FORENSIC SOIL BACTERIAL PROFILING USING 16S rRNA GENE SEQUENCING AND DIVERSE STATISTICS

By

James MacKenzie Hopkins

A THESIS

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Forensic Science—Master of Science

2014

ABSTRACT

FORENSIC SOIL BACTERIAL PROFILING USING 16S rRNA GENE SEQUENCING AND DIVERSE STATISTICS

By

James MacKenzie Hopkins

Evidentiary soil in an investigation can link an individual with the scene of a crime since the diversity and geospatial distribution of soils can make it highly probative. Recently, advanced techniques have been developed that allow a deeper investigation into bacterial communities and produce considerably more data than previous methods. This research used next-generation sequencing and statistical analyses to identify factors influencing soil bacterial communities and assess the feasibility for their use forensically. Soil samples were collected from a variety of habitats over different distances, depths, and times, DNAs were extracted, the *16S rRNA* gene amplified, and DNAs sequenced on a Roche 454 platform. Five statistical procedures—nonmetric multidimensional scaling, hierarchical cluster analysis, integral library shuffle, unique fraction method, and *k*-Nearest Neighbor—were used to compare differences or changes in bacterial communities. Multiple similar and diverse habitats were differentiated with both multivariate statistics and pairwise comparisons. Additionally, changes in communities were indicated over time, horizontal space, and depth. Multivariate statistics generally suggested similar relationships though not always consistent with pairwise comparisons, which showed analogous results though the unique fraction method always found fewer differences. *k*-Nearest Neighbor could be forensically useful based on the correct classification accuracy of 'unknown' samples from a non-ideal training set. This research elucidates the potential of next-generation sequencing for soil investigation, how samples should be collected, and what statistics would be useful to analyze the data.

ACKNOWLEDGEMENTS

There are many individuals who deserve thanks for helping with the research and completion of this thesis; the first, and most valiant, my advisor Dr. David Foran. Without his tireless hours of research meetings and thesis edits, this project and manuscript would be a far cry from what it is today. Additionally, my other committee members, Dr. Ruth Smith and Dr. Merry Morash, deserve thanks for their time that was dedicated for meetings and suggestions. Further thanks to my colleagues in the Michigan State University Forensic Biology Laboratory and Forensic Science program, especially my fellow second year Masters students, Ashley, Ashley, and Jordyn, as well as my protégé Ellen, whose hands this project is safely in. A special thank you to Michelle and Lisa for their help with sampling is also needed. I would like to thank Kylie Farrell for her help in sequencing my samples. Finally, without my friends and family, specifically Jason, Liz, Sherry, Terri, and my mummy, I would have never weathered this storm as well as I did. Thank you for keeping me grounded and focused. This project was supported by grant numbers 2011-DN-BX-K560 and 2013-R2-CX-K010, awarded by the National Institute of Justice, Office of Justice Programs, U.S. Department of Justice. Points of view in this document are those of the author and do not necessarily represent the official position or policies of the U.S. Department of Justice.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

Forensic Soil Investigation

Soil in an investigation can prove an invaluable evidentiary source for linking a suspect or victim with a crime. Potentially found on shoes, tires, shovels, or other objects, the virtually unlimited types of soil and their geospatial distribution can make such evidence highly probative (Saferstein, 2002). The utility of these properties for linking an individual with a geographic location, though explored in the works of Sir Arthur Conan Doyle (Alden, 2014), were not implemented into a forensic context until the early 20th century.

The death investigation of Eva Disch in 1908 was the first documented use of soil evidence in solving a criminal case (Bressan, 2010). Disch was found strangled in a bean field in Frankfurt, Germany. Crucial evidence found at the scene was a soiled handkerchief with particles of hornblende, snuff, and coal (Bergslien, 2012). After Disch's identity had been established, local authorities identified Karl Laubach as a suspect. Investigators enlisted Georg Popp, a chemist, to examine the soil found on Laubach's clothing. Popp identified two distinct layers of sediment in the pant cuffs worn by Laubach the day Disch was murdered. One was consistent with the soil at the crime scene. The other contained mica, which was consistent with the path between the scene of the crime and Laubach's home. Combining the two pieces of evidence and challenging Laubach with it, he confessed to murdering Disch.

While the admissibility of Popp's analyses in the Disch case would be questionable today, forensic scientists have continued developing more precise, accurate, and acceptable methodologies for the examination of soil evidence. However, the recent National Academy of Sciences report (National Research Council, 2009) has called many of the practices used in forensics into question, soil examination included, requiring a reassessment of what is currently being done and how it can be improved. Additionally, the *Daubert* ruling has elucidated the need

for forensic science to have accepted and peer reviewed procedures, with established error rates (Daubert v. Marrell Dow Pharmaceuticals). These requirements have pressured forensic scientists to develop more resilient techniques that incorporate the use of powerful statistics.

Classic Soil Analyses

Expanding upon the work of Popp and others, forensic geologists in the 20th century aimed to utilize the multitude of soil characteristics to classify, compare, and identify them (Saferstein, 2002). A collection of tests exists to analyze attributes of soil, and while they singly focus on its physical properties, they can be broken into four broad categories: general, microscopic, non-microscopic, and chemical (Saferstein, 2002).

The easiest examination for the comparison of soils, found in the general category and requiring no special equipment, is color (Saferstein, 2002). The organic and inorganic components of soil as well as moisture influence its visual appearance (Coyle, 2008). Samples may be dried or moistened, with other pretreatments available for normalization, before comparison against standard color charts, commonly Munsell (Saferstein, 2002). In this way known and unknown samples can be treated the same for an accurate visual comparison.

The broad grouping of microscopic analysis includes particle size distribution, which is considered the most useful physical property for soil examination (Saferstein, 2002). Further, automated image analysis systems can count large amounts of particles and generate soil profiles for comparisons against database samples. Though the light microscope is the most useful and cost-effective microscopic tool, newer technologies, including phase contrast, confocal, and electron microscopy are effective if available (Saferstein, 2002).

Diffraction techniques, in the non-microscopic category, include x-ray diffraction, which is particularly useful in detecting chemical compounds. Additionally, crystal compounds can be quantified and clay species identified (Saferstein, 2002). The accuracy, however, is limited by machine detection thresholds, variability of chemical and crystalline compounds, as well as other factors.

Finally, chemical methods revolve around elemental or organic compound analysis including x-ray fluorescence, inductively coupled plasma mass spectrometry, and energy/wavelength dispersive x-ray. Each method suffers from two major disadvantages; first, detection limits vary among machines and even elements, and second, they are destructive in nature (Saferstein, 2002). Infrared microspectroscopy bypasses the second limitation and is capable of identifying organic compounds; however, it is weaker in analyzing inorganic components.

Though the aforementioned techniques are established in geological soil examination, many are lacking when transitioning into forensics. A shortcoming of these techniques is they require large amounts of soil for testing, which is often unrealistic in a forensic scenario. Each, with the exception of a rare compound or element, measures class characteristics, leading to a general association of known and unknown samples. A match cannot be made with non-unique characteristics of soil, and greatly lowers the value of the evidence. Finally, the subjectivity in interpretation, *e.g.* matching soil color to a Munsell color chart, and difficulty in attribution of statistical significance (Pye, 2007) are serious limitations. Similar types of soils may not be differentiable from each other using these techniques, increasing the possibility of a false association among evidentiary samples. It is clear there is a need for techniques that capture the

unique characteristics of soil for better characterization and identification of this complex medium.

Molecular Analyses of Soil Bacteria

It has been estimated that $4 \times 10^7 - 2 \times 10^9$ prokaryotic cells are present in one gram of soil, representing up to 18,000 different genomes, which may themselves be underestimations (Daniel, 2005). The potential breadth of microbial diversity in soil, considering only the prokaryotic contribution, is staggering. Recent advances have allowed forensic scientists to assay the bacterial metagenome with the goal of using bacterial communities to link evidentiary and known samples. Several techniques exist to assay bacterial communities in soil; however, only a few have gained footing in the forensic sciences. The commonly enlisted techniques include denaturing gradient gel electrophoresis (DGGE), amplicon length heterogeneity-polymerase chain reaction (ALH-PCR), and, most popularly, terminal restriction fragment length polymorphism (T-RFLP).

DGGE requires the input of melted and re-annealed amplified product into a gel containing a gradient of a denaturing chemical (*e.g.* urea or formamide) (Muyzer and Smalla, 1998). The DNAs migration on a gel is influenced by the level of mismatch between the strands and thus where in the gradient the strands denature, which functionally halts their migration. The complex banding pattern of the fragments (Figure 1) can be compared to assess similarity among samples. Additionally, statistical procedures can be applied to do the same, commonly by first calculating (dis)similarity between samples. A major complication of this technique comes from slight gel to gel gradient differences, introducing artificial differences among samples, which can

make replication difficult. Further, different fragments can co-migrate, reducing the resolving power of the technique.

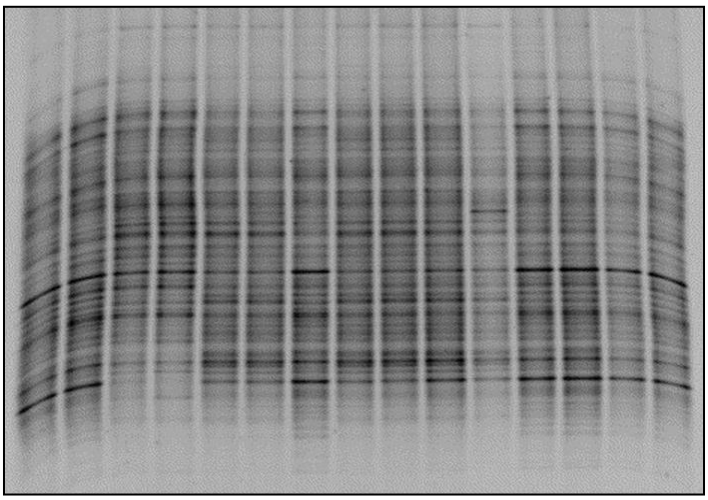

Figure 1. A DGGE acrylamide gel showing the complex banding patterns developed when using this technique. Taken from Niemi *et al.* (2001).

Brons and van Elsas (2008) investigated whether the binding site of the reverse primer utilized in the PCR step of DGGE influenced which bacterial DNAs amplified. Thirteen reverse primers were tested in triplicate and run through DGGE to assess reproducibility. Two broad clusters of band patterns were observed that differed in a single nucleotide in the primer sequence, while replicate samples were indistinguishable. Clone libraries were further developed using the four reverse primers that had the highest calculated diversity. Pairwise comparisons of the clone libraries, using the library shuffle (LIBSHUFF) statistic, revealed significant differences ($p = 0.0043$) between some of the libraries. Nakatsu *et al.* (2000) showed a similar pattern when comparing primer sets used to amplify DNA from the same soil sample. Also, in soils with high levels of bacterial diversity, a smear was produced on the gel, hindering the comparison of samples since similarity could not be calculated. Replicate samples were identical

and more similar to each other than to the other locations. Statistical evaluation consisted of the calculation and comparison of the similarity coefficients. Lerner *et al.* (2006), who simulated a forensic scenario, collected evidentiary soil from a shoe, as well as known samples from the scene, an alibi location proposed by the suspect, the suspect's home, and from similar soil types in surrounding locations. The DGGE patterns were compared using cluster analysis from calculated similarities. The authors were able to distinguish between the crime scene samples and alibi samples; however, the suspect's home and crime scene samples clustered closely.

ALH-PCR is less often seen in the forensics literature. This method first employs amplification of a bacterial DNA locus with fluorescently labeled primers. The amplicons are capillary electrophoresed and the resulting electropherograms are compared for similarities in shared peaks (Moreno *et al.*, 2006). T-RFLP analysis is similar to ALH-PCR, although it differs in that it employs endonuclease digestion of the labeled amplicon before electrophoresis (Heath and Saunders, 2006). The electropherogram output for both of these methods describes the variable length of the amplified/digested products. Next, procedures, like analysis of similarity (ANOSIM) or nonmetric multidimensional scaling (NMDS), can be applied to assess statistical differences between and among the samples. Both techniques have a limited resolving power (Liu *et al.*, 1997), and distinct bacterial species can, by chance, have the same length product and thus be indistinguishable. Given this, the statistical analysis coupled with these techniques are often limited in their ability to differentiate between habitats, though successes have been reported (*e.g.* Macdonald *et al.*, 2011; Sullivan *et al.*, 2013).

Moreno *et al.* (2006) utilized ALH-PCR in conjunction with NMDS to assess the differences in soil bacteria during the dry and wet seasons in Miami-Dade County. Four soil types could not be fully differentiated in multidimensional space for the wet season, though dry

season samples tended to associate better within each habitat depending on the combination of 16S ribosomal RNA (rRNA) gene hypervariable regions (V1 – V3) amplified. Further, when wet and dry season samples were plotted together in a pairwise fashion for each site, there was a variable amount of clustering of the samples from the same season, from very little to close association, though the seasons could be readily differentiated. Yang *et al.* (2006) used two supervised classification techniques for the identification of ALH profiles: support vector machines and k-Nearest Neighbor (KNN). Testing a combination of different primer sets, correct classification frequency ranged from ~33% up to 100% depending on the land usage. Additionally, surface and depth samples were both correctly classified with above 66% accuracy. Though the two supervised classification techniques performed equally well, the reason for misclassified samples was not identified. Supervised classification techniques are powerful ways to designate class membership; however, small sets of known samples can limit the accuracy of the classifications.

A preliminary investigation into the use of T-RFLPs in forensic case work, published by Horswell *et al.* (2002), simulated two forensic scenarios. In the first, soil was collected from the bottom of a shoe, its corresponding print, the exact location eight months later, and four other locations. The second mimicked soil being deposited onto clothing, where samples were taken from the soil used to dirty the clothing and the clothing itself. T-RFLP analysis was conducted and profiles were compared using Sørensen's similarity index. A high similarity was found between the soil from the bottom of the shoe and the footprint sample taken at the same. The eight month sample was 'moderately' similar to the previous two. The four other locations had lower values, but were also moderately similar to the soil from the shoe. The DNAs collected from the clothing were highly similar to the soil that was used to stain the clothing. Though able

to correlate samples, the analysis used by Horswell *et al.* (2002) was highly subjective with no assessment of statistical significance and must be interpreted cautiously. Heath and Saunders (2006) took T-RFLP profile analysis a step further by utilizing pairwise comparisons (*t*-test and Mann-Whitney), as well as hierarchical cluster analysis (HCA) to visualize the relationships of samples taken from three different habitats. The samples within a habitat were statistically more similar to each other than they were to the samples taken from the other habitats. These relationships were also represented by HCA, with samples from the same habitat clustering on the dendrogram. Finally, the effectiveness of T-RFLPs in discriminating among similar soil types was described by Macdonald *et al.* (2011). Enlisting analysis of variance (ANOVA) and ANOSIM, eight of the ten sites were statistically different; however, replicate samples taken from the same site were as well. The NMDS configurations demonstrated some separation of sites, though poor resolution of samples in two dimensions was evident.

Analysis of Soil Bacteria at the Michigan State University Forensic Biology Laboratory

Forensic biologists at Michigan State University have been studying various methodologies for identifying soil samples based on their microbial populations for the last eight years. The goal, through utilization of T-RFLPs, was to characterize how bacterial populations differ within the same habitat over time and space, as well as among habitats. These initial research questions needed to be addressed to fully realize how feasible the use of bacterial profiling is in a forensic context.

First, Meyers and Foran (2008) addressed spatiotemporal considerations. Soil samples were collected from five habitats: an agricultural field, a marsh edge, a yard, a deciduous forest, and a sandy woodlot approximately 100 miles away from the other sites. Sampling included a

central location once every month for a year, with auxiliary samples taken ten feet in each cardinal direction every three months. This sampling scheme aimed to address changes in bacterial composition month-to-month and over short distances. DNAs were extracted and the entirety of the *16S rRNA* gene was amplified, incorporating an end-labeled primer. Amplicons were digested with *MspI* and capillary electrophoresed. Normalized similarity indices were calculated for each electropherogram and analyzed using single factor ANOVA as well as multivariate ANOVA. The authors found that among habitats there was no significant difference when considering the entire year, with only the agricultural field showing significant differences from month-to-month. Conjointly, the extent of within-habitat temporal change was significant when compared to the other habitats. Lastly, there was no significance in the intra-habitat variability for any of the five habitats.

Lenz and Foran (2010) sought to differentiate among the same five habitats, using T-RFLPs, through a more focused approach. *Rhizobia* DNAs were amplified using *recA* gene specific primers, with amplicons subjected to *RsaI*, *MspI*, or *DpnII* digestion and capillary electrophoresis. Relationships among the samples' T-RFLP profiles were evaluated with NMDS. In two dimensional multivariate space the deciduous forest and sandy woodlot could almost always be differentiated regardless of the restriction enzyme used, while the other three habitats were heavily comingled when all five were plotted together. Accurate differentiation of sites, except for the agricultural field, was accomplished when pairwise comparisons were projected into two dimensions. The introduction of questioned samples had variable success depending on the endonuclease used and habitats being compared; however, the appropriate association was more often seen than not. These results further support the idea that bacterial communities can be used to differentiate unrelated habitats. The use of NMDS does not allow for the attribution of

statistical significance, though it can represent the underlying patterns within these data with useful information displayed in the ordination plots.

Taken together, these studies have shown that T-RFLP analysis is a valuable tool for the study of microbial populations in soil. Recently, more powerful technologies have come into use that allow for an even greater understanding of soil bacterial metagenomics. A promising technique developed in the last ten years is massively parallel sequencing.

Theory of Next-Generation Sequencing

Introduced in 2005, next-generation sequencing, also known as next-gen, massively parallel, high-throughput, or 2nd generation sequencing, is an alternative to automated Sanger sequencing (Margulies *et al.*, 2005; reviewed by Shokralla *et al.*, 2012). These platforms have the ability to generate vast amounts of data in short periods of time, and do not require the creation of clone libraries (MacLean *et al.*, 2009), which facilitates metagenomic analysis of complex substrates like soil. A great number of next-gen sequencing platforms exist, each with their own chemistries and detection systems; however, they can be broken down into two major groups: PCR-based sequencing, which include Roche 454, Illumina MiSeq, and Applied Biosystems SOLiD, and single-molecule based technologies, which include Helicos Bio-Sciences HeliScope and PacBio RS SMRT (MacLean *et al*., 2009; Metzker, 2010; Shokralla *et al.*, 2012). The Roche platform was used for this research and will be discussed in detail.

The 454 sequencing platform was the first of the aforementioned to be introduced (MacLean, 2009) and is based on pyrosequencing (Figure 2). After amplified template DNA is introduced to the sequencer, nucleotides are released one at a time in a given order (A-T-G-C in Figure 2). When a nucleotide is incorporated, pyrophosphate (PPi) is released by the polymerase,

and used by sulfurylase with adenosine 5'phosphosulfate to generate ATP. Luciferase then produces light through the reaction of luciferin and ATP. The light is captured by a camera and recorded on a pyrogram, where the height of the peak is proportional to the number of nucleotides incorporated. The pyrogram is finally converted into a nucleotide sequence. Apyrase is used to digest unincorporated nucleotides and ATP before the next nucleotide in the sequence is released. This platform is advantageous over most in that read lengths can range from 400 to 600 base pairs, with over 100,000 sequence reads possible per run (www.454.com). This amount of data is vastly greater and more precise (less anonymous) than what can be recovered using T-RFLP, ALH-PCR, or DGGE analysis, and bacterial identification on a large scale is possible. Furthermore, computer programs have been developed for easy processing of the sequences and attribution of statistics, including multidimensional and pairwise, with some being modified specifically for next-gen data (discussed below).

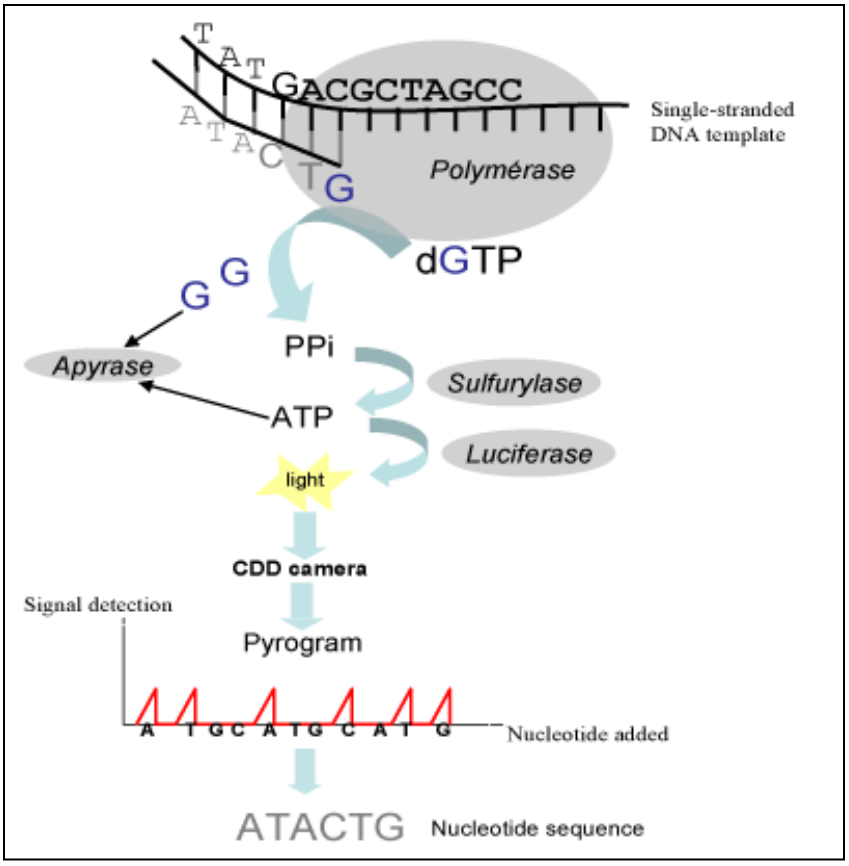

Figure 2. The addition of a guanine nucleotide into the growing DNA strand in a 454 pyrosequencing reaction releases pyrophosphate (PPi). The PPi is then used to convert adenine 5' phosphosulfate into ATP through sulfurylase and reacts with luciferin to produce light by luciferase. The light is captured by a camera and recorded on a pyrogram, where the height of the peak is proportional to the number of nucleotides incorporated. The pyrogram is finally converted into a nucleotide sequence. Apyrase is used to digest the remaining unincorporated nucleotides and ATP before the next nucleotide in the sequence is released. Taken from Armougom and Raoult (2009).

Introduction of Statistical Methods

Two β-diversity indices and five statistical procedures are presented below. The first two statistics included an unconstrained ordination (NMDS) and unsupervised classification technique (agglomerative hierarchical cluster analysis), which both aim to identify relationships among the samples without imposing constraints such as class membership. Clusters, or groupings, of samples can be identified by the graphical outputs they produce. Both NMDS and

HCA are descriptive statistics, meaning no significance can be attributed to the groupings. The second two statistics encompassed two pairwise comparisons, ∫-LIBSHUFF and the Unique Fraction Metric (UniFrac), which can detect statistical differences between sequence libraries. The final statistic was a supervised classification technique (KNN) which uses a predetermined set of samples with assigned class membership (a training set) to classify unknown samples.

Beta-Diversity Indices

Beta (β)-diversity was defined by Whittaker (1960) as "the extent of change of community composition, or degree of community differentiation, in relation to a complex-gradient of environment, or a pattern of environments". Two commonly used β-diversity indices, the Bray-Curtis dissimilarity index (BCDI) (1957), and Sørensen-Dice coefficient (SDC; described independently by Sørensen (1948) and Dice (1945)), can be used to investigate the diversity in bacterial communities among the habitats sampled in this study. The pairwise distances developed by each index were the input for NMDS, HCA, and KNN.

BCDI and SDC are popular ways for calculating measurements of dissimilarity in ecological data and can be used to reduce large or complex data sets into distance measurements. BCDI calculates the structural dissimilarity between communities, meaning not only is shared membership considered but also the number of individuals in the populations. On the other hand, SDC calculates community membership differences by only assessing shared membership of populations. This concept is illustrated in Table 1 with the two communities being compared, A and B, having three species in common. The BCDI value would reflect the large difference in individuals of species 2 because it considers the number of individuals of each population. SDC would not detect this difference, calculating a value of zero, since both communities have

members from each species. Note that comparing A to B and B to A will result in the same dissimilarity measurement, so the final matrix is square symmetric.

Table 1. Example communities for the explanation of BCDI and SDC.

Community	Number of Individuals of Species 1	Number of Individuals of Species 2	Number of Individuals of Species 3
A	5	45	19
B	5	1	18

Multidimensional Scaling

MDS is a procedure used to visualize and explore the patterns or structure of complex data sets (Borg and Groenen, 2005). The goal of MDS is to plot data in a low, multidimensional space, making it an ordination technique, where each data point represents a single sample and the spread of the data approximates the originally imputed (dis)similarities (Borg and Groenen, 2005; Cox, 2001). Data inputted into MDS take the form of a square symmetric matrix of (dis)similarities. The final configuration of data points illustrates the correlations among proximities (*i.e.* closer data points are more highly correlated). In MDS, all data points are randomly plotted in a given number of dimensions. Those points are then systematically adjusted in relation to each other to reduce the amount of stress, which is a measure of how accurately the plot is representing the data. When the global minimum stress is achieved further iterations are discontinued. Figure 3 illustrates a global minimum, where point Y has the lowest stress. Point X is similar to point Y and is considered a local minimum. Depending on the computing power used for analysis and the complexity of the proximities, a local minimum may be found instead of the global minimum, causing the stress to be higher and a less accurate MDS solution to be developed.

The analysis of additional plots (i.e., Scree or Shepard diagrams) can be used to identify the reliability of the final configuration. A stress diagram, or Scree plot, is a measure of the badness-of-fit of the MDS configuration to the given proximities (Borg and Groenen, 2005). The lower the stress the better the configuration is fitting the data. For ordinal MDS, stress (σ) decreases as dimensionality increases until the number of dimensions (m) is equal to the number of samples (n) minus two ($m = n - 2$). However, as the number of dimensions increases, the interpretability of the MDS plot becomes more difficult. An adequate number of dimensions needs to be identified so that the stress is low and the plot is understandable; generally two dimensions are used. There is no globally accepted level of stress for a MDS plot and thus, acceptance is at the discretion of the analyst, although Kruskal (1964) introduced the idea that a final configuration can be chosen where an increase in m does not greatly reduce stress. This is often referred to as the 'elbow' in the stress diagram which can be seen at two dimensions in Figure 4.

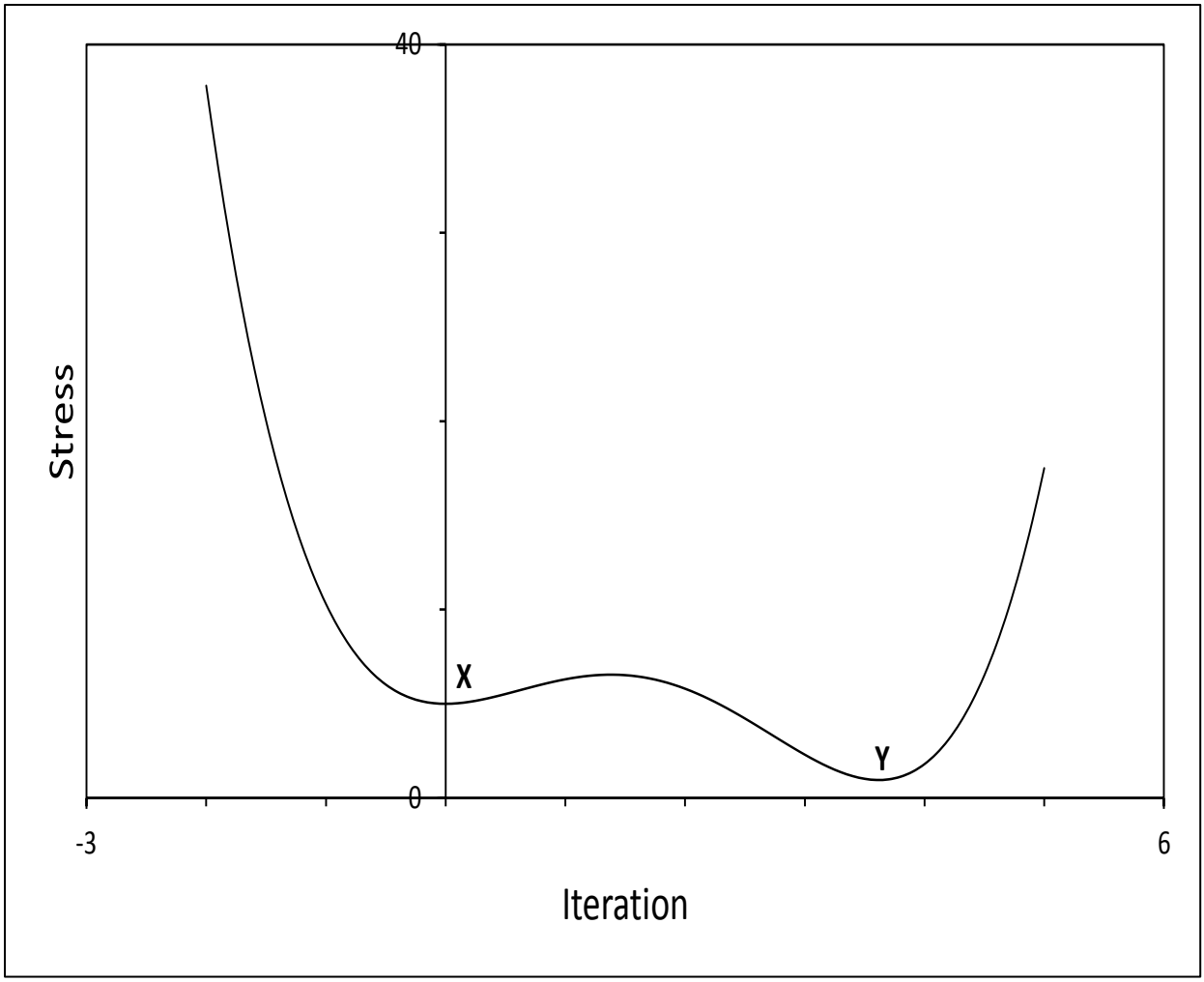

Figure 3. Graph illustrating the concept of local and global minimum with regards to stress for multidimensional scaling. Position Y indicates the global minimum having the lowest stress for the entire graph. Position X is similar to Y and is considered a local minimum. Multidimensional scaling aims to attain the global minimum when plotting proximities.

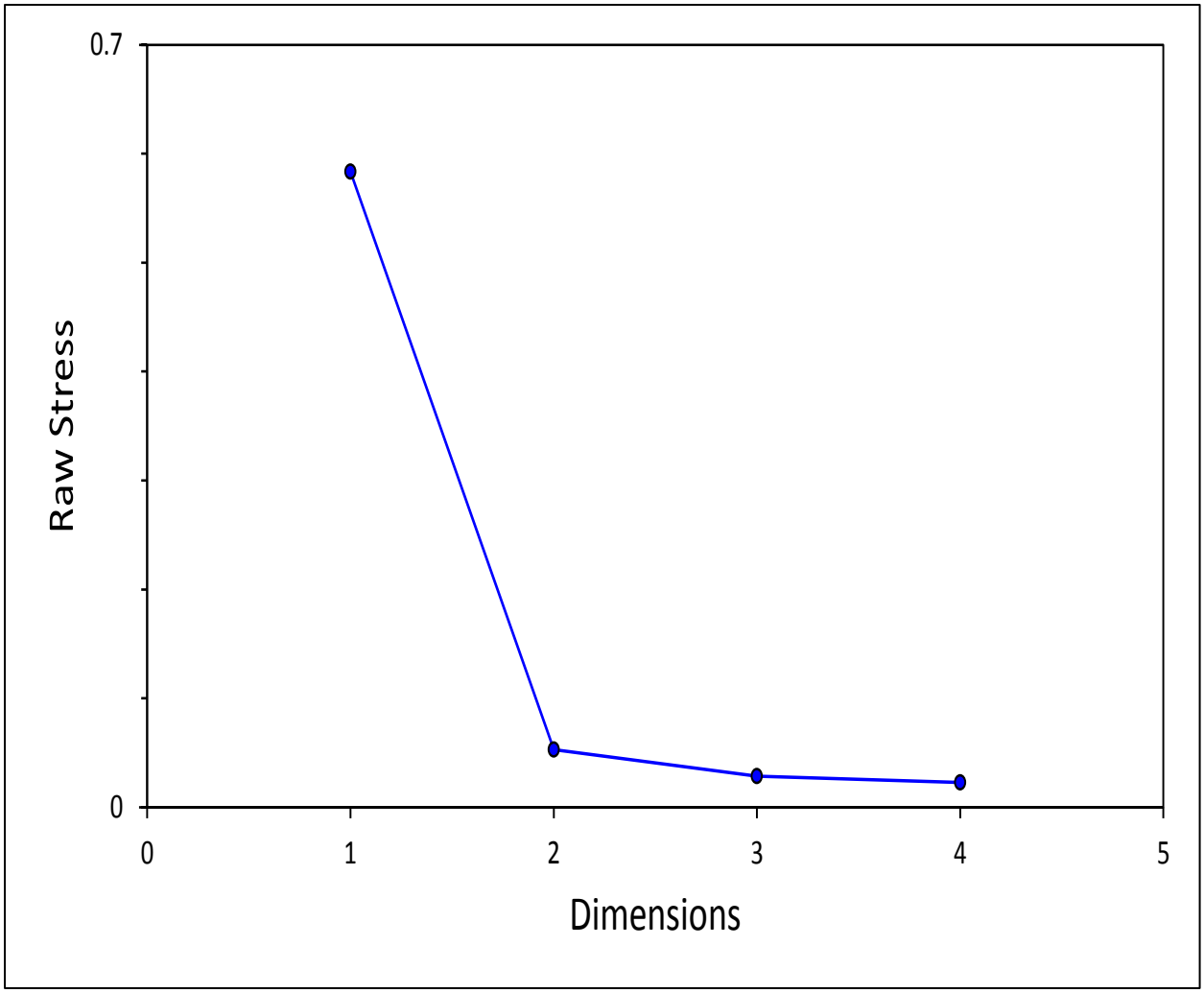

Figure 4. Typical Scree diagram for multidimensional scaling. Stress is high in one dimension followed by a substantial decrease at two dimensions. Stress continues to decrease into higher dimensionality, though not appreciably, creating the elbow, after which little additional information is gained.

Another test for the acceptability of stress at a given dimension was described by Spence (1979), who proposed that random stress, or stress produced by random data for a given number of samples and dimensions, could be approximated using

$$\sigma_1 = 0.001[a_0 + a_1 m + a_2 n + a_3 \ln(m) + a_4 \sqrt{\ln(n)}]$$

where, a_0 = -524.25, a_1 = 33.8, a_2 = -2.54, a_3 = -307.26, and a_4 = 588.35. This estimation of random stress is accurate for the range n = 10 – 60 and m = 1 – 15. Random stress can be used as a threshold for the acceptance of the MDS plot with relation to its associated stress diagram.

Similar to stress diagrams, Shepard diagrams (Figure 5) are an indicator of the badness-of-fit for the final configuration of the data. They plot disparities (open circles) as a monotonic function over the range of inputted proximities on the X-axis. The monotonic regression in a Shepard diagram varies depending on the type of MDS being used. Additionally, approximated distances (filled circles) are also plotted onto this function (Borg and Groenen, 2005). A plot with a perfect stress of zero would have disparities and distances sitting atop each other. In cases where stress is nonzero, the vertical distance between each disparity and distance is the error of representation for that pair. The comparison of these points allows for the identification of outliers and possible sources of high stress. The larger the deviation of distances from disparities, the worse MDS is at explaining the original proximities, and the larger the stress.

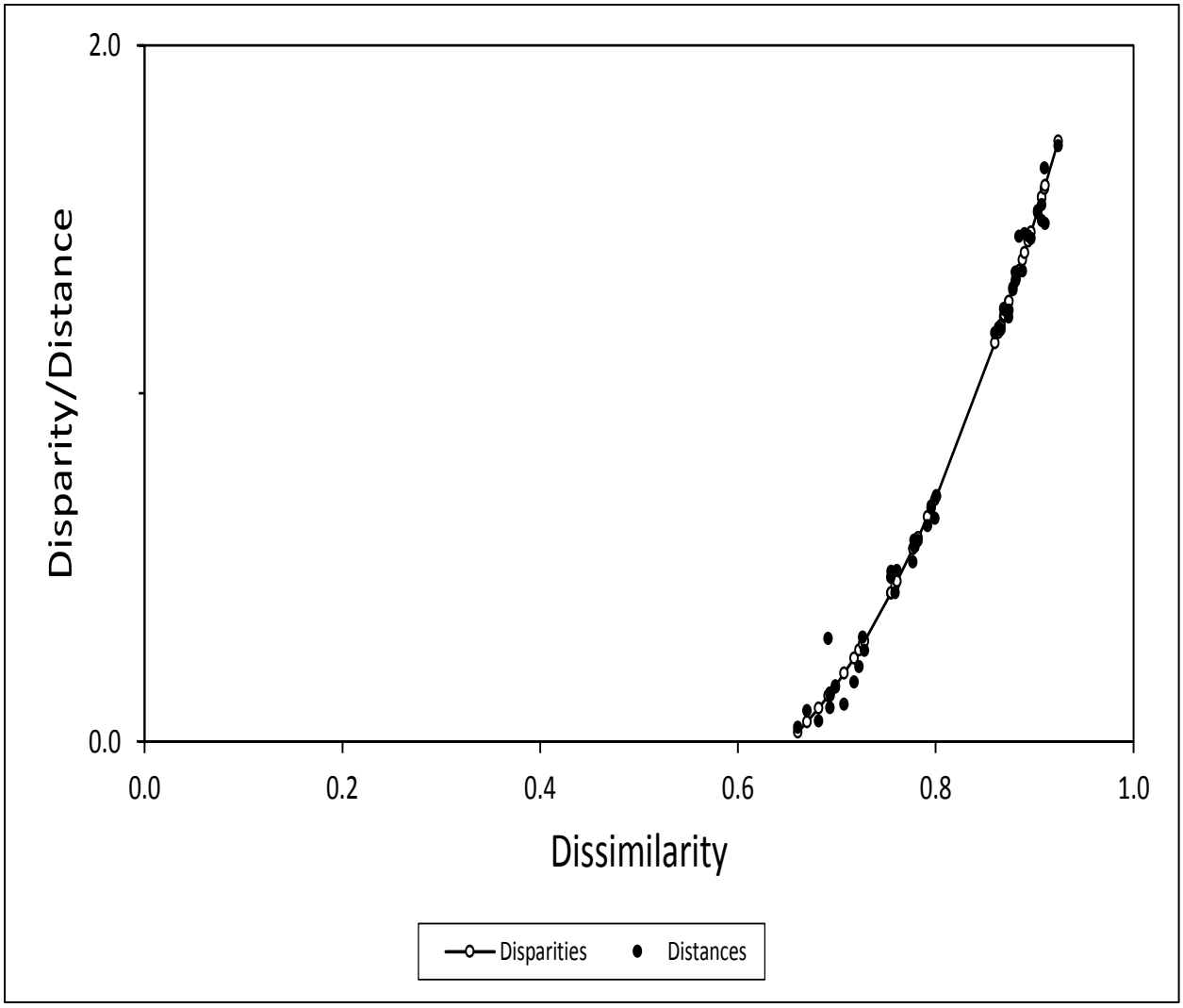

Figure 5. A Shepard diagram with low stress and good association of disparities and distances. The closer the association of the filled circles (representing approximated distances) and open circles (representing disparities) the better multidimensional scaling is representing the data. This Shepard diagram shows very close relationship between the two using a polynomial monotonic function.

Agglomerative Hierarchical Cluster Analysis

Similar to MDS, agglomerative hierarchical cluster analysis is an unsupervised cluster technique that allows for the visualization of distances among samples in a two dimensional dendrogram (Figure 6) (Beebe *et al.*, 1998). A square symmetric matrix of (dis)similarities is first developed using a distance calculation. The largest or smallest value between two samples,

depending on the input, is identified as being the most similar and combined into a cluster (Gemperline, 2006). Agglomerative HCA begins with each sample 'existing as its own cluster' with subsequent clusters being formed from the combination of samples/clusters based on nearness or similarity (Dougherty, 2013). HCA updates the matrix by calculating the distances from the new cluster to all other points, replacing all data related to the original points (Gamperline, 2006). This process is repeated until all samples are clustered together with samples closer in multivariate space represented with shorter lines on the dendrogram. This is exemplified in Figure 6 where samples 212 and 214 are closer to each other in multivariate space than they are to the rest of the samples. The distance at which clusters are formed is based on what method is used to calculate nearness in multidimensional space. Three linkage methods, single, complete, and unweighted pair-group method using arithmetic averages (UPGMA), differ in how distance is calculated between established clusters and unclustered samples (Beebe *et al.*, 1998).

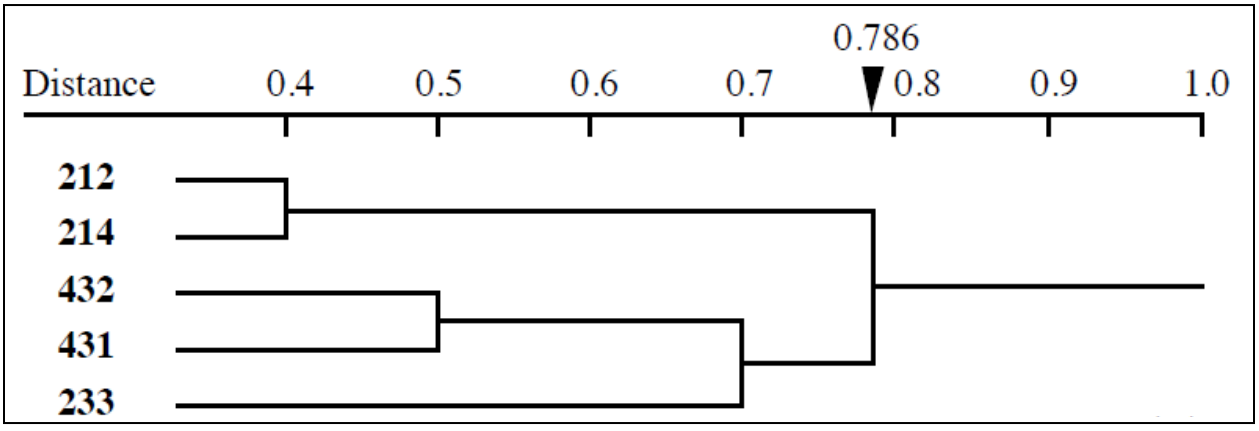

Figure 6. A representative dendrogram of five samples. The axis along the top represents the distance the samples in each cluster are from each other in multivariate space. The final distance to cluster all samples is represented by the black wedge at a distance of 0.786. Figure taken from Legendre and Legendre (2012).

