Global Journals LaTeX JournalKaleidoscope™
Artificial Intelligence formulated this projection for compatibility purposes from the original article published at Global Journals. However, this technology is currently in beta. *Therefore, kindly ignore odd layouts, missed formulae, text, tables, or figures.*

Multiple Perceptual Map Generation using MDSvarext

Moacyr Machado Cardoso Junior[1]

[1] Instituto TecnolAgico de AeronAutica

Received: 7 December 2013 Accepted: 31 December 2013 Published: 15 January 2014

Abstract

Perceptual mapping is widely spread to assess perception in different areas, like marketing, political and social sciences, psychology and others. One opportunity for development is to add statistical inference on final configuration in order to consider inherent differences of a group of evaluators. The main objective is to produce multiple perceptual maps from focal panel and to incorporate the confidence regions of different evaluators into the visual representation using MDSvarext. The algorithm represents a joining of non metric multidimensional scaling, shape statistical tool, clustering techniques and non parametric estimation of variance-covariance matrix to generate a visual representation of object's perception and its confidence regions. An experiment to assess occupational risk perception has been run in order to demonstrate the method. The results showed that different perceptual maps are needed to encompass the variability of a focal group. The generated perceptual maps have different interpretations since the objects may be on opposite sides of the graph. The solution generated by MDsvarext was effective and statistical inference could be done. To explore the variability in focal groups is very important, and MDSvarext represents a path to be followed, since it was possible to visualize the differences that are statistical significant.

Index terms— clustering, multidimensional scaling, perception, perceptual map, statistical inference.

1 Introduction

erceptual maps are very useful and widely used among researchers of different areas, like marketing, behavioral sciences, econometrics, social and political sciences and risk perception (Moreira, 2006; **??**lovic, 2001; **??**anlaar and Yannis, 2006; Cardoso-Junior and Scarpel, 2010).

Perceptual maps are obtained by multidimensional scaling (MDS), which is a statistical tool for dimensional reduction and visual representation of multivariate data.

Starting with a dissimilarity matrix MDS solves the problem of representing data in low dimensional space by making the inter-objects distance in low dimensional space as close as possible to the initial dissimilarity.

Statistical inference for MDS problems have been well debated in the past. Some researchers suggested that MDS should remain only as an exploratory technique or a visual representation of data. Nevertheless other researchers state that some efforts Author ? ?: Instituto Tecnológico de Aeronáutica -ITA, Praça Mal. Eduardo Gomes, 50, Vila das Acácias -São José dos Campos -SP -12.228-900 -Brazil. e-mail: Moacyr@ita.br should be done to incorporate statistical inference in MDS models. (Cox and Cox, 2001) One relevant question that arises refers to uncertainty of the final position of objects in MDS representation, especially if one is dealing with threeway MDS, which considers a group of different persons, assessing several objects on many attributes.

This paper presents the Multidimensional Scaling External Variability (MDSvarext) algorithm developed by Cardoso-Junior and Scarpel, (2012) that is an alternative to solve the problem of representing data originated in focal group studies which involves ordinal scales of judgment and inherent subjectivity.

The expected contribution of the work is to produce multiple three-way perceptual maps using visualization techniques of non metric multidimensional data, aided by a statistical shape tool. The methodological approach employed in this study was an exploratory research. This paper aims to: i) obtain multiple perceptual maps using MDSvarext, ii) Present an experimental data set collected within a focal group and to represent it in a multiple perceptual map, iii) Split the focal group into homogeneous clusters, iii) to test statistical differences between intra-clusters objects.

This paper is organized as follows: the motivation and objectives for development of this work are presented in section 1. In Section 2 the theoretical framework of MDSvarext and three-way perceptual map generation are shown. In Section 3 we present the data and results obtained in this study. Section 4 presents the final considerations.

2 II.

3 Theoretical Framework a) MDSvarext algorithm

The MDSvarext algorithm is used to incorporate the variability inherent to group of evaluators into the perceptual map obtained via non metric multidimensional scaling (NMDS). The MDSvarext method has four phases: Dimension reduction, Configuration alignment and Clusterization, and Inferential analysis, as proposed by Cardoso-Junior and Scarpel, (2012).

In the first phase, based on individual dissimilarity matrices D, a SMACOF solution algorithm, proposed by De **??**eeuwn (1977) is applied in order to reduce dimensions.

The SMACOF -Scaling by MAjorizing a COmplicated Function algorithm is used because we can guarantee a solution with monotone convergence of the Stress function, (Eq.1), proposed by Kruskal (1964).?? ?? (??) = ?? ???? ???? ???? ? ?? ???? ? 2 ?? ??<?? (1)

where ?? ???? represents weights, n denotes the number of empirical objects, ?? ???? is the dissimilarity between i and j, ?? ???? is the Euclidian distances between objects i and j in final space.

