

# PRACE NAUKOWE

Uniwersytetu Ekonomicznego we Wrocławiu

# RESEARCH PAPERS

of Wrocław University of Economics

Nr 328

**Taksonomia 23**

**Klasyfikacja i analiza danych –  
teoria i zastosowania**

Redaktorzy naukowci

Krzysztof Jajuga, Marek Walesiak



Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu  
Wrocław 2014

Redaktor Wydawnictwa: Barbara Majewska

Redaktor techniczny: Barbara Łopusiewicz

Korektor: Barbara Cibis

Łamanie: Beata Mazur

Projekt okładki: Beata Dębska

Publikacja jest dostępna w Internecie na stronach:

[www.ibuk.pl](http://www.ibuk.pl), [www.ebscohost.com](http://www.ebscohost.com),

w Dolnośląskiej Bibliotece Cyfrowej [www.dbc.wroc.pl](http://www.dbc.wroc.pl),

The Central and Eastern European Online Library [www.ceeol.com](http://www.ceeol.com),

a także w adnotowanej bibliografii zagadnień ekonomicznych BazEkon

[http://kangur.uek.krakow.pl/bazy\\_ae/bazekon/nowy/index.php](http://kangur.uek.krakow.pl/bazy_ae/bazekon/nowy/index.php)

Informacje o naborze artykułów i zasadach recenzowania znajdują się  
na stronie internetowej Wydawnictwa

[www.wydawnictwo.ue.wroc.pl](http://www.wydawnictwo.ue.wroc.pl)

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

Kopiowanie i powielanie w jakiegokolwiek formie  
wymaga pisemnej zgody Wydawcy

© Copyright by Uniwersytet Ekonomiczny we Wrocławiu  
Wrocław 2014

**ISSN 1899-3192** (Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu)

**ISSN 1505-9332** (Taksonomia)

Wersja pierwotna: publikacja drukowana

Druk: Drukarnia TOTEM

## Spis treści

<b>Wstęp</b> .....	11
<b>Małgorzata Rószkiewicz</b> , Wykorzystanie metaanalizy w budowaniu modelu pomiarowego w przypadku braku niezmienniczości zasad pomiaru na przykładzie pomiaru zadowolenia z życia.....	13
<b>Elżbieta Sobczak</b> , Harmonijność inteligentnego rozwoju regionów Unii Europejskiej .....	21
<b>Ewa Roszkowska, Renata Karwowska</b> , Analiza porównawcza województw Polski ze względu na poziom zrównoważonego rozwoju w roku 2010.....	30
<b>Tadeusz Kufel, Magdalena Osińska, Marcin Błażejowski, Paweł Kufel</b> , Analiza porównawcza wybranych filtrów w analizie synchronizacji cyklu koniunkturalnego.....	41
<b>Marcin Salamaga</b> , Próba konstrukcji tablic „wymierania scenicznego” spektakli operowych na przykładzie Metropolitan Opera.....	51
<b>Iwona Foryś</b> , Wykorzystanie analizy dyskryminacyjnej do typowania rynków podobnych w procesie wyceny nieruchomości niemieszkalnych .....	59
<b>Jerzy Korzeniewski</b> , Selekcja zmiennych w klasyfikacji – propozycja algorytmu .....	69
<b>Sabina Denkowska</b> , Testowanie wielokrotne przy weryfikacji wieloczynnikowych modeli proporcjonalnego hazardu Coxa.....	76
<b>Ewa Chodakowska</b> , Teoria równań strukturalnych w klasyfikacji zmiennych jawnych i ukrytych według charakteru ich wzajemnych oddziaływań .....	85
<b>Iwona Konarzewska</b> , Model PCA dla rynku akcji – studium przypadku .....	94
<b>Katarzyna Wójcik, Janusz Tuchowski</b> , Dobór optymalnego zestawu słów istotnych w opiniach konsumentów na potrzeby ich automatycznej analizy	106
<b>Aleksandra Łuczak</b> , Zastosowanie metody AHP-LP do oceny ważności determinant rozwoju społeczno-gospodarczego w jednostkach administracyjnych .....	116
<b>Aleksandra Witkowska, Marek Witkowski</b> , Klasyfikacja pozycyjna banków spółdzielczych według stanu ich kondycji finansowej w ujęciu dynamicznym .....	126
<b>Adam Depta</b> , Zastosowanie analizy korespondencji do oceny jakości życia ludności na podstawie kwestionariusza SF-36v2 .....	135
<b>Marek Lubicz, Maciej Zięba, Konrad Pawelczyk, Adam Rzechonek, Marek Marciniak, Jerzy Kołodziej</b> , Indukcja reguł dla danych niekompletnych i niezbalansowanych: modele klasyfikatorów i próba ich zastosowania do predykcji ryzyka operacyjnego w torakochirurgii .....	146

