Original Research Article

Spatial Characterization of Cone Index and some Nutrients in a Sandy Loam Soil (*EutricLeptosol*) Using the Principal Component Analysis

David Lomeling[1], Sebit MathewOtwari[2] and Yahya Mohammed Khater[2]

[1]Department of Agricultural Sciences, College of Natural Resources and Environmental Studies (CNRES) University of Juba. P.O. Box 82 Juba, South Sudan

[2]Graduate Students, Department of Agricultural Sciences, College of Natural Resources and Environmental Studies (CNRES), University of Juba, P.O. Box 82 Juba, South Sudan

Authors´ Contribution

Author DL designed the study, performed the model simulations and statistical analysis and wrote the first draft of the manuscript. Authors SMO and YMK conducted the field and laboratory tests. All authors read and approved the final manuscript.

ABSTRACT

A Principal Component Analysis (PCA) was performed on multi-nutrient soil and Cone Index (CI) data from the research and demonstration farm of the Dept. of Agricultural Sciences, University of Juba in South Sudan. The main objective of the study was to characterize the spatial distribution of the soil nutrients: N, P, K, Fe and Mn as well as soil penetration resistance CI. A Gaussian Mixture Model (GMM) and Hierarchical Cluster Analysis (HCA) were performed on the analyzed samples. Eigenvalues of the three major principal components F1, F2 and F3 accounted for 75.67% of the total variance of the data. From hierarchical clustering, P was observed to cluster with Fe, Mn with N which at second level clustered with K then with CI. The results of the PCA showed that Nitrate-N, Mn and Kwere strongly influenced by CI and so determining their spatial distribution. This could be associated mainly to earlier anthropogenic activities on the soil. The results of this study also showed spatial relationships between individual soil nutrients with both K and P mutually antagonistic with Nitrate-N, whereas between K and P where mutually synergistic. While P was strongly adsorbed to Fe, this was associated to lithogenic soil materials and therefore interpreted as derived from natural sources of the *EutricLeptosol.*The goodness-of-fit test using the Kolmogorov-Smirnov (KS) showed that the values of the variables: CI, K, P and Fe were significant at $p \leq 0.05$ and that the data followed normal distribution, whereas Mn and Nitrate-N were not. The KS test also corroborated the results of strong spatial dependency of each variable at less than 25%.

Keywords*: dendrogram,Gaussian Mixture Model,hierarchical clustering, Kolmogorov-Smirnov test, Principal Component Analysis, spatial distribution*

1. INTRODUCTION

Principal component analysis (PCA) is a multivariate analysis method~~which is~~ also known as eigenvector analysis. It has been applied in environmental studies in; assessing heavy metals in urban soils [1],estimating heavy metals in agricultural soils [2],assessing earthworm populations in oil plantation soils [3],on genetic diversity of tomato germplasm[4],on effects of soil parameters and environmental factors on flavonoid contents of medicinal plants [5],soil factors influencing heavy metals in medicinal plants [6],and in combination with geo-statistical tools in assessing fertility of humicrhodichapludox[7], on distribution of trace elements in unsaturated soil profile [8].

Similarly, PCA has been performed to investigate management impacts on soil quality [9], on chemical and microbial properties in histosols as influenced by land-use types [10] and microbial communitystructure and function [11, 12].

PCA is a technique that reduces the number of variables and eliminates the relations among input variables by developing a set of new variables that are linear functions of the original variables. This new variables are denoted as F1, F2, F3 etc.Whereas F1 will give the direction with the greatest variation, F2 orthogonal to F1 will give the direction with the maximum variation left in data. The variables are multiplied by loadings that are vectors of constantsgenerated during PCA and whose values reflect the importance of original variables in the direction of each PC[13]. The resulting PC can be used to project the originally multidimensional data into only a two- or three dimensionalspaceknown as a score plot [14].

Hierarchical Cluster Analysis (HCA) is another method applied for geological/hydrological analysis and looks for groups of samples according to their similarities. HCA is a powerful tool for analyzing data sets for expected or unexpected clusters including the presence of outliers. In HCA, each point forms, initially, one cluster, and the preliminary matrix is analyzed. The most similar points are grouped forming one cluster and the process is repeated until all points belong to one cluster [15]. HCA examines distances between samples and datasets. The result obtained could be presented in a two-dimensional plot called dendrogram which illustrates the fusions or divisions made at each successive stage of analysis.Typically, a hierarchical agglomerative cluster analysis of the studied variables is performed and the square of the Euclidean distance taken as a measure of similarity between samples [16]. Samples that are similar will lie close to one another, whereas dissimilar ones will lie distant from each other. Several methods may be applied in decidingsimilarity or dissimilarity between two groups: (i) single linkage, which uses the minimum distance between points in different groups; (ii) complete linkage, which uses the maximum distance between the furthest points; (iii) mean linkage, which uses the average of all distances between points in the two groups and (iv) centroid linkage, which uses the distances between group centroids (*e.g.* group means).

The PCA and HCA are both methods for uncovering relationships in large multivariate datasets. However, they are not sufficient for developing a classificationrule that can accurately predict the class-membership of unknown

sample [17]. Classification or pattern recognition is a method that seeks to develop class-membership based on some specific and measured property within the variables. For cluster classification, the similarity-based classifier e.g. *k-nearest neighbor* (KNN) may be chosen under the premise, that distances between points in the measurement spacewill be inversely related to their degree of similarity.

Another tool for multivariate analysis is the Gaussian Mixture Model (GMM). This is a parametric probability density function (pdf) and is often applied as a model for the probability distribution of continuous measurements ofvariables in a dataset. As opposed to histogram analysis, which is non-parametric, GMM provides greater flexibility and precision in modeling the underlying statistics and smoothening gaps resulting from sparse sample data by assigning variable to specific membership. Based on the premise of a vector-based variable x, a dataset is assumed to have a univariate normal (*Gaussian*) distribution if its probability density function is given by:

$$p(x,\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) \text{(Eqn. 1)}$$

It is also said to have a multivariate normal (Gaussian) distribution if its probability density function is given by:

$$p(x,\mu,\Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right) \quad \text{(Eqn. 2)}$$

Where μ is the mean (location)and Σ the covariance matrix of the Gaussian that incorporates the variance σ; and therefore dispersion. Under the GMM, the random variable x is generated from several distinct random and independent processes each generating a specific pdf. A GMM is often a mixture of two or more independent or un-imodalpdfs combined into a single bi- or multimodal pdfs. The concept of clustering using the GMM is based upon the fact that individual data points are produced by choosing from one of a set of multivariate Gaussians and the GMM parameters estimated using the iterative Expectation-Maximization (EM) algorithm or Maximum A-Posteriori (MAP).Although the EM algorithm has some limitations (e.g. it is not guaranteed to converge to a global rather than a local maximum of the likelihood), it is generally efficient and effective for theparameters' estimation of GMM [18].

