DE GRUYTER
OPEN

Ekológia (Bratislava) Vol. 35, No. 2, p. 173–190, 2016
DOI:10.1515/eko-2016-0014

IMPACT OF SAMPLE SIZE ON PRINCIPAL COMPONENT ANALYSIS ORDINATION OF AN ENVIRONMENTAL DATA SET: EFFECTS ON EIGENSTRUCTURE

S. SHAHID SHAUKAT[1], TOQEER AHMED RAO[2*], MOAZZAM A. KHAN[1]

[1]Institute of Environmental Studies, University of Karachi, Karachi-75270, Pakistan
[2]Department of Botany, Federal Urdu University of Arts, Sciences & Technology, Karachi-75300, Pakistan; e-mail: toqeerahmedrao@fuuast.edu.pk, toqeerrao@yahoo.com

* Author for correspondence

Abstract

Shaukat S.S., Rao T.A., Khan M.A.: Impact of sample size on principal component analysis ordination of an environmental data set: effects on eigenstructure. Ekológia (Bratislava), Vol. 35, No. 2, p. 173–190, 2016.

In this study, we used bootstrap simulation of a real data set to investigate the impact of sample size (N = 20, 30, 40 and 50) on the eigenvalues and eigenvectors resulting from principal component analysis (PCA). For each sample size, 100 bootstrap samples were drawn from environmental data matrix pertaining to water quality variables (p = 22) of a small data set comprising of 55 samples (stations from where water samples were collected). Because in ecology and environmental sciences the data sets are invariably small owing to high cost of collection and analysis of samples, we restricted our study to relatively small sample sizes. We focused attention on comparison of first 6 eigenvectors and first 10 eigenvalues. Data sets were compared using agglomerative cluster analysis using Ward's method that does not require any stringent distributional assumptions.

Key words: eigenstructure, environmental data, ordination, PCA.

Introduction

Ordination methods belong to the group of multivariate analytical methods that are primarily used by ecologists for exploratory data analysis. Many informal and formal ordination techniques, including polar ordination (Bray, Curtis, 1957), principal component analysis (PCA) (Goodall, 1953; Orloci, 1966), correspondence analysis (Hill, 1973), detrended correspondence analysis (DCA) (Hill, Gauch, 1980), canonical correspondence analysis (CCA) (ter Braak, 1986) and so on have been proposed. Ecologists have often compared the results of different ordination methods (Gauch, Whittaker, 1972; Fasham, 1977; Gauch et al. 1977, 1981; Minchin, 1987; Whittaker, 1987; Shaukat, Uddin, 1989a,b; Anderson, Willis, 2003; Shaukat et al. 2005; Hirst, Jackson, 2007; Legendre, Birks, 2012) and have pointed out their advantages and disadvantages (Orloci, 1978; Shaukat, Siddiqui, 2005).

However, of all the ordination techniques developed so far, PCA continues to be the most popular technique in a number of biological sciences and other fields besides ecology and environmental sciences. The properties as well as the interpretation of the components of PCA have been investigated extensively (Rao, 1964; Jackson, 1991; Joliffe, 2002). PCA is a formal ordination technique basically used by ecologists for the purpose of parsimony and exploration of data in order to comprehend data and to seek the underlying trends and gradients in the data structure (James, McCulloch, 1990; Walker, Jackson, 2011).

In the context of environmental studies, although PCA is often useful for the analysis of samples in site space, it is still quite appropriate for the analysis of samples in environmental space. This is because it is likely for most environmental variables to be monotonically related to underlying factors and to each other. Also, PCA allows the use of variables that are not measured in the same units (e.g. salinity, biological oxidation demand (BOD), concentration of nutrients, temperature and pH). Conceptually, PCA exposes the underlying common covariance structure inherent in the data matrix resulting in a new set of coordinates called principal components. The Principal components (Y_1, …, Y_t) are orthogonal to each other and reflect different dimensions of the data. PCA essentially identifies the direction of maximum variation contained in the multivariate hyperspace of data points. It partitions the total variance inherent within rows (variables) of a data matrix into new set of derived variables (Y_i), which are linear combination of original variables. Thus the model of PCA is as follows:

$$\mathbf{Yij = Bi1\ A1j + Bi2A2j + ….+ BipApj}$$

i = 1, 2, …, p;
j = 1, 2,…, n, where $\mathbf{B_i}$
are the eigenvectors and **Aij** are the (trivially transformed) observations on variable i and object j.

The components are such that

$$\mathbf{Var(Y_1) \geq Var(Y_2) \geq Var(Y_3),\ldots \geq Var(Y_t)}$$

Or

$$\boldsymbol{\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \lambda_4 \ldots\ldots \geq \lambda_t}$$

which essentially shows the order of their importance in terms of explained variance of each component.

where λ_i are the standardized values (eigenvalues) of the original variables, and B_{ik} are the eigenvector coefficients. The first transformation converts the raw data matrix **X** or **A** (standardized form) into a variance–covariance matrix **S** (or correlation matrix **R).** The second transformation involves deriving the principal components Y_{ij} from the variance–covariance matrix **S** such that

$$\mathbf{Y} = \mathbf{B}' \mathbf{X}$$

where **B** is a matrix of eigenvectors $\mathbf{b}_i$ as follows:

$$\mathbf{B} = [\, \mathbf{b}_1 \mathbf{b}_2 \mathbf{b}_3 \ldots \mathbf{b}_t \,] .$$

The eigenvalues and eigenvectors are related as follows:

$$\mathbf{SB} = \lambda \mathbf{B}$$

where λ is a matrix of eigenvalues λ_i. The eigenvalues are the generalized variances related to individual variances of the variables as follows:

$$\Sigma \lambda_{ii} = \Sigma S_{ii}$$

An eigenanalysis is performed to obtain eigenvalues and eigenvectors. This involves solving the determinantal equation as follows:

$$| \mathbf{S} - \lambda \mathbf{I} | = 0$$

where **I** is an identity matrix.

The dimensions of the data set are defined to be equal to the number of principal components. Subsequently, the set of t principal components are reduced to a set of size k, where $1 \leq k << t$. The major advantage of parsimony (dimension reduction) is that it renders the analysis and interpretation easier whilst retaining most of the variation inherent in the data structure. Evidently, the closer the value of k is to t, the PCA model will more effectively fit the data because it would retain greater information contained in the data set.

To interpret the PCA axes effectively, it is pertinent to identify which of the multitude of variables are associated with specific axes (components). Ecologists have often correlated the environmental variables with the principal components using standard univariate tests, which often suffer from different violations of the underlying assumptions, to say the least (e.g. Swan, Dix, 1970; Wikum, Wali, 1974). Others have relied on loadings (eigenvector coefficients) with larger magnitude to disclose the importance of variables with respect to contributing the variance associated with each of the component (Hirosawa et al., 1996; Gehlhausen et al., 2000). The effects of resampling and randomization procedures have been examined in a few studies (Diaconis, Efron, 1983; Knox, Peet, 1989; Stauffer et al., 1985; de Pillar, 1999; Chateau, Lebart, 1996). The bootstrap resampling technique can be used to generate large number of samples that may provide the means of evaluating the sampling error. Peres-Neto et al. (2003) presented some new Monte Carlo approaches for testing the significance of eigenvector coefficients. There are only a few studies of sample size (or implied sample size) on the results of PCA or other eigenvector ordination techniques. Manjarres-Martinez et al. (2012) tested the performance of three ordination methods in terms of their stability to bootstrap-generated sampling variance. Gamito and Raffaelli (1992) found that first axis of various ordinations including PCA were insensitive to sample replication but

some applications of DCA, NMDS and HMDS appeared sensitive to second axis though PCA largely remained unaffected in this respect.

Burd et al. (1990) demonstrated that because of subtle changes in the multivariate procedures, such as PCA-ordination, substantially different results can be obtained, leading to varied interpretation and even differing conclusions. The users of multivariate methods such as factor analysis (FA) and PCA widely believe that the use of larger sample sizes tends to provide factor loadings and eigenvalues that are more precise estimates of population values and are also more stable across repeated sampling. Goff and Mitchell (1975) performed a comparison of species ordination results from plot and stand data (implied difference in sample size) and found that once a reliable set of species adaptation values are produced for a region, data from various plot sizes can be ordered. Okland et al. (1990) sampled the boreal conifer forest using plots of different sizes (implied sample size) and subjected the data to DCA ordination. It was found that the eigenvalues of axes invariably increased upon lowering sample plot size. Otypkova and Chytry (2006) examined the effect of plot size on PCA ordination of vegetation samples and concluded that smaller plot size produces less stable ordination pattern particularly when beta diversity is low. Dengler et al. (2009) found that a two- to fourfold increase in plot size resulted in an increase in constancy of 20 percent or more.

