AAB BIOFLUX

Advances in Agriculture & Botanics-
International Journal of the Bioflux Society

Agri-systems variations determined through Principal Component Analysis

Dennis A. Apuan, Mary A. T. Mercurio

Department of Agricultural Sciences, College of Agriculture, Xavier University, Cagayan de Oro City, Philippines. Corresponding author: D. A. Apuan, dennis_apuan@yahoo.com

Abstract. The present study tested the capability of Principal Component Analysis (PCA) in determining variations of the agri-systems landscape using ten qualitative characters *viz.* levels of nitrogen, phosphorus, potassium and pH, as well topography, vegetation type, presence of rocks, and characters of color such as hue, chroma and value. The 60 hectare farm in Manresa was used as model where assessment of variations using PCA was done. Exactly 103 random samples were collected from 11 sites within the model farm, and soils were analyzed and characterized using numerical coding technique. There were 1030 coded data generated and PCA was implemented on these data using PAST (Paleontological Statistics) software version 1.78. The similarity of sites was tested using cluster analysis and validation of results was made using Discriminant Function Analysis using SPSS ver. 17. Here, we found that PCA effectively deciphered variation existing in the agri-systems landscape, and dominant factors contributing to such variations identified. Practical applications of the method in agriculture is discussed.

Key Words: Agriculture, agri-system, landscape variation, Principal Component Analysis, cluster analysis.

Introduction. Precision agriculture requires that variations in the farm landscape be determined for site specific treatment and differential fertilizer application. This can be done easily when only one variable is considered, but the farm situation is complex. In nutrient requirement alone, a single plant requires sixteen essential nutrients, which variation over the landscape must be determined. Physical variations such as slope, soil color and the presence/absence of rocks may vary over the landscape, and more so on biological variables which interact with the physical environment. Clearly, a very complex situation in the farm exist that integrative and objective method is needed to summarize information’s altogether and variations determined and measured, including the identification of dominant factors.

The present study explored the use of Principal Component Analysis (PCA) to address this problem using ten qualitative variables construed as the characters, each with different levels or categories called character states. The problem on statistical operations concerning the use of qualitative variables has recently been resolved by Apuan (2011) through categorical data transformation method.

With the application of categorical data transformation, information’s obtained from qualitative measurements, can be processed statistically using non-parametric test. PCA is in fact one of those that used the non-parametric statistical principles. The concept of non-parametric test is popular in the field of biology. Somehow, the application was extended to the field of soil science in the second half of 20th century. Sarkar et al (1966) enumerated those characteristics for inclusion in numerical taxonomy of soils, Rayner (1966) and Grigal & Arneman (1969) used numerical classification of soils in forest areas. Cipra et al (1970) used such method with the use of Centroid Component Analysis (CCA). Goodall (1954) in his ecological studies pioneered the use of factor analysis – a non-parametric and multivariate technique.

The current study explored the use of Principal Component Analysis (PCA) - a new mathematical tool capable of detecting variations and classification of groups based on components. This tool is very much related to Centroid Component Analysis (CCA), used by Cipra et al (1970).

The general objective was to assess variations of the farm landscape using Principal Component Analysis. Specifically, it aims to determine the amount and patterns of farm landscape variations, and to identify dominant contributing factor to such variation.

Material and Method. Manresa farm, a 60 ha property of Xavier University in Cagayan de Oro, was selected as a model landscape in the study because of its varying and sharp contrasting topography, and each was treated differently. The relatively flat landscape also had different divisions based on the history of use and treatments. There were a total of 11 divisions of this agricultural landscape where random sampling was done.

A total of 103 soil samples were obtained from all these divisions at an effective depth of 10 inches, and samples were air dried for soil analysis in the laboratory. Data on qualitative variables (Table 1) were obtained. Soils were characterized based on pH, nitrogen, phosphorus and potassium using the “Soil Test Kit”. Readings were in terms of magnitude and were categorized as low, medium and high. A numerical coding scheme was used similar to the scoring method of Dixon et al (2005) in their TRARC (Tropical Rapid Appraisal of Riparian Condition) program and Jansen et al (2006) in their tool known as RARC (Rapid Appraisal of Riparian Condition). Each category of the variable was given a code, for example, a code of 1 was being denoted as “low”, followed by “medium” given a code of 2 and “high” a numerical code of 3. Although the numerical codes do not represent magnitude, but the code assignment follows a logical pattern from less desirable character state to a more desirable character state. This facilitates understanding and analysis when all samples were projected in a scatter plot. The same coding scheme was also applied to other qualitative soil variable, which are summarized in Table 1.

