International Journal of Biological, Life and Agricultural Sciences
ISSN: 2415-6612
Vol:5, No:9, 2011

Landscape Data Transformation: Categorical Descriptions to Numerical Descriptors

Dennis A. Apuan

***Abstract*—**Categorical data based on description of the agricultural landscape imposed some mathematical and analytical limitations. This problem however can be overcome by data transformation through coding scheme and the use of non-parametric multivariate approach. The present study describes data transformation from qualitative to numerical descriptors. In a collection of 103 random soil samples over a 60 hectare field, categorical data were obtained from the following variables: levels of nitrogen, phosphorus, potassium, pH, hue, chroma, value and data on topography, vegetation type, and the presence of rocks. Categorical data were coded, and Spearman's rho correlation was then calculated using PAST software ver. 1.78 in which Principal Component Analysis was based. Results revealed successful data transformation, generating 1030 quantitative descriptors. Visualization based on the new set of descriptors showed clear differences among sites, and amount of variation was successfully measured. Possible applications of data transformation are discussed.

***Keywords*—**data transformation, numerical descriptors, principal component analysis

I. Introduction

QUALITY descriptions during rapid appraisal of the landscape is common. Qualitative variables such topography has categories *e.g.* flat, slightly rolling, hilly and steep. A particular spot of the landscape can be described as having low, medium or high in phosphorus. As such, bio-chemical and physical variables in nature can be described qualitatively, but often encounters problems when data are used in classification and in delineating boundaries. This is not a problem when landscape is classified based on a single variable; complications only appear when all variables are integrated, which often be the case in landscape evaluation.

There are a number of limitations identified from using qualitative descriptions. First, the information does not yield itself to statistical testing; second, patterns of variations cannot be deciphered and measured objectively; third, there is difficulty in identifying factor contributing to large spatial variation; and lastly, hyper-variation may result if several qualitative variables are included in landscape classification; every variable added contributes to variations.

With the application of numerical coding technique, categorical data obtained from qualitative measurements, can be processed statistically using non-parametric test. These concepts of non-parametric test and numerical taxonomy are 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 enumerated those characteristics for inclusion in numerical taxonomy of soils [1]; Rayner and Grigal used numerical classification of soils in forest areas [2, 3]. Goodall in his ecological studies pioneered the use of factor analysis – a non-parametric and multivariate technique [4].

The current study deals with transformation of categorical data generated from qualitative measurements into quantitative descriptors. The technique involves numerical coding of categorical data similar to dummy variable described by Field [5]. The quantitative descriptors are a new set of data which are generated based on Wilcoxon's ranking and the use of least squares method. Through a non-parametric multivariate test known as Principal Component Analysis (PCA), patterns of variations can be observed and measured using the new numerical descriptors.

The study specifically shows the numerical coding scheme applied, and the exportation of coded data to a platform of PAST (Paleontological Statistics) software version 1.78. Extraction of new descriptors and the implementation of PCA are through the use of this software, including estimates of variation and its pattern.

The applications of the technique in rapid appraisal of the ecosystems landscape and in the field of agriculture are stressed out; especially its possible utilization for site specific intervention.

II. Materials and Methods

The 60 hectare field of Manresa Research Station in Cagayan de Oro, Philippines was chosen as a sampling site due to the natural contrasting variation of the landscape, and variations caused by agricultural treatments and landuse. Eleven different sites were sampled where a total of 103 random soil samples were obtained.

The soil test kit was used to analyze the samples and obtained qualitative measurements on the following variables: amount of nitrogen, phosphorus, potassium and pH. The kit has limitations and can only give categorical readings such as low, medium and high. The color variables such as the chroma, value and hue were measured using the soil Munsell color chart. Other variables measured were the topography, presence or absence of rocks and kinds of vegetations. Categories within each of these variables were recorded during the field visits.

D. A. Apuan is with the Department of Agricultural Sciences, College of Agriculture, Xavier University, Cagayan de Oro City, Philippines. Telephone number (088) 858 3116 loc. 3100 (dennis_apuan@yahoo.com)

A numerical coding scheme was applied for each category of nutrient levels and pH, in which a code of 1 is 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 categories to a more desirable categories. This facilitates understanding and analysis when all samples are projected in a scatter plot. The same coding scheme was also applied to other qualitative soil variables, which are summarized in the table 1.

