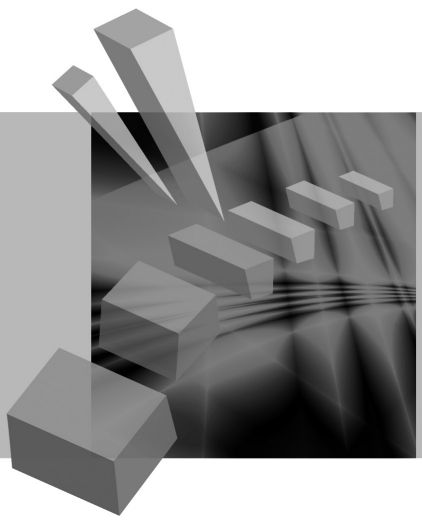
**PRACE NAUKOWE**
Uniwersytetu Ekonomicznego we Wrocławiu
**RESEARCH PAPERS**
of Wrocław University of Economics

**279**

# Taksonomia 21

## Klasyfikacja i analiza danych – teoria i zastosowania



Redaktorzy naukowi
**Krzysztof Jajuga**
**Marek Walesiak**

Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2013

Redaktor Wydawnictwa: Aleksandra Śliwka
Redaktor techniczny: Barbara Łopusiewicz
Korektor: Barbara Cibis
Łamanie: Małgorzata Czupryńska
Projekt okładki: Beata Dębska

Publikacja jest dostępna w Internecie na stronach:
www.ibuk.pl, www.ebscohost.com,
The Central and Eastern European Online Library www.ceeol.com,
a także w adnotowanej bibliografii zagadnień ekonomicznych BazEkon
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

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



**ISSN 1899-3192** (Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu)
**ISSN 1505-9332** (Taksonomia)

Wersja pierwotna: publikacja drukowana
Druk: Drukarnia TOTEM

## Spis treści

## Summaries

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS nr 279 • 2013

Taksonomia 21. Klasyfikacja i analiza danych – teoria i zastosowania ISSN 1899-3192

**Wojciech Roszka**
Uniwersytet Ekonomiczny w Poznaniu

# SZACOWANIE ŁĄCZNYCH CHARAKTERYSTYK CECH NIEOBSERWOWANYCH ŁĄCZNIE

**Streszczenie:** Zwiększające się zapotrzebowanie na aktualne komunikaty statystyczne stanowi rosnące wyzwanie dla instytucji badawczych, zarówno państwowych, jak i prywatnych. Duże koszty przeprowadzenia nowych badań powodują stosunkowo niedużą częstotliwość ich realizacji. Wykorzystanie metod statystycznej integracji danych umożliwia łączenie dostępnych repozytoriów danych w sposób umożliwiający szacowanie łącznych charakterystyk cech nieobserwowanych łącznie w pojedynczych źródłach. Metody te mogą stanowić dodatkowe źródło informacyjne dla badań społeczno-ekonomicznych.

**Słowa kluczowe:** statystyczna integracja danych, badania społeczno-ekonomiczne, zasilanie informacyjne gospodarki.

## 1. Wstęp

Informacja w dzisiejszym społeczeństwie pełni istotną funkcję, w szczególności jako podstawa podejmowania decyzji zarówno administracyjnych, społecznych (np. kierowanie inwestycji w rejony najbardziej ich potrzebujące), jak i biznesowych (np. kierowanie kampanii marketingowych do odpowiednich segmentów rynkowych). Dlatego też podmioty zgłaszające popyt na informacje oczekują, by była ona rzetelna oraz aktualna. Przeprowadzenie badania specjalnego bardzo kosztownego i trwającego wiele dni, a nawet tygodni, powoduje utratę aktualności informacji, zmniejszając jej użyteczność.

Rozwiązaniem problemu dostępności informacji spełniającej wymogi określone nie tylko przez niezależne organizacje międzynarodowe i instytuty statystyki publicznej, ale przede wszystkim formułowane przez gospodarkę wydają się metody statystycznej integracji danych. Polegają one na łączeniu informacji z dostępnych źródeł danych w taki sposób, by możliwa była łączna obserwacja cech nieobserwowanych łącznie w pojedynczych repozytoriach danych. Wykorzystanie różnorodnych źródeł danych nie tylko pozwala na oszczędność kosztów i czasu, ale umożliwia również łączenie zasobów informacyjnych już istniejących baz, generując efekt synergii informacyjnej.

