Uniwersytet Ekonomiczny w Katowicach

blisko

międzynarodowo

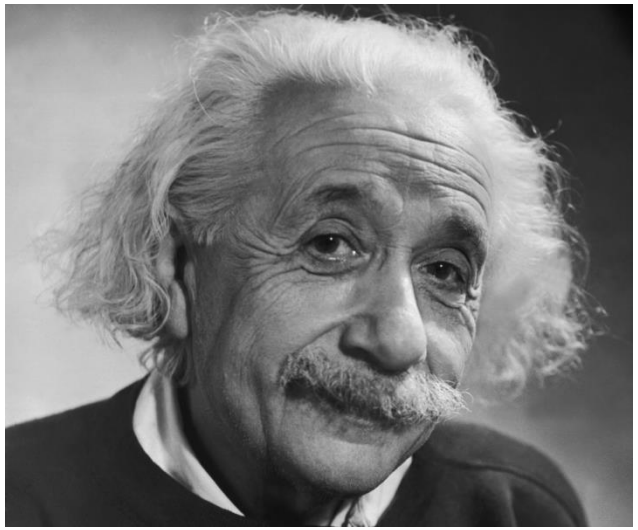
przez całe życie

Metody wizualizacji danych jakościowych w programie R

Justyna Brzezińska

Uniwersytet Ekonomiczny w Katowicach
Wydział Finansów i Ubezpieczeń
Katedra Analiz Gospodarczych i Finansowych

„If I can`t picture it,
I can`t understand it”



Albert Einstein

Plan prezentacji

1. Wprowadzenie
2. Metody analizy danych jakościowych
3. Nowoczesne metody wizualizacji
4. Zastosowanie metod wizualizacji w programie R
5. Podsumowanie

Wprowadzenie

- Metody zaprezentowane w niniejszej prezentacji należą do grupy metod analizy danych jakościowych (*categorical data analysis*)
- Szczególny nacisk położony zostanie na analizę zmiennych mierzonych na słabych skalach pomiaru zapisanych w postaci tablic kontyngencji
- Dane opisane w taki sposób są danymi przekrojowymi, a obserwacje dotyczą jednostek w określonym momencie czasu
- Zaprezentowane metody statystyczne są metodami niemodelowymi (*non-model-based analysis of association*)
- Pojęcie to wprowadził do literatury Agresti [2010] w celu określenia niemodelowych metod analizy asocjacji zmiennych o charakterze porządkowym

Pomiar zależności

- Do podstawowych mierników statystycznych, pozwalających na analizę zależności pomiędzy zmiennymi nominalnymi, należy współczynnik chi-kwadrat, McNemara, Fishera, Cohena, iloraz szans, współczynnik koligacji i korelacji Yule`a, a także oparte na chi-kwadrat miary takie, jak współczynnik Yule`a, Czuprowa, Pearsona oraz Cramera
- Mierniki te są wykorzystywane w przypadku dwuwymiarowej tablicy kontyngencji
- Nie pozwalają one jednak na pełną analizę zależności w przypadku, gdy zmienne mają charakter porządkowy oraz gdy analiza dotyczy więcej niż dwóch cech
- Spośród nich jedynie iloraz szans oraz współczynniki oparte na statystyce chi-kwadrat pozwalają na zbadanie siły zależności między większą liczbą zmiennych nominalnych

Tablice kontyngencji (1)

- Tablice, które stanowią podstawową formę zapisu zmiennych niemetrycznych, znane były w historii już ponad 2000 lat przed naszą erą
- Babilończycy wykorzystywali je do przedstawienia zależności w pewnym systemie liczbowym
- Matematycy chińscy używali tablic liczbowych w obliczeniach [Crilly 2008], które niewiele różniły się od znanej dziś tabliczki mnożenia
- Część etymologów uważa za źródłosłów terminu „tablica” słowo „stół” (*table*), który w czasach średniowiecznych wykorzystywany był do układania na nim należności podatkowych od obywateli danego państwa
- W XVIII wieku, kiedy rozwinęła się statystyka państwowa, tablice były wykorzystywane do opisu zasobów państwa

Tablice kontyngencji (2)

- Kluczowym okresem z punktu widzenia statystyki jako nauki jest przełom XIX i XX wieku, kiedy zaczęto analizować formalne własności tablic
- Pionierem w tym zakresie był Karl Pearson, który wprowadził po raz pierwszy pojęcie korelacji należące do najbardziej fundamentalnych narzędzi opisu i interpretacji zjawisk w wielu dyscyplinach naukowych, a także zdefiniował pojęcie tablicy kontyngencji
- Pearson, zainspirowany problemem losowości wyników ruletki Monte Carlo, zdefiniował także jako pierwszy współczynnik chi-kwadrat, dzięki czemu analiza zmiennych niemetrycznych wkroczyła w epokę rozwoju i zainteresowania naukowego, która nadal trwa

Tablice kontyngencji (3)

- Kolejne prace z zakresu analizy danych niemetrycznych dotyczyły liczebności oczekiwanych w tablicy kontyngencji [Galton 1892]
- W latach 1900-1912, równolegle do Pearsona, prace nad analizą tablic kontyngencji prowadził także Yule, który zdefiniował miarę zależności zwaną współczynnikiem Yule`a
- Bartlett [1935] jako pierwszy zaproponował metodę estymacji największej wiarygodności (*maximum likelihood, ML*), a w latach następnych Deming i Stephan [1940] wykorzystanie algorytmu dopasowania iteracyjno-proporcjonalnego (*iterative proportional fitting*)
- Wilks [1938] natomiast zaproponował iloraz największej wiarygodności (*likelihood ratio*), który jest alternatywą dla statystyki chi-kwadrat Pearsona, natomiast jego modyfikację zaproponował Neyman [1949]

Dwuwymiarowe tablice kontyngencji 2x2

- Tablica o wymiarach 2x2 jest tablicą kwadratowa, dla której dwie zmienne X i Y mają odpowiednio po dwie kategorie $\{X_1, X_2\}$ oraz $\{Y_1, Y_2\}$
- Empiryczne (zaobserwowane) liczebności w *h*-tym wierszu i *j*-tej kolumnie oznaczone są przez n_{hj} i oznaczają liczbę jednoczesnych wystąpień *h*-tej kategorii cechy X oraz *j*-tej kategorii cechy Y

Kategorie zmiennej X	Kategorie zmiennej Y		
	Y_1	Y_2	$n_{h.}$
X_1	n_{11}	n_{12}	$n_{1\bullet}$
X_2	n_{21}	n_{22}	$n_{2\bullet}$
$n_{\bullet j}$	$n_{\bullet 1}$	$n_{\bullet 2}$	n

Analiza zależności

- Iloraz szans: $\theta = \frac{p_{11}p_{22}}{p_{12}p_{21}}$
- Współczynnik korelacji Yule`a: $Q = \frac{p_{11}p_{22} - p_{12}p_{21}}{p_{11}p_{22} + p_{12}p_{21}} = \frac{\theta - 1}{\theta + 1}$
- Współczynnik koligacji Yule`a: $Y = \frac{\sqrt{p_{11}p_{22}} - \sqrt{p_{12}p_{21}}}{\sqrt{p_{11}p_{22}} + \sqrt{p_{12}p_{21}}} = \frac{\sqrt{\theta} - 1}{\sqrt{\theta} + 1}$
- Współczynnik korelacji: $\rho = \frac{p_{11}p_{22} - p_{21}p_{12}}{\sqrt{p_{1\bullet}p_{2\bullet}p_{\bullet 1}p_{\bullet 2}}}$
- Współczynnik chi-kwadrat: $\chi^2 = \frac{n(n_{11}n_{22} - n_{12}n_{21})^2}{(n_{11} + n_{12})(n_{21} + n_{22})(n_{11} + n_{21})(n_{12} + n_{22})}$

Dwuwymiarowe tablice kontyngencji HxJ

- Empiryczne liczebności w *h*-tym wierszu i *j*-tej kolumnie w tablicy H x J oznaczone są przez n_{hj} i są one liczbą jednoczesnych wystąpień *h*-tej kategorii cechy *X* oraz *j*-tej kategorii cechy *Y*
- Dwuwymiarowa tablica kontyngencji H x J dla zmiennych X oraz Y zbudowana jest w następujący sposób

Kategorie zmiennej *X*	Kategorie zmiennej *Y*			$n_{h\bullet}$
	Y_1	...	Y_J	
X_1	n_{11}	...	n_{1J}	$n_{1\bullet}$
⋮	⋮	⋮	⋮	⋮
X_H	n_{H1}	...	n_{HJ}	$n_{H\bullet}$
$n_{\bullet j}$	$n_{\bullet 1}$	...	$n_{\bullet J}$	n

Analiza zależności

- Iloraz szans: $$\theta_{hj} = \frac{p_{hj} p_{(h+1)(j+1)}}{p_{h(j+1)} p_{(h+1)j}}$$

- Współczynnik chi-kwadrat: $$\chi^2 = \sum_{h=1}^{H} \sum_{j=1}^{J} \frac{\left(n_{hj} - m_{hj}\right)^2}{m_{hj}}$$

