Communications in Physics, Vol. 27, No. 4 (2017), pp. 291-300
DOI:10.15625/0868-3166/27/4/10499

DEVELOPMENT OF A PC PROGRAM FOR MULTIVARIATE STATISTICAL ANALYSIS

PHAM NGOC SON[†], CAO DONG VU AND MAI QUYNH ANH
Nuclear Research Institute, 1 Nguyen Tu Luc, Dalat, Vietnam

[†]*E-mail:* pnson.nri@gmail.com

Received 7 July 2017
Accepted for publication 11 September 2017
Published 18 November 2017

Abstract. *This report introduces a new computer program, so-called MSAP-1.0, which has been developed at the Dalat Nuclear Research Institute, for data processing and interpretation of the experimental data sheets based on the multivariate data analysis techniques. In this preliminary version of the program, the dimensions of a given data set to be analyzed are up to 50 variables and thousands of observations. The main functions in this version are principal component analysis, cluster analysis, standardization and output data plot. In comparison with other well-known statistical analysis software programs, the same results are very well reproduced with MSAP-1.0. The format of the input data file was designed in a way convenient for the management of experimental survey data or preparation from other analysis procedures at the Institute.*

Keywords: multivariate data analysis, principal component analysis, cluster analysis.

Classification numbers: 29.85.-c, 02.50.Sk.

I. INTRODUCTION

In the modern trend of survey data, multi-dimension information is often observed in many cases of research and application. In these situations, the data reduction process or multivariate statistical analysis method should be performed in order to determine whether any distinct groups or distributions are present in the input data sheet that supports a meaningful interpretation in the research topic such as archeological, geological, environmental, etc.

Based upon a data set of elemental concentrations, for a particular specimen, questions that we hope to answer are: (i) Where did the specimen come from? (ii) From what raw materials was the specimen made? (iii) Was the specimen made at the same time and in the same place as other specimens? [1]. In our case, the archaeological brick and clay samples have been collected from a number of archaeological sites in Vietnam, and the concentrations of more than 28 elements in these specimens are determined by the reactor based neutron activation analysis (NAA) method. Accordingly, the principal component analysis (PCA) and cluster analysis (CA) methods are required to investigate the correlation of the brick and clay samples, and also to identify the similarities and differences between groups of brick specimens. The variations of an original multivariate data set are usually under indirect observation, but after being treated by PCA, the optimal visualization of variance can be presented with a high degree of explanation [2], in which a new space of principal component scores with maximal variance is obtained within the linear combination of the initial variables and eigenvectors as the following expressions [3]. Suppose that the initial data sheet is prepared as an $n \times p$ matrix **A**, where n is the number of the observation vectors, and p is the number of variables of each vector.

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1p} \\ a_{21} & a_{22} & \dots & a_{2p} \\ & & & \\ \vdots & & & \\ & & & \\ a_{n1} & a_{n2} & \dots & a_{np} \end{bmatrix} \qquad (1)$$

The covariance matrix **S** of **A**, and the eigenvector matrix **X** are defined as:

$$\mathbf{S}_{jk} = \frac{1}{n-1} \sum_{i=1}^{n} (a_{ij} - \overline{a_j})(a_{ik} - \overline{a_k}) \qquad (2)$$
$$\{j = 1, \dots, p;\ k = 1, \dots, p\},$$

$$(\mathbf{S} - \lambda \mathbf{I})\mathbf{X} = 0 \qquad (3)$$

where λ is the eigen-value (column matrix), and **I** is the unity matrix.

We can determine the λ and **X** matrices from the condition of $|\mathbf{A} - \lambda \mathbf{I}| = 0$, and then the principal components (PCA scores) can be calculated as the following linear combinations:

$$y_j = (x_{j1}, x_{j2}, x_{j3}, \dots, x^T_{jp)}(a_{j1}, a_{j2}, a_{j3}, \dots, a_{jp}) = \sum_{i=1}^{p} x_{ji} a_{ji} \qquad (4)$$

$$\begin{aligned} y_1 &= x_{11}a_1 + x_{12}a_2 + x_{13}a_3 + \ldots + x_{1p}a_p \\ y_2 &= x_{21}a_1 + x_{22}a_2 + x_{23}a_3 + \ldots + x_{2p}a_p \\ y_3 &= x_{31}a_1 + x_{32}a_2 + x_{33}a_3 + \ldots + x_{3p}a_p \\ y_4 &= x_{41}a_1 + x_{42}a_2 + x_{43}a_3 + \ldots + x_{4p}a_p \end{aligned}$$

where y: principal component scores, a: linear combination parameter (eigenvector), p: size of variable.