The Gaussian model is then fitted to the data points that is either spherical (circular with off-diagonal correlations equal to zero) symmetry or elliptical (off-diagonal correlation equal to non-zero). To fit a Gaussian model to given data points, the sample mean and variance is computed and the resulting Gaussian model is superimposed as a contour map.

GMM, a parametric probability density function (pdf) was employed and provided the basis for partitioning and fuzzy classification of the data into different clusters. Arguably, this probabilistic method partitions the databy considering that each component represents a cluster through the *k-means* clustering that tends to minimize the *Euclidian distance* or the within-cluster sum of squares (WCSS). A

measure of dissimilarity or similarity between observed data sets was used to combine specific clusters (agglomerate) or split (divide) them altogether. This hierarchical clustering is achieved by use of measuring the *Euclidian distance* between pairs of observations and a linkage criterion which specifies the dissimilarity of sets as a function of the pairwise distances of observations in the sets.We used complete linkage as the agglomeration method as suggested by [19]for clustering the same dataset.

Formatted: Font: (Default) Arial

The primary goal of any soil scientist is the ability to adequately characterize and describe major soil factors, parameters and processes as well as their spatio-temporal distribution within the context of precise farming. Increasing environmental awareness in trying to reduce the application of especially costly inorganic fertilizers and use of heavy farm implements by farmers has prompted the need for more precise and sustainable farming methods. Introducing precise farming especially in developing countries like South Sudan would not only reduce unnecessary production costs, but promote a more sustainable and environmentally-friendly farming. As South Sudan intends to embark on large scale commercial farming to address food security, such tools as PCA would be utilized in designing precise and sustainable farming methods that would compromise both economic and ecological imperatives.

The objective of this studywas to assess how best the PCA in combination with GMM and HCA tools could be used to adequately characterize the spatial distribution of some major soil nutrients such as Nitrate-N, P, K and micro-nutrients: Mn and Fe as well as the soil penetration resistance.

2. MATERIALS AND METHODS

2.1 Study area

The study was conducted at the demonstration farm of the Department of Agricultural Sciences, College of Natural Resources and Environmental Studies (CNRES), University of Juba, South Sudan. The study area lies within the green belt agro-ecological zone of Central Equatoria State (CES), South Sudan and lies between latitude 4°50’28’’ and longitude 31°35’24. Theaverage annual rainfall is about 650 mm during the months from April to October with dry spell in July. The climate of the area is tropical wet during the rainy season with average temperatures at around 27 to 30℃ and over 35℃ dur ing the dry season of November to March. The soils can be predominantly classified as *EutricLeptosol* with less associated *EutricGleysol* as shown in Table 1.

Table 1.Average values of some selected physical and chemical properties of*EutricLeptosol*

Soil property	Description
Soil Texture Classification*	Sandy loam
Drainage Class (0-0.5%)*	Moderately well
Sand	47.6%
Silt	45.1%
Clay	7.3%
pH (LaMotte STH Test Method)	7.2
Vol. Water Content	18.4%
Bulk Density (gm/cm³)	1.34
Humus Content	2.95%

*__Source__: Harmonized World Soil Data Viewer Version 1.2

The experimental area was the 40 m x 80 m demonstration farm or nursery of the Department of Agricultural Sciences, Universityof Juba and divided in to 32 plots each 10 m x 10 m. Four samples were extruded from each 12 out of the 32 randomly chosen plots for the NO_3-, P, K, Mn and Fe analysis. Similarly, four penetrations at a separation distance of 1 to 2 m per plot to maximum depth of 80 cm were determined using a hand push electronic cone penetrologger (*EijkelkampPenetrologger SN*) as reported by [20, 21] with a cone type 1.0 cm^2, 60° and a penetration speed of 2 cm/s. Processed so il samples were analyzed for various soil nutrients: nitrate-nitrogen, phosphorus, potassium, manganese and ferric iron using the LaMotte basic Model STH-4 Outfit (Code 5029) and values expressed in kg/ha except for Mn which was expressed in ppm.

2.2 Principal Component Analysis (PCA)

Datasets consisting of (n=48) from 12 randomly selected plots were analyzedby PCA. The PCs were calculated using a covariance matrix. A class model related to the classification of soil nutrients according to their spatial distribution and variability was developed by applying the statistical software package XLSTAT 2014.4.06 (Addinsoft SARL, Paris, France). PCA was used to establish the simplest mathematical model capable of describing the dataset satisfactorily.

2.3Geo-statistical analysis

Spatial variability of soil nutrients as well as penetration resistancemeasured as cone index were analyzed using geo-statistics. Geo-statistical software GS+^{TM} Version 9 (GAMMA DESIGN SOFTWARE, LLC, PLAINWELL MICHIGAN,USA, 2001) was applied to quantify the isotropic spatial variability and semi-variogram models. Several models: spherical, exponential, Gaussian and linear semi-variogram were considered in selecting the best fitting model based on the values of weighted residual sums of squares, regression coefficient (r^2) and relative spatial structure or dependency measured as the ratio of the nugget to sill.

For the GS+ Version 9, the semi-variance is defined by the following equation:

$$\gamma(h) = \sum_{i=1}^{N(h)} [2(x_i + h)+2(x_i)]^2 \qquad \text{(Eqn. 3)}$$

whereγ(h) is the experimental semi-variogram value at distance interval h; N(h) is number of sample value pairs within the distance interval h; and z (x_i+h) is sample value at two points separated by the distance interval h. All pairs of points separated by distance h (lag h) were used to calculate the experimental variogram. Several variogram functions were evaluated to choose the best fit with the data. After several simulation runs, the Gaussian model was found and chosen to be adequate in describing the measured data and was fitted to the empirical semi-variogram as:

$$\gamma(h) = C_0 + C\ (1- \exp\ (\frac{-h}{A_0})] \qquad \text{(Eqn. 4)}$$

whereγ(h) = semi-variance for interval distance class h, h= lag interval, C_0 = nugget variance ≥ 0, C = structural variance ≥ C_0, and A_0 = range parameter. In the Gaussian model, both the effective range A = $3^{0.5}A_0$and sill (C+C_0 that lies within 5% of the asymptote) are less discernible owing to the gradual and asymptotic rise of γ(h).

