Volume: 10 Number: 02 Category: 02 Page: 73-82 ISSN No: 1303-6319 Cilt: 10 Sayı: 02 Kategori: 02 Sayfa: 73-82

AN APPLICATION OF COPLOT ON TURKEY DEMOGRAPHIC AND HEALTH SURVEY 2008 DATA

Yasemin KAYHAN* **Süleyman GÜNAY****

ABSTRACT

CoPlot method, an extension of multidimensional scaling, gives opportunity to investigate the relations between the observations, and between the variables on the same map. CoPlot consists of two graphs drawn on each other. First graph represents the distribution of n multivariate observations in a two dimensional space. The second graph consists of p arrows each representing a variable. By means of CoPlot, researchers can make more detailed comments about the multivariate data set with a single map. Since CoPlot is easy to understand, it has been used in various disciplines such as socioeconomic, economics and medicine but not in the demographic studies. In this study, CoPlot is briefly explained and a simple application of the method on a part of "Turkey Demographic and Health Survey, 2008" data set is presented. Superiority of CoPlot about visually interpreting the multivariate data set is emphasized by using an exceptional data set.

Keywords: Kruskal raw stress, Multidimensional scaling, Scree plot.

1. INTRODUCTION

In multivariate data analysis literature, there are various graphical representation techniques used to visualize a multivariate dataset. The main objective of these methods is to reduce multidimensional data into a lower dimension and then discover the hidden structure by means of a graphical representation of the data. One of the multivariate data analysis techniques which has been used with this purpose is multidimensional scaling (MDS). However, like many multivariate analysis methods, MDS investigates the relations between the observations, and between the variables by producing two different graphs.

MDS analysis requires only the proximities, the dissimilarities or similarities between the pairs of observations (Hastie et al., 2008). MDS tries to find the appropriate low dimensional graphical representation of the observations, so that the distances between the observations match the proximities as close as possible. In MDS analysis, similarities or dissimilarities between the observations are transformed into distances by using some specific distance functions, such as Euclidean distance, City-Block distance. Once the proximities are determined, MDS representation can be produced by using different possible optimization algorithms. The obtained graph shows that the higher the dissimilarity (similarity) measures, the larger (smaller) the corresponding distances (Borg, 2005). Although MDS can be performed either on the observations or the variables, it cannot produce a map to analyze the variables and observations, simultaneously.

CoPlot analysis, an extension of MDS, enables the simultaneous investigation of the relations between the observations, and variables with a single map. This map consists of two graphs drawn on top of each other. The first graph is obtained from MDS, and

*Yrd. Doç. Dr., Hacettepe University, Department of Statistics, Ankara, e-mail: ykayhan@hacettepe.edu.tr
**Prof. Dr., Hacettepe University, Department of Statistics, Ankara, e-mail: sgunay@hacettepe.edu.tr

represents the distribution of the *p*-dimensional observations over two dimensional space. On the second graph, the relations between variables are shown by vectors, and the location of each vector is determined by mapped observations. With this main feature, CoPlot gives researcher an opportunity to make deeper and richer interpretations of the multivariate data (Lipshitz and Raveh, 1998).

Several disciplines such as econometrics (Huang and Liao, 2012), medicine (Bravata et al., 2008), computer science (Talby et al., 1999), management science (Weber et al., 1996) require the analysis of complex multivariate data often coming from large data sets describing numerous variables for many subjects, and the multivariate nature of these data make it difficult to assess the associations of the predictors and the outcomes of interest. So, CoPlot can be seen as a valuable exploratory analysis tool in such analyses.

The objectives of this paper are simply trying to introduce the CoPlot and presenting an application of this method on a demographic data. Although CoPlot have been used previously in several disciplines, it has not been used on demographic data sets. So, this study can be considered as useful in the sense that there are not many applications of CoPlot available in the literature. Analyzing Turkey Demographic and Health Survey 2008 data set with only CoPlot map does not give enough evidence to make any conclusion in general. To get deeper understanding on the issues investigated, more works on detailed statistical analysis and inferences would be needed.