Single linkage or single-link clustering merges clusters based on the distance between nearest neighbors of all possible pairs of members between the two groups (Beebe *et al.*, 1998; Dougherty, 2013). This method can induce an undesirable feature termed chaining (Figure 7), where clusters are merged because individual points in each cluster are close in multivariate space though the rest of the samples in the cluster are not (Dougherty, 2013). Chaining can be avoided if complete linkage is used (Legendre and Legendre, 2012), which merges clusters based on the largest distance between all samples of the groups, joining the two with the smallest of these distances. Since the largest distance is calculated between members of two clusters, all members of one cluster are already linked to all members of the other before merging them, *i.e.* all other distances between members of the two clusters are smaller than the one used to link them (Dougherty, 2013; Legendre and Legendre, 2012). Finally, UPGMA merges clusters by calculating the average distances between all members of the two clusters and combining groups with the smallest of those averages (Legendre and Legendre, 2012). While *a priori* estimation of which linkage method will perform better is not possible, the goal of HCA is to understand the underlying patterns in the data and enlisting all three methods can elucidate these relationships better than one alone (Beebe *et al.*, 1998).

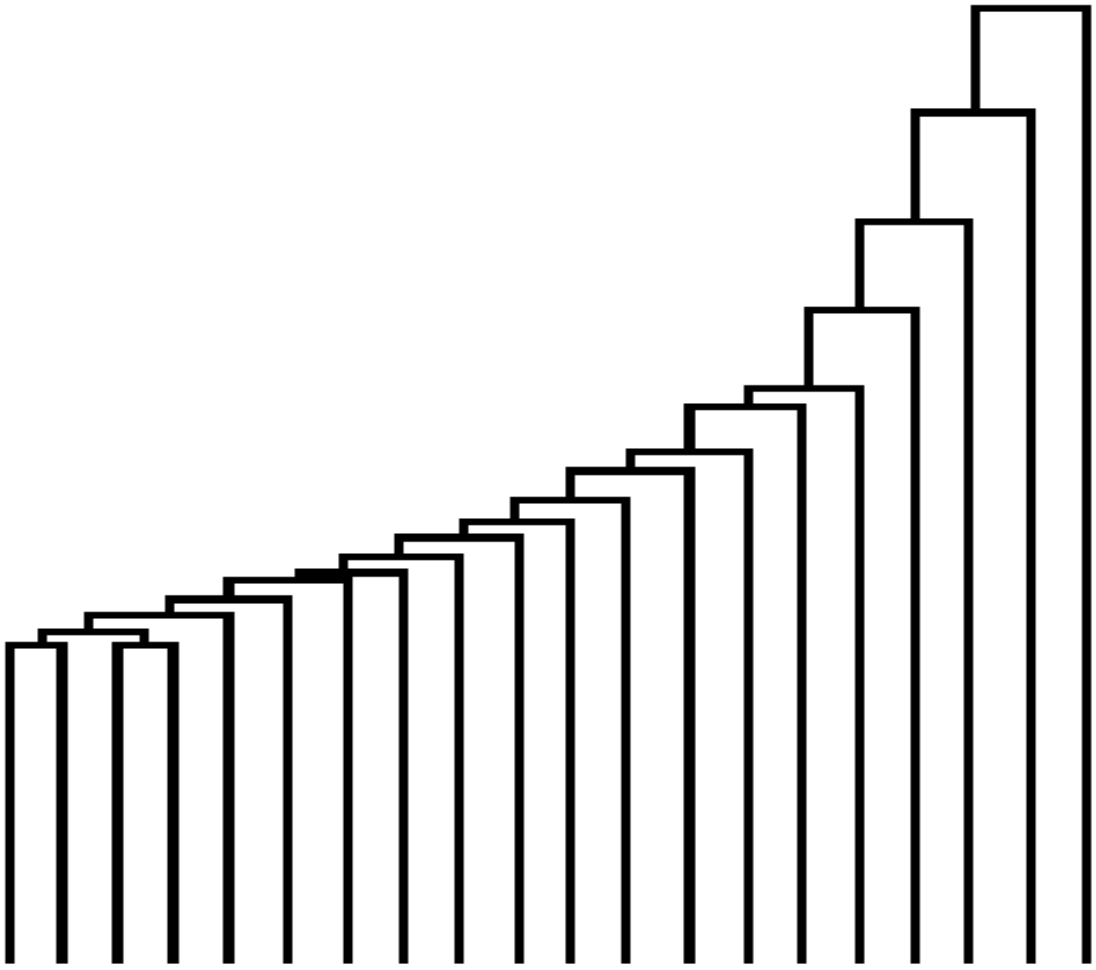

Figure 7. This dendrogram illustrates the idea of chaining. Clusters are formed by the sample next in the chain being linked to the previous one because they are close in proximity, though the rest of the cluster may not be. This is undesirable in ecological data since no real understanding of the relationships can be determined. Taken from Legendre and Legendre (2012).

Integral Library Shuffle

The LIBSHUFF statistic was introduced by Singleton *et al.* (2001) for the statistical comparison of *16S rRNA* clone libraries using the approximation form of the Cramér-von Mises test statistic for curve fitting and Monte Carlo simulations to calculate significance. An updated version of LIBSHUFF, ∫-LIBSHUFF, was published by Schloss *et al.* (2004) which uses the exact and integral form of the Cramér-von Mises test statistic.

This statistic measures the quantity of singleton sequences found in a library compared to another, or library coverage. Library coverage is defined as the percentage of the library that is composed of non-singletons, where a singleton is a sequence whose genetic distance to every other sequence in the dataset is as large or larger than the defined distance. This is accomplished through the calculation of the difference in library coverage using the following equation

$$\Delta C_{XY} = \int_0^{\infty} [C_X(D) - C_{XY}(D)]^2 dD$$

where $C_X(D)$ and $C_{XY}(D)$ are the library coverage at different genetic distances (D). Coverage is calculated using the method described by Good (1953), where

$$C_X(D) = 1 - \left[\frac{N_X(D)}{n_X}\right]$$

such that $N_X(D)$ is the number of singleton sequences in library X for various genetic distances (D) and n_X is the total number of sequences in library X. $C_{XY}(D)$, or the percentage of sequences in X that have a correspondingly similar sequence in library Y, is calculated with

$$C_{XY}(D) = 1 - \left[\frac{N_{XY}(D)}{n_X}\right]$$

where $N_{XY}(D)$ is the number of sequences which are distance D or greater in library X from library Y. The difference in coverage, $C_x(D) - C_{XY}(D)$, indicates the percentage of sequences in library X that are not singletons and are also not found in library Y. The squared difference in coverage is then integrated for all values of D, from 0 – 0.5. Homologous and heterologous coverage curves can be plotted to visualize how coverage is changing as D increases (Figure 8).

For the calculation of significance, the libraries being compared, X and Y, are combined and randomly split into two new libraries of equal size to the originals. ΔC_{XY} is calculated for the randomized libraries over multiple iterations and a random distribution is constructed. The

proportion of ΔC_{XY} values in the random distribution greater than the original ΔC_{XY} is the p-value. The reverse comparison, ΔC_{YX}, is then calculated in the same way. If either is found to be statistically significant, the libraries are composed of significantly different bacterial communities.

In the case where multiple pairwise comparisons are made, *e.g.* a data set of many samples is being analyzed, a Bonferroni correction can be applied to correct for the large number of tests conducted. This correction aims to preserve a family-wise error rate, *i.e.* user-defined p-value, while reducing the probability of a false positive (statistical significance) (Kaltenbach, 2012). The corrected p-value can be approximated with the following equation

$$p \approx \frac{\alpha}{k}$$

where α is the family-wise error rate and *k* is the number of comparisons being made. Logically, as the number of tests increases, p can become incredibly small to the point where no meaningful results can be determined; therefore, the Bonferroni correction should only be used for a small number of tests (Kaltenbach, 2012). The corrected p-value is then used to assess statistical significance for the pairwise comparisons made.

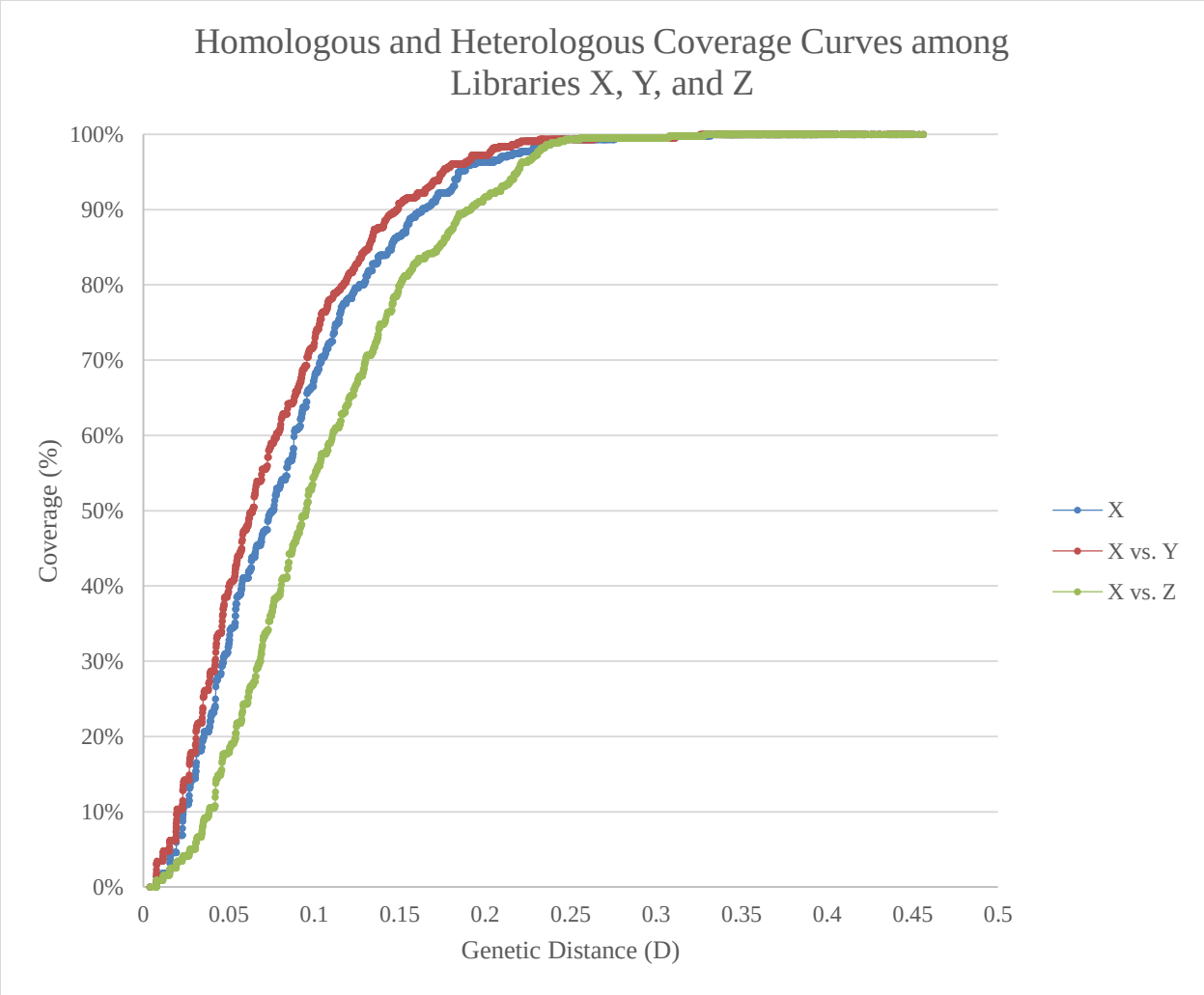

Figure 8. Homologous coverage curve ($\Delta C_X(D)$) and heterologous coverage curves ($\Delta C_{XY}(D)$ and ($\Delta C_{XZ}(D)$) for the hypothetical comparison of libraries X, Y, and Z developed by ∫-LIBSHUFF. Each line represents the change in coverage over genetic distance from 0 – 0.5 on the X-axis. When comparing libraries X and Y no significant difference was found between the bacterial communities (p-value > 0.025). This is understandable as the coverage curves fit well with each other; however, a significant difference (p-value < 0.025) was found between the X and Z libraries which is evident in the distance, or non-fitting, between the curves.

Unique Fraction Metric

UniFrac is a technique for pairwise comparisons introduced by Lozupone and Knight (2005) using phylogenetic distance. A tree is developed for the environments being compared in which UniFrac measures the branch length for descendants, sequences at the end of branches,

from one or the other, but not both, and thus measures the unique evolutionary adaptation that occurred in each environment individually. Samples taken from similar environments with similar evolutionary pressures would share more populations, *i.e.* have high occurrence of branch lengths leading to shared descendants. This is represented in the bolded branches in Figure 9A. The two communities being compared, where squares and circles represent sequences originating from one or the other, share a high amount of branch length with many of the nodes leading to descendants from both communities. This indicates similar evolutionary adaptation occurred in those environments. However, if two environments are so different that members of one community would not survive in the other, there would be little branch length shared between them. Figure 9B demonstrates this point with the square and circle communities sharing zero branch length. UniFrac also uses a Monte Carlo procedure to calculate the statistical significance between samples by randomizing the sequences at the ends of the branches while keeping the tree constant. The p-value is defined as the percent of random trees that share the same or greater fraction of unique branch length with the original tree. This concept is illustrated in Figure 9C, where r = 1, 2, and 3 are randomized trees and the histogram represents the normal distribution of unique branch lengths from the Monte Carlo simulations. The arrow in the histogram indicates a predetermined cutoff value for assessing significance, and the star the calculated p-value. In this case the calculated p-value is less than the cutoff, so the square and circle communities are significantly different. Again like ∫-LIBSHUFF, a Bonferroni corrected p-value should be used when multiple comparisons are conducted to preserve the family-wise error rate.

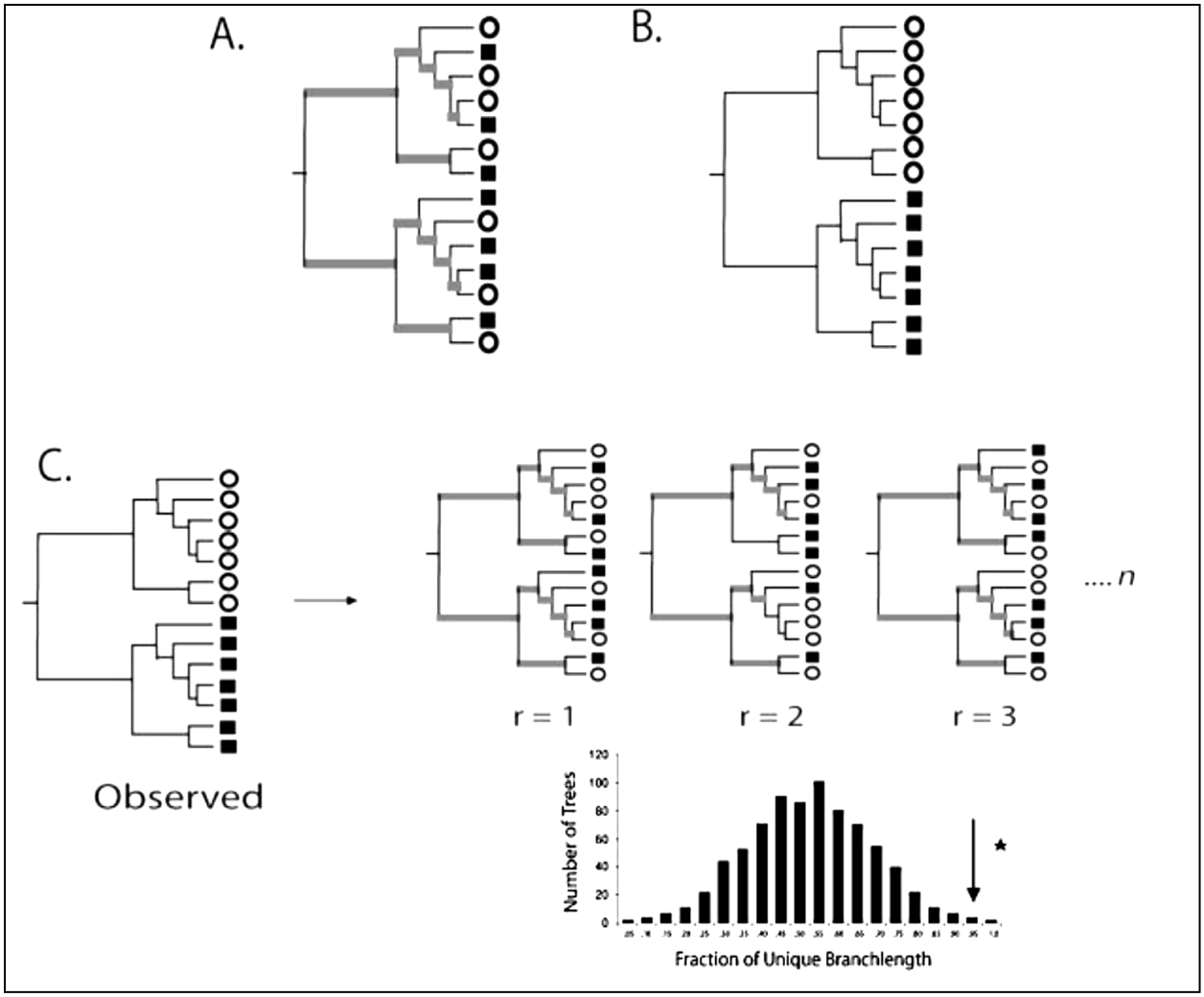

Figure 9. (A) and (B) illustrate the concept of heavily shared or completely unique branch length respectively. The intermingling of circles and squares in (A) represents similar phylogenies in both environments, where the opposite is true in (B). (C) demonstrates the Monte Carlo iterations for assessment of significance between the circle and square communities with randomly produced trees labeled r = 1, 2, and 3. The histogram is a composite of all the random trees with the star indicating the p-value, or proportion of random trees that had an equal or greater fraction of unique branch length as the original tree. The desired threshold value, the arrow, is then compared to the calculated p-value, the star. In the case of this example, the p-value is less than the threshold, so the two communities are significantly different. Taken from Lozupone and Knight (2005)**.**

k-Nearest Neighbor

Class membership of unknown samples can be determined using KNN based on their nearness in multivariate space to known samples in the training set (Beebe *et al.*, 1998). For the

purposes of training set development known samples are assigned a class, where the pairwise distance of all training set members is calculated, that is used to predict membership of unknown samples. To validate the model, *a priori* assignment of class membership is checked against assignment by KNN (Pirouette user guide, version 4.0). This is accomplished by individually classifying each training set sample based on the remaining training set members. Validation continues over a range of nearest neighbors (*k*), ultimately outputting the number of correctly/incorrectly classified individuals in the training set over the values of *k*. Classification is based on each nearest neighbor contributing a single vote for its class, with the majority vote determining the final class membership of the sample. For the range of *k* values, good association of *a priori* class assignment and predicted validation assignment are indicators that sample designations are realistic and supported by the training set (Beebe *et al.*, 1998; Pirouette user guide, version 4.0). Having found good association, the optimal *k* to use for unknown sample classification is identified as the largest odd value of *k* for which minimal/no misclassifications were made. Choosing an odd numbered *k* eliminates the need to enlist tie-breaking procedures for samples that might fall between two different classes. The highest value of *k* that can be used is equal to the class with the smallest number of members. The higher the *k*, the more confidence can be placed in the classification of unknown samples (Beebe *et al.*, 1998).

When classifying samples, KNN forces a classification regardless of how far away the samples are in multivariate space; however, how good those classifications are can be assessed with a calculated 'goodness value' (G) or 'class fit' (Beebe *et al.*, 1998; Pirouette user guide). G is calculated using the following equation

$$G = \frac{d_{unk} - \bar{d}_X}{sd(d_X)}$$

where d_{unk} is the distance of the unknown sample to the nearest neighbor of the class KNN assigned it, while $\bar{d}_X$ and $sd(d_X)$ are the mean distance and standard deviation, respectively, of the single neighbor distances of the class X the unknown sample was classified into. These values assess the spread of the points in the class or cluster diameter. The smaller the value of G the more confidence can be placed in the classification, with the inverse also holding true (Beebe *et al.*, 1998). Additionally, a threshold for G can be established with a definitive cutoff for whether the sample is a member of that class or not identifying poor classifications (Pirouette user guide).

When an unknown sample is classified, the distance of that sample to all samples in the training set is calculated. The distances are then ranked smallest to largest identifying the nearest neighbors. Based on the chosen *k*, *n* number of the smallest distances are used with each neighbor contributing one vote to the classification of the unknown sample. The majority vote from the nearest neighbors decides the class membership of the sample (Gemperline, 2006).

Feasibility of Next-Gen Sequencing and Statistical Analysis in Forensic Soil Analysis

The need for statistical methods that can identify, differentiate, and/or associate soils is evident. The aim of the research presented here was to identify variables that might influence the accuracy of identifying where a soil sample originated. Several factors were considered: habitat (ten different and similar habitats were sampled), time (changes over a year with sites sampled every three months), within habitat heterogeneity, and depth. 454 sequencing was utilized because the amount of data developed with this technique is substantially greater than older techniques and previous research in the ecological community has shown successes in describing the microbial communities of soil (Hollister *et al.*, 2010; Lauber *et al.*, 2009; Roesch *et al.*,

2007). Further, five statistical procedures were applied to the sequencing data for the investigation of sample association in each study.

MATERIALS AND METHODS

Soil Sampling Schemes

Biological Replicate Study Collection

Soil samples were collected from three locations in the Fenner Nature Center in Lansing, MI in 2012 (Table 2). The sampling locations represent three distinct habitats: a yard, a marsh edge, and a deciduous woods. GPS coordinates were taken for each site. Three soil samples were collected from the surface of each site less than a meter apart and used as replicates. Samples were taken using a gardening spade, which was rinsed with water between collections, and the collected soil was stored at -20°C within an hour of collection. A mixture (1:1:1) of the yard samples (Ymix) was also processed. Photographs of the sampling locations are in Appendix 1.

Table 2. Sites of biological replicate sampling and corresponding GPS coordinates at the Fenner Nature Center in Lansing, MI.

Site Name	Abbreviation	GPS Coordinates
Deciduous Woods	W	42° 42′ 33″ N 84° 31′ 00″ W
Marsh Edge	M	42° 42′ 32″ N 84° 30′ 53″ W
Yard	Y	42° 42′ 39″ N 84° 30′ 54″ W

Habitat Distance Study Collection

Soil samples were collected at three habitats, the Fenner Nature Center yard and deciduous woods, as well as a yard treated with herbicides, insecticides, and fertilizer on the Michigan State University campus in 2013 (Table 3). At each site a main sample was taken with four additional samples collected at 5, 10, 50, and 100 feet in each cardinal direction resulting in 17 samples per location. Soils were taken as three spot samples mixed into a single sample of soil.

Table 3. Sites of habitat distance sampling and corresponding GPS coordinates of the main sample.

Site Name (Location)	Abbreviation	GPS Coordinates
Fenner Nature Center, Deciduous Woods (Lansing, MI)	WM	42° 42′ 33″ N 84° 31′ 00″ W
Michigan State University, Treated Yard (East Lansing, MI)	TYM	42° 43′ 27″ N 84° 28′ 03″ W
Fenner Nature Center, Yard	YM	42° 42′ 39″ N 84° 30′ 54″ W

Depth Study Collection

Depth samples were taken at a different treated yard, from the aforementioned one, on the Michigan State University campus in 2013 (Table 4) with a soil corer and mud auger (AMS, Inc. American Falls, ID) that were rinsed with water between samplings. A surface sample was taken with additional samples collected at 1, 2, 5, 10, 20, and 36 inches below it.

Table 4. Site of depth sampling and corresponding GPS coordinates.

Site Name (Location)	Abbreviation	GPS Coordinates
Michigan State University, Treated Yard (East Lansing, MI)	TY	42° 43′ 45″ N 84° 28′ 26″ W

Time Study Collection

Soils were collected as single spot samples once every three months (August, November, February, and May) for a year (2012 – 2013) from the marsh edge, deciduous woods, and yard sites in Table 2. The biological replicates acted as the August samples for this study.

Similar Habitat Study Collection

Soil samples were collected from ten yards in the Greater Lansing area in 2012. The yards were treated (chemicals had been applied) or untreated. The central (main) location was on

the Michigan State University campus (MSUM), with nine additional sites at various distances from there. Figure 10 is a map of the sampling locations. These were collected in a manner similar to the time study. The location names, abbreviations, and GPS coordinates are shown in Table 5.

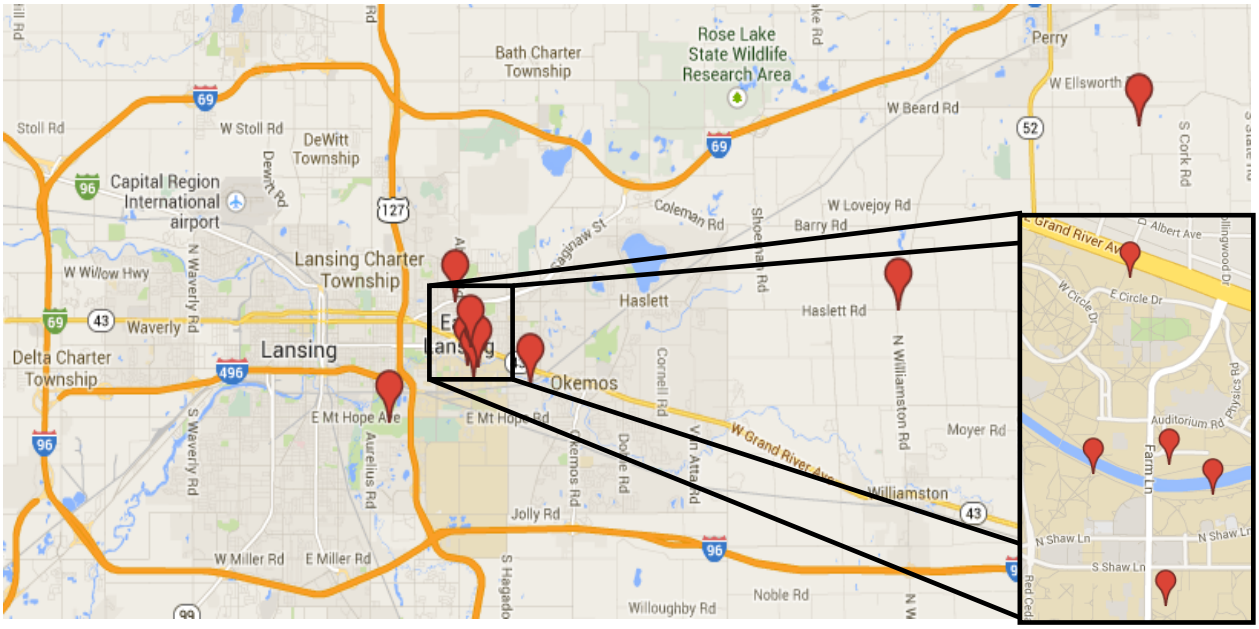

Figure 10. Map of sampling locations for similar habitat study. The cluster in the middle are the five samples taken from the Michigan State University campus and are magnified on the right.

Table 5. Sites of similar habitat sampling and corresponding GPS coordinates.

Site Name (Location)	Abbreviation	GPS Coordinates
Fenner Nature Center, Yard (Lansing, MI)	F	42° 42′ 39″ N 84° 30′ 54″ W
Foran Lawn (Williamston, MI)	Fo	42° 44′ 41″ N 84° 17′ 12″ W
Lisa Lawn (East Lansing, MI)	Lisa	42° 44′ 50″ N 84° 29′ 08″ W
Michelle Lawn (East Lansing, MI)	M	42° 43′ 19″ N 84° 27′ 08″ W
Michigan State University, East Site (East Lansing, MI)	MSUE	42° 43′ 38″ N 84° 28′ 32″ W
Main Site	MSUM	42° 43′ 41″ N 84° 28′ 38″ W
North Site	MSUN	42° 43′ 60″ N 84° 28′ 43″ W
South Site	MSUS	42° 43′ 27″ N 84° 28′ 38″ W
West Site	MSUW	42° 43′ 40″ N 84° 28′ 48″ W
Perry Lawn (Perry, MI)	P	42° 48′ 00″ N 84° 10′ 44″ W

Diverse Habitat Study Collection

Soils were collected from ten different habitats in the Greater Lansing area in 2013. Samples were taken in a similar manner as those collected in the replicate collection over two days; however, the three samples taken per site were mixed together before freezing to make one composite sample. If soils could not be frozen within an hour of collection they were temporarily kept on ice. The location names, abbreviations, and GPS coordinates are shown in Table 6. Figure 11 is a map of the sampling locations. Photographs of locations can be found in Appendix 1.

Table 6. Sites of diverse habitat sampling and corresponding GPS coordinates.

Site Name (Location)	Abbreviation	GPS Coordinates
Fenner Nature Center, Field (Lansing, MI)	F	42° 42′ 39″ N 84° 31′ 16″ W
Fenner Nature Center, Marsh Edge (Lansing, MI)	M	42° 42′ 32″ N 84° 30′ 53″ W
Fenner Nature Center, Deciduous Woods (Lansing, MI)	W	42° 42′ 33″ N 84° 31′ 00″ W
Fenner Nature Center, Yard (Lansing, MI)	Y	42° 42′ 39″ N 84° 30′ 54″ W
Lake Lansing, Beach (Haslett, MI)	LL	42° 45′ 14″ N 84° 24′ 16″ W
Michigan State University, Corn Agricultural Field (East Lansing, MI)	CAF	42° 42′ 33″ N 84° 28′ 18″ W
N. Canal Rd, Road Side (Lansing, MI)	RS	42° 45′ 04″ N 84° 39′ 43″ W
Fallow Agricultural Field (Perry, MI)	FAF	42° 48′ 04″ N 84° 11′ 10″ W
S. Cork Rd, Dirt Road (Perry, MI)	DR	42° 48′ 17″ N 84° 09′ 34″ W
Woldumar Nature Center, Coniferous Forest (Lansing, MI)	CF	42° 41′ 12″ N 84° 38′ 05″ W

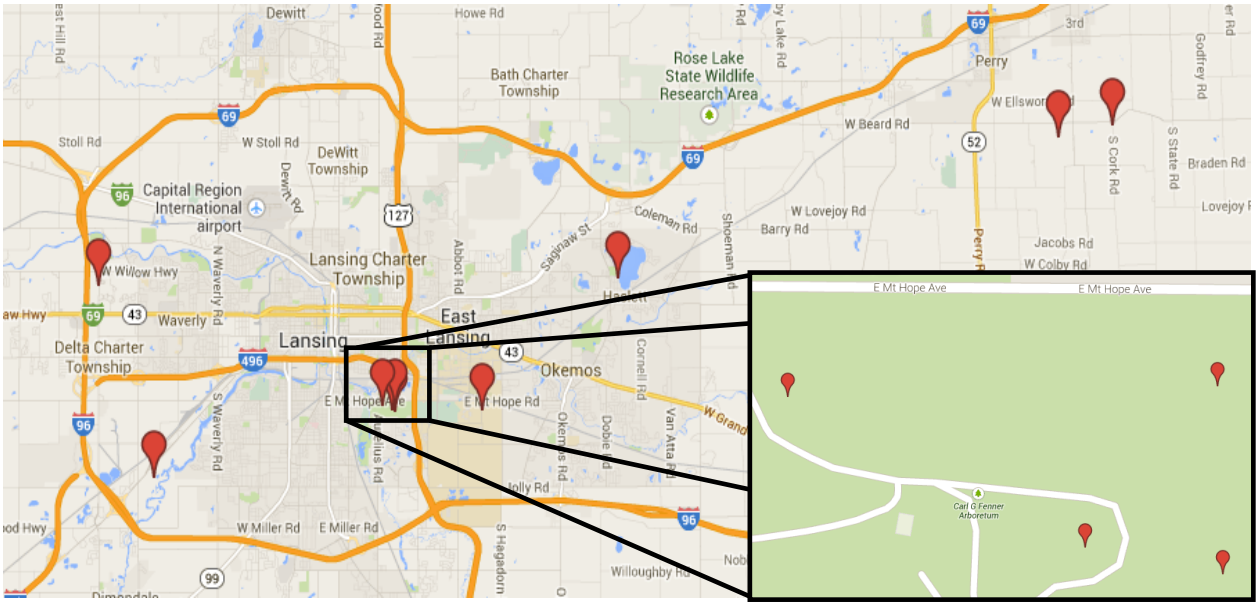

Figure 11. Map of sampling locations for diverse habitat studies. The cluster of spots on the left are the four sampling locations at the Fenner nature center and are magnified on the right.

DNA Techniques

DNA Extraction

Micropipette tips and tubes were UV irradiated in a Spectrolinker XL-1500 UV Crosslinker (Spectronic Corporation Lincoln, NE) for rounds of 5 min (~ 2.5 J/cm^2). DNA was extracted from soil samples using a PowerSoil® DNA Isolation Kit (MoBio Carlsbad, CA) with two minor modifications of the manufacturer's protocol: an additional wash was conducted after step 16 by adding 500 µL of 70% ethanol and centrifuged for 30 s at 10,000 x g, and DNA was eluted using 100 µL of solution C6 that had been heated to 55°C. Reagent blanks were processed with every extraction.

PCR Amplification of 16S rRNA Hypervariable Regions 4 – 6

Reagents suitable for UV irradiation and all micropipette tips and tubes were UV irradiated for rounds of 5 min. Hypervariable regions 4 – 6 of bacterial *16S rRNA* genes were amplified with conserved bacterial primers 518F and 1064R (Table 7) (Filkins *et al.*, 2012). A PCR master mix was generated, with final concentrations of 1X AmpliTaq Gold buffer (Life Technologies Carlsbad, CA), 2.5 mM $MgCl_2$, 0.2 mM nucleotide triphosphates, 1.2 µL of the 10 µM forward primer, 0.4 µg/µL bovine serum albumin, and 3U AmpliTaq Gold (Life Technologies). The master mix was aliquoted into ten PCR tubes to a final volume of 53.4 µL per tube. Six microliters of soil DNA extract was added to an aliquot, along with 1.2 µL of one of the 10 µM reverse primers (each reverse primer, while identical in binding region, had a unique DNA barcode that was used for downstream analysis). The 60 µL reaction was mixed, then equally aliquoted into three PCR tubes to help avoid stochastic sampling of template DNA. DNAs were amplified on an Applied Biosystems® 2720 thermal cycler (Life Technologies) under the conditions noted in Table 8, for 30 cycles. Identical 20 µL reactions were combined

into a 60 µL pool and 5 µL was electrophoresed on a 1% agarose gel followed by ethidium bromide visualization.

Table 7. 518F and 1064R primer sequences, adaptor sequences, and barcodes. Degenerate nucleotides bind the following ways: N with purines and pyrimidines, R with purines, and Y with pyrimidines.

Forward or Reverse	Adaptor Sequence	Barcode	Binding Sequence
Forward	CCTATCCCCTGTGTGCCTTGGCAGTCTCAG		CCAGCAGCYGCGGTAAN
Reverse A1	CCATCTCATCCCTGCGTGTCTCCGACTCAG	AATGGTAC	CGACRRCCATGCANCACCT
Reverse A2	CCATCTCATCCCTGCGTGTCTCCGACTCAG	TCTCCGTC	CGACRRCCATGCANCACCT
Reverse A3	CCATCTCATCCCTGCGTGTCTCCGACTCAG	AACCTGGC	CGACRRCCATGCANCACCT
Reverse A4	CCATCTCATCCCTGCGTGTCTCCGACTCAG	ACGAAGTC	CGACRRCCATGCANCACCT
Reverse A5	CCATCTCATCCCTGCGTGTCTCCGACTCAG	TTCGTGGC	CGACRRCCATGCANCACCT
Reverse A6	CCATCTCATCCCTGCGTGTCTCCGACTCAG	AACACAAC	CGACRRCCATGCANCACCT
Reverse A7	CCATCTCATCCCTGCGTGTCTCCGACTCAG	TTCTTGAC	CGACRRCCATGCANCACCT
Reverse A8	CCATCTCATCCCTGCGTGTCTCCGACTCAG	TCCAAGTC	CGACRRCCATGCANCACCT
Reverse A9	CCATCTCATCCCTGCGTGTCTCCGACTCAG	TTCGCGAC	CGACRRCCATGCANCACCT
Reverse A10	CCATCTCATCCCTGCGTGTCTCCGACTCAG	CCGGTCGC	CGACRRCCATGCANCACCT

Table 8. PCR cycling parameters.

PCR Step	Temperature (°C)	Time (s)
Initial Heating	94	600
Denaturation	94	30
Annealing	60	45
Extension	72	60
Final Extension	72	120

PCR Product Purification

Forty microliters of the remaining pooled amplification reactions were purified using Agencourt® AMPure® XP (Beckman Coulter Brea, CA). The bottle containing the beads was

vortexed briefly and 30 µL was aliquoted into a 1.5 mL micro-centrifuge tube. DNA was added and the mixture was vortexed and incubated at room temperature for 15 min. The beads were bound to a MagnaRack™ (Life Technologies) for a minimum of five min. The supernatant was aspirated from the beads and discarded. Undisturbed beads were washed with 500 µL 70% ethanol for 30 s. The supernatant was again aspirated and the beads were washed an additional time. Beads were then dried on the magnet for 30 min at 37°C. DNA was eluted by adding 100 µL of 10 mM Tris, pH 8 and vortexing the tubes for at least 10 s. The tubes were returned to the magnet and beads were bound for at least five min. Supernatant was aspirated away from the pellet and saved in a 1.5 mL micro-centrifuge tube.

PCR Quantification and Equimolar Pooling

Purified PCR product, from the biological replicate sampling, was quantified using a Quant-iT™ dsDNA High-Sensitivity Assay Kit (Life Technologies) following the manufacturer's protocol. The ten quantified samples were pooled so that 25 ng of DNA from each was in the final pool and brought to a final concentration of 1 ng/µL.

All additional amplified samples were quantified with a Qubit® fluorometer (Life Technologies) using a dsDNA High-Sensitivity Assay Kit (Life Technologies) following the manufacturer's protocol. Samples were pooled but not diluted to 1 ng/µL due to the already low final concentration of the pool.

Sequencing Purified PCR Product

The pooled DNAs were sequenced on a Roche GS Junior 454 sequencer following the manufacturer's protocols using a titanium emPCR kit (Lib-L), sequencing kit, and PicoTiter

plate kit (Roche, San Francisco, CA) by Kylie Farrell in the Department of Large Animal Clinical Sciences at Michigan State University.

Gene Sequence Data Pretreatment

Sequencing data output for the 454 sequencer was processed using open-source mothur software (Schloss *et al.*, 2009; www.mothur.org). The program input codes and explanations for processing sequence data, along with a sample file, are given in Appendix 2. Bacterial sequences were also classified using the SILVA bacterial reference alignment provided on the mothur website with input codes given in Appendix 3. Sequence libraries were subsampled to the group with the smallest number of sequences for that experiment. Additionally, the replicate samples taken from the marsh edge, yard, and deciduous woods were merged into single habitat sequence libraries and then processed. The compiled yard and Ymix samples were further compared to assess whether mixing soil before extraction or merging sequence files (compiled yard) influenced sample association.

Statistical Procedures

Gene Sequence Data Analysis

Dissimilarity values for NMDS, HCA, and KNN were calculated using BCDI and SDC (Appendix 4). The equation below is the one used in the mothur software to calculate BCDI

$$D_{Bray\text{-}Curtis} = 1 - 2\frac{\sum \min(S_{A,i}, S_{B,i})}{\sum S_{A,i} + \sum S_{B,i}}$$

Dissimilarity is calculated by subtracting from one the sum of the minimum number of DNA sequences seen in a single operational taxonomic unit (OTU) between two samples (S_A and S_B) divided by the sum of the total number of sequences for each sample multiplied by two. This

process is conducted for each pair of samples until all pairwise comparisons have been made. The final output for BCDI is a square, symmetric matrix.

SDC was calculated using the following equation in the mothur software

$$D_{Sørensen\text{-}Dice} = 1 - \frac{2S_{AB}}{S_A + S_B}$$

This equation outputs dissimilarity data by subtracting double the number of shared OTUs between the samples (S_{AB}) divided by the sum of the total number of OTUs in each sample from one. The final result from SDC analysis is a square symmetric matrix.