The main idea of SMACOF is to replace, iteratively, the original complicated function f(x) by an auxiliary function g(x;z). z is a known constant. The g function should comply with the following requisites, in order that g(x;z) be a majorizing function for f(x):

? auxiliary function g (x, z) should be easier to minimize than f (x);

? the original function must be less than or at most equal to the auxiliary function f (x)<= g (x, z);

? auxiliary function must touch the surface at the point of support z: f (z)<= g (z, z);

A low dimensional configuration is obtained in this phase.

In the second phase, the final individual configurations obtained from each judge, which are invariant to rotation, reflection and translation are submitted to Generalized Procrustes Analysis (GPA) in order to align the different configurations and obtain a consensual configuration through admissible rigid transformations. The final configurations are aligned according to a criterion of coordinate error minimization.

According to Brombin and Salmaso (2009), GPA is a statistical shape tool. The term shape is defined by the authors as associating the geometric properties of a configuration of points that are invariant to changes in translation, rotation, and scale. Direct analysis of a set of points is not ideal because of the presence of systematic errors, such as position, orientation and size, and GPA is usually used to conduct reliable statistical analysis, eliminating factors that are not related to the shape and aligning the configurations for a system of common coordinates.

GPA is a multivariate statistical technique involving three empirical dimensions: the objects studied, the people judging the objects, and the attributes by which the objects are judged. GPA is ideal for analyzing data from different individuals according to Dijksterhuis and Gower (2010).

The transformations allowed in GPA are translation, rotation/reflection, and isotropic scaling, so that the relative distances between the objects remain unchanged as cited by Rodrigue, (1999). Gower (1975) proposed an algorithm for the iterative solution of GPA, which has the following steps:

? Centralize each input matrix and then scale it by a constant?? = ?? ? ????(?? ?? ?? ?? ??) ?? ?? =1

where m is the number of configurations.

? Using the centered and scaled configurations, rotate X2 to X1, (??2 * ?? , H is obtained by svd decomposition of ?? 2 ?? ?? 1) and compute Z as the average of the two the current configurations, (??1+??2 * ?? 2). The configuration X3 is then rotated to Z. Z is then updated. This procedure is used until the final configuration is reached. The final value of Z is named the Configuration or space consensus.

? Calculate the sum of squared residuals as ?? ?? = ???1 ? ????(?? ?? ??)? and adjust ?? 0 = 1(?? = 1, ? , ??); This step is repeated until the criterion of tolerance is achieved, usually 0.0001.? For j = 1,..m,

In the third phase of MDSvarext, clusters shall be generated using the non-hierarchical K-means method, which seeks to maximize the distance between different clusters and to minimize the intra-cluster distances.

The K-means algorithm follows the formulation by mathematical programming, referred by Webb **??**2002).?? ? = = = ? ? ? ? ? ? ? n 1 i 1 2 1 p 1 j 2) (k c cj ij ic m x z Min (2) ????. ? ?? ???? = 1, ?? = 1, ? , ?? ?? ??=1 ?? ???? = ? 1, ???? ??????????????? ?? ?????????????? ???? ?????????????? ?? 0, ?????????????????

Where in p 1,..., j and k 1,..., c1 1 = = = ? ? = = n i ic n i ij ic cj z x z m

In which m cj is the cluster´s centroid c in dimension j, k is the number of clusters and p the number of dimensions considered.

After separation of the judges into clusters, the algorithm MDSvarext obtains the consensus solutions for each group and uses a nonparametric method to estimate the variance-covariance matrices (fourth phase) and to represent the confidence regions for each cluster generated. The Bootstrap method was used, as it is not necessary to hypothesize about the coordinate's probability distribution obtained via MDSvarext.

Bootstrap is based on intensive computation in place of theoretical analysis, providing answers to problems that are too complex for traditional approaches as well as to simpler problems. It has been made a suitable option by the sharp decline in computational costs (Efron;Tibshirani, 1986).

The Bootstrap strategy is implemented by the construction of B random samples of equal size to the original set with replacement. The Monte Carlo algorithm is then executed in three phases: In this work we used B=10.000, in order to ensure the convergence of the real value of the covariance-variance matrix.? A random

4 b) Clustering Judgment

The metrics used to validate the number of clusters or classes in which data are partitioned can be divided into two major groups: Internal and stability, as proposed by Brock et al. (2008).

For the purposes of this work we selected only internal validation metrics. We selected measures that reflect compaction, connectivity and separation of the generated clusters. Connectivity refers to the extent to which an instance is allocated to the same cluster of its immediate neighbors. Compaction evaluates the homogeneity of cluster usually calculated with intracluster variance, while separation quantifies the degree of separation of clusters, usually by measuring the distance between centroids. Since compaction and separation have opposite tendencies, namely, compaction increases with the number of groups, and the separation decreases, one option is to combine the two metrics. Two measures that represent a non-linear combination of compaction and separation are represented by Dunn index and Silhouette's width. (Everitt et al., 2001) The connectivity is defined by:??????(??) = ? ? ?? ??, ???? ??(??) ?? ?? =1 ?? ??=1 (3)

where N represents the total number of observations or instances and M is the number of dimensions.

