Title: Clustering in non-parametric multivariate analyses.

Article Type: Full Length Article

Corresponding Author: Dr. Paul J Somerfield,

Corresponding Author's Institution: Plymouth Marine Laboratory

First Author: K. Robert Clarke

Order of Authors: K. Robert Clarke; Paul J Somerfield; Raymond N Gorley

Abstract: Non-parametric multivariate analyses of complex ecological
datasets are widely used. Following appropriate pre-treatment of the
data inter-sample resemblances are calculated using appropriate measures.
Ordination and clustering derived from these resemblances are used to
visualise relationships among samples (or variables). Hierarchical
agglomerative clustering with group-average (UPGMA) linkage is often the
clustering method chosen. Using an example dataset of zooplankton
densities from the Bristol Channel and Severn Estuary, UK, a range of
existing and new clustering methods are applied and the results compared.
Although the examples focus on analysis of samples, the methods may also
be applied to species analysis. Dendrograms derived by hierarchical
clustering are compared using cophenetic correlations, which are also
used to determine optimum β in flexible beta clustering. A plot of
cophenetic correlation against original dissimilarities reveals that a
tree may be a poor representation of the full multivariate information.
UNCTREE is an unconstrained binary divisive clustering algorithm in which
values of the ANOSIM R statistic are used to determine (binary) splits in
the data, to form a dendrogram. A form of flat clustering, k-R
clustering, uses a combination of ANOSIM R and Similarity Profiles
(SIMPROF) analyses to determine the optimum value of k, the number of
groups into which samples should be clustered, and the sample membership
of the groups. Robust outcomes from the application of such a range of
differing techniques to the same resemblance matrix, as here, result in
greater confidence in the validity of a clustering approach.

PML | Plymouth Marine Laboratory

Listen to the ocean

16 June 2016

Dr. Sandra E. Shumway
Editor, Journal of Experimental Marine Biology and Ecology
Department of Marine Sciences
UCONN
1080 Shennecossett Road
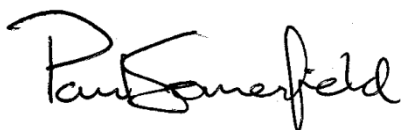Groton, CT 06340

Dear Sandy,

It gives me great pleasure to submit for consideration for publication in the journal our latest manuscript, entitled "Clustering in non-parametric multivariate analyses", quite out of character for Bob in the brevity of its title but, we believe, an important advance in terms of ecological methods for complex data.

Obviously, as required by the journal, all authors are happy to declare that they agree to this submission. I believe that it is all in order, and look forward to hearing from you in due course.

Actually, Bob asked me to submit this while he is on a much needed and hard-earned vacation, and in the course of doing so I nominated him as the corresponding author. I then found that I couldn't complete the submission as a key step was unavailable to me as a non-corresponding author. I tried twice before I realised this. Thus elsewhere there are two versions of this manuscript in the system, which need to be cleared. Hopefully that will be sorted once he returns. In the meantime, I'm happy to correspond and hope you like the manuscript.

Yours sincerely,

Paul J. Somerfield
Biodiversity
+44 (0) 1752 633100
http://www.pml.ac.uk/about_us/pml_people/paul_somerfield.aspx
http://scholar.google.co.uk/citations?user=ESTzLfAAAAAJ&hl=en&oi=ao

HIGHLIGHTS

Dendrograms may be poor representations of inter-sample dissimilarities

ANOSIM R and SIMPROF are combined to generate new methods of clustering

UNCTREE is a binary divisive clustering algorithm

k-R clustering is a flat clustering algorithm

Robustness of clustering is assessed by applying different methods to example data

1   Clustering in non-parametric multivariate analyses.

2

3   K. Robert Clarke [a,b], Paul J. Somerfield [a], Raymond N. Gorley [b]

4

5   [a] Plymouth Marine Laboratory, Prospect Place, West Hoe, Plymouth PL1 3DH, UK

6   [b] Primer-E Ltd, c/o Plymouth Marine Laboratory, Prospect Place, West Hoe, Plymouth PL1 3DH, UK

7

8   K. Robert Clarke, krc@pml.ac.uk +44 1752 633100, corresponding author

9   Paul J. Somerfield, pjso@pml.ac.uk

10   Raymond N. Gorley, tech@primer-e.com

11

12   HIGHLIGHTS

13   Dendrograms may be poor representations of inter-sample dissimilarities

14   ANOSIM R and SIMPROF are combined to generate new methods of clustering

15   UNCTREE is a binary divisive clustering algorithm

16   *k-R* clustering is a flat clustering algorithm

17   Robustness of clustering is assessed by applying different methods to example data

18

19

20   ABSTRACT

21   Non-parametric multivariate analyses of complex ecological datasets are widely used. Following

22   appropriate pre-treatment of the data inter-sample resemblances are calculated using appropriate

23   measures. Ordination and clustering derived from these resemblances are used to visualise

24   relationships among samples (or variables). Hierarchical agglomerative clustering with group-

25   average (UPGMA) linkage is often the clustering method chosen. Using an example dataset of

26   zooplankton densities from the Bristol Channel and Severn Estuary, UK, a range of existing and new

27   clustering methods are applied and the results compared. Although the examples focus on analysis

28   of samples, the methods may also be applied to species analysis. Dendrograms derived by

29   hierarchical clustering are compared using cophenetic correlations, which are also used to

30   determine optimum β in flexible beta clustering. A plot of cophenetic correlation against original

31   dissimilarities reveals that a tree may be a poor representation of the full multivariate information.

32   UNCTREE is an unconstrained binary divisive clustering algorithm in which values of the ANOSIM R

33   statistic are used to determine (binary) splits in the data, to form a dendrogram. A form of flat

34   clustering, k-R clustering, uses a combination of ANOSIM R and Similarity Profiles (SIMPROF)

35   analyses to determine the optimum value of k, the number of groups into which samples should be

36   clustered, and the sample membership of the groups. Robust outcomes from the application of such

37   a range of differing techniques to the same resemblance matrix, as here, result in greater confidence

38   in the validity of a clustering approach.

39

40   KEYWORDS: Non-parametric multivariate; divisive clustering; flat clustering; SIMPROF; cophenetic

41   correlation; cophenetic distance

42

1. Introduction

Field et al. (1982) described a robust non-parametric multivariate strategy for the analysis of biological assemblage data, such as the abundance or biomass of taxa in samples. Collins and Williams (1982) present one of the first applications of the strategy, to plankton data from the Bristol Channel and Severn Estuary. In essence the strategy, expanded and clarified by Clarke (1993), is to display patterns among samples determined by appropriate resemblance measures (Clarke et al., 2006) using clustering and ordination, and to analyse these patterns using a range of hypothesis tests and associated analyses, primarily based on ranked resemblances. Additional analyses are constantly added to the framework. Clarke et al. (2008) described a method for divisive clustering constrained by thresholds in explanatory variables, Linkage Trees, and Similarity Profiles analysis (SIMPROF) which tests for multivariate structure within groups of samples. The latter was further discussed by Somerfield and Clarke (2013) in the context of species (r-mode) analysis.

Literally hundreds of clustering methods exist, some of them operating on resemblance matrices whilst others are based on the original data (Legendre and Legendre, 2012). Everitt (1980) and Cormack (1971) give excellent and readable reviews, while Clifford and Stephenson (1975) is another well-established text from an ecological viewpoint. To cope with this variety a widely adopted approach has been to use a single technique that has been found to be of widespread utility in ecological studies while recommending the need to perform a cluster analysis in conjunction with a range of other techniques (e.g. ordination, statistical testing) to obtain balanced and reliable conclusions (Clarke et al., 2014).

Hierarchical clustering with group-average linking, based on sample similarities or dissimilarities such as the Bray-Curtis coefficient, has proved a useful technique in many ecological studies over the past half-century. As with clustering methods in general, it is appropriate for delineating groups of sites with distinct community structure. It is an agglomerative method. Agglomerative methods are bottom-up and 'see' only the nearby points throughout much of the process. When reaching the top of the dendrogram no possibility of taking a different view, of the main merged groups that have formed, remains. Binary divisive methods, however, are potentially advantageous for some clustering situations. They take a top-down view of the samples, so that the initial binary splits should (in theory) be better able to respect any major groupings in the data, since these are found first. However, as with all hierarchical methods, once a sample has been placed within one initial group it cannot jump to another at a later stage. Whilst divisive methods have the potential to produce marginally better solutions in practice, there is a counterbalancing downside to their

75 algorithms, in that they can be computationally intensive and complex (Gower, 1967), so iterative

76 approaches are generally required. The agglomerative approach, in contrast, is simple and entirely

77 determined, requiring nothing more than simple numerical operations based on values of

78 resemblance measures.

79 The purpose of this paper is to compare and discuss methods for, and associated with, cluster

80 analysis in a non-parametric multivariate framework. We look at some existing methods and

81 consider how their success, in conserving the inter-sample patterns in the underlying resemblance

82 matrix, may be assessed and compared. We go on to introduce new clustering methods and

83 compare their results. For the purpose of the paper we restrict ourselves to examples based on

84 analyses among samples, though it should be remembered that clustering of variables (taxa,

85 functional groups, OTUs, environmental measurements) is often entirely appropriate following

86 suitable pre-treatment of the data (Somerfield and Clarke, 2013).

