim Weighting with IBM SPSS Statistics: Theory and Practice*, IBM Corporation, Armonk, NY.

Raessler S., 2002, *Statistical Matching. A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches*, Springer, New York.

Raessler S., 2004, *Data fusion: Identification problems, validity, and multiple imputation*, Austrian Journal of Statistics, vol. 33, no. 1/2, s. 153–171.

Rao J.N.K., 2003, *Small Area Estimation*, Wiley & Sons, Hoboken, NJ.

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

Szymkowiak M., Beręsewicz M., Józefowski T., Klimanek T., Małasiewicz A., Młodak A., Wawrowski Ł., 2013, *Mapy ubóstwa na poziomie podregionów w Polsce z wykorzystaniem estymacji pośredniej*, Urząd Statystyczny w Poznaniu, Ośrodek Statystyki Małych Obszarów, GUS, Warszawa.

Wawrowski Ł., 2014, *Wykorzystanie metod statystyki małych obszarów do tworzenia map ubóstwa w Polsce*, Wiadomości Statystyczne, nr 9.