TABLE I
NUMERICAL CODING SCHEME OF DIFFERENT CATEGORIES OF QUALITATIVE VARIABLES

The categories of each variable presented in table 1 were construed as the different states of such variables. The coded states or categories of the soil samples were then entered into the matrix of Paleontological Statistics (PAST) version 1.78 developed by Hammer and Harper [6]. Using the PAST software as the platform, non-parametric multivariate analysis was performed, specifically the Principal Component Analysis (PCA) based on correlation matrix.

PCA formed a series of linear least square orthogonal axes, which are a combination of the original variables. The new quantitative descriptors of every sample, known as scores, are its coordinates relative to the axes which can be viewed and retrieved from the PCA score panel of the PAST software. These new set of data were then used as basis in visualizing their distribution along principal axes in the scatter diagram.

Using the same software, the amount of variations explained by the PC axes was estimated based on the distance of scores from these axes.

III. Results And Discussion

The numerical coding of all categorical data for 103 soil samples generated 1030 coded states. Results of the non-parametric multivariate analysis based on these codes revealed successful transformation of all categorical data into numerical descriptors. Ten selected samples out of 103 are shown here in table 2 and table 3. For example, sample number 1 from forest has high nitrogen level, while sample number 2 has low nitrogen. These were respectively transformed numerically into -4.204 and -3.7994 (table 3).

These new quantitative data are coordinate positions of the sample points as shown in figure 1. The distribution of sample points in the scatter plot projects a pattern of variation in the landscape influenced by a latent variable represented by PC 1 and PC 2 in figure 1.

Coordinates of all sample points, as a new data set that described numerically the state of the samples were used to measure amount of variation. The distance of samples points from the PC axis was a measure of variance explained by that axis. Table 4 gives variance contribution of the four Principal Components (PC 1 to PC 2). PC 1 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.

The uniqueness of the current study in areas of rapid appraisal lies on the method that transforms categorical data that enables one to decipher variations existing in the landscape, and provides a way to measure amount of variation. The pattern of variation observed in figure 1 could have been difficult to see when all categorical data are integrated for analysis. In natural conditions, samples overlap in one or more variables since environmental factors follow some gradients criss-crossing in any directions in space. For instance, sample 5 from ricefield and sample 1 from forest are separated in the scatter diagram (figure 1) but they have same conditions in terms of hue and levels of N, K and pH (table 2). In contrast, samples coming from cornfield and orchard plain are group together while they differ in 60% of the variables considered viz. levels of N, P, K and color variables like hue, value and chroma.

The mathematical and statistical limitations of categorical data are overcome through the assignment of codes serving as bridge to extract the numerical descriptors (tables 2 and 3). These new descriptors, often called scores, are coordinate points of the samples. Consequently, these descriptor coordinates, made the samples become separated or clustered along an axis as projected in the scatter diagram.

TABLE II
CATEGORICAL DATA SET OF TEN QUALITATIVE VARIABLES FROM TEN SELECTED SAMPLES SUBJECTED TO CODING AND TRANSFORMATION

Sites	Categorical Data of Original Variables									
	N	**P**	**K**	**pH**	**Hue**	**Value**	**Chroma**	**Topo**	**Vege**	**Rocks**
forest1	high	medium	low	high	10YR	4	2	hilly	forest	present
forest2	low	high	low	High	10YR	6	1	hilly	forest	present
pomegrnt2	low	high	medium	Medium	5YR	4	3	flat	crops	present
pomegrnt3	low	high	medium	Medium	5YR	4	4	flat	crops	present
corn1	medium	low	medium	medium	10YR	4	3	flat	crops	absent
corn2	low	medium	low	medium	10YR	3	4	flat	crops	absent
orchardP1	medium	low	medium	medium	10R	6	1	flat	crops	absent
orchardP2	medium	low	medium	medium	10R	2	1	flat	crops	absent
ricefield1	medium	low	medium	medium	10YR	5	2	flat	crops	absent
ricefield5	high	low	low	high	10YR	2	1	flat	crops	absent

