# Analyzing abundance of different earthworm species with respect to different months of the year in a reserve forest area of Manipur: Application of Principal Component Analysis.

Ksh. Anand Singh*
Assistant Professor
Department of Statistics
Manipur University
Imphal, India.

Abstract

*The technique of Principal Component Analysis (PCA) is generally applied to condense information contained in a large number of variables into a smaller set of new composite dimensions with minimum loss of information. Ecologists have long been using the technique of PCA to effectively reduce the dimensionality of large ecological datasets. The application of PCA to analyse the abundance of earthworm species in a mixed reserve forest area of Manipur has been carried out in this paper. Data on 12 species of earthworm collected during the 12 different months of the year from six replicates are used in the analysis. Necessary steps for the analysis are systematically highlighted and performed to reach to a valid conclusion. The result of the eigenanalysis using covariance matrix shows that the first and second components accounted for nearly 95% of the total variability in the original data so that the first two components will be enough to retain for further study. Graphical interpretations are also presented accordingly.*

Key words: PCA, dimensionality, abundance, correlation matrix, eigen-analysis.

## 1. Introduction

Ecologists have long been using the technique of PCA to analyze large ecological dataset with the purpose of effectively reducing the dimensionality of the dataset. PCA as a statistical multivariate technique uses orthogonal transformation to convert a set of correlated observations into a set of orthogonal uncorrelated axes called principal components (James & McCulloch[1] 1990; Robertson, et al[2]., 2001). PCA tries to condense large ecological datasets without compromising much of the information contained in the original data set. It reduces P-original variables (dimensions) of the data set into fewer number of dimensions as principal components where each dimension is defined by a normalized linear combination of the p-original variables.

Ecologists are generally interested in understanding patterns in the distribution and abundance of organisms and the factors (environmental) that control the pattern. In the most common use of PCA in ecology the investigator collects a set of abundance or importance values for many species over many samples and then organizes the data into a matrix. The elements of the data matrix may be converted, rescaled or transformed as appropriate; either a variance- covariance or correlation matrix is then computed. The eigenvalues and eigenvectors of this matrix is then computed by solving the matrix equation

$$A\underset{\sim}{u} = \lambda \underset{\sim}{u} \qquad \ldots\ldots \quad (1)$$

Where $\underset{\sim}{u}$ is the eigenvector and $\lambda$ is a scalar and A is the sample covariance or correlation matrix. There are as many $\lambda$'s and $\underset{\sim}{u}$'s as there are rows (species) in A.

Geometrically, PCA will return scatter plots on a new set of axes established by a rigid rotation of the original species axes. The eigenvectors are the direction cosines relating the species axes to the component axes.

Each sample point has a score on each axis, calculated as the sum (over all species) of the products of each species importance value for that sample times that species' eigenvector. As such the principal component scores are linear combination of species importance scores and will behave much like species importance scores. The main attempt in principal component analysis is the examination of the eigenvector coefficient s which defines the extracted axes. The eigen vector components are called the loadings with respect to each principal component. The $i^{th}$ component of the $r^{th}$ coefficient is the loading of the $r^{th}$ principal component on the $i^{th}$ response variate. That is, if $v_r = \alpha^{(r)/}X$ is the $r^{th}$ Principal Component, then $i^{th}$ component of $\alpha^{(r)}$ gives the loading of $v_r$ on the $i^{th}$ response variate $X_i$ of $X^{/} = (X_1, X_2,\ldots..,X_p)$.

## 2. Source of Data

Earthworms are widely distributed in most ecosystems in natural and plantation forest, grasslands and agro-ecosystem. Earthworms represent a major portion of (>80%) the soil invertebrate biomass and involve in the process of soil formation and maintenance of soil fertility. Distribution and abundance of earthworms are governed by several ecological factors viz temperature, moisture, pH, available organic matter etc. The number of species in a given earthworm community, which is the simplest measure of species diversity range from 1 to 15 species (Edwards and Bohlen[3], 1996).

The diversity of the earthworm community at a given locality is influenced by the characteristics of the soil, climate and organic resources of the locality as well as its history of land use and soil disturbance. Earthworms perform several beneficial functions which include decomposition of organic matter that helps in increasing soil nutrients, increase air water infiltration, soil aggregation, increase the availability of plant nutrients, worm cast as biofertlilizer etc.

