한국통계학회논문집
2012, 19권, 4호, 559–569

DOI: http://dx.doi.org/10.5351/CKSS.2012.19.4.559

# Comparison of Methods for Reducing the Dimension of Compositional Data with Zero Values

Taegyoun Song[a], Byungjin Choi[1,b]

[a]LRIS Consulting; [b]Department of Applied Information Statistics, Kyonggi University

**Abstract**

Compositional data consist of compositions that are non-negative vectors of proportions with the unit-sum constraint. In disciplines such as petrology and archaeometry, it is fundamental to statistically analyze this type of data. Aitchison (1983) introduced a log-contrast principal component analysis that involves logratio transformed data, as a dimension-reduction technique to understand and interpret the structure of compositional data. However, the analysis is not usable when zero values are present in the data. In this paper, we introduce 4 possible methods to reduce the dimension of compositional data with zero values. Two real data sets are analyzed using the methods and the obtained results are compared.

Keywords: Compositional data, dimension-reduction, log-contrast principal component analysis, correspondence analysis, ranked data, quantification method.

## 1. 서론

암석학에서는 암석들의 지구화학적 구성물의 분석이 기본이 되고 고고학에서는 도기와 유리 유물과 같은 고고학적인 제재의 화학적 구성물의 분석이 기본이 된다. 그런 구성물은 일반적으로 각 표본에 포함된 전체 성분에서 관심의 대상이 되는 중요한 성분들이 차지하는 정도를 비율로 나타내게 된다. 합이 1이 되는 $p$개의 양의 성분 $x_1, x_2, \ldots, x_p$들로 구성되는 $\mathbf{x} = (x_1, \ldots, x_p)'$는 구성비율벡터(composition)가 되고 이런 구성비율벡터를 $n$개의 표본으로부터 측정해서 얻은 다변량 자료를 구성비율자료(compositional data)라고 한다.

차원축약에 의한 구조적 단순화와 요약을 통해 저차원상에서 구성비율자료를 해석하고 변동을 설명하고자 Webb과 Briggs (1966), Le Maitre (1968)와 Butler (1976)는 주성분분석을 시도한 바가 있다. Le Maitre (1968)와 Butler (1976)의 방법은 모든 성분들의 공분산행렬에 기초를 두고 있고 제약조건에 의해 공분산들이 음의 편향을 가짐으로 인해 유도된 주성분들이 독립이 되지 않는 문제를 안고 있다. 또한 구성비율자료는 Gnanadesikan (1977)의 관념에서 선형이 아닌 패턴을 보이는 경향이 있고 원 성분들의 선형결합으로 주어지는 주성분들이 비선형적인 변이를 잘 포착하지 못하게 된다. 변환된 성분 $y_i = x_i/x_p$, $i = 1, \ldots, p-1$들의 공분산행렬에 기초를 두고 있는 Webb과 Briggs (1966)의 방법은 Le Maitre (1968)와 Butler (1976)의 방법에서 제약조건으로 인한 문제들이 해결되고 있지만 변환에서 공통분모로 사용되는 $x_p$의 선택에 따라 서로 다른 주성분들이 유도가 되는 문제가 있다.

Aitchison (1982)은 구성비율자료의 통계분석에서 제약조건으로 인해 야기되는 여러 문제들을 지적하면서 문제 해결을 위해 성분들의 로그비 변환에 기초한 새로운 분석 방법론을 제시했다. 제안한

[1] Corresponding author: Associate Professor, Department of Applied Information Statistics, Kyonggi University, Iui-Dong, Yeongtong-Gu, Suwon-Si, Gyeonggi-Do 443-760, Korea. E-mail: bjchoi92@kyonggi.ac.kr

방법론의 유용성을 보이기 위한 노력의 일환으로 Aitchison (1983)은 구성비율자료에 대한 로그대비 주성분분석(log-contrast principal component analysis; LPCA)를 소개했다. 이 방법은 $z_i = \log\{x_i/g(\mathbf{x})\}$, $i = 1, \ldots, p$로 변환된 성분들의 공분산행렬에 기초를 두고 있고 변환에서 사용한 분모는 원 성분들의 기하평균으로 $g(\mathbf{x}) = (\prod_{i=1}^{p} x_i)^{1/p}$이다. Aitchison (1983)은 로그대비 주성분분석의 다양한 이점들을 언급함과 동시에 유도된 주성분들이 구성비율자료가 보이는 비선형적인 변이를 포착하는데 있어 다른 방법들보다 우월함을 보여주는 실제 자료분석의 결과를 제시했다. 그러나, 로그대비 주성분분석이 기존의 주성분분석에서 나타나는 문제들을 해결하기 위한 대안적 방법으로 여러 장점을 가지고 있다고는 하지만 원 성분들이 영값을 가지는 경우에는 처리할 수 없는 단점이 있다.

실제 응용에서 얻게 되는 구성비율자료에는 영값들이 나타나는 경우가 흔히 발생한다. 이런 구성비율자료의 구조를 저차원상에 표현된 형상을 통해 탐색하고 해석하고자 한다면 적절한 차원축약기법을 사용해야만 한다. 그런 기법들 중에서 로그대비 주성분분석을 적용하고자 한다면 영값들의 처리가 필수적이다. Aitchison (1986)은 $c$개의 영값들이 있는 구성비율벡터에서 영값들에는 동일한 작은 양수를 더해 주고 나머지 $(p-c)$개의 값들은 동일한 값을 빼주어서 조정된 성분들의 합이 1이 되게 하는 가법교체법을 소개했다. 가법교체법은 단순하지만 이렇게 해서 얻은 구성비율벡터는 영값들의 대치에 사용된 양수의 선택에 따라 달라지게 되므로 어떤 양수를 사용해야 하는지에 대한 판단이 어려운 문제가 있다. Martin-Fernandez 등 (2003)은 가법교체법의 대안으로 승법교체법(multiplicative replacement)을 제안했다. 이 방법은 구성비율벡터가 $c$개의 영값들을 가지는 경우에 $c$개는 각각 $\delta_1, \ldots, \delta_c$로 대치하고 $(p-c)$개는 각각의 원래 값에다가 $1 - \sum_{i=1}^{c} \delta_i$를 곱한 값으로 대치하여 영값이 없는 구성비율벡터를 만들게 된다. 승법교체법 또한 가법교체법과 마찬가지로 영값들의 대체에 사용할 $\delta_1, \ldots, \delta_c$를 객관적으로 선택하기가 쉽지 않다.

