EKONOMETRIA ECONOMETRICS 3(37) • 2012

ISSN 1507-3866

**Agnieszka Stanimir**
Uniwersytet Ekonomiczny we Wrocławiu

# RÓŻNE TECHNIKI PREZENTACJI POWIĄZAŃ KATEGORII ZMIENNYCH NIEMETRYCZNYCH

**Streszczenie:** Celem artykułu jest zaprezentowanie oraz sprawdzenie użyteczności stosowania jednocześnie kilku metod wielowymiarowej analizy statystycznej w badaniu powiązań zmiennych niemetrycznych. Przy wyborze prezentowanych metod głównym aspektem była możliwość znajdowania relacji między kategoriami zmiennych. W artykule omówiono algorytmy analizy korespondencji, wykresów mozaikowych, wykresów powiązań oraz drzew klasyfikacyjnych. W pracy wyróżniono aspekt aplikacyjny. Przeprowadzenie badań na rzeczywistych danych dotyczących wiedzy i umiejętności gimnazjalistów umożliwiło sprawdzenie użyteczności wspomnianych narzędzi. Dzięki zastosowaniu wybranych metod wyróżniono najbardziej i najmniej charakterystyczne poziomy wyniku egzaminacyjnego wśród gimnazjalistów z uwzględnieniem podziału ze względu na płeć i miejsce zdawania egzaminu.

**Słowa kluczowe:** zmienne niemetryczne, wykresy mozaikowe, wykresy powiązań, analiza korespondencji, drzewa klasyfikacyjne.

## 1. Wstęp

Celem artykułu jest zaprezentowanie oraz sprawdzenie użyteczności stosowania jednocześnie kilku metod wielowymiarowej analizy statystycznej w badaniu powiązań lub zależności zmiennych niemetrycznych. Analiza powiązań nie będzie dotyczyła tylko zmiennych, ale również będzie pogłębiona do analizy współwystąpień kategorii tych zmiennych. W pracy zdefiniowano również cel metodyczny, a mianowicie opis sposobu postępowania w rzadko opisywanych w polskiej literaturze metodach, tj. wykresach mozaikowych i wykresach powiązań. W tej części artykułu będą również zaprezentowane sposoby interpretacji wyników uzyskanych na postawie przeprowadzonych analiz.

W prezentowanej pracy ważny jest również aspekt aplikacyjny. Przeprowadzenie badań na rzeczywistych danych dotyczących wiedzy i umiejętności gimnazjalistów daje możliwość sprawdzenia użyteczności wspomnianych narzędzi.

Przystępując do przedstawienia proponowanych metod analitycznych, należy omówić, a raczej jedynie przypomnieć[1] definicje zmiennych niemetrycznych.

[1] Stwierdzenie to jest podyktowane dużą liczbą opracowań dotyczących skal pomiarowych, np. autorów takich jak: S. Mynarski [2000], E. Gatnar [1998], M. Walesiak [1996a; 1996b], K. Jajuga [1993] czy S.S. Stevens [1959].

Zmienne niemetryczne często bywają również nazywane zmiennymi jakościowymi i obejmują zmienne, których pomiaru dokonano na skalach nominalnej lub porządkowej. Zmienna zmierzona na najsłabszej skali pomiaru jest opisana dwiema lub kilkoma rozłącznymi kategoriami. Jeśli w badaniu wystąpi obiekt zmierzony na tego typu skali, to można go przyporządkować tylko do jednej z kategorii danej zmiennej. Kategoriom takiej zmiennej można przypisać symbole (można przypisać również liczby, ale w takiej sytuacji uznaje się je za symbole). Na podstawie pomiaru wykonanego na skali nominalnej nie można wykonać wielu działań matematycznych. Najczęściej dokonuje się zliczenia wystąpień kategorii, a następnie wyznaczane są częstości lub proporcje dla każdej kategorii. Miarą położenia jest modalna. Zmienne zmierzone na skali porządkowej to druga grupa zmiennych, które zalicza się do zmiennych niemetrycznych. Skala porządkowa powstaje przez wzbogacenie skali nominalnej o relację porządku (dla kategorii). Między kategoriami można ustalić relację mniejszości lub większości, jednak nie jest możliwe określenie wielkości różnic między kategoriami. Miarą położenia jest mediana. Skalą porządkową jest skala rangowa. Wzmacnianie skal o kolejne relacje powoduje powstawanie skal porządkowej i ilorazowej, dla których zakres stosowania metod matematycznych czy statystycznych jest bardzo szeroki.

Wyróżniając skale pomiarowe, należy pamiętać o zasadzie kumulatywności skali i sposobach transformacji skali. Kumulatywność dotyczy dostępnych na skalach relacji i działań. Działania możliwe do wykonania na skali nominalnej są dostępne na skalach porządkowej oraz silniejszych. Natomiast działania dostępne na skali porządkowej można wykonać na skalach metrycznych, ale nie można wykonać na skali nominalnej. Transformacja skali polega na przekształcaniu pomiaru dokonanego na skali silniejszej na pomiar wykonany na skali słabszej. Działanie to jest niezmiernie użyteczne w przypadku konieczności jednoczesnego zanalizowania kilku zmiennych, które są zmierzone na różnych skalach pomiarowych. Sprowadzenie wykonanych pomiarów do poziomu najniższego umożliwia łączną analizę wszystkich zmiennych.

W literaturze dotyczącej wielowymiarowej analizy statystycznej coraz częściej omawiane są zaawansowane metody analizy zmiennych nominalnych lub zmiennych mieszanych, takie jak np. analiza korespondencji czy skalowanie wielowymiarowe. Jednak w badaniach praktycznych, aplikacyjnych zainteresowanie badacza kształtowaniem się zmiennych nominalnych koncentruje się wokół rozpoznawania zależności między dwiema zmiennymi albo wokół jednowymiarowej analizy zmiennych. Powszechnie znane są współczynniki mierzące zależności między dwiema zmiennymi nominalnymi oparte na statystyce $\chi^2$, np. $\varphi$ Yule'a, $V$ Cramera, $Q$ Kendalla, lub zmiennymi porządkowymi: współczynnik korelacji rang Spearmana, korelacji rang Kendalla, konkordancji Kendalla. Po zastosowaniu tych miar zależności pozostaje jednak pytanie, czy występują jakieś relacje między kategoriami zmiennych oraz czy możliwe jest wprowadzenie do analizy większej liczby zmiennych? Prezentowane w tej pracy metody umożliwiają jednoczesną analizę dwóch lub kilku zmiennych.