???? ??(??) is the j-th nearest neighbor of instance i in dimension j, and ?? ??, ???? ??(??) = 0 if i and j are in the same cluster and 1/j otherwise.

Connectivity range is 0 ? con(C) ? ? , and it is a metric that should be minimized, that is, the lower the value the better the structure proposed by the algorithm, as cited by Everitt et al. (2001).

The Dunn index is the ratio of the shortest distance between instances that are not in the same cluster and maximum distance intra-cluster. The Dunn index value varies from 0 to 1 and the closer to 1 the better the result, according to Brock et al. (2008). The Silhouette's width was proposed by Kauffman and Rousseeuw, (1990) and recommended by Everitt et al. (2001). For each instance i an index S(i) ? [-1,1] is calculated. S(i) measures the difference between b(i) and a(i), where a(i) is the mean dissimilarity of instance i in relation to their cluster and b(i) is the average dissimilarity of the instance i for all instances in the nearest cluster. When S(i) is close to 1 instance i is closer of its cluster than to the nearest neighbor cluster, and thus represents a good allocation. When S(i) is close to -1, the instance is poorly allocated. The authors of the proposal also indicates that values above 0.5 represent a good result and values below 0.2 may indicate the absence of clear structure of the data. Finally, Everitt et al. (2001) warn that it is not prudent to rely on only one of the metrics to select the optimal number of clusters.

5 III.

6 Results

In order to verify the results obtained by the MDSvarext we collected data to establish the multiple perceptual map. The main objective was to assess the perception of a focal group of safety engineer´s students regarding occupational risks. For this purpose a questionnaire was applied. The questionnaire listed 10 objects. The objects represent occupational risks that are classified into two major groups: physical and chemical agents as shown in Table 1. For each object the respondents were asked to assign scores on a Likert scale from 1 to 7 in nine dimensions, as Table 2. The forms provided to respondents contained objects arranged in a random way, aiming to eliminate any possibility of systematic error in data collection. Respondents were only given instructions on how to fill the form, using the Likert scale, with no explanation of the meaning of each object. The focal group comprised 14 students from a Safety Engineering course. The raw matrices, with 10 objects and 9 dimensions, obtained from students were then submitted to MDSvarext. The first checkpoint is to verify what is the ideal number of dimensions. Usually 2 or 3 dimensions are recommended for better visualization of data.

7 Global

Figure **??** shows the Scree Plot of the adjustment of Stress function and dimension. What can be seen is that for two dimensions we obtain the greatest decrease in stress function, thus, this is the dimension to be adopted for data representation. The SMACOF solution was obtained with SMACOF package of statistical software, R, version 2.15.0. implemented by De Leeuwn and Mair (2009). GPA was performed using the statistical software R and the SHAPES package written by Dryden, (2009).

After running the algorithm MDSvarext the perceptual maps obtained can be seen in Figures **??** and 3. The clustering process generated two perceptual maps. These clusters have different interpretations, but overall both groups perceive the physical and chemical hazards as different. As can be seen in Figures **??** and 3 groups are on different sides in the generated maps. One exception is ”mercury” in cluster 1.

The number of generated clusters was validated with the assistance of clValid package implemented on software R by Brock, (2011). The results are shown in Table 3. We can extract from Table 3 that the results obtained with two different clustering algorithms, one hierarchical and the other non-hierarchical, the last one proposed to run along with the algorithm MDSvarext. In both cases the optimal number of clusters to be generated was two.

Finally the last phase of the algorithm generates the confidence regions using a nonparametric technique, Bootstrap. To demonstrate the solution obtained for the two clusters, we selected only two risks. They are on opposite sides of the map, and belong to different groups of risks. The depiction of only two risks has the sole purpose of generating a clean map.

In Figure 4 it could be seen the confidence regions representation for ”NOISE” and ”MANGANESE” in cluster 1. What can be observed is that the two risks are perceived differently for cluster 1. The students belonging to cluster #2 were not able to discriminate the two risks statistically, as can be seen by the overlapping ellipses.

It should be emphasized that this analysis using focal group, with only 14 students, or sometimes even less is quite usual due to the costs for obtaining a very large sample. And the method proposed is suitable from the statistical point of view, to deal with small samples.

IV.

8 Conclusions

The main conclusion of this paper is that perceptual maps generated from individual subjectivity analysis have large variance due different perception inherent to a focal group. The proposed MDSvarext deals with that problem by generating more than one perceptual map, which reduces variability by splitting the data in more homogeneous subgroups. This finding could help researchers to better interpret data from subjectivity studies.