<b>Małgorzata Misztal</b> , Wybrane metody oceny jakości klasyfikatorów – przegląd i przykłady zastosowań.....	156
<b>Anna M. Olszewska</b> , Wykorzystanie wybranych metod taksonomicznych do oceny potencjału innowacyjnego województw .....	167
<b>Iwona Bąk</b> , Porównanie jakości grupowań powiatów województwa zachodniopomorskiego pod względem atrakcyjności turystycznej.....	177
<b>Agnieszka Kozera, Joanna Stanisławska, Romana Głowicka-Wołoszyn</b> , Segmentacja gospodarstw domowych według wydatków na turystykę zorganizowaną.....	186
<b>Agnieszka Wałęga</b> , Podejście syntetyczne w analizie spójności ekonomicznej gospodarstw domowych.....	196
<b>Joanna Banaś, Małgorzata Machowska-Szewczyk, Bożena Mroczek</b> , Zastosowanie analizy korespondencji do badania wpływu elektrowni wiatrowych na jakość życia ludności .....	205
<b>Joanna Banaś, Krzysztof Małecki</b> , Klasyfikacja punktów pomiarów ankietowych kierowców na granicy Szczecina z wykorzystaniem zmiennych symbolicznych.....	214
<b>Aneta Becker</b> , Wykorzystanie informacji granularnej w analizie wymagań rynku pracy.....	222
<b>Katarzyna Cheba, Joanna Holub-Iwan</b> , Wykorzystanie analizy korespondencji w segmentacji rynku usług medycznych.....	230
<b>Adam Depta, Iwona Staniec</b> , Identyfikacja czynników decydujących o jakości życia studentów łódzkich uczelni.....	238
<b>Katarzyna Dębowska, Jarosław Kilon</b> , Reguły asocjacyjne w analizie wyników badań metodą Delphi.....	247
<b>Anna Domagała</b> , O wykorzystaniu analizy głównych składowych w metodzie <i>Data Envelopment Analysis</i> .....	254
<b>Alicja Grześkowiak</b> , Analiza wykluczenia cyfrowego w Polsce w ujęciu indywidualnym i regionalnym.....	264
<b>Anna M. Olszewska, Anna Gryko-Nikitin</b> , Pomiar postrzegania jakości kształcenia uczelni wyższej na danych porządkowych z wykorzystaniem środowiska R.....	273
<b>Karolina Paradysz</b> , Hierarchiczna metoda grupowania powiatów jako podejście benchmarkowe w ocenie bezrobocia według BAEL-u w wybranych typach małych obszarów .....	282
<b>Radosław Pietrzyk</b> , Porównanie metod pomiaru efektywności zarządzania portfelami funduszy inwestycyjnych.....	290
<b>Agnieszka Przedborska, Małgorzata Misztal</b> , Wybrane metody statystyki wielowymiarowej w ocenie skuteczności terapeutycznej głębokiej stymulacji elektromagnetycznej u pacjentów z chorobą zwyrodnieniową stawów.....	299

<b>Wojciech Roszka, Marcin Szymkowiak</b> , Podejście kalibracyjne w statystycznej integracji danych .....	308
<b>Iwona Skrodzka</b> , Zastosowanie wybranych metod klasyfikacji do analizy kapitału ludzkiego krajów Unii Europejskiej .....	316
<b>Agnieszka Stanimir</b> , Wielowymiarowa analiza czynników sprzyjających włączeniu społecznemu .....	326
<b>Dorota Strózik, Tomasz Strózik</b> , Przestrzenne zróżnicowanie poziomu życia w województwie wielkopolskim.....	334
<b>Izabela Szamrej-Baran</b> , Identyfikacja przyczyn ubóstwa energetycznego w Polsce przy wykorzystaniu modelowania miękkiego.....	343
<b>Janusz Tuchowski, Katarzyna Wójcik</b> , Klasyfikacja obiektów w systemie Krajowych Ram Kwalifikacji opisanych za pomocą ontologii .....	353
<b>Aleksandra Matuszewska-Janica</b> , Grupowanie krajów Unii Europejskiej ze względu na poziom feminizacji sektorów gospodarczych .....	361
<b>Monika Rozkrut, Dominik Rozkrut</b> , Identyfikacja strategii innowacyjnych przedsiębiorstw usługowych w Polsce .....	369

## Summaries

<b>Małgorzata Rószkiewicz</b> , The use of meta-analysis in building the measurement model in case of the absence of measurement invariance on the example of measuring of life satisfaction.....	20
<b>Elżbieta Sobczak</b> , Harmonious smart growth of European Union regions.....	29
<b>Ewa Roszkowska, Renata Karwowska</b> , The comparative analysis of Polish voivodeships with respect to sustainable development in 2010.....	40
<b>Tadeusz Kufel, Magdalena Osińska, Marcin Błażejowski, Paweł Kufel</b> , Comparative analysis of chosen filters in business cycles analysis .....	50
<b>Marcin Salamaga</b> , The attempt of construction of the life tables for opera works on the example of the Metropolitan Opera .....	58
<b>Iwona Foryś</b> , Using discriminant analysis to select similar markets in non-residential property valuation process.....	68
<b>Jerzy Korzeniewski</b> , Variable selection in classification – algorithm proposal .....	75
<b>Sabina Denkowska</b> , Multiple testing in the verification process of multifactorial Cox proportional hazards models .....	84
<b>Ewa Chodakowska</b> , The theory of structural equations modelling in the classification of observed variables and latent constructs according to the character of their relationship.....	93
<b>Iwona Konarzewska</b> , Modelling stock market by PCA factor model – case study .....	105