W celu zobrazowania proponowanych graficznych metod prezentacji zmiennych niemetrycznych wykorzystano wyniki egzaminu gimnazjalnego przeprowadzonego w 2010 r. Dane, które wykorzystano, dotyczą całościowych wyników z egzaminu oraz płci i podziału terytorialnego województwa dolnośląskiego.

## 2. Źródło danych i zakres badania

Opis metodologii proponowanych narzędzi analizy danych niemetrycznych będzie o wiele bardziej czytelny, gdy będzie poparty przykładem aplikacyjnym. W tym celu wykorzystane będą rzeczywiste dane dotyczące wiedzy i umiejętności uczniów gimnazjum. Zgromadzone dane dotyczą województwa dolnośląskiego podlegającego Okręgowej Komisji Egzaminacyjnej we Wrocławiu. Badaniem objęto 15 111 dziewcząt i 15 058 chłopców, którzy przystąpili do egzaminu gimnazjalnego[2] w 2010 r. Dane, którymi dysponowano, umożliwiały podział populacji badanej ze względu na płeć, umiejscowienie szkoły, w której uczeń przystępował do egzaminu. Ponadto dysponowano szczegółową punktacją uzyskaną przez poszczególnych uczniów z egzaminu. Liczba punktów uzyskana przez ucznia z egzaminu jest zmienną, której pomiaru dokonano na skali porządkowej w przedziale wartości $\langle 0; 100 \rangle$.

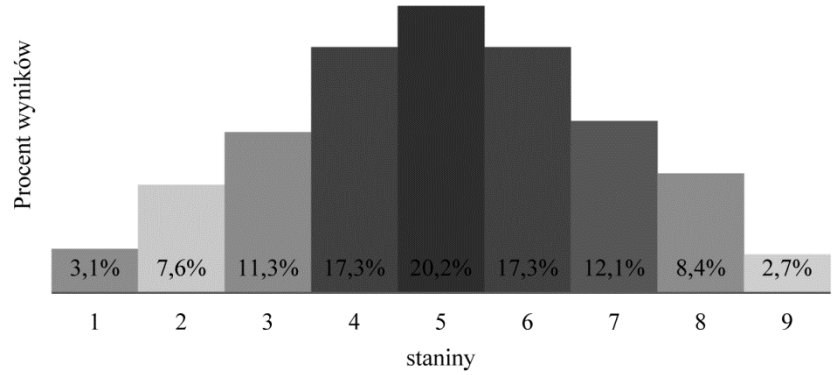
W celu przeprowadzenia analiz za pomocą wybranych metod oraz zwiększenia czytelności porównań efektów kształcenia w obydwu województwach konieczne było zmniejszenie liczby kategorii zmiennej „wynik egzaminu". W tym celu dokonano przyporządkowania uzyskanych przez uczniów wyników z egzaminu do klas skali staninowej. Skala ta jest stosowana przez Centralną Komisję Egzaminacyjną. Jest to skala znormalizowana nazywana również „standardową dziewiątką". Przedziały skali konstruowane są według następującego algorytmu, na podstawie uzyskanych przez uczniów punktów:

- ustalenie indywidualnej punktacji za egzamin dla uczniów;
- uporządkowanie niemalejące wyników;
- rozkład liczebności wyników;
- stworzenie szeregu kumulacyjnego;
- wyznaczenie procentów skumulowanych;
- podział wyników na 9 przedziałów po: 4, 7, 12, 17, 20, 17, 12, 7, 4 procent wyników;
- numerowanie przedziałów od 1 do 9 (1 – wynik najniższy, 2 – bardzo niski, 3 – niski, 4 – wynik niżej średniej, 5 – średni, 6 – wynik wyżej średniej, 7 – wysoki, 8 – bardzo wysoki, 9 – najwyższy)[3].

Poziomy skali staninowej określono na podstawie rozkładu normalnego surowych wyników uporządkowanych w kolejności nierosnącej (rys. 1).

[2] Dane udostępniła Okręgowa Komisja Egzaminacyjna we Wrocławiu.

[3] Zgodnie z materiałami opracowanymi w Centralnej Komisji Edukacyjnej w 2010 r. [*Osiągnięcia uczniów*… 2010].




**Rys. 1.** Procentowy rozkład wyników na skali staninowej

Źródło: opracowanie własne na podstawie [*Osiągnięcia uczniów*… 2010].

W prowadzonych analizach wykorzystywane będą następujące zmienne:

- wyniki uczniów Dolnego Śląska uzyskane z egzaminu gimnazjalnego w 2010 r. zmierzone na skali staninowej (Wynik1 – najniższy, Wynik2, Wynik3, Wynik4, Wynik5, Wynik6, Wynik7, Wynik8, Wynik9 – najwyższy);
- miejsce zdawania egzaminu: GM – gmina miejska, GMW – gmina miejsko-wiejska, GW – gmina wiejska, MJG – Jelenia Góra, ML – Legnica, MW – Wrocław;
- płeć: D – dziewczęta, CH – chłopcy.

W prowadzonych analizach uwzględniono również zmienną kombinowaną miejsce zdawania egzaminu/płeć. Dla tej zmiennej kategorie oznaczono następująco: GM_D (dziewczęta z gmin miejskich), GM_CH (chłopcy z gmin miejskich), GMW_D, GMW_CH, GW_D, GW_CH, MJG_D, MJ_CH, ML_D, ML_CH, MW_D, MW_CH.

## 3. Analiza korespondencji

Analiza korespondencji[4] należy do technik eksploracji danych polegających na redukcji wielowymiarowości. Efektem przeprowadzenia analizy korespondencji jest prezentacja współwystąpień kategorii dwóch lub kilku zmiennych nominalnych w przestrzeni dwu- lub trójwymiarowej. Ocenienie położenia na wykresie punktów obrazujących kategorie zmiennych jest podstawą do wnioskowania o wzajemnych interakcjach tych kategorii.

[4] Metodologię analizy korespondencji szczegółowo omawia w swoich pracach M. Greenacre [1984; 1993; 1994]. Publikacje polskojęzyczne przybliżające algorytm i przykłady aplikacyjne tej metody to m.in. monografia autorki [Stanimir 2004] oraz jej późniejsze opracowania.

Analiza korespondencji w klasycznej odmianie opiera się na analizie liczebności jednoczesnych wystąpień kategorii dwóch zmiennych zapisanych w tablicy kontyngencji. W celu dokonania graficznej prezentacji wyników analizy w przestrzeni o niskim wymiarze wykorzystuje się rozkład macierzy według wartości osobliwych. Dla dwóch zmiennych *A* oraz *B* stworzona będzie tablica kontyngencji:

$$\mathbf{N} = [n_{ij}] , \tag{1}$$

gdzie: $n_{ij}$ – liczebności jednoczesnych wystąpień *i*-tej kategorii zmiennej *A* ($i = 1, ..., r$) *j*-tej kategorii zmiennej *B* ($j = 1, ..., c$).

