A Classification of Obsidian Artifacts by Applying Pattern Recognition to Trace Element Data

Chul Lee*, Myung-Zoon Czae, Seungwon Kim, Hyung Tae Kang†, and Jong Du Lee‡

Department of Chemistry, Hanyang Univ., Seoul 133-791
†National Research Institute of Cultural Properties, Seoul 110-050
‡Korean Atomic Energy Research Institute, Seoul 139-240. *Received July 2, 1990*

Fifty eight obsidian artifacts and four obsidian source samples have been analyzed by instrumental neutron activation analysis. Artifact samples have been classified into classes by unsupervised learning techniques such as eigenvector projection and nonlinear mapping. The source samples have thereafter been connected to the classes by the supervised learning techniques such as SLDA and SIMCA so as to characterize each class by possible source sites. Some difference attributable to different nonlinear mapping techniques and the elemental effects on the separation between classes have been discussed.

Introduction

Considerable studies have been made on obsidian samples because, as a volcanic glass, the overall composition within a given flow tends to be homogeneous[1]. Therefore, in many cases the elemental composition is characteristic of the individual source. The samples are taken from sources as well as from archaeological sites. The multivariate statistical analyses of the elemental composition of various obsidian samples have been found to provide diagnostic patterns suggesting possibilities of classification of obsidian artifacts and hence the correlation of these artifacts with their source and history[1-3].

The multivariate statistical methods of analysis employed in this study are based on both supervised and unsupervised pattern recognition techniques. The unsupervised portion of the pattern recognition is constituted of mapping methods such as the eigenvector projection[4,5] and the nonlinear method[1,4,6]. The supervised pattern recognition employs a program called SIMCA, a program using disjoint principal component method to obtain models of known classes of objects, to which unclassified samples can be compared and identified[7]. This supervised pattern recognition also employs Fisher's discriminant analysis, a territorial mapping method of unclassified samples[7,8].

The classification of fifty eight obsidian artifacts have first been performed in this study by a nonlinear method in combination with the eigenvector projection. This part of study thus has been used as the unsupervised pattern recognition method. The results thus obtained have been used as a preliminary identification of the class membership. The supervised pattern recognition has thereafter been employed in this study for analyzing other objects in order to relate them to the predetermined classes.

Experimental

Apparatus. Gamma-ray counting was done with a HPGe detector(GEM-15180) with a dimension of 46.6 mm diameter×53.3 mm length, which is coupled with an ORTEC 8192 channel analyzer. Gamma-ray energy and peak areas have been calculated by ADCAM 100 soft wares stored in the computing system.

Table 1. Sampling Sites and Their Corresponding Symbols for Obsidian

Symbol	Series	Number of Samples	Sites	Feature
	1-8	8	Yondae-do, Tongyung, Kyungnam	Scraper
△	9, 10	2	Sangnodae-do, Tongyung, Kyungnam	Scraper
	11-19	9	Yokchi-do, Tongyung, Kyungnam	Scraper
	20-22	3	Dongsamdong, Pusan	Scraper
▲	23, 24	2	Yongsundong, Pusan	Scraper
	25	1	Pusan	Scraper
	26	1	Shinam-ni, Uljoo, Kyugnam	Scraper
▼	27	1	Angang, Wolsong, Kyungbuk	Unknown
	28-30*	3	Mt. Kumsung, Euiseong, Kyungbuk	
▽	31[a]	1	Osan-ni, Yangyang, Kangwon	
	32-55	24	Sangmooryong-ni, Yangku, Kangwon	Scraper Burin, Awl
■	56, 57	2	Yupan, Nanam, Hambuk	Scraper
□	58	1	Jiandao, Manzhou, China	Scraper
†	59*	1	Kyushu, Japan	
	60-62	3	Kagoshima, Kyushu, Japan	Scraper

Series numbers marked by asterisk mean obsidian sources. (a) means number

Sampling and Pretreatments. Obsidian artifacts and source samples collected for the elemental analyses are given in Table 1 together with corresponding symbols. The samples were washed with acetone as well as with dilute nitric acid. The samples were finally washed with distilled water and dried at 110 °C for 2 hours. An amount of 100-200 mg was cut by a tungsten-carbide edge from each sample and pounded to powder in an agate mortar. An about 40 mg

amount of each sample was a accurately weighed and sealed in a silica glass vial.