The problem of optimal number of clusters to be generated is not overcome, and more efforts should be done in order to fulfill this gap, but it is beyond the scope of this paper.

9 J

Figure 1:

[1]

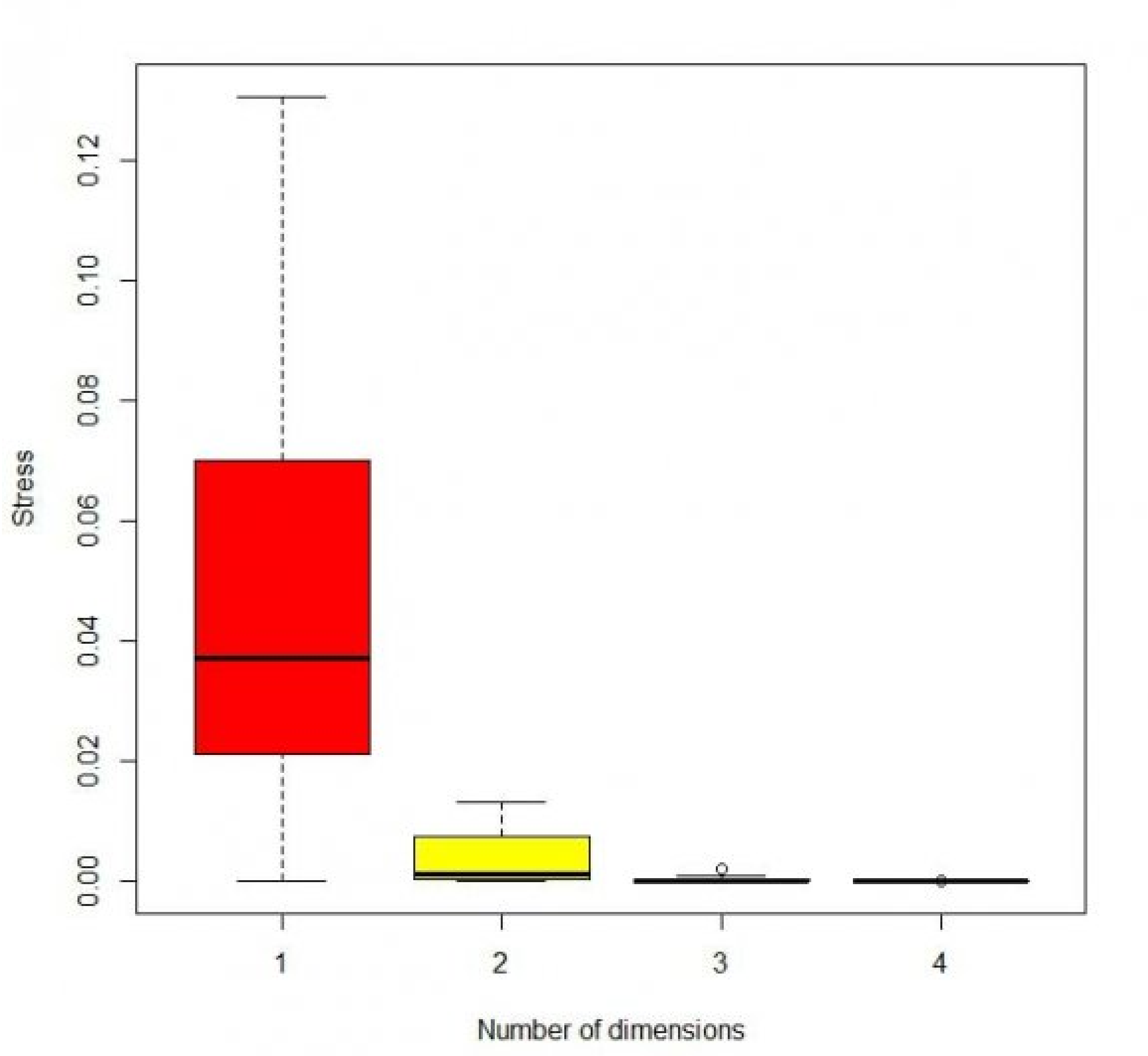
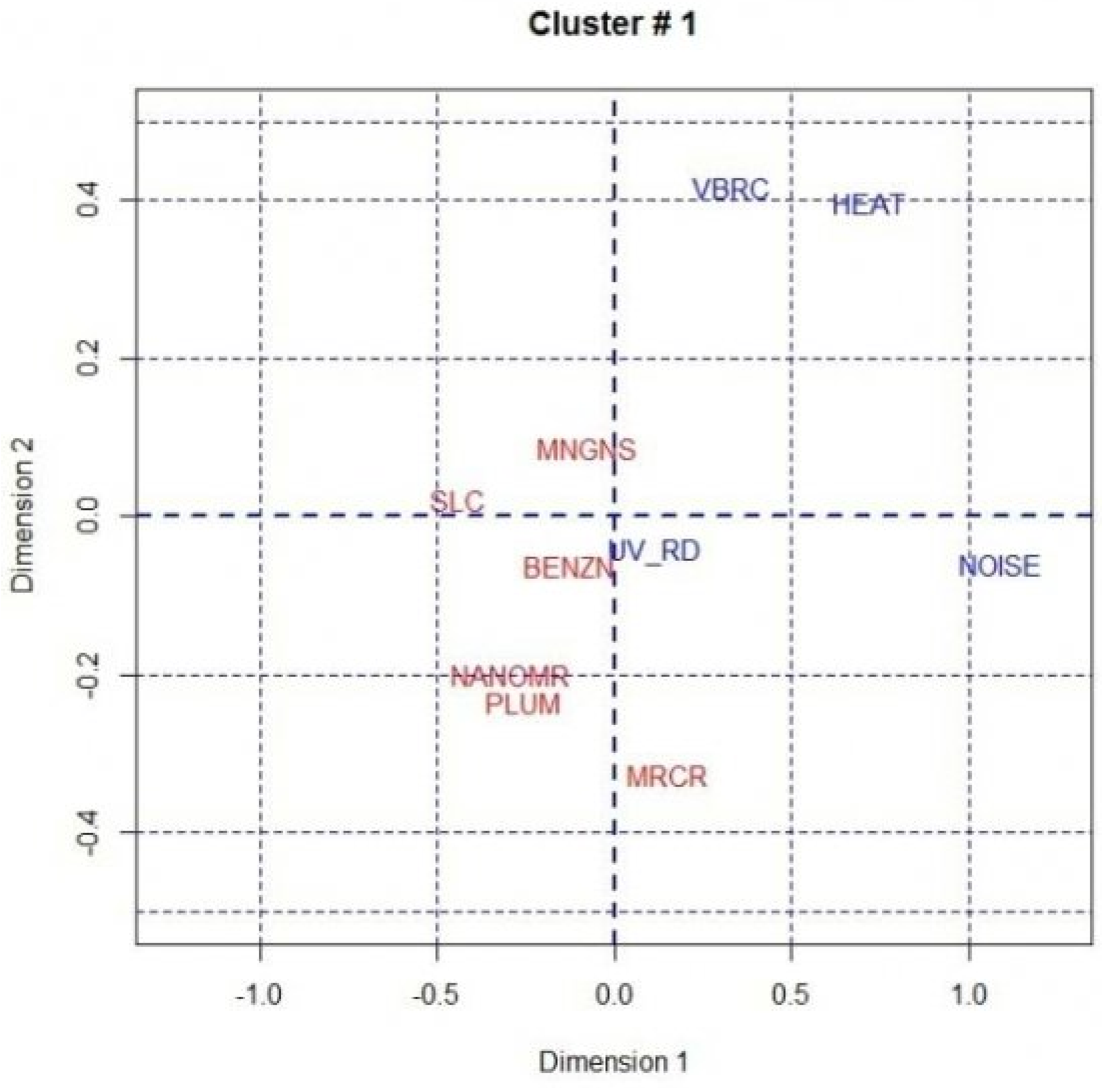
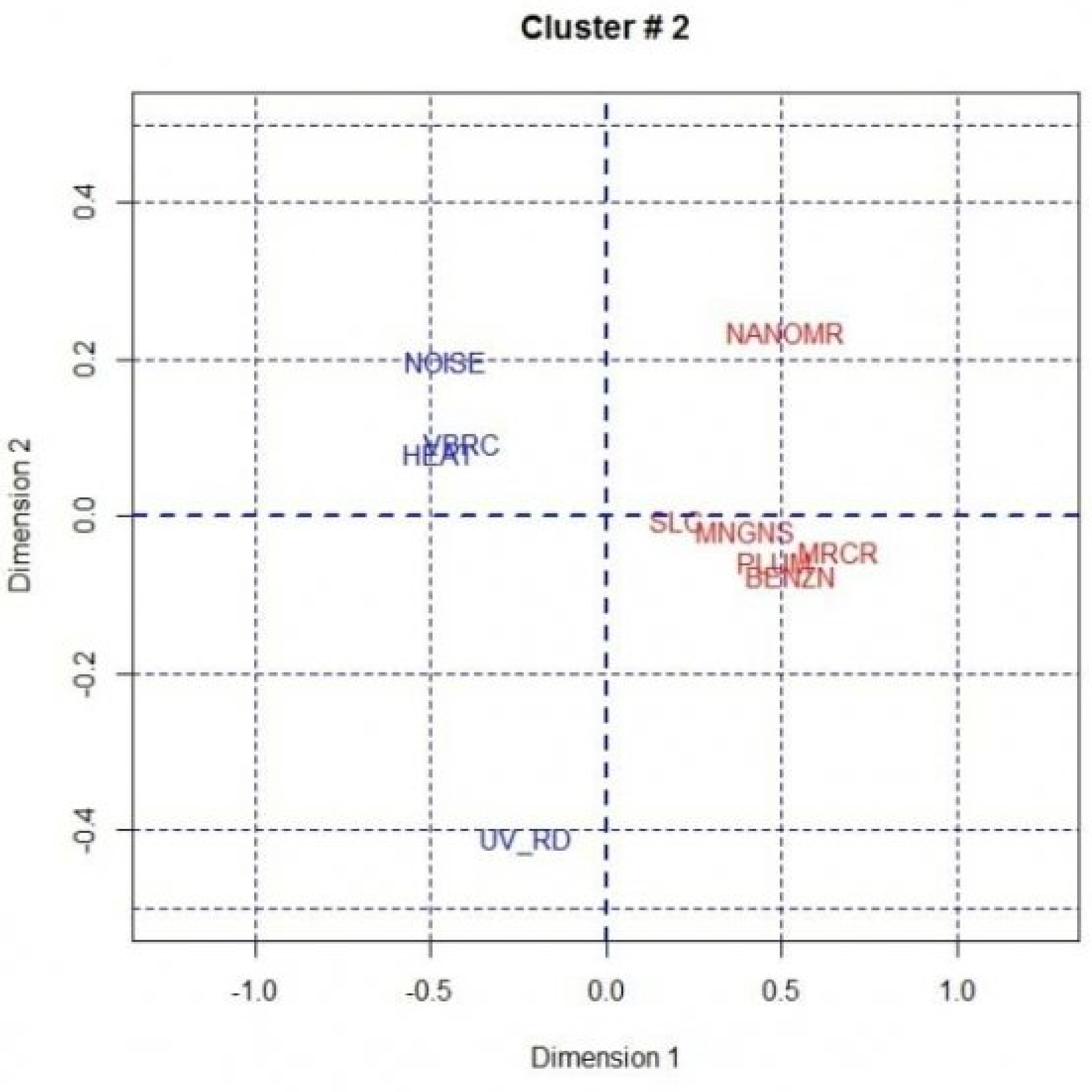