W analizie korespondencji przeprowadzany jest test niezależności $\chi^2$, by sprawdzić, czy między zmiennymi występuje zależność, a następnie na tej podstawie można określić, jak silna jest ta zależność. W celu przeprowadzenia testu niezależności $\chi^2$ niezbędne jest określenie hipotezy zerowej:

$$H_0: \; p_{ij} = p_{i\bullet} \cdot p_{\bullet j},$$

stwierdzającej, że cechy są niezależne, oraz hipotezy alternatywnej:

$$H_1: \; p_{ij} \neq p_{i\bullet} \cdot p_{\bullet j},$$

która wskazuje, że zmienne są zależne. Sprawdzianem hipotezy zerowej jest statystyka:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(n_{ij} - n \cdot p_{i\bullet} \cdot p_{\bullet j})^2}{n \cdot p_{i\bullet} \cdot p_{\bullet j}}, \tag{2}$$

gdzie:

- $p_{i\bullet} = \sum_{j=1}^{c} p_{ij} = \sum_{j=1}^{c} \frac{n_{ij}}{n} = \frac{n_{i\bullet}}{n}$ to częstości brzegowe wierszy ( $p_{ij} = \frac{n_{ij}}{n}$ to częstości zaobserwowane, $n_{i\bullet} = \sum_{j=1}^{c} n_{ij}$ liczebności brzegowe wierszy);
- $p_{\bullet j} = \sum_{i=1}^{r} p_{ij} = \sum_{i=1}^{r} \frac{n_{ij}}{n} = \frac{n_{\bullet j}}{n}$ to częstości brzegowe kolumn ( $n_{\bullet j} = \sum_{i=1}^{r} n_{ij}$ liczebności brzegowe kolumn).

Jeżeli liczebności oczekiwane różnią się znacznie od liczebności zaobserwowanych, to znaczy, że cechy są zależne. Empiryczna wartość statystyki $\chi^2$ jest porównywana z wartością krytyczną $\chi^2_\alpha$ wyznaczoną dla poziomu istotności $\alpha$ oraz $(r-1)(c-1)$ stopni swobody. Jeżeli $\chi^2 \leq \chi^2_\alpha$, to wskazując brak podstaw do od-

rzucenia hipotezy $H_0$, należy stwierdzić, że cechy są niezależne. Gdy $\chi^2 > \chi^2_\alpha$, hipoteza $H_0$ jest odrzucana na rzecz alternatywnej, a między cechami występuje zależność.

W następnym kroku konieczne jest wyznaczenie macierzy standaryzowanych różnic:

$$\mathbf{A} = \mathbf{D}_r^{-1/2}(\mathbf{P} - \mathbf{r}\mathbf{c}^{\mathrm{T}})\mathbf{D}_c^{-1/2}, \tag{3}$$

gdzie: $\mathbf{D}_r$ oznacza macierz częstości wierszowych; $\mathbf{D}_c$ oznacza macierz częstości kolumnowych; **P** to macierz częstości zaobserwowanych; **r** to wektor częstości brzegowych wierszy, **c** – wektor częstości brzegowych kolumn.

Macierz **A** jest poddawana dekompozycji według wartości osobliwych:

$$\mathbf{A} = \mathbf{U}\boldsymbol{\Gamma}\mathbf{V}^{\mathrm{T}}, \tag{4}$$

gdzie: $\boldsymbol{\Gamma}$ to macierz diagonalna ($k \times k$) niezerowych wartości osobliwych $\gamma_k$ ($k = 1, \ldots, K$) macierzy **A**, ułożonych w porządku nierosnącym, $K$ jest rzędem macierzy **A** oraz $K \le \min(r-1; c-1)$;

**U** jest macierzą (($r$ – 1) × $k$) lewych wektorów osobliwych;

**V** to macierz (($c$ – 1) × $k$) prawych wektorów osobliwych.

Na podstawie wartości osobliwych, lewych i prawych wektorów osobliwych wyznacza się współrzędne rzutowania kategorii zapisanych w wierszach (**F**) i kolumnach (**G**) tablicy kontyngencji:

$$\mathbf{F} = \mathbf{D}_{\mathrm{r}}^{-1/2}\mathbf{U}\boldsymbol{\Gamma}, \tag{5}$$

$$\mathbf{G} = \mathbf{D}_{\mathrm{c}}^{-1/2}\mathbf{V}\boldsymbol{\Gamma}. \tag{6}$$

Kolejne kolumny macierzy **F** (**G**) zawierają współrzędne kategorii z wierszy (kolumn) tablicy kontyngencji na kolejnych osiach głównych rzutowania. Do oceny jakości odwzorowania w analizie korespondencji najczęściej wykorzystuje się wskaźniki oparte na inercji całkowitej $\lambda$:

$$tr\mathbf{A}^{\mathrm{T}}\mathbf{A} = tr\mathbf{A}\mathbf{A}^{\mathrm{T}} = tr\boldsymbol{\Lambda} = \frac{\chi^2}{n} = \lambda = \sum_{k=1}^{K}\gamma_k^2. \tag{7}$$

Wartość inercji całkowitej jest sumą inercji głównych, czyli kwadratów wartości osobliwych wyznaczonych w trakcie dekompozycji macierzy **A**. Kwadraty wartości osobliwych to wartości własne. Pierwszą oś główną tworzy się, korzystając z najwyższej wartości własnej $\lambda_1$. Zatem jej udział w wyjaśnieniu inercji całkowitej jest największy.

Po dokonaniu graficznej prezentacji wyników analizy korespondencji oceniane jest położenie punktów obrazujących kategorie zmiennych:

- bliskie położenie kategorii dwóch różnych zmiennych oznacza, że ich współwystępowanie jest znaczące;
- bliskie ułożenie na wykresie kategorii należących do tej samej zmiennej wskazuje, że w tablicy kontyngencji można te kategorie sprowadzić do wspólnej kategorii, gdyż ich oceny są do siebie bardzo podobne;
- punkty położone blisko centrum rzutowania obrazują kategorie, które w najmniejszym stopniu przyczyniają sie do odrzucenia hipotezy o niezależności zmiennych.

Zgodnie z opisanym postępowaniem w analizie korespondencji dwóch zmiennych nominalnych można przeprowadzić badanie dla większej liczby zmiennych, które zapisano w wielowymiarowej tablicy kontyngencji. Tablica tego typu powstaje przez dodanie w wierszach bądź kolumnach tablicy kontyngencji (dla dwóch zmiennych) warstw z kategorii trzeciej zmiennej. Powstają wtedy zmienne kombinowane. Warstwy można wprowadzać również jednocześnie w wierszach i kolumnach, wtedy analiza jest wzbogacana o dwie zmienne. Dalsze postępowanie jest identyczne z opisanym algorytmem klasycznej analizy korespondencji.

