Editorial

The first three articles in this issue all develop clustering methodology, but each originates from quite a different setting and uses a different conceptual framework. In the first article, Contreras and Murtagh present a new approach to hierarchical clustering that is substantially faster than the usual agglomerative hierarchy-constructing algorithms, and that can handle in the millions of objects. They capitalize on the Baire metric, which compares pairs of sequences in terms of the longest common prefix, and which induces naturally an ultrametric on the set of objects. The authors establish a link with formal concept analysis, and demonstrate the power of their approach with a huge astronomical case study.

The setting of the second article by Palarea-Albaladejo, Martín-Fernández and Soto involves compositional data, in which the data elements are non-negative and sum to a constant. Geometrically, the sample space of such data is the interior of a simplex, which leads to the natural requirement that data analysis must focus on ratios between components, rather than suppose that data points vary freely in an unbounded space. The natural distance to use here turns out to be the Aitchison distance, and the paper shows how to build a coherent framework for (fuzzy, non-hierarchical) clustering in this metric. In the third article, Giordani and Kiers work in a setting of three-way proximity data, and use the fuzzy clustering framework as well. They define and motivate a fuzzified alternative to Carroll and Arabie's INDCLUS model for overlapping clustering, and provide an alternating least squares algorithm for fitting this new FINDCLUS model.

After these clustering articles, Morlini and Zani present a new class of general-purpose indices measuring the similarity between subjects or objects on the basis of polytomous variables, taking into account the marginal frequency of the categories (or levels) involved. They consider two versions of their index, one that ignores possible associations among the variables—that they compare with the well-known chi-squared distance of correspondence analysis—and another that includes these associations. In the last paper of this issue, D'Ambrosio, Aria and Siciliano introduce an approach to missing data imputation and data fusion that does not follow Rubin's classical stochastic paradigm, but aims at minimizing the generalization (or prediction) error in Vapnik's deterministic paradigm of statistical learning. Their incremental tree-based methodology works one variable at a time, processes multiple imputation via boosting, and turns out to perform well in simulations.

Willem J. Heiser

Institute of Psychology,
Leiden University,
Leiden, The Netherlands