In addition, the cluster analysis is another multivariate analysis approach to measure the dissimilarity between specimens using Euclidean distances. The goal is to find an optimal grouping for which the specimens within each cluster are similar, but the clusters or groups are dissimilar to each other. The result of cluster analysis is generally presented in form of a dendrogram that shows the order and level of specimen clustering. Because the interpretation from a specific dendrogram is highly subjective, it is normally only used to identify possible groups, and after that other techniques are employed for group refinement and classification [1,4]. The Euclidean distance between two vectors $\mathbf{x} = (x_1, x_2, \ldots, x_p)$ and $\mathbf{y} = (y_1, y_2, \ldots, y_p)$ is defined as:

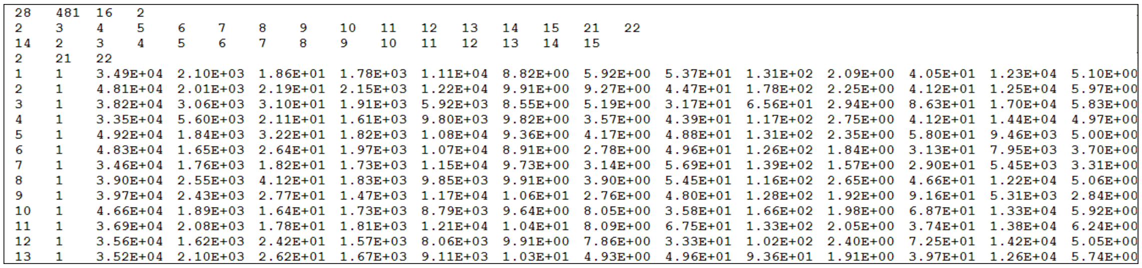
$$d(x,y) = \sqrt{\sum \frac{1}{S}(x_i - y_i)^2} \tag{5}$$

where S is the covariance matrix of A and $d(x,y)$ is Euclidean distance between vectors **x** and **y**.

In this work, we developed a new computer program, so-called MSAP-1.0 to perform the above-mentioned calculations within the multivariate data analysis techniques. In this preliminary version of the program, the dimensions of a given data set to be analyzed are up to 50 variables and thousands of observations. The main functions in this version are principal component analysis, cluster analysis, standardization and output data plot. In comparison with the SPSS program, a well-known statistical analysis software, the same results are very well reproduced with MSAP-1.0. The format of the input data file was designed in a way that is convenient for the management of experimental survey data or preparation from other analysis procedures at the Dalat Nuclear Research Institute.

