MATEMATIKA, 2013, Volume 29, Number 1c, 151-160

An Improved Pseudo Distance Scale for Measuring One-to-many Relationships in Multidimensional Scaling

[1]Norliza Adnan, [2]Norhaiza Ahmad and [3]Maizah Hura Ahmad
[1,2,3]Department of Mathematical Sciences, Faculty of Science, Universiti Teknologi Malaysia, 81310 UTM Skudai, Johor
e-mail: [1]orlyy2liza@gmail.com, [2]norhaiza@utm.my, [3]maizahutm@gmail.com

Abstract In this study, we compared different pseudo distance measures used on a set of rainfall data recorded between the years 1968 and 2003 at nine rain gauge stations. The data is mapped on 2D plots where Euclidean distance and two different units of pseudo-scale distance are applied through Multidimensional Scaling visualization techniques. The results show that a higher unit of Pseudo distance scale gives the smallest STRESS value when compared to the others. This implies that this type of distance measure reduces the mismatch between the distance rank order in the data and the rank order of distances in the ordinations.

Keywords ANOVA-Tukey; Euclidean-distance; Pseudo-distance; Multidimensional Scaling.

2010 Mathematics Subject Classification 62H30, 65F35

1 Introduction

The main goal of data visualization is to communicate information clearly and effectively through graphical means [1]. Multidimensional scaling (MDS) is a statistical technique that visualizes the proximity of points in multidimensional plots [2]. Possible inputs for MDS involve relationships between pairs of objects, which often indicate dissimilarities, or similarities that can be translated by a proximity matrix or a distance matrix. Although many studies have demonstrated that MDS with various distance measurements can best visualize the data, many of these studies relate to one-to-one relationships between the objects.

A typical pairwise distance measure, d_{ij} between two objects, i and j in MDS is the Euclidean distance [3]. However, this type of distance measure lacks in the ability to extract the relation between other objects in the data [4]. The Euclidean distance shows a direct one-to-one relationship between objects and can be demonstrated by considering the pythagorean theorem in equation (1).

$$d_{AB} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \tag{1}$$

Equation (1) shows the distance between points A and B are being measured where the coordinates of each points are (x_1, y_1) and (x_2, y_2), respectively. The Euclidean type of distance value between A and B shows one-to-one inter-point relationship between two objects. This means that d_{ij} elements in a Euclidean distance matrix relates only two objects at one time.

As the main objective of MDS is to produce a data configuration that can show the structure of similarity or dissimilarity among objects, the idea of identifying one-to-many relationship using pseudo-scale approach is a useful alternative in identifying the proximity between objects based on the data.

Previously, a 3-point pseudo distance scale was highlighted in [5] in an attempt to discover multi-point relationship among objects. However, the distance matrix shows the problem of ties in the distance value among most of the objects. The number of ties arises due to the restrictive relationship between objects highlighted at two scale, while the rest of the objects outside this category is classified into scale 3.

This paper aims to refine the visualization of MDS geometrical representation by comparing a 3-point pseudo distance scale measurement with a higher point pseudo distance scale, specifically a 5-point pseudo distance scale aided by the ANOVA-Tukey test approach. The objective is to improve the proximity data matrix that exhibits the relation between more than two objects concurrently.

2 Methodology

2.1 ANOVA-Tukey test

In this study, the pseudo distance scaling measures was applied to a metric type of data unlike that of [6] and [7] which focused on categorical i.e. preference data.

The multiple comparison of means among objects is first investigated using the Tukey's HSD (honestly significant difference). The Tukey's test is a statistical test that is generally used in conjunction with ANOVA to find which means are significantly different. The aim of using Tukey's test in the procedure is to determine the significance difference between objects understudy in order to construct the distance matrix, d_{ij}. In Tukey's method, the difference between any two mean scores is compared against HSD. A mean difference is statistically significant only if it exceeds HSD.

The Tukey's test begins with carrying out ANOVA test to select the appropriate means in order to calculate Tukey's test for each mean comparison. Then, the Tukey's score is checked if it is statistically significant with Tukey's probability or critical value table taking into account appropriate df_{within} and number of treatments. The equation of of HSD value is shown as follows :

$$HSD = q\sqrt{\frac{MS_{\text{withingroup}}}{S}} \tag{2}$$

where, q is the table value at a given level of significance for the total number of group means being compared. MS_{within} is a within-group mean square that is obtained from the analysis of variance and S is the group sample size.

