

CAFCA

Chapter 4

Group Compatibility - Secondary Analysis

© M. Zandee 1989,1996 All Rights Reserved.

February 1996

THIS PAGE INTENTIONALLY LEFT BLANK

IV. GROUP COMPATIBILITY - SECONDARY ANALYSIS

INTRODUCTION

The use of a secondary analysis is to resolve to greater extent, if necessary, a cladogram resulting from a primary analysis. Sometimes the number of characters, or the pattern of interesting clada build from these characters, does not allow for a complete dichotomous resolution of cladograms. In that case at least one, but sometimes many nodes in a cladogram do not support a dichotomy but a trichotomy or even worse. The resolving power of the available character states within the constraints governing the recognition of clada simply is not sufficient relative to the number of terminal taxa.

You can improve resolving power by relaxing the rigour by which clada are defined and thereby increase their number and the chance of mutual fit. One of the options is to go from a partial definition of clada (PMS; cladon option 1) to a strict one (SMS; cladon option 2). You can also use PMS plus all additive binary coding (PMS + ABC; cladon option 4), or PMS plus all clada from valid three-taxon-statements (PMS + TTSP; cladon option 5). However, that does not always prove a workable solution because of limited available memory, for instance. How to strike a balance between these options is something the user should find out for himself as foolproof guide-lines are very difficult to derive.

So in many cases limiting the scope to a restricted number of taxa is the only way out and indeed will result in an improved resolution of the branching pattern. In doing so, however, one loses sight of character state distributions in the other taxa beyond the constrained view.

A secondary analysis can be run on unresolved cladograms selected as the best ones in a previous primary or secondary analysis. In a secondary analysis the program will check all nodes for dichotomy. All leaf-side neighbouring nodes are isolated if a polytomous node is found. They are isolated either from the data matrix (in case of single terminal taxa) or from the list of character states on internal nodes (in case of groups of terminal taxa) that CAFCA generates during its analyses.

On this selection of clada a primary analysis is run using all characters (but always *undirected* and *unordered*). The new branching patterns found for this selection are kept in memory while the program checks the other nodes for dichotomy.

For all other polytomous nodes this process is repeated. Multiple solutions for the total cladogram will result if for each of these nodes more than one completely resolved solution is possible.

As the search for polytomies and their resolution is completed the newly found extended cladogram(s) will be evaluated against the **complete** data matrix (with direction and ordering intact, eventually). Thus a secondary analysis comprises all (restricted) primary analyses plus the evaluation of the new cladogram. I will now give you a hands-on guided tour through the secondary analysis menu and then discuss the results for the example data matrix.

AN EXAMPLE.

Before you can run a secondary analysis you must have the results of a primary analysis available first. That's what we are going to do right now.

PRIMARY ANALYSIS

TUTORIAL

1. Select **Primary Analysis** from the **Run** menu.
2. Select **1 (From an ASCII file)** in the **Load a data matrix** dialog. Click **OK**.
3. In the next file selector box select **SECOND.BIN** from the example data on your distribution disk and click **Load File**.
4. Enter 'Second' (without quotation marks!), for example, as a name for your data matrix in the next dialog. Click **OK**.
5. Click **all-columns-are-equivalent** in the **No (correct) partition vector present** dialog. Click **OK**.
6. Click **No** in the **Do you want to edit your data** dialog.
7. Click **No** in the **Data matrix needs clipping ?** dialog. Click **OK**.
8. Click **Yes** for the **Ancestral state indicated by zero** option in the **Set CAFCA Parameters** dialog. Leave all other options in their default settings. Click **OK**.
9. Wait till the analysis is finished.
10. Select **All Results** from the **Print** menu.
11. Click **File** in the **Select Print Device** dialog box. Click **OK**.
12. Enter the filename in the next dialog and click **Save**. You can view or print the file by means of your favourite word processor.

The next two steps are optional:

13. After printing your results to file, select **Save and Resume** in the **OutputFile** menu.
14. Select an Outputfile system (**CAFCA.IO**) in the file select box to save your results to.
15. Click **OK** in the **Results for Second will be saved to OutputFile** dialog.

