ISSN 1392 – 124X INFORMATION TECHNOLOGY AND CONTROL, 2006, Vol. 35, No. 1

MULTIDIMENSIONAL DATA PROJECTION ALGORITHMS SAVING CALCULATIONS OF DISTANCES

Rasa Karbauskaitė, Gintautas Dzemyda

Institute of Mathematics and Informatics
Akademijos St. 4, 08663 Vilnius, Lithuania

Abstract. In this paper, the triangulation method, the classic algorithm for Sammon's projection and the combination of Sammon's algorithm with the triangulation method are examined for mapping new points in detail. A new realization of the combination of Sammon's algorithm and the triangulation method is proposed. These algorithms are analyzed and compared in the following respects: visual evaluation of data projection, evaluation of data mapping time and evaluation of projection error.

Keywords: triangulation; Sammon's projection; visualization; mapping.

1. Introduction

Data perception is frequently a complex problem, especially when data point to a complicated phenomenon described by many parameters, i.e., multidimensional data are analyzed. In order to better perceive multidimensional data, to establish their interrelations, and the groups (clusters) formed, we often have to visualize them. A human being is capable to perceive visual information much faster than textual.

Visualization methods of multidimensional data can be partitioned into several groups: direct visualization methods, projection methods, clustering methods and artificial neuron networks. Projection methods are frequently used the aim of which is to present multidimensional data in a space of smaller dimension so as to preserve the data structure analyzed as precisely as possible. There are the linear projection methods (Principal Component Analysis (PCA) [11], Projection Pursuit [4] and so on) and nonlinear projection methods (Multidimensional Scaling (MDS): Sammon's algorithm, SMACOF [3], etc.; Principal Curves [6]; the Triangulation method [8] etc.).

This paper deals with two nonlinear projection methods of multidimensional data: Sammon's projection and the triangulation method as well as their combination. The main differences between these methods are as follows: Sammon's method is a method of simultaneous mapping where all of the distances are tried to preserve relatively, while the triangulation is a method of sequential mapping. Here a new point is mapped and its distances to two points previously mapped are exactly preserved.

A new realization of the combination of Sammon and triangulation methods has been proposed which enables us to map multidimensional points rather precisely and fast without Sammon mapping of the whole (updated) data set.

2. The triangulation method

Consider three high-dimensional points P_i^*, P_j^*, and P_k^*. Suppose on the two-space, P_i and P_j exactly preserve the distance between P_i^* and P_j^*. That is, $d_{ij} = d_{ij}^*$. Then the third point P_k^* can be mapped to a point P_k in the two-space such that the distances among P_i^*, P_j^*, and P_k^* are all exactly preserved. This can be done by drawing two circles with P_i and P_j as centers and d_{ik}^* and d_{jk}^* as radii (Figure 1).

Note that because of the triangle inequality, the circles either intersect at two points, or are tangent. Let P_k^1 and P_k^2 be two possible locations of P_k on the two-space. For every possible location, we can now find out its nearest neighbor (excluding P_i and P_j) on the map already obtained. Let us denote the next nearest neighbor of P_k^1 and P_k^2 by Q_1 and Q_2, respectively. We cannot preserve the original distances between P_k and Q_1 and between P_k and Q_2 exactly. However, we now calculate the error to Q_1 caused by putting P_k at location P_k^1 and the error to Q_2 caused

by putting P_k at P_k^2. We then put P_k at the location where the smaller error will be caused. On the resulting map, for every point P_k, there exist two points P_i and P_j such that the distances among P_i, P_j, and P_k are all exactly preserved.