In the following section, general description and methodology of CoPlot are given. In section 3, an application of CoPlot on a part of Turkey Demographic and Health Survey 2008 data set is presented. By using an original data set, interpretive superiority of CoPlot over MDS is displayed. Some concluding remarks are given in the final part.

2. GENERAL DESCRIPTION OF COPLOT

The final product of CoPlot is a simple picture of the multidimensional dataset. With this plot, one can observe: the similarity between the observations, the correlations among the variables and the mutual relationships between the observations and the variables. The main advantage of CoPlot over a MDS is that the clusters of observations which are highly characterized by a particular variable is mapped together and located in the same direction as that of the variable's vector (Bravata et al., 2008). CoPlot analysis is performed in four steps. MDS analysis is executed in the first three steps, and in the last step the variables' vectors are placed on top of the obtained MDS graph.

Let's assume that X is an *nxp* data matrix. The rows of this matrix are assumed as the observations and will be denoted as *n* points on two dimensional space. The columns of the data matrix correspond to the variables which are exhibited by *p* vectors on CoPlot map.

Step 1: Standardization of the Data Matrix

The dissimilarities between the observations are transformed into distances by using City-Block distance in CoPlot. That is why the variables measured on different scales affect the distance values evaluated by City-Block in such a way that the largest variance variable will dominate the distance measure (Borg, 2005). So the data matrix X is transformed into matrix Z as follows,

$$Z_{ij} = \frac{\left(x_{ij} - \overline{X}_j\right)}{S_j}, \qquad i = 1,2,\ldots,n \; ; \; j = 1,2,\ldots,p \tag{1}$$

where $\overline{X}_j$ and S_j represent mean and standard deviation of the *j*-th column of X, respectively.

Step 2: Creating distance matrix

At this step, dissimilarities between the observations are transformed into distances by using a proper distance function. The distance between each pair of observations is denoted in a symmetric $D_{nxn}\left(d_{ij}\right)$ matrix. To generate D_{nxn}, 2-combination of *n*, $C(n,2)$, different d_{ij} distances are calculated as follows,

$$d_{ij} = \sum_{k=1}^{p} \left| Z_{ik} - Z_{jk} \right| > 0 \tag{2}$$

Step 3: Mapping Distances

The representation of *n* observations on two-dimensional space is generated during this step. If the rank-order of the proximities implies that the distances must have the same rank-order as proximities, we speak of non-metric MDS. In non-metric MDS, it is accepted that rank-order of the distances between the observations are informative. In this case, disparities $(\hat{d}_{ij})$ which are obtained from Euclidean distances of MDS coordinate matrix *Y* are assigned to the proximities in such a way that these values display the same rank-order as the data (Kruskal, 1964a; Borg, 2005).

The aim of non-metric MDS is to obtain a coordinate matrix *Y* such that the distance between rows *i* and *j* of *Y*, $d_{ij}(Y)$, matches disparities as closely as possible. Namely, MDS tries to minimize the following cost function,

$$\sigma^2\left(\hat{d}, Y\right) = \sum_{i=2}^{n} \sum_{j}^{i-1} w_{ij} \left(\hat{d}_{ij} - d_{ij}(Y)\right)^2 \tag{3}$$

This function is known as raw-Stress introduced by Kruskal (1964a), and the value of the function measures the quality of MDS representation. The w_{ij} is a user defined weight that should be non-negative and relates to missing information. The minimization of this function has no closed form solution, so it must be solved by iterative algorithms. Within these iterative algorithms, one of them is known as SMACOF (De Leeuw, 1977). By using this iterative majorization algorithm, optimal *Y* can be obtained.

Step 4: Adding Vectors

At this step, vectors corresponding to the variables are drawn onto the map obtained from Step 3. For each variable, CoPlot produces a vector coming up from the center of mass of the points denoted on the MDS map.

To find the direction of the vector *j*, initial angle between the *j*-th vector and the *x* axis is assumed to be zero. Then projections of all points on to the vector are calculated. The goal is to find the angle which maximizes the correlation between *n* projected scores and the *z*-score values for variable *j*. The angle which makes the correlation maximum decides the direction of the vector for variable *j*. This procedure is separately performed for all variables in the data set (Lipshitz and Raveh, 1998).