DISCUSSION OF RESULTS

Table 6 shows the results of a primary analysis on *Second*. *Second* is a data matrix with many taxa relative to the number of characters.

Data Matrix (binary) : SECOND (Columns represent character states)

	1	2	3	4	5	6	7	8	9	10	11
1	1	0	0	0	0	1	1	0	0	1	1
2	1	0	0	0	0	1	1	0	1	1	0
3	0	1	0	0	0	0	0	1	1	0	0
4	0	1	0	0	1	0	0	1	1	0	0
5	0	0	1	0	0	0	0	0	1	0	0
6	0	0	0	1	0	1	0	0	1	1	1
7	0	0	0	1	0	1	0	0	0	0	1
8	0	0	0	1	0	1	1	0	0	1	0
9	0	0	1	0	1	0	0	0	1	1	0
10	0	0	0	0	0	0	0	0	0	0	0

Column Partitioning Vector :
 1 1 1 1 1 1 1 1 1 1 1

Table 4.1 Data matrix to be used for a primary and secondary analysis.

RQ_C indicates that cladogram # 2 fits the data slightly better than # 1 does (table 4.4), thus cladogram # 2 is selected to do a secondary analysis on.

This cladogram (table 4.3) has two trichotomies. The one close to the root (cladon 21) involves three branches leading to 3 internal nodes, {1 2 6 7 8}, {3 4}, and {4 5}, respectively. The second trichotomy (cladon 18) involves 2 terminal taxa, # 6, 7, as well as an internal node {1 2 8}.

Selection criteria for cladograms of: Second				
Column numbers refer to numbers of cladograms				

Row 1 :	Total number of homoplasous events			
Row 2 :	Total number of single origins (Support)			
Row 3 :	Corrected Extra Length (x1000; CEL: Turner + Zandee)			
Row 4 :	Total number of state changes (S: Steps)			
Row 5 :	Redundancy Quotient (x1000; RQ: Zandee + Geesink)			
Row 6 :	Rescaled Redundancy Quotient (x1000; RQc)			
Row 7 :	Consistency Index (x1000; CI), with autapomorphy correction			
Row 8 :	Rescaled Consistency Index (x1000; RC: Farris)			
Row 9 :	Average Unit Character Consistency (x1000; AUCC: Sang)			
Row 10:	Homoplasy Distribution Ratio (x1000; HDR: Sang)			
Row 11:	Compatible Character State Index (x1000; CCSI: Zandee)			
	1	2		

1	7	8		
2	6	6		
3	9273	9273		
4	20	20		
5	504	505		
6	117	120		
7	550	550		
8	325	325		
9	720	720		
10	377	377		
11	273	273		
No-Order Limit for Steps, Extra Steps, RQ, and CI:				
	S	ES	RQ	CI

	33	22	438	333

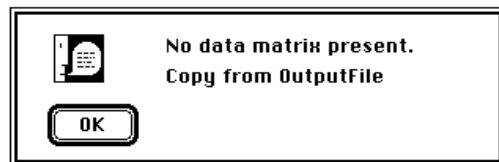
Table 4.4 Selection criteria for cladograms in Second.

In the secondary analysis each of these trichotomies is subjected to a primary analysis, thus each involving a data matrix with 3 nodes (rows) and all characters (11). Let's run the secondary analysis now.

