

A peer-reviewed version of this preprint was published in PeerJ on 25 February 2016.

[View the peer-reviewed version](https://peerj.com/articles/1720) (peerj.com/articles/1720), which is the preferred citable publication unless you specifically need to cite this preprint.

Vavrek MJ. 2016. A comparison of clustering methods for biogeography with fossil datasets. PeerJ 4:e1720 <https://doi.org/10.7717/peerj.1720>

A comparison of clustering methods for biogeography with fossil datasets

Matthew J Vavrek

Cluster analysis is one of the most commonly used methods in palaeoecological studies, particularly in studies investigating biogeographic patterns. Although a number of different clustering methods are widely used, the approach and underlying assumptions of many of these methods are quite different. For example, methods may be hierarchical or non-hierarchical in their approaches, and may use Euclidean distance or non-Euclidean indices to cluster the data. In order to assess the effectiveness of the different clustering methods as compared to one another, a simulation was designed that could assess each method over a range of both cluster distinctiveness and sampling intensity. Additionally, a non-hierarchical, non-Euclidean, iterative clustering method implemented in the R Statistical Language is described. This method, Non-Euclidean Relational Clustering (NERC), creates distinct clusters by dividing the data set in order to maximize the average similarity within each cluster, identifying clusters in which each data point is on average more similar to those within its own group than to those in any other group. While all the methods performed well with clearly differentiated and well-sampled datasets, when data are less than ideal the linkage methods perform poorly compared to non-Euclidean based *k*-means and the NERC method. Based on this analysis, Unweighted Pair Group Method with Arithmetic Mean and neighbor joining methods are less reliable with incomplete datasets like those found in palaeobiological analyses, and the *k*-means and NERC methods should be used in their place.

A comparison of clustering methods for biogeography with fossil datasets

Matthew J. Vavrek¹

¹Royal Ontario Museum, Department of Natural History, 100 Queen's Park, Toronto,
Ontario M5S 2C6, Canada

Abstract

Cluster analysis is one of the most commonly used methods in palaeoecological studies, particularly in studies investigating biogeographic patterns. Although a number of different clustering methods are widely used, the approach and underlying assumptions of many of these methods are quite different. For example, methods may be hierarchical or non-hierarchical in their approaches, and may use Euclidean distance or non-Euclidean indices to cluster the data. In order to assess the effectiveness of the different clustering methods as compared to one another, a simulation was designed that could assess each method over a range of both cluster distinctiveness and sampling intensity. Additionally, a non-hierarchical, non-Euclidean, iterative clustering method implemented in the R Statistical Language is described. This method, Non-Euclidean Relational Clustering (NERC), creates distinct clusters by dividing the data set in order to maximize the average similarity within each cluster, identifying clusters in which each data point is on average more similar to those within its own group than to those in any other group. While all the methods performed well with clearly differentiated and well-sampled datasets, when data are less than ideal the linkage methods perform poorly compared to non-Euclidean based k -means and the NERC method. Based on this analysis, Unweighted Pair Group Method with Arithmetic Mean and neighbor joining methods are less reliable with incomplete

26 datasets like those found in palaeobiological analyses, and the *k*-means and NERC methods
27 should be used in their place.

28

29 Keywords: Adjusted Rand Index, biogeography, cluster analysis, ecological similarity,
30 palaeoecology

31

32 Introduction

33 Clustering, defined as “a classificatory method which optimizes intra-group
34 homogeneity” (Lance and Williams, 1967), is one of the most frequently used forms of
35 multivariate analysis in palaeoecology (Hammer et al., 2001). One of the areas in which cluster
36 analysis is commonly used is studying patterns of biogeography amongst species assemblages.
37 Cluster analysis has been used in palaeoecological studies on groups as diverse as vertebrates
38 (Shubin and Sues, 1991; Holtz, Jr. et al., 2004; Fröbisch, 2009; Gates et al., 2010; Noto and
39 Grossman, 2010; Donohue et al., 2013), invertebrates (Schwimmer, 1975; Clapham and James,
40 2008), foraminifera (Collins, 1993) and plants (LePage et al., 2003), and assemblages spanning
41 the Ediacaran (Clapham et al., 2003) to the Pleistocene (Wolfe, 2000). With the rise of large
42 datasets of fossil species occurrences [e.g. Paleobiology Database, MioMAP (Carrasco et al.
43 2005), FAUNMAP (Graham and Lundelius, Jr. 2010), NOW (Fortelius 2015); see Uhen et al.,
44 2013 for recent review] with hundreds or thousands of records, semi-automated methods such as
45 clustering are becoming more and more necessary to find underlying patterns in these highly
46 complex collections. As the use of cluster analysis in palaeobiology has steadily expanded, so
47 too have the types of methods used. Although the underlying purpose of these methods is the
48 same (i.e. to delimit different groups from one another), their approaches and assumptions are
49 often quite different. For example, some cluster analysis methods (e.g. Unweighted Pair Group
50 Method with Arithmetic Mean/UPGMA, neighbour-joining) use a hierarchical approach to
51 grouping data (James and McCulloch, 1990; Shi, 1993).

