Creation Research Society Quarterly 2024 61:101–111

Simulation of the Pearson Method and the Spearman Method Coefficients in Morphology-Based Baraminological Research

Harry F. Sanders, III

Abstract

Statistical baraminology has been a common research topic in creation biology for the last several decades. Several statistical methods have been employed but none have been tested against simulated data where the result is known. Therefore, both the original Pearson method and newly deployed Spearman method are tested in this paper against simulated data where the existing patterns were known. Multiple patterns of varying strengths were tested. In every case, the Pearson method outperformed the Spearman method. Based on these preliminary results, the Pearson method (aka the Pearson correlation coefficient) should be the preferred method of statistical baraminology. However, much more research is needed to ensure that baraminology is placed on a strong empirical foundation.

Introduction

Statistical baraminology is the methodology of choice for determining baramins for many in the creation science community. Statistical baraminology uses a distance correlation equation drawn from cladistics (Sokal and Sneath, 1963) to measure the distance between taxa, assuming the more closely related taxa have a lower distance (Robinson and Cavanaugh, 1998). These measurements are done using statistical algorithms, and the data is drawn from the secular phylogenetic literature. Such data is assumed to be completely free of bias (Wood, 2011), despite the evolutionary literature being open about possible biases (Winsor, 1994). The data is analyzed in the web applications BDISTMDS and/or BARCLAY and visualized in graphs and multidimensional scaling plots. For an extensive review of how statistical baraminology works, see Sanders and Cserhati, 2022.

The web applications BDISTMDS and BARCLAY are the primary tools used by baraminologists. The original

---

* Accepted for publication February 9, 2024

method, BDISTMDS (Wood, 2001), has recently been superseded by BARCLAY (Wood, 2021), though the original method's results have been deemed close enough to be accurate (Wood, 2021). BDISTMDS used the Pearson coefficient as its statistical base, while BARCLAY originally defaulted to the Spearman coefficient (Wood, 2020). BARCLAY now defaults to Pearson despite the original paper claiming Pearson did not meet the required assumptions for use (Wood, 2020). The current recommendation is to use multiple clustering techniques in baraminological analysis (Wood, 2021). However, no simulation studies testing either coefficient have ever been published. While the desire to apply statistical baraminological methods to real-life data is understandable, knowing whether the methods work as intended requires careful testing and simulations where the outcome is known. Therefore, this paper will perform simulation studies on both the Pearson method and the Spearman method to determine if they work as intended and what situations, if any, each one excels at.

Importantly, this is not intended to be the final word on simulation studies in baraminology. This study does not simulate MDS plots, nor does it deal with either PAM or FANNY clustering. Instead, it is intended to provide a foundation for future testing and provide a test for the statistical methods which has, as yet, not been done. Follow-up testing should be done to further tune and test statistical baraminology.

Methods

To perform these experiments, a dataset was generated in Microsoft Excel using the "RANDBETWEEN" function. The dataset was created using 100 simulated taxa and 1,000 simulated characters and randomized before each trial. Character states were permitted to be integers in a range from 0–4. This range of five numbers was selected to allow a wide range of possible characters that would match a wild-type dataset as well as make pattern detection easier. No question marks (representing unknown character states) were introduced into the data, as the desire was to give both methods a best-case scenario. Groups of ten, twenty, fifty, and one hundred taxa were compared using 50, 100, 200, 500, and 1,000 characters. These numbers were chosen as rough approximates of what is available in the mainstream scientific literature. It is rare that a dataset will have more than 100 taxa or 1,000 characters, therefore these were selected as the upper boundaries. It is also uncommon but not unheard of for a dataset to have less than ten taxa or 50 characters, therefore these served as the lower bounds.

The dataset was used in seven separate experiments. First, both Pearson and Spearman were exposed to purely random data. In the succeeding experiments, patterns were introduced to the data. For the second experiment, two taxa were given identical characters for X number of traits. For the third, a second pattern of equal strength but completely different from the first, was given to two different taxa. Effectively, this should have created two clusters. For the fourth experiment, a single pattern with three taxa was used. For the fifth experiment, two patterns with three taxa each were created. In the sixth experiment, there were three patterns given to three taxa each, which should create three clusters. In the final test, both methods were presented with patterns that were identical for all characters in the dataset at varying numbers of patterns and taxa.

For the purposes of this experiment, a pattern is a preset sequence of characters shared across multiple taxa. The pattern represents similarity due to potential ancestry. Taxa where a pattern was introduced might also share random characters generated due to the dataset, but these characters do not represent ancestry. As an example, two species might share number of vertebrae, dentition, eye structure, and specialized structures because of ancestry, but share similar diets, gut length, and size with a third species based purely on chance. The pattern based on ancestry is the pattern baraminologists want to find in the data.

Patterns varied from 0.5% to 50%. In practice, this means that if the experiment used 50 characters, no more than 25 were deliberately made similar between two or more simulated taxa. If an experiment used 1,000 characters, no less than five were deliberately made similar between two or more simulated taxa. When a pattern was introduced, it was given to a select number of taxa, varying from two to three. In some tests several different patterns were introduced to create multiple clusters. To introduce a pattern, the randomized character data was altered so that the states of anywhere from 5–500 characters were identical for the selected taxa. Each dataset was tested at the aforementioned number of characters and taxa. As an example, if 10 taxa were selected for analysis, along with 100 characters, a 10% strength of pattern would be 10 characters being placed in states that were identical. Character 1 for taxa A and B might be simulated as having a value of "1" while Character 3 for the same taxa might have a value of "0." No matter how many patterns were introduced, the same strengths of patterns were used for each number of characters. The tests were identical for each method, and each pair of tests was run using identical datasets. Each unique combination of number of taxa, number of characters, strength of pattern, and number of pattern was run once for the Pearson method and once for the Spearman method. In total, 520 unique combinations were submitted to both algorithms. In total, 1040 tests were performed.

Note that for the purposes of this study, multidimensional scaling, a visualization tool used to represent calculated distances, was not performed. The goal here was not to interpret the results of the baraminic distance calculation, merely to determine how well the algorithms performed at detecting patterns and rejecting noise. Further, given the subjectivity of the MDS plots, they would not have been helpful in this scenario. FUZZY and PAM plots were also not performed as the point was to determine how well the Pearson method and the Spearman method coefficients performed compared to one another when the correct result was known. The Spearman method tests were performed using the BARCLAY algorithm, while the Pearson method tests were performed with the BDISTMDS algorithm. It is possible not using BARCLAY for both may have impacted the results slightly as the two algorithms are not identical. BARCLAY no longer provides the bootstrapping results (Wood, 2020), that BDISTMDS did (Wood, 2008) as an example. However, given that Wood has deemed the results of the two methods comparable (Wood, 2021) this seems unlikely.