W tabeli 1 zaprezentowano wielowymiarową tablicę kontyngencji zawierającą liczebności jednoczesnych wystąpień kolejnych poziomów wyniku egzaminacyjnego oraz miejsca zdawania egzaminu w województwie dolnośląskim w podziale na płeć.

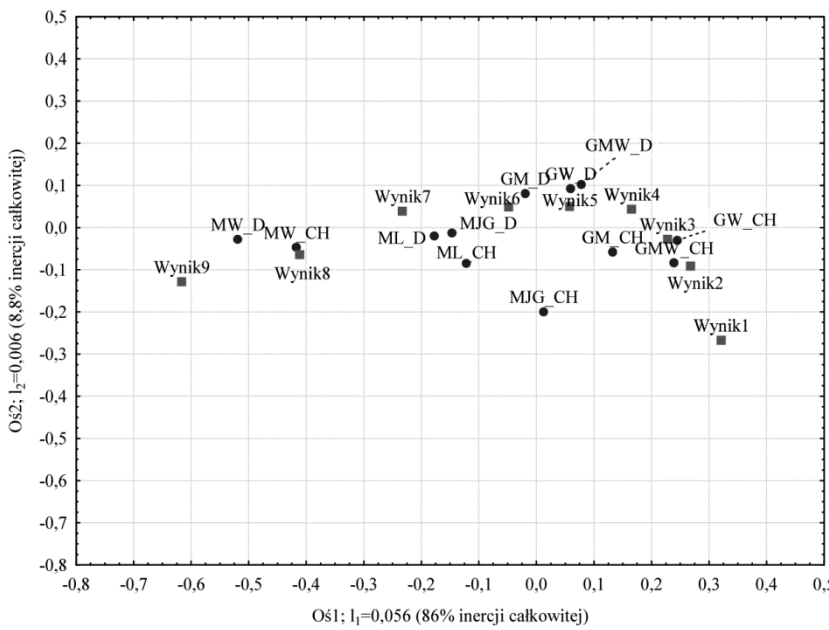
**Tabela 1.** Liczebności wyników egzaminu gimnazjalnego w staninach wraz z podziałem na płeć i miejsce zdawania egzaminu w 2010 r.

| | Wynik1 | Wynik2 | Wynik3 | Wynik4 | Wynik5 | Wynik6 | Wynik7 | Wynik8 | Wynik9 |
|---|---|---|---|---|---|---|---|---|---|
| GM_D | 124 | 263 | 473 | 724 | 901 | 831 | 562 | 355 | 150 |
| GM_CH | 261 | 397 | 556 | 794 | 881 | 789 | 492 | 299 | 127 |
| GMW_D | 111 | 304 | 491 | 791 | 880 | 769 | 492 | 265 | 115 |
| GMW_CH | 238 | 430 | 606 | 820 | 769 | 649 | 362 | 226 | 100 |
| GW_D | 81 | 197 | 316 | 570 | 540 | 476 | 383 | 178 | 82 |
| GW_CH | 134 | 242 | 395 | 546 | 542 | 393 | 225 | 131 | 62 |
| MJG_D | 22 | 16 | 51 | 65 | 77 | 90 | 76 | 44 | 23 |
| MJG_CH | 35 | 36 | 51 | 59 | 78 | 67 | 61 | 44 | 16 |
| ML_D | 20 | 42 | 45 | 86 | 110 | 99 | 98 | 53 | 39 |
| ML_CH | 19 | 40 | 68 | 87 | 102 | 89 | 66 | 65 | 37 |
| MW_D | 31 | 88 | 166 | 288 | 402 | 505 | 469 | 424 | 258 |
| MW_CH | 57 | 112 | 168 | 310 | 447 | 442 | 443 | 347 | 246 |

Źródło: opracowanie własne na podstawie danych Okręgowej Komisji Egzaminacyjnej we Wrocławiu.

Dla danych zapisanych w tab. 1 wyznaczono wartość statystyki $\chi^2 = 1988{,}6$. Dla stopni swobody wynoszących 88 oraz przy dowolnie wybranym poziomie istotności hipotezę o niezależności zmiennych należy odrzucić. Rzeczywista przestrzeń współ-

wystąpień zmiennych jest równa $K = min\{12 - 1; 9 - 1\} = 8$. Inercja całkowita $\lambda = 0{,}067$. Natomiast pierwsza inercja główna wynosi 0,057 i stanowi 86% inercji całkowitej, a druga inercja główna wynosi 0,006 i stanowi 8,8% inercji całkowitej. Zatem przy rzutowaniu punktów na przestrzeń dwuwymiarową zostanie zachowane 95% informacji o rzeczywistych powiązaniach kategorii analizowanych zmiennych. Rysunek 2 prezentuje wynik przeprowadzonej analizy.




**Rys. 2.** Współwystąpienia osiąganego wyniku z egzaminu gimnazjalnego na Dolnym Śląsku w 2010 r. oraz miejsca zdawania egzaminu z podziałem na płeć – wynik analizy korespondencji

Źródło: opracowanie własne z wykorzystaniem pakietu Statistica 10.

Ułożenie punktów na rys. 2 wskazuje, że wyniki najwyższe i bardzo wysokie są najbardziej charakterystyczne dla chłopców i dziewcząt z Wrocławia. Dziewczęta z Legnicy osiągają wysokie wyniki z egzaminu gimnazjalnego. Wyniki wyżej średniej, średnie i niżej średniej osiągają dziewczęta z gmin miejskich, miejsko-wiejskich oraz wiejskich. Niskie i bardzo niskie wyniki z egzaminu gimnazjalnego są charakterystyczne dla chłopców z gmin miejskich, miejsko-wiejskich i wiejskich. Punkty obrazujące dziewczęta z gmin miejsko-wiejskich i wiejskich są położone bardzo blisko siebie. Możliwe jest zatem ponowne przeprowadzenie analizy, ale dla tablicy, w której dokonano by sumowania liczebności tych dwóch kategorii dla poszczególnych poziomów wyników egzaminu. Dzięki takiemu postępowaniu udałoby się jeszcze zwiększyć jakość odwzorowania. Do odrzucenia hipotezy o niezależności zmiennych najbardziej przyczyniły się najniższa i najwyższa kategoria wyniku eg-

zaminacyjnego (najbardziej oddalone od centrum rzutowania), a najmniej kategorie: GM_D, Wynik6, Wynik5 (punkty te są położone bardzo blisko centrum rzutowania).

