A data mining approach to improve multiple regression models of soil nitrate concentration predictions in *Quercus rotundifolia* montados (Portugal).

Jorge Nunes • Manuel Madeira • Luíz Gazarini • José Neves • Henrique Vicente

Abstract The changes in the soil nitrate concentration were studied during 2 years in a ''montado'' ecosystem, in the South of Portugal. Total rainfall, air and soil temperature and soil water content under and outside *Quercus rotundifolia* canopy were also evaluated. A cluster analysis was carried out using climatic and microclimatic parameters in order to maximize the intraclass similarity and minimize the interclass similarity. It was used the k-Means Clustering Method. Several cluster models were developed using k values ranging between 2 and 5. Thereafter, in each cluster, the data were divided according to their origin (soil under canopy and open areas, and from surface and deep layers). Multiple regression models were tested for each cluster, to assess the relationship between soil nitrate concentration and a set of climatic and microclimatic parameters and the results were compared with models assessed without clustering. The models achieved with data grouped in result of clustering analysis showed better performance than the models achieved without clustering, mostly for data from open areas soils. When temperature is low and/or water presents excess or scarcity levels, the data from soils in under canopy areas, give rise to models with worst performance than models from open soil areas data. The results obtained for under canopy area suggest that nitrification process in soil under *Quercus rotundifolia* trees influence is more complex than for open areas, and subject to other relevant factors beyond water and temperature.