EKONOMETRIA ECONOMETRICS 34 • 2011

ISSN 1507-3866

Agnieszka Stanimir
Uniwersytet Ekonomiczny we Wrocławiu

WIZUALIZACJA ZMIENNYCH NOMINALNYCH – ANALIZA KORESPONDENCJI A WYKRESY MOZAIKOWE

Streszczenie: Celem artykułu jest zaprezentowanie graficznych metod analizy wykorzystywanych w badaniu zmiennych nominalnych. Analiza tego typu zmiennych wykonywana pojedynczo dla każdej z nich jest najrzadziej pożądana przez badacza. Rozpoznanie zależności i jej siły dla pary zmiennych nominalnych pozwala już szerzej wnioskować o tych najsłabszych zmiennych. Wszystkie zaprezentowane techniki i metody umożliwiają rozpatrywanie zmiennych nominalnych również przez pryzmat kategorii i interakcji między kategoriami pochodzącymi z różnych cech. W artykule zaprezentowano analizę profili, skumulowane wykresy słupkowe, wykresy mozaikowe oraz analizę korespondencji.

Słowa kluczowe: analiza korespondencji, wykresy mozaikowe, analiza profili, skumulowane wykresy słupkowe.

1. Wstęp

Określenie „zmienne nominalne" dotyczy zmiennych zmierzonych na najsłabszej skali pomiaru. Zgodnie z opisami pomiarów prowadzonych na skali nominalnej (zob. [Walesiak 1996; Stevens 1959]) między kategoriami zmiennej nominalnej można doszukiwać się jedynie różnic lub równości. Dopuszczalne jest zliczanie wystąpień poszczególnych kategorii i obliczanie na tej podstawie częstości, proporcji. Miarą położenia dla tak przeprowadzonego pomiaru jest modalna. Charakterystyka skali nominalnej dopuszcza zastosowanie wybranych testów i metod statystycznych. Najczęściej analizę zmiennych nominalnych przeprowadza się oddzielnie dla każdej z nich lub dla par zmiennych. Po zbudowaniu tablicy kontyngencji, czyli tablicy jednoczesnych wystąpień dwóch kategorii opisujących dwie różne zmienne, można przeprowadzić test niezależności χ^2, a na podstawie wyznaczonej wartości statystyki χ^2 określić siłę zależności dwóch zmiennych nominalnych. Współczynniki badające zależność zmiennych na skali nominalnej opisują tylko zmienne, nie wskazując w żaden sposób na interakcje między ich kategoriami. Badając współwystąpienia kategorii analizowanych zmiennych, należy skorzystać z analizy korespondencji. Wyniki analizy są prezentowane graficznie, a odległość punktów obrazujących poszczególne kategorie wskazuje na ich częste bądź sporadyczne współwystępowanie.

Analiza korespondencji nie ogranicza badania tylko do dwóch zmiennych. Możliwe jest zastosowanie jej odmian do rozpoznawania współwystąpień kategorii wielu zmiennych nominalnych. Parę zmiennych nominalnych można również analizować, wykorzystując wykresy mozaikowe. Metoda ta, tak jak analiza korespondencji, bazuje również na tablicy kontyngencji, ale jej efektem nie jest rozrzut punktów w układzie współrzędnych. Wykresy mozaikowe są to wykresy słupkowe o szerokości i wysokości zależnej od częstości zapisanych w tablicy kontyngencji. Wykresy mozaikowe można tworzyć również dla wielu zmiennych nominalnych.

Celem artykułu jest zaprezentowanie metod graficznych umożliwiających zobrazowanie interakcji między kategoriami zmiennych nominalnych. Przedstawiono wykresy profili, wykresy słupkowe, analizę korespondencji i rzadko wykorzystywane w polskojęzycznej literaturze wykresy mozaikowe.

W celu zobrazowania prowadzenia analizy i wnioskowania na podstawie wykresów mozaikowych i analizy korespondencji posłużono się danymi zaczerpniętymi z rocznika statystycznego *Aktywności ekonomicznej ludności Polski. II kwartał 2011*. Wybrano dane dotyczące:

- stanu aktywności zawodowej: pracujący w pełnym wymiarze czasu pracy (A1), pracujący w niepełnym wymiarze czasu pracy (A2), bezrobotni (A3), bierni zawodowo (A4);
- poziomu wykształcenia: wyższe (W1), policealne (W2), średnie zawodowe (W3), średnie ogólnokształcące (W4), zasadnicze zawodowe (W5), gimnazjalne, podstawowe i niepełne podstawowe (W6).

Aktywność ekonomiczną Polaków w II kwartale 2011 r. prezentuje tab. 1.

Tabela 1. Aktywność ekonomiczna Polaków (II kwartał 2011 r., tys. os.)

		Aktywność ekonomiczna			
		Aktywni zawodowo			
		A1	**A2**	**A3**	**A4**
Wykształcenie	**W1**	4230	288	247	1149
	W2	537	43	64	298
	W3	3741	260	345	2109
	W4	1233	153	209	1708
	W5	4189	317	556	2874
	W6	949	223	269	5863

Źródło: opracowanie własne na podstawie *Aktywności ekonomicznej ludności Polski. II kwartał 2011*.