## 4. Wykresy mozaikowe i wykresy powiązań

W celu przeprowadzenia analizy zmiennych nominalnych za pomocą wykresów mozaikowych lub wykresów powiązań konieczne jest, tak jak w analizie korespondencji, stabelaryzowanie danych. Wskazane jest zatem zbudowanie tablicy kontyngencji oraz przeprowadzenie testu niezależności $\chi^2$.

Wykresy mozaikowe[5] składają się z płytek obrazujących wszystkie pola tablicy kontyngencji. Jako podstawę konstrukcji wykresów mozaikowych można uznać skumulowane wykresy kolumnowe. Dodatkowo każdy słupek kategorii jednej zmiennej takiego wykresu jest podzielony pionowo zależnie od liczebności kategorii drugiej zmiennej.

Pola płytek wykresu mozaikowego są proporcjonalne do liczebności poszczególnych komórek tablicy kontyngencji ($n_{ij}$). Szerokość płytek jest proporcjonalna do liczebności brzegowych każdej kolumny tabeli. Wysokość płytek jest proporcjonalna do warunkowych częstości wierszy:

$$n_{i|j} = \frac{n_{ij}}{n_{\bullet j}}. \tag{8}$$

W kolejnym kroku budowy wykresu mozaikowego wykonywane jest cieniowanie powstałych płytek. Wybór koloru i intensywności cieniowania zależy od standaryzowanych odchyleń od niezależności [Friendly 1994], które są obliczane jako:

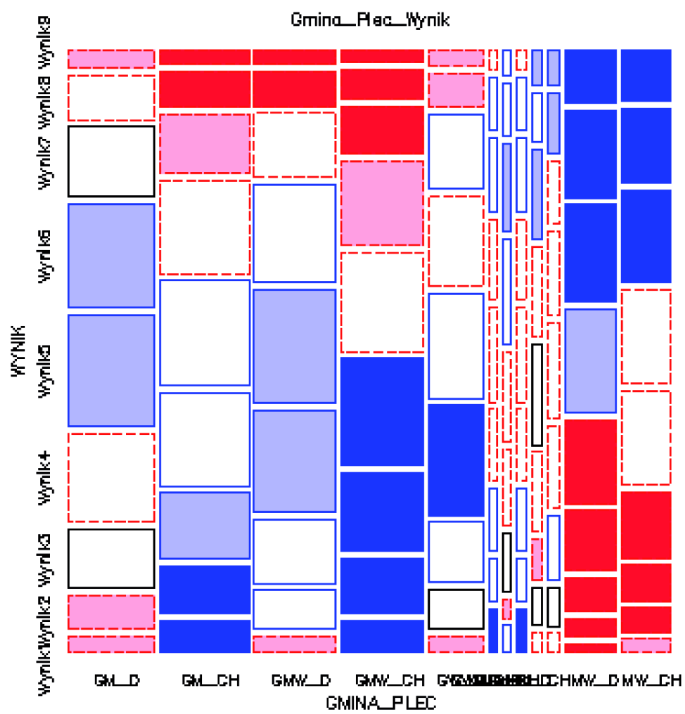
$$m_{ij} = \frac{(n_{ij} - n\hat{p}_{ij})}{\sqrt{n\hat{p}_{ij}}}. \tag{9}$$

Wartości $m_{ij}$ są udziałem komórki $ij$ w wartości statystyki $\chi^2$. Jeśli odchylenia wyznaczone wzorem (9) uzyskają w określonej komórce tablicy kontyngencji wartości dodatnie, to cieniowanie wykonuje się kolorem niebieskim. Gdy $m_{ij} < 0$, cieniowanie jest wykonywane kolorem czerwonym. Wartość bezwzględna odchyleń jest prezentowana intensywnością cieniowania. „Komórki o wartości bezwzględnej mniejszej niż 2 są puste, komórki, gdzie $|m_{ij}| \geq 2$, są wypełnione, a te, gdzie $|m_{ij}| \geq 4$, są wypełnione ciemniejszym wzorem" ([Friendly 1994], tłum. własne).

Na rysunku 3 zaprezentowano wykres mozaikowy stworzony dla danych zapisanych w tab. 1. Dwie ostatnie kolumny odpowiadają dziewczętom i chłopcom zdającym egzamin we Wrocławiu. Dla tych uczniów najbardziej charakterystyczne są wysokie, bardzo wysokie i najwyższe wyniki z egzaminu (płytki intensywnie niebie-

[5] Największy wkład w rozwój i upowszechnienie stosowania wykresów mozaikowych ma M. Friendly [1992a; 1992b; 1994 i in.]. W literaturze polskojęzycznej można znaleźć opis tej metody w pracach A. Stanimir, np. [2011].

skie), natomiast najmniej charakterystyczne są wyniki niżej średniej, niskie, bardzo niskie i najniższe (intensywnie czerwony kolor płytek). Kategorie wyniku egzaminu: Wynik1, Wynik2, Wynik3, Wynik4 (czyli niższe niż średnie) są charakterystyczne dla chłopców z gmin miejsko-wiejskich. Dla dziewcząt z gmin miejskich i miejsko-wiejskich oraz chłopców z gmin miejsko-wiejskich płytki określające bardzo wysokie i najwyższe wyniki mają intensywnie czerwony kolor, czyli takie wyniki z egzaminu nie są charakterystyczne dla tej grupy uczniów. Najmniej intensywne zabarwienie płytek lub brak zabarwienia można zaobserwować dla większości kategorii wyników egzaminu w grupie gimnazjalistów z Jeleniej Góry i Legnicy oraz chłopców z gmin wiejskich. Te kategorie w niewielkim stopniu wpływają na odrzucenie hipotezy o niezależności zmiennych.




**Rys. 3.** Wykres mozaikowy wyników egzaminu gimnazjalnego oraz płci i miejsca zdawania egzaminu

Źródło: opracowanie własne z wykorzystaniem programu Mosaic Displays.

Wykresy powiązań (asocjacji)[6] mają zbliżony algorytm do algorytmu wykresów mozaikowych. Ponownie punktem wyjścia jest zapisanie liczebności jednoczesnych wystąpień kategorii zmiennych w tablicy kontyngencji.

[6] W literaturze angielskojęzycznej ta metoda jest określana jako *association plot* [Friendly 1992a].