Before going any further, it is important to define some terms. The first is the strength of pattern. When used, it refers to how strong of a pattern was introduced into the dataset. A higher strength of pattern means that more matching characters were placed in the data. The introduction of a pattern requires predefining a certain number of characters so that a given percentage of characters in taxa A and taxa B are identical. Thus, if taxa A and B have a 50% strength of pattern, they are predefined to be identical in 50% of the characters used for analysis. The remaining characters are allowed to vary randomly. If it is assumed that taxa with similar characters are likely related, then changing the strength of pattern reveals which of the current statistical programs is better at detecting those relationships.

The second important term is noise. For this study, noise was defined as any positive or negative correlation appearing on the graph that should not have been there based on the pattern placed in the data. No noise was deliberately introduced into the data. Instead, any character states not predefined in the experiment varied randomly. Effectively noise was the presence of a false positive or negative correlation between two taxa. The more false positives or negatives that are present, the larger the amount of noise. A false negative was marked as a failed test because it indicated that two taxa that should be seen as similar were being read as dissimilar. A false positive was marked as a failed test because it indicated that two taxa that should not be similar were found to be similar. Just one of these errors in a given test resulted in a failed test. The goal is to determine what was required to produce 100% accuracy in the results. If a negative correlation (discontinuity) was not present, but the positive correlation (continuity) was all correct, the test was marked as passed. Missing discontinuity was ignored because only continuity was built into the data and the presence of continuity does not necessarily imply discontinuity (Wood et al., 2003). In other words, just because two organisms are similar to each other, it is not necessarily implied that they are dissimilar to a third organism. However, if continuity was absent from where it should have been present, this was marked as a failed test, because continuity was built into the data. If discontinuity was present when continuity should have been present, this was also marked as a failed test. The goal was to determine how well each method detected continuity patterns in the data.

Results and Discussion

Random Data Experiment

When presented with purely random data, no matter how what combination of characters and taxa numbers were used, the Spearman method always found patterns of continuity. The more characters and taxa were added, the worse at filtering out noise the Spearman method became. As shown in Figure 1a and b, the Spearman method had some noise when combining ten taxa with fifty characters, but a much larger amount when comparing 100 taxa with 1,000 characters. The Pearson method showed no noise at the ten and twenty taxa levels. However, when the character to taxa ratio met or dropped below the 2:1 threshold, the Pearson method began to present noise as well.

The fact that the Spearman method always finds patterns in purely randomized data is disconcerting. That patterns potentially exist in randomized data is possible, given each cell in the dataset is filled randomly. If this were the case, however, it is concerning that the Pearson method does not also find these patterns. Further, given the data is created randomly for each test, if the problem were patterns within a given dataset, the Spearman method would not be expected to find patterns on every occasion. Given that only the Spearman method finds these patterns, it raises the possibility that the switch from the Pearson method to the Spearman method coefficients made when Wood (2020) introduced the BARCLAY algorithm has created a tendency to find much weaker patterns than had been done previously.

Single Pattern

For this test, a single pattern was introduced into the dataset. Taxa A

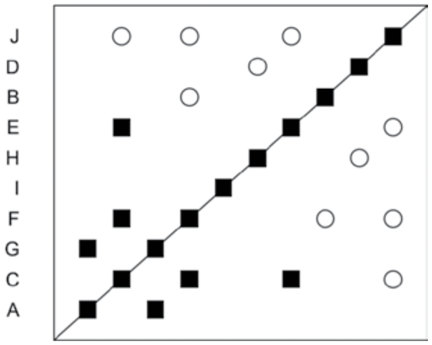

Figure 1A. Spearman coefficient result when presented with randomly generated data. Ten taxa and fifty characters were used for analysis.

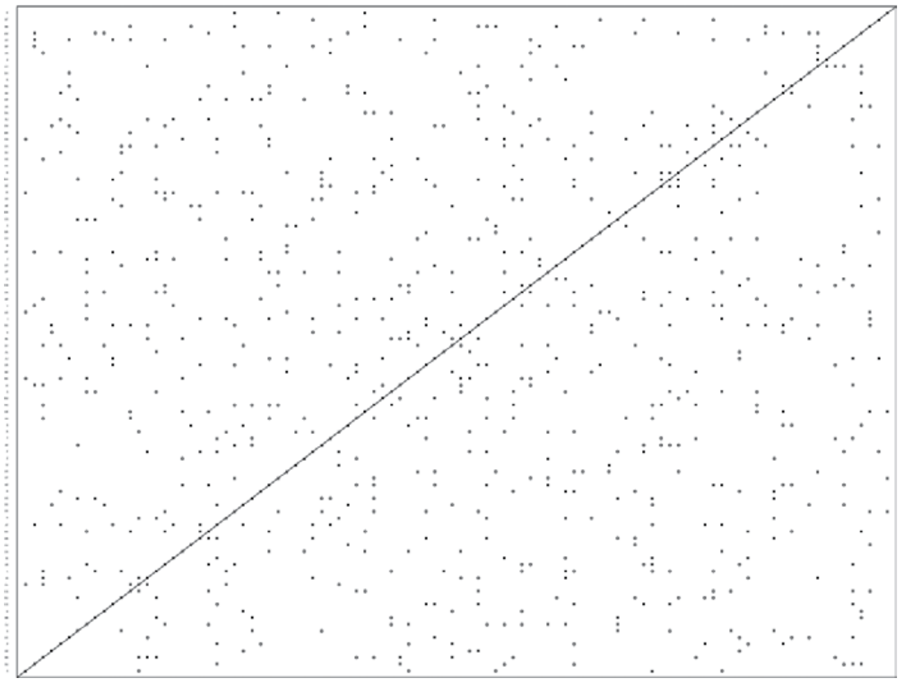

Figure 1B. Spearman coefficient result when presented with randomly generated data. 100 taxa and 1000 characters were used in this analysis.

and B were given a pattern that varied in strength as described above. Both methods performed poorly, when the number of taxa was equal to or greater than the number of characters, which is unsurprising. However, the Spearman method always produced much more noise than the Pearson method in this scenario (See Figure 2a and 2b for an example)[1].

When 50 characters were used, the Pearson method detected the pattern when it was fifty percent of the dataset and characters were more than the number of taxa (Figure 3a). The Spearman method never successfully detected the pattern with no noise (Figure 3b). In fact, when only one pattern was used, the Spearman method never successfully separated the pattern from the random background noise.