2. Test niezależności χ^2

Zagadnienia omówione w tej części pracy są powszechnie znane (por. np. [Stanimir 2006; Rószkiewicz 2002; Walesiak 1996; Jobson 1992; Rönz, Förster 1992; Greń 1984; Blalock 1975; Yule, Kendall 1966; Cramér 1958]), jednak ich przypomnienie jest istotne dla wyjaśnienia konstrukcji wykresów mozaikowych. Zmienne nominal-

ne są najczęściej analizowane za pomocą testu niezależności χ^2. Test pozwala stwierdzić, czy dwie zmienne nominalne (A oraz B) są niezależne, czy też hipotezę o niezależności można odrzucić. Jeśli przyjęte będą oznaczenia, że zmienna A ma r kategorii, a zmienna B – c kategorii, to tablica kontyngencji będzie wypełniona elementami n_{ij}, czyli liczebnościami jednoczesnych wystąpień i-tej kategorii zmiennej A ($i = 1, ..., r$) oraz j-tej kategorii zmiennej B ($i = 1, ..., c$). W celu przeprowadzenia testu niezależności χ^2 należy kolejno wyznaczyć:

– liczebności brzegowe wierszy:

$$n_{i\bullet} = \sum_{j=1}^{c} n_{ij}, \tag{1}$$

– liczebności brzegowe kolumn:

$$n_{\bullet j} = \sum_{i=1}^{r} n_{ij}, \tag{2}$$

– częstości zaobserwowane:

$$p_{ij} = \frac{n_{ij}}{n}, \tag{3}$$

– częstości brzegowe wierszy:

$$p_{i\bullet} = \sum_{j=1}^{c} p_{ij} = \sum_{j=1}^{c} \frac{n_{ij}}{n} = \frac{n_{i\bullet}}{n}, \tag{4}$$

– częstości brzegowe kolumn:

$$p_{\bullet j} = \sum_{i=1}^{r} p_{ij} = \sum_{i=1}^{r} \frac{n_{ij}}{n} = \frac{n_{\bullet j}}{n}, \tag{5}$$

– częstości oczekiwane:

$$\hat{p}_{ij} = p_{i\bullet} \cdot p_{\bullet j}, \tag{6}$$

– liczebności oczekiwane:

$$n \cdot \hat{p}_{ij} = n \cdot p_{i\bullet} \cdot p_{\bullet j} = n \cdot \frac{n_{i\bullet}}{n} \cdot \frac{n_{\bullet j}}{n} = \frac{n_{i\bullet} \cdot n_{\bullet j}}{n}. \tag{7}$$

Liczebności brzegowe to liczba wystąpień poszczególnych kategorii rozpatrywanej zmiennej. Częstości zaobserwowane p_{ij} są procentowym udziałem jednoczesnego wystąpienia w badaniu i-tej kategorii zmiennej A oraz j-tej kategorii zmiennej B. Wartości te są elementami macierzy **P**. Częstości brzegowe $p_{i\bullet}$ oraz $p_{\bullet j}$ wskazują procentowy udział wystąpień kategorii określonej cechy w liczbie wszystkich badanych jednostek. Częstości brzegowe wierszy zapisuje się w wektorze **r**, a kolumn w wektorze **c**.

W celu zbadania niezależności cech weryfikuje się hipotezę zerową:

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

stwierdzającą, że cechy są niezależne, wobec hipotezy alternatywnej:

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

która wskazuje, że zmienne są zależne. Sprawdzianem hipotezy zerowej jest statystyka $\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(n_{ij} - n \cdot p_{i\bullet} \cdot p_{\bullet j})^2}{n \cdot p_{i\bullet} \cdot p_{\bullet j}}$.

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

Dla zaprezentowanych w tab. 1 danych statystyka $\chi^2 = 6352,3$ a $\chi^2_{\alpha=0,01} = 30,6$. Na tej podstawie hipotezę o niezależności cech należy odrzucić, czyli aktywność ekonomiczna zależy od poziomu wykształcenia.

3. Analiza profili lub skumulowany wykres słupkowy jako graficzna prezentacja zmiennych nominalnych

Na podstawie tablicy kontyngencji można przeprowadzić analizę kategorii zmiennych w sposób graficzny.

Jedną z tego typu analiz jest prezentacja profili wierszy i kolumn. W celu wyznaczenia profili należy skorzystać ze wzorów (1)-(6). Profile wierszowe wyznacza się jako:

$$\mathbf{R} = \left[\frac{n_{ij}}{n_{i\bullet}}\right] = \left[\frac{p_{ij}}{p_{i\bullet}}\right] = \mathbf{D}_r^{-1}\mathbf{P}, \tag{8}$$

a profile kolumnowe jako:

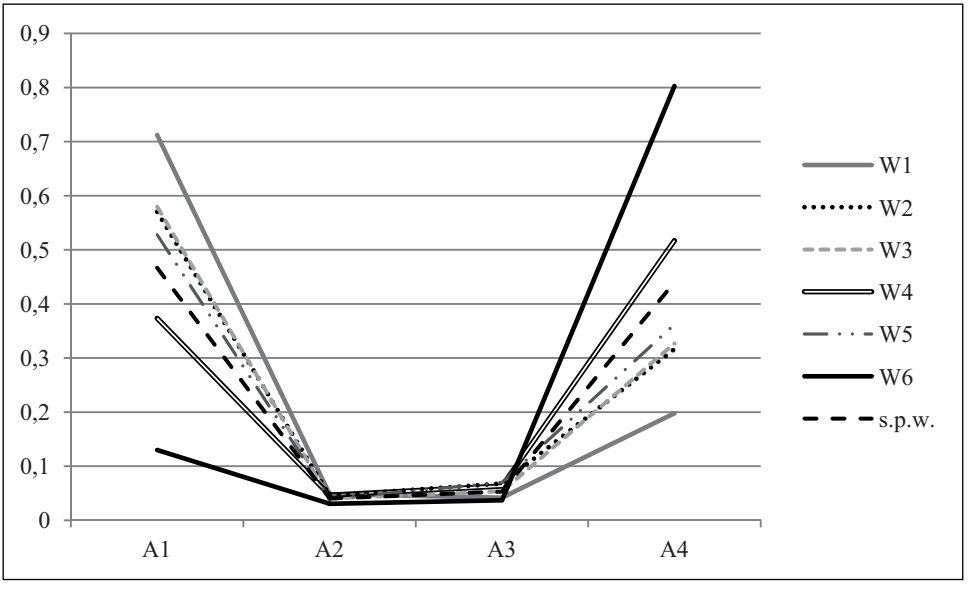
$$\mathbf{C} = \left[\frac{n_{ij}}{n_{\bullet j}}\right] = \left[\frac{p_{ij}}{p_{\bullet j}}\right] = \mathbf{D}_c^{-1}\mathbf{P}. \tag{9}$$