<b>Katarzyna Wójcik, Janusz Tuchowski</b> , Selection of the optimal set of relevant words in consumers opinions in the context of the opinion mining ..	115
<b>Aleksandra Łuczak</b> , Application of AHP-LP to the evaluation of importance of determinants of socio-economic development in the administrative units .....	125
<b>Aleksandra Witkowska, Marek Witkowski</b> , A dynamic approach to the ranking of cooperative banks by their financial condition .....	134
<b>Adam Depta</b> , Application of correspondence analysis for the measurement of quality of life – questionnaire SF-36v2 based research .....	145
<b>Marek Lubicz, Maciej Zięba, Konrad Pawelczyk, Adam Rzechonek, Marek Marciniak, Jerzy Kołodziej</b> , Classification rules extraction for missing and imbalance data: models of classifiers and initial results in the rules-based thoracic surgery risk prediction.....	155
<b>Małgorzata Misztal</b> , Selected methods for assessing the performance of classifiers – an overview and examples of applications.....	166
<b>Anna M. Olszewska</b> , The application of selected quantitative methods to the evaluation of voivodeship innovation level potential.....	176
<b>Iwona Bąk</b> , The comparison of the quality of groupings of poviats of West Pomeranian Voivodeship in terms of tourism attractiveness .....	185
<b>Agnieszka Kozera, Joanna Stanisławska, Romana Głowicka-Wołoszyn</b> , Household segmentation with respect to the expenditure on organized tourism.....	195
<b>Agnieszka Wałęga</b> , Synthetic approach in the analysis of economic coherence of households .....	204
<b>Joanna Banaś, Małgorzata Machowska-Szewczyk, Bożena Mroczek</b> , Using the correspondence analysis to examine the impact of wind turbines on the quality of life.....	213
<b>Joanna Banaś, Krzysztof Małecki</b> , Classification of measurement survey points of drivers on the boundary of Szczecin using symbolic variables...	221
<b>Aneta Becker</b> , The use granular information in the analysis of the requirements of the labor market.....	229
<b>Katarzyna Cheba, Joanna Hołub-Iwan</b> , The application of the correspondence analysis of patients segmentation on the medical service market .....	237
<b>Adam Depta, Iwona Staniec</b> , Identification of the factors that determine the quality of students life at universities in Lodz.....	246
<b>Katarzyna Dębkowska, Jarosław Kilon</b> , Association rules in the analysis of research results the Delphi method .....	253
<b>Anna Domagała</b> , About using Principal Component Analysis in Data Envelopment Analysis .....	263
<b>Alicja Grześkowiak</b> , Analysis of the digital divide in Poland at the individual and regional level .....	272

<b>Anna M. Olszewska, Anna Gryko-Nikitin</b> , Assessment of perception of quality of teaching at an institution of higher learning based on the ordinal data with the utilization of R environment.....	281
<b>Karolina Paradysz</b> , The hierarchical method of grouping poviats as a benchmark approach in the assessment of unemployment by BAEL in selected types of small areas .....	289
<b>Radosław Pietrzyk</b> , Comparison of methods of measuring the performance of investment funds portfolios.....	298
<b>Agnieszka Przedborska, Małgorzata Misztal</b> , Selected multivariate statistical analysis methods in the evaluation of efficacy of deep electromagnetic stimulation in patients with degenerative joint disease .....	307
<b>Wojciech Roszka, Marcin Szymkowiak</b> , A calibration approach in statistical data integration .....	315
<b>Iwona Skrodzka</b> , Application of some methods of classification to the analysis of human capital in the European Union.....	325
<b>Agnieszka Stanimir</b> , Multivariate analysis of social inclusion factors.....	333
<b>Dorota Strózik, Tomasz Strózik</b> , Spatial differentiation of the standard of living in Great Poland Voivodeship .....	342
<b>Izabela Szamrej-Baran</b> , Identification of fuel poverty causes in Poland using soft modelling .....	352
<b>Janusz Tuchowski, Katarzyna Wójcik</b> , Classification of objects in the National Classification Framework described by the ontology.....	360
<b>Aleksandra Matuszewska-Janica</b> , Clustering of European Union states taking into consideration the levels of feminization of economic sectors..	368
<b>Monika Rozkrut, Dominik Rozkrut</b> , Identification of service sector innovation strategies in Poland.....	379