52 Other common methods include partitioning techniques, such as *c*-means or *k*-means,
53 which may try to optimize groups by minimizing relative distances based on a chosen index
54 (Hartigan and Wong, 1979). Although clustering methods may be widely used, their
55 effectiveness relative to one another is less well known, in particular with the often sparse

56 datasets used in palaeobiological studies. In order to examine the relative efficacy of these
57 different clustering methods with species occurrence data, a dataset where the "true" clustering
58 relationship is known is required. To generate multiple simulated datasets with established
59 clustering relationships, I created an R function which could create a species occurrence database
60 that could then be used to test the efficiency of the methods over a large number of trials.

61 In addition to the analysis of the various clustering methods commonly used, I also
62 describe here an R function for a non-Euclidean, non-hierarchical clustering method termed here
63 Non-Euclidean Relational Clustering (NERC), an iterative method that uses agglomerative
64 clustering with post-clustering optimization. The efficacy of this function is tested in comparison
65 to the more traditional methods.

66

67 **Materials and Methods**

68 **The NERC Function**

69 The NERC Function The algorithm's execution can be broken down into three distinct
70 steps [after Lance and Williams (1967)]: the initialization of clusters; the allocation of new
71 elements to a cluster; and finally an iterative reallocation process whereby the clusters are
72 optimized. The first step, initialization of the clusters, begins by sampling a number of elements
73 equal to the requested number of final clusters. Each of these selected samples is assigned
74 randomly to a different initial cluster. In the second step, the function searches for the greatest
75 similarity (smallest value in a dissimilarity matrix) between any unassigned sample and any
76 assigned sample. The unassigned sample with the highest similarity is assigned to the same
77 group as that which it shares the greatest similarity, similar to Single Linkage Clustering
78 Analysis (Gower and Ross, 1969). This process then repeats, until all samples are assigned to a
79 cluster. At the end of the second step, if any group has only one member the process restarts
80 from the first step.

81 As a final step, an optimization of the clusters is performed. To begin, each individual
82 sample within the entire set is assessed for its average similarity to every cluster. The similarity
83 is based on the average pairwise distance from a sample to every member of a cluster (excluding
84 the sample itself in the case of the cluster it had been assigned to). If a sample has a greater

85 similarity to another cluster other than the one it has been assigned to, the optimization routine
86 will reassign the sample to the cluster that it had the most similarity to. If more than one sample
87 is in a suboptimal cluster, only one sample, chosen at random, will be reassigned at a time. After
88 a sample has been reassigned, the average pairwise distances will be calculated again before
89 another sample is reassigned (if necessary). If all the samples are in the cluster with which they
90 have the greatest average similarity then the cycle is complete. At present, an upper limit of 1000
91 reassignments has been set so as to avoid an infinite loop if there is no solution where every
92 sample is in its optimal grouping. The process will find a local, but not necessarily global,
93 optimum by minimizing the overall dissimilarity within clusters. Because the method is heuristic
94 in nature, it is best to repeat the clustering process many times.

95

96 **Implementation of NERC**

97 The R Statistical Language (R Development Core Team, 2012) was used to implement the
98 NERC function. The R Language is cross platform, Open Source and free to use, and is widely
99 used in statistical research, making it easy to extend with new functions and packages. The
100 package fossil (Vavrek, 2011) with all of the functions discussed in this paper is available
101 through the Comprehensive R Archive Network (CRAN) at [http://cran.r-](http://cran.r-project.org/web/packages/fossil/)
102 [project.org/web/packages/fossil/](http://cran.r-project.org/web/packages/fossil/). All data analysis and figure creation was done using R v3.2.1
103 on a Mac OS X 10.10 system. For a full copy of the R code used in the calculations and figures,
104 please consult the Supplementary Materials.