- Współczynnik Cressie-Reada: $$CR = \frac{2}{\lambda(\lambda+1)} \sum_{h=1}^{H} \sum_{j=1}^{J} n_{hj} \left[\left(\frac{n_{hj}}{m_{hj}} \right)^{\lambda} - 1 \right]$$

Wielowymiarowe tablice kontyngencji

- W trójwymiarowej tablicy kontyngencji zakłada się, że analizie poddane są trzy zmienne: X, Y oraz Z
- Liczebności empiryczne oznaczone są jako n_{hjk}
- Poprzez dodawanie odpowiednich elementów względem odpowiedniej kategorii otrzymuje się liczebności brzegowe poszczególnej zmiennej, zdefiniowane jako:

$$n_{h\bullet\bullet} = \sum_{j=1}^{J}\sum_{k=1}^{K} n_{hjk} \qquad n_{\bullet j\bullet} = \sum_{h=1}^{H}\sum_{k=1}^{K} n_{hjk} \qquad n_{\bullet\bullet k} = \sum_{h=1}^{H}\sum_{j=1}^{J} n_{hjk}$$

- Współczynnik chi-kwadrat:

$$\chi^2 = \sum_{h=1}^{H}\sum_{j=1}^{J}\sum_{k=1}^{K} \frac{\left(n_{hjk} - m_{hjk}\right)^2}{m_{hjk}}$$

Zapis tablic w programie R

- Program R umożliwia zapis tablic kontyngencji w kilku postaciach:

Nazwa tablicy kontyngencji	Przykładowy zbiór danych w R	Pakiet	Opis tablicy kontyngencji
Case form	Arthritis	vcd	**wiersze odpowiadają jednemu respondentowi, a kolumny zmiennym niemetrycznym**
Frequency form	housing	MASS	**której kolumny odpowiadają zmiennym niemetrycznym, a ostatnia o nazwie Freq lub count przedstawia liczebność obiektów dla każdej z kombinacji kategorii zmiennych**
Table form	**HairEyeColor**	**MASS**	**budowanych jest wiele dwuwymiarowych tablic kontyngencji**

Zmiana sposobu zapisu tablicy w programie R

- Wybór odpowiedniej formy tablicy kontyngencji powinien odpowiadać metodzie analizy danych
- Sprawdzenie formatu i klasy obiektu umożliwia funkcja `class`
- Funkcje pozwalające na zmianę zapisu tablicy kontyngencji w programie **R:**

Zamiana z/na	Case form	Frequency form	Table form
Case form	---	`xtabs(~A+B)`	`table(~A,B)`
Frequency form	`expand.dft()`	---	`xtabs(count~A+B)`
Table form	`expand.dft()`	`as.data.frame()`	---

Metody wizualizacji danych jakościowych w R

- Metody wizualizacji wielowymiarowych tablic kontyngencji prezentuje Hartigan i Kleiner [1981, 1984], Friendly [1991, 1994, 1995, 1999, 2000, 2012] oraz Hornik, Mayer, Zeileis [2006]
- Wizualizacja zmiennych niemetrycznych, które zapisane są w postaci tablicy kontyngencji dostępna jest w pakietach:
  - `vcd`
  - `vcdExtra`
  - `ca`
  - `graphics`
  - `extracat`
- Zaprezentowane zostaną metody wizualizacji przeznaczone odpowiednio dla dwuwymiarowych tablic o wymiarach 2x2, HxJ, a także dla wielowymiarowych tablic kontyngencji

Tablice dwuwymiarowe

Dane – Główny Urząd Statystyczny – Wymiar pracy względem płci w 2016 r.

```
> library(vcd)
> dane<- c(24652, 32760, 3060, 1779)
> dane<- array(dane, dim=c(2,2))
> dimnames(dane)<- list(Plec=c("Mezczyzna",
"Kobieta"),wymiar=c("Pelny", "Niepelny"))
```

```
 wymiar
Plec         Pelny Niepelny
  Mezczyzna 24652     3060
  Kobieta   32760     1779
```

Wykres czteropolowy (1)

- Jednym z prostszych wykresów przeznaczonym dla tablic kontyngencji o wymiarach jest wykres *czteropolowy*(*fourfold display*)
- Na wykresie tym liczebności każdej komórki n_{hj} przedstawione są w postaci ćwiartki koła, którego promień jest proporcjonalny do $\sqrt{n_{hj}}$
- Wykres ten jest odpowiednikiem wykresu kołowego, jednak różnicą jest kąt koła pomiędzy wycinkiem koła, który w wykresie kołowym jest zmienny, a w wykresie *fourfold* stały (90°)
- Ponadto, promień koła, który na wykresie kołowym jest stały, a na wykresie kwadratowym jest zmienny
- W postaci łuków wewnętrznych i zewnętrznych przedstawione są przedziały ufności ilorazu szans θ na poziomie ufności $\gamma = 0{,}95$, przy czym wartości te mogą być zmieniane
- W rogach ćwiartek przedstawione są liczebności z tablicy kontyngencji, dzięki czemu możliwe jest wyznaczenie ilorazu szans

Wykres czteropolowy (2)

- Wykres czteropolowy stanowi graficzną prezentację hipotezy o niezależności postaci:

  $H_0 : \theta = 1$ - zmienne są niezależne, łuki ćwiartek okręgów nachodzą na siebie,

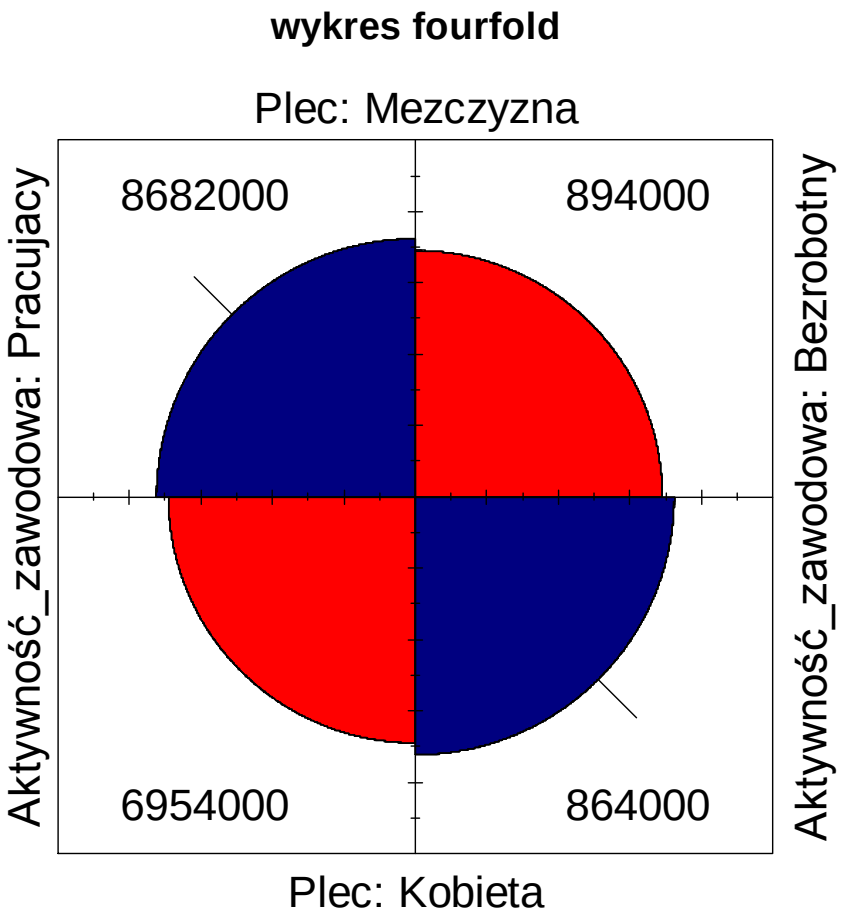
  $H_1 : \theta \neq 1$ - zmienne są zależne, łuki ćwiartek okręgów nie nachodzą na siebie

- Jeśli liczebności empiryczne są większe od teoretycznych, wówczas ćwiartka koła oznaczona jest kolorem niebieskim
- W przeciwnym wypadku tj. gdy liczebności empiryczne są mniejsze od teoretycznych, ćwiartka koła oznaczona jest kolorem czerwonym

- W programie **R** wykres czteropolowy otrzymuje się dzięki funkcji `fourfold()`