II. DEVELOPMENT OF THE COMPUTER PROGRAM MSAP-0.1

Based on the multivariate statistical analysis method, a new C++ computer program, so-called MSAP-1.0 (Multi-variable Statistical Analysis Program), has been developed for principal component analysis, cluster analysis, standardization, and output data plot. The program was tested and applied in the processing and interpretation of multi-dimensional measured data from NAA studies of archaeological samples collected from archaeological sites in Vietnam. The format structure of an input file and the window interface are presented in Fig. 1 and Fig. 2.



```
28   481  16   2
2    3    4    5    6    7    8    9    10   11   12   13   14   15   21   22
14   2    3    4    5    6    7    8    9    10   11   12   13   14   15
2    21   22
1    1    3.49E+04  2.10E+03  1.86E+01  1.78E+03  1.11E+04  8.82E+00  5.92E+00  5.37E+01  1.31E+02  2.09E+00  4.05E+01  1.23E+04  5.10E+00
2    1    4.81E+04  2.01E+03  2.19E+01  2.15E+03  1.22E+04  9.91E+00  9.27E+00  4.47E+01  1.78E+02  2.25E+00  4.12E+01  1.25E+04  5.97E+00
3    1    3.82E+04  3.06E+03  3.10E+01  1.91E+03  5.92E+03  8.55E+00  5.19E+00  3.17E+01  6.56E+01  2.94E+00  8.63E+01  1.70E+04  5.83E+00
4    1    3.35E+04  5.60E+03  2.11E+01  1.61E+03  9.80E+03  9.82E+00  3.57E+00  4.39E+01  1.17E+02  2.75E+00  4.12E+01  1.44E+04  4.97E+00
5    1    4.92E+04  1.84E+03  3.22E+01  1.82E+03  1.08E+04  9.36E+00  4.17E+00  4.88E+01  1.31E+02  2.35E+00  5.80E+01  9.46E+03  5.00E+00
6    1    4.83E+04  1.65E+03  2.64E+01  1.97E+03  1.07E+04  8.91E+00  2.78E+00  4.96E+01  1.26E+02  1.84E+00  3.13E+01  7.95E+03  3.70E+00
7    1    3.46E+04  1.76E+03  1.82E+01  1.73E+03  1.15E+04  9.73E+00  3.14E+00  5.69E+01  1.39E+02  1.57E+00  2.90E+01  5.45E+03  3.31E+00
8    1    3.90E+04  2.55E+03  4.12E+01  1.83E+03  9.85E+03  9.91E+00  3.90E+00  5.45E+01  1.16E+02  2.65E+00  4.66E+01  1.22E+04  5.06E+00
9    1    3.97E+04  2.43E+03  2.77E+01  1.47E+03  1.17E+04  1.06E+01  2.76E+00  4.80E+01  1.28E+02  1.92E+00  9.16E+01  5.31E+03  2.84E+00
10   1    4.66E+04  1.89E+03  1.64E+01  1.73E+03  8.79E+03  9.64E+00  8.05E+00  3.58E+01  1.66E+02  1.98E+00  6.87E+01  1.33E+04  5.92E+00
11   1    3.69E+04  2.08E+03  1.78E+01  1.81E+03  1.21E+04  1.04E+01  8.09E+00  6.75E+01  1.33E+02  2.05E+00  3.74E+01  1.38E+04  6.24E+00
12   1    3.56E+04  1.62E+03  2.42E+01  1.57E+03  8.06E+03  9.91E+00  7.86E+00  3.33E+01  1.02E+02  2.40E+00  7.25E+01  1.42E+04  5.05E+00
13   1    3.52E+04  2.10E+03  2.62E+01  1.67E+03  9.11E+03  1.03E+01  4.93E+00  4.96E+01  9.36E+01  1.91E+00  3.97E+01  1.26E+04  5.74E+00
```

Fig. 1. The input file format for the MSAP-1.0 program.

II.1. Test of the program

The MSAP-1.0 program has been tested for validation by making comparisons under the same input data with the well-known SPSS-16.0 program [5]. The testing input data was selected from the elemental concentrations of five elements (As, La, Lu, Nd and Sm) for 61 geological samples, as seen in Table 1. The input file for MSAP-1.0 is shown in Fig. 2. The output results of eigenvalues, eigenvectors and principal components (PCA scores) obtained in the testing process in comparison with the SPSS-16.0 program are presented in Tables 2-3 and Figures 3-4.

Table 1. The input data of elemental concentrations used in testing the program.