87 **2. Material and methods**

88 *2.1 Hierarchical agglomerative clustering*

89 The most commonly used clustering techniques are hierarchical agglomerative methods. These

90 usually take a resemblance matrix (Clarke et al., 2006) as their starting point and successively fuse

91 the samples into groups, and the groups into larger clusters, starting with the highest mutual

92 similarities then lowering the similarity level at which groups are formed, ending when all samples

93 are in a single cluster. The result of a hierarchical clustering is generally presented as a tree diagram

94 or dendrogram. There is no firm convention for which way up a dendrogram should be portrayed

95 (increasing or decreasing resemblance values) or even whether the tree can be placed on its side,

96 but we will refer to the *x* axis as representing the full set of samples and the *y* axis defining a

97 resemblance level at which two samples or groups are considered to have fused. Neither is there

98 anything sacrosanct about the ordering of samples along the x axis, with the exception of constraints

99 imposed by the grouping structure among samples at higher levels in the tree.

100 *2.2. Linkage options*

101 Within hierarchical agglomerative clustering a range of linkage/sorting/joining options are defined

102 which determine how resemblances between samples and groups of samples are recalculated

103 following fusion of samples into a group. For single linkage (also called nearest-neighbour joining)

104 the dissimilarity of groups A and B, $\delta_{A\text{-}B}$, is the minimum across all dissimilarities between pairs of

105 samples with the first in A and the second in B. The dissimilarity of a group C to two merged groups

106 A and B, δ_{C-AB}, is therefore just the minimum of δ_{C-A} and δ_{C-B}. For complete linkage (also called

107 farthest-neighbour joining), δ_{C-AB} is the maximum of δ_{C-A} and δ_{C-B}. In group-average linkage δ_{A-B} is the

108 simple (unweighted) average over all dissimilarities from A to B pairs, leading to the acronym

109 UPGMA, Unweighted Pair Group Method with Arithmetic mean. When A and B are of different sizes,

110 it follows that, under UPGMA, δ_{C-AB} is a weighted average of δ_{C-A} and δ_{C-B}, e.g. giving more weight to

111 δ_{C-A} if there are more samples in A than B. Somewhat confusingly, the simple average of δ_{C-A} and δ_{C-B}

112 is then referred to as weighted linkage, WPGMA, since it weights the original dissimilarities between

113 samples in C and those in the combined group A and B unequally.

114 Other linkage options have been suggested. One is the flexible beta method of Lance and Williams

115 (1967), in which $\delta_{C-AB} = (1 - \beta) [(\delta_{C-A} + \delta_{C-B})/2] + \beta\delta_{A-B}$. Only negative values of β in the range (-1, 0)

116 make much sense in theory, the effect of including the δ_{A-B} term then being to make the merged AB

117 group more likely to join with the group C, the further A and B themselves are from each other. That

118 is, there will be a tendency to merge loosely bound samples or groups with each other, leaving

119 tightly bound groups separate. Lance and Williams (1967) suggest the use of β = -0.25, for which the

120 flexible beta has affinities with Gower's median method (Gower 1967). If β = 0, $\delta_{C-AB} = (\delta_{C-A}+\delta_{C-B})/2$,

121 which is the WPMGA method given above, also known as McQuitty's (1967) linkage.

122 Within a non-parametric multivariate analytical framework it might be expected that a linkage

123 option that is a function only of the ranks in the underlying resemblance matrix would be preferred.

124 Single linkage does this, but experience shows that it leads to 'chaining' in the resulting dendrogram,

125 with samples continuously joined to the next most similar sample without forming discrete clusters.

126 Complete linkage, conversely, tends to result in starkly separated, compact clusters. Group average

127 linkage will find a seemingly reasonable balance between the two. In order to choose between

128 linkage methods and their associated dendrograms a more objective means than simple visual

129 comparison of dendrograms is clearly needed.

2.3. Cophenetic correlation

131 One objective approach is provided by cophentic correlation, which is a (Pearson) matrix correlation

132 between each original dissimilarity and the (vertical) distance through a dendrogram to the common

133 node of the corresponding pair of samples (Jain and Dubes, 1988). If the *y*-axis of the dendrogram is

134 a dissimilarity scale then, naturally, these vertical distances are also dissimilarities. A dendrogram is a

135 good representation of the dissimilarity matrix, therefore, if the cophenetic correlation is close to 1.

136 As such the correlation may be seen as a way to compare different dendrograms, to assess the

137 performance of different analysis choices starting from the same dissimilarity matrix. In particular,

138 the correlation may also be used to determine β for the flexible beta method, computing a range of

139 values and choosing that which maximises the cophenetic correlation.

140 *2.4. Binary divisive clustering*

141 In hierarchical agglomerative clustering, samples start in separate groups and are successively

142 merged until, at some level of similarity, all are considered to belong to a single group. Hierarchical

143 divisive clustering does the converse operation: samples start in a single group and are divided into

144 two sub-groups, which may be of quite unequal size, each of those being further sub-divided into

145 two (i.e. binary division), and so on. Ultimately, all samples become singleton groups unless

146 (preferably) some criterion is applied to stop further sub-division of any specific group. Clarke et al.

147 (2008) describe such a clustering method (LINKTREE), which derives a binary divisive dendrogram

148 from a resemblance matrix. Divisions are constrained by thresholds on individual explanatory

149 variables, and the criterion for assessing the 'best' division at each step is to choose the one which

150 maximises the ANOSIM R statistic, defined as the difference between the average of the rank

151 dissimilarities between the (two) groups and within the groups. This is suitably scaled by a divisor of

152 $M/2$, where $M = n(n-1)/2$ is the total number of dissimilarities calculated between all the n samples

153 currently being split. This divisor ensures that R takes its maximum value of 1 when the two groups

154 are perfectly separated, defined as all between-group dissimilarities being larger than any within-

155 group ones. R will be approximately zero when there is no separation of groups at all, but this will

156 never occur in this context since the groups are chosen to maximise the value of R. There is an

157 important point not to be missed here: R is in no way being used as a test statistic, the reason for its

158 development (Clarke and Green, 1988). Instead, its value provides a pure measure of separation of

159 groups of points represented by the high-dimensional structure of the resemblances (here perhaps

160 Bray-Curtis, but any coefficient can be used with R, including Euclidean distance). This provides a

161 universal scaling of between vs. within group dissimilarities/distances (whatever their measurement

162 scale) through their reduction to simple ranks. A stopping rule is provided naturally here by the

163 SIMPROF test: if there is no demonstrable structure within a group, i.e. the null hypothesis for a

164 SIMPROF test cannot be rejected, then that group is not further sub-divided (Clarke et al., 2008).

165 The constrained divisive LINKTREE approach (typically used to 'explain' a biotic clustering by a series

166 of inequalities on individual environmental variables) permits examination of only a relatively small

167 number of the possible ways of splitting a single group into two sub-groups. A division is only

168 considered if one of the constraining variables takes larger values for all the samples in one sub-

169 group than the other. This reduces the number of choices at each step to a maximum of $p(n-1)$,

170 where p is the number of constraining variables and n the number of samples in the group currently

171 under division, and full calculation for all valid choices is straightforward. In contrast, an

172 unconstrained divisive approach needs, for each of the current groups, a full (binary) flat clustering,

173 namely computation of some optimality criterion (here the maximisation of ANOSIM R) for all

174 possible binary divisions of that group of n samples. There are $2^{n-1} - 1$ possibilities and for even quite

175 modest n (say >25) evaluating all of them quickly becomes prohibitive. This necessitates an iterative

176 search procedure (a variant of a standard "k-means" algorithm; MacQueen 1967), starting from an

177 arbitrary allocation of samples to the two sub-groups. In turn, each sample is then re-allocated to

178 the opposite group if this increases R, or left where it is if that would reduce R, and that process

179 continues until an entire loop has passed in which no sample changes its allocation. Such a

180 converged solution will typically vary depending on the starting configuration, so the whole iteration

181 must be repeated many times from different starting configurations. The 'best' of the divisions from

182 these different random restarts is then selected as likely, though not guaranteed, to be the optimal

183 solution. The binary divisive procedure (UNCTREE) starts with all samples in a single group, and if a

184 SIMPROF test provides evidence that the group has structure which can be further examined, a

185 search is made for an optimal split of those samples into two groups, maximising R. This could

186 produce anything from splitting off a singleton sample through to an even balance of the sub-group

187 sizes. The SIMPROF test is then repeated for each sub-group and this may justify a further split, again

188 based on maximising R, but now calculated having re-ranked the dissimilarities in that sub-group.

189 The process repeats until SIMPROF cannot justify further binary division on any branch: groups of

190 two are therefore never split as there is no sensible SIMPROF test for two samples (Clarke et al.,

191 2008).