105 The R implementation of the NERC function has one required and three optional
106 arguments, and takes the form:

107

```
108 rclust(dist, clusters = 2, rand = 1000, counter = FALSE)
```

109

110 The only required argument is a distance or dissimilarity matrix (the dist argument), either as a
111 full matrix or lower triangle. The first optional argument (clusters) is the number of groups to be
112 created. The number of groups used must be a positive integer equal to or greater than 2 but no
113 greater than 1/2 the total number of samples. The minimum value represents the smallest number
114 of clusters without placing all samples within one group, and the maximum value prevents

115 clusters of one. The default value for the number of clusters is set to 2. The second optional
116 argument gives the number of times the clustering process should be run. Because the method
117 should be run many times to have a better chance of finding the global optimal solution, this
118 option has a default value of 1000. The last optional argument (counter) specifies whether to
119 print the current run. Note that at this point the R function returns only the result with the
120 smallest average within group distances overall.

121

122 **Data Simulation and Comparisons**

123 In order to test the efficacy of NERC in comparison to several other cluster methods, I
124 also created a simple function to simulate a species abundance data set. This function, called
125 `sim.occ()`, creates a matrix of sites (columns) and species (rows) with a known clustering
126 solution. The number of species, localities, regions (clusters), sample size and proportion of
127 regional endemism can all be adjusted. Each specific 'region' in the simulated set contains a
128 number of 'cosmopolitan' species that are found in every region, as well as 'endemic' species
129 that are found in only that particular region. To obtain a sample for a single locality, a
130 randomized log-normal distribution is applied to the total possible species pool for a given
131 region; the parameters are set so that any given locality will have several abundant species, a
132 large number of less common species, and some species which are not present. A log-normal
133 distribution was used as it is one of the most common species abundance distributions found in
134 empirical samples of modern habitats (Preston 1962; Gaston and Blackburn 2000; Magurran
135 2004). For every sample a new randomized log-normal distribution was created from the parent
136 region species pool. The average number of specimens can be varied to simulate different
137 sampling intensities. The full R code for the function can be found within the fossil package.

138 The simulated data was clustered using 6 different combinations of methods and input
139 matrices: single linkage, complete linkage, UPGMA, *k*-means on a db-RDA ordination using
140 both Euclidean and a non-Euclidean distance measure, and NERC. For those methods that
141 provide hierarchical clusters, discrete clusters were made using the `cutree` function. The db-RDA
142 ordination was performed using the `capscale` function in the `vegan` (Oksanen et al., 2011)
143 package.

144 Most functions used require a distance matrix as input, rather than raw species values. In

145 order to convert the occurrence matrices to dissimilarity matrices, the `ecol.dist()` function was
146 used, with the Sørensen (sometimes called Dice) dissimilarity index used to calculate pairwise
147 dissimilarities. The Sørensen dissimilarity index was used because it is one of the most
148 commonly used indices, and is regarded as one of the most effective presence/absence
149 dissimilarity measures (Southwood and Henderson 2000; Magurran 2004). Although the
150 `sim.occ()` function did create abundance-based occurrence matrices, the use of the Sørensen
151 dissimilarity index is presence/absence based, in effect converting the data. Although discarding
152 abundance data is not generally recommended in actual analyses, presence/absence data is
153 typically more common in palaeontological datasets, so using the Sorenson dissimilarity index
154 created a more realistic scenario.

155 The six methods were tested to see how well they performed both with varying levels of
156 endemicity (or differentiation between clusters; Fig. 1) as well as with varying levels of sampling
157 intensity. A simulated occurrence matrix was created 1000 times for each level of differentiation
158 or sampling intensity, and then clustered to obtain averaged performance values for all five
159 clustering methods. Each of the simulations consisted of 30 samples from 3 different endemic
160 regions, for a total of 90 samples to be used in the cluster analysis. Because of the parallel nature
161 of this simulation, the `multicore` (Urbanek, 2011), `foreach` (Revolution Analytics, 2011b), and
162 `doMC` (Revolution Analytics, 2011a) parallel computing packages for R were also used. The
163 visualization of cluster distinctiveness in Fig. 1 was created using the `NMDS` function provided
164 by the `ecodist` package (Goslee and Urban, 2007).

165 For the simulated biogeographic datasets, the “true” clustering was known, and so the
166 results of each clustering method could be compared to this *a priori* grouping. The Rand Index
167 (Rand, 1971; Hubert and Arabie, 1985) is method to compare two clustering outcomes and
168 calculates an index of similarity, with a value of 1 being a perfect match. The original formula
169 for this index, however, had a lower bound that fluctuated, depending on group sizes and
170 numbers (Hubert and Arabie, 1985). A modification of this original formula, given by Hubert
171 and Arabie (1985), scaled the value so that the greatest mathematically possible difference would
172 always be 0, with the upper bound still set to 1. This modification is referred to as the Adjusted
173 Rand Index (ARI). In the `fossil` package, both functions are provided, although only the ARI is
174 used to calculate the effectiveness of the clustering methods in this paper.

