The 13th Annual General Assembly of the IAMU

Expanding Frontiers -
Challenges and Opportunities in Maritime Education and Training

Application of an Internet-based Tool for Visualization of Multi-Dimensional Objects with Missing Data in Maritime Education

Natalia Nikolova
Assoc. Prof. PhD., N. Vaptsarov Naval Academy, V. Drumev 73, Varna 9026, Bulgaria,
natalianik@gmail.com

Daniela Toneva Zheynova
Assoc. Prof. PhD., Technical University – Varna, 1 Studentska Str., Varna, Bulgaria,
d_toneva@abv.bg

Kalin Kalinov
Assoc. Prof. DSc, Captain, N. Vaptsarov Naval Academy, V. Drumev 73, Varna 9026,
Bulgaria, va_vms@yahoo.com

Boyan Mednikarov
Prof. DSc, Captain, N. Vaptsarov Naval Academy, V. Drumev 73, Varna 9026, Bulgaria,
bobmednikarov@abv.bg

Kiril Tenekedjievld
Prof. DSc, N. Vaptsarov Naval Academy, V. Drumev 73, Varna 9026, Bulgaria, Kiril.
Tenekedjiev@fulbrightmail.org

Abstract: In certain disciplines of maritime education it is necessary to operate with multi-dimensional data that is difficult for students to comprehend. Such disciplines happen to be environmental monitoring of water, technical diagnostics of ship machines and equipment using statistical pattern recognition, decision analysis of alternatives with multi-dimensional consequences, etc. Worst still, some of the multi-dimensional vectors are not known completely. Even though there are some mathematical algorithms developed to tackle the problem stated, in some countries maritime education suffers certain limitations caused by the high prices of specialized software. Here, an Internet based tool is presented which solves two interconnected problems. The first problem is the visualization of multi-dimension environmental vectors. That problem is addressed with two methods: The Principal Component solution of orthogonal Factor Analysis model and with Multi-Dimensional Scaling procedure. The second problem is the generation (the imputation) of the missing data in multi-dimension environmental vectors. The implemented solution uses modified Roweis algorithm for Expectation Maximization (EM) algorithm for Principal Component solution of orthogonal Factor

Analysis method. The functionality of the system is described along with environmental examples.

Keywords: multi-dimensional scaling, principle component solution, Roweis algorithm,

1. General Set Up of the Visualization Problem

The records in an environmental database normally contain several environmental variables measured at approximately the same time point. Let us assume that we have an environmental database with n records each of which is represented by a p-dimensional column vector $\vec{x}_j = \left(x_1^{(j)}, x_2^{(j)}, \ldots, x_p^{(j)}\right)^T$, where T stands for the transpose operator and i=1,2,...,n. The whole data base can be written in a p x n matrix $D = \left(\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n\right)$.

If all the measurements of the environmental variables are known and recorded into the data base D then we can denote this matrix with $D_{full.}$ The problem which we face is to compress the p-dimensional vectors in $D_{full.}$ into m-dimensional vectors $\vec{f}_j$ for j=1, 2, ...,n. If m=2, the compressed vectors can be depicted by a scatter plot of the first against the second compressed coordinates. If m=3, the compressed vectors can be depicted by three scatter plots which show respectively the first against the second, the second against the third and the first against the third compressed coordinates. We call that a visualization problem. The result of the visualization problem is a compressed database in form of a m x n matrix $D_{compr} = \left(\vec{f}_1, \vec{f}_2, \ldots, \vec{f}_n\right)$.

Two classical algorithms are employed in the Internet based tool for solution of the visualization problem (that is for the data reduction from the p-dimensional to the m-dimensional space):

- Multi-Dimensional Scaling Procedure (MDS) [4]
- Principal Component Solutions of Exploratory Orthogonal Factor Analysis Model (PCA_EOFA) [6]

So the visualization module of the Internet-based tool has two sub-modules; the MDS sub-module and the PCA_EOFA sub-module.

The MDS sub-module uses 10 different distances as a measure of dissimilarity between the vectors:

- Manhattan distance: $\delta_{k,j} = \sum_{i=1}^{p}\left|x_i^{(k)} - x_i^{(j)}\right|$
- Euclidean Distance: $\delta_{k,j} = \sqrt{\sum_{i=1}^{p}\left(x_i^{(k)} - x_i^{(j)}\right)^2}$
- Minkovski r-distance (with r=3, 4, 5, 10, 20) : $\delta_{k,j} = \sum_{i=1}^{p}\left(\left|x_i^{(k)} - x_i^{(j)}\right|^r\right)^{1/r}$
- Hamilton (or Chebishev) distance: $\delta_{k,j} = \underset{i}{max}\left(\left|x_i^{(k)} - x_i^{(j)}\right|\right)$

- Normalized Euclidean Distance: $\delta_{k,j} = \sqrt{\sum_{j=1}^{p}\left(\frac{x_i^{(k)} - x_i^{(j)}}{s_i}\right)^2}$, where s_i is the sample standard deviation of the coordinate i.
- Mahalonobis distance: $\delta_{k,j} = \left(\vec{x}_k - \vec{x}_j\right)^T K^{-1}\left(\vec{x}_k - \vec{x}_j\right)$, where K is the sample covariance matrix of the data.

