Handling preferential sampling in areal summary statistics

Abstract

When an environmental (real valued) resource \( Y \) is monitored in a finite region \( D \subset \mathbb{R}^2 \) it is often required to estimate some spatial-aggregated characteristics of the distribution of \( Y \) such as the total, \( \int_D Y(x) \, dx \), or the mean, \( \bar{Y}(D) = Y(D) / |D| \), where \( |D| \) is the area of \( D \). In cases where \( D \) is a continuum, the sampling design is defined similarly to the usual approach adopted in the finite population framework except that continuous probability distributions are used instead of sampling probabilities. In this case the sampling design is defined by a joint distribution of \( n \) random variables and a sample of the surface \( Y(x) \) over a region \( D \) is chosen by selecting randomly a set \( x_1, \ldots, x_n \) of \( n \) locations and measuring \( Y(x) \) on it. More specifically, assuming that \( f(x_1, \ldots, x_n) \), the joint probability density function (PDF) of the sample locations, and \( f_i(x) \), the marginal PDF of \( x_i \), all exist, the inclusion density function is defined by \( \pi(x) = \sum_{i=1}^{n} f_i(x) \) for \( x \in D \).

By preferential sampling we mean that the sampling design, \( \pi(x) \), is dependent on \( Y(x) \) which is the object of the inference (McArthur, 1987). Figure 1 exemplifies the situation. In environmental monitoring, preferential designs mostly arise when the sampling locations are deliberately concentrated in subregions of \( D \) where the underlying values of \( Y \) are thought likely to be larger (or smaller) than average. Whatever the reason of this was, the selected sample is no longer representative of the region of interest and area-aggregated statistics are likely to be biased.

Under the hypothesis that the sampling design is known a Horvitz-Thompson-like estimator (Horvitz and Thompson, 1952) for the total has been proposed (Cordy, 1993) where the inclusion probabilities used in finite population sampling are replaced with inclusion densities i.e. \( \hat{Y}(D) = \sum_{i=1}^{n} Y(x_i) / \pi(x_i) \). Stevenson (1997) discussed several designs where the Horvitz-Thompson may be applied and Barabesi et al. (2012) investigated its statistical properties.

However, in many practical situations, although the data are sampled preferentially, the sampling design is not known analytically, i.e. the function \( \pi(x) \) is not known in advance and has to be estimated \emph{a posteriori} on the data at hand.

In this paper we estimate the sampling design for irregularly sites preferentially sampled using the Voronoi tessellation of the region \( D \) (Okabe 2012) generated by the sampling locations as \( \hat{\pi}(x_i) = 1 / |V(x_i)| \) where \( V(x_i) \) is the polygon generated by the sample location \( x_i \) i.e. the region consisting of all points closer to \( x_i \) than to any other \( x_j \neq x_i \), as it is exemplified by figure 1.

We investigate the properties of the above estimator and compare it to other procedures currently utilized to calculate area-aggregated statistics in presence of preferential samples.
Figure 1. Preferential sample: small black circles represent the sampling locations whereas the white-red thematism represents the surface \(Y(x)\) to be monitored on a squared region of side 1. Sample sites tend to cluster on white-yellow areas where the value of the field is higher. Polygons represent the Voronoi tessellation generated by the sample.

References