Celem niniejszego artykułu jest weryfikacja możliwości wykorzystania metod statystycznej integracji danych w celu zapewnienia informacji o łącznych charakterystykach cech niewystępujących łącznie w pojedynczym źródle danych. Cel zostanie osiągnięty poprzez badanie empiryczne, w którym zintegrowane zostaną zbiory Badania Budżetów Gospodarstw Domowych oraz Badania Dochodów i Jakości Życia EU-SILC z 2005 r. Przeprowadzone zostanie badanie współzależności między cechami niewystępującymi łącznie w żadnym z badań. Przy spełnieniu określonych założeń oszacowany zostanie przedział możliwych wartości współczynnika korelacji.

## 2. Idea statystycznej integracji danych

Metodyka statystycznej integracji danych polega na łączeniu dwóch (lub więcej) źródeł danych niezawierających unikatowego klucza połączeniowego w sposób umożliwiający oszacowanie łącznych charakterystyk cech z obu zbiorów [Raessler 2002; Di Zio i in. 2006]. Metodologia ta jest szeroka i zawiera techniki łączenia zbiorów danych zarówno zawierających informacje o tych samych jednostkach (probabilistyczne łączenie rekordów, *probabilistic record linkage*), jak i nie zawierających takich informacji (parowanie statystyczne[1], *statistical matching*).

Parowanie statystyczne to grupa metod służących do integracji dwóch (lub więcej) źródeł danych zwykle pochodzących z badań próbkowych odnoszących się do tej samej populacji generalnej. Ponieważ prawdopodobieństwo wylosowania tej samej jednostki do dwóch różnych badań reprezentacyjnych jest bardzo małe (zbliżone do zera), zakłada się, że integrowane zbiory są rozłączne w sensie pokrycia. W każdym zbiorze (oznaczono je jako $A$ i $B$) znajduje się zwykle pewien wspólny wektor zmiennych o tych samych lub zbliżonych definicjach i wariantach. Nazywa się je zmiennymi wspólnymi (oznaczonymi jako $\boldsymbol{X}$). Zbiór $A$ zawiera wektor zmiennych obserwowanych wyłącznie w nim, oznaczony jako $\boldsymbol{Y}$, natomiast zbiór $B$ zawiera analogiczny wektor – $\boldsymbol{Z}$ (por. rys. 1). Celem parowania statystycznego jest analiza związków pomiędzy zmiennymi $\boldsymbol{Y}$ i $\boldsymbol{Z}$.

Algorytm statystycznej integracji danych metodą parowania statystycznego inicjowany jest poprzez identyfikację wektora zmiennych wspólnych $\boldsymbol{X}$. Są to zmienne występujące w obu zbiorach charakteryzujące się takimi samymi lub podobnymi definicjami. W przypadku braku pełnej spójności definicji zmiennych wspólnych należy przeprowadzić etap ich harmonizacji.

Dalszym elementem algorytmu parowania statystycznego jest wybór zmiennych parujących. Wektor zmiennych wspólnych $\boldsymbol{X}$ może zawierać wiele zmiennych o różnej mocy predykcyjnej wyjaśniającej związek ze zmienną (zmiennymi) dołączanymi $\boldsymbol{Y}$ lub $\boldsymbol{Z}$. Zastosowanie byt wielu zmiennych w procesie łączenia baz danych może

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

Zbiór A

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

Zbiór B

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

**Rys. 1.** Dane wejściowe w parowaniu statystycznym

Źródło: opracowanie własne.

prowadzić do błędnego odzwierciedlenia łącznego rozkładu ($\boldsymbol{XYZ}$) [D'Orazio 2012]. Praktyka pokazuje, że zmiennych parujących powinno być „mało" (optymalnie 4-6), co w znaczny sposób nie tylko przyspiesza proces integracji w sensie obliczeniowym, ale również ułatwia interpretację otrzymanych modeli [Di Zio i in. 2006]. Wybór zmiennych parujących ze zbioru wektora zmiennych wspólnych $\boldsymbol{X}$ może zostać dokonany dwojako: w sposób ekspercki przez specjalistów lub za pomocą metod statystycznych.