**Wojciech Roszka, Marcin Szymkowiak**

Uniwersytet Ekonomiczny w Poznaniu

---

## **PODEJŚCIE KALIBRACYJNE W STATYSTYCZNEJ INTEGRACJI DANYCH**

---

**Streszczenie:** Statystyczna integracja danych jest grupą metod umożliwiających łączną obserwację cech nieobserwowanych wspólnie w żadnym z dostępnych źródeł danych. Efektem zastosowania tych metod jest utworzenie jednostkowego zbioru danych zawierających informacje o zmiennych ze wszystkich integrowanych źródeł. W konsekwencji możliwa jest analiza zmiennych pochodzących z różnych zbiorów danych. Dzięki takiemu podejściu istnieje na przykład możliwość konstrukcji tabeli kontyngencji zawierającej łączny rozkład zmiennych nieobserwowanych wspólnie w żadnym ze zbiorów. W artykule przedstawiono sposób konstrukcji tabel kontyngencji zmiennych nieobserwowanych jednocześnie w dwóch zbiorach danych z wykorzystaniem techniki parowania statystycznego uwzględniającego metody kalibracji.

**Słowa kluczowe:** kalibracja, statystyczna integracja danych, parowanie statystyczne.

### **1. Wstęp**

Przeprowadzane przez organy statystyki publicznej badania reprezentacyjne odpowiadają na zapotrzebowanie informacyjne różnych instytucji państwowych i prywatnych. Ich zawartość merytoryczna wynika nie tylko z potrzeb odbiorców, ale również z konieczności prowadzenia analiz porównawczych różnych zjawisk społeczno-ekonomicznych w krajach Unii Europejskiej. Jednocześnie, ze względu na bardzo duże koszty, jak również obciążenie respondentów skutkujące zwiększoną liczbą odmów i braków odpowiedzi, nie przeprowadza się badań kompleksowo ujmujących zjawiska społeczno-gospodarcze. Z tych powodów obecny proces modernizacji infrastruktury statystycznej obejmuje m.in. zwiększenie wydajności systemów sprawozdawczości statystycznej poprzez integrację informacji z dostępnych źródeł danych [Leulescu, Agafitei 2013, s. 28-30, 70].

Parowanie statystyczne (*statistical matching, data fusion*) jest techniką zapewniającą uzyskanie łącznej informacji statystycznej w oparciu o zmienne i wskaźniki poddane pomiarowi w dwóch lub więcej większej liczby rozłącznych źródeł. Potencjalne korzyści z tego podejścia wynikają w możliwości zwiększenia zakresu



informacyjnego istniejących źródeł danych bez zwiększania kosztów badań i dodatkowych obciążeń respondentów.

Celem niniejszego artykułu jest opis wybranych metod statystycznej integracji dla danych uzyskanych w oparciu o złożone schematy losowania (jak to ma miejsce w przypadku badań reprezentacyjnych statystyki publicznej). Jednocześnie przedstawiony zostanie empiryczny przykład zastosowania opisywanych metod poprzez integrację informacji z Badania Dochodów i Warunków Życia (EU-SILC) oraz Badania Budżetów Gospodarstw Domowych.

## 2. Statystyczna integracja danych

Najczęściej opisywanym w literaturze celem statystycznej integracji danych jest zwiększenie zakresu informacyjnego różnych źródeł informacji. Wymienia się przy tym dwa podstawowe podejścia metodologiczne:

- probabilistyczne łączenie rekordów (*probabilistic record linkage*) – łączenie repozytoriów danych nieposiadających unikatowego klucza połączeniowego, zawierających informacje o tych samych jednostkach;
- parowanie statystyczne – integracja dwóch (lub więcej) rozłącznych (w sensie pokrycia) źródeł danych (zwykle pochodzących z badań próbkowych) odnoszących się do tej samej populacji generalnej.

Pierwsze z podejść wykorzystywane jest najczęściej do integracji repozytoriów administracyjnych i badań pełnych (np. spisów ludności); rekordy w każdym ze źródeł charakteryzują te same jednostki. Możliwa jest więc identyfikacja określonych jednostek w integrowanych zbiorach i połączenie informacji je charakteryzujących.

W przypadku integracji danych pochodzących z badań próbkowych, ze względu na ich rozłączność<sup>1</sup>, integruje się jednostki podobne pod względem wybranych cech. Integracji można dokonać dwojako: tworząc pełny, syntetyczny zbiór danych, zawierający łączną obserwację cech nieobserwowanych wspólnie w pojedynczych źródłach (tzw. podejście mikro) lub tworząc szacunki określonych związków (np. korelacji, współczynników regresji, tabeli kontyngencji) między zmiennymi łącznie nieobserwowanymi (tzw. podejścia makro)<sup>2</sup>. Celem artykułu jest utworzenie tabeli kontyngencji cech występujących w rozłącznych zbiorach.

Niech zbiór  $A$  zawiera zmienne  $X = [1, \dots, I]$  oraz  $Y = [1, \dots, J]$ . Zbiór  $B$  natomiast zmienne  $X = [1, \dots, I]$  oraz  $Z = [1, \dots, K]$ , gdzie  $i, j, k$  to warianty zmiennych, odpowiednio,  $X, Y$  i  $Z$ . Szacowanym parametrem jest wówczas:

$$\theta_{ijk} = P(X = i, Y = j, Z = k), \quad (1)$$

gdzie  $0 \leq \theta_{ijk} \leq 1$  oraz  $\sum_{i,j,k} \theta_{ijk} = 1$ .