In this paper data on 12 species of earthworm collected from mix reserve sub-tropical forest ecosystem located at Koirengei about 12 km from Imphal city is used for analysis. The collection site lies at $14^0\ 54^{/}\ 49.74^{//}$ longitude and it is protected from various biotic interference. It has a moderate to steep slopes at certain sites. Numbers of species of earthworms are collected from six different replicates at the study site. Each replicate has a depth of 10 cm inside the soil from the surface. In the present study the maximum number of counts from six replicates is considered for analysis thinking that the abundance of species will be largely depend on the maximum total count of that species. Data on the number of counts of species are thus obtained for the 12 months of the year during 2012 and 2013(Sharon Haokip[4], 2015).

The following types of earthworm species are found from the study site- Drawida sp (Type-1), Drawida Japouica (Type-2), Drawida nepalensis(type-3), Eutyphoeus sp.(type-4), Pontoscolex corethrurus (Type-5), Kanchuria sumerianus (Type-6), Dichogaster bolani (Type-7), Amynthasa corticis (Type-8), Metaplure anomala (Type-9), M. Houlleti (Type-10), Periyonyx Shillongensis (Type-11) and Enchytracidae (Type-12).

## 3. Objectives of the study:

1) To explain the variability in the abundance of species of earthworms with respect to different months of the year.
2) To employ the technique of PCA to achieve the above and thereby reducing the dimensionality of the data set by interpreting only the first few components.

## 4. Analysis and results

PCA requires a number of essential steps to reach to valid conclusions. First the row data has to be set up in a matrix form (e.g. species by months) in order to make it suitable for input to an existing program (Software).

### 4.1 Multivariate Normality

PCA assumes that the underlying data structure is multivariate normal. Geometrically a multivariate normal distribution exists when the data cloud is hyperellipsoid with normally varying density around the centroid

(Beals[5], 1973). Such a distribution exists when each variable has a univariate normal distribution about fixed values on all others (Blalock[6], 1979). If the dataset has a multivariate normal, then the linear axes (i.e. P.C's) will adequately display the data. Further, second and subsequent component axes maintain strict independence (orthogonality) only when the underlying structure of the data is multivariate normal. When the datasets are not multivariate normal there is usually some redundancy in the principal components.

In this paper normality assumption is checked and verified for all the variables (months) by using graphical methods. Four different plots viz. Histogram, boxplots, density plot and quantile-quantile plot (qq plot) are drawn and verified that we can without much difficulty assume that the variables are coming from a nearly normal distribution. The graphs are shown in Fig.1 and 2 for two months (May and December).

## 4.2 Outliers

In Fig.1 and 2, we use the qq plot to detect any outlier. It is quite apparent from the two month's graph of the qq plot that no significant outlier is present in the data.