Matrices developed from BCDI and SDC were input into the Addinsoft© XLSTAT Pro (New York, NY) expansion for Microsoft Excel for analysis by NMDS and HCA. NMDS was run for four dimensions with 500 iterations, each stopping at a convergence of 0.00001 using the Scaling by Majorizing a Complicated Function algorithm. Random starting configurations were used with five repetitions. In this research, Kruskal's Stress-1 was calculated for all plots using,

$$\text{Stress-1} = \sigma_1 = \sqrt{\frac{\sum[f(p_{ij}) - d_{ij}(\mathbf{X})]^2}{\sum d_{ij}^2(\mathbf{X})}}$$

where, *f* is a representation function that establishes the MDS model, p_{ij} is the proximity for point *i,j*, and $d_{ij}(\mathbf{X})$ is the corresponding distance in the MDS solution **X** (Borg and Groenen, 2005). The stress is expected to decrease as the number of dimensions increases. Any deviation from this expectation indicated errors in the way MDS plots the proximities, exemplified in Figure 12 where the raw stress is extremely low for all dimensions; however, the increase in stress from one to three dimensions calls the final configuration into question. Logically, this should not happen and the MDS plot can be disregarded as being erroneous. Two dimensional MDS plots were analyzed along with Shepard and Scree diagrams, the latter of which were compared against the random stress plot described by Spence (1979). If the Kruskal's stress was less than

random stress, the NMDS plot was accepted at that dimension. Finally, standard error bars were applied to NMDS configurations of the habitat distance and depth samples. HCA was run using the classical AHC algorithm and three linkage methods: single, complete, and UPGMA, clustering rows using either Bray-Curtis or Dice proximities. If chaining was present in any dendrogram it was not analyzed, as per Legendre and Legendre (2012).

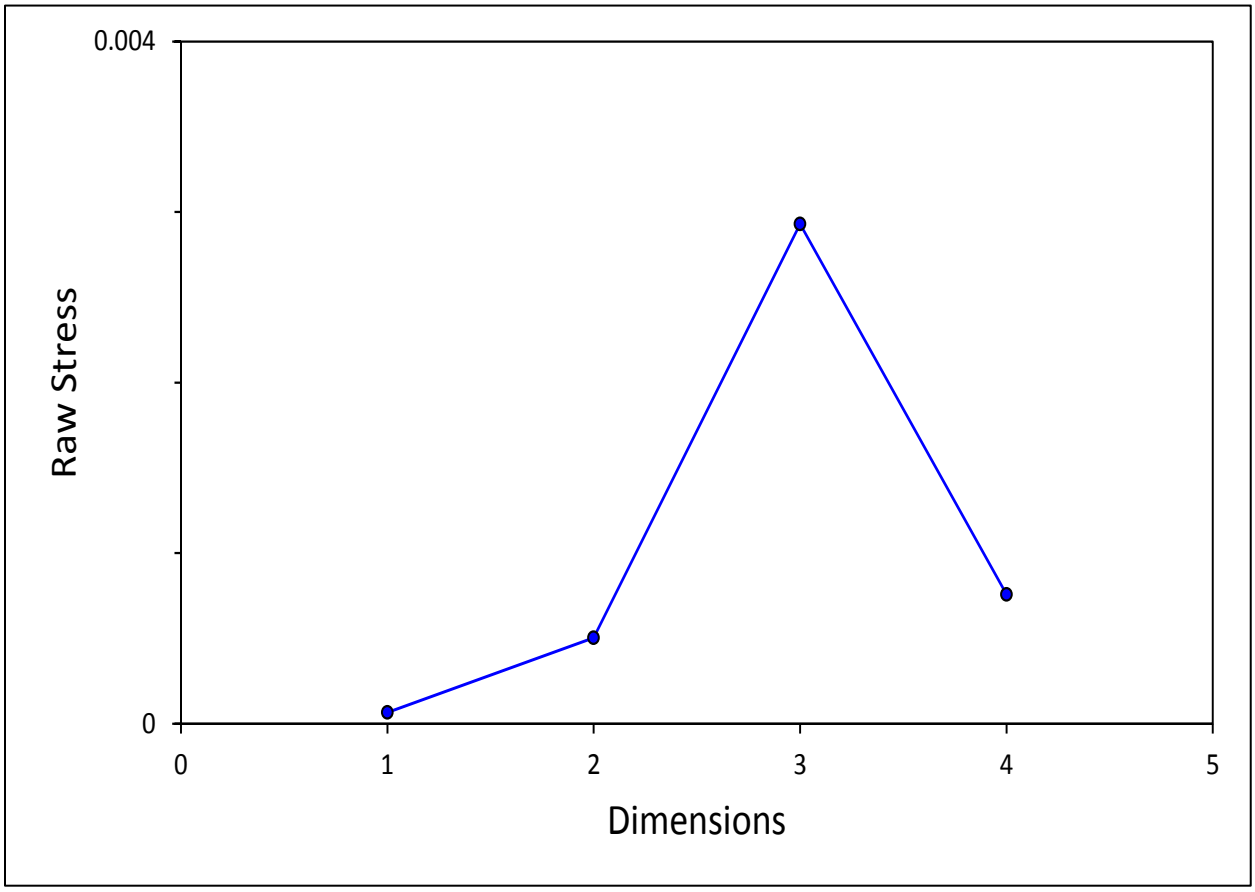

Figure 12. Erroneous Scree diagram. The plot exhibits non-normal stress changes with increasing dimensionality. In one dimension, the stress is lowest, with an increase to three dimensions followed by a decrease into four. This does not follow the expected relationship and would indicate that the multidimensional scaling configurations associated with the plot could be misrepresenting inputted data.

The ∫-LIBSHUFF and UniFrac pairwise comparisons were completed in mothur. The input codes and file names are given in Appendix 5. A Bonferroni correction was applied to all pairwise comparisons due to the large number of comparisons being made, starting with a family-wise error rate of $p = 0.05$. Results from both procedures were graphed in Microsoft Excel.

KNN was run in Pirouette 4.0 (Infometrix, Inc.©, Bothell, WA) for both diversity indices. Habitat distance samples were used as the training set (n=50) with assigned class membership. Any sample misclassified during validation was excluded. The unknown soils were the compiled biological replicates, time series samples from the marsh edge, yard (including Ymix), and deciduous woods, along with the treated yard depth samples. The accuracy of the classifications was calculated as percent correct classification.

RESULTS

Amplification, Sequencing, and Processing of *16S rRNA* Hypervariable Regions 4 – 6

All samples amplified the first time attempted except the deciduous woods west 100' sample. This sample was not amplified again and no sequences were obtained. Most of the remaining samples were successfully sequenced and recovered over 100,000 sequences per pool and more than 10,000 raw sequences from each extract. Ten samples had recovery lower than 6,000. All February, May, and November time points for the marsh edge, yard, and deciduous woods had around 5,000 sequences. The number of raw sequences for that run was just over 50,000, so those lower numbers can be attributed to machine variability. The treated yard south 5' sample had the least sequences, 3181, which resulted from poor amplification.

The processing of libraries in mothur culminated in the removal of 82 – 89% of total sequences, averaging around 87%. There were five steps during processing that removed 'erroneous' or repetitive sequences. The trim sequences command removed between 4 – 23%, averaging 9%, while the unique sequence command merged 3 – 17%, averaging about 11%. The first screen eliminated 18 – 71% based on a length of 250 bp, the average was 34%, while the second deleted 6 – 47% based on 300 bp, averaging 12%. Finally, the precluster step merged between 52 – 80% of the sequences, averaging around 70%.

Relative Abundance of Bacterial Classes

Ninety bacterial classes were identified among all soil samples, ranging from 27 in the dirt road sample to 78 in the Lake Lansing beach sample (Table 9). The same samples had the lowest and highest number of classes when considering only the least abundant 5%, 10 and 49 respectively. The depth samples were the only to show a pattern, where more classes were found as depth increased.

Biological replicate samples shared most bacterial classes, though the abundances of each were visually variable (Figure 13). The least abundant 5% of bacterial classes (Figure 14) seemed to show more diversity than the remaining 95% among replicate samples. The dirt road (Figure 15) from the diverse habitat study was the only obviously divergent sample, which differed from all other habitat types. Variability among bacterial classes was seen in relative abundance charts for all samples in this work (Appendix 6 Figures 71 – 76).

Table 9. Total and least abundant 5% of bacterial classes identified in all soil samples. Location names corresponding to sample abbreviations can be found in Tables 2 – 6.

Sample	Total Bacterial Classes	Least Abundant 5% of Bacterial Classes	Sample	Total Bacterial Classes	Least Abundant 5% of Bacterial Classes
M1	31	13	TYN50’	45	17
M2	33	15	TYS50’	45	20
M3	38	20	TYE50’	51	23
W1	32	15	TYW50’	46	19
W2	28	11	TYN100’	47	19
W3	30	13	TYS100’	47	18
Y1	32	14	TYE100’	43	15
Y2	32	14	TYW100’	46	18
Y3	37	19	WM	59	33
Ymix	35	17	WN5’	56	31
CAF	66	37	WS5’	51	25
CF	52	26	WE5’	51	27
DR	27	10	WW5’	57	31
FAF	60	31	WN10’	54	29
F	66	37	WS10’	60	34
LL	78	49	WE10’	54	28
M	53	24	WW10’	51	25
RS	51	22	WN50’	59	33
W	58	30	WS50’	60	34
Y	65	37	WE50’	57	31
M-Feb	48	23	WW50’	65	39
M-May	52	28	WN100’	45	20
M-Nov	49	24	WS100’	50	24
W-Feb	35	13	WE100’	48	22
W-May	53	29	YM	61	32
W-Nov	44	19	YN5’	64	35
Y-Feb	38	17	YS5’	60	31
Y-May	37	14	YE5’	59	30
Y-Nov	52	28	YW5’	61	32
F	61	34	YN10’	59	30
Fo	57	30	YS10’	63	34
Lisa	55	28	YE10’	61	32
M	59	32	YW10’	63	34
MSUE	51	24	YN50’	61	32
MSUM	50	23	YS50’	66	37
MSUN	55	28	YE50’	63	34
MSUS	61	34	YW50’	63	36
MSUW	54	27	YN100’	66	37
P	56	29	YS100’	66	37
TYM	44	16	YE100’	62	33
TYN5’	46	17	YW100’	60	31
TYS5’	43	16	TY Surface	47	20
TYE5’	46	19	TY1”	53	26
TYW5’	53	24	TY2”	52	25
TYN10’	46	17	TY5”	54	27
TYS10’	44	17	TY10”	54	27
TYE10’	47	19	TY20”	54	27
TYW10’	55	26	TY36”	59	32

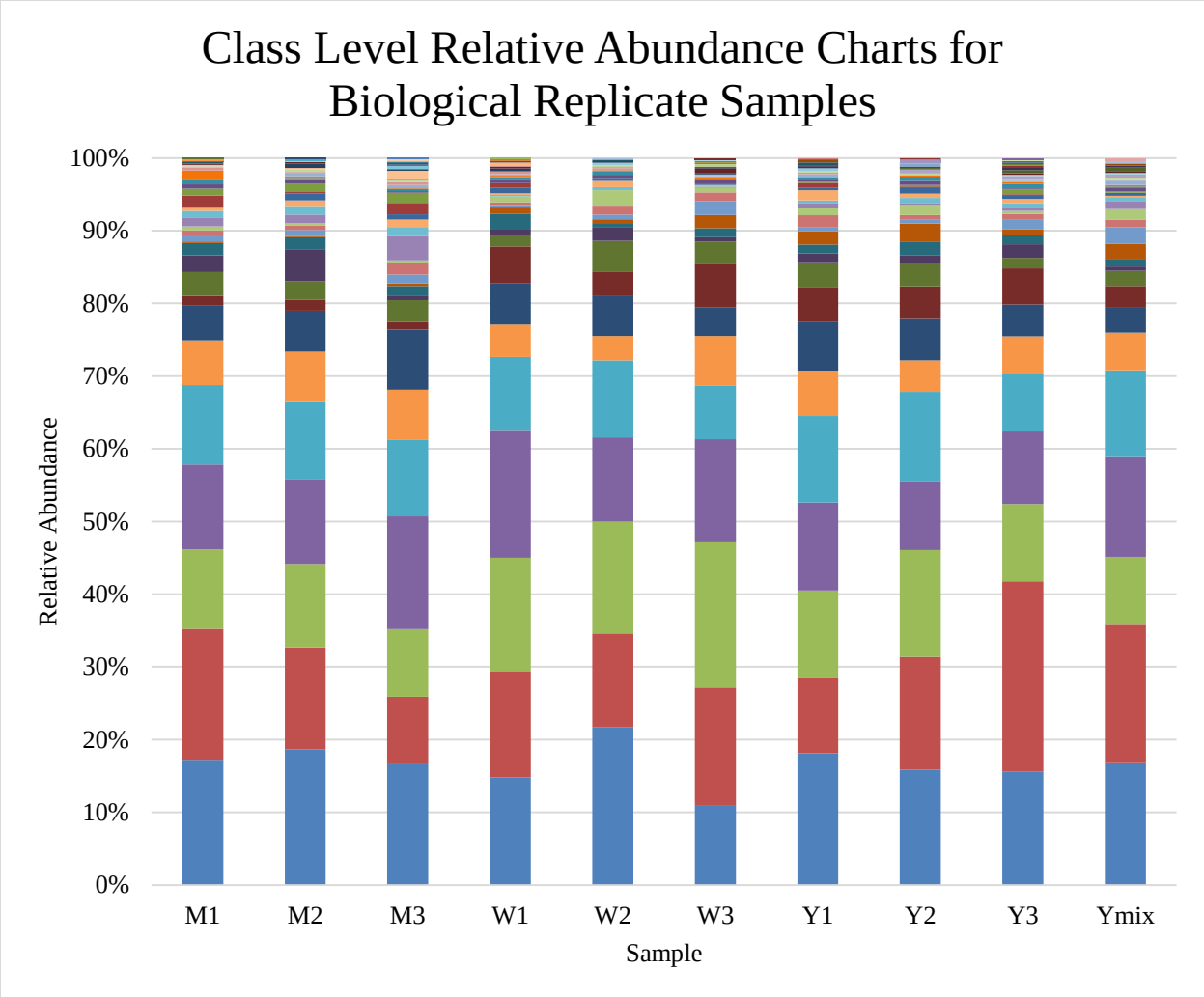

Figure 13. Class level relative abundance charts for biological replicate soil samples representing 63 bacterial classes. Samples share bacterial classes up to 95% total relative abundance though variability is evident among them. See Table 2 for site names corresponding to abbreviations and Appendix 6 Figure 77 for legend of bacterial classes.

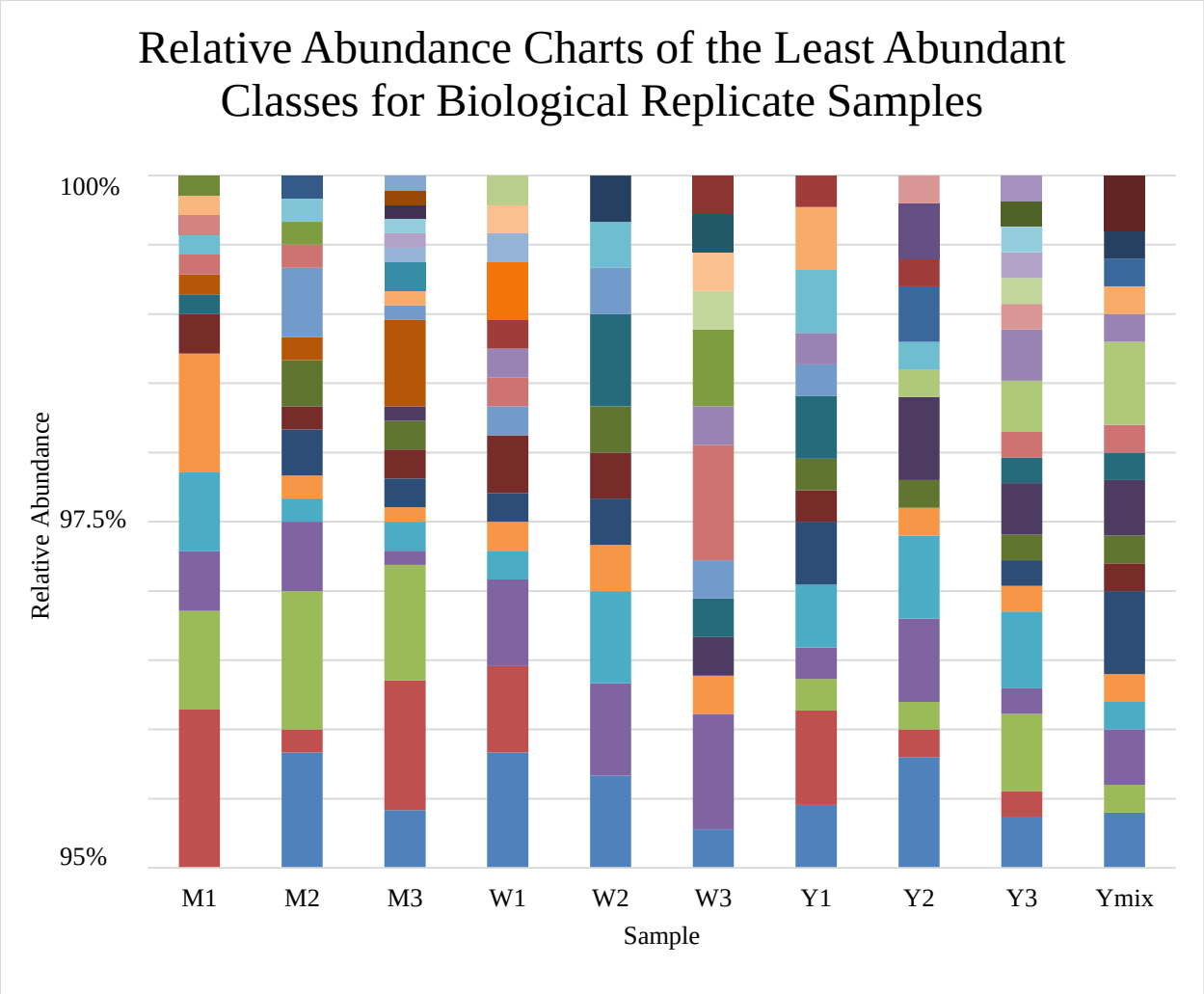

Figure 14. Relative abundance charts for the least abundant 5% of bacterial classes for biological replicate soil samples representing 45 classes. There is a great amount of bacterial diversity for each sample. See Table 2 for site names corresponding to abbreviations and Appendix 6 Figure 78 for legend of bacterial classes.

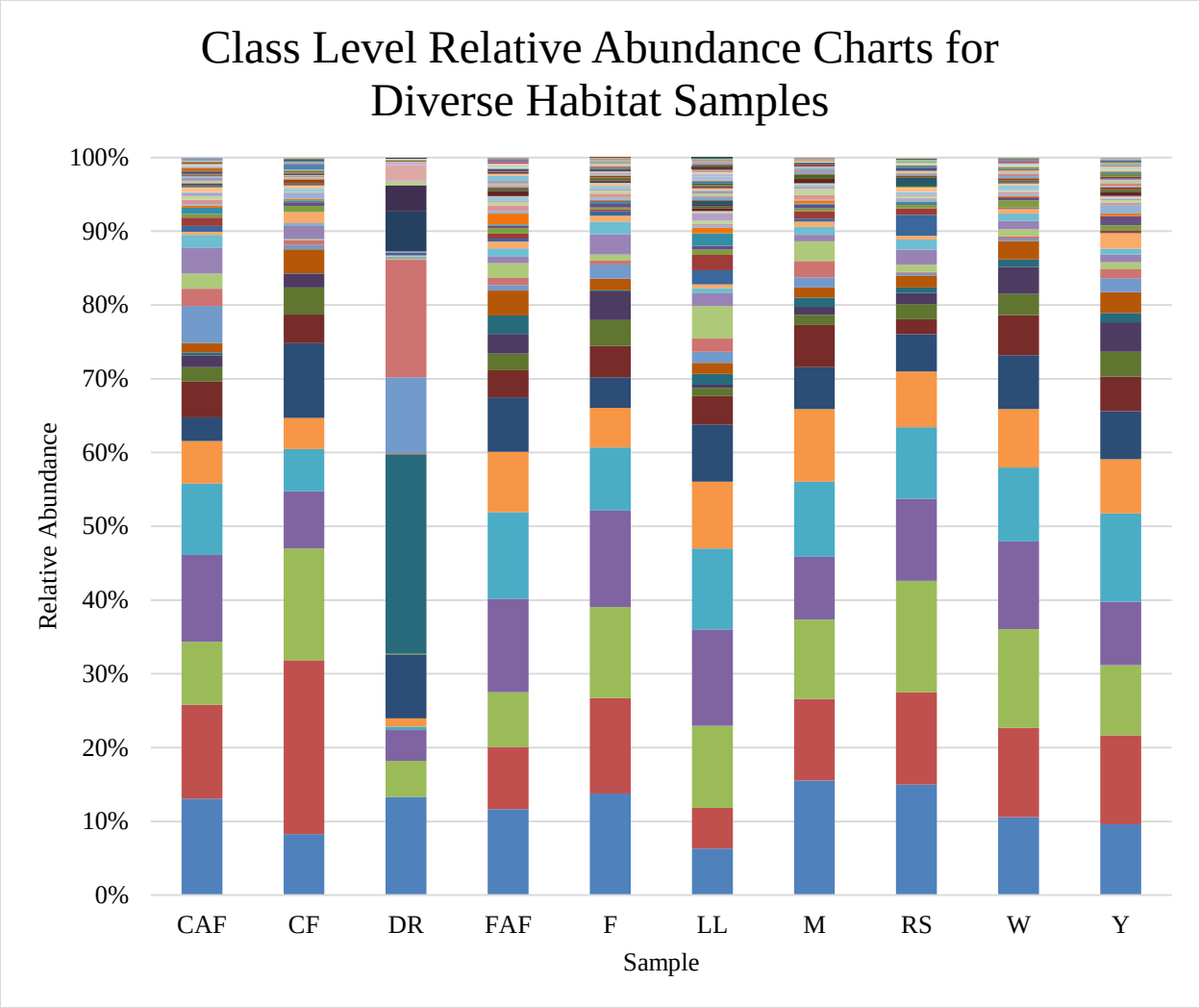

Figure 15. Class level relative abundance charts for diverse habitat soil samples representing 90 bacterial classes. Samples share bacterial classes up to 95% total relative abundance except the dirt road, which shows a much different pattern of abundances compared to the others. See Table 6 for site names corresponding to abbreviations and Appendix 6 Figure 77 for legend of bacterial classes.

Biological Replicate Samples Analysis

Nonmetric Multidimensional Scaling

Scree diagrams developed for the ten biological replicate samples from BCDI and SDC (Figure 16 and Appendix 7 Figure 79) had a high stress in one dimension with a decrease

(elbow) at two dimensions and little further reduction into higher dimensionality. Both plots failed random stress at one dimension; however, they fell below at two or more dimensions.

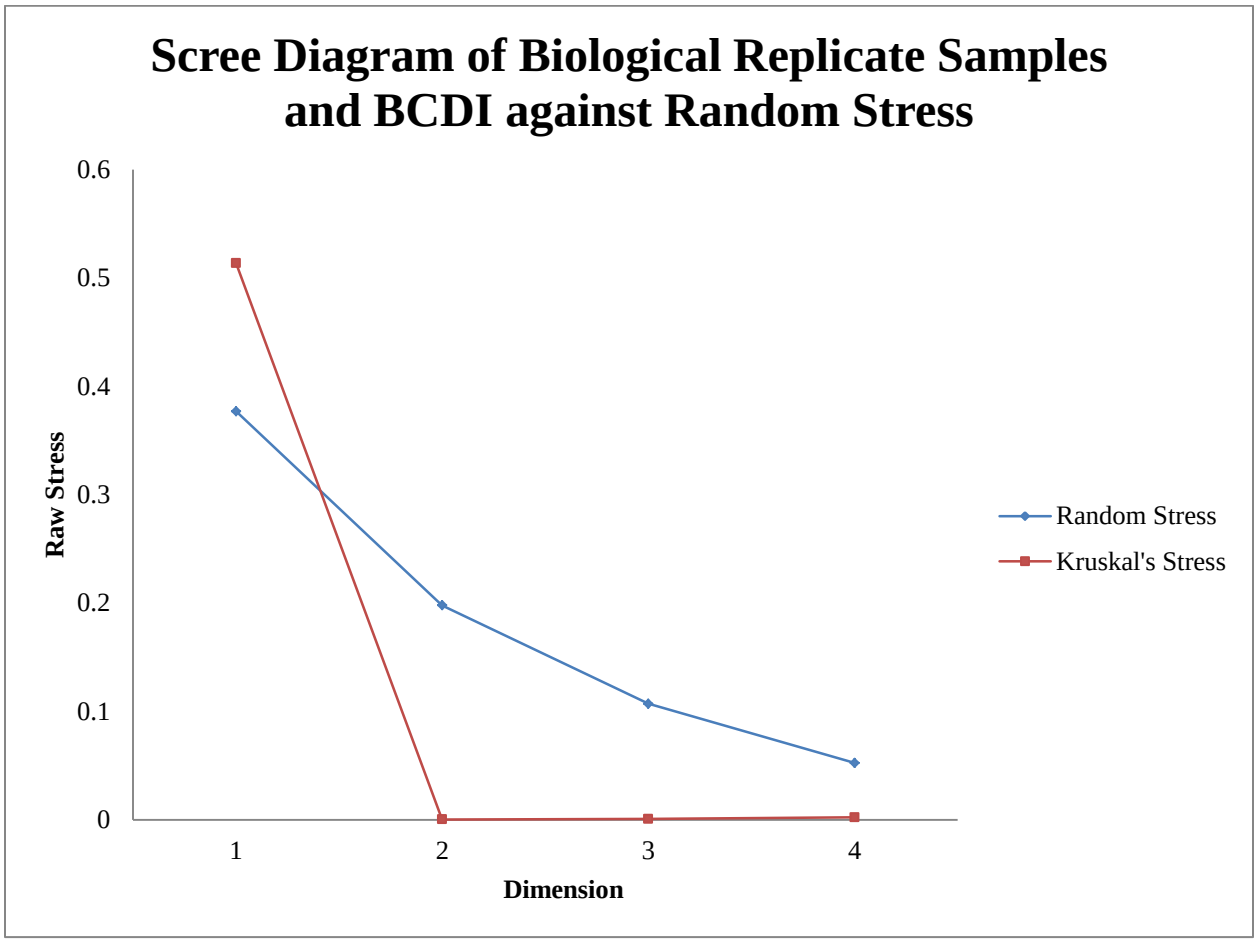

Figure 16. Scree diagram developed by NMDS for the final configurations of biological replicate soil samples over four dimensions from BCDI. Random stress was used as a threshold for the acceptance of final configurations. The one dimension configuration was the only to exceed the threshold and was rejected. The elbow in the curve is at two dimensions.

Shepard diagrams developed for final configurations in two dimensions (Figure 17 and Appendix 7 Figure 80) had good association of distances and disparities. This corresponded well with the low stress at that dimension.

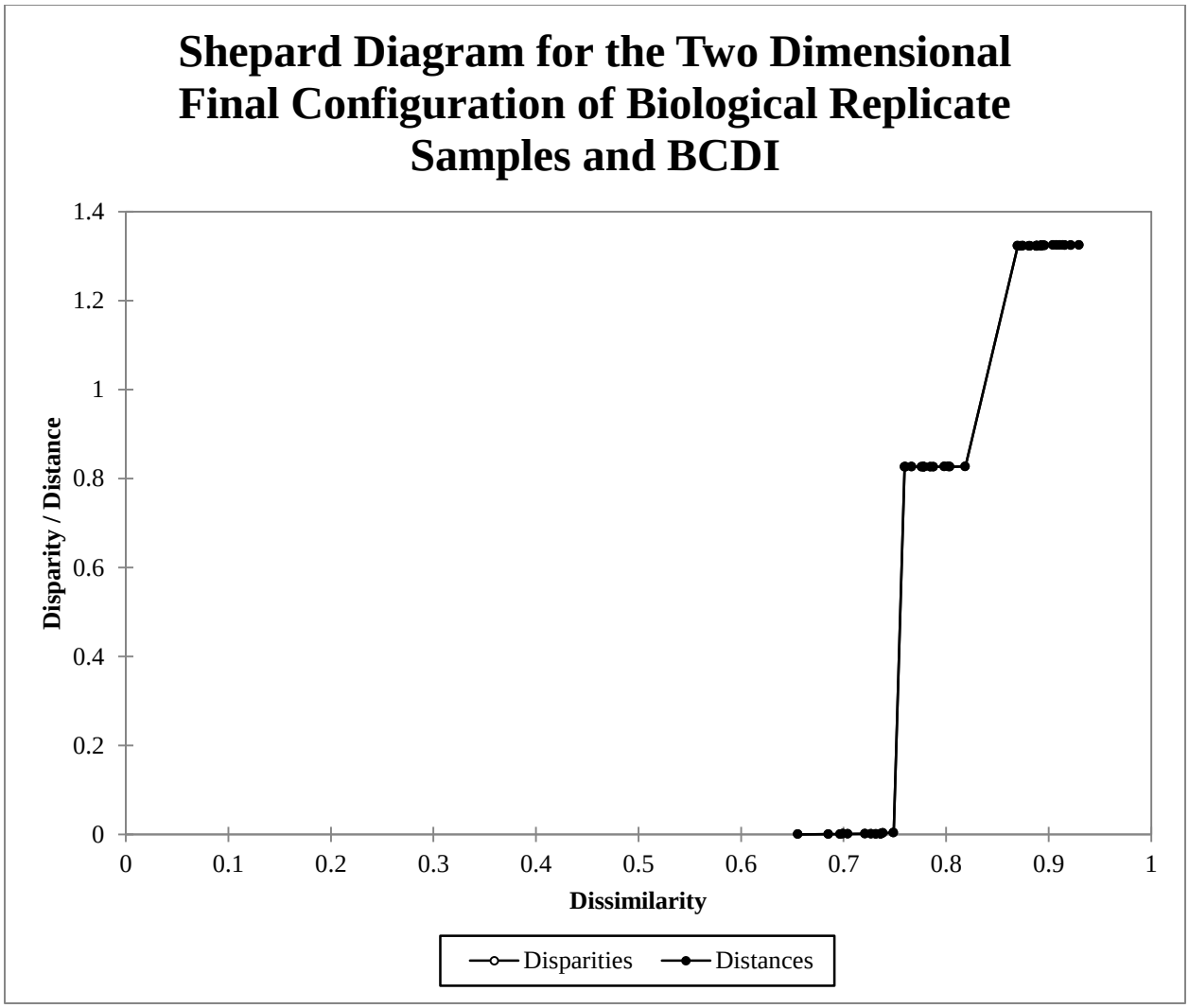

Figure 17. Shepard diagram for the two dimensional final configuration developed from BCDI of biological replicate soil samples. All distances fall nearly on top of their corresponding disparities indicating good correlation of the two in the final configuration.

The final configurations for both diversity indices (Figures 18 and 19) showed complete separation of the three habitats with replicate samples clustered closely. The configurations developed from BCDI and SDC were not exactly the same, though the relative positions of the habitats were similar.

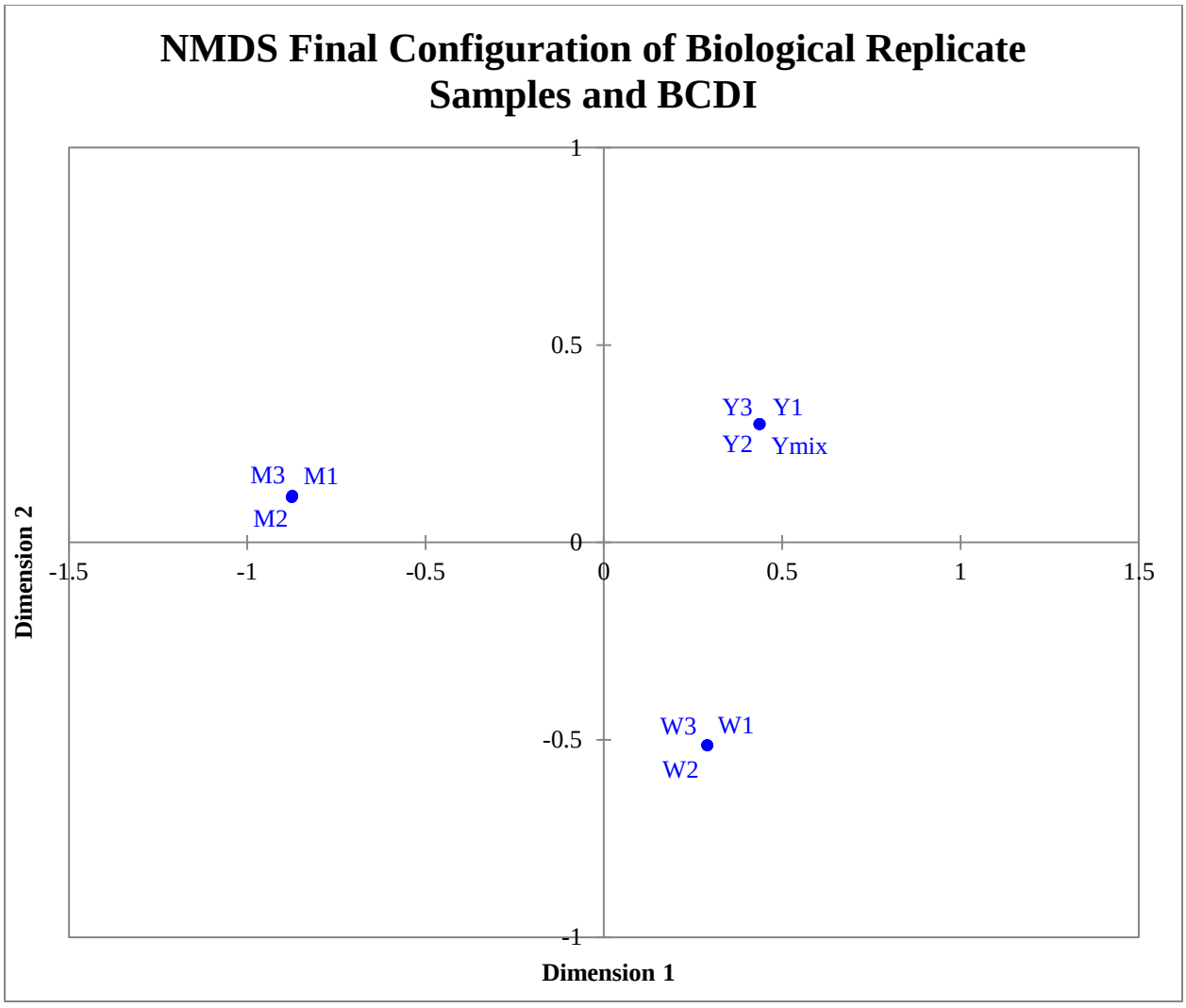

Figure 18. NMDS final configuration of biological replicate soil samples from BCDI. Replicate samples cluster very closely within their respective habitat while separating from the other habitats. See Table 2 for site names corresponding to abbreviations.

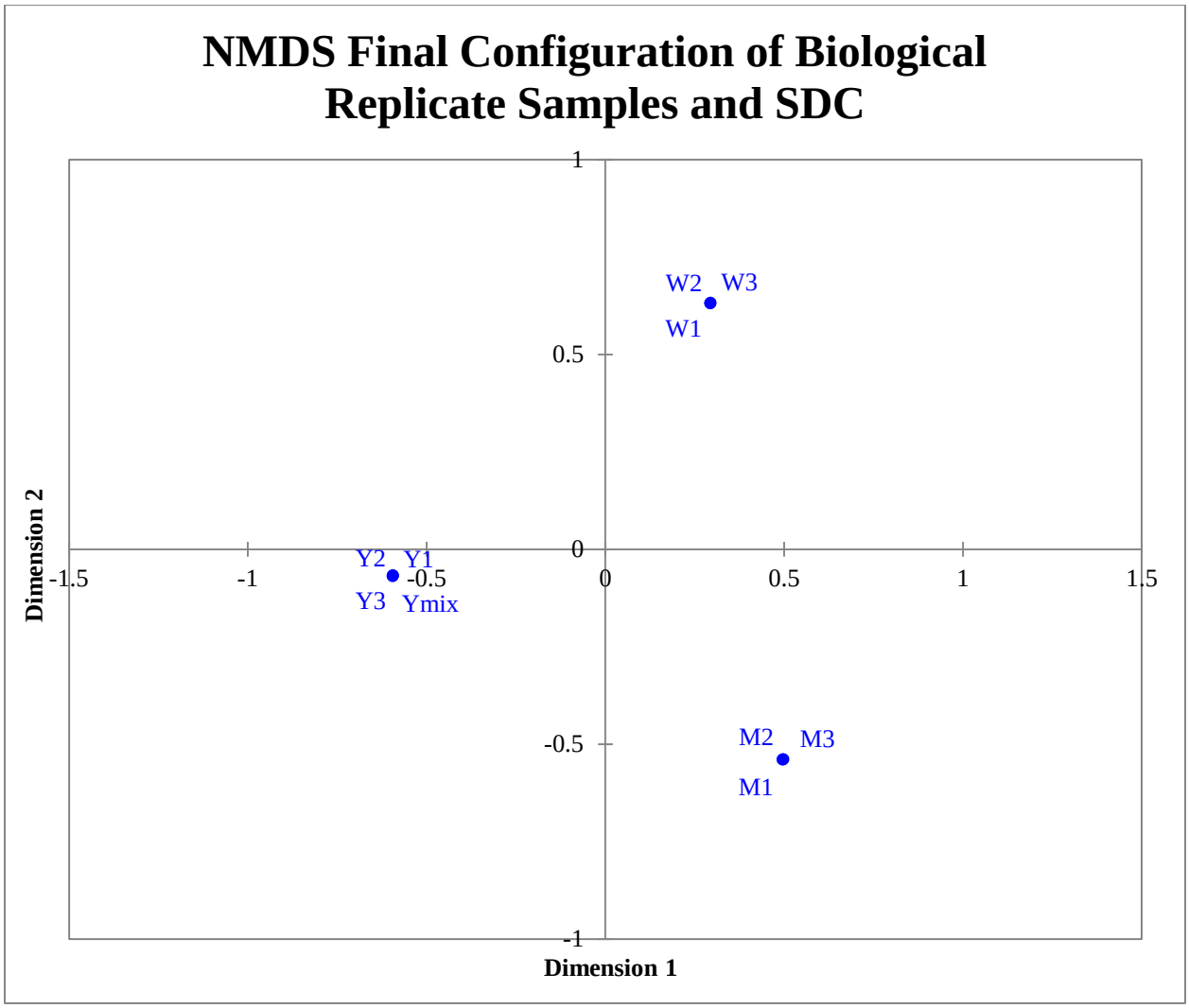

Figure 19. NMDS final configuration of biological replicate soil samples from SDC. Replicate samples cluster very closely within their respective habitat while separating from the other habitats. See Table 2 for site names corresponding to abbreviations.

Hierarchical Cluster Analysis

Dendrograms for biological replicate samples showed they clustered by habitat for both BCDI and SDC and all three linkage methods (Figures 20 – 22 and Appendix 8 Figures 81 – 83).Within the yard samples cluster, Ymix and Y3 were always the most similar. Y2 was the next most similar followed by Y1 in all dendrograms except Figure 20. This cluster was formed between 0.704 – 0.797. Deciduous woods replicates clustered in all between 0.669 – 0.795 with

W1 and W2 being more similar than W3. Similarly, the marsh edge samples clustered the same for all linkage methods and indices between 0.738 – 0.796, with M2 and M3 being more similar followed by M1. The dissimilarity values where clusters formed were similar between complete and UPGMA linkage methods for both BCDI and SDC. In all dendrograms the deciduous woods and yard samples were more similar to each other than they were to the marsh edge. The dissimilarity value for the formation of one cluster encompassing all habitats was above 0.869 for all dendrograms.

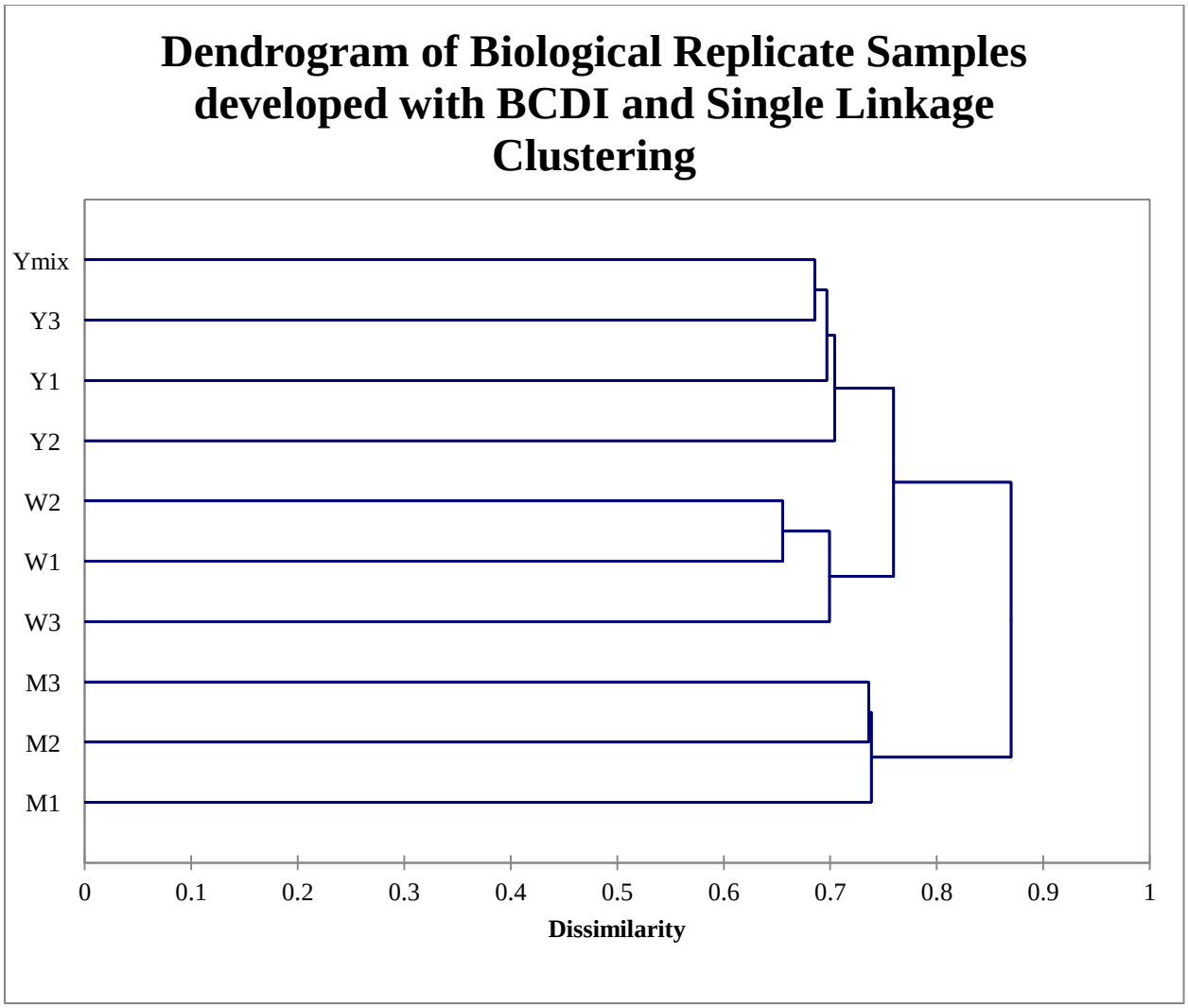

Figure 20. Dendrogram of biological replicate soil samples developed with BCDI and single linkage clustering. Three clusters are formed by habitat with replicate samples being more similar to each other than to other habitats. A cluster of yard samples is formed at 0.704 with one of the deciduous woods samples at 0.699. The marsh edge samples cluster at 0.738 and are most dissimilar from the other two habitats forming a cluster with them at 0.870. See Table 2 for site names corresponding to abbreviations.