Figure 2:



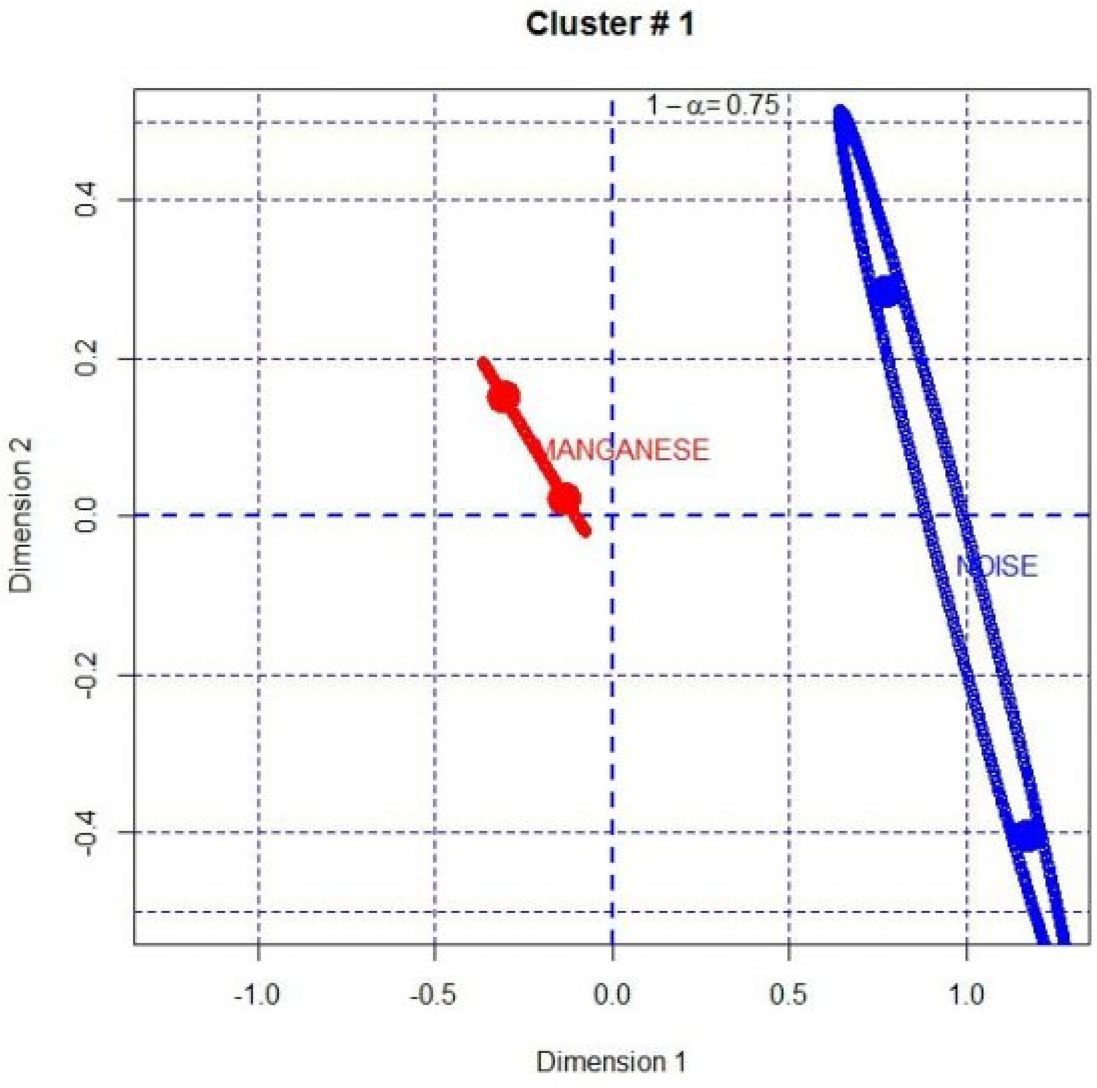
14

Figure 3: Figure 1 : 4 J



