Mean-min distance sampler

Ty Tuff for Bruno Vilela

project began: Fall 2017, document updated: 08 December 2017

We begin with a dataset that we would like to subsample to remove some bias. In the example below, these position data are and coordinate values that we could measure the Euclidean distance between. However, this method can work for any data where individual row indicate a unique position relative to the other data so, distances can be calculated between those points. Some examples of positional data include geographic, genetic, environemntal, or phylogenetic distances.

as.data.frame(matrix(c("x1", "x2", "x3", "x4", "y1", "y2", "y3", "y4") ,4,2, byrow=FALSE, dimnames = list(c("S1", "S2", "S3", "S4"), c("x","y"))))

## x y  
## S1 x1 y1  
## S2 x2 y2  
## S3 x3 y3  
## S4 x4 y4

The length of this dataset is defined as and the index values for each row define the vector .

Randomly remove one datum from , called , and use that index to start a second vector, .

defines the number of samples to be drawn

Only the first sample is chosen at random, all subsequent samples are chosen according to the sampling procedure described below and continued without replacement until reaches length

is the distance between a point in and a point in . This can be any type of distance measure that can be reported as a number, including: geographic, phylogenetic, genetic, and environmental distances. the distance between points and .

Calculate two weights, and , values from each row, , of the matrix

This creates two weights for each row in matrix . For each pair of weights representing each row of , raise and to the power of and sum them to define .

weighting exponent. This variable can range from to . Negative values of will promote overdispersed sampling and positive values of promote clumped sampling around the initial, randomely drawn, sample.

Create a new vector, , by scaling each element of to fall within the range of 0 1.

The weights described by correspond element-for-element with . Sample one point from , without replacement, according to probabilities defined by . Add that sampled point to the vector and create a new matrix, , with 2 columns and rows.

Repeat this process of weighting samples, sampling them according to that weight, moving that sampled point from to , and then recalculating the distance matrix until the length of to define .