192 The resulting groupings may be plotted in a dendrogram, but it is not readily apparent what values

193 on the y axis should be assigned to the various divisions. Unlike hierarchical agglomerative

194 clustering, in which fusions occur at increasing levels of similarity, each division in UNCTREE is

195 determined by a maximum value of R, which may be higher or lower than values at other levels in

196 the same analysis. There are two sensible approaches here. The first is to simply plot each level on

197 an arbitrary scale with equal spacing between levels and the level of subsequent divisions set so that

198 they sum to 100. This A% scale allows the plotting of the dendrogram, though gives little sense of

199 the relative importance of the divisions. An alternative is to use the B% scale introduced by Clarke et

200 al. (2008) for LINKTREE, calculating the average of between-group dissimilarities for each division

201 and dividing this by the maximum value it can take for a perfect split (R = 1) on the first division, and

202 then multiplying by 100 to give a positive scale which never exceeds 100. This scale measures how

203 well separated the two groups of samples are in the current division relative to the maximum

204 separation possible in the first division.

205 *2.5. k-R clustering (non-hierarchical)*

206 Another major class of clustering techniques is non-hierarchical, referred to above as flat clustering.

207 The desired number of clusters (k) must be specified in advance, and an iterative search attempts to

208 divide the samples in an optimal way into k groups, in one operation rather than incrementally. The

209 classic method is k-means clustering, which seeks to minimise within-group sums of squares about

210 the k group centroids. This idea, appropriate only for Euclidean distance matrices, can again be

211 generalised to apply to any resemblance measure, e.g. Bray-Curtis, by maximising ANOSIM R, which

212 measures (non-parametrically) the degree of overall separation of the k groups, formed from the

213 ranks in the full resemblance matrix. By analogy with k-means clustering, the principle of maximising

214 R to obtain a k-group division of the samples is referred to as k-R clustering, and it again involves an

215 iterative search, from several different random starting allocations of samples to the k groups.

216 The k-group solution will not, of course, simply split one of the groups in the ($k − 1$)-group solution:

217 there could be a wide-scale rearrangement of many of the points into different groups. A widely

218 perceived disadvantage of the k-means idea is the need to specify k before entering the routine.

219 Another, if it is re-run for many different k values, is the absence of a convenient visualisation of the

220 clustering structure for differing values of k, analogous to the hierarchy of a dendrogram. The output

221 of the solution is simply a factor denoting which samples belong to the same groups. This has tended

222 to restrict its use to cases where there is a clear *a priori* idea of the approximate number of groups

223 required, perhaps for operational reasons (e.g. in a quality classification system). However, the

224 SIMPROF test can also come to the rescue here, to provide an objective choice of k. Starting from a

225 low value for k (say 2) the two groups produced by k-R clustering are tested for evidence of within-

226 group structure by SIMPROF. If either of the tests are significant, the routine increments k (to 3),

227 finds the 3-group solution and retests those groups by SIMPROF. The procedure is repeated until a

228 value for k is reached in which none of the k groups generates significance in their SIMPROF test,

229 and the process terminates with that group structure as the best solution. (This will not, in general,

230 correspond to the maximum R when these optima for each k are compared across all possible k; R

231 must increase to its maximum of 1 as k approaches n, the number of samples.)

232 A variation of this flat-clustering procedure, rather than based on the R statistic, utilises the concept

233 of group-average linking met earlier, though still in a non-parametric setting. For a pre-specified

234 number of groups (k), each stage of the iteration process involves removing each sample in turn and

235 then allocating it to one of the other ($k − 1$) groups currently defined, or returning it to its original

236 group. In k-R clustering it is re-allocated to the group which produces the highest R value for the

237 resulting full set of groups. In the group-average rank variation, the sample is re-allocated to the

238 group to which it has the best rank similarity, defined as the average of the pair-wise rank

239 similarities between it and all members of that group (or all remaining members, in the case of its

240 original group). The process is then iterated until it converges and repeated a fair number of times

241 from different random starting allocations to groups, as before. The choice of k can use the same

242 SIMPROF procedure as previously.

2.6. Data

244 The data used here are from Collins and Williams (1982). Full details of the sampling method are

245 given in Collins and Williams (1981). In brief, the set of samples used here were collected in April

246 1974 from 57 sites (numbered 1-58, site 30 was not sampled) in the Bristol Channel and Severn

247 Estuary, UK (Fig. 1) by means of double-oblique plankton hauls. 24 holoplanktonic taxa were

248 identified and counted in the samples, and abundances were converted to densities (numbers.m^{-3})

249 using standard conversion factors. As in Collins and Williams (1982), data were then fourth-root

250 transformed and inter-sample pairwise resemblances were calculated using the Bray-Curtis similarity

251 measure.

3. Results and specific discussion

3.1. Hierarchical agglomerative clustering

254 Collins and Williams (1982) presented cluster analyses derived from the Bray-Curtis similarities using

255 hierarchical agglomerative clustering with group-average sorting. The resulting dendrogram (Fig.

256 2A) divides the samples into 4 groups; once each of the main groups has formed it remains separate

257 from other groups over a relatively large drop in similarity. Collins and Williams (1982) selected the

258 four groups determined at a 55% similarity level and characterised these as true estuarine (sites 1-8,

259 10, 12), estuarine and marine (9, 11, 13-27, 29), euryhaline marine (28, 31, 33-35, 42-44, 47-50, 53-

260 55) and stenohaline marine (32, 36-41, 45, 46, 51, 52, 56-58). It is not clear from the dendrogram

261 alone whether there is any natural sequence of community change across the four main clusters

262 (implicit in the designations true estuarine, estuarine and marine, euryhaline marine, stenohaline

263 marine). For example, the stenohaline marine group could just as correctly have been rotated to lie

264 between the estuarine and marine and euryhaline marine groups. In fact, there is a strong (and

265 more-or-less continuous) gradient of community change across the region, associated with changing

266 salinity levels.

267 More importantly, there is a degree of subjectivity in deciding the level of similarity at which

268 apparent divisions among groups of samples make sense. A visual inspection of the dendrogram

269 (Fig. 2A) suggests a sensible separation is at 55%, dividing the samples into 4 groups. An equally

270 justifiable decision could have been made to slice the dendrogram at the 50% level (Fig. 2A), giving 3

271 groups of samples. While there are methods available to decide at what level of similarity a division

272 of a dendrogram may be justified, it is not always the case that dividing all groups of samples at the

273 same level of similarity makes sense. It was to address these issues in an objective and robust

274 statistical manner that Clarke et al. (2008) initially developed Similarity Profiles (SIMPROF) analysis.

275 According to the terminology of Somerfield and Clarke (2013), the appropriate test here is Type 1

276 SIMPROF, and carrying out such a test at each node in the resulting dendrogram does indeed show

277 that the samples divide into 4 internally-homogeneous groups (Fig. 2B). It is convenient to represent

278 all splits down to single points, but the grey dashed lines indicate divisions where SIMPROF provides

279 no support for that sub-structure.

280 *3.2. Linkage options and cophenetic correlations*

281 This dendrogram has a cophenetic correlation of 0.797. A corresponding dendrogram constructed

282 using single linkage gives a correlation of 0.633, and with complete linkage a value of 0.722. Thus

283 the cophenetic correlations provide objective evidence for the guidance based on experience, that

284 group-average linkage usually provides a better view of the inter-sample relationships than analyses

285 based on other linkage options.

286 It is instructive to see how reliable this view may be. A plot of cophenetic distances (the vertical

287 distance to the first shared node between pairs of points; Jain and Dubes, 1988) against the

288 underlying dissimilarities (Fig. 3) shows that the overall pattern of dissimilarities are generally poorly

289 retained in the cluster analysis. Although, as indicated by the cophenetic correlation, increasing

290 dissimilarity tends to result in increasing distance (itself a dissimilarity on the same measurement

291 scale), there is also a strongly increasing trend in variability with the larger cophenetic distances

292 representing very wide ranges of dissimilarities in the dendrogram.

293 An analysis of cophenetic correlations for a range of β (Fig. 4) shows no compelling evidence for

294 preferring Lance and Williams' (1967) suggestion of a value of -0.25, for these data at least. In fact,

295 the optimum value of β, -0.04, is very close to zero suggesting that WPMGA is an appropriate linkage

296 choice. The cophenetic correlation from the resulting dendrogram is 0.793, slightly less than that
297 from UPMGA, and the analyses differ primarily in the placement of one sample, 23, from the
298 estuarine and marine group to the euryhaline marine group. Empirical evidence suggests that an
299 optimum flexible beta solution is usually inferior to group average linkage, perhaps as a result of the
300 failure of WPGMA-type solutions to weight δ_{CA} and δ_{CB} appropriately under averaging, for groups A
301 and B of very different sizes.