Przykład w R

```
> fourfold(dane, main="wykres fourfold")
```

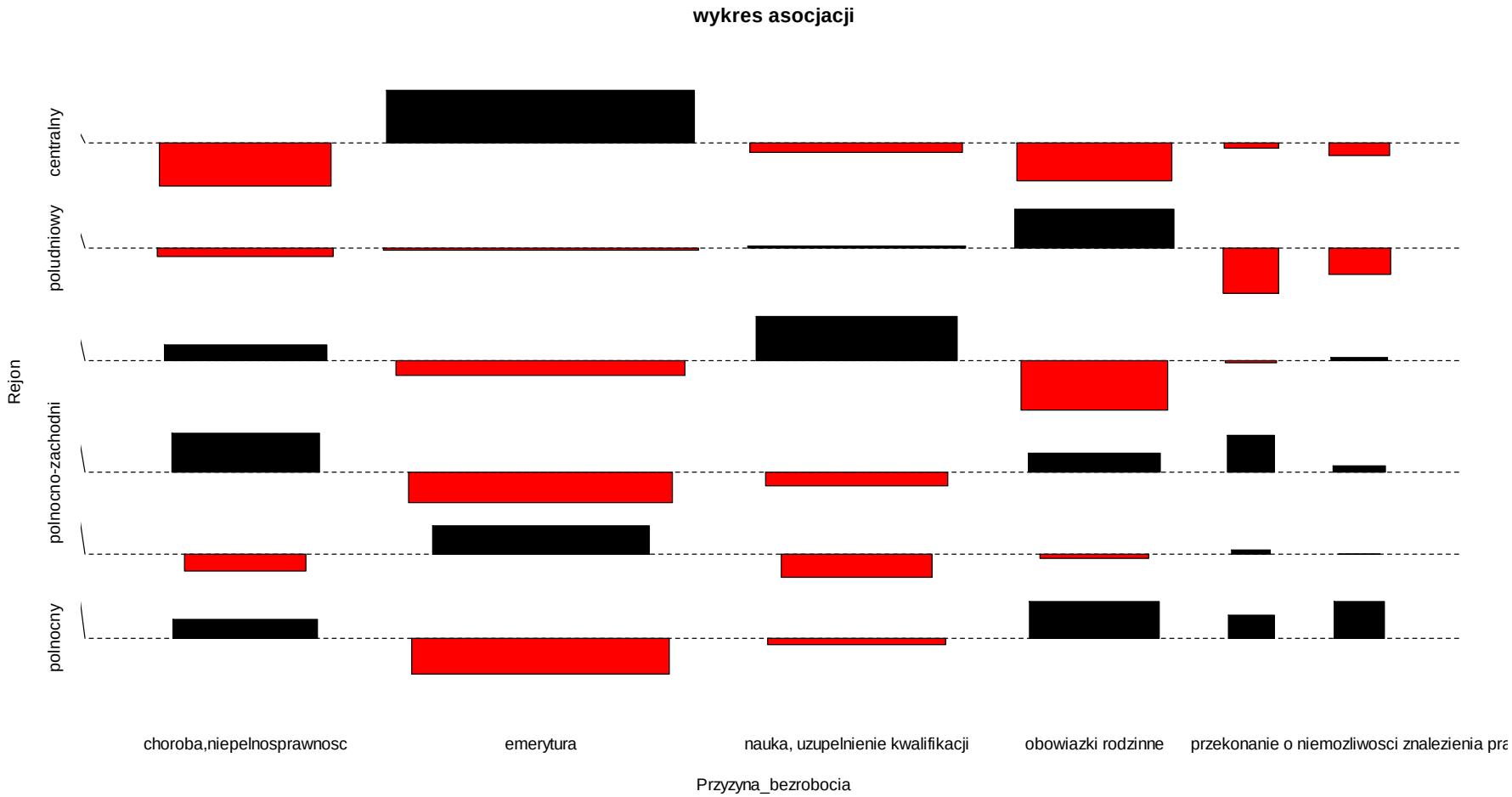


Wykres asocjacji

- Wykres asocjacji (*association plot*) jest kolejnym narzędziem wizualizacji zmiennych niemetrycznych, na którym prostokąty są proporcjonalne do liczebności teoretycznych
- Odchylenia liczebności empirycznych od teoretycznych zaznaczone są kolorami
- Jeśli różnica ta jest ujemna, wówczas prostokąt jest czerwony i znajduje się poniżej linii, jeśli różni ca ta jest dodatnia, wówczas prostokąt jest czarny i usytuowany jest powyżej linii
- Wysokość prostokąta jest proporcjonalna do standaryzowanej reszty Pearsona d_{hj}, a szerokość do $\sqrt{m_{hj}}$

- W programie **R** wykres asocjacji otrzymywany jest dzięki funkcji `assocplot()`

Przykład w R


wykres asocjacji
centralny
poludniowy
polnocno-zachodni
polnocny
Rejon
choroba,niepelnosprawnosc
emerytura
nauka, uzupelnienie kwalifikacji
obowiazki rodzinne
przekonanie o niemozliwosci znalezienia pra
Przyzyna_bezrobocia

Tablice wielowymiarowe

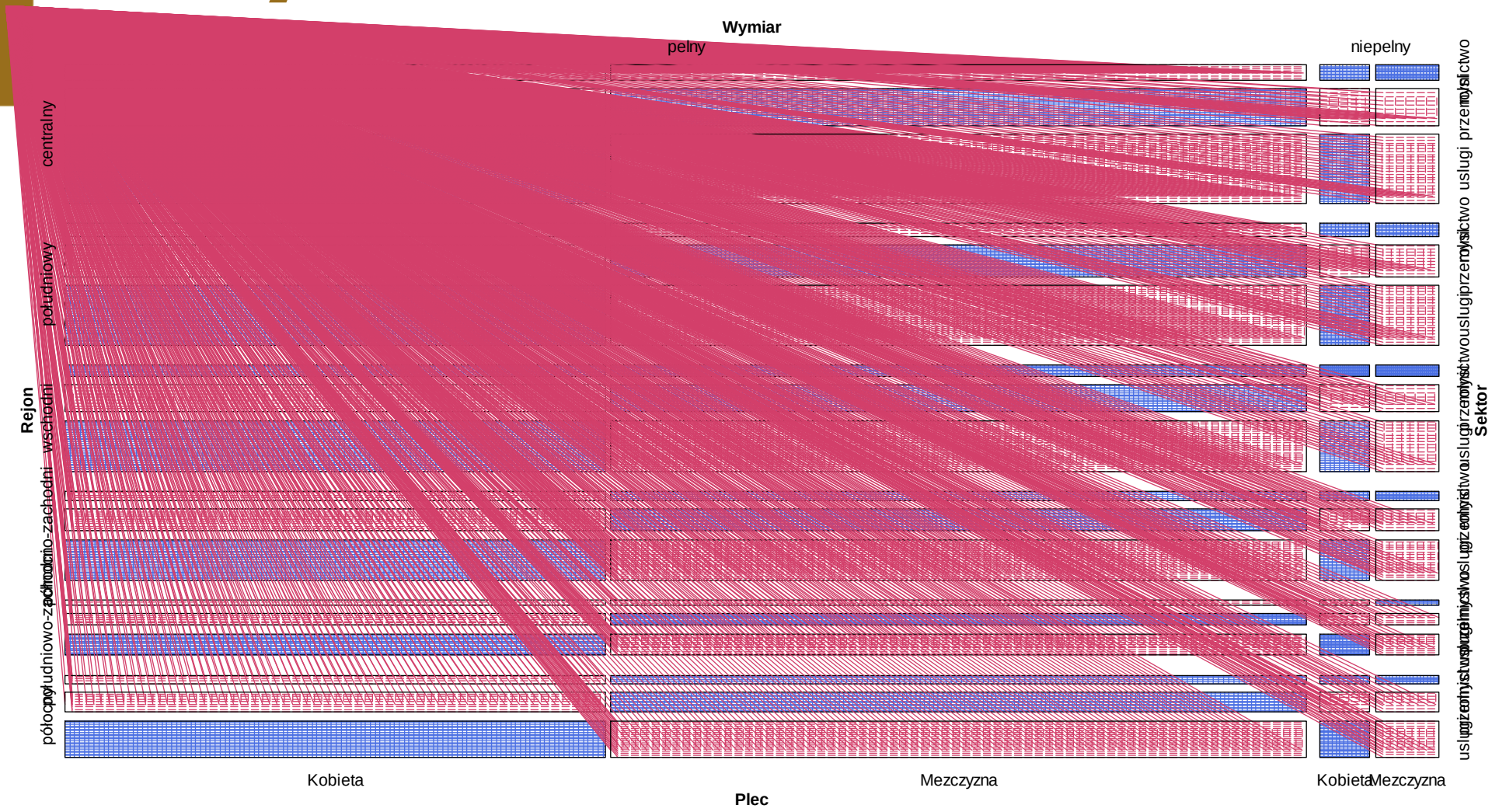
Dane – Główny Urząd Statystyczny – Aktywność zawodowa ludności w 2016 r.

```
> library(vcd)
> dane <- c(551,243,758,348,129,304,133,116,316,49,24,61,953,932,
575,727,463,586,42,62,18,38,18,32,4531,3777,2968,2602,1635,2543,492,56
7,270,303,221,298,909,380,1168,603,229,577,119,89,268,36,10,46,2660,35
80,2202,2328,1450,2025,55,84,40,53,37,36,3943,2818,2355,2151,1294,2088
,228,211,116,134,81,136)
> dane <- array(dane, dim=c(6,2,3,2))
> dimnames(dane)<- list(Rejon=c("centralny","południowy","wschodni",
"północno-zachodni", "południowo-zachodni","półocny"),
Wymiar=c("pelny","niepelny"),Sektor=c("rolnictwo","przemysl","uslugi")
, Plec=c("Kobieta", "Mezczyzna"))
```