Metoda ekspercka uwzględnia wiedzę merytoryczną z danej dziedziny, natomiast wykorzystując metody statystyczne, przeprowadza się analizę współzależności[2] między cechami $\boldsymbol{X}$ a $\boldsymbol{Y}$ w zbiorze A oraz $\boldsymbol{X}$ i $\boldsymbol{Z}$ w zbiorze B. Jako zmienne parujące wyznacza się podzbiór cech $\boldsymbol{X}$ istotnie korelujący z cechami zarówno $\boldsymbol{Y}$, jak i $\boldsymbol{Z}$ [Singh i in. 1990; Cohen 1991].

Wybierając metodę integracji danych, należy rozważyć cel integracji, stawiane założenia, charakter dołączanych zmiennych, dostępność informacji dodatkowych oraz możliwość wykorzystania informacji płynącej ze schematu losowanie próbek. W parowaniu statystycznym zasadniczo wyróżnia się dwa główne podejścia metodologiczne [Di Zio i in. 2006]:

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

[2] Współzależność cech najczęściej rozpatruje się wielowymiarowo, np. za pomocą drzew klasyfikacyjnych i regresyjnych, analizy skupień, analizy czynnikowej, a także metod eliminujących współliniowość wektora cech $\boldsymbol{X}$, np. metody odwróconej macierzy korelacji.

W artykule rozważane będzie podejście makro.

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

$$f(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}) = f_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x}) f_{\boldsymbol{Z}|\boldsymbol{X}}(\boldsymbol{z}|\boldsymbol{x}) f_{\boldsymbol{X}}(\boldsymbol{x}),\ \forall\, \boldsymbol{x} \in \mathcal{X}, \boldsymbol{y} \in \mathcal{Y}, \boldsymbol{z} \in \mathcal{Z}, \tag{1}$$

gdzie $f_{\boldsymbol{Y}|\boldsymbol{X}}$ to warunkowa funkcja gęstości dla $\boldsymbol{Y}$ przy danym $\boldsymbol{X}$, $f_{\boldsymbol{Z}|\boldsymbol{X}}$ to warunkowa funkcja gęstości dla $\boldsymbol{Z}$ przy danym $\boldsymbol{X}$, a $f_{\boldsymbol{X}}$ to gęstość brzegowa $\boldsymbol{X}$. Przy prawdziwości założenia o warunkowej niezależności do oszacowania (1) wystarczą informacje o brzegowym rozkładzie $\boldsymbol{X}$**,** a także o związkach pomiędzy $\boldsymbol{X}$ i $\boldsymbol{Y}$ oraz $\boldsymbol{X}$ i $\boldsymbol{Z}$. Informacje te dostępne są w zbiorach, odpowiednio, $A$ i $B$.

Założenie warunkowej niezależności jest trudne do spełnienia w rzeczywistości, a jednocześnie jego zweryfikowanie nie jest możliwe przy użyciu informacji płynących z $A \cup B$. W takim przypadku należy przeprowadzić analizę niepewności umożliwiającą wyznaczenie przedziału wiarygodnych łącznych charakterystyk cech nieobserwowanych łącznie.

## 3. Analiza niepewności

Jeżeli założenie warunkowej niezależności jest nieprawdziwe i nie występują dodatkowe informacje, których można by użyć w toku integracji, należy przeanalizować tzw. przestrzeń niepewności. Jest to zbiór wszystkich możliwych rozkładów zmiennych losowych $(\boldsymbol{Y},\boldsymbol{Z}|\boldsymbol{X})$ zgodnych z dostępną informacją, tj. obserwowanym brzegowym rozkładem $(\boldsymbol{Y}|\boldsymbol{X})$ oraz $(\boldsymbol{Z}|\boldsymbol{X})$ [D'Orazio 2012].

Produktem zastosowania metod parowania statystycznego przy niepewności dla podejścia makro są przedziały wiarygodnych wartości szacowanych parametrów (np. wariancji, kowariancji, korelacji). Przy braku dodatkowej informacji o wartości $\rho_{YZ}$ lub $\rho_{YZ|X}$ i przy braku założenia o warunkowej niezależności jedyną dostępną informacją jest [Kadane 1978; Rubin 1986; Moriarity, Scheuren 2001; 2003]:

$$\rho_{XY}\rho_{XZ} - \sqrt{[(1-\rho_{XY}^2)(1-\rho_{XZ}^2)]} \le \rho_{YZ} \le \rho_{XY}\rho_{XZ} + \sqrt{[(1-\rho_{XY}^2)(1-\rho_{XZ}^2)]} \tag{2}$$

ze względu na fakt, że macierz korelacji musi być dodatnio półokreślona ($\det\rho \ge 0$). Szacunek $\rho_{YZ} = \rho_{XY}\rho_{XZ}$ jest centralnym punktem przedziału.