Bacon-Shone (1992)은 영값들을 처리하기 위해서 순위에 기초하여 새로운 구성비율자료를 만들고 이 자료에 Aitchison (1982)의 방법론을 적용할 것을 제안했다. 이 방법에서는 적절한 순위부여방식에 따라 각 자료값들을 순위로 바꾼 다음, 각 순위를 대응되는 행의 합으로 나누어서 행의 합이 1이 되는 구성비율자료를 얻는다. Bacon-Shone (1992)은 실제 자료분석의 결과를 통해 Aitchison (1986)이 고려한 방법의 대안으로 가능한 새로운 접근방법이 될 수 있음을 언급했지만 주성분분석에는 직접 적용해 보지는 않았다. Baxter 등 (1990)은 대응분석에서 분석 대상이 되는 분할표가 구성비율자료와 유사한 점에 착안하여 로그대비 주성분분석이 적절한 방법으로 변환된 구성비율자료의 근사적인 대응분석으로 간주될 수 있음을 밝힌 바가 있다. 영값이 없는 실제 구성비율자료를 분석한 결과를 통해 그들의 방법이 로그대비 주성분분석과 유사함을 보였고 자료에 영값들이 있을 때에도 적용 가능하기 때문에 로그대비 주성분분석보다는 더 선호될 것이라고 주장했다. Huh (1999)는 순위자료에 대한 다변량 수량화 방법을 소개한 바가 있고 이 방법은 각각의 원 자료값을 행내에서의 순위로 대체시킨 순위자료에 대한 주성분분석으로 볼 수가 있다. 영값들이 있는 구성비율자료에도 쉽게 적용이 가능하므로 위에서 열거한 방법들의 대안으로 고려해 볼 수 있다.

본 논문에서는 영값들이 있는 구성비율자료의 차원축약을 위해 사용 가능한 방법들을 비교해 보고자 한다. 고려한 방법들은 Aitchison (1986)의 가법교체법을 이용한 로그대비 주성분분석, Bacon-Shone (1992)의 순위변환을 통한 구성비율자료를 이용한 로그대비 주성분분석, Baxter 등 (1990)의 대응분석과 순위자료에 대한 수량화 방법들이고 승법교체법을 이용한 로그대비 주성분분석은 가법교체법과 달리 영값이 여러 개가 있을 경우에 각각의 대치값으로 사용할 $\delta$들을 선정하는 것이 쉽지 않아서 제외하기로 한다. 본 논문의 구성은 다음과 같다. 2절에서는 4가지 분석 방법의 각각을 간략히 기술하고 3절에서는 비교를 위해서 실제 자료에 대해 각 방법을 적용해서 얻은 분석 결과를 제시한다. 끝으로 4절에서는 결론을 맺는다.

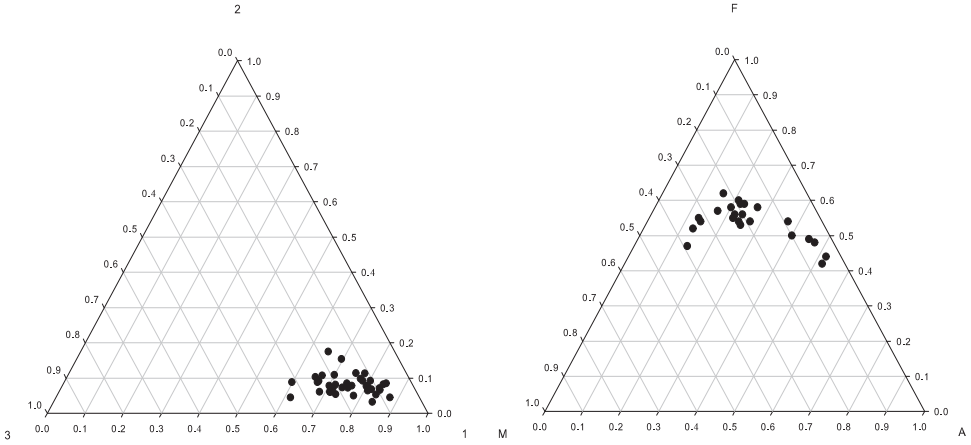


Figure 1: *Ternary diagrams for steroid and Skye lava compositions*

## 2. 분석방법들

고차원적인 구성비율자료를 저차원상에 표현된 형상을 통해 자료의 구조를 해석하기 위한 차원축약방법으로 가장 많이 사용되는 것이 주성분분석이다. 주성분분석은 대수적으로는 $p$-차원 유클리드 표본공간 $R^p$에서 $p$개의 변수들의 선형결합으로 표현되는 성분들의 분산을 최대화하는 정규직교선형결합의 결정이며, 기하적으로는 $p$-차원 공간에 흩어져 있는 점들을 가장 잘 적합시키는 평면을 찾는 것을 의미한다. 구성비율벡터가 가지는 제약조건은 주성분분석에서 별 문제를 일으키지 않을 것처럼 보인다. 하지만, 구성비율벡터의 표본공간인 심플렉스공간 $S^{p-1} = \{(x_1, \ldots, x_p) : x_1 > 0, \ldots, x_p > 0; \sum_{i=1}^{p} x_i = 1\}$이 유클리드 표본공간 $R^p$의 부분집합으로 생각하여 유클리드 거리와 직교성을 이용해 기하적인 면에서 주성분분석을 사용한다 해도 실질적인 해석상의 어려움이 있음을 Gower (1967)는 지적한 바가 있다. 직교성은 원 변수들의 선형결합으로 이루어지는 주성분들의 상관계수는 0이 되는 독립성과 관계가 있다. 그러나, 구성비율벡터의 제약조건은 변수들간의 상관이 음의 값쪽으로 편향을 가지게 하여 주성분들간의 상관계수가 0이라는 것이 곧 독립성을 의미하지는 않음을 Aitchison (1983)은 역설한 바가 있다.