Wykres sitkowy

- Riedwyl i Schüpbach [Riedwyl, Schüpbach 1983, 1994] wprowadzili do literatury pojęcie wykresu sitkowego (*sieve diagram*), zwanym także wykresem parkietowym (*parquet diagram*)
- Na wykresie tym powierzchnia każdego prostokąta jest proporcjonalna do liczebności oczekiwanych m_{hj} , przy czym liczebność empiryczna odpowiada liczbie kwadratów w danym prostokącie [Friendly 2000]
- Szerokość każdego prostokąta jest proporcjonalna do liczebności brzegowych kolumn $n_{\bullet j}$, a jego wysokość do liczebności brzegowych wierszy $n_{h\bullet}$
- Odchylenia liczebności empirycznych od teoretycznych ($n_{hj} - m_{hj}$) przedstawione są w postaci kolorowych linii
- Jeśli różnica ta jest ujemna, wówczas linia tworząca kwadraty jest czerwoną linią ciągłą; jeśli różnica jest dodatnia, wówczas linia w danym prostokącie jest przerywaną niebieską
- W programie **R** wykres sitkowy otrzymuje się dzięki funkcji `sieve()`

Przykład w R


Wymiar
pelny
niepelny
Rejon
centralny
południowy
wschodni
zachodni
Sektor
Kobieta
Mezczyzna
Kobieta
Mezczyzna
Plec

Wykres mozaikowy (1)

- Wykres mozaikowy jest graficzną prezentacją liczebności tablicy kontyngencji przedstawione w odpowiednich proporcjach
- Dzięki wykresowi mozaikowemu możliwa staje się także graficzna ocena modelu w analizie logarytmiczno-liniowej
- Wykresy mozaikowe mają charakterystyczny kształt, który nie zależy od struktury danych, lecz od postaci danego modelu
- Kształt ten odzwierciedla strukturę modelu zależną od występowania lub braku w równaniu modelu danego parametru
- Wykresy mozaikowe składają się z prostokątnych płytek (*tile, bin, box, rectangle*), których pole jest proporcjonalne do liczebności empirycznych n_{hj} , szerokość do liczebności brzegowych $n_{h\bullet}$, a wysokość do ilorazu $\frac{n_{hj}}{n_{h\bullet}}$

Wykres mozaikowy (2)

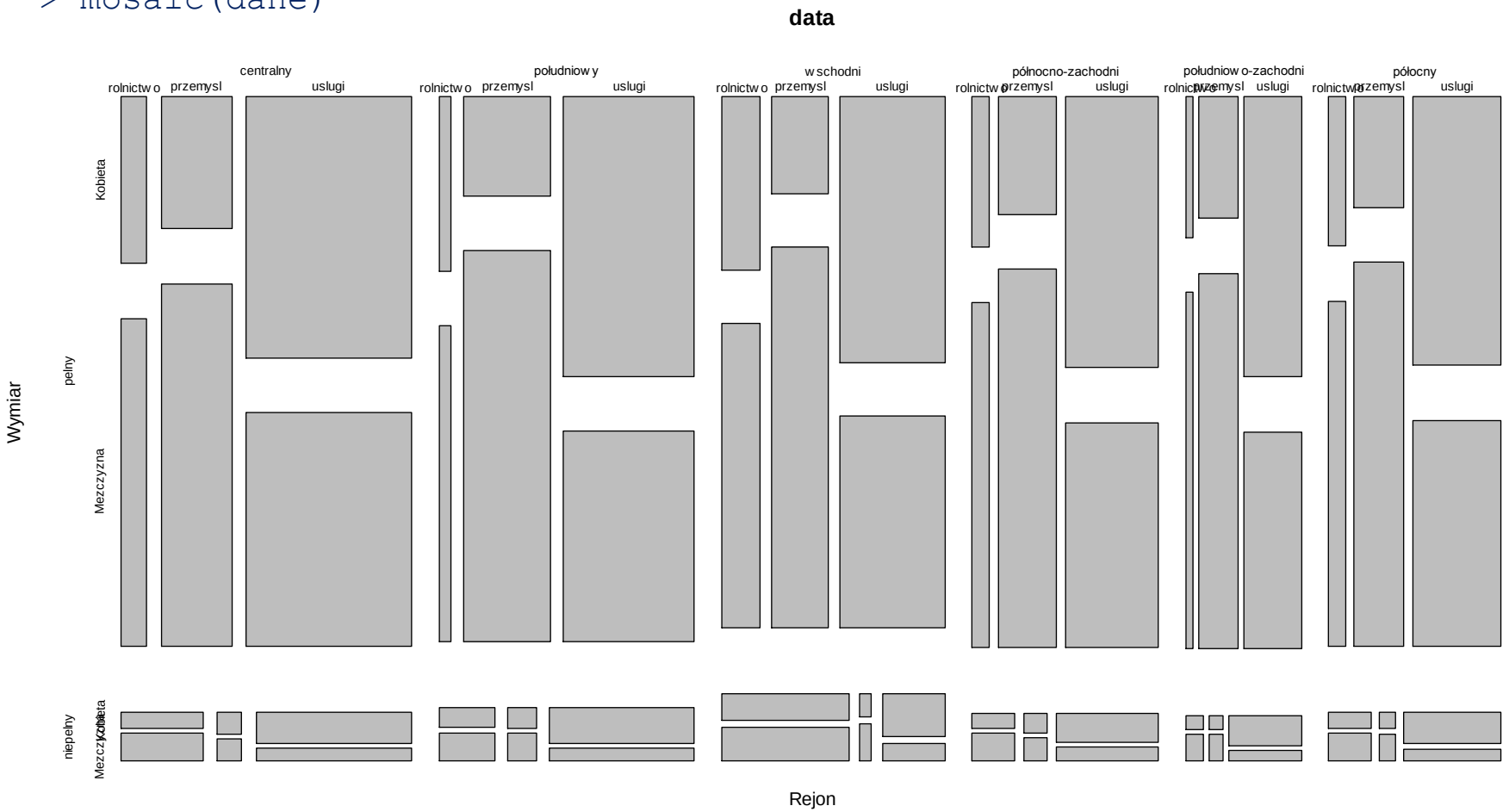
- Budowa wykresu mozaikowego oparta jest na standaryzowanych resztach Pearsona postaci:

$$\tilde{r}_{hj} = \frac{n_{hj} - \hat{m}_{hj}}{\sqrt{\hat{m}_{hj}}}$$

- Jeśli reszta jest dodatnia, dany prostokąt oznaczony jest kolorem niebieskim, jeśli ujemna – kolorem czerwonym
- Przedziały, w których znajdują się reszty w miarę wzrostu wartości d_{hj} ($|d_{hj}| > 0,2,4,...$) oznaczone są coraz ciemniejszym kolorem
- Pierwotnie wykresy mozaikowe były czarno-białe jednak zastosowanie kolorystyki w oznaczeniu reszt Pearsona spowodowało zmianę ich nazwy na wykresy udoskonalone, uzupełnione (*enhanced mosaic plot*)
- W programie **R** dwuwymiarowy wykres mozaikowy otrzymywany jest dzięki funkcji `mosaic()`

