## Clusterings should not be compared by visual inspection: response to Gagné & Proulx

### ABSTRACT

In Heikinheimo *et al.* (*Journal of Biogeography*, 2007, **34**, 1053–1064) we used clustering to analyse European land mammal fauna. Gagné & Proulx criticized our choice of the Euclidean distance measure in the analysis, and advocated the use of the Hellinger distance measure, claiming that this leads to very different clustering results. The criticism fails to take into account the probabilistic nature of the methods used and the fact that in this case the similarity measures correlate strongly. Gagné & Proulx used subjective inspection as the criterion of similarity between clusterings. We show that this is insufficient and misleading. Namely, owing to the local minimum problem, two clustering runs rarely give identical results. In the case of our study, the measured similarity (using the kappa statistic) between the Euclidean- and Hellinger-based clusterings is roughly equal to the similarity between two clusterings that both use the Hellinger distance but different random initialization points.

**Keywords** Clustering, double-zero problem, Euclidean distance, Europe, Hellinger distance, kappa, objective comparison, optimization with randomization, species abundance paradox.

In our recent paper (Heikinheimo *et al.*, 2007) we used clustering to analyse European land mammal fauna. In a following correspondence, Gagné & Proulx (2008) criticized our choice of the Euclidean distance measure in the analysis and advocated the use of the Hellinger distance measure, claiming that this leads to very different clustering results.

We want to thank Gagné and Proulx for drawing attention to an aspect of our study that we had not considered, namely that our cluster maps might be used as a basis for conservation policy decisions. We agree that this would be a dubious use of them for reasons that we discuss below. The purpose of our study was not to investigate the exact distribution boundaries of taxa, threatened or otherwise. Instead, we wished to study the nature of the spatial distribution pattern of community-like species assemblages, a goal clearly reflected in our discussion in the paper (Heikinheimo *et al.*, 2007).

Which similarity measure is the best for the purpose is no doubt debatable, and a full discussion is beyond the scope of this response. The properties of the measure depend heavily on the type of data. We are preparing a research paper in which we will test and discuss at length the differences and choices between the similarity measures (H. Heikinheimo, M. Fortelius, J. Eronen & H. Mannila unpublished data). Here we simply report what seems to us the most important result, that the choice between Euclidean and Hellinger distances has a very small effect on the results.

The cluster analysis of our study (Heikinheimo *et al.*, 2007) was conducted using two distinct clustering methods: *k*-means (also known as ISODATA) (Legendre & Legendre, 1998; Duda *et al.*, 2000; Theodoridis & Koutroumbas, 2003); and a probabilistic technique, Bernoulli mixture modelling, using the expectation maximization (EM) algorithm (Everitt & Hand, 1981; Cadez *et al.*, 2000; McLachlan & Peel, 2000; Hand *et al.*, 2001). Both methods are based on an iterative process with random initialization points.

For the data in our study, the Euclidean distance and the Hellinger distance are in fact very similar (Table 1). The correlation between distances is strong (> 0.75) for all the subsets considered. The small values of the distance measures agree especially well (Fig. 1). As clustering methods search for clusters with small diameter, the behaviour of the distance measure on small distances is more important to the clustering outcome than the behaviour of the distance measure on large distances. We compared the results of the two clustering methods using the kappa statistic (Monserud & Leemans, 1992). According to the guidelines in Monserud & Leemans (1992), the clusterings of all of the species subsets have either good or very good agreement (kappas from 0.52 to 0.81) between the results of the two methods (Table 2 in Heikinheimo *et al.*, 2007).

Gagné and Proulx do not present an objective or systematic evaluation of the level of similarity in support of the claim that their clusterings using the Hellinger distance differ considerably from our results. To test their claim, we repeated the *k*-means cluster analysis using the Hellinger distance and compared the results with the *k*-means clusterings presented in our original study (Heikinheimo *et al.*, 2007) with the kappa statistic. All species subsets show either good agreement or very good agreement between clusterings based on Euclidean vs. Hellinger distance (Table 2). Thus, the differences cannot be described as large, contradicting the claim of Gagné and Proulx based on subjective visual inspection of map outputs.

