Short Research Article

DISCRIMINANT ANALYSIS AND IT'S APPLICATION TO THE OIL PALM CULTIVATION IN NIGERIA.

ABSTRACT

There are various environmental factors that need to be considered when assessing the suitability of site for oil palm cultivation some of which are; climate, vegetation, and the soils. Using soil morphology and degree of profile development or the nature of the parent bedrock and the vegetation formed grouped the soil supporting oil palm as generally belonging to five parent materials. The various soil groups includes:- Crystalline metamorphic and Igneous rocks, Shale mixed with sand stone and clay, Coastal plain sand, Coastal alluvium and Fresh water swamp, these groups are presumed to differ on several physiochemical properties and formed the basis on which land is being selected for oil palm cultivation. The classification of soil location in future into any of the five soil types on the basis of the soil characteristics can be facilitated using different approach, but in this study, descriminant analysis will be used to measure the success rate of classifying the soil types by using the physiochemical properties soils in the Raphia growing zone of Nigeria. The research is aimed at applying discriminant analysis thereby satisfying the following objectives: Use discriminant analysis to predict soil group membership in order to correctly classify future unknown observation into any of the five soil groups based on the observed predictors (soil characteristics) in soils supporting Raphia palms of Southarn Nigeria, Form linear combinations of the discriminating predictor variables that differs significantly in their group means, Identify the soil properties that best discriminates among the soil types. The results have demonstrated the use of some multivariate statistical methods as useful tool for soil study, also permitted a better understanding of soil variability within the oil palm belt of Nigeria. This knowledge will be useful for assessing soil diversity for crop improvement and as a key for designing agricultural management practices, especially for the oil palm belt of Nigeria.

Keywords: [Discriminant Function; Soil Class; Soil Characteristics; Raphia growing zone; Southern Nigeria]

1. INTRODUCTION

Climate, vegetation and the soil are the three key environmental factors routinely considered when assessing the suitability of site for oil palm cultivation (Paramananthan, 2003 [1]). For soil, the topography and shape, moisture availability, soil physical and chemical properties are most essential. However, the soil chemical properties are more important consideration than soil physical properties in assessing the suitability of soil for oil palm cultivation and the requirement for mineral fertilizers. AS observed by Paramanthan (2003), soil chemical characteristics are more easily changed compared to soil physical properties. Some very important chemical properties useful to oil palm are ; Organic matter, Nitrogen (N) content, Soil Phosphorus (P), Effective Caution Exchange Capacity (ECEC), Potassium (K), Calcium (Ca), Magnesium (Mg), Soil pH and micronutrient toxicity. The knowledge of these soil properties and their response to management practices is an essential requirement by any land users or oil palm growers.

The oil palm belt of Nigeria occupies a wide expanse of land, and each soil location has its own particular characteristics. Knowledge of the levels and patterns of the soils diversity in the site or location with similar soil properties. When soil series are relatively homogeneous, it permits some specific management decisions, For example, an experiment can be conducted, information is not only useful for the particular location, but other locations within the groups or cluster. Such group specific investigation will help proffer management techniques and recommendations that

will lead to significant reduction in the cost of conducting research and provide maximum benefits and other stake holders compare to site-by-site investigation.

Discriminant analysis is a multivariate statistical procedure which can be used to predict group membership from a set of predictors (variables) (Tabachnick and Fidell,1989 [2]). This statistical method appears to have important applications for soil Classification, but has received little attention from pedologist. (Norris, 1970 [3]) described important justifications for the application of multivariate statistical method to the study of soil science. These methods interrelated variables which define soils, Multivariate analysis allows an objective, unbiased extermination of the variables, thus ensuring that a priori perceptions do not lead to incomplete or faulty conclusions. Finally, the knowledge of statistical methods required by proper application of multivariate analysis should result in a precise and repeatable conclusion not possible with non-numeric methods (Norris,1970). Classification of soil developmental sequences in sand dunes (Berg, 1980[4]). Edmands and Lentner (1987[5]) reported that discriminant analysis was better to predict soil response classes than soil Taxonomy. Lentz and Simonson (1987[6]) used discriminant analysis to classify soils associated with sagebrush communities. The analysis reveal that soil properties other than those used in soil Taxonomy were important discriminators among soil classes. We are not aware of any attempts to use a large data set in conjunction with discriminant analysis to examine the relationship of soil landforms (Land types).