Tukey's HSD is used as a statistical analysis to investigate which means are significantly different from one another. These results are very important in order to develop the distance matrix, d_{ij} as a basis for constructing a pseudo distance matrix.

2.2 Pseudo Scaling Visualization

2.2.1 3-point pseudo distance

Pseudo distance scale was initiated by Kendall's in [7] which focused on categorical based preferences types of data sets. The approach was extended by [8] and applied to policy selection experiment as one of the problems in decision analysis. The scales represent the samples' preferences on the specific policy selection.

Previous work in [5] focuses on 3-point pseudo distance scale to set up the distance matrix, d_{ij}. Given that A, B and C represent the objects, the distance measures is as follows :

$\delta_{AB} = 1$: if A ≈ B

$\delta_{AB} = 2$: if A ≈ C and C ≈ B with A ≈ B

$\delta_{AB} = 3$: for all other cases

where $\delta_{AB} = 1$ indicates that A and B are similar to each other, $\delta_{AB} = 2$ indicates that A is similar to B, A is similar to C and C is similar to B, and $\delta_{AB} = 3$ indicates the conditions for all other cases.

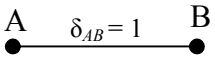
Here, as the relationship between objects is restricted only to two major cases, while the others are considered to be in all other cases, the distance value may lead to the problem of tie because most of the objects share the same scale. This issue is described further in the numerical analysis in Table 6.

2.2.2 5-point pseudo distance

A pseudo distance can be represented as *n*-point pseudo-scale that consists of a penalty score which is based on the Kendall's Primary (PTT), Secondary (STT) and Tertiary (TTT) treatment of ties [7]. In this study, the pseudo distance scaling measures was applied to a metric type of data by considering the penalty score based on the specific treatment of ties.

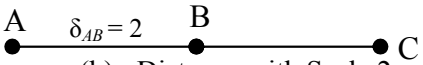
For the purpose of the current study, we extended our 3-point pseudo scale in [5] as stated in section 2.2.1 into a 5-point based on the 5-types of inter-point relationships that may exist as shown in Figure 1 as follows :

$\delta_{AB} = 1$: if $A \approx B$



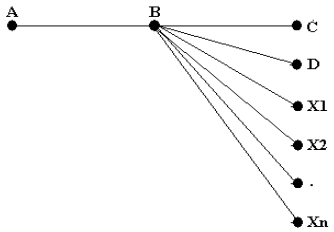
(a) : Distance with Scale 1

$\delta_{AB} = 2$: if $A \approx B$ and $B \approx C$ with $A \neq C$



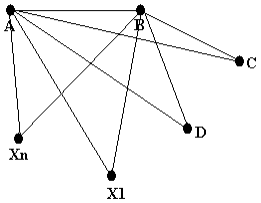
(b) : Distance with Scale 2

$\delta_{AB} = 3$: if $A \approx B$ and $B \approx C$, $B \approx D$, $B \approx X_1$, $B \approx X_2$, ..., $B \approx X_n$ with $A \neq C \neq D \neq X_1 \neq X_2 ... \neq X_n$



(c) : Distance with Scale 3

$\delta_{AB} = 4$: if $A \approx B \approx C \approx X_1 \approx X_2 \approx ... \approx X_n$



(d) : Distance with Scale 4

$\delta_{AB} = 5$: for all other cases

Figure 1 Possible Inter-point Relationships

The determination of the relationships in Figure 1 is based on the results from the Tukey's test analysis. The sign ≈ shows similarity between objects.

Scale 1 shows one-to-one inter-point relationship which refers to only two objects having a relationship at one time and shows no similarity pattern with any other objects. Scale 2 refers to relationships in terms of similarity involving 3 objects where B being the middle object is connected to the other two objects that have no relations at all with each other.

Scale 3 represents the same idea of relationships as scale 2 but the point B in the middle shows the similarity with more than one object that have no relations at all with the first object. Scale 4 shows the relationships that may exist among few objects that correlate to each other or in other words, the inter-point relationships that are also similar to more than one object. Other cases are categorized in scale 5.

The 3-point pseudo scale in section 2.2.1 consist of scale 1, scale 2 and all other cases for scale 3. Since the type of relationships is limited when using 3-scale, the multi-point distances showed ties when most of them have a value between 2 and 3 only. Therefore, as we add on additional scales, the relationships between objects is further refined.

