Analysis, Resolution Differentiation of Principal Component Analysis and Correspondent Analysis

Edmund Baffoe-Twum

[1]Msc, Ph.D. Student, North Dakota State University, Department of Civil and Environmental Engineering, Dept. 2470, P.O. Box 6050, Fargo, ND, 58108-6050, USA

Abstract: ***In recent times precision and effective use of techniques at solving problems have become very important. Consequently, the idea of using real data to determine and resolve the resolution between two most use statistical technique at addressing similar situations. This paper looks at the princicpal component analysis and correspondences analysis at solving the same question. This is a test to identify the highest of resolutions of the two techniques. With the data considered, percentages ofthe correspondence analysis showed it isof much higher in value with similar factors, indicating a higher resolution compared to principal component analysis.It is therefore probably the best in explaining the relationship between/among variables in large and as well as multivariate dataset.***

Keywords: Princicpal Component Analysis, Correspondences Analysis, Precision

1. Introduction

This paper seeks to analysis, interpret and compare two statistical methods on data set acquired from the data base of the Ocean Drilling Project, Leg 138, on Hole 844B(http://brg.ldeo.columbia.edu/data/odp/leg138/844B/).

The methods are;
1) Principal Component Analysis.
2) Correspondence Analysis.

These two statistical methods are a part of a broad statistical method known as Principal Components. Other methods under the Principal Components are Standardized Principal Component and Factor analysis (Carr, 2002). Principal component analysis uses the sample variance and covariance whereas Correspondence analysis uses the chi-square distance between each data entry and its expected value as means of generating the end product. The both rely on the eigen disintegration of some data likeness matrix with the development of two dimensionalplots as the eventual objective (Carr, 2002). These statistical approaches are used to give an account of multi-variable data (Lynn and McCulloch, 2000) or determining similarity among data (Carr, 2002). Both are used to determine the relationship (correlation) and variations within data set.

The main objective is to determine the relationship and variability of the data set. The second objective is to compare these methods visually (plots) and statistically to arrive at the better of two. This is to help make informed decision based on the basic principles of these methods and the conditions under which each could be used appropriately. There may be though, unsure a possibility of adapting the best of these methods for the analysis of the relationship of data set of major elemental composition to be acquired from core logs for my intended dissertation work.

Project area (ODP Leg 138)

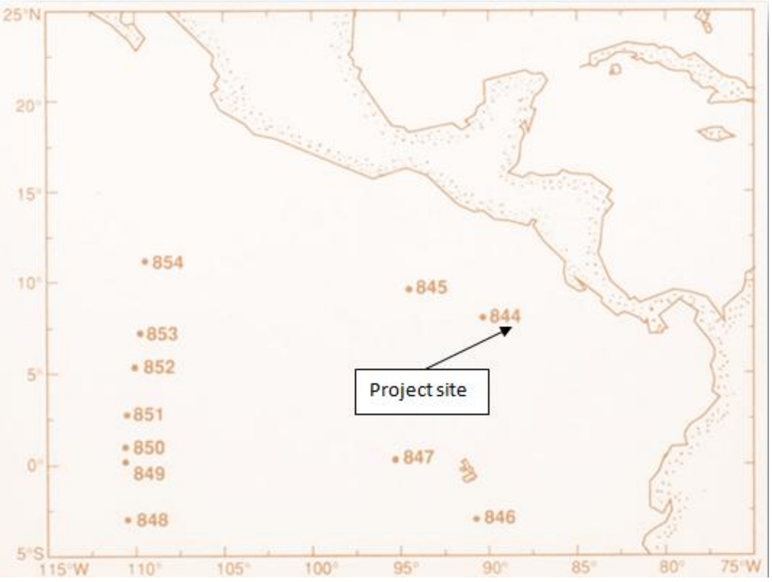


Figure 1: Picture of area data was collected (adopted from ODP map data base)

Data for this paper is acquired from the geochemical data base of the ODP Leg 138 hole number 844B data. Hole is located at 7^o 55.279’N and 90^o 28.846’W. The total depth of hole 844B is 290.8 meters. The data constitute the major oxide composition of the cored rock samples from hole number 844B. The data consist of 5 variables (oxides) with 1462 observations for each variable through the entire depth of the hole. The variables are SIO_2, $CACO_3$ (CAO), FEO, TIO_2, $K\ _2O$ and AL_2O_3.TIO_2 had null marginal sum given that all measurement made were zero for the 1462 observations,consequently eliminated from the analysis.

