A strategy for re-mapping the Murray-Darling Basin with a minimum of fieldwork

Elisabeth N. Bui and Christopher J. Moran

CSIRO Land and Water, GPO Box 1666, Canberra, ACT 2601, Australia

Presented at SIRC 99 – The 11th Annual Colloquium of the Spatial Information Research Centre
University of Otago, Dunedin, New Zealand
December 13-15th 1999

ABSTRACT

Until now, the only complete basin-wide soil coverage available for the Murray-Darling Basin (MDB) was the *Atlas of Australian Soils* at a scale of 1:2 000 000. The MDB Soil Information Strategy project has now produced a new soil map of the MDB that can be used at a scale of 1:250 000. All existing digital maps with any soil information were collated regardless of initial purpose for mapping, scale, and age. Thus much legacy data, some maps as old as 1931, had to be assembled. Many different legend schemes had been used to identify soil types. Legend differences, edge-matching, and variable scales raised major problems for re-mapping the MDB but these were not addressed directly. Rather than try to adapt different legend schemes to our purpose, a new framework for producing a unified, seamless coverage was devised. All soil types were converted to Principal Profile Forms (PPF) (Northcote, 1979) to describe soils in a consistent manner in the new soil-landforms map of the MDB.

Several gaps in detailed soil coverage became apparent when the soil data distribution was mapped, especially in the irrigated areas of northern NSW and in the Murray-Murrumbidgee Riverine plain. Where soil maps were completely lacking, soil distribution had to be determined by extrapolation using computer-based methods rather than field work (due to the project's limited budget for field work as well as time and personnel constraints).

Using existing soils maps, rules integrating DEM, geology, and MSS were developed to extend the soil mapping to neighbouring unmapped areas. The assumptions underlying rule development are that soil distribution reflects the long-term interactions between terrain variables, geology, and vegetation in landscapes and that the existing soil maps have captured those interactions. These rules were created using the C5.0 data mining software.

MSS imagery, geology, DEM, and terrain attributes were integrated into a large data set. The slope of the variograms for these input data was used to vary the size of a window over which the range (max-min), the variance, and the mean were computed for each band of MSS, the DEM and terrain attributes. The data layers thus created were used as inputs into C5.0 which was used to build decision trees to predict (1) regions over which environmental conditions were similar to those of the training maps and (2) soil distribution. In other words, 2 tree building procedures were performed. In the first, the training areas were used to identify similar environmental conditions to build "region trees". Next, trees were built for each training area to predict their soil distribution. Then prediction of soil distribution into unmapped areas occurred: the trees from each training area were applied over the appropriate region as identified in the first step to predict soil distribution over that region.

To build "region trees", random sampling was performed as a function of scale in keeping with the field sampling density guidelines outlined in the Australian Soil and Land Survey Handbook (Gunn et al., 1988). Thus to identify regions over which rules from a training set could be extended, 1:500 000 maps were randomly sampled at 2%, 1:250 000 maps at 8%, and 1:100 000 maps at 50%. To build trees for predicting soil distribution, where better precision was desirable, 1:500 000 maps were randomly sampled at 15%, 1:250 000 maps at 25%% (Manilla) or 20% (Bathurst, Forbes, Dubbo, and Goulburn), and 1:100 000 maps at 35%.

The consistency of a prediction can be used as an indication of its (un)certainty. With this principle in mind, ten classification trees were generated—as the training set is randomly sampled each time, the trees differ slightly in their predictions every time—and the frequency of class predicted was mapped. Where the same class is consistently predicted, the likelihood is greater that the prediction is correct. The modal class prediction was used to make the final map.