175

176 Results

177 Overall, the NERC and non-Euclidean k -means methods were the most effective at
178 recovering the original groupings across the different levels of regional endemism (Fig. 2), with
179 the NERC slightly outperforming the non-Euclidean k -means. Using a Euclidean distance metric
180 for the k -means method, even when the rest of the method and dataset are kept the same, led to a
181 notable reduction in performance. Complete linkage and UPGMA were readily able to recover
182 the correct clusters when the groups were relatively distinct. However, when the simulated
183 clusters were less distinct their effectiveness quickly declined. Single linkage clustering was least
184 effective and, produced unreliable results even at levels where all the other methods easily found
185 the proper clustering arrangement.

186 For the differing levels of sampling intensity (Fig. 3), the NERC method and non-
187 Euclidean k -means methods were again the most effective at recovering an accurate signal,
188 although in this instance the k -means was slightly more effective. Overall, complete linkage and
189 UPGMA gave accurate results when sampling intensity was high, but their performance was
190 very poor with sparsely sampled data. Single linkage was again the least effective of all the
191 methods tested.

192

193 Discussion

194 All cluster methods performed well when clusters were very distinct and sampling
195 intensity was high. However, in cases where biogeographic clusters were less distinct or
196 sampling was poor, the db-RDA/ k -means and the NERC methods were best able to recover the
197 original clusters compared to the other tested clustering methods. Among the other clustering
198 methods, single linkage performed the poorest of any of the methods. The notably poor
199 performance of the single linkage method was likely the result of individual samples that were
200 extremely distant from all others placed at the base of the tree, and because I applied a strict tree
201 cutting method with the hierarchical methods to obtain discrete clusters, the tree cutting method
202 then identified this single distant sample as an individual cluster. However, the treatment of
203 outliers is challenging in all clustering approaches, and their exclusion may not be possible or
204 desirable. A similar situation, where outliers have an undue influence on group composition, is

205 likely why complete linkage and UPGMA are also less effective than k -means or NERC.

206 These hierarchical methods are well suited to applications such as phenetic analyses or
207 phylogenetics, where a single ancestor (theoretically) gives rise to multiple descendants.
208 However, this one-to-many structure often translates poorly to species occurrence data sets like
209 those commonly used in biogeographic studies, where individual lineages may be operating in
210 parallel and independently (Brown, 1999). Individual species may originate in different locations
211 and disperse by various methods to new regions (Brown, 1999), leading to a more reticulate,
212 many-to-many relationship. In this case, a method that does not enforce a hierarchy may better
213 represent the relationships present.

214 Further, species occurrence data is typically non-Euclidean in nature. Whereas all the
215 cells in a phylogenetic data matrix represent a directly observed value, in a species occurrence
216 matrix any cell that has a zero value may be due to either the species not occurring in that area or
217 incomplete sampling, two possibilities that may be indistinguishable from one another. To deal
218 with incomplete sampling, most species occurrence data sets are converted into a distance
219 matrix, where the species composition of each sample is compared to every other sample using
220 an index of similarity (or dissimilarity); yet, while most of these measures provide some measure
221 of distance, these distances are not necessarily Euclidean (Gower and Legendre, 1986). The
222 benefit of using non-Euclidean measures over Euclidean distances is readily observable in this
223 study, with the non-Euclidean based k -means outperforming the Euclidean based k -means.

224 Although for this study, the Sørensen dissimilarity index was used, the choice of which
225 non-Euclidean dissimilarity index to use is not necessarily straightforward (e.g. Shi 1993;
226 Magurran 2004; Alroy 2015). By some counts, dozens of different dissimilarity indices have
227 been proposed in the literature (Hubálek 1982; Pielou 1984; Shi 1993), although only a handful
228 of these have entered into common use (Magurran 2004). While alternative methods, such as a
229 recent modification to the Forbes metric (Alroy 2015), have been proposed as replacements to
230 more traditional dissimilarity metrics, the choice of measure is a separate question to the issue in
231 the present study. Although using other dissimilarity measures may have changed the individual
232 effectiveness of the different clustering methods, the relative performance of the clustering
233 methods to each other is unlikely to change, as even with different measures the problems of
234 outliers and hierarchical/non-hierarchical methods would persist.

235 Both poor differentiation between clusters and inadequate sampling are common

236 problems with palaeobiological data. No method is entirely immune to either of these issues, but
237 overall, based on these simulations, *k*-means and NERC give more reliable and accurate results
238 when data are less than robust. Using these methods still does make one strong assumption about
239 the underlying data - namely, that true divisions within the data exist. Unfortunately, with the
240 often muddled and noisy nature of biogeographic data, this assumption is also the hardest to
241 objectively determine.

