ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 216, 2008

*Iwona Kasprzyk**

VISUALISATION OF A TWO –WAY CONTINGENCY TABLE IN R

ABSTRACT. The contingency table is one of the most popular ways of presenting categorical data. We can make a visualisation of data contained in the two – way contingency table using the *vcd* and *graphics* packages in the R sofware. The main aim of this paper is to show the use of various types of plots: the fourfold display, the mosaic display, the sieve diagram and the association plot. In addition to that, we can describe the relations among different categories of variables by applying the correspondence analysis.

Key words: contingency table, correspondence analysis, fourfold display, association plot, mosaic display, sieve diagram

I. INTRODUCTION

This paper provides various types of plots of visualization of a contingency table, especially the two-way table.

As an example, we present the analysis of the unemployment in the city of Bytom – a place strongly affected by the issue. The unemployment rate has been one of the highest in the Silesia area. In first half of 2006, it was over 23%. The unemployment analysis is shown on the strength of variables: time without work, age, level of education and job seniority.

II. THE ASSOCIATION PLOT

The association plot has been proposed by Cohen (1980). The height of each rectangle is proportional to the Pearson residual e.t.:

$$r_{ij} = \frac{n_{ij} - e_{ij}}{\sqrt{e_{ij}}} \tag{1}$$

* PhD Student, Department of Statistics, The Karol Adamiecki University of Economics, Katowice.

where: $e_{ij} = \frac{n_{i+} n_{+j}}{n}$.

The width of each rectangle is proportional to $\sqrt{e_{ij}}$, and the area of the rectangle is proportional to $n_{ij} - e_{ij}$. If the difference is positive, the rectangle is filled with black colour, if negative – the colour is grey.

Figure 2 presents the unemplyment analysis for Bytom. In the R software, the commands can be saved as follows:

```
> library(graphics)
> dat<-read.table("dane-Bytom.R", header=FALSE)
> rownames(dat)<-c("to 1","1-3","3-6","6-12","12- 24","over 24")
> colnames(dat)<-c("18-24","25-34","35-44","45-54","55-59","60-
64")
> dat1<-as.matrix(dat)
> assocplot(dat1)
```

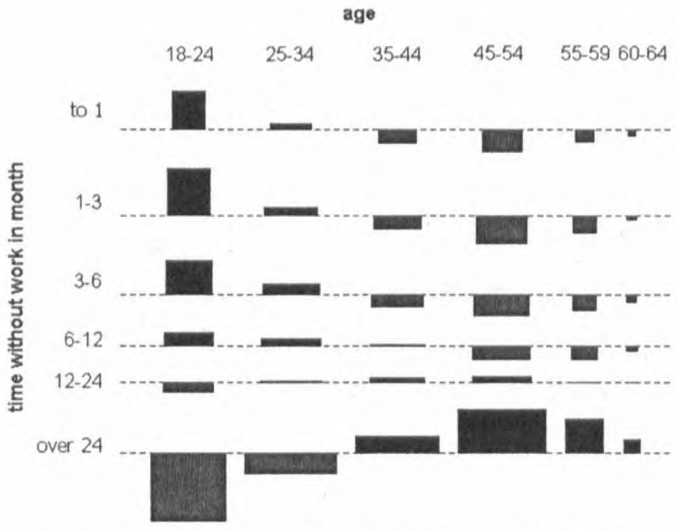

Figure 1: The association plot for age and time without work
Source: Own research.

III. THE SIEVE DIAGRAM

The sieve diagram has been proposed by Riedwyl and Schüpbach (1983) and in 1994 it was called a parquet diagram. This kind of plot divides a square unit into rectangles. The height of each rectangle is proportional to the row marginal frequency (n_{i+}), the width of each is proportional to the column marginal

frequency (n_{+j}). Hence, the area of each rectangle is proportional to the expected frequency (e_{ij}).

If the difference between the observed and expected frequency is positive, the rectangle is filled with a dark grey colour, but if it is negative, the rectangle is a light grey. Using these colours one can indicate whether the devation from independence is positive or negative. Inside each of the rectangles are drawn squares, which reflect the observed frequency contained in the contingency table.