구성비율자료의 비선형성을 확인하기 위한 일반적인 방법은 없지만 $p = 3$인 경우에는 삼각좌표(triangular coordinate)체계를 이용한 플롯을 통해서 파악할 수가 있다. Figure 1의 왼쪽은 37명의 성인으로부터 소변배설물에서 측정한 세 종류의 스테로이드 성분(1: total cortisol metabolites, 2: total corti-costerone metabolites, 3: pregnanetriol+$\Delta$-5-pregnentriol)에 대한 구성비율자료를 나타낸 것이고 오른쪽은 스코틀랜드 서부에 위치한 Skye섬의 시신세기에 형성된 23개의 화산암 표본으로부터 측정한 세 성분(A: $Na_2O$, F: $Fe_2O_3$, M: $MgO$)에 대한 구성비율자료를 나타낸 것이다. 왼쪽 그림은 Gnanadesikan (1977)에 의하면 비교적 타원의 형태를 가지는 선형적인 모습을 보이는 반면, 오른쪽 그림은 어떤 가상적인 곡선의 형태로 변이가 나타나는 비선형의 형태를 보인다. Aitchison (1983)은 이 그림들을 통해서 비선형성을 보이는 구성비율자료의 변이를 원 변수들의 선형결합으로 이루어지는 주성분들로는 설명할 수 없음을 언급하면서 새로운 형태의 주성분의 구축이 필요함을 주장하였다.

구성비율벡터 $\mathbf{x}$에 대한 주성분분석을 수행했을 때 나타나는 문제들을 해결하기 위한 노력의 일환으로 Aitchison (1983)에 의해 제안된 로그대비 주성분분석은 구성비율벡터 $\mathbf{x}$ 대신에 $z_j = \log x_j - \sum_{i=1}^{p} \log x_i/p$, $j = 1, \ldots, p$를 원소들로 가지는 변환된 새로운 벡터 $\mathbf{z}$에 대해 수행하는 것이다. $z_j$들의 선

형결합으로 주어지는 주성분 $y = a_1 z_1 + \cdots + a_p z_p$는 주성분계수들의 제약으로 $\sum_{i=1}^{p} a_i = 0$을 주게 되면 $y = a_1 \log x_1 + \cdots + a_p \log x_p$로 표현이 된다. $\mathbf{z}$의 공분산행렬 $\mathbf{\Sigma_z}$는 양반정치 행렬로 $p$개의 고유값들 중에서 $(p-1)$개의 고유값들은 $\lambda_1 > 0, \ldots, \lambda_{p-1} > 0$이며 가장 작은 고유값은 $\lambda_p = 0$이 된다. 주성분계수들은 이들 고유값들에 대응하는 고유벡터를 사용하게 된다.

그러나, 로그대비 주성분분석은 구성비율벡터의 원소가 영값을 가지는 경우에는 적용할 수가 없는 문제가 있다. 이 문제에 대한 해결책으로 고려할 수 있는 방법은 영값을 적당하게 작은 값으로 교체하여 만든 조정된 구성비율벡터를 사용하는 것이다. 이런 목적으로 Aitchison (1986)에 의해 제안된 가법교체법(additive replacement)은 구성비율벡터가 $c$개의 영값과 $(p-c)$개의 0이 아닌 값을 가진다면 영값들에는 $\delta(c+1)(p-c)/p^2$을 더해주고 0이 아닌 값들에는 $\delta c(c+1)/p^2$을 빼주는 것으로 $\delta$는 허용이 가능한 최대마무리오차이다. 예를 들면, $p = 3$인 구성비율벡터가 $\mathbf{x} = (0.55, 0.45, 0.00)'$로 주어진 경우에 조정된 구성비율벡터는 다음과 같이 얻을 수 있다: $c = 1$, $p - c = 2$이므로 영값에는 $4\delta/9$를 더해주고 영이 아닌 값에는 $2\delta/9$를 빼주면 된다. 그러므로, 합이 1이 되는 조정된 구성비율벡터는 $\mathbf{x}^* = (0.55 - 4\delta/9, 0.45 - 4\delta/9, 4\delta/9)'$가 됨을 알 수 있다. 가법교체법은 단순하지만 영값들의 대치에 사용할 양수를 선택할 객관적인 기준이 없음에 따라 어떤 값을 사용해야 하는지에 대한 판단이 어려운 문제가 있다. Martin-Fernandez 등 (2003)은 가법교체법의 대안으로 승법교체법(multiplicative replacement)을 제안했다. 이 방법은 $c$개의 영값과 $(p-c)$개의 0이 아닌 값을 가지는 구성비율벡터에서 $c$개는 각각 $\delta_1, \ldots, \delta_c$로 대치하고 $(p-c)$개는 각각의 원래 값에다가 $1 - \sum_{i=1}^{c} \delta_i$를 곱한 값으로 대치하여 영값이 없는 구성비율벡터를 만들게 된다. 승법교체법 역시 가법교체법과 마찬가지로 $\delta_1, \ldots, \delta_c$를 객관적으로 선택하기가 쉽지 않다.

Bacon-Shone (1992)은 영값들을 처리하기 위해서 순위에 기초한 변환된 구성비율자료를 만든 다음 이 자료를 Aitchison (1982)의 방법론에 따라 분석할 것을 제안했다. Bacon-Shone의 방법(이하 BS의 방법)은 열 또는 행과 자료 전체에 대한 순위부여방식을 사용하여 각 자료값을 순위로 바꾼 순위자료를 만들고 이것을 각 행의 합으로 나누어서 변환된 구성비율자료를 얻게 된다. Bacon-Shone (1992)은 실제 자료에 3가지 순위부여방식을 적용하여 얻은 각각의 변환된 구성비율자료에 대한 여러 측면에서의 분석 결과를 제시했다. 이들의 비교를 통해 자료 전체에 대한 순위부여방식을 적용하는 것이 가장 바람직한 결과를 제공해 줌을 보고했다. BS의 방법에 따라 변환된 구성비율자료를 얻는 것은 어렵지 않다. 예를 들면, $n = 3$이고 $p = 3$인 아래의 구성비율자료

$$X = \begin{pmatrix} 0.2 & 0.7 & 0.1 \\ 0.0 & 0.6 & 0.4 \\ 0.3 & 0.5 & 0.2 \end{pmatrix}$$

로부터 자료 전체에 대한 순위부여방식을 사용하여 만든 순위자료는

$$R_{total} = \begin{pmatrix} 3.5 & 9 & 2 \\ 1 & 8 & 6 \\ 5 & 7 & 3.5 \end{pmatrix}$$

가 된다. 각각의 순위를 대응되는 행의 합으로 나누게 되면 변환된 구성비율자료로

$$X_T = \begin{pmatrix} \frac{3.5}{14.5} & \frac{9}{14.5} & \frac{2}{14.5} \\ \frac{1}{15} & \frac{8}{15} & \frac{6}{15} \\ \frac{5}{15.5} & \frac{7}{15.5} & \frac{3.5}{15.5} \end{pmatrix} = \begin{pmatrix} 0.2414 & 0.6207 & 0.1379 \\ 0.0667 & 0.5333 & 0.4000 \\ 0.3226 & 0.4516 & 0.2258 \end{pmatrix}$$

를 얻게 된다.