2. Methodology

Principal Component Analysis

Principal component analysis is a technique used in establishing observable characteristics in data. This helps in depicting observable distinction and comparability. The information from the data is displayed graphically for easy interpretation (Carr, 2002). This is usually good when data to be analyzed are huge and by visual inspection of data no inferences of correlation can be made (Swan *et al*, 1995). The advantage of principal component analysis over others is its ability to compress the data into a much smaller dimensions and not losing part of the resolution.

Principal component analysis results are attained by the use several statistical techniques. Some of these are summary statistics (mean, standard deviation etc) and weights (also know as loadings). The weights are deduced from the correlation or covariance matrix of the data (Swan *et al*, 1995), while correlation matrix is much suited for data with variable units (PH, %, PPM etc) of measurement, covariance matrix is best for a single unit measurement. Thus weights derived from covariance and correlation matrix are different. The weights are also used in generating the eigenvalues and vectors. The eigenvalues are then used to obtain percentage variability (factors).

The following are the procedures used in generating the end product of principal component analysis.

A) Determining the summary statistics (mean and Standard deviation).
B) Generate a square matrix of the variables measured using covariance and variances generated from the data. Therefore the size of the matrix is dependent on the number of variables measured. With the covariance matrix, the variances form the diagonal and the covariance forms the other members of the matrix. The correlation Matrix approach is obtained using the correlation coefficient of data and replacing all the diagonal variances with the number 1. This square matrix in principal component analysis is generated from the original data (Carr, 2002).

$$[S\ ]=\text{MxM matrix}=\begin{pmatrix} \mathrm{var}_{11} & \mathrm{cov}_{12} & \mathrm{cov}_{13} \\ \mathrm{cov}_{21} & \mathrm{var}_{22} & \mathrm{cov}_{23} \\ \mathrm{cov}_{31} & \mathrm{cov}_{32} & \mathrm{var}_{33} \end{pmatrix} \text{-1}^{\text{st}} \text{ approach}$$

$$[S\ ] = \text{MxM matrix}=\begin{pmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{pmatrix} - \ 2^{\text{nd}} \text{ approach}$$

where r is the correlation coefficient. r ranges between -1 and 1.

C) After the matrix is obtained as above, it is used to generate the eigenvectors and eigenvalues.

$$\left[[S\ ] - \lambda \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \right] \{X\ \} = \{0\}$$

λ is the determinant of the matrix (eigenvalue). The decomposition of this matrix generates the eigenvalues and vectors.

D) Finally the percentage variability (% factor) for the individual eigenvalue is obtained using the ratio of the eigenvalue to total sum of all eigenvalues multiplied by 100.

$$\frac{\lambda_i}{\lambda_T} * 100 = \frac{\lambda_i}{\sum_{i=1}^{n} \lambda_i} * 100$$

$\frac{\lambda_i}{\sum_{i=1}^{n} \lambda_i}$ is defined as the factor/ variability loading.

E) The variability (factor) defines the coordinates for the plot of the data. The higher the eigenvalue the higher the percentage variability (factor). The percentage variability sum up to 100% (Carr, 2002).

3. Correspondence Analysis

Correspondence analysis is a technique for demonstrating the correspondence between rows and columns of multivariate data matrix as points in twofold low-dimensional vector spaces. The matrix is principally a dual-way incident table (Greenacre, 1984). The name correspondence analysis was coined from the fact that, the geometry of the row profiles to and the columns relates directly in some many ways. Thus, in most cases the data matrix is evaluated along the profiles of the columns and rows for a clearer picture of the relationship (Greenacre, 1984). The columns are the variables and the rows are the samples. The similarities in the data matrix are based on the chi-square distance between each expected values and its data entry (Carr, 2002).

$$X\ \hat{}\ 2 = \frac{(\mathrm{O} - \ \mathrm{E})\ \hat{}\ 2}{\mathrm{E}} = \frac{(Observeddifference - \ Expected)\ \hat{}\ 2}{Expected}$$

Carr 2002, suggest the following steps for establishing the chi –square values in correspondence analysis.

