IMP:DisparityBox6

H. David Sheets, Dept. of Physics, Canisius College, Buffalo, NY 14208,

nual, resampling statistics)

Miriam L. Zelditch, Museum of Paleontology, University of Michigan, Ann Arbor, Michigan 48109, zelditch@umich.edu (conceptualization)

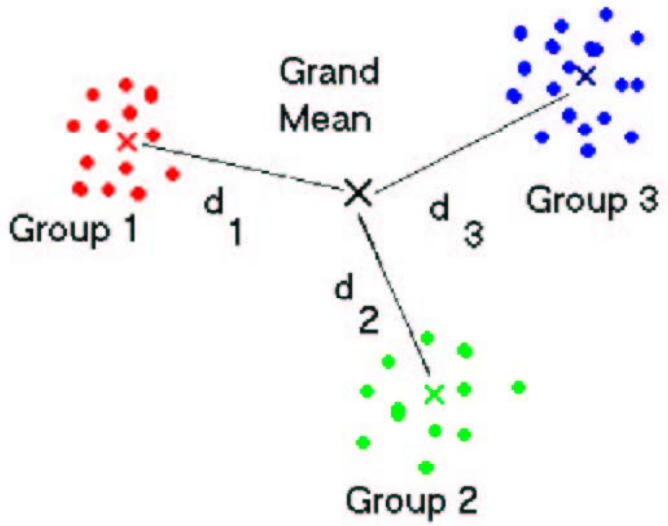
Introduction

DisparityBox6 is a tool for calculating disparity (morphological diversity of a group or clade) based on landmark data in the geometric morphometrics framework. The measurement of disparity follows that established by Foote (1993)

$$D = \Sigma (d_i^2)/(N-1)$$

where di represents the distance of the centroid of group i from the centroid of all N groups. The distance metric we have adopted is the Procrustes distance. This is calculated from the mean specimen of each group to the average of all the group means(which we call the grand mean shape). The “cartoon” shown below illustrates), which is the centroid of that group. The centroid of the group centroids, called epresenting a species (or group) of organisms. Each data set must use the same set of landmarks; each is loaded into DisparityBox in turn. The disparity is calculated and a confidence interval for it may be determined using a bootstrapping procedure. In this

bootstrapping, the original species data sets are resampled with replacement to determine the range of disparity values possible under resampling. This allows determination of the 95% confidence interval. Note that the bootstrapping is done at the specimen level, not at the group or species level. The same approach may be used to determine the disparity within a single species.



One of the difficulties encountered in computing disparity is that differences in shape due to size or age can contribute to disparity. DisparityBox is capable of fitting a regression model to the data and determining the residuals to the regression model. The regression model and residuals can then be used to produce a "size standardized" or "age-standardized" data set, by using the residuals and the regression model evaluated at a specified age or size.

Some researchers may also be interested in disparity calculations based on traditional morphometric length measurements. DisparityBox is capable of calculating disparity based on distance measures between landmarks. The user inputs a length protocol of desired interlandmark distances desired in the calculation. Several size standardization procedures are included, including one that standardizes by geometric scale and one that standardizes by allometric size.

DisparityBox may also be used to determine the Procrustes distance between two species, or between two developmental stages of a single species, providing confidence interval on that distance. To determine the distance between two species, load two species into DisparityBox and determine the disparity (and confidence interval). The distance between the pair of species is the square root of this pairwise disparity, and the confidence interval of the distance between means is the square root of the corresponding confidence intervals on disparity. To calculate the distance between two developmental stages of a single species, load the *same* file twice, inputting the two sizes to which the data should be standardized. The distance between the two stages is the square root of the of the pairwise disparity and the confidence interval of the distance is the square root of the corresponding confidence interval.

A Note on this Manual

DisparityBox is changing rapidly (as of 8/15/02), and I am behind in keeping the manual up to date. Contact me (HDS) or Miriam directly if needs be, we will try to answer questions, or update the manual for you. Thanks for your patience...

Operation of Disparity Box

Calculating Disparity of Two or More Groups

Input:

You will need a separate data file for each group or species in the $X_1,Y_1,X_2,Y_2\ldots X_L,Y_L$, CS format used throughout the IMP system. If you are not familiar with this format, you need to read the manual for the program CoordGen before proceeding. You may also want to create a *Length Protocol File* and/or a *Log Size Target File* as discussed below, but you don't need them if you are using DisparityBox for the first time.