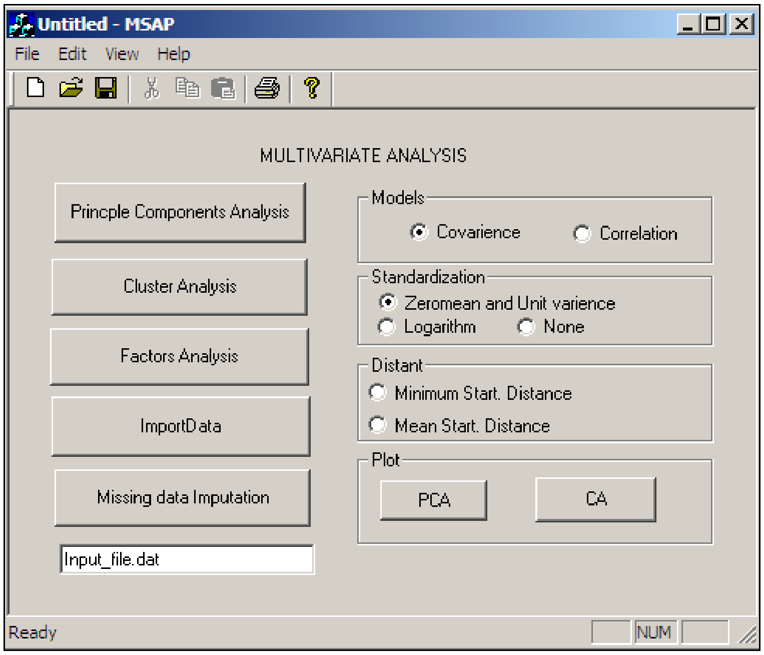
Sample	**As**	**La**	**Lu**	**Nd**	**Sm**	**Sample**	**As**	**La**	**Lu**	**Nd**	**Sm**
ID	**(ppm)**	**(ppm)**	**(ppm)**	**(ppm)**	**(ppm)**	**ID**	**(ppm)**	**(ppm)**	**(ppm)**	**(ppm)**	**(ppm)**
M1	6.902	35.822	0.3111	28.978	5.716	M32	7.404	35.28	0.2951	29.31	5.43
M2	8.898	34.58	0.3202	29.78	5.795	M33	7.474	33.66	0.303	30.02	5.481
M3	8.861	33.18	0.2924	32.444	5.534	M34	8.607	35.32	0.3226	31.95	5.851
M4	5.732	39.5	0.3218	31.35	6.141	M35	10.139	35.52	0.3261	30.98	5.702
M5	7.28	39.48	0.3203	32.91	6.161	M36	7.799	34.74	0.3132	30.39	5.627
M6	6.452	37.59	0.3469	31.31	5.973	M37	6.982	33.15	0.2945	28.19	5.247
M7	6.773	35.31	0.3026	30.29	5.485	M38	6.007	38.03	0.324	31.31	5.916
M8	8.507	35.17	0.3231	30.5	5.64	M39	10.213	36.7	0.3252	31.16	5.949
M9	7.122	31.93	0.2588	27.08	4.888	M40	7.651	34.52	0.2935	29.93	5.55
M10	5.663	36.4	0.2947	32	5.482	M41	6.599	33.72	0.275	26.46	5.184
M11	9.29	39.66	0.3186	33.41	5.929	M42	9.334	36.75	0.3308	28.84	5.942
M12	8.603	34.74	0.3215	30.63	5.664	M43	13.092	39.85	0.4643	36.02	6.879
M13	8.713	36.63	0.3049	32.46	5.604	M44	14.267	43.35	0.4602	36.28	7.37
M14	7.32	32.24	0.2925	27.77	5.189	M45	15.641	39.54	0.4097	33.68	6.559
M15	9.794	36.47	0.3047	31.01	5.692	M46	11.852	42.73	0.4461	33.33	6.956
M16	8.488	36.66	0.3313	32.39	5.808	M47	15.991	42.83	0.452	37.09	7.208
M17	7.764	37.37	0.3003	31.48	5.837	M48	12.598	39.59	0.4517	33.74	6.802
M18	7.061	33.43	0.3096	31.39	5.44	M49	15.394	42.73	0.4584	35.51	7.141
M19	8.677	36.63	0.3625	29.63	5.838	M50	14.755	42.57	0.4406	33.57	7.171
M20	8.399	35.78	0.337	30.06	5.627	M51	11.783	43.22	0.3836	36.97	6.994
M21	7.737	36.05	0.3185	28.55	5.576	M52	12.144	45.8	0.436	45.65	7.451
M22	6.458	29.46	0.2868	25.3	4.806	M53	13.392	46.42	0.4155	42.41	7.635
M23	7.195	31.74	0.2858	26.9	5.034	M54	10.401	49.9	0.4247	42.11	7.692
M24	10.135	38.33	0.3192	31.51	5.983	M55	15.039	52.74	0.4003	47.88	7.769
M25	8.5	36.84	0.3187	29.29	5.656	M56	13.576	44.4	0.4228	38.91	7.41
M26	8.802	35.88	0.3172	30.99	5.661	M57	11.513	44.91	0.4179	42.04	7.365
M27	8.953	35.67	0.3115	31.48	5.964	M58	10.522	41.39	0.4432	40.81	6.933
M28	7.923	35.09	0.309	28.15	5.682	M59	8.661	43.7	0.3953	44.17	7.23
M29	9.755	36.89	0.3147	32.37	6.021	M60	12.011	43.46	0.3932	38.67	7.229
M30	9.118	37.28	0.3305	32.89	6.044	M61	11.801	47.13	0.4294	43.01	7.723
M31	11.367	43.3	0.4781	39.39	7.492						