Neutron Activation Analysis. Each silica glass vial was attached on its surface by known amounts of Au and Co as single comparators[9]. Use of two nuclides with different nuclear properties facilitates the evaluation of effective cross sections of all nuclides involved in activation for a given conditions[10].

The vials were irradiated in the rotary specimen rack of TRIGA MARK III reactor for about 10 hours. After irradiation the samples were allowed to cool for 2 days. The surface of each silica vial was cleaned with dilute nitric acid and the sealed vial was placed on a given geometry of detector (the background of the vial was negligible). The same vial was recounted at another given geometry for longer nuclides after 1 weeks' cooling.

Gamma-ray energy and peak areas were calculated by ADCAM 100 soft wares as described above. Calculation of elemental contents was carried out as shown in the procedure described in the previous papers[9-11] by using flux indices at the irradiation condition, nuclear data given in references and counting efficiency curves at given geometries.

A total of 14 elements were determined. Those element analyzed and used for this study were sodium, potassium, samarium, lanthanum, cerium, scandium, thorium, cesium, hafnium, lutetium, rubidium, iron, terbium and antimony.

The Pattern Recognition Approach.

The classification of obsidian artifacts has generally been carried out by previous authors on the basis of their origin[1] *i.e.*, geographical location of obsidian sources. Unfortunately obsidian sources origined from South Korea are scarcely found, because it is located outside volcanic regions.

The first aim of this study was therefore put on the possible classification of artifacts on the basis of unsupervised learning techniques. For the purpose, display methods were applied. The reason for this is that the human is the best pattern recognizer in the familiar two- or three-dimensional space.

Unsupervised Learning.

Eigenvector Projection. The set of data on m samples with n variables measured ($m \times n$ data) can be represented as a set of m points in n dimensional space. The method starts by calculation of n variable means $\bar{X}_k (k = 1 \text{ to } n)$ as

$$X_k = \frac{1}{m} \sum_{i=1}^{m} X_{ik} \tag{1}$$

Next the covariance matrix C is generated, each element C_{ij} of whch is given as

$$C_{ij} = \sum_{i=1}^{m} (X_{ii} - \bar{X}_i)(X_{ji} - \bar{X}_j) \tag{2}$$

The eigenvalues λ_k and eigenvectors μ_k for $k = 1$ to n are then calculated by solving

$$C \cdot \mu_k = \lambda_k \cdot \mu_k \tag{3}$$

For display purposes in two dimensions, the basis vectors μ_1 and μ_2 corresponding to the two largest eigenvalues λ_1 and λ_2 are used as a projection plane. The eigenvector plots are the linear projections because the new two-space coordinates are the linear combinations of all the original coordinates[5]. Additionally, the projections are the best linear projections which can be obtained. Since such an eigenvalue is proportional to the variance along its corresponding eigenvector, the percent variance retained by the eigenvector projection can be calculated. This percent variance is useful for determining the reliability of interpretations made using a projection.

Nonlinear Mapping(NLM). A nonlinear map is a compression of n-dimensional hyperspace into two dimensions brought about with a minimum of distortion in the interpoint distances. This method, suggested by Sammon[6] and applied by Kowalski and Bender[4] is presented as follows.

The eigenvector plot described above is used as the starting configuration for NLM map.[4] All of the n-space interpoint distance, d^*_{ij} are calculated as

$$d^*_{ij} = [\sum_{k=1}^{n} (X_{ik} - X_{jk})^2]^{1/2} \tag{4}$$

and all of the two-space interpoint distance, d_{ij} are calculated as

$$d_{ij} = [\sum_{q=1}^{2} \{Y_{iq}(l') - Y_{jq}(l')\}^2]^{1/2} \tag{5}$$

where the Y's are found by the rotation matrix that diagonalizes C in equation (3). The object here is iteratively to change the two coordinates (Y_{i1} and Y_{i2}) for each point Y_i so as to minimize an error function E, defined as

$$E(\rho) = \sum_{i<j} \frac{\{d^*_{ij} - d_{ij}(l')\}^2}{(d^*_{ij})^\rho} \tag{6}$$

The minimization has been done to preserve interpoint distances by finding d_{ij}'s that are as close as possible to d_{ij}*'s. The value of ρ in equation (6) is fixed to 2 in this study[6]. The unknows in this error function are the two-space coordinates from equation (5).