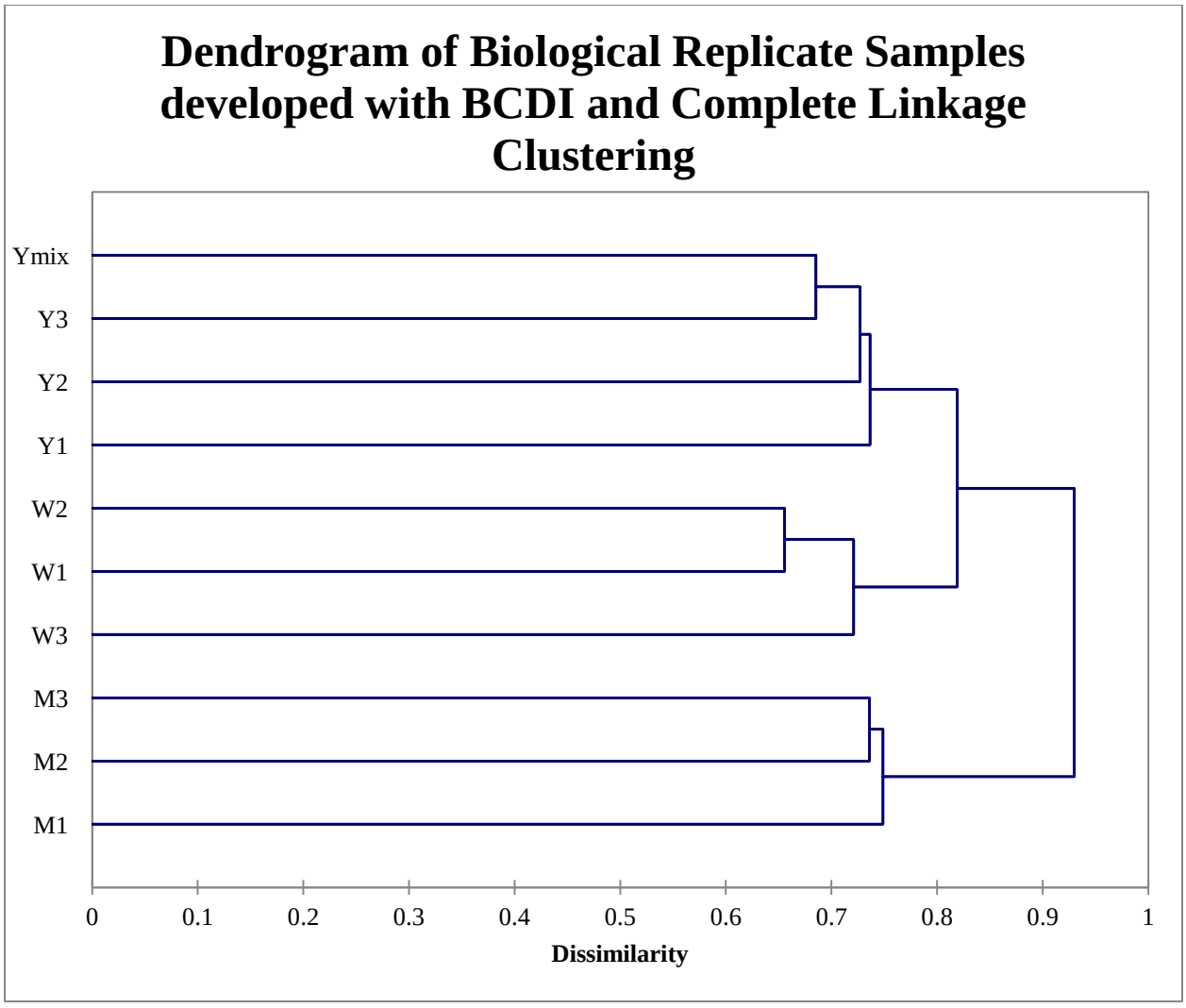

Figure 21. Dendrogram of biological replicate soil samples developed with BCDI and complete linkage clustering. Three clusters are formed by habitat with replicate samples being more similar to each other than to other habitats. The yard samples cluster at 0.737 while the deciduous woods group at 0.721. The marsh edge samples form a cluster at 0.749 and are most dissimilar from the other two habitats forming a cluster with them at 0.930. See Table 2 for site names corresponding to abbreviations.

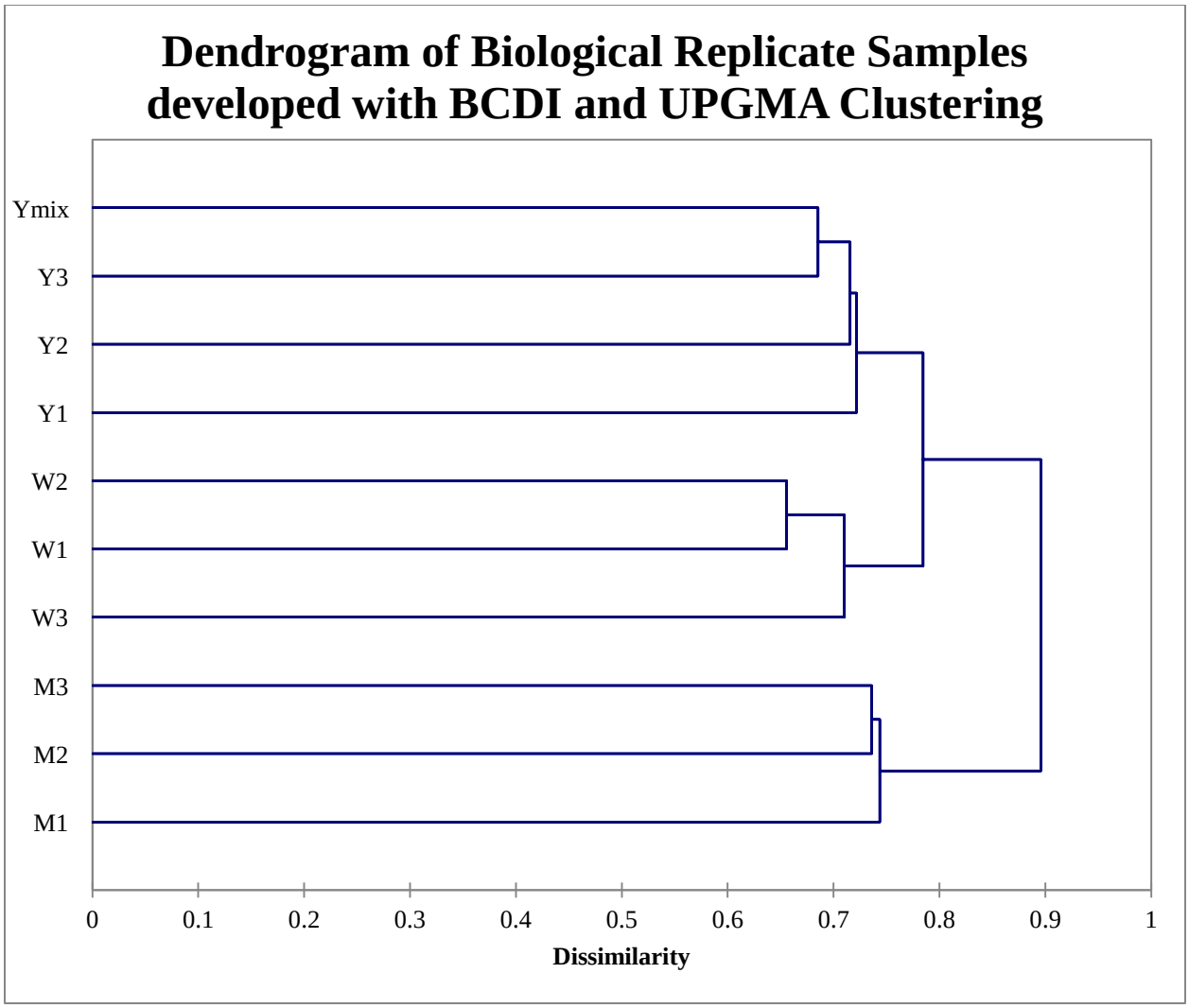

Figure 22. Dendrogram of biological replicate soil samples developed with BCDI and UPGMA clustering. Three clusters are formed by habitat with replicate samples being more similar to each other than to other habitats. Yard samples form a cluster at 0.722 while the deciduous woods group at 0.710. The marsh edge samples cluster at 0.744 and are most dissimilar from the other two habitats forming a cluster with them at 0.896. See Table 2 for site names corresponding to abbreviations.

Pairwise Comparisons

Ninety comparisons were made with the ∫-LIBSHUFF statistic for the biological replicate samples. This test revealed significant differences between habitats when using a Bonferroni corrected p-value of p=0.00056 (Figure 23). Further, M2 and M3 were statistically different,

while the other within marsh edge comparisons were not. All marsh edge replicates were different than the deciduous woods and yard samples. There were no within habitat significant differences for either the deciduous woods or yard sites; while, 75% of deciduous woods and yard samples differed.

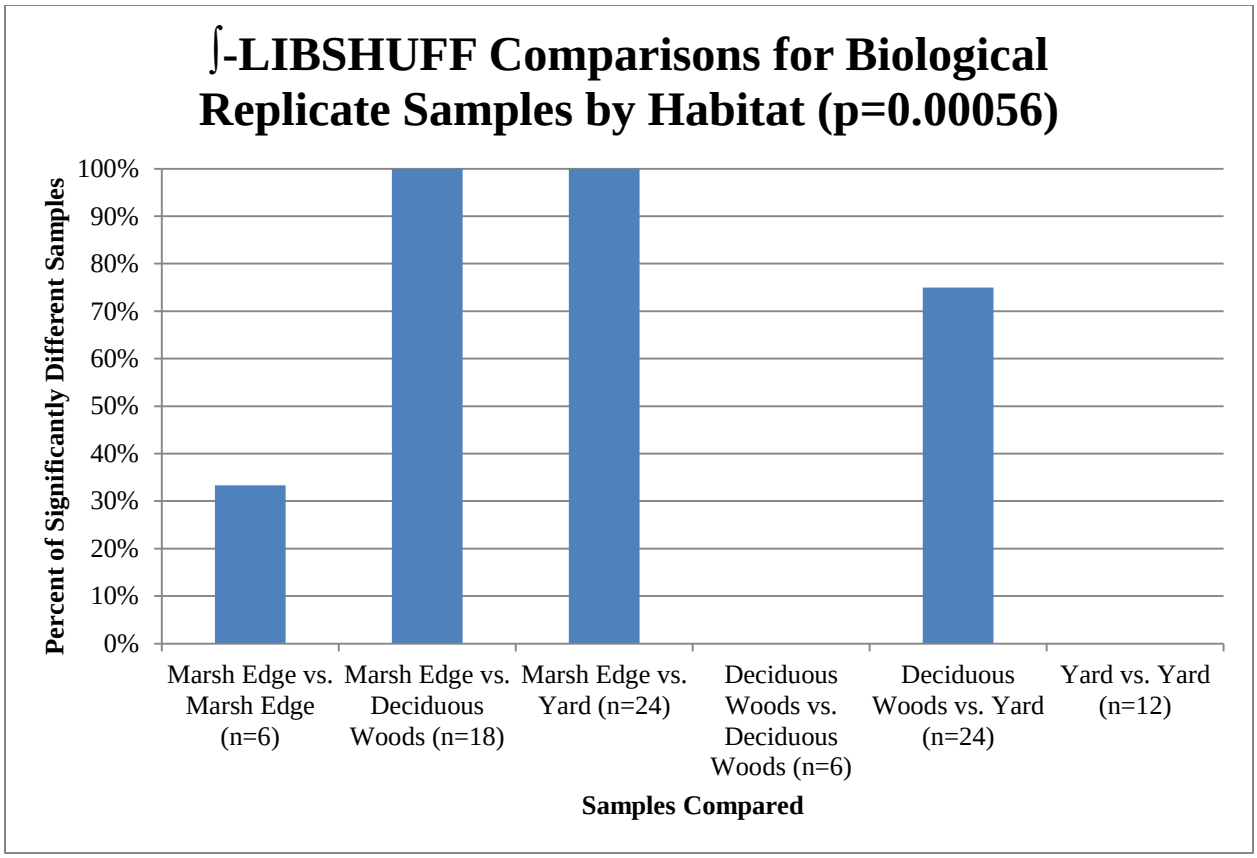

Figure 23. Pairwise comparisons of biological replicate soil samples using ∫-LIBSHUFF. Each bar represents the percent of samples that were statistically different. Thirty-three percent of marsh edge samples (M2 vs. M3) were different when compared to themselves; however, no deciduous woods or yard replicates were. All marsh edge samples were significantly different from the other habitats. Seventy-five percent of deciduous woods and yard samples differed statistically.

UniFrac revealed significant differences between habitats but not within them for the 45 comparisons made when using a Bonferroni corrected p-value of 0.0011 (Figure 24). All marsh

edge replicates differed significantly from the deciduous woods and yard samples. Further, 25% of comparisons between deciduous woods and yard habitats were statistically different.

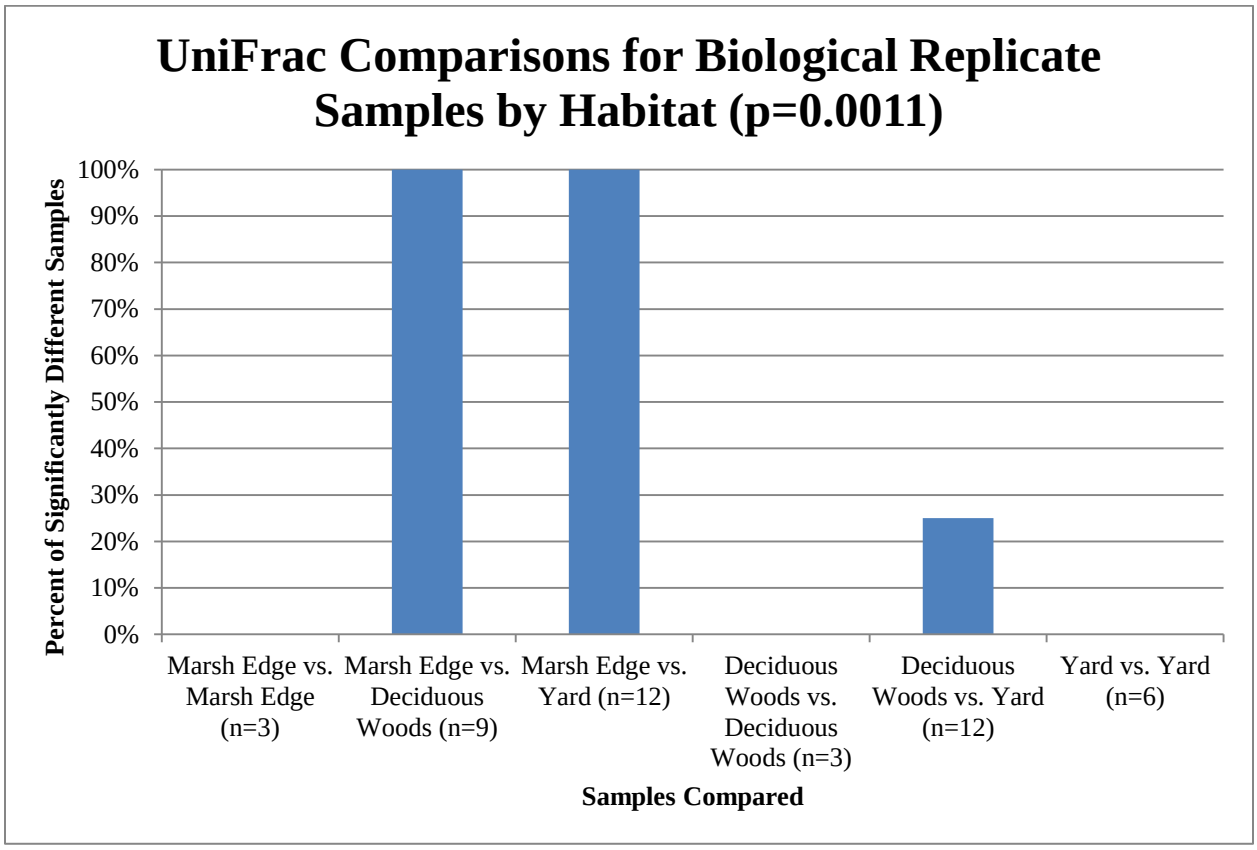

Figure 24. Pairwise comparisons of biological replicate soil samples using UniFrac. Each bar represents the percent of samples that were statistically different. No within habitat differences were seen; however, all marsh edge samples were significantly different from the other two habitats. Additionally, 25% of deciduous woods and yard samples differed.

Habitat Distance Samples Analysis

Nonmetric Multidimensional Scaling

Treated Yard Samples

The Scree diagram for treated yard distance samples and BCDI (Appendix 9-1 Figure 84) had high stress in one dimension with a decrease (elbow) in higher dimensions. The Scree

diagram for SDC dissimilarities (Figure 25) had no elbow, only a gradual decrease in stress as dimensions increased. Shepard diagrams for both indices (Appendix 9-1 Figures 85 and 86) had disparities and distances plotting closely. Final configurations for both diversity indices (Figures 26 and 27) showed similar relationships. Loose clusters were formed in two dimensions. The first included the main sample as well as all 5’ and 10’ distances and south 50’. Not all samples fell within standard error of each other though they plotted closely. The second cluster was comprised of the north 50’ and 100’ samples, which were within standard error in the BCDI plot. The last was formed by the east 50’ and 100’ distances with west 50’ and 100’. The east 100’ and west 50’ were within standard error while the other two samples were not but fell close in the BCDI plot while the east samples were within standard error as were the west ones in the SDC configuration. The south 100’ distance was outside standard error of all other samples.

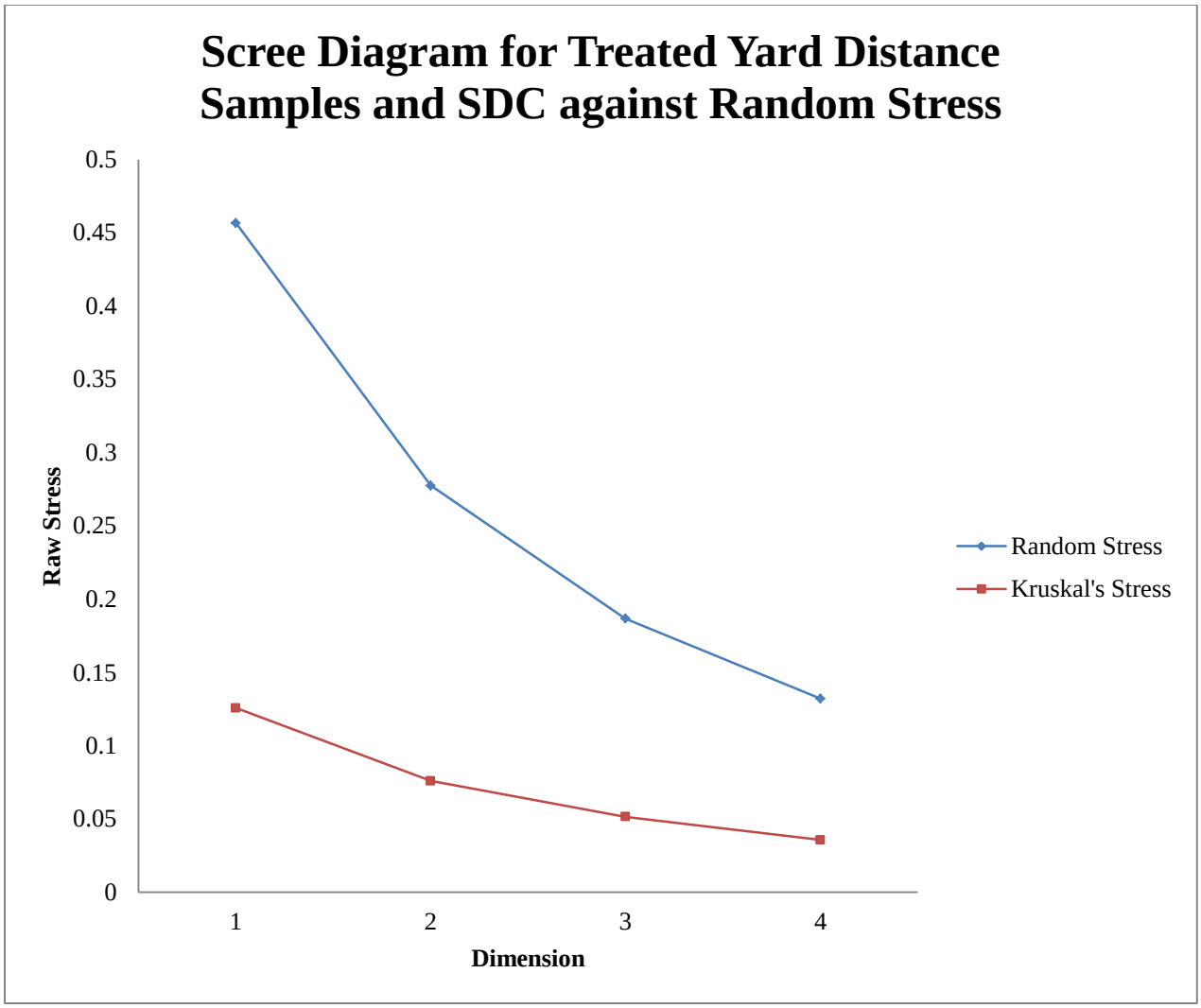

Figure 25. Scree diagram developed by NMDS for the final configurations of treated yard distance soil samples over four dimensions from SDC. Random stress was used as a threshold for the acceptance of final configurations. No configuration exceeded the threshold though the stress for the one dimension plot was higher than the others. There is no elbow at any dimension; however, for consistency two dimensional plots were analyzed.

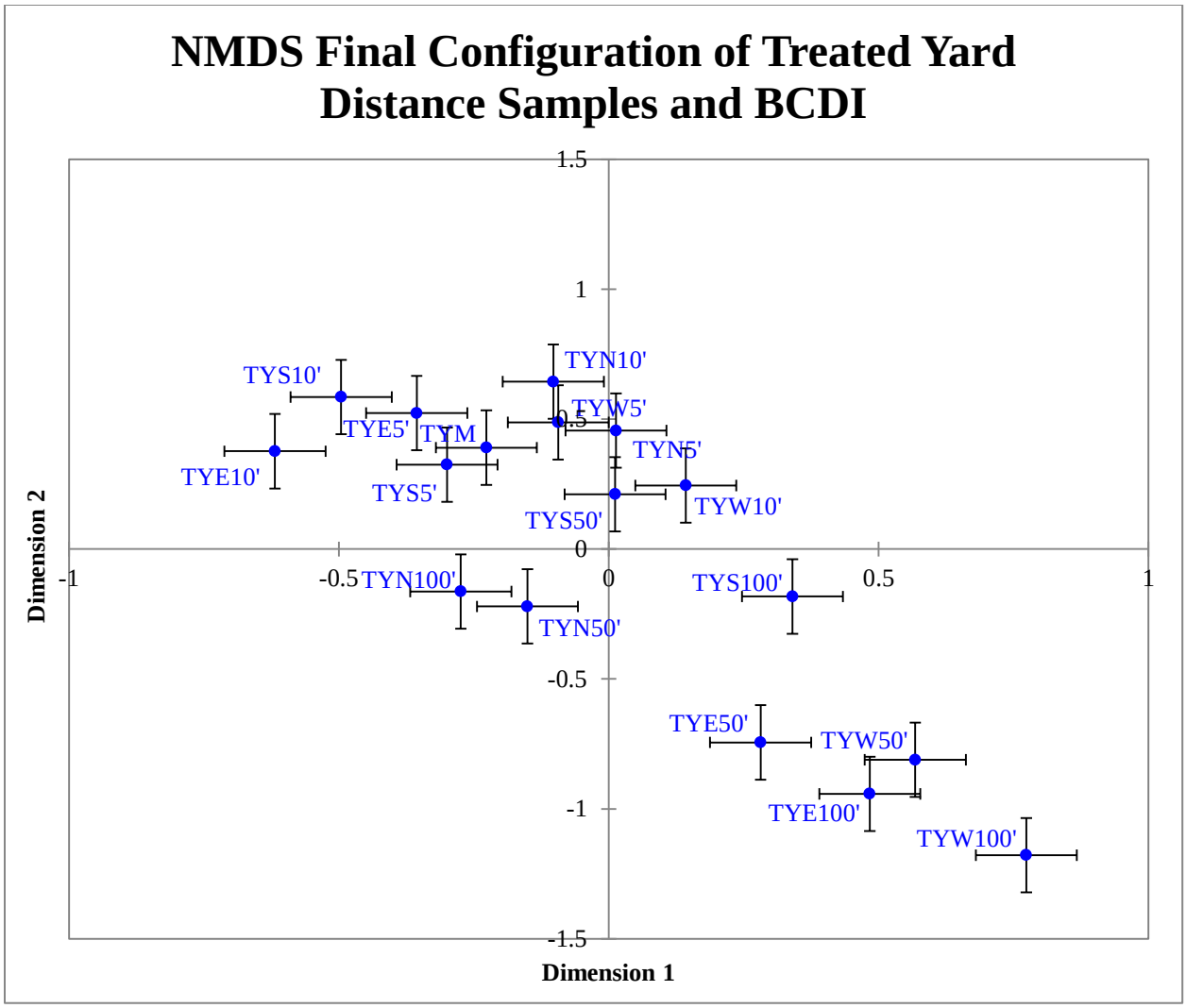

Figure 26. NMDS final configuration of treated yard distance soil samples from BCDI. Loose clusters are present in two dimensions. The first has the main sample as well as all 5’ and 10’ distances and south 50’ plotting closely but not all within standard error. The second cluster contains only the north 50’ and 100’ samples, which are within standard error but not close. The last has the east 50’ and 100’ as well as west 50’ and 100’ samples though the east 100’ and west 50’ distances are only within standard error. The south 100’ distance is outside standard error of all other samples. See Table 3 for site names corresponding to abbreviations.

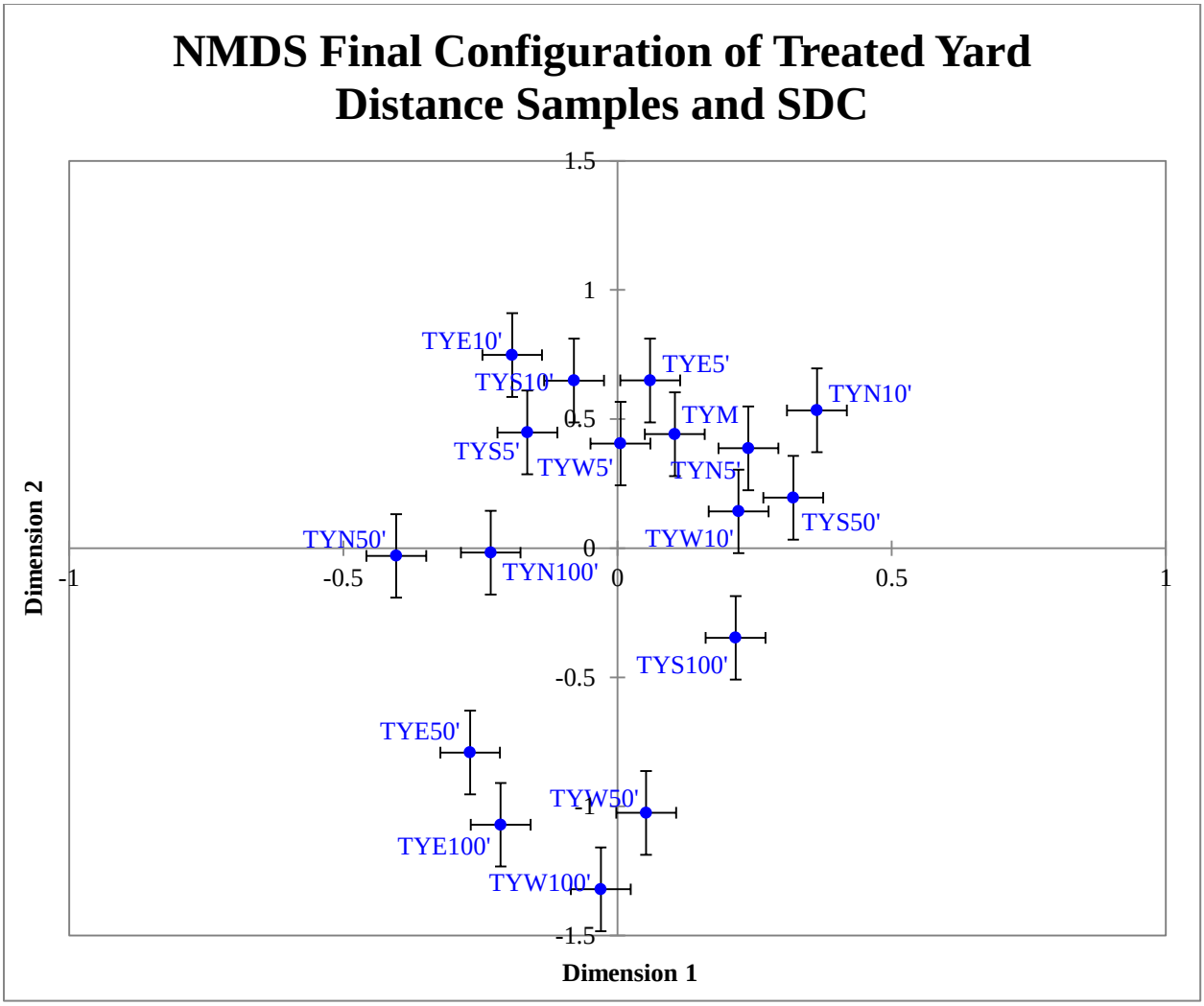

Figure 27. NMDS final configuration of treated yard distance soil samples from SDC. Loose clusters are present in two dimensions. The first has the main sample as well as all 5' and 10' distances and south 50' plotting closely but not all within standard error. The second cluster contains only the north 50' and 100' samples, which are not within standard error. The last has the east 50' and 100' as well as west 50' and 100' samples. The east distances are within standard error as are the west ones. The south 100' distance is outside standard error of all other samples. See Table 3 for site names corresponding to abbreviations.

Yard Samples

The Scree diagram for BCDI dissimilarities of yard distance samples (Appendix 9-2 Figure 87) had no elbow, only a gradual decrease in stress as dimensionality increased. However, the SDC Scree diagram (Appendix 9-2 Figure 89) did have an elbow formed by a large decrease

in stress from one to two dimensions with a further decrease in higher dimensions. Shepard diagrams (Appendix 9-2 Figures 88 and 90) had distances plotting closely with disparities. Final configurations (Figures 28 and 29) had one main cluster, which was comprised of the main sample with all 5’ and 10’ distances. All 50’ distances were associating with the main cluster, though outside standard error. The 100’ samples were outside standard error of all other samples and plotted varying distances from the main cluster.

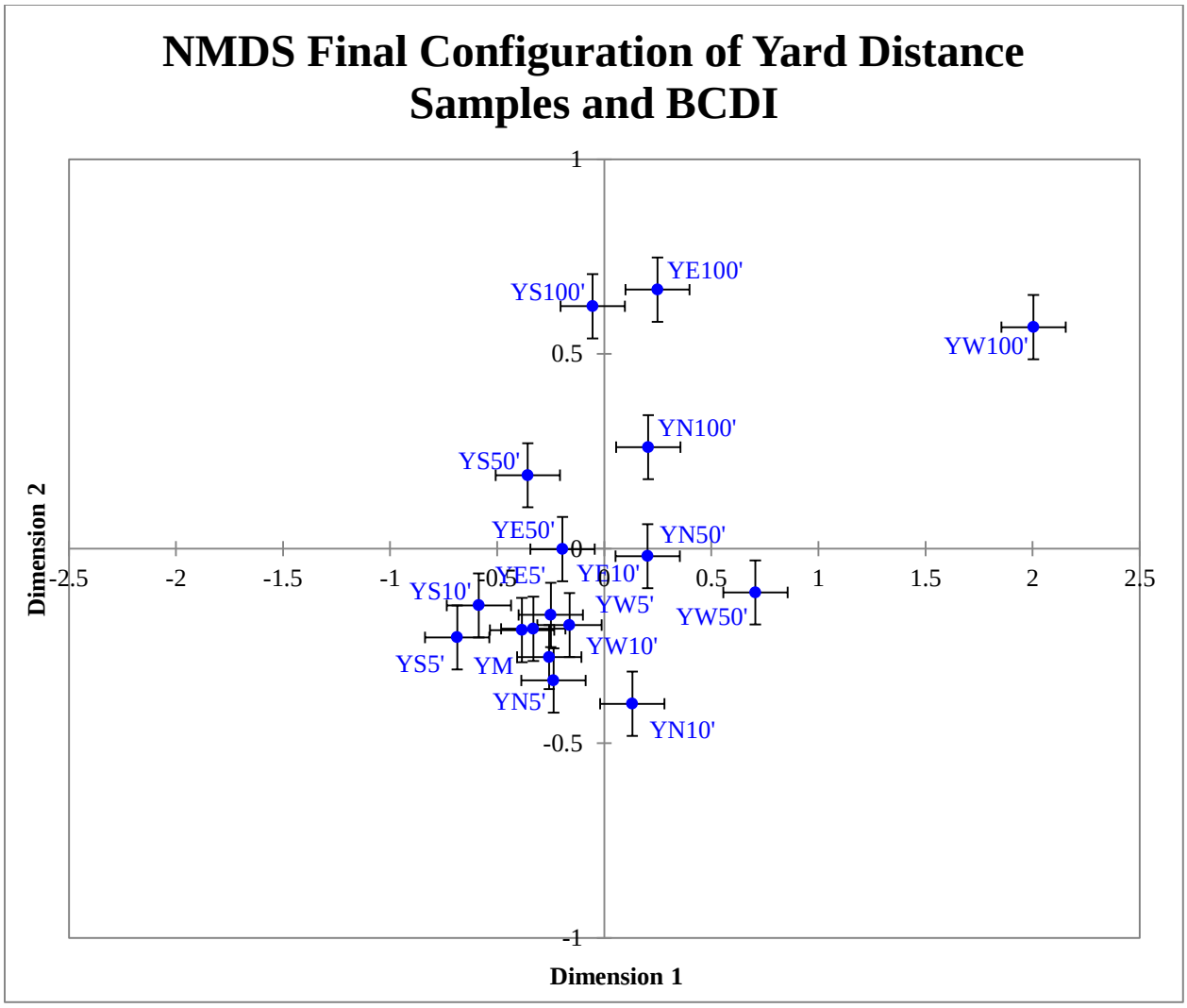

Figure 28. NMDS final configuration of yard distance soil samples from BCDI. A single main cluster is present that includes the main sample and all 5’ and 10’ distances. The 50’ distances associate with the main cluster, though outside standard error of it and each other. All 100’ samples are outside standard error of all other samples. They plot varying distances from the main cluster with the west 100’ being the furthest away. See Table 3 for site names corresponding to abbreviations.

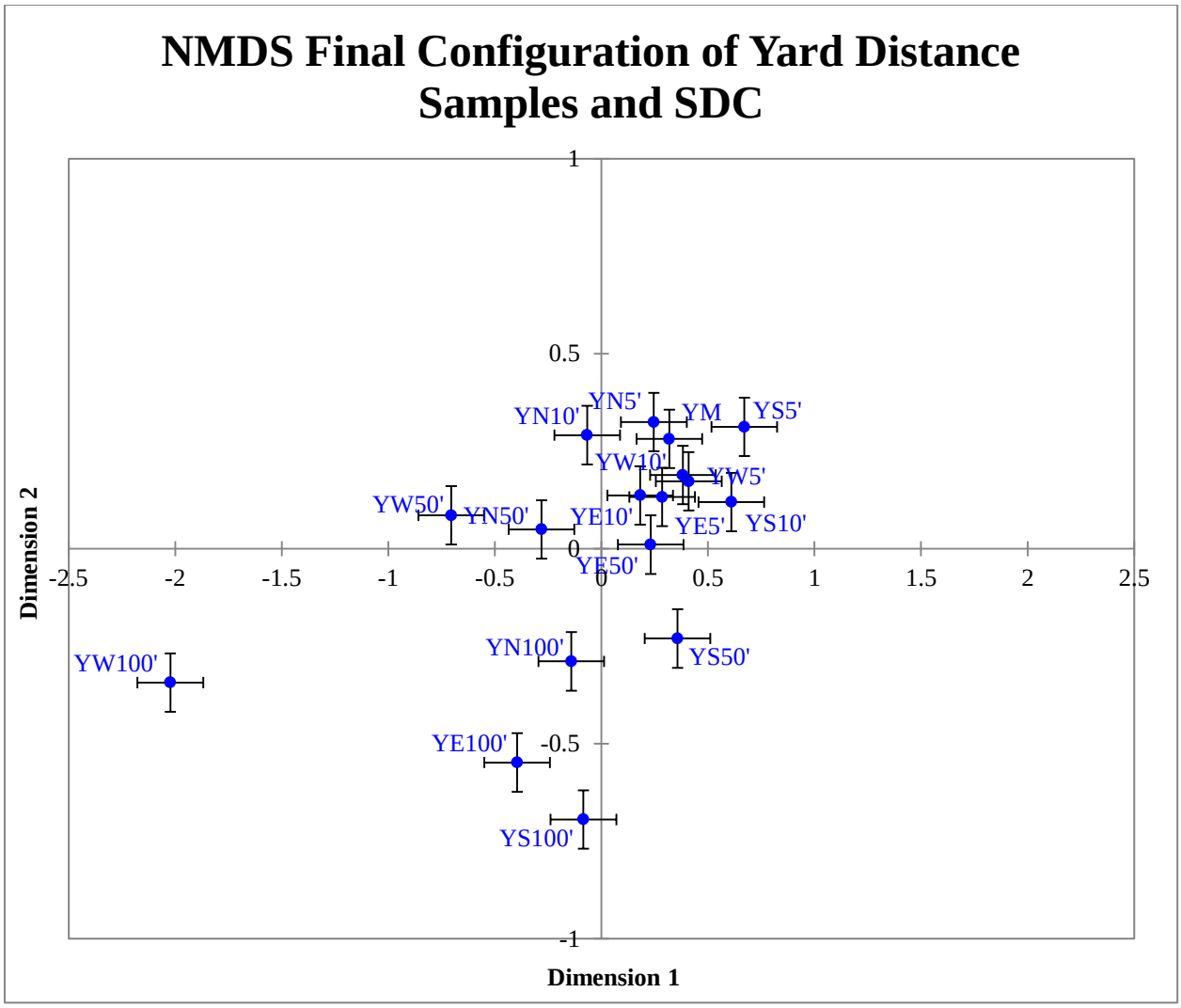

Figure 29. NMDS final configuration of yard distance soil samples from SDC. A single main cluster is present that includes the main sample and all 5' and 10' distances. The 50' distances associate with the main cluster, though all are outside standard error of it except the east 50' sample. All 100' samples are outside standard error of all other samples. They plot varying distances from the main cluster with the west 100' being the furthest away. See Table 3 for site names corresponding to abbreviations.

Deciduous Woods Samples

Scree diagrams of the deciduous woods distance samples for both diversity indices (Appendix 9-3 Figures 91 and 93) had high stress in one dimension followed by decreasing stress (elbow) as dimensions increased. Shepard diagrams (Appendix 9-3 Figures 92 and 94) had

close association of disparities and distances. Final configurations developed for both BCDI and SDC (Figures 30 and 31) had no discernible clusters and distances appeared randomly assorted. Samples not geographically close often plotted near each other (*e.g.* east 10' and north 10'). The main sample did not cluster with any other sample in both configurations.