This vector representation has some advantages for the interpretation of the data. By considering the correlation values of the vectors, it may be decided that which variables should be kept in graphical representation or discarded. The vectors for highly correlated variables are located in the same direction. The vectors for highly negatively correlated variables are located along the same axis but in opposite directions. Two vectors which are orthogonal to each other imply that corresponding variables are not correlated. Obviously, the main advantage of this representation is to allow the simultaneous consideration of both variables and observations.

3. APPLICATION

For a better explanation of the procedure, some applications of the use of CoPlot are given in this section. The used data set, Turkey Demographic and Health Survey - 2008, is taken from Hacettepe University Institute of Population Studies (TDHS-2008, with the permission number 2012/8). 2,473 women who have terminated their pregnancy within the duration of their marriage are selected. Respondents (observations) are separated with respect to 12 regions. Table 1 represents the number of respondents within corresponding regions.

Table 1. Number of respondents from each region

Regions	Number of Respondents
Istanbul	187
West Marmara	134
East Marmara	199
Aegean	199
Central Anatolia	191
Mediterranean	333
West Black Sea	232
East Black Sea	124
West Anatolia	160
North East Anatolia	188
Central East Anatolia	204
South East Anatolia	322

Ten variables such as; respondent's current age (1), number of household member (2), wealth index which is an indicator of the level of wealth (3), total children ever born (4), age of respondent at 1st birth (5), age of respondent at first marriage (6), years since first marriage (7), partner's educational attainment (8), partner's age (9), number of living children (10) are selected from the survey data set. Observations are classified on the MDS and CoPlot maps with respect to the respondent's educational attainment as

follows; not educated women – square shaped points, highest educated women – cross shaped points, and the rest of the women – circle shaped points. Obtained Figure 1 and Figure 2 for the first region (Istanbul) represent distinction between MDS and CoPlot, and display what is gained in interpretability. Figure 1 shows the map produced by MDS of the ten variables describing 187 respondents using City-Block distance to calculate the dissimilarities between the observations. Figure 1(a) shows the embedding of 187 observations in a two dimensional space, and high educated women and not educated women generate two different clusters. In Figure 1(b), each variable is shown as a point and the points are arranged in such a way that their distances correspond to the correlations. That is, two points are close to each other (such as Wealth Index and Number of Household Members), if their corresponding correlation is high. However, one cannot decide the direction of this correlation. Conversely, Wealth Index and Number of Household Members are found as highly negatively correlated variables, see Figure 2. With MDS analysis, it is possible to obtain a graph that displays the relations between either the observations or the relations between variables, but seeing the observations and variables at the same graph simultaneously is not possible.