In order to iteratively change the two-space coordinates and minimize E, a gradient method can be used. Sammon used the method of steepest descent. A software has been developed in this laboratory according to Sammon's suggestion and used in this study. The new two-space configuration at time $l' + 1$ is given by

$$Y_{nq}(l' + 1) = Y_{nq}(l') - (MF) \cdot \Delta_{nq}(l') \tag{7}$$

where

$$\Delta_{nq}(l') = \frac{\partial E(l')}{\partial Y_{nq}(l')} \Big/ \frac{\partial^2 E(l')}{\partial Y_{nq}(l')^2} \tag{8}$$

and MF is magic factor which is empirically given to be MF≈0.3~0.4. The partial derivatives are given by Sammon[6].

For NLM map, as described above, a plotting has been carried out by using two eigenvectors which are corresponding to two largest eigenvalues as the starting configuration[4]. However Sammon's original paper suggested to use two original coordinates with largest variances as starting configurations and the suggestion was once adopted by

Table 2. Mean Values (ppm) and Standard Deviations of Elemental Contents in Obsidian

Elements	Na	K	Sm	La	Ce	Sc	Th
Mean values	3.79×10^4	2.57×10^4	6.61	47.6	78.4	2.47	22.1
Standard deviations	7.04×10^4	1.04×10^4	6.51	36.4	58.1	4.04	15.3
Elements	**Cs**	**Hf**	**Lu**	**Rb**	**Fe**	**Tb**	**Sb**
Mean values	6.36	5.09	0.413	225	1.14×10^4	1.08	0.337
Standard deviations	4.48	4.29	0.292	98.2	6.00×10^4	1.07	0.331

Boulle and Peisach[12] for the visual display of the archaeological artifacts.

In this study both processes have been adopted as the nonlinear mapping approaches and have been compared. The results showed some difference between processes as discussed below under "Results and Discussion".

Supervised Learning.

SIMCA(Statistical Isolinear Multiple Component Analysis). From the unsupervised learning, similar obsidian artifacts that may represent a class can be identified. These artifacts can then be modeled as linear structures using SIMCA[7]. The concept is to fit each class with a minimal set of principal components sufficient to reproduce the variance structure within that class. By comparing the residuals of the fit of an object to the class model with the average value of that fit for objects belonging to the class, the probability of the particular object to belong to the particular class can be tested. This modeling approach permits the testing of the class assignments derived from the unsupervised learning and provides an added measure of confidence in the validity of the assigned classes.

SLDA (Statistical Linear Discriminant Analysis). The discriminant function Y_i is a linear combination of original or autoscaled variables and is given in the form:

$$Y_i = v_1 z_{i1} + v_2 z_{i2} + \cdots + v_j z_{ij} + \cdots + v_n z_{in} \tag{9}$$

In vector notation, equation (9) can be written as

$$Y_i = V_j \cdot Z_{ij} \tag{10}$$

where V_j is the set of discriminant coefficients and Z_{ij} is the sample vector of individual i. The projection of variables Z_{ij} onto the vector described by discriminant coefficients V are given by $Y = V \cdot Z$. In order to determine the discriminant coefficients, the following equation (11) can be used according to Fisher's criterion.

$$W^{-1}BV = LV \tag{11}$$

where W and B are the within-class and between-class dispersions for the projected points. From this equation, it is recognized that the coefficients V are given by eigenvector coefficients of matrix $W^{-1}B$ and L is the corresponding eigenvalue. Since L is defined as the ratio of the between-class dispersion to the within-class dispersion to obtain the maximum discrimination, the eigenvector associated with the largest eigenvalue of matrix $W^{-1}B$ should be used as discriminant coefficients.