The Java realization of the MDS sub-module has been based on [1] where the 'stress' criterion (formula (8) in the accompanying paper [8]) is optimized. The mathematical algorithms of PCA_EOFA sub-module are explained in the accompanying paper [8].

Sometimes the record $\vec{x}_j$ of the data matrix D is not a full vector. Instead, part of the coordinates $x_1^{(j)}, x_2^{(j)}, \ldots, x_p^{(j)}$ are missing due to various reasons. In this case the same data matrix is called $D_{miss} = \left(\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n\right)$. For example, let p=5 and for the third observation $\vec{x}_3$, the second and the fifth coordinates are missing (denoted as NaN), whereas the first the third and the forth coordinates are -17, 24 and 6 respectively. Then:

$$\vec{x}_3 = \begin{pmatrix} x_1^{(3)} \\ x_2^{(3)} \\ x_3^{(3)} \\ x_4^{(3)} \\ x_5^{(3)} \end{pmatrix} = \begin{pmatrix} -17 \\ NaN \\ 24 \\ 6 \\ NaN \end{pmatrix}$$

There are many reasons for the missing data, especially when social surveys are involved, because people do not always want to answer all questions. As long as the environmental data is gathered by direct measurements with some measurement devices we have accepted the hypothesis that the data is missing at random [7]. The problem that we face is to restore the values of the missing part in D_{miss}. Of course the missing part of D_{miss} is an unobservable quantity which has to be estimated from the observable quantities (the known data) in the data base. We call that the Missing Data Imputation Problem. The result of the missing data imputation problem is an imputed database in the form of a p x n matrix $D_{imp} = \left(\vec{x}_1^{(imp)}, \vec{x}_2^{(imp)}, \ldots, \vec{x}_n^{(imp)}\right)$. D_{imp} has the same values as D_{miss} but the missing values are substituted with values generated by a mathematical algorithm called imputed values.

A powerful expectation maximization algorithm for estimation of unobservable quantities was proposed in the seminal paper [3]. Six years later that algorithm was properly proven [10]. In [9], the algorithm was used for imputation of missing data in the context of a factor analysis method. In the accompanying paper [8] a modification of the Roweis algorithm is proposed which quickly converges for medium sized data points. In the developed missing data imputation module of the Internet based tool the linear system $T_j \vec{z}_j^{(nimp)} = \vec{t}_j$ with p equations and N_j unknowns in point 7) of the algorithm in [8] is solved by QR decomposition of the matrix T_j [5].

Once the missing data imputation problem is solved and D_{imp} is generated then (having in mind that there are no more missing position in D_{imp}) it can be put that $D_{full} = D_{imp}$ and the imputed data can be passed to the visualization problem for data reduction.

2. Examples and demonstration

An Internet based system is developed to realize procedures for visualization of multi-dimensional data and for imputing missing data into multi-dimensional data sets. The system is available through the webpage www.ubss-tuv.com and it has been developed under the activities of the UPGRADE Black Sea Scientific Network (financed by the Seventh Framework Program of the EU). The main window of the platform is shown in fig. 1. Here we focus on the options “Visualization” and “Missing Values”, the others being described in [3]. The procedures to be commented on may run on three file formats: text document (*.txt), Excel document (*.xls) and comma delimited text (*.csv).

2.1 Visualizing multi-dimensional data

A test file is created, containing 70 numbers of 7-dimensional test measurements. The structure of the file is shown in fig. 2 (for the sake of all examples, only the operation with Excel files shall be discussed here, all the others being very similar). The names of the coordinates (i.e. the environmental variables) are given on the second row, columns from 2 to p+1=8 of the file. The third row (same columns) contains the variable dimensions. The fourth row, columns from 2 to p+1=8 contains the values of $\vec{x}_1$, and the other vectors are written consecutively on the next rows. The last one is on row n+3=70+3=73. The first column of the file contains the consecutive number of the vectors.

The visualization section of the UBSS system is given in fig. 3. Let us use factor analysis to perform the visualization. After selecting the file (fig. 4) the user might choose from the panel shown in fig. 5:

1. whether to use standardized data or not (the first option is selected in this example), i.e. whether to replace the original data $\vec{x}_j$ in the input by the data normalized in the first place in formulae (30) and (31) in [8];
2. the output dimension to be 2 or to be 3 (m=3 is selected in this example);
3. the type of compressed data that is ordinary least square factor score using formula (28) or weighted least square factor score (29) of [8] (the first option is selected in this example).

