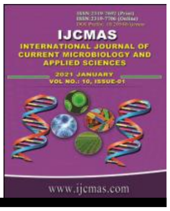
EXCELLENT PUBLISHERS

International Journal of Current Microbiology and Applied Sciences
ISSN: 2319-7706 **Volume 10 Number 01 (2021)**
Journal homepage: http://www.ijcmas.com


IJCMAS
INTERNATIONAL JOURNAL OF CURRENT MICROBIOLOGY AND APPLIED SCIENCES
2021 JANUARY
VOL NO.: 10, ISSUE-01
www.ijcmas.com

Original Research Article https://doi.org/10.20546/ijcmas.2021.1001.362

Characterization of Environmental Covariates of Coimbatore District using Principal Component Analysis

R. Priyadharshini[1], M. Radha[*], R. Kumaraperumal[2],
G.Vanitha[3] and Balaji Kannan[4]

[1]Agricultural Statistics, [2]Remote Sensing & GIS, [3]Computer Science, [4]Soil and Water Conservation Engineering, Tamil Nadu Agricultural University, Coimbatore, India

**Corresponding author*

ABSTRACT

Keywords

Principal component analysis (PCA), Eigenvalues, Eigenvectors, correlation plot

Article Info

Accepted:
20 December 2020
Available Online:
10 January 2021

The principal component analysis (PCA) is used to identify the most influencing variable. It is one of the statistical techniques for reducing the dimension of the data. The study was conducted in Coimbatore district, Tamil Nadu with 340 profile points. More than 30 environmental covariates are available for this analysis. To make the analysis easier and accurate the data has to be reduced. The principal components (PC1, PC2, PC3 and PC4) are selected for further analysis which accounts for 53.84% of variation. From the selected four principal components the variables which are having higher percentage of variation were identified. Hence it is one of the easiest methods to predict the most influencing variable using R software.

Introduction

The environmental covariates are a key approach in spatial prediction of soil properties which represent the soil forming factors. The environmental covariates are classified into five categories based on CLORPT model, namely climate, organism, relief or topography and parent material (14) and these variables are briefly described in Table 1. Among all categories, climate is the most important factor. Topography tends to be a passive factor for soil formation, as it has a major influence on the soil distribution and vegetation. Each category has different set of environmental covariates. It is very difficult to predict the results with larger set of environmental covariates. To identify the most influencing environmental covariates, principal component analysis (PCA) is used (4). PCA is used for different purposes such as interpreting and visualizing data, finding interrelations between variables in the data, decreasing the number of variables for making further analysis simpler (3) and for many other similar reasons**.** PCA is a very versatile method that enables an interpretation of datasets that can include, for example, multilinearity, missing values, categorical data, and imprecise measurements. "PCA was first coined by Pearson (1901), and developed independently by Hotelling (1933)". It is one of the methods for reducing variables without much loss of information. The main use of principal component analysis (PCA) is to define trends in the data and to guide the data by emphasizing their similarities and differences.

Materials and Methods

The study was conducted in Coimbatore district of Tamil Nadu located between 11°24'23" to 10°13'12" N Latitude and 76°39'20" to 77°18'00" E longitude with an area of about 4721.28 sq.km as shown in Figure 1. The data was obtained from the Department of Remote Sensing and GIS, Tamil Nadu Agricultural University. The environmental covariates such as Satellite data (Green, Blue, NIR, Red), Agro Climatic Zones (ACZ), Agro Ecological Zones (AEZ), Western Ghats (WG), Maximum Temperature, Minimum Temperature, Rainfall, Land Use and Land Cover (LUCU), Elevation, Hill shading, Aspect, Convergence Index, Climate, General curvature, Longitudinal Curvature, LS factor, Maximum Curvature, Mid Slope Position, Minimum Curvature, Physiography, Plan Curvature, Profile Curvature, Slope, Tangential Curvature, TRI, TWI, TCA, Total Curvature Valley, Depth, Geomorphology and Geology (14) are used. Some of the independent variable such as Land use & land cover, Physiography, Geomorphology, Physiography, Western Ghats, Geology, ACZ and AEZ are in shape file format (Polygon feature). Hence, these variables are converted to raster format using Feature to Raster tool in ArcGIS software. The extracted environmental covariate includes remotely sensed spectral data and derivatives of terrain attributes. Totally 33 terrain attributes were layer stacked using R software. In order to indentify the most influencing variables, the selected 340 points with corresponding 33 layer staked variables are used to run the principal component analysis (PCA) in R software.