3.3. Binary divisive clustering (UNCTREE)

303 The tree diagram which results from unconstrained binary divisive clustering of the Bray-Curtis
304 resemblances for the 57 Bristol Channel zooplankton samples is given in Fig. 5, showing the two
305 alternative scalings (A% and B%) of the *y*-axis (Fig. 5A and 5B). As with the comparative
306 agglomerative clustering (Fig. 2) it is convenient to represent all splits down to single points, but the
307 grey dashed lines indicate divisions where SIMPROF provides no support for that sub-structure.
308 Visual comparison of the divisive and agglomerative trees (e.g. Fig. 2B and Fig. 5B) is not particularly
309 easy, though they have been manually rotated to aid this (a dendrogram is only defined down to
310 arbitrary rotations of its branches, in the manner of a 'mobile'). Clearly, however, only four groups
311 have been identified by the SIMPROF tests in both cases. The group constitutions have much in
312 common, though they are not identical. This is more readily seen from Fig. 6A & B, which use a non-
313 metric MDS plot to represent the community sample relationships in 2-d ordination space. It is clear
314 that only sites 9, 23 and 24 change groups between the two hierarchical clustering methods and
315 these all appear at the edges of their groups in both plots, which are thus reassuringly consistent
316 (bear in mind also that a 2-d MDS plot gives only an approximation to the true sample relationships
317 in higher dimensions, the MDS stress of 0.11 here being low but not negligible).

3.4. k-R clustering

319 Fig. 6C shows the optimum grouping produced by *k-R* clustering, superimposed on the same MDS
320 plot as for Figs 6A and B. The SIMPROF routine has again terminated the procedure with *k*=4 groups
321 (a to d), which are very similar to those for the two hierarchical methods, but with the three sites 9,
322 23 and 24 allocated to the four groups in yet a third way. This appears to be at least as convincing an
323 allocation as for either of the hierarchical plots (though do not lose sight of the fact that the MDS
324 itself is only an approximation to the real inter-sample resemblances). The group-average rank
325 variation of *k-R* clustering produces exactly the same four groups as seen in Fig 6C. This will not
326 always be the case, but it should be expected that these two variations will generally give closer
327 solutions to each other than to the hierarchical methods.

3.5. Numerical comparison of methods

329   As already seen, agglomerative hierarchical methods with different linkage options, which result in a

330   dendrogram with *y*-axis on a (dis)similarity scale, can be compared by means of a Pearson matrix

331   correlation between their cophenetic distances (dissimilarities) and the original dissimilarity matrix.

332   Here, as almost without exception, this demonstrated the superiority of group average (UPGMA)

333   over single or complete linkage strategies, and also over flexible beta, albeit more marginally.

334   However, comparison of UPGMA with the unconstrained binary divisive (UNCTREE) clustering is less

335   straightforward, because the latter can have somewhat arbitrary scaling on the y-axis of its tree

336   diagram. For example, for clarity of visualising the sequence of binary divisions, it is sometimes

337   preferable to use the equi-stepped scale (A%) for divisions on the *y*-axis (Fig. 5A), particularly where

338   there are reversals of direction in branch structures when using the B% scale. This does not happen

339   in Fig. 5B, but can occur, particularly with the constrained form of binary divisive clustering

340   (LINKTREE), if a clear division in the community samples has no constraining "explanation" in terms

341   of any of the recorded environmental variables (Clarke et al. 2008). This is avoided by the A% scaling

342   but *y*-axis values are now no longer comparable across separate branches, being dependent only on

343   the number of samples in that section of the tree. Contrasting Figs. 5A & B, it is clear that on the A%

344   scale even the rank order of the divisions on the *y*-axis is not maintained, so a non-parametric

345   correlation (e.g. Spearman) between cophenetic *y*-axis distances and original dissimilarities is not a

346   viable answer here for head-to-head comparison of these metric agglomerative and non-metric

347   divisive solutions.

348   Instead, there are two possible approaches, characterisable as "home" and "away" matches. Given

349   that both techniques have resulted in four SIMPROF groups, a "home" match (non-metric) for the

350   UNCTREE divisive method might compare the global ANOSIM *R* value computed from all four groups

351   of zooplankton samples displayed in Fig. 6A with that in Fig. 6B (this *R* value is computed from the

352   original high-d dissimilarity ranks of course, and not the 2-d nMDS ordination). Not unexpectedly,

353   this results in a "home" win, with *R* = 0.880 for the non-metric divisive UNCTREE and *R* = 0.870 for

354   the metric agglomerative UPGMA. The "away" match for the non-metric UNCTREE is more decisive.

355   This is to compare the cophenetic correlation for UPGMA, previously seen to be 0.797, with that for

356   the (rank-based) divisive tree displayed with (metric) group average dissimilarities computed at

357   every division, as if it were an agglomerative dendrogram. In other words, the *y* axis value for the

358   node at the top of each group in the divisive tree is the unweighted average of all pair-wise

359   dissimilarities between the two sub-groups of samples joined at that node. This is certainly playing a

360   match on the agglomerative method's home turf but the outcome is a marginal win for the divisive

361   method, with a (Pearson) cophenetic correlation of 0.812 with the original (metric) dissimilarities.

362   Can such comparisons be extended to the non-hierarchical k-R clustering method? Clearly, any

363   concept of cophenetic distance is impossible for a flat clustering method, with no form of sequential

364   clustering of the individual samples. But the comparison of global ANOSIM R statistics for the four

365   (slightly differing) groups identified by all three methods is perfectly viable, albeit it must be

366   considered the "home" match for the flat clustering method, since that should have optimised the

367   global ANOSIM R statistic among the k groups which result from the sequence of SIMPROF tests.

368   And indeed, for k-R clustering, $R = 0.884$, in comparison to $R = 0.880$ for UNCTREE and $R = 0.870$ for

369   UPGMA.

370   These numerical comparisons are therefore consistent with the earlier contention of a performance

371   ranking, with flat>divisive>agglomerative, though the fact that they rank in precisely the opposite

372   order in terms of ease of calculation, and the very marginal differences seen here and in Fig. 6,

373   suggest that such a ranking should not be taken too seriously.

374

4. General discussion

376   In this paper we introduce some methods that use ANOSIM R to determine the relative strengths of

377   different clustering of the same dissimilarities, in order to construct group structures. There are, of

378   course, classical equivalents of each method. For binary divisive clustering the criterion for

379   quantifying a good binary division is clearly central. Classically (e.g. Edwards and Cavalli-Sforza,

380   1965), ordinary (Euclidean) distance is regarded as the relevant resemblance measure, and the total

381   sums of squared distances of all points about the grand mean (overall centroid) is partitioned into a

382   combination of sums of squares within the two groups about their group centroids, and that

383   between the group centroids about the overall centroid (the same principles apply to one or many

384   dimensions and two or more groups). By minimising the within-group sums of squares, that between

385   groups is maximised, since the total sums of squares is fixed. For each group, Huygens theorem (e.g.

386   see Anderson *et al*, 2008) expresses those within-group sums of squares as simply the sum of the

387   squared Euclidean distances between every pair of points in the group, divided by that number of

388   points. In other words, the classic criterion minimises a weighted combination of within group

389   resemblances, defined as squared Euclidean distances. The classic flat clustering method is k-means

390   clustering (MacQueen, 1967; Steinhaus, 1957), which seeks to minimise within-group sums of

391 squares about the *k* group centroids (Lloyd, 1982). This is also equivalent to minimising some

392 weighted combination of within-group resemblances between pairs of samples, as measured by a

393 squared Euclidean distance coefficient. Thus the classical methods are really only appropriate in

394 situations where (squared) Euclidean distance is an appropriate measure of resemblance, such as

395 analyses of normalised environmental variables. For community analyses, appropriate measures of

396 resemblance such as those in the Bray-Curtis family (Clarke et al., 2006) are required, and

397 partitioning sums of squares is no longer a possibility. The new methods proposed here, UNCTREE

398 and *k-R* clustering, address the problem by maximising ANOSIM *R*, which measures (non-

399 parametrically) the degree of overall separation of the groups, formed from the ranks in the full

400 resemblance matrix. Thus they are appropriate for any data type and resemblance measure.

401 A 'take-home' message from Fig. 6 is that clustering rarely escapes a degree of arbitrariness: the

402 data simply may not represent clearly separated clusters. For the Bristol Channel sites, where there

403 certainly are plausible groups but within a more or less continuous gradation of change in plankton

404 communities (strongly correlated with increased salinity of the sites, Fig. 6B), different methods

405 must be expected to divide this continuum up in slightly different ways. It is important to remember

406 that, as applied here, SIMPROF tests groupings that are determined by the clustering algorithm and,

407 under different clustering techniques, samples can move from one group to another, both deemed

408 homogeneous by SIMPROF, so neither should be regarded as absolute. In spite of the above

409 evidence that the top-down methods can lead to marginally better solutions (notwithstanding the

410 vagaries of optimal search methods), use of a specific grouping from a bottom-up, group-average

411 hierarchy should probably be viewed operationally as on a par with that from a divisive hierarchy or

412 from the non-hierarchical *k-R* clustering, in either form. And, certainly, the group average (UPGMA)

413 approach appears uniformly superior to other linkage options in agglomerative clustering. It is

414 notable here that SIMPROF supports four very similar groups for all three methods shown in Fig. 6,

415 though this degree of consistency is not guaranteed for analyses of other sets of data. In fact,

416 especially in cases where a low-dimensional MDS plot is not at all reliable because of high stress, the

417 plurality of clustering methods may provide insight into the robustness of conclusions that can be

418 drawn about group structures from the (high-dimensional) resemblance matrix. Such comparisons of

419 differing clustering methods need to 'start from the same place', namely using the same

420 resemblance matrix, otherwise an inferred lack of a stable group structure could be due to the

421 differing assumptions being made about how the (dis)similarity between two samples is defined (e.g.