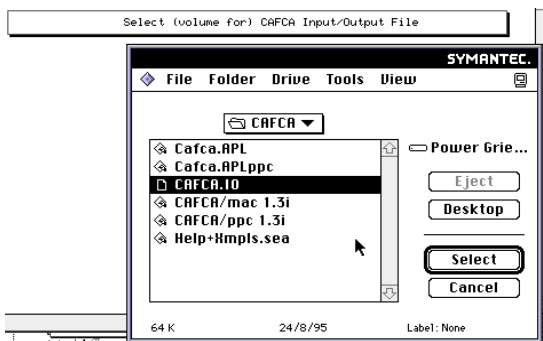
SECONDARY ANALYSIS

TUTORIAL

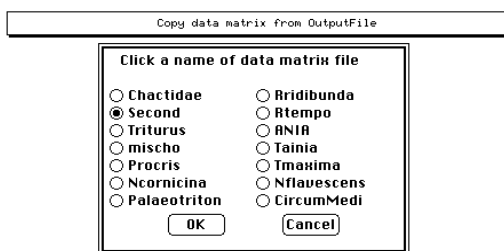
1. Click **Secondary Analysis** in the **Run** menu. You get the following message. If you skipped step 12 and 13 (**Save and Resume**) in the primary analysis, your data will still be present in the workspace and CAFCA starts a secondary analysis with step 4 from this tutorial.



2. Click **Read & Continue** from the **OutputFile** menu.
3. Select an Outputfile system (**CAFCA.IO**) in the next file select box.

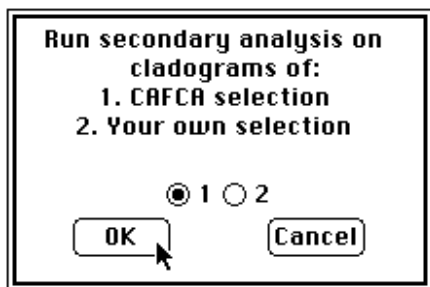


4. Click **Second** in the next dialog box. Note that you may have just one file, **Second**, in your OutputFile system instead of many as shown in this particular example

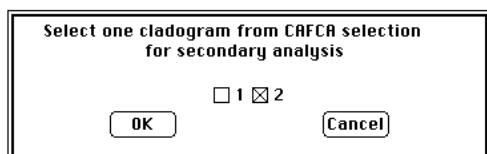


All the elements from the primary analysis on **Second** that are needed to run a secondary analysis are now loaded into the workspace.

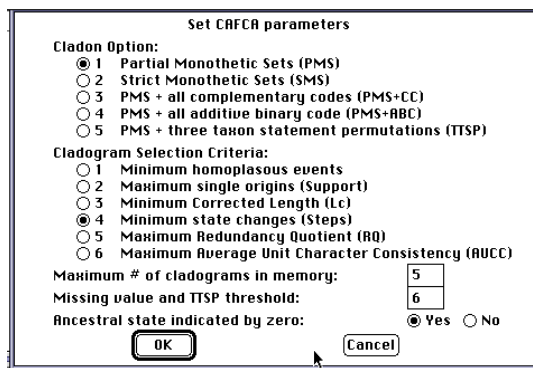
5. Click **1 (CAFCA selection)** in the next dialog box.



Remember that **two** minimum step cladograms were selected by CAFCA. Because a secondary analysis is restricted to only one cladogram at the time the following dialog appears next.

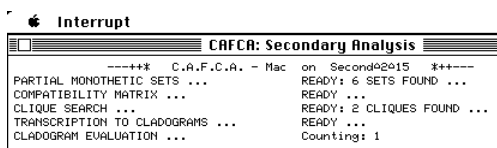


6. Click **2** (= cladogram # 2) in this dialog box, and then **OK**.



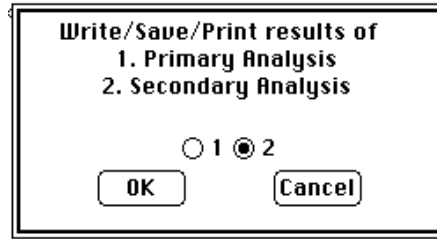
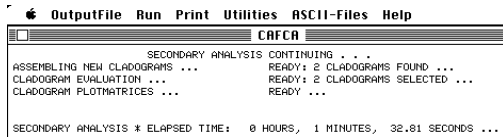
7. The next dialog is for setting the CAFCA parameters to be used in this secondary analysis. The main point to consider here is the maximum number of cladograms that will be retained in memory for each of the nodes to be analysed. If this number is too high, a combinatorial explosion may occur when all equally parsimonious possibilities for each node are combined to give all possible full resolutions for the original cladogram, and consequently, memory will be swamped. Another remedy against this combinatorial explosion is provided by RQ as a selection criterion for cladograms, because it offers more resolving power and is therefore more restrictive than the CI (minimum steps).