TABLE III
NEW QUANTITATIVE DATA SET AFTER TRANSFORMATION OF CATEGORICAL DATA FROM ORIGINAL VARIABLES

Sites	axis 1	axis 2	axis 3	axis 4	axis 5	axis 6	axis 7	axis 8	axis 9	axis 10
forest1	-4.204	0.42042	0.41666	0.89925	0.71873	0.97505	-0.48135	0.68153	-0.61746	-0.15735
forest2	-3.7994	-1.1427	1.8914	-0.82287	0.58797	-0.98	0.68591	0.73464	-0.22435	-0.049954
pomegrnt2	1.4524	-2.5767	1.0504	0.83743	0.25719	0.93837	0.5524	-0.0053548	0.024096	0.54349
pomegrnt3	1.5444	-2.8726	0.79104	1.0362	0.63455	0.81121	0.15942	0.036371	-0.10345	0.52836
corn1	0.89203	0.80047	-0.89663	0.45374	-0.42589	-0.45172	-1.1586	-0.28522	-0.27612	0.27258
corn2	0.70191	0.059926	-0.88002	0.76882	1.6056	-1.4775	0.35837	-1.0328	-0.2946	0.13805
orchardP1	1.1992	0.55282	-0.20202	-1.0986	-1.0226	0.56095	-0.387	0.87142	0.32209	-0.0075208
orchardP2	1.2436	1.0987	-0.36964	1.587	-1.0692	0.60123	0.42607	-0.26977	0.5836	-0.22891
ricefield1	0.78885	0.9599	-0.59534	-0.41642	-0.79161	-0.33462	-0.96884	-0.041647	-0.21395	0.34306
ricefield5	-0.9143	3.2617	-0.049123	1.8746	0.65336	0.38795	0.51179	0.87608	-0.83295	0.2509

TABLE IV
VARIANCE CONTRIBUTION FROM THE FIVE PRINCIPAL COMPONENTS OF 100 SAMPLES FROM TEN 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

The use of codes is similar to the scoring method used by Dixon *et al*. in their TRARC (Tropical Rapid Appraisal of Riparian Condition) program [7]. They used the scoring technique to describe riparian conditions. The same scoring method was used by Jansen *et al*. as index describing riparian condition in their tool known as RARC (Rapid Appraisal of Riparian Condition) [8]. Site comparison was made based on sum of scores; sites with higher scores are Riparians with good conditions, while those with lower scores are in poor conditions. Further, comparisons were made based on frequency of sites with particular range of scores [8]. This is purely descriptive that could have been advanced further for inferential testing using non-parametric approach.The use of the scoring method advocated by Dixon *et al*. and Jansen *et al*., and the scores as codes for data transformations are foreseen to have bright potential for better resolution of sites with variations and the identification of sites with similarities [7, 8]. Furthermore, amount of variations can also be measured.In the agricultural landscape for crop production, data transformation could be useful in precision agriculture. A large span of land that seems uniform may have variation brought by past treatments and by differences in landuse. Determining and measuring variations through transformed data set are useful for site specific interventions and crop selections.

