A New Approach To Cluster Validation in Experimental Investigations of (Geo)Spatial Concepts

J. O. Wallgrün[1], A. Klippel[1], D. Mark[2]

[1]The Pennsylvania State University, University Park, PA 16802
Email: {wallgrun; klippel}@psu.edu

[2]NCGIA & Department of Geography, University at Buffalo, Buffalo, NY 14228
Email: dmark@buffalo.edu

1. Introduction

Cluster analysis is a popular method across many disciplines and frequently employed in the spatial sciences to analyze observed or experimentally collected data. Cluster analysis either operates on entities in an *m*-dimensional feature space based on some distance measure (e.g., Euclidean distance) representing the dissimilarity of features, or, alternatively, directly on a given proximity matrix representing the similarity / dissimilarity between pairs of entities. Cluster analysis approaches can be distinguished into partitioning and hierarchical methods. While partitioning cluster methods identify cluster membership at a single level, e.g., on the basis of a predefined number of clusters, hierarchical methods iteratively join entities and clusters based on different algorithms resulting in a tree structure (dendrogram).

One of the greatest challenges of hierarchical cluster analysis is to decide how to interpret the clustering process and dendrogram, i.e., deciding on the right/best number of clusters (cluster validation). This problem is aggravated by different clustering methods being available which potentially offer different interpretations on both the number of clusters and the cluster-membership of entities.

This paper advances cluster validation for grouping experiments (category construction / free classification) in which participants organize stimuli on the basis of perceived similarity. Such experiments typically aim at shedding light on cognitive concepts and are applied in many areas, from human-computer-interaction design (e.g., Roth et al. 2011) to the assessment of qualitative spatial calculi (Mark and Egenhofer 1994, Knauff et al. 1997, Klippel et al. 2013). We propose a novel cluster validation method that determines a) the best cluster solution; b) the required number of participants, and c) enables (meta) comparisons across different experiments. We discuss results of applying this method to data collected in several geospatial behavioral studies.

2. Cluster Validation

Three cluster validation approaches are prominent in the literature: First, comparing the results of different clustering methods. This approach is sometimes referred to as confirmatory cluster analysis (Fisher and Ransom 1995). Reanalyzing the data using different methods can determine the extent to which solutions converge.

Second, indices (e.g., Rand Statistics, Jaccard Coefficient; see Halkidi et al. 2002) are used to assess to which degree two partitions (i.e., sets of clusters) *G* and *H*, match. These indices allow for comparing the clustering of stimuli resulting from a grouping experiment with an assumed theoretical partition or to assess how similar the results of different clustering methods are. They are based on the number of pairs of stimuli that belong to the same group in both *G* and *H* (*SS*), the number of pairs that belong to the same group in *G* but different groups in *H* (*SD*), the number of pairs that belong to the same group in *H* but

different groups in G (DS), and the number of pairs that belong to different groups in both G and H (DD). The Jaccard Coefficient, for instance, is computed as

$$J(G,H) = \frac{SS}{SS+SD+DS}$$

Third, if the sample is large enough, it can be randomly split and each half can be analyzed separately and solutions can be compared (Mandara 2003).

3. Sampling-based Cross-Method Validation Approach

Combining and extending the above mentioned validation approaches, our cluster validation method described in the following is based on the general idea of sampling from the pool of participants and computing a similarity index called *cross-method similarity index* (CMSI) for the groups obtained for a given number of clusters by applying three different clustering methods: Ward's method, average linkage, and complete linkage. The CMSI value is computed for sample sizes up to the number of available participants and different numbers of clusters c. Results are plotted (see Figure 4) which allows for addressing the issues described above.

The CMSI is computed as follows: Participant pool P consists of m participants $p_1, \ldots p_m$ and the grouping created by participant p_i is given by an individual similarity matrix (ISM) ISM_i which contains a '1' for each pair of stimuli put into the same group by p_i and a '0' for pairs put into different groups.

Given the sample size n, a random sample S_n of n participants is drawn from P. An overall similarity matrix (OSM) is then computed by adding up the ISMs for the participants contained in the sample (Figure 1). Three cluster analyses using three different methods (see above) are performed on the OSM.

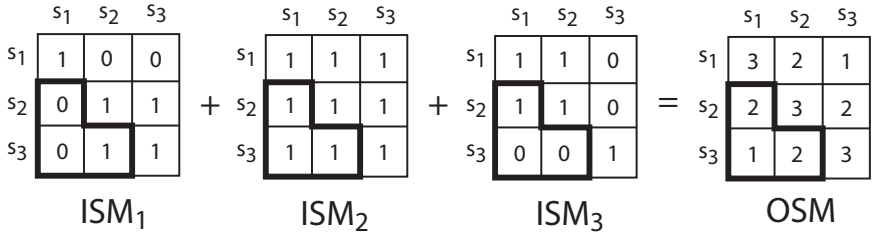

Figure 1. Computation of the OSM matrix.