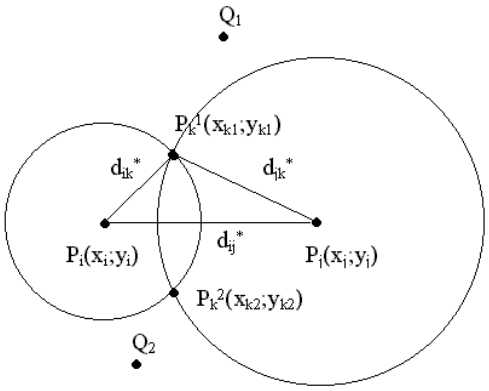


Figure 1. Mapping of a point on the plain

The mapping is based on the distances of a minimal spanning tree constructed from the points. All of the distances on the MST are exactly preserved. If we have m points, then only $m-1$ distances will be preserved. However, the triangulation method can preserve $2m-3$ distances, therefore we must provide additional information. To this end, two methods are used: the second nearest neighbor approach and the reference point approach. When a point is being mapped to a plane, we always use two points P_i and P_j as reference ones. When applying the second nearest neighbor approach, P_j denotes the first closest neighbor of point P_k mapped, and P_i stands for the second closest neighbor of point P_k. In the case of the reference point approach, P_i is the selected reference point, and P_j is the first closest neighbor. The reference point approach tends to preserve global information, while the second nearest neighbor approach tries to reproduce local information from the original set of points [8].

To realize the triangulation method, the following algorithm was used in this article:

1. Based on Prim's algorithm [9], a MST is built from a given set of points.
2. Any point is chosen as the root of the MST, i.e., a directed tree is formed.
3. Ordering of points being mapped is maintained. The breadth first approach is used for tree searching in the realization.
4. The analyzed points are being mapped.

3. Sammon's algorithm

The Sammon projection (algorithm, method) [10] is a nonlinear mapping method of multidimensional objects on a lower measurable space. This is one of the best methods of the multidimensional scaling group (MDS) [1]. Let us consider a case where the dimension of a projection space, onto which *n*-dimensional vectors are being mapped, is 2, i.e., we map on a plane.

Suppose we have multidimensional vectors $\boldsymbol{X}^i = (x_1^i, x_2^i, ..., x_n^i)$, $i = 1,...,m$, belonging to space R^n. The problem to be solved is to map these *n*-dimensional vectors $\boldsymbol{X}^i = (x_1^i, x_2^i, ..., x_n^i)$, $i = 1,...,m$ on the plane R^2. Two–dimensional vectors $\boldsymbol{Y}^1, \boldsymbol{Y}^2, ..., \boldsymbol{Y}^m \in R^2$ will correspond to them. Here $\boldsymbol{Y}^i = (y_1^i, y_2^i)$, $i = 1,...,m$. Let us denote by d_{ij}^* the distance between the multidimensional vectors $\boldsymbol{X}^i$ and $\boldsymbol{X}^j$, and d_{ij} – the distance between the two-dimensional vectors $\boldsymbol{Y}^i$ and $\boldsymbol{Y}^j$ corresponding to the vectors $\boldsymbol{X}^i$ and $\boldsymbol{X}^j$ ($i, j = 1,..,m$). The Sammon's algorithm minimizes the distortion (error) of projection E_S:

$$E_S = \frac{1}{\sum\limits_{\substack{i,j=1 \\ i<j}}^{m} d_{ij}^*} \sum_{\substack{i,j=1 \\ i<j}}^{m} \frac{(d_{ij}^* - d_{ij})^2}{d_{ij}^*} \qquad (1)$$

Sammon's error E_S is a measure that shows how exactly the distances among vectors are preserved in the transition from a higher dimensional space to a lower one. The main task is to minimize the function E_S of this error. Various optimization methods can be used for this purpose.

4. Combination of the triangulation method with sammon's projection

In [2], the authors suggest joining the triangulation method with Sammon's algorithm. The triangulation method is fast, but only $(2m-3)$ distances among the points analyzed can be preserved by means of it. Sammon's algorithm tries to preserve all the distances $\frac{m(m-1)}{2}$ among the points, however it is rather slow. That is why it is worth using a combination of both these methods with a view to accelerate computations, though losing certain exactness. The Sammon-triangulation algorithm creates a frame from the data by first projecting $\hat{m}$ of the m points ($\hat{m} < m$) onto a plane using Sammon's algorithm. The $\hat{m}$ projected points are then fixed in the plane (they are called basic points) and the remaining $(m-\hat{m})$ points are projected

sequentially, using the triangulation method, so that the distances from each point not in the frame to two points in the frame are exactly preserved. The two points selected from the frame are the two nearest neighbors of that particular point. The triangle inequality may not be maintained through Sammon's method for all points in the frame. This causes difficulties because the triangulation method relies on the triangle inequality. Biswas, Jain, Dubes (1981) overcome this problem by checking the triangle inequality for every point projected, and when not satisfied, they successively pick the third nearest neighbor, fourth nearest neighbor, and so on from the frame until the inequality is satisfied. The question arises what number $\hat{m}$ should be. The point classes being known, several representatives from each class have to be basic points. In the opposite case, the data are clustered before analyzing and their centers are assumed as basic points. In [2] it has been shown that, as a result of this combination, this method is rather fast, in comparison with other projection methods, and the accuracy is lost but insignificantly.