Step 1: The data set are first grouped into a matrix *NXM* represented by [Y]

$$[Y] = [NCM]$$

Step 2: All the data values of Y in $[Y]$ are sum up to generate the total sum

Step 3: Each of the individual values in the matrix $[Y]$ is divided by the total sum to obtain a new matrix $[Y']$

$$[Y] / TotalSum \; [Y'] \Longrightarrow$$

The $[Y']$ looks more of a probability function.

Step 4: After the generation of $[Y']$, two new vectors are created. These vectors are represented by {W} for (Nx1) for each row in $[Y']$ and {T} for (1xM) for each column in $[Y']$. The steps 2 to 4 are replicated iteratively till no modifications can be made to the vectors {W} and {T}.

Step 5: With the above steps as the foundation, a matrix NxM $[S]$ which is eigen decomposable is formed. Each Si is an approximate measure of chi-square

$$X \wedge 2 = \frac{(O - E) \wedge 2}{E} = \frac{(Observeddifference - Expected) \wedge 2}{Expected}$$

Where $O = y_{ij}$ **and** $E = W_i T_j$

Software Used

The software used for this analysis is a 30 day free trial version of a complete excels data analysis program called XLSTAT. The XLSTAT is a product of Addinsoft, privately-owned company managed by Thierry Fahmy (PhD).

4. Results

Principal Component Analysis

Table PCA:1 Summary statistics

Variable	Observations	Obs. With missing data	Obs. Without missing data	Minimum	Maximum	Mean	Std. deviation
SIO2	1462	0	1462	-0.002	92.944	18.605	19.053
CACO3-CAO	1462	0	1462	0.000	95.929	65.255	26.228
FEO*	1462	0	1462	0.000	55.636	5.320	8.102
K2O	1462	0	1462	0.000	2.813	0.364	0.379
AL2O3	1462	0	1462	1.035	32.881	5.394	5.211

Table PCA:2 Correlation matrix (Pearson (n-1))

Variables	SIO2	CACO3-CAO	FEO*	K2O	AL2O3
SIO2	1	-0.709	-0.067	0.015	0.048
CACO3-CAO	-0.709	1	-0.623	-0.539	-0.649
FEO*	-0.067	-0.623	1	0.571	0.676
K2O	0.015	-0.539	0.571	1	0.868
AL2O3	0.048	-0.649	0.676	0.868	1

Table PCA: 3 Eigenvalues and their cumulative variability

	F1	F2	F3	F4	F5
Eigenvalue	3.030	1.361	0.490	0.118	0.000
Variability (%)	60.604	27.229	9.797	2.369	0.000
Cumulative %	60.604	87.834	97.631	100.000	100.000

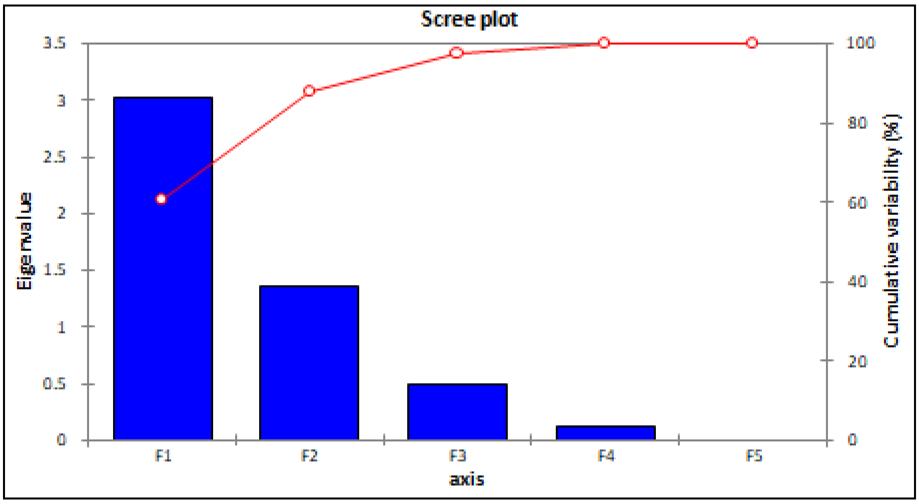


Figure PCA 1: Scree Plot showing a histogram of eigenvalues and their percentage cumulative variability