Character and numerical character coded states of the soil samples were then entered into the matrix of Paleontological Statistics (PAST) version 1.78 developed by Hammer et al (2001). Using the PAST software as the platform, Principal Component Analysis was then used based on correlation matrix to explore site variation in the landscape described by the principal axis. These axes were selected based on the amount of loading which also corresponds to the amount of variation in sites that it can explain.

On the other hand, similarity of sites was measured using Cluster Analysis following algorithm of the Unweighted Pair Group Method Average (UPGMA) based on Spearman’s rho correlation matrix of soil characters. Output of such analysis was a Dendogram showing clusters of sites with similar characteristics and clusters that differ among each other.

Finally, Discriminant Function Analysis (DFA) was implemented using Statistical Package for Social Sciences (SPSS) version 17, to check the classification of sample groups.

Table 1

Numerical coding scheme of different categories of qualitative variables

Code	*N*	*P*	*K*	*pH*	*Hue*	*Value*	*Chroma*	*Topography*	*Vegetation*	*+/- rocks*
1	Low	Low	Low	Low	7.5R	1	1	Hilly	Grassland	Absent
2	Medium	Medium	Medium	Medium	10R	2	2	Moderately rolling	Forest	Present
3	High	High	High	High	2.5YR	3	3	Slightly rolling	Fruit trees	
4					5YR	4	4	Flat	Cropped area	
5					7.5R	5	5			
6					10YR	6	6			
7					2.5YR	7	7			
8					5Y	8	8			
9					7.5Y					
10					10Y					

Results and Discussion. Subjecting the 103 randomly taken samples of Manresa Agri-system to Principal Component Analysis revealed the variations in the Bio-Physico-Chemical components of the agri-systems (Figure 1). From among the ten dimensions (character combinations) considered in the correlation data matrix, four came out to have large contributions to variations.

The four dimensions or Principal Components showing varying levels of eigenvalues and degrees of variance is presented in Table 2. Principal Component 1 (PC1) has the largest eigenvalue, and can explain 28.27 % of the total variance in the samples. This is followed by PC 2, which can explain 17.57 % of the variance, PC 3 which accounts for 12.195 %, and PC 4 with 10.984 % having the lowest value.

Among the ten variables (characters) in Figure 2, it turned out that character like vegetation, topography and presence/absence of rocks being part of the Bio-Physical components of the Agri-systems were highly correlated to PC 1. With the scatter plot in Figure 1, the evident separation of samples from forest, pomegranate and orchard-hilly sites are clearly depicted with 2.82698 eigenvalue. The site of orchard hilly was different to forest site in terms of topography - the former was moderately rolling and the later was hilly. The two also differ in terms of vegetation type.