대응분석은 2원 분할표의 행과 열을 저차원상의 그래프로 나타내어 행과 열의 특성을 한 눈에 파악하고자 하는 차원축약방법으로 2원 분할표에 대한 주성분분석으로 볼 수가 있다. 대응분석과 주성분분석은 이론적인 측면에서 많은 유사성을 가지고 있고 자세한 내용은 Jolliffe (2002)를 보기 바란다. 구성비율자료 또한 행의 합이 1로 일정하므로 각 자료값은 정수가 아닐 뿐이지 각 행의 비중이 동일한 2원 분할표(개체는 행이고 변수는 열)로 생각할 수 있기 때문에 대응분석의 직접적인 적용을 고려할 수가 있다. Baxter 등 (1990)에 따르면 구성비율자료의 대응분석은 $y_{ij} = (x_{ij} - \bar{x}_j)/\sqrt{\bar{x}_j}$, $i = 1, \ldots, n$, $j = 1, \ldots, p$로 변환된 자료의 주성분분석과 같게 된다. 만일 각 변수들의 평균과 분산이 선형적인 관계를 가진다면 구성비율자료의 대응분석은 상관행렬에 기초한 주성분분석이 된다. Aitchison (1986)은 여러 이유를 들어 상관행렬에 기초한 구성비율자료의 주성분분석을 강하게 비판한 바가 있다. 대안적 방법으로 Baxter 등 (1990)은 $y_{ij} = (w_{ij} - \bar{w}_j)/\sqrt{\bar{w}_j}$, $i = 1, \ldots, n$, $j = 1, \ldots, p$로 변환된 자료의 대응분석을 제안했다. 여기서 $w_{ij} = x_{ij}/\bar{x}_j$이고 $\bar{w}_j = \sum_{i=1}^{p} w_{ij}/p$이다. 주성분분석의 목적 중의 하나가 고차원상에 있는 임의의 두 점간의 거리를 저차원상에 정보의 손실없이 그대로 재현하는 것이고 이런 관점에서 $y_{ij}$들에 기초한 대응분석(이하 BCH의 방법)은 Aitchison의 로그대비 주성분분석과 근사적으로 같은 방법이 된다. Baxter 등 (1990)은 영값이 없는 실제 구성비율자료에 적용했을 때 그들이 방법이 Aitchison의 방법과 유사한 결과를 제공함을 보고하면서 영값이 있는 경우에도 적용이 가능함으로 로그대비 주성분분석보다 더 선호될 것이라고 주장했다.

Huh (1999)에 의해 소개된 순위자료에 대한 다변량 수량화 방법(이하 순위자료 수량화)는 각각의 원 자료값을 행내에서의 순위로 대체시킨 순위자료에 대한 주성분분석으로 볼 수가 있다. 영값들이 있는 구성비율자료에도 쉽게 적용이 가능하므로 위에서 열거한 방법들의 대안으로 고려해 볼 수 있다. 영값들이 있는 구성비율자료의 차원축약을 위한 순위자료의 수량화 방법을 간략히 설명하면 다음과 같다. $r_{ij}$를 $x_{ij}$에 부여한 순위라고 하면 구성비율자료 $X$로부터 순위자료로 변환한 자료행렬 $R = (r_{ij})$, $i = 1, \ldots, n$, $j = 1, \ldots, p$를 얻는다. 순위자료에서 각 행의 평균은 모두 같은 반면 열의 평균들은 모두 다르게 되며 그것 자체가 중요하기 때문에 행 중심화 처리를 한 $Y = R - \bar{r}J_{n,p}$를 차원축약을 위한 자료행렬로 사용하게 된다. 여기서 $\bar{r} = (p+1)2$이고 $J_{n,p}$는 모든 원소값이 1인 $n \times p$ 행렬이다. 최적의 차원축약을 위한 수량화의 목표는 $Y$의 행공간에 있는 단위벡터 $\mathbf{v}$에 $Y$의 각 행 $\mathbf{y}_i$를 사영시켜서 얻은 $\sum_{i=1}^{n}\{\mathbf{y}_i - (\mathbf{y}_i'\mathbf{v})\mathbf{y}\}/n'$를 제약조건 $\mathbf{v}'\mathbf{v} = 1$과 $\mathbf{v}'\mathbf{1}_p = 0$하에서 최소화하는 $\mathbf{v}$를 찾는 것이다. 여기서 $\mathbf{1}_p$는 모든 원소값들이 1인 $p \times 1$ 벡터이고 $n' = n - 1$이다. $\mathbf{v}$의 결정은 라그라지 승수법을 사용해서 대수적으로 풀어 보면 $Y'Y$에 대한 고유체계 $(Y'Y/n')\mathbf{v} = \lambda\mathbf{v}$의 해가 된다. 따라서 순위자료의 수량화에서도 $Y'Y/n'$에 대한 분광분해 $Y'Y/n' = VD_\lambda V'$의 계산이 요구되므로 주성분분석과 다를 바가 없다. 여기서 $V'V = I_p$이며 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p = 0$이다. 차원축약을 위해 사용할 첫 $s$개의 축은 $\lambda_1, \ldots, \lambda_s$에 대응하는 고유벡터에 의해 결정되고 $s$차원 수량화에 대한 근사도, 즉 설명력은 주성분분석에서와 같이 고유값들의 점유비가 된다.

## 3. 방법들의 비교

이 절에서는 앞에서 소개한 Aitchison의 가법교체법을 이용한 로그대비 주성분분석, BS의 방법, BCH의 방법과 순위자료 수량화들을 실제자료에 적용해 본다. 비교를 위해서 설명력과 개체플롯을 제시하고 Sibson (1978)의 $\gamma$를 통해서 각 방법에 따라 작성된 개체플롯의 유사성을 알아본다.

