**The Hopkins statistic**

The Hopkins statistic (H) tells you how well your data will produce clusters. The null-hypothesis is that the data points were generated by a Poisson point process, and thus, the data points are distributed randomly.

To measure the Hopkins statistic, the same number of random data points are generated to match the number of points in your data set. If your data set is made up of meaningless clusters, there should be no difference between your data points and the randomly generated data points. Conversely, if your data set is real, then there should be a statistically significant difference between your data points and the random ones.

Thus:

* Let X be the set of n data points in your data set.
* Generate a set of m << n randomly selected data points from your data set, denoted as xi.
* Generate a set Y of uniformly distributed data points.
* Define two distance measures:
  + ui, the distance of yi from its nearest neighbor in X
  + wi, the distance of m randomly chosen xi, xi data points in X from its nearest neighbor in X

The Hopkins statistic is defined as

The Hopkins statistic is interpreted in the following way:

* 0.75 < H : good clustering
* 0.5 < H < 0.75 : acceptable clustering
* H = 0.5 : random clustering
* H = 0 : uniform data points

Generally, the higher the Hopkins statistic, the better quality your data set is. Data sets with a Hopkins statistic above 0.75 are worth analyzing further.

**References:**

1. Wikipedia: en.wikipedia.org/wiki/Hopkins\_statistic
2. Banerjee, A. (2004). "Validating clusters using the Hopkins statistic". *IEEE International Conference on Fuzzy Systems*: 149–153.