2. MATERIAL AND METHODS

The data used for the study are from Chemistry Division, Nigerian Institute for oil Palm Research (NIFOR). The soil samples were taken from 57 locations belonging to five soil types in the southern part of Nigeria, an area believed to be suitable for Raphia palm cultivation. For each of the 57 samples the physiochemical properties were obtained from five soil types. Five soil types and its physiochemical properties are listed in the table 1 and 2 respectively.

Given a set of W independent variables $\boldsymbol{X_1, X_2, \ldots, X_K}$ (Soil characteristics in this case), Multiple Discriminant Analysis (MDA) attempts to find the linear combination (Discriminant Function) of these variables that best separate the groups of cases $\boldsymbol{Z_1, Z_2, \ldots, Z_5}$ (Soil types in this case). The functions are generated from a sample of cases for which group membership is known; the functions can then be applied to new cases with measurements for the predictors variables, but unknown group membership.

In general form, the discriminant function is expresses as:

$$\boldsymbol{Z = a + W_1X_1 + W_2X_2 + \cdots + W_KX_K} \qquad \boldsymbol{3.1}$$

Where

Z = Discriminant Score, $\boldsymbol{X_{k=}}$ an independent variable or predictor variables.

a= Discriminant Constant, $\boldsymbol{W_k}$= Discriminant weight or Coefficients.

Table 1; Parent material (soil type) and group code.

Group code	Soil types
1	Crystalline Metamorphic and Igneous rocks
2	Share Mixed with Sandstone
3	Coastal Plain Sand
4	Coastal Alluvium
5	Fresh Water Swamps

Table 2: Soil- Site Characteristics Measured.

Variable Code	Code of soil properties	Soil properties
$\boldsymbol{X_1}$	Ph	Soil pH

X_2	Orgm	Organic matter
X_3	N	Nitrogen
X_4	P	Phosphorus
X_5	K	Potassium
X_6	Ca	Calcium
X_7	Mg	Magnesium
X_8	S	Sulphur
X_9	Fe	Iron
X_{10}	Mn	Manganese
X_{11}	Cu	Copper
X_{12}	Zn	Zinc
X_{13}	B	Boron

The procedure automatically chooses a first function that will separate the groups as much as possible, it then chooses a second function that is both uncorrelated with the first function and provide as much further separation as possible. The procedure continues adding functions in this way until reaching the maximum number of functions as determined by the number of predictors and groups in the dependant variables. In two group discriminant function, there is only one discriminant function. But for higher order discriminant function, the number of functions each with is own cut-off poin value is the lesser of g-1, where g is the number of categories in the grouping variables. Each discriminant function is orthogonal to the other.

The discriminant ability of the functions will be determined by the following statistic, The Eigen-value, Canonical correlation and The Wilks' Lambda.

i. Eigen-value ($\boldsymbol{\lambda}$) also called the characteristics root indicates the relative discriminating power of the discriminant functions. There is one Eigen value for each discriminant function. With more than one function, the first function will be the largest and the most important, the second next most importance in explanatory power and so on. This can be simply defined as:

$$\boldsymbol{\lambda = BSS/WSS = [\sum(\overline{Z_J} - \overline{Z})^2 / \sum(Z_{1J} - \overline{Z_J})^2]} \qquad \mathbf{3.2}$$

Where BSS is between group variance and WSS is within group variance:

If $\boldsymbol{\lambda}$=0, the model has no discriminant power. The larger the value of ⟍, the greater the discriminant power.

ii. The Conical Correlation ($\boldsymbol{\pi}$) eta is a measure of the association between groups formed by the dependent variable and the given discriminant function and this can be defined as:

$$\boldsymbol{\pi = \sqrt{\lambda/(1+\lambda)} = \sqrt{BSS + TSS}} \qquad \mathbf{3.3}$$

Where BSS is between group variance and TSS is total variance.