3. RESULTS

3.1. Principal Component Analysis

Application of the PCA three principal components were extracted thatcumulatively explained for 75.67 % of the total data variability with an eigenvalue of greater than 1 (Table 2). The other three components had an eigenvalue less than 1 and were not considered. In our study, the first two components accounted for 56.65% of total data variability as represented the single largest variability contained in the data set [22].

Table 2. Principal Component Analysis

	F1	F2	F3	F4	F5	F6
Eigenvalue	1.881	1.517	1.142	0.671	0.484	0.305
Variability (%)	31.356	25.291	19.025	11.187	8.063	5.078
Cumulative %	31.356	56.647	75.672	86.859	94.922	100.000

The first component F1 correlated with 3 of the 6 soil nutrients analyzed. The nutrients Nitrate-N, P and Fe correlated positively with the F1, while it negatively correlated with Mn and CI and neither with K (Fig. 1).

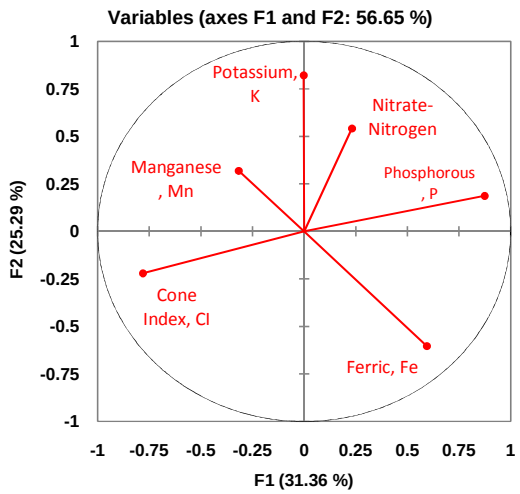

Fig. 1.Loading plots of principal components analysison soil nutrients: Nitrate-N, P, K, Fe, Mn as well as soil penetration resistance measured as Cone Index, CI

Similarly, the F1 *versus* F2 loading plot showed that, variable K was not in any way correlated with CI, however, the variable CI was negatively correlated with both Nitrate-N and P and positively correlated with Mn.Variables with significant loadings on F1 Nitrate-N (factor loading= 0.231), P (factor loading = 0.875), Fe (factor loading = 0.594), while Mn (factor loading= -0317) and CI (factor loading=-0.781) had negative loadings. Interestingly, variable CI had negative loadings on both F1 and F2, whereas Nitrate-N, P, K and Mn had positive loadings on F2 and Fe negative loading on F2.

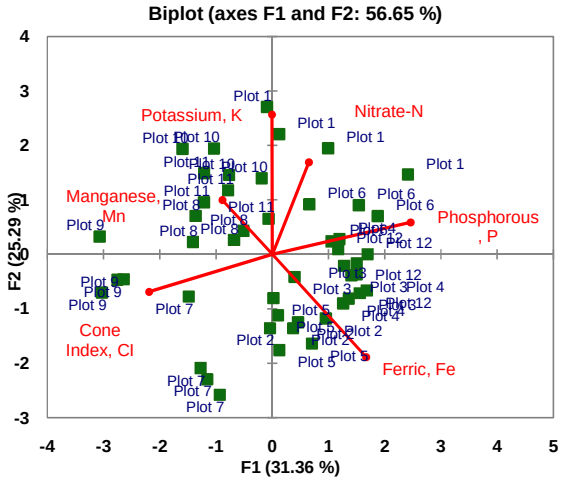

Fig.2.Biplot generated by the principal component analysis for some soil nutrients and CI of the sandy loam soil (*EutricLeptosol*)

Biplot generated by the PCA in Fig. 2 showed the spatial distribution of the different variables in the different plots. For example, the amounts of P, $N0_3$-N, K and Fe in kg/ha were higher in plots 1, 3, 4, 5, 6 and 12 than in the rest plots. Similarly, CI and Mnwas higher in plots 7, 8, 9, 10 and 11 than in the rest plots. The CI as first principal component F1 significantly influenced the spatial distribution of the soil nutrients.

Ward's method of hierarchical agglomerative clustering was used to characterize soil parameter CI and soil nutrients N P K, Fe and Mn. The method is based on minimum variance in which groups are formed so that the pooled within-group sum of square is minimized. Fig. 3 shows the hierarchical agglomerative clusters. The first clusters (Phosphorous, P and Ferric, Fe) and (Manganese. Mn and Nitrate-Nitrogen, N) are fused at 0.3412 and 0.2542 respectively. The second cluster (Mn, N and Potassium, K) are fused at around 0.1042. The third cluster (Mn, N, K and CI) is fused slightly above the truncation level of 0.0542.

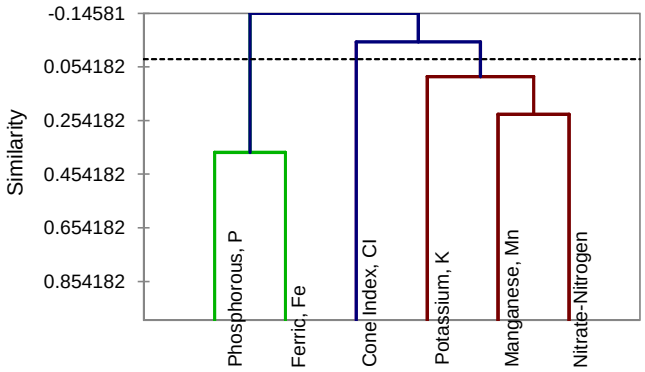

Fig.3.Dendrogram generated by the agglomerative Hierarchical Cluster Analysis (HCA) for some soil nutrients and CI of a sandy loam soil (*EutricLeptosol*)

For all the measured soil variables, there was only one class of spatial dependency as shown in Table 3. Spatial dependency was classified on the basis on nugget/sill ratio. Spatial class ratio between 0 to 25% were considered strongly spatially dependent; those between 25 to 75% moderately spatially dependent; and greater than 75% as weakly spatially dependent [23]. The structural variance of all estimated parameters showed a narrow range of variability within the plots between 8.6 and 15.5% thereby indicating a strong spatial dependency.