Principal component analysis

Principal component analysis (PCA) is a statistical tool for dimensional reduction that is often used by converting a large set of variables into a smaller one that also includes much of the information in the larger set to reduce the dimensionality of larger data sets. Principal component analysis (PCA) uses single value decomposition (SVD) to reduce the dimension of the data. PCA is derived from the decomposition of a covariance or a matrix of correlation. It uses orthogonal transformation (15) to translate a set of measurements of potentially associated variables into a set of values of linearly uncorrelated variables called principal components,

Principal components are the linear combinations of the initial variables that are squeezed to contain maximum information in the first principal component. The first few principal components hold maximum variability of the model. The second and third component explains less variation compared to first component. These are the uncorrelated variable by discarding the component containing the lowest information. By ranking the eigenvector of the covariance matrix that explains the variation of the principal components gives the order of significance. This analysis was used on the environmental covariates to reduce the dimension and to identify the most influencing variable of the data.

Steps involved in Principal Component Analysis (PCA)

Step 1: Standardization of the dataset.

The Principal Component Analysis initially standardizes the data to remove the scale difference between the variables and convert to z values. Scaling is done to remove the difference in their range of the variable that affects the performance of the analysis.

$$z = \frac{variable\ value - mean}{Standard\ deviation}$$

Mean is computed by dividing sum of the observations with the number of observations. Let x_1, x_2, x_3,…, x_n be the number of observations. The mean $\bar{x}$ is calculated by

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Standard Deviation is the best measure of dispersion. It is calculated from mean of squared deviation of individual values from their mean. It is always positive ranges from zero to infinity.

$$\sigma_x = \sqrt{\frac{(x_i - \mu)^2}{N}}$$

Step 2: Calculation of covariance matrix for the features in the dataset

The analysis computes the covariance (pxq) symmetric matrix that explains the correlation of the variables where p is the number of dimensions and q is the number of variables. The covariance matrix was calculated using the matrix equation

$$\Sigma = \frac{1}{n-1}\left((x - \bar{x})^T (x - \bar{x})\right)$$

Where $\bar{x}$ is the mean vector, $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Step 3: Calculation of eigenvalues and eigenvectors for covariance matrix.

The eigenvectors of the covariance matrix are referred as principal components (5). The eigenvectors and eigenvalues are constructed from the covariance matrix. They explain the percentage of variation of each principal component.

det (λI - A) = 0 & (λI - A) v = 0I

Where λ is eigenvalue, v is eigenvector, A is square matrix and det is the determinant of the matrix. The eigenvectors of a matrix are perpendicular to each other. The eigenvectors provides the information about the pattern of the given data.

Step 4: Picking k eigenvalues and formation of matrix of eigenvectors.

For n variable, there will be n eigenvalues and eigenvectors. The eigenvalues are ordered from largest to smallest to select the components in the order of significance. The eigenvalues likely to be greater the one are selected to form the principal components which explains the maximum variation. Mostly first k eigen values (7) are selected to reduce the dimension of the data.

Step 5: Transformation of the original matrix

The data has to be transformed by multiplying the k eigenvectors with feature matrix. The feature matrix is a matrix of vectors.

The principal component can be selected based on the scree plot (13), where X axis represent the principal components, Y axis represents the variations explained by each component. The scree plot shows the variation captured by each principal component. In Scree plot, the knee point or bending point shows the number of principal component to be selected. An eigenvalue greater than one indicates that PCs account for more variance than accounted by one of the original variables in standardized data. This is commonly used as a cutoff point for which PCs are retained. When the eigenvectors are plotted in scatter plot, the principal eigenvectors fits well with the data. The loading plot shows how each variable characterizes the principal component. The PCA plot shows the clusters of samples based on their similarities.

Results and Discussion

The proportion of variation explained by each eigenvalue is given in the second column in table 2. It shows that the first nine principal components represent an eigenvalues of more than one which accounts for about 74.7% of variation. The first component accounts for maximum total variation as possible, the second component accounts for the remaining variation. The first principal component explains 22.58% variation followed by second, third and fourth component explains 15.72, 9.772 and 5.769 respectively. The scree plot shows the proportion of information retained by each principal component (Figure 2). The bend or knee point in scree plot indicated that the first four principal component was selected for further analysis. The first four principal components account for 53.8% of variation.