$\boldsymbol{\pi}$ =The correlation of the predictor(s) with the discriminant scores produced by the model (measure the association between discriminant scores and the groups).

$\boldsymbol{\pi^2}$ = Coefficients of determination.

$\boldsymbol{1 - \pi^2}$ = Coefficient of non-dettermination.

when the canonical correlation is large ($\boldsymbol{\text{near } 1}$), it indicate that there is high correlation between the discriminant functions and the groups (i.e the function discriminates well among the groups. But when the correlation is zero. It means there is no correlation between the groups and the functions.

iii. The Wilks Lambda used to test the significant of the discriminant function as a whole and this can be defined as:

$\boldsymbol{\Lambda}=(\mathbf{1}-\mathbf{n}^{\mathbf{2}})=[\mathbf{1}/(\mathbf{1}+\boldsymbol{\lambda})]=\boldsymbol{WSS/TSS}$ 3.4

Where WSS is within variance sum of square and TSS is the total sum of square.

A significant lambda, means one can reject the null hypothesis, which says that the two or more groups have the same mean discriminant function score. The chi-square ($\boldsymbol{X^2}$) is then used to tests the significance of the difference in the mean discriminant score between groups and is defines as:

$\boldsymbol{X^2}=-[(\boldsymbol{n}-\mathbf{1})-\mathbf{0.5}(\boldsymbol{m}+\boldsymbol{k}+\mathbf{1})]\boldsymbol{ln}\,\boldsymbol{\Lambda}$ 3.5

$\boldsymbol{k}-\mathbf{1}$ =degree of freedom (df) of a given function. It is based on the number of groups present in the categorical variables and the number of continuous discriminant variables.

The chi-square statistic is compared to a chi-square distribution with the degree of freedom stated here. However, the p-value associated with the chi-square statistic is given. For a given alpha level say 0.05, if the p-value is less than alpha, the null hypothesis that a given function canonical correlation and all smaller canonical correlations are equal to zero is rejected. If not, then we fail to reject the null hypothesis.

In testing the classification performances of the discriminant function, we used the overall hit ratio which is the same thing as percentage of the original group cases correctly classified. Three benchmarks are used. The Maximum Chance Criterion (MCC), Proportional Chance Criterion and Press's Q statistic are used to test the significance of the Hit ratio. If the hit ratio exceed the groups maximum and the proportional chance value, the model is said to be significant better than chance. This statistic can be defined as follows:

- Maximum Chance Criterion

  MCC=$(\boldsymbol{N_I}/\boldsymbol{N_L})(100)$ 3.6

  Where

  $\boldsymbol{N_I}$=number of subjects in the largest group

  $\boldsymbol{N_L}$=total number of subjects in the combined

- Proportional Chance Criterion

  C pro=$\sum\boldsymbol{P_j}^{\mathbf{2}}$ 3.7

  Where

  $\boldsymbol{P_j}$ = Proportional of subjects in each group.

- Press Q statistic

  $\boldsymbol{Q}=[\boldsymbol{N}-(\boldsymbol{n})(\boldsymbol{g})]^{\mathbf{2}}\,/\,[\boldsymbol{N}-(\boldsymbol{g}-\mathbf{1})]$ 3.8

  $\boldsymbol{N}$= total number of subjects

  $\boldsymbol{n}$ = number of cases correctly classified

  $\boldsymbol{g}$ =number of groups

  $\boldsymbol{Q}\sim\boldsymbol{x^2}_{\boldsymbol{g-1}}$

  The value of Q is compared to the chi-square distribution at g-1 degree of freedom since Q is approximately the chi-square value and if $\boldsymbol{Q}<x^2_{\boldsymbol{g-1}}$, reject the null hypothesis that the model hit ratio is not significantly better than chance, otherwise accept.