422 Bray-Curtis vs squared Euclidean distance). This is also a point to bear in mind in comparing

423 ordination methods: a primary difference between them is often not the way they choose to

424 represent high-dimensional information in lower dimensional space but how they define that higher-

425 dimensional information differently, in their choice of explicit or implicit resemblance measure

426 (Clarke et al., 2006). Comparison of results from clustering methods that utilise axis scores from

427 ordination methods (e.g. Hill, 1979; Lefkovitch, 1976; Noy-Meir, 1973) will be particularly influenced

428 by this issue. Given that the appropriate choice of resemblance is fundamental to a successful

429 analysis, many existing methods which use inappropriate measures, either implicitly or explicitly, are

430 to be avoided, or at best treated with caution.

431 It is often recommended that cluster analysis is best used in conjunction with ordination. Indeed,

432 that has been done here (Fig. 6) by labelling samples in the ordination plot with their group

433 membership from cluster analyses, a simple technique to apply in either 2- or 3-d ordination space.

434 Widely used alternatives are to draw smoothed convex hulls (contours) on a 2-d ordination,

435 indicating the samples grouped together at different levels in the dendrogram (nested contours), or

436 groups of samples which are not separated by SIMPROF, and so on. The implication is that both

437 clustering and ordination provide equally valid views of the relationships among samples, so if the

438 two methods being used in conjunction support each other then the investigator is in a position to

439 trust both. A novel alternative view of the relationship between a cluster analysis and an ordination

440 (Fig. 7) shows graphically the complex interrelationships between the linking of samples in the

441 former and the placement of samples in the latter. It is readily apparent that the beguiling simplicity

442 of a dendrogram may give a highly distorted view of true distances among objects.

443 In the statistical analytical framework under discussion the ordination method of choice is often

444 nMDS, which arrives at its solution through an iterative procedure which is not guaranteed to

445 achieve the optimal (global) solution with minimum stress. It is normal to rerun the procedure many

446 times and to check numerically (stress values, a measure which reflects lack of agreement in the

447 rank order of the distances among points in the ordination and the rank order of points in the

448 resemblance matrix) and graphically (Shepard diagrams, plots of distances in the solution against

449 original resemblances) before accepting that a nMDS plot is an accurate reflection of the true

450 multivariate information in the underlying resemblance matrix. It is interesting that, to date, such

451 rigour has rarely been applied to clustering. In part this reflects the numerical simplicity of widely

452 used methods, in that there is only one solution. What is rarely questioned is the adequacy of that

453 solution. Here we show how the adequacy of a hierarchical agglomerative clustering may be

454 checked numerically (cophenetic correlation) and graphically (plotting cophenetic distances against

455 dissimilarities, Fig. 3). It may surprise some to see how poor a representation of the true

456 relationships a dendrogram or tree diagram may be, and perhaps consideration should be given to

457 more appropriate methods. This may be particularly relevant in some areas of science where trees

458   are routinely used to represent relationships, such as in phylogenetics and microbial studies,

459   especially under conditions in which a pure hierarchical structure is not mandated (e.g. in genetics,

460   because of possibilities of recombination etc.), or in the derivation of hierarchical groupings of taxa

461   based on taxonomic information or traits for the calculation of certain indices of relatedness which

462   are defined in terms of branch lengths in a tree (see Somerfield et al., 2008; Warwick and Clarke,

463   2001).

464   The terms clustering and classification were originally synonymous and that is the way they still tend

465   to be used by ecologists, but in statistical language the methods considered here are all clustering

466   techniques. The term classification is now usually reserved for classifying unknown new samples into

467   known prior group structures. Nearly half a century ago, Cormack (1971) warned against the

468   indiscriminate use of cluster analysis, saying "availability of … [clustering] techniques has led to the

469   waste of more valuable scientific time than any other 'statistical' innovation". The ever larger

470   number of clustering techniques and their increasing accessibility on modern computer systems

471   makes this warning no less pertinent today. With that in mind it may seem rather unnecessary to

472   consider adding to the clustering techniques available, but the thrust of this paper is to demonstrate

473   how useful it can be to compare a range of different, but *compatible*, techniques when assessing the

474   robustness of a clustering solution. And, in order to facilitate that, we have introduced two new

475   methods: a divisive hierarchical and a non-hierarchical (*k*-means type) method, the latter in two

476   variants. These are now entirely compatible with a range of other widely-employed non-parametric

477   multivariate methods (e.g. nMDS ordination, ANOSIM tests, BEST or LINKTREE approaches to linking

478   community structure to environmental variables; Clarke, 1993; Clarke et al., 2008, 2014) because

479   they all start from exactly the same information, the rank orders of the entries in a defined

480   resemblance matrix. As a result, multiple clustering methods in conjunction with ordination have, for

481   the Bristol Channel zooplankton analyses, both directed attention to those few sites which are not

482   firmly allocated to one of the groupings, and instilled a greater degree of confidence in the

483   robustness of the main groupings produced.

484

495 **References**

496 Anderson, M.J., Gorley, R.N., Clarke, K.R., 2008. Permanova+ for Primer: guide to software and

497   statistical methods. PRIMER-E, Plymouth.

498 Clarke, K.R., 1993. Non-parametric multivariate analyses of changes in community structure. Aust. J.

499   Ecol. 18, 117-143.

500 Clarke, K.R., Green, R.H., 1988. Statistical design and analysis for a 'biological effects' study. Mar.

501   Ecol. Progr. Ser. 46: 213-226.

502 Clarke, K.R., Somerfield, P.J., Chapman, M.G., 2006. On resemblance measures for ecological studies,

503   including taxonomic dissimilarities and a zero-adjusted Bray–Curtis coefficient for denuded

504   assemblages. J. Exp. Mar. Biol. Ecol. 330, 55-80.

505 Clarke, K.R., Somerfield, P.J., Gorley, R.N., 2008. Testing of null hypotheses in exploratory community

506   analyses: similarity profiles and biota-environment linkage. J. Exp. Mar. Biol. Ecol. 366, 56-69.

507 Clarke, K.R., Gorley, R.N., Somerfield, P.J., Warwick, R.M., 2014. Change in marine communities: an

508   approach to statistical analysis and interpretation, 3rd edn. PRIMER-E, Plymouth.

509 Clifford, D.H.T., Stephenson, W., 1975. An introduction to numerical classification. Academic Press,

510   New York.

511 Collins, N.R., Williams, R., 1981. Zooplankton of the Bristol Channel and Severn Estuary. The

512   distribution of four copepods in relation to salinity. Mar. Biol. 64: 273-283.

513 Collins, N.R., Williams, R., 1982. Zooplankton communities in the Bristol Channel and Severn Estuary.

514   Mar. Ecol. Progr. Ser. 9: 1-11.

515 Cormack, R.M., 1971. A review of classification. J. R. Statist. Soc. Ser. A. 134: 321-367.

516 Edwards, A.W.F., Cavalli-Sforza, 1965. A method for cluster analysis. Biometrics 21: 362-375.

517 Everitt, B., 1980. Cluster Analysis, 2nd edn. Heinemann, London.

518 Field, J.G., Clarke, K.R., Warwick, R.M., 1982. A practical strategy for analysing multispecies

519   distribution patterns. Mar. Ecol. Progr. Ser. 8, 37-52.

520 Gower, J.C., 1967. A comparison of some methods of cluster analysis. Biometrics 23: 623-637.

521 Hill, M.O., 1979. TWINSPAN – A FORTRAN program for arranging multivariate data in an ordered

522   two-way table by classification of the individuals and attributes. Cornell University, Ithaca. 49 pp.

523 Jain, A.K., Dubes, R.C., 1988. Algorithms for clustering data. Prentice Hall, Englewood Cliffs. 320 pp.

524 Lance, G.N., Williams, W.T., 1967. A general theory of classificatory sorting strategies. I. Hierarchical

525   systems. Comp. J. 9: 373–380.

526 Lefkovitch, L.P., 1976. Hierarchical clustering from principle coordinates: an efficient method for

527   small to very large numbers of objects. Math. Biosci. 31: 157-174.

528 Legendre, P., Legendre, L., 2012. Numerical Ecology, 3rd Engl edn. Elsevier, Amsterdam.

529 Lloyd., S.P., 1982. Least squares quantization in PCM. IEEE Transactions on Information Theory 28:

530   129–137.

531 MacQueen, J.B., 1967. Some Methods for classification and Analysis of Multivariate Observations.

532   Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability 1. University

533   of California Press. pp. 281–297.

534 McQuitty, L.L., 1967. A mutual development of some typological theories and pattern-analytic

535   methods. Ed. Psychol. Measurement 27: 21-46. *[Check]*

536 Noy-Meir, I., 1973. Divisive polythetic classification of vegetation data by optimized division on

537   ordination components. J. Ecol. 61: 753-760.

538 Somerfield P.J., Clarke, K.R., 2013. Inverse analysis in non-parametric multivariate analyses:

539   distinguishing groups of associated species which covary coherently across samples. J. Exp. Mar.