Table PCA: 4 Eigenvectors of the variables

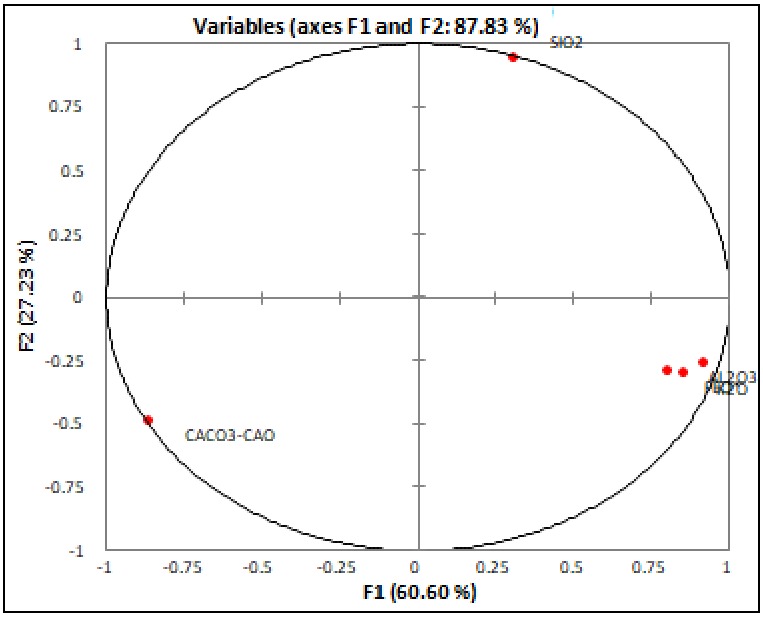
	F1	F2	F3	F4	F5
SIO2	0.175	0.813	-0.116	-0.086	0.536
CACO3-CAO	-0.498	-0.412	-0.185	-0.051	0.738
FEO*	0.458	-0.243	0.753	-0.205	0.349
K2O	0.486	-0.253	-0.551	-0.629	0.005
AL2O3	0.524	-0.215	-0.285	0.743	0.213

Table PCA:5 Factor loadings

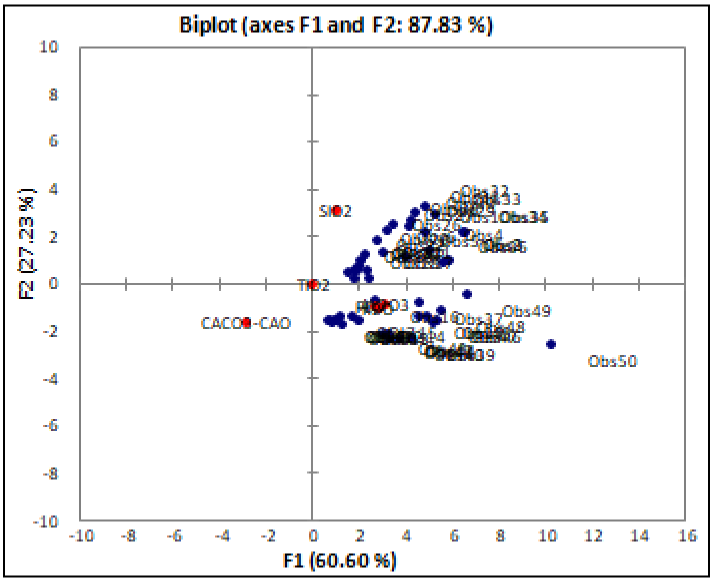
	F1	F2	F3	F4	F5
SIO2	0.305	0.949	-0.081	-0.030	0.000
CACO3-CAO	-0.867	-0.480	-0.130	-0.017	0.000
FEO*	0.798	-0.284	0.527	-0.070	0.000
K2O	0.847	-0.295	-0.386	-0.216	0.000
AL2O3	0.912	-0.251	-0.199	0.256	0.000

Table PCA: 6 Correlations between variables and factors:

	F1	F2	F3	F4	F5
SIO2	0.305	0.949	-0.081	-0.030	0.000
CACO3-CAO	-0.867	-0.480	-0.130	-0.017	0.000
FEO*	0.798	-0.284	0.527	-0.070	0.000
K2O	0.847	-0.295	-0.386	-0.216	0.000
AL2O3	0.912	-0.251	-0.199	0.256	0.000



FigurePCA2: Principal component analysis of the variables explained on by the first two factors



FigurePCA3: Biplot – simultaneous graphic view of observations and variables

Table PCA: 7 Contribution of the variables (%)