  In constructing the classification matrix, the cutting score was used. For equal groups, it is half way between the two groups centroid. This is defined as:

$$\boldsymbol{Zcs}=\boldsymbol{N_AZ_B}+\boldsymbol{N_BZ_A}/\boldsymbol{N_A}+\boldsymbol{N_B} \qquad \mathbf{3.9}$$

$\boldsymbol{W}$here

Zcs= Optimum cutting score between group A and B

$\boldsymbol{N_A}$= Number of observation in group A

$\boldsymbol{N_B}$= Number of observation in group B

$\boldsymbol{Z_A}$= Centroid for group A

$\boldsymbol{Z_B}$= Centroid for group B

$$\boldsymbol{Zcs = Z_A + Z_B/2} \qquad 3.10$$

$\boldsymbol{W}$here

$\boldsymbol{Zcs}$ = Optimal cutting score for equal size

$\boldsymbol{Z_A}$ = Centroid for group A

$\boldsymbol{Z_B}$ = Centroid for group B

In order to determine the significant of the predictors, the differences among the groups are tested. In this case we will test the following hypothesis.

$\boldsymbol{H_0}: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5.$

$\boldsymbol{H_1}: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \boldsymbol{\mu_5}.$

Where

$\boldsymbol{\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5}$ are the population means of group 1,2,3,4, and 5 respectively.

The above hypothesis can be tested using the wilk's lambda test statistic which may be defined as:

$$\boldsymbol{\lambda = WSS/TSS} = [\sum(\overline{\boldsymbol{Z_j}} - \overline{\boldsymbol{Z}})^2 / \sum(\boldsymbol{Z_{1j}} - \overline{\boldsymbol{Z_j}})^2] \qquad 3.11$$

Wilks' lambda ranges from 0 to 1. A small value indicates strong group differences. Values close to 1 indicates no group differences. Also the smaller the lambda for an independent variable, the more the variable contributes to the discriminant function.

The F-test of Wilks' lambda shows which variables contributions are significant. The F-test can be defined as:

$$\boldsymbol{F} = [\boldsymbol{1 - (\lambda_k - 1)} \,/\, (\boldsymbol{\lambda_k}) \,/\, (\boldsymbol{\lambda_k - 1}) \,/\, (\boldsymbol{\lambda_k}) \,/\, [\boldsymbol{N - g - 1/g - 1}] \qquad 3.12$$

Where,

N = total sample size,

G = number of defended variables (group),

d.f (N-g-1) and (g-1)

If the significant value is small (< 0.05), this indicates that the variable contributes to group differences.

3. RESULTS AND DISCUSSION

The result obtained from the analysis for the four discriminant functions is shown in table 3.1 below. The coefficients of the discriminant functions provide an index of the importance of each predictor while the sign indicates the direction of the relationship in the first function which was significant at ($p< 0.05$), N was the strongest predictor while pH was next importance as predictor. Both predictors have positive relationship with the soil types. For the remaining three functions. K, N and pH are strong predictors for the second function while K and N were the only strong predictors for the third and fourth functions respectively. The implication of this is that these variable (N,pH,k) with large coefficients stand out as

those that strongly predict allocation to the soil types or groups and that the score of the other soil characteristics were less successful predictors.

Table 3.2 shows the following statistic, the Eigen values, Canonical correlation, Wilks' lambda, its Chi-square and the associated p-value (sig). These statistics help to describe the discriminating abilities of the function. Four functions (table 3.1) were estimated, and only one function with the higher Eigen-value (1.315), which explain the variation of the soil group types by 55.6% (table 3.2) was significant and feasible for further used. The estimated Canonical correlation to measure the association between the groups and the disceriminant score is 0.754, this shows that the model discriminates well among the five groups. The wilks' lambda is 0.178, which implies that there is significant variation in the soil types grouped in the discriminant model. Approximately, 55.6% of the variation is explained by the differences between the groups (1,2,3,4, and 5). The Chi-square statistic have the value (82.503) and significant at ($p< 0.05$) indicating that, there is significant difference between group centroids. From the wilks' lambda and Chi-square results, we reject the null hypothesis of the equal mean between the groups. The estimated Press's Q statistic is 8.96 and this greater than chance value, implying higher power of the discriminant model is established for grouping the soil types in the study area, hence more credit to the model for further analysis.

Table 3.1: The Linear Discriminant Function Coefficients for the Soil Groups.