By using the following commands in the R software, one receives the sieve diagram for the two variables: the age and the time without work are shown in Figure 2.

```
> library(vcd)
> dat<-read.table("data-Bytom.R", header=FALSE)
> rownames(dat)<-c("to 1","1-3","3-6","6-12","12-24","over 24")
> colnames(dat)<-c("18-24","25-34","35-44","45-54","55-59","60-
64")
> dat1<-as.matrix(dat)
> sieve(dat, shade=TRUE)
```

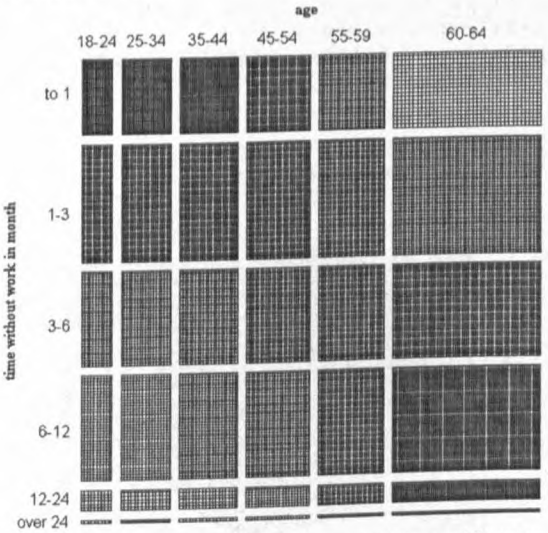

Figure 2: The sieve diagram for age and time without work
Source: Own research.

IV. THE FOURFOLD PLOT

A fourfold display has been proposed by Friendly (1994). It can be used for a 2×2 and $2\times2\times k$ table. In this kind of plot, the radius of a quarter – circle is proportional to $\sqrt{n_{ij}}$. Here, the odds ratio is used as the measure of the strength of association between the two variables contented in the contingency table ($\Theta = (n_{11}/n_{12})/(n_{21}/n_{22})$).

In Bytom, women constituted about 32% of all the registered unemployed who had been without work for over 12 months in the first half of 2006. The odds ratio is 1,5, indicating that men were 1,5 times more likely to stay without work for 12 months than women. Since the odds ratio is not 1, sex and time without work are dependent.

The fourfold display discribes the unemplyment analysis for Bytom.To present this the following listing can be derived:

```
> library(vcd)
> dat<-c(3242,2767,3540,4580)
> dim(dat)<-c(2,2)
> rownames(dat)<-c("to 12 month","over 12 month")
> colnames(dat)<-c("men","women")
> names(dimnames(dat)) <- c("time without work", "sex")
> fourfold(dat,  fontsize = 10)
```

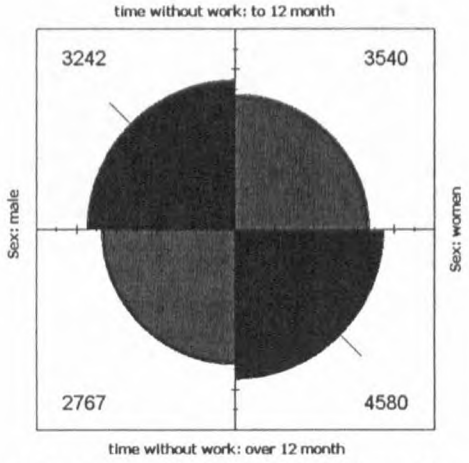

Figure 3: The fourfold display for sex and time without work
Source: own research.

V. THE MOSAIC DISPLAY

The mosaic display has been proposed by Hartigan and Kleiner (1981) and later considered by Friendly (1994). This plot is a graphical method for visualizing n-way contingency table.

For the two-way table, the width of each rectangle is proportional to the marginal probabilities ($p_i = n_{i+}/n$) and the heigth of the rectangle is proportional to the conditional probabilities for the columns given rows i ($p_{j/i} = n_{ij}/n_{i+}$).

The area of the rectangle is proportional to the observed frequency and the given probabilities:

$$p_{ij} = p_i \cdot p_{j/i} = \frac{n_{i+}}{n} \cdot \frac{n_{ij}}{n_{i+}} = \frac{n_{ij}}{n} \tag{2}$$

In the mosaic displaycolour is of great significance. The $|r_{ij}| < 2$ cells are filled with white colour and the $2 < |r_{ij}| < 4$ cells are filled with a light grey. It is very specific for this kind of plot. Then the $|r_{ij}| \geq 4$ cells are filled with a dark grey colour.