We wzorze (8) $\mathbf{D}_r$ oznacza macierz częstości wierszowych (analogicznie we wzorze (9)). Macierze $(\mathbf{R} - \mathbf{1}_r\mathbf{c}^{\mathbf{T}})$ oraz $(\mathbf{C} - \mathbf{r}\mathbf{1}_c^{\mathbf{T}})$ są miarą stopnia odchylenia zmiennych

od niezależności. Częstości brzegowe wierszy i kolumn, które są składowymi wektorów **r** oraz **c**, stanowią jednocześnie średnie profile wierszowe i kolumnowe.

Porównując profil wybranej kategorii z odpowiednim profilem średnim, można stwierdzić, czy dana kategoria ma duży wpływ na odrzucenie hipotezy o niezależności zmiennych. Im profil kategorii jest bardziej zbliżony do profilu średniego, tym większy wpływ kategorii zmiennej na stwierdzenie o niezależności zmiennych. Porównanie kategorii między sobą ze względu na ich profile pozwala stwierdzić, czy kategorie te są do siebie podobne[1].

Na rysunkach 1 oraz 2 przedstawiono profile wierszowe i kolumnowe zmiennych dotyczących stanu aktywności ekonomicznej. Na tych rysunkach przyjęto skrócone oznaczenia dla średnich profili wierszowych – s.p.w. oraz kolumnowych – s.p.k.



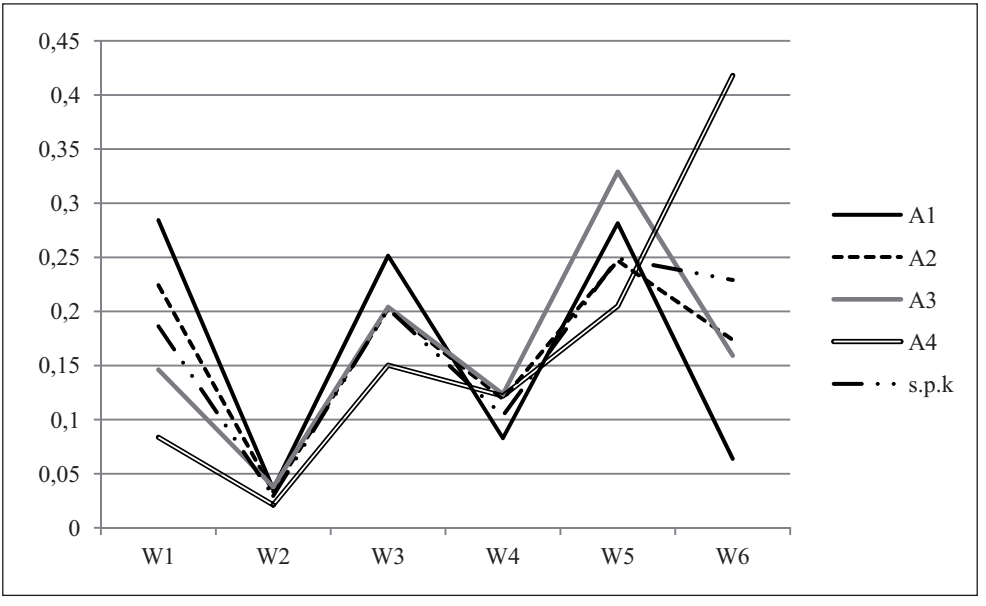
Rys. 1. Profile aktywności ekonomicznej Polaków (profile wierszowe)

Źródło: opracowanie własne na podstawie danych GUS.

Profile przedstawione na rys. 1 wskazują, że największe rozbieżności między aktywnością zawodową Polaków występują w grupach osób pracujących w pełnym wymiarze czasu pracy i osób biernych zawodowo. W tych grupach do profilu średniego najbardziej zbliżone są osoby z wykształceniem średnim ogólnokształcącym (W4) i zasadniczym zawodowym (W5). Na rysunku 2 widoczne jest podobieństwo

[1] Taka analiza jest istotna ze względu na redukcję liczby kategorii zmiennych, a tym samym obniżenie rzeczywistego wymiaru powiązań. Jeśli kategorie mają zbliżone profile, to ich liczebności można połączyć i stworzyć jedną kategorię.

profilu osób pracujących w niepełnym wymiarze do profilu średniego. Największe odchylenie od profilu średniego wykazuje profil osób biernych zawodowo. To ta cecha ma bardzo duży wpływ na odrzucenie hipotezy o niezależności analizowanych cech.



Rys. 2. Profile poziomu wykształcenia Polaków (profile kolumnowe)

Źródło: opracowanie własne na podstawie danych GUS.