Variables	Z_1	Z_2	$\boldsymbol{Z_3}$	$\boldsymbol{Z_4}$
Constant	-8.403	4.949	-5.567	-1.930
Ph	1.324	-1.104	0.608	0.399
Orgm	-0.032	-0.147	0.023	-0.193
N	3.386	1.029	0.382	-2.702
P	0.052	-0.058	0.177	0.005
K	0.546	-2.083	-2.174	0.897
Ca	0.0232	0.600	0.123	-0.632
Mg	0.324	0.449	-0.550	-0.632
S	-0.062	0.023	0.009	0.046
Fe	0.003	0.008	0.003	0.001
Mn	0.024	-0.005	0.009	0.018
Cu	0.199	-0.195	0.072	0.052
Zn	0.005	0.049	0.035	-0.003
B	0.625	0.090	1.084	0.689

Table 3.2: Test of Significant for Discriminant Function.

Function	Eigen-values	Percentage of Variance	Canonical Correlation	Wilks' Lambda	Chi-square	Significant value
1	1.315	55.6	0.754	0.178	82.503	0.004*
2	0.491	20.7	0.574	0.415	42.211	0.221
3	0.433	18.3	0.550	0.619	23.039	0.400
4	0.128	5.4	0.336	0.887	5.767	0.834

Table 3.3 shows a comparison of the actual groupings of the five soil types to the predicted groupings generated by the discriminant functions. In the group 1 and 5, 75% of the cases were correctly classified, indicating a very good performance of the discriminant functions, while groups 2 and 4 predictions were barely less than 60%. The percentage of the cases correctly classified (hit ratio) was 67.7% while about 33% of the cases were classified wrongly. This result compare favorable with 67% recorded by (Fincher and Marie-Louise Smith in 1993 [7]) and 72% by (Stancy cambell, 2004 [8]) when discriminant predictive method is used for soil group membership prediction. But when compare with the results from other areas of the applications such as education and social sciences, the result could not be said to be remarkable. The reason for this could be due to insufficient variations in soil characteristics between groups to distinguish one source from the other. Another problem might be the number of groups. As (Stancy, 2004) suggested, the larger the number of groups, the chance of misclassification.

The predictive accuracy of the model was measure by the above values (Table 3.4). The overall hit ratio was 67.7% and this exceeded both maximum chance and proportional chance values of 27.59 and 21.22 percent respectively. The Press's Q statistic of 347.57 was greater than the tabulated value (11.1), indicating that the prediction accuracy is greater than that expected by chance.

Table 3.3: Classification performance of the estimated Discriminant Function

Actual Group	Number of Cases	Predicted	Group	Memberships		
		1	2	3	4	5
1	16	12(75%)	2(12.5%)	0(0%)	0(0%)	2(12.5%)
2	12	1(8.3%)	7(58.3%)	2(16.7%)	0(0%)	2(16.7%)
3	13	0(0%)	1(7.7%)	9(64.2%)	1(7.7%)	2(15.4%)
4	9	1(11.1%)	1(11.1%)	1(11.1%)	5(56.8%)	1(11.1%)
5	8	0(0%)	0(0%)	2(25.0%)	0(0%)	6(75.0%)

67.7% of the cases was correctly classified.

32.3% of the cases not correctly classified.

Table 3.4: Comparison of goodness of results.

Measure	Value
Maximum Chance	27.59
Proportional Chance	21.22
Hit Ratio	67.7
Press' Q calculated value	347.57
Press' Q table value	11.1

Table 3.5 shown below contains Wilks' lambda, the F statistics and its significance level. From the table, seven (7) predictor variables (pH, P, K, Ca, Mg, Mn and Cu) were significant of the thirteen (13) variables used in the model. The result shown that Mg is the best in discriminant among the groups, and closely followed by pH, P, Ca, Mn, Cu and finally K. The implication is that soil types are determined by these seven variables, and the inter group difference among the remaining six variables are not wide enough in discriminating among the groups. This means that the Orgm, N, S, Fe, Zn and B within the groups in the study areas are almost the same.

Table 3.5: Characteristics of soil locations that best discriminate among the five soil parents material.