242 **Acknowledgments**

243 I would like to thank Caleb M. Brown, Nicolás E. Campione, Luke B. Harrison, Pat A. Holroyd
244 and Brian D. Rankin for their critical feedback which improved the quality of this paper, and
245 Callum G. Vavrek and Ada E. Vavrek for their assistance in editing the manuscript.

246

247 **References**

- 248 Alroy, J. (2015). A new twist on a very old binary similarity coefficient. *Ecology*, 96: 575–586.
249 doi: 10.1890/14-0471.1.
- 250 Brown, J. H. (1999). Macroecology: Progress and Prospect. *Oikos*, 87(1):3–14.
- 251 Carrasco, M. A., Kraatz, B. P., Davis, E. B., and Barnosky, A. D. (2005). Miocene Mammal
252 Mapping Project (MIOMAP). University of California Museum of Paleontology,
253 <http://www.ucmp.berkeley.edu/miomap/>
- 254 Clapham, M. E. and James, N. P. (2008). Paleoecology Of Early-Middle Permian Marine
255 Communities In Eastern Australia: Response To Global Climate Change In the Aftermath
256 Of the Late Paleozoic Ice Age. *Palaios*, 23(11):738–750.
- 257 Clapham, M. E., Narbonne, G. M., and Gehling, J. G. (2003). Paleoecology of the oldest known
258 animal communities: Ediacaran assemblages at Mistaken Point, Newfoundland.
259 *Paleobiology*, 29(4):527–544.
- 260 Collins, L. S. (1993). Neogene Paleoenvironments of the Bocas del Toro Basin, Panama. *Journal*
261 *of Paleontology*, 67(5):699–710.
- 262 Donohue, S. L., Wilson, G. P., and Breithaupt, B. H. (2013). Multituberculates of the Black
263 Butte Station local fauna (Lance Formation, southwestern Wyoming), with implications for
264 compositional differences among mammalian. *Journal of Vertebrate Paleontology*,
265 33(May):677–695.
- 266 Fortelius, M. (coordinator). (2015). New and Old Worlds Database of Fossil Mammals (NOW).
267 University of Helsinki. <http://www.helsinki.fi/science/now/>.
- 268 Fröbisch, J. (2009). Composition and similarity of global anomodont-bearing tetrapod faunas.
269 *Earth- Science Reviews*, 95(3-4):119–157.
- 270 Gaston, K. J., and Blackburn, T. M. (2000). Pattern and Process in Macroecology. Blackwell
271 Science, Ltd. doi: 10.1002/9780470999592.
- 272 Gates, T. A., Sampson, S. D., Zanno, L. E., Roberts, E. M., Eaton, J. G., Nydam, R. L.,
273 Hutchison, J. H., Smith, J. A., Loewen, M. A., and Getty, M. A. (2010). Biogeography of
274 terrestrial and freshwater vertebrates from the late Cretaceous (Campanian) Western
275 Interior of North America. *Palaeogeography, Palaeoclimatology, Palaeoecology*, 291(3-
276 4):371–387.

- 277 Goslee, S. and Urban, D. (2007). The ecodist package for dissimilarity-based analysis of
278 ecological data. *Journal of Statistical Software*, 22(7):1–19.
- 279 Gower, J. C. and Legendre, P. (1986). Metric and Euclidean properties of dissimilarity
280 coefficients. *Journal of Classification*, 3(1):5–48.
- 281 Gower, J. C. and Ross, G. J. S. (1969). Minimum Spanning Trees and Single Linkage Cluster
282 Analysis. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 18(1):54–
283 64.
- 284 Graham, R. W., and Lundelius, Jr., E.L. (2010). FAUNMAP II: New data for North America
285 with a temporal extension for the Blancan, Irvingtonian and early Rancholabrean.
286 FAUNMAP II Database, <http://www.ucmp.berkeley.edu/faunmap/index.html>
- 287 Hammer, Ø., Harper, D. A. T., and Ryan, P. D. (2001). PAST: paleontological statistics software
288 package for education and data analysis. *Palaeontologia Electronica*, 4(1):9.
- 289 Hartigan, J. A. and Wong, M. A. (1979). A k -means clustering algorithm. *Applied Statistics*,
290 28(1):100– 108.
- 291 Holtz, Jr., T. R., Chapman, R. E., and Lamanna, M. C. (2004). Mesozoic Biogeography of
292 Dinosauria. In: Weishampel, D. B., Dodson, P. and Osmólska, H., eds. The Dinosauria,
293 Second Edition. University of California Press, 627–642.
- 294 Hubálek, Z. (1982). Coefficients of association and similarity, based on binary (presence-
295 absence) data: an evaluation. *Biological Reviews*, 57:669–689.
- 296 Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2(1):193–218.
- 297 James, F. C. and McCulloch, C. E. (1990). Multivariate analysis in ecology and systematics:
298 panacea or Pandora’s Box? *Annual Review of Ecology and Systematics*, 21(1):129–166.
- 299 Lance, G. N. and Williams, W. T. (1967). A general theory of classificatory sorting strategies: II.
300 Clustering systems. *The Computer Journal*, 10(3):271–277.
- 301 LePage, B. A., Beauchamp, B., Pfefferkorn, H. W., and Utting, J. (2003). Late Early Permian
302 plant fossils from the Canadian High Arctic: a rare paleoenvironmental/climatic window in
303 northwest Pangea. *Palaeogeography, Palaeoclimatology, Palaeoecology*, 191(3-4):345–
304 372.
- 305 Magurran, A. E. (2004). *Measuring Biological Diversity*. Blackwell, Oxford.
- 306 Noto, C. R. and Grossman, A. (2010). Broad-scale patterns of Late Jurassic dinosaur
307 paleoecology. *PLOS ONE*, 5(9):e12553.