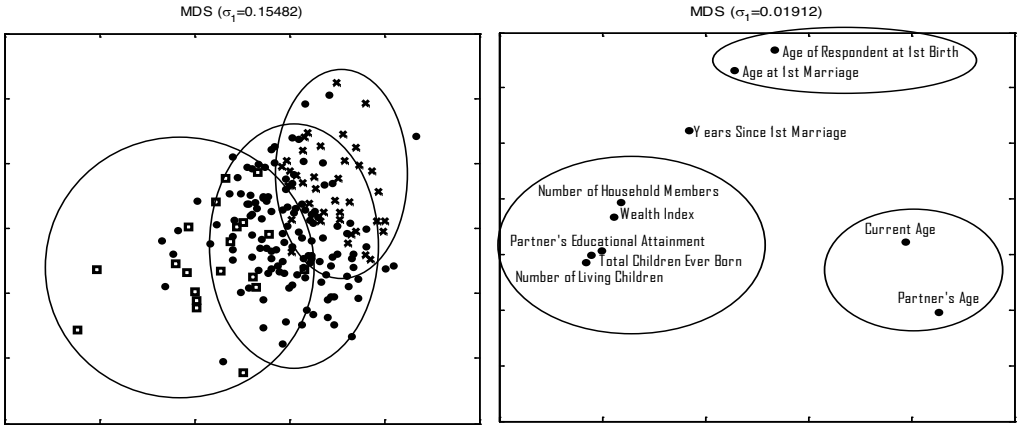

Figure 1. MDS representations of observations (a) and variables (b) for Istanbul

The map of Coplot is a simple picture of the multivariate data. From Figure 2, the following results can be concluded: Ages of the women who live in Istanbul are generally the same with the ages of their husband (1 and 9). Number of living children is highly correlated with the number of total children ever born (4 and 10). So it can be thought that child mortality is not high in this region. Age of 1-st marriage and age of 1-st birth are close to each other (5 and 6). So it can be said that women got pregnant when they got married. From the vectors 2 and 8, it is said that well educated husband do not prefer to live with big families. As the distribution of the observations and the directions of the vector are considered together, it can be said that, well educated women prefer to become mother at later age, and well educated women are richer. Not educated women have more children, and they live with crowded families. It is obvious that CoPlot map of the data set gives much more information.

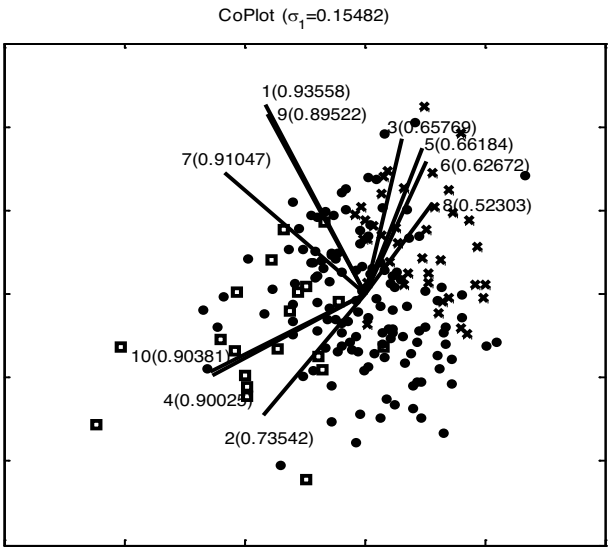

Figure 2. CoPlot representation of Istanbul

The following findings with CoPlot about rest of the regions are indicated by Figure 3 and Figure 4: For Aegean region, obtained map, therefore the comment, is nearly same as Istanbul. For Central Anatolia, variables 1, 7 and 9 are correlated. It may be concluded that not educated women got married at an early age with peers. For Central East Anatolia, not educated women rate is high relative to Istanbul, Aegean and Central Anatolia. For East Marmara, high educated women's number of living children and number of total children ever born are low (4 and 10 are in the opposite direction of cross-shaped cluster). Similar to Central East Anatolia, in Mediterranean region, the number of living children and the number of total children ever born are high for not educated women (4 and 10 are in the same direction of square-shaped cluster). For East Blacksea, high educated women get married and be mother at a late age. Similar comments can be given for the rest of the regions. We do not claim that a single map can be sufficient for these comments. These results need to be discussed sociologically and/or psychologically in details and need to be extended with more comparative statistical analysis. However, CoPlot is useful to give a researcher insight into understanding the possible relations in the data and for further analysis.