Table 2. Results of calculated eigenvalues (variance explained).

Component	SPSS-16.0 program			MSAP-1.0 program		
	Eigenvalue	% of variance	Cumulative %	Eigenvalue	% of variance	Cumulative %
1	4.3239	86.477	86.477	4.2530	86.477	86.477
2	0.4355	8.710	95.187	0.4284	8.710	95.187
3	0.1483	2.966	98.153	0.1459	2.966	98.153
4	0.0774	1.548	99.701	0.0761	1.548	99.701
5	0.0149	0.299	100.0	0.0147	0.299	100.0

Table 3. Results of calculated eigenvectors in comparison with SPSS-16.0 program.

Eigenvectors calculated with SPSS-16.0					Eigenvectors calculated with MSAP-1.0				
-0.413	-0.698	0.582	-0.053	-0.027	0.413	-0.698	-0.582	0.053	-0.027
-0.458	0.350	0.136	0.679	-0.434	0.458	0.350	-0.136	-0.679	-0.434
-0.447	-0.333	-0.747	-0.180	-0.313	0.447	-0.333	0.747	0.180	-0.313
-0.440	0.519	0.243	-0.683	-0.110	0.440	0.519	-0.242	0.683	-0.110
-0.476	0.102	-0.158	0.193	0.837	0.476	0.102	0.158	-0.193	0.837



Models
● Covarience ○ Correlation

Standardization
● Zeromean and Unit varience
○ Logarithm ○ None

Distant
○ Minimum Start. Distance
○ Mean Start. Distance

Input_file.dat

Fig. 2. The window interface of MSAP-1.0.

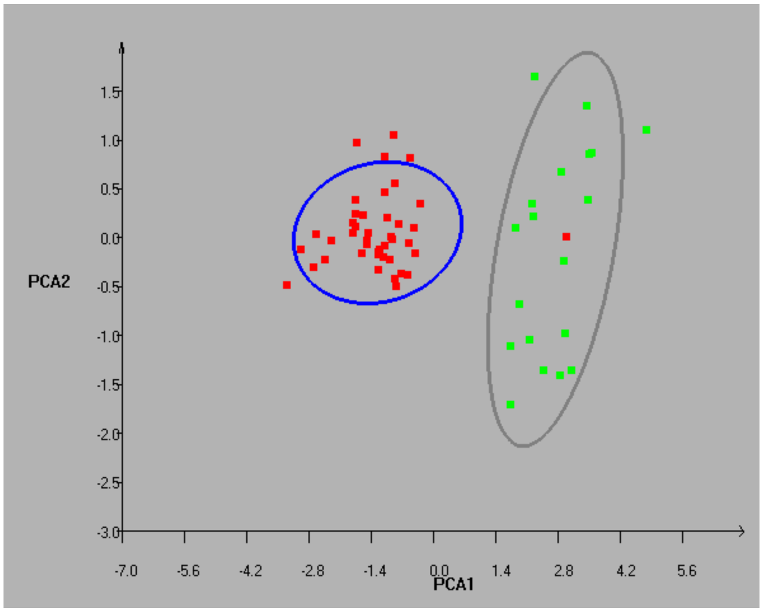


Fig. 3. Result of PCA scores plot by MSAP-1.0.