In the context of PCA and FA, some workers have proposed rules of thumb for minimum sample size in relation to number of variables or correlation structure. Gorsuch (1983) recommended at least 100 samples. Hatcher (1994) recommended that the sample size should be larger than five times the number of variables (p). Hutcheson and Sofroniou (1999) suggested the sample size of 150 or more for highly correlated data. Cattell (1978) recommended a sample size (N) of 250. Garson suggested a sample size of 300 as appropriate for FA. Comrey and Lee (1992) recommended N to be 300 as good and 500 as very good (see MacCallum et al. 1999). Generally, ecological data, which is always difficult to gather, has much smaller sample size. Velicer and Fava (1998) and MacCallum et al. (1999) in their comprehensive simulation studies showed that rules of thumb are not valid and that the lowest sample size depends on other aspects of sampling design. Levels of communality (correlation structure) have a great bearing on sample size. When this structure is strong, greater recovery can be achieved with a relatively small sample size (N). Various rules of thumb are given pertaining to objects-to-variable ratio. Bryant and Yarnold (1995) recommended a ratio of not lower than 5. Cattell (1978) suggested a subject-to-variable ratio of 3:1 up to 6:1. A ratio of 2 was suggested by Kline (1979). Osborne and Castello (2004) in an intensive simulation study of PCA found an interaction between the sample size and the subject-to-variable ratio and showed that the best outcomes occurred in analyses where large Ns and high ratios were used. Bandalos and Boehm-Kaufman (2009) recently suggested that a suitable sample size would depend on the number of factors (Y,) the number of variables p associated with factors and to what extent the set of factors explain the variance inherent in the variables. Forcino (2012) concluded that a too small sample size would obviously lead to erroneous ecological conclusions, whilst an increasingly large sample size would follow the law of diminishing returns. Using paleoecological data sets of various sizes, Forcino (2012) demonstrated that a sample size for a multivariate analysis on a paleocommunity is around 25–50.

Recently, Dochtermann and Jenkins (2011) challenged the conventional views on sample size limitations in multivariate analysis. Using computer simulations, it was claimed that a model comparison procedure can correctly rank alternative models in about 90% of cases with the sample size of 19. Based on uncorrelated random normal deviates, they fitted a structural equation model, assuming that all variables were independent, and a similar model, assuming that all variables were intercorrelated because of an underlying latent construct. With respect to PCA, MacCallum et al. (2001) obtained good results with extremely small sample sizes and even for data with $p > n$, whilst Mundfrom et al. (2005) found some cases where large sample sizes ($n > 100$) were necessary. They also found that if the number of underlying factors stays the same, *more* variables (and not fewer, as implied by guidelines based on the observations-to-variables ratio) could lead to better results with small samples of observations.

The interpretation of PCA is, in general, subjective in ecology, environmental sciences, biology and many other disciplines (Orloci, 1978; Kendall, 1980; Shaukat, 1985; Shaukat, Uddin, 1989a; Cadima, Joliffe, 1995). The major reason behind it is that the ordination tools (such as PCA) are used as data exploratory techniques rather than as hypothesis testing procedures.

The principal objective of this study was to examine the influence of sample size on the eigenstructure, that is, eigenvalues and eigenvector coefficients using Monte Carlo simulation used on a real data set pertaining to the field of environmental science. The water quality data of a river in Baluchistan was used. Therefore, the current study, as opposed to many other studies on sample size, is based on a real ecological data set pertaining to an aquatic environment.

Material and methods

The Data Set: The data set comprised of 55 water samples (N=55) collected from different locations (sites) of Hingol river. The surface water was collected in clean plastic bottles previously rinsed with nitric acid. The quantitative water quality variables were analysed using the procedures described in the American Public Health Association (APHA, 1992). Twenty-two water quality variables were analysed. Thus our complete data matrix was 22×55.

Simulations: Principal component analysis (PCA) was performed on randomly selected samples of sizes N =20, 30, 40 and 50 from complete data set. PCA of complete data set N = 55 was also performed. For each sample size, 100 bootstrap simulations were performed (Knox, Peet, 1989; Manly, 1998; Pillar, 1999; Peres-Neto et al., 2005) and eigenvalues and the associated eigenvector were retained. This involved resampling the original data with replacement.

Means and standard errors of the first 10 eigenvalues and the first 6 eigenvectors were computed. The effect of sample size (N = 20, 30, 40 and 50) on the stability of component patterns in PCA was investigated. The sample sizes are those typically used by environmental scientists and ecologists. The eigenvalues and eigenvectors for each sample set were compared using a hierarchical agglomerative cluster analysis, specifically Ward's method of minimum within group variance clustering strategy in conjunction with Euclidean distance as the resemblance function. Because of the exploratory nature of PCA components in ecological or environmental sciences, we used this approach that does not involve any distributional assumptions regarding the data. Multivariate inferential tests, though available in this respect (e.g. chi-square, Lawley's test, Bartlett's test, (see Lawley, Maxwell, 1971; Jackson, 1993; Peres-Neto et al., 2003), were not used because of their stringent underlying assumptions that are hardly ever met by the real data sets and also because the intermediate or final results of statistical analysis do not yield independent random variables following particular distributions assumed by the multivariate statistic. Most such techniques either suffer from an inherent subjectivity or have a tendency to yield under estimate or over estimate of the true dimensions

of the data (Jackson, 1993). Thus, these tests have limited utility (Dochtermann, Jenkins, 2011). Loadings are often considered as significant when their absolute value is larger than a certain pre-selected arbitrary value, for example, 0.25 (Chatfield, Collins, 1980) and 0.3–0.5 (Richman, 1988). However, in ecological data sets, the loadings are usually smaller and their sizes depend on the characteristics of the data sets including their correlation structure and no rule of thumb can be realistically applied. They are interpreted according to their relative magnitude and direction with respect to the associated variable. Furthermore, in the present study, the object was to compare the PCA results pertaining to various sample sizes with those of the complete data set, rather than testing the significance of eigenvalues and eigenvectors. Cluster analysis was chosen for comparison of eigenvectors because it is flexible and capable of depicting similarities without the need to invoke strict distributional assumptions. To compare the set of eigenvalues ($\lambda_1, \ldots, \lambda_{10}$), scree plots (Cattell, 1966) were developed to compare the ranked eigenvalues resulting from PCA of various sample sizes. In addition, a Euclidean distance matrix was computed between sets of eigenvalues to elucidate the differences across various sample sizes. Likewise, Pearson correlation coefficient matrix was also computed to assess the similarities between sets of eigenvalues. The latter was used as a measure of similarity; therefore, no significance is attached. Cluster analysis was also used to evaluate the similarities between sets of eigenvalues pertaining to data of different sample sizes.

Results and discussion

The first 10 eigenvalues pertaining to data sets of different sizes were compared by using scree plots, Euclidean distance (D), correlation coefficient (r) and Ward's agglomerative clustering. The scree plots given in Fig. 1a–e show slight differences in their shapes.

The scree plots for sample sizes N = 20 and 30 (Fig. 1a and b) are closely similar, which is also depicted by their Euclidean distance ($D_{12} = 1.661$) and correlation coefficient (r = 0.999), whilst the scree plots of both these sample sizes exhibit marked difference with that of N = 40, particularly with respect to second eigenvalue (λ_2) (Fig. 1c), which is also shown by Euclidean distance (D) and correlation coefficient (r) ($D_{14} = 4.968$, $D_{24} = 4.629$). The scree plot resulting from PCA of complete data set (N = 55) exhibited slight but discernable differences with those resulting from various sample sizes (N = 20, 30, 40 and 50), which can also be confirmed by relatively greater values of Euclidean distance ($D_{15} = 6.190$, $D_{25} = 4.826$, $D_{35} = 6.839$, $D_{45} = 3.408$). The scree plot for sample size of 50 (Fig. 1d) is surprisingly more similar to that for N = 20 and 30 than that for N = 40, which can also be seen by the relatively lower values of Euclidean distances ($D_{14} = 3.292$ and $D_{24} = 2.276$) (Table 1) and also by the relatively greater values of correlation coefficients ($r_{14} = 0.998$, $r_{24} = 0.999$) (Table 2). The set of eigenvalues for complete data set showed relatively greater distances (and lower correlation coefficient) with the sets of eigenvalues resulting from various sample sizes (Tables 1 and 2).

