The dramatic effect of preferential sampling designs on sampling variance in environmental applications

Dr David Clifford, Research Scientist, CSIRO Division of Mathematics, Informatics and Statistics and the Sustainable Agriculture Flagship.

Classic probability based designs (simple random sampling and stratified sampling) are widely used for spatial sampling in environmental research. In practice however, when the sampling space is large, researchers may wish to preferentially sample some sites over others due to ease of access or because of previously established relationships with landowners. Non-standard probability designs can be implemented and Horvitz-Thompson analysis provides unbiased estimates for spatial means and variances in such cases provided first and second order inclusion probabilities can be evaluated. However the effect of preferential sampling on the sampling variance can be quite dramatic, even with minor departures from standard designs. We examine this effect through a study of percentage soil organic carbon (%SOC) as predicted across a large part (150,000 squared-km) of New South Wales in Australia using a Cubist based data-mining model of legacy %SOC data. Basic forms of preferential sampling are compared with standard designs. These forms have a single parameter that controls the level of preferential sampling. Plots of sampling variance against this parameter highlight significant increases in sampling variance as sites are more preferentially sampled. Our work shows that non-standard designs can result in significantly weaker performance of soil sampling.

This is collaborative research with P Kuhnert, M Dobbie, J Baldock, N McKenzie and B Harch from CSIRO and I Wheeler and A McBratney from the University of Sydney.