Dla przypadku z wieloma zmiennymi macierz korelacji ma postać:

$$\boldsymbol{\Sigma} = \begin{pmatrix} \Sigma_{XX} & \Sigma_{YX} & \Sigma_{ZX} \\ \Sigma_{XY} & \Sigma_{YY} & \Sigma_{ZY} \\ \Sigma_{XZ} & \Sigma_{YZ} & \Sigma_{ZZ} \end{pmatrix}. \tag{3}$$

Wartość wektora współczynników korelacji $\boldsymbol{YZ}$ wyznacza się ze wzoru [Kiesl, Raessler 2006]:

$$\Sigma_{YZ} = \Sigma_{ZX}\Sigma_{XX}^{-1}\Sigma_{XY}, \tag{4}$$

natomiast przedziały niepewności[3] dla (4) wyznacza się w dwóch krokach [Kiesl, Raessler 2006]:

Wyznaczenie wektorów własnych macierzy:

$$\tilde{C} = (I - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY})^{-1}(\Sigma_{ZZ} - \Sigma_{ZX}\Sigma_{XX}^{-1}\Sigma_{XZ})^{-1}. \tag{5}$$

Wyznaczenie długości półosi elipsoidy prawdopodobnych korelacji $\boldsymbol{YZ}$: $\frac{1}{\sqrt{\lambda_i}}$, gdzie $\lambda_i$ to *i*-ta wartość własna.

Przedział niepewności dla (4) przyjmuje więc postać:

$$\Sigma_{YZ} - \frac{1}{\sqrt{\lambda_i}} \leq \Sigma_{YZ} \leq \Sigma_{YZ} + \frac{1}{\sqrt{\lambda_i}}. \tag{6}$$

Im węższy jest przedział (6), tym mniejsza jest niepewność i tym lepiej wektor wybranych zmiennych parujących wyjaśnia nieznane powiązania między integrowanymi cechami $\boldsymbol{Y}$ i $\boldsymbol{Z}$.

## 4. Badanie empiryczne

Celem badania symulacyjnego jest integracja zbiorów danych zawierających informacje o gospodarstwach domowych: Badania Budżetów Gospodarstw Domowych oraz Badania Dochodów i Jakości Życia EU-SILC. Oba zbiory pochodzą z badań przeprowadzonych w 2005 r. Integracja umożliwi oszacowanie korelacji między zmienną rozchody netto gospodarstwa domowego (obserwowaną wyłącznie w BBGD, oznaczoną jako *Y*) i zmienną dochody głowy gospodarstwa domowego (wyłącznie w EU-SILC, oznaczoną jako *Z*). Celem szczegółowym integracji jest oszacowanie współczynnika korelacji między zmiennymi nieobserwowanymi łącznie w żadnym ze zbiorów przy wykorzystaniu wybranych metod wyznaczania zmiennych parujących.

Zmiennymi wspólnymi ($\boldsymbol{X}$) zawartymi w obu zbiorach były: typ własności gospodarstwa domowego, aktywność zawodowa głowy gospodarstwa domowego, płeć głowy gospodarstwa domowego, wykształcenie głowy gospodarstwa domowego, stan cywilny głowy gospodarstwa domowego, wielkość ekwiwalentna gospodarstwa domowego, wiek głowy gospodarstwa domowego, dochód ekwiwalentny głowy gospodarstwa domowego. Zmienne wspólne zharmonizowano w taki sposób, by ich kodowanie i rozkłady w obu zbiorach były analogiczne.

[3] Przedziały możliwych wartości $\Sigma_{YZ}$ zapewniających dodatnią półokreśloność macierzy korelacji Σ.

Liczebność zbioru danych BBGD wynosiła 34 767 gospodarstw domowych, zaś EU-SILC – 16 263. Liczebność populacji generalnej ustalono na 13 167 722 gospodarstw[4].

Wśród zmiennych wspólnych wybrano zmienne parujące metodami:

– eksperckimi:
  - ze względu na potrzebę oszacowania współczynnika korelacji[5] – tylko zmienne ciągłe,
  - wszystkie zmienne wspólne[6];
– statystycznymi:
  - eliminacja współliniowości w wektorze zmiennych wspólnych – metoda odwróconej macierzy korelacji[7],
  - wybór najsilniejszych predykant – metoda drzewa klasyfikacyjnego i regresyjnego[8] (CART).