T a b l e 1. Euclidean distances between the set of first 10 eigenvalues of PCA for various sample sizes and complete data sets. The key for labels 1–5 are as follows: 1, N = 20; 2, N = 30; 3, N = 40; 4, N = 50 and 5 is complete data set (N = 55).

Size	1	2	3	4	5
1	X				
2	1.661	X			
3	4.968	4.629	X		
4	3.292	2.276	5.998	X	
5	6.190	4.826	6.839	3.408	X

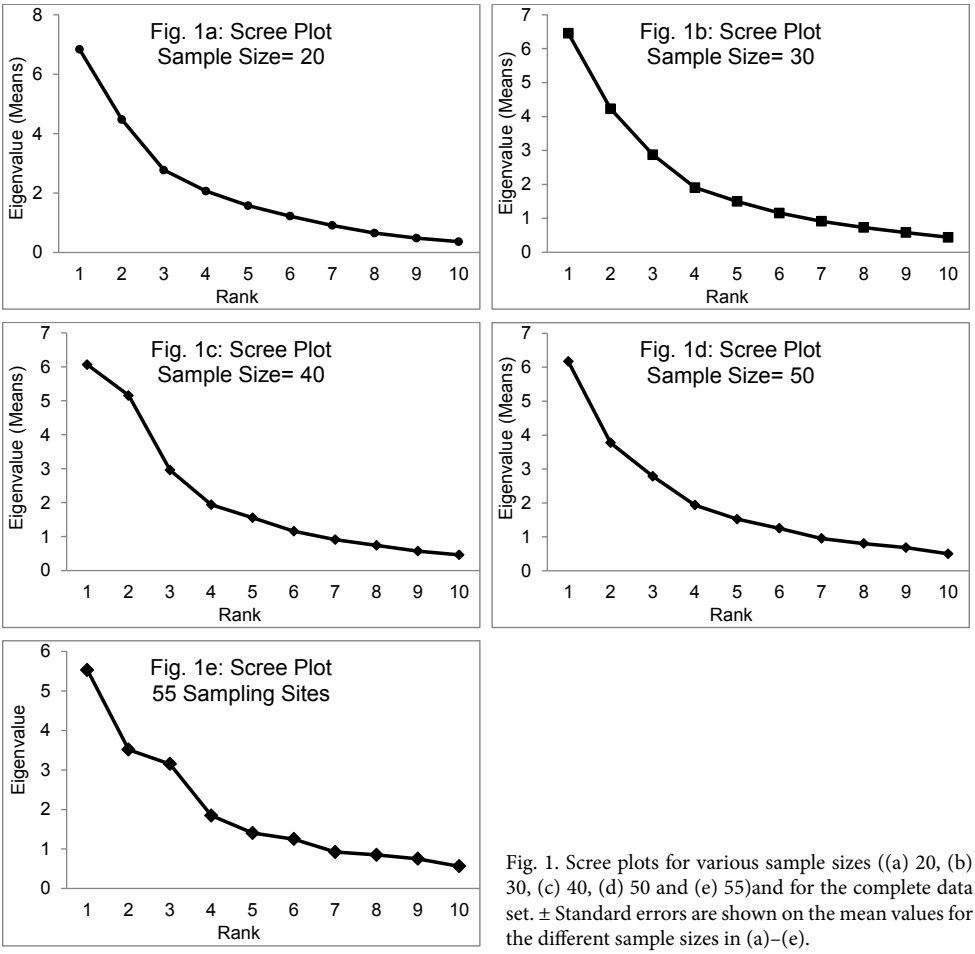

Fig. 1. Scree plots for various sample sizes ((a) 20, (b) 30, (c) 40, (d) 50 and (e) 55)and for the complete data set. ± Standard errors are shown on the mean values for the different sample sizes in (a)–(e).

T a b l e 2. Pearson correlation coefficients between the first 10 eigenvalues of PCAs for various sample sizes and complete data sets: 1, N = 20; 2, N = 30; 3, N = 40; 4, N = 50 and 5 is complete data set (N = 55).

Size	1	2	3	4	5
1	X				
2	0.999	X			
3	0.986	0.986	X		
4	0.998	0.999	0.976	X	
5	0.987	0.992	0.973	0.994	X

Statisticians have focused attention on the question of sample size with respect to multivariate analysis such as FA and PCA for decades. Some have looked specifically at sample size (N), whilst others at the ratio of sample to variable (N:p). In case of real data sets (e.g. in ecology or environmental sciences), the number of variables chosen are generally those

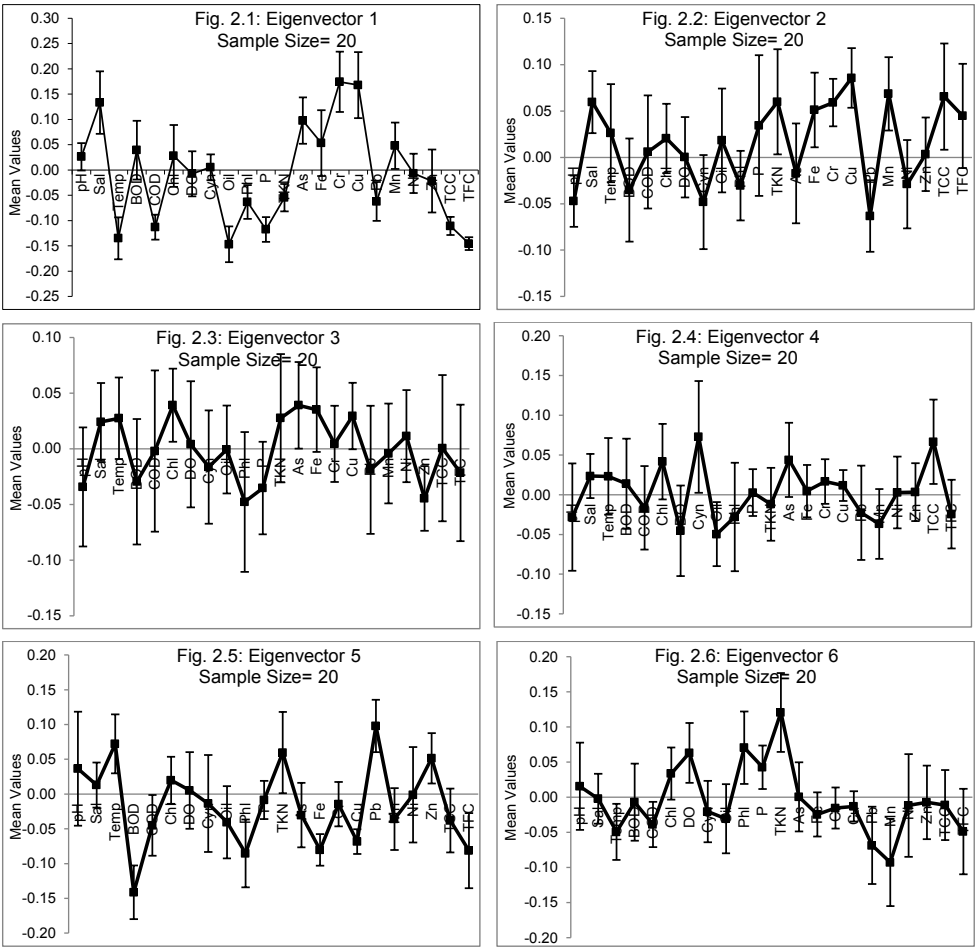

Fig. 2. (1–6) First six eigenvector loading at various sample sizes N = 20,30.40, 50 and for the complete population (55 sites). For variables associated with each eigenvector, coefficient standard symbols are used. Key to symbols: pH, water pH; sal, salinity; Temp, temperature; BOD, biological oxidation demand; COD, chemical oxidation demand; Chl, Chlorine conc.; DO, dissolved oxygen; Oil, Oil and Grease; Cyn, cynide; Phl, phenol; TKN, total Kjeldahl nitrogen; As, arsenic; Pb, lead; Cu, copper; Zn, zinc; Fe, iron; Mn, manganese; Ni, nickle; Cr, chromium; P, phosphorus; Tcc, total coliform bacteria; TFC, total faecal coliform.