<sup>1</sup> Prawdopodobieństwo wylosowania tej samej jednostki do dwóch badań jest bardzo małe i zwykle zakłada się, że jest zerowe [D’Orazio i in. 2006, s. 158-159].

<sup>2</sup> Opis procedury parowania statystycznego opisany został w [Roszka 2013, s. 175-181].

Dla zmiennych jakościowych szukany parametrem jest częstość (1). Przy założeniu o warunkowej niezależności (*Conditional Independence Assumption* – CIA) można wyznaczyć [D’Orazio et al. 2006, s. 13, 23-24]:

$$P(X = i, Y = j, Z = k) = P(Y = j|X = i)P(Z = k|X = i)P(X = i), \quad (2)$$

$$\theta_{ijk} = \theta_{j|i}\theta_{k|i}\theta_{i..} = \frac{\theta_{ij..}\theta_{i..k}}{\theta_{i..}\theta_{ij.}}\theta_{i..} = \frac{\theta_{ij..}\theta_{i..k}}{\theta_{i..}}, \quad (3)$$

gdzie „kropka” oznacza liczebność brzegową z wyłączeniem wariantu odpowiedniej zmiennej.

Wartości brzegowe tabeli  $Y \times Z$  uzyskiwane są z:

$$\sum_i \theta_{ijk} = \sum_{i=1}^I \frac{\theta_{ij..}\theta_{i..k}}{\theta_{i..}}. \quad (4)$$

Niech  $n_{A,ij}$  będą liczebnościami w tabeli  $X \times Y$  uzyskanej ze zbioru  $A$ ,  $n_{B,i.k}$  – liczebnościami w tabeli  $X \times Z$  uzyskanej ze zbioru  $B$ . Wykorzystując estymator największej wiarygodności<sup>3</sup> [Anderson 1957, s. 200-203], otrzymuje się:

$$\hat{\theta}_{i..} = \frac{n_{A,i.} + n_{B,i.}}{n_A + n_B}, \quad (5)$$

$$\hat{\theta}_{j|i} = \frac{n_{A,ij.}}{n_{A,i.}}, \quad (6)$$

$$\hat{\theta}_{k|i} = \frac{n_{B,i.k}}{n_{B,i.}}. \quad (7)$$

### 3. Podejście Renssena

Metodą, w której wykorzystuje się informacje pochodzące ze schematu losowania próby, jest podejście kalibracyjne Renssena [1998, s. 171-183]. Oparte jest ono na algorytmie kalibracji wag analitycznych wynikających ze schematu losowania, oddzielnie dla  $A$  i  $B$ . Wynikiem procedury Renssena jest tabela kontyngencji  $Y \times Z$ .

Niech  $d_k$  oznacza wagi początkowe, a  $w_k$  finalne wagi kalibracyjne. Wagi finalne uzyskiwane są jako rozwiązanie zagadnienia optymalizacji

$$\min[\sum_{k \in S} D(d_k, w_k)],$$

gdzie  $D(d, w)$  to miara odległości, z zastrzeżeniem, że  $\sum_{k=1}^n w_k x_k = \sum_{k=1}^n d_k x_k$  oraz  $\sum_{k=1}^n w_k = N$ . Szczegółowy opis podejścia kalibracyjnego można znaleźć w pracy [Särndal, Lundström 2005; Szymkowiak 2009, s. 90-105].

Pierwsza faza polega na harmonizacji wag w integrowanych zbiorach. Wybiera się podzbiór zmiennych  $X_1 \subseteq X$ , dla których znane są liczebności w populacji generalnej:

<sup>3</sup> Raessler [2002] zaproponowała szacowanie tego parametru z wykorzystaniem KMNK.

- wagi  $w_a$  w zbiorze  $A$  są kalibrowane w taki sposób, by wagi kalibracyjne  $w_a^{(1)}$  spełniały warunek

$$\sum_{a \in A} w_a^{(1)} x_{1a} = t_1,$$

gdzie  $t_1$  oznacza wektor wartości globalnych w populacji,

- wagi  $w_b$  w zbiorze  $B$  są kalibrowane w taki sposób, by wagi kalibracyjne  $w_b^{(1)}$  spełniały warunek

$$\sum_{b \in B} w_b^{(1)} x_{1b} = t_1.$$

Jeżeli istnieją jakieś zmienne  $X_2 \subseteq X$ , dla których wartości globalne w populacji nie są znane, w kolejnym kroku wyznaczany jest łączny estymator (*pooled estimate*):

$$\hat{t}_2 = \lambda \sum_{a \in A} w_a^{(1)} x_{2a} + (1 - \lambda) \sum_{b \in B} w_b^{(1)} x_{2b}, \quad (8)$$

gdzie  $0 \leq \lambda \leq 1$ . Następnie wagi  $w_a^{(1)}$  i  $w_b^{(1)}$  są rekalkulowane w taki sposób, że:

- w zbiorze  $A$  powstają wagi  $w_a^{(2)}$  spełniające warunek

$$\sum_{a \in A} w_a^{(2)} x_{1a} = t_1 \text{ oraz } \sum_{a \in A} w_a^{(2)} x_{2a} = \hat{t}_2,$$

- w zbiorze  $B$  powstają wagi  $w_b^{(2)}$  spełniające warunek

$$\sum_{b \in B} w_b^{(2)} x_{1b} = t_1 \text{ oraz } \sum_{b \in B} w_b^{(2)} x_{2b} = \hat{t}_2.$$

W drugim etapie wagi kalibracyjne  $w_a^{(2)}$  i  $w_b^{(2)}$  mogą zostać użyte do wyznaczenia estymatorów łącznych rozkładów w  $A$  i  $B$ . Dla zmiennych jakościowych, przy założeniu CIA, łączny rozkład  $Y$  i  $Z$  może zostać wyznaczony za pomocą (3).

W praktyce zdarzają się sytuacje, w których procedura kalibracji w podejściu Renssena jest nieskuteczna (tzn. algorytm nie osiąga zbieżności, pojawiają się ujemne wagi itp.). Ma to miejsce zwłaszcza w przypadku, gdy wektor  $X$  zawiera zmienne mierzone na różnej skali lub (i) gdy zmienne jakościowe charakteryzują się dużą liczbą wariantów. W takich przypadkach należy grupować warianty cech jakościowych lub (i) kategoryzować zmienne ilościowe.

#### 4. Badanie empiryczne

Głównym celem badania była integracja informacji pochodzących z dwóch badań reprezentacyjnych prowadzonych przez Główny Urząd Statystyczny, tj. Badania Budżetów Gospodarstw Domowych (BBGD) i Badania Dochodów i Warunków Życia (EU-SILC). Na potrzeby przykładu empirycznego wykorzystano jednostkowe zbiory danych dla gospodarstw domowych z 2005 roku.

Integracja miała na celu stworzenie dwuwymiarowej tabeli kontyngencji pomiędzy zmiennymi:  $Y$  – Wydatki ogółem gospodarstw domowych i  $Z$  – Czy gospodarstwo stać na tygodniowy urlop poza miejscem zamieszkania, zgodnie z podejściem zaproponowanym przez Renssena [1998]. Pierwsza z wymienionych zmiennych była obserwowana wyłącznie w zbiorze BBGD, a druga wyłącznie w EU-SILC. Oznacza to, że utworzenie tabeli kontyngencji nie byłoby możliwe przy wykorzystaniu każdego zbioru z osobna.

W charakterze zmiennych wspólnych, składających się na wektor  $X$  i występujących w obydwu zbiorach, przyjęto:  $X_1$  – Region (NUTS1),  $X_2$  – Rodzaj budynku,  $X_3$  – Tytuł prawny do zajmowanego mieszkania,  $X_4$  – Typ biologiczny gospodarstwa domowego,  $X_5$  – Czy jest ustęp splukiwany,  $X_6$  – Czy jest łazienka,  $X_7$  – Czy gospodarstwo posiada TV,  $X_8$  – Czy gospodarstwo posiada komputer,  $X_9$  – Czy gospodarstwo posiada samochód,  $X_{10}$  – Ekwiwalentna wielkość gospodarstwa domowego,  $X_{11}$  – Ekwiwalentny dochód gospodarstwa domowego. Wektor zmiennych wspólnych  $X$  poddano procesowi harmonizacji w taki sposób, że zapewniono zgodność ich wariantów i sposobu kodowania.