Output:

Unlike most programs in the IMP system, DisparityBox provides largely textual output in the *Results Box* which is in the center-left of the screen. The contents of this box can be written to a file using the Save Results Box option on the file menu, or the Append Results Box to File option on the file menu. Note that Append Results Box to File option will allow you to add many calculations to a single file. It will not overwrite existing files, only add to them, so **do not worry about the warning message**. (I can't figure out how to turn it off!) The append option is often useful.

Procedure:

1.) Load each species data file one at a time using the *Load Data* set button. When each set is loaded, the file name, number of specimens and number of landmarks will be displayed in the results window. The distribution of specimens (in GLS Procrustes superposition) will be shown. Load all your data files before proceeding. You can use the *List Loaded Sets* button to see a summary listing of all files loaded at any time.

2.) Once you have loaded the data files, compute the Grand Consensus Mean shape. There are two options here (a) the Grand Consensus Mean based on Group Means or (b) the Grand Consensus Mean based on all the specimens. These will produce identical results if all spccies have the same number of specimens. In general you should compute the Grand Consensus Mean using the group means.

3.) If you want to calculate disparity based on geometric measurements only, go on to step 4. If you wish to calculate disparity using traditional morphometric measurements, (as well as, or intead of, geometric data), you now need to load a length protocol that will be used to generate the traditional data set. Refer to the users manual for TradMorphGen to learn about how to create the "length protocol" file you will need to do this.

4.) If you wish to calculate disparity without "size-" or "age-standardization", go on to step 5. To size- or age-standardize, you will need to load a file specifying the sizes that

will be used for the standardizations. DisparityBox always regresses shape on ln (natural log) of the last column in the data file, which is assumed to be centroid size. If you need to size or age standardize on something other than ln centroid size, see the discussion later in this manual. When the "Log Size Targets" file is loaded, a summary of the files loaded, the specimen counts in each and the ln size target to be used will appear in the results box on the screen.

The file giving the sizes must be a text (ASCII) file, listing the desired value of ln centroid size of each group. The ln centroid size values must listed (one per line) in the same order as the data files were entered into disparity box (see the discussion later on size standardization for an example of such a file). When the length protocol is loaded, a diagram showing the lengths, plotted on the Grand Consensus Mean, specimen will be displayed.

5.) Specify the number of bootstrap sets you desire. Start with 100 to get a rough idea of how long the calculations are going to take on your data and computer.

6.) You are now ready to calculate disparity. The disparity calculations are executed using the menus at the top of the screen. The 1-group analysis computes the disparity (variance) within a single group; this option is specified using the "Active Group box" in the bottom-middle of the screen to select a single file. The Multi-group analysis routines are executed using all loaded data sets.

Types of Disparity Calculations

One-Group Analysis

The group to be used is specified using the up and down buttons in the active set box. The landmark configuration of this group can be plotted using the buttons in this box. All One-Group Analysis options are accessed through the menu bar.

The *Within-Group Disparity* option calculates a number of different distances, some of which are relics of earlier versions of the program. The within-group disparity is the last value in the table, labeled as "Foote Disparity". This option does not calculate a confidence interval. Use the *Bootstrap Within-Group Disparity* option to calculate both disparity and its confidence interval by bootstrapping (with replacement).

The *Traditional Measures Disparity* calculates disparity within the group based on the variance-covariance measures of the ln values of lengths specified by the length protocol. It also calculates disparity based on standardized traditional measures, obtained by dividing the lengths by the square root of the summed squared length for that specimen. This approach removes geometric scale, one aspect of size. Allometric standardization, which removes the effects of geometric scale on shape is not implemented for the within-group analysis. There is no bootstrapping of the rescaled measures currently implemented. They should be treated with some degree of caution because the properties of this type of data have not been fully characterized yet.

The *Bootstrap Size Corrected, Within Species Disparity* calculates the within species disparity of the active group, using the log size standardization target entered for that species.

Multi-Group Analysis:

These options compute the Between-Group and Among-Group disparity, using all groups loaded.

