PRACE NAUKOWE
Uniwersytetu Ekonomicznego we Wrocławiu

RESEARCH PAPERS
of Wrocław University of Economics

Nr 384

Taksonomia 24
Klasyfikacja i analiza danych – teoria i zastosowania

Redaktorzy naukowi
Krzysztof Jajuga
Marek Walesiak

Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2015

Redaktor Wydawnictwa: Aleksandra Śliwka

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 stronie internetowej Wydawnictwa
www.pracenaukowe.ue.wroc.pl
www.wydawnictwo.ue.wroc.pl

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

Wersja pierwotna: publikacja drukowana

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

Druk i oprawa: TOTEM

Spis treści

Summaries

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS nr 384 • 2015
Taksonomia 24 ISSN 1899-3192
Klasyfikacja i analiza danych – teoria i zastosowania e-ISSN 2392-0041

Wojciech Roszka
Uniwersytet Ekonomiczny w Poznaniu
e-mail:wojciech.roszka@ue.poznan.pl

KONSTRUKCJA SYNTETYCZNYCH ZBIORÓW DANYCH NA POTRZEBY ESTYMACJI DLA MAŁYCH DOMEN

Streszczenie: Rozwijająca się gospodarka wymaga wsparcia informacyjnego w szczegółowych przekrojach. Liczebność próby w badaniach reprezentacyjnych uniemożliwia spełnienie tego postulatu, natomiast badania pełne, ze względu na koszty i czas realizacji, przeprowadzane są rzadko. Metody statystycznej integracji danych umożliwiają łączenie informacji z wielu źródeł. Dzięki ich zastosowaniu możliwe jest połączenie bogatych merytorycznie źródeł pochodzących z badań reprezentacyjnych ze zbiorami o pełnym pokryciu, zapewniającym możliwości estymacji dla małych domen. W artykule przedstawiony zostanie proces łączenia informacji z publikacji spisowych z informacjami z badań reprezentacyjnych w celu estymacji stopy bezrobocia w ujęciu powiatów.

Słowa kluczowe: statystyka małych obszarów, integracja danych, badania reprezentacyjne, spis powszechny.

DOI: 10.15611/pn.2015.384.27

1. Wstęp

Statystyka publiczna (jak również inne instytucje państwowe i prywatne) przeprowadzają badania reprezentacyjne pokrywające bardzo wiele aspektów społeczno-gospodarczych. Jednak ze względu na wysokie koszty przeprowadzenia takich badań, jak również czas ich trwania, liczebność próby jest zbyt mała[1], by możliwe było stosowanie estymatorów bezpośrednich dla małych domen. Badania pełne, takie jak spisy powszechne, umożliwiają tworzenie rzetelnych szacunków dla małych domen. Jednak ze względu na wysiłek organizacyjny, jak również koszty, przeprowadzany jest stosunkowo rzadko – z częstotliwością około dziesięcioletnią.

[1] Największe względem pokrycia badanie reprezentacyjne w Polsce – Badanie Aktywności Ekonomicznej Ludności – pokrywa około 1% populacji generalnej w ujęciu rocznym i jego budżet według PBSSP na 2015 rok wynosi 39 586 850 zł.

Wykorzystanie rejestrów administracyjnych jako źródła zasilania informacyjnego gospodarki jest coraz częściej podnoszonym postulatem [Dygaszewicz 2010]. Jednak ich integracja i harmonizacja z populacjami, definicjami i okresami referencyjnymi stosowanymi w statystyce publicznej jest zadaniem trudnym i wymagającym szeregu rozwiązań nie tylko technicznych, ale również prawnych.

Statystyczna integracja danych jest metodą umożliwiającą łączenie informacji z różnych źródeł w celu łącznej obserwacji wielu cech. Jednocześnie dzięki ich zastosowaniu możliwe jest połączenie bogatych merytorycznie źródeł pochodzących z badań reprezentacyjnych ze zbiorami o pełnym pokryciu, zapewniającym możliwości estymacji dla małych domen.

Celem niniejszego opracowania jest próba konstrukcji syntetycznego repozytorium danych jednostkowych o pełnym pokryciu na podstawie publikacji spisowych i zbiorów danych pochodzących z badań reprezentacyjnych w celu estymacji różnych charakterystyk dla małych domen. W artykule przedstawiona zostanie idea konstrukcji takiego zbioru, etapy harmonizacji i integracji różnych źródeł, a także badanie symulacyjne wraz z oceną jakości uzyskanych oszacowań.