When 100 characters were used, the Pearson method again successfully detected the pattern at the ten and twenty taxa levels when the pattern reached 50% of the total characters. When 200 characters were used, the Pearson method successfully detected the pattern at the 50% mark for the 10, 20, and 50 taxa levels (Figure 4a). It also successfully detected the pattern when the pattern was only 25% strength, and 50 taxa were used. When 500 and 1000 characters were used, the Pearson method always discovered the pattern at the 50% mark, regardless of how many taxa were included. The Spearman method never successfully detected the pattern (Figure 4b).

Two Patterns

For this test, an additional pattern was introduced into the dataset. Taxa C and D were given a pattern that was equal in strength, but completely discontinuous from Taxa A and B's pattern. Again, both methods performed poorly when the number of taxa was greater than or equal to the number of characters. As before, the Spearman method performed much worse in this area than the Pearson method, recording much higher false positives and negatives (Figure 5a and b).

When 50 characters were used, the Pearson method and the Spearman method both found the patterns at the 10 taxa level with 50% strength of pattern and the Spearman method even found discontinuity between the groups (Figure 6a). However, when more taxa were introduced, the Spearman method began producing noise (Figure 6b). The Pearson method however, successfully recovered the pattern at the 20 taxa level when the pattern was 40% of the dataset or stronger. At the 100 character level, the Spearman method does not find either pattern without noise. By contrast, the Pearson method finds both patterns at the 30% range and above when 50 taxa or less are used. At the 200 character level, the Spearman method again failed to detect the patterns without noise. The Pearson method was successful at detecting the pattern at the 50% strength

---

[1] Note that all results are not shown as there are too many to fit in an article. Those selected should serve as examples. If a reader wishes to duplicate these result, they are welcome to contact the author for assistance should it be required.

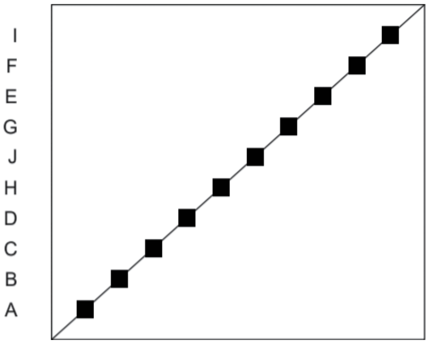

Figure 2A. Pearson coefficient result when presented with randomly generated data. Ten taxa and fifty characters are used for this analysis.

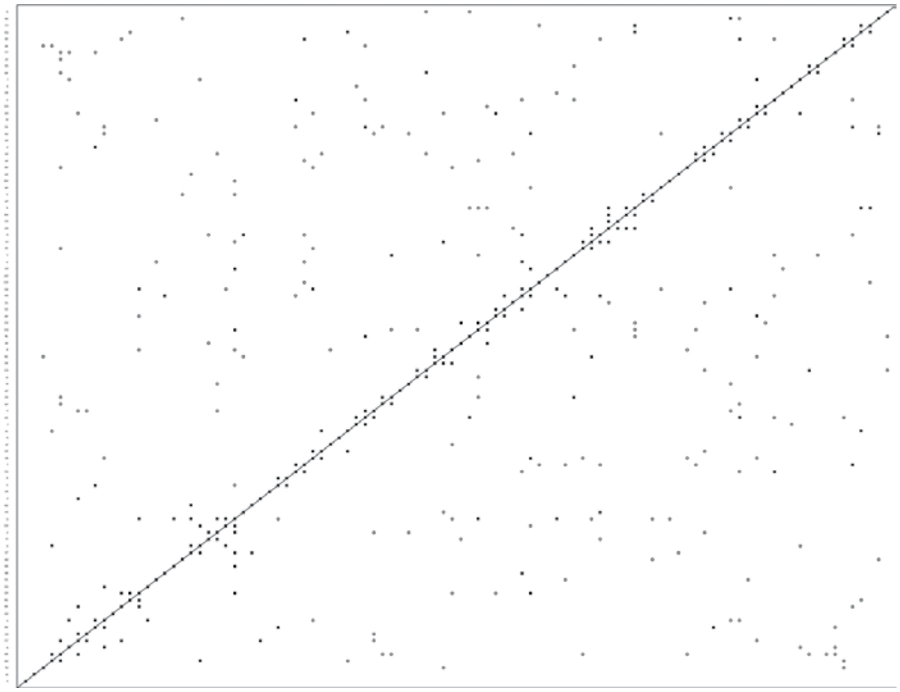

Figure 2B. Pearson coefficient result when presented with randomly generated data. 100 taxa and 1000 characters were used in this analysis.

mark when up to 50 taxa were used and successfully detected the pattern at 25% strength when 50 taxa were used. At the 500 character level, the Spearman method fails to find the patterns through the noise. The Pearson method results are similar to the 200 character level except they now find patterns at the 50 and 100 taxa level with 20% strength and above, and find the patterns at the 40% level no matter what the taxa number. At 40% strength, the Pearson method also sometimes finds appropriate discontinuity between the groups at the 50 taxa and above level. At the 1,000 character level, a very similar pattern was held. The Spearman method could not find the patterns without noise and at the 50% level, the

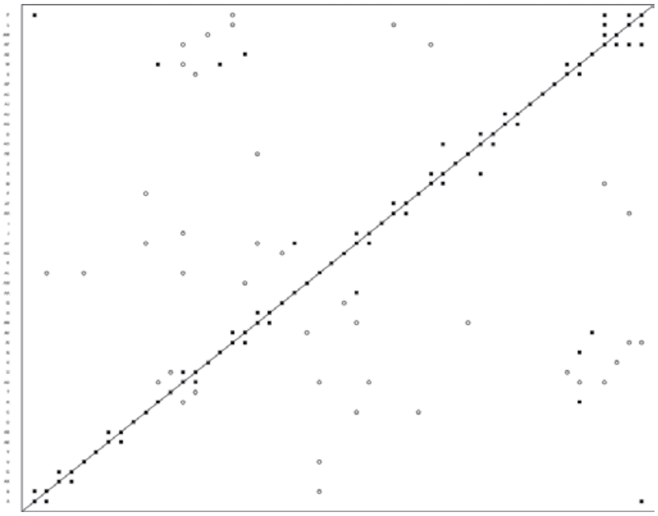

Figure 3A. Pearson coefficient result when presented with a single two taxa pattern. 50 characters and 50 taxa were used for this analysis.