### 3.1. 유공충 자료

이 자료는 서로 다른 깊이에서 추출한 32개의 퇴적층 표본에 포함되어 있는 4종류의 유공충(1:

Table 1: Results for foraminiferal compositions

| Method | Eigenvalue | Variable | | | | Eigenvalue | Cumulative percentage of total variance |
|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | | |
| M1 | $\mathbf{e}_1$ | −0.028 | −0.149 | 0.782 | −0.605 | 4.297 | 83.39 |
| | $\mathbf{e}_2$ | 0.321 | 0.635 | −0.342 | −0.614 | 0.709 | 97.15 |
| M2 | $\mathbf{e}_1$ | −0.045 | −0.127 | 0.781 | −0.609 | 7.078 | 82.04 |
| | $\mathbf{e}_2$ | −0.388 | −0.586 | 0.361 | 0.613 | 1.392 | 98.18 |
| M3 | $\mathbf{e}_1$ | −0.029 | −0.122 | 0.773 | −0.622 | 1.195 | 83.72 |
| | $\mathbf{e}_2$ | 0.374 | 0.600 | −0.374 | −0.600 | 0.210 | 98.42 |
| M4 | $\mathbf{e}_1$ | 0.043 | 0.114 | 0.736 | −0.666 | 0.401 | 68.00 |
| | $\mathbf{e}_2$ | 0.449 | 0.546 | −0.442 | −0.553 | 0.148 | 93.17 |
| M5 | $\mathbf{e}_2$ | 0.741 | 0.173 | −0.414 | −0.499 | 4.021 | 78.79 |
| | $\mathbf{e}_2$ | 0.101 | −0.283 | 0.759 | −0.577 | 0.866 | 95.76 |

M1: LPCA using additive replacement($\delta = 10^{-3}$), M2: LPCA using additive replacement($\delta = 10^{-4}$),
M3: BS method, M4: BCH method, M5: quantification for ranked data

Neogloboquadrina atlantica, 2: Neogloboquadrina pachyderma, 3: Globorotalia obesa, 4: Globigerinoides triloba)의 비율을 측정해서 얻은 $n = 30$, $p = 4$인 구성비율자료이다 (Aitchison, 1986). 30개의 개체 중 7, 17, 21, 25, 30번 개체에 영값이 포함되어 있고 전체 120개의 자료값들 중 영값은 5개로 나타나고 있어 그 비율은 약 4%가 된다.

Table 1은 각 방법을 사용해서 분석한 결과로 $\mathbf{e}_1$과 $\mathbf{e}_2$는 각각 첫 번째 고유값과 두 번째 고유값에 대응하는 고유벡터를 나타낸다. 표에서 보는 바와 같이 누적 설명력이 가장 높게 나온 것은 BS의 방법으로 98.42%이고 가장 낮게 나온 것은 BCH의 방법으로 93.17%이다. 가법대치법을 이용한 로그대비 주성분분석의 경우 영값의 보정에 사용한 $\delta$를 작게 할 수록 설명력이 높아지는 것을 알 수 있다. 각 방법에 의한 설명력 모두가 90% 이상으로 높게 나타나므로 첫 두 개의 고유벡터로 이루어지는 새로운 축이 자료의 변이를 잘 포착함을 알 수 있다.

첫 번째 고유벡터를 보면 가법대치법을 이용한 로그대비 주성분분석과 BS의 방법이 동일한 부호패턴을 보이고 두 번째 고유벡터에서는 $\delta = 10^{-3}$인 가법대치법을 이용한 로그대비 주성분분석, BS의 방법과 BCH의 방법이 동일한 부호를 가짐을 알 수 있다. 첫 번째 고유벡터의 원소들을 절대값 크기순으로 비교해 보면 순위자료 수량화를 제외한 나머지 방법들은 3번, 4번, 1번, 2번 변수 순으로 동일한 크기패턴을 가진다. 두 번째 고유벡터의 경우는 $\delta = 10^{-3}$인 가법대치법을 이용한 로그대비 주성분분석과 BS의 방법, 그리고 $\delta = 10^{-4}$인 가법대치법을 이용한 로그대비 주성분분석과 BCH의 방법에서 같은 크기패턴을 보인다.

가장 많은 변이를 보유하고 있는 제 1축은 가법대치법을 이용한 로그대비 주성분분석, BS의 방법과 BCH의 방법에서는 3번 변수와 나머지 1번, 2번, 4번 변수들의 대비로 이루어지는 형태적 특성을 나타내는 반면에 순위자료 수량화에서는 1번, 2번 변수들과 3번, 4번 변수들의 대비로 이루어지는 형태적 특성을 나타내고 있다. 이와 함께, 가법대치법을 이용한 로그대비 주성분분석, BS의 방법과 BCH의 방법의 경우 제 1축은 3번과 4번 변수의 선형결합에 의해 대부분 결정이 되어지지만 순위자료 수량화의 경우는 2번 변수를 제외한 나머지 3개 변수들의 선형결합에 의해 대부분 결정이 됨을 볼 수 있다. 제 2축을 보면 $\delta = 10^{-3}$인 가법대치법을 이용한 로그대비 주성분분석, BS의 방법과 BCH의 방법에서는 1번과 2번 변수 대 3번과 4번 변수의 대비를 나타내고 있다. $\delta = 10^{-4}$인 가법대치법을 이용한 로그대비 주성분분석에서는 3번과 4번 변수 대 1번과 2번 변수들의 대비를, 순위자료 수량화에서는 1번과 3번 변수들과 2번과 4번 변수들의 대비를 나타내고 있다.

Figure 2는 각 방법별로 얻은 첫 두 개의 축으로 이루어진 2차원 평면에 30개의 개체들을 타점한 개체플롯으로 $\delta = 10^{-3}$과 $\delta = 10^{-4}$인 가법대치법을 이용한 로그대비 주성분분석에 의한 플