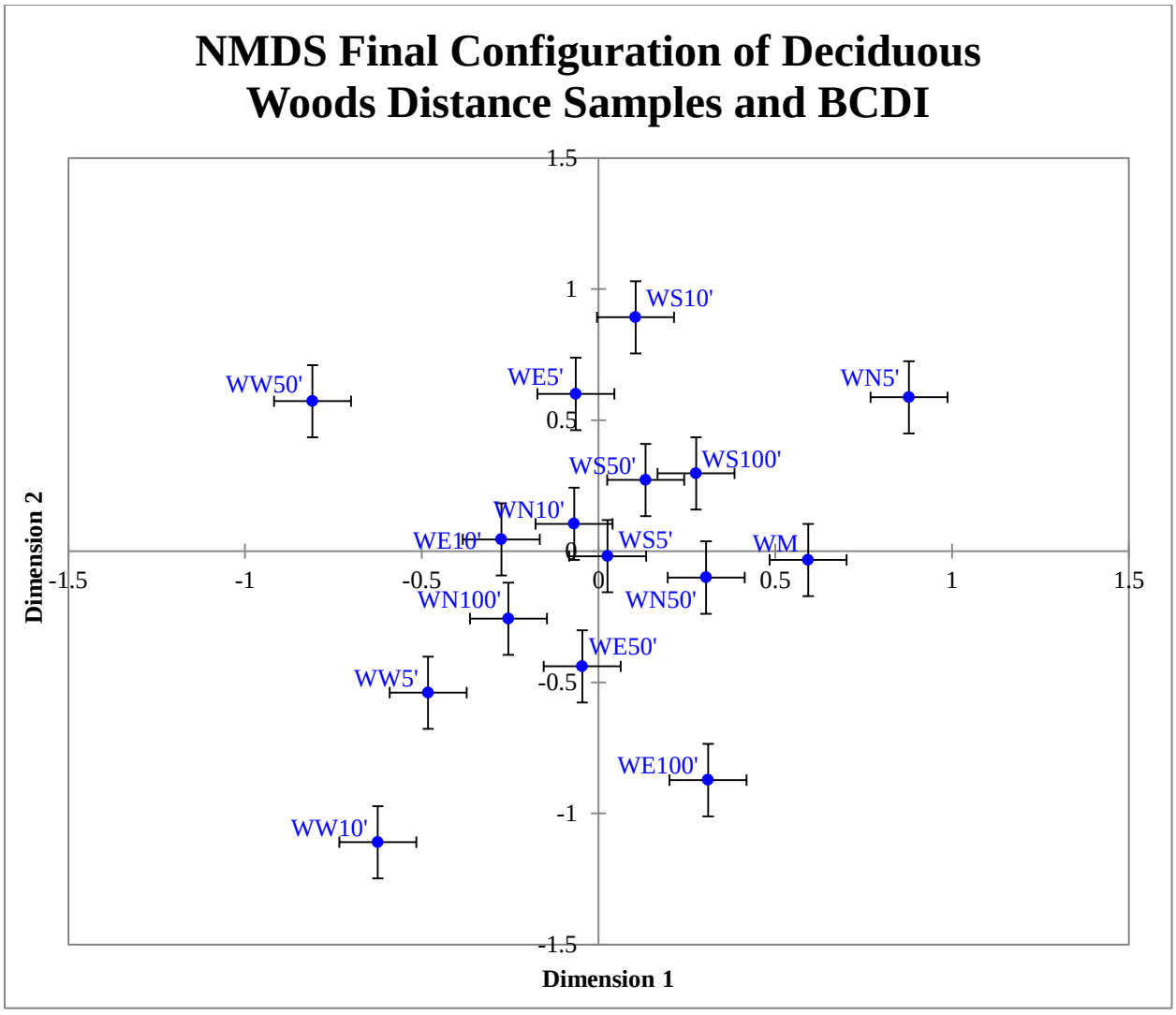

Figure 30. NMDS final configuration of deciduous woods distance soil samples from BCDI. No discernible clusters are evident in two dimensions. Geographically close samples are not clustering together and the main sample is not associated with any other samples. See Table 3 for site names corresponding to abbreviations.

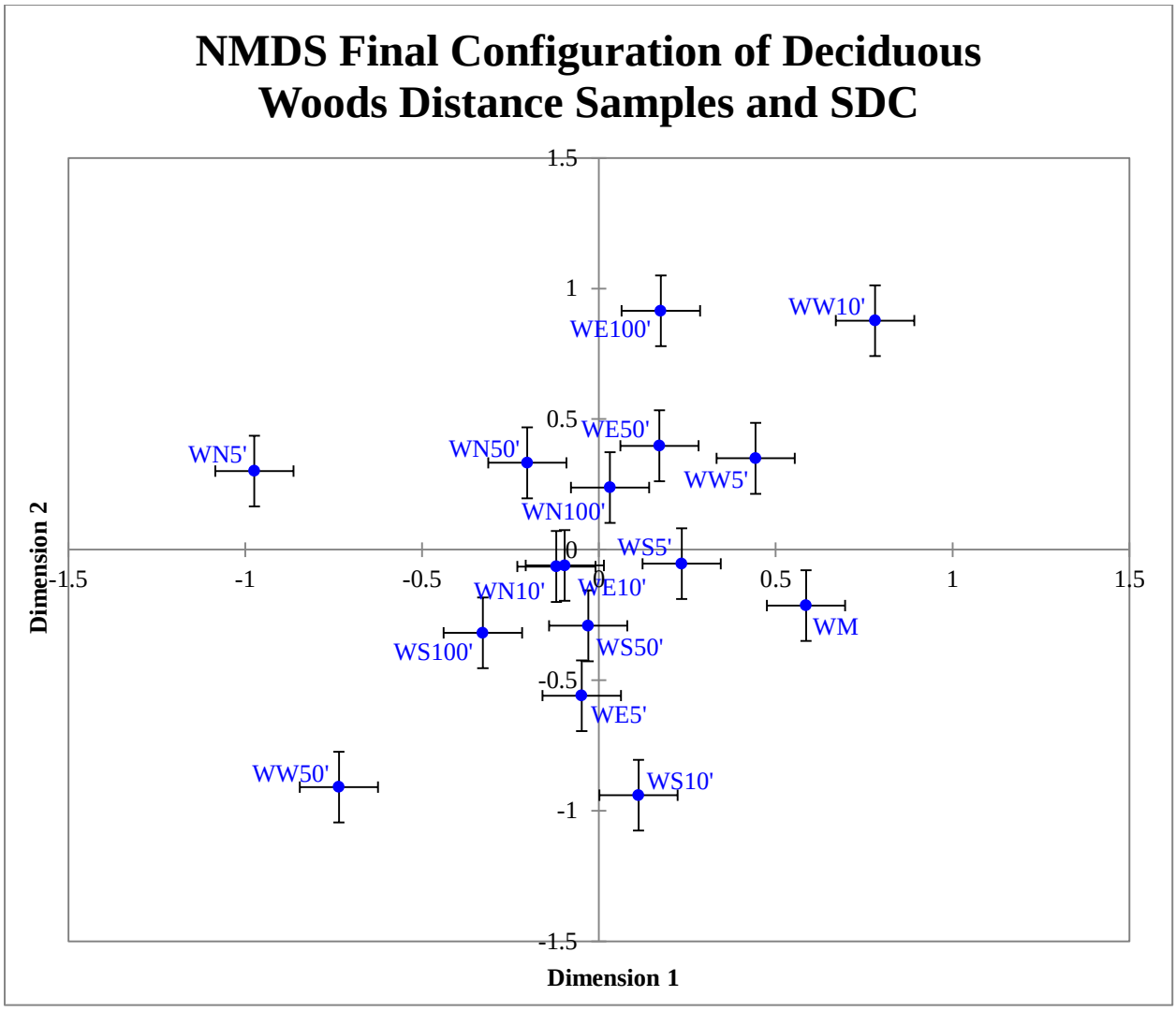

Figure 31. NMDS final configuration of deciduous woods distance soil samples from SDC. No discernible clusters are evident in two dimensions. Geographically close samples are not clustering together and the main sample is not associated with any other samples. See Table 3 for site names corresponding to abbreviations.

Hierarchical Cluster Analysis

Treated Yard Samples

Dendrograms developed from treated yard distance samples with single linkage for both diversity indices (Figure 32 and Appendix 10-1 Figure 96) had extensive chaining and thus no strong conclusions could be drawn. The complete linkage dendrogram for BCDI (Figure 33) had

three clusters. The first consisted of two smaller clusters that merged at 0.745. This cluster contained the north 50’ and 100’ distances as well as the west 10’, and south 50’ and 100’ samples. The second cluster was formed first by the main and all 5’ distances, followed by the remaining 10’ samples. The final cluster was composed of the east 50’ and 100’ samples as well as the west 50’ and 100’ distances at 0.756. This cluster joined the others at 0.910.

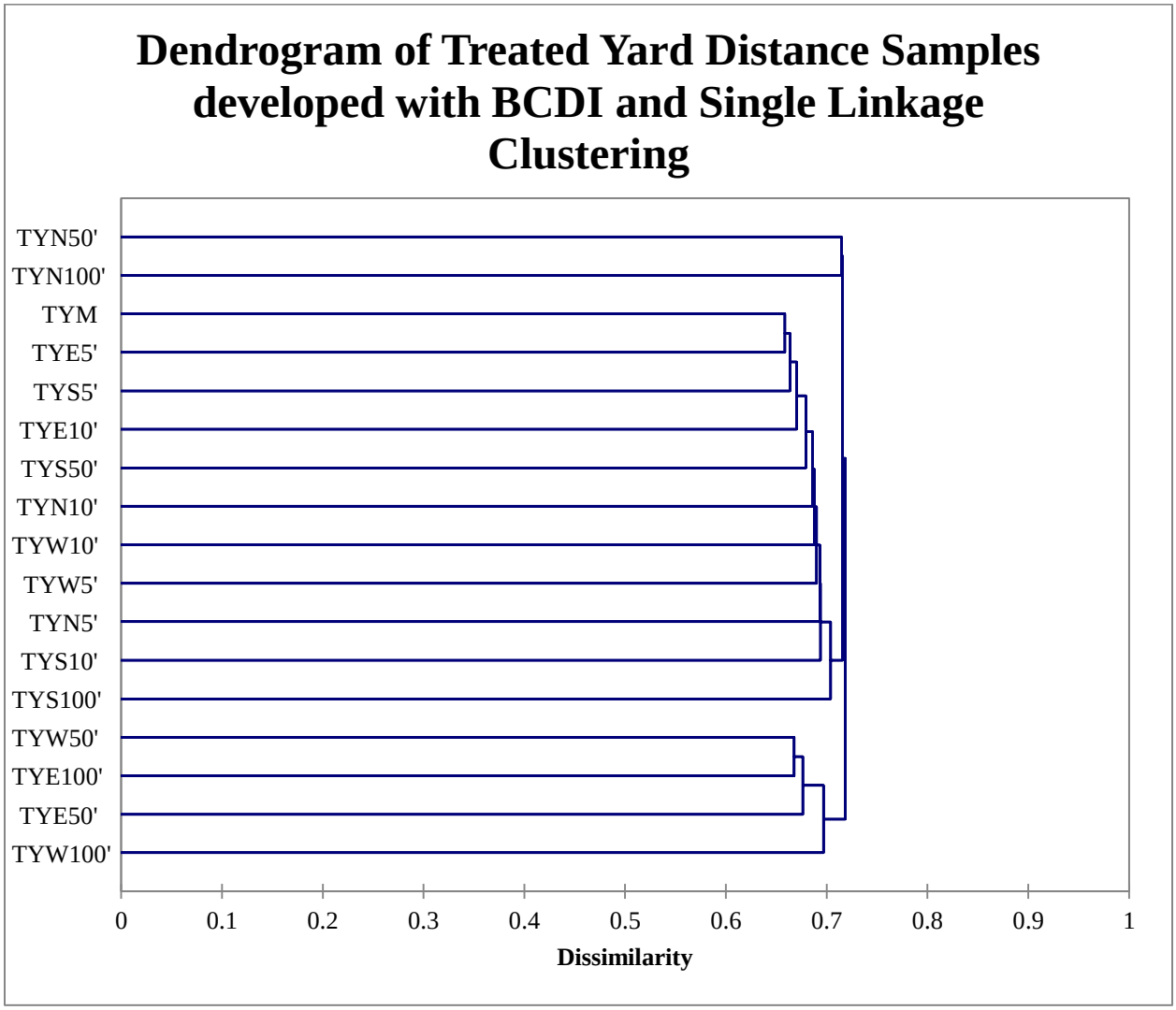

Figure 32. Dendrogram of treated yard distance soil samples developed with BCDI and single linkage clustering. There are three possible clusters present; however, the largest of them shows chaining for the short distances analyzed. All samples in this cluster are grouped at very similar dissimilarities with little structure within it. See Table 3 for site names corresponding to abbreviations.

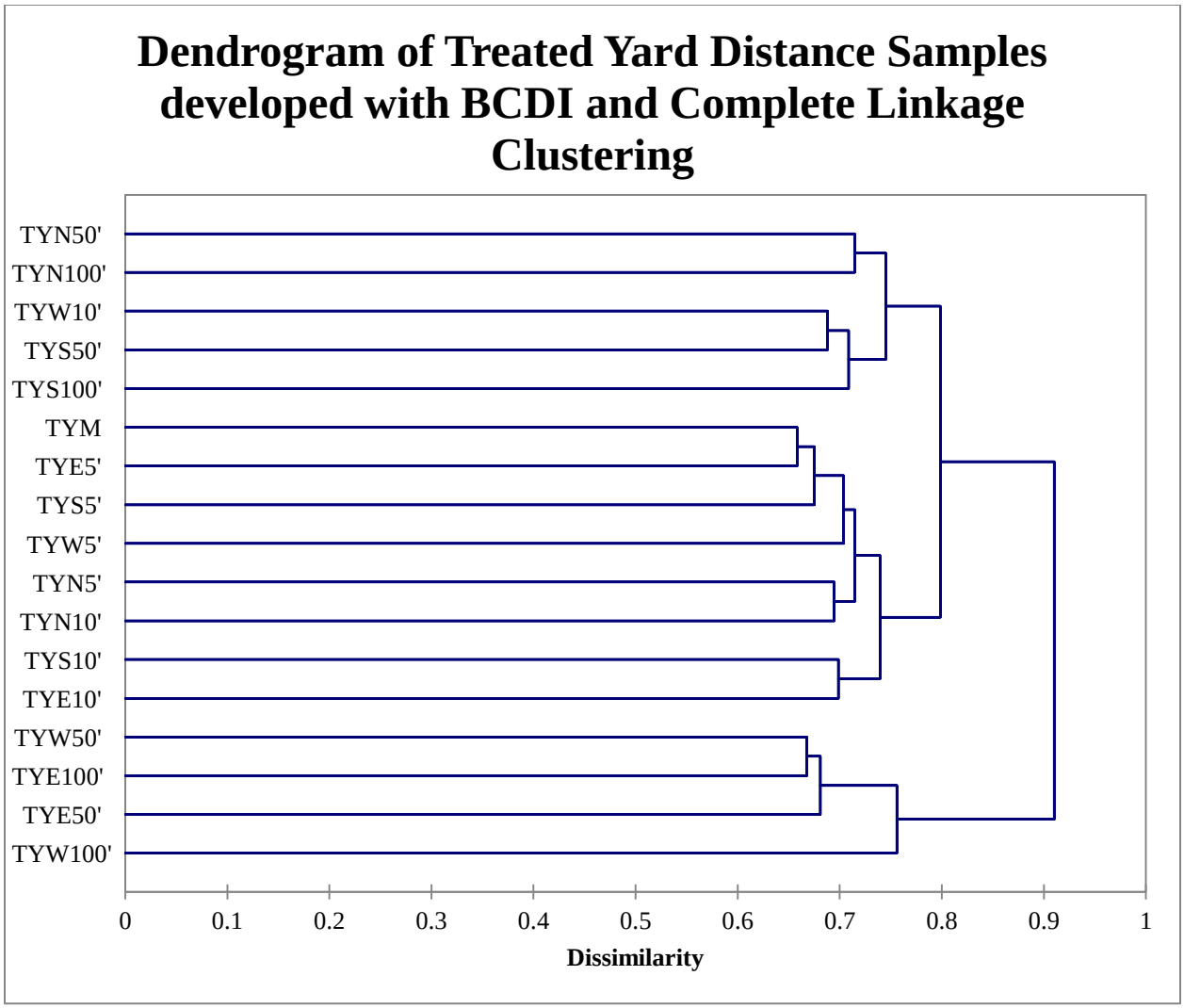

Figure 33. Dendrogram of treated yard distance samples developed with BCDI and complete linkage clustering. Three clusters are present. The first consists of two smaller clusters which are merged at 0.745. This cluster contains the north 50' and 100' distances as well as the west 10', and south 50' and 100' samples. The second cluster is formed first by the main and all 5' distances, followed by the remaining 10' samples grouping at 0.739. The final cluster is composed of the east 50' and 100' samples as well as the west 50' and 100' distances clustered at 0.756. This cluster joins the others at 0.910. See Table 3 for site names corresponding to abbreviations.

Dendrograms developed with SDC and complete linkage (Appendix 10-1 Figure 97) and UPGMA for both indices (Appendix 10-1 Figures 95 and 98) were similar, though dissimilarity values where clusters formed were different. Three loose clusters were present. The first had the

north 50’ and 100’ samples followed by the south 100’ distance between 0.724 – 0.760. The second was formed between 0.712 – 0.757 by two five member clusters and contained the main, all the 5’ and 10’ distances along with the south 50’ samples. The final cluster was composed of the east and west 50’ and 100’ samples between 0.720 – 0.734. This cluster joined the others at 0.812 or higher.

Yard Samples

Dendrograms of yard distance samples developed with single linkage for both diversity indices (Figure 34 and Appendix 10-2 Figure 100) had extensive chaining and thus no strong conclusions could be drawn. The complete linkage BCDI dendrogram (Figure 35) had two broad clusters. The first was a joining of two three member groups at 0.725 that contained the north 10’ and 50’ and west 50’ and north 100’, south 100’, and east 100’ samples, respectively. The second cluster was also a grouping of two smaller clusters at 0.630. The three member one contained the south 5’, 10’, and 50’ samples while the other had the main, the three remaining 5’ distances, east 10’ and 50’, and west 10’ samples. The west 100’ distance joined the two clusters at 0.923. Similarly, the SDC complete linkage dendrogram (Appendix 10-2 Figure 101) had two large clusters. The first had a two and three member group formed at 0.736 that contained the north 50’ and west 50’ and north 100’, south 100’, and east 100’ samples, respectively. The second cluster was also a grouping of two smaller clusters at 0.734. The three member one contained the south 5’, 10’, and 50’ samples while the other had the main, the three remaining 5’ distances, east 10’ and 50’, and west 10’ samples. The west 100’ distance joined the two clusters at 0.917.

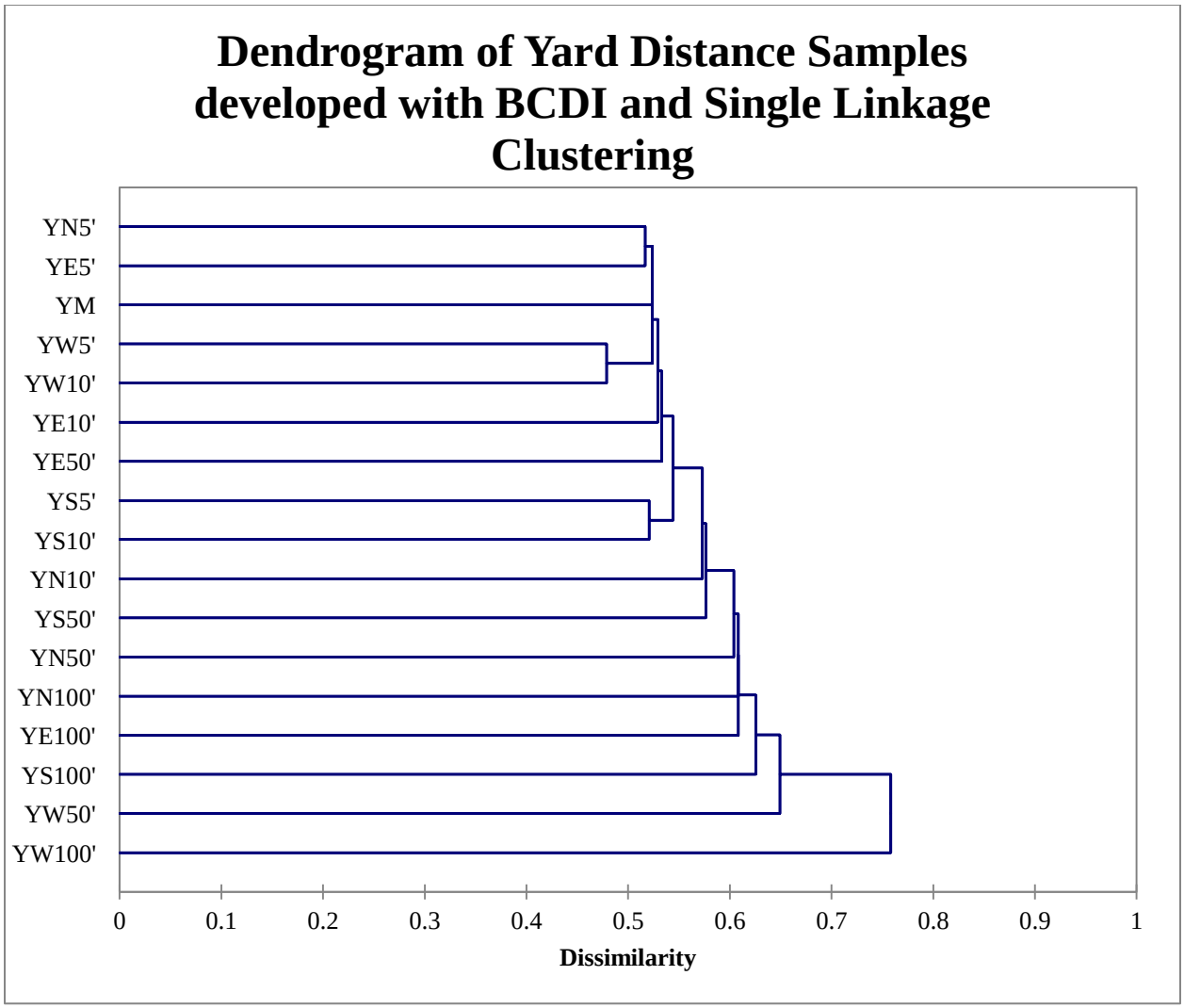

Figure 34. Dendrogram of yard distance samples developed with BCDI and single linkage clustering. There are possible clusters present; however, extensive chaining of samples makes interpretation difficult. This dendrogram was not analyzed. See Table 3 for site names corresponding to abbreviations.

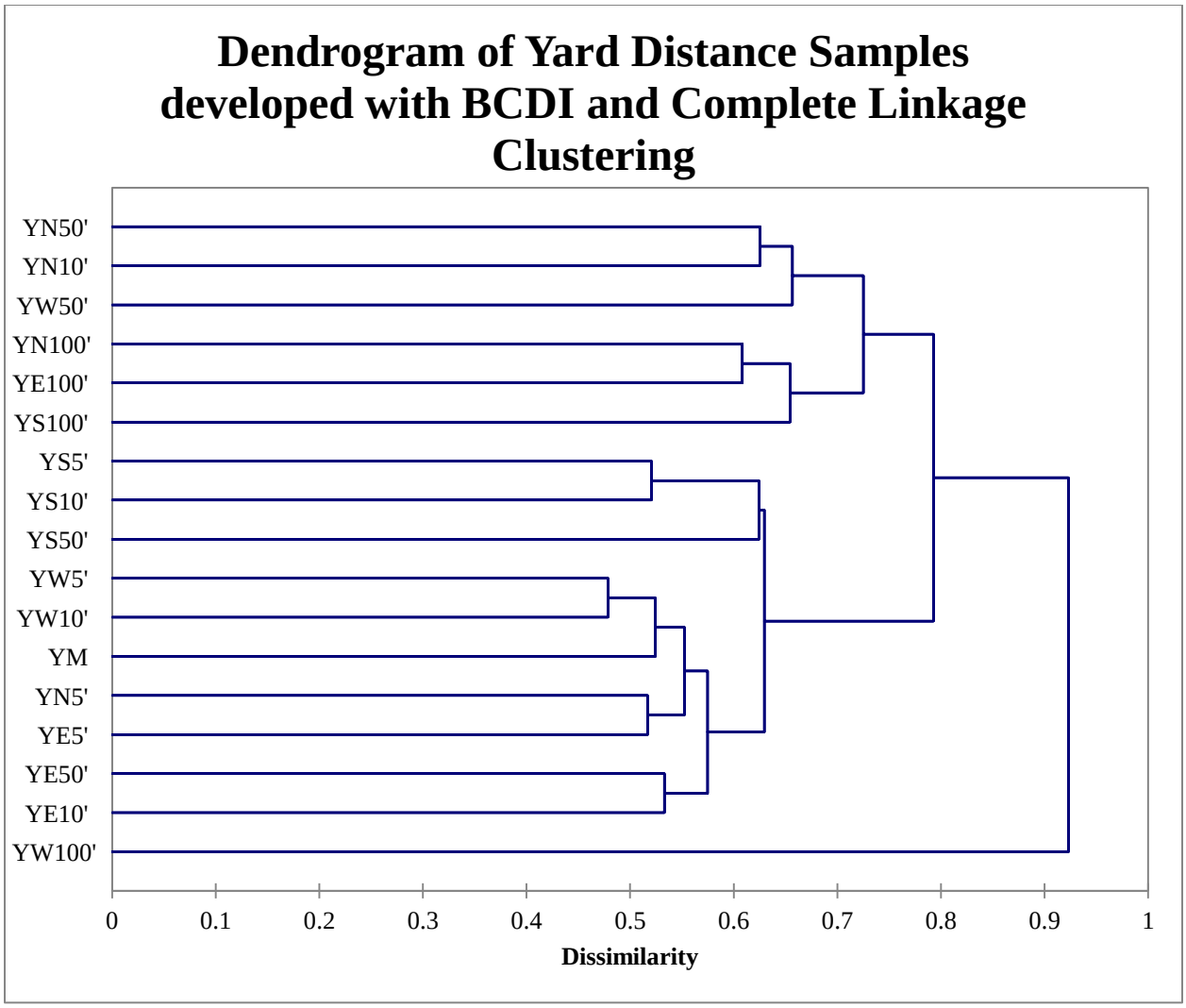

Figure 35. Dendrogram of yard distance soil samples developed with BCDI and complete linkage clustering. Two clusters are formed. The first is a joining of two three member groups that contain the north 10' and 50' and west 50' and north 100', south 100', and east 100' samples, respectively, at 0.725. Additionally, the second cluster is a grouping of two smaller clusters at 0.630. The three member one contains the south 5', 10', and 50' samples while the other has the main, the remaining 5' distances, east 10' and 50', and west 10' samples. The west 100' distance joins the two clusters at 0.923. See Table 3 for site names corresponding to abbreviations.

The UPGMA clustering dendrograms (Appendix 10-2 Figures 99 and 102) had similar overall structure; though clustering within them was different. The BCDI one had two large clusters. The first was comprised of the north 100', south 100', east 100', and west 50' distances

at 0.688. The second cluster was formed at 0.638 by smaller two sample groups and single samples. The west 100’ distance joined the two clusters at 0.861. Similarly, the SDC dendrogram had two broad clusters. The first was comprised of a two and three member group at 0.719 that contained the north 50’ and west 50’ and north 100’, south 100’, and east 100’ samples, respectively. The second cluster was formed by smaller two sample groups and single samples at 0.682. The west 100’ distance joined the two clusters at 0.860.

Deciduous Woods Samples

Dendrograms of deciduous woods soils developed with single linkage for both diversity indices (Figure 36 and Appendix 10-3 Figure 104) had extensive chaining and thus no strong conclusions could be drawn. Two distinct clusters were present in the BCDI dendrogram developed with complete linkage (Figure 37). The first was comprised of the north 5’ and 10’, south 10’, 50’, and 100’, east 5’ and 10’, and west 50’ distances at 0.701. The other was formed by the remaining samples as two smaller clusters joined at 0.670. The two larger clusters grouped together at 0.768. Similarly, the SDC dendrogram developed with complete linkage clustering (Appendix 10-3 Figure 105) had two large groupings. The first was comprised of the north 5’ and 10’, south 10’, 50’, and 100’, east 5’ and 10’, and west 50’ distances at 0.718. The other was formed by the remaining samples as two smaller clusters joined at 0.702. The two larger clusters grouped together at 0.778. The final dendrograms, UPGMA clustering for BCDI and SDC dissimilarities (Appendix 10-3 Figures 103 and 106), had possible clustering of samples; however, none were distinct. There were small groupings of samples but no pattern of how they were associating was evident.

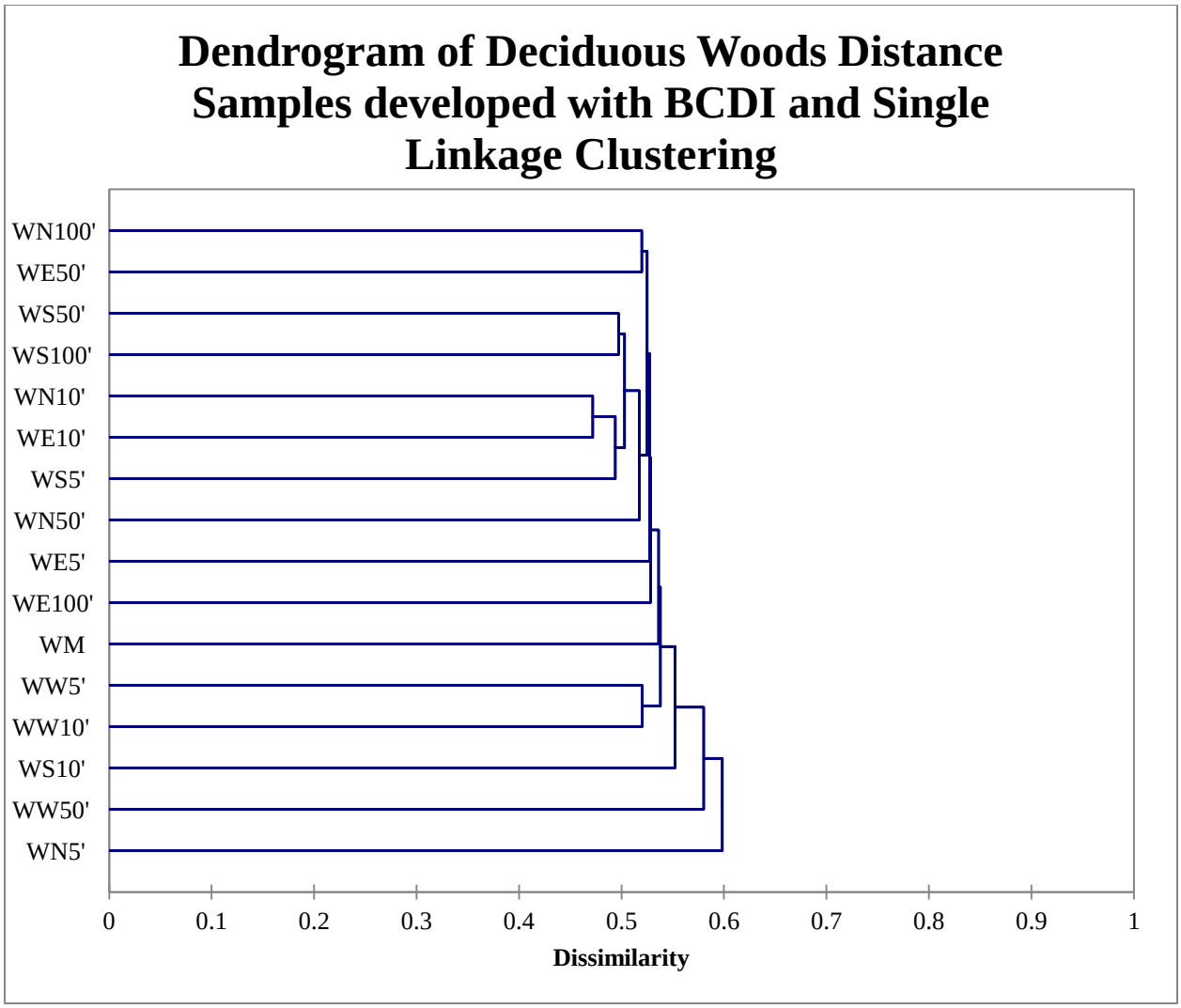

Figure 36. Dendrogram of deciduous woods distance samples developed with BCDI and single linkage clustering. There are possible clusters present; however, extensive chaining of samples makes interpretation difficult. This dendrogram was not analyzed. See Table 3 for site names corresponding to abbreviations.

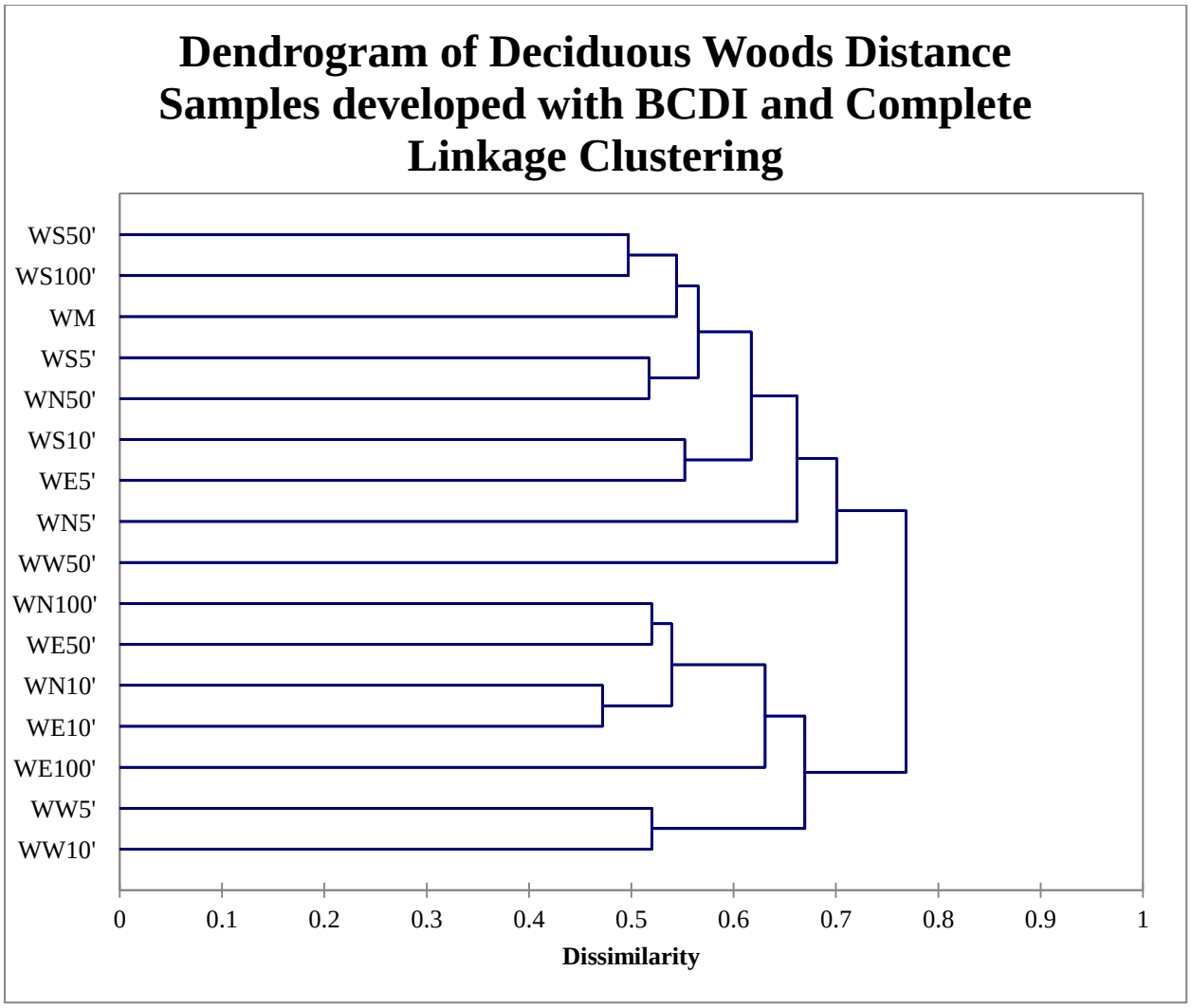

Figure 37. Dendrogram of deciduous woods distance soil samples developed with BCDI and complete linkage clustering. Two distinct clusters are present. The first is formed by some two member groups as well as single samples at 0.701. This cluster is comprised of the main sample as well as the north 5' and 50', all south distances, east 5', and the west 50' sample. The other cluster consists of the remaining samples and is formed by three two member groups and a single sample a 0.670. The two clusters group around 0.768. See Table 3 for site names corresponding to abbreviations.

Pairwise Comparisons

<u>Treated Yard Samples</u>

∫-LIBSHUFF comparisons for treated yard distance samples differed significantly at distances greater than 10 feet (Figure 38). The number of significantly different samples rose

with increased distance, finally leveling off around 50% for samples separated by more than 50 feet. Interestingly, UniFrac comparisons revealed statistical differences starting at distances greater than 20 feet (Figure 39). The number of statistically different samples rose to about 15% in the 51 – 100 feet range.

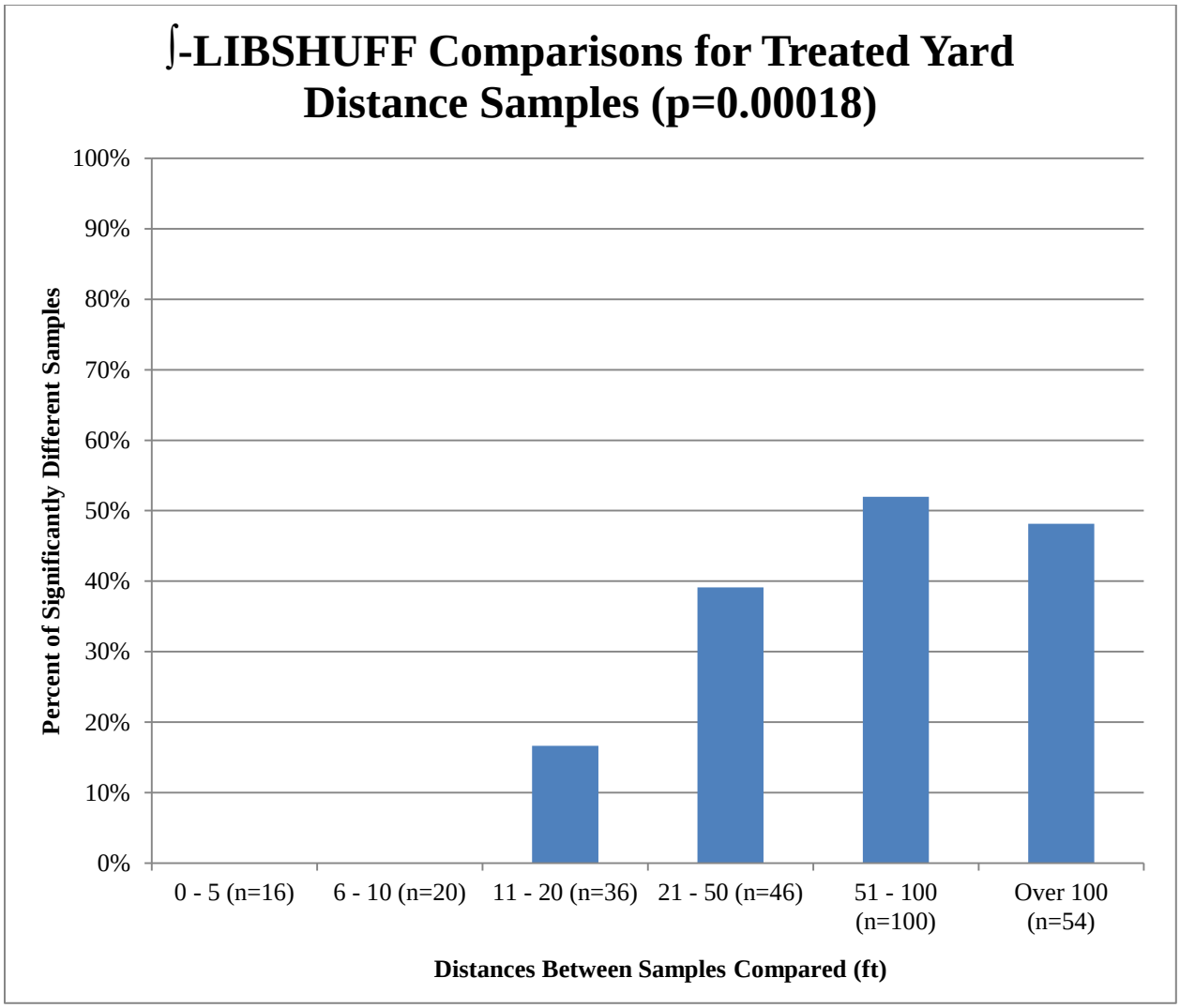

Figure 38. ∫-LIBSHUFF comparisons for treated yard distance soil samples show statistically significant differences at distances greater than 10 feet. Less than 20% of samples between 11 – 20 feet are significantly different. The percentage of different samples increases as the distance between them increase, finally leveling off around 50% for samples separated by 50 or more feet.