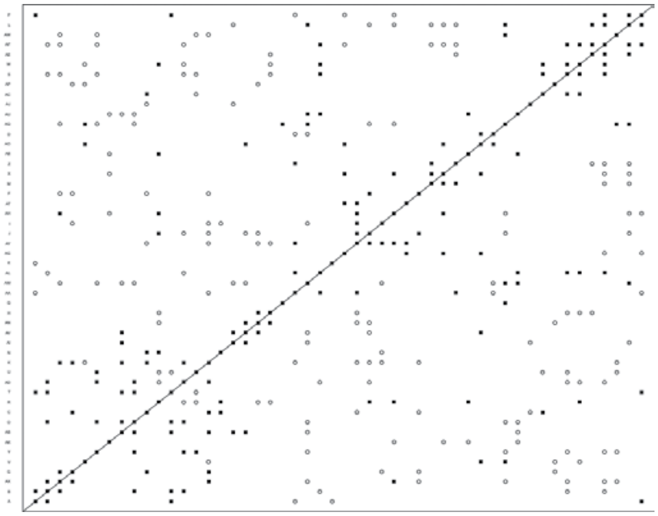

Figure 3B. Spearman coefficient result when presented with a single two taxa pattern. 50 characters and 50 taxa were used for this analysis.

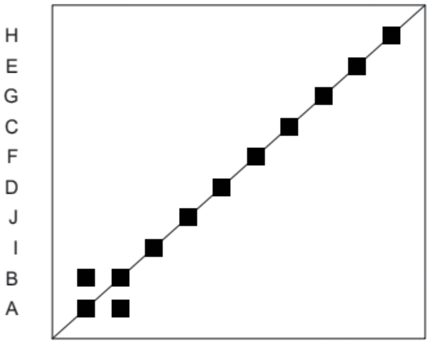

Figure 4. Pearson coefficient result when presented with a single two taxa pattern. 50 characters and 10 taxa were used for this analysis. Strength of pattern was 50%.

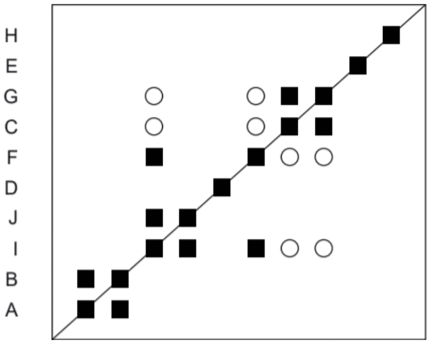

Figure 5. Spearman coefficient result when presented with a single two taxa pattern. 50 characters and 10 taxa were used for this analysis. Strength of pattern was 50%.

Pearson method always was able to detect the pattern. At 50 or more taxa, 20% was enough to detect the pattern.

Three Taxa Pattern

To test whether the prevalence of the pattern had any role in the Pearson method and the Spearman method successfully detecting patterns, a single pattern of three taxa, A, B, and C, was introduced into the dataset. As before, the pattern varied in strength depending both on the number of characters used and the number of taxa tested.

As in the previous single pattern scenario, the Spearman method never successfully detected the pattern without also detecting background noise. In some cases, the Spearman method was all noise, failing to distinguish the pattern at all. This outcome was most common at high numbers of taxa and characters.

The Pearson method performed significantly better. At the 50 character level, the Pearson method detected the pattern at the 50% level when ten and twenty taxa were used, and the 40% level when twenty taxa were used. As before, when characters were at less than 2:1 ratio to taxa, the Pearson method produced statistical noise, though not in the same quantities as the Spearman method.

At the 100 character level, 50% strength of pattern allowed the Pearson method to pick up on the pattern at the 10 and 20 taxa levels. This was also true at the 200 character level, with the pattern also being detected at the 50 taxa level. At the 500 character level, 50% strength of pattern was detected in all taxic levels. At the 20% strength of pattern level, the Pearson method detected the pattern from 20 taxa upwards (Figure 7). As before, 50% strength of pattern allowed for pattern detection for all taxic numbers at the 1,000 character level and 25% strength

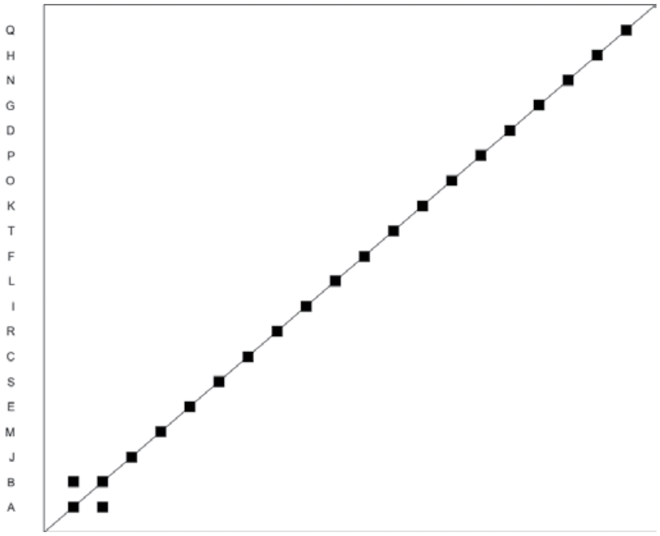

Figure 6A. Pearson coefficient result when presented with a single two taxa pattern. 200 characters and 20 taxa were used for this analysis. Strength of pattern was 50%.

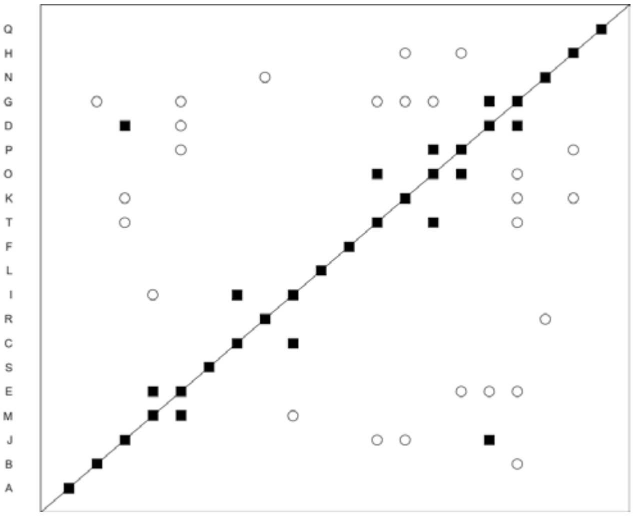

Figure 6B. Spearman coefficient result when presented with a single two taxa pattern. 200 characters and 20 taxa were used for this analysis. Strength of pattern was 50%.

