Appendix H

Stratified random sampling with disproportionate allocation

When a multi-scale decomposition is applied to the scalar field from which the structures are educed, large differences among the number of structures obtained for each scale are to be expected. Larger scales will generally have a smaller number of educed structures than smaller scales. This difference in number can sometimes be of several orders of magnitude, particularly when analyzing fields with high grid resolution that results in a larger number of scales.

Thus, when structures of all scales are considered in the clustering algorithm, after the geometrical characterization, those structures (and their geometries) corresponding to the largest scale can be under-represented owing to the much smaller population they have compared to the others. In such a scenario, it can be beneficial to apply, prior to the clustering algorithm itself, a sampling of the population that takes into account the uneven sizes of the strata in which it can be divided.

We use a disproportionate stratification that considers the variance of the mutually exclusive strata to determine the sample size for each stratum. If n_o is the sample size of the stratum with the minimum standard deviation, $\sigma_o = \min\{\sigma_h, \forall h\}$, then the sample size, n_h, of any other stratum, h, with standard deviation σ_h will be proportional to $(\sigma_h/\sigma_o)\ n_o$. Therefore, those strata with higher variances will have also a higher number of elements to represent them in the clustering algorithm, accounting for their higher diversity. We take n_o as the population size of that stratum with the minimum standard deviation, N_o, since the purpose of this sampling is not to reduce the global population size, but to have a more balanced representation of the different groups present in it for

a better clustering.

After the disproportionate stratification, for those strata with $n_h < N_h$, where N_h is the population size of the stratum h, we take a random sample of n_h out of the N_h elements. Otherwise, the complete population is considered for that stratum.