What are the true clusters?

Finding clusters in data, often referred to as "usupervised classification", is a central task in statistical learning. Although technically it may look like a minor nuisance that in clustering there is no "training set" with class labels, actually there is a fundamental difference between clustering and supervised classification.

The key difference is that it is not clear in cluster analysis whether there is any such thing as a "true clustering", and if there is one, there is no guarantee whatsoever that this is unique. This issue is ignored in much literature about cluster analysis; routinely cluster analysis methods are compared based on "misclassification rates", which implicitly assume that there is such a unique true clustering, and this is known for the datasets used for comparison.

I will argue that, depending on the aim of clustering, there are several legitimate clusterings on the same data. Furthermore, I will argue that many of the decisions made when clustering data, such as transformations of variables, selection of variables, and definition of distances, come with different implicit definitions of what kind of clusters can be found. Ultimately, the user has to decide what kinds of clusters are of interest in the given application, and this decision cannot be made automatically by the data.

I will present some different ways of defining "true clusters" and some criteria to measure the quality of a clustering, and will give some guidance in what kind of different situations these could be relevant.

Because different cluster analysis methods are implicitly based on different definitions of clusters, choosing a definition and quality criteria can then be used to compare and choose dofferent clustering methods in a given application or potentially even define new ones that are better adapted to the problem in hand.

Reference

Hennig, C. (2015) What are the true clusters? Pattern Recognition Letters 64, 53-62.