In a two class problem, it is not necessary to solve for the eigenvalues of $W^{-1}B$, as the vector given by BV is in the same direction as the vector D of difference between two means[13]. If the vector D of the difference is defined as

$$D = \mu_2 - \mu_1 \tag{12}$$

a new set of discriminant coefficients V_2 is obtained by

$$V_2 = W^{-1} D \tag{13}$$

A computational example of how to determine V_2 in practice is given by Kendall[14]. Obsidian sources have been classified according to their positions on the discriminant axis as compared to the position of centroids of the classes of obsidian artifacts.[7,15]

A pattern consisting of many parameters often contains a lot of noise, *i.e.*, redundant parameters. To trace redundant variables several criteria are available[15]. Criteria based on discriminant functions give a large importance to a variable when the absolute value of the corresponding weight coefficient is higher, provided that the variable have been standardized. A direct method is to determine the contribution percentage of each variable to the total distance D^2 in the discriminant space, which is the distance between the centroids of the groups considered. The contribution percentage of variable j is given as

$$100\times | V_j \cdot \delta_j | / D^2 \tag{14}$$

where V_j is the weight coefficient of the discriminant function for jth variable and

$$\delta_j = \frac{x_{p\cdot j} - x_{r\cdot j}}{\sigma_{\cdot\cdot j}} \tag{15}$$

where $\sigma_{\cdot\cdot j}$ is the overall standard deviation of the jth variable. And $x_{p\cdot j}$ and $x_{r\cdot j}$ are mean values of jth variable in group p and r, respectively.

Thus

$$D^2 = \sum_{j=1}^{n} | V_j \cdot \delta_j | \tag{16}$$

Results and Discussion

Forteen elements which were analyzed and used in this pattern recognition are given in Table 2 together with the means and standard deviations of the elemental contents for overall samples. The effect of very differing data ranges and variances of the various measured variables have been com-

Table 3. Eigenvalues and Their Contribution (%)

Components	1	2	3	4	5	6
Eigenvalues	6.18	2.65	1.90	1.07	0.66	0.53
Cumulative Contribution (%)	44.1	63.1	76.6	84.3	89.0	92.8

pensated by autoscaling the measured variables to produce "features" with means of zero and variances of unity[8] using the data in Table 2.

The covariance matrix in equation (2) has been generated from the autoscaled data set. Forteen eigenvalues have been found by equation (3). The eigenvalues which are supposed to be important are given in Table 3 together with the contribution of each principal component to total variance. The eigenvector corresponding to each eigenvalue, normalized with the sum of the squares being one, has been calculated and are given in Table 4. The communalities of the input variables in each of the derived components, *i.e.*, the fraction of total variance accounted for in each component, are also given in Table 4 to show the importance of a given element in a component.

An eigenvector projection plot is given in Figure 1 which is a plot of component 1 *vs.* component 2. Even though the two different classes are clearly shown, the used percent variance, given as $\%V = 100(\lambda_1 + \lambda_2)/\sum_{i=1}^{n} \lambda_i$ is only 63% and remaining 37% of the variance is lost in the projection.

The nonlinear mapping by error minimization has been supposed to be the most useful display of multidimensional data in a two-dimensional space. Two space coordinates are derived iteratively by equation (7). Figure 2 is the NLM of obsidian artifacts and source samples after allowing a certain number of iteration. The mapping error defined by equation (6) is also given in the figure. For this NLM, the eigenvector plot described for Figure 1 has been used as the starting configuration as applied by Kowalski and Bender[4]. Another NLM has been done by using two coordinates with largest variances as starting configurations[6] and the results are given in Figure 3 for the comparison. In both figures classes are separated but classification of unknown is more reliable using Figure 2 because the actual data structure is more faithfully represented in this figure.

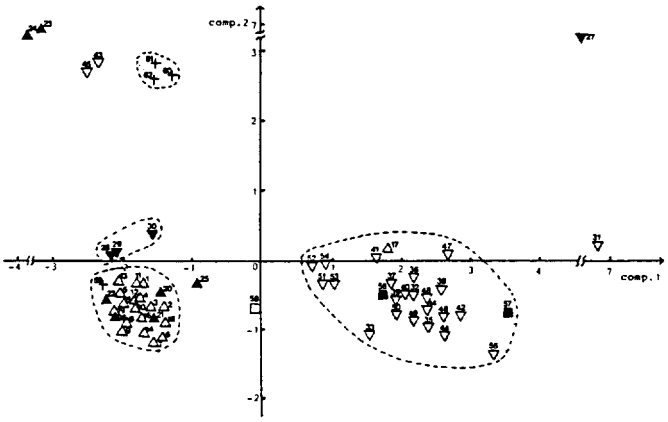

Figure 1. The result of eigenvector projections.