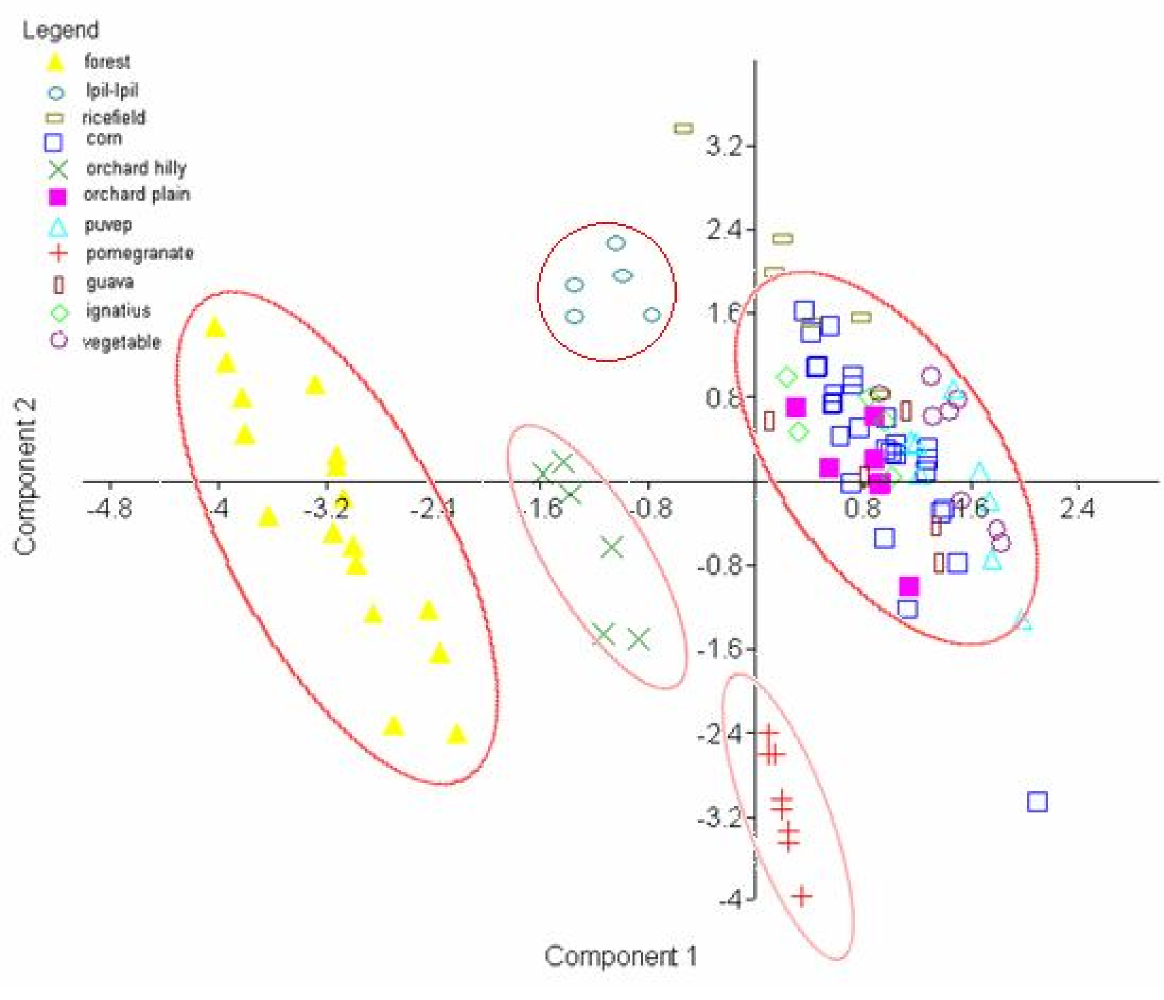

Figure 1. Variations in bio-physico-chemical components of Manresa Agri-Systems Model Using Principal Component Analysis (PC 1 & 2).

Table 2

Variance contribution from the four principal components of 103 samples from 11 sampling sites

Principal Components	*Eigenvalue*	*% Variance*	*Cumulative Variance*
1	2.82698	28.27	28.27
2	1.75696	17.57	45.84
3	1.21954	12.195	58.035
4	1.09844	10.984	69.019

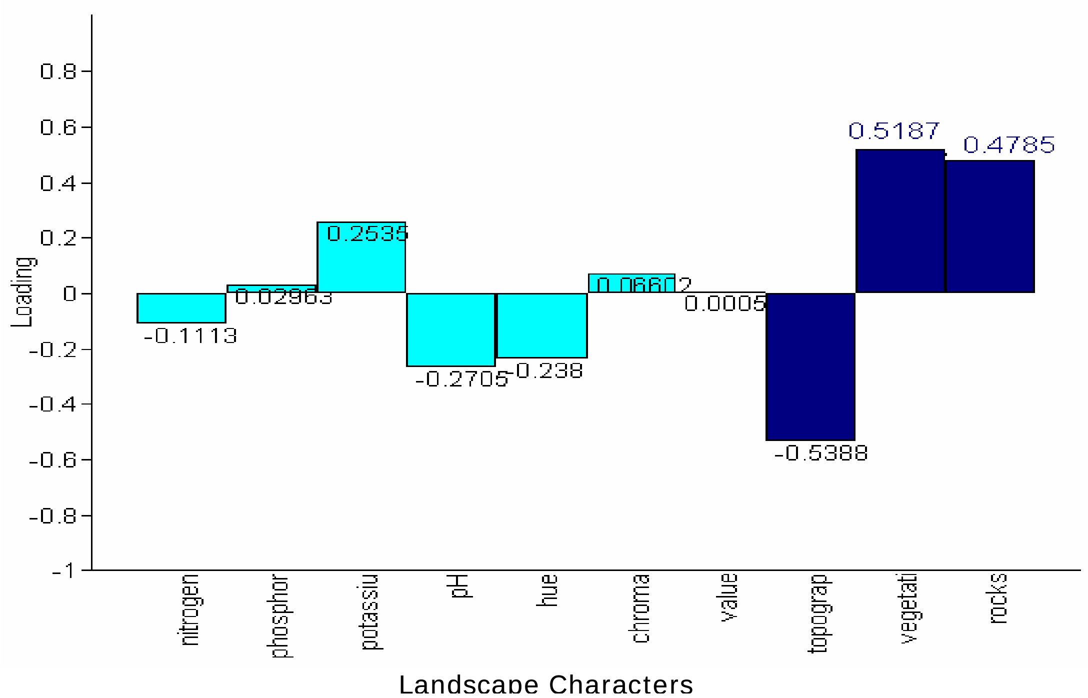

Figure 2. Component loading for Principal Component 1 with highly correlated characters highlighted.