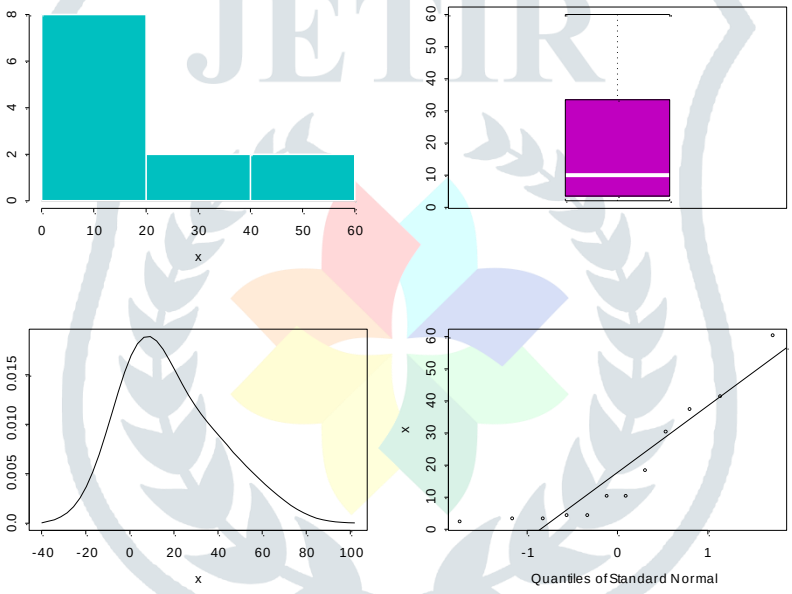


**Fig. 1: Test for normality (**Month = May)

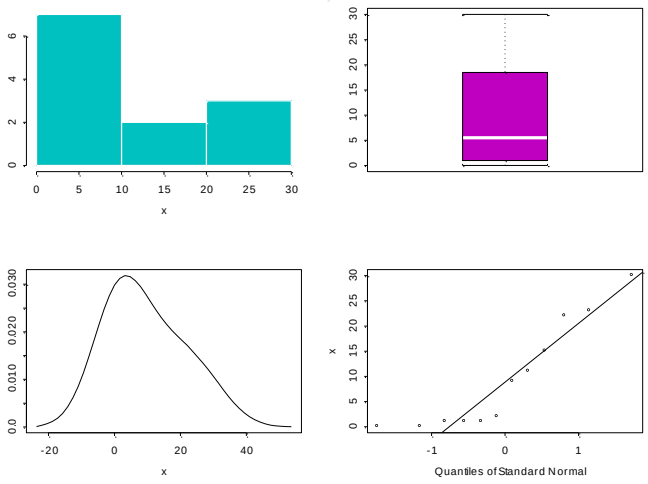


**Fig. 2: Test for normality (**Month = December)

### 4.3 Extracting the principal components

Principal Components are not directly extracted from the original raw data matrix. In general the covariance matrix or the correlation matrix is used to extract the principal components. A component analysis using the covariance matrix gives more weight to the variables with larger variance whereas analysis based on the correlation matrix gives equal weight to all the variables. PCA using the two different matrices are different. Use of correlation matrix is always more appropriate if the scale or unit of measurement differ among the variables (Noy-Mier et al[7]; 1975). However if the variables share a common measurement scale, the covariance matrix could be more appropriate or desirable. The present analysis which consists of dataset with month by species abundance, the use of covariance matrix will be more appropriate as more abundant species will have greater absolute variances.

### 4.4 The Eigenvalues

Computationally, PCA is essentially solving the characteristics equation

$$|\sum - \lambda I| = 0 \qquad \ldots. (2)$$

Where $\sum$ is the covariance or correlation matrix, $\lambda$ is a scalar (eigenvalue) and I is the identity matrix. For a pxp matrix $\sum$ we get p eigenvalue solutions of (2).

As for the present analysis having 12 months (Jan. – Dec.) and 12 different species we obtain a 12x12 covariance matrix from which we extracted 12 eigenvalues. The variances (Std. deviations), the proportion of variance, and cumulative proportion for each of the first six components are presented in Table 1.

**Table 1: Standard deviations, proportion of variance and Cumulative proportion for the first six principal components**

| | Comp. 1 | Comp. 2 | Comp. 3 | Comp. 4 | Comp. 5 | Comp. 6 |
|---|---|---|---|---|---|---|
| **Standard deviation** | 3.373583 | 0.549664 | 0.412373 | 0.295789 | 0.168436 | 0.148585 |
| **Proportion of Variance** | 0.948422 | 0.025178 | 0.014171 | 0.007291 | 0.002364 | 0.00184 |
| **Cumulative Proportion** | 0.948422 | 0.973599 | 0.98777 | 0.995061 | 0.997425 | 0.999265 |

All the 12 eigenvalues are all non-negative, the larger the value is, the greater the sample variability on that principal component. Thus the component corresponding to the largest variance (**λ**) captures the maximum variation on the sample variance-covariance. A principal component with lesser and lesser eigenvalues reduces the explanatory power of variability among sample points.

The first eigenvalue is always the largest. Therefore the first principal component defines the dimension or gradient with the single highest variance. The second eigenvalue and its corresponding principal component represent the largest variance in a dimension orthogonal to the first principal component. Thus the second component provides the greatest explanation of the sample variability after the first component has done its best. And so on so forth the remaining principle components.

### 4.5 Principal component loadings

As a byproduct of the principal component analysis (PCA) we obtained eigenvectors associated with each eigenvalues by solving the matrix equation in (2). The number of components of each eigenvector equals number of the variables in the data matrix A. Each component of the eigenvector defines the loading of the corresponding principal component on the particular response variate. The strength of the monthly species contribution to a component axis is indicated by the magnitude of the eigenvector components.