2.2 Assessing the reliability of MDS

The square root of a normalized residual sum of squares (STRESS) is used as a reliability measurement to identify the deviation from monotonicity between distances and the observed similarities. The STRESS evaluation measures the mismatch between the rank order of distances in the data and in the ordinations. The configuration approaches a perfect fit to the observed similarity when the STRESS value is minimized. In other words, STRESS evaluation aims the best configuration as the value of STRESS approaches zero.

The STRESS evaluation is defined by equation (3)

$$S = \sqrt{\frac{\sum_{i \neq j} \left(d_{ij} - \hat{d}_{ij} \right)^2}{\sum_{i \neq j} d_{ij}^{\,2}}} \tag{3}$$

The interpretation of the STRESS measure suggested by [9] is shown in Table 1.

Table 1 The 'Goodness of fit' for STRESS value

Stress (in %)	Goodness of fit
20 or 0.2	Poor
10% or 0.1	Fair
5% or 0.05	Good
2.5% or 0.025	Excellent
0	Perfect

Based on Table 1, the value 0.20 means 80% of the variance of the $\hat{d}_{ij}$ is explained by the distances. Therefore, it is considered as poor fit compared to other STRESS values such as 0.1, 0.05, 0.025 and 0.

3 Numerical Results

To illustrate the technique described in this paper, we use 3 and 5 point pseudo distance scales on the Malaysian rainfall data sets and the graphical results and compare that with the MDS approach using Euclidean distance measurement.

The 36 years durations from 1968 to 2003 of rainfall data over nine rain-gauge stations located in Peninsular Malaysia is used to illustrate the proposed distance measurement as discussed in the previous sections. The amount of rainfall consists of daily amount in *mm* unit for four regions in Peninsular Malaysia. The stations are indicated in Table 2 and the notations are used as illustration purposes in the MDS 2D plots:

Table 2 Notations of the stations

Notations	Stations	Height Above
AOS	AlorStar	3.9 *mtr*
BYN	BayanLepas	2.8 *mtr*
IPH	Ipoh	40.1 *mtr*
MLK	Malacca	8.5 *mtr*
SWN	Setiawan	7.0 *mtr*
SBG	Subang	16.5 *mtr*
KBH	Kota Bharu	4.6 *mtr*
KTN	Kuantan	15.3 *mtr*
MSG	Mersing	43.6 *mtr*

The analysis starts by setting all the stations as objects. Then, the distance matrix, d_{ij} is developed based on the practical algorithm of MDS as mentioned in the previous sections and then the calculation procedures are applied for finding the coordinates.

Results for the Tukey's score (sect. 2.1) for each pairs of objects can be seen in Table 3. From the Tukey's test result, the significant difference between the objects can be determined. The purpose of using this method is to identify which pairs have no significant difference between objects. Then, the distance matrix, d_{ij}, is set up prior to the coordinates of the configuration to be calculated.

Table 3 Score for Tukey's test significant difference

	AOS	BYN	IPH	MLK	SWN	SBG	KBH	KTN	MSG
AOS	-								
BYN	1.32	-							
IPH	0.00	0.01	-						
MLK	0.01	0.04	0.88	-					
SWN	0.00	0.00	0.11	0.22	-				
SBG	0.03	0.02	1.05	2.33	0.34	-			
KBH	0.00	0.02	0.00	0.00	0.00	0.03	-		
KTN	0.00	0.01	0.01	0.03	0.00	0.00	0.23	-	
MSG	0.00	0.75	0.00	1.37	1.11	0.00	0.00	0.00	-

Table 4 shows the significant difference among rain-gauge stations for rainfall data based on the results from Tukey's test. The lower triangle matrix shows that the a_{ij} and a_{ji} values have no difference in meaning because it carries the same values. The similarity among the stations is denoted by the symbol '≈' which means, no significant difference between stations. The Tukey's test results show the significance difference at $p < 0.05$ value where the station is dissimilar to each other.

Table 4 Results for Tukey's test significant difference

	AOS	BYN	IPH	MLK	SWN	SBG	KBH	KTN	MSG
AOS	-								
BYN	≈	-							
IPH			-						
MLK			≈	-					
SWN			≈	≈	-				
SBG			≈	≈	≈	-			
KBH							-		
KTN							≈	-	
MSG		≈		≈	≈				-

There are eleven pairs with the "≈" sign indicating that there is no difference (as in Table 4) based on the Tukey's hypotheses testing. For example, the rain-gauge station in MERSING shows no significant difference with BAYAN LEPAS, MALACCA and SITIAWAN. There is no judgement made to itself and this is denoted by the "-" entries.