Choosing to proceed, the system generates output as follows:

1. scatter plot of the first two coordinates of the compressed data (see fig. 6 for one of the three scatter plots, because the compressed data is three-dimensional);
2. the m-dimensional compressed data values on screen;

3. an excel file with five sheets, containing the compressed data points, the covariance matrix, the absolute error, the relative error and the cumulative percentage of the total variance explained (fig. 7) (which are formulae respectively (21), (25), (26) and (27) from [8]).

If we choose to visualize using multi-dimensional scaling, then the partial differences would appear in the input and in the output of the system. The parameters which determine the course of calculation of the multi-dimensional scaling are defined in a panel shown in fig. 8.

1. 1) whether to use standardized data or not (the first option is selected in this example), as in the visualization with factor analysis.
2. 2) the output dimension to be 2 or to be 3 (m=2 is selected in this example);
3. 3) the type of distance measure (Manhattan distance in this example).

Choosing to proceed, the system generates output as follows:

1. 1) scatter plot of the first two coordinates of the compressed data (as in fig. 6 but here this is the only scatter plot because the compressed data is two-dimensional);
2. 2) the compressed m-dimensional data values on screen;
3. 3) an excel file with four sheets (fig.9), containing the compressed m-dimensional data points, ($\vec{f}_j$), the distances in the original space ($\delta_{k,j}$), the distances in the compressed space ($\delta_{k,j}$) and the difference between the corresponding pairs of distances ($\delta_{k,j} - d_{k,j}$).

Such a procedure might be employed to identify outliers or certain internal connections (e.g. sometimes data would lie on the same straight line, which would indicate high level of multicolinearity in the data).

2.2 Imputing Missing Values

A test file is created containing 500 numbers of 10-dimensional test measurements, where some of the values are missing. The structure of the file is shown in fig. 10. The names of the coordinates (i.e. the environmental variables) are given on the second row, columns from 2 to p+1=11 of the file. The third row (same columns) contains the variable dimensions. The fourth row, columns from 2 to p+1=11 contains the values of $\vec{x}_1$, and the other vectors are written consecutively on the next rows. The last one is on row n+3=500+3=503. The first column of the file contains the consecutive number of the vectors.

The missing values section of the system is given in fig. 11. After choosing the *.xls format and the file itself (fig. 12) the user might choose:

1. the dimension of the compressed space (m=5 in the example)

2. whether to use standardized data or not (the first option in this example), i.e. whether to divide by $\sqrt{k_{i,i}}$ or not in step 3) of the algorithm in [8];

3. the drifting moments (either moving or fixed moments on X on each iteration, i.e. whether to execute step 9) in the algorithm of [8] or not; it is executed for this example);

4. whether to have additional fixing of the recovered data or not (whether to execute step 17) in the algorithm of [8] or not; it is executed for this example).

Choosing to proceed, the system generates output as follows:

1. graphics of the imputed data (see fig. 13 for the first coordinate, where the imputed points are in blue; such a plot is created for each coordinate);

2. all the imputed values on the screen;

3. an excel file with a single sheet containing the imputed set D_{imp} with structure the same as the input data, however the imputed values are placed instead of the missing ones in red (fig. 14).

2.3 Real Data Analysis

A file with real environmental data is created, related to chemical characteristics of sea water, containing 110 numbers of 13-dimensional measurements. Some of the values in the dataset are missing. The structure of the file is shown in fig. 15. The names of the coordinates (i.e. the chemical parameters) are given on the second row, columns from 2 to p+1=14 of the file. The third row (same columns) contains the variable dimensions. The fourth row, columns from 2 to p+1=14 contains the values of $\vec{x}_1$, and the other vectors are written consecutively on the next rows. The last one is on row n+3=110+3=113. The first column of the file contains the consecutive number of the vectors.

We set the following input parameters for the missing value procedure:

1. 6-dimensional compressed space;

2. standardized data;

3. moving drifting moments on X on each iteration;

4. additional fixing of the recovered data.

The system generates the following result:

1. graphics of the imputed data (see fig. 16 for the second coordinate; such a plot is created for each coordinate);

2. the imputed data set on screen;

3. an excel file with a single sheet structured the same as the input data, but the imputed values are placed instead of the missing ones in red (fig. 17).

The resulting file is sent for visualization with multi-dimensional scaling at m=3 using Mahalonobis distance measure. As an output, the system generated the following:

1. 1) graphics of the compressed data (three two-dimensional plots; fig. 18 gives the plot for the second and third compressed coordinates);
2. 2) the m-dimensional compressed data values on the screen;
3. 3) an excel file with four sheets as in section II.1 (see fig. 19).