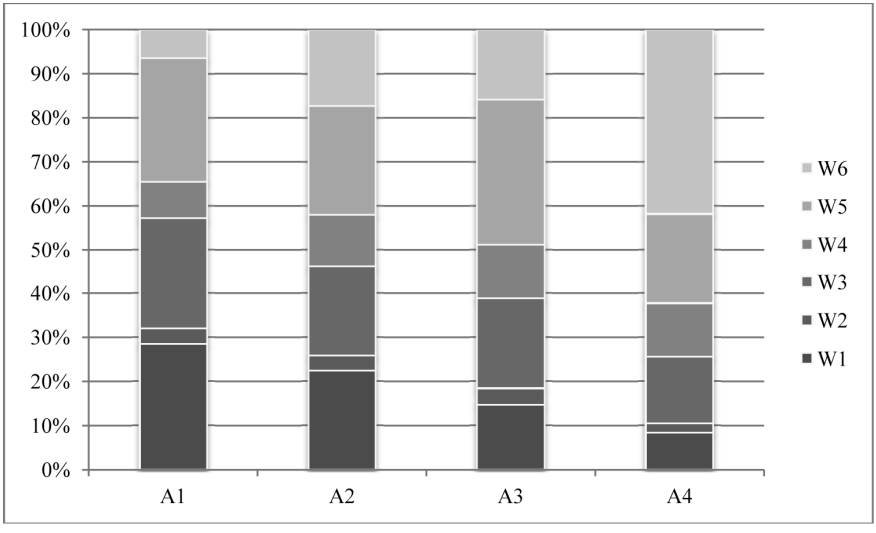
Na podstawie tablicy kontyngencji przeprowadza się również inną graficzną prezentację kategorii zmiennych. Dane z tablicy kontyngencji można przedstawić na wykresie słupkowym, gdzie liczebności kategorii są przeliczone jako procentowe udziały w odpowiedniej liczebności brzegowej.

„Wykres słupkowy jest … używany, gdy chcemy przedstawić wartości zmiennych, które to zmienne są… jakościowe lub dyskretne" (zob. [Wallgren i in. 1996, s. 28]). Na wykresie słupkowym porównywane są liczebności bezwzględne. Jeśli natomiast wykres słupkowy ulegnie modyfikacji do postaci skumulowanego wykresu słupkowego, to można zaprezentować proporcje liczebności poszczególnych kategorii jednej zmiennej w określonej kategorii drugiej zmiennej. W przypadku danych pochodzących z tablicy kontyngencji możliwe jest zbudowanie dwóch skumulowanych wykresów słupkowych.

Dla rozpatrywanego przykładu zbudowano skumulowane wykresy słupkowe aktywności ekonomicznej (rys. 3) oraz wykształcenia (rys. 4).

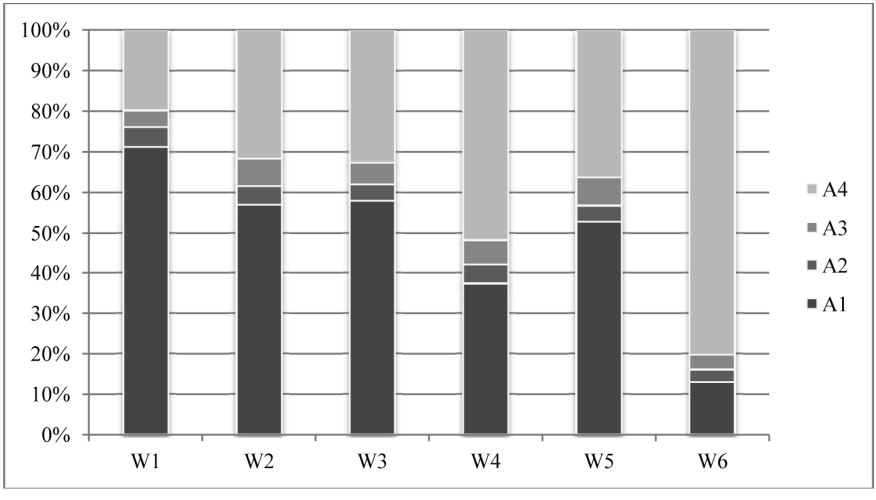
Proporcja udziału osób z wykształceniem wyższym (W1), średnim zawodowym (W3) oraz zasadniczym zawodowym (W5) w grupie osób pracujących w pełnym

wymiarze czasu pracy (A1) jest zbliżona i jednocześnie wyższa niż pozostałych trzech grup wykształcenia (W2, W4, W6). W grupie osób biernych zawodowo znaczny udział mają osoby z najniższym poziomem wykształcenia (W6). Wynika to z faktu, że badanie obejmuje osoby od 15 roku życia, a zatem również gimnazjalistów, którzy podlegają obowiązkowi kształcenia.



Rys. 3. Skumulowany wykres słupkowy aktywności ekonomicznej Polaków

Źródło: opracowanie własne na podstawie danych GUS.



Rys. 4. Skumulowany wykres słupkowy poziomu wykształcenia Polaków

Źródło: opracowanie własne na podstawie danych GUS.

W trzech pierwszych grupach osób o najwyższym poziomie wykształcenia (W1--W3) można zaobserwować najwyższy udział osób zatrudnionych w pełnym wymiarze czasu pracy (A1). Wśród osób z najniższym wykształceniem (W6) największy udział mają bierni zawodowo (A4).