The output of the clustering methods are dendrograms (see Figure 2). For a given number of clusters c, the respective groupings are derived which corresponds to cutting the dendrograms at a particular height (Figure 2). The resulting groupings are referred to as $G_{c,S_n}^{[ward]}$, $G_{c,S_n}^{[avg]}$, and $G_{c,S_n}^{[comp]}$. Finally, the Jaccard Coefficient is used to compute the similarity of the groupings for each pair of methods; the average is taken as the CMSI value:

$$CMSI_{c,S_n} = \frac{J(G_{c,S_n}^{[ward]},G_{c,S_n}^{[avg]}) + J(G_{c,S_n}^{[ward]},G_{c,S_n}^{[comp]}) + J(G_{c,S_n}^{[avg]},G_{c,S_n}^{[comp]})}{3}$$

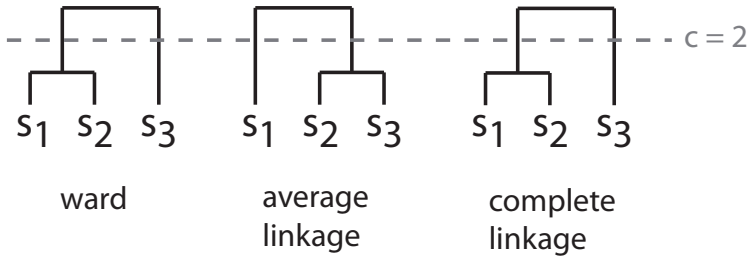

Figure 2. Dendrograms resulting from different clustering methods and cutting them to get groupings with c=2 clusters.

The CMSI value provides a measure of how well the different clustering methods agree for a given sample and number of clusters. It can easily be generalized to include other clustering methods as well. Computing the average CMSI value over many samples and for different numbers *n* and *c* and plotting it as in Figure 4 allows for a statistical perspective on the grouping behavior.

4. Results

We used the CMSI approach to analyze and compare the results from several grouping experiments we conducted in the past on human conceptualizations of spatial relations. We here only provide a demonstration by comparing two experiments. The first used icons showing different topological relations based on Galton's overlap relations (Galton 1998); the second used icons showing two airplanes in different direction relations (see Figure 3).

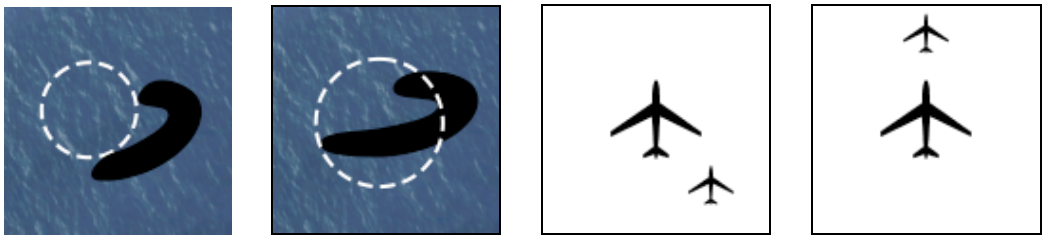

Figure 3. Icons from the two grouping experiments (left: a protected habitat in relation to an oil spill; right: two airplanes).

Figure 4 shows the resulting CMSI plots for these two experiments. In the Overlap experiment even a small numbers of participants reached the optimal CMSI score for a three-cluster solution. This corroborated findings reported in Wallgrün et al. (2013) that a three-cluster solution that separates non-overlapping, overlapping, and proper-part relations is the cognitively adequate model for modes of overlap.

In contrast, the CMSI plot for the Directions experiment does not show a perfect score at all. The conclusion is that participants adopted competing and mutually exclusive direction concepts: some used half planes and separated directions into quadrants, other used a cone shape approach. However, splitting the participants according to their strategies and applying the CMSI approach to these subgroups results in perfect scores.