4.1. Realization of the combination of Sammon's and triangulation methods

A new realization of the combination of Sammon's mapping and the triangulation method is proposed in this paper. In difference to [2], in our realization of the combination of Sammon's and the triangulation methods, we do not require for satisfication of the triangle inequality when we perform the sequential mapping. However, we put a new point in the close area of its two nearest neighbors. We provide the details of such combination and sequential mapping below.

Sammon's algorithm tries to preserve all the relative distances among points, however there appears a certain error. Therefore, to map a new point by means of the triangulation method, three cases occur: the circles are tangent, the circles are intersecting, and they are neither tangent nor intersecting. The third case is most probable and it is most complicated when mapping new points. A new way of solution is proposed in the paper.

Suppose we have to map point P_k^* to $P_k(x_k; y_k)$ on the plain. So we find two nearest neighbors of this point, for instance P_1^* and P_2^*, in an *n*-dimensional space. In the plain of their projection there will be points $P_1(x_1; y_1)$ and $P_2(x_2; y_2)$. Let us denote: $d(P_1, P_2) = d$; $d(P_k^*, P_1^*) = r_1$; $d(P_k^*; P_2^*) = r_2$.

Let us draw two circles with $P_1(x_1; y_1)$ and $P_2(x_2; y_2)$ as centers and r_1 and r_2 as radii. The circles are neither tangent nor intersecting if:

a) $r_1 > d + r_2$; b) $r_2 > d + r_1$; c) $d > r_1 + r_2$.

We analyze case a). Let $r_1 > d + r_2$. The situations may be as those shown in Figure 2. Let us derive formulas to find the coordinates $(x_k; y_k)$ of the point P_k. Denote:

$$\hat{d} = \frac{r_1 - d - r_2}{2}.$$

The equation of a straight line leading through two points $P_1(x_1; y_1)$ and $P_2(x_2; y_2)$ (Figure 2a) is:

$$\frac{x - x_1}{x_2 - x_1} = \frac{y - y_1}{y_2 - y_1}. \quad (2)$$

By rewriting this equation, we obtain:

$$y = y_1 + \frac{(x - x_1)(y_2 - y_1)}{x_2 - x_1},\ x_2 \neq x_1. \quad (3)$$

$\Delta P_1 A P_2 \sim \Delta P_1 B P_k$. Therefore their respective segments are proportional:

$$\frac{x_2 - x_1}{d} = \frac{x_k - x_1}{r_1 - \hat{d}}. \quad (4)$$

By rewriting formula (4), we have:

$$x_k = x_1 + \frac{(x_2 - x_1)(r_1 - \hat{d})}{d} \quad (5)$$

By substituting (5) into formula (3), we get point $P_k(x_k; y_k)$ of the plain.

Suppose there exists a situation as shown in Figures (2b, 2c). Then the coordinates of point $P_k(x_k; y_k)$ are found from the formulas:

$$x_k = x_1,\ y_k = y_2 - r_2 - \hat{d} \text{ (Figure 2b);} \quad (6)$$

$$x_k = x_1,\ y_k = y_2 + r_2 + \hat{d} \text{ (Figure 2c).} \quad (7)$$

Sometimes it is possible to obtain concentric circles (Figure 2d). Then we find the third nearest neighbor P_3^* of point P_k^* in an *n*-dimensional space (P_1^* and P_2^* are the first and the second nearest neighbors, respectively). Its projection on a plane will be point $P_3(x_3; y_3)$. We draw a straight line $P_1 P_3$. Its equation will be:

$$\frac{x - x_1}{x_3 - x_1} = \frac{y - y_1}{y_3 - y_1}. \quad (8)$$

By rewriting this equation, we obtain:

$$x = x_1 + \frac{(x_3 - x_1)(y - y_1)}{y_3 - y_1},\ y_3 \neq y_1. \quad (9)$$

Afterwards we draw a circle with the center $P_1(x_1; y_1)$ and a radius $r = r_2 + \frac{r_1 - r_2}{2}$. The equation of the circle will be:

$$(x - x_1)^2 + (y - y_1)^2 = r^2. \quad (10)$$

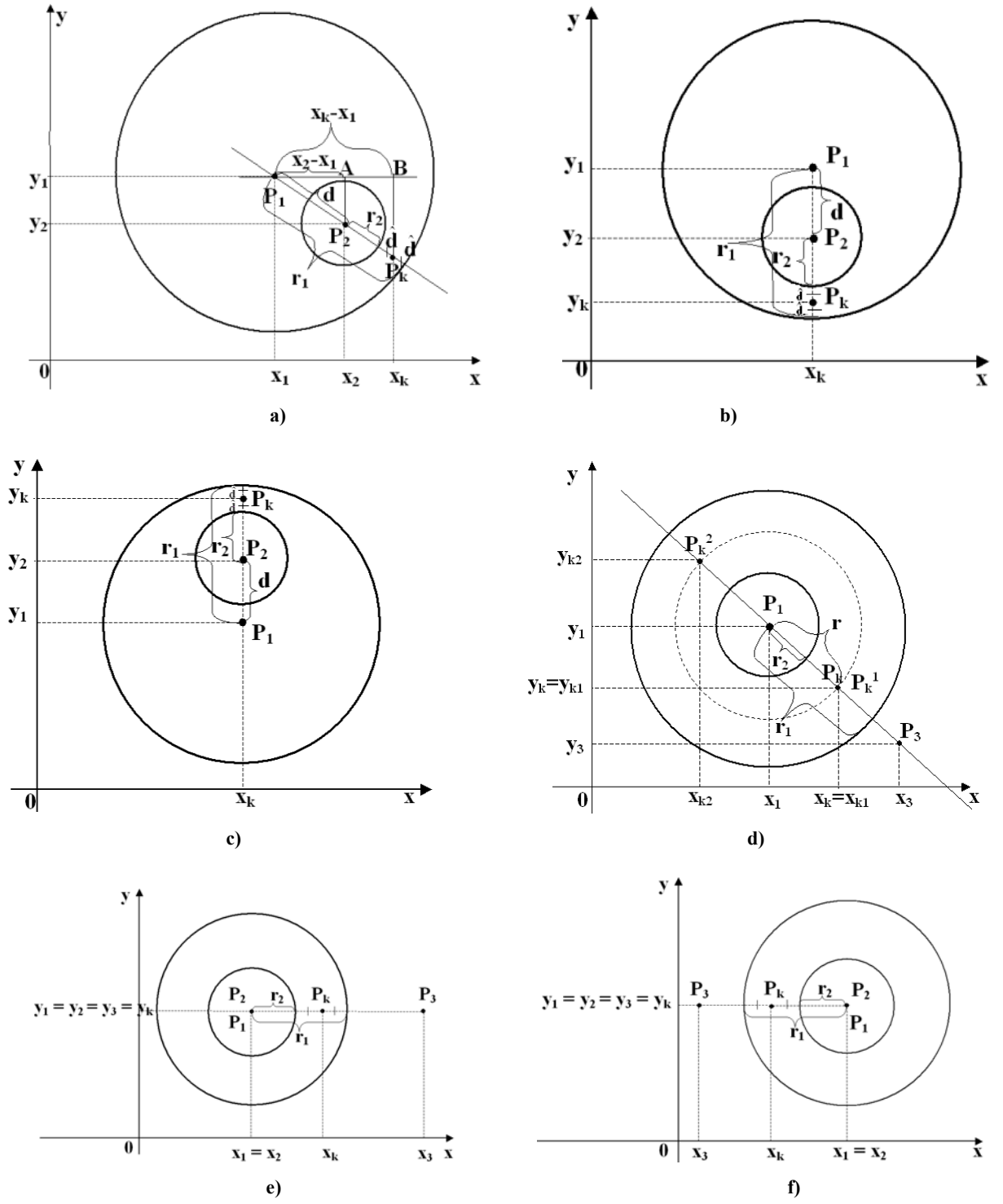


Figure 2. Mapping of a point in case the circles being one inside the other

The point of intersection of the straight line with the circle will be just the point in quest $P_k(x_k; y_k)$. The ordinate of this point is found by the formula:

$$y = y_1 \pm \sqrt{\frac{r^2 (y_3 - y_1)^2}{(y_3 - y_1)^2 + (x_3 - x_1)^2}}. \qquad (11)$$