4. Klasyczna analiza korespondencji[2]

Klasyczna analiza korespondencji jest metodą stosowaną w analizie dwóch zmiennych nominalnych, z których każda opisana jest kilkoma kategoriami. Liczebności jednoczesnych wystąpień zapisuje się w tablicy kontyngencji. Rzeczywisty wymiar współwystąpień kategorii dwóch zmiennych nominalnych jest równy $\min\{r - 1; c - 1\}$. Jeśli zatem analizowane zmienne mają po więcej niż cztery kategorie, to graficzna prezentacja ich rzeczywistych powiązań jest niemożliwa. Analiza korespondencji daje możliwość prezentacji współwystąpień kategorii zmiennych w przestrzeni o niskim wymiarze z zachowaniem jak najpełniejszej informacji o zmiennych. Działanie to jest możliwe dzięki zastosowaniu rozkładu macierzy według wartości osobliwych, a na tej podstawie wyznaczeniu współrzędnych rzutowania kategorii zmiennych. Macierzą, która w analizie korespondencji podlega dekompozycji, jest macierz standaryzowanych różnic **A**, czyli ważonych odchyleń profili od średnich profili wierszowego i kolumnowego (centrum wierszowego i kolumnowego):

$$\mathbf{A} = \mathbf{D}_r^{-1/2}(\mathbf{P} - \mathbf{r}\mathbf{c}^{\mathrm{T}})\mathbf{D}_c^{-1/2} = \mathbf{U}\Gamma V^{\mathrm{T}}, \tag{10}$$

gdzie: Γ – to macierz diagonalna ($k \times k$) niezerowych wartości osobliwych γ_k ($k = 1, \ldots, K$) macierzy **A**, ułożonych w porządku nierosnącym, K jest rzędem macierzy **A** oraz $K \leq \min(n;m)$; **U** – jest macierzą ($n \times k$) lewych wektorów osobliwych; **V** – to macierz ($m \times k$) prawych wektorów osobliwych.

Wartości osobliwe, lewe i prawe wektory osobliwe są wyznaczane w celu obliczenia współrzędnych rzutowania kategorii zapisanych w wierszach (**F**) i kolumnach (**G**) tablicy kontyngencji:

$$\mathbf{F} = \mathbf{D}_{\mathrm{r}}^{-1/2}\mathbf{U}\Gamma, \tag{11}$$

$$\mathbf{G} = \mathbf{D}_{\mathrm{c}}^{-1/2}\mathbf{V}\Gamma. \tag{12}$$

Wektory macierzy **U** (**V**) są nazywane głównymi osiami rzutowania kategorii zapisanych w kolumnach (wierszach). Kolejne kolumny macierzy **F** (**G**) zawierają współrzędne kategorii z wierszy (kolumn) tablicy kontyngencji na kolejnych osiach głównych. Do oceny jakości odwzorowania w analizie korespondencji najczęściej wykorzystuje się wskaźniki oparte na inercji całkowitej λ:

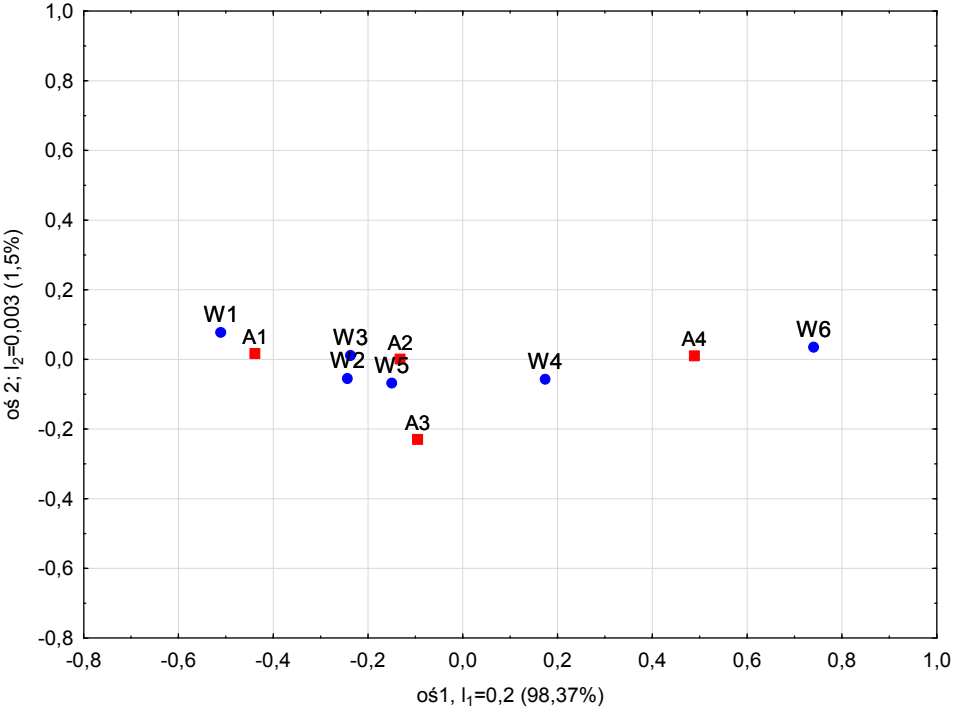
[2] W tej części pracy nie omówiono wszystkich aspektów analizy korespondencji, gdyż nie jest to konieczne do dokonania porównań między prezentowanymi metodami. Szczegółowy opis postępowania w analizie korespondencji można znaleźć w pracy A. Stanimir [2005].

$$tr\mathbf{A}^{\mathrm{T}}\mathbf{A} = tr\mathbf{A}\mathbf{A}^{\mathrm{T}} = tr\Lambda = \frac{\chi^2}{n} = \lambda = \sum_{k=1}^{K}\gamma_k^2. \tag{13}$$

Inercja całkowita rozkłada się na wszystkie osie rzeczywistych powiązań ($\lambda_k = \gamma_k^2$). Pierwsza oś główna jest tworzona z wykorzystaniem najwyższej wartości własnej λ_1. Zatem jej udział w wyjaśnieniu inercji całkowitej jest największy.