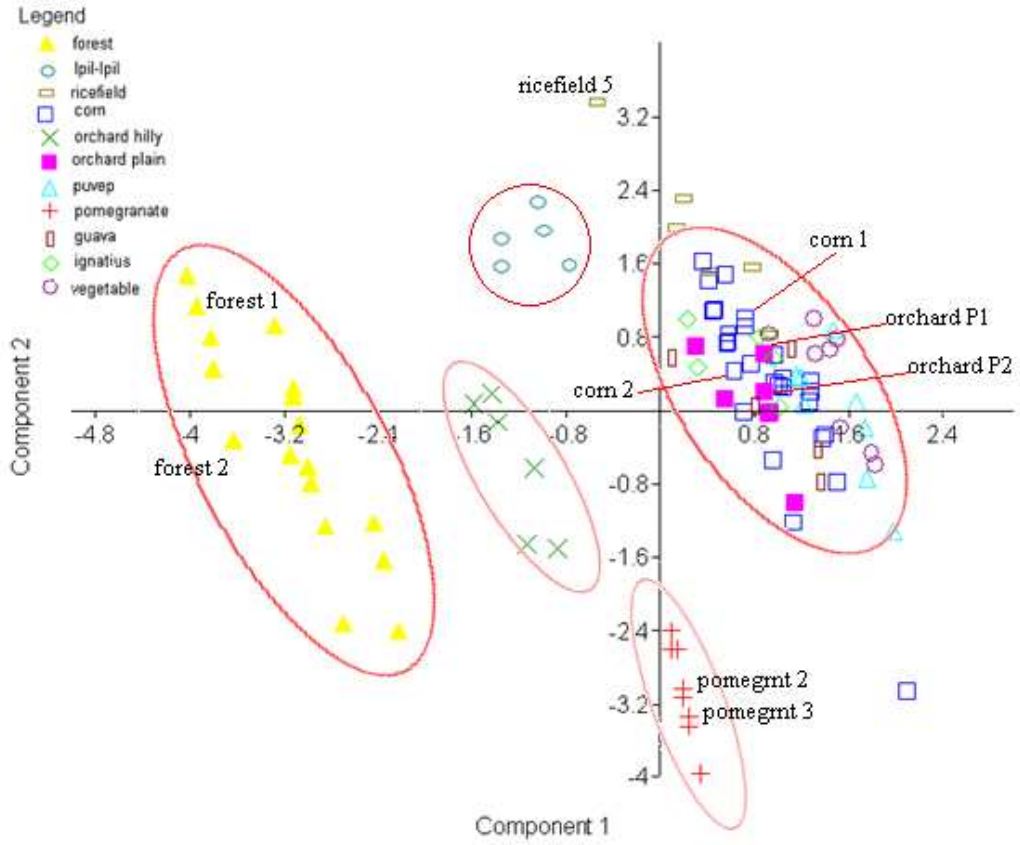

Fig. 1 Coordinate positions of the 103 sample points along the Principal Component 1 and Principal Component 2 axes

IV. Conclusion

Categorical data from qualitative descriptions and measurements were successfully transformed into numerical descriptors through assignment of codes and the use of non-parametric statistical approach. The use of PAST (Paleontological Statistics) ver 1.78 was useful in extracting the new data set, and in the implementation of Principal Component Analysis, in which visualization of sample point distribution and measurement of amount of variation was made.

Acknowledgment

I would like to thank my students who helped in data collection: Mr. Paul Geromg, Edouard Buyan and Jose Sacal, and my colleague at the crop science unit for their logistical support, Prof. Mary Ann Mercurio and Prof. Floro Dalapag.

References

[1] O. P. K. Sarkar, O. W. Bidwell, L. F. Marcus, "Selection of Characteristics for numerical classification of soils". Soil Science Society of America Proceedings, 30:269-272, 1966.
[2] J. H. Rayner, "Classification of Soils by Numerical Taxonomy" Journal of Soil Science, 17:79-92, 1966.
[3] D. F. Grigal, H. F. Arneman, "Numerical Classification of Some Forested Minnesota Soils", Soil Science Society of America Proceedings, 33: 433-438, 1969.
[4] D. W. Goodall, "Objective methods for the classification of vegetation. III. An essay in the use of factor analysis", Australian Journal of Botany, 2(3): 304 – 324, 1954.
[5] A. Field, "Discovering Statistics using SPSS", Sage Publication, London, pp. 619-679, 2005.
[6] O. Hammer, D. A. T. Harper, P. D. Ryan, "PAST: Paleontological Statistics for Educational and Data Analysis", Paleontologia Electronica 4 (1):9, 2001.
[7] I. H. Dixon, M. M. Douglas, J. L. Dowe, D. W. Burrows, S. A. Townsend, "A Rapid Method for Assessing the Condition of Riparian Zones in the Wet/Dry Tropics of Northern Australia", In: I. D. Rutherfurd, I. Wiszniewski, M. A. Askey-Doran, R. Glazik (eds.), Proceedings of the 4th Australian Stream Management Conference; Linking Rivers to Landscapes, Launceston, Tasmania, Department of Primary Industries, Water and Environment, Hobart, pp. 173-178, 2005.
[8] A. Jansen, A. Robertson, L. Thompson, A. Wilson, K. Nicholls, "Rapid Appraisal of Riparian Condition" In: Technical Guide for the Mid North of South Australia. Land, Water and Wool. pp. 1-17, 2006.