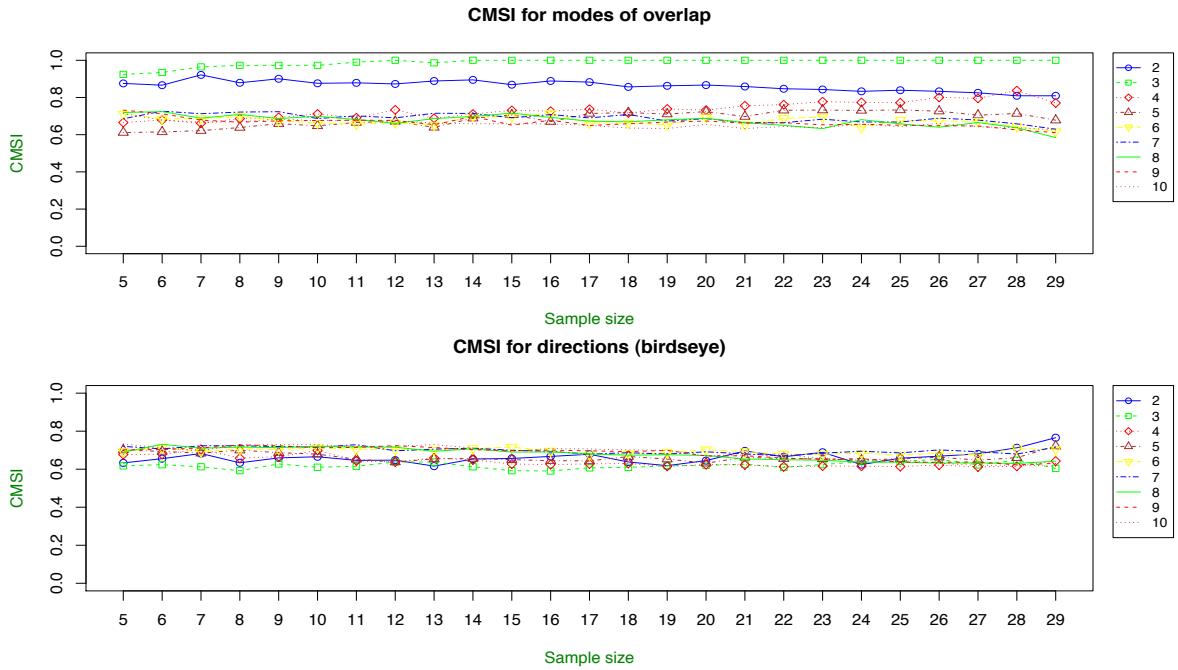

Figure 4. Plots of CMSI values for cluster numbers c=2 to 10 (see legends on the right).

5. Conclusions

The sampling based cross-method similarity validation method proposed in this paper is a tool to support the statistical analysis of experimental data from behavioral studies on human spatial cognition. It has the potential to reveal whether a common conceptualization of the stimuli exists or whether there is a need to acknowledge competing perspectives—crucial questions in cognition and ontology engineering. The CMSI is applicable in other domains where cluster analysis plays a role. In addition to what we could show in this abstract, we have applied it to a much larger number of experiments and to subgroups of participants resulting from a participant similarity analysis to validate clustering results.

Acknowledgements

This research is funded by the National Science Foundation under grant #0924534.

References

Fisher, L and Ransom, D C, 1995, An empirically derived typology of families: I. Relationships with adult health. *Family Process*, 34(2):161–182.

Galton, A, 1998, Modes of overlap. *Journal of Visual Languages and Computing, 9*:61–79

Halkidi, M, Batistakis, Y, and Vazirgiannis, M, 2002, Cluster validity methods: Part I. *ACM SIGMOD Record, 31*(2):40–45.

Klippel, A, Li, R, Yang, J, Hardisty, F, and Xu, S, 2013, The Egenhofer-Cohn hypothesis or, topological relativity? In Raubal, M, Frank, A, and Mark, D (eds.), *Cognitive and Linguistic Aspects of Geographic Space - New Perspectives on Geographic Information Research*, pp. 195–215.

Knauff, M, Rauh, R, Renz, J, 1997, A cognitive assessment of topological spatial relations: Results from an empirical investigation. In: Hirtle, Frank (eds.) *Spatial Information Theory*, pp. 193-206.

Mandara, J, 2003, The typological approach in child and family psychology: A review of theory, methods, and research. *Clinical Child and Family Psychology Review*, 6(2):129–146.

Mark, D M and Egenhofer, M J, 1994, Calibrating the meanings of spatial predicates from natural language: Line-region relations. In: Waugh, Healey, (eds.) *Advances in GIS Research*, pp. 538-553.

Roth, R E, Finch, B G, Blanford, J I, Klippel, A, Robinson, A C, and MacEachren, A M, 2011, The card sorting method for map symbol design. *Cartography and Geographic Information Science, 38*(2):89–99.

Wallgrün, J O, Yang, J, Klippel, A, 2013, Investigating intuitive granularities of overlap relations. In *12th IEEE International Conference on Cognitive Informatics and Cognitive Computing*, 2013.