540   Biol. Ecol. 449: 261–273.

541 Somerfield, P.J., Clarke, K.R., Warwick, R.M., Dulvy, N.K., 2008. Average functional distinctness as a

542   measure of composition in assemblages. ICES J. Mar. Sci. 65:1462-1468.

543 Steinhaus, H., 1957. Sur la division des corps matériels en parties. Bull. Acad. Polon. Sci. 4: 801–804.

544 Warwick, R.M., Clarke, K.R., 2001. Practical measures of marine biodiversity based on relatedness of

545   species. Oceanography and Marine Biology: an Annual Review, 39: 207–231.

546

Figure legends

Fig 1. Map of the Bristol Channel and Severn Estuary showing the locations of the 57 sites from which zooplankton samples were collected in April 1974.

Fig 2. Dendrogram from hierarchical agglomerative clustering of the 57 sites using group average linking of Bray-Curtis similarities calculated from √√-transformed holoplankton densities showing: A) The three groups produced by applying an abritrary cut at 50% similarity; B) the results of successive SIMPROF tests on groups of samples defined by nodes of the tree, beginning at the top. Only the first three tests showed significant multivariate structure in the samples below that point , so there is no evidence from SIMPROF that the detailed clustering structure (grey dashed lines) within each of the 4 main groups is interpretable.

Fig 3. Variation in the cophenetic correlation between the original resemblances from the zooplankton data and cophenetic distances from cluster analyses using flexible beta linking with different values of β. The highest correlation (0.794) is from the analysis in which β = -0.04.

Fig 4. Relationship of the original Bray-Curtis dissimilarities between pairs of samples and cophenetic distances between those same pairs of samples in the dendrogram shown in Fig 2.

Fig 5. Dendrogram from unconstrained binary divisive clustering of 57 sites maximising ANOSIM R at each binary split (UNCTREE) plotted using: A) the A% scale in which steps are arbitrarily equi-spaced; B) the B% scale which indicates the relative strength of each division. As in Fig. 2B continuous lines indicate structure supported by SIMPROF and grey dashed lines indicate no interpretable structure.

Fig 6. Non-metric MDS ordination of the 57 sites derived from Bray-Curtis similarities calculated from √√-transformed holoplankton densities. Symbols indicate the groups found by SIMPROF tests for each of 3 clustering methods: A) agglomerative hierarchical with group-average linking; B) hierarchical divisive clustering; C) non-hierarchical *k-R* clustering. Labels are site numbers (A, C) or salinity scores (B, from 1: <26.3, …, 9: >35.1 ppt, see Clarke et al., 2014 for full details).

Fig 7. 2-d nMDS ordination presented in a '3-d project' with the dendrogram from group-average linking. Contours indicate samples joined with a similarity of 55%. Sample symbols denote SIMPROF groups.
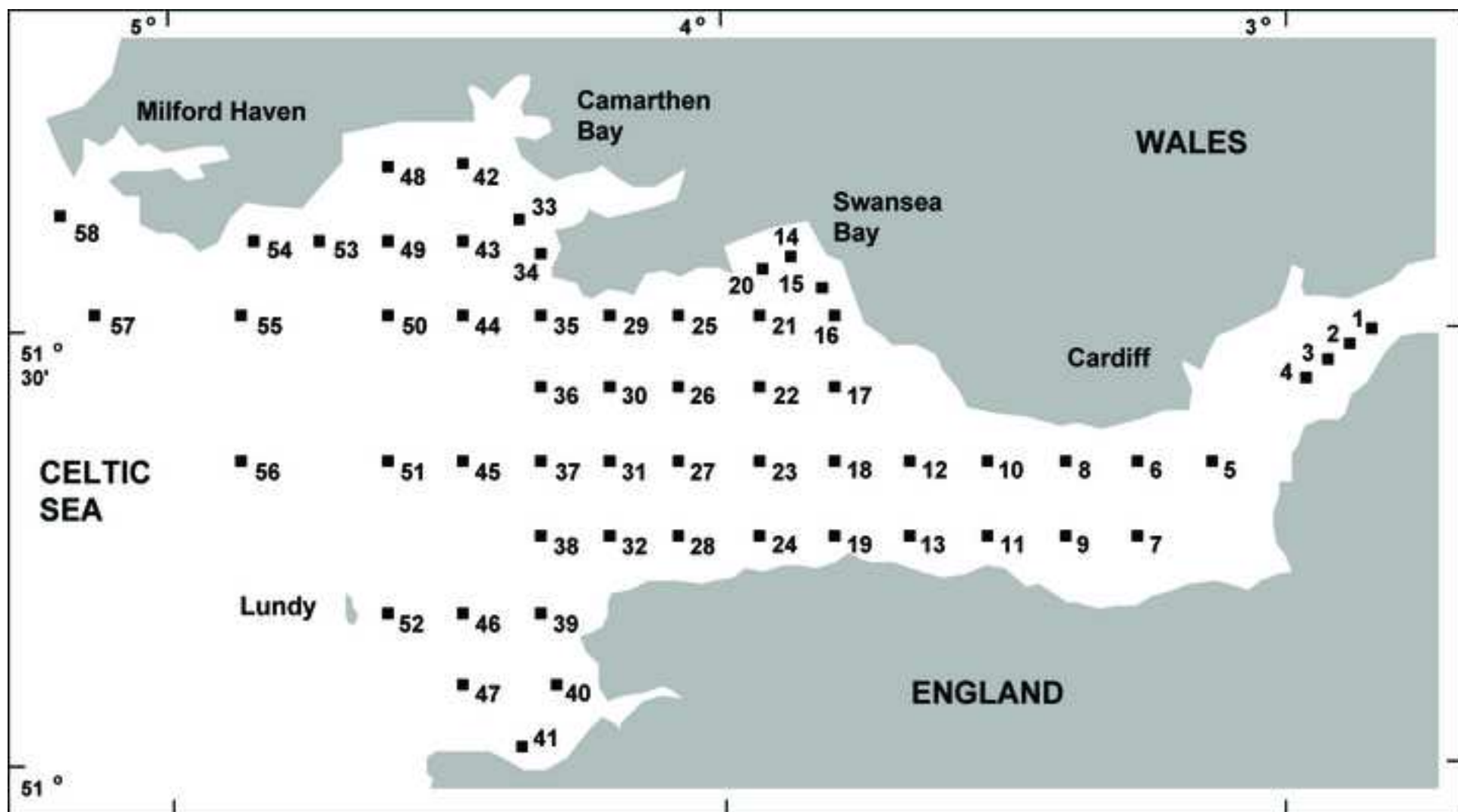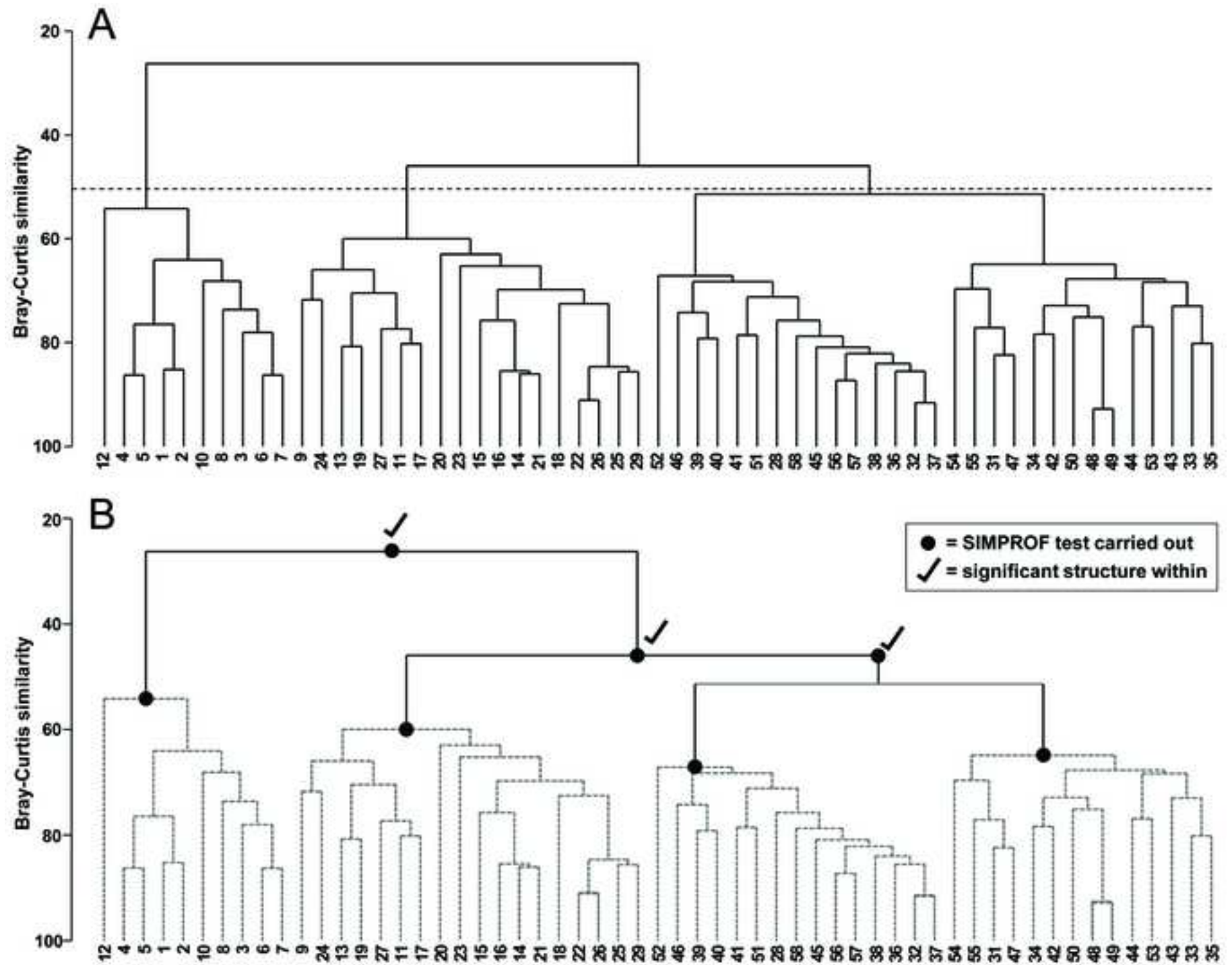
Figure 1
Click here to download high resolution image