The values for the upper triangle are similar to that for the lower-triangle matrix. Based on the results shown in Table 3, the stations that have significant difference are reflected in the form of lower-triangle matrix with the same pairs and have no sign at all with each other.

Table 5 shows the use of pseudo-scale as the measurements for the similarity matrix based on the results in Table 4. The level of distance scale amongst stations values are shown in the entries. In contrast with the Euclidean distance as distance measurement, the level of scales in the entries shows the similarity amongst stations and no judgement made to itself is indicated by the 0 entries in the matrix, which corresponds to no meaning at all for the d_{ii}. The scales between stations in columns as reflected in rows are done based on the 5-pseudo distance scales stated in Figure 1.

Table 5 5-point pseudo distance scale in the similarity matrix, d_{ij}.

$$d_{ij} = \begin{bmatrix} 0 & & & & & & & & \\ 2 & 0 & & & & & & & \\ 5 & 5 & 0 & & & & & & \\ 5 & 5 & 4 & 0 & & & & & \\ 5 & 5 & 4 & 4 & 0 & & & & \\ 5 & 5 & 4 & 4 & 4 & 0 & & & \\ 5 & 5 & 5 & 5 & 5 & 5 & 0 & & \\ 5 & 5 & 5 & 5 & 5 & 5 & 1 & 0 & \\ 5 & 3 & 5 & 3 & 3 & 5 & 5 & 5 & 0 \end{bmatrix}$$

The distance values stated in Table 5 show the 5-point pseudo distance measurements among objects. The values in the entries shows the level of scales of the distances amongst the stations. In contrast to the distance measurement using Euclidean distance, the entries are the levels of scales that show similarity amongst stations. Therefore, scale 0 corresponds to no meaning.

To illustrate, for example scale 2 in coordinate $i = 2$ and $j = 1$, indicates the distance between ALOR SETAR and BAYAN LEPAS where the relationship between both stations follows the rules of scale 2, that is, A ≈ B and B ≈ C with A ≠ C. This means that, ALOR SETAR is similar to BAYAN LEPAS, while at the same time BAYAN LEPAS is similar to MERSING, but, MERSING is dissimilar to ALOR SETAR.

However, the distance matrix shows most of the pairs have scale 5 as distance values because of only five cases are being considered as scales that have been discussed in the previous section.

Table 6 3-point pseudo distance scale in the similarity matrix, d_{ij}.

$$d_{ij} = \begin{bmatrix} 0 & & & & & & & & \\ 2 & 0 & & & & & & & \\ 3 & 3 & 0 & & & & & & \\ 3 & 3 & 2 & 0 & & & & & \\ 3 & 3 & 2 & 2 & 0 & & & & \\ 3 & 3 & 2 & 2 & 2 & 0 & & & \\ 3 & 3 & 3 & 3 & 3 & 3 & 0 & & \\ 3 & 3 & 3 & 3 & 3 & 3 & 1 & 0 & \\ 3 & 2 & 3 & 2 & 2 & 3 & 3 & 3 & 0 \end{bmatrix}$$

Table 6 shows the similarity matrix based on the 3-point pseudo distance scale as the measurements. Scale 1 corresponds to "similar" (for example, KOTA BHARU rain-gauge station is similar to KUANTAN rain-gauge station) as defined in the pseudo-scales. Scale 2 corresponds to for example, ALOR SETAR rain-gauge station is similar to BAYAN LEPAS rain-gauge station, at the same time, BAYAN LEPAS rain-gauge station is similar to MERSING rain-gauge station. Scale 3 shows other cases which are not considered in scale 1 and scale 2. However, since only three cases are considered in the scales, the distance measurement shows a tie, especially for scale 2. The ties in Table 6 occur because the number of possible types of relationship among objects was very limited.

The similarity matrix, d_{ij} for Euclidean distance is shown in Table 7.