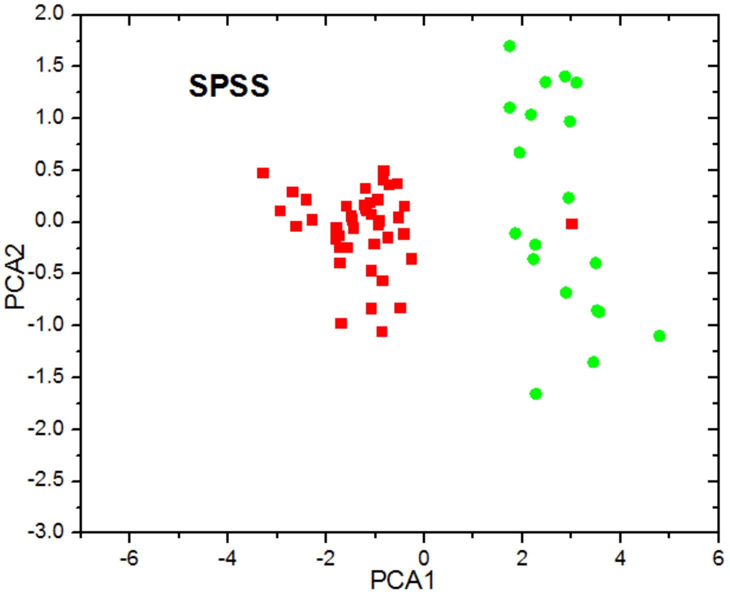


Fig. 4. Result of PCA sores plot from SPSS-16.0.

From the output values of the testing result above, both the programs SPSS-16.0 and MSAP-1.0, produce the similar output results of eigenvalues, eigenvectors and PCA scores. In some cases, the signs of eigenvectors from the two programs are different, but they are also identical because of random sign property of the eigenvector matrix. Based upon the comparison results, we can state that the MSAP-1.0 program can be used to conduct the principal component analysis, cluster analysis, standardization and output data plot for interpretations with multivariate data source.

II.2. Application of the program

The MSAP-1.0 program was applied for the PCA analysis and meaning interpretation of experimental multivariate data from NAA studies of archaeological samples collected from the archaeological sites in Lam Dong (Cat Tien) and Quang Nam (My Son) provinces of Vietnam. The results of principal component analysis for NAA data of the brick specimens and the clay samples collected in Cat Tien – Lam Dong province and the clay samples from Duy Xuyen – Quang Nam province are presented in the Figs. 5-6.

The results of calculations within PCA method are given in Fig. 5. From this figure, we can recognize the difference between the Group1 and Group2 which were from different provenances. In the projection view on the plane of PC1 versus PC3, Fig. 6, the difference between the clay group (group 2) and the brick group (group 1) is presented. The distinction can be explained by the fact that the concentrations of some elements were changed during burning process for brick. These elements could be Al, Mn, K, Ga, Fe, Co and Hf (PC3). The PC1 is contributed by the 8 main elements as follows: Al, Ti, V, Rb, Sc, Cr, Cs and Th; as the same way PC2 (7 elements): Dy, Na, La, Sm, Nd, Ce and Eu; PC3 (7 elements): Al, Mn, K, Ga, Fe, Co and Hf; and PC4 (4 elements): K, Lu, Tb and Yb. The values of loading factors larger than 0.25 are chosen (the bold values in Table 4).