The resulting file from the missing values procedure is also sent for visualization with factor analysis at m=2. As an output, the system generated the following:

1. graphics of the compressed data (see fig. 20);
2. the m-dimensional compressed data values on screen;
3. an excel file with five sheets, containing the compressed data points, the covariance matrix, the absolute error, the relative error and the cumulative percentage of the total variance explained (fig. 21) as in section II.1.

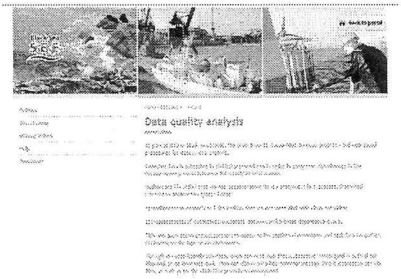
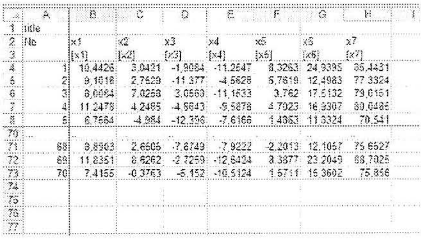


Fig. 1



	A	B	C	D	E	F	G	H	I
1	title								
2	No	x1	x2	x3	x4	x5	x6	x7	
3		[x1]	[x2]	[x3]	[x4]	[x5]	[x6]	[x7]	
4	1	10,4426	3,0421	-1,9084	-11,2647	8,3263	24,9395	86,4431	
5	2	9,1818	2,7629	-11,377	-4,5628	6,7619	12,4983	77,3324	
6	3	8,0884	7,0258	3,0668	-11,1633	3,762	17,5132	79,0151	
7	4	11,2478	4,2485	-4,9643	-9,5878	4,7023	16,9307	80,0486	
8	5	6,7664	-4,984	-12,396	-7,6166	1,4363	11,6324	70,541	
70	...	...	...	...	...	...	...	...	
71	68	8,8903	2,6606	-7,8749	-7,9222	-2,2013	12,1057	75,6527	
72	69	11,8351	8,6262	-2,7259	-12,8434	8,3877	23,2049	88,7025	
73	70	7,4155	-0,3763	-5,152	-10,5124	1,5711	16,3602	75,856	
74									
75									
76									
77									

Fig. 2

Fig. 3

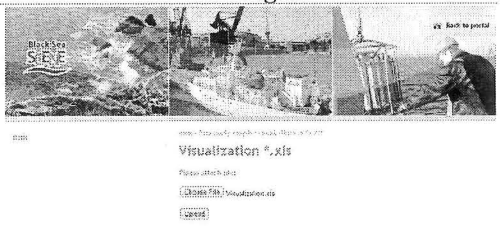
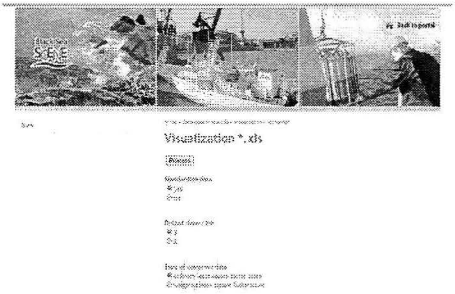


Fig. 4



Visualization *.xls

[illegible]

Standardize data
● yes
○ no

Output dimension
● y
○ x

Type of [illegible]
● [illegible]
○ [illegible]

Fig. 5

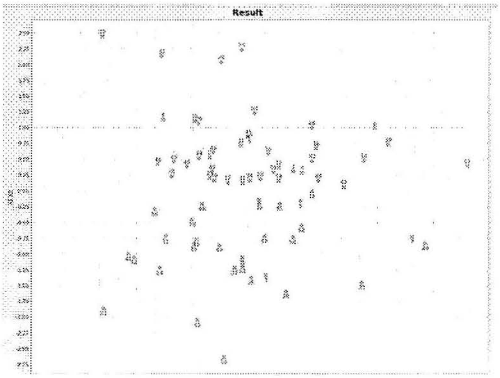
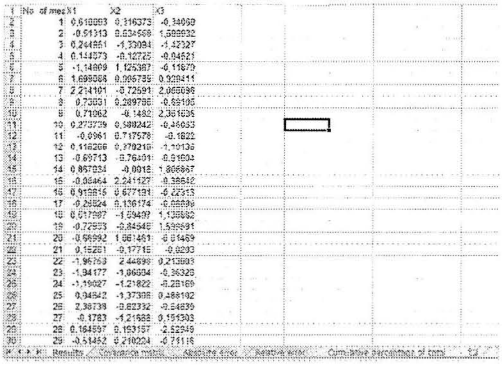
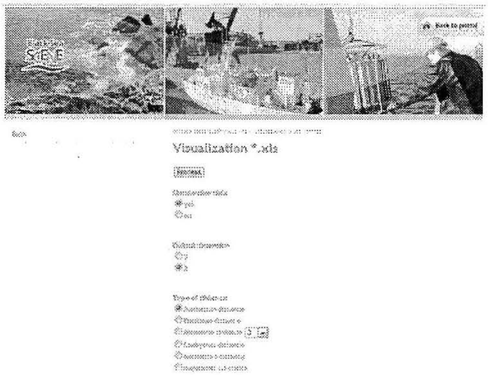