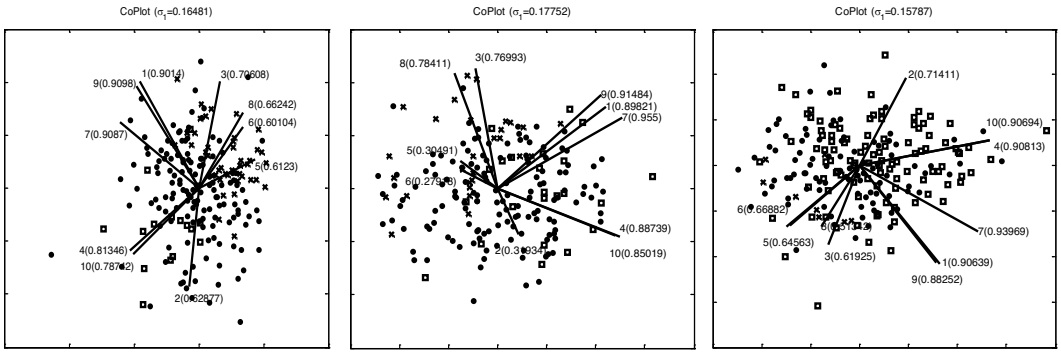

Figure 3. CoPlot maps of Aegean, Central Anatolia and Central East Anatolia

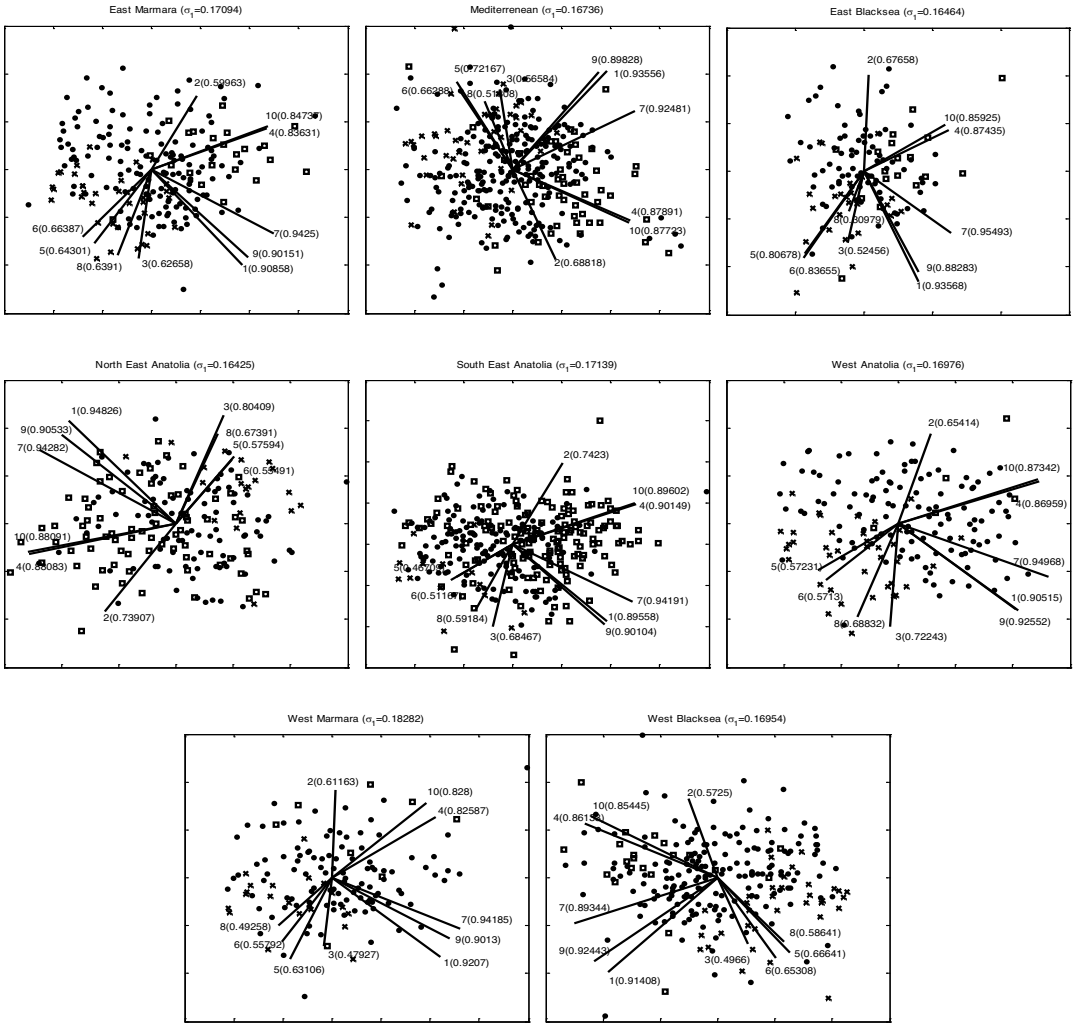

Figure 4. CoPlot maps of 8 regions

Kruskal-stress, σ_1, describes how well the map represents the observations in lower dimension. Generally MDS representations having Kruskal stress value 0.10 are accepted as fair and over 0.20 as poor (Kruskal, 1964b). To improve the goodness of fit, investigating the scree plot may be helpful. For example, from Figure 4, West Marmara has the highest σ_1 value. Figure 5 is the scree plot of this region for 30, 90 and 134 observations. For 30 observations, since σ_1 is almost 0.15, two dimensional MDS representation of the region can be accepted as fair. For 134 observations, to reach a fair representation at least 3 dimensions are needed. It is obvious that, σ_1 decreases with decreasing number of observations.