The *Bootstrap Disparity Measures* option computes the disparity based on geometric data, and determines the confidence interval of the disparity via bootstrapping. This is going to take a while, particularly if you have a lot of species and specimens and want to use a lot of bootstraps. Try it with a couple of species and 100 bootstraps to get some idea of the time involved (this would be a good time get coffee). The *Bootstrap Size-Corrected Disparity* regresses shape on ln centroid size, calculates the expected shape at the desired size, and adds the residuals of the regression to that expected value. Each species can be standardized to a different target size. Disparity is then calculated from the standardized data; the confidence interval is determined by bootstrapping residuals from the regression model (with replacement) and repeating the entire procedure. The *Size-Corrected Disparity* option computes the size corrected disparity without the confidence interval.

The *Bootstrap Geometric Disparity (MD,PD)* and *Bootstrap Size Corrected MD,PD* options carry out the same computations of geometric morphometric based disparity as the *Bootstrap Disparity Measures* and *Bootstrap Size-Corrected Disparity*, but also display the contribution to the disparity from each file (species or group) loaded, and a bootstrap estimate of the standard deviation of this contribution. To determine the Partial Disparity (PD) of a subclade, one simply sums the contributions to the disparity of all the species in that clade. To estimate the standard deviation of the Partial Disparity of a subclade, one can take the square root of the summed squared standard deviations of each species contribution to the disparity. From this one can then estimate a confidence interval for the Partial Disparity. Note that this estimate of the confidence interval for the Partial Disparity of the subclade is based on a normal model of the distribution, unlike the direct bootstrap estimates of confidence intervals used in the rest of DisparityBox. The normal model assumption allows one to compute the standard deviation of the Partial Disparity from the individual species contributions, which is not possible without the assumption of normality.

The *Bootstrap Traditional Measures Disparity* option computes disparity based on traditional measures, along with a bootstrap estimate of the confidence interval for this disparity. The *Bootstrap Size-Corrected Traditional Disparity* carries out the size correction using the regression model and residuals as discussed earlier. The *Bootstrap Dual-Size- Corrected Traditional Disparity* calculates disparity using length measures (standardized by both geometric scale and the regression model).

The *Calculate Within-Species Size-Corrected Disparity, PW+Traditional* option is vestigial; of no use to anyone but the author.

Other Operations

The *Axis* menu allows one to turn the axis on or off and clear the image plotted.

File Menu

The File menu allows saving contents of the Results box to an ASCII text file, or to append these contents to an existing text file, as discussed earlier. The *Load Log Size Targets* file option is also repeated here. There is an option to save the Grand Consensus Mean to a file, which may be useful as a reference form for use in other programs within IMP.

The *Output Size-Corrected Traditional Morphometrics, All groups*, option outputs a series of files (one per species) of the size-standardized traditional measures, which are not available (at present) from any other program in the IMP series.

DisparityBox can also be used to concatenate data sets. This file may then be loaded into the PCAGen or CVAGen programs within IMP. This file is created using the *Output Concatenated list of all loaded data.* PCAGen and CVAGen also require use of a group list. Such a group list corresponding to the concatenated data set may also be created,

using the *Output Group List for all files* option under the file menu. The first file loaded into distance box will have a group code of 1, the second will have a group code of 2 and so on. These options allow DistanceBox to create files rapidly for use with PCAGen and CVAGen (further information on these programs is available in their manuals). Note that if you want to alter codes in the group file, this can be done using the "Find and Replace" option in Excel or a word processor.

Literature Cited:

Foote, M. 1993. Contributions of individual taxa to overall morphological disparity. Paleobiology 19:403-419.

Carrying Out Hierarchical Calculations in DisparityBox

The term Hierarchial Disparity refers to the partitioning of disparity into contributions from two or more subclades. Foote (1993) developed the approach of partitioning the total morphological disparity into the contributions due to each subclade. We (some permutation of Sheets, Swiderski and Zelditch) have developed an approach that allows the partitioning of the partial disparity of each subclade into a withing subclade contribution and a between subclade contribution. The within subclade contribution is due to the distances of the mean shapes of each group within a subclade to the mean shape of all group means within a subclade (a shape we call the subclade mean shape). The between subclade contribution of each subclade is due to the distance of the subclade mean form from the grand mean form (the grand mean form is the mean of all

group mean forms). See the current version of the primer chapter on disparity for a discussion of how this partitioning of disparity has been derived.