By substituting (11) into formula (9), we obtain two possible locations $P_k^1(x_{k1}; y_{k1})$ and $P_k^2(x_{k2}; y_{k2})$ of point $P_k(x_k; y_k)$. We determine which of the two points obtained suits better: we compute the distances from these points to the projection of the third closest neighbor, i.e., $d\left(P_k^{\,1}, P_3\right)$ and $d\left(P_k^{\,2}, P_3\right)$, and we choose the point from which the distance to the projection of the third closest neighbor is shorter.

Figures 2e and 2f illustrate a specific case of the situation considered before, where the ordinates of projections of all the three nearest neighbors are equal,

i.e., $y_1 = y_2 = y_3$. In this case, the coordinates of point $P_k(x_k; y_k)$ are found by the formulas:

$$y_k = y_1,$$

$$x_k = x_2 + r_2 + \frac{r_1 - r_2}{2}, \text{ if } x_3 \geq x_1 \text{ (Figure 2e),}$$

$$x_k = x_2 - r_2 - \frac{r_1 - r_2}{2}, \text{ if } x_3 < x_1 \text{ (Figure 2f).}$$

For $r_2 > d + r_1$, we get an algorithm analogous to the case considered. The only difference is that the circle with the center $P_1(x_1; y_1)$ and radius r_1 will be inside of the circle with the center $P_2(x_2; y_2)$ and radius r_2. The search for the coordinates of a new point, for $d > r_1 + r_2$, is based on a similar principle. We can see that graphically in Figure 3.

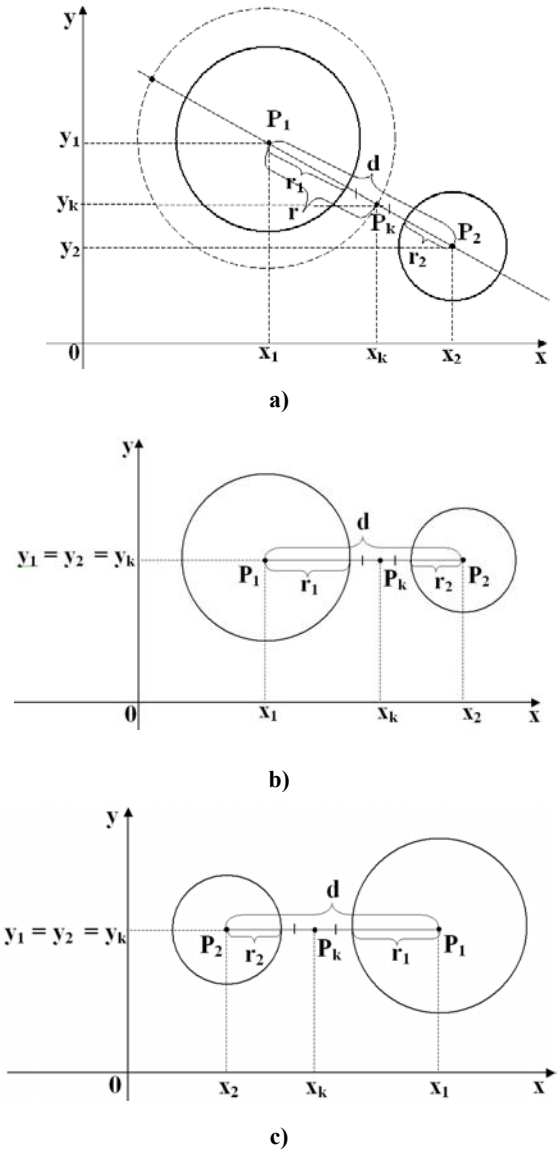


Figure 3. Mapping of a point in case the circles being one outside the other

5. Results of experimental investigation of the triangulation and sammon methods

Test data – 100 10-dimensional points that form 5 clusters – were used for research. An array of 5 normally distributed 10-dimensional vectors was generated. In the vicinity of each of these vectors 20 random 10-dimensional vectors were generated. Thus we have an array of 100 vectors. We choose 50 points – 10 of each class – from the 100 available. These are regarded as the initial points, while the remaining 50 make up a set of new points.