The clustering of crop cereals and crop vegetable samples on the positive side of the PC 1 axis depicts uniformity in terms of topography, absence of rocks and type of vegetation. These sites include the corn field, rice field, a portion of the orchard, and the vegetable gardens of the different departments *viz.* Crop Science, St. Ignatius, PuVep, and the Guava Area.

Principal Component 2 (PC 2) had large loadings on the character nitrogen, and characters related to soil color such as hue and chroma (Figure 3). The high loading of these characters on PC 2 was based on their high correlations on this component.

The observed spreading of all samples along the axis of PC 2 was attributed to variations in the levels of nitrogen in the soil, as well as on chroma (purity of soil color). It was on this Principal Component that within site variation was evident. Within the forest site for example, there were 7 out of 17 specific points where nitrogen was low, and 7 specific points where nitrogen was medium and only 3 where nitrogen was high. Similar pattern was observed in the vegetable sites, corn field, rice field, Ipil-ipil and orchard site. This pattern also holds true to chroma character.

Principal Component 3 in Figure 4 was related to chemical properties of the agri-system since phosphorus, potassium and pH have relatively large loading in this dimension of data as shown in Figure 5. Among those three characters however, phosphorus was the most influential since it had the highest loading on PC 3 axis. The obvious separation of pomegranate and ipil-ipil sites from orchard hilly and corn sites in

Figure 5 were based on the levels of phosphorus element. Pomegranate and orchard hilly were high in phosphorus while the later groups were low in this element.

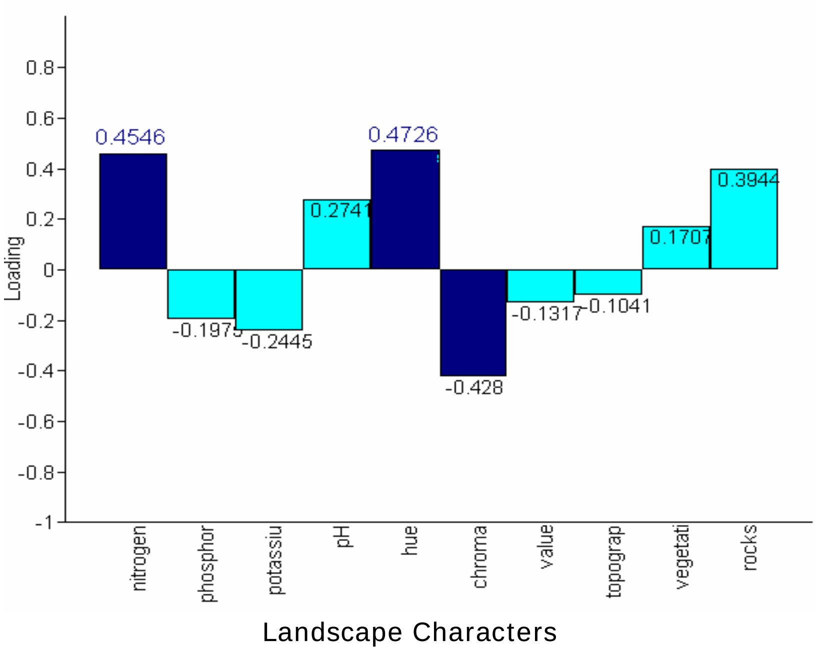

Figure 3. Component loading for Principal Component 2 with highly correlated characters highlighted.