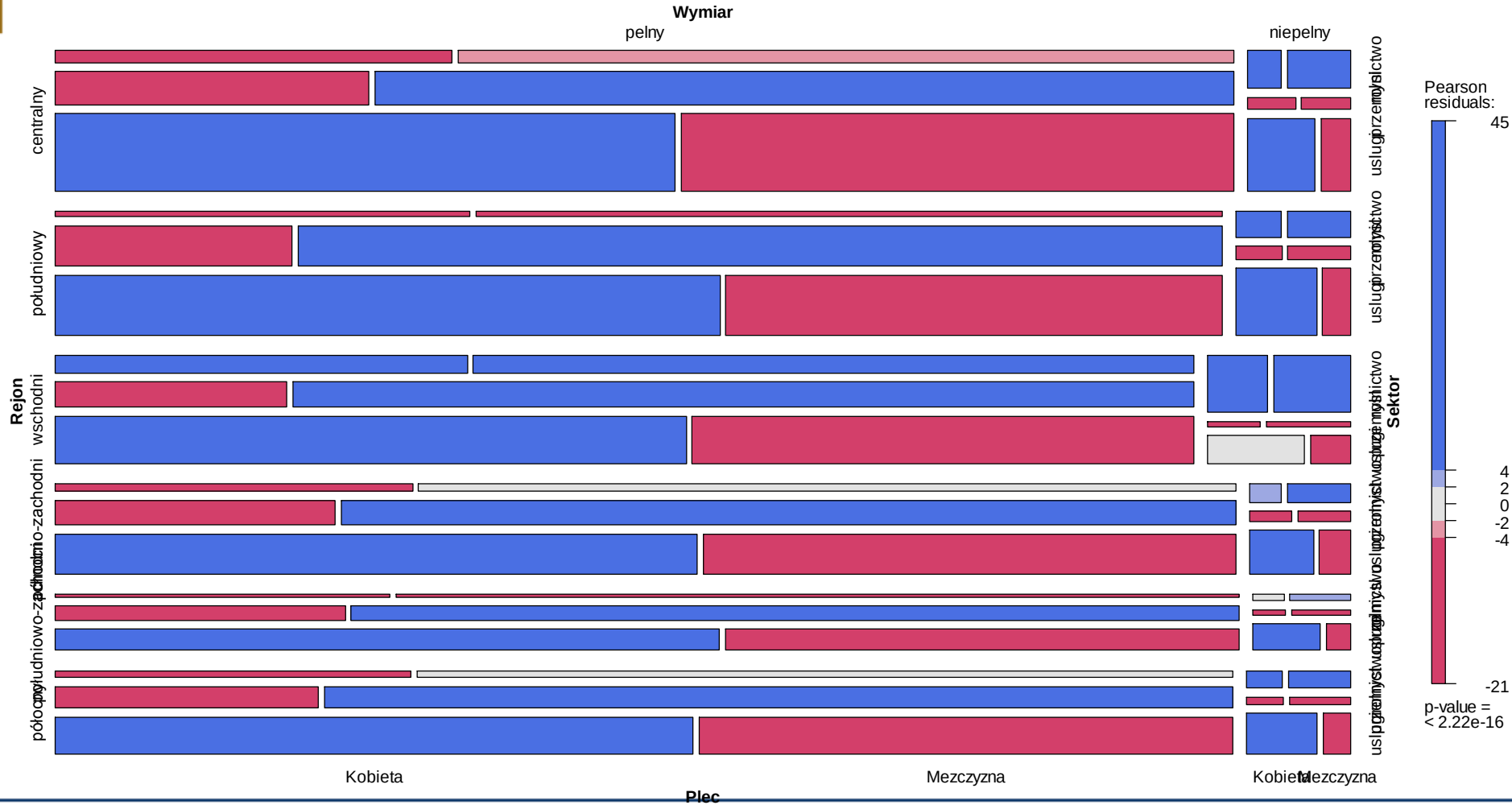
Przykład w R

```
> mosaic(dane)
```

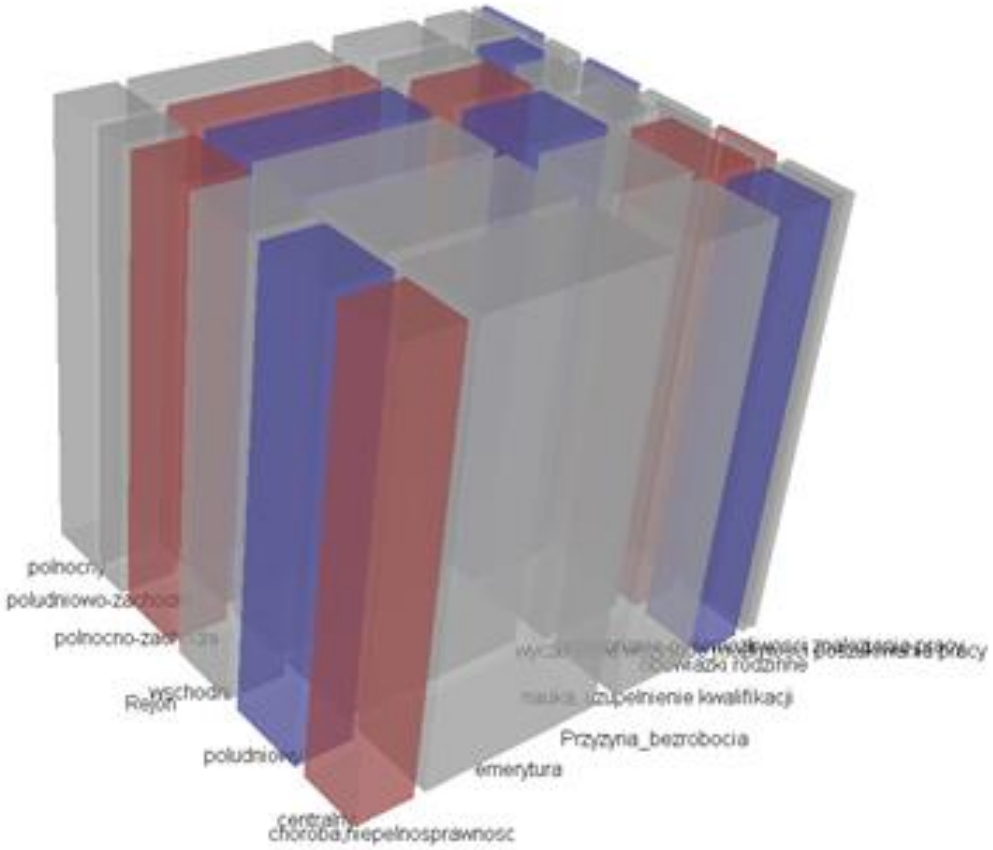


Przykład w R

```
> mosaic(dane, shade=T)
```



Przykład w R



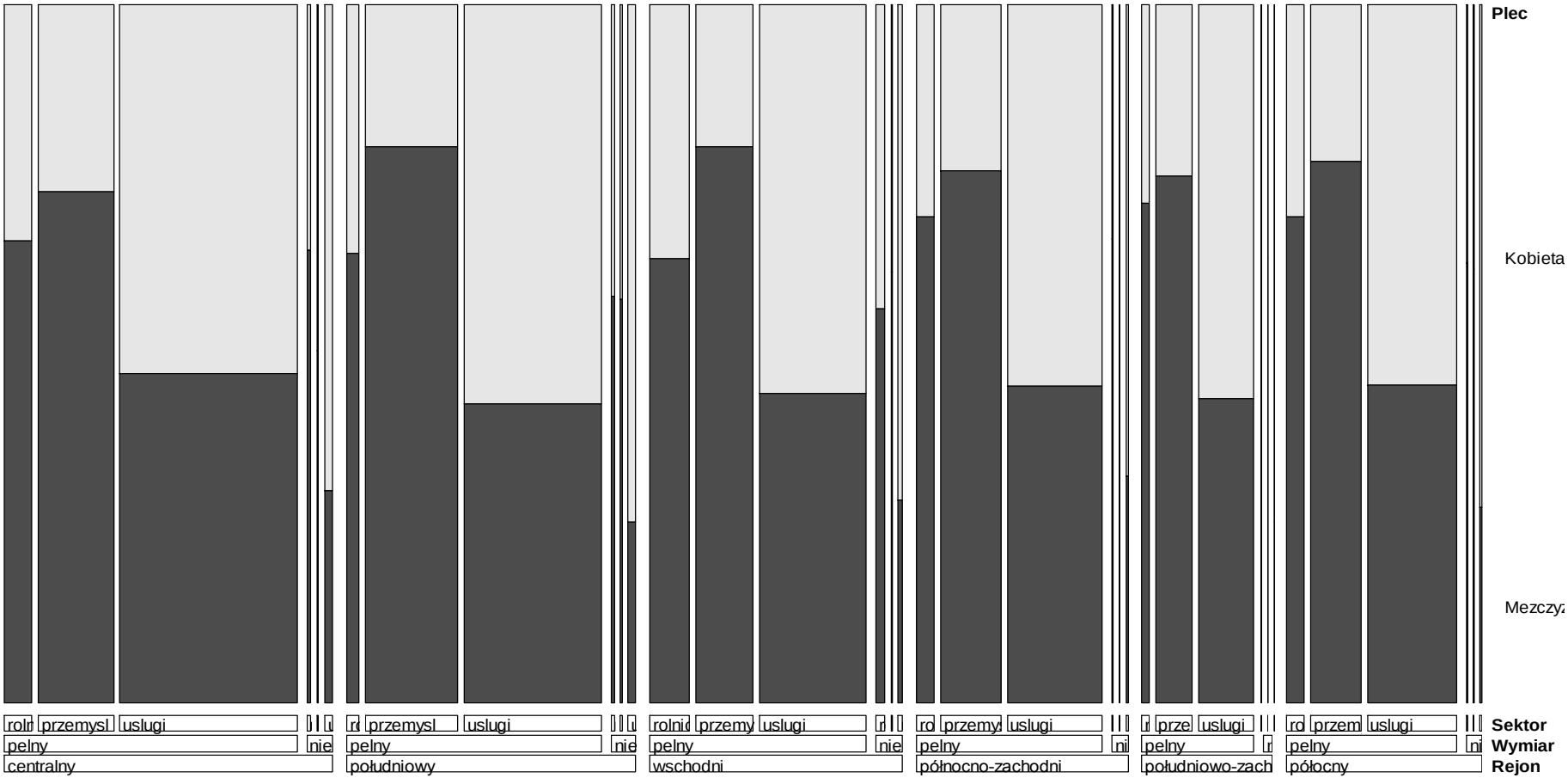
Wykres dwuwarstwowy

- W celu reprezentacji struktury dwu– i wielowymiarowych tablic kontyngencji wykorzystywany jest także wykres dwuwarstwowy (*doubledecker display*)
- Jest on podobny do histogramu dla zmiennych łączonych
- Liczebności poszczególnych kategorii przedstawione są postaci słupków dla każdej ze zmiennych