8. After the parameters are set the secondary analysis will start running. You can follow its progress on a sequence of screens, each showing a primary analysis on each separate node that constitutes a polytomy.



If for whatever reason you want to stop this run, use the **Interrupt** menu (see ch. 2).

The solutions resulting from these restricted primary analyses are combined and merged (with each other and with the resolved part of the old cladogram) to form better resolved cladograms. The new cladogram is evaluated (table 4.6) with respect to the **complete** data matrix (table 4.1)



9. When the secondary analysis is finished you can print the results by selecting **All Results** from the **Print** menu. In the following dialog you must select option **2** (Secondary analysis).

DISCUSSION OF RESULTS

I will now present a discussion of the results.

The layout of the output corresponds with that of a primary analysis so almost all explanation given in chapter 3 also applies here. I will only pin-point the most interesting parts.

First, note that output of a secondary analysis is always recognisable by the addendum **sec** (+ number) in the name of the data matrix (Secondsec2). The number refers to the cladogram that was subjected to a secondary analysis.

Second, take good notice of the fact that the clada mentioned in this output (table 4.5) refer only to the ones derived in the secondary analysis. Their numbers therefore do not correspond with the numbers of the clada derived during the primary analysis (table 4.2).

Nevertheless, you can easily pick out the clada that are new now in comparison with the primary analysis, viz. # 14, {6 7}, # 16 {1 2 6 8}, and # 17, {3 4 5 9},

Partial Monothetic Sets of terminal taxa in Secondsec2

1 :	1								
2 :	2								
3 :	3								
4 :	4								
5 :	5								
6 :	6								
7 :	7								
8 :	8								
9 :	9								
10 :	10								
11 :	1	2							
12 :	3	4							
13 :	5	9							
14 :	6	7							
15 :	1	2	8						
16 :	1	2	6	8					
17 :	3	4	5	9					
18 :	1	2	6	7	8				
19 :	1	2	3	4	5	6	7	8	9
20 :	1	2	3	4	5	6	7	8	9 10

Table 4.5 List of clada for secondary analysis on Second.

First, we take a look at the values for the selection criteria (table 4.6) and notice that the new cladograms perform better than the old ones we started with, with respect to the number of steps (20 vs. 18) as well as to the value of RQ (.505 vs. .527).

Second, we can easily see by comparing the diagrams of the primary (table 4.4) and secondary analysis (table 4.7) that the cladograms are now completely resolved indeed.

Selection criteria for cladograms of: Secondsec2		
Column numbers refer to numbers of cladograms		

Row 1 :	Total number of homoplasous events	
Row 2 :	Total number of single origins (Support)	
Row 3 :	Corrected Extra Length (x1000; CEL: Turner + Zandee)	
Row 4 :	Total number of state changes (S: Steps)	
Row 5 :	Redundancy Quotient (x1000; RQ: Zandee + Geesink)	
Row 6 :	Rescaled Redundancy Quotient (x1000; RQc)	
Row 7 :	Consistency Index (x1000; CI), with autapomorphy correction	
Row 8 :	Rescaled Consistency Index (x1000; RC: Farris)	
Row 9 :	Average Unit Character Consistency (x1000; AUCC: Sang)	
Row 10:	Homoplasy Distribution Ratio (x1000; HDR: Sang)	
Row 11:	Compatible Character State Index (x1000; CCSI: Zandee)	

	1	2

1	6	6
2	6	6
3	7220	7197
4	18	18
5	520	527
6	146	159
7	611	611
8	417	417
9	742	742
10	338	338
11	273	273

No-Order Limit for Steps, Extra Steps, RQ, and CI:			
S	ES	RQ	CI

33	22	438	333

Table 4.6 Selection criteria for the cladograms resulting from the secondary analysis on Second.

