KNOWLEDGE DISCOVERY FROM DATA MINING OF THE ASRIS POINT DATABASE: SOIL NUTRIENTS

Elisabeth Bui[1], *Brent Henderson*[2], *Karin Viergever*[3]

[1]CSIRO Land and Water, GPO Box 1666, Canberra ACT 2601, Australia
[2]CSIRO Mathematical & Information Sciences, GPO Box 664, Canberra ACT 2601, Australia
[3]Ecometrica, Top Floor, Unit 3B, Kittle Yards, Edinburgh, EH9 1PJ, UK

Data mining can be used for making predictions and/or for uncovering patterns in large datasets, the knowledge discovery component of DM&KDD. During the first phase of the National Land and Water Resources Audit (NLWRA) the Australian Soil Resources Information Systems (ASRIS) project, a relatively large point database of soil properties was created by collating various databases into a single Oracle database [1]. Depending on the soil property, 5,000 to 24,000 points had useful data. Using this point database linked to national environmental data for climate (19 continuous variables), geology (23 discrete classes), land use (14 discrete classes), 4 Landsat MSS bands, and topography (14 continuous terrain variables), rule induction using Cubist (http://www.rulequest.com) decision trees was used to predict the spatial distribution of soil properties across the intensively used agricultural areas of Australia [2].

Cubist models are presented as a series of rules, each with a linear regression model. Continuous predictor variables can feature as splitting criteria in sub-setting the dataset and in the linear regressions at each leaf of the piecewise linear decision trees but categorical predictors can only be used to subset the data. Models were constructed with a 70:30 training to test data split: 70% of the observations were used to construct the model in the model development stage. 30% were held back in order to assess the performance of each model. Once the strongest possible model according to performance on the test data was identified, it was refitted using all the data to maximize the use of the relatively sparse data over Australia, with the same model form and options. The performance of the model on the full data set was assessed by 10-fold cross validation. The data were randomly split into 10 partitions or folds; at each step, nine of these partitions were used to fit the model and the performance assessed on the remaining partition held back as the test data. This procedure was repeated for each partition sequentially. The performance, averaged over all 10 partitions held back, delivers the cross-validated performance assessment. The performance of models was also assessed in terms of a number of key indicators: the number of points used in the model, the R^2 between measured and predicted values, the (rank) correlation, the RMSE (root mean square error), which gives an estimate of the standard deviation of the errors, the average error, and the relative error.

The data mining application is not explicitly spatial, i.e., it does not use geographical coordinates as predictors, rather, spatial structure is introduced implicitly by reliance on predictors that are available spatially extensively. Therefore, evaluation of the models in a spatial context has to proceed via an evaluation of the spatial distribution of the predictors in the context of model structure [3]. ASRIS maps were thus assessed against expert knowledge in natural sciences using visualization of model rules and of patterns of usage of predictor variables—what variables were important in models, whether consistent patterns emerged in their thresholds, and the spatial pattern defined by these thresholds.

In this talk, ASRIS data and models for soil organic C, total N and P will be presented and discussed. The Cubist model for soil organic C had 29 rules whereas the model for total P had 18 rules. Total N was predicted as a function of soil organic C. The model for total P performed better than that for C in terms of model evaluation statistics however both appear reasonable in terms of their predicted spatial patterns and what is known about the soil processes driving these soil nutrient patterns. Climatic variables alone were the most important predictors in the soil organic C model whereas lithology was also important in the total P model. Visualization of model rules showed a spatial correspondence between extent of rules and bioregions of Australia, as independently determined by the Interim BioRegionalization of Australia expert committee. A major spatial pattern in climatic thresholds seemed to correspond to the distribution of rainforests and Eucalyptus forests along the Australian coasts. Further investigation of associations between rules and vegetation pattern was undertaken using the National Vegetation Information Systems (NVIS), another NLWRA project that mapped vegetation across Australia, by estimating soil C:N:P relationships for the major vegetation groups of the NVIS. Vegetation groups rich in Proteaceae have a much higher soil C:N:P than other NVIS classes.

Canopy reflectance based on near infra-red spectroscopy has been used to link foliar chemistry to forest floor C and N stocks and soil C:N ratios [4, 5]. Differences in vegetation structure and reported foliar chemistry in the literature are sufficient over the NVIS classes to expect that hyperspectral remote sensing will be able to discriminate them. The important next step will be to verify the relationship between foliar chemistry, litter chemistry, and soil organic matter elemental composition. These results are thus potentially important for monitoring biomass and soil C in post-Kyoto Protocol agreements on greenhouse gas abatement.

REFERENCES

[1] R.M. Johnston, S.J. Barry, E. Bleys, E.N. Bui, C.J. Moran, D.A.P. Simon, P. Carlile, N.J. McKenzie, B. Henderson, G. Chapman, M. Imhoff, D. Maschmedt, D. Howe, C. Grose, N. Schokneckt, B. Powell, and M. Grundy, "ASRIS: The database," *Australian Journal of Soil Research,* vol. 41, no. 6, pp. 1021-1036, 2003.

[2] B.L. Henderson, E.N. Bui, C.J. Moran, and D.A.P. Simon, "Australia-wide predictions of soil properties using decision trees," Geoderma, vol. 124, no.3-4 , pp. 383-398, Feb. 2005.

[3] E.N. Bui, B.L. Henderson, and K. Viergever, "Knowledge discovery from models of soil properties developed through data mining," Ecological Modelling, vol.191, no. 3-4, pp. 431-446, Feb. 2006.

[4] J.A. Aitkenhead-Peterson, J.E. Alexander, J. Albrechtova, P. Kram, B. Rock, P. Cudlin, J. Hruska, Z. Lhokatova, R. Huntley, F. Oulehle, T. Polak, and W.H. McDowell, "Linking foliar chemistry to forest floor solid and solution phase organic C and N in Picea abies [L.} karst stands in northern Bohemia," *Plant and Soil*, vol. 283, no. 1-2, pp. 187-201, May 2006.

[5] S.V. Ollinger, M.L. Smith, M.E. Martin, R.A. Hallett, C.L. Goodale, and J.L. Aber, "Regional variation in foliar chemistry and N cycling among forests of diverse history and composition," *Ecology*, vol. 83, no. 2, pp.339-355, Feb. 2002.