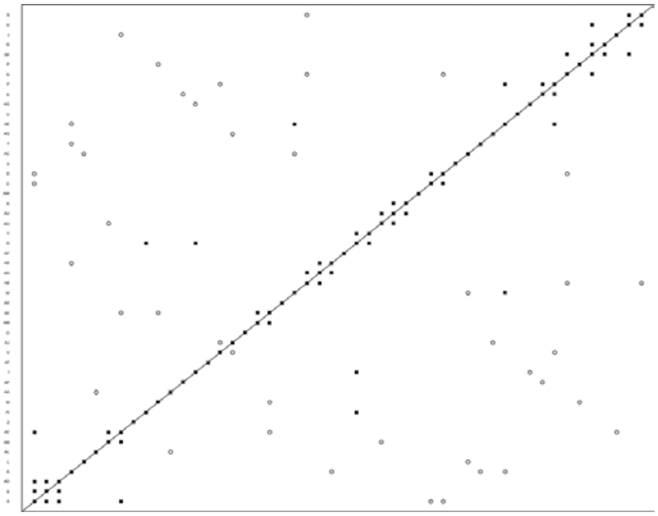

Figure 7A. Pearson coefficient result when presented with two, two-taxa patterns. 50 characters and 50 taxa were used for this analysis. Strength of pattern was 50%.

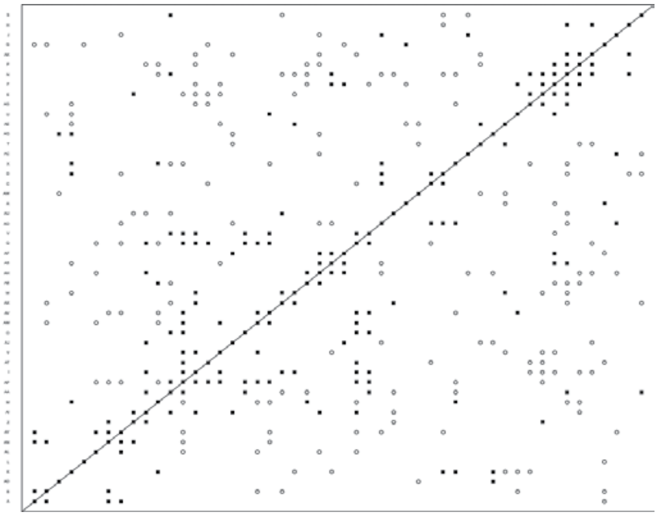

Figure 7B. Spearman coefficient result when presented with two, two-taxa patterns. 50 characters and 50 taxa were used for this analysis. Strength of pattern was 50%.

of pattern was enough at the 50 and 100 taxic levels.

Two 3 Taxa Patterns

To test if prevalence and number of patterns had an effect on the Pearson method and the Spearman method outcomes, two patterns, A, B, C and D, E, F were introduced into the dataset. The introduced patterns were discontinuous with one another and, as before, varied in strength depending on the number of characters and taxa used in the analysis. They are meant to represent two internally continuous, externally discontinuous patterns. They could potentially be holobaramins, but, as discontinuity was ignored (unless it appeared where continuity should be), it cannot be stated definitely.

The Spearman method was able in this scenario to detect the two patterns, but only did so without noise twice out of twenty different scenarios. One time was with 10 taxa, 50 characters, and a 50% pattern strength. In that same scenario, the Pearson method also detected the two patterns, but the Spearman method detected more expected discontinuity (Figure 8a and 8b). The second was at 1,000 characters, 10 taxa, with 25% strength of pattern. Given it failed at the 50% strength of pattern, this result is probably a result of random variation in the dataset as characters randomly changed to strengthen or reduce patterns.

The Pearson method outperformed the Spearman method again in this scenario. At the 50 character level, it correctly determined the patterns at the 50% strength of pattern for the 10 and 20 taxa level and at 40% for the twenty taxa level. Above that, as before, was simply no

At the 100 character level, 50% matching was required for the Pearson method to detect the patterns without noise at the 10 and 20 taxa level. These results carried over into the 200-character level but extended to the 50 taxa level. Further, at the 50 taxa level, 25% strength of pattern was enough for the Pearson method to correctly determine patterns. At the 500-character level, the results were similar. All taxa numbers detected the pattern at 50% strength of pattern and the 50 and 100 taxa levels also detected the pattern at 20% strength of pattern. At the 1,000 taxa level, 50% strength of pattern allowed the pattern to be detected across all taxa numbers tested. Further, 25% strength of pattern was enough from 50 taxa upwards.

3 Patterns 3 Taxa

As a final test, 3 patterns, A,B,C, D,E,F, and G,H,I, were introduced into the dataset. At the ten taxa level, this is equivalent to three groups with an outgroup, something not uncommon in real biological data so this test may be regarded as the closest to an actual dataset. As before, strength of pattern varied with the number of characters and taxa used in the analysis. As before, the Spearman method performed poorly in this test. The Pearson method performed much better as in previous experiments, provided the ratio of character to taxa was kept 2:1 or higher. At the 50 character level, it correctly determined the patterns for the 10 (Figure 9) and 20 taxa levels at 50% strength of pattern, as well as 40%

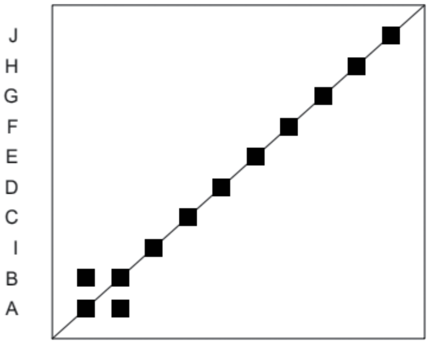

Figure 8A. Pearson coefficient result when presented with two, two-taxa patterns. 50 characters and 10 taxa were used for this analysis. Strength of pattern was 50%.

