Abstract

Let P be a set of points in $\mathbb{R}^d$, and let M be a function that maps any subset of P to a positive real number. We examine the problem of computing the exact mean and variance of M when a subset of points in P is selected according to a well-defined random distribution. We consider two distributions; in the first distribution (which we call the *Bernoulli* distribution), each point $p \in P$ is included in the random subset independently, with probability $\pi(p)$. In the second distribution (the *fixed-size* distribution), a subset of exactly s points is selected uniformly at random among all possible subsets of s points in P. This problem is a crucial part of modern ecological analyses; each point in P represents a species in d-dimensional trait space, and the goal is to compute the statistics of a geometric measure on this trait space, when subsets of species are selected under random processes. We present efficient exact algorithms for computing the mean and variance of several geometric measures when point sets are selected under one of the described random distributions. More specifically, we provide algorithms for the following measures: the bounding box volume, the convex hull volume, the mean pairwise distance (MPD), the squared Euclidean distance from the centroid, and the diameter of the minimum enclosing disk. We also describe an efficient $(1 - \varepsilon)$-approximation algorithm for computing the mean and variance of the mean pairwise distance. We implemented three of our algorithms: an algorithm that computes the exact mean volume of the 2D bounding box in the Bernoulli distribution, an algorithm that computes the exact mean and variance of the MPD for d-dimensional point sets in the fixed-size distribution, and an $(1 - \varepsilon)$-approximation algorithm for the same measure. We conducted experiments where we compared the performance of our implementations with a standard heuristic approach used in ecological applications. We show that our implementations can provide major speedups compared to the standard approach, and they produce results of higher precision, especially for the calculation of the variance. We also compared the implementation of our exact MPD algorithm with the corresponding $(1 - \varepsilon)$-approximation method; we show that the approximation method performs faster in certain cases, while also providing high-precision approximations. We thus demonstrate that, as an alternative to the exact algorithm, this method can also be used as a reliable tool for ecological analysis.