The functions within DisparityBox under the Hierarchial Disparity menu will allow you to determine:

1.) The total disparity of the clade as a whole, and the amount of that disparity due to within subclade contributions and between subclade contributions . The total disparity is the sum of the within and between contributions.

2.) For each subclade, the Partial Disparity of the subclade is determined, and the between subclade and within subclade components of that Partial Disparity are determined.

3.) Confidence interval estimates for each measurement are obtained via bootstrap.

Currently, the size-standardized forms of the partial disparity are not calculated. If we need these, implementation should not be difficult.

Input Files

The approach used in DisparityBox6 is to represent each group by a sample of an number of specimens from that group (not a single mean specimen from that group). In addition to an IMP format file of specimens for each group, your will need a *Subclade List File*, which specifies how many subclades there are within your clade, the names of the subclades and a listing of which groups belong in each subclade. You will need to create this file using a text editor. Create this file before attempting to begin the analysis.

The first line of the file contains a single integer indicating which indicates how many subclades are present. The next lines each contain the name of the subclades (one name per line, and the name of each subclade must included). Following the list of subclades, are paired integer numbers indicating which subclade each file loaded into DisparityBox is a member of. The first number is an ordinal listing (always starting with 1 and increasing) of the order in which files are to be loaded into DisparityBox. The second number on each line indicates which subclade the group belongs to. Note that all groups loaded must be included in one of the subclades. {I'm not sure right now if the software will work if you include a subclade with only group in it. It might, might not. Let me know if it doesn't and I'll fix it….}.

The listing below is an example of a *Subclade List File*, which is called exampSCa.txt and which has two subclades {Bills and Dolphins, as I couldn't remember the correct genera in this case, so the example is purely fictitious…}

Example *Subclade List File* {exampSCa.txt}

```
2        % first line is the numuber of subclades
Bills
Dolphins
1 1                    % dent
2 1                    % el is in subclade 1
3 1                    % car is in 1
4 2                    % pir  next three are all in subclade 2
5 2                    % nat
6 2              % man, note necessity of line feed after this line!  Unlike most IMP files...
```

Note that the % starts a comment on each line which will not be loaded into Disparity Box. Comments can only a appear on lines which have only numerical values on them, ie. comments cannot appear on the lines with subclade names. In this example, there are 2 subclades named Bills and Dolphins. The first three groups loaded will be assigned to the Bills subclade, while the second three (less fortunate) groups will be assigned to the Dolphins subclade {yes, I'll fix this later…}.

The comment lines identify the group name corresponding to each group. DisparityBox refers to groups by the ordinal number in which groups are loaded, not by name, so the comments are for the users convenience only.

Running the Calculation

1.) Create the *Subclade File List*.

2.) Start DisparityBox6 and load the files of each of the groups, in the same order they appear in the Subclade File List you have created. It may help to print out the subclade file list, so that you can get the loading order correct.

3.) Hit the **Find Grand Consensus Mean (Groups)** button.

4.) Hit the **List Loaded Sets** button to verify that you have the loading order correct.

5.) Use the **Hierarchial Disparity** menu to **Load Subclade File**, and load the Subclade file list you made earlier.

6.) Use the **Calculate Hierarchial Disparity** option under Hierarchial Disparity menu or the **Bootstrap Hierarchial Disparity** as desired.

Nearest Neighbor Analysis of Disparity

This is a discussion of how one might implement the Nearest-Neighbor Analysis of disparity using geometric morphometric methods. The initial approach is to follow the method outline by Foote (1990). In Foote's approach each species is represented by a single specimen (which may be an individual or a mean of a group).

Summary of Foote's Approach

The initial step is to determine the nearest-neighbor distance d_i for each of the n species (groups) in the study. For a geometric morphometrics approach to this problem, we would form a matrix of all pairwise Procrustes distances between specimens, and then find the minimum of this distance matrix along each row, yielding the nearest-neighbor distance for each specimen. Foote's original approach uses a Fourier measure based on an outline, which has 12 components which describe 99% of the shape information.

The next step in the procedure is to form a Monte-Carlo set of simulated data, based on the ranges of the observed values in the variables measured. This is done by estimating the mean value and the range of each variable (Foote) or landmark coordinate (geometric morphometrics). A computer random number generator is used to generate n-1 specimens with coordinates or variables drawn randomly from the observed range. The nearest neighbor distance r_i is computed for each of the observed forms (the ith specimen) to each of the randomly drawn forms. Note here the nearest neighbor distance

is computed for each *observed specimen* to each *Monte-Carlo specimen,* this distance r_i is not among Monte-Carlo specimens.