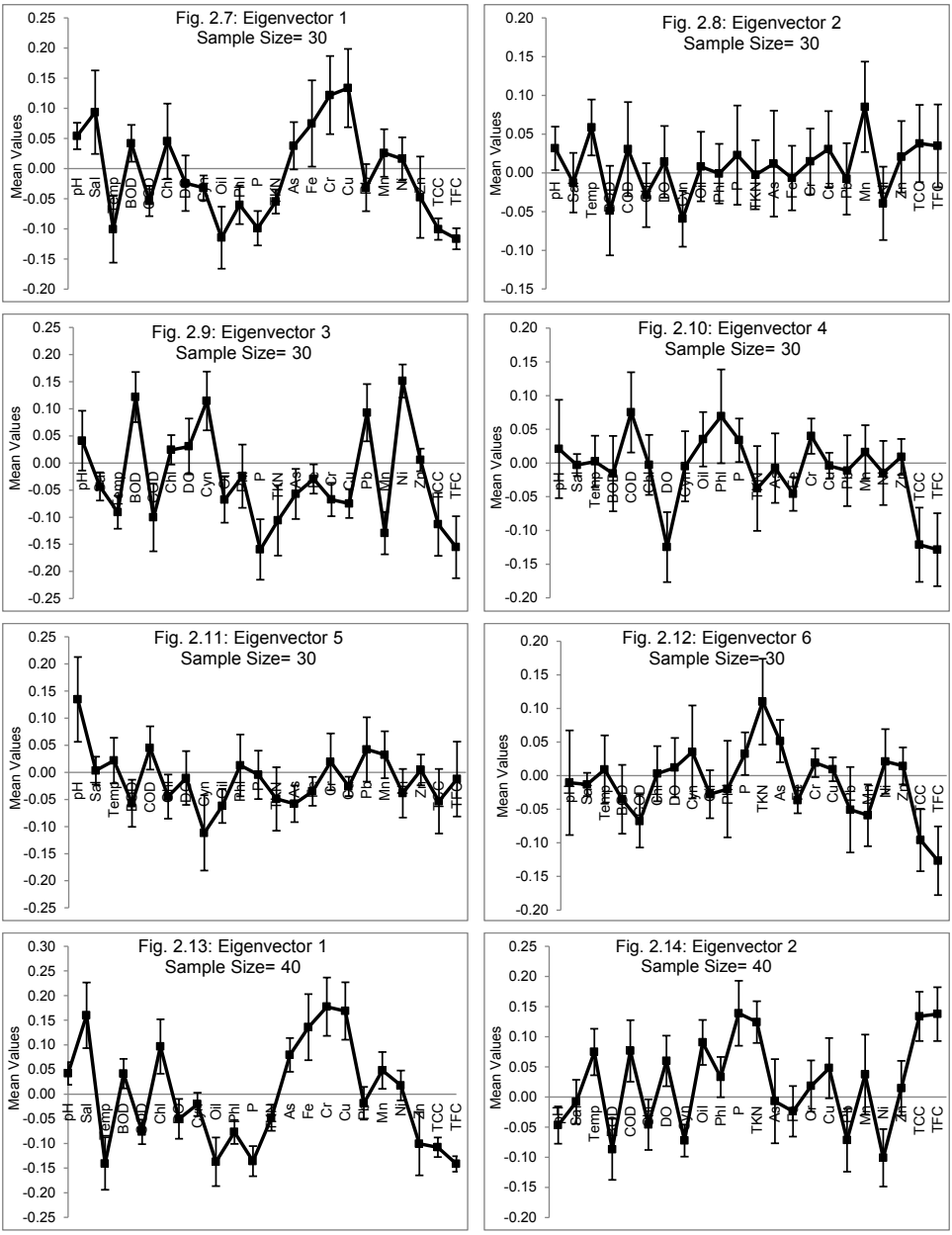

Fig. 2. (7–14) First six eigenvector loading at various sample sizes N = 20,30.40, 50 and for the complete population (55 sites). For variables associated with each eigenvector, coefficient standard symbols are used. Key to symbols: pH, water pH; sal, salinity; Temp, temperature; BOD, biological oxidation demand; COD, chemical oxidation demand; Chl, Chlorine conc.; DO, dissolved oxygen; Oil, Oil and Grease; Cyn, cynide; Phl, phenol; TKN, total Kjeldahl nitrogen; As, arsenic; Pb, lead; Cu, copper; Zn, zinc; Fe, iron; Mn, manganese; Ni, nickle; Cr, chromium; P, phosphorus; Tcc, total coliform bacteria; TFC, total faecal coliform.

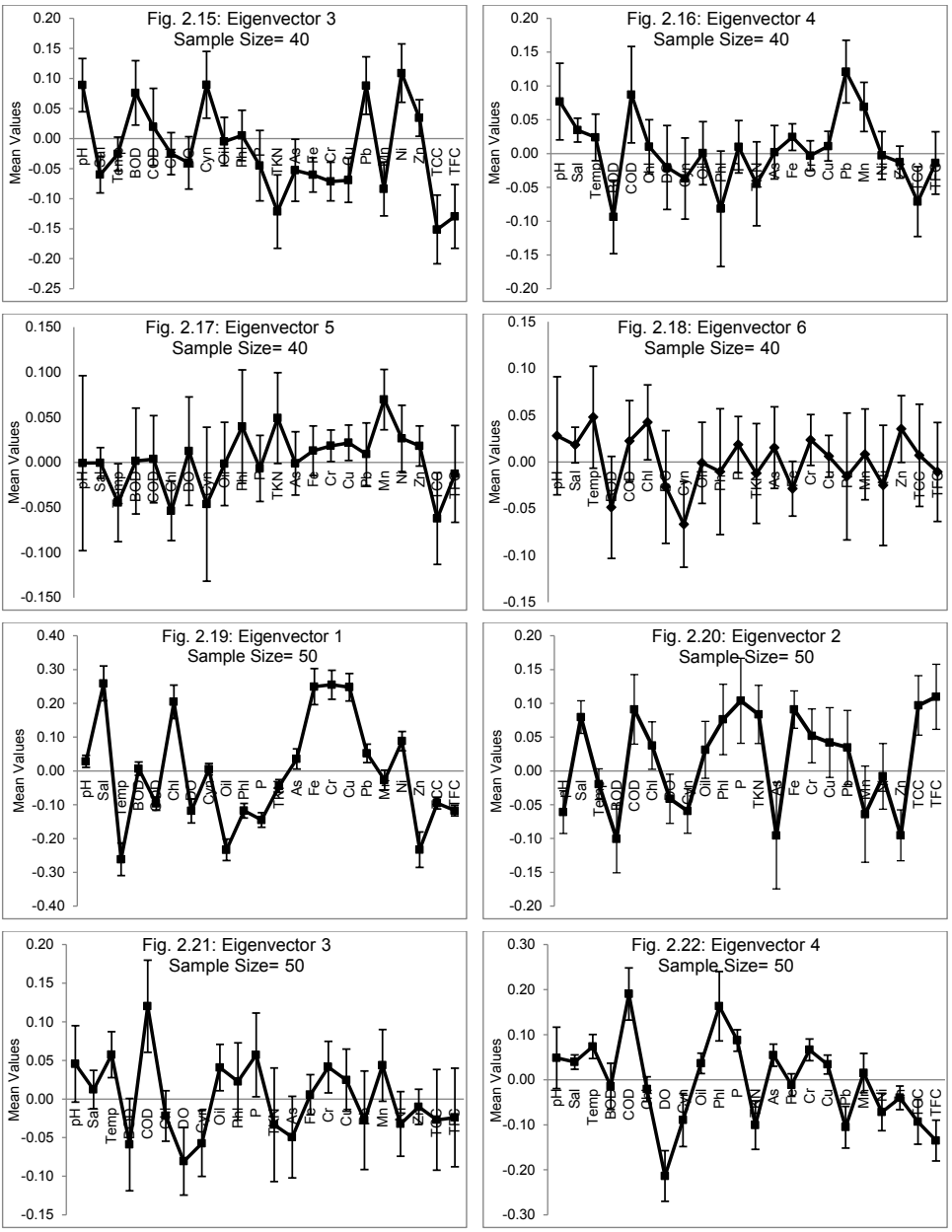

Fig. 2. (15–22) First six eigenvector loading at various sample sizes N = 20,30.40, 50 and for the complete population (55 sites). For variables associated with each eigenvector, coefficient standard symbols are used. Key to symbols: pH, water pH; sal, salinity; Temp, temperature; BOD, biological oxidation demand; COD, chemical oxidation demand; Chl, Chlorine conc.; DO, dissolved oxygen; Oil, Oil and Grease; Cyn, cynide; Phl, phenol; TKN, total Kjeldahl nitrogen; As, arsenic; Pb, lead; Cu, copper; Zn, zinc; Fe, iron; Mn, manganese; Ni, nickle; Cr, chromium; P, phosphorus; Tcc, total coliform bacteria; TFC, total faecal coliform.