Figure 3 and 4 explains the contribution of each variable to first four principal components. Variables that are correlated with PC1 (i.e., Dim.1) and PC2 (i.e., Dim.2) are the most important in explaining the variability in the data set. Variables that do not correlated with any PC or correlated with the last dimensions are variables that are with low contribution and might be removed to simplify the overall analysis. WG, ACZ, Tmax, TWI, DEM, Slope are the variables that contribute more for the first component. GC, Tang Curv, TRI, Max Curv, LC, Min Curv are the variable that contribute for second component.

Table.1 Parameters of Environmental Covariates

Relief/Topography		Climate	Organism	Parent material
Elevation **Hill shading** **Aspect** **Convergence Index** **General Curvature** **Longitudinal Curvature** **LS factor** **Max Curvature** **Mid Slope Position** **Min Curvature** **Physiography** **Plan Curvature**	Profile Curvature Slope Tangential Curvature TRI TWI TCA Total Curvature Valley Depth Western Ghats	Max Temperature Min Temperature Annual Rainfall ACZ AEZ	LULC Satellite data: NIR Green Blue Red	Geomorphology Geology

Table.2 Results of Eigen values and Percentage of variance

Principal Components	Eigen value	Percentage of variance	Cumulative percentage of variance
1	**7.454**	**22.589**	**22.589**
2	**5.188**	**15.721**	**38.31**
3	**3.225**	**9.772**	**48.082**
4	**1.904**	**5.769**	**53.852**
5	1.766	5.352	59.203
6	1.476	4.474	63.677
7	1.307	3.961	67.638
8	1.178	3.57	71.209
9	1.162	3.521	74.729
10	0.914	2.768	77.498
11	0.89	2.698	80.196
12	0.819	2.482	82.678
13	0.8	2.424	85.102
14	0.723	2.19	87.292
15	0.637	1.93	89.222
16	0.596	1.806	91.028
17	0.547	1.656	92.684
18	0.523	1.585	94.269
19	0.412	1.248	95.517
20	0.355	1.077	96.593
21	0.303	0.918	97.512
22	0.26	0.789	98.301
23	0.206	0.623	98.924
24	0.14	0.425	99.35
25	0.086	0.26	99.609
26	0.078	0.238	99.847
27	0.024	0.074	99.921
28	0.019	0.058	99.979
29	0.005	0.014	99.993
30	0.002	0.007	100
31	0	0	100
32	0	0	100
33	0	0	100

Table.3 Results of Eigen vectors of first four principal components

Parameters	PC1	PC2	PC3	PC4
Blue	-0.188	-0.048	**0.408**	-0.047
Green	-0.182	-0.059	**0.411**	-0.059
Red	-0.043	-0.035	**0.125**	-0.155
NIR	-0.145	-0.086	**0.422**	-0.133
ACZ	-0.322	0.026	-0.097	0.072
AEZ	-0.209	-0.01	0.133	**0.219**
WG	**0.322**	-0.026	0.097	-0.072
Tmax	-0.309	0.001	0.039	**0.162**
Tmin	-0.268	0.025	-0.2	**0.168**
Rainfall	**0.261**	0.001	-0.109	-0.213
LULC	-0.126	-0.037	0.032	-0.002
Hillshade	-0.051	-0.043	0.014	**0.278**
Aspect	0.006	0.05	-0.009	-0.226
CI	-0.009	0.106	0.033	**0.344**
DEM	**0.29**	-0.007	0.205	-0.099
GC	-0.023	**0.423**	0.07	-0.087
LC	-0.035	**0.364**	0.075	-0.068