Fig. 6



1	No. of mes	X1	X2	X3
2	1	0,610093	0,316373	-0,34069
3	2	-0,51313	0,634568	1,589932
4	3	0,244851	-1,33084	-1,42327
5	4	0,144573	-0,12725	-0,94521
6	5	-1,14809	1,125387	-0,11679
7	6	1,699068	0,096739	0,928411
8	7	2,214101	-0,72591	2,065096
9	8	0,73031	0,289796	-0,69106
10	9	0,71062	-0,1482	2,361606
11	10	0,273739	0,588242	-0,46053
12	11	-0,0961	0,717578	-0,1822
13	12	0,116206	0,378218	-1,10136
14	13	-0,69713	-0,76401	-0,91604
15	14	0,867034	-0,0018	1,806867
16	15	-0,08464	2,241127	-0,38642
17	16	0,919816	0,677191	-0,22313
18	17	-0,26824	0,136174	-0,08699
19	18	0,617987	-1,59497	1,138862
20	19	-0,72933	-0,84546	1,599691
21	20	-0,66992	1,861461	-0,01469
22	21	0,15281	-0,17715	-0,0293
23	22	-1,96768	2,44896	0,213883
24	23	-1,94177	-1,06604	-0,36326
25	24	-1,19027	-1,21822	-0,28169
26	25	0,94642	-1,37396	0,488192
27	26	2,38738	-0,82332	-0,84839
28	27	-0,1783	-1,21688	0,151303
29	28	0,164597	0,193157	-2,52949
30	29	-0,51452	0,210224	-0,73116

Fig. 7



Visualization *.xls

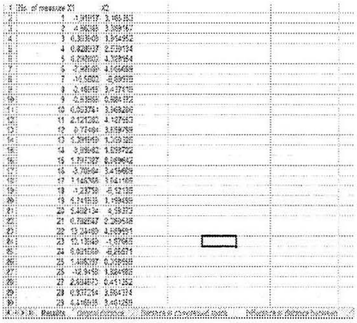
[illegible]

Standardize data
● yes
○ no

Output dimension
○ y
● x

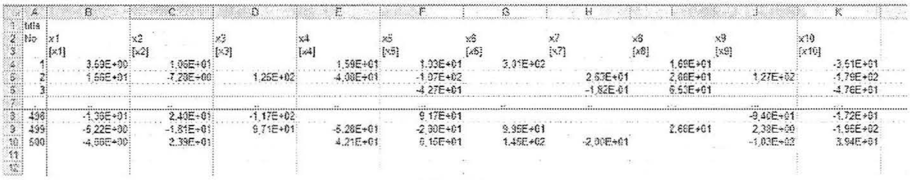
Type of distance
● [illegible]
○ [illegible]
○ [illegible] 3
○ [illegible]
○ [illegible]
○ [illegible]

Fig. 8



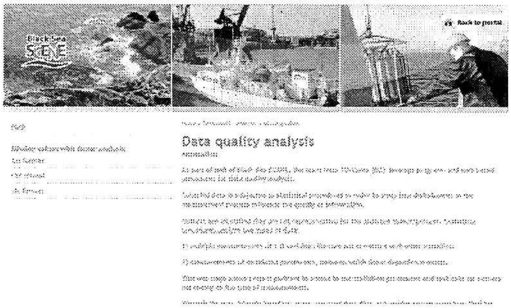
1	No. of measure	X1	X2
2	1	-1,91117	3,165353
3	2	4,56348	3,389167
4	3	0,383403	3,954952
5	4	0,828937	2,539134
6	5	0,292802	4,328954
7	6	-7,92169	4,166499
8	7	-14,5802	-6,69978
9	8	-2,46915	3,437478
10	9	-2,53958	0,984172
11	10	0,053741	3,968286
12	11	2,121282	4,127453
13	12	0,72494	3,159759
14	13	1,391159	1,359326
15	14	3,39642	1,529722
16	15	1,797327	6,369642
17	16	-3,70494	3,415409
18	17	1,146768	3,041105
19	18	-1,23758	-6,12135
20	19	4,741535	1,199459
21	20	5,402134	4,59373
22	21	0,782547	2,269548
23	22	13,24489	4,169491
24	23	10,13549	-1,67063
25	24	6,021569	-6,26671
26	25	1,486797	0,359558
27	26	-12,9458	1,884986
28	27	2,834870	0,411352
29	28	0,937214	3,864774
30	29	4,416835	3,461289