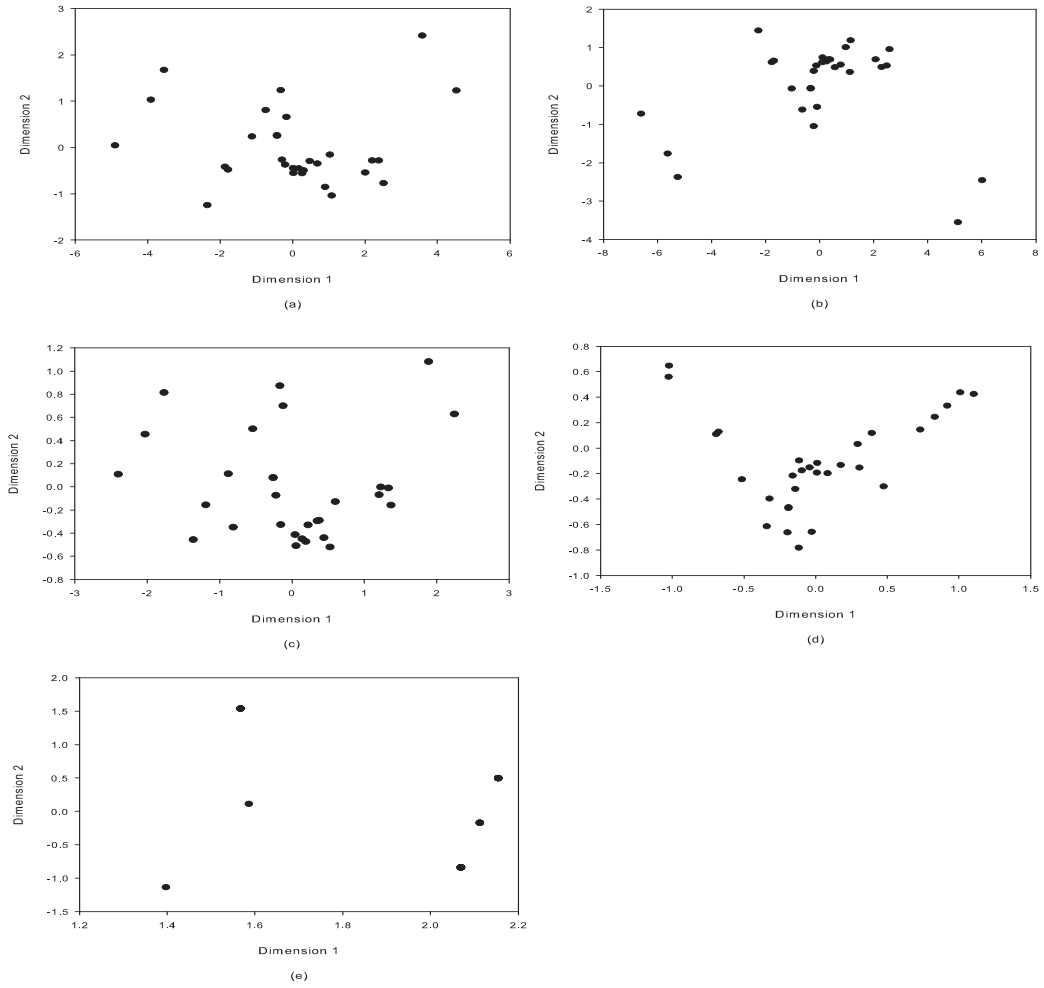


Figure 2: *Observation plots for foraminiferal compositions: (a) and (b) LPCA using additive replacement ($\delta = 10^{-3}$, $\delta = 10^{-4}$), (c) BS method, (d) BCH method, (e) quantification for ranked data*

롯 (a)와 (b)와 BS의 방법에 의한 플롯 (c)가 유사하게 나타남을 볼 수 있다. 개체플롯의 유사성을 객관적인 수치를 통해서 비교하기 위해 Sibson (1978)이 제안한 $\gamma$ 측도를 사용하기로 한다. 두 개체플롯의 작성을 위해 사용한 좌표값 행렬을 각각 $Z_1$과 $Z_2$라 하면 $\gamma$ 측도는 다음과 같이 $\gamma = 1 - \{\mathrm{tr}(Z_2'Z_1Z_1'Z_2)^{1/2}\}^2/\{\mathrm{tr}(Z_1'Z_1)\mathrm{tr}(Z_2'Z_2)\}$로 정의된다. $\gamma = 0$이면 두 개체플롯은 동일한 형상을 가지게 되고 $\gamma \leq 0.10$이면 두 개체플롯의 유사성이 높음을 나타낸다. Table 2는 각 방법별 개체플롯의 유사성을 보기 위해 $\gamma$를 계산한 결과이다. $\gamma$가 0.1 이하의 값을 가지는 것은 (M1, M2), (M1, M3)과 (M2, M3)으로 이들 세 개 개체플롯들은 서로 유사함을 알 수 있고, 특히 (M1, M3)의 $\gamma$는 0.024로 두 개체플롯은 매우 높은 유사성을 가진다. BCH의 방법과 순위자료 수량화의 경우는 전반적으로 $\gamma$가 큰 값을 가지는 것으로 나타나서 이들 개체플롯은 전반적으로 다른 방법들과는 매우 상이한 형상을 보여줌을 알 수 있다.

Table 2: $\gamma$ values calculated based on observation plots for foraminiferal compositions

| | M2 | M3 | M4 | M5 |
|---|---|---|---|---|
| M1 | 0.029 | 0.024 | 0.355 | 0.901 |
| M2 | | 0.071 | 0.495 | 0.919 |
| M3 | | | 0.362 | 0.893 |
| M4 | | | | 0.756 |

M1: LPCA using additive replacement($\delta = 10^{-3}$), M2: LPCA using additive replacement($\delta = 10^{-4}$),
M3: BS method, M4: BCH method, M5: quantification for ranked data

Table 3: Results for compositions of 92 glacial tills

| Method | Eigenvector | Variable | | | | Eigenvalue | Cumulative percentage of total variance |
|---|---|---|---|---|---|---|---|
| | | A | B | C | D | | |
| M1 | $\mathbf{e}_1$ | −0.414 | −0.081 | −0.343 | 0.839 | 3.173 | 50.01 |
| | $\mathbf{e}_2$ | −0.489 | 0.835 | −0.132 | −0.214 | 1.897 | 79.90 |
| M2 | $\mathbf{e}_1$ | −0.292 | −0.195 | −0.372 | 0.859 | 7.301 | 61.07 |
| | $\mathbf{e}_2$ | −0.277 | −0.582 | 0.758 | 0.101 | 2.841 | 84.83 |
| M3 | $\mathbf{e}_1$ | −0.306 | −0.201 | −0.354 | 0.861 | 0.543 | 61.47 |
| | $\mathbf{e}_2$ | −0.355 | −0.505 | 0.783 | 0.078 | 0.231 | 87.69 |
| M4 | $\mathbf{e}_1$ | 0.758 | −0.303 | 0.113 | −0.567 | 0.314 | 40.89 |
| | $\mathbf{e}_2$ | −0.051 | 0.810 | −0.215 | −0.544 | 0.302 | 80.16 |
| M5 | $\mathbf{e}_2$ | 0.591 | 0.399 | −0.470 | −0.520 | 4.071 | 81.50 |
| | $\mathbf{e}_2$ | −0.548 | 0.685 | −0.401 | 0.264 | 0.501 | 91.53 |