2. Idea konstrukcji syntetycznego zbioru danych

Publikacje spisowe są powszechnie dostępne w formie predefiniowanych tablic zawierających informacje o różnych zjawiskach społeczno-ekonomicznych. Posiadając zagregowaną informację o liczbie jednostek współdzielących warianty określonych zmiennych, replikuje się taką kombinację charakterystyk N razy jako N rekordów w jednostkowym, syntetycznym zbiorze danych (por. tab. 1).

Problemem w takim podejściu jest fakt, że tablice spisowe bardzo często nie zawierają łącznej obserwacji wielu zmiennych. W Banku Danych Lokalnych, jak również w innych formach publikacji wyników publikowana jest łączna obserwacja co najwyżej trzech-czterech zmiennych (np. ludność według grup wieku i poziomu wykształcenia w ujęciu powiatu zamieszkania). Charakterystyki łącznego rozkładu większej liczby zmiennych nie są publikowane.

W celu utworzenia możliwie kompleksowego jednostkowego zbioru danych, zawierającego możliwie wiele charakterystyk, wykorzystuje się zbiory danych pochodzące z badań reprezentacyjnych, zawierające łączną obserwację długiego wektora cech. W celu dostosowania rozkładu analizowanych cech do ograniczeń spisowych stosuje się różne miary przeważania [Anderson 2013]. Następnie dokonuje się replikacji przeważonych i zagregowanych zbiorów danych podobnie jak pokazano w tab. 1.

Utworzenie syntetycznego zbioru danych jednostkowych umożliwia nie tylko możliwość bardziej elastycznego tworzenia zestawień, ale również może służyć jako podstawa dołączania informacji z badań reprezentacyjnych [Rahman 2008].

Tabela 1. Dezagregacja tablic spisowych w jednostkowy zbiór danych

Zagregowana tabela spisowa

Powiat	Płeć	Wiek	Wykszt.	*N*
3064	K	20-24	wyższe	100
2861	M	30-35	zawodowe	50
…	…	…	…	…

↓

Syntetyczny jednostkowy zbiór danych

Lp.	Powiat	Płeć	Wiek	Wykszt.
1	3064	K	20-24	wyższe
2	3064	K	20-24	wyższe
…	…	…	…	…
100	3064	K	20-24	wyższe
101	2861	M	30-35	zawodowe
…	…	…	…	…
150	2861	M	30-35	zawodowe
…	…	…	…	…

Źródło: opracowanie własne.

Tabela 2. Imputacja metodą syntetycznej rekonstrukcji

Kroki	1.	2.	…	Ostatni
Wiek, płeć, stan cywilny	Wiek: 18 Płeć: M St. cyw.: żonaty	Wiek: 34 Płeć: K St. cyw.: rozwiedziona	...	Wiek: 87 Płeć: M St. cyw.: wdowiec
Prawdopodobieństwo rodzaju aktywności ekonomicznej dla danych wieku, płci i stanu cywilnego	Prawd. / Przedziały skum. prawd. Prac.: 0,4 (0,0 - 0,4) Bezr.: 0,3 (0,4 - 0,7) Bier.: 0,3 (0,7 - 1)	Prawd. / Przedziały skum. prawd. Prac.: 0,6 (0,0 - 0,6) Bezr.: 0,3 (0,6 - 0,9) Bier.: 0,1 (0,9 - 1)	...	Prawd. / Przedziały skum. prawd. Prac.: 0,0 (0,0 - 0,0) Bezr.: 0,0 (0,0 - 0,0) Bier.: 1 (0,0 - 1)
Liczba losowa	0,281	0,709	…	0,481
Imputowana aktywność ekonomiczna	Pracujący	Bezrobotna	…	Bierny

Źródło: opracowanie własne na podstawie [Williamson 2013].

Dołączanie informacji z informacji z innych źródeł w literaturze (m.in. [Rahman 2008; Williamson 2013]) przedstawia się jako problem imputacji. Jednym z podejść dołączenia informacji z badania reprezentacyjnego do syntetycznej bazy spisowej jest syntetyczna rekonstrukcja (*synthetic reconstruction*). Polega ono na losowaniu wariantów cech z warunkowych prawdopodobieństw (dla domeny) oszacowanych na podstawie badania reprezentacyjnego (por. tab. 2).