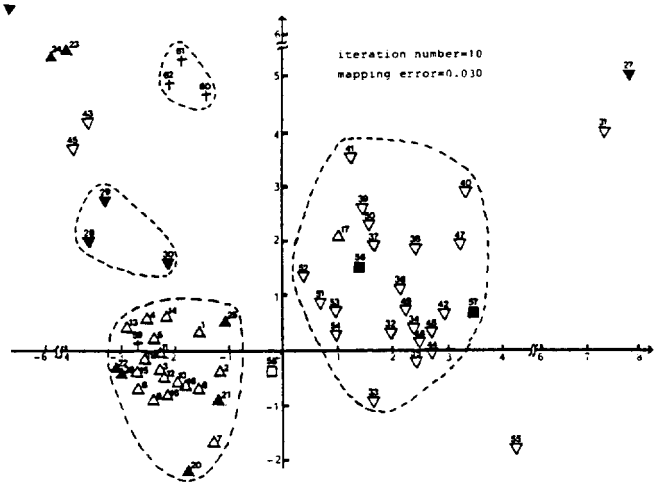

Figure 2. Nonlinear mapping from eigenvector projections using 14 elements.

Table 4. Eigenvector Coefficients, Communalities and Percentage of Variance for Each Element

Elements	Eigenvector coefficients		Communalities		Percentage of variance accounted for in	
	Comp. 1	Comp. 2	Comp. 1	Comp. 2	Comp. 1	Comp. 2
Na	-0.121	0.221	0.090	0.130	9.0	13.0
K	0.238	-0.344	0.350	0.314	35.0	31.4
Sm	0.345	0.252	0.737	0.169	73.8	16.9
La	0.368	0.099	0.921	0.026	92.2	2.6
Ce	0.342	-0.179	0.720	0.085	72.0	8.5
Sc	0.049	0.548	0.015	0.797	1.5	79.8
Th	0.270	-0.211	0.451	0.118	45.1	11.8
Cs	-0.174	-0.249	0.187	0.164	18.7	16.4
Hf	0.333	-0.055	0.684	0.008	68.5	0.8
Lu	0.345	0.221	0.734	0.130	73.4	13.0
Rb	0.279	-0.404	0.482	0.433	48.2	43.3
Fe	0.055	-0.055	0.019	0.170	1.9	17.0
Tb	0.351	0.199	0.759	0.106	76.0	10.6
Sb	-0.065	-0.036	0.026	0.003	2.6	0.3

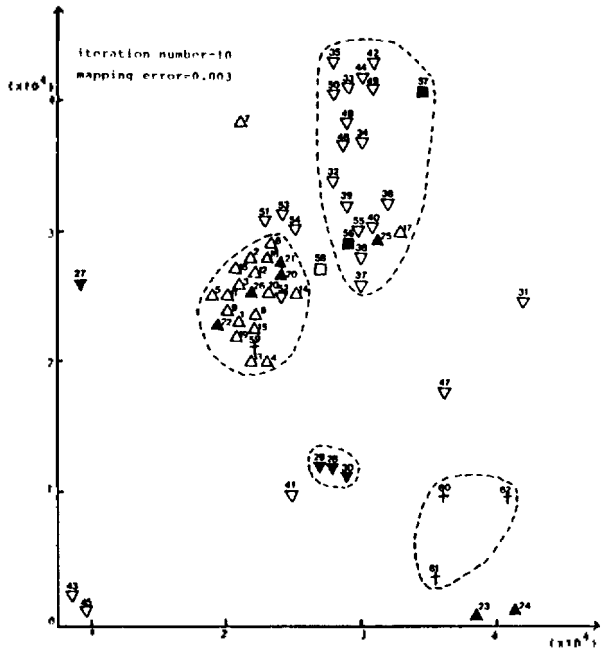

Figure 3. Nonlinear mapping using two coordinates with largest variances.