- Wykres dwuwarstwowy w programie R uzyskuje się dzięki funkcji `doubledecker()`

Przykład w R

```
> doubledecker(data)
```



Wykres zgodności

- Zgodność oceny produktu przez dwóch respondentów mierzona jest zazwyczaj za pomocą współczynnika κ-Cohena
- Spełnione powinno być założenie, że tablica kontyngencji ma tyle wierszy, ile kolumn, oraz że zmienne w wierszu, jak i w kolumnie, maja taki sam porządek kategorii
- Wykres zgodności (*agrement plot*) wprowadzony został do literatury przez Bangdiwalę [Bangdiwala 1987] i pozwala on na graficzna prezentację siły zgodności w tablicy kontyngencji
- Pomiar zgodności odbywa się poprzez analizę przekątnej tj, elementów n_{hh} , a każdy prostokąt na wykresie ma wymiar $n_{h\bullet} \times n_{\bullet h}$
- Duży prostokąt oznacza największą możliwą zgodność respondentów przy danych liczebnościach brzegowych
- Siła zgodności:

$$B_N = \frac{powierzchnia \quad ciemnych \quad kwadratów}{powierzhcnia \quad prostokontów} = \frac{\sum_{h=1}^{k} n_{hh}^2}{\sum_{h=1}^{k} n_{h\bullet} \cdot n_{\bullet h}}$$

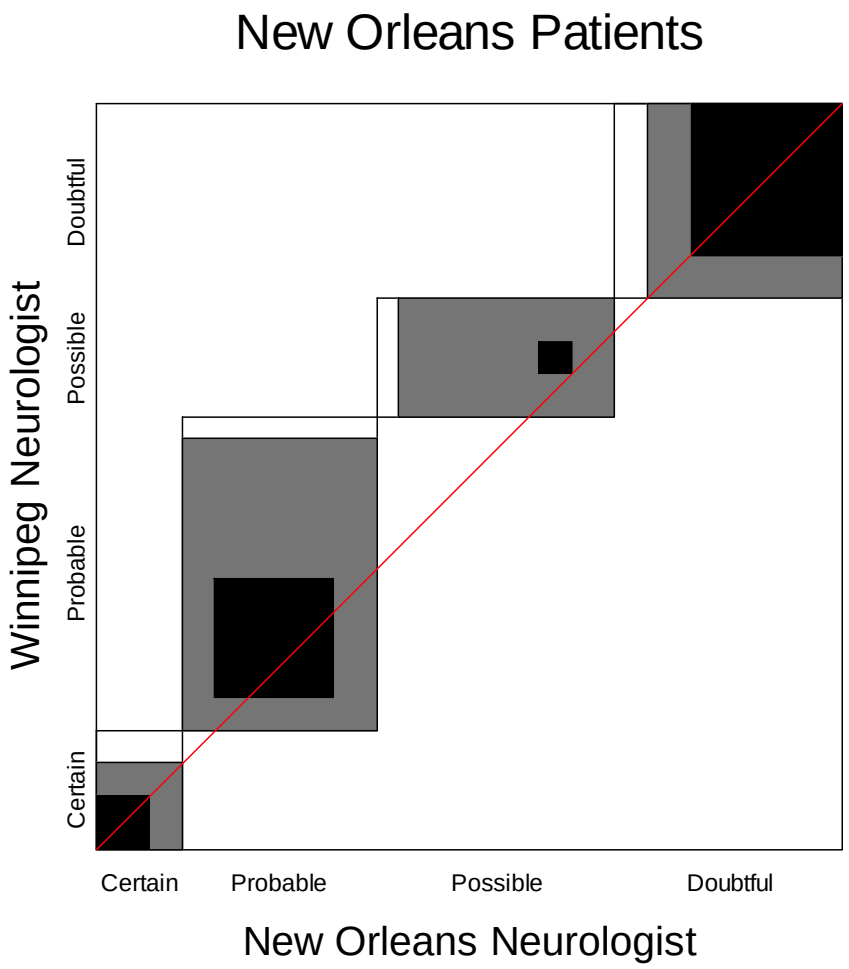
Przykład w R

- Wykres zgodności zaprezentowano na zbiorze danych `MSPatients`, który jest dostępny w pakiecie `vcd` [Westlund, Kurland 1953]
- Zbiór danych dotyczących sklerozy w postaci trójwymiarowej tablicy dla zmiennych liczący 218 pacjentów ocenione były przez dwóch niezależnych neurologów (każdy z innego miasta) stawiających diagnozę, klasyfikując chorobę do jednej z czterech kategorii

No	Name	Levels
1	New Orleans Neurologist	Certain, Probable, Possible, Doubtful
2	Winnipeg Neurologist	Certain, Probable, Possible, Doubtful
3	Patients	Winnipeg, New Orleans

- Wykres zgodności w programie **R** dostępny jest dzięki funkcji `agreementplot()`

Wykres zgodności w R



Analiza korespondencji

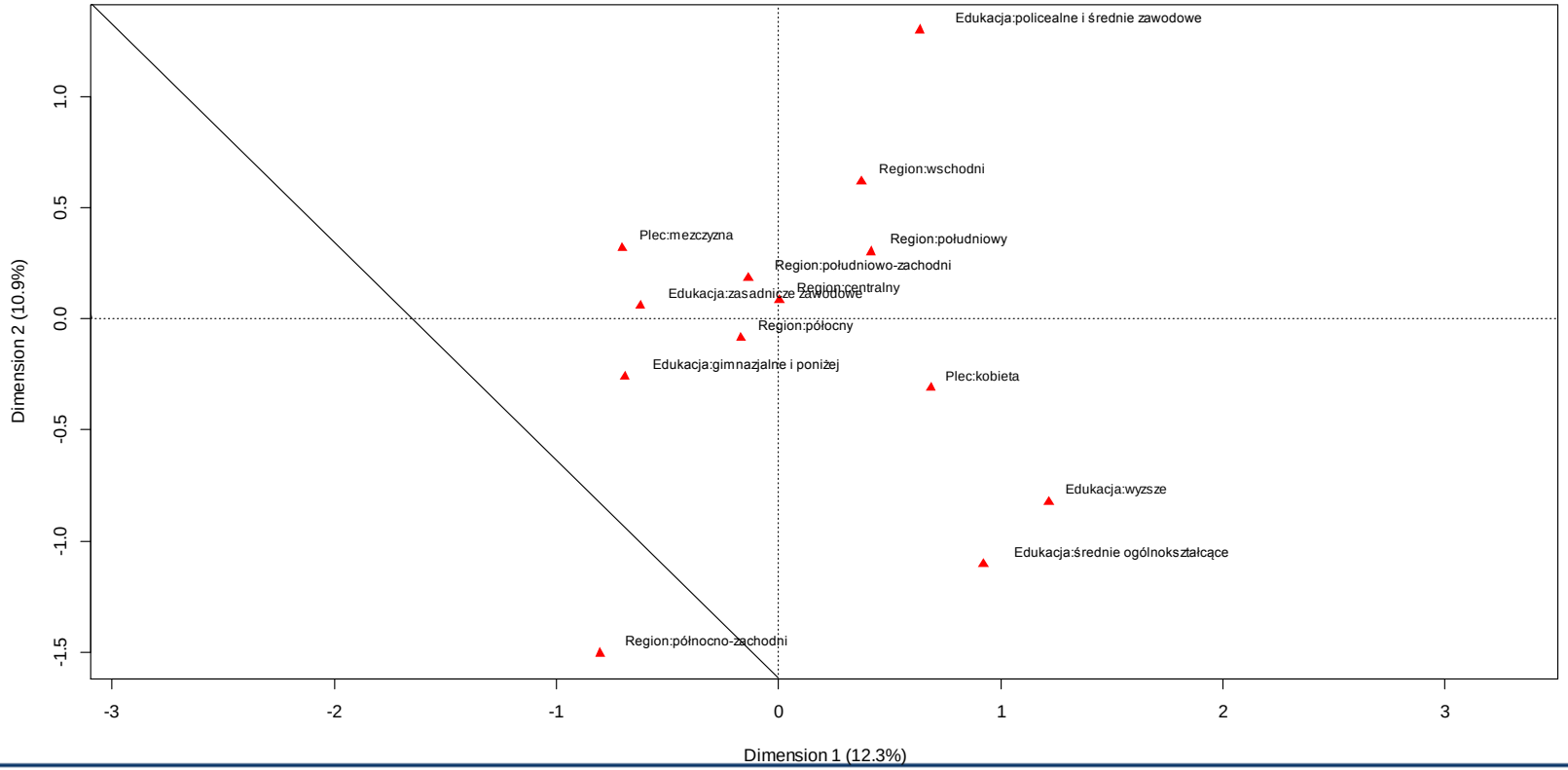
- Punktem wyjścia w analizie korespondencji jest tablica kontyngencji, na podstawie której budowana jest tablica (macierz) korespondencji
- Zakłada się, że sumy elementów w poszczególnych wierszach, jak i w kolumnach są niezerowe
- Celem analizy korespondencji jest graficzne przedstawienie wyników na tzw. mapie percepcji
- Gdy badamy dwie zmienne mamy do czynienia z klasyczną analizą korespondencji; gdy analizie poddanych jest więcej niż dwie zmienne, mówimy o wielowymiarowej analizie korespondencji
- W wielowymiarowej analizie korespondencji dane zapisane są w postaci tablic umożliwiających jednoczesną analizę wielu zmiennych: tablicy Burta, złożonej macierzy znaczników, wielowymiarowej tablicy kontyngencji lub łączonej tablicy kontyngencji