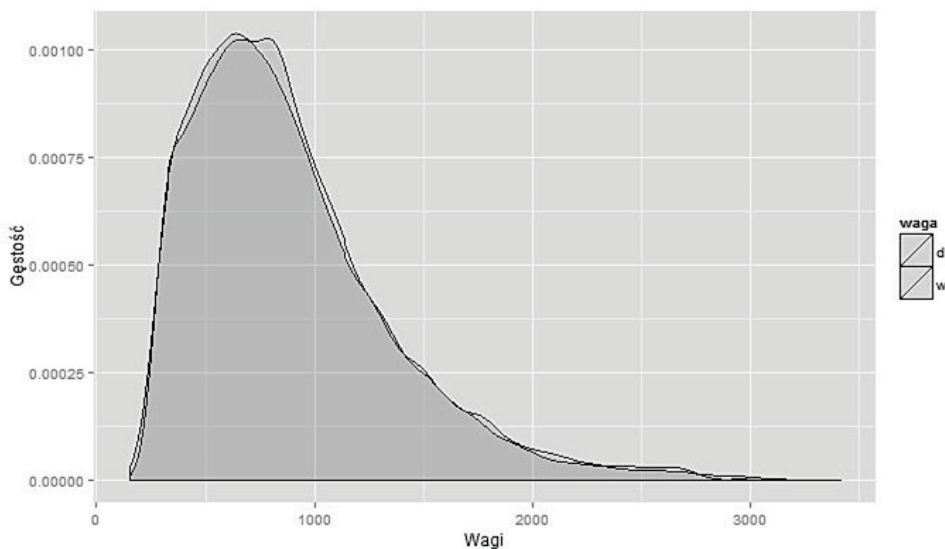
W następnym kroku dokonano harmonizacji wag w integrowanych zbiorach. Ponieważ nie były znane wartości globalne poszczególnych zmiennych wspólnych, dokonano ich oszacowania zgodnie ze wzorem (8). Parametr  $\lambda$  wyznaczono zgodnie ze wzorem  $\lambda = \frac{n_A}{n_A+n_B}$ , gdzie  $n_A$  to liczebność zbioru BBGD, a  $n_B$  to liczebność zbioru EU-SILC. W dalszym etapie prac dokonano rekaliibracji wag z BBGD i EU-SILC zgodnie ze schematem opisanym w punkcie 3 artykułu. Jak pokazują wykresy na rys. 1 i 2, rozkłady wyznaczonych wag kalibracyjnych, zarówno w badaniu EU-SILC, jak i BBGD, w niewielkim stopniu odbiegają od rozkładów wag wejściowych wynikających ze schematu losowania próby.

Jest to zgodne z ideą kalibracji, tj. wagi kalibracyjne nie tylko odtwarzają oszacowane wartości globalne dla wszystkich zmiennych wspólnych w poszczególnych zbiorach, ale również nieznacznie różnią się, w sensie przyjętej funkcji odległości, od wag wejściowych (por. tabela 1).

**Tabela 1.** Charakterystyki rozkładu wag wejściowych i kalibracyjnych

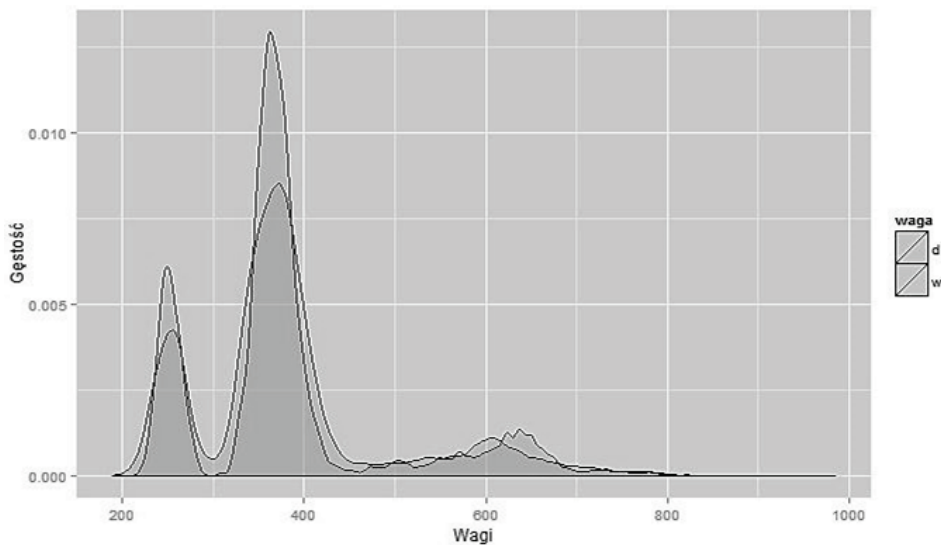
Statystyka	Wagi wejściowe		Wagi kalibracyjne	
	BBGD	EU-SILC	BBGD	EU-SILC
Liczebność	34 767	14 914	34 767	14 914
Średnia	383,5	893,0	383,6	893,6
Odch. stand.	114,1	456,2	113,7	472,5
Mediana	365,3	806,4	367,1	793,9
Minimum	214,5	268,0	189,5	152,4
Maksimum	882,5	3029,1	984,9	34 15,7
Skośność	1,3	1,2	1,3	1,3
Kurtoza	1,6	1,9	1,7	2,2

Źródło: opracowanie własne.



**Rys. 1.** Rozkład wag wejściowych (d) i kalibracyjnych (w) w EU-SILC

Źródło: opracowanie własne.



**Rys. 2.** Rozkład wag wejściowych (d) i kalibracyjnych (w) w BBGD

Źródło: opracowanie własne.

Dla tak wyznaczonych wag kalibracyjnych w dalszym kroku zbudowano tabelę kontyngencji dla zmiennych Y i Z (por. tabela 2) przy założeniu warunkowej niezależności (wzory 2-7).