Fig. 9



	A	B	C	D	E	F	G	H	I	J	K
1	tdla										
2	No	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10
3		[x1]	[x2]	[x3]	[x4]	[x5]	[x6]	[x7]	[x8]	[x9]	[x10]
4	1	3.69E+00	1.06E+01		1.59E+01	1.03E+01	3.01E+02		1.69E+01		-3.61E+01
5	2	1.66E+01	-7.29E+00	1.25E+02	-4.08E+01	-1.07E+02		2.63E+01	2.88E+01	1.27E+02	-1.79E+02
6	3					-4.27E+01		-1.82E-01	6.63E+01		-4.76E+01
7	...	...	...	...	...	...	...	...	...	...	...
8	498	-1.36E+01	2.40E+01	-1.17E+02		9.17E+01				-9.40E+01	-1.72E+01
9	499	-5.22E+00	-1.81E+01	9.71E+01	-5.28E+01	-2.90E+01	9.95E+01		2.66E+01	2.38E+00	-1.95E+02
10	500	-4.66E+00	2.39E+01		4.21E+01	6.15E+01	1.46E+02	-2.00E+01		-1.03E+02	3.94E+01
11											
12											

Fig. 10



Back to portal

Data quality analysis

[illegible]

Fig. 11

Back to portal

Missing values *.xls

[illegible]

Fig. 12

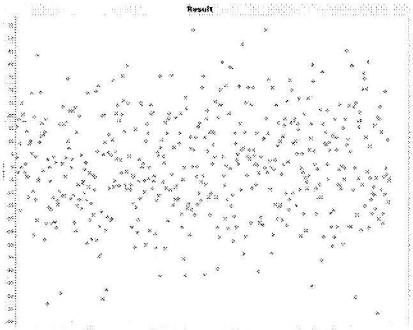
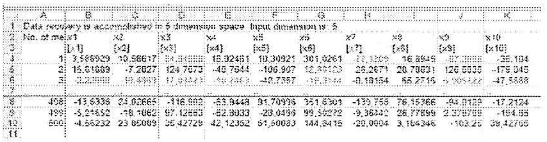
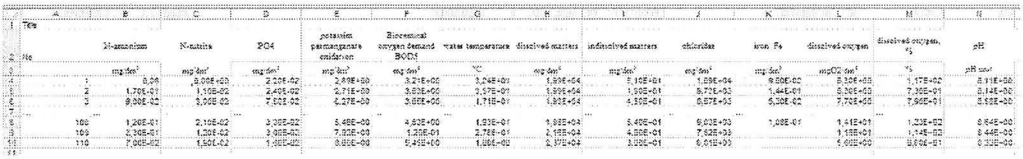


Fig. 13



	A	B	C	D	E	F	G	H	I	J	K
1	Data recovery is accomplished in 5 dimension space. Input dimension is 5										
2	No. of me	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10
3		[x1]	[x2]	[x3]	[x4]	[x5]	[x6]	[x7]	[x8]	[x9]	[x10]
4	1	3,586929	10,58617	[illegible]	16,92461	10,30921	301,0261	[illegible]	16,8945	[illegible]	-35,104
5	2	15,61889	-7,2827	124,7673	-48,7644	-106,907	[illegible]	26,2671	28,78631	126,5835	-179,045
6	3	[illegible]	[illegible]	[illegible]	[illegible]	-48,7357	[illegible]	-0,18154	55,2716	[illegible]	-47,5888
7											
8	498	-13,6336	24,02665	-116,802	-63,9448	91,70996	351,6301	-139,758	76,15366	-94,0129	-17,2124
9	499	-5,21652	-18,1062	97,12663	-62,8833	-23,0496	99,50272	-9,36442	26,77899	2,378709	-184,85
10	500	-4,56232	23,85069	35,42729	42,12362	61,60083	144,9415	-20,0004	3,184346	-103,25	39,42765
11											