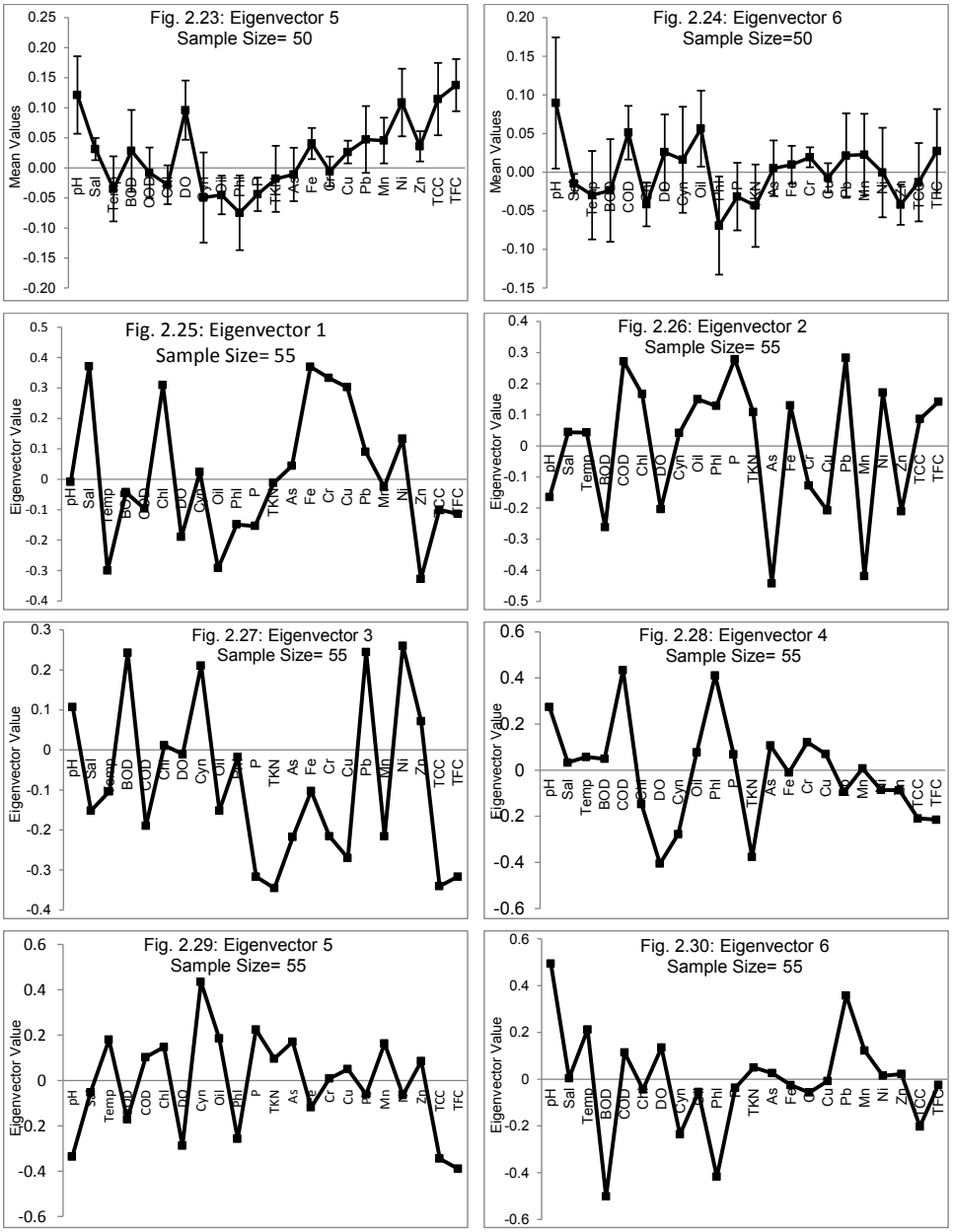

Fig. 2. (23–30) First six eigenvector loading at various sample sizes N = 20,30.40, 50 and for the complete population (55 sites). For variables associated with each eigenvector, coefficient standard symbols are used. Key to symbols: pH, water pH; sal, salinity; Temp, temperature; BOD, biological oxidation demand; COD, chemical oxidation demand; Chl, Chlorine conc.; DO, dissolved oxygen; Oil, Oil and Grease; Cyn, cynide; Phl, phenol; TKN, total Kjeldahl nitrogen; As, arsenic; Pb, lead; Cu, copper; Zn, zinc; Fe, iron; Mn, manganese; Ni, nickle; Cr, chromium; P, phosphorus; Tcc, total coliform bacteria; TFC, total faecal coliform.

It is also evident from Fig. 2 that the standard deviations of loadings were high because of the melded fluctuations owing to sampling and to the mixture of factor solutions. The interpretation of these standard deviations is not straightforward. Therefore, future investigations are required to further explore this area.

The dendrograms derived from cluster analysis of the first six eigenvectors of the sets pertaining to different sample sizes are given in Figs 3–7. Figure 3 shows the dendrogram based on cluster analysis using the first two components (first two eigenvectors) of all sample sizes and that of complete data set. The first and second eigenvectors are clearly separated out.

A perusal of dendrograms based on eigenvectors 3–6 (Figs 3–7) disclosed that the fist eigenvector of all the data sets was neatly segregated out and formed a compact group in each of the cluster analysis depicting its stability. Where the first three eigenvectors were used (Fig. 4), the second eigenvector tended to be separated from the third with slight intermixing. In other dendrograms (Figs 5–7), the first eigenvector was well separated but there was some amalgamation of higher order eigenvector (3–6) particularly the third eigenvector exhibited low-order segregation. However, the second eigenvector was separated to a considerable extent.

Evidently, PCAs based on N = 40 and N = 50 returned stable and consistent eigenvectors (Figs 6 and 7) compared to those of sample sizes N = 20 and N =3 0.

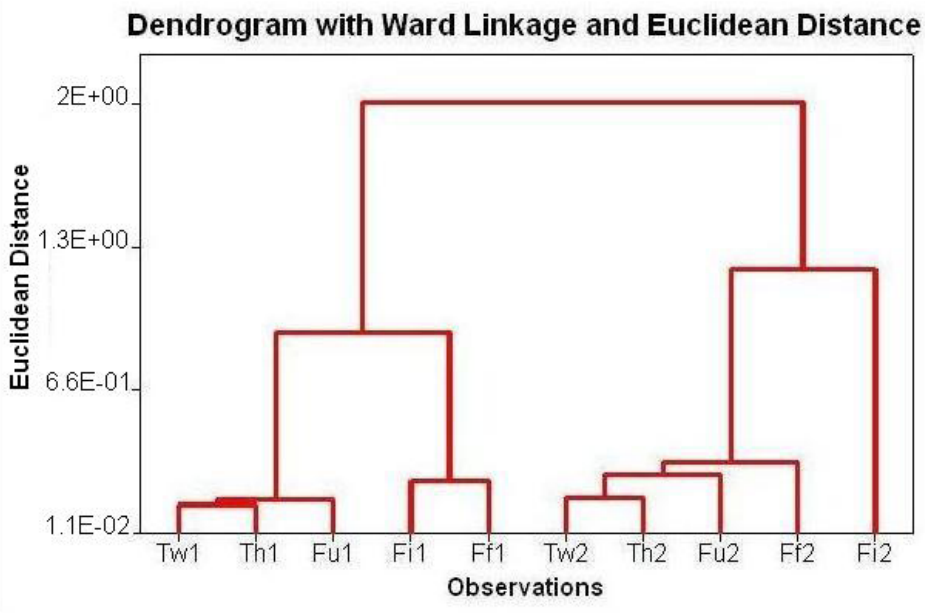

Fig. 3. Dendrogram derived from first two eigenvectors for various sample sizes and complete data set. The symbols Tw, Th, Fu, Fi and Ff represent N = 20,30,40,50, and 55, respectively. The associated letters 1 or 2 indicate eigenvector 1 and 2, respectively.