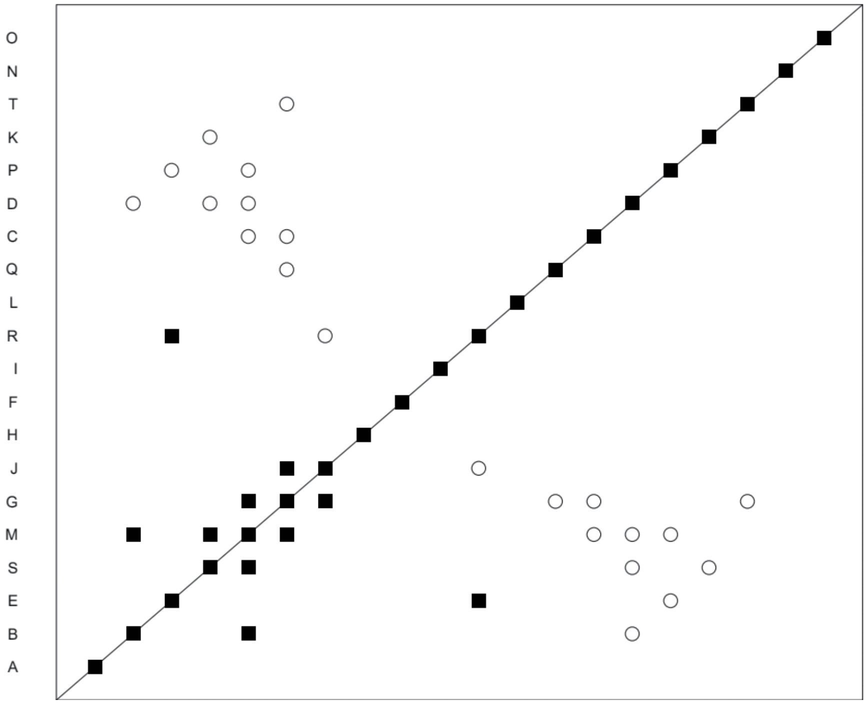

Figure 8B. Spearman coefficient result when presented with two, two-taxa patterns. 50 characters and 20 taxa were used for this analysis. Strength of pattern was 50%.

strength of pattern at the 20 taxa level. The Spearman method includes the outgroup taxa J as part of one of the groups (Figure 10).

At the 100 character level, The Pearson method successfully detected the pattern at 50% strength of pattern up to the 50 taxa level. At 30% strength of pattern, it detected the pattern at the 20 and 50 taxa level. At the 200 character level, 50% strength of pattern was required to detect the pattern up to the 50 taxa level. At 500 characters, 50% strength of pattern was required to detect patterns at all taxic levels. However, 20% was enough at 50 taxa and above. At 1,000 characters, the same was true but, instead of 20%, 25% strength of pattern was enough to detect the pattern when the taxa numbers were 20 and above.

100% strength

Just to ensure that it was possible to get a correct answer, both the Pearson method and the Spearman method were given a 100% strength of pattern for two taxa, two pairs of taxa, three taxa, and two trios of taxa at the ten taxa, 100 character level. The same pat-

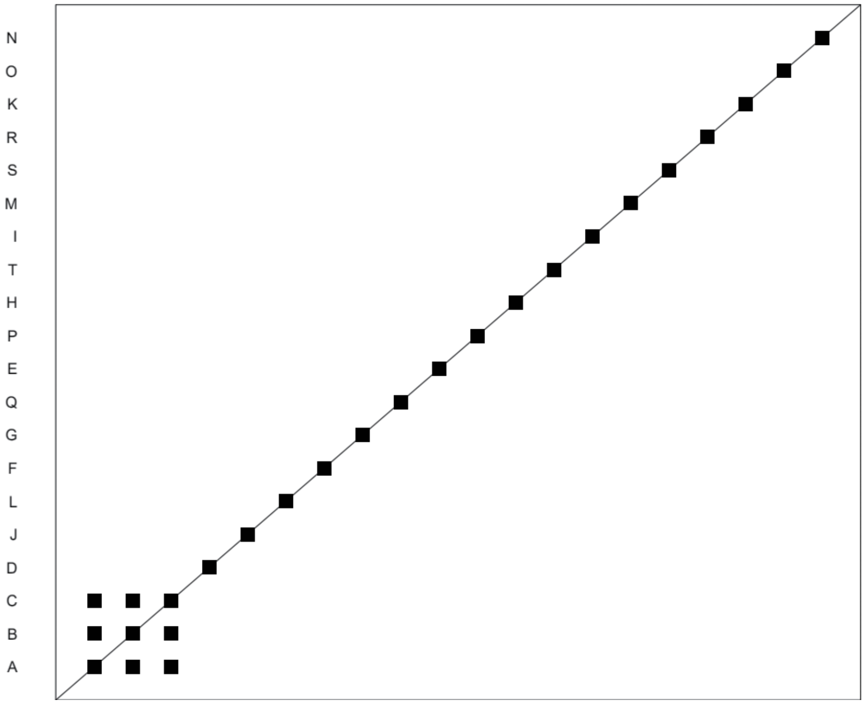

Figure 9. Pearson coefficient result when presented with one three-taxa patterns. 500 characters and 20 taxa were used for this analysis. Strength of pattern was 40%.

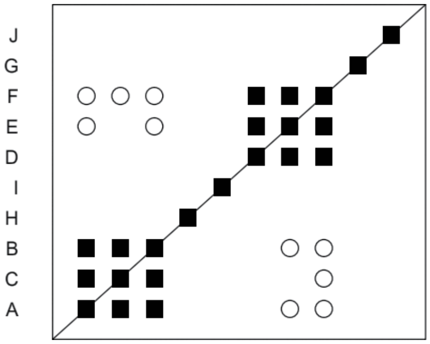

Figure 10A. Pearson coefficient result when presented with two, three-taxa patterns. 50 characters and 10 taxa were used for this analysis. Strength of pattern was 50%.

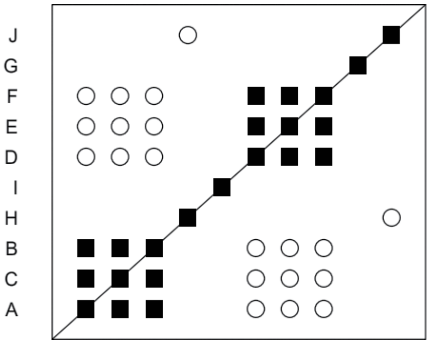

Figure 10B. Spearman coefficient result when presented with two, three-taxa patterns. 50 characters and 10 taxa were used for this analysis. Strength of pattern was 50%.

tern as before was followed. A,B was the first pattern introduced, then C,D was added. For three taxa, the groups were A,B,C and D,E,F. The Pearson method passed every test (Figures 11, 13, 15, 17). The Spearman method failed every test (Figures 12, 14, 16, 18). Even when given a 100% strength of pattern, the Spearman method failed to correctly distinguish pattern from noise. In some cases the Spearman method found whole groups that were not present in the data. In others, it joined outgroups to existing groups. By contrast, the Pearson method successfully found the patterns on every occasion. This evidence alone should be enough to cause concern about the Spearman method coefficient.