- 308 Oksanen, J., Blanchet, F. G., Kindt, R., Legendre, P., O'Hara, R. G., Simpson, G. L., Solymos,
309 P., Stevens, M. H. H., and Wagner, H. (2011). *vegan*: Community Ecology Package.
- 310 Pielou, E. C. (1984). *The interpretation of ecological data*. Wiley, New York.
- 311 Preston, F. W. (1962). The canonical distribution of commonness and rarity: Part I. *Ecology*, 43:
312 185–215. doi: 10.2307/1931976.
- 313 R Development Core Team (2012). *R: A Language and Environment for Statistical Computing*.
314 R Foundation for Statistical Computing, Vienna, Austria. <http://www.R-project.org/>.
- 315 Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the*
316 *American Statistical Association*, 66(336):846–850.
- 317 Revolution Analytics (2011a). *doMC*: Foreach parallel adaptor for the multicore package.
- 318 Revolution Analytics (2011b). *foreach*: Foreach looping construct for R.
- 319 Schwimmer, D. R. (1975). Quantitative taxonomy and biostratigraphy of Middle Cambrian
320 trilobites from Montana and Wyoming. *Journal of the International Association for*
321 *Mathematical Geology*, 7(2):149–166.
- 322 Shi, G. R. (1993). Multivariate data analysis in palaeoecology and palaeobiogeography – a
323 review. *Palaeogeography, Palaeoclimatology, Palaeoecology*, 105(3-4):199–234.
- 324 Shubin, N. H. and Sues, H.-D. (1991). Biogeography of Early Mesozoic continental tetrapods:
325 Patterns and implications. *Paleobiology*, 17(3):214–230.
- 326 Uhen, M. D., Barnosky, A. D., Bills, B., Blois, J., Carrano, M. T., Carrasco, M. A., Erickson, G.
327 M., Eronen, J. T., Fortelius, M., Graham, R. W., Grimm, E. C., O'Leary, M. A., Mast, A.,
328 Piel, W. H., Polly, P. D., and S^ˆail^ˆa, L. K. (2013). From card catalogs to computers:
329 databases in vertebrate paleontology. *Journal of Vertebrate Paleontology*, 33(1):13–28.
- 330 Urbanek, S. (2011). *multicore*: Parallel processing of R code on machines with multiple cores or
331 CPUs.
- 332 Vavrek, M. J. (2011). *fossil*: palaeoecological and palaeogeographical analysis tools.
333 *Palaeontologia Electronica*, 14(1):1T.
- 334 Wolfe, A. P. (2000). Paleoecology of a 90,000-year lacustrine sequence from Fog Lake, Baffin
335 Island, Arctic Canada. *Quaternary Science Reviews*, 19(17-18):1677–1699.
- 336

337 **Figure captions**

338 **Figure 1.** Visualization of the changing endemism of clusters (i.e. distinctiveness) and how it
339 alters the clustering of sites in an NMDS plot for the simulated biogeographic data sets. 'e' is the
340 proportion of all species that are endemic to only one biogeographic region. A higher proportion
341 of endemics results in more distinctive clusters, while a lower proportion of endemics results in
342 less distinctive clusters.

343

344 **Figure 2.** Response of various clustering methods to the distinctiveness of clusters as given by
345 the proportion of endemics (i.e. a higher endemism creates more highly differentiated clusters).
346 The values for each method at any given level of endemism is the average Adjusted Rand Index
347 comparing the known solution and the calculated solution over 1000 simulations.