	F1	F2	F3	F4	F5
SIO2	3.061	66.091	1.34	0.739	28.769
CACO3-CAO	24.817	16.955	3.438	0.256	54.533
FEO*	21.011	5.911	56.73	4.187	12.161
K2O	23.663	6.413	30.376	39.544	0.003
AL2O3	27.448	4.63	8.116	55.273	4.533

Table PCA:8 Squared cosines of the variables:

	F1	F2	F3	F4	F5
SIO2	0.093	0.9	0.007	0.001	0
CACO3-CAO	0.752	0.231	0.017	0	0
FEO*	0.637	0.08	0.278	0.005	0
K2O	0.717	0.087	0.149	0.047	0
AL2O3	0.832	0.063	0.04	0.065	0

Correspondence Analysis

Table CA: 1 Eigenvalues, Inertia and cumulative percentage of dimensions

	F1	F2	F3
Eigenvalue	0.350	0.049	0.002
Inertia (%)	87.434	12.172	0.394
Cumulative %	87.434	99.606	100.000

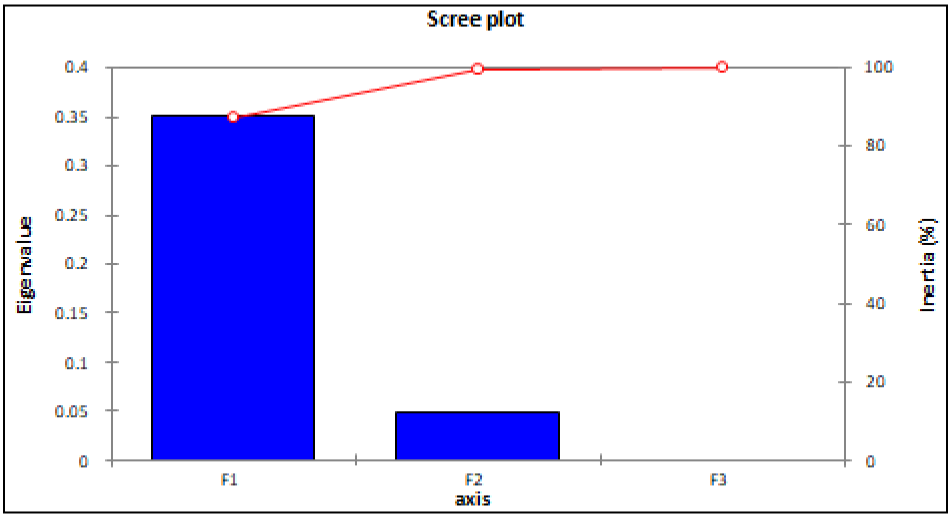


Figure CA 1: Scree Plot showing histogram of eigenvalues and their percentage cumulative variability

Table CA:2 Weights, distances and squared distances to the origin, inertias and relative inertias (columns):

	Weight (relative)	Distance	Sq-Distance	Inertia	Relative inertia
CACO3-CAO	0.855	0.240	0.058	0.049	0.123
FEO*	0.070	1.782	3.176	0.221	0.553
K2O	0.005	1.368	1.872	0.009	0.022
AL2O3	0.071	1.308	1.711	0.121	0.302

Table CA: 3 Chi squared distances (Columns)

	CACO3-CAO	FEO*	K2O	AL2O3
CACO3-CAO	0	2.009	1.568	1.528
FEO*	2.009	0	1.444	1.276
K2O	1.568	1.444	0	0.600
AL2O3	1.528	1.276	0.600	0

Table CA: 4 Principal coordinates (Columns)

	F1	F2	F3
CACO3-CAO	-0.239	0.017	0.000
FEO*	1.713	0.492	0.002
K2O	1.056	-0.668	0.557
AL2O3	1.137	-0.646	-0.038

Table CA: 5 Standard coordinates (Columns)

	F1	F2	F3
CACO3-CAO	-0.405	0.077	-0.003
FEO*	2.895	2.229	0.041
K2O	1.785	-3.024	14.023
AL2O3	1.921	-2.926	-0.947

Table CA: 6 Contributions (Columns)

	Weight (relative)	F1	F2	F3
CACO3-CAO	0.855	0.140	0.005	0.000
FEO*	0.070	0.584	0.346	0.000
K2O	0.005	0.015	0.044	0.937
AL2O3	0.071	0.261	0.605	0.063