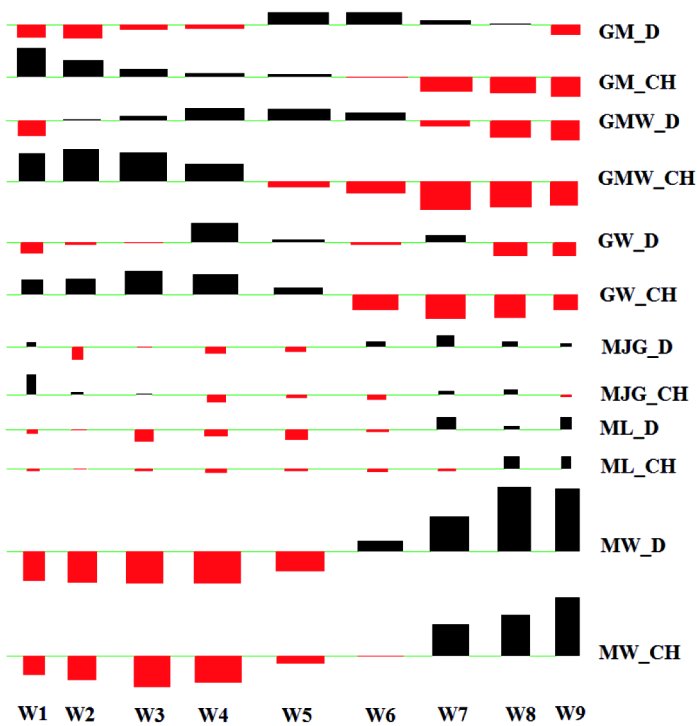
Obrazem poszczególnych komórek tablicy są prostokąty. W celu prawidłowego określenia wielkości prostokątów konieczne jest wyznaczenie oczekiwanych liczebności komórek tablicy kontyngencji:

$$\hat{n}_{ij} = n \cdot \hat{p}_{ij} = n \cdot p_{i\bullet} \cdot p_{\bullet j} = n \cdot \frac{n_{i\bullet}}{n} \cdot \frac{n_{\bullet j}}{n} = \frac{n_{i\bullet} \cdot n_{\bullet j}}{n}. \tag{10}$$

Pola prostokątów są proporcjonalne do różnic między obserwowanymi i oczekiwanymi liczebnościami komórek analizowanej tablicy. Wysokość prostokąta dla określonej komórki tabeli odpowiada wartości $m_{ij}$ obliczanej zgodnie ze wzorem (9), natomiast szerokość prostokąta jest równa $\sqrt{\hat{n}_{ij}}$.

Na wykresie dla każdego wiersza umieszczana jest linia oznaczająca niezależność kategorii: $m_{ij} = 0$. Jeśli dla rozpatrywanej komórki wiersza $n_{ij} > \hat{n}_{ij}$, to czarny prostokąt jest umieszczany nad linią. W przypadku przeciwnym czerwony prostokąt jest umieszczany pod linią.

Na rysunku 4 zamieszczono wykres powiązań dla danych zapisanych w tab. 1.




**Rys. 4.** Wykres powiązań wyników egzaminu gimnazjalnego oraz płci i miejsca zdawania egzaminu

Źródło: opracowanie własne.

Podobnie jak w przypadku wyników poprzednich analiz również wykres powiązań wskazuje, że wysokie wyniki egzaminu gimnazjalnego są charakterystyczne dla dziewcząt i chłopców uczących się we Wrocławiu. Dla tej grupy uczniów wyniki średnie i niższe nie są najbardziej charakterystyczne. Na rysunku 4 wiele czarnych prostokątów pojawiło się w górnym lewym rogu, wskazując na duże powiązanie osób uczących się w gminach miejskich, miejsko-wiejskich i wiejskich z niskimi wynikami egzaminu. Z tej grupy uczniów można wykluczyć dziewczęta z gmin miejskich, gdyż dla nich najbardziej charakterystyczne są wyniki przeciętne.

Rozpatrując wielkość i kolorystykę prostokątów wykresu powiązań, można również stwierdzić, że najmniejszy udział w odrzuceniu hipotezy o niezależności analizowanych zmiennych miały kategorie opisujące dziewczęta i chłopców z Legnicy i Jeleniej Góry.

## 5. Drzewa klasyfikacyjne

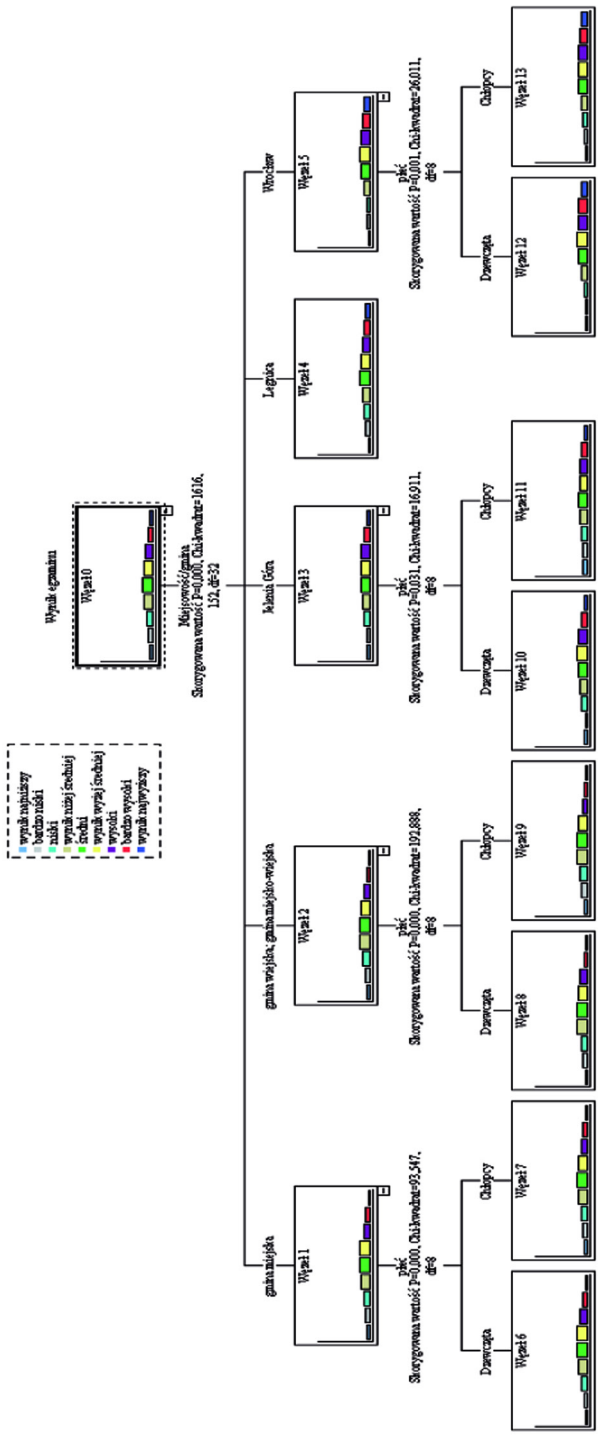
Drzewa klasyfikacyjne to metoda znacznie różniąca się od poprzednich trzech omówionych technik. Punktem wyjścia w prezentowanej metodzie jest wskazanie zmiennej zależnej i zmiennych niezależnych. Następnie przystępuje się do podziału zgromadzonych danych na jednorodne grupy ze względu na wartości przynależności do określonej kategorii zmiennej zależnej. Podział danych ma charakter rekurencyjny, czyli w każdym kroku wyjściowy zbiór jest dzielony na dwie lub więcej części za pomocą jednej ze zmiennych niezależnych [Gatnar 2004; Kurzydłowski 2002]. Na tym etapie analizy konieczne jest określenie skal pomiarowych, na których dokonano pomiaru zmiennych. Działanie to jest istotne, gdyż dalsze postępowanie zależy od rodzaju zmiennych. Jeśli zmienna zależna jest zmienną metryczną, to budowany jest model regresyjny, który reprezentuje drzewo regresyjne. W przypadku, gdy zmienna jest zmienną nominalną, mówimy o modelu klasyfikacyjnym (dyskryminacyjnym) z drzewem klasyfikacyjnym.