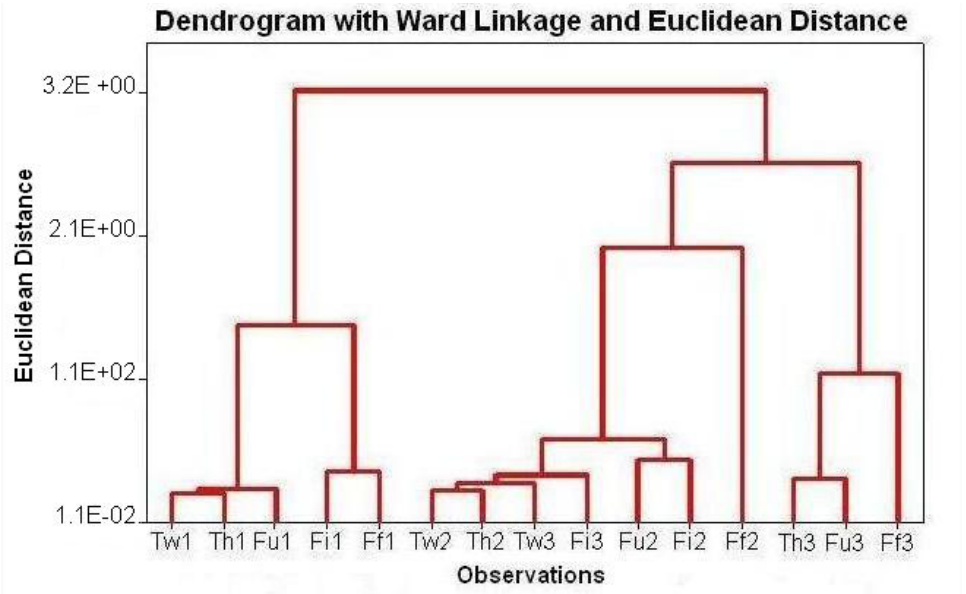

Fig. 4. Dendrogram derived from first three eigenvectors for different sample sizes and complete data set. The symbols Tw, Th, Fu, Fi and Ff represent N = 20,30,40,50, and 55, respectively. The associated letters 1, 2 or 3 indicate eigenvectors 1, 2 and 3, respectively.

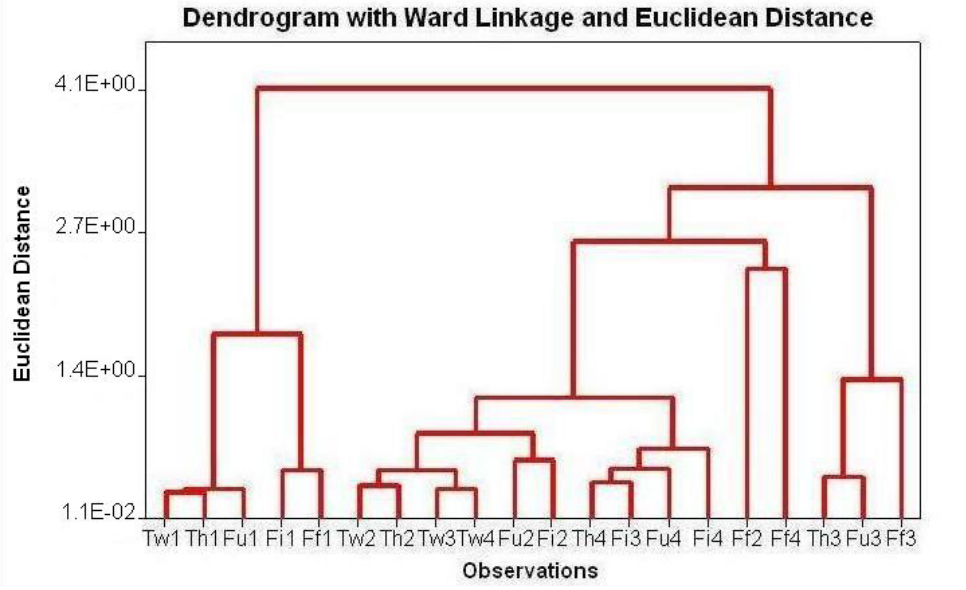

Fig. 5. Dendrogram derived from first four eigenvectors for different sample sizes and complete data (see Fig. 3). The associated numbers with the symbols 1, 2, 3 and 4 indicate eigenvectors 1, 2, 3 and 4, respectively.

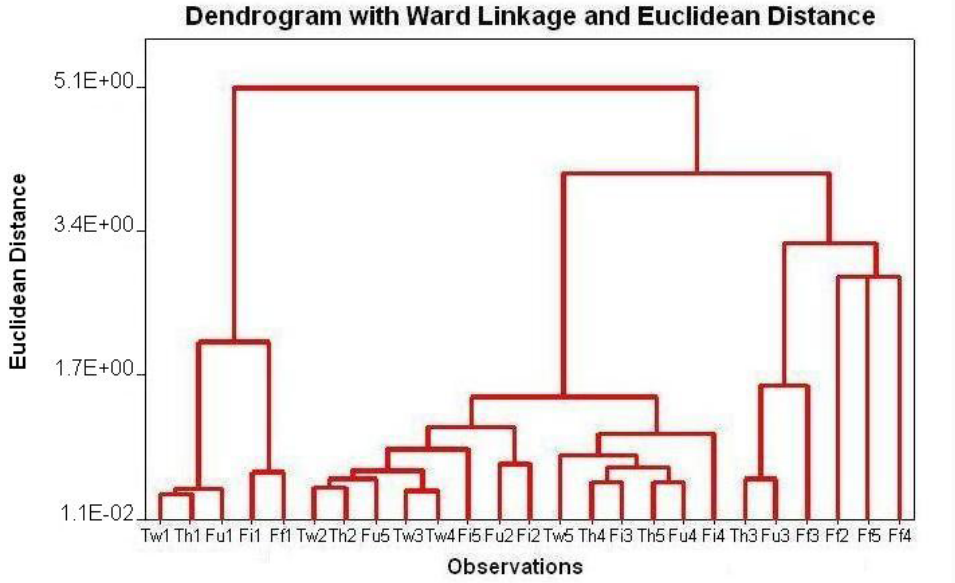

Fig. 6. Dendrogram derived from first five eigenvectors of various sample sizes and complete data set. Symbols as in Fig. 3. The associated numbers 1–5 represent eigenvectors 1–5, respectively.

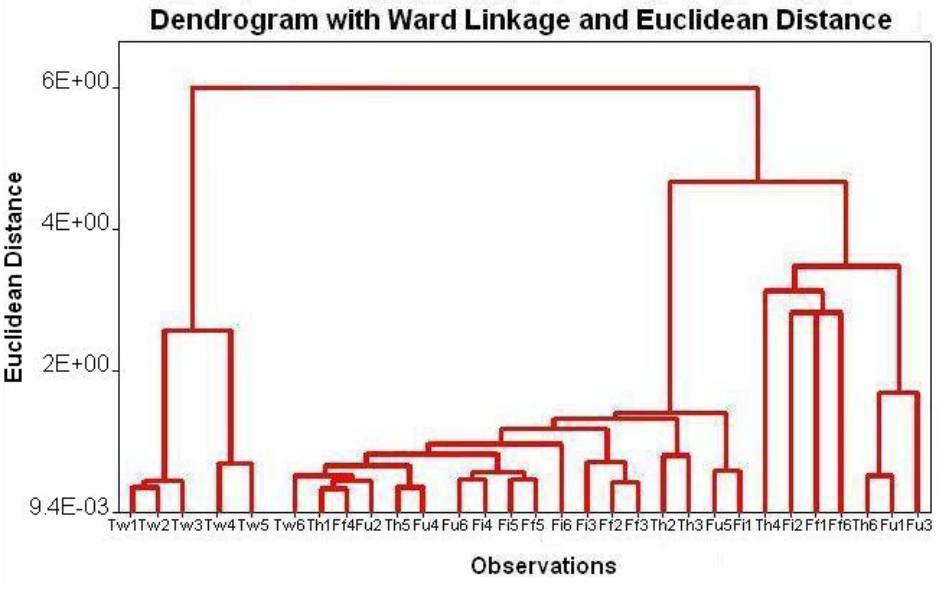

Fig. 7. Dendrogram derived from first six eigenvectors of various sample sizes and complete data set. Symbols as in Fig. 3. The associated numbers 1–6 represent eigenvectors 1–6, respectively.