Table CA: 7 Square cosines (Columns)

	F1	F2	F3
CACO3-CAO	0.995	0.005	0.000
FEO*	0.924	0.076	0.000
K2O	0.596	0.238	0.166
AL2O3	0.755	0.244	0.001

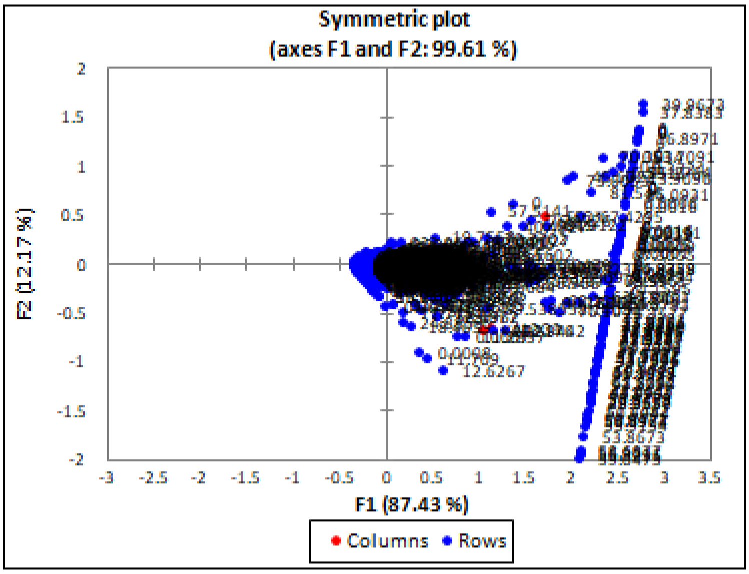


Figure CA 2: Symmetric plot of analysis of association between variables and observations (rows and columns)

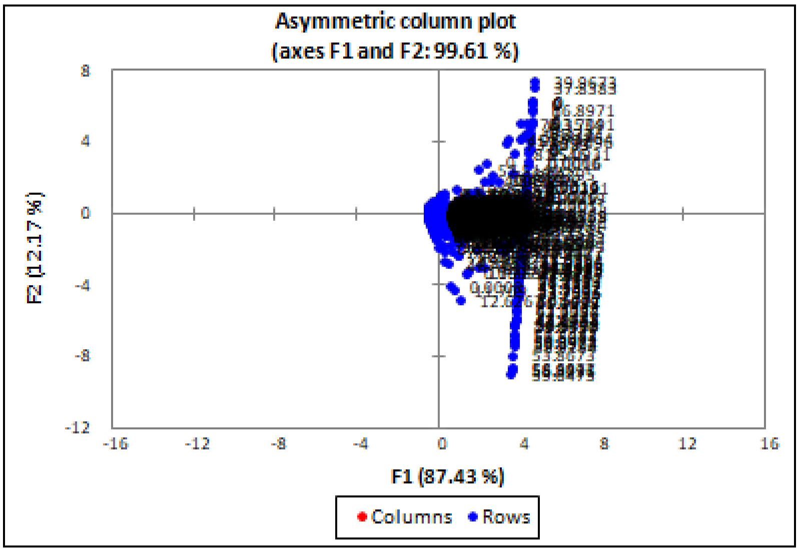


Figure CA 3: Asymmetric column plot of analysis of association between variables and observations

5. Interpretations and Discussions

Principal Component Analysis

In the introduction, principal component analysis is discussed as a method to reducing data sets and as well as helping in the application of multivariate statistical methods like analysis of variance or regression analysis. The data used have been reduced and presented in the tables and plots under the sub heading “results”. Table PCA 1 expresses the summary statistics of the data employed in the analysis. The table shows the means, standard deviation, minimum and maximum observation values of the variables under consideration. The minimum value measured is -0.002 corresponding to SIO_2 and the maximum of 95.929 corresponds to $CACO_3$-CAO. The least deviation among the observations of the measured parameters corresponded to K_2O at 0.379 though is in conformity with that mean of its data set.Table PCA 2 outlines the correlation of the variances between and among variables in the matrix generated. The number of variables equals the total variance in the matrix. The significant correlations are exhibited in bold face which in other words expresses self correlation of the measured parameters. The consideration of the correlation is based on the relative reflectance of independence of the factors (high and low correlations between and among variables). A strong correlation isexperiential between AL_2O_3& FEO, and AL_2O_3& K_2O, indication a close similarity between/among the variables of the data set while an anti correlation is observed between $CACO_3$-CAO and SIO_2as wells as with FEO,AL_2O_3, K_2O.