From Figure 4, it can be seen that, some of these ten variables do not have high correlation coefficient values for every region. For example, at East Black Sea, correlation coefficient values of vectors 3 and 8 are low (0.52456 and 0.30979, respectively). In CoPlot map, low correlation coefficient value, namely less than 0.70, implies that that variable is not interpretive. From Figure 6, when these two vectors are

discarded from the CoPlot analysis, it is seen that correlation coefficient values of the rest of the variables are increasing and the σ_1 is reducing. Therefore, it can be suggested that vectors with low correlation coefficients should be omitted from the analysis.

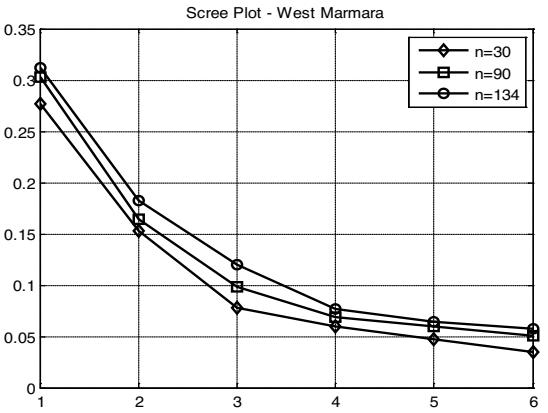

Figure 5. Scree plot for West Marmara

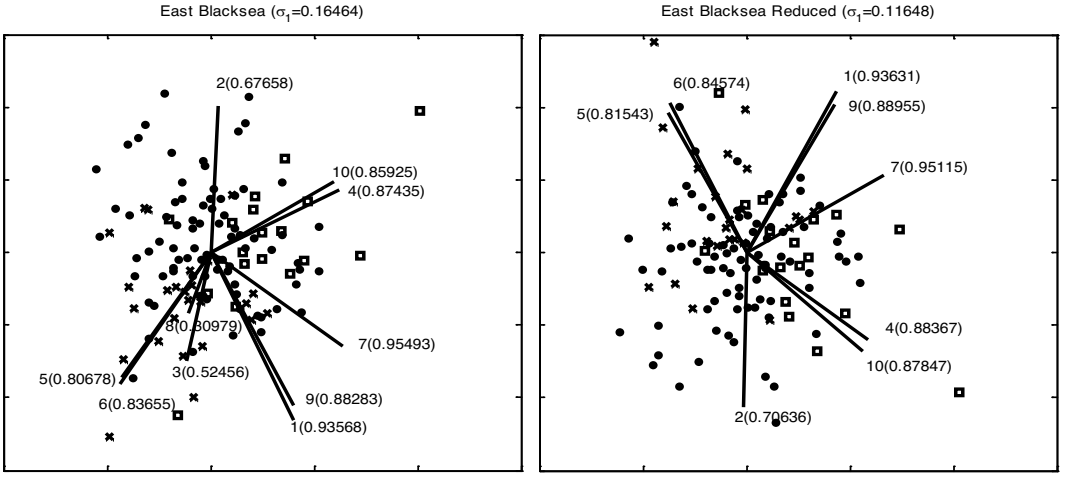

Figure 6. Effect of reducing the low correlated variables on CoPlot for East Black Sea