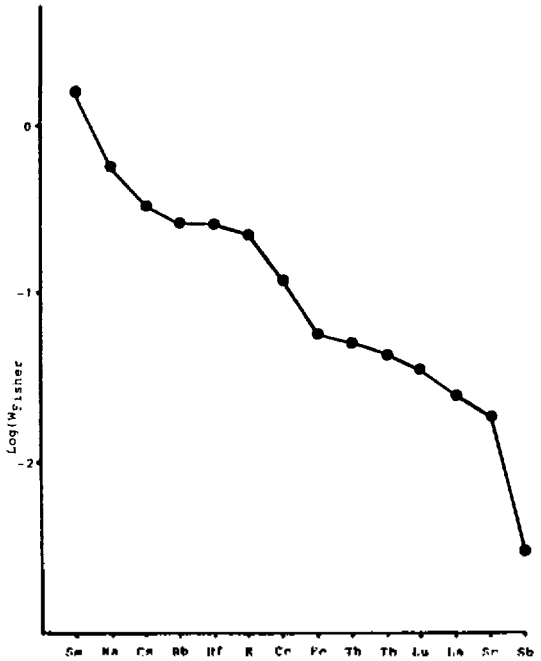

Figure 4. Fisher weights defined by equation (14).

As shown in Figure 2 and Table 1, obsidian artifacts are separated into two classes, *i.e.*, the class △(1–19) and ▲(20–26) from Kyungnam–Pusan and the class▽(31–35) from Kangwon. Figure 2 shows also that three source samples † (60–62) from Kyushu, Japan and three source samples ▼ (28–30) from Kyungbuk are supposedly making respective isolated classes but need a further study because of small number of source samples.

SLDA has been carried out for the linear separation of artifacts into two main classes (*i.e.*, Kyungnam–Pusan and Kangwon) as well as for possible similarity between source samples and classes. Since two class problem is involved here, the discriminant coefficient V_2 was estimated by equation (13). Obsidian source samples have been classified according to their position on the discriminant axis as compared to the positions of centroids of the classes. As the results a source sample († 59 in Table 1) from Kyushu, Japan was found to belong to the artifact class from Kyungnam–Pusan. The results show also that two artifacts ■ (56, 57) from Hambuk belong to the artifact class from Kangwon.

To trace redundant variables, Fisher weights of elements defined as equation (14) have been calculated for the separation between two main classes in Figure 2. For this purpose, mean values and eigenvector coefficients used in the calculation are given in Table 5. Fisher weights against elements are given in Figure 4. From the figure, seven elements such as Sm, Na, Cs, Rb, Hf, K and Ce have been selected as important elements, *i.e.*, giving high contribution to the total distance D^2 between classes. Using these elements, all obsidian samples have been reclassified for NLM and the results are given in Figure 5. It can be noted that an almost identical display of the sample is obtained as Figure 2 which is obtained using all elements.

The two main classes of artifacts given in Figure 2 have been modeled as linear structures using SIMCA[7]. By comparing the residuals of the fit of source samples with the class model the probability of the particular source samples to belong to the particular class has further been estimated. The results showed that three source samples † (60–62) from Kyushu, Japan and three source samples ▼ (28–30) from Kyungbuk were found to be outliers from both classes and were forming respective isolated classes. These results are in accord with those obtained by NLM.

A source sample † (59) from Kyushu, Japan was found to belong to the artifact class from Kyungnam–Pusan. The results showed also that two artifacts ■ (56, 57) from Hambuk were belong to the artifact class from Kangwon. These re-

Table 5. Data to Calculate Fisher Weights

Elements		Na	K	Sm	La	Ce	Sc	Th
Mean values	Kyungnam	2.30×10^4	2.61×10^4	2.48	21.4	30.9	1.33	15.5
(ppm)	Kangwon	2.82×10^4	3.25×10^4	9.67	72.7	143	1.12	30.0
Eigenvector coefficients		−0.404	−0.262	−0.828	0.313	−0.063	0.018	−0.023
Elements		Cs	Hf	Lu	Rb	Fe	Tb	Sb
Mean values	Kyungnam	9.51	2.07	0.255	201	7020	0.459	0.470
(ppm)	Kangwon	3.24	8.03	0.521	302	12100	1.66	0.286
Eigenvector Coefficients		0.178	−0.134	−0.020	0.159	−0.034	0.030	0.005