Table 3.Geo-statistical parameters for the different soil nutrients including the Cone Index of a sandy loam soil (*EutricLeptosol*) (Adapted from Lomeling[21])

Parameter	NO_3^--N	P	K	Mn	Fe	CI
Model	Gaussian	Gaussian	Gaussian	Gaussian	Gaussian	Gaussian
Nugget variance, C_0	104.24	2402.50	1200.00	0.065	0.089	0.0378
Sill, $C+C_0$	827.00	27930.00	7758.00	0.674	0.774	0.282
Range (m), A_0	14.85	8.50	16.93	98.76	1.95	8.80
R^2	0.709	0.865	0.793	0.532	0.304	0.232
RSS	904998	4.04E+08	2.14E+07	0.0425	0.136	0.261
$C_0/C+C_0$	0.126 (12.6%)	0.086 (8.6%)	0.155 (15.5%)	0.096 (9.6%)	0.115 (11.5%)	0.134 (13.4%)
0-0.25 or 0-25% strong dependency; 0.25-0.75 or 25-75% moderate dependency; >0.75 or >75% weak dependency						

The results of the Pearson's correlation matrix have been presented in Table 4. The correlation matrix identified the relationship among the soil nutrients and CI with values between +1 or -1 revealing either a positive or negative relation between the variables. In general, the CI showed a negative correlation with most soil nutrients: $N0_3$-Nitrogen, P, K and Fe (*p<0.0001*) and a weak positive correlation with Mn. However, between the individual soil nutrients, P showed a weak positive correlation with K, Nitrate-Nitrogen Fe. Similarly, Mn showed low negative correlation with bothP and Fe (*p<0.0001*) as well as with K and Fe (*p<0.0001*).

Table 4.Pearson correlation coefficient matrix (r) of CI and some soil nutrients

Variables	Cone Index, CI	Manganese, Mn	Phosphorous, P	Potassium, K	Ferric, Fe	Nitrate-Nitrogen
Cone Index, CI	**1**					
Manganese, Mn	0.138	**1**				
Phosphorous, P	-0.566	-0.172	**1**			
Potassium, K	-0.139	0.032	0.155	**1**		
Ferric, Fe	-0.208	-0.127	0.373	-0.455	**1**	
Nitrate-Nitrogen	-0.117	0.232	0.231	0.148	-0.025	**1**

The results of factor loadings are presented in Table 5. The first three principal components F1, F2 and F3 had eigenvalues greater than 1 and accounted for 75.67% of total variation of the data with values at 31.36%, 25.29% and 19.03% respectively. F1 showed high loadings of P, Fe and low Nitrate-Nitrogen with positive effect and CI, Mnand K with negative effect. F2 showed high loadings of K, Nitrate-Nitrogen and moderate loadings of Mn and P with positive effects whereas it showed negative effects with CI and Fe. F3 showed high loadings of Mn, Nitrate-Nitrogen and moderate to low loadings of Fe and CI with positive effects with negative effects of both P and K.

Table 5. Factor loadings of the first 5 components of some soil nutrients and CI of a sandy loam soil (*EutricLeptosol*)

	F1	F2	F3	F4	F5
Cone Index, CI	-0.781	-0.221	**0.130**	0.333	0.443
Manganese, Mn	-0.317	**0.320**	**0.717**	-0.524	0.082
Phosphorous, P	**0.875**	**0.188**	-0.002	-0.020	0.265
Potassium, K	-0.002	**0.821**	-0.395	-0.066	0.329
Ferric, Fe	**0.594**	-0.603	**0.326**	-0.020	0.290
Nitrate-Nitrogen	**0.231**	**0.542**	**0.591**	0.530	-0.135

We employed a Gaussian mixture model of the analyzed dataset as criterion for estimating the probability density function. 2-D probability density function ellipsoids are shown in Fig. 4 represented as two clouds and centered on the means: μ_1 at 2 and $\mu_2$at3 for cluster classes 1 and 2 respectively. Similar clustering is based on the k-means which seeks to cluster data point closest to some centroid. A closer look at this figure gives shows two pdf peaks at 0.9 and 0.1 for F1 and F2 respectively.The classification into different cluster classes was computed according to the probabilities via the MAP (Maximum A Posteriori) rule.

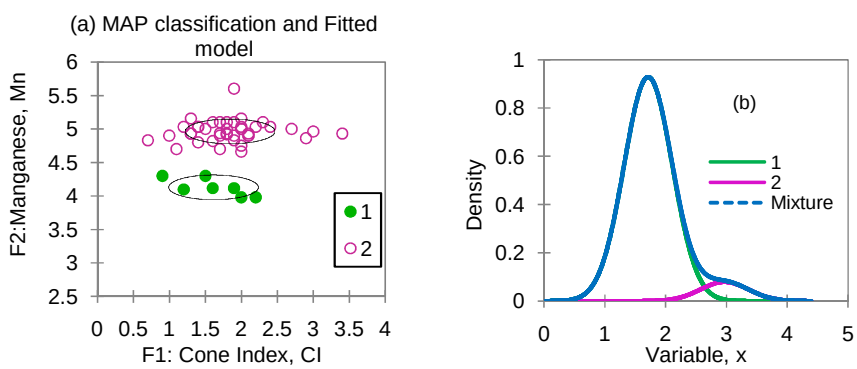

Fig.4.(a) Scatter plots of a two dimensional data between F1 and F2 generated by a mixture of Gaussians. Ellipsoids corresponding to the region with 95% of the componentdensity and (b) example of densities modeled by mixtures of two Gaussians pdfs. Green and green lines indicate the individual component densities and blue line the mixture of both densities (*bimodal density*)

The results of the GMM also showed the loadings of F1 on F2. Class 1 denoted as F1 (*associated with*CI) with smaller dataset was clustered around 1.0-2.0 MPa had low loading on F2 (*associated with*Mn) at around 4 ppm with positive effect. Similarly, Class 2, F1 had a more wider clustering between 1.2-2.5 MPa with outliers up to 3.5 MPa had high loading on F2 at about 5 ppm with positive effect. This shows that increase of F1 positively correlated with F2 suggesting that increase in the CI tended to increase Mn in soil. Such positive correlation was also

reported by [20].Low loading and negative effects of F1 component on F5 (*associated with* Fe) was investigated and shown as in Fig. 5. In class 2 and 4, F1 showed low loading on F5 at around 1.0-1.5 MPa with negative effect and several outliers at around 5-5.5 ppm. Class 1 and 3 showed fewer data points of F1 clustered between 1.5-3.0 MPa with F2 at around 4.5 ppm. The results showed the negative correlation between F1 and F5 suggesting that low CI tended to favor increased Fe. Increasing the CI led to a more reduced Fe [20].