An aspect that Gagné and Proulx omit in their critique is the inherent stochastic nature of any high-dimensional clustering method: for each clustering run, the *k*-means method is initialized using a random assignment of cluster centres. This causes results to fluctuate somewhat from one clustering run to the next. This is also referred to as the 'local minimum' problem by Legendre & Legendre (1998, p. 350). However, by repeating the procedure several times and selecting the best among the resulting outcomes, algorithms of this kind are known to give modelling results close to the best possible solution (Motwani & Raghavan, 1995). This is why for all data sets in our original paper we present the clustering with the smallest error out of 100

**Table 1** Concordance between the Euclidean and the Hellinger distance measure for each species subset studied in Heikinheimo *et al.* (2007).

| Species subset | All species | Small species | Large species | Herbivora | Omnivora | Carnivora | Present 10% | Present 20% | Present 30% | At risk | Not at risk |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Correlation | 0.78 | 0.75 | 0.82 | 0.76 | 0.82 | 0.83 | 0.78 | 0.80 | 0.87 | 0.85 | 0.79 |

The values in the table are the Pearson correlations between the Euclidean distance and the Hellinger distance for the entire set of 2,381,653 grid-point pairs in the mammal data with each species subset. Present 10%, Present 20% and Present 30% refer to species with a grid cell coverage of 10%, 20% and 30% or higher.

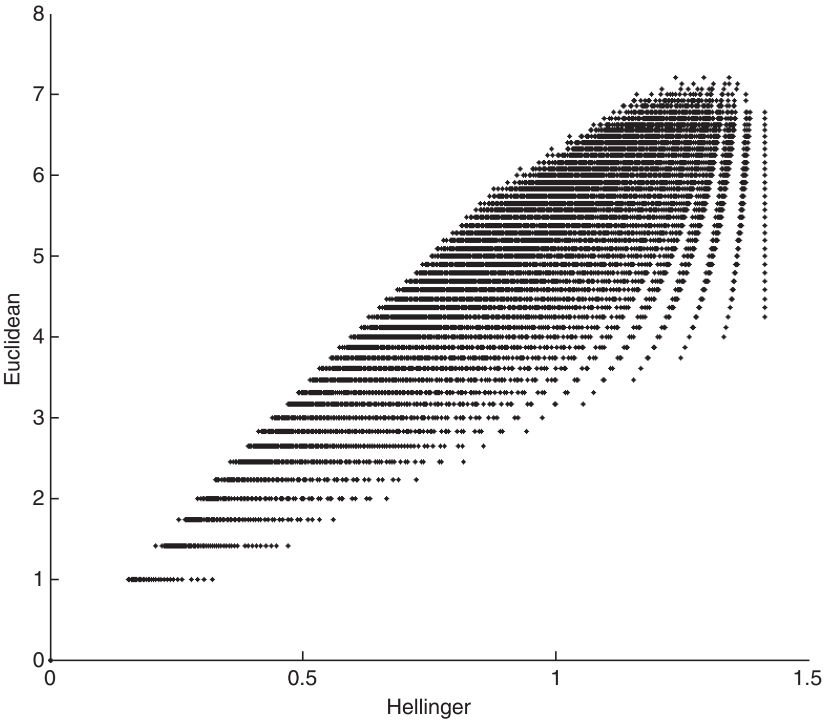

**Figure 1** Comparison between the Hellinger distance measure and the Euclidean distance measure for the 'all species' data set studied in Heikinheimo *et al.* (2007). Each point in the scatter plot represents a grid-point pair such that the respective Hellinger distance value is plotted on the *x*-axis and the corresponding Euclidean distance value is plotted on the *y*-axis. The number of grid-point pairs in the comparison is 2,381,653.

**Table 3** Average agreement between two *k*-means clustering runs both using the Hellinger distance measure but a different initial assignment of cluster centres.

| Species subset | Kappa average | Kappa standard deviation |
|---|---|---|
| All species | 0.70 | 0.08 |
| Small species | 0.69 | 0.09 |
| Large species | 0.60 | 0.08 |
| Herbivora | 0.63 | 0.07 |
| Omnivora | 0.61 | 0.08 |
| Carnivora | 0.64 | 0.08 |
| Present 10% | 0.71 | 0.09 |
| Present 20% | 0.70 | 0.08 |
| Present 30% | 0.62 | 0.09 |
| At risk | 0.59 | 0.09 |
| Not at risk | 0.70 | 0.09 |

The table shows the average kappa value with standard deviation for a set of 100 clustering result pairs. Each clustering in the comparison has been initialized with a different set of 12 random cluster centres. Present 10%, Present 20% and Present 30% refer species with a grid cell coverage of 10%, 20% and 30% or higher.

runs. As no two clustering outcomes using the same data and metric can be expected to be identical, visual inspection is an insufficient basis for meaningful comparison. We would like to take this opportunity to discuss further the probabilistic nature of the clustering methods used.