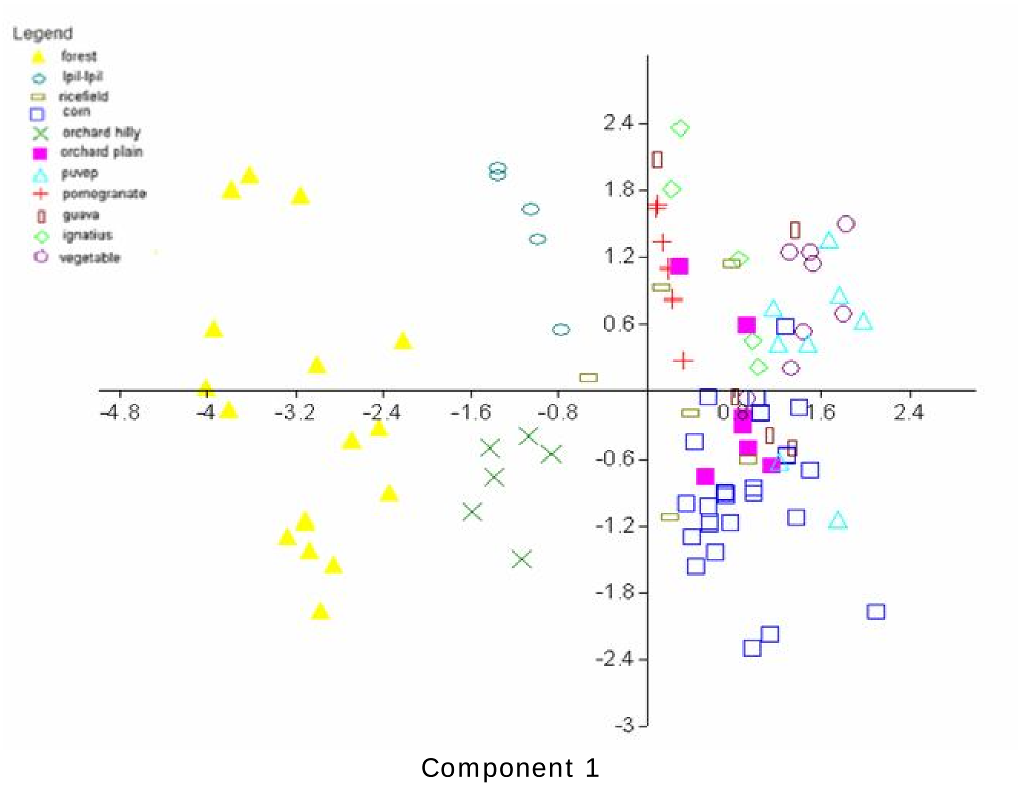

Figure 4. Variations in the chemical component of the agri-systems model along PC 3 axis.

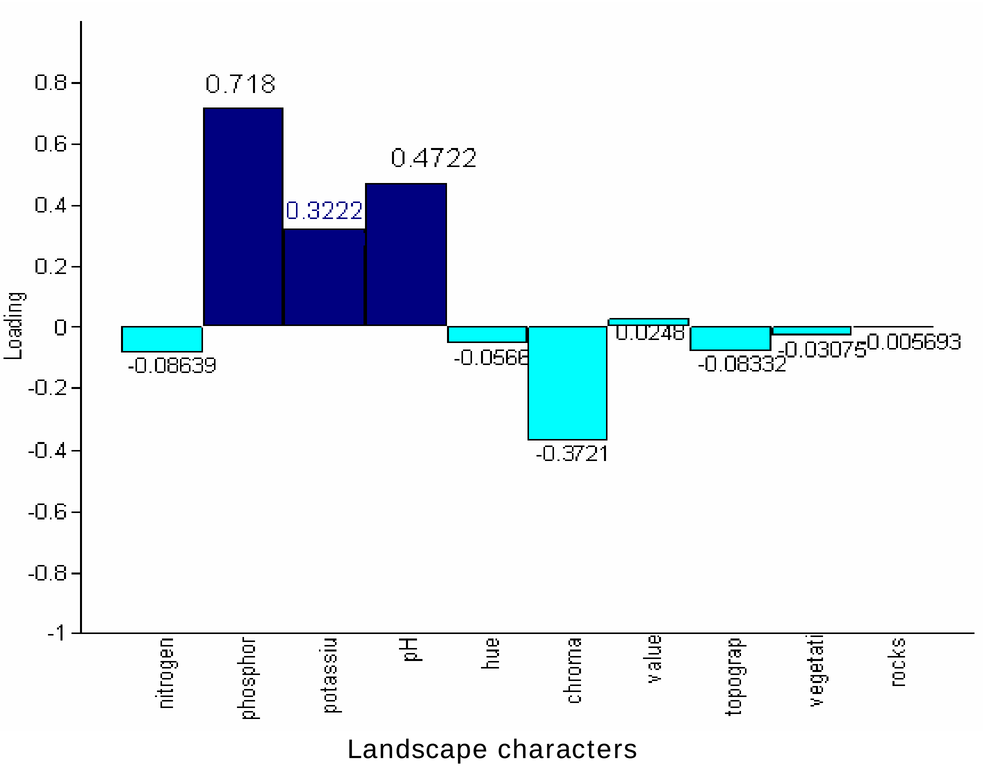

Figure 5. Component loading for Principal Component 3 with highly correlated characters highlighted.

Principal Component 4 was related to the intensity of soil color (value) as shown in Figure 6. It should be noted that this dimension of the data accounts only 10.98 % of the total variance, but it successfully revealed that even within every site soils vary.