Ponieważ celem niniejszego artykułu jest zaprezentowanie wybranych metod analizy i prezentacji zmiennych niemetrycznych, to w centrum zainteresowania pozostaje możliwość przypisania obiektów do znanych kategorii nominalnej zmiennej zależnej.

Przedstawioną ideę drzew klasyfikacyjnych można zapisać następującym modelem [Gatnar 2004]:

$$y = \sum_{k=1}^{K} a_k I(\mathbf{x}_i \in R_k), \tag{11}$$

gdzie: – $y$ to zmienna zależna przyjmująca wartości $l = 1, \ldots, L$;
– $(\mathbf{x}_i, \ldots, \mathbf{x}_m)$ to klasyfikowane obiekty znajdujące się w przestrzeni zmiennych $\mathbf{X}^m$;




**Rys. 5.** Drzewo klasyfikacyjne wyników egzaminu gimnazjalnego oraz płci i miejsca zdawania egzaminu

Źródło: opracowanie własne z wykorzystaniem IBM SPSS Statistics 19.

– $R_k$ to rozłączne fragmenty przestrzeni $\mathbf{X}^m$; w tych częściach dzielonej przestrzeni znajdują się obiekty należące do tej samej klasy reprezentowanej przez zmienną zależną;
– $a_k$ to parametry modelu;

$$I(q)=I(q)=\begin{cases}1, \text{gdy } q \text{ jest prawdziwe,}\\ 0, \text{w przeciwnym przypadku.}\end{cases}$$

W modelu klasyfikacyjnym parametry są wyznaczane w następujący sposób:

$$a_k = \arg\max_l\{p(l|k)\}, \tag{12}$$

gdzie: $p(l|k)$ to prawdopodobieństwo, że obiekt z fragmentu $R_k$ należy do klasy $l$.

W swojej pracy Gatnar [1998] wskazuje, że w drzewie klasyfikacyjnym[7] „(…) wewnętrzne węzły opisują sposób dokonania podziału (w oparciu o wartości cech obiektów), a liście odpowiadają klasom, do których należą obiekty. Z kolei krawędzie drzewa reprezentują wartości cech, na podstawie których dokonano podziału".

Na rysunku 5 zaprezentowano wyniki analizy przeprowadzonej z wykorzystaniem drzew klasyfikacyjnych. Jako zmienną zależną wybrano wyniki egzaminu (na skali staninowej).

Z drzewa klasyfikacyjnego zaprezentowanego na rys. 5 wynika, że niezależnie od poziomu drzewa wyniki egzaminu gimnazjalnego rozkładają się w podobny sposób. Najwięcej jest wyników przeciętnych, a najmniej najwyższych i najniższych. Na uwagę zasługują jednak węzeł 12 i 13 odpowiadające, odpowiednio, wynikom osiąganym przez dziewczęta i chłopców we Wrocławiu – nastąpiło tu przesunięcie kategorii dominujących w stronę lepszych wyników w porównaniu z sytuacją w pozostałych węzłach. Procentowy udział wyników najlepszych jest w tych dwóch węzłach najwyższy: 26% dziewcząt z Wrocławia i 23% chłopców z Wrocławia uzyskało bardzo dobre wyniki z egzaminu (Wynik8 i Wynik9). Dla porównania w gminach miejsko-wiejskich i wiejskich 9,1% dziewcząt i 7,6% chłopców uzyskało wysokie wyniki z egzaminu gimnazjalnego. Zupełnie przeciwna sytuacja występuje, gdy analizie poddane zostaną wyniki najsłabsze z egzaminu (Wynik1 i Wynik2). Najwyższy procentowy udział wystąpienia tych wyników można zauważyć wśród chłopców z gmin miejskich, miejsko-wiejskich i wiejskich (14,3% oraz 15,2%). Natomiast najniższy udział wystąpień najsłabszych wyników odnotowano we Wrocławiu wśród dziewcząt.

## 6. Podsumowanie

Analiza korespondencji, mimo skomplikowanego algorytmu, umożliwia graficzną prezentację współwystąpień kategorii zmiennych zapisanych w nawet bardzo skomplikowanych tablicach kontyngencji. Dobierając odpowiedni stopień powiązania zmiennych, można uzyskać wysoką jakość prezentacji zmiennych (wyrażoną np.

[7] W cytowanym fragmencie określenie *cecha* jest równoznaczne ze *zmienną*.

stopniem wyjaśniania inercji całkowitej w wybranym wymiarze). Możliwe jest tu wskazanie zarówno kategorii, których współwystępowanie jest istotne (por. na rys. 1 punkty blisko położone na wykresie, np. Wynik9, Wynik8 są charakterystyczne dla dziewcząt i chłopców z Wrocławia (MW_CH, MW_D), jak i takich kategorii, których współwystąpienia nie można uznać za charakterystyczne (zob. na rys. 1 punkty bardzo oddalone od siebie, np. Wynik9, Wynik8 nie są charakterystyczne dla chłopców z gmin wiejskich i miejsko-wiejskich (GW_CH, GMW_CH).