**Table 2: Loadings for the first three components**

| Month | Comp. 1 | Comp. 2 | Comp. 3 |
|---|---|---|---|
| **January** | 0.000 | 0.103 | 0.000 |
| **February** | 0.000 | 0.000 | 0.193 |
| **March** | 0.148 | -0.154 | 0.491 |
| **April** | 0.177 | 0.216 | 0.142 |
| **May** | 0.178 | 0.212 | 0.207 |
| **June** | 0.251 | 0.265 | 0.322 |
| **July** | 0.337 | 0.444 | -0.287 |
| **August** | 0.491 | -0.307 | 0.000 |
| **September** | 0.495 | 0.268 | 0.000 |
| **October** | 0.407 | -0.624 | -0.146 |
| **November** | 0.253 | 0.000 | 0.401 |
| **December** | 0.000 | 0.205 | -0.182 |

### 5. Assessing the importance of the principal components

An important decision in principal component analysis (PCA) is to determine on how many principal components to retain for interpretation and use for later analyses. There are in the literature some approaches to judge the number of components to retain but each of them have their own merits and demerits.

The latent root criteria (Guttman[8], 1954; Cliff[9], 1988) are most reliable when the number of variables is 20 to 50. Another criterion is the scree plot criterion in which eigenvalues are plotted against the component number in the order of extraction. The shape of the resulting curve (Fig.3) is used to evaluate the appropriate number of components to retain. The Broken stick criteria proposed by Frontier[10] (1976) assumes that if the total variance is distributed randomly among the components, then the scree plot will show a broken stick distribution (Fig. 3).

Relative percent of the total variability explained by each component gives another important criteria in determining the number of components to retain for further analysis. The relative percent explained by the $i^{th}$ principal components is defined by

$$\Phi_i = \lambda i / \sum \lambda i \qquad \ldots \quad (3)$$

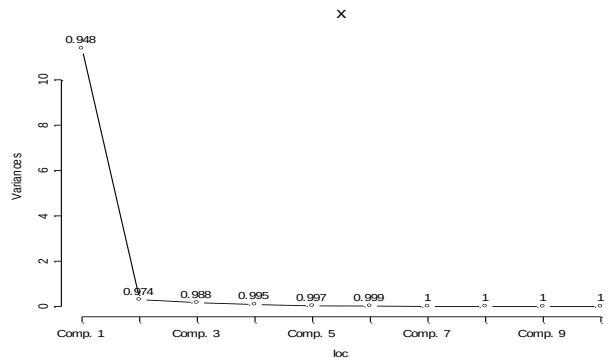


Fig. 3: Scree plot

It measures how much of the total variance is accounted for by each of the principal components. The cumulative percent variance of all principal components is 100%. Generally, the cumulative percent variance of the first few components will be high indicating that the data structure could be effectively summarized by the first few components. Unfortunately there is no standard value on how much of the total variance should the first few components explain 90% is generally agreed upon.

Referring to Table 1 the percentage of variance accounted for by the first principal component is approximately 94.8%. This shows that the total variability in the original variables (months) could be effectively explained by the first component alone. However, the scree plot criteria in Fig.3 suggests that we could retain the first and second principal components together accounting for nearly 97% of the total variability in the data. The remaining components contribute very little information and therefore we retain only the first two components without compromising much information of the original variables.

## 6. Interpreting the principal components

The extracted principal components can be interpreted by (i) examining the relationship between the individual variables and the principal components and (ii) examining the relative positions of the sampling entities in the ordination space. The principal component loadings in Table 2 play an important role in interpreting the significant components. As a rule of the thumb and without any mathematical proposition, principal component loadings greater than 0.3 (absolute value) is considered significant (Hair, Anderson, and Tatham[11], 1987). Loadings greater that 0.4 (absolute value) is considered more significant and thus more important.

Several months are important in each of the first two components as evident from the loadings table in Table 2. We interpret the components on those months with loadings greater than 0.3 or less than -0.3. The first principal component gives maximum loadings to September, August, October and July showing that this component represents a gradient fro m these four months. All the values of these loadings are positive indicating that all these months have positive correlation. Variability in the abundance of earthworm species could be explained during these four months as explained by the first principal component. These four months include part of the rainy season and summer with hot and moist climate. The second component represents a gradient from October, July and August whereas the third component represents November and June. We notice the inverse relation between October, August and July indicating that large number of species during July is associated with low number of species during October and August.