Performance Analysis: The Spearman Method

Wood (2020) introduced the Spearman method coefficient into baraminology and performed studies that demonstrated it produced comparable results to the Pearson method. Those results may be correct, but only on a classification level. The actual distance matching between taxa is much different as has been shown above. It is important to point out that the datasets Wood used are unknowns. We do not know the correct answers. That is why simulated data is so useful. The correct answers are known before the tests are run.

The Spearman method performed poorly when attempting work with datasets where correct answers are known. Even when strength of pattern was very weak and matching was not expected (0.5%) (i.e., 5 matching characters per 1,000 characters), the Spearman method always found patterns. Of the 520 tests run during this simulation, the Spearman method correctly reported the patterns without noise, false positives or negatives, or missing connections just 5 times. Every time it did so, ten taxa were in use, along with a small number of characters, usually 50. These facts would seem to indicate that the Spearman method coefficient is inefficient at baraminological analysis and should be immediately withdrawn from use as results obtained using it are likely inaccurate. If use must be made of it, limiting it to small datasets, with few taxa and characters is probably best as it seems to perform best in those circumstances.

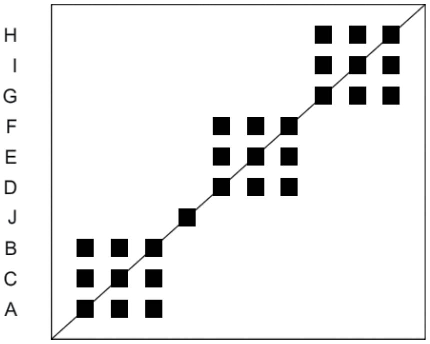

Figure 11. Pearson coefficient result when presented with two three-taxa patterns. 50 characters and 10 taxa were used for this analysis. Strength of pattern was 50%.

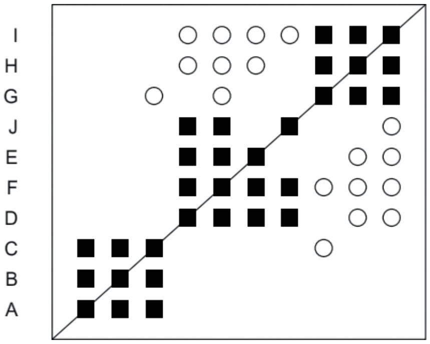

Figure 12. Spearman coefficient result when presented with two three-taxa patterns. 50 characters and 10 taxa were used for this analysis. Strength of pattern was 50%.

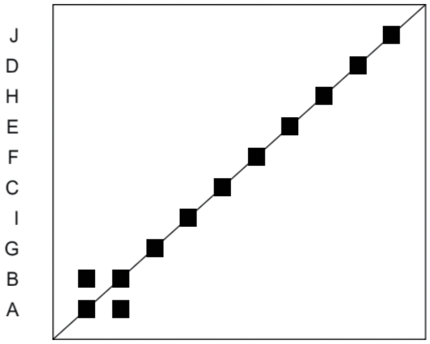

Figure 13. Pearson coefficient result when presented with one, two-taxa pattern. 100 characters and 10 taxa were used for this analysis. Strength of pattern was 100%.

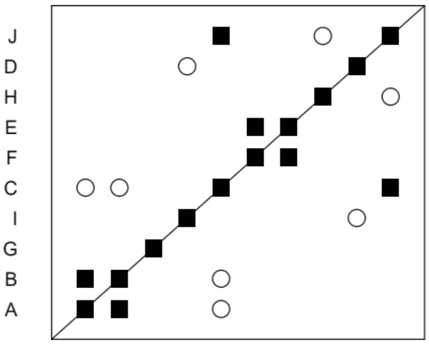

Figure 14. Spearman coefficient result when presented with one, two-taxa pattern. 100 characters and 10 taxa were used for this analysis. Strength of pattern was 100%.

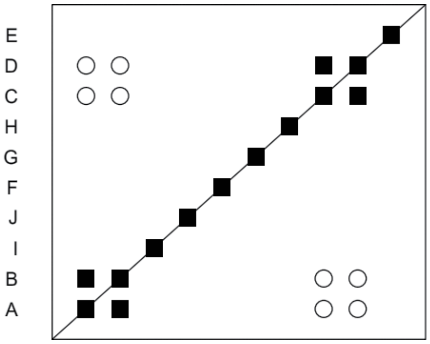

Figure 15. Pearson coefficient result when presented with two, two-taxa patterns. 100 characters and 10 taxa were used for this analysis. Strength of pattern was 100%.

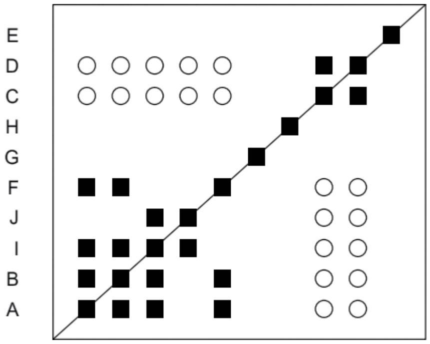

Figure 16. Spearman coefficient result when presented with two, two-taxa patterns. 100 characters and 10 taxa were used for this analysis. Strength of pattern was 100%.

Performance Analysis: The Pearson Method

The Pearson method coefficient (aka the Pearson correlation coefficient) has been criticized recently (Reeves, 2021a, 2021b) which prompted Wood (2020) to release BARCLAY incorporating the Spearman method coefficient. While Reeves's criticisms are fair and well-argued, the proposed cure seems much worse than the disease. Of the 520 tests run in this simulation, 123 times the Pearson method correctly identified the pattern without noise, false positives or negatives or missing connections. That number looks even better if the 15 times (of 20) the Pearson method coefficient correctly found no pattern in purely random data are added in. 138 correct answers out of 520 is much better than 5 out of 520. The strength of pattern both methods could detect was unknown and therefore very weak patterns were used as a baseline and increased up to 50%. Therefore, while 520 tests were done, many of those were expected to not detect patterns and, in many cases, the Pearson method did not produce patterns. In fact, only 126 times of 520 was noise present, compared to 515 times for the Spearman method. Noise was typically only present when the character to taxa ratio dropped below 2:1 in the Pearson method.

Further, the Pearson method correctly determined patterns at all taxic levels and all character counts. However, strength of pattern could be weaker at higher character and taxa counts and still be determined. When taxa counts increased individually, weaker patterns could be detected. This was also true of character counts. Therefore, the Pearson method seems to perform best when character and taxa counts are maximized. However, since no question marks were in the data to simulate absent values, it is impossible to know how the datasets would react to lower character relevance that might be required to increase taxa and character counts. This is an area where further research is required.