It is then possible to form a proportional distance p_i for the ith specimen

$$p_i = (d_i - r_i)/r_i$$

and p_{mean} is the average over all p_i. If p_{mean} is less than zero, the points are clustered, if p_{mean} is greater than zero, there is a repulsion effect. To get an estimate of the range of p_{mean}, the Monte-Carlo simulation is run many times (20 times in Foote's paper). Foote indicates little variation in p_{mean} over these 20 trials.

To determine the range of variable values (coordinate positions) , Foote uses estimates of the "true" minimum Y and the true maximum X of a distribution as

$$Y=(nA-B)/(n-1)$$

$$Z=(nB-A)/(n-1)$$

Where A is the lowest observed value in n specimens, and B is the highest observed (Strauss and Sadler, 1989).

Rather than rely directly on this approach to estimating the observed minimum and maximum, Foote determines the mean and the standard deviation of a normal distribution fitted to the data. Working from Feller (1968), he notes the relationship

$$x_{mean} = Y + (Z-Y)/2$$

$$SD_x = \{(Z-Y)^2/12\}^{1/2}$$

and so uses the mean and standard deviation of each variable to determine the range parameters.

$$Y = x_{mean} - 3^{1/2} SD_x$$

$$Z = x_{mean} + 3^{1/2} SD_x$$

Extension of Foote's Approach to geometric morphometrics

The extension is quite straightforward. The distance used would be a Procrustes distance in the tangent plane (probably a Partial Procrustes). The distances d_i would be determined first. The mean and standard deviation of each coordinate at each landmark would be determined, and used to estimate the range of observed values for each coordinate.

A series of Monte-Carlo data sets would then be formed and r_i determined for each specimen. The p_{mean} value would be determined for each Monte-Carlo set, to form a distribution of p_{mean} values over the Monte-Carlo set. It would then be possible to carry out all the usual statistical tests using this distribution.

This is unlike our work, in that it requires a few specimens from many species, rather than many specimens from a few species. I think this method requires a large number of species to "fill" the space in cleanly, many missing groups could easily lead to very erroneous conclusions, particularly if there was a bias in the sampling. But this isn't really news to anybody, is it?

Foote (1990) has n (species counts) ranging from 22 to 87, and uses 12 variables. Increasing the number of variables (landmarks) will probably increase the number of species needed to get a good "coverage" of the space.

Distribution Options in Nearest Neighbor Analysis

Note: This manual on nearest-neighbor models in Disparity Box is not complete. Disparity Box allows the use of two different models of the distribution of points in the nearest neighbor calculations, either uniform (ala Foote) or gaussian. It also allows for use of the Strauss-Sadler range estimates, or use of the usual standard deviation estimates assuming a normal distribution.

Contact either Dave (sheets@canisius.edu) or Miriam (zelditch@umich.edu) if you really need to use these models and need more info, and one of us will write up a more complete explanation, as time allows.

Additions 8/15/02

I have added an option which allows use of a chain file to load the input files into DisparityBox, which means one does not have to load files in individually.

Creating a Chain File

A chain file is an ascii text file, listing all the data files you want to load into disparitybox. Create this file using Word, Wordpad, Notepad or some other text editor. Each file you want to load into Disparity Box must appear on a single line in the file, with it's file extension included. Save the chain files and all the files that are listed in the chain file in a single directory. Start DisparityBox, go to the File menu and chose the option *Load from Chain File List.* When you do this, all files listed in your chain file will be loaded into DisparityBox, in exactly the same fashion is when you use the *Load Data*

Set button, except that all files will be loaded automatically. When the files are loaded, Disparity Box will display the names and specimen counts of the files you have loaded.

There are two examples of Chain File included in the compressed version of DisparityBox, and the specimen files to go with those chainfiles. The chain files included are called *disparitybox_chain.txt* and *disparitybox_chain2.txt*. Try loading them and looking at them using a text editor/word processor, so that you can construct your own chain files. Note that the file *disparitybox_chain2.txt* has a deliberate error included in it, there is a non-existent file listed in it. This was done so that the error-trapping in the chain file will be demonstrated when you load this file. Do try it.