Figure 2
Click here to download high resolution image

Figure 6
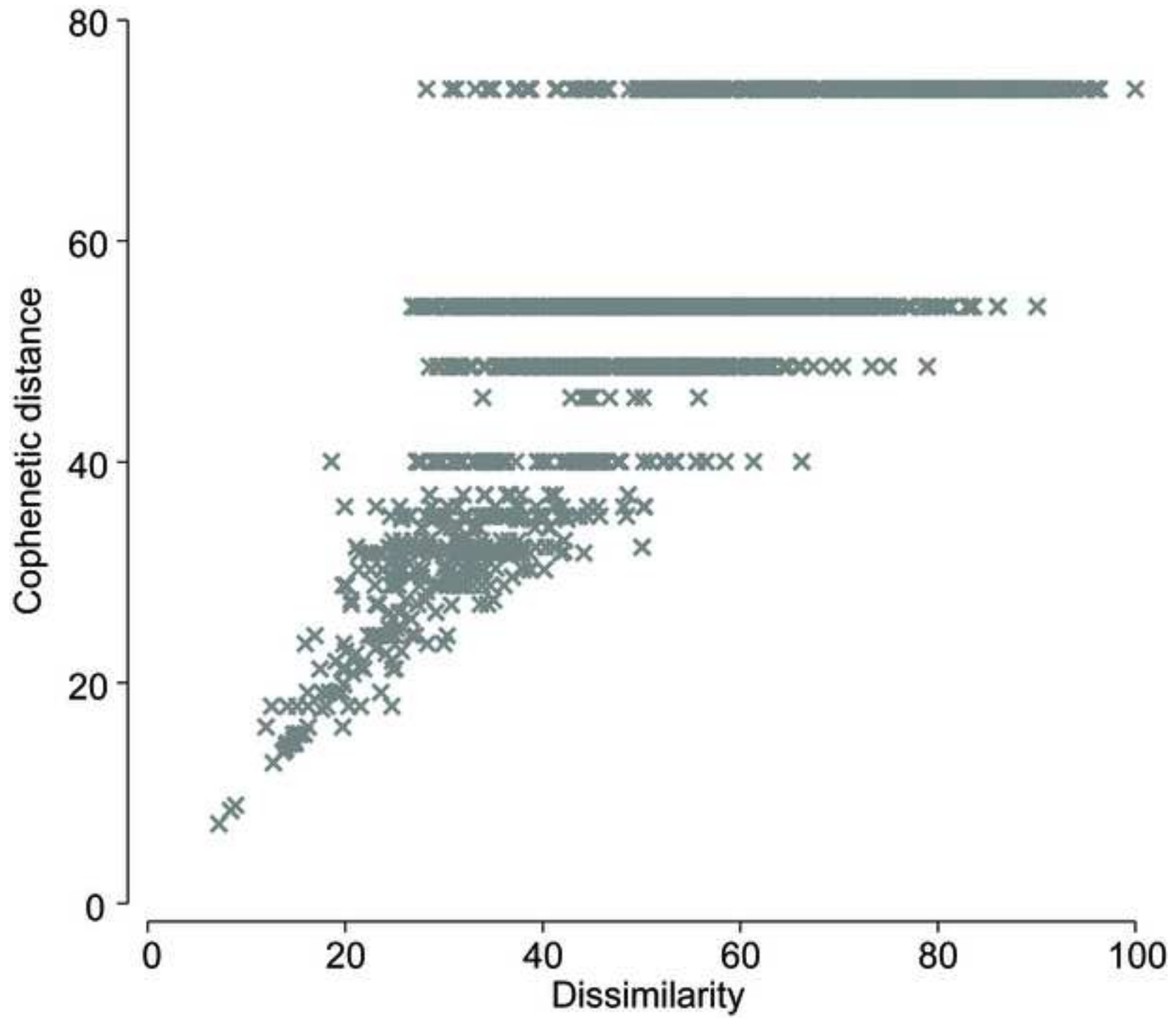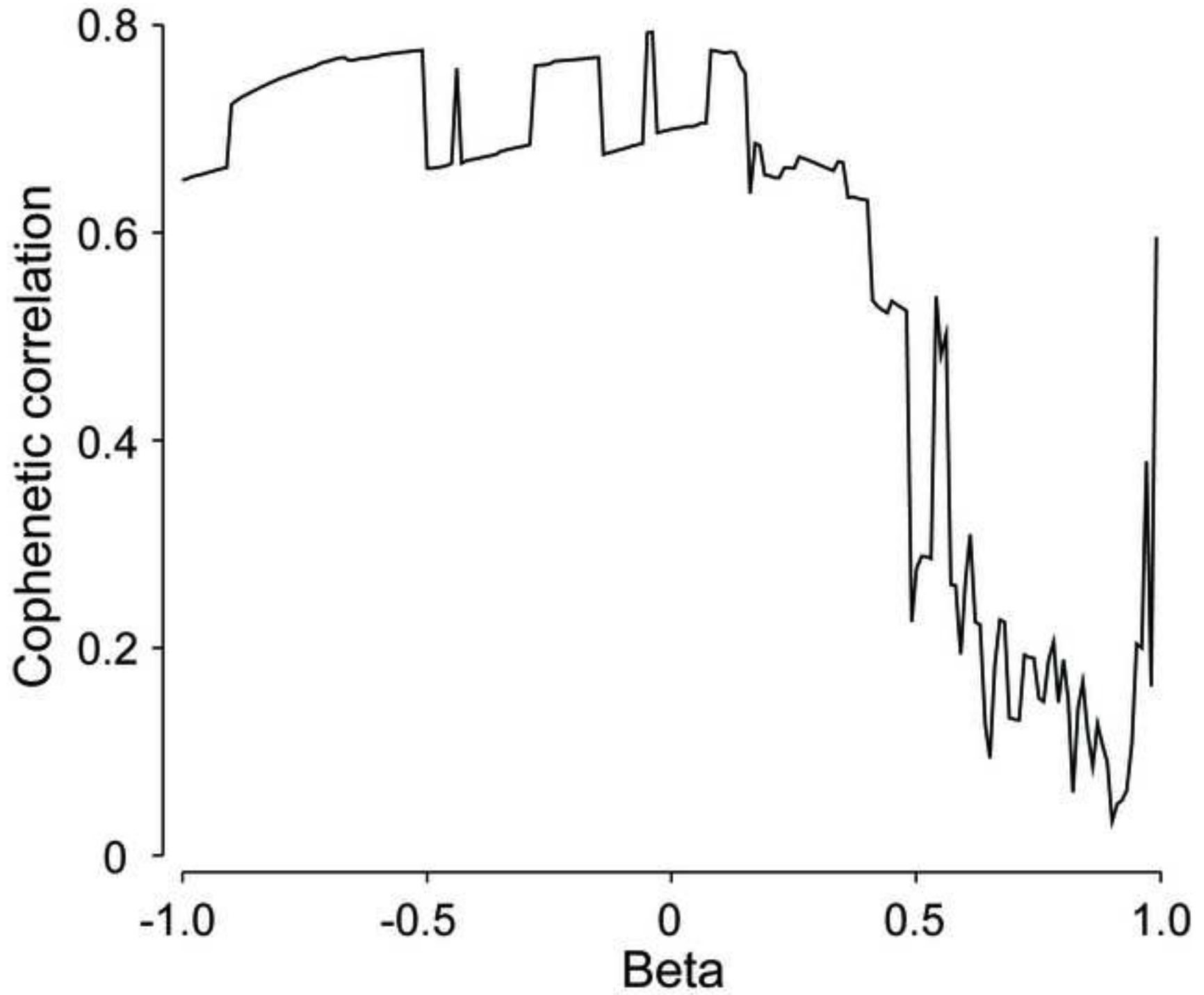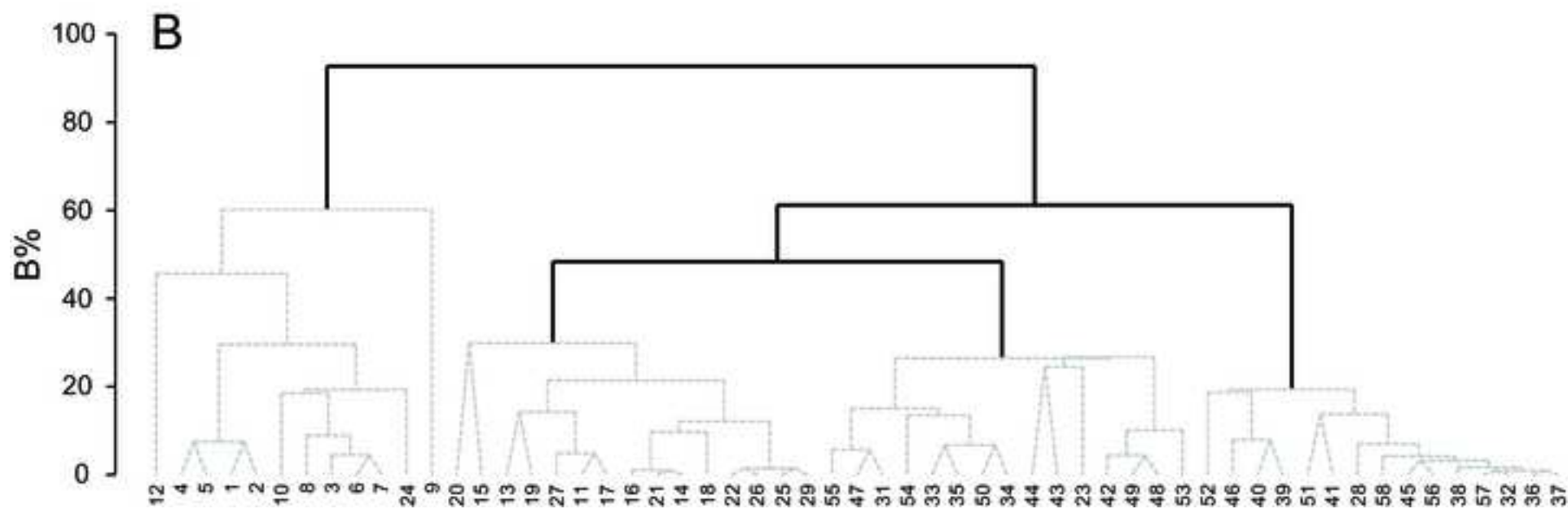