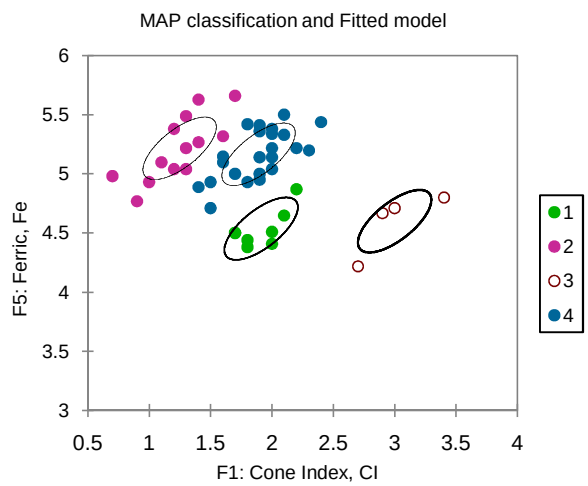

Fig. 5. Scatter plots of a two dimensional data between F1 and F5 generated by a mixture of Gaussians with diagonal covariance matrices

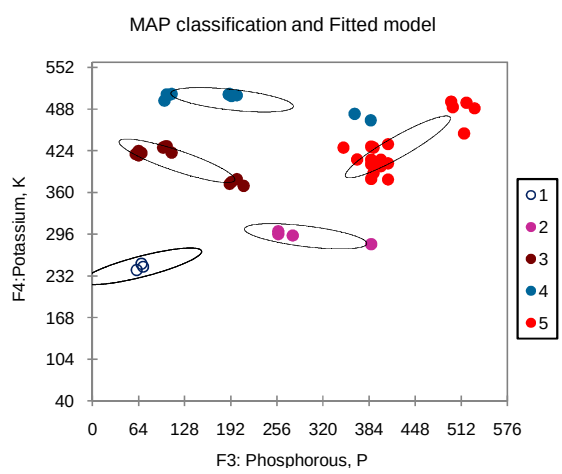

Fig. 6. Scatter plots of a two dimensional data between F3 and F4 generated by a mixture of Gaussians with diagonal covariance matrices

Figure6 shows a group of 5 clusters (*representing 5 probability density functions*) classes between the F3 and F4. This showed that within the complete dataset five center locations for both point P and K were evident with lowest P concentrations at 64 k/ha and for K at around 232 kg/ha respectively. The fifth cluster and highest concentrations for P and K were at around 512kg/ha and 488 kg/ha respectively suggesting a more synergistic relationship between both elements. Between 95-99% of soil P is insoluble P andunavailable for plants [24, 25].

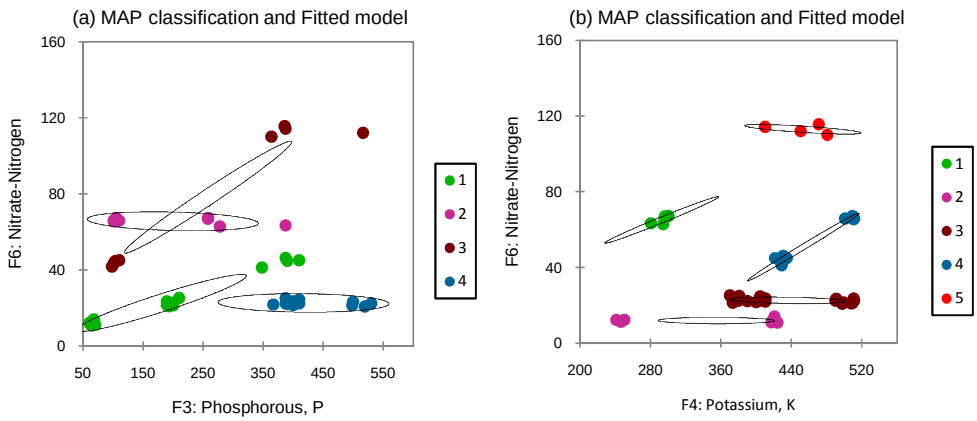

Fig. 7. Scatter plots of a two dimensional data between (a) F3 and F6 and (b) between F4 and F6 generated by a mixture of Gaussians with diagonal covariance matrices

Figure 7(a) showed the F3 loading on F6 with four cluster classes (*representing four probability density functions*). Although not quiet discernible, a more accurate and clearer antagonistic relationship especially between classes 2 and 4 was noticeable. At lower P between 60-250 kg/ha, the Nitrate-N was around 70 kg/ha and reduced to about 30 kg/ha with increased P at between 350-550 kg/ha. Similarly, Figure 7(b) showed a less discernible relationship between F4 and F6 with both antagonistic and synergistic features, however, with a predominating antagonistic relationship. In class 1, with F4 at around 280 kg/ha, the F6 was at around 60kg/ha. In class 3 and 4, F4 was at 380-520 kg/ha and 440-520 kg/ha respectively. There was generally a negative correlation between F4 and F6.

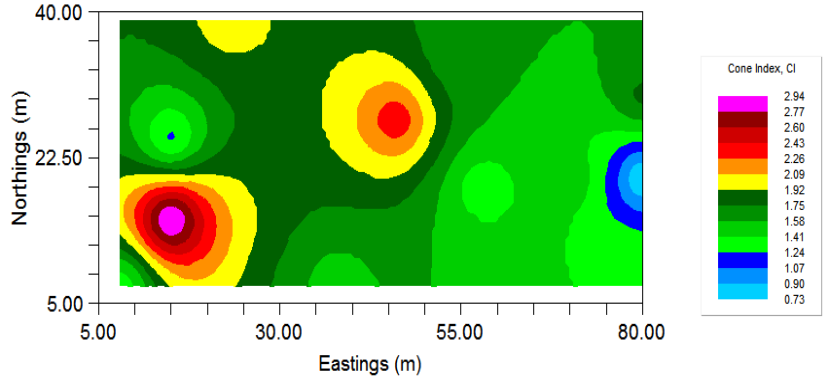

Fig. 8.IDW contour maps showing the distribution of the Cone Index CI in a *EutricLeptosol*(Adapted from Lomeling[21])

Contour maps were generated by the interpolation method (Inverse Distance Weighting, IDW)as in Fig. 8. Although the plots between 1, 3 and 9 remained fallow during the last five years, CI values were high between 2.26 to 2.94 MPa due to the presence of tree roots and termite anthills. CI values tended to decrease in an easterly direction to values as low as between 0.73 – 1.07 MPa.Generally, the CI values in all plots were below the critical value of 3 MPa on sandy loam soil as observed by [26].