4. CONCLUSION

CoPlot is a graphical representation of a multivariate data set by projecting onto the two-dimensional space. It provides the possibility to analyze the observations and variables on the same graph. Previously, CoPlot has been applied to data sets from different disciplines (Raveh, 2000; Bravata et al., 2008). In this study, it has been applied to a demographic data for the first time. The purpose of this study is to show how to use CoPlot for a demographic data set. Our analysis of the demographic data suggests that CoPlot may be a useful analyzing tool for exploring the relation between observations and variables in multivariate data. At the application part of this study, it can be seen that with a simple analysis one can roughly get the picture of the regions,

and with these pictures similarities or dissimilarities between the regions would be observed. In CoPlot analysis, there are two goodness of fit measures such as Kruskall raw stress value and Pearson correlation coefficient. These measures enable the user whether to keep or delete specific variables and observations from the map. Possible problems and solutions of the problems such as low goodness of fit or correlation coefficient value are also discussed in the application part. Besides, the superiority of CoPlot on the visual interpretation of multivariate data is emphasized.

5. REFERENCES

Borg, I., Groenen, P. J. F., 2005. Modern Multidimensional Scaling, 2nd edition, Springer.

Bravata, D. M., Shojania, K. G., Olkin, I., Raveh, A., 2008. CoPlot: A Tool for Visualizing Multivariate Data in Medicine, Statistics in Medicine, 27, 2234-2247.

De Leeuw, J., 1977. Application of Convex Analysis to Multidimensional Scaling, In: J. Barra et al. Recent Developments in Statistics, 133-145.

Hastie, T., Tibshirani, R., Friedman, J., 2008. The Elements of Statistical Learning, 2nd edition.

Huang, H., Liao, W., 2012. A CoPlot-based Efficiency Measurement to Commercial Banks, Journal of Software 7:10, 2247-2251.

Kruskal, J. B., 1964a. Nonmetric Multidimensional Scaling: A Numerical Method, Psychometrika, 29:2, 115-129.

Kruskal, J. B., 1964b. Multidimensional Scaling by Optimizing Goodness of Fit to A Nonmetric Hypothesis, Psychometrika, 29:1, 1-27.

Lipshitz, G., Raveh, A., 1998. Socio-economic Differences Among Localities: A New Method of Multivariate Analysis, Regional Studies, 32:8, 747-757.

Raveh, A., 2000. CoPlot: A Graphic Display Method for Geometrical Representations of MCDM, European Journal of Operational Research, 125, 670-678.

Talby, D., Feitelson, D. G., Raveh, A., 1999. Comparing Logs and Models of Parallel Workloads Using the CoPlot Method, Lecture Notes in Computer Science 1659, 43-66.

Weber, Y., Shenkar, O., Raveh, A., 1996. National and Corporate Cultural Fit in Mergers/Acquisitions: An Exploratory Study, Management Science 42:8, 1215-1227.

TÜRKİYE NÜFUS VE SAĞLIK ARAŞTIRMASI 2008 VERİSİ ÜZERİNDE BİR COPLOT UYGULAMASI

ÖZET

Çok boyutlu ölçeklemenin bir uzantısı olan CoPlot yöntemi, gözlemler arasındaki ve değişkenler arasındaki ilişkileri aynı grafik üzerinde inceleme fırsatı verir. CoPlot, birbiri üzerine çizdirilen iki grafikten oluşur. İlk grafik n sayıda çok değişkenli gözlemin iki boyutlu uzaydaki dağılımını gösterir. İkinci grafik herbiri bir değişkeni temsil eden p sayıda oktan oluşur. CoPlot sayesinde araştırmacılar tek bir grafik ile çok değişkenli veri kümesi hakkında daha detaylı yorumlar yapabilirler. CoPlot kolay anlaşılabilir olduğu için sosyo-ekonomi, ekonomi, tıp gibi çeşitli disiplinlerde kullanılmıştır, ancak demografik çalışmalarda kullanılmamıştır. Bu çalışmada, CoPlot kısaca açıklanacak ve "Türkiye Demografik ve Sağlık Anketi 2008" veri kümesinin bir parçası üzerinde basit bir uygulaması sunulacaktır. CoPlot yönteminin çok değişkenli veri kümesini görsel olarak yorumlama üstünlüğü bu özgün veri kümesi ile vurgulanacaktır.

Anahtar Kelimeler: Kruskal raw stress, Çok boyutlu ölçekleme, Scree plot.