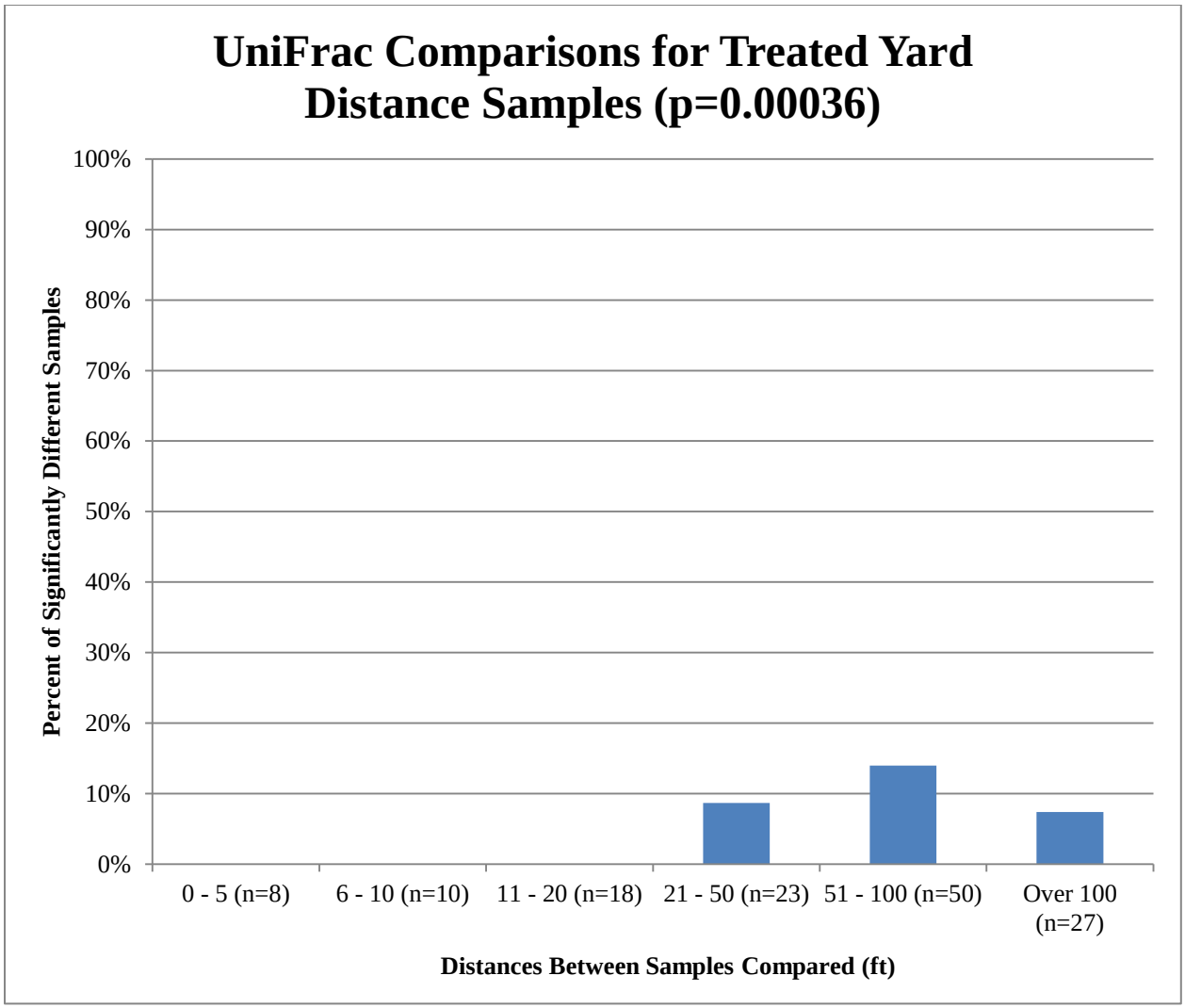

Figure 39. UniFrac comparisons for treated yard distance soil samples reveal statistical differences for distances greater than 20 feet. The number of significantly different distances rose to about 15% in the 51 – 100 feet range.

Yard Samples

∫-LIBSHUFF comparisons for yard distance samples showed significant differences beginning at distances of five feet (Figure 40). The number of different samples rose from 50% in the first distance range to 100% for all distances over 100 feet. Contrarily, UniFrac comparisons of yard distance samples were only statistically different for distances greater than 10 feet (Figure 41). The percent of different samples rose to 80% for distances over 100 feet.

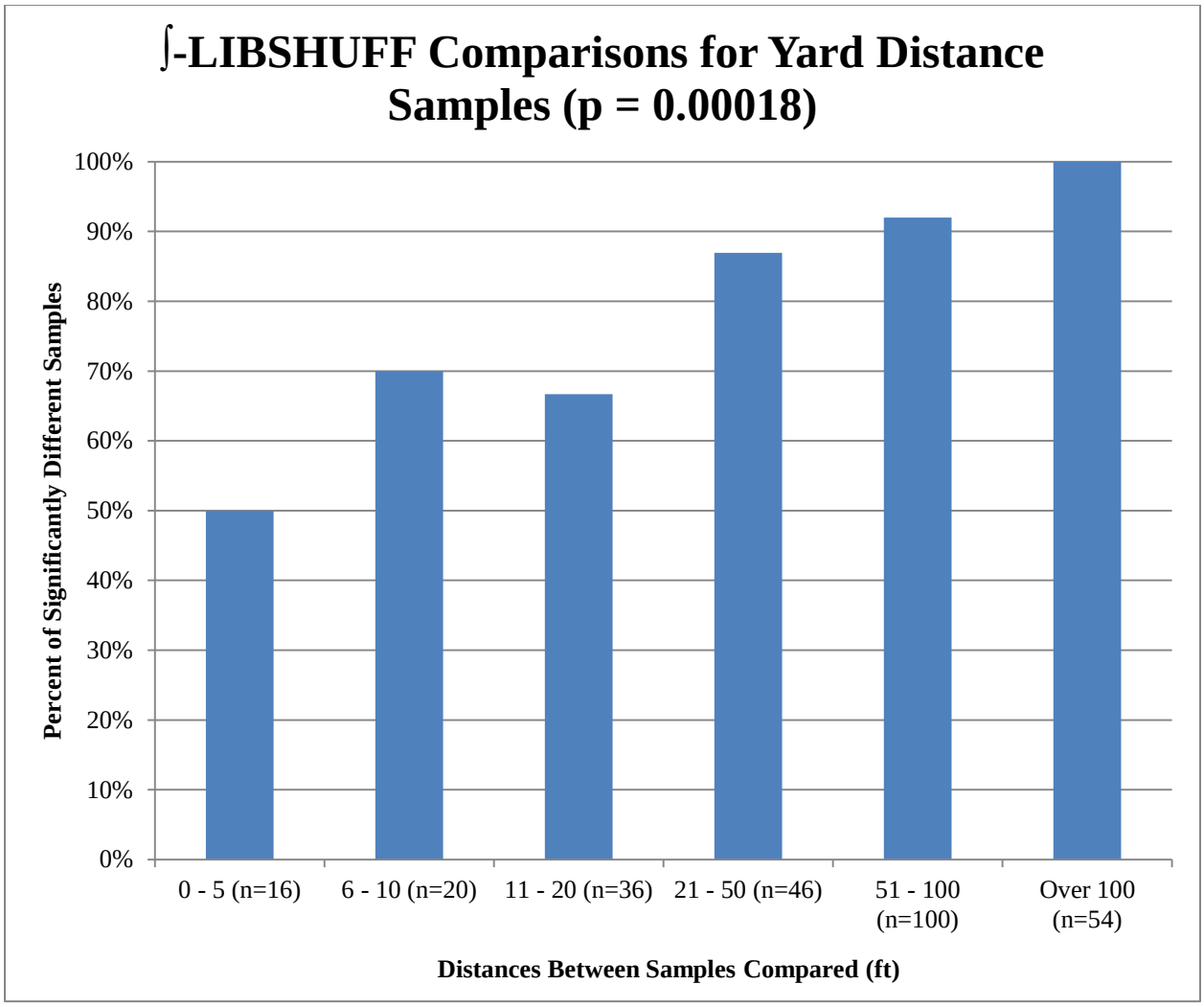

Figure 40. ∫-LIBSHUFF comparisons for yard distance soil samples show statistically significant differences at all distances. The percent of significantly different samples rose from 50% in the first distance range to 100% for all distances over 100 feet.

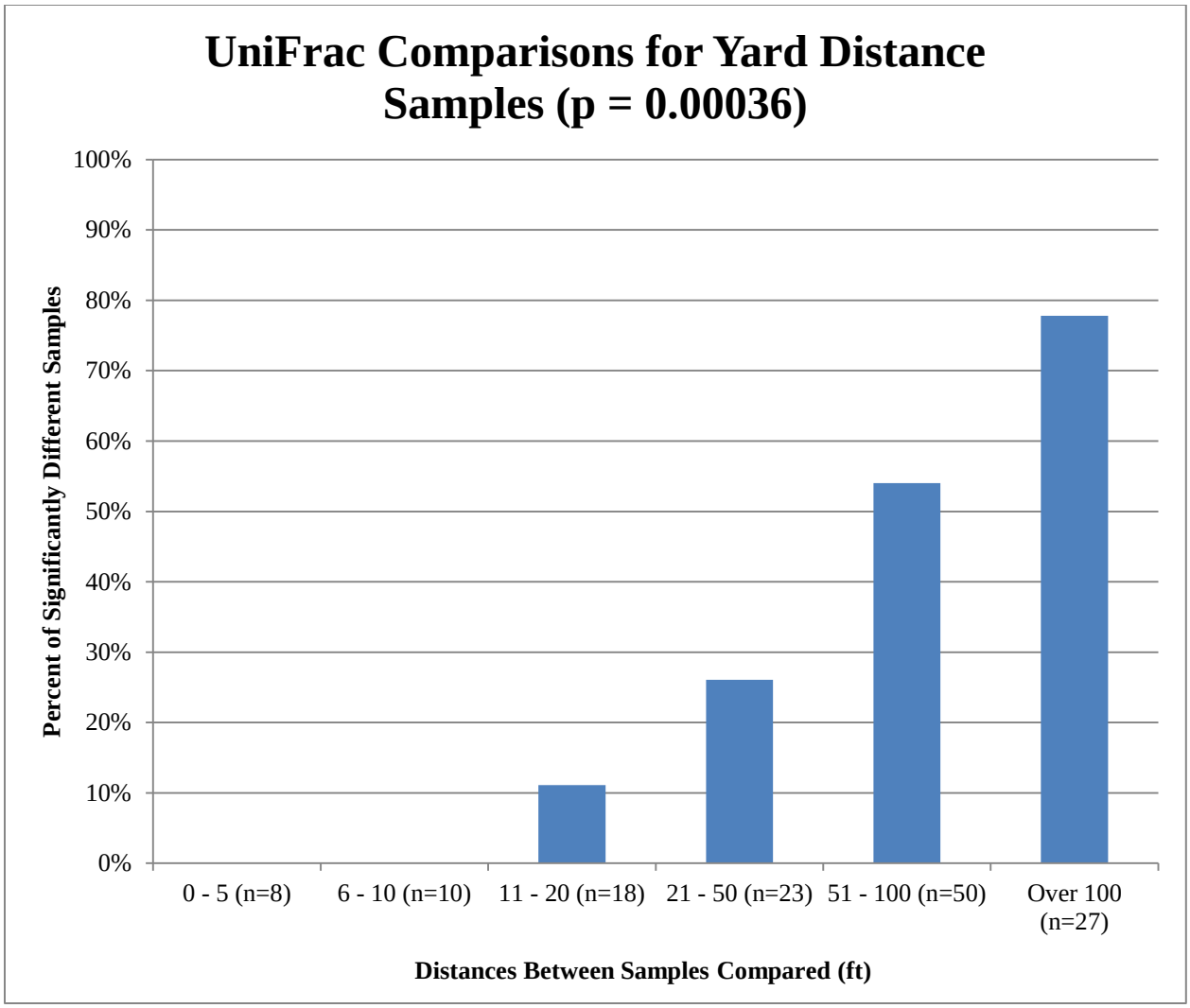

Figure 41. UniFrac comparisons for yard distance soil samples show statistical differences for distances greater than 10 feet with the percentage of significantly different samples rising to 80% for distances greater than 100 feet.

Deciduous Woods Samples

∫-LIBSHUFF comparisons of deciduous woods distance soils showed significant differences beginning at distances of five feet (Figure 42). The percent of different samples was above 70% for all distance ranges. Similarly, some samples differed at all distances in UniFrac comparisons for deciduous woods (Figure 43). The number of different samples was highest

(50%) in the 11 – 20 feet range and gradually declined to about 15% at distances greater than 100 feet.

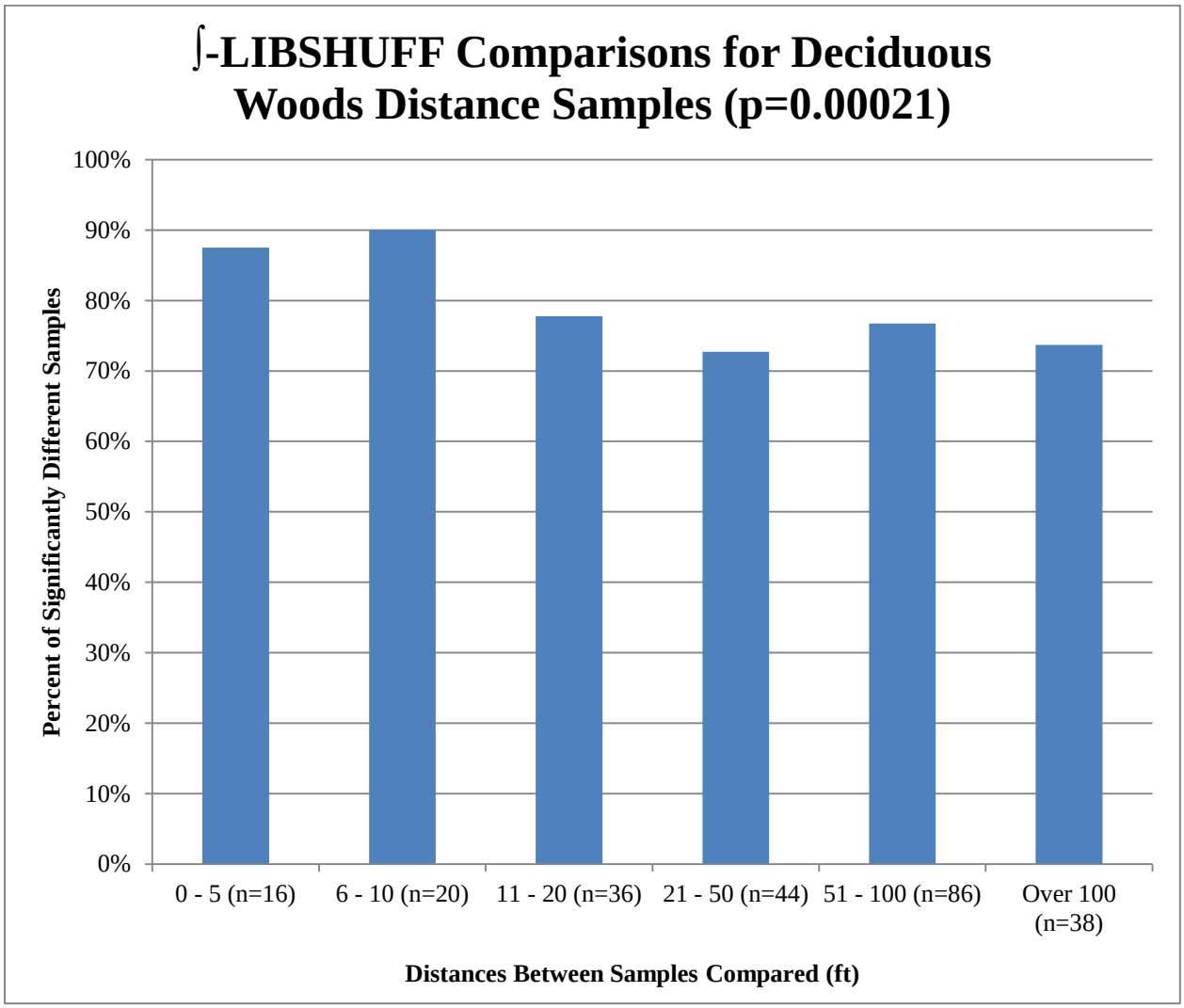

Figure 42. ∫-LIBSHUFF comparisons for deciduous woods distance soil samples show statistically significant differences for distances five feet and greater. The percent of different samples is above 70% for all distance ranges.

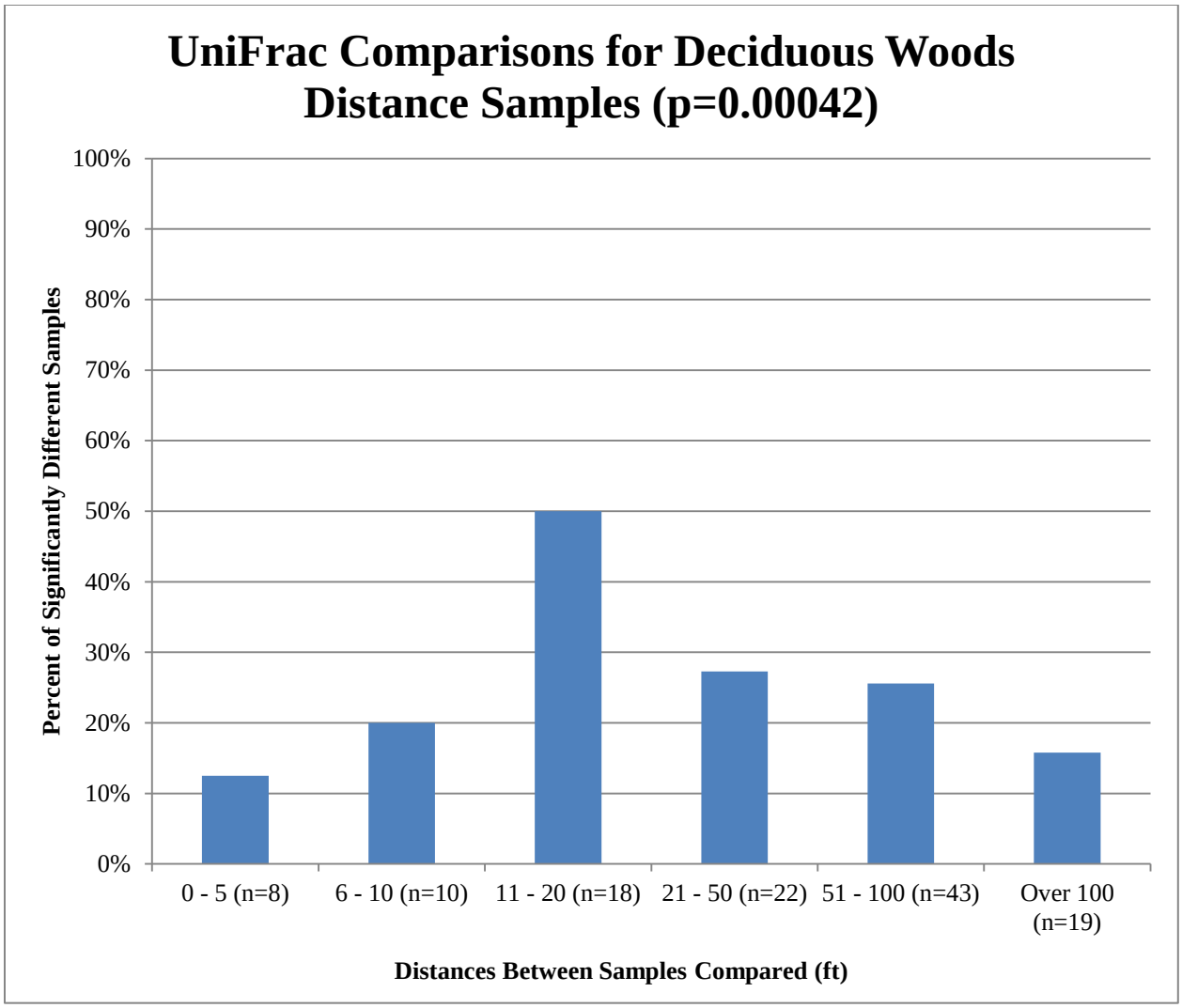

Figure 43. UniFrac comparisons for deciduous woods distance soil samples reveal statistical differences for all distances. The percent of different samples rose to 50% in the 11 – 20 feet range and gradually declined for larger distances.

Depth Samples Analysis

Nonmetric Multidimensional Scaling

Scree diagrams of depth samples for both diversity indices (Appendix 11 Figures 107 and 109) had high stress in one dimension with an elbow at two dimensions. Shepard diagrams (Appendix 11 Figures 108 and 110) showed a close association of distances and disparities. The surface and 1" samples clustered nearly on top of each other for final configurations of both

diversity indices (Figures 44 and 45). The BCDI plot also had these depths within standard error of the 2” sample. The 2” sample was also within standard error of the 5” and 10” samples, which fell on top of each other. The 20” and 36” were removed from the other samples. The SDC plot showed the 2” depth plotting close to the surface and 1” samples, though outside standard error. The 5” and 10” samples were the next closest to the cluster, followed by the 20” and 36” depths; however, all fell outside standard error of all other samples.

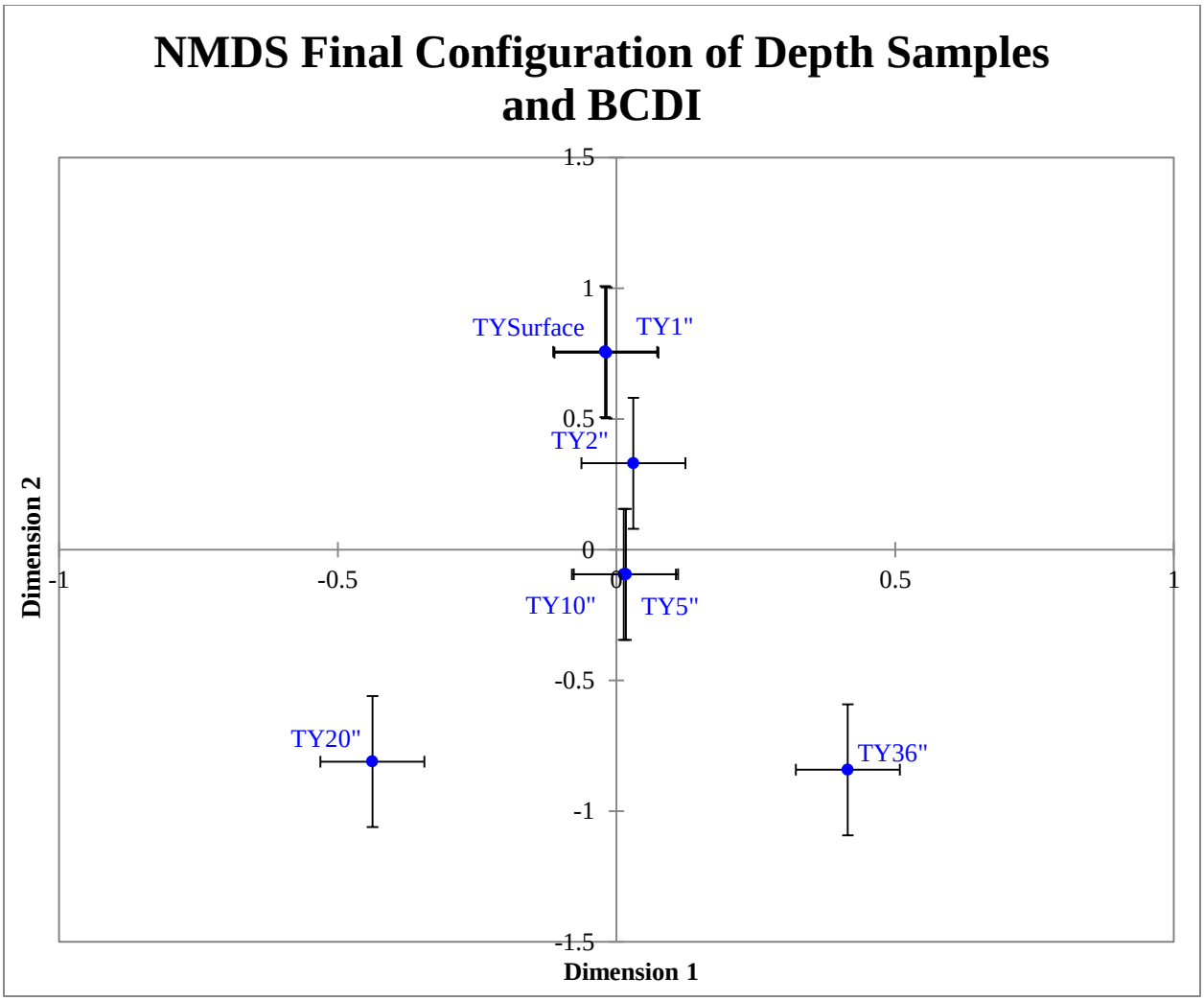

Figure 44. NMDS final configuration of depth soil samples from BCDI. The surface and 1" samples plot almost on top of each other and are within standard error of the 2" sample. The 5" and 10" samples, which also fall perfectly on top of each other, are within standard error of the 2" sample. The 20" and 36" samples plot further away from the rest of the samples and outside standard error. See Table 4 for site names corresponding to abbreviations.

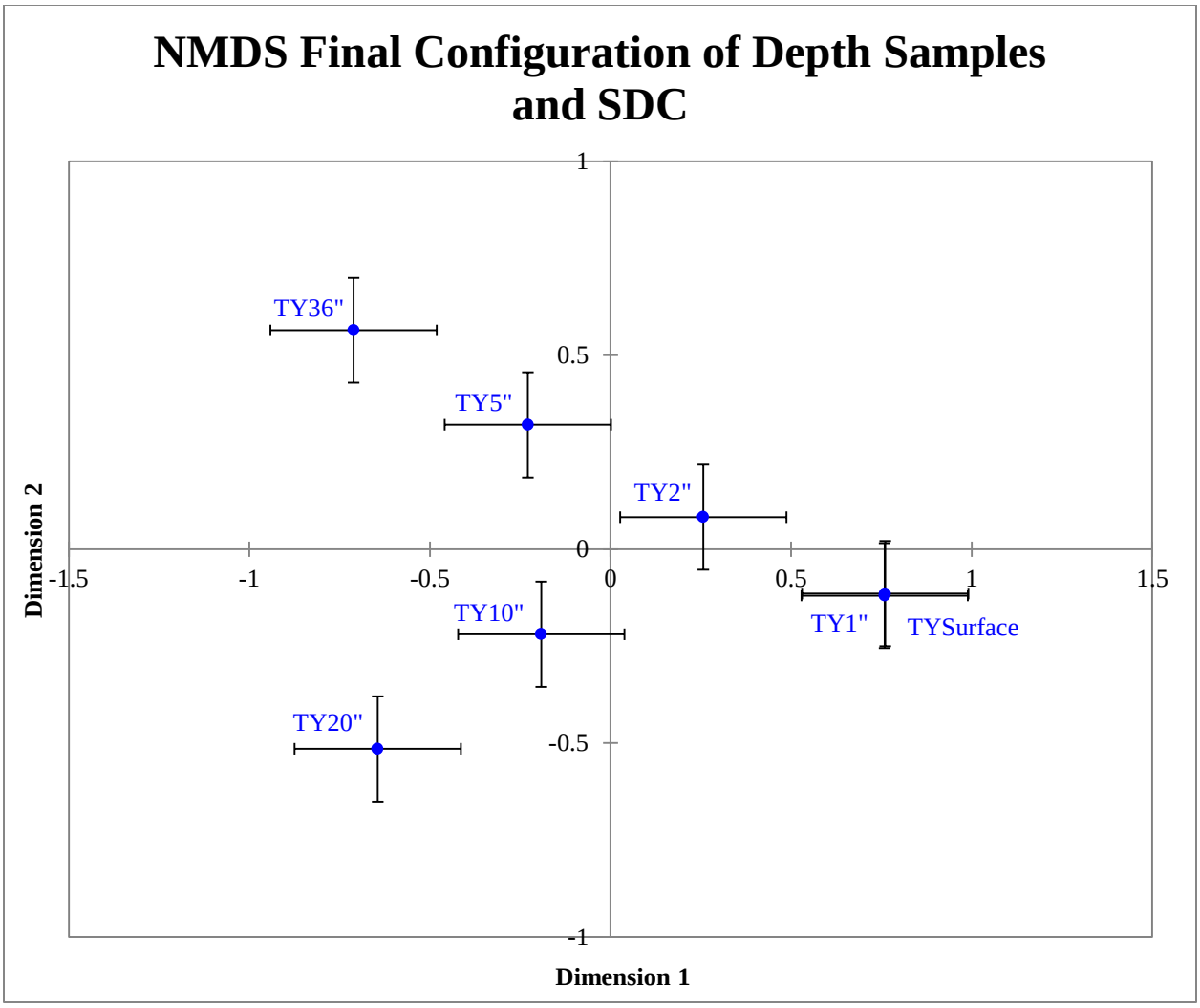

Figure 45. NMDS final configuration of depth soil samples from SDC. The surface and 1" samples fall almost perfectly on top of each other in two dimensions. The remaining samples fall outside of standard error of all other samples; however, the 2", 5", and 10" are closer to the surface and 1" cluster than are the 20" and 36" samples. See Table 4 for site names corresponding to abbreviations.

Hierarchical Cluster Analysis

Dendrograms developed with single linkage clustering for both diversity indices were similar (Figure 46 and Appendix 12 Figure 113) though the dissimilarities where clusters formed were different. Both had two clear clusters of the surface, 1", and 2" samples and 5" and 10" depths at either 0.532 or 0.631. The 20" and 36" samples joined the two groups at a dissimilarity

of 0.617 or higher. The BCDI complete linkage dendrogram was similar to the UPGMA dendrograms for both indices (Appendix 12 Figures 111, 112, and 115), though like before the dissimilarities where clusters formed were different. In these dendrograms three clusters were present. The first was a two member group formed by the surface and 1" samples between 0.468 – 0.576. This cluster was joined by a three member group of the 5", 10", and then 2" depths between 0.626 – 0.694. The final grouping was the 20" and 36" depths which clustered with the other two at 0.724 or higher. Finally, the SDC complete linkage dendrogram (Appendix 12 Figure 114) showed two clusters. The first was made up of the surface and 1" depths at 0.576. The other was formed initially at 0.632 by the 5" and 10" depths followed by the 2", 20", and 36" samples. The two are grouped together at a dissimilarity of 0.796.

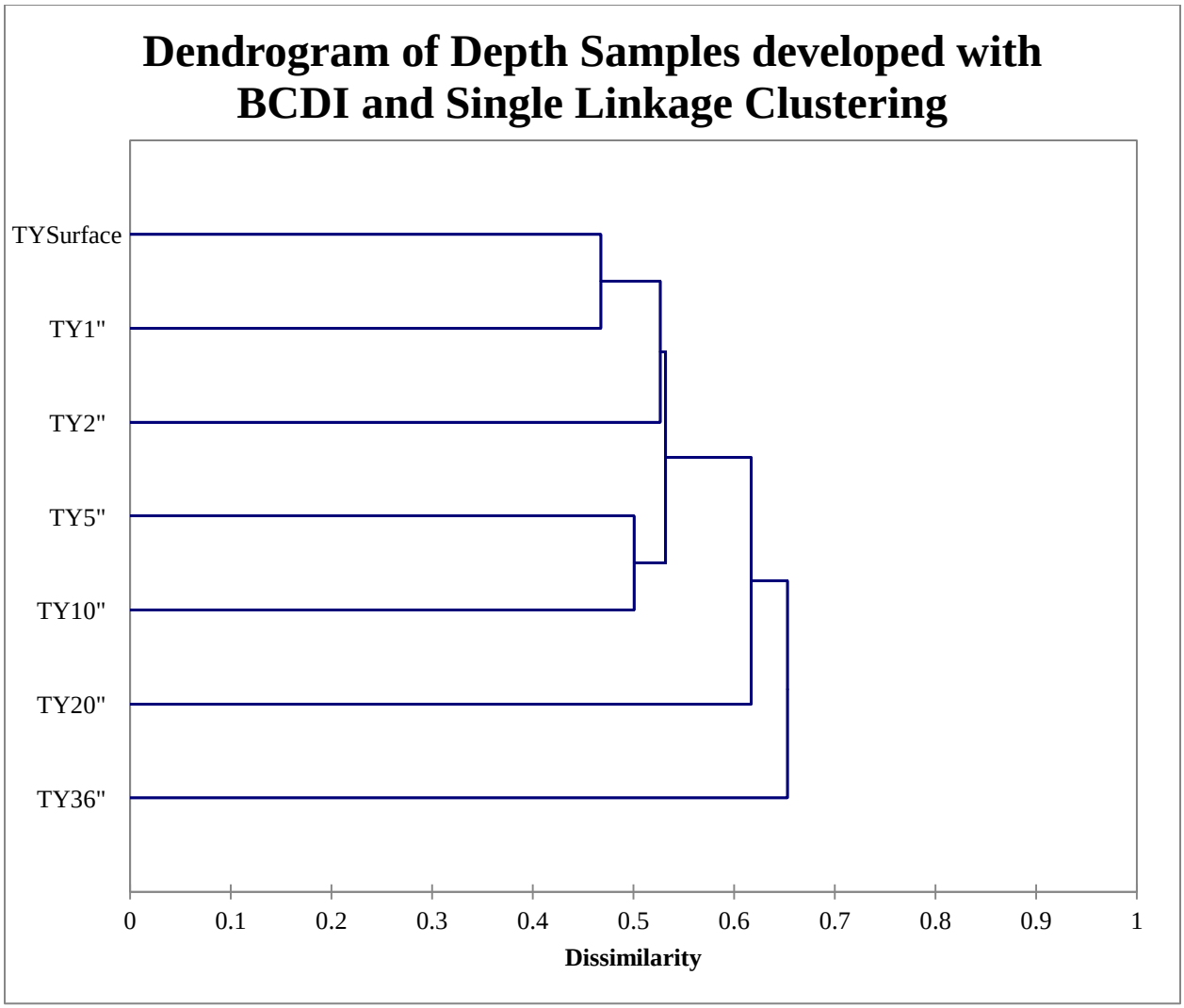

Figure 46. Dendrogram of depth soil samples developed with BCDI and single linkage clustering. Two clusters are present. The first is a three member group formed first by the surface and 1" samples followed by the 2" depth at 0.527. This cluster is joined by a two member group of the 5" and 10" depths at 0.532. The 20" and 36" depths group with the rest at 0.617 or greater. See Table 4 for site names corresponding to abbreviations.

Pairwise Comparisons

∫-LIBSHUFF revealed all pairwise comparisons for depth samples to be significantly different except for the 5" and 10" depths (corrected p=0.0012). The surface and 1" and 5" and 10" samples were not significantly different with UniFrac (corrected p=0.0024).

Time Series Samples Analysis

Nonmetric Multidimensional Scaling

Scree diagrams of seasonal samples from the marsh edge, yard, and deciduous woods for both BCDI and SDC (Appendix 13 Figures 116 and 117) had no elbow in the curve though all dimensions were below random stress. Two dimensional final configurations were analyzed to remain consistent with other studies. Additionally, distances and disparities did not associate well in the Shepard diagrams (Figure 47 and Appendix 13 Figure 118) agreeing with the higher stress at that dimension for both indices.

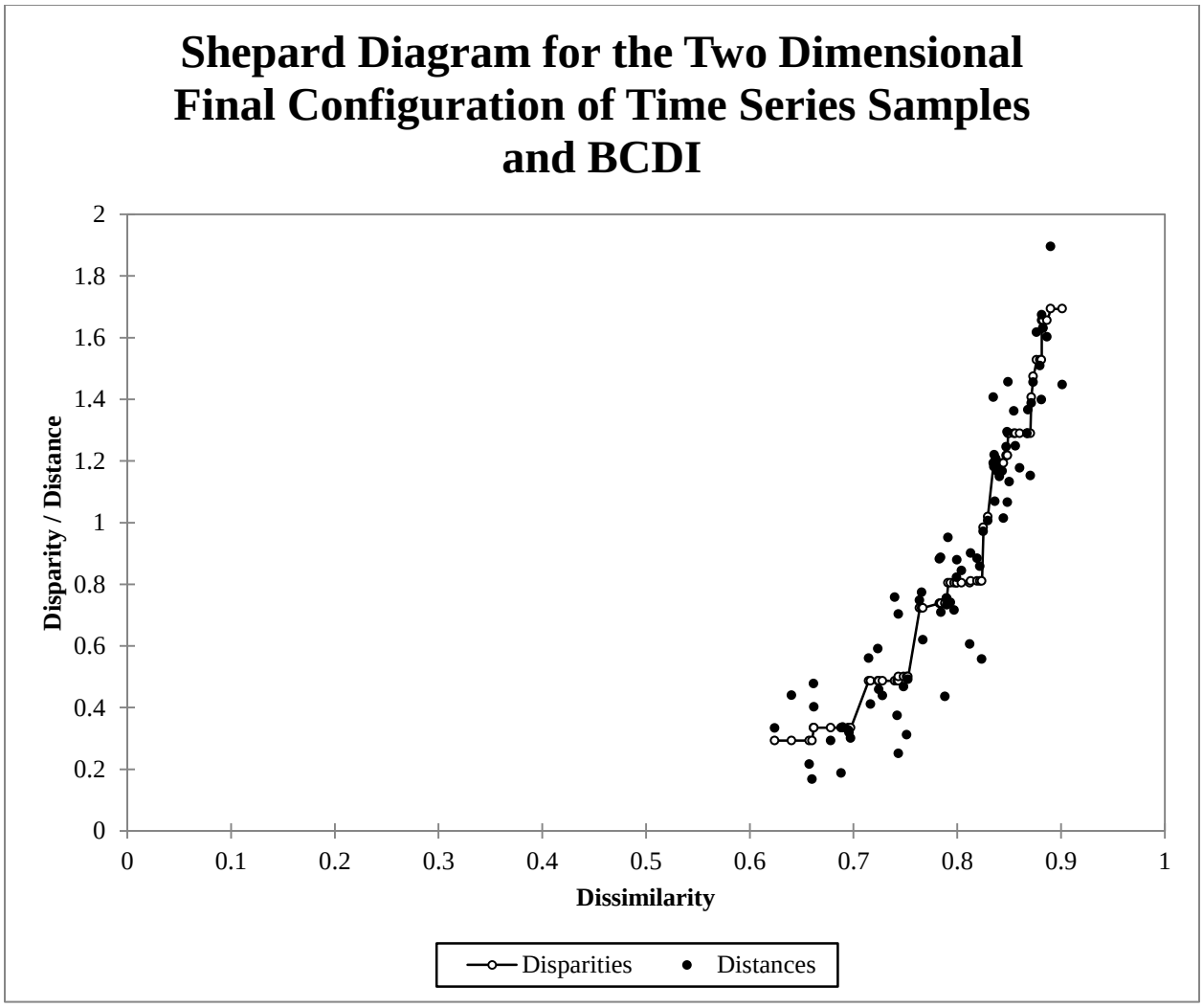

Figure 47. Shepard diagram for the two dimensional final configuration developed from BCDI of time series soil samples. Distances do not associate well with their corresponding disparities, agreeing with the higher stress in the Scree diagram.

The final configuration for time series samples developed from BCDI (Figure 48) had no samples that clustered closely by habitat; however, the marsh edge samples were removed from the others in quadrant four. The August and November marsh edge samples were close, though not tightly clustered, while the May sample plotted equidistant from both. The marsh edge February sample was near the other marsh edge samples but the least associated. The deciduous woods and yard samples intermingled in quadrants one and two. The yard May and deciduous

woods February samples plotted the closest in quadrant one and weakly clustered with the yard February sample. The remaining samples formed a loose cluster in the second quadrant. The deciduous woods May and August samples were close as were the May and yard August samples. The yard November associated with the Ymix August sample as well as the yard August sample. The deciduous woods November sample was furthest from all samples in quadrant two.

The final configuration for SDC dissimilarities (Figure 49) had three loose clusters formed by habitat. The marsh edge May sample was closer to the February and November samples, though the August sample plotted closely with them. The deciduous woods November sample was closest to the deciduous woods February and May samples, but the August sample associated with them as well. Finally, the yard August and Ymix August samples plotted closely, as did the August and February samples. The May and November samples were the furthest from the other yard time points.

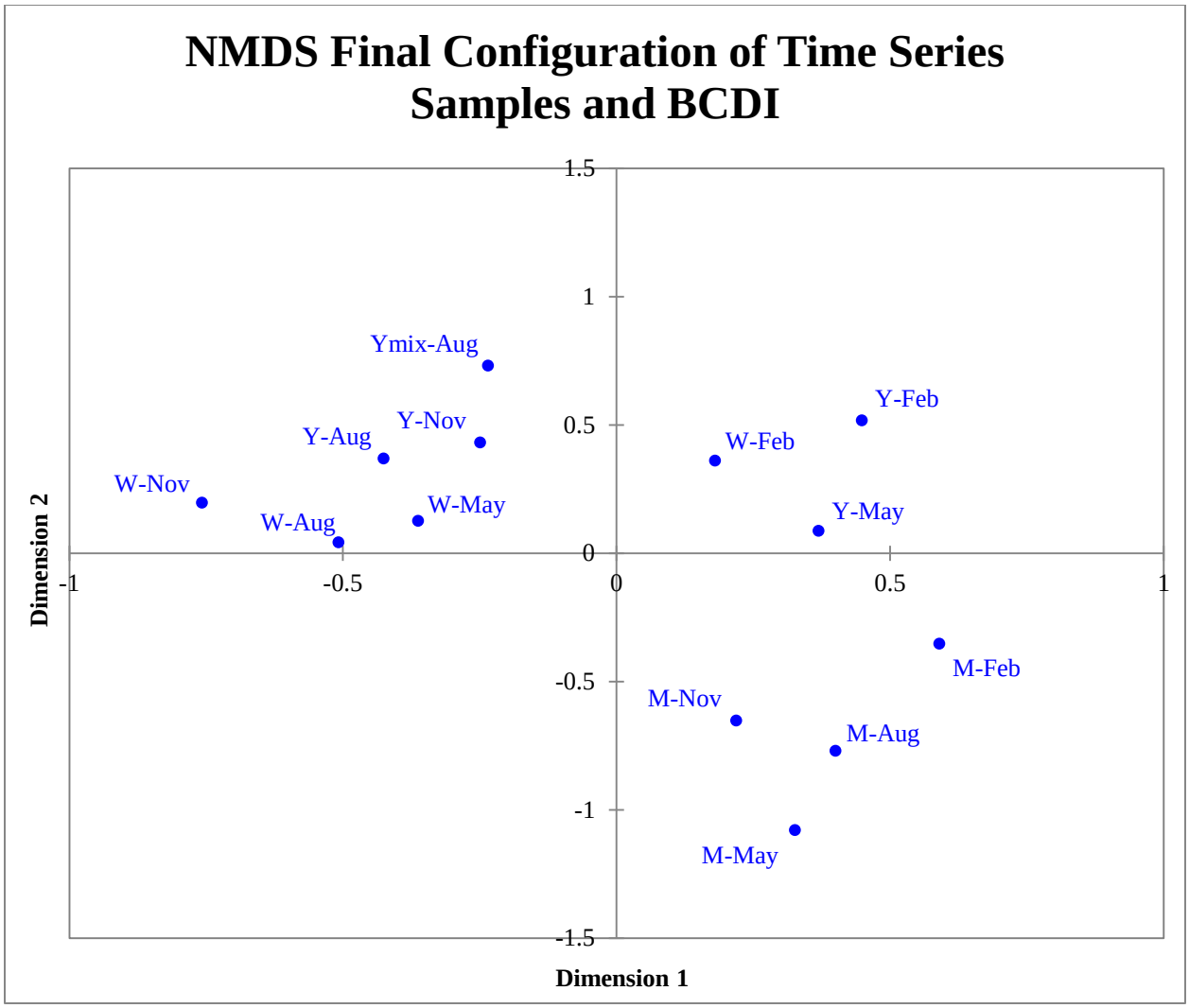

Figure 48. NMDS final configuration of time series soil samples from BCDI. Samples do not cluster tightly by habitat; however, the marsh edge samples inhabit a single quadrant. The August and November samples are close, though not well clustered. The May sample plots near the November and August samples while the February sample is associated with the other marsh edge samples but is the furthest away. The deciduous woods and yard samples intermingle in quadrants one and two. The yard May and deciduous woods February samples are loosely associated with the yard February sample. The remaining samples form a loose cluster in the second quadrant. The deciduous woods May and August samples are associated as are the May and yard August samples. The deciduous woods November sample is the furthest from all samples in the second quadrant. The yard November associated with both the Ymix August and yard August samples. See Table 2 for site names corresponding to abbreviations.