Table 6. Statistical parameters of some soil nutrients and CI of a *EutricLeptosol*

Variable	Mean	Variance	Skewness (Pearson)	Kurtosis (Pearson)	Kolmogorov-Smirnov Test	Chi-Test
CI	1.840	0.270	0.587	0.866	0.368*	0.348*
K	416.307	5576.883	-0.708	-0.132	0.443*	<0.0001*
Mn	4.842	0.112	-1.173	1.139	0.008*	<0.0001*
P	260.021	23846.11	0.136	-1.465	0.121*	<0.0001*
Nitrate-N	39.087	832.009	1.317	0.906	0.002*	<0.0001*
Fe	5.027	0124	-0.353	-0.721	0.912*	0.277*

** significant at $p \leq 0.05$*

Table 6 shows some statistical parameters that were tested using the Kolmogorov-Smirnov (KS) and Chi-tests. The test using the (KS) showed that the values of the variables: CI, K, P and Fe were significant at $p \leq 0.05$ and that the data followed normal distribution, whereas Mn and Nitrate-N were not. Conversely, the Chi-test showed that all values except of the CI and Fe variables were not significant at $p \leq 0.05$ and so did not follow normal distribution. No reasons could be given for this contradicting fit-of-goodness of both test procedures.

3.2 Nutrient distribution and spatial dependency

It was reported that, soil nutrients generally tended to increase with the sand fraction whereas this was the reverse trend for silt and clay fractions. Sand particles have relatively higher friction coefficients than silt and clay particles, which would explain the increasing trends of CI with sand fraction [20].

From the dendrogram diagram as in Fig. 3 it may be seen that the first four variables (P, Mn, Fe and Nitrate-N) are clustered together at first level with strong spatial dependency at 8,6%, 9.6%, 11.5% and 12.6% respectively. Meanwhile, K is then clustered to Nitrate-N and Mn at second level also with strong spatial dependency at 15.5%. Results of our study showed that lower spatial dependency values implied higher similarities between the compared variables.

4. DISCUSSION

The results of PCA and cluster analysis showed that the spatial distribution of soil nutrients was most strongly correlated with soil CI within the different plots. The results further showed that CI was the most important major factor that affected the spatial distribution of the soil nutrients. These results are in agreement with those reported by [20] on *EutricLeptosol*. Increasing CI in the different plots was a major aspect that influenced the spatial distribution of soil nutrients. With an increase in the CI, the available $N0_3$-N, P, K and Fe decreased correspondingly as shown by the negative Pearson´s correlation. For example, the amounts of P, $N0_3$-N, K and Fe in kg/ha were higher in plots 1, 3, 4, 5, 6 and 12 than in the rest plots. Similarly, CI and Mnwas higher in plots 7, 8, 9, 10 and 11 than in the rest plots. The CI as first principal component F1 significantly influenced the spatial distribution of the soil nutrients.

Although the first threefactors F1, F2 and F3explained most of the variance, components F4 to F6 were also chosen to explain any interactions and loading effects of each individual variable. The idea was to assess, if there were any antagonistic or rather synergistic effects of any of the variables on the other. The study also tried to assess, if indeed anthropological influences had any significant roles on spatial distribution of some soil nutrients, or rather if spatial distribution was random and only dependent on lithogenic influences. Associating the respective variables to each component, (CI to F1, F2 to Mn, F3 to P, F4 to K, F5 to Fe and F6 to Nitrate-N) and incorporating geo-statistical parameters obtained from Gaussian model (Table 3), the short-range component (at 8.8 m) of F1 positively correlated with F2 (at range 98.76 m). Conversely, F1 negatively correlated with F5 (at range 1.95 m). Component F3 positively correlated with F4 with ranges at 8.5 and 16.93 m respectively.

The plots on the research and demonstration farm showed heterogeneous distribution and spatial variations of soil nutrients and cone index. The results also indicated that not all soil nutrients and cone index values were homogenously distributed in all plots. Plots in the same quadrant with similar predominating soil nutrients could practically be managed alike. It also implied that, similar plot

management practices would be tailored on each quadrant within the context of precision farming. Kriged maps reported by [21]showed more CI, Nitrate-N and P on the left-hand side gradually decreasing in the easterly direction, whereas Mn and Fe tended to increase in the opposite direction. Implicitly, this did not suggest any negative correlations between CI and both Nitrate-N and P, on the contrary, high CI negatively correlated with both Nitrate-N and P as reported by [20].

The CI values were generally high(2.94 MPa) at the left-hand side of the farm and gradually decreased in an easterly direction to as low as 0.73 MPa at the extreme right edge of the farm. One reason for this is probably the differentiated tillage activities on the farm. There was hardly any tillage activity on the left hand side of the farm in the last 3 years as opposed to much soil tillage on the right hand side which must have significantly altered the structural and aggregative nature of the soil.

The interaction CI and Mn (or F1 and F2) found in the present study could be described as synergistic. Increase in CI would increase inter-particle contacts with correspondingly reduced pore space. This would tend to favor anaerobic soil conditions thereby increasing Mn availability. In the present study, we also found out that on average, increasing P concentration in soil did not accentuate K availability in soils. Although the *EutricLeptosol* has a low organic carbon (0.72% by weight) and low $CaC0_3$ (0.2% weight), there appeared to be specific preference for Ca^{2+}as opposed to K^+ ions in soils in formingbothin- and organic complexes with phosphate ions. Such mutually synergistic relationship would suggest that increasing concentration of P ions would enhancefurther adsorption to Ca^{2+} and so increasing the concentration of K^+. Conversely, a mutually antagonistic relationship was also found between P and Nitrate-N whereby there was specific and preferential adsorption of the phosphate than Nitrate ions [27]. Furthermore, our work showed that Nitrate-N was clustered with Mn. This mutually synergistic relationship would be the result of high water table and poor drainability of the *EutricLeptosol*following the heavy rains between April and November. In the absence of cultivated crops (*that would utilize the Nitrate-N*), much of the Nitrate-N was stored within the rhizophere. Equally, the high CI values tended to reduce soil drainability thereby enhancing the development of anaerobic conditions which favored reduction processes and thereforeMn increase [20]. Based on these groups, the clustering of these nutrients in thestudied soilcould be ascribed to two different components: P and Fe that have a common lithogenic origin and the CI, K, Mn and Nitrate-N to have a common anthropogenic influence.