Table 7 The similarity matrix, d_{ij} using Euclidean-distance

$$d_{ij} = \begin{bmatrix} 0 & & & & & & & & \\ 1977.138 & 0 & & & & & & & \\ 2008.230 & 2209.841 & 0 & & & & & & \\ 2870.478 & 3075.857 & 2961.379 & 0 & & & & & \\ 1977.628 & 2217.737 & 2057.046 & 2907.457 & 0 & & & & \\ 2848.449 & 2975.806 & 2884.959 & 3363.941 & 2821.691 & 0 & & & \\ 2727.521 & 2891.553 & 2755.077 & 3327.927 & 2653.590 & 2783.549 & 0 & & \\ 1881.863 & 2019.229 & 1865.443 & 2822.736 & 1900.715 & 2696.037 & 2568.321 & 0 & \\ 2076.899 & 2278.759 & 2032.612 & 2958.499 & 1983.051 & 2854.808 & 2711.109 & 1913.074 & 0 \end{bmatrix}$$

The similarity matrix in Table 7 shows the distance measurement based on Euclidean distance. The values in the entries show the distance measures amongst the stations. For example, the distance between ALOR SETAR and MERSING is 2076.899 units while the distance between ALOR SETAR and KUANTAN is 1881.863 units. These mean that KUANTAN is similar to ALOR SETAR as compared to MERSING.

The results in Table 5 and Table 6 are used to construct a similarity matrix using the proposed pseudo-distance as shown in Table 7.

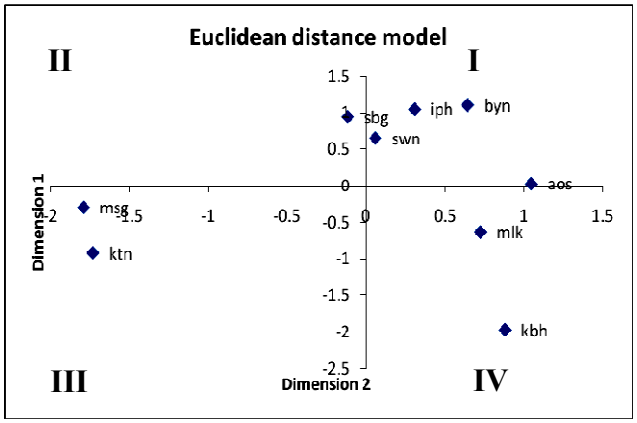


Figure 2 Euclidean distance in MDS plot.

Figure 2 shows the visual plot of Euclidean distance in MDS algorithm. By considering the quadrant plot, SITIAWAN, IPOH, BAYAN LEPAS and ALOR SETAR's stations are embedded together in the very first quadrant, where the only station plotted in the second quadrant is SUBANG. Both MERSING and KUANTAN lie in the third quadrant and only MALACCA and KOTA BHARU are plotted in the fourth quadrant.

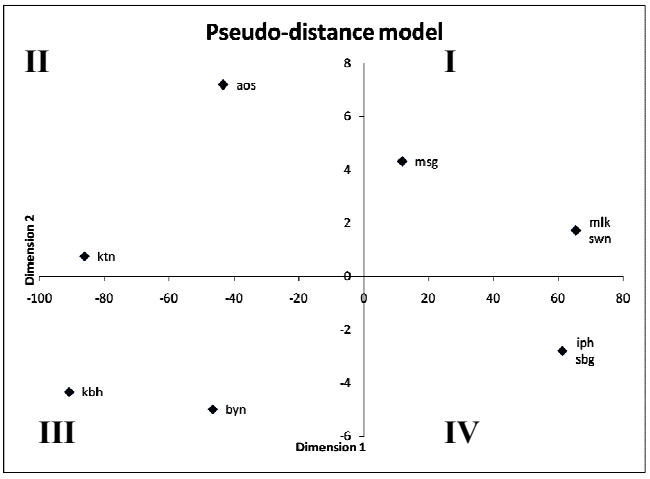


Figure 3 3-point pseudo scale in MDS plot.

Figure 3 shows that MERSING, MALACCA and SITIAWAN are in the first quadrant. ALOR SETAR and KUANTAN are in the second quadrant. While KOTA BHARU and BAYAN LEPAS are in the third quadrant, IPOH and SUBANG are in the last quadrant.