3. Badanie symulacyjne

Badanie symulacyjne przeprowadzono w celu zaprezentowania idei tworzenia syntetycznych, jednostkowych zbiorów danych o pełnym pokryciu, jak również empirycznej weryfikacji hipotezy o możliwości dołączenia informacji z badań reprezentacyjnych w celu estymacji dla małych domen. Badanie przeprowadzono na podstawie rezultatów Narodowego Spisu Powszechnego Ludności i Mieszkań 2011 dla ludności oraz wybranych badań reprezentacyjnych[2].

Moment referencyjny spisu został ustalony na 31 marca 2011. W celu zachowania zgodności okresów referencyjnych zbiory danych zostały ograniczone do jednostek, które podlegały pomiarowi w I i II kwartale 2011 roku[3]. Wybrano sześć zmiennych zgodnych co do definicji ze zmiennymi spisowymi: powiat zamieszkania, płeć, wiek, stan cywilny, wykształcenie, klasa miejscowości zamieszkania (miasto-wieś). Przyjęto założenie, że zbiory danych pochodzące z badań reprezentacyjnych są rozłączne[4]. Jednocześnie w celu zwiększenia liczebności próby oraz zmniejszenia prawdopodobieństwa wystąpienia „zerowych" domen[5] dokonano konkatenacji zbiorów danych (por. tab. 3). Wagi analityczne[6] zharmonizowano, korzystając ze wzoru:

$$w'_i = \frac{w_i}{\sum_{i=1}^{n} w_i} N, \tag{1}$$

gdzie: w'_i – zharmonizowana waga analityczna dla *i*-tej jednostki w zintegrowanym zbiorze,
w_i – oryginalna waga analityczna,
N – liczebność populacji generalnej.

Tabela 3. Konkatenacja zbiorów danych pochodzących z badań reprezentacyjnych

Zbiór danych	n
BAEL	139 694
BBGD	44 754
DS	31 401
Ogółem	**215 849**

Źródło: opracowanie własne.

[2] Ze względu na dostępność danych wybrano Badanie Aktywności Ekonomicznej Ludności, Badanie Budżetów Gospodarstw Domowych oraz Diagnozę Społeczną.

[3] Pomiar w Diagnozie Społecznej przeprowadzony został w marcu i kwietniu 2011roku. BAEL i BBGD charakteryzują się pomiarem kwartalnym. Jako populację wybrano osoby w wieku 15 lat i więcej.

[4] Liczebność badań częściowych stanowi zwykle bardzo niewielki odsetek liczebności całej populacji, więc prawdopodobieństwo wylosowania jednej jednostki do dwóch badań jest zbliżone do zera.

[5] Domeny, które nie dostały się do próby.

[6] Wagi wynikające ze schematu losowania w poszczególnych badaniach.

Uzyskano zbiór danych o liczebności 215 849 rzeczywistych jednostek. W celu 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)[7]. Liczebności cząstkowe[8] zostały rozszacowane z wykorzystaniem modelu logliniowego [Peck 2011]:

$$N_{ij} = a_i b_i n_{ij} \tag{2}$$

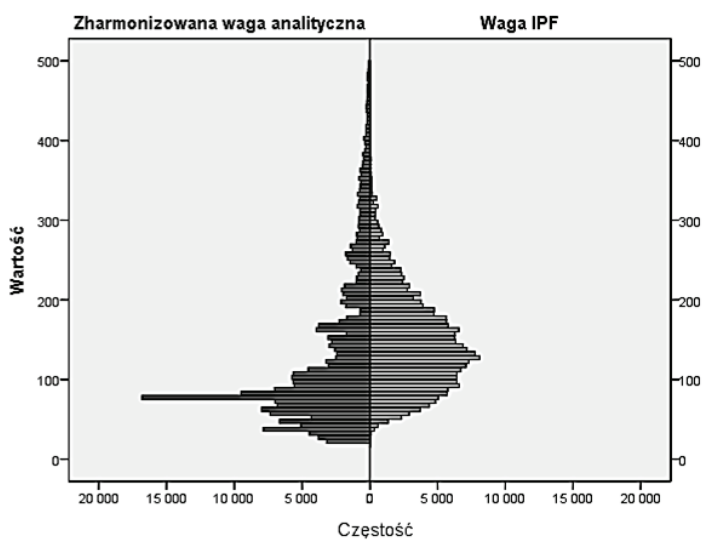
zapisanego jako prawdopodobieństwa:

$$\pi_{ij} = a_i b_i p_{ij}, \tag{3}$$

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

$$\log\left(\frac{\pi_{ij}}{p_{ij}}\right) = \log(a_i) + \log(b_i) + \epsilon_{ij}. \tag{4}$$

Zakłada się, że liczebności empiryczne są zmiennymi niezależnymi o rozkładzie Poissona. Dopasowanie modelu przeprowadzane jest metodą największej wiarygodności przy użyciu algorytmu Newtona-Raphsona. Rozkład zmodyfikowanych wag analitycznychw_i^{IPF} ilustruje rys. 1.