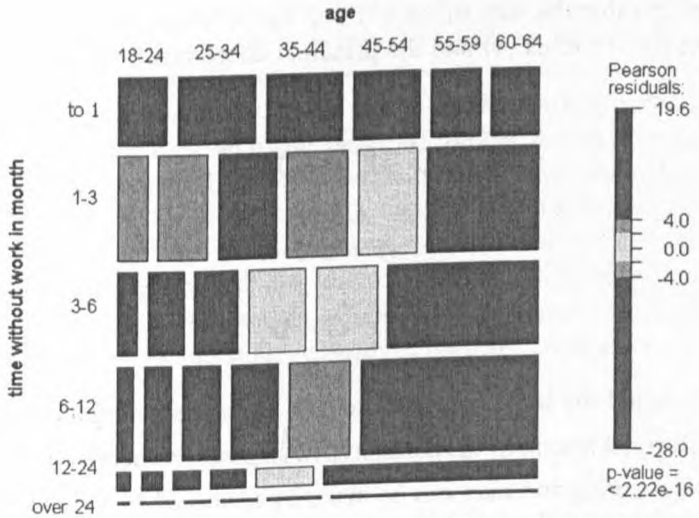

Figure 4: The mosaic display for age and time without work.
Source: own research.

The mosaic display in Figure 4, can be obtained using the following commands in the R software:

```
> library(vcd)
> dat<-read.table("dane-Bytom.R", header=FALSE)
> rownames(dat)<-c("to 1","1-3","3-6","6-12","12-24","over 24")
> colnames(dat)<-c("18-24","25-34","35-44","45-54","55-59","60-
64")
> dat1<-as.matrix(dat)
> mosaic(dat1, shade=TRUE)
```

Analyzing the example of the unemployment in Bytom, one can observe that the unemployed, aged between 60 to 64 stayed without work for 12 to 24 months. Moreover, it can be concluded that the unemployed between the age of 25 to 34 had been without work for only up to 1 month.

VI. CORRESPONDENCE ANALYSIS

The correspondence analysis is a multivariate method for categorical data. This technique analyzes the association between two or more categorical variables.

The contingency table is the starting point for the method. The next step is to create the correspondent matrix, which is defined as the matrix of the element of the contingency table divided by the size: $P = n_{ij} / n$. Using the generalized singular value decomposition (3) one can calculate the principal coordinates for the row profiles (4) and the principal coordinates for the column profiles (5):

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

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

$$\mathbf{F} = \mathbf{D}_r^{-1/2}\mathbf{U}\boldsymbol{\Gamma} \quad \mathbf{G} = \mathbf{D}_c^{-1/2}\mathbf{U}\boldsymbol{\Gamma}, \tag{5}$$

where: $\mathbf{U}$ is called the left singular vectors, $\mathbf{V}$ – the right singular vectors, $\mathbf{D}_r$, $\mathbf{D}_c$ is the diagonal matrix of the column (row) masses, respectively.

The below perception map can be made by means of MASS package in the R software using the following listing:

```
> library(MASS)
> dat<-read.table("dane-Bytom.R", header=FALSE)
> biplot(corresp(dat,nf=2),xlab = "dim 1",ylab="dim 2", cex=0.8)
```

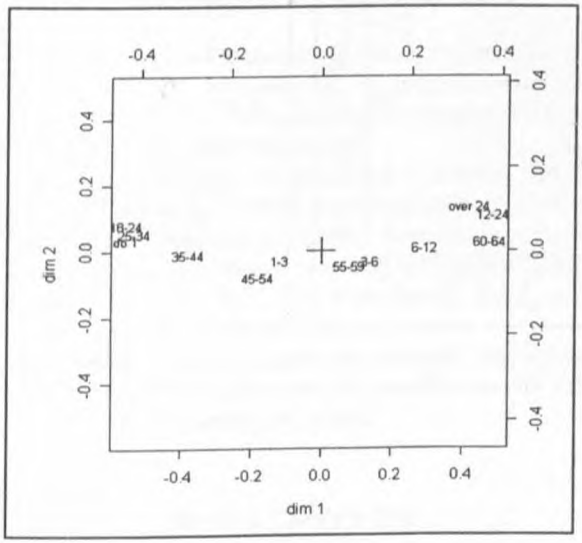

Figure 5: Perception map for age and time without work in the city of Bytom
Source: Own research.