Eigenvalues of the data set acquire with the use of factors in variance extraction (principal component analysis) are elaborated in Table PCA 3. The eigenvalues are scalar values obtained in the reduction and extraction processes and the match up vector component are revealed in Table PCA 4. The table displays the eigenvalues to correspond to factors (F) - (principal components extracted) and percentages. A factor with eigenvalue greater than 1.00 indicatessignificance (Kaiser 1960) and therefore explains an important amount of variability in the data whereas a factor of the eigenvalue which is less than 1.00 is considered to be too weak (Kaiser 1960) and thus not able to explain a significant portion of the data variability. Consequently the factors F1 (3.030) and F2 (1.361) with eigenvalues greater than 1.00 are considered for explaining the significance of the data variability.Thus these two factors are retained in considering other processes for the final decision.

According to Raykov and Marcoulides, 2008, the scree plot(proposed by Cartel 1966) characterizes the consecutive sequence of significance of eigenvalues pictographically and the term scree is used in geologicperceive for a steep mountain slope with debrisor fragments at itsbase. In this manner, the debris are detected on the plot with an ‘’elbow’’ shaped plot of the cumulative eigenvalue percentage contributions which show a slightly tilted flat pattern at emergence. The plot helpssettle on the number of factors to use in an analysis and accordingly emphasizes the choice of F1 and F2 in Table PCA 3.

Table PCA 5,presents summary of factor loadings of the data set which are the results of the factor score using the theory of factor analysis in examining the variability among observed variables. The correlation between/among variables and factors as displayed are similar to the factor loading in this data analysis, thus the similar results. Factor 1 generally appears to show a strong correlation with the variablesfollowed by factor 2.The contributions from both are enough for the classification of the variables and best suites the generation of Figure PCA 2 rather than the combination of others.Hence the factor loading of the six variables are reduced to the specific factors displayed on the plot. In Figure PCA 2, AL_2O_3, FEO and K_2O are closer in the domain and therefore exhibits and translates their correlation as well as significance of the components while $CACO_3$-CAO and $SIO_2$are separated from the variables with partial significance. The contributions in percentage are again expressed in Table PCA 7. The square cosines help emphasize variable contributions. Low square distances are sometimes not interpreted though the values also tell the contribution within each factor. Overall, as shown in the graphic view of the biplot are the variables and observations. It is significant and enough to validates the interpretations the first two factors by also showing higher percentages on the plot axis.

Correspondence Analysis

The correspondence analysis is an evocative technique intended to examine a two way or multi-way tables which has some quantity of correspondence between columns and rows. Though a result from this technique has similarity with factor analysis, the correspondence analysis makes it possible to investigate the structure of definite variables in the table(Greenacre, 1984).The process establishes cross tabulation of frequencies such that the sum of all relative frequencies (mass) equals 1.00. The tables also represent the distances between the individual columns and rows in a low dimensional space. The computation of the relative frequencies for 1462 observation of the six variables gives a large data set thus the mean relatives frequencies are shown in the table below;

Variables	$CACO_3$-CAO	FEO	K_2O	AL_2O_3	Total
Mean	0.816	0.088	0.006	0.089	1.00

The above is achieved with each element divided by the total. The relative frequencies are named column or row mass depending on the emphasis. Table CA1 illustrates the eigenvalues, inertia, cumulative percentages andtheir corresponding dimensions (factors).The two way table generates a maximum number of eigenvalues which are equivalent to minimum number of columns minus 1 and rows minus 1. The dimensions are similar to the extracted principal components. The inertia is equivalent to “moment of inertia” which is computed as the squared distance to the centroid times the integral of the mass (Greenacre, 1984). Therefore the inertia in the table is defined as the total Pearson Chi-Square of a two-way table divided up by the overall sum (inertia = chi-Square/Total N). The first column of the table shows a single dimension (F1) of 87.43%, implying 87.43% of the inertia is explained which in other words the relative frequency of the values from a single dimension of the chi-square value that can be recreated. The second column explains 12.17 % of the inertia, thus both F1 and F1 explains 99.60% of the inertia. They are therefore used in the analysis of the variables. The selection is confirmed with the scree plot in Figure CA 1 where the “elbow” flattens (smoothens) up. The weights assigned, distances and the distances squares generated the processes of the Correspondence analysis are as shown in table CA 2 whereas the chi-square distances, principal and standard coordinates are expressed in tables CA3, CA4, and CA5 respectively. The weighteddistances are weightsput on variable observations for the computation of the factors. Figures CA 2 and CA 3 express the symmetric plot and asymmetric plot of the variables with the observations with respect to distances of the centriod respectively. In all cases the F1 and F2 are used for the analysis.