348

349 **Figure 3.** Accuracy of various clustering methods in response to changing levels of sampling
350 intensity (coverage). Overall, as sampling intensity decreases (to the right), clustering becomes
351 less reliable. The values for each method at any given level of sampling is the average Adjusted
352 Rand Index comparing the known solution and the calculated solution over 1000 simulations.

Figure 1(on next page)

Variation in group distinctiveness for simulated data.

Visualization of the changing endemism of clusters (i.e. distinctiveness) and how it alters the clustering of sites in an NMDS plot for the simulated biogeographic data sets. 'e' is the proportion of all species that are endemic to only one biogeographic region. A higher proportion of endemics results in more distinctive clusters, while a lower proportion of endemics results in less distinctive clusters.

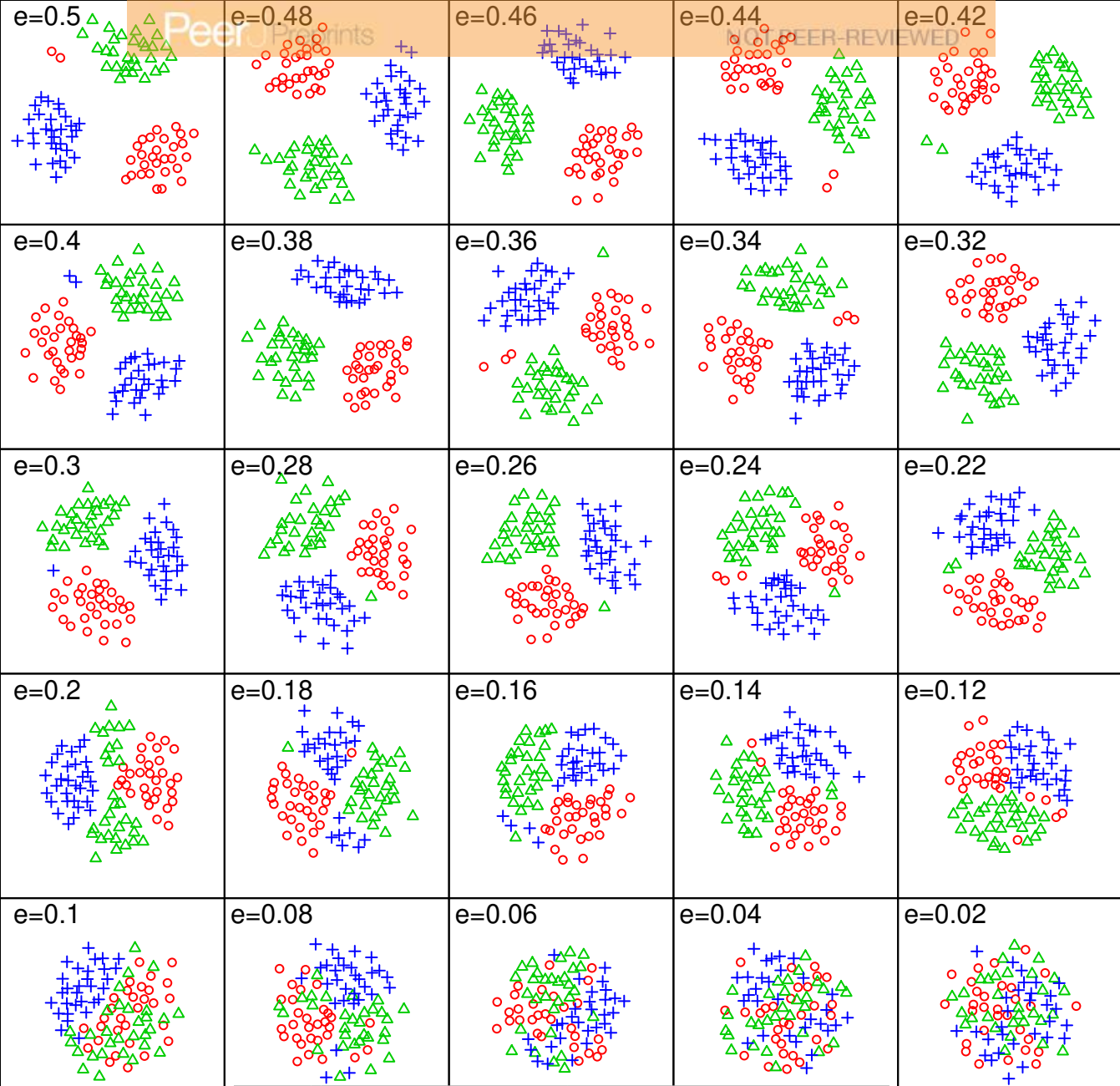


Figure 2(on next page)

Comparison of cluster methods with varying group distinctiveness.

