Standard errors in systematic sampling

Chris Glasbey and Stijn Bierman

Biomathematics and Statistics Scotland, Edinburgh, UK

There are many designs for sampling a region in order to estimate the mean value of a variable [5]. Of these, systematic sampling is known to be the most efficient in a broad range of situations [2]. Unfortunately, there is no generally accepted way to estimate the precision of the estimate obtained from this sampling scheme [3,1,4]. Because the data are essentially a sample of size one, no design-based standard error can be derived; and the alternative, a model-based approach, is time consuming to implement and relies on subjective assumptions that are difficult to validate.

Conservative standard errors can be obtained by pretending that the data had been obtained using a different, less efficient design, such as simple random sampling. If there is no indication of spatial correlation in the data, then such a standard error is also realistic for a systematic sample, but it can be a severe over-estimate in other cases. A tighter upper bound can be obtained by assuming a less inefficient design using strata [5,3]. Given prior knowledge of the direction of maximum variation, blocks of size two of adjacent points are formed orthogonal to this direction. In the presence of spatial correlation, results based on this form of pseudo-stratification are likely to over-estimate the true standard error for a systematic sample, but they are well established in the literature, are simple to compute and do not rely on unsubstantiated modelling assumptions.

We will explore the potential for improving on this established approach, using tree count data from Barro Colorado Island for illustration (http://ctfs.si.edu/datasets/bci/). In increasing levels of generality, we will consider whether: 1) standard errors obtained from simple, non-parametric modelling of data, can perform more accurately and robustly; 2) there are better estimators of the regional mean than the sample mean; 3) less efficient designs than systematic sampling are preferable because they permit quantification of the precision of results.

[1] Aubry P and Debouzie D (2001) Estimation of the mean from a two-dimensional sample: The geostatistical model-based approach. Ecology 82:1484-1494.

[2] Bellhouse DR (1977) Some optimal designs for sampling in two dimensions. Biometrika 64:605-611.

[3] Dunn R and Harrison AR (1993) Two-dimensional systematic-sampling of land-use. Applied Statistics 42:585-601.

[4] D'Orazio M (2003) Estimating the variance of the sample mean in two-dimensional systematic sampling. JABES 8: 280-295.

[5] Ripley BD (1981) Spatial Statistics (chapter 3). Wiley: NewYork.