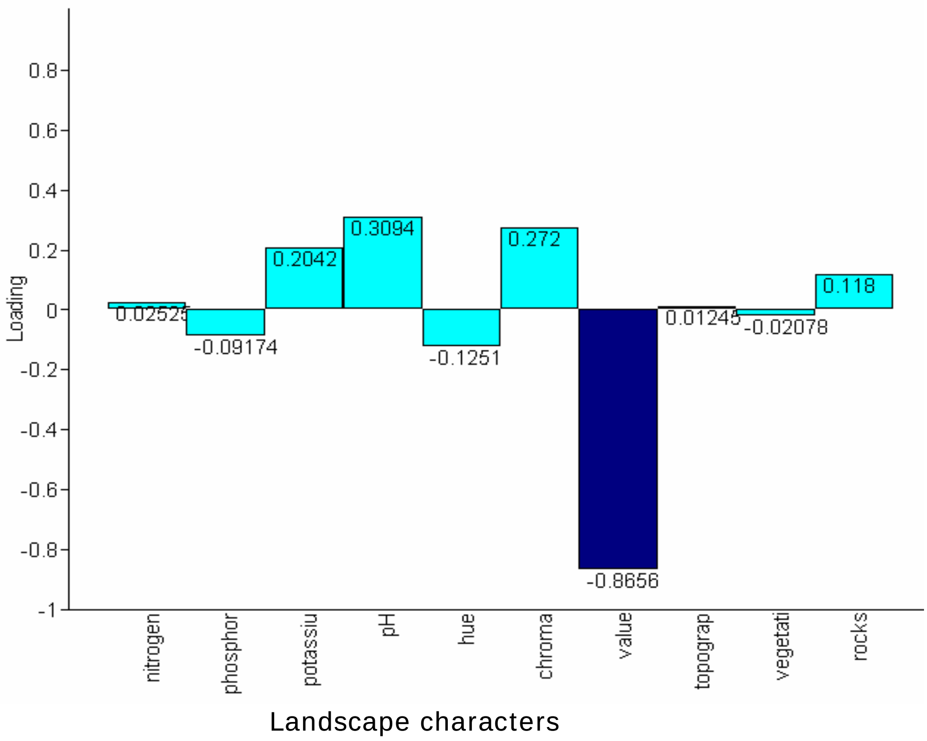

Figure 6. Component loading for Principal Component 4 with highly correlated characters highlighted.

Results of cluster analysis projected in Figure 7 revealed five major groups of sites that differ in characteristics. The grouping pattern observed was similar in Figure 1. It then becomes obvious that the clustering of samples was highly influence by PC 1.

