iEMSs 2008: International Congress on Environmental Modelling and Software
Integrating Sciences and Information Technology for Environmental Assessment and Decision Making
4th Biennial Meeting of iEMSs, http://www.iemss.org/iemss2008/index.php?n=Main.Proceedings
M. Sànchez-Marrè, J. Béjar, J. Comas, A. Rizzoli and G. Guariso (Eds.)
International Environmental Modelling and Software Society (iEMSs), 2008

Data mining using k-means clustering and classification and regression trees (CART) as post-processing methods: identifying management and environmental factors for explaining sugarcane yield in Northern Argentina (1971-2005)

D.O. Ferraro

Instituto de Investigaciones Fisiológicas y Ecológicas Vinculadas a la Agricultura (IFEVA), Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET) / Facultad de Agronomía, Universidad de Buenos Aires, Av. San Martín 4453, C1417DSE Buenos Aires, Argentina.

***Keywords*:** modelling; crop yield; agriculture; sugarcane; Argentina.

Final crop yield in an agroecosystem is determined by agronomic management and environmental factors interactions. Gains in understanding the magnitude and nature of these interactions are keys for the design of efficient and productive systems. Traditionally, this knowledge is acquired for a particular area indirectly, by means of simulation growth models (Lisson, et al., 2005). These models deliberately reduce the system complexity for identifying the key aspects as well as to reduce the dataset required for parameterization. Alternatively, direct information of crop production can be stored in databases, which document what has actually happened in the farming systems, capturing large scale information on a wide range of variables that may potentially influence crop yield. However, the analysis of these large databases requires statistical methods capable of dealing with multivariate and nonlinear data structures.

Mathematical tools based on data mining provide an adequate framework for extracting useful information from large databases as well as they can also lead to the knowledge discovery for understanding environmental patterns and process. Recently, the use of learning algorithms of data mining has increased for analyzing small scale information (e.g hyper spectral remote sensing data) in order to provide crop management information for use in precision farming (Waheed, et al., 2006). However, the analyses of crop yield variability related to soil properties, agronomic practices or farmers' resource allocation decisions (e.g. agronomic factors) is usually carried out through traditional techniques such as linear regression and correlation.

This work proposes the use of two data mining techniques, k-means clustering (Jain and Dubes, 1988) and classification and regression trees (CART) (Breiman, et al., 1984) in order to identifying management and environmental factors for explaining sugarcane yield in Northern Argentina from a large database over the period of 1971 through 2005. The database contains information (i.e. field attributes) on the variety, crop age, area, farm of origin, and rainfall amount for every individual field of sugarcane harvested in each year. Sugarcane yield variable recorded was cane yield (TCH, t cane ha−1).

Firstly, TCH is analyzed through a k-means cluster analysis to group the sugarcane fields increasing cluster internal homogeneity and external or between-group heterogeneity. Then,

a classification tree partitions the space of all possible field attributes, starting with all field attributes (at the root of the tree) and successively splitting that space in subsets in which each subset is more likely to be assigned to one of the k-means clusters than the subset from which it is split. Also, the classification procedure is capable of provide information about the importance score (i.e. importance ranking) of the independent variables for explaining crop yield, which are ranked in descending order of their contribution to tree construction. The variable with the highest sum of improvements is scored 100, and all other variables have lower scores ranging downwards towards zero (Steinberg and Colla, 1995).

The methodology presented here is able to deal with large datasets in a robust way for detecting data patterns without the usual assumptions of data balance or distribution. Also, although CART are not inference tools, the information from this kind of studies could be useful to elaborate on testable hypothesis; pinpoint future mechanistic studies and designing more comprehensive management decisions for farming systems.

REFERENCES

Breiman, L., Friedman, R., Olshen, R., and Stone, C., *Classification and regression trees*, Pacific Grove, California, 1984.

Jain, A. K., and Dubes, R. C., *Algorithms for Clustering Data*, Prentice-Hall, Englewood Hills, NJ, 1988.

Lisson, S. N., Inman-Bamber, N. G., Robertson, M. J., and Keating, B. A., The historical and future contribution of crop physiology and modelling research to sugarcane production systems, *Field Crops Research, 92(2-3)*, 321-335, 2005.

Steinberg, D., and Colla, P., *CART: Tree-Structured Non-Parametric Data Analysis*, San Diego, 1995.

Waheed, T., Bonnell, R. B., Prasher, S. O., and Paulet, E., Measuring performance in precision agriculture: CART--A decision tree approach, *Agricultural Water Management, 84(1-2)*, 173-185, 2006.