Fig. 14



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Title													
2	No	N-ammonium	N-nitrite	PO4	potassium permanganate oxidation	Biochemical oxygen demand BOD5	water temperature	dissolved matters	indissolved matters	chlorides	iron Fe	dissolved oxygen	dissolved oxygen, %	pH
3		mg/dm^3	mg/dm^3	mg/dm^3	mg/dm^3	mg/dm^3	°C	mg/dm^3	mg/dm^3	mg/dm^3	mg/dm^3	mgO2/dm^3	%	pH unit
4	1	8,06	2,00E+00	2,20E-02	2,89E+00	3,21E+00	3,24E+01	1,59E+04	5,10E+01	1,09E+04	9,50E-02	8,20E+00	1,17E+02	8,11E+00
5	2	1,70E-01	1,10E-02	2,40E-02	2,71E+00	3,63E+00	3,57E+01	1,59E+04	1,90E+01	9,73E+03	1,44E-01	8,30E+00	7,30E+01	8,14E+00
6	3	9,00E-02	3,00E-02	7,50E-02	6,27E+00	3,65E+00	1,71E+01	1,92E+04	4,50E-01	8,57E+03	6,30E-02	7,70E+00	7,90E+01	8,58E+00
7	...	...		...	...	...	...		...	...	...		...	
8	108	1,20E-01	2,10E-02	3,30E-02	5,48E+00	4,60E+00	1,93E+01	1,98E+04	5,40E-01	9,03E+03	1,08E-01	1,41E+01	1,20E+02	8,64E+00
9	109	2,30E-01	1,20E-02	3,08E-02	7,92E+00	1,20E-01	2,78E+01	2,15E+04	4,80E-01	7,62E+03		1,15E+01	1,14E+02	8,44E+00
10	110	7,00E-02	1,50E-02	1,60E-02	8,60E+00	5,45E+00	1,80E+00	2,37E+04	3,00E-01	6,01E+03		5,60E+00	8,80E-01	8,33E+00
11														

Fig. 15

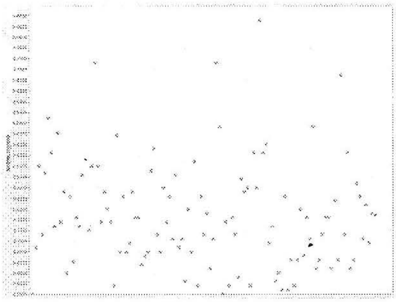
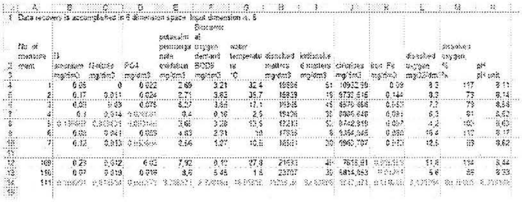


Fig. 16



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Data recovery is accomplished in 6 dimension space. Input dimension is 6													
2	No. of measurement	N-ammonium	N-nitrite	PO4	potassium permanganate oxidation	Biochemical oxygen demand BOD5	water temperature	dissolved matters	indissolved matters	chlorides	iron Fe	dissolved oxygen	dissolved oxygen, %	pH
3		mg/dm3	mg/dm3	mg/dm3	mg/dm3	mg/dm3	°C	mg/dm3	mg/dm3	mg/dm3	mg/dm3	mgO2/dm3	%	pH unit
4	1	0.06	0	0.022	2.89	3.21	32.4	15896	51	10932.99	0.09	8.2	117	8.11
5	2	0.17	0.011	0.024	2,71	3.63	35.7	15939	19	9732.616	0,144	8.3	73	8.14
6	3	0.09	0.03	0.075	6.27	3.65	17.1	[illegible]	45	8570.868	0.063	7.7	79	8.58
7	4	0.1	0.014	[illegible]	0.4	0.16	2.5	19426	38	8986.646	0.091	6.3	91	8.62
8	5	[illegible]	[illegible]	[illegible]	3.66	3.28	13.5	12213	52	8742.919	0.097	4.2	100	8.63
9	6	0.09	0.041	0.009	4.83	2.31	10	17836	8	9364.645	0.090	15.4	117	8.17
10	7	0.12	0.013	[illegible]	2.56	1.27	10.5	16661	20	9960.707	[illegible]	12.5	89	8.62
11														
12	109	0.23	0.012	6.03	7,92	0.12	27,8	21693	48	7615,81	[illegible]	11,8	114	8,44
13	110	0.07	0.019	0.019	8,6	5.45	1.8	23707	30	6814,063	[illegible]	5.6	86	8.33
14	111	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
15														

Fig. 17

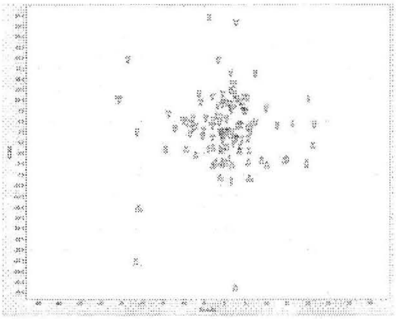
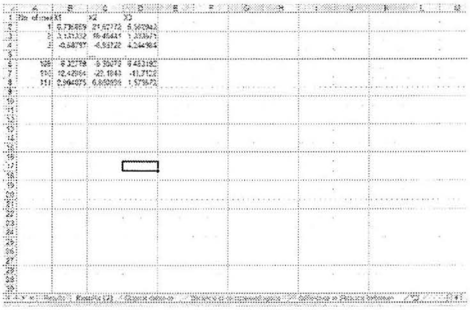