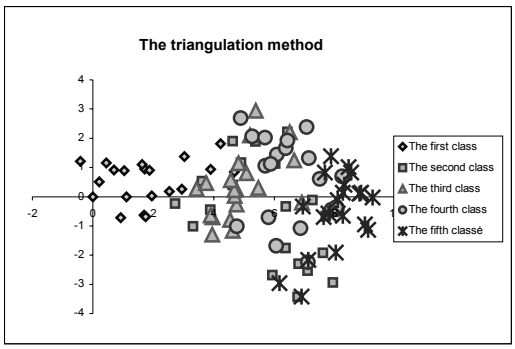
When mapping points by the triangulation method, we build a minimal spanning tree (MST) from the initial points and they are visualized on the plain. The remaining 50 points are sequentially mapped onto the plain, exactly preserving the distances up to the two nearest points mapped before (Figure 4a). These 100 points are also mapped by Sammon's method (Figure 4b). In order to prove an advantage of the combination of Sammon's and triangulation methods, the initial points are visualized on the plain by Sammon's method, while the remaining 50 points are sequentially mapped onto the plain by the triangulation method (Figure 4c).

The pictures obtained (Figure 4) as well the estimates of time necessary for mapping and the projection error (Table 1) illustrate that the triangulation method is fast enough, however the projection error is large. The projection error obtained by Sammon's algorithm is small, however this method is rather slow. Meanwhile, the triangulation and Sammon methods together operate rather fast and the accuracy is lost but insignificantly.

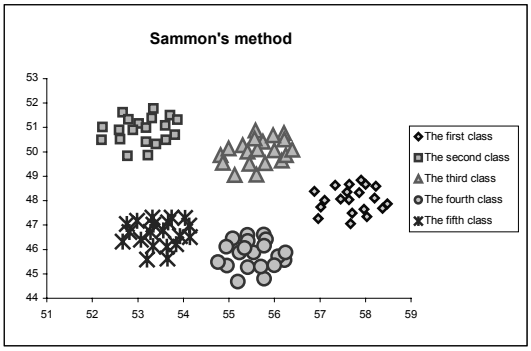
This paper also deals with the issue how time and the projection error depend on the fact how many initial points and the new ones have been mapped. The research was pursued as follows: the number of initial points was increased little by little, while the number of new ones was decreased. The projection error is calculated after mapping all the data available. The triangulation method may be realized using two approaches: the second nearest neighbor approach and the reference point approach.

Carrying out investigations with random numbers, it has been noticed that the projection error, obtained when visualizing data by the second nearest neighbor approach or the reference point approach, is variable, - no regularity can be established. It is not clear when it is higher: either by increasing number of initial points or by decreasing it (Figure 7).

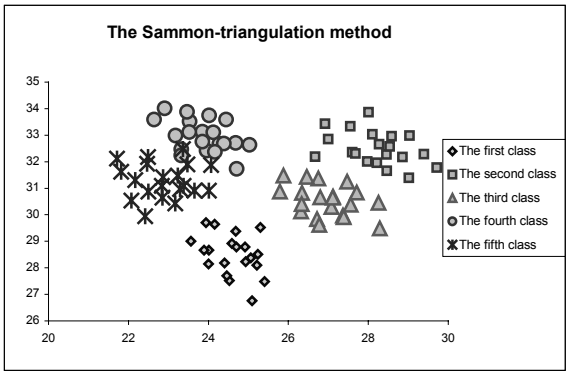
The data of coastal dunes were investigated. These are ecological data that define seaside dunes of Finland and their vegetation [7]. There are sixteen 16-dimensional vectors got by analyzing the correlation matrix of parameters that characterize the dunes [5].