Po dokonaniu graficznej prezentacji wyników analizy korespondencji oceniane są położenia punktów obrazujących kategorie zmiennych. Jeśli kategorie dwóch różnych zmiennych są położone blisko siebie, to znaczy, że ich współwystępowanie jest znaczące. Kategorie należące do tej samej zmiennej, których punkty są położone w niewielkiej odległości, można sprowadzić do wspólnej kategorii, gdyż ich profile są do siebie bardzo podobne. Punkty położone blisko centrum rzutowania obrazują kategorie, które nie przyczyniają się do odrzucenia hipotezy o niezależności cech.

Efektem przeprowadzenia analizy korespondencji danych zapisanych w tab. 1 jest wykres prezentujący rozrzut punktów w przestrzeni R^2 (rys. 5).



Rys. 5. Graficzna prezentacja wyników analizy korespondencji aktywności ekonomicznej względem wykształcenia

Źródło: opracowanie własne na podstawie danych GUS.

Rzeczywista przestrzeń współwystąpień kategorii analizowanych zmiennych to R^3. Zatem prezentacja w przestrzeni R^2 nie spowoduje zbyt dużej utraty informacji. W przestrzeni dwuwymiarowej zachowano 98,4% + 1,5% = 99,9% informacji o powiązaniach między poziomem wykształcenia a aktywnością ekonomiczną. Na rysunku 5 można zaobserwować, że kategorie, których profile były zbliżone do profili średnich, są położone blisko centrum rzutowania (środka układu współrzędnych): A2, W4, W5. Najdalej od centrum rzutowania umiejscowione są kategorie, które mają największy wpływ na odrzucenie hipotezy o niezależności, czyli W6, A4 (po prawej stronie wykresu), A1, W1 (po lewej stronie wykresu). Interpretując rozrzut punktów na wykresie, należy stwierdzić, że dla osób z wykształceniem wyższym najbardziej charakterystyczne jest zatrudnienie w pełnym wymiarze czasu pracy. Osoby, których wykształcenie oznaczono symbolami W2, W3, W5, są najczęściej zatrudniane w niepełnym wymiarze czasu pracy. Osoby bierne zawodowo to osoby, które deklarują najniższy poziom wykształcenia.

5. Wykresy mozaikowe

Wykresy mozaikowe stanowią alternatywę dla analizy korespondencji prowadzonej dla tablicy kontyngencji. Ich stworzenie nie wymaga zastosowania skomplikowanych metod numerycznych opartych na redukcji wymiarowości. Ten typ wykresów wprowadzili Hartigan i Kleiner w 1981 r. [Hofmann 2000; Friendly 1992].

Wykresy mozaikowe są rozszerzeniem wykresów parkietowych (*sieve* lub *parquet diagram*). Angielskojęzyczne nazwy tych ostatnich wykresów są trudne do przetłumaczenia i zaakceptowania w języku polskim. Diagramy przesiewowe kojarzą się raczej z odrzucaniem obserwacji czy kategorii, a wykresy parkietowe mogłyby być błędnie kojarzone ze stosowaniem jedynie w celu zobrazowania zjawisk finansowych. Zatem w całej pracy przyjęto nazwę *wykresy mozaikowe* jako graficzną prezentację powiązań w dwuwymiarowej tablicy kontyngencji.

Wykresy mozaikowe składają się z płytek. W literaturze angielskojęzycznej można znaleźć wiele określeń, np. *tile, bin, box, rectangle* (zob. [Hofmann 2000; Friendly 1992]) oznaczających prostokąty, które tworzą wykres. Za podstawę wykresów mozaikowych można uznać skumulowane wykresy kolumnowe. Każda płytka (słupek) jest podzielona pionowo zależnie od liczebności drugiej zmiennej.

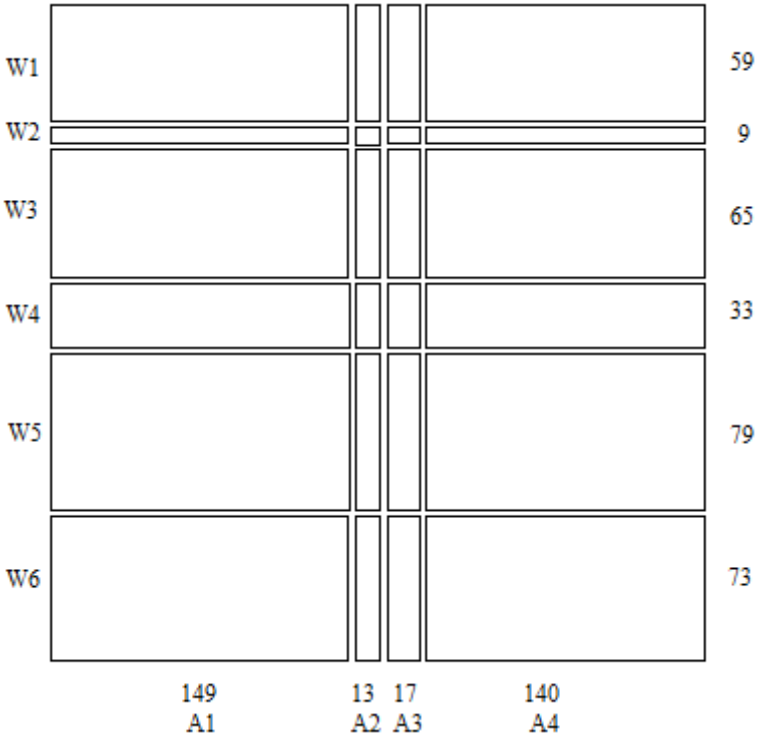
W przypadku analizy tablicy kontyngencji pole każdej płytki na wykresie jest proporcjonalne do liczebności oczekiwanej dla danej komórki. Natomiast liczebności obserwowane są przedstawiane jako określona liczba kwadratów wypełniających każdą płytkę. Różnica między obserwowaną a oczekiwaną liczebnością komórki jest prezentowana za pomocą cieniowania lub faktury linii. Dodatnie wartości odchyleń są prezentowane jednym kolorem lub liniami ciągłymi, a ujemne – innym kolorem lub linią przerywaną.