M1: LPCA using additive replacement($\delta = 10^{-3}$), M2: LPCA using additive replacement($\delta = 10^{-4}$),
M3: BS method, M4: BCH method, M5: quantification for ranked data

### 3.2. 빙력토 자료

이 자료는 호적세의 빙하기에 빙하에 의해 침식되어 운반된 분급되지 않은 91개의 빙력토(till) 표본들에 포함되어 있는 적사암(A: red sandstone), 회색사암(B: gray sandstone), 결정체(C: crystalline)와 기타 물질(D: miscellaneous)들을 무게 비율로 측정한 $n = 92$, $p = 4$인 구성비율자료이다 (Kaiser, 1962). 92개의 개체 중 42개의 개체에 영값이 적어도 하나 이상 포함되어 있으며 전체 368개의 자료값들 중 영값은 47개로 그 비율은 약 13%가 된다.

Table 3은 각 방법을 사용해서 분석한 결과로 $\mathbf{e}_1$과 $\mathbf{e}_2$는 각각 첫 번째 고유값과 두 번째 고유값에 대응하는 고유벡터를 나타낸다. 첫 2개의 축에 의한 설명력이 가장 높게 나온 것은 순위자료 수량화로 91.53%이고 가장 낮게 나온 것은 $\delta = 10^{-3}$인 가법대치법을 이용한 로그대비 주성분분석으로 79.90%이다. 두 고유벡터들의 부호를 살펴보면 $\delta = 10^{-4}$인 가법대치법을 이용한 로그대비 주성분분석과 BS의 방법이 동일한 패턴을 가진다. 첫 번째와 두 번째 고유벡터에서 원소들을 절대값 크기순으로 비교해 보면 $\delta = 10^{-4}$인 가법대치법을 이용한 로그대비 주성분분석과 BS의 방법은 동일한 크기패턴을 가지지만 다른 방법들은 서로 다른 크기패턴을 보인다.

제 1축과 2축 모두는 빙력토에 포함된 성분들의 형태적 특성에 관한 것임을 알 수 있다. 제 1축의 경우에는 가법대치법을 이용한 로그대비 주성분분석과 BS의 방법에서는 D 변수와 A, B, C 변수들의 대비를 나타내고 있다. 반면에 BCH의 방법에서는 C 변수와 A, B, D 변수들과의 대비를, 순위자료 수량화에서는 A와 B 변수들과 C와 D 변수들과의 대비를 나타내고 있다. 제 1축의 구성에서 주도적인 역할을 하는 변수, 즉 절대값 기준으로 가장 큰 계수를 가지는 변수를 보면 가법대치법을 이용한 주성분분석과 BS의 방법은 D인 반면에 BCH의 방법과 순위자료 수량화는 A이다. 제 2축은 $\delta = 10^{-3}$인 가법대치법을 이용한 로그대비 주성분분석과 BCH의 방법에서는 동일하게 B 변수와 나머지 세 변수들의

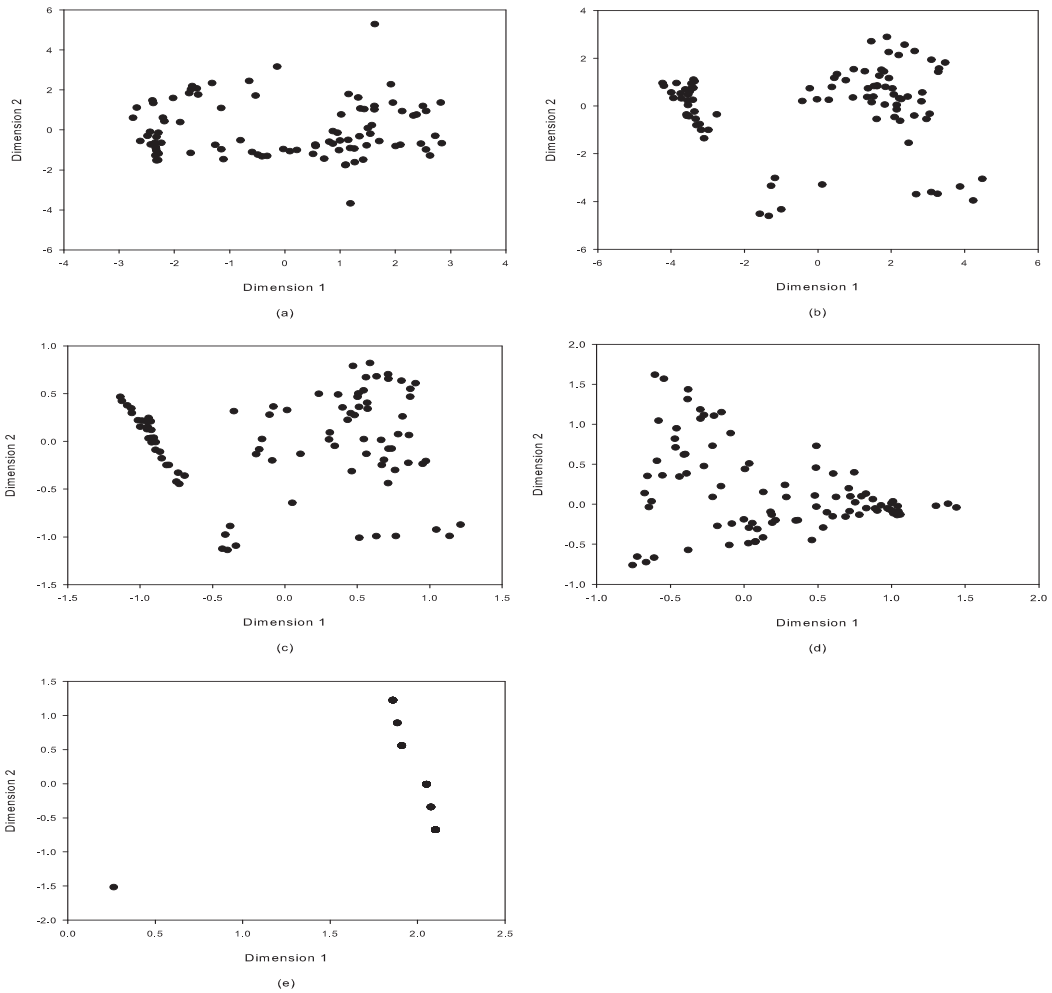


Figure 3: *Observation plots for compositions of 92 glacial tills: (a) and (b) LPCA using additive replacement($\delta = 10^{-3}$, $\delta = 10^{-4}$), (c) BS method, (d) BCH method, (e) quantification for ranked data*

대비를 나타낸다. 반면에 $\delta = 10^{-4}$인 가법대치법을 이용한 로그대비 주성분분석과 BS의 방법에서는 C와 D 변수들과 A와 B 변수들의 대비를 나타내고 순위자료 수량화의 경우는 B와 D 변수들과 A와 C 변수들의 대비임을 알 수 있다. 제 2축의 결정을 주도하는 변수를 보면 $\delta = 10^{-4}$인 가법대치법을 이용한 로그대비 주성분분석과 BS의 방법은 C 변수이며 나머지 방법들은 B 변수임을 알 수 있다.