a) 100 points mapped by the triangulation method



b) 100 points mapped by Sammon's method

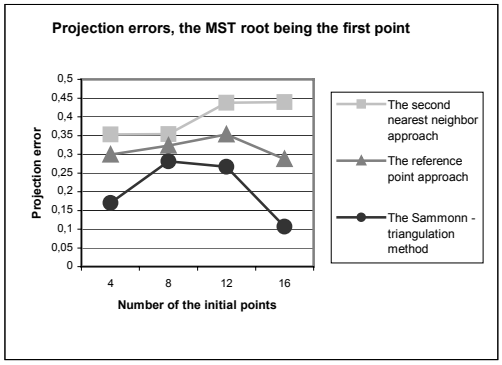


c) 50 initial points mapped by Sammon's method and 50 new points mapped by the triangulation method

Figure 4. Visualization of the test data

Table 1. Estimates of time and the projection error obtained by mapping clusters [100x10] by the triangulation method, Sammon's method, and by their combination

Method	Time, ms	Projection error (Sammon's error)
Triangulation	1639	0.2039847
Sammon	16589	0.0374195
Sammon and triangulation methods in combination	4889	0.0639497



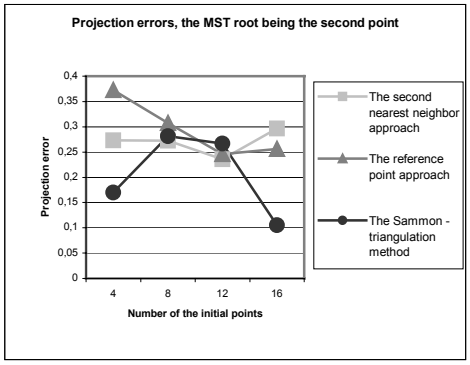


Figure 5. Graphs of projections errors obtained in the investigation of dune data by the second nearest neighbor approach and that of the reference point, the MST root being: a) the first point b) the second point, - and by Sammon's method

When investigating the data of dunes, different points of a given set were chosen as the root of the MST (reference point). It has been noticed that different projection errors are obtained even with the same number of initial points (Figure 5).

Consequently, the projection error depends on the order of the points to be mapped. Therefore, in further experiments we will consider only the projection error, obtained after visualizing the initial data by Sammon's algorithm, and the new ones – by the triangulation method (the second nearest neighbor approach).

Quite a few experiments have been done with random data. One test has been performed with 100 15-dimensional random data. Random numbers were generated in the interval (0; 1). 10, 20, 30, 40, 50, 60, 70, 80, 90, and 100 were selected as initial points. In total, 100 points have been mapped.

Figure 6 demonstrates that in mapping points either by the second nearest neighbor approach, or the reference point approach, or the combination of Sammon's and triangulation methods, an increasing time function is obtained. The larger the number of initial points, the more time it takes to map the points.

Figure 7 shows that the projection error, obtained by combining Sammon's and triangulation methods, decreases with an increase in the number of initial points.

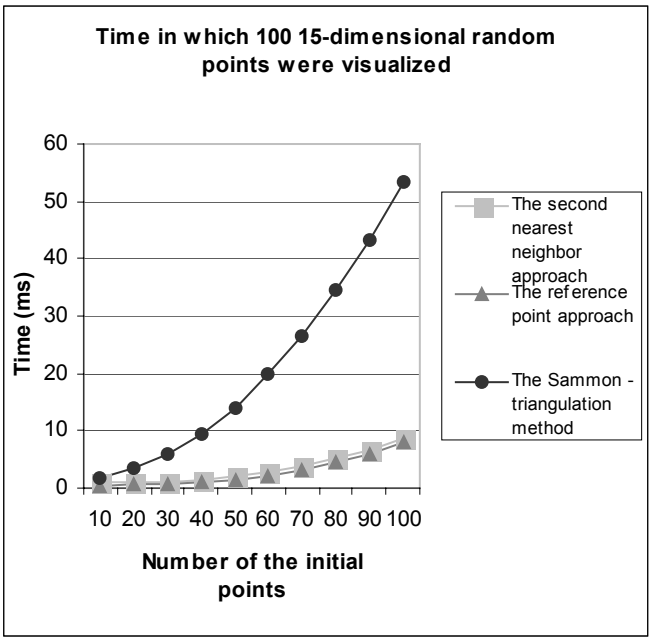


Figure 6. Graphs of time variation in which 100 15-dimensional random points were visualized