It should be emphasized that the dependence on random starting points is typical of all clustering methods for high-dimensional data: the optimization tasks in clustering are computationally intractable. [They belong to the class of nondeterministic polynomial-time hard, more often refferred to as NP-hard (Garey & Johnson, 1979), computational problems; such problems are not known to be efficiently solvable.] Thus methods that can give only locally optimal solutions have to be used (McLachlan & Peel, 2000; Hand *et al.*, 2001), and if a high-dimensional clustering method gives deterministic answers, it will report suboptimal solutions for some inputs.

To put the difference in the clustering results reported by Gagné and Proulx into perspective, we ran the *k*-means clustering algorithm (Matlab standard *k*-means) 100 times using the Hellinger distance with 12 clusters and compared the similarity of consecutive clustering runs using the kappa statistic. The idea was to see how much the results fluctuated within clustering runs using the Hellinger distance. The results show that the average kappa similarity between two clustering runs is from 0.6 to 0.7, with a standard deviation of 0.07 to 0.09 (Table 3). This represents good to very good agreement, according to the guidelines of Monserud & Leemans (1992). Recall that this is the amount of difference that is also found between the clusterings obtained using the Euclidean distance and Hellinger distance (Table 2), as well as between the Bernoulli mixture modelling and the *k*-means using the Euclidean distance.

To conclude, we agree with Gagné and Proulx that clusterings such as these should not be frivolously applied to conservation policy decisions. To do so would obviously be inappropriate, not only because of the uncertainty that is inherent in the methods, but also because in this case the raw data have a resolution of 50 km, which is too coarse to capture the mosaic nature of endangered species habitats and constrained ranges. We also agree with Gagné and Proulx

**Table 2** Agreement between clustering results using the Euclidean and the Hellinger distance measure for each species subset studied in Heikinheimo *et al.* (2007).

| Species subset | All species | Small species | Large species | Herbivora | Omnivora | Carnivora | Present 10% | Present 20% | Present 30% | At risk | Not at risk |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Kappa | 0.72 | 0.75 | 0.72 | 0.55 | 0.63 | 0.73 | 0.68 | 0.82 | 0.75 | 0.77 | 0.74 |

The table shows the kappa values computed between each *k*-means clustering presented in Heikinheimo *et al.* (2007) and the corresponding clustering using the Hellinger distance with the smallest error out of 100 clustering runs. Present 10%, Present 20% and Present 30% refer to species with a grid cell coverage of 10%, 20% and 30% or higher.

that the rigorous application of appropriate statistical techniques is a crucial concern in quantitative biogeographical analysis.

H. Heikinheimo[1]*,
M. Fortelius[2],
J. Eronen[2] and
H. Mannila[1,3]

*[1]HIIT Basic Research Unit, Department of Information and Computer Science, Helsinki University of Technology, PO Box 5400, FI-02015 TKK, Espoo Finland, [2]Department of Geology and Institute of Biotechnology, FIN-00014 University of Helsinki, PO Box 64, Helsinki, Finland and [3]HIIT Basic Research Unit, Department of Computer Science, FIN-00014 University of Helsinki,*
*PO Box 68, Helsinki, Finland*
**E-mail: hannes.heikinheimo@tkk.fi*

## REFERENCES

Cadez, I.V., Gaffney, S. & Smyth, P. (2000) A general probabilistic framework for clustering individuals and objects. *Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ed. by R. Ramakrishnan and S. Stolfo), pp. 140–149. ACM Press, New York.

Duda, R.O., Hart, P.E. & Stork, D.G. (2000) *Pattern classification*, 2nd edn. John Wiley & Sons, New York.

Everitt, B.S. & Hand, D.J. (1981) *Finite mixture distributions*. Chapman & Hall, London.

Gagné, S.A. & Proulx, R. (2008) Accurate delineation of biogeographical regions depends on the use of an appropriate distance measure. *Journal of Biogeography*, doi: 10.1111/j.1365-2699.2008.01990.x.

Garey, M.R. & Johnson, D.S. (1979) *Computers and intractability: a guide to the theory of NP-completeness*. W. H. Freeman and Company, New York.