Rys. 1. Rozkład wag: analitycznych i IPF w zbiorze danych z badań reprezentacyjnych

Źródło: opracowanie własne.

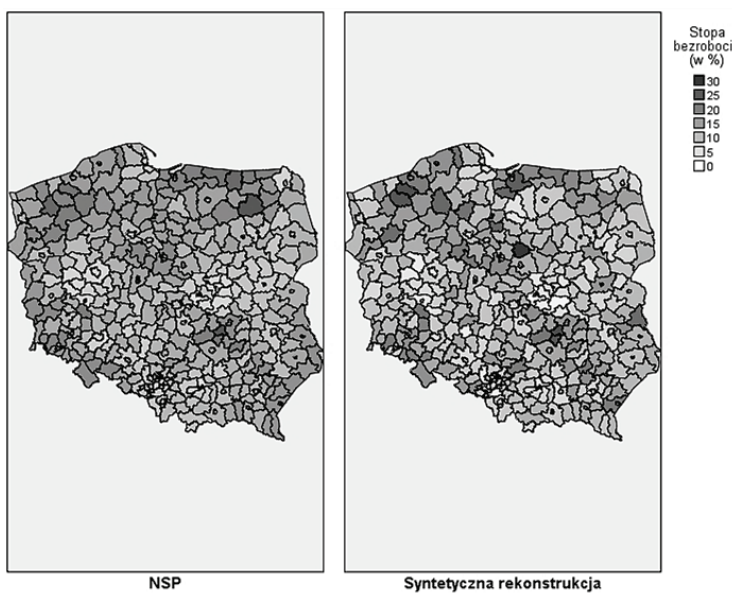
[7] Inną metodą dostosowania wag do określonych wartości globalnych jest m.in. kalibracja [Józefowski, Szymkowiak 2012].

[8] Algorytm metody IPF został opisany w [Roszka 2011].

Zbiór danych przeważono według wag IPF. Następnie, zgodnie z metodą przedstawioną w tab. 1, utworzono syntetyczny zbiór danych o liczebności $N = 32\,679\,635$ zawierający łączną obserwacje wszystkich analizowanych cech.

W kolejnym kroku do syntetycznego, pełnego zbioru danych dołączono informację o aktywności ekonomicznej ludności z BAEL. Zbiór danych z badań reprezentacyjnych podzielono na 55 170 warstw (domen) na podstawie wariantów analizowanych cech i w każdej warstwie obliczono warunkowe prawdopodobieństwa dla każdego rodzaju aktywności zawodowej. Następnie w syntetycznym zbiorze danych dla każdej jednostki losowano liczbę pseudolosową z rozkładu jednostajnego z przedziału $\langle 0,1 \rangle$ i w zależności od przedziału wartości wynikającego ze skumulowanych warunkowych prawdopodobieństw dla danej domeny imputowano określony rodzaj aktywności.

Ocenę jakości uzyskanych szacunków przeprowadzono poprzez porównanie stopy bezrobocia uzyskanej z syntetycznego zbioru z rzeczywistymi wynikami spisowymi. W pierwszej kolejności porównano przestrzenny rozkład prawdziwej i teoretycznej (wynikającej z modelu imputacji) stopy bezrobocia (por. rys. 2).

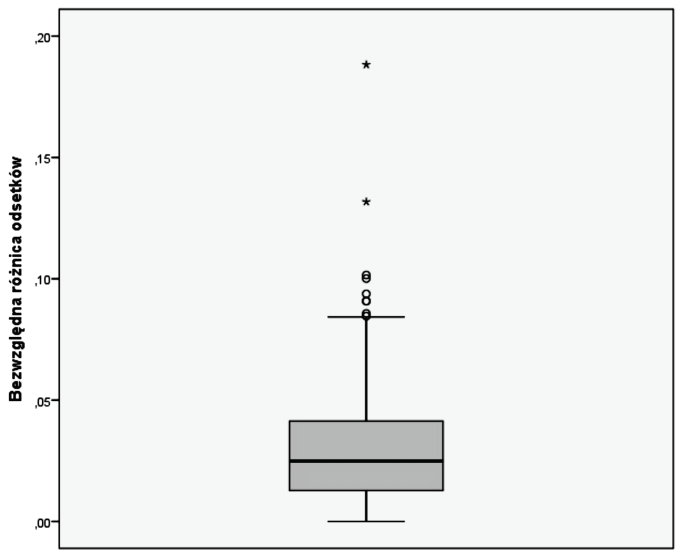


Rys. 2. Rozkład stopy bezrobocia w ujęciu powiatów na podstawie NSP 2011 i syntetycznego zbioru danych

Źródło: opracowanie własne.