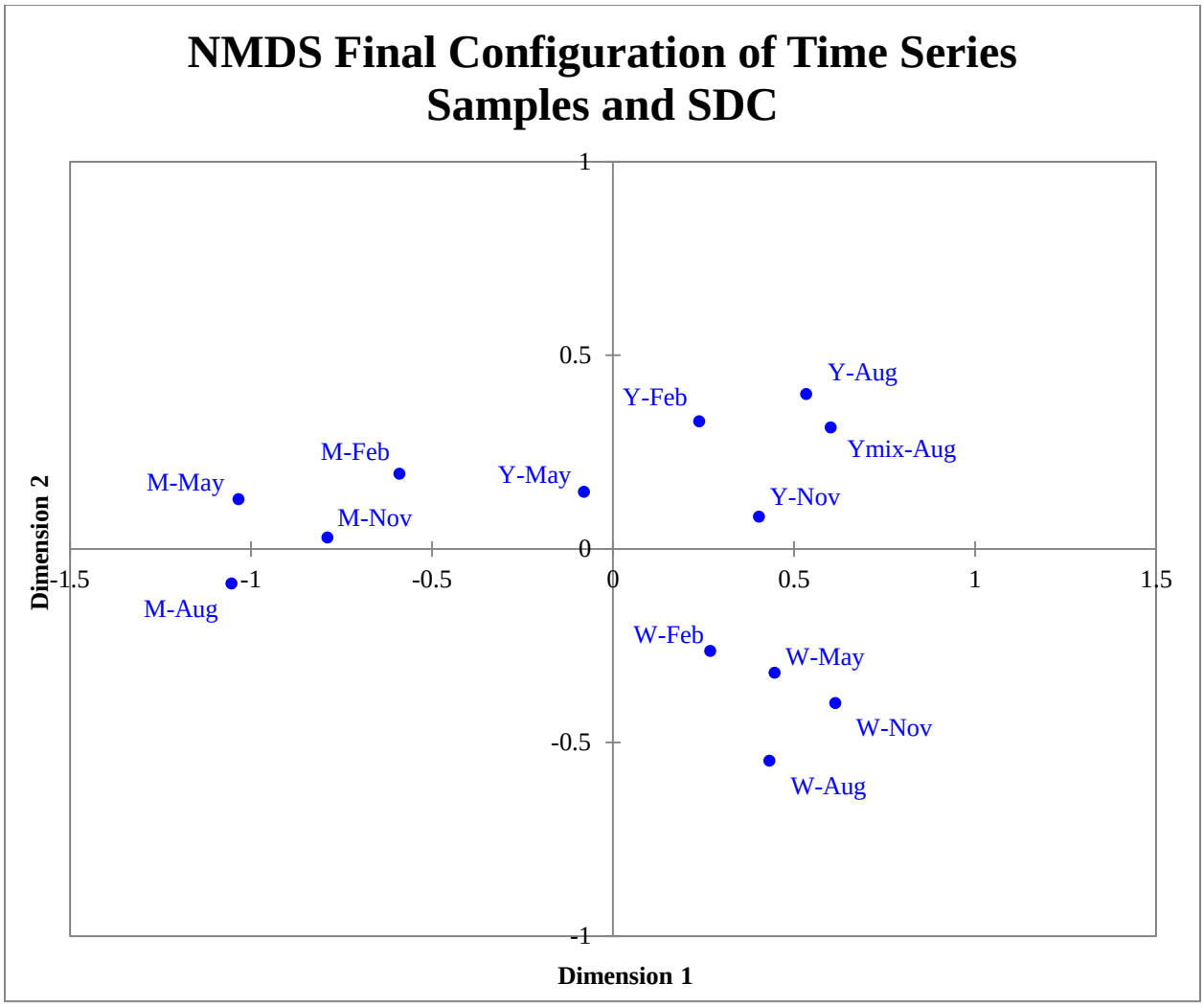

Figure 49. NMDS final configuration of time series soil samples from SDC. Three loose clusters are formed by habitat. The marsh edge May sample was closer to the February and November samples, though the August sample plots closely with them. The deciduous woods samples also plot closely where the May sample is associated with the November and February samples. The August sample is the furthest from the other deciduous woods samples. Finally, the yard August and Ymix August samples plot closely as do the August and February samples. The May and November samples are the furthest from the other yard time points. See Table 2 for site names corresponding to abbreviations.

Hierarchical Cluster Analysis

The dendrogram developed with single linkage clustering and BCDI (Appendix 14 Figure 119) had four clusters. The first was a two member group at 0.690 of the yard August and November samples while the second was composed of the deciduous woods May and August

samples followed by the November time point at 0.662. The third cluster was another two member grouping of the yard May and deciduous woods February samples at 0.624. The final cluster at 0.662 was comprised of all the marsh edge samples where the November and May samples were the most similar followed by August then February. The Ymix August sample clustered with the first two groups at 0.697 while the yard February sample clustered with the first three groups at 0.728. The marsh edge cluster grouped with the rest of the samples at 0.753.

The dendrogram developed with single linkage clustering and SDC (Appendix 14 Figure 122) had three clusters differentiated by habitat. The first at 0.749 contained the deciduous woods November and May samples clustered followed by the August and February ones. The second had the yard November and February samples being more similar than the May, August, and Ymix August time points also at 0.749. The third cluster at 0.733 was comprised of the marsh edge November and February samples followed by the May then August ones. The first and second clusters grouped at 0.749 with the third at 0.771.

The complete linkage and UPGMA clustering dendrograms for BCDI (Appendix 14 Figures 120 and 121) were similar. Four clusters were present in both; however, the dissimilarities where they clustered were different though close. The first consisted of the yard November and August time points followed by the Ymix August sample at either 0.717 or 0.707. The second cluster at 0.678 or 0.670 had the deciduous woods May and August samples grouped followed by the November one. The third grouping was comprised of the yard May and deciduous woods February time points followed by yard February at either 0.752 or 0.740. The final cluster at 0.766 or 0.717 was comprised of all the marsh edge samples where the November and May samples were the most similar followed by August then February. All clusters were joined at a dissimilarity of 0.901 for complete linkage and 0.849 for UPGMA.

Four clusters were present in the SDC complete linkage dendrogram (Appendix 14 Figure 123), three of which contained samples from only one habitat. The first at 0.757 was a two member group composed of the yard May and deciduous woods February samples. The second had the deciduous woods November and May time points clustered followed by the August sample at 0.743. The first and second clusters joined at a dissimilarity of 0.831. The third cluster was two two member groups: the yard August and Ymix August samples and the yard November and February samples, respectively. These joined together at 0.776 and then clustered with the first and second groups at 0.837. The final cluster was another formed of two two member groups: the marsh edge May and August samples and the marsh edge November and February samples. They joined together at 0.775 then with the others at 0.900.

Finally, the SDC UPGMA clustering dendrogram (Figure 50) showed three clusters formed by habitat. The first was a two member cluster, the Ymix August and yard August samples, grouped with a three member cluster including the yard November, February, and May time points at 0.800. The second at 0.763 contained the deciduous woods November and May samples clustered first followed by August and February. The third was comprised of the marsh edge November and February samples followed by May then August at 0.753. The first and second clusters grouped at 0.800 with the third at 0.861.

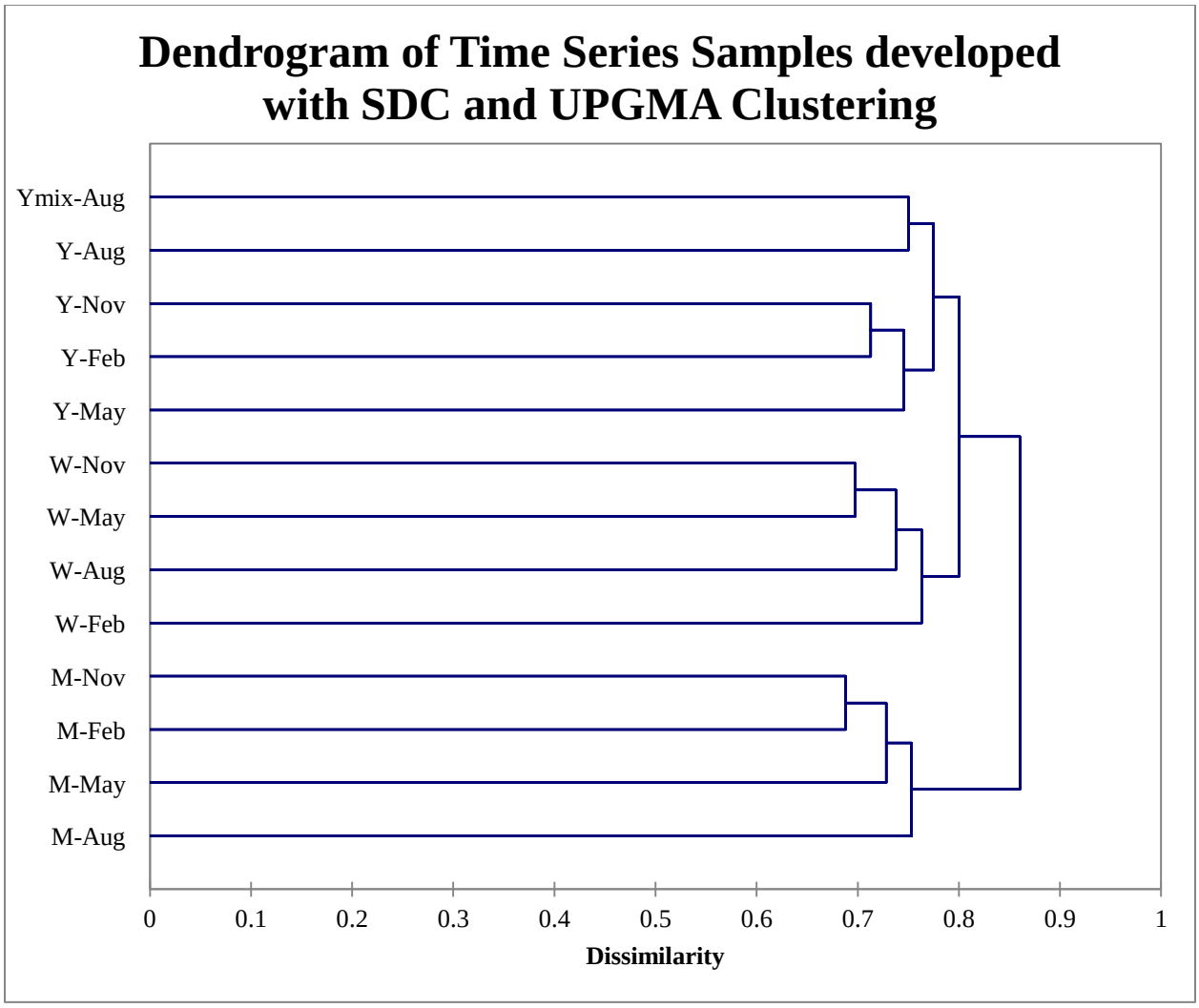

Figure 50. Dendrogram of time series soil samples developed with SDC and UPGMA clustering. Three clusters are formed by habitat at 0.800, 0.763, and 0.753. The first cluster is formed by a two member cluster, the Ymix August and yard August samples, grouping with a three member cluster including the yard November, February, and May time points. The second contains the deciduous woods November and May samples clustering first followed by the August and February ones. The third comprises the marsh edge November and February samples followed by the May then August ones. The first and second clusters group at 0.800 with the third at 0.861. See Table 2 for site names corresponding to abbreviations.

Pairwise Comparisons

Time series samples differed significantly within and between habitats using ∫-LIBSHUFF (corrected p=0.00032) (Figure 51). Half of the marsh edge samples were statistically different when compared to other marsh edge samples. The May and August time points differed

significantly from the February one as did the May and August samples. All of the marsh edge samples were statistically different from the deciduous woods and yard time points. Half of within deciduous woods samples were different including all comparisons against the February sample. Eighty-five percent of the deciduous woods and yard samples were statistically different; the three that were not included the deciduous woods May sample and the Ymix August or yard November, as well as the deciduous woods November and Ymix August. Finally, 70% of yard samples were significantly different were the August and Ymix August or November as well as Ymix August and November were not.

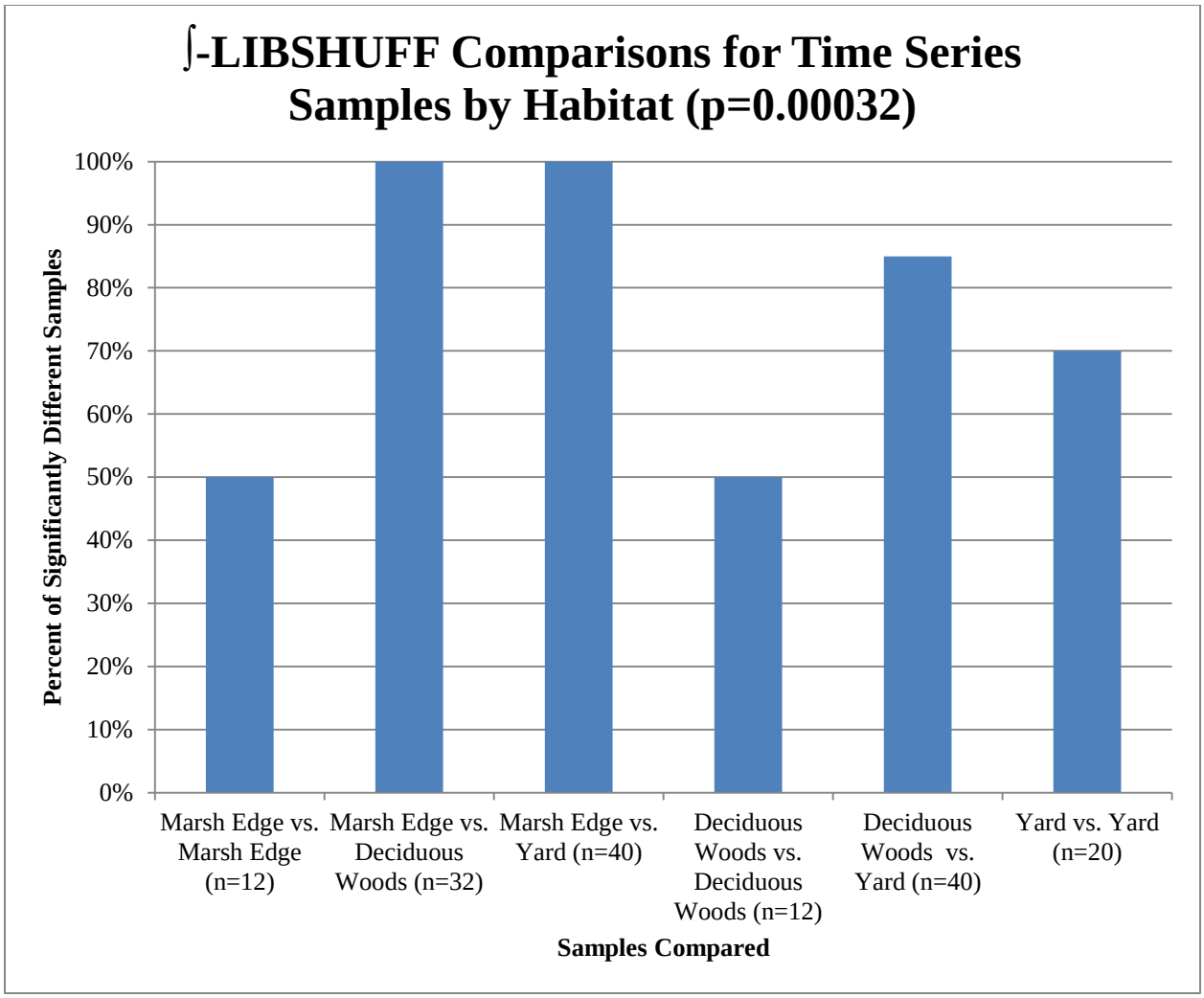

Figure 51. ∫-LIBSHUFF comparisons for time series soil samples were statistically different within and between habitats. Half of the marsh edge samples differed statistically when compared to other marsh edge samples where the May and August time points differed from the February one as did May and August. All marsh edge samples were significantly different from the deciduous woods and yard soils. Similarly, half of deciduous woods samples were different from each other, where the February sample was statistically different from the rest. The deciduous woods and yard samples differed 85% of the time where the three that did not include the May sample and the Ymix August or yard November, as well as the November and Ymix August. Finally, 70% of yard samples were significantly different. The samples that did not differ were the August and November ones.

UniFrac comparisons for time series samples also revealed statistical differences within and between habitats (corrected p=0.00064) (Figure 52). Half of the within marsh edge samples were statistically different, including the February and May or August as well as May and

August comparisons. All marsh edge and deciduous woods samples differed significantly while 85% of marsh edge and yard did as well. The marsh edge February and yard February or May and marsh edge November and yard May samples were not statistically different. Sixty-seven percent of within deciduous woods samples were different. All time points compared to the August sample differed statistically as did February and November. Further, 60% of the deciduous woods and yard samples were statistically different. The samples that were not included deciduous woods August and yard August or Ymix August, deciduous woods February and yard February, May, or November, deciduous woods May and yard May or November, as well as deciduous woods November and yard November. Finally, 60% of within habitat yard samples were significantly different which included the August and February, May, or November as well as Ymix August and February, May, or November time points.

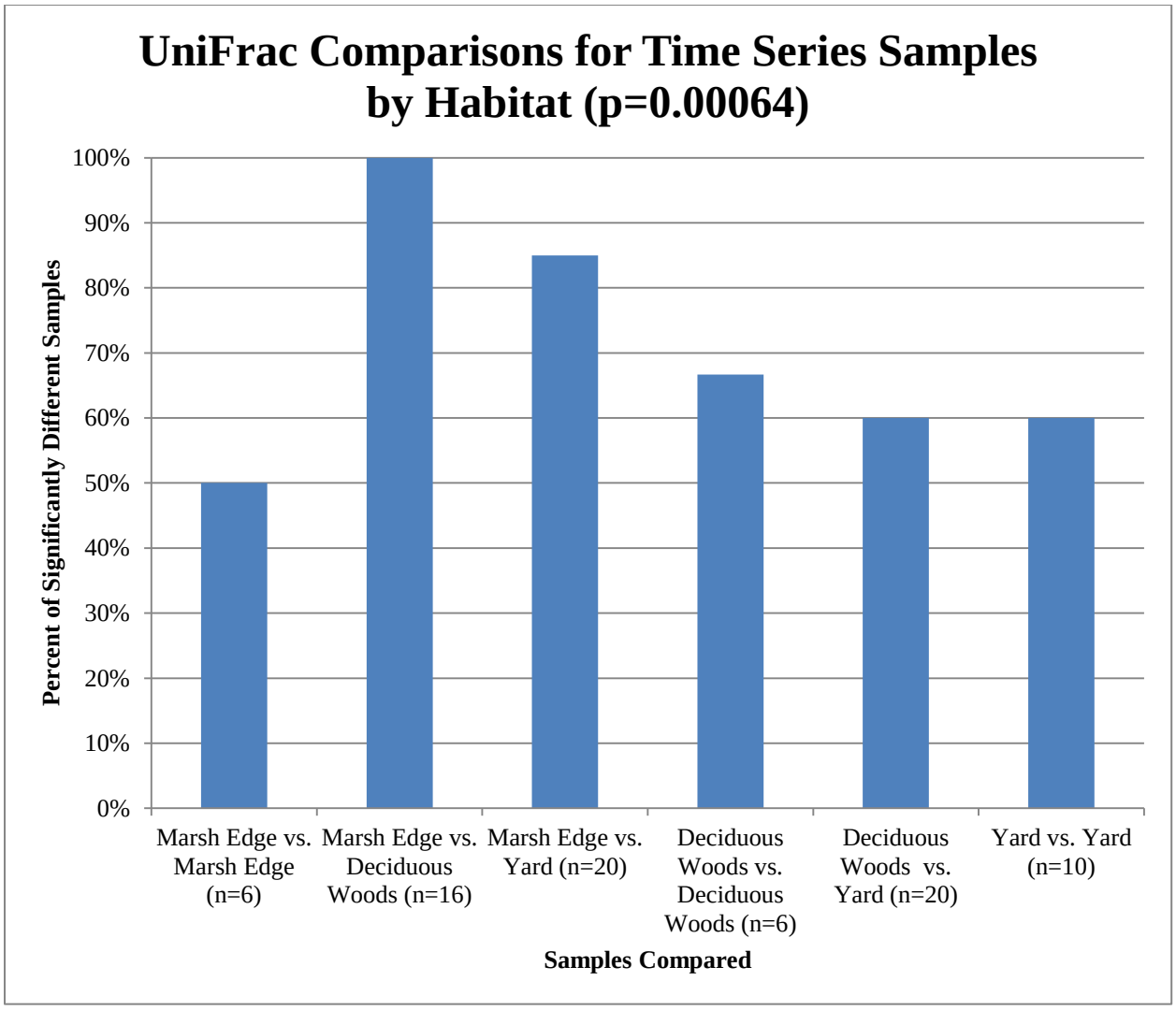

Figure 52. UniFrac comparisons for the time series soil samples showed statistical differences within and between habitats. Half of the within marsh edge samples were statistically different. Those that were different include the February and May or August as well as May and August time points. All marsh edge and deciduous woods samples were different while 85% of marsh edge and yard ones were as well. The marsh edge February and yard February or May samples as well as the marsh edge November and yard May samples were not different. Deciduous woods samples differed 67% of the time. All time points were different from the August sample as were the February and November ones. Sixty percent of the deciduous woods and yard samples were different. The samples that were not different include the deciduous woods August and yard August or Ymix August, deciduous woods February and yard February, May, or November, deciduous woods May and yard May or November, as well as deciduous woods November and yard November. Finally, 60% of within yard samples were significantly different. The August samples were different from all other months but not each other.

Similar Habitat Samples Analysis

Nonmetric Multidimensional Scaling

The Scree diagrams of yard samples for both BCDI and SDC (Appendix 15 Figures 124 and 126) had a high stress in one dimension that decreased (elbow) into higher dimensions. Shepard diagrams (Appendix 15 Figures 125 and 127) showed close association of most distances and disparities for the two dimensional configurations agreeing with the low stress in the Scree diagrams. Neither final configuration showed a trend of geographically closer yards plotting with each other. The two dimensional plot of all ten samples from BCDI (Figure 53) had multiple sets of samples clustering, though samples were not closely associated. Four pairs of yard samples were close: the Perry and Michigan State University west, the Michigan State University main and east, the Michigan State University main and Lisa, and the Michigan State University west and north. The remaining samples were not associated with any others. The configuration developed from SDC (Figure 54) had a different clustering of yards. Again four pairs fell close: the Perry and Michigan State University west, the Lisa and Michigan State University west, the Michigan State University west and east, and the Perry and Michigan State University north samples. Remaining samples plotted further away from the others.

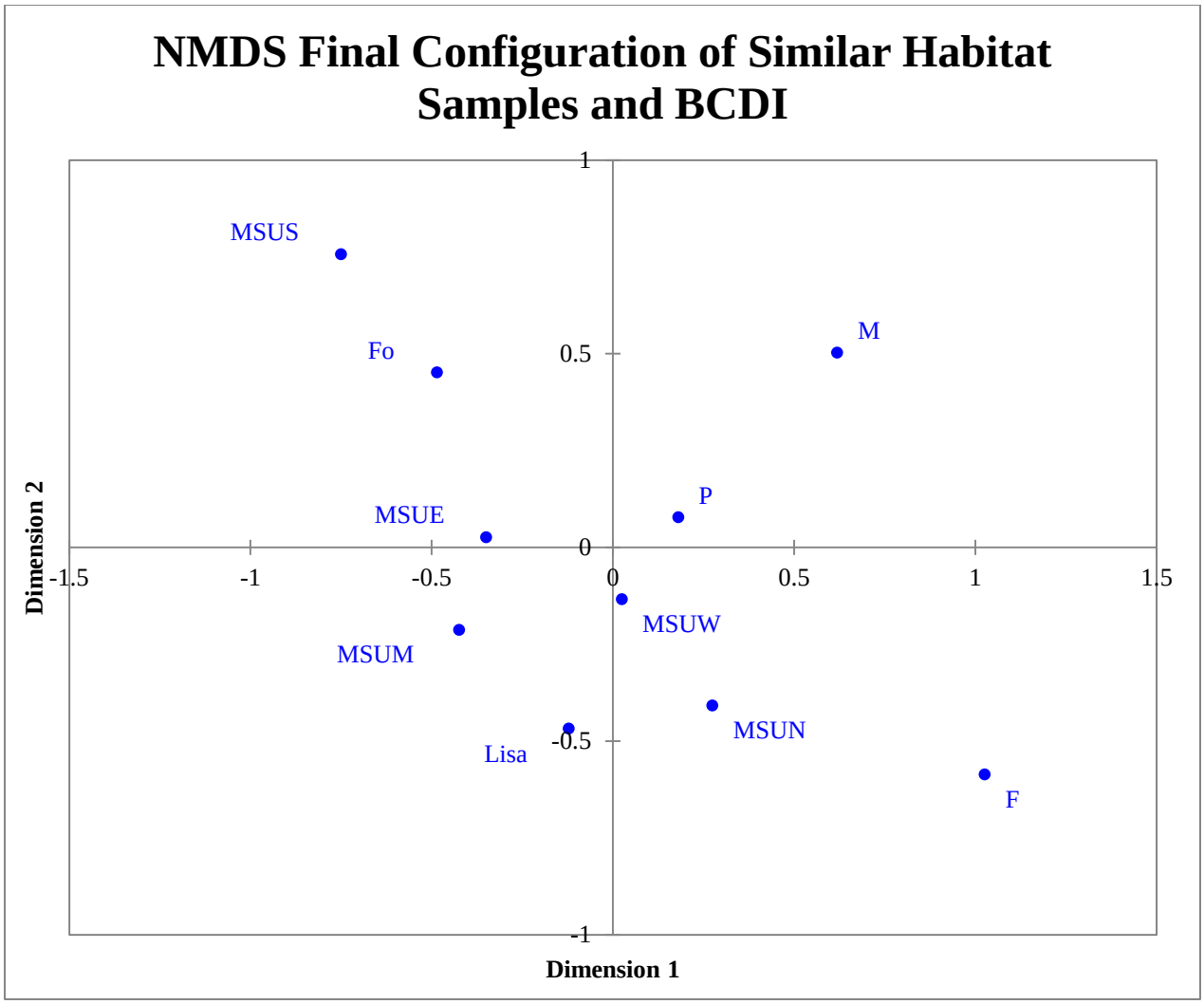

Figure 53. NMDS final configuration of similar habitat (yard) soil samples from BCDI. Multiple sets of samples fall close to each other. They include the Perry yard and Michigan State University west yard, the Michigan State main and east yards, the Michigan State University main and Lisa yards, and the Michigan State University west and north samples. The remaining samples are not associated with any other samples. See Table 5 for site names corresponding to abbreviations.

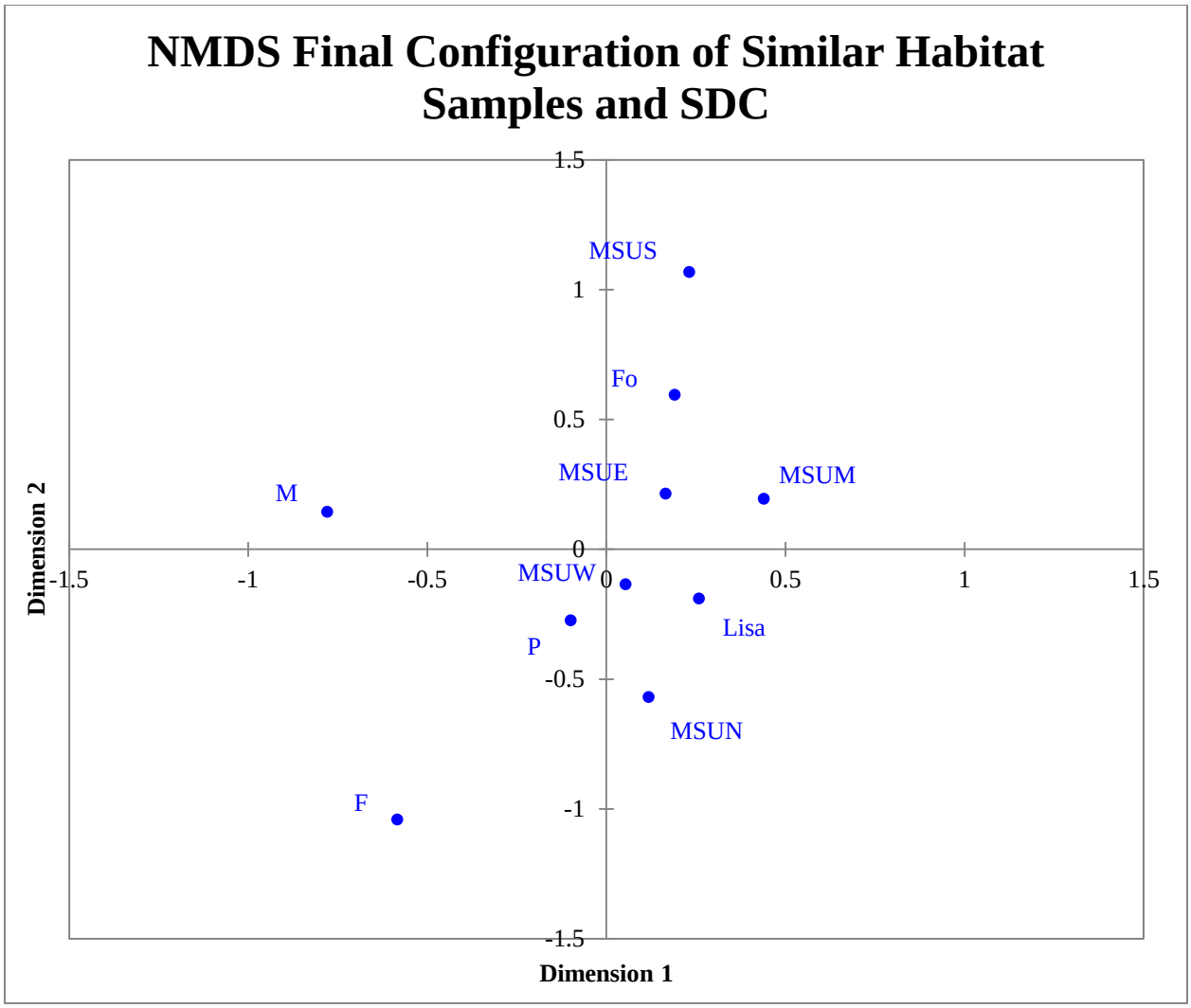

Figure 54. NMDS final configuration of similar habitat soil samples from SDC. Multiple sets of samples fall close to each other. They include the Perry and Michigan State University west yards, the Michigan State University west and Lisa yards, the Perry and Michigan State University north yards, and the Michigan State University west and east yards. The remaining samples are not associated with any other samples. See Table 5 for site names corresponding to abbreviations.

Hierarchical Cluster Analysis

The dendrograms of similar habitat soils developed with single linkage from BCDI and SDC (Appendix 16 Figures 128 and 130) were not consistent with each other. The BCDI dendrogram had two distinct three member groups at 0.619 and 0.608. The first had the

Michigan State University main and east samples clustered followed by the Lisa yard, while the second had the Michigan State University west and north yards grouped before the Perry sample. The remaining four samples clustered with the first two groups at a dissimilarity of 0.634, with the most dissimilar, the Michelle yard, grouped at 0.758. The SDC dendrogram also had two clusters at 0.676 and 0.650 though the membership was slightly different. The first was composed solely of the Michigan State University main and east yards while the other had the Perry and Michigan State University west samples clustered first followed by the north and Lisa yards. Similarly to the BCDI dendrogram, the remaining samples clustered with the two groups at 0.704 and greater, with the Michelle yard being the most dissimilar.

The dendrograms developed with complete linkage from BCDI and SDC (Figure 55 Appendix 16 Figure 131) were similar; however, some of the cluster orders were different. Three clusters were present, two three member and one two member. The first three member group at 0.645 or 0.715 was comprised of the Michigan State University main and east samples followed by the Lisa yard. The second was formed by the Michigan State University west and north then Perry yard for BCDI at 0.625. The SDC dendrogram had the west and Perry yards clustered followed by the north sample at 0.688. The final group at either 0.680 or 0.719 was composed of the Michigan State University south and Foran yards. The remaining samples clustered with the first two groups at a dissimilarity of 0.805 or greater followed by the final cluster at 0.905 or greater.

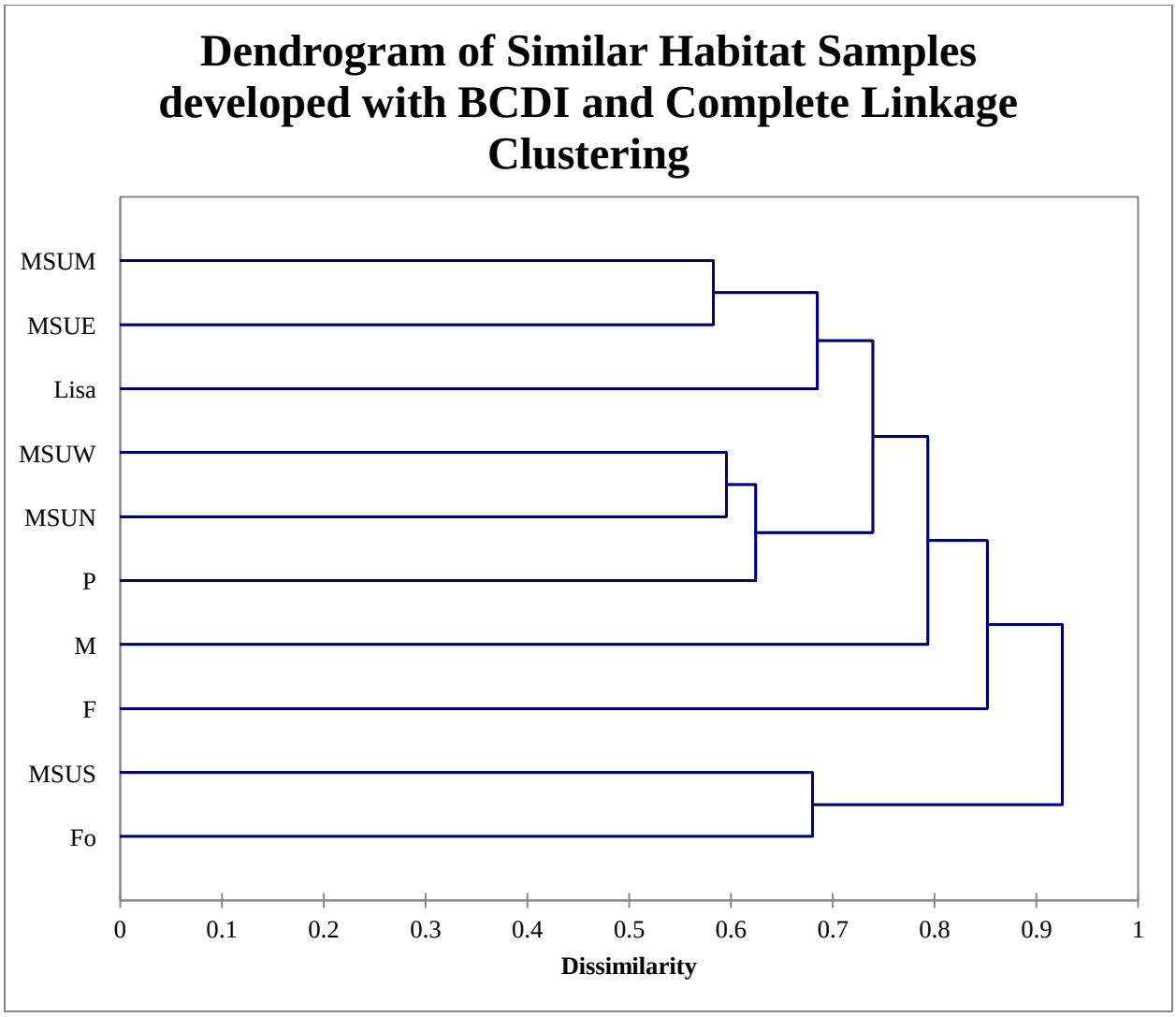

Figure 55. Dendrogram of similar habitat soil samples developed with BCDI and complete linkage clustering. Three clusters are formed at 0.645, 0.625, and 0.680. The first has the Michigan State University main and east samples being the most similar followed by the Lisa yard. The second is comprised of the Michigan State University west and north samples clustering first then the Perry yard. The final cluster contains the Michigan State University south and Foran yards. The two remaining samples cluster with the first two groups at a dissimilarity of 0.793. The third cluster groups with the rest at 0.925. See Table 5 for site names corresponding to abbreviations.

The dendrograms for BCDI and SDC developed with UPGMA (Appendix 16 Figures 129 and 132) showed differences in the groups formed. The BCDI dendrogram had three clusters at 0.652, 0.616, and 0.680. The first contained the Michigan State University main and east samples

then the Lisa yard. The second consisted of the Michigan State University west and north yards followed by the Perry sample. The third cluster was formed by the Michigan State University south and Foran yards. The Michelle yard clustered with the first two groups at a dissimilarity of 0.772, closely followed by the third cluster. The Fenner yard grouped with the rest at 0.822. The SDC dendrogram had two distinct clusters at 0.718 and 0.697. The first was formed by the Michigan State University main and east samples then the Foran yard. The second cluster had the Perry and Michigan State University west samples being the most similar followed by the Michigan State University north and Lisa yards. The three remaining samples clustered with the two groups at 0.794 or greater.

Pairwise Comparisons

All pairwise comparisons of similar habitat samples for both ∫-LIBSHUFF and UniFrac differed statistically except for the Lisa and Michigan State University east comparison (corrected p=0.00056 and p=0.0011 respectively). These samples were, however, borderline significant (p=0.0007 and p=0.0015 respectively).

Diverse Habitat Samples Analysis

Nonmetric Multidimensional Scaling

The Scree diagrams (Appendix 17 Figures 133, 134, and 137) for diverse habitat samples had high stress in one dimension with a decrease (elbow) at two and higher dimensions. Shepard diagrams for both BCDI and SDC (Appendix 17 Figures 135, 136, and 138) had close association of distances and disparities affirming the low stress seen at that dimension. When all ten samples were plotted in two dimensions from BCDI (Figure 56) the dirt road sample oriented far from all other samples in quadrant two; while, the remaining samples formed a tight cluster in

quadrant four. The same samples plotted very differently when the first and forth dimensions were examined together (Figure 57). The yard, deciduous woods, and field samples were associated as were the marsh edge, fallow agricultural field, and corn agricultural field. The coniferous forest and roadside plotted closest in quadrant one and associated with the Lake Lansing beach sample. The dirt road was the most dissimilar, plotting the furthest from all other samples. Configurations when samples were plotted into eight dimensions had erroneous stress plots and were not considered. Analyzing the samples with the dirt road excluded (Figure 58) resulted in a similar configuration to Figure 57 having the yard and deciduous woods samples clustered tightly, as were the marsh edge and fallow agricultural field. The yard and deciduous woods samples were also close to the field sample. The remaining samples fell further from all other samples. The configuration developed from SDC (Figure 59) of all samples showed similar clusters, with the marsh edge and fallow agricultural field associated as well as the yard and deciduous woods. Additionally, the corn agricultural field clustered with the marsh edge and fallow agricultural field samples while the field clustered with the yard and deciduous woods samples. The field sample was associated with the corn agricultural field like the yard and deciduous woods samples were with the fallow agricultural field. All remaining samples were further from all other samples.