Fig. 18



No. of me	X1	X2	X3
1	6.736859	21.67772	5.562942
2	3.131332	16.46441	1.233671
3	-0.58797	-6.93722	4.244984
...			
109	6.32778	-9.30072	9.483192
110	12.42864	-22.1843	-11.7122
111	2.964875	6.860896	1.573672

Fig. 19

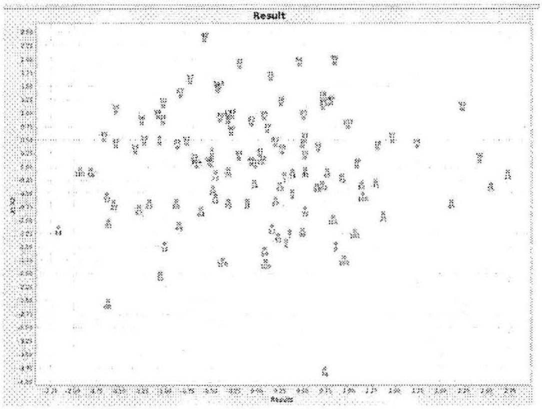
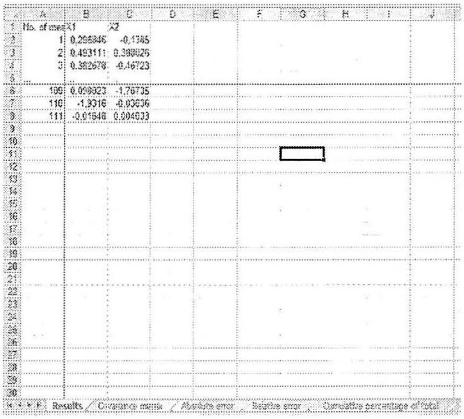


Fig. 20



No. of me	X1	X2
1	0,298846	-0,1385
2	0.493111	0.388026
3	0.382678	-0,46723
...		
109	0,098023	-1,76735
110	-1,3316	-0,03636
111	-0,91648	0,004033

Fig. 21

3. Conclusions

This paper presented the operation of an Internet based tool which solves two interconnected problems – visualization of multi-dimensional environmental vectors and imputation of missing values in such data sets. The procedures use modern algorithms and techniques to generate results. The tool is available free of charge to all users.

4. Acknowledgments

This paper has been supported by the UPGRADE Black Sea SCENE project (project No. 226592, Seventh Framework Program), and the INPORT project (project No. DVU01-0031, Bulgarian National Science Fund).

References

[1] Algorithmic Group, "*MDSJ: Java Library for Multidimensional Scaling (Version 2)*", University of Konstanz (2009)

[2] Bechev, Ch., Nikolova, N.D., Teohareva, M., and Tenekedjiev, K., "Testing the Quality of Environmental Data by an Internet-Based Tool", *Journal of Environmental Protection and Ecology*, Vol. 10, No. 3, (2009), pp. 877-888

[3] Dempster, A.P., Laird, N.M., and Rubin, D.B., "Maximum Likelihood from Incomplete Data via the EM Algorithm", *Journal of the Royal Statistical Society, Series B*, Vol. 39, (1977), pp. 1-39

[4] Duda, R., Hart, P, and Stork, D. "*Pattern Classification and Scene Analysis*", A Wiley Interscience Publication, (1973)

[5] Golub, G.H., and Van Loan, Ch. "*Matrix Computations*", Third Edition, Johns Hopkins (1996)

[6] Johnson R., and Wicherin D., "*Applied Multivariate Statistical Analysis*", 6th ed., Prentice Hall, (2007)

[7] Little, R. J.A., "A Test of Missing Completely at Random for Multivariate Data with Missing Values", *Journal of the American Statistical Association*, Vol. 83, No. 404, (1988), pp. 1198-1202

[8] Nikolova, N.D., Toneva-Zheynova, D., Naydenov, D., and Tenekedjiev, K. "Imputing missing values of environmental multi-dimensional vectors using a modified Roweis algorithm", *International Workshop on Dynamics and Control in Agriculture and Food Processing*, (2012), Plovdiv, Bulgaria, (in print)

[9] Roweis, S., "EM Algorithms for PCA and SPCA", *Neural Informatics Processing Systems, NIPS'1997*, (1977), pp. 626-632

[10] Wu, J., "On the Convergence Properties of the EM Algorithm", *The Annals of Statistics*, Vol. 11, No. 1, (1983), pp. 95-103