Na wykresie mozaikowym każda komórka tablicy kontyngencji jest prezentowana jako płytka, której pole odpowiada liczebności komórki tabeli. Szerokość każdej

płytki jest proporcjonalna do liczebności brzegowej każdej kolumny tabeli, a wysokość do warunkowej częstości wiersza:

$$n_{i|j} = \frac{nij}{n \bullet j}. \tag{14}$$

Taki sposób budowy wykresu mozaikowego zaproponował Friendly [1994], jako modyfikację wykresów Hartigana i Keinera, gdzie wysokość płytki była proporcjonalna do liczebności brzegowej wiersza [Friendly 1994]. Wykres mozaikowy dla podejścia stosowanego przez Hartigana i Keinera zaprezentowano na rys. 6.



Rys. 6. Wykres mozaikowy według podejścia Hartigana i Kleinera powiązań aktywności ekonomicznej i poziomu wykształcenia

Źródło: opracowanie własne na podstawie danych GUS.

W przeciwieństwie do wykresu Hartigana i Kleinera, gdzie płytki w wierszu mają tę samą wysokość, na wykresach Friendly'ego wysokości płytek dla kategorii zapisanej w wierszu są różne w kolejnych kategoriach przedstawionych w kolumnach. Na wykresach Friendly'ego łatwiej zaobserwować niezależność zmiennych,

gdyż w przypadku ich całkowitej niezależności wysokość płytek w każdym wierszu będzie jednakowa.

Kolejna modyfikacja, jaką proponuje Friendly, to zastosowanie kreskowania i zmiany kolejności kategorii (zarówno w wierszach, jak i w kolumnach), by wykres stał się bardziej czytelny i spójny. Friendly [1994] proponuje, by odpowiednie użycie kolorów i kreskowania odpowiadało standaryzowanym odchyleniom od niezależności, które są obliczane jako (oznaczenia zgodne ze wzorami (1)-(7)):

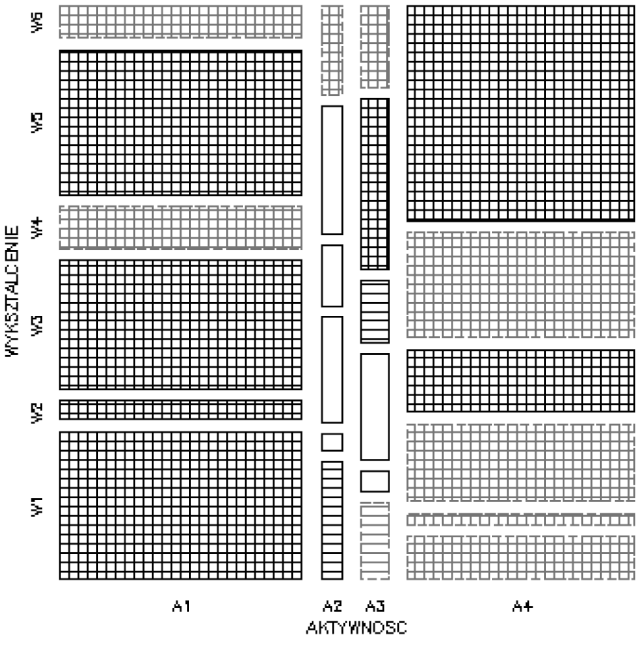
$$m_{ij} = \frac{(n_{ij} - n\hat{p}_{ij})}{\sqrt{n\hat{p}_{ij}}}. \tag{15}$$

Jeśli odchylenia wyznaczone wzorem (15) uzyskają w określonej komórce wartości dodatnie, to kreskowanie wykonuje się czarnymi ciągłymi liniami od górnego lewego do prawego dolnego rogu płytki. Gdy $m_{ij} < 0$, kreskowanie jest wykonywane za pomocą czerwonej, przerywanej linii od prawego górnego do lewego dolnego rogu. Wartość bezwzględna odchyleń jest prezentowana gęstością zamieszczanych kresek na płytce. „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].

W celu przeprowadzenia analizy danych z tab. 1 za pomocą wykresów mozaikowych skorzystano z programu Mosaic Displays proponowanego przez M. Friendlyego na stronie http://euclid.psych.yorku.ca/cgi/mosaics. Wynik zaprezentowano na rys. 7.