LS factor	**0.258**	-0.004	0.116	0.253
Max Curvature	0.032	**0.381**	0.075	0.014
Midslope	0.163	-0.009	-0.033	-0.073
Min Curvature	-0.074	**0.339**	0.044	-0.165
Physio	**0.183**	-0.021	0.071	-0.045
Plan Curv	0.032	0.181	0.044	**0.37**
Profile Curvature	-0.008	0.138	0.075	**0.321**
Slope	**0.289**	0.01	0.068	0.269
Tang Curvature	0.006	**0.392**	0.042	-0.09
TRI	0.006	**0.392**	0.042	-0.09
TWI	**0.29**	0.039	0.071	0.275
TCA	-0.016	-0.033	-0.022	-0.016
Total Curvature	0.073	0.141	0.018	0.139
Valley depth	-0.083	0.072	-0.329	-0.042
Geomorphology	-0.059	0.07	-0.32	-0.039
Lithology	0.024	0.055	-0.259	0.001

Green, Blue, NIR, Red, Agro Climatic Zones (ACZ), Agro Ecological Zones (AEZ), Western Ghats (WG), Maximum Temperature(Tmax), Minimum Temperature(Tmin), Rainfall, Land Use and Land Cover (LUCU), Elevation(DEM), Hill shading, Aspect, Convergence Index(CI), General curvature(GC), Longitudinal Curvature(LC), LS factor, Maximum Curvature(Max Curv), Mid Slope Position, Minimum Curvature(Min Curv), Physiography, Plan Curvature, Profile Curvature, Slope, Tangential Curvature, Terrain Ruggedness Index (TRI), Topographic Wetness Index(TWI), TCA, Total Curvature Valley Depth, Geomorphology and Lithology