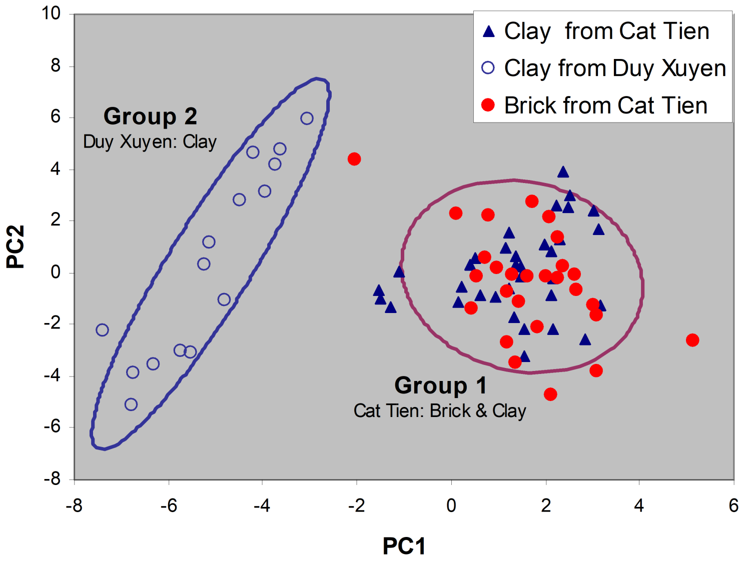


Fig. 5. Scattering plot of PC1 and PC2 for NAA data of Brick and Clay from Cat Tien and Clay from Duy Xuyen, Ellipses indicate 95% confidence limits.

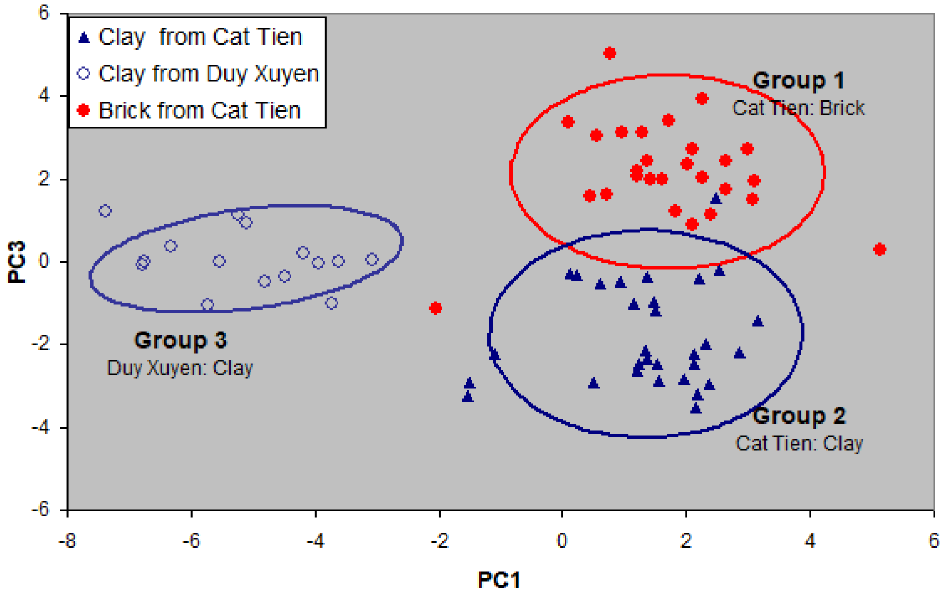


Fig. 6. Scattering plot of PC1 and PC3 for NAA data of brick and clay from Cat Tien and Duy Xuyen, Ellipses indicate 95% confidence limits.