Variables	Group 1 Mean	Group 2 Mean	Group 3 Mean	Group 4 Mean	Group 5 Mean	Wilks' Lambda	F-value	P-value
pH	5.7812	5.2250	5.0538	5.0111	4.9000	0.802	3.268	0.018*
Orgm	3.2200	3.5850	2.8092	5.8366	4.7512	0.912	1.278	0.290
N	0.1695	0.1605	0.1261	0.2223	0.2024	0.934	0.937	0.450
P	5.3750	8.4167	8.9231	14.000	5.5000	0.785	3.628	0.011*
K	0.3000	0.1458	0.1315	0.1856	0.0975	0.843	2.465	0.056*
Ca	2.4037	2.1916	0.4961	1.1167	0.2938	0.810	3.117	0.022*
Mg	1.6968	0.8292	0.3015	0.6989	0.2238	0.596	8.986	0.000**
S	11.0312	14.7750	16.8077	17.5000	16.2125	0.890	1.641	0.178
Fe	77.5625	133.500	75.4615	68.7778	79.000	0.907	1.364	0.259
Mn	26.7188	14.5333	8.5692	9.8111	3.7125	0.810	3.102	0.023*
Cu	4.3438	1.8917	1.5231	3.5777	0.5375	0.787	3.590	0.012*
Zn	8.4000	10.0500	3.7846	7.7333	2.5500	0.913	1.265	0.295
B	0.6150	0.4975	0.5046	0.5422	0.1938	0.937	0.895	0.474

*Significant ($p< 0.05$)

From the analysis, the following findings and conclusion were drawn. Discriminant analysis completely reduced thirteen (13) physical characteristics used in the study to seven (7) component indicating that fewer variables such as Mg, P, Cu, pH, Ca, Mn, and K are sufficient to explain the nutrient requirement of the oil palm within the belt. This shown that with fewer soil properties, we can sufficiently describe the nutrient required and information about the relative importance of each of the soil properties in characterizing the soil. The cluster analysis completely grouped the 57 agricultural fields into five groups of similar pattern indicating that the diversity of soil within the belt can be grouped into five according to the soil properties.

Four Discriminant functions were extracted in the analysis but only the first function was significant, of the variance explained by the four functions, 55.6% was explained by the first function while 20.7%, 18.3% and 5.4% were explained by the second, third, and fourth functions respectively. The canonical correlation indicate a very strong correlation between the discriminant score and the groups using the first function while the correlation between the other three functions were weak. The wilks' lambda for the first function was significant while the overall hit ratio of the model is 67.7% and this was shown to be significantly better than chance. The test for the equality of group means shown that seven out of the thirteen predictor variables were significant, indicating that these variables will be useful in classifying new observations drown from the same populations as the original sample groups.

4. CONCLUSION

Finally, the result has shown that the discriminant analysis is a fairly good method for predicting new cases into any of the five soil types in the Raphia growing zone of southern Nigeria and that Mg, P, Cu, pH, Ca, Mn, and K were identified as the soil properties that best discriminates among the soil types in the zone studied.

REFERENCES

1. 1. Berg. R.C 1980 Use of stepwise discriminant analysis to assess soil genesis in a youthful standy environment. Soil science, Vol. 129:353-365.
2. Edmonds and Lantner, 1987. Alliance of crop, soil and environmental science societies.
3. Fincher, J. and Marie-Louise, S 1994. A Discriminant Approach to Ecological Site Classification in Northern New English, a publication of the United States Department of Agriculture.

4. Lentz and Simonson, 1987 *Journal of Soil Properties and land types classification and identification with Discriminant Analysis*.
5. Norris J.M 1970 Multivariate Methods in the study of soil, soil and fertiliz. 33:313.318
6. Paramananthan, S. 2003. Land selection for oil palm. Chapter of the book, Oil palm Management for large and sustainable yields. *Publication of International Potash Institute*.
7. Stancy Cambel (2004) Discriminant analysis of heavy metal concentration in the soils of St. John, New Found Land.
8. Tabacchenick and Fidell, 1989. Using Multivariate Statistics, New York, Harper and Row.