The K^+ is required for cells to maintain the osmotic balance and regulating stomatal opening and closure during photosynthesis [26, 28]. It is also an essential co-factor for many enzymes. Similarly, P concentration in plant cells is closely associated with photosynthetic rate [29] and; therefore, decrease of P in leaves will reduce the plant growth. By implications, soils low in K^+ would not only reduce stomatal activity, but also low available P would lead to reduced photosynthesis and hence reduced plant growth rates. Analysis of thedendrogram confirmed that the CI is an

independent variable that tended to influence spatial variations of K and Nitrate-N suggesting that both nutrients reduced with increased CI whereas Mn increased with increasing CI and *vice-versa.*

P showed a significant and positive relationship with Fe (r = 0.37, $p < 0.0001$). The significant relationship between these nutrients could be as a result of adsorption resulting from degradation of manure or organic material and mineralized Fe-oxides.Especially under anaerobic conditions influenced by both high soil compaction (high CI values) and high water table, this would favor reduction processes and enhancement of *Vivianite*$Fe_3(P0_4)_2.8H_20$ [30]. Clustering and the mutually synergistic relationship between P and Fe was to be expected as much of the P ions ($H_2PO_4^-$) were easily adsorbed to Fe ions (Fe^{2+}) especially under poor drainable conditions(*conditioned by high CI values or high soil penetration resistance*) where Fe(III)-oxide was reduced to Fe^{2+}.

The set of 6 original soil properties as in Fig. 3 with low similarity measure at 0.054 (*truncation or cut-off threshold level*), created three classes that explained 75.67 % of the data variability. Increase of the cut-off level would not only imply increase in the number of classes but also represent percentage increase in similarities suggesting that comparison was better under a large number of classes. For example, increasing similarity at cut-off threshold level between 0.2542 and 0.4518 would increase the number of classes to 5 therefore enhancing better comparison of the variables.

In our study,the Euclidian distance was utilized as a basis for classifying the clusters that demonstrated both mutual antagonistic and synergistic relationships between the different principal components. However, the Euclidean distance, suffers from the so-called "*scaling effect*"[17] due to the inadvertent weighting of the variables in theanalysis that can occur due to differences in magnitudeamong the measurement variables. For example, the measured concentration of two variables: P and Nitrate-N is soils with varying ranges. The measured amount of P in the studied soil varied between 61 and 511 kg/ha, whereas that of Nitrate-N between 11 and 116 kg/ha approximately one-fifth of P. Afive-fold increase in the P concentrationwill have a greater effect on Euclidean distance than a similar increase in the Nitrate-N concentration.However, effect of variablescaling on the Euclidean distance by the different measured variables may be mitigated bystandardizing themeasurement variables, so that each variable has a meanof zero and a standard deviation of 1.

One of the major problems in hierarchical clustering is setting a threshold value or truncation level that would define similarity within a cluster. This value is often subjective [17].By varying the truncation or cut-off level, similarity within a cluster may increase or decrease. Thisis conditioned by practical considerations as to whether or not one needs a high degree of similarity within a given cluster. Ifthe similarity value is substantially larger, the random variation and dissimilarity within the data isprobably negligible and *vice-versa*. Our study also showed that clustering of the different variables gave rise to correspondingly different number of classes, e.g. between F1 and F2 gave 3 cluster classes, whereas between F4 and

F6 gave 5 cluster classes. Practically, smaller number of classes e.g. 2 to 3 would suggest lesser variability with a more homogenous spatial distribution than a larger number of cluster classes that would show more variability and heterogeneity.

5. CONCLUSION

The application of PCA, HCA and GMMwas used to grouping multi-nutrient and Cone Index data on sandy loam soil (*EutricLeptosol*) from the University of Juba demonstration farm, in South Sudan. The first principal component F1, explained 31.36%, the second principal component F2, accounting for 25.29% and F3 19.06%. GMM analysis for the rest components F3 to F6 showed that these components representing individual variables had either antagonistic or synergistic relationships, which explained their spatial distribution and variations. The results of PCA made it possible for the initial six variables to be reduced to three factors representing 75.67% of the total variance. From hierarchical clustering, CI was observed to be clustered with K, Mn and Nitrate-N whereas P was clustered with Fe. It can be inferred from these results that, the spatial distributionof K, Mn and Nitrate-N was predominantly due to anthropological influences as conditioned by CI and less lithogenic as determined by nature of*EutricLeptosol*.

The PCA is a standard tool in modern data analysis owing to its simplicity, non-parametric method for extracting relevant information from large and confusing data sets. [31] in his unpublished work describes how with minimal effort,the PCA can provide a roadmap for reducing complex data set to a lower dimension that can reveal sometimes hidden, simplified structures. However, PCA is not a statistical method from the viewpoint that there is no probabilitydistribution specified for the observations. Therefore, it is important that only in combination with other pattern recognition tools does it best serve to explain underlying relationships between the different variables within a data set.

The PCA, HCA and GMM methods used herein have been found to bemost useful in identifyingthe structure and inherent inter-linkages within the variables contained in the dataset. The approaches described in these methods rely heavily on mathematical algorithms whose results are presented in graphical form as clusters, scatterplots and dendrograms for easy patter recognition and interpretation. For two independent variables or components, amixture of Gaussian probability density functionsis best modeled under the assumption that the distribution is bi- or multimodal as it is poorly approximated using a single Gaussian.

In conclusion, the PCA, HCA and GMM helped reveal some underlying relationships between soil nutrients and soil parameter CI. They were shown to be useful tools for studying the spatial distribution and variations of soil nutrients and variables either due to intrinsic factors (*lithogenic*) or anthropological influences. More research on wider range of soil mechanical-physical as well as chemical parameters would provide more knowledge on their interactions and inter-linkages.