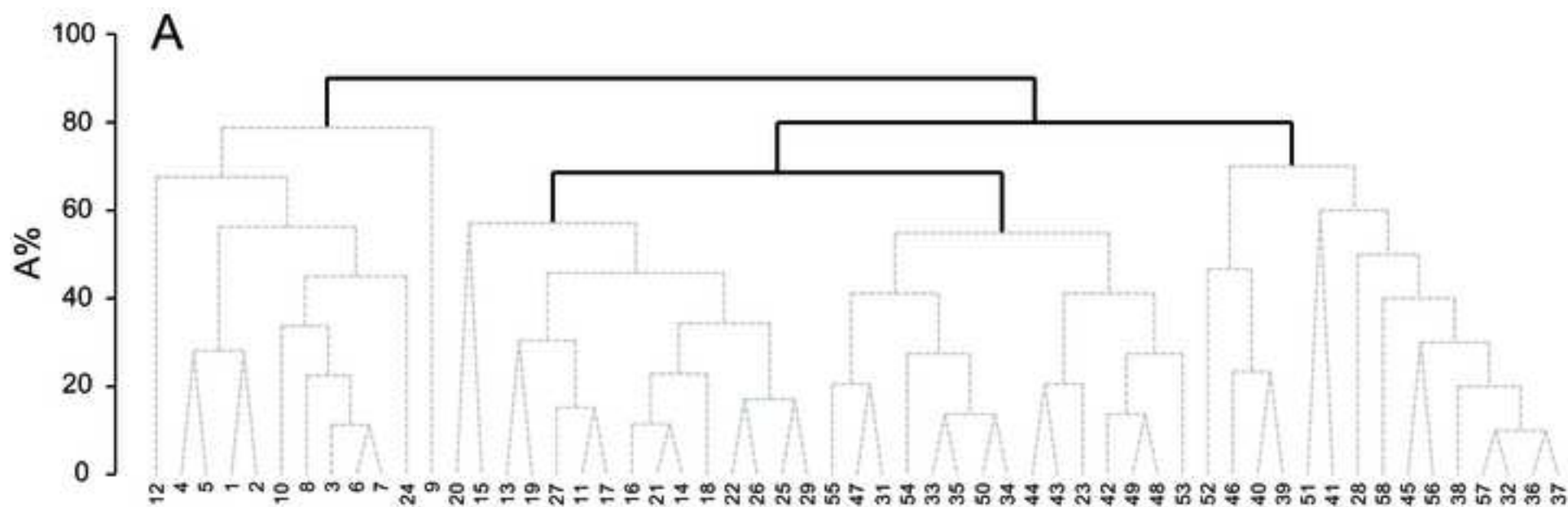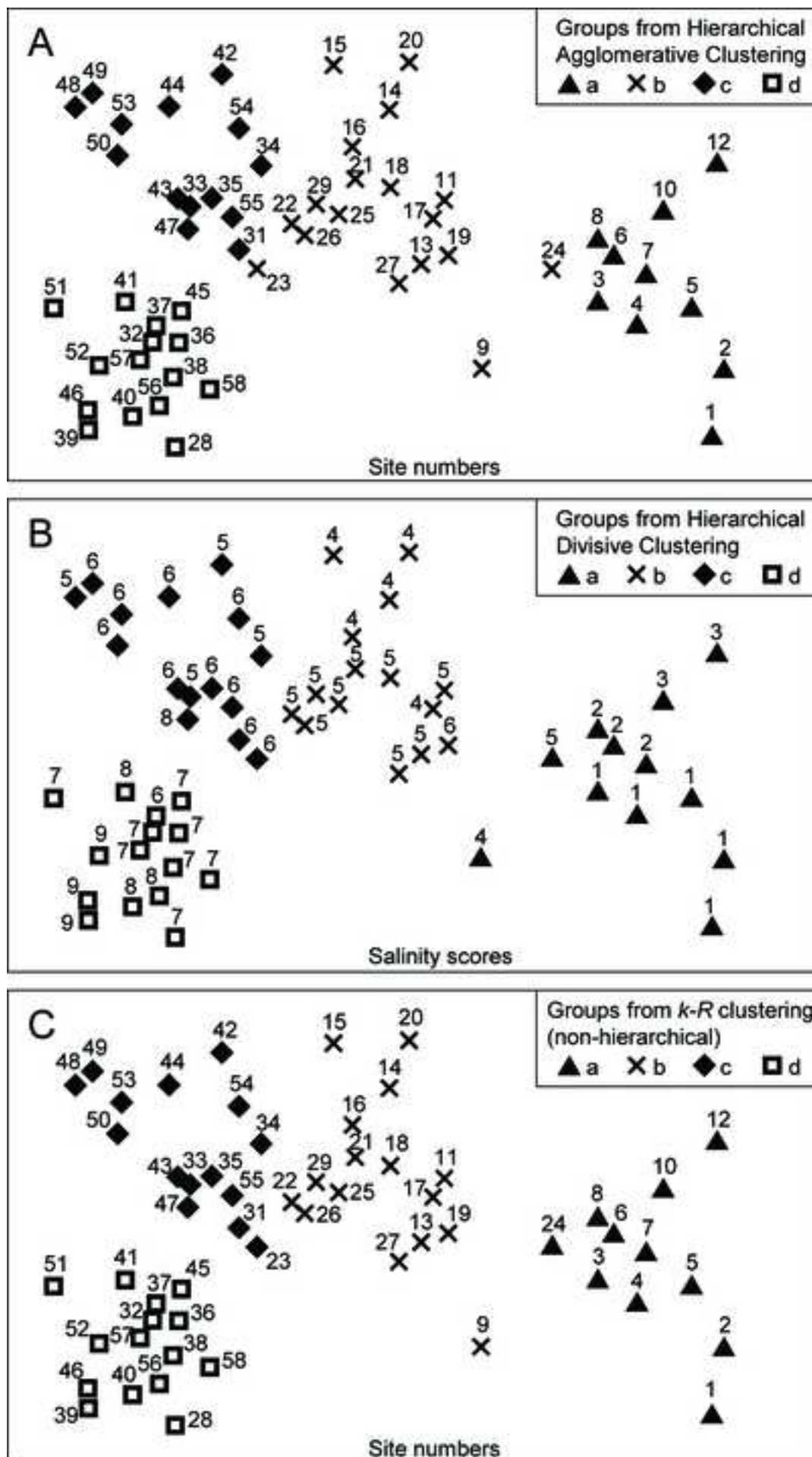**Click here to download high resolution image**

Figure 6

Figure 7
[Click here to download high resolution image](#)