**Tabela 2.** Tabela kontyngencji cech Y i Z (liczebności i odsetki wierszowe)

Wydatki	Czy stać na urlop?			Wydatki	Czy stać na urlop?		
	tak	nie	ogółem		tak	nie	ogółem
do 1000	415 618	2 363 869	2 779 487	do 1000	15%	85%	100%
1000-1500	724 001	2 430 836	3 154 837	1000-1500	23%	77%	100%
1500-2000	749 870	1 743 275	2 493 145	1500-2000	30%	70%	100%
2000-2500	624 005	1 089 996	1 714 001	2000-2500	36%	64%	100%
powyżej 2500	1 631 404	1 453 827	3 085 232	powyżej 2500	53%	47%	100%
Ogółem	4 144 898	9 081 803	13 226 701	Ogółem	31%	69%	100%

Źródło: opracowanie własne.

Dla utworzonej tabeli kontyngencji przeprowadzono test niezależności  $\chi^2$ . *P-value* < 0,001 oznacza istotną w sensie statystycznym zależność pomiędzy badanymi cechami. Z analizy danych zawartych w tabeli 2 wynika, że wzrostowi wydatków towarzyszy wzrost zdolności gospodarstwa do sfinansowania urlopu wypoczynkowego. Dla tak wyznaczonej tabeli kontyngencji zostały ponadto zachowane rozkłady brzegowe cech Y i Z.

## 5. Podsumowanie

Zaprezentowana w artykule metoda wyznaczania tabeli kontyngencji dla cech nieobserwowanych łącznie może stanowić cenne źródło zasilania informacyjnego. Oszacowanie łącznych charakterystyk takich zmiennych nie wymaga bowiem przeprowadzenia dodatkowych badań. W konsekwencji możliwa jest redukcja kosztów i zmniejszenie obciążeń respondentów. Przedstawiona w artykule idea łącznego wykorzystania informacji pochodzących z różnych źródeł wpisuje się w rozwijający się w świecie nurt statystycznej integracji danych. Techniki te będą odgrywały coraz większą rolę w praktyce badań prowadzonych przez urzędy statystyczne [U.S. Bureau of the Census 1999]. W środowisku statystyków zajmujących się metodami integracji danych pochodzących z różnych źródeł panuje bowiem powszechne przekonanie, że XXI wiek zdominowany zostanie przez techniki statystycznej integracji danych [Zhang 2011, s. 445].

## Literatura

- Anderson T.W. (1957), *Maximum likelihood estimates for a multivariate normal distribution when some observations are missing*, „Journal of the American Statistical Association” 52.
- D’Orazio M. (2011), *Statistical Matching and Imputation of Survey Data with the Package StatMatch for the REEnvironment*, Italian National Institute of Statistics (Istat), Rome, Italy.
- D’Orazio M., Di Zio M., Scanu M. (2006), *Statistical Matching. Theory and Practice*, John Wiley & Sons Ltd., England.

- Leulescu A., Agafitei M. (2013), *Statistical matching: a model based approach for data integration*, Eurostat Methodologies and Working Papers. [http://epp.eurostat.ec.europa.eu/cache/ITY\\_OFFPUB/KS-RA-13-020/EN/KS-RA-13-020-EN.PDF](http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-RA-13-020/EN/KS-RA-13-020-EN.PDF).
- Raessler S. (2002), *Statistical Matching: A Frequentist Theory, Practical Applications and Alternative Bayesian Approaches*, Springer, New York.
- Renssen R.H. (1998), *Use of statistical matching techniques in calibration estimation*, Survey Methodology 24.
- Roszka W. (2013), *Szacowanie łącznych charakterystyk cech nieobserwowanych łącznie*, [w:] *Klasyfikacja i analiza danych – teoria i zastosowania*, red. K. Jajuga, M. Walesiak, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu nr 279, Taksonomia 21, Wydawnictwo UE, Wrocław.
- Särndal C.E., Lundström S. (2005), *Estimation in Surveys with Nonresponse*, Wiley.
- Szymkowiak M. (2009), *Imputacja i kalibracja – nowe możliwości estymacji w badaniach statystycznych z brakiem odpowiedzi*, Zeszyty Naukowe nr 116, Wydawnictwo Uniwersytetu Ekonomicznego w Poznaniu, Poznań.
- U.S. Bureau of the Census (1999), *Integrated Information Solutions – The Future of Census Bureau Data Access and Dissemination*, Statistical Commission and Economic Commission for Europe, Conference of European Statisticians.
- Zhang Li-Chun (2011), *Topics of statistical theory for register-based statistics*, 58<sup>th</sup> World Statistical Congress, Dublin, <http://2011.isiproceedings.org/papers/450014.pdf>.

## A CALIBRATION APPROACH IN STATISTICAL DATA INTEGRATION

**Summary:** Statistical data integration comprises a group of methods enabling joint observation of variables which are not observed together in any of the available data sources. Depending on the approach adopted, these methods make it possible to create a dataset of units combining information about variables from all integrated sources or a contingency table containing a joint distribution of variables which are not observed together in any dataset. The article presents a method of constructing contingency tables of variables which are not observed together in two datasets by applying methods of statistical matching and calibration.

**Keywords:** calibration, statistical data integration, statistical matching.