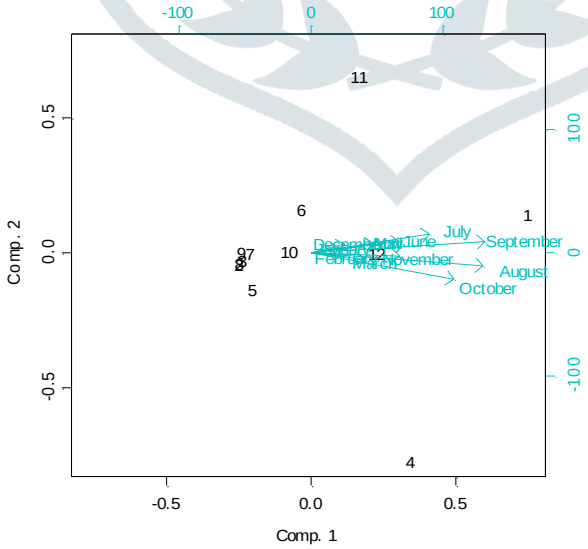


Fig. 4 Bi-plot showing graphical representation of the principal component loadings

To have a comprehensive view of both the principal components and the original variables we use the biplot in fig. 4. The biplot (Gabriel,[12] 1971) allows us to represent the original variables and the transformed observations on the principal component axes. In the biplot, the x-axis represents the scores for the first principal components and the y-axis represents the scores for the second principal components. The original variables are represented by arrows which graphically indicate the proportion of the original variance explained by the first two principal components. The direction of the arrows indicates the relative loadings of the first and second components. The month September has the highest loading followed by August in the first principal component. This is indicated by longer arrows for September and August. The months of October and August have negative signs of their loadings on the second component which is indicated by slightly downward slopes in fig. 4.

## 7. Conclusion

Because of the availability of efficient software packages that enable computation of PCA, it has become one of the most useful tools in ecological data analysis. However, one should understand the limitations of PCA before taking further applications of the technique. For example, PCA is only capable of detecting gradients that are intrinsic to the data set and thus more important gradients that are not measured using the selected variables may distort the relationships that are intrinsic to the data (Gauch,[13] 1982). Another limitation is that by looking at the first few components one may overlook a later axis that explains most of the variations in some variables. The present paper is primarily aimed at illustrating the necessary steps in taking up PCA applied to ecological data without considering much of the limitations of the method. Thus the results obtained above may be subject to further investigation for obtaining more accurate results of future use.

## References

[1] Beals, E.W., Ordination: mathematical elegance and ecological naivete. *Journal of Ecology* **61**(1973): 23-35.

[2] Blalock, H.M., Jr.: Social Statistics (book). *New York:* McGraw Hill (1979).

[3] Cliff, N. The eigenvalues-greater-than-one rule and the reliability of components. *Psychological Bulletin*(1988), 103:276-179.

[4] Edwards, C. A. and Bohlen, P. J. Biology and ecology of earthworm. (book 3rd edn.), Chapman and Hall, London(1996).

[5] Frontier, S. Etude de la decroissance des valeurs propres dans une analuze en composantes principles: comparision avee le modele de baton brise. Journal of Experimental Marine Biology and Ecology(1976), 25:67-75.

[6] Gabriel,K.R. The biplot graphical display of matrices with applications to principal component analysis (1971). *Biometrika,* 58:453 – 467.

[7] Gauch, H.G. Noise reduction in eigenvector ordination. (1982). *Ecology*, 63 1643 – 1649.

[8] Guttman, L. Some necessary conditions for common factor analysis. *Psychometrika*(1954), 19:149-161.

[9] Hair, J.F., Anderson, R.E. and Tatham, R.L. Multivariate Data Analysis 2nd Edition (book) (1987) New York: Macmillan.

[10] James, F.C. & McCulloch, C.E.. Multivariate Analysis in Ecology and Systematics: Panacea or Pandora's Box? *Annual*

*Review of Ecology and Systematics*, **Vol. 21**(1990),129−166.

[11] Noy-Mier, I., Walker, D., and Williams, W.T., Data Transformation in ecological ordination. II On the meaning of data standardization. *Journal of Ecology* **63**(1975): 779-800.

[12] Robertson, M.P., Caithness, N. & Villet, M.H. A PCA-based modelling technique for predicting environmental suitability for organisms from presence records. *Diversity and Distributions*, **Vol. 7** (2001), 15−27.

[13] Sharon Haokip. Ecological study of Earthworm in a subtropical forest ecosystem, Manipur. Thesis submitted to Manipur University, (2014).