Fig.1 Location of the Study area

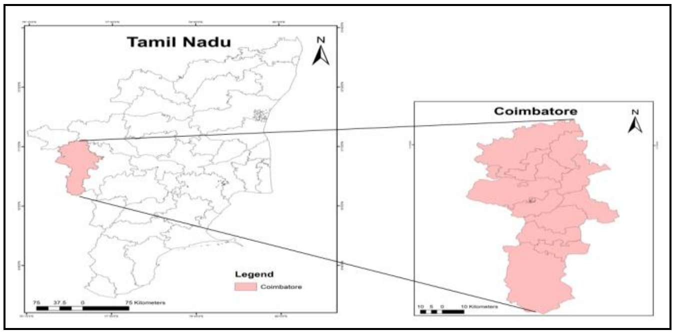


Fig.2 Scree plot of Principal Component Analysis

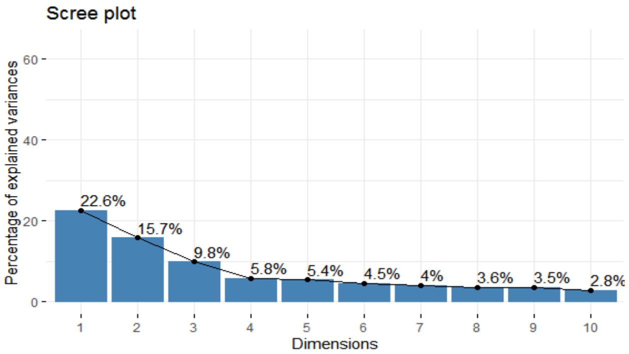


Fig.3 Contribution charts of PC1 – PC2

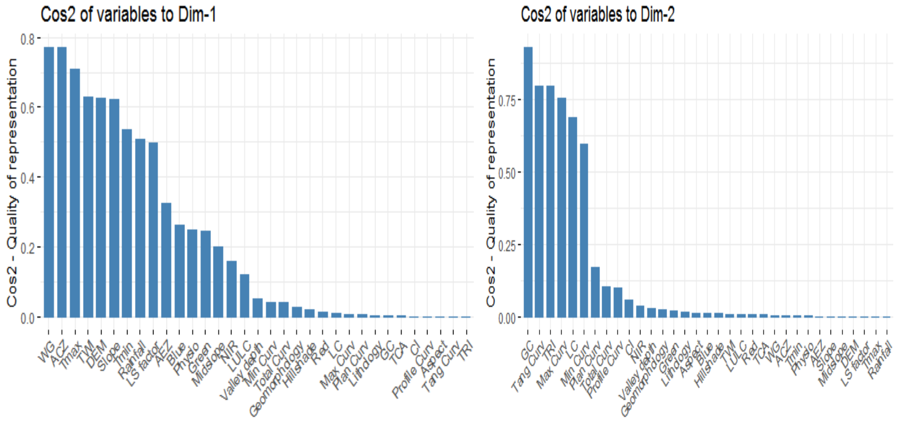


Fig.4 Contribution charts of PC3 – PC4

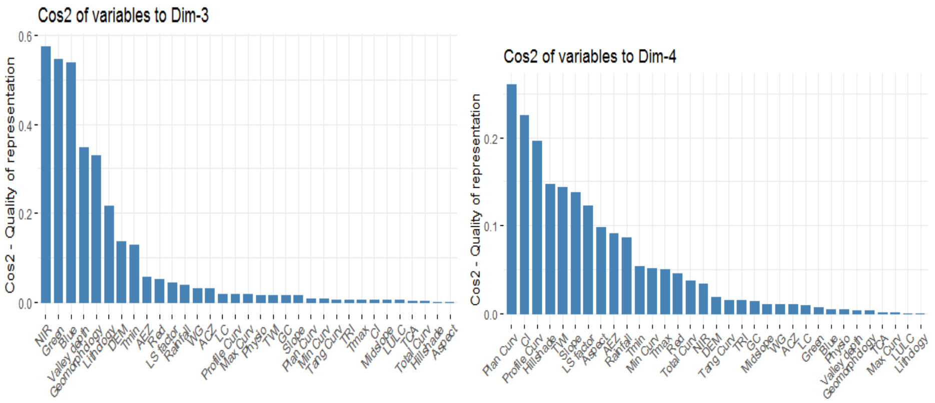


Fig.5 PCA plot of variable

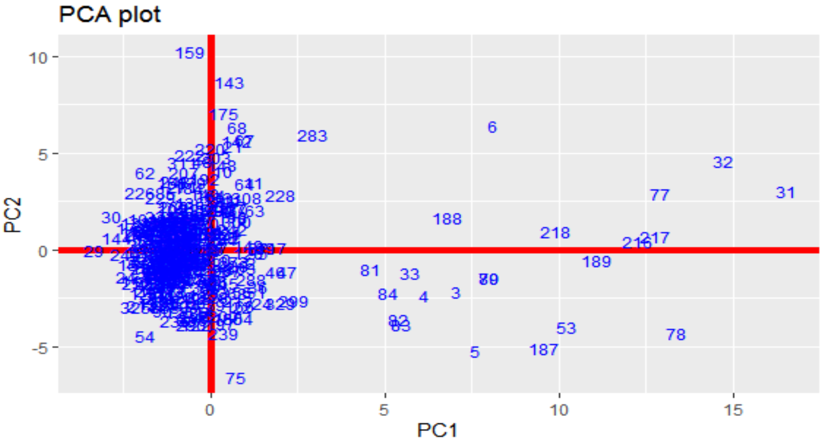


Fig.6 Correlation

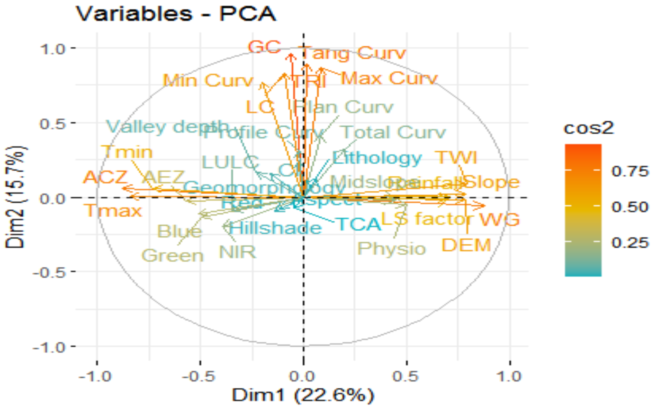
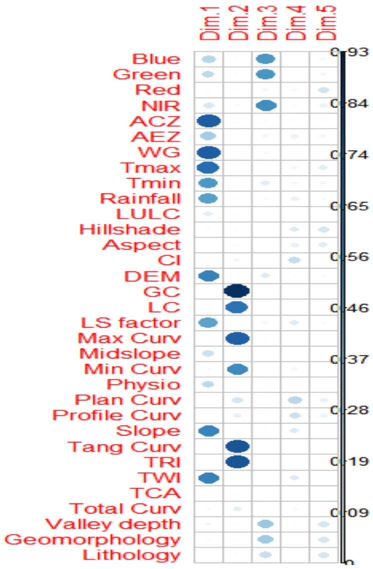


Fig.7 Contribution of each variable



From figure 6 the cos2 values are used to estimate the quality of the representation. A high cos2 indicates a good representation of the variable on the principal component. In this case the variable such as WG, ACZ, Tmax, TWI, DEM and Slope are positioned close to the circumference of the correlation circle. A low cos2 indicates that the variable is not perfectly represented by the PCs. The variables such as Lithology, Midslope, Geomorphology, LUCU, CL and Aspect are close to the center of the circle.