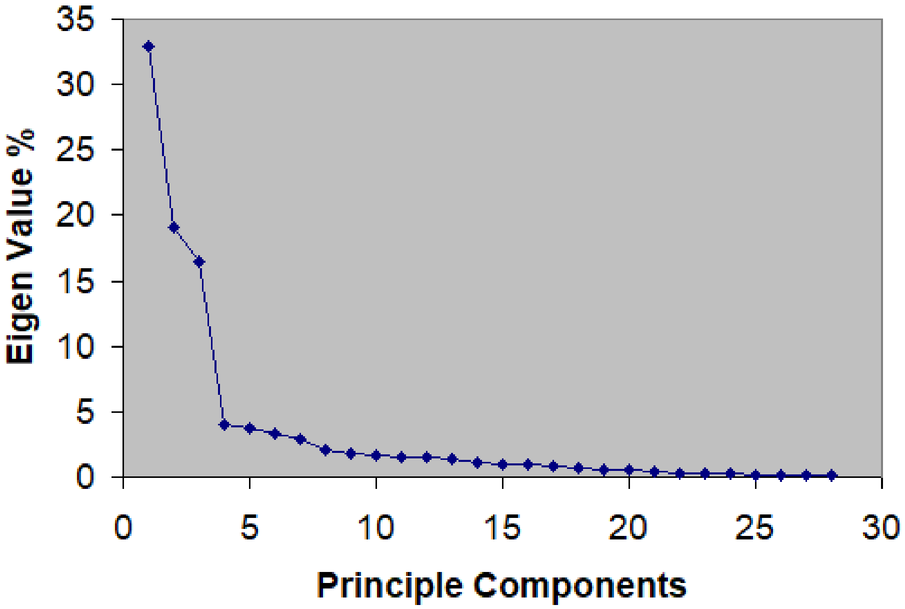


Fig. 7. Plot for eigenvalues of NAA data from Cat Tien’s brick & clay and Duy Xuyen’s clay.

Table 4. Contribution parameters for each element to the PC1, PC2, PC3 and PC4.

Elements	**PCA1**	**PCA2**	**PCA3**	**PCA4**
Al	**0.2554**	-0.0078	**-0.2493**	-0.0230
Ti	**0.2651**	-0.0089	0.1384	0.0202
V	0.2757	-0.0379	-0.1109	0.0299
Mn	0.0833	0.0577	**0.3619**	0.1811
Dy	0.0517	**0.3060**	0.1283	0.0668
Na	0.0605	**0.2618**	0.1424	0.2326
K	0.1600	0.1002	-0.2979	**0.2543**
Ga	0.1969	-0.0122	-0.2786	-0.0341
As	0.2374	-0.0343	0.2247	0.0692
La	-0.0871	**0.3380**	-0.1548	0.1540
Sm	0.0052	**0.3748**	-0.0655	0.1739
Rb	**0.2531**	-0.0493	-0.0052	0.1757
Ba	0.1740	0.0031	-0.0726	0.0303
Nd	-0.0382	**0.3661**	-0.0698	0.1563
Lu	0.0070	0.2428	0.2172	**-0.2504**
Sc	**0.2936**	0.0509	-0.1464	-0.0747
Cr	**0.2724**	0.0406	-0.0183	-0.1049
Fe	0.2124	0.0851	**0.2644**	-0.0321
Co	0.2246	0.0183	**0.2898**	0.0957
Sb	0.2364	-0.0109	-0.0548	-0.1139
Cs	**0.2911**	-0.0355	-0.1514	0.0695
Ce	-0.1224	**0.3530**	-0.0660	-0.1480
Eu	0.0162	**0.3830**	-0.0529	0.0589
Tb	0.0173	0.2239	-0.0587	**-0.2545**
Yb	0.0809	0.1650	-0.0003	**-0.7209**
Hf	-0.0078	0.0616	**0.4076**	-0.0011
Ta	0.2477	-0.0340	0.2305	-0.0547
Th	**0.2663**	0.0674	-0.0691	-0.0771

III. CONCLUSION

A new computer program, called MSAP-1.0, has been developed for multi-dimension data analysis and data reduction. The program has been tested by comparison with the SPSS-16.0 program. This new program was successfully applied for multivariate statistical analysis of INAA data in archeological studies at the Dalat Nuclear Research Institute.

REFERENCES

[1] A. C. Rencher, *Methods of Multivariate Analysis*, 2 End, Wiley-Interscience, A John Wiley & Sons, Inc. Publication, ISBN 0-471-41889-7, 2002.

[2] W. Hardle and L. Simar, *Applied Multivariate Statistical Analysis*, version 29th, MD-TECH, 2003.

[3] A. Buhagiar, *Exploration and reduction of data using principal component analysis*, Malta Medical Journal **14** (01) (2002) 27-35.

[4] P. Delicado, *Statistics in Archaeology: New directions, Computer Applications in Archaeology meeting (CAA 1998),* New Techniques For Old Times, 1998.

[5] SPSS Inc, Released 2007, SPSS for Windows, Version 16.0, Chicago, 2007.