which are important for a particular study. Therefore, they are more or less fixed. On the other hand, the number of samples is deterministically chosen on the criteria of time and cost of collection and analysis of samples. Field data collection is often considerably time consuming, and the analysis of collected samples such as water or soil samples is expensive. This puts constraints on sample size. Thus, sampling is restricted and sample sizes are usually small even for large areas. This paper attempts to examine the effect of sample size on eigenstructure in the context of a real data set pertaining to environmental science. It must be mentioned that most of the studies of the impact of sample size on PCA ordinations have been conducted on simulated data sets with different properties such as standard deviation of variables and correlation structure of the resemblance matrix. Therefore, evidently, many of the results are dependent on the specific properties of simulated data sets. The sample-to-variable ratio has not been investigated here, though obviously it will vary with the sample size. The conclusions of different workers that are mostly based on simulated data are highly contradictory, and most of such studies usually recommend a sample size of 200–500, which is not realistic for ecologists or environmentalists. The ecological data sets because of time and economic constraints (high cost of sample collection and analysis) are much smaller (e.g. N = 20–80). These data sets are often subjected to PCA as a data exploratory technique. Our study showed that a sample size N = 40 was sufficient to achieve the stability of eigenvalues and eigenvectors of PCA. Barrett and Kline (1981) recommended a minimum sample size N = 50 for behavioural studies. Forcino (2012) prescribed a sample size N = 50 in paleocommunity research with regard to the recovery of first few components that are most often used for the explanation of trends in the data structure. Because of a fair bit of consistency in PCA ordinations of various sample sizes compared to that of complete data set (N = 55), particularly with respect to first two components (which are generally interpreted by ecologists or environmental scientists), we conclude that small sample sizes (e.g. $N \geq 40$) may be used when sampling and analysis of collected samples (e.g. water or soil samples) is expensive. However, larger sample sizes could be more reliable (e.g. N = 50–80), though there must be a trade-off between sampling effort and cost, on one hand, and the quality of information extracted from PCA, on the other hand. Nonetheless, it should be mentioned here that the study was based on a real environmental data set that comprised of continuous variables without zero entries. Whilst ecologists or environmentalists often use species data sets in vegetation analysis studies (or other communities) for which the data matrix usually contains excessive zero entries (sparse matrix). Such data sets would probably require greater sample sizes to attain stability of eigenstructure.

Therefore, we are led to conclude that a sample size of 40 or 50 is sufficient in ecological and environmental studies to recover the first few components that are necessary to explore, comprehend and summarise the multivariate data.

References

Anderson, M.J. & Wilis T.J. (2003). Canonical analysis of principal coordinates: a useful method of constrained ordination for ecology. *Ecology*, 84, 511–525. DOI: 10.1890/0012-9658(2003)084[0511:CAOPCA]2.0.CO;2.

APHA, (1992). *Standard methods for the examination of water and waste water*. American Washington: Public Health Association.

Bandalos, D.L. & Boehm-Kaufman M.R. (2009). Four common misconceptions in exploratory factor analysis. In

C.E. Lance & R.J. Vandenberg (Eds.), *Statistical and methodological myths and urban legends* (pp. 61–87). New York: Routledge Publisher.
Barrett, P.T. & Kline P. (1981). The observation to variable ratio in factor analysis. *Personality Study and Group Behaviour, 1*, 23–33.
Bray, J.R. & Curtis J.T. (1957). An ordination of the upland forest communities of Southern Wisconsin. *Ecol. Monogr.*, 27, 325–349. DOI: 10.2307/1942268.
Bryant, F.B. & Yarnold P.R. (1995). Principal components analysis and exploratory and confirmatory factor analysis. In L.G. Grimm & R.R. Yarnold (Eds.), *Reading and understanding multivariate statistics (p*p. 99–136). Washington: American Psycholgical Association.
Burd, B.J.A., Nemec, A. & Brinkhurst R.O. (1990). The development and application of analytical methods in benthic marine faunal studies. *Adv. Mar. Biol.*, 26, 169–247. DOI: 10.1016/S0065-2881(08)60201-1.
Cadima, J. & Jolliffe I.T. (1995). Loadings and correlations in the interpretation of principal components. *Journal of Applied Statistics*, 22, 203–214. DOI: 10.1080/757584614.
Cattell, R.B. (1966). The Scree test for the number of factors. *Multivariate Behavioral Research*, 1, 245–276. DOI: 10.1207/s15327906mbr0102_10.
Cattell, R.B. (1978). *The scientific use of factor analysis in behavioral and life sciences*. New York: Plenum Press.
Chateau, F. & Lebart L. (1996). Assessing sample variability in the visualization techniques related to principal component analysis: Bootstrap and alternative simulation methods. In A. Prats (Ed.), *Proceedings of COMPSTAT 2006*. Heidelberg: Physica Verlag.
Chatfield, C. & Collins A.J. (1980). *Introduction to multivariate analysis*. London, New York: Chapman & Hall.
Comrey, A.L. & Lee H.B. (1992). *A first course in factor analysis*. London: Taylor and Francis.
de Winter, J.C.F., Dodou, D. & Wieringa P.A. (2009). Exploratory factor analysis with small sample sizes. *Multivariate Behavioral Research,* 44, 147–181. DOI: 10.1080/00273170902794206.
Dengler, J., Lobel, S. & Dolnik C. (2009). Species constancy depends on plot size a problem for vegetation classification and how it can be solved. *J. Veg. Sci.*, 20, 754–766. DOI: 10.1111/j.1654-1103.2009.01073.x.
Diaconis, P. & Efron B. (1983). Computer-intensive methods in statistics. *Sci. Am.*, 248, 116–130. doi:10.1038/scientificamerican0583-116
Dochtermann, N.A. & Jenkins S.H. (2011). Multivariate methods and small sample sizes. *Ethology*, 117, 95–101. DOI: 10.1111/j.1439-0310.2010.01846.x.
Fasham, M.J.R. (1977). The comparison of nonmetric multidimensional scaling, principal component analysis and reciprocal averaging for the ordination of simulated coenocline and coenoplanes. *Ecology*, 58, 551–561. DOI: 10.2307/1939004
Forcino, F.L. (2012). Multivariate assessment of the required sample size for community paleoecological research. *Palaeogeo. Palaeoclimatol. Palaeoecol.,* 315–316, 134–141. DOI: 10.1016/j.palaeo.2011.11.019.
Gamito, S. & Raffaelli D. (1992). The sensitivity of several ordination methods to sample replication in benthic surveys. *J. Exp. Mar. Biol. Ecol.*, 164, 221–232. DOI: 10.1016/0022-0981(92)90176-B.
Gauch, H.G. & Whittaker R.H. (1972). Comparison of ordination techniques. *Ecology*, 53, 868–875. DOI: 10.2307/1934302.
Gauch, H.G., Whittaker R.H. & Wentworth T.R. (1977). A comparative study of reciprocal averaging and other ordination techniques. *J. Ecol.*, 65, 157–174. DOI: 10.2307/2259071.
Gauch, H.G., Whittaker R.H. & Singer S.B. (1981). A comparative study of nonmetric ordinations. *J. Ecol.*, 69, 135–152. DOI: 10.2307/2259821
Gehlhausen, S.M., Schwartz, M.W. & Augspurger C.K. (2000). Vegetation and microclimatic edge effects in two mixed mesophytic forest fragments. *Plant Ecol.*, 147, 21–35. DOI: 10.1023/A:1009846507652.
Goff, F.G. & Mitchell R. (1975). A comparison of species ordination results from plot and stand data. *Vegetatio*, 31, 15–22. DOI: 10.1007/BF00127871.
Goodall, D.W. (1953). Objective methods for the classification of vegetation. III. An essay in the use of factor analysis. *Aust. J. Bot.*, 1, 39–63. DOI: 10.1071/BT9530039.
Gorsuch, R.L. (1983). *Factor analysis*. Hillsdale NJ: Lawrence Erlbaum Associates.
Hatcher, L. (1994). *A step-by-step approach to using the SAS system for factor analysis and structural equation modeling*. Cary: SAS Institute.
Hill, M.O. (1973). Reciprocal averaging: an eigenvector method of ordination. *J. Ecol.*, 61, 237–249. DOI: 10.2307/2258931.
Hill, M.O. & Gauch H.G. (1980). Detrended correspondence analysis: an improved technique. *Vegetatio*, 42, 47–58.