Hand, D., Mannila, H. & Smyth, P. (2001) *Principles of data mining*. MIT Press, Cambridge, MA.

Heikinheimo, H., Fortelius, M., Eronen, J. & Mannila, H. (2007) Biogeography of European land mammals shows environmentally distinct and spatially coherent clusters. *Journal of Biogeography*, **34**, 1053–1064.

Legendre, P. & Legendre, L. (1998) *Numerical ecology*, 2nd edn. Elsevier, Amsterdam.

McLachlan, G. & Peel, D. (2000) *Finite mixture models*. John Wiley & Sons, New York.

Monserud, R.A. & Leemans, R. (1992) The comparison of global vegetation maps. *Ecological Modelling*, **62**, 275–296.

Motwani, R. & Raghavan, P. (1995) *Randomized algorithms*. Cambridge University Press, Cambridge.

Theodoridis, S. & Koutroumbas, K. (2003) *Pattern recognition*, 2nd edn. Elsevier Academic Press, New York.

Editor: Bradford Hawkins

doi:10.1111/j.1365-2699.2008.02054.x

## Panbiogeographical study of hagfishes: an anachronistic analysis

### ABSTRACT

In a recent paper by M. J. Cavalcanti and V. Gallo, 'Panbiogeographical analysis of distribution patterns in hagfishes (Craniata: Myxinidae)' (*Journal of Biogeography*, 2008, **35**, 1258–1268), the authors studied the biogeography of an ancient fish family (Myxinidae) in the hope that the contemporary distributions of the species would reveal their past history and that of the ocean basins where they reside. In order to accomplish this task, there are several criteria that should have been met: (1) the ages of the taxa utilized (species) would have to be old enough to reflect the history of the areas where they are found, (2) the identification of the species as listed in the databases would have to be accurate, (3) the geographical locations indicated on the figures would have to be consistent with the statements in the text, and (4) the significance of the vicariant patterns would have to depend on evidence pertaining to the ages of such patterns. Unfortunately, it appears that none of these conditions has been met. It seems apparent that faith in an antiquated method of analysis led to neglect of the necessary steps in the analysis. This leaves little justification for publication of the paper, except to show that hagfishes are very widely distributed.

**Keywords** Cladistic biogeography, hagfishes, ocean basins, panbiogeography, vicariance.

In a recent paper published in *Journal of Biogeography*, Cavalcanti & Gallo (2008) chose to analyse the global distribution of hagfishes (Myxinidae) using a biogeographical method proposed by Croizat (1958, 1964). For most biogeographers, that method has long been superseded by others. At the American Museum in New York, in the early 1970s, panbiogeography was combined with part of Hennig's phylogenetic method to give birth to vicariance biogeography. After about 10 years, the name was changed to cladistic biogeography and the latter remained the preferred approach by those biogeographers who did not recognize dispersal as an important process in the formation of biogeographical patterns (Briggs, 2007).

Cladistic biogeography was a relatively popular movement until the late 1990s, when an outpouring of work on molecular genetics began to have its effect. In more recent years, it has become obvious that most of the distributions of contemporary clades, which vicarianists had attributed to the fractionation of Gondwana, had actually taken place via dispersal in the Tertiary or in more recent times. Cladistic (vicariant) biogeography has declined, primarily because its followers do not recognize the kind of allopatric speciation that takes place when members of a population migrate across a barrier to colonize a new area. The modern approach to biogeography is an eclectic one, recognizing the importance of both vicariance and dispersal, and is based on clues to be found within the relationship of the group concerned and in the history of its territory.

The aim of the authors was to correlate the hagfish distribution patterns with the tectonic history of the ocean basins. Why shouldn't they do this? Granted, Myxinidae is a very old family extending back some 400 Myr, but does this mean that they could examine the databases for locality records of the *living species*, draw lines between those that occupy certain regions, and come up with information that reflects the history of the ocean basins? Certainly, the ages of the species that have been connected by the lines are critical. The molecular relationship suggests that the split between the two hagfish subfamilies (Myxininae and Eptatretinae) took place in the late Cretaceous or early Tertiary (Kuraku & Kuratani, 2006). The phylogeny published by Møller & Jones (2007), based on original and published DNA sequences, clearly indicates that the ages of the genera and species must be considerably younger than those of the subfamilies.

Data on the identification and location of the various species were extracted from portals such as FishBase and Ocean