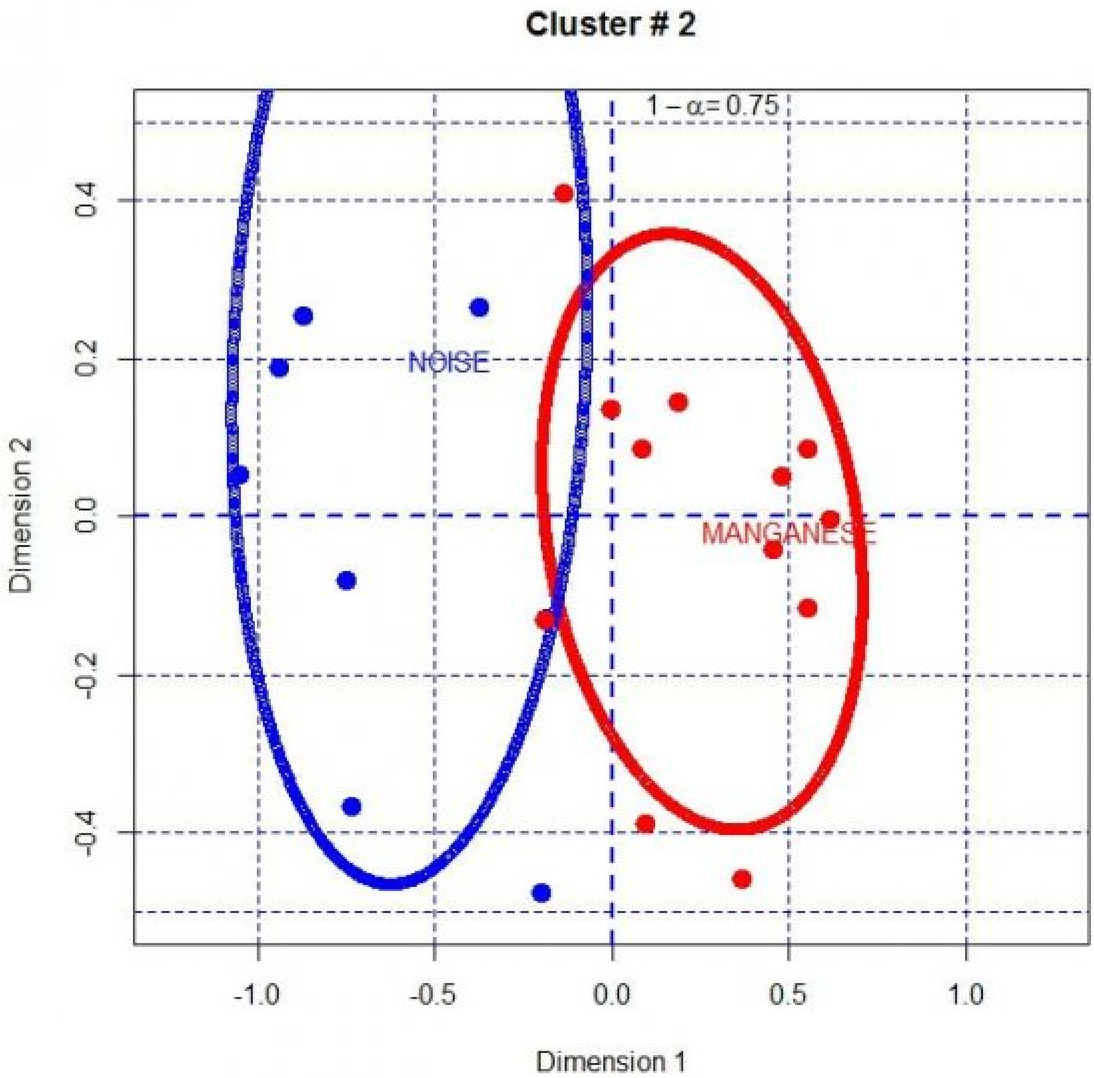
213

Figure 4: Figure 2 : 1 Figure 3 :