Dodatkowo, ze względu na bardzo silną asymetrię rozkładu cech: rozchody gospodarstw domowych oraz dochody głów gospodarstw domowych[9], dokonano analogicznej analizy dla cech poddanych transformacji logarytmicznej.

Dla postaci oryginalnych współczynnik korelacji przy założeniu warunkowej niezależności cech $\boldsymbol{Y}$ i $\boldsymbol{Z}$ przy danym zestawie cech $\boldsymbol{X}$, w zależności od metody doboru cech parujących, wahał się w przedziale od 0,41 do 0,43 (por. tab. 1). Przedziały

**Tabela 1.** Szacunki $\Sigma_{YZ}$ oraz przedziały niepewności przy różnych zestawach zmiennych parujących dla oryginalnych postaci zmiennych dołączanych

| Metoda doboru zmiennych | $\Sigma_{YZ}$ | $\Sigma_{YZ} - \frac{1}{\sqrt{\lambda_i}}$ | $\Sigma_{YZ} + \frac{1}{\sqrt{\lambda_i}}$ | Szerokość przedziału |
|---|---|---|---|---|
| Ciągłe | 0,4157 | –0,1576 | 0,9890 | 1,1467 |
| Wszystkie | 0,4284 | –0,1217 | 0,9784 | 1,1001 |
| Odwrócona macierz korelacji | 0,4330 | –0,1160 | 0,9819 | 1,0978 |
| CART | 0,4338 | –0,1144 | 0,9819 | 1,0963 |

Źródło: opracowanie własne.

[4] Suma wag analitycznych w każdym ze zbiorów wejściowych.

[5] Formalne założenia współczynnika korelacji liniowej mówią, że analizowane cechy powinny mieć charakter ciągły [Aczel 2000].

[6] Cechy jakościowe zdychotomizowano i potraktowano jako ciągłe. Takie upraszczające podejście w szacowaniu macierzy korelacji między cechami jakościowymi i ilościowymi zaproponowano w [Kiesl, Raessler 2006; Di Zio i in. 2006].

[7] Metoda odwróconej macierzy korelacji opisana jest szczegółowo w [Witkowski, Klimanek 2006].

[8] Algorytm drzewa klasyfikacyjnego i regresyjnego opisany jest szczegółowo w [Gatnar, Walesiak (red.) 2009; Rószkiewicz 2002].

[9] Współczynnik asymetrii dla rozchodów wynosił 8,8, natomiast dla dochodów głów był równy 9,0.

niepewności, a więc wartości zapewniające dodatnią półokreśloność macierzy korelacji, były bardzo szerokie i zawierały wartość zerową. Taki szacunek współzależności między cechami nieobserwowanymi łącznie nie jest akceptowalny.

Transformacja logarytmiczna cech $\boldsymbol{Y}$ i $\boldsymbol{Z}$ umożliwiła utworzenie węższych przedziałów niepewności (por. tab. 2). Nie zawierają one wartości zerowej współczynnika korelacji. Jednocześnie wzrosła ocena wartości współczynnika korelacji między cechami do przedziału od 0,68 do 0,69, w zależności od doboru cech parujących.

**Tabela 2.** Szacunki $\Sigma_{YZ}$ oraz przedziały niepewności przy różnych zestawach zmiennych parujących dla zmiennych dołączanych poddanych transformacji logarytmicznej

| Metoda doboru zmiennych | $\Sigma_{YZ}$ | $\Sigma_{YZ} - \frac{1}{\sqrt{\lambda_i}}$ | $\Sigma_{YZ} + \frac{1}{\sqrt{\lambda_i}}$ | Szerokość przedziału |
|---|---|---|---|---|
| Ciągłe | 0,6826 | 0,4135 | 0,9517 | 0,5382 |
| Wszystkie | 0,6946 | 0,4476 | 0,9415 | 0,4939 |
| Odwrócona macierz korelacji | 0,6941 | 0,4450 | 0,9431 | 0,4981 |
| CART | 0,6953 | 0,4476 | 0,9430 | 0,4954 |

Źródło: opracowanie własne

Metoda CART, zapewniająca wybór predykant najbardziej wyjaśniających zmienność zmiennej objaśnianej, zapewniała jedne z najwęższych przedziałów niepewności. Nie uwzględnia ona jednak współliniowości wektora zmiennych parujących. Do wyznaczenia optymalnego wektora zmiennych parujących, uwzględniających zarówno moc predykcyjną, jak i współliniowość cech, potrzebne są dalsze badania.