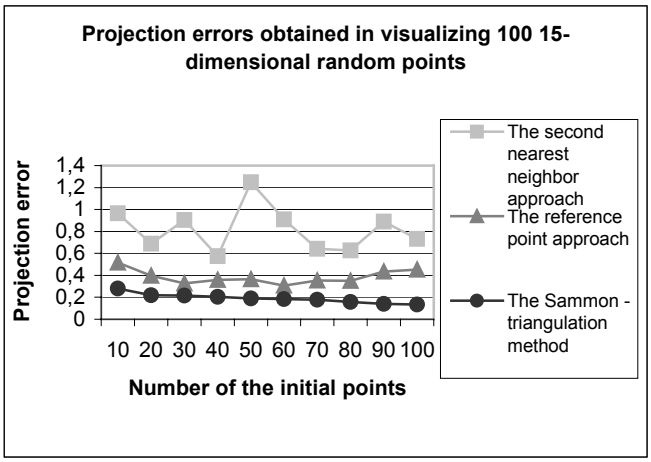


Figure 7. Graphs of projection error E_s (1) variation, obtained in visualizing 100 15-dimensional random points

6. Conclusions

The triangulation method, Sammon's algorithm, and the combination of both of them have been considered in this paper. Realizations of the triangulation method have been explored experimentally, using the method of the second nearest neighbor and that of the reference point to select the reference points. It has been established that in both cases the projection error strongly depends on the order of points to be mapped, which proves ones again that it

does not suffice to use the triangulation method alone for data visualization.

The triangulation method is fast enough, however it can preserve only $(2m-3)$ distances among the points analyzed. Sammon's algorithm tries to preserve all $\frac{m(m-1)}{2}$ relative distances among data points, however it is rather slow: to map a new point, the mapping procedure should be repeated once again.

Combining the triangulation method with Sammon's algorithm solves this problem. This method is helpful if we have to promptly map new points of the set analyzed. It acts rather fast, and the loss of accuracy is insignificant.

In general, any other MDS – type method may be used in the combination instead of the Sammon's method. Of course, the used computer time estimates differ.

References

[1] **J.C. Bezdek, N.R. Pal.** An index of topological preservation for feature extraction. *Pattern Recognition, Vol*.28, 1995, 381-391.

[2] **G. Biswas, A.K. Jain, R.C. Dubes.** Evaluation of Projection Algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol*.3, *No*.6, 1981, 701-708

[3] **I. Borg, P.J.F. Groenen.** Modern multidimensional scaling. 2*nd edition. New York: Springer*, 2005.

[4] **C. Brunsdon, A.S. Fotheringham, M.E. Charlton,** An Investigation of Methods for Visualising Highly Multivariate Datasets in Case Studies of Visualization in the Social Sciences, D. Unwin and P. Fisher (eds.) *Joint Information Systems Committee, ESRC, Technical Report Series* 43, *ISSN* 1356-9066, 1998, 55-80.

[5] **G. Dzemyda.** Visualization of correlation – based environmental data. *Environmetrics, Vol*.15, 2004, 827-836.

[6] **T. Hastie.** Principal Curves and Surfaces, PhD Dissertation. *Stanford Linear Accelerator Center, Stanford University, Stanford, California*, 1984. *<http://www.slac.stanford.edu/pubs/slacreports/slac-r-276.html>*.

[7] **P. Hellemaa.** The Development of Coastal Dunes and Their Vegetation in Finland. Dissertation. *Fenia* 176: 1, *Helsinki, ISSN* 0015-0010, 1988. *<http://ethesis.helsinki.fi/julkaisut/mat/maant/vk/hellemaa/>*.

[8] **R.C.T. Lee, J.R.Slagle, H. Blum.** A triangulation method for the sequential mapping of points from N-space to two-space. *IEEE Transactions on Computers, Vol*.26, 1977, 288-292.

[9] **R.C. Prim.** Shortest connection networks and some generalizations. *Bell System Technical Journal, Vol*.36, 1957, 1389-1401.

[10] **J.W. Sammon.** A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers, Vol*.18, 1969, 401-409.

[11] **P. Taylor.** Statistical Methods. Intelligent Data Analysis: an Introduction. *Edited by M. Berthold, D. J. Hand. Springer-Verlag*, 2003, 69-129.

Received January 2006.