Figure 3은 첫 2개의 축으로 이루어진 평면 위에 92개의 개체들을 타점한 개체플롯들이다. 그림에서 보는 바와 같이 $\delta = 10^{-4}$인 가법대치법을 이용한 로그대비 주성분분석에 의한 플롯 (b)와 BS의 방법에 의한 플롯 (c)가 매우 유사함을 볼 수 있다. Table 4는 Sibson (1978)의 $\gamma$를 계산한 것으로 (M2, M3)에 대한 $\gamma$값이 0.1 이하로 두 개체플롯들의 유사성이 매우 높음을 알 수 있다. 그리고 이들 개체플롯을 제외한 나머지 개체플롯들은 전반적으로 $\gamma$값이 크게 나오므로 Figure 3에 보는 바와 같이 서로 상이한 형상을 가짐을 알 수 있다.

Table 4: $\gamma$ values calculated based on observation plots for compositions of 92 glacial tills

| | M2 | M3 | M4 | M5 |
|---|---|---|---|---|
| M1 | 0.341 | 0.395 | 0.361 | 0.958 |
| M2 | | 0.028 | 0.730 | 0.986 |
| M3 | | | 0.737 | 0.985 |
| M4 | | | | 0.744 |

M1: LPCA using additive replacement($\delta = 10^{-3}$), M2: LPCA using additive replacement($\delta = 10^{-4}$),
M3: BS method, M4: BCH method, M5: quantification for ranked data

## 4. 결론

본 논문에서는 영값이 있는 구성비율자료에 대한 차원축약의 목적으로 사용 가능한 방법으로 Aitchison (1986)의 가법대치법을 이용한 로그대비 주성분분석, Bacon-Shone (1992)의 순위변환에 기초한 로그대비 주성분분석, Baxter 등 (1990)의 대응분석과 순위자료에 대한 수량화 방법을 소개했다. 각 방법의 비교를 위해 실제 자료 분석을 수행했고 이들 결과를 요약하면 다음과 같다.

가법대치법을 이용한 로그대비 주성분분석은 영값 보정을 위해 사용한 $\delta$값에 따라 서로 다른 결과를 보여 주게 되고 $\delta$값을 작게 할수록 누적 설명력은 높아지는 현상을 볼 수 있었다. 순위변환에 기초한 로그대비 주성분분석은 다른 방법들에 비해서 대체적으로 높은 설명력을 가지는 것으로 나타났고 가법대치법을 이용한 로그대비 주성분분석과 유사한 결과를 제공함을 알 수 있었다. Baxter 등 (1990)의 대응분석과 순위자료에 대한 수량화 방법은 다른 방법들과는 대체로 상이한 결과를 주지만 누적 설명력은 높게 나타났다. 분석자료에서 차지하는 영값들의 비율이 커짐에 따라 모든 방법들은 누적 설명력이 떨어지는 경향을 보였다.

실제 응용에서 순위변환에 기초한 로그대비 주성분분석은 가법대치법을 이용한 로그대비 주성분분석과 Baxter 등 (1990)의 대응분석에 대한 대안적 방법으로 사용이 될 수 있을 것으로 판단이 된다. 또한 하나의 분석 방법을 선택하여 자료 분석을 하기 보다는 다른 방법들도 같이 사용하여 비교 및 검토하는 것이 유용할 것으로 생각된다.

## References

Aitchison, J. (1982). The statistical analysis of compositional data(with discussion), *Journal of the Royal Statistical Society, Series B,* **44**, 139–177.

Aitchison, J. (1983). Principal component analysis of compositional data, *Biometrika,* **70**, 57–65.

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data,* Chapman and Hall, New York.

Bacon-Shone, J. (1992). Ranking methods for compositional data, *Applied Statistics,* **41**, 533–537.

Baxter, M. J., Cool, H. E. M. and Heyworth, M. P. (1990). Principal component and correspondence analysis of compositional data: Some similarities, *Journal of Applied Statistics,* **17**, 229–235.

Butler, J. C. (1976). Principal component analysis using the hypothetical closed array, *Journal of Mathematical Geology,* **8**, 25–36.

Gnanadesikan, R. (1977). *Methods for Statistical Data Analysis of Multivariate Observations,* Wiley, New York.

Gower, J. C. (1967). Multivariate analysis and multidimensional geometry, *Statistician*, **17**, 13–28.

Huh, M.-H. (1999). *Quantification Methods for Multivariate Data,* Freedom Academy, Seoul.

Jolliffe, I. T. (2002). *Principal Component Analysis, 2nd Edition* Springer, New York.

Kaiser, R. F. (1962). Composition and origin of glacial till, Mexico and Kasoag quadrangles, New York, *Journal of Sedimentary Petrology,* **32**, 502–513.

Le Maitre, R. W. (1968). Chemical variation within and between volcanic rock series - a statistical approach, *Journal of Petrology,* **9**, 220–252.

Martin-Fernadez, J. A., Barcelo-Vidal, C. and Pawlowsky-Glahn, V. (2003). Dealing with zeros and missing values in compositional data sets using nonparametric imputation, *Journal of Mathematical Geology,* **35**, 253–278.

Sibson, R. (1978). Studies in the robustness of multidimensional scaling, *Journal of the Royal Statistical Society, Series B,* **40**, 234–238.

Webb, W. N. and Briggs, L. I. (1966). The use of principal component analysis to screen mineralogical data, *Journal of Geology,* **74**, 716–720.

2012년 5월 2일 접수; 2012년 6월 13일 수정; 2012년 6월 16일 채택