## 5. Podsumowanie

Metody parowania statystycznego umożliwiły oszacowanie współczynnika korelacji liniowej między cechami obserwowanymi oddzielnie w rozłącznych w sensie pokrycia zbiorach danych. Wykorzystanie informacji uzyskanych ze zbiorów *A* i *B* dało możliwość utworzenia przedziałów wiarygodnych wartości współczynników korelacji analizowanych cech. Dzięki wykorzystaniu metod doboru zmiennych parujących, w tym taksonomicznych, możliwe było wybranie predykant najlepiej wyjaśniających związki między integrowanymi zmiennymi.

Brak dodatkowych informacji o łącznym rozkładzie $(\boldsymbol{XYZ})$ lub $(\boldsymbol{YZ})$ prowadzi do powstania szerokich przedziałów niepewności, często uniemożliwiających rzetelne określenie związków między integrowanymi cechami. W takich przypadkach konieczna może okazać się transformacja wejściowych danych.

Jako dalsze kierunki badań można wskazać wykorzystanie informacji dodatkowej o łącznych charakterystykach integrowanych zbiorów.

## Literatura

Aczel A.D. (2000), *Statystyka w zarządzaniu. Pełny wykład*, Wydawnictwo Naukowe PWN, Warszawa.

Cohen M.L. (1991), *Statistical Matching and Microsimulation Models,* [w:] *Improving Information for Social Policy Decisions, the Use of Microsimulation Modeling*, *Technical Papers*, vol. II, National Academy Press.

Di Zio M., D'Orazio M., Scanu M. (2006), *Statistical Matching. Theory and Practice*, John Wiley & Sons Ltd., England.

D'Orazio M. (2012), *Statistical Matching and Imputation of Survey Data with the Package StatMatch for the R Environment*, Italian National Institute of Statistics (Istat), Rome, Italy.

Gatnar E., Walesiak M. (red.) (2009), *Statystyczna analiza danych z wykorzystaniem programu R*, Wydawnictwo Naukowe PWN, Warszawa.

Kadane J.B. (1978), *Some Statistical Problems in Merging Data Files,* [w:] *Department of Treasury, Compendium of Tax Research*,. US Government Printing Office, Washington, DC.

Kiesl H., Raessler S. (2006), *How Valid Can Data Fusion Be?*, IAB Discussion Paper 15/2006, Nürnberg, Deutschland.

Moriarity C. (2009), *Statistical Properties of Statistical Matching. Data Fusion Algorithm*, VDM Verlag Dr. Mueller, Saarbrucken, Deutschland.

Moriarity C., Scheuren F. (2001), *Statistical matching: a paradigm for assessing the uncertainty in the procedure*, "Journal of Official Statistics" 17.

Moriarity C., Scheuren F. (2003), *A note on Rubin's statistical matching using file concatenation with adjusted weights and multiple imputation*, "Journal of Business and Economic Statistics" 21.

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

Rószkiewicz M. (2002), *Metody ilościowe w badaniach marketingowych*, Wydawnictwo Naukowe PWN, Warszawa.

Rubin D.B. (1986), *Statistical matching using file concatenation with adjusted weights and multiple imputations*, "Journal of Business and Economic Statistics".

Singh A.C., Mantel H., Kinack M., Rowe G. (1990), *On methods of statistical matching with and without auxiliary information*, Technical Report, DDMD-90-016, Statistics Canada.

Witkowski M., Klimanek T. (2006), *Prognozowanie gospodarcze i symulacje w przykładach i zadaniach*, Wydawnictwo Akademii Ekonomicznej w Poznaniu.

## JOINT CHARACTERISTICS' ESTIMATION OF VARIABLES NOT JOINTLY OBSERVED

**Summary:** Increasing demand for up-to-date statistical information is an increasing challenge for research institutions, both public and private. High costs of new studies result in relatively small frequency of their implementation. The use of statistical data integration methods allows to combine the available datasets in order to estimate the joint characteristics of variables not jointly observed. These methods may be the source of additional information for socio-economic research.

**Keywords:** statistical data integration, socio-economic research, information supply for economy.