Table 3 shows the results of eigen vectors for each components. The first four components were selected to identify the most influencing variable. The variable which has higher value contributes more for the variation.

From the study, it was concluded that PC1, PC2, PC3 and PC4 contributed much of the variations; hence these four components were taken for further analysis. The variables that contribute more variations are Blue, Green, Red, NIR, Agro Climatic Zones (ACZ), Agro Ecological Zones (AEZ), Western Ghats (WG), Maximum Temperature, Minimum Temperature, Rainfall, Hillshade, Elevation, General curvature, Longitudinal Curvature, LS factor, Maximum Curvature, Minimum Curvature, Plan Curvature, Profile Curvature, Physiography, Slope, Tangential Curvature, TRI and TWI. These variables further used to produce and validate the accuracy of soil map of Coimbatore district.

References

1. Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4), 433-459.
2. Bhuyan, K. C. (2005). Multivariate Analysis & Its Applications. New Central Book Agency.
3. Bro, R., & Smilde, A. K. (2014). Principal component analysis. *Analytical Methods*, 6(9), 2812-2831.
4. Demšar, U., Harris, P., Brunsdon, C., Fotheringham, A. S., & McLoone, S. (2013). Principal component analysis on spatial data: an overview. *Annals of the Association of American Geographers*, 103(1), 106-128.
5. Fox, G. A., & Metla, R. (2005). Soil property analysis using principal components analysis, soil line, and regression models. Soil Science Society of America Journal, 69(6), 1782-1788.
6. Ivosev, G., Burton, L., & Bonner, R. (2008). Dimensionality reduction and visualization in principal component analysis. *Analytical chemistry*, 80(13), 4933-4944.
7. Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065), 20150202.
8. Karamizadeh, S., Abdullah, S. M., Manaf, A. A., Zamani, M., & Hooman, A. (2013). An overview of principal component analysis. *Journal of Signal and Information Processing*, 4(3B), 173.
9. Kambhatla, N., & Leen, T. K. (1997). Dimension reduction by local principal component analysis. *Neural computation*, 9(7), 1493-1516.
10. Kooch, Y., Jalilvand, H., Bahmanyar, M. A., & Pormajidian, M. R. (2008). The use of principal component analysis in studying physical, chemical and biological soil properties in southern Caspian forests (north of Iran). *Pakistan Journal of Biological Sciences*, 11(3), 366-372.
11. Liu, Y., Singleton, A., & Arribas-Bel, D. (2019). A Principal Component Analysis (PCA)-based framework for automated variable selection in geodemographic classification. *Geo-spatial Information Science*, 22(4), 251-264.
12. McBratney, A. B., Santos, M. M., & Minasny, B. (2003). On digital soil mapping. *Geoderma*, *117*(1-2), 3-52.
13. Paul, L. C., Suman, A. A., & Sultan, N. (2013). Methodological analysis of principal component analysis (PCA) method. *International Journal of Computational Engineering & Management*, 16(2), 32-38.
14. Samuel-Rosa, A., Heuvelink, G. B. M., Vasques, G. M., & Anjos, L. H. C. (2015). Do more detailed environmental covariates deliver more accurate soil maps?. Geoderma, 243, 214-227.
15. Skrbic, B., & Durisic-Mladenovic, N. (2007). Principal component analysis for soil contamination with organochlorine compounds. Chemosphere, 68(11), 2144-2152.

How to cite this article:

Priyadharshini, R., M. Radha’ R. Kumaraperumal, Tmt.G.Vanitha and Balaji Kannan. 2021. Characterization of Environmental Covariates of Coimbatore District using Principal Component Analysis. *Int.J.Curr.Microbiol.App.Sci.* 10(01): 3114-3123.
doi: https://doi.org/10.20546/ijcmas.2021.1001.362