Furthermore, creating two perception maps (Figure 6), which present the relationship assosiation between the seniority and time without work. The second map shows the relationship between time without work and the level of education among the unemployed.

On the basis of the overall analysis of the unemployment in the city of Bytom, we can conclude that the major group of the unemployed staying without work are the people between the 18–34 years of age. Those unemployed for 3 to 6 months are between 55–59 years of age, the people about the job seniority 1–5 years and 30 years of age or more are unemployed for 6–12, from 12 to 24 months – the people between 60–64 years of age and with basic vocational education and the people in 60–64 years of age, but with lower secondary, primary and incomplete primary education are unemployed over 24 months.

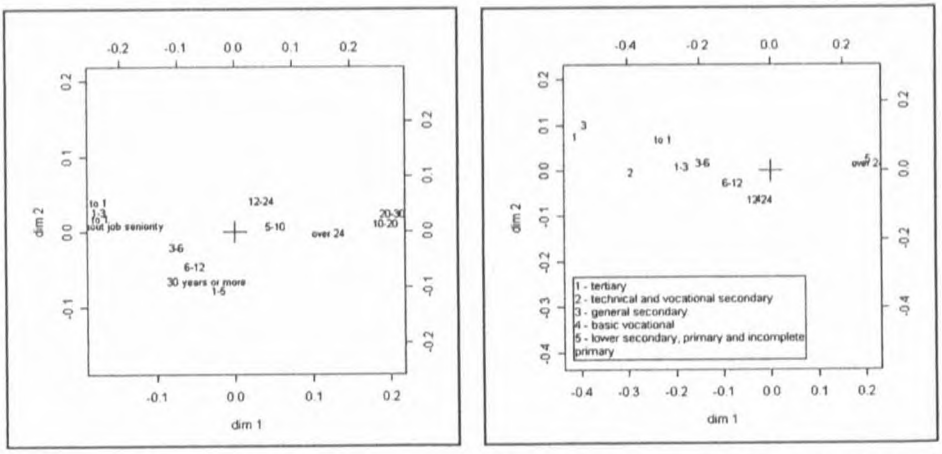

Figure 6: On the left side – the perception map for time without work and job seniority, on the right side – the perception map for time without work and level of education
Source: own research.

VII. CONCLUSION

All types of plots that have been shown in this article present the degree to which the variables in the two-way contingency table are independent or not. The correspondence analysis, as the method, allows for more profound data analysis. The main of aim this method is not only to research association between the two variables, but also to disclose the relationship between categories of variables.

REFERENCES

Clausen S.E.(1998), *Applied Correspondence Analysis. An Introduction.* Sage: University Paper 121.

Friendly M. (1994). Mosaic displays for multi-way contingency tables. *Journal of the American Statistical Association*, *89*, p. 190–200

Friendly M. (1998), *Conceptional Models for Visualizing Contingency Table Data*, in: Blasius J., Greenacre M. (eds.), *Visualization of Categorical Data*, Academic Press.

Friendly M.(1999), Extending Mosaic Displays: Marginal, Partial, and Conditional Views of Categorical Data, *Journal of Computational and Graphical Statistics*, 8, p.373–395.

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

Hartigan, J. A., and Kleiner, B. (1984). A mosaic of television ratings. *The American Statistician*, *38*, p. 32–35.

Iwona Kasprzyk

WIZUALIZACJA DWUWYMIAROWYCH TABLIC KONTYNGENCJI W PAKIECIE STATYSTYCZNYM R

Tablica kontyngencji jest częstym sposobem przedstawiania danych mierzonych zarówno na skali nominalnej jak i porządkowej. W artykule zostanie przeprowadzona analiza bezrobocia na terenie Śląska, ze szczególnym uwzględnieniem obszaru Bytomia tj. miasta szczególnie dotkniętego tą problematyką.

Za pomocą pakietu *vcd* i *graphics* w programie *R* zostanie dokonana wizualizacja danych zawartych w dwuwymiarowej tablicy kontyngencji przy pomocy kilku sposobów graficznej prezentacji, w tym za pomocą wykresu mozaikowego, wykresu siatkowego oraz wykresu zależności. W celu dokładniejszej analizy danych, wyniki zostaną przedstawione również za pomocą analizy korespondencji, która pozwala na opisanie zależności pomiędzy kategoriami zmiennych.