Przykład w R

Dane – Główny Urząd Statystyczny – Liczba bezrobotnych w 2016 r.

```
> library(ca)
> plec<-c("kobieta","mezczyzna")
> Plec<- factor(rep(plec, rep(30, length(plec))), levels=plec)
> Region<- factor(rep(rep(c("centralny","południowy","wschodni",
"północno-zachodni", "południowo-zachodni","półocny"),
c(5,5,5,5,5,5)),2),levels=c("centralny","południowy","wschodni",
"północno-zachodni","południowo-zachodni","półocny"))
> Edukacja<- factor(rep(c("wyzsze", "policealne i średnie zawodowe",
"średnie ogólnokształcące","zasadnicze zawodowe", "gimnazjalne i
poniżej"), 12), levels=c("wyzsze", "policealne i średnie zawodowe",
"średnie ogólnokształcące","zasadnicze zawodowe", "gimnazjalne i
poniżej"))
> n<- c(35422,  49924, 29828, 42707, 47007, 32417, 52396, 25608,45365,
40707, 42026,59220,30943,47084,37160, 22617,4151, 21451,43160,43059,
15116,26318,13217,25559,27058, 22916,47134,26733,51459,54064, 17755,
41308, 18530, 65691, 74791, 14945, 34860, 12357, 59345,48620, 20024,
50310,18315,73531,61732,10223, 24931, 9864, 53807, 50414, 7270, 17419,
6561, 35150, 35476, 10258, 27129, 12863, 61054, 63096)
```

```
> Freq<-n/1000
> dane<-data.frame(Plec, Region, Edukacja, Freq)
> data<-expand.dft(dane)
> mjca(data)
> summary(mjca(data))
> data.mjca<-mjca(data, lambda="indicator")
> plot(data.mjca, what=c("none", "all"))
```



Pakiet `extracat`

- Dotychczas prezentowane wykresy pozwalają na szczegółową ocenę struktury danych o charakterze jakościowym
- Nie dostarczają jednak one na informacji na temat rozkładu prawdopodobieństwa badanych zmiennych
- Szczegółowe informacje, które nie były dotychczas zapewniane przez narzędzia wizualizacji takie jak wykres sitkowy, mozaikowy, czteropolowy, czy asocjacji zapewniają nowoczesne wykresy dostępne w pakiecie `extracat`:

1. **`rmb`** [Meyer, Zeileis, Hornik 2006,]
2. **`cpcp`** [Uwin, Volinsly, Winkler 2003]

Dane

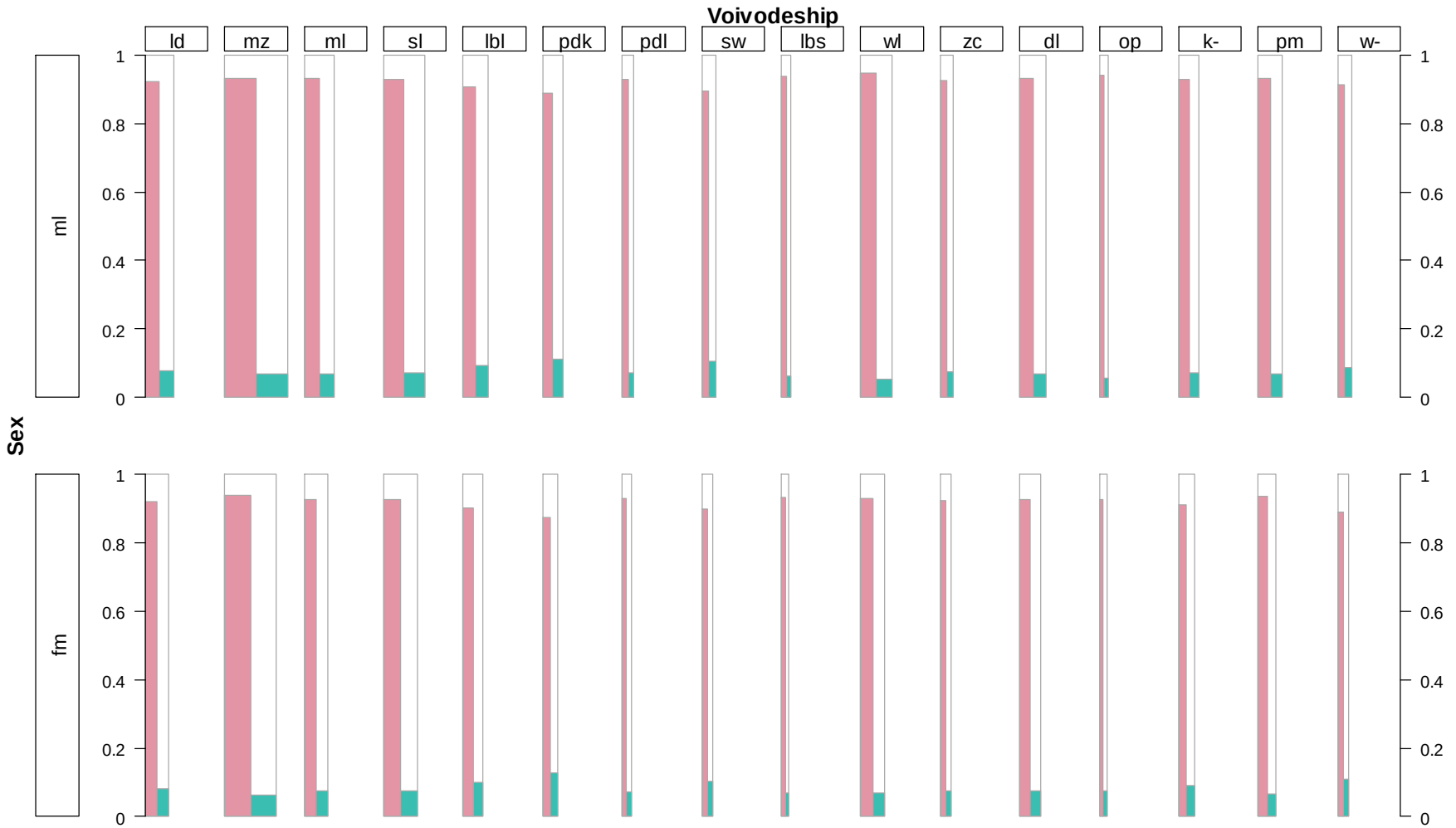
Dane – Główny Urząd Statystyczny – Aktywność Ekonomiczna w 2015 r

Zmienne	Kategorie
Województwo	łódzkie, mazowieckie, małopolskie, śląskie, lubelskie, podkarpackie, podlaskie, świętokrzyskie, lubuskie, wielkopolskie, zachodniopomorskie, dolnośląskie, opolskie, kujawsko-pomorskie, pomorskie, warmińsko-mazurskie
Płeć	kobieta, mężczyzna
Aktywność	Pracujący, bezrobotny