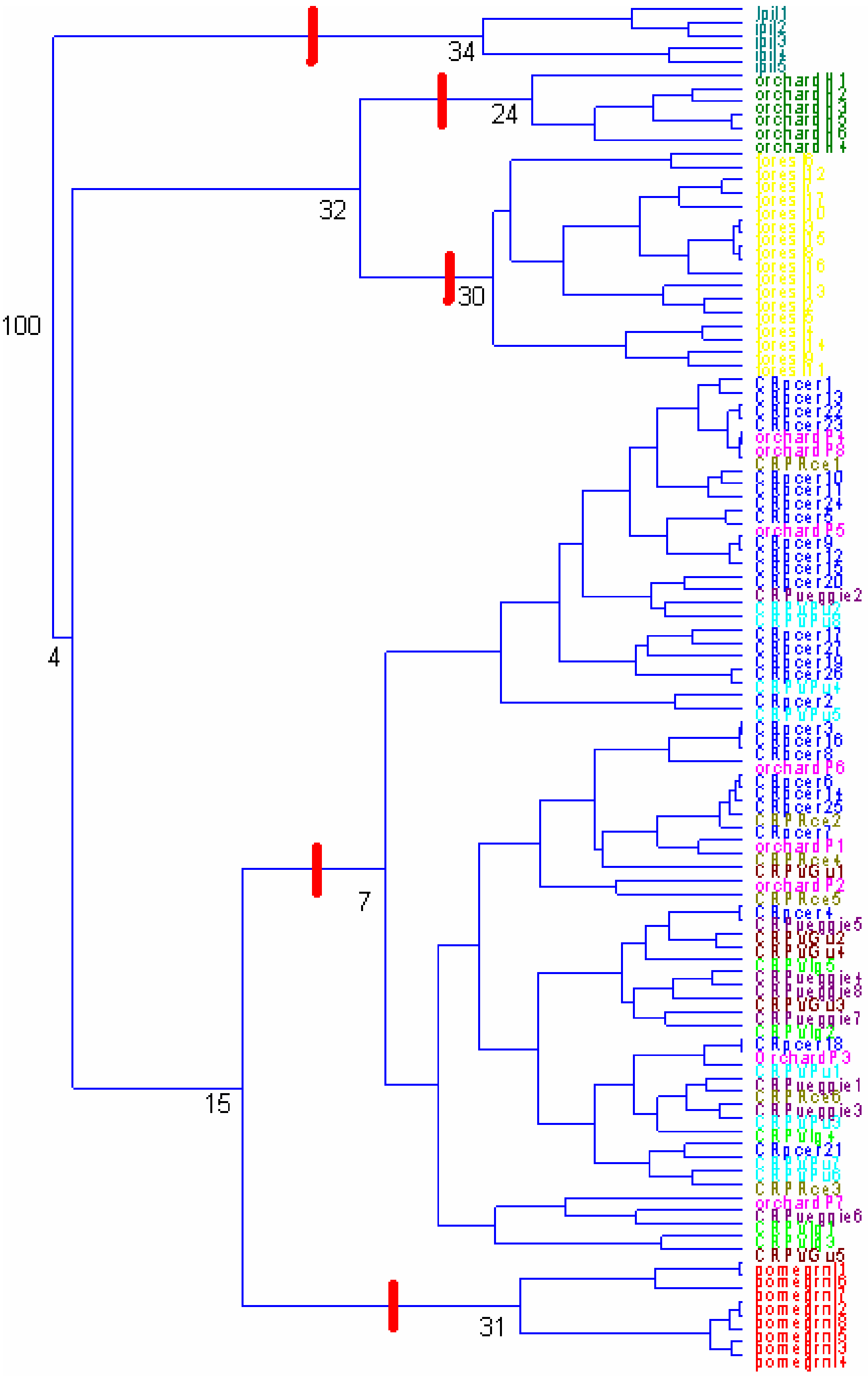

Figure 7. Dendogram showing cluster of 103 random samples collected from 11 sites in the 60 hectare model farm after cluster analysis.

Dendogram in Figure 7 shows that samples from Ipil-ipil site formed a single group – although supported only by a bootstrap value of 34, but its similarity index among 5 samples within this site was 75 %. However, in comparison with the other sites, its similarity was only 32 %. Orchard hilly and forest sites were 62 % similar since they both have rocky soils and sloping topography, but they split further due to differences in vegetation type and formed two groups with respective similarity index of 80 % and 78 %. Both sites were weakly supported by bootstrap value of 24 and 30 respectively. All samples taken from sites planted with vegetables, cereals, including orchard with flat topography and site for pomegranate formed a huge cluster weakly supported by bootstrap value of 15, and similarity of samples within these sites was only 50 %. The proposed pomegranate site split further by having rocks on its soil and from having different vegetation type (pasture grasses). Samples within this site were 78 % similar and supported by a bootstrap value of 31.

Discriminant function analysis however, found only two major groups in which its separation was primarily influenced by characters from Principal Component 1 such as topography, rocks and vegetation type. Figure 8 projects this grouping with the yellow bars representing all samples from the forest site, and notably separated from the rest of the samples. This result of DFA practically discriminate the small groups in cluster analysis, and this is logically reasonable since the observed groupings in CA are not supported highly by bootstrap values. This suggests that resolution of groups in CA is poor. Variations in the agri-systems model may have been resolve better if more characters were included. Nevertheless, it effectively demonstrates that PCA can determine variations of the agri-systems component.