Comparison of methods

Both methods are compared using their percentage variability and percentage inertia given that they are generated using the eigenvalues obtained from the matrix decomposition.

Table CM: 1

Method / Factor	F1 (%)	F2 (%)	F3 (%)	F4 (%)	Total (%)
Principal Component Analysis	60.60	27.23	9.80	2.37	100.00
Correspondence Analysis	87.43	12.17	0.39	0.00	100.00

Table CM: 2

Factor Contribution	Principal Component Analysis	Correspondence Analysis
F1 (%)	60.60	87.43
F2 (%)	27.23	12.17
Total %	87.83	99.60

The correlation matrix (Table PCA: 2) of the principal component analysis indicates a strong a correlation between AL_2O_3 and K_2O whereas a good but less strong relationship between AL_2O_3 and FeO. This association is again depicts in the plot of the column variables (figure PCA2). The AL_2O_3, FeO and K_2O are illustrated as similar to each other whereas the SiO_2 and $CaCO_3$-CAO are far apart. The strong association of AL2O3, FeO and K_2O are also represented in the table of factors and variables (Table PCA: 6). The factor loading relating to these variables are dependent on the eigenvalue/ vectors (Table PCA: 4 and 5).

In the correspondence analysis, a similar association is noted for the principal coordinates (column) and the standard coordinates (columns) in tables CA 4 and 5 respectively. The symmetric (columns) and asymmetric (rows) again portray the relationship. These are as shown in figures: CA 2 and 3.

In summary there are similarities and differences amongst the variables and the samples.

6. Conclusions

The factor loadings percentages from both techniques are summarized in tables CM 1 and 2 above. Considering factors 1 and 2 (F1 and F2) under each of the techniques, Principal component analysis had F1 and F2 contribute 60.60 and 27.23 % respectively, while the correspondence analysis F1 had percentage 87.43 and F2 at 12.17. The total for both F1 and F2 for the principal component and correspondence analysis are 87.83% and 99.61% respectively. With these percentages the correspondence analysis showed a much higher percentage value with similar factors, indicating a higher resolution compared to principal component analysis and probably the best in explainingthe relationship between/among variables in large and as well as multivariate dataset.

Conflict of Interest: There is no conflict of interest.

References

[1] Albarède, F, (1995), “Introduction to Geochemical Modeling”, pp 237-243.
[2] Carr, J. R., (2002), “Data Visualization in the Geosciences”, pp 65-77.
[3] Greenacre, M.J., (1984), “Theory and Applications of Correspondence Analysis”, pp, 34-35, 55-67
[4] Lynn, H.S., and McCulloch, C.E. (2000), “Using Principal Component Analysis and Correspondence Analysis for Estimation in Latent Variable Models”, Journal of the American Statistical Association”, Vol.95, No. 450, pp 561-572.
[5] Ocean Drilling Project website, (Leg 138),http://brg.ldeo.columbia.edu/data/odp/leg138/844B/
[6] Raykov, T. and Marcoulides, G.A., (2008), “An Introduction to Applied Multivariate Analysis”, pp 211-263.
[7] Swan, A.R.H, Sandilands, M., and McCabe P. (1995), “Introduction to Geological data analysis”, pp 360-370.

Author Profile

Edmund Baffoe-Twum received aB.Sc (Hons). in Geology from the University of Ghana, Legon in1998, an M.Sc.in Hydrogeology from the University of Nevada, Ren in 2007 and currently PhD Student at North Dakota State University with the Civil and Environmental Engineering with concentration in Construction Management and Engineering .He has since April 2011 been a member of the faculty of Engineering and Technology,Civil Engineering Department at Kumasi Technical University, Kumasi-Ghana.