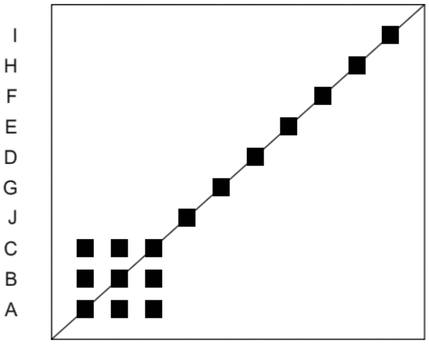

Figure 17. Pearson coefficient result when presented with one, three-taxa pattern. 100 characters and 10 taxa were used for this analysis. Strength of pattern was 100%.

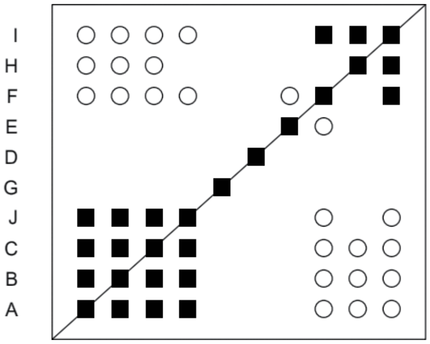

Figure 18. Spearman coefficient result when presented with ones, three-taxa pattern. 100 characters and 10 taxa were used for this analysis. Strength of pattern was 100%.

Future

The Pearson method tests should ensure the character to taxa ratio of 2:1 is used when selecting datasets. If the ratio of characters to taxa is allowed to drop below that level, noise is almost

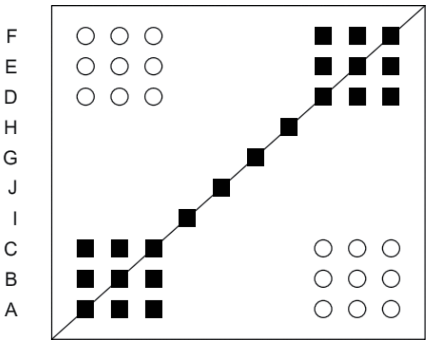

Figure 19. Pearson coefficient result when presented with two, three-taxa patterns. 100 characters and 10 taxa were used for this analysis. Strength of pattern was 100%.

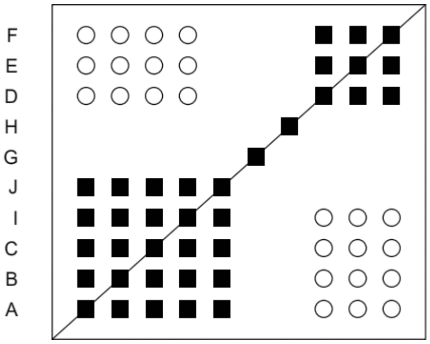

Figure 20. Spearman coefficient result when presented with two, three-taxa patterns. 100 characters and 10 taxa were used for this analysis. Strength of pattern was 100%.

always introduced into the results. Keeping the ratio at 2.5:1 or higher is probably optimal as in some cases even 2:1 produces noise. The 2.5:1 ratio was the lowest ratio that did not produce statistical noise in random data. If the character to taxa ratio is maintained above 2:1, then a strength of pattern of 60% or higher should reveal the correct patterns in most scenarios. However, lower strengths of pattern will work if the character and taxa counts are high enough. Generally, this means 50 taxa and at least 200 characters, which is not always feasible in real datasets. Lower numbers of taxa reduce strength of pattern required, if there are enough characters, usually 500 or more. Importantly, maintaining high levels taxa and characters should be done without lowering taxic or character relevance as neither of these parameters was tested in this study and thus changing them may have unknown effects on study results.

Conclusions

BARCLAY and BDISTMDS are popular with many in the statistical baraminological community because they are easy to use and produce relatively easy to interpret results. However, this basic simulation study produces some concerning results. When using the Spearman method coefficient, statistical baraminology is very unreliable, producing a correct result less than one percent of the time. Because the Spearman method coefficient results are riddled with false continuity and discontinuity, it should be considered unreliable as a method. The Spearman method coefficient, whatever statistical promise it may have brought to the table, must be abandoned. The Pearson method coefficient, however, shows much better results. While the Pearson method coefficient is open to question and may need to be replaced, it is significantly better than the current alternative. For those who wish to practice statistical baraminology, the Pearson method coefficient should be the preferred method for now as the alternative is currently unworkable.

References

Kaufman, L., and P.J. Rousseeuw. 1990. *Finding Groups in Data*. John Wiley & Sons, New York, NY.

Robinson, D.A., and D.P. Cavanaugh. 1998. A quantitative approach to baraminology with examples from the catarrhine primates. *Creation Research Society Quarterly* 34(4):196–208.

Reeves, C.R. 2021a. A critical evaluation of statistical baraminology: Part 1—Statistical principles. *Answers Research Journal* 14:261–269.

Reeves, C.R. 2021b. A critical evaluation of statistical baraminology: Part 2—Alternatives and conceptual and practical issues. *Answers Research Journal* 14:271–282.

Sanders, H., and M. Cserhati. Statistics, baraminology, and interpretations: A critical evaluation of current-morphology-based baraminology methods. *Creation Research Society Quarterly* 58:175–192

Sokal, R.R., and P.H.A. Sneath. 1973. *Numerical Taxonomy*. W.H. Freeman & Co., San Francisco, CA.

Winsor, M.P. 1994. "The Lessons of History." In *Models in Phylogeny Reconstruction*. Scotland, R.W., D.J. Siebert, and D.M. Williams (editors). Clarendon Press, Oxford, UK.

Wood, T.C. 2001. *BDIST Software, v1.0*. Center for Origins Research and Education. Bryan College, Dayton, TN. Distributed by the author.

Wood, T.C. 2005. Visualizing baraminic distances using classical multidimensional scaling. *Origins (GRI)* 57:9–29.

Wood, T.C. 2008. Baraminic Distance, bootstrap, and BDISTMDS. *Occasional Papers of the BSG* 12:1–17.

Wood, T.C. 2011. Baraminology, the image of God, and *Australopithecus sebida*. *Journal of Creation Theology and Science, Series B: Life Science* 1:6–14.

Wood, T.C. 2020. Expanding the toolkit of statistical baraminology with BARCLAY: baraminology and cluster analysis.

Wood, T.C. 2021. Baraminology by cluster analysis: A response to Reeves. *Answers Research Journal* 14:283–302.

Wood, T.C., K.P. Wise, R. Sanders, and N. Doran. 2003. A refined baramin concept. *Occasional Papers of the BSG (Baraminology Study Group)* 3:1–14.