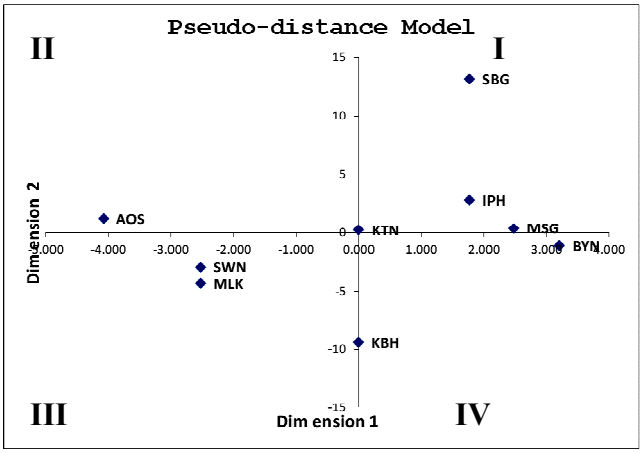


Figure 4 5-point pseudo scale in MDS plot.

In contrast to Figure 2 and Figure 3, the visual plot of MDS using 5-point pseudo scales in Figure 4 shows that SUBANG, IPOH and MERSING are in the first quadrant, followed by both ALOR SETAR and KUANTAN in the second quadrant, SITIAWAN, MALACCA and KOTA BHARU in third quadrant while BAYAN LEPAS is the only station in the fourth quadrant.

The 3-point pseudo distance plot in Figure 3 shows an overlapping point between MALACCA and SITIAWAN stations and also IPOH and SUBANG which lie in the first and fourth quadrant respectively. In contrast to 5-point pseudo distance scale, the plot in Figure 4 shows that MALACCA and SITIAWAN are plotted separately in third quadrant likewise IPOH and SUBANG in the first quadrant.

Table 8 STRESS values

MDS Approach	STRESS values
Euclidean distance	0.3287
3-point pseudo scale	0.1826
5-point pseudo scale	0.1785

The STRESS evaluation is used to validate the results. The STRESS value in Table 8 shows that MDS using Euclidean distance is 0.3287 while the MDS using 3-point pseudo distance is 0.1826 and 0.1785 for 5-point pseudo distance scale. The smallest STRESS value showed by 5-point pseudo scale implies that 82.15% of the variance in $\hat{d}_{ij}$ was explained by the distances. In other word, the mismatch between the distance rank order in the data and the rank order of distances in the ordinations is smaller compared to 3-point pseudo distance and Euclidean distance approach.

4 Conclusions

The effects of using indirect distance measurement to identify one-to-many relationships in MDS technique have been studied in the numerical analysis. The analysis compared three different distance measures namely Euclidean distance, 3-point pseudo distance scale and 5-point pseudo distance scale. We have found that

i) The STRESS evaluation of a higher point pseudo scale showed a smaller STRESS values when compared to a lower point pseudo distance scale and Euclidean distance. This implies a reduction of mismatch between the distance rank order in the data and the rank order of distances in the

ordinations.

ii) The higher pseudo distance scale is able to :
   a. further refine the ordination of the MDS configuration in terms of overlapping objects plotted using lower-point pseudo scale.
   b. overcome the problem of ties in the 3-point pseudo scale.

Adding to existing literature, the present study shows a higher pseudo based distance is a tool to determine one-to-many relationships among objects and could refine the ordination of metric MDS configuration.

Acknowledgement

Universiti Teknologi Malaysia is gratefully acknowledged for providing the financial support under FRGS Vote 78487.

References

[1] Friedman, V. Data Visualization and Infographics. In: *Graphics*, Monday Inspiration, January 14th, 2008.
[2] Torgerson, W. S. Multidimensional scaling of similarity. *Psychometrika*. 1965. 30: 379-393.
[3] DeJordy, R., Borgatti, S.P., Roussin, C. and Halgin, D.S. Visualizing proximity data. *Field Methods*. 2007. 19(3): 239-263.
[4] Dattorro, J. *Convex Optimization and Euclidean Distance Geometry*. United State of America: MEO Publishing. 2005.
[5] Adnan, N., Ahmad, N. and Ahmad, M. A 3-point pseudo scale distance measure for measuring indirect proximity. *Applied Mathematical Sciences*. 2003. **7** (25): 1239 – 1247.
[6] Rivett, B.H.P. Policy selection by structural mapping. *Proc. Royal Society of London*. 1977. A354: 407-423.
[7] Kendall, D.G. The recovery of structure from fragmentary information. *Philos. Trans. Royal Society of London.*1975. A279: 547-582.
[8] Roy G.G. The use of muti-dimensional scaling in policy selection. *Journal Optimization Research Society.*1982. 33: 239-245.
[9] Kruskal, J. and Wish, M. *Multidimensional Scaling*. Beverly Hills : SAGE. 1978.