4

Figure 5: Figure 4 :



5

Figure 6: Figure 5

*[Note: * = ?? ?? ? ??(?? * ?? ?? * ? ?? ?? ??). Make ?? ?? * * = ?? ?? * . ? If the change in residuals ?? ?? ? ?? ?? * * ? ?? , adjust ?? ?? * * = ?? ?? * , and go back to 4.]*

Figure 7:

1

XIV Issue III Version I
( ) Volume
of Researches in Engineering
Journal

Figure 8: Table 1 :

2

Dimensions	Scale	
Willingness to risk.	Voluntary Involuntary	
People "take" this risk voluntarily	1 2 3 4 5 6 7	
	Immediate	Late
	1 2 3 4 5 6 7	
	Known	Not Known
	1 2 3 4 5 6 7	
Knowledge of Risk. -Science		
To what degree the risk is known to science.		

Figure 9: Table 2 :

3

Clustering Methods/ Metrics	clValid package Optimal # of clusters	hierarchical	Optimal # of clusters	kmeans
Connectivity	2	3.8579	2	3.8579
Dunn	2	1.4407	2	1.4407
Silhouette	2	0.7115	2	0.7115

Figure 10: Table 3 :

XIV Issue III Version I
( ) Volume
of Researches in Engineering
Global Journal

Figure 11:

[Rodrigue ()] *A comparison of the performance of generalized procrustes analysis and the intraclass coefficient of correlation to estimate interrater reliability*, N Rodrigue . 1999. McGill. (Master Thesis)

[Leeuw ()] 'Applications of Convex Analysis to Multidimensional Scaling'. De Leeuw , J . *Recent Developments in Statistics*, (Amsterdam) 1977. North Holland Publishing Company. p. .

[Efron and Tibshirani ()] 'Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy'. B Efron , Tibshirani . *Statistical Science* 1986. 1 (1) p. .

[Everitt et al. ()] *Cluster Analysis. 4*, B S Everitt , S Landau , L Morven . 2001. Arnold.

[Brock ()] 'clvalid, an r package for cluster validation'. G Brock . *Journal of Statistical Software* 2008. 25 (4) .

[Kaufman and Rousseeuw ()] *Finding Groups in Data: An Introduction to Cluster Analysis*, L Kaufman , P J Rousseeuw . 1990. New York: John Wiley & Sons.

[Gower ()] 'Generalized procrustes analysis'. J Gower . *Psychometrika* 1975. 40 p. .

[Moreira ()] *Mapas Perceptuais e Variações na Participação de Mercado*, J A S Moreira . 2006. FGV), Rio de Janeiro. (Master Thesis. Fundação Getúlio Vargas)

[Brombin and Salmaso ()] 'Multi-aspect permutation tests in shape analysis with small sample size'. C Brombin , L Salmaso . *Comput. Stat. Data Anal* 2009. 53 p. .

[Cox and Cox ()] *Multidimensional Scaling*, T F Cox , M A A Cox . 2001. Chapman & Hall/CRC. 2.

[Leeuw and Mair ()] 'Multidimensional scaling using majorization: Smacof in R'. De Leeuw , J Mair , P . *Journal of Statistical Software* 2009. 31 p. .

[Kruskal ()] 'Nonmetric multidimensional scaling: a numerical method'. J Kruskal . *Psychometrika* 1964. 29 p. .

[Cardoso-Junior and Scarpel ()] 'Proposal for generation of the three-way perceptual map using non-metric multidimensional scaling with clusters'. M M Cardoso-Junior , R A Scarpel . *Journal of Mathematics and System Science* 2012. 2 p. .

[R: A language and environment for statistical computing. R Foundation for Statistical Computing ()] *R: A language and environment for statistical computing. R Foundation for Statistical Computing*, `http://www.R-project.org/` 2012. Vienna, Austria.

[Cardoso-Junior and Scarpel ()] *Riscos Ocupacionais: Elaboração do mapa perceptual de riscos utilizando o paradigma psicométrico e métodos multivariados*, M M Cardoso-Junior , R A Scarpel . 2010. Baurú, Brazil. p. . XVII Simpósio de Engenharia de Produção

[Dryden ()] *Shapes: Statistical shape analysis*, I Dryden . `http://-CRAN.R-project.org/package=shapes` 2009. (R package version 1.1-3. Available at)

[Dijksterhuis and Gower ()] *The interpretations of generalized Procrustes analysis and allied methods. Oliemans Punter and Partners*, G Dijksterhuis , J Gower . 2010. Urtecht.