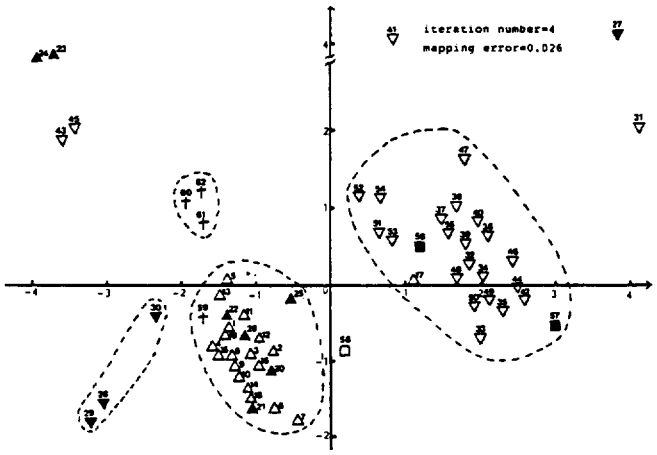

Figure 5. Nonlinear mapping from eigenvector projections using 7 elements.

sults are in accord with those obtained by SLDA as well as NLM.

Acknowledgement. The present studies are supported by Basic Science Research Institute Program, Ministry of Education, 1989.

References

1. B. R. Kowalski, T. F. Schatzki, and F. H. Stross, *Anal. Chem.*, **44**, 2176 (1972).
2. Ph. K. Hopke, R. C. Martin, and M. A. Evins, *J. Radioanal. and Nucl. Chem. Articles*, **112**, 215 (1987).
3. J. R. Bird, L. H. Russell, M. D. Scott, and W. R. Ambrose, *Anal. Chem.*, **50**, 2082 (1978).
4. B. R. Kowalski and C. F. Bender, *J. Am. Chem. Soc.*, **95**, 686 (1973).
5. C. Lee, O. C. Kwun, and H. T. Kang, *Bull. Kor. Chem. Soc.*, **7**, 73 (1986).
6. J. W. Sammon, Jr., *IEEE Trans. Comput.*, **C-18**, 401 (1969).
7. C. Lee, H. T. Kang, and S. Kim, *Bull. Kor. Chem. Soc.*, **9**, 223 (1988).
8. C. Lee, O. C. Kwun, S. Kim, I. C. Lee, and N. B. Kim, *Bull. Kor. Chem. Soc.*, **7**, 347 (1986).
9. C. Lee, *J. Kor. Nucl. Soc.*, **5**, 137 (1973).
10. J. I. Kim, I. Feidler, H. J. Born, and D. Lux, *Internat. J. of Environ. Anal. Chem.*, **10**, 135 (1981).
11. J. I. Kim, *J. Radioanal. Chem.*, **63**, 121 (1981).
12. G. J. Boulle and M. Peisach, *J. Radioanal. Chem.*, **50**, 205 (1979).
13. W. W. Cooley and P. R. Lohnes, "Multivariate Data Analysis", Wiley, New York, 1971.
14. M. Kendall, "Multivariate Analysis", Charles Griffin and Co., London, 1975.
15. D. Coomans, D. L. Massart, and L. Kaufman, *Anal. Chim. Acta.*, **112**, 97 (1979).

A Synthetic Study on *Trans*-2,5-Disubstituted Tetrahydrofurans via Phenylselenoetherification

Sung Ho Kang*

Department of Chemistry, Korea Advanced Institute of Science and Technology, Taejon 305-701

Tae Seop Hwang and Joong Ki Lim

College of Pharmacy, Sung Kyun Kwan University, Suwon 440-746

Wan Joo Kim

Korea Research Institute of Chemical Technology, Taejon 305-606. *Received July 6, 1990*

2,5-Disubstituted tetrahydrofurans **11-13** were prepared by phenylselenoetherification of 1-alkyl-4-phenyl-(3E)-butenols **8-10** under kinetic conditions. Their stereochemical outcome and reactivity were controlled by solvent, reaction temperature and the alkyl substituent. While the cyclization was stereorandom in dichloromethane, its stereoinduction was moderate to good in propionitrile and good to excellent in diethyl ether. The reaction went to completion in dichloromethane and propionitrile, but it did not in diethyl ether. The results can be rationalized by the degree of reversibility in the formation of episelenonium cation and 1,3-diaxial interactions in the transition state of the formation of tetrahydrofuranonium cation.

Introduction

Since 2,5-disubstituted tetrahydrofurans are crucial structural units in many natural products such as polyether antibiotics,[1] furanoterpenes[2] and polyene mycotoxins,[3] there has been increasing interest in the synthesis of the ring sys-