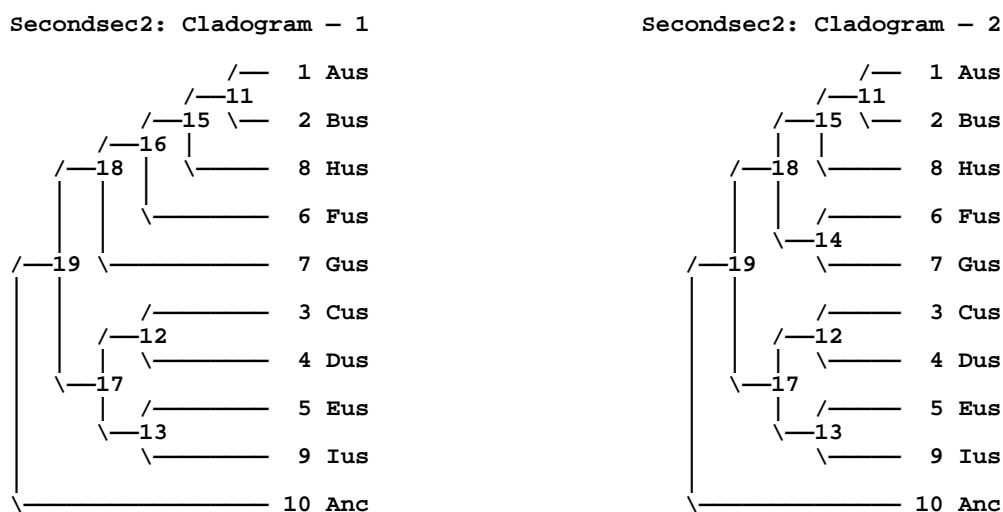


Table 4.7 Fully resolved cladograms resulting from secondary analysis on cladogram in table 4.4

In cladograms 1 and 2 greater resolution results from the recognition of two extra groups. First, group {3 4 5 9} is new in both cladograms, at the cost of two parallelisms for character 9 in taxon 2 and 6. As the distribution of states for this character also takes 3 steps in the less resolved cladogram **no extra** penalty is due. Second, in cladogram # 2 group {6 7} is now recognised on the basis of character 11. In cladogram # 1 group {1 2 8 6} is new. It is recognised on the basis of character 10. The second cladogram has a slightly better RQ value and also a lower (= better) value for the CEL. In this particular case CEL as well as RQ discriminate among MPT's where AUCC does not.

When, given the very minor difference in RQ for the primary cladograms #1 and #2, we run a secondary analysis on cladogram #1 (from table 4.3) as well, we get two cladograms one of which has 18 steps and a RQ of .527 (table 4.8), the same values as we got for the second secondary cladogram from cladogram #2.

Secondsec1: Cladogram - 2

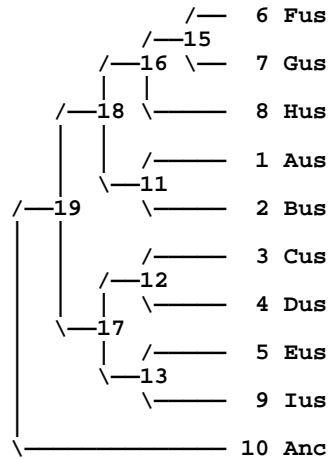


Table 4.8. Fully resolved cladogram resulting from secondary analysis on cladogram #1 from primary analysis (table 4.3)

USING OTHER OPTIONS INSTEAD OF SECONDARY ANALYSIS.

Applying a secondary analysis on both cladograms that resulted from the primary analysis finally gave us three most parsimonious cladograms with 18 steps. The question is whether a secondary analysis is an effective and efficient way of finding more resolved resolutions for polytomous primary cladograms, compared with the other options available in CAFCA.

What would the results be if we do not stop with PMS to analyse the data matrix in table 4.1, and use other options first to see if fully resolved cladograms can be obtained?

First we try PMS with the option **Ancestral condition indicated by zero** set to **No** (default). This may seem silly as an all-zero outgroup (ancestor) is present in the data matrix. Nevertheless, this option allows CAFCA to find sets of taxa based on reversed character states (1->0) within the in-group, and shorter cladograms, eventually. We find 10 cladograms, two of them with 19 steps. Both have 1 trichotomy left. Thus this option also requires a secondary analysis of the primary cladograms. If we actually run the secondary analysis, the same three 18-step cladograms result from the two primary cladograms.