Wypełnienia dostępne w autorskim oprogramowaniu Friendlyego rozmijają się nieznacznie z opisem w cytowanych już opracowaniach Friendlyego kreskowania płytek w celu zaznaczenia nasilenia i kierunku odchylenia standaryzowanych różnic od niezależności dla określonej celi tabeli. Jak widać na rys. 7, jedną z opcji jest zaznaczenie kolorem czarnym odchyleń dodatnich, a czerwonym ujemnych, oraz pozostawienie niewypełnionych płytek, gdy $|m_{ij}| < 2$, wypełnienie dużymi kratkami, gdy $|m_{ij}| \geq 2$, natomiast gdy $|m_{ij}| \geq 4$, wypełnienie gęstą kratką.

Na podstawie rys. 7 można stwierdzić, że osoby zatrudnione w niepełnym wymiarze czasu pracy i mające wykształcenie W2-W5 oraz osoby bezrobotne o wykształceniu W2-W3 mają najmniejszy wpływ na odrzucenie hipotezy o niezależności analizowanych zmiennych. Największe odchylenia od niezależności prezentują dwie kategorie aktywności ekonomicznej (A1 oraz A4) we wszystkich poziomach wykształcenia. Na rysunku 7 powierzchnie płytek odpowiadają liczebnością komórek tablicy kontyngencji. Największą liczebność zaobserwowano dla komórki W6A4, następnie W1A1, W5A1, W3A1 itd. Dodatnie odchylenia od niezależności wskazują, że kategorie cech występują często łącznie, czyli należy odnotować współwystępowanie kategorii A1 i W1, A4 i W6. Jednak dla kategorii A1 należy zwrócić uwagę na fakt rzadkiego współwystępowania z kategoriami W4 i W6, gdyż odchylenie od



Rys. 7. Wykres mozaikowy wzajemnych powiązań aktywności ekonomicznej i poziomu wykształcenia

Źródło: opracowanie własne na podstawie danych GUS z wykorzystaniem programu Mosaic Displays.

niezależności jest tu zaznaczone gęstą kratką w kolorze czerwonym, czyli jest wysokie i ujemne. Również przykładowo dla kategorii A4 można wskazać, że wykształcenie W1, W2, W3, W5 nie występuje często (kolor płytek czerwony z silnym kratkowanym wzorem).

6. Podsumowanie

Zaprezentowane w artykule różne techniki i metody graficznej prezentacji powiązań między zmiennymi nominalnymi i ich kategoriami dały możliwość szerokiego zdiagnozowania zależności zmiennych.

Wszystkie zaprezentowane w artykule metody można wykorzystać również w analizie wielu zmiennych nominalnych. W przypadku wykresów profili, skumulowanych wykresów słupkowych oraz wykresów mozaikowych analiza wielu zmiennych wymaga stworzenia wielowymiarowej tablicy kontyngencji. Natomiast analiza korespondencji może zbadać wiele zmiennych nominalnych nie tylko na podstawie tablicy wielowymiarowej, ale również z wykorzystaniem macierzy Burta lub łączonej macierzy kontyngencji [Stanimir 2005]. Aspektom wykorzystania wykresów mozaikowych w analizie wielu zmiennych będzie poświęcone kolejne opracowanie autorki.

Literatura

Aktywności ekonomicznej ludności Polski. II kwartał 2011, Główny Urząd Statystyczny, Warszawa 2011 r.
Blalock H.M., *Statystyka dla socjologów*, PWN, Warszawa 1975.
Cramér H., *Metody matematyczne w statystyce*, PWN, Warszawa 1958.
Friendly M., *Mosaic display for multi-way contingency tables*, „Journal of the American Statistical Association" 1994, vol. 89, no 425.
Friendly M., *Mosaic displays for loglinear models*, „Proceedings of the Statistical Graphic Section" 1992.
Greń J., *Statystyka matematyczna – modele i zadania*, PWN, Warszawa 1984.
Hofmann H., *Exploring categorical data: interactive mosaic plots*, „Metrika" 2000 no 51.
Jobson J.D., *Applied Multivariate Data Analysis. Vol. II: Categorical and Multivariate Methods,* Springer-Verlag, New York 1992.
Rönz B., Förster E., *Regressions- und Korrelationsanalyse. Grundlagen. Methoden. Beispiele*, Gabler-Verlag, Wiesbaden 1992.
Rószkiewicz M., *Metody ilościowe w badaniach marketingowych*, Wydawnictwo Naukowe PWN, Warszawa 2002.
Stanimir A. (red.), *Analiza danych marketingowych. Problemy, metody, przykłady*, AE, Wrocław 2006.
Stanimir A., *Analiza korespondencji jako narzędzie do badania zjawisk ekonomicznych*, AE, Wrocław 2005.
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., *Metody analizy danych marketingowych*, PWN, Warszawa 1996.
Wallgren A., Wallgren B., Persson R., Jorner U., Haaland J.-A., *Graphing Statistics & Data. Creating Better Charts*, SAGE Publications, Thousand Oaks, London, New Delhi 1996.
Yule G.U., Kendall M.G., *Wstęp do teorii statystyki*, PWN, Warszawa 1966.

VISUALIZATION OF NOMINAL VARIABLES – CORRESPONDENCE ANALYSIS VERSUS MOSAIC DISPLAY

Summary: The aim of this article is to present graphical methods used in the study of nominal variables. The analysis of these variables carried out for each variable separately, is rarely craved by a researcher. The diagnosis of dependence and its strength for both nominal variables at the same time can wider describe these variables. All presented techniques and methods enable to consider nominal variables from the perspective of the category and the interaction between categories of different variables. This article presents an analysis of the profiles, the cumulative bar charts, mosaic display and correspondence analysis.

Keywords: correspondence analysis, mosaic display, profiles, cumulative bar charts.