ACKNOWLEDGMENTS

This research was partly supported by the Rebuilding Higher Education in Agriculture (RHEA) Project funded by USAID (HED072-9742-SDN-11-01) through Virginia Tech at the College of Natural Resources and Environmental Studies (CNRES), University of Juba as well as purchase of the geo-statistical software GS+^{TM} Version 9. We equally thank the Norwegian funded project NORAD for the purchase of penetrologger and soil testing kit. Appreciation is also extended to Sebit Mathew Otwari and Yahya Mohammed Khaterfor conducting field trials, analyzing plant, soil physical and chemical properties.

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

1. Cleto MS de Moura, et al. Analysis and assessment of heavy metals in urban surface soils of Teresina, Piauí State, Brazil: a study based on multivariate analysis. Comunicata Scientiae.2010;1(2):120-127.
2. Abollino OM, et al. Heavy metals in agricultural soils from Piedmont, Italy. Distribution, speciation and chemometric data treatment. Chemosphere. 2002;49:545–557.
3. Sabrina T, et al.Earthworm Populations and Cast Properties in the Soils of Oil Palm Plantations. Malaysian Journal of Soil Science. 2009;13: 29-42.
4. OseiMK, et al. Genetic Diversity of Tomato Germplasm in Ghana using Morphological Characters. International Journal of Plant & Soil Science. 2014; 3(3):220-231.
5. Reimberg, MC et al. Multivariate analysis of the effects of soil parameters and environmental factors on the flavonoid content of leaves of *Passiflora incarnate* L., Passifloraceae.Brazilian Journal of Pharmacognosy. 2008;19(4):853-859.
6. Dayang SN, Fauziah CI. Soil Factors Influencing Heavy Metal Concentrations in Medicinal Plants. Pertanika J. Trop. Agric. Sci. 2013;36(2):161–178.
7. Silva SA, Lima JSS. Multivariate Analysis and Geostatistics of the Fertility of a HumicRhodicHapludox under Coffee Cultivation.R. Bras. Ci. Solo. 2011;36:467-474.
8. Bam EK, et al. Multivariate cluster analysis of some major and trace elements distribution in an unsaturated zone profile, Densu river basin, Ghana. African Journal of Environmental Science and Technology. 2011;5(3):155-167.
9. Wander MM, Bollero GA. Soil quality assessment of tillage impacts in Illinois. Soil Sci. Soc. Am. J. 1999;63:961–971.
10. Ye RZ, Wright AL. Multivariate analysis of chemical and microbial properties in histosols as influenced by land-use types.Soil & Tillage Research.2010; 110: 94–100.

11. Allison VJ, et al. Changes in enzyme activities and soil microbial community composition along carbon and nutrient gradients at the Franz Josef chronosequence, New Zealand. Soil Biol. Biochem. 2007;39:1170–1781.
12. Cookson WR, et al. Controls on soil nitrogen cycling and microbial community composition across land use and incubation temperature. Soil Biol. Biochem. 2007;39:744–756.
13. Smejkalova D, et al. Multivariate analysis of CPMAS ^{13}C-NMR spectra of soils and humic matter as a tool to evaluate organic carbon quality in natural systems. European Journal of Soil Science.2008;59:496–504.
14. Brereton RG. Chemometrics.Data Analysis for the Laboratory and Chemical Plant.John Willey and Sons, West Sussex; 2003.
15. Davis JC. Statistics and Data Analysis in Geology, 2nd ed., John Willey and Sons, New York; 1987.
16. Everitt B, et al.Cluster Analysis, 4th ed., Oxford University Press, Inc., New York; 2001.
17. Lavine BK. Clustering and Classification of Analytical Data, in: Robert A Meyers (Ed.),Encyclopedia of Analytical Chemistry. John Wiley & Sons Ltd, Chichester, USA; 1992.
18. Gu J, et al. Gaussian Mixture Model of Texture for Extracting Residential Area from High-Resolution Remotely Sensed Imagery. ISPRS Workshop on Updating Geo-spatial Databases with Imagery &The 5th ISPRS Workshop on DMGISs. Hannover, Germany, 17-20, May, 2005.
19. Clarkson D, et al. S+ Functional Data Analysis User Guide, Springer, New York; 2005.
20. LomelingD, Abakr AA. Variability of Cone Index on Seedling Emergence Rate and Growth Establishment of Cowpea in a Sandy Loam Soil (Eutric Leptosol).Int. J. of Sci.: Basic and Applied Res. 2014;14(1):34-48.
21. Lomeling D. Correlating the spatial distribution of some macro- and micro nutrients with Cone Index in a sandy loam soil (EutricLeptosol). Int. J. of AgriScience. 2014; 4(2): 89-101.
22. Johnson RA, Wichern DW. Applied multivariate statistical analysis, 5thed. New Jersey, Prentice Hall; 2002.
23. Cambardella CA, et al. Field-scale variability of soil properties in central Iowa soils. Soil Sci. Soc. of Amer. J. 1994;58:1501-1511.
24. Sridevi M, et al. Isolation and Characterization of Phosphate- solubilizing bacterial species from different crop fields of Slem, Tamil Nadu, India. Int. J. of Nutrition, Pharmacology, Neurological disease. 2013; (3): 1.
25. Gyaneshwar P, et al. Role of soil microorganisms in improving P nutrition of plants. Plant Soil.2002; 245:83–93.
26. Hâkansson I, Lipiec J. A review of the usefulness of relative bulk density values in studies of structure and compaction. Soil Till. Res. 2000;53:71-85.
27. Yeo A. Molecular biology of salt tolerance in the context of whole-plant physiology. J. Exp. Botany.1998; 49(323): 915-929.
28. Chow W, et al. Growth and photosynthetic responses of spinach to salinity: implications of K+ nutrition for salt tolerance. Funct. Plant Biol. 1990;17(5):563-578.

Formatted: Font: (Default) Arial

29. Overlach S, et al. Phosphate translocator of isolated guard-cell chloroplasts from Pisumsativum L. transports glucose-6-phosphate. Plant Physiol. 1993;101(4): 1201-1207.
30. Schachtschabel, et al. Lehrbuch der Bodenkunde. 12th ed. Ferdinand Enke Verlag, Stuttgart Germany; 1989.
31. Shlens J. ATutorial on Principal Component Analysis.Unpublished manuscript; 2009.