Porównując rozkład stopy bezrobocia w ujęciu powiatów, stwierdzono pewne podobieństwo, sugerujące, że imputacja metodą syntetycznej rekonstrukcji w dobry sposób odtwarza rozkład imputowanej cechy.

Następnie obliczono bezwzględne różnice między stopą bezrobocia obliczoną na podstawie NSP a stopą oszacowaną na podstawie syntetycznego zbioru danych. Średnia bezwzględna różnica wynosiła 2,98% przy odchyleniu standardowym równym 2,31%. Najmniejsza różnica wynosiła zaledwie 0,002%, największa zaś aż 18,82% (por. rys. 3).



Rys. 3. Rozkład bezwzględnych różnic NSP i syntetycznego zbioru danych

Źródło: opracowanie własne.

Estymacja stopy bezrobocia dla małych obszarów (powiatów) charakteryzowała się umiarkowaną jakością, jednak dla wielu domen uzyskano szacunki zbliżone do wartości rzeczywistych. Jako główny problem w analizowanych przykładzie można wskazać niewielką liczbę zmiennych w syntetycznym zbiorze danych, które dodatkowo są cechami niemierzalnymi. Utworzenie dobrego modelu imputacji na podstawie stosunkowo niewielu informacji jest utrudnione. Należy jednak podkreślić, że pomimo umiarkowanej jakości szacunków, sposób konstrukcji syntetycznego zbioru danych wraz z dołączaniem do niego informacji z innych źródeł może, przy dostępie większej liczby zmiennych, w tym ilościowych, zapewnić tworzenie dobrych jakościowo szacunków dla małych domen.

4. 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 do dołączania informacji z innych źródeł, a dzięki pełnemu pokryciu możliwa jest estymacja dla małych domen.

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

Anderson B., 2013, *Estimating Small Area Income Deprivation: An Iterative proportional Fitting Approach,* [w:] Tanton R., Edwards K.L. 2013, *Spatial Microsimulation: A Reference Guide for Users*, J. Bus. Econ. Stat. 4, s. 87-94.

Dygaszewicz J., 2010, *Integracja rejestrów publicznych*, Główny Urząd Statystyczny, Warszawa.

Józefowski T., Szymkowiak M., 2009, *Estymatory kalibracyjne w badaniach statystycznych*, Wiadomości Statystyczne, 2009, | nr 1.

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

Rahman A., 2008, *A Review of Small Area Estimation Problems and Methodological Developments*, Discussion Paper 66, NATSEM, University of Canberra.

Rao J.N.K., 2003, *Small Area Estimation*, Wiley and Sons.

Roszka W., 2011, *Iteracyjne dopasowanie proporcjonalne jako metoda poprawiania wyników w badaniach sondażowych,* [w:] Garczarczyk J., Skikiewicz R., *Metody pomiaru i analizy rynku usług. Dylematy badawcze*, Zeszyty Naukowe UEP, 201.

Williamson P., 2013, *An Evaluation of Two Synthetic Small-Area Microdata Simulation Methodologies: Synthetic Reconstruction and Combinatorial Optimisation,* [w:] Tanton R., Edwards K.L. 2013, *Spatial Microsimulation: A Reference Guide for Users*, J. Bus. Econ. Stat. 4, s. 8794.

CONSTRUCTION OF SYNTHETIC DATA SETS FOR SMALL AREA ESTIMATION

Summary: A growing economy requires the information support in the specific cross-sections. The sample size in sample surveys makes the fulfillment of this demand impossible, while the full study, due to the cost and time of implementation, is rarely conducted. Statistical methods of data integration allow to combine information from multiple sources. Thanks to their application, it is possible to combine sample surveys and full studies, which makes estimation for small domains possible. In this article the process of combining information from the census publication and information from sample surveys to estimate the rate of unemployment in terms of counties will be presented.

Keywords: statistical data integration, small area estimation, sample survey, census.