Rozpatrując wykres powiązań, uzyskuje się podobną informację jak w przypadku analizy korespondencji. Wysokość prostokąta pokazuje, jak bardzo dwie kategorie wpływają na odrzucenie hipotezy o niezależności zmiennych w teście $\chi^2$. Kolor prostokąta odzwierciedla wartość różnicy między liczebnościami obserwowanymi a oczekiwanymi, pokazując tym samym, czy dwie kategorie różnych zmiennych są lub nie są ze sobą powiązane. Prezentacja powiązań dla wybranego przykładu określa, że największy wpływ na odrzucenie hipotezy o niezależności mają kategorie W7-W8 zaobserwowane wśród dziewcząt i chłopców z Wrocławia, najmniejszy wpływ na odrzucenie tej hipotezy mają wszystkie kategorie wyników egzaminacyjnych dziewcząt i chłopców z Jeleniej Góry i Legnicy oraz prawie wszystkie wyniki uczniów z gmin miejskich. Dla chłopców i dziewcząt z Wrocławia występuje silne powiązanie z wysokimi wynikami egzaminacyjnymi. Natomiast wyniki niskie są najmniej charakterystyczne dla tej grupy uczniów.

Podobną prezentację do wykresu powiązań można uzyskać, stosując wykresy mozaikowe. Łatwo zauważyć, które pary kategorii różnych zmiennych mają największy i najmniejszy wkład w odrzucenie hipotezy o niezależności. Ponadto właściwe kolorowanie z wykorzystaniem intensywności zabarwienia płytek również umożliwia sprawne ocenienie powiązań między kategoriami zmiennych.

Dużą niedogodnością stosowania wykresów powiązań i mozaikowych jest zacieśnianie prezentacji w przypadku dużej liczby kategorii zmiennych, np. w sytuacji, gdy tworzona jest wielowymiarowa tablica kontyngencji. Zatem te dwa typy wykresów można potraktować jako alternatywę dla analizy korespondencji w przypadku badania współwystąpień niezbyt dużej liczby kategorii. Ponadto analiza korespondencji ma jeszcze jedną zaletę, a mianowicie umożliwia wskazanie kategorii należących do tej samej zmiennej, których wystąpienia są oceniane w podobny sposób i można dokonać ich kumulacji w jedną kategorię.

Jednoczesne zastosowanie analizy korespondencji wraz z wykresami powiązań lub wykresami mozaikowymi ułatwia percepcję prezentowanych wyników, szczególnie gdy w grę wchodzi analiza zmiennych z wieloma kategoriami i poszukiwanie powiązań kategorii na wykresach jest trudne.

Skalowanie wielowymiarowe pozwala uzupełnić analizę korespondencji, wykresy mozaikowe oraz wykresy powiązań np. o liczebności wystąpień w kategoriach zmiennej zależnej kategorii wyróżnionych w końcowych węzłach drzewa.

Przedstawiony przykład aplikacyjny wykorzystany w badaniu poziomu osiąganych wyników egzaminacyjnych przez gimnazjalistów pozwolił zobrazować sposób interpretacji wyników uzyskanych po zastosowaniu opisanych metod.

## Literatura


Friendly M., *Graphical Methods for Categorical Data*, W Proceedings of the SAS User's Group International Conferences, 17, 1992a.

Friendly M., *Mosaic Display for Loglinear Models*. *American Statistical Association*, *Proceedings of the Graphics Section*, 1992b.

Friendly M., *Mosaic display for multi-way contingency tables*. *American Statistical Association*, "Journal of the American Statistical Association – Theory and Methods" 1994, vol. 89, nr 425.

Gatnar E., *Drzewa klasyfikacyjne*, [w:] *Metody statystycznej analizy wielowymiarowej w badaniach marketingowych*, red. E. Gatnar, M. Walesiak, Wydawnictwo Akademii Ekonomicznej, Wrocław 2004.

Gatnar E., *Symboliczne metody klasyfikacji danych*, Wydawnictwo Naukowe PWN, Warszawa 1998.

Greenacre M., *Correspondence Analysis in Practice,* Academic Press, London 1993.

Greenacre M., *Multiple and Joint Correspondence Analysis*, [w:] *Correspondence Analysis in Social Sciences. Recent Developments and Applications,* red. M. Greenacre, J. Blasius, San Diego 1994.

Greenacre M., *Theory and Applications of Correspondence Analysis*, Academic Press, London 1984.

Jajuga K., *Statystyczna analiza wielowymiarowa*, Wydawnictwo Naukowe PWN, Warszawa 1993.

Kurzydłowski A., *Ocena jakości podziału w wybranych algorytmach drzew klasyfikacyjnych*, [w:] Ekonometria 9, Prace Naukowe Akademii Ekonomicznej we Wrocławiu nr 935, Wydawnictwo Akademii Ekonomicznej, Wrocław 2002.

Mynarski S., *Praktyczne metody analizy danych rynkowych i marketingowych*, Zakamycze, Kantor Wydawniczy, 2000.

*Osiągnięcia uczniów kończących gimnazjum w roku 2010. Sprawozdanie z egzaminu gimnazjalnego 2010,* Centralna Komisja Egzaminacyjna we współpracy z okręgowymi komisjami egzaminacyjnymi w Gdańsku, Jaworznie, Krakowie, Łodzi, Łomży, Poznaniu, Warszawie i Wrocławiu, Warszawa 2010.

Stanimir A., *Analiza korespondencji jako narzędzie do badania zjawisk ekonomicznych*, Wydawnictwo UE we Wrocławiu, Wrocław 2004.

Stanimir A., *Wizualizacja zmiennych nominalnych – analiza korespondencji a wykresy mozaikowe*, [w:] Ekonometria 34, Wydawnictwo UE we Wrocławiu, Wrocław 2011.

Stevens S.S., *Measurement, Psychophysics and Utility*, [w:] *Measurement. Definitions and Theories*, red. C.W. Churchman, P. Ratoosh, John Willey & Sons, Inc., New York 1959.

Walesiak M., *Dopuszczalne działania na liczbach w badaniach marketingowych z punktu widzenia skal pomiarowych*, [w:] Informatyka i ekonometria 1, Wydawnictwo AE we Wrocławiu, Wrocław 1996a.

Walesiak M., *Metody analizy danych marketingowych*, PWN, Warszawa 1996b.

## DIFFERENT TECHNIQUES OF GRAPHICAL PRESENTATION OF NON-METRIC DATA CATEGORIES

**Summary:** The aim of this article is to present and examine the usefulness of several methods of multivariate statistical analysis in the study of dependencies and relationship of non-metric variables. The main aspect of selection of the presented methods was the possibility of finding connections between categories of variables. The article discusses algorithms of correspondence analysis, mosaic displays, association plots and classification trees. In the present paper the application aspect is also very important. Conducting researches on real data of knowledge and skills of gymnasium students made it possible to verify the usefulness of those methods. Thanks to the use of selected methods it was possible to identify the most and least characteristic levels as a result of the examination among pupils by sex and place of taking the exam.

**Keywords:** non-metric data, mosaics displays, association plots, correspondence analysis, classification trees.