Next we try SMS (strict monothetic sets) with the option **Ancestral condition indicated by zero** set to **Yes**. Five cladograms are found, three of which have 19 steps and one trichotomy left. This option too is thus in itself less effective than a secondary analysis. Again, when running a secondary analysis on the three 19-step primary cladograms, the same three 18-step cladograms as found earlier result.

Third, we use SMS plus all complementary sets. We find 57 sets from which 52 cladograms can be derived. Five of these are MPC with 19 steps.

Next, we try SMS with the option **Ancestral condition indicated by zero** set to **No** (default). CAFCA finds 99 strict monothetic sets, and 2637 cladograms can be based on these sets. Out of the 2637 there are three most parsimonious completely resolved cladograms with 18 steps. These cladograms are identical to the ones found by applying a secondary analysis to both primary cladograms. This analysis took almost 11 hours of computation on a Mac SE/30, and 73 minutes on a Power PC (Power Macintosh 6100/66 with native CAFCA version). It may be as effective as the two secondary analyses, but we can hardly call it an efficient way to find the same three cladograms.

Fifth, we use the Three-Taxon permutation possibilities of CAFCA to build sets of terminal taxa, with the option **Ancestral condition indicated by zero** set to **Yes**. On the basis of 77 of these sets CAFCA finds 855 cladograms, three of them the shortest with 18 steps, and the same as those from the secondary analyses (tables 4.7 and 4.8) As this analysis took less than 4 hours to complete on a Mac SE/30 and 22 minutes on a PPC 6100/66, it is more efficient than the one above, but not nearly as efficient as the secondary analyses.

The cladograms from the analyses discussed above are the same 3 most parsimonious cladograms as found by PAUP (with much more ease and very, very much faster).

Three-Taxon Statements according to Nelson & Platnick (1991) generates a data matrix with 214 characters. Again using PAUP, at least 9500 most parsimonious cladograms are found (length 214 steps; PAUP stops at 9500 cladograms due to a *maxtrees* limit when running out of memory). I did not check whether the topologies found earlier by CAFCA and PAUP are among the 9500. It appears rather unconvincing that for the data in table 4.1 the Three Taxon Statement approach according to Nelson & Platnick indeed implies a more *precise* use of parsimony.

A COMPARISON WITH CHARACTER COMPATIBILITY

Strauch (1984) analysed his data on the Alcidae (Aves) by means of character compatibility analysis, applying a secondary analysis on the incompletely resolved primary cladogram based on 23 compatible characters. Secondary analysis in the context of character compatibility implies looking for additional compatible characters in only those subsets of taxa that build the unresolved branches in the primary cladogram, while adjusting the direction and order of the additional characters.

Running a secondary analysis by means of *group* compatibility on the single most parsimonious cladogram that resulted from the primary analysis (67 steps; table 3.32) generates 18 almost completely resolved cladograms, 9 of which have 66 steps. It appears that holding on to group {18-21} as a sister-group of {1-17} makes it impossible to find cladograms any shorter than 66 steps, let alone the 64 steps that PAUP finds for its most parsimonious cladograms (without this particular sistergroup relation). In this case, secondary analysis is not as effective as PAUP (and certainly not as efficient), unless one finds the loss of the aforementioned sistergroup relation unacceptable.

The shortest cladograms found by CAFCA in this secondary analysis still have 19 compatible characters; in this case the reduction in length by one step apparently has no effect on the number of *fully* compatible characters (it does have an effect for the cladograms with length 64 found by PAUP). There is only one trichotomy left, in contrast to the cladograms that PAUP found, where 3 trichotomies remained as a result of collapsing empty branches.

One of the cladograms found by CAFCA in its secondary analysis of the primary cladogram from table 3.32 is shown below.

As can be seen in the data matrix for the *Alcidae* (table 3.31), there are no characters that differentiate the members of the group {15-17}; these taxa are identical. All other branches are resolved, in contrast to Strauch's cladogram found after a secondary analysis of *character* compatibility.

AlcidaeBsec1: Cladogram -15