DOI: 10.1007/BF00048870.
Hirosawa, Y., Marsh, S.E. & Kliman D.H. (1996). Application of standardized principal component analysis to land-cover characterization using multi temporal AVHRR data. *Remote Sens. Environ*., 58, 267–281. DOI: 10.1016/S0034-4257(96)00068-5.
Hirst, C.N. & Jackson D.A. (2007). Reconstructing community relationships: the impact of sampling error, ordination approach and gradient length. *Divers. Distrib*., 13, 361–371. DOI: 10.1111/j.1472-4642.2007.00307.x.
Hutcheson, G. & Sofroniou N. (1999). *The multivariate social scientist: Introductory statistics using generalized linear models*. London: Sage Publication.
Jackson, D.A. (1993). Stopping rules in principal components analysis: A comparison of heuristical and statistical approaches. *Ecology*, 74, 2204–2214. DOI: 10.2307/1939574.
Jackson, J.A. (1991). *A user's guide to principal component analysis*. New York: Wiley Inter Science.
James, F.C. & McCulloch C.E. (1990). Multivariate analysis in ecology and systematics: panacea or Pandoras box. *Annu. Rev. Ecol. Evol. Syst.*, 21, 129–166. DOI: 10.1146/annurev.es.21.110190.001021.
Joliffe, I. (2002). *Principal component analysis*. New York: Springer-Verlag.
Kendall, M. (1980). *Multivariate analysis*. London: Charles Griffin.
Kline, P. (1979). *Psychometrics and psychology*. London: Academic Press.
Knox, R.G. & Peet R.K. (1989). Bootstrapped ordination: a method for estimating sampling effects in indirect gradient analysis. *Vegetatio*, 80, 153–165. DOI: 10.1007/BF00048039.
Lawley, D.N. & Maxwell A.E. (1971). *Factor analysis as a statistical method*. New York: Macmillan.
Legendre, P. & Birks H.J.B. (2012). Clustering and partitioning. In H.J.B. Birks, A.F. Lotter, S. Juggins & J.P. Smol (Eds.), *Tracking environmental change using lake sediments Vol.* 5: *Data handling and numerical techniques* (pp. 167–200). Dordrecht: Springer. DOI: 10.1007/978-94-007-2745-8_7.
MacCallum, R.C., Widaman, K.F., Zhang, S. & Hong S. (1999). Sample size in factor analysis. *Psychological Methods*, 4, 84–99. DOI: 10.1037/1082-989X.4.1.84.
MacCallum, R.C., Widaman, K.F., Preacher, K.J. & Hong S. (2001). Sample size in factor analysis: The role of model error. *Multivariate Behavioral Research*, 36, 611–637. DOI**:** 10.1207/S15327906MBR3604_06.
Manjarres-Martinez, L.M., Gutiérrez-Estrada, J.C., Hernando, J.J.A. & Soriguer M.C. (2012). The performance of three ordination methods applied to demersal fish data sets: stability and interpretability. *Fish. Manag. Ecol*., 19, 200–213. DOI: 10.1111/j.1365-2400.2011.00817.x.
Manly, B.F.J. (1998). *Randomization, bootstrap and Monte Carlo methods in biology*. London: Chapman & Hall.
Minchin, P.R. (1987). An evaluation of the relative robustness of techniques for ecological ordination. *Vegetatio*, 69, 89–107. DOI: 10.1007/BF00038690.
Mundfrom, D.J., Shaw, D.G. & Ke T.L. (2005). Minimum sample size recommendations for conducting factor analyses. *International Journal of Testing,* 5, 159–168. DOI: 10.1207/s15327574ijt0502_4.
Okland, R.H., Eilersten, O. & Okland T. (1990). On the relationship between sample size and beta diversity in boreal coniferous forests. *Vegetatio*, 87, 187–190. DOI: 10.1007/BF00042954.
Orloci, L. (1966). Geometric models in ecology 1.The theory and application of some ordination methods. *J. Ecol*., 54, 193–215. DOI: 10.2307/2257667.
Orloci, L. (1978). *Multivariate analysis in vegetation research*. The Hague: Junk.
Osborne, J.W. & Costello A.B. (2004). Sample size and subject to item ratio in principal components analysis. *Practical Assessment Research & Evaluation*, 9, 15–23.
Otypkova, Z. & Chytry M. (2006). Effects of plot size on the ordination of vegetation samples. *J. Veg. Sci*., 17, 465–472. DOI: 10.1111/j.1654-1103.2006.tb02467.x.
Peres-Neto, P.R., Jackson, D.A. & Somers K.M. (2003). Giving meaningful interpretation to ordination axes: assessing loading significance in principal component analysis. *Ecology*, 84, 2347–2363. http://www.jstor.org/stable/3450140
Peres-Neto, P.R., Jackson, D.A. & Somers K.M. (2005). How many principal components? Stopping rules for determining the number of non-trivial axes revisited. *Computational Statistics and Data Analysis*, 49, 974–997. DOI: 10.1016/j.csda.2004.06.015.
Pillar, V. de P. (1999). The bootstrapped ordination re-examined. *J. Veg. Sci*., 10, 895–902. DOI: 10.2307/3237314.
Preacher, K.J. & MacCallum R.C. (2002). Exploratory factor analysis in behavioral genetics research: Factor recovery with small sample sizes. *Behav. Genet.,* 32, 153–161. DOI: 10.1023/A:1015210025234.
Rao, C.R. (1964). The use and interrelation of principal component analysis in applied research. *Sankhya* (*Ser. A*), 26, 329–358. http://www.jstor.org/stable/25049339

Richman, M.B. (1988). A cautionary note concerning a commonly applied eigen analysis procedure. *Tellus B*, 40, 50–58. DOI: 10.1111/j.1600-0889.1988.tb00212.x.
Shaukat, S.S. (1985). *Approaches to the analysis of ruderal weed vegetation*. PhD. thesis, University of Western Ontario, London, Canada.
Shaukat, S.S. & Uddin M. (1989a). A comparison of principal component and factor analysis as an ordination model with reference to desert ecosystem. *Coenoses*, 4, 15–28. http://www.jstor.org/stable/43461254
Shaukat, S.S. & Uddin M. (1989b). An application of canonical and principal component analysis to the study of desert environment. *Abstracta Botanica (Budapest)*, 13, 17–45. http://www.jstor.org/stable/43519176
Shaukat, S.S. & Siddiqui I.A. (2005). Essentials of Mathematical Ecology: Computer Programs in BASIC, FORTRAN and C++. Karachi: Farquan Publishers.
Shaukat, S.S., Sheikh I.H. & Siddiqui I.A. (2005). An application of correspondence analysis, Detrended correspondence analysis and Canonical correspondence analysis to the vegetation and environment of calcareous hills around Karachi. *Int. J. Biol. Biotechnol*., 2, 617–627.
Stauffer, D. F., Garton E.O. & Steinhorst R.K. (1985). A comparison of principal component from real and random data. *Ecology*, 66, 1693–1698. DOI: 10.2307/2937364.
Swan, J.M.A. & Dix R.L. (1966). The phytosociological structure of upland forest at Candle Lake, Saskatchewan. *J. Ecol*., 54, 13–40. DOI: 10.2307/2257657.
Ter Braak, C.J.F. (1986). Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. *Ecology,* 67, 1167–1179. DOI: 10.2307/1938672.
Velicer, W.F. & Fava J.L. (1998). The effects of variable and subject sampling on factor pattern recovery. *Psychological Methods*, 3, 231–251. DOI: 10.1037/1082-989X.3.2.231.
Walker, S.C. & Jackson D.A. (2011). Random-effects ordination: describing and predicting multivariate correlations and co-occurrences. *Ecol. Monogr*., 81, 635–663. http://www.jstor.org/stable/23208478
Whittaker, R.J. (1987). An application of detrended correspondence analysis and nonmetric multidimensional scaling to the identification and analysis of environmental factor complexes and vegetation structures. *J. Ecol*., 75, 363–376. DOI: 10.2307/2260424.
Wikum, D.A. & Wali M.K. (1974). Analysis of a North Dakota gallery forest: Vegetation in relation to topographic and soil gradients. *Ecol. Monogr*., 44, 441–464. DOI: 10.2307/1942449.