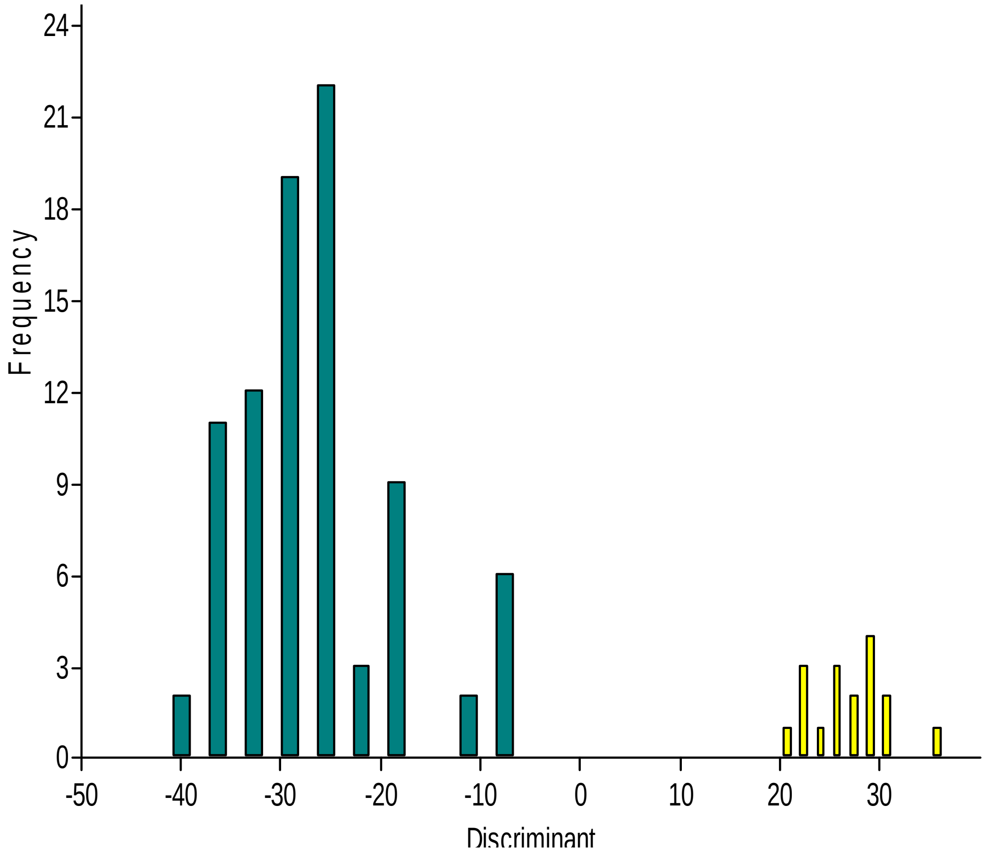

Figure 8. Results of Discriminant Function Analysis showing two major groups.

This pattern of results was similar to the pioneering works in numerical taxonomy by Cipra et al (1970). Here, we differ by introducing Principal Component Analysis as a statistical tool to decipher variations. The current study also differs to them in the sense that qualitative variables as characters were used.

Qualitative characters always generate categorical data, and thus the landscape data transformation method of Apuan (2011) was applied such that categories can be subjected to non-parametric test (Field 2005).

In data transformation method (Apuan 2011), numerical codes were assigned to character states construed here as the categories. This is important in the sense that these coding are used in ranking of samples from lowest to highest, to which the calculation of median scores for sample comparison was based. The determination of the Spearman's product-moment correlation (Spearman's rho correlation) was also based on the ranks.

Determining variations through the Principal Component Analysis as shown in Figure 1 and 4, was in fact based on the Spearman's rho correlation matrix. The same data set in the matrix was used in the cluster analysis and discriminant function analysis.

Conclusions. The study demonstrated that Principal Component Analysis (PCA) shows degree of capability in determining variations in the landscape of the Agri-systems model. With PCA analysis, the evident separation of samples from different sampling sites can be visualized in a scatter plot. Cluster analysis and discriminant function analyses were able to objectively summarize the overall results by categorizing it into groups. The method may have potential use in agriculture for site specific interventions.

References

Apuan D. A., 2011 Landscape data transformation: categorical descriptions to quantitative descriptors. World Academy of Science, Engineering and Technology 81:143-146.

Cipra J. E., Bidwell O. W., Rolf F. J., 1970 Numerical taxonomy of soils from nine orders by Cluster and Centroid Analyses. Soil Sci Soc Am J 34(2):281-287.

Dixon I. H., Douglas M. M., Dowe J. L., Burrows D. W., Towsend S. S., 2005 A rapid method for assessing the condition of riparian zones in the wet/dry tropics of northern Australia. In: Proceedings of the 4th Autralin Stream Management Conference, Linking Rivers to landscapes. Rutherfurd I. D., Wiszniewski M. A., Askey-Doran R., Glazik R. L. (eds), pp. 173-178, Department of Primary Industries, Water and Environment, Hobart, Tazmania.

Field A., 2005 Discovering statistics using SPSS. Sage Publication, London, pp. 619-679.

Goodall D. W., 1954 Objective methods for the classification of vegetation. III. An essay in the use of factor analysis. Aust J Bot 2(3):304–324.

Grigal D. F., Arneman H. F., 1969 Numerical classification of some forested Minnesota soils. Soil Sci Soc Am J 33:433-438.

Hammer O., Harper D. A. T., Ryan P. D., 2001 PAST: Paleontological Statistics for Educational and Data Analysis. Paleontologia Electronica 4(1):1-9.

Jansen A., Robertson A., Thompson L., Wilson A., Nicholls K., 2006 Rapid appraisal of riparian ondition. In: Technical guide for the mid north of South Australia. Land Water and Wool pp. 1-17.

Rayner J. H., 1966 Classification of soils by numerical taxonomy. Journal of Soil Science 17:79-92.

Sarkar P. K., Bidwell O. W., Marcus L. F., 1966 Selection of characteristics for numerical classification of soils. Soil Sci Soc Am J 30:269-272.

Received: 27 December 2013. Accepted: 28 January 2014. Published online: 24 February 2014.

Authors:
Dennis Alalim Apuan, Xavier University, College of Agriculture, Department of Agricultural Sciences, Philippines, Cagayan de Oro City, e-mail: dennis_apuan@yahoo.com
Mary Ann Taran Mercurio, Xavier University, College of Agriculture, Department of Agricultural Sciences, Philippines, Cagayan de Oro City, e-mail: annmercurio2006@yahoo.com.ph

How to cite this article:
Apuan D. A., Mercurio M. A. T., 2014 Agri-systems variations determined through Principal Component Analysis. AAB Bioflux 6(1):1-10.