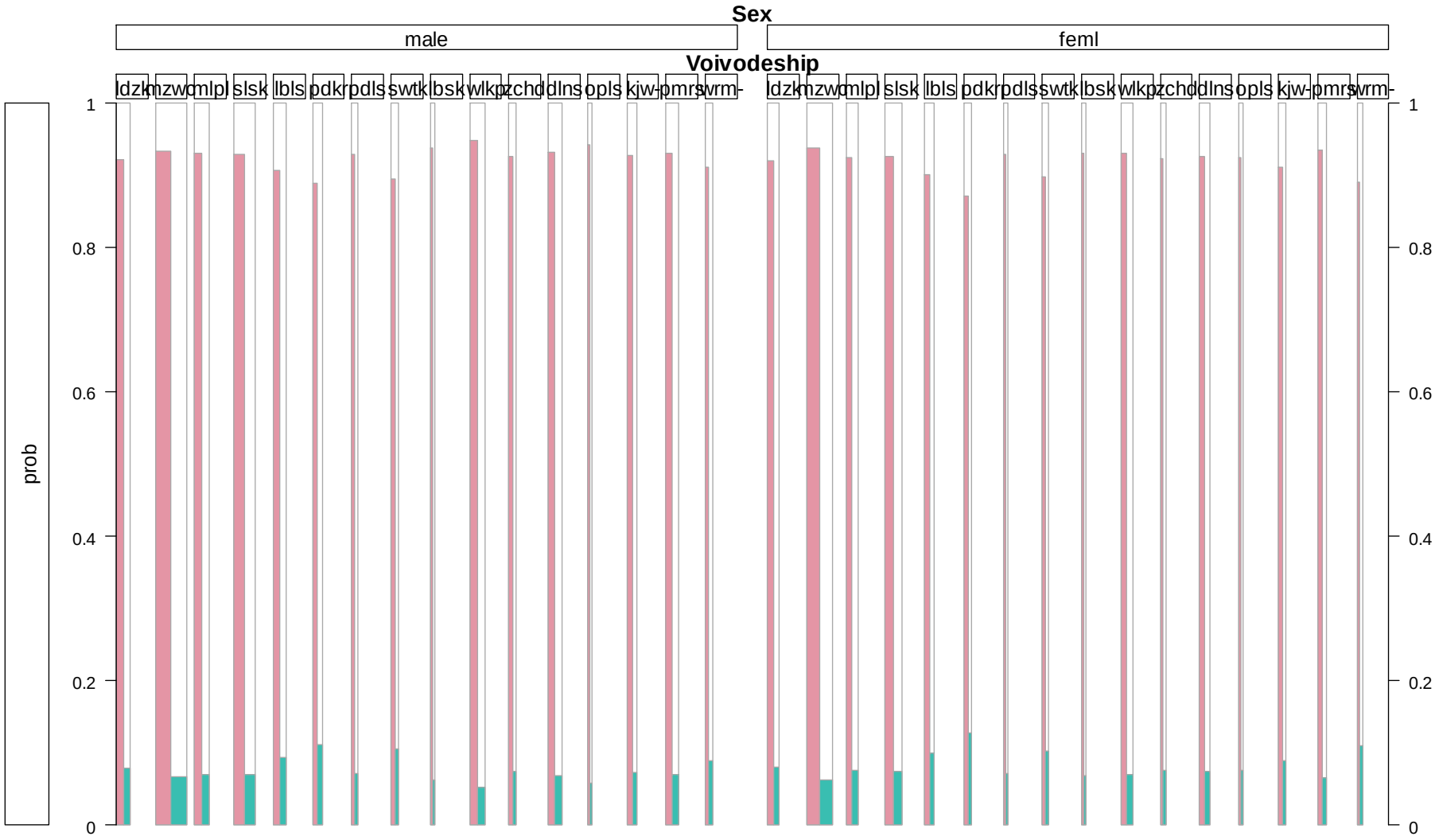
Wykres `rmb`

- Budowa wykresu `rmb` oparta jest na tych samych zasadach, co wykresu mozaikowego
- Funkcja `rmb` umożliwia generowanie wielokrotnego wykresu dla względnych częstotliwości niektórych kategorii docelowych w obrębie każdej kombinacji zmiennych objaśniających
- Wagi tych kombinacji (czyli częstotliwości bezwzględne) są reprezentowane w sumie każdego wykresu
- Budowa wykresu `rmb` możliwa jest dzięki funkcji `rmb()` w pakiecie `extracat`

Przykład `rmb` w R



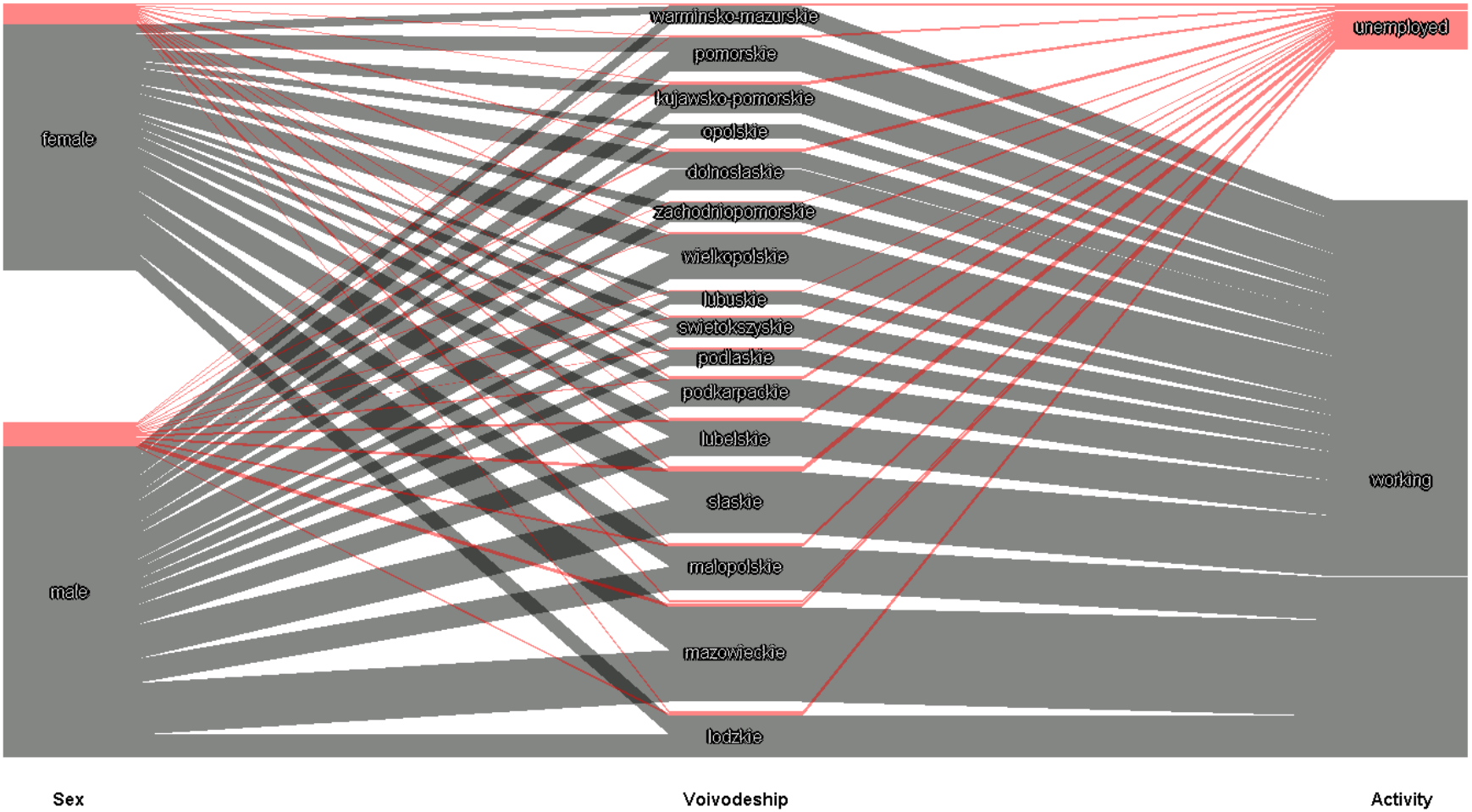
Przykład `rmb` w R



Wykres cpcp

- Koncepcja osi równoległej (*parallel coordinates plot*) powstała pod koniec XIX wieku , a jej twórcami byli Unvin, Volinsky i Winkler
- Wykresy równoległe stanowią jedno z najbardziej użytecznych rozwiązań graficznych, dzięki którym możliwa jest graficzna prezentacja wielu zmiennych na jednym wykresie
- Pierwotne pojęcie to nie pozwoliło na analizę zmiennych kategorycznych, co było jego poważną wadą
- Bendix, Kosara i Hauser [2005] opracowali aplikację danych jakościowych, która została wdrożona najpierw w oprogramowaniu Parallel Sets (wersja 2.1), a później w oprogramowaniu R [Pilhöpher, Unwin 2013]
- Dzięki ich pracy możliwe jest wykorzystanie wykresu **cpcp** zarówno dla zmienne numerycznych, jak i jakosciowych

Przykład cpcp w R



Podsumowanie (1)

- Zaawansowane programy komputerowe przyczyniły się do wzrostu zainteresowania metodami analizy danych jakościowych, które przez długi czas pozostawały w cieniu metod przeznaczonych dla danych ilościowych
- Dane jakościowe, które mierzone na słabych skalach pomiaru (nominalna lub porządkowa), zapisywane są zazwyczaj w formie tablic kontyngencji (dwu- lub wielowymiarowych)
- Wizualizacja tego rodzaju danych będąca tematem niniejszego artykułu daje szerokie możliwości określenia rodzaju zależności między zmiennymi, przedstawiając tym samym w szczegółowy sposób strukturę badanego zjawiska
- Jest to szczególnie przydatne w sytuacjach, gdy analizie poddanych jest kilka zmiennych jednocześnie

Podsumowanie (2)

- Metody wizualizacji danych niemetrycznych zaprezentowane w niniejszym artykule z powodzeniem wykorzystywane mogą być jako uzupełnienie klasycznej analizy danych jak np. analiza zależności, analiza korespondencji, czy też analiza logarytmiczno-liniowa
- Dzięki odpowiednim wykresom jak np. wykres sitkowy, mozaikowy czy też wykres asocjacji, możliwe jest przedstawienie odchyleń liczebności empirycznych od teoretycznych w danej tablicy kontyngencji w sposób graficzny, a co za tym idzie, ocena jakości dopasowania
- Narzędzia wizualizacyjne są szczególnie przydatne w sytuacjach, gdy formalny model jest skomplikowany, a interpretacja jego parametrów trudna

Dziękuję za uwagę!

Uniwersytet
Ekonomiczny
w Katowicach

www.ue.katowice.pl