Response of various clustering methods to the distinctiveness of clusters as given by the proportion of endemics (i.e. a higher endemism creates more highly differentiated clusters). The values for each method at any given level of endemism is the average Adjusted Rand Index comparing the known solution and the calculated solution over 1000 simulations.

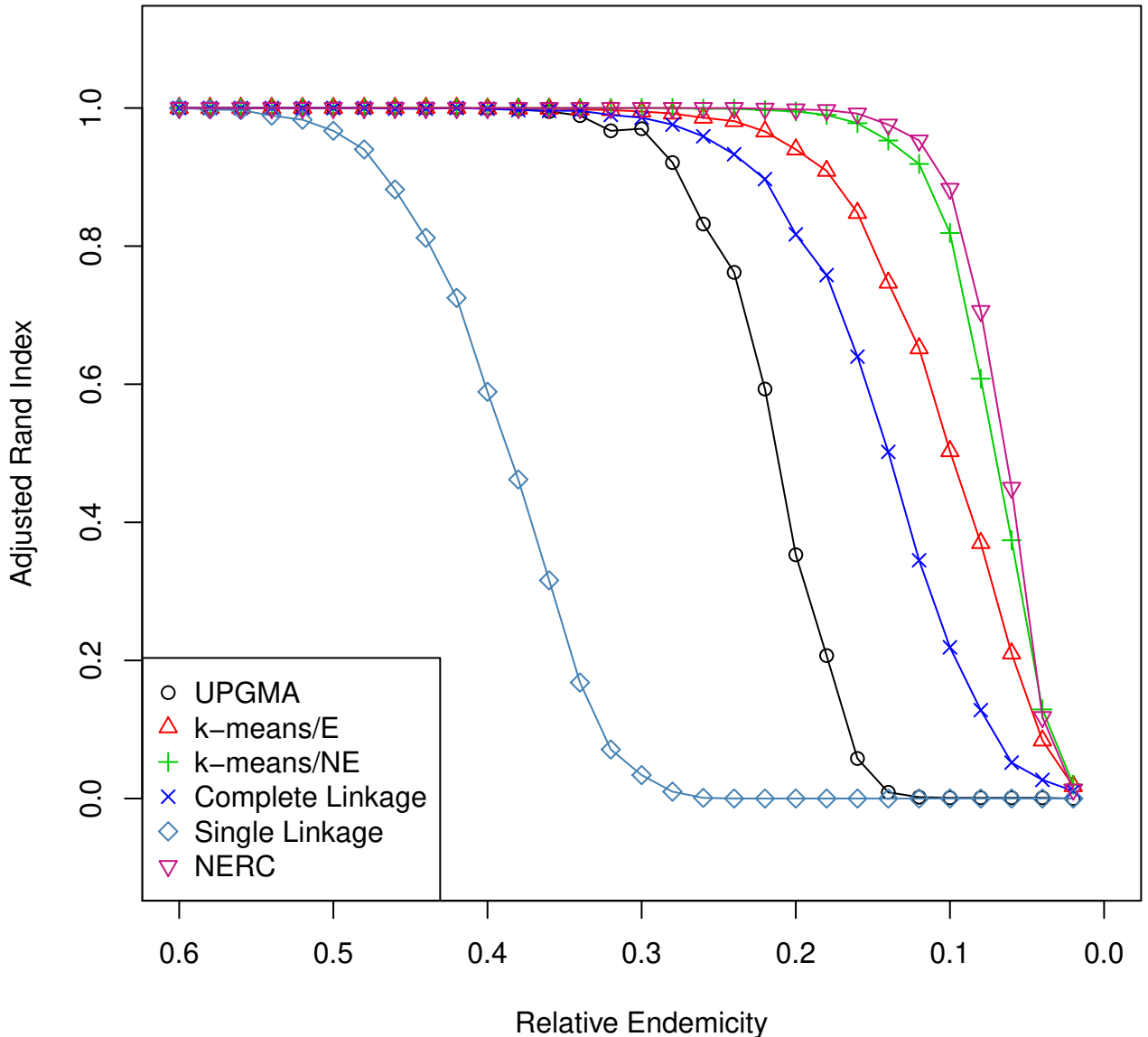


Figure 3(on next page)

Comparison of cluster methods with varying sampling intensity.

Accuracy of various clustering methods in response to changing levels of sampling intensity (coverage). Overall, as sampling intensity decreases (to the right), clustering becomes less reliable. The values for each method at any given level of sampling is the average Adjusted Rand Index comparing the known solution and the calculated solution over 1000 simulations.