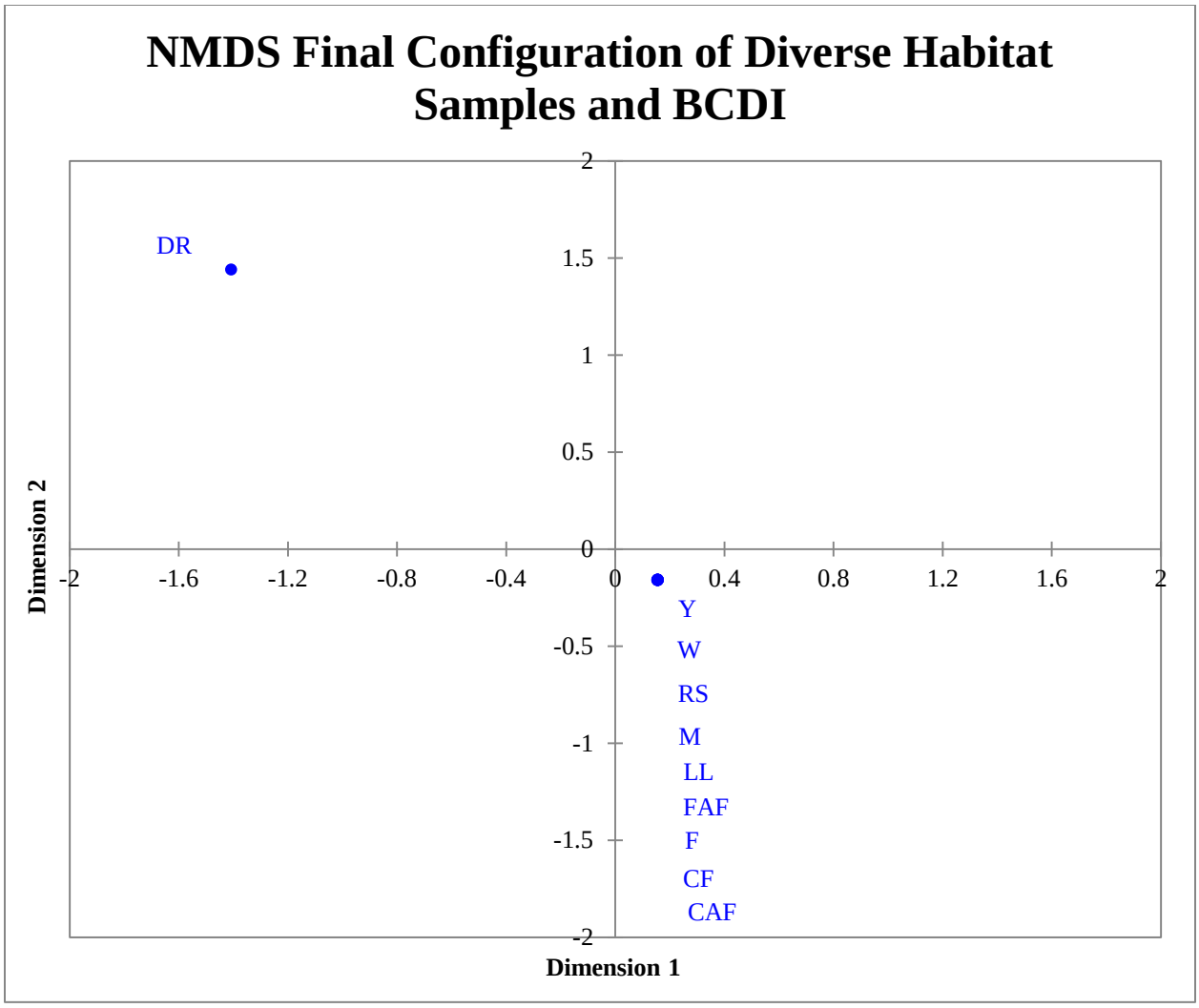

Figure 56. NMDS final configuration of diverse habitat soil samples from BCDI. All samples except the dirt road sample cluster closely in quadrant four. The dirt road sample falls far outside this tight clustering. See Table 6 for site names corresponding to abbreviations.

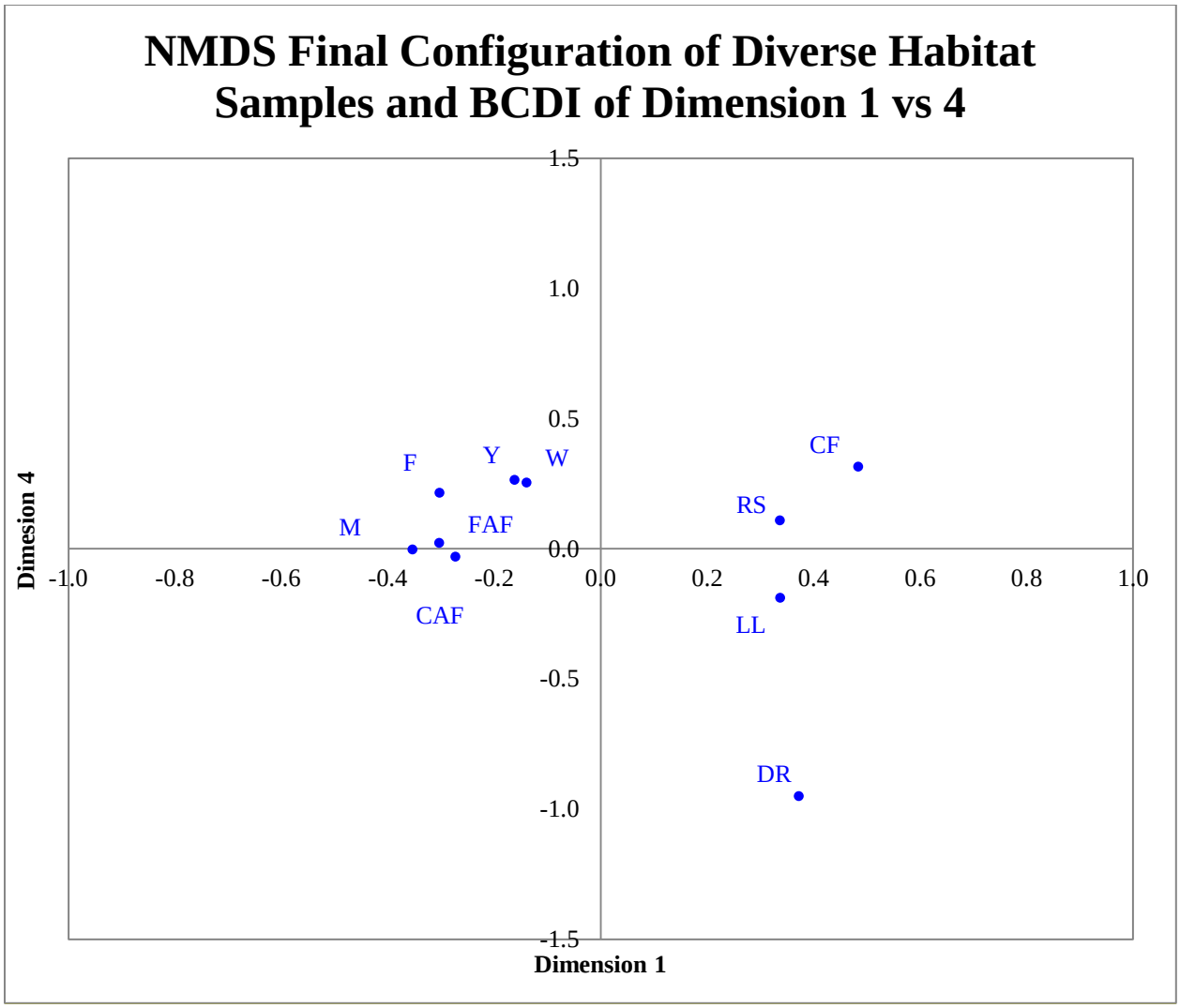

Figure 57. NMDS final configuration of diverse habitat soil samples from BCDI of dimensions 1 and 4. The yard and deciduous woods samples plot very closely and associate with the field. The marsh edge, fallow agricultural field, and corn agricultural field all cluster. The coniferous forest and roadside associate in quadrant one and plot close to the Lake Lansing beach sample. The dirt road sample is the most dissimilar from the other samples and plots the furthest away.

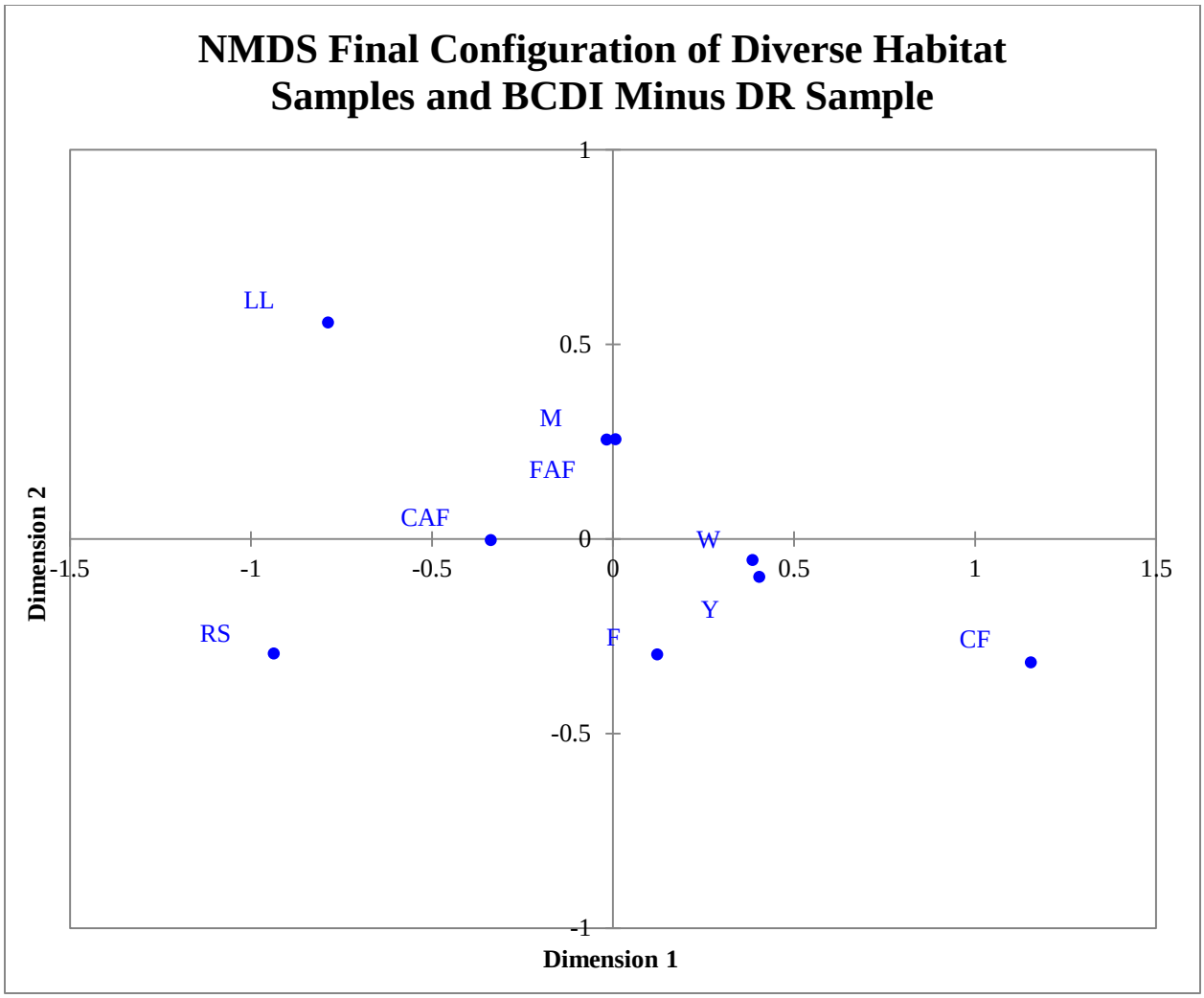

Figure 58. NMDS final configuration of diverse habitat soil samples minus the dirt road sample from BCDI. With the dirt road sample removed there is more spread to the samples in two dimensions. The marsh edge and fallow agricultural field cluster as do the deciduous woods and yard samples. The field weakly associates with the yard and deciduous woods. The remaining samples plot further from the others. See Table 6 for site names corresponding to abbreviations.

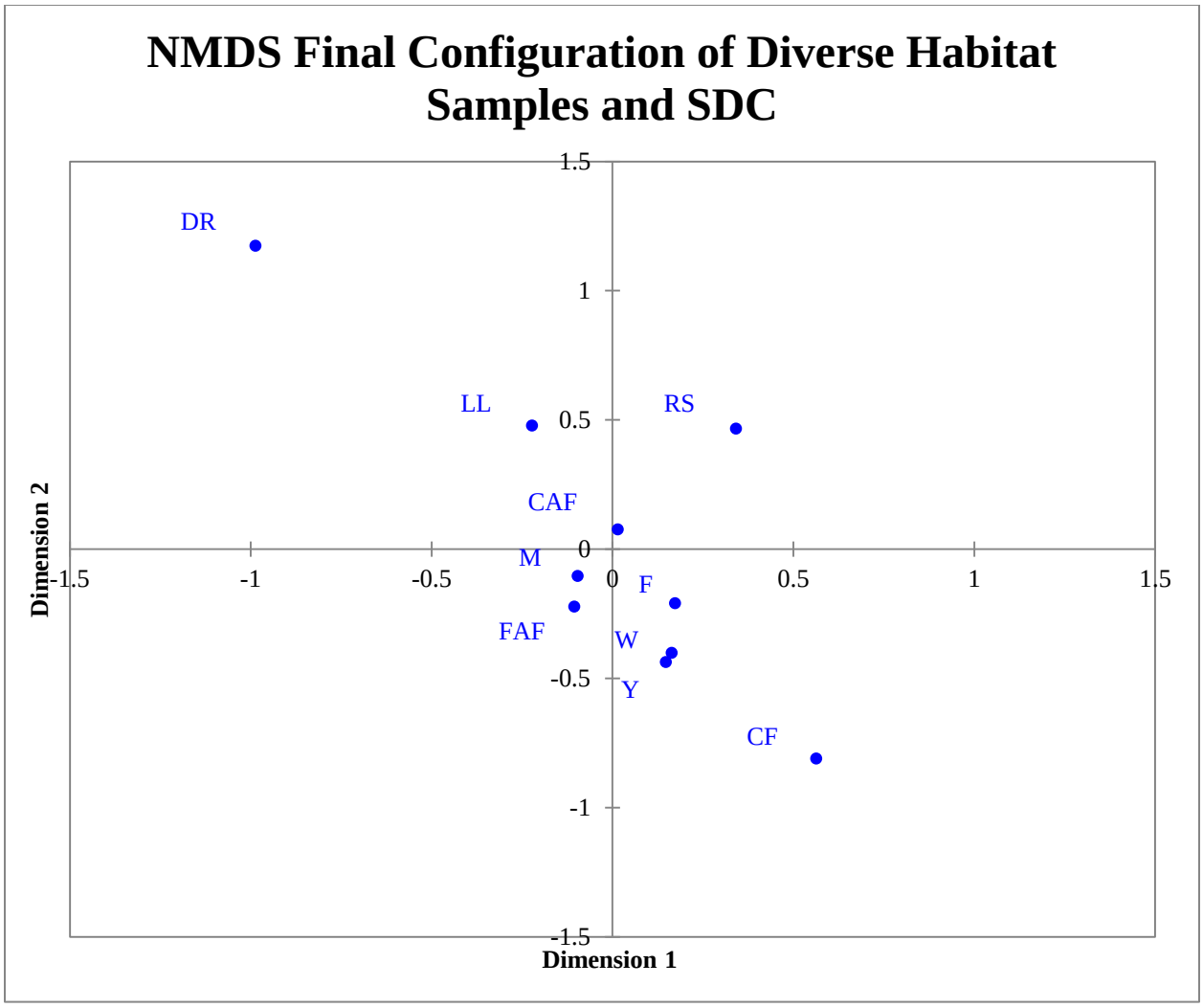

Figure 59. NMDS final configuration of diverse habitat soil samples from SDC. The dirt road, roadside, coniferous forest, and Lake Lansing beach plot the furthest from all other samples. Additionally, the field, deciduous woods, and yard are close as are the marsh edge, fallow agricultural field, and corn agricultural field. The deciduous woods and yard samples are also associated with the fallow agricultural field. See Table 6 for site names corresponding to abbreviations.

Hierarchical Cluster Analysis

Dendrograms for diverse habitat samples developed from both BCDI and SDC with single linkage clustering (Appendix 18 Figures 139 and 142) had two clusters. The first at either 0.708 or 0.750 had the yard and deciduous woods samples being the most similar followed by the field. The second had the marsh edge and fallow agricultural field clustered at 0.666 or 0.728.

The remaining samples were a dissimilarity of 0.795 or greater from the two clusters with the dirt road being the most dissimilar. The dendrograms produced from BCDI and SDC with complete and UPGMA linkage methods (Figure 60 and Appendix 18 Figures 140, 141, and 143) showed three clusters. The first between 0.728 – 0.791 had the yard and deciduous woods samples clustered followed by the field. The second had the marsh edge and fallow agricultural field clustered followed by the corn agricultural field between 0.800 – 0.809. The third cluster between 0.900 – 0.923 contained the Lake Lansing beach and roadside samples. The coniferous forest and dirt road samples were the most dissimilar from the rest.

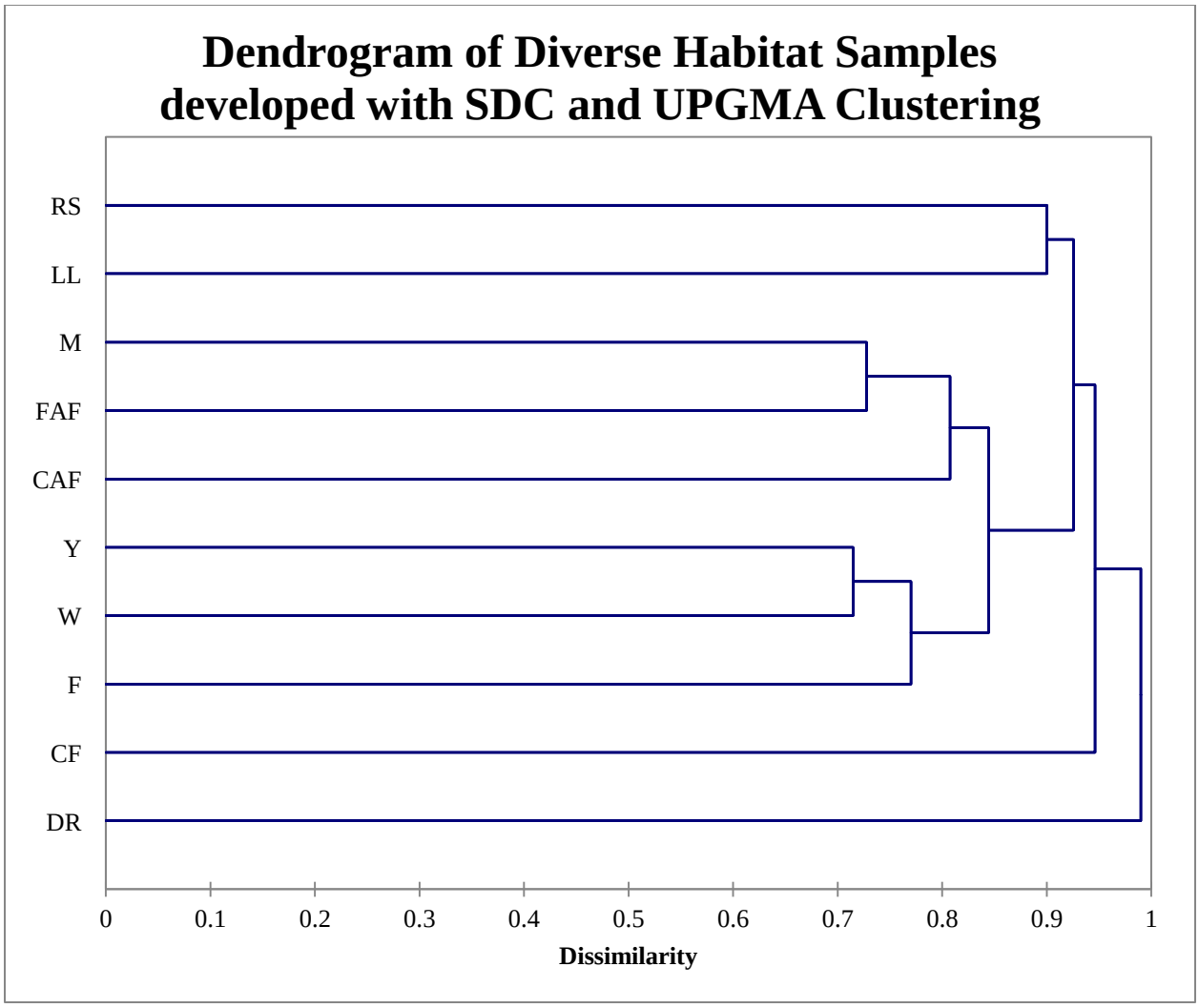

Figure 60. Dendrogram of diverse habitat soil samples developed with SDC and UPGMA clustering. Three clusters are present. The first has the yard and deciduous woods samples being most similar followed by the field joining at 0.770. The second has the marsh edge and fallow agricultural field clustering followed by the corn agricultural field at 0.807. The third cluster contains the Lake Lansing beach and roadside samples grouping at 0.900. The coniferous forest and dirt road samples were the most dissimilar from the rest. See Table 6 for site names corresponding to abbreviations.

Pairwise Comparisons

All pairwise comparisons for diverse habitat samples were significantly different for both ∫-LIBSHUFF and UniFrac (corrected p=0.00056 and p=0.0011 respectively).

k-Nearest Neighbor Classification

The yard west 100 feet sample misclassified for both diversity indices over all values of nearest neighbors during training set validation. This sample was routinely the most dissimilar from the rest of the yard samples, thus was excluded and the training sets were revalidated. The new training sets had 100% training and 98% validation accuracies. Classifications of unknown samples were the same for all odd values of nearest neighbors from 1 – 10 for both BCDI and SDC except for the yard February sample. All deciduous woods and treated yard depth samples classified to their corresponding knowns (Table 10). The remaining unknowns were not well classified. Three of the five yard samples were classified appropriately: yard August, Ymix August, and November samples. The two remaining were misclassified as deciduous woods. The four marsh edge samples classified as treated yard and were misclassified as expected since they did not have a corresponding set of knowns. These classifications were not considered towards the accuracy of this technique. Overall, using BCDI KNN classified samples with an 87.5% accuracy. The results of KNN using SDC had the same classifications with the only difference being the February yard sample classified as a yard. Overall with SDC, there was a 93.75% accuracy.

Table 10. The classification of 'unknown' samples using KNN.

Sample	Classification (BCDI)	Classification (SDC)
M-Feb	Treated Yard	Treated Yard
M-May	Treated Yard	Treated Yard
M-Aug	Treated Yard	Treated Yard
M-Nov	Treated Yard	Treated Yard
W-Feb	Deciduous Woods	Deciduous Woods
W-May	Deciduous Woods	Deciduous Woods
W-Aug	Deciduous Woods	Deciduous Woods
W-Nov	Deciduous Woods	Deciduous Woods
Y-Feb	Deciduous Woods	Yard
Y-May	Deciduous Woods	Deciduous Woods
Y-Aug	Yard	Yard
Ymix-Aug	Yard	Yard
Y-Nov	Yard	Yard
TYSurface	Treated Yard	Treated Yard
TY1"	Treated Yard	Treated Yard
TY2"	Treated Yard	Treated Yard
TY5"	Treated Yard	Treated Yard
TY10"	Treated Yard	Treated Yard
TY20"	Treated Yard	Treated Yard
TY36"	Treated Yard	Treated Yard

DISCUSSION

Technological advances have made it easier to study the bacterial makeup of complex substrates like soil. Some of the forces driving bacterial diversity include pH, plant species presence, soil type, and management procedures for human manipulated soils (Garbeva *et al.*, 2004). Surveys of the dominant soil bacterial taxa have been consistent, with major phyla including Acidobacteria, Actinobacteria, Bacteroidetes, Firmicutes, and Proteobacteria (Fierer *et al.*, 2007; Janssen, 2006). These findings were reproduced in this research. Additionally, the phylum Planctomycetes was discovered in similar quantities to the aforementioned phyla. These bacteria were long thought to primarily reside in aquatic environments (Buckley *et al.*, 2006), but since 1992 have been shown to also exist in soil of diverse habitats. Janssen *et al.* (2006) reported that Planctomycetes range from 0 – 8% of the total bacterial abundance in soil, which is consistent with the findings in this work. Class level diversity could not be verified from the primary literature, though the majority of those identified in this research were members of the major phyla noted above.

Evaluating the number of bacterial classes identified for each sample, the Lake Lansing beach and the biological replicates stand out as being different from the others in their respective studies. First, the Lake Lansing beach sample was found to contain more total (78) and least abundant 5% (49) bacterial classes. The texture of soil has been shown to influence bacterial diversity (Chau *et al.*, 2011), where coarser soils have higher levels of diversity. These soils, like sand, can have lower water retention leading to isolated bacterial communities (Carson *et al.*, 2010). If the communities are disconnected, bacterial diversity can flourish because motility is decreased, reducing competition for nutrients and allowing 'less competitive' species to thrive. Total bacterial diversity was also shown to increase as distance from the ocean increased in Hawaiian beaches (Cui *et al.*, 2013). The backshore locations had less tidal activity, causing the

sand to be drier and potentially increasing bacterial diversity. The sampling location at Lake Lansing was near the edge of the beach a few feet from the water and could be drier than the other soils in this study. Additionally, Lake Lansing beach is artificial, where the creation of the beach, human/animal activity, and its maintenance could be influencing the amount of bacterial diversity. Second, biological replicates from the marsh edge, yard, and deciduous woods had fewer bacterial classes than most of the time series samples. Comparing the total and least abundant bacterial classes of these samples to the others (Table 9) both were greatly reduced. These differences could be a result of sequencing variability, as they were processed slightly differently from the others, where each sample was diluted to one nanogram of DNA in the combined pool for sequencing. It is possible this dilution caused the rarer classes to be underrepresented in the sequencing libraries. The remaining pools were undiluted and had a higher number of classes identified, so reducing the total DNA from each sample in the pool would be unadvisable if it is lowering the amount of diversity in the sequence libraries.

Library processing followed a version of the 454 protocol found on the mothur website that had been slightly modified in the Schmidt laboratory, and their SOP was used. The average loss of 87% of total sequences during mothur processing was surprising but not completely unexpected when using the strict parameters in each step. The first step removes sequences that have any ambiguous bases, they must have a high average quality, and less than eight of the same base consecutively. There can be only one difference between the sequence and barcode and two differences for primers as well. It is not unreasonable then that a large portion could be removed because of sequencing errors. The second step eliminates any sequence under 250 bp. Short sequences reads are not uncommon and are less informative than longer ones, so another screen to remove reads less than 300 bp is utilized later in processing. It would make sense that if

the longer threshold was imposed earlier more erroneous sequences would be removed sooner in processing, quickening it while maintaining only those that are desirable. The unique sequences command removes repetitive sequences for the purposes of faster processing; however no information is lost as the abundance of each sequence is maintained through the remaining steps. Finally, the precluster step attempts to remove any sequences believed to be from pyrosequencing errors by merging them with more abundant sequences they are thought to originate from. It is not surprising then that a large number of sequences (*ca.* 70%) were 'lost' at this stage since this step clusters sequences with a one base pair difference.

All libraries were subsampled to the group with the lowest number of sequences in an attempt to reduce the influence of unequal number of sequences rather than biological differences. However, in doing so it is possible rare sequences could be underrepresented or even completely eliminated. Subsampling is random, though it is conceivable that it will not accurately represent the diversity of the original community. If this is true, it would follow that anything measuring these, *e.g.* BCDI, SDC, ∫-LIBSHUFF, and UniFrac, will be influenced by under or overrepresentations of species, especially when these are artificially different in samples based on subsampling. Investigating the influences of subsampling would be useful to understand if results from downstream analyses are affected by it, which could be easily performed through direct comparison of sampled and unsampled libraries. It might also be worthwhile to determine if a single subsampling event accurately captures the species and their abundances in the libraries or if they should be sampled multiple times to overcome stochastic differences of the process.

Perhaps the most important aspect of the work presented here is obtaining a better understanding of how the different statistical techniques reflected the bacterial sequence data

obtained. The samplings undertaken—biological replicates, within and among habitat variability, and temporal—all have the potential to result in data sets that respond differently to the various statistical manipulations, thus each will be considered individually below. In some cases additional experiments and/or statistical considerations may be necessitated, especially when addressing the forensic implications of these studies.

β-diversity was used to assess how dissimilar bacterial communities were among all samples in this research. Bray-Curtis dissimilarity index and Sørensen-Dice coefficient are commonly used in ecological research and were investigated in this work. BCDI measures dissimilarity based on both shared membership and differences in abundance, so logically it can be affected by both variability in PCR amplification and sequencing as well as natural fluctuations of species. If sequences are equally represented in two samples, BCDI will find little dissimilarity between them; however, if large abundance differences exist, a high level of dissimilarity could be developed even if the samples, such as biological replicates, are expected to be similar. The specificity at which BCDI measures dissimilarity might not be useful in a forensic setting, especially if species abundances fluctuate regularly. Samples taken from the same habitat could have high levels of dissimilarity, which might influence their association to unknown samples in later analyses.

On the other hand, SDC only measures shared membership of sequences; it does not consider abundances. While this could still be influenced by PCR and sequencing variability as well as changes in bacterial composition, it would not be expected to be affected as much as BCDI. This broader measure might be forensically more applicable if the number of species is not fluctuating but their abundances are; however, the resolution of samples would be lower using SDC since samples need only share at least one sequence in common in every OTU to not

be different. It should be noted that both BCDI and SDC consistently calculated dissimilarities above 0.6, higher for biological replicates, which intuitively seems high. These values could be affected by many rare sequences in the samples being compared. Such sequences could be underrepresented or lost completely because of subsampling (as noted above) leading to high calculations of dissimilarity. It is also possible the amount of data next-generation sequencing can produce allows for the representation of rare populations specific to certain soils and these high dissimilarities are accurate representations of that. The rare populations were usually represented in these data as single sequences, and calculation of dissimilarity without inclusions of these singleton sequences could be used to determine if they were the cause of the high dissimilarity values among the samples.

Because the two indices assess bacterial populations differently, they have the potential to generate conflicting results, which was exemplified in the research presented here. NMDS final configurations showed differences between the diversity indices for the diverse habitat, similar habitat, time series, and depth studies. The most obvious example was from the time series samples. NMDS using BCDI did not differentiate the deciduous woods and yard habitats, while SDC did (forensic implications discussed below). The remaining inconsistent NMDS configurations varied more in the distance between samples than in cluster membership, which forensically is less of a problem. If one diversity index routinely causes samples from the same habitat to not cluster closely, its use for association of samples would be very limited. The distances between samples from the same habitat were usually smaller with SDC than BCDI, indicating the former might be better for forensic implementation. HCA dendrograms were also divergent in cluster membership for the time series, similar habitat, and depth studies, along with the habitat distance samples. Again, the time series differences were the most obvious, showing

similar results to the NMDS configurations. Comparisons of final configurations and dendrograms showed inconsistencies when the same diversity index was used, the majority of which were present in dendrograms developed from complete linkage clustering, though UPGMA dendrograms were contradictory as well. The classification of unknown samples using KNN had only one difference between the diversity indices, indicating either could be utilized.

NMDS is a useful tool for finding patterns or relationships within nonparametric data sets using any 'distance' measure by plotting samples in *m*-dimensional space, where the number of dimensions of the plot are user defined. To help an analyst choose an 'appropriate' number of dimensions for an accurate representation of samples, Scree and Shepard diagrams are utilized. Though these graphs attempt to portray how well the final configuration is representing the rank-order of dissimilarities, their interpretation is subjective (Borg *et al.*, 2013). There are no set standards of how high a stress can be that makes the configuration uninterpretable or how far away a distance has to be from its corresponding disparity to make it an outlier. Spence's random stress was used to help eliminate subjectivity in Scree plot interpretation. Plots generally fell well below what is expected for random stress, though seven had a stress above the threshold in one dimension, yet these too fell below random stress when a second dimension was added. The low number of rejections indicates that either NMDS is representing the data well enough to be below the threshold or Spence's random stress is a very conservative choice for comparison. Likely, it is a combination of both; however, random stress consideration is still useful for reducing subjectivity by acting as a 'worst-case scenario' for comparison against the Scree diagram. Currently, there is not a similar test for Shepard diagrams, though to standardize interpretation, thresholds could be developed or set that require distances to be within a certain vertical measure from disparities.

The interpretation of the NMDS final configurations is subjective as well. Standard error bars were used to help assign membership to clusters; however, it is obvious from the final configurations that samples were often outside this measure. Standard error is not often applied to NMDS configurations and with only a small number of samples plotting within error bars, they were not very useful for interpretation. The assignment of cluster membership needs to be standardized for uniform interpretation and to reduce analyst bias, as subjectivity is a major limitation for using this technique forensically. If samples are clustered well, as in the biological replicates, interpretation is easy, but if they are not, as in the time series study, the assessment of which samples are associated becomes considerably harder, especially when a point is equidistant between two others. It would be helpful to define a maximum distance between points to include them in the same cluster; however this should be done cautiously because the distance among samples can change depending both on how many samples are being plotted and how dissimilar they are. This was exemplified in the diverse habitat study, where the dirt road sample in the BCDI plot forced all other samples to fall almost directly on top of each other. With this in mind, comparing multiple dissimilar habitats at the same time is unadvisable. Reducing the number of habitats and increasing the number of samples could add strength to associations, especially if all samples cluster within their respective habitat.

HCA is similar to NMDS in trying to identify natural groupings or clusters of samples that are present in complex multivariate data sets; though, conflicting dendrograms can be developed using different linkage methods (Everitt *et al.*, 2011). Regardless of the linkage method, well clustered data will form the same groupings but possibly at different dissimilarities (*e.g.* the biological replicates) however, for highly dissimilar samples (*e.g.* the time series ones), dendrograms can be inconsistent. Again, since dissimilarities tended to be high for all studies,

groupings often differed based on linkage method, thus the use of HCA to find clusters of samples was limited for analysis of these data. When clusters were well defined, the groupings were more informative and allowed a better understanding of sample association and possible forensic application, but the high number of differing dendrograms decreases the usefulness of HCA for this type of data.

The two pairwise statistical techniques used in this research, ∫-LIBSHUFF and UniFrac, are considered divergence-based methods because they do not make the assumption that all species in the sample are equally related (Lozupone and Knight, 2008). The divergence among community members is calculated differently depending on the test being used: ∫-LIBSHUFF calculates sequence distance, while UniFrac calculates phylogenetic distance. Since the statistics are assessing different conditions in each community it was not surprising they produced different results. ∫-LIBSHUFF is sensitive to differences in abundances of sequences while UniFrac to variations in evolution (*e.g.* selection for/against species or sequence mutations). Communities that have evolved similarly, like those from the same environment, would not be significantly different using UniFrac but because of natural variations in environmental conditions species abundances could be very different. If this were the case, ∫-LIBSHUFF would show the communities to differ statistically where UniFrac would not. This was seen throughout this research (discussed further below). Finally, ∫-LIBSHUFF calculated two p-values for a single pair of samples, that were rarely the same. The calculation of both a significant and non-significant p-value for the same pair of samples introduced an aspect of subjectivity to their interpretation. A decision had to be made as to whether both values needed to be significant (‘strict’) or only one (‘relaxed’) to determine if the samples were statistically different. UniFrac, however, only calculates a single p-value for each pair of samples, removing this requirement.

The variation in interpretations of p-values resulted in some inconsistencies between the two statistics.

Differences between the statistics themselves can explain some of the relationships seen among the samples; however, the underlying nature of bacterial communities also has important implications on the association of samples. There are two major competing ideas of prokaryotic biogeography: cosmopolitan or endemic distribution of taxa (Ramette and Tiedje, 2005). It is well accepted that bacterial classes including Actinobacteria, β-Proteobacteria, Cyanobacteria, and Flavobacteria have a global terrestrial cosmopolitan distribution. Further, genus level taxa are believed to follow a similar distribution within habitats (Ramette and Tiedje, 2005). Contrarily, it is becoming better understood that prokaryotes at lower taxonomic levels also exist in non-random biogeographic patterns (Martiny *et al.*, 2006; Ramette and Tiedje, 2005), which are thought to mainly develop from a combination of speciation (divergent evolution of a species due to physical barriers), extinction, and dispersal limitations from one location to another. Ambient temperature, geographic latitude, and distance among habitats have been proposed to not influence bacterial communities (Fierer and Jackson, 2006); however, more recently the spatial distribution and abundance of bacterial taxa were found to be positively correlated with environmental variables influencing plant diversity (Griffiths *et al.*, 2011). These factors can cause microhabitats in soil, where slightly different environmental conditions create the potential for bacterial diversity to vary over small spatial scales. Additionally, current and historical conditions, *e.g.* weather events or dispersal limitations, need consideration to fully understand what influences prokaryotic biogeography (Martiny *et al.*, 2006).

In the current study, differences in communities were evaluated over distances smaller than a foot, since the aforementioned factors have the potential to influence bacterial diversity

over similar spatial scales. Multivariate statistical analysis of biological replicate samples, with their reduced class numbers, showed the same relationships using both BCDI and SDC: habitats were completely separated and replicate samples were more similar to each other than to the other habitats. Microhabitats, however, can potentially explain the statistical differences seen using ∫-LIBSHUFF in the marsh edge replicates. Two of these samples (M2 and M3) differed statistically, which is reflected in their relative abundance charts (Figure 13), while none of the deciduous woods or yard replicates differed within their habitat. Grundmann and Debouzie (2000), using an antibody assay, found that *Nitrobacter* formed groupings over distances as small as 2 – 4 mm. If bacterial populations can differ in abundance over such a small distance, this would be reflected using a statistic, *e.g.* ∫-LIBSHUFF, or a diversity measurement, *e.g.* BCDI, sensitive to differences in abundance. The finding of significant differences between replicate samples signals potential forensic complications, as bacterial communities differing significantly over small spatial scales could severely limit the ability to associate evidentiary soil with a location. If the known sample came from a microhabitat different from where the unknown sample originated, they could differ significantly and lead to rejection of that location as a source of the evidence, accentuating the need for multiple samples to be collected from each location. The potential inability of ∫-LIBSHUFF to associate knowns and an evidentiary sample could make it forensically limited (discussed below).

In contrast, 80% of comparisons between deciduous woods and yard samples did not differ statistically using UniFrac, while ∫-LIBSHUFF showed no differences for 25% of the same samples. These findings have a different forensic implication. If ∫-LIBSHUFF and UniFrac cannot differentiate between distinct habitats, their forensic utility is dubious, calling into question any non-significant differences between known and evidentiary soils. In the late 1950s,

both the yard and deciduous woods were part of a large farm according to the Fenner Nature Center Staff, and the resolving power of the tests could be limited for historically similar soils. Though over 60 years have passed since each habitat was established, it is possible the bacterial communities have not wholly diverged from each other evolutionarily. This seems unlikely however, as bacterial populations have been show to diverge over as few as 1,100 generations (24 hours) in experimentally controlled environments (MacLean and Bell, 2003). It is also possible the low amount of diversity in the sequence libraries of deciduous woods and yard replicate samples could be reducing the ability of ∫-LIBSHUFF and UniFrac to show significant differences between them. The processing of these samples could have influenced the rarer sequences of the samples and underrepresented their differences, as noted above. Though the multivariate statistics were able to differentiate the three habitats, a deeper investigation into their bacterial communities would need to be conducted to better understand the implications of these results.

Environmental conditions have also been reported to influence bacterial composition in soil over short distances (Martiny *et al.*, 2006), though said distances were larger than those investigated with the biological replicates. Horner-Devine *et al.* (2004) found a distance-decay relationship from as little as a few centimeters to over a hundred meters when sequencing clone libraries of Proteobacteria. Communities geographically closer to each other were similar, becoming less so as distance increased. When evaluating the cause of this, the authors found differences in environment, such as nutrient availability and moisture content, rather than distance or plant composition, played the more important role. A distance-decay relationship was seen in the distance samples from the treated and untreated yards in this study, most notably with NMDS (Figures 26 – 29), HCA (Figure 33 and Appendix 10-1 Figures 95, 97, and 98; Figure 35

and Appendix 10-2 Figures 99, 101, and 102), and ∫-LIBSHUFF (Figures 38 and 40), while UniFrac (Figure 41) only suggested one for the untreated yard. Of the yards, the untreated one is a more 'natural' environment with only minor maintenance of grass height and human perturbation, and it exhibited a dramatic distance-decay relationship congruent with those of Horner-Devine *et al.* (2004). The prevalence of statistically significant differences was found to increase as distance increased for both ∫-LIBSHUFF and UniFrac; however, the former showed them first at five feet while the latter at greater than 10. ∫-LIBSHUFF comparisons for the treated yard showed statistical differences starting at distances greater than 10 feet, with more found over larger distances, which is also reflected in the relative abundance charts (Appendix 6 Figure 71). A reason for the different distance-decay relationship between the yard and treated yard could be a reduction of environmental factors that influence small spatial distances with the introduction of non-naturally occurring chemicals (*e.g.* pesticides and fertilizers). In this regard, pesticide treatments have been found to reduce bacterial diversity in field soil when investigated with TGGE and clone libraries (Engelen *et al.*, 1998). This yard is also fertilized, which has been shown to reduce the number of some bacterial species (Sarathchandra *et al.*, 2001) but increase the abundance of others (Bittman *et al.*, 2004). Both environmental conditions and chemical treatments could explain why distances less than 20 feet were not significantly different. The UniFrac comparisons did not show a strong distance-decay relationship in the treated yard (Figure 39), instead displaying relatively few statistically different samples, all restricted to distances greater than 20 feet. This suggests that chemicals (both pesticides and fertilizers) may be acting to 'homogenize' the soil with regard to species presence, though not necessarily abundance.

Interestingly, NMDS (Figures 30 and 31), HCA (Figure 37 and Appendix 10-3 Figures 103, 105 and 106), and ∫-LIBSHUFF (Figure 42) comparisons for distance samples from the deciduous woods showed no distance-decay relationship, even at the closest distances. ∫-LIBSHUFF results revealed more than 70% of samples differed statistically for all distances. Additionally, both NMDS and HCA seemingly had a random distribution of samples. UniFrac comparisons (Figure 43) potentially showed a distance-decay relationship up to 20 feet; however, that relationship did not continue out to greater distances. These findings could be explained by environmental heterogeneity. The forest floor was littered with decaying leaves, fallen trees, a large variety of plant life, and patchy areas of sunlight that could be causing substantial heterogeneity over short distances, as suggested by Meyers and Foran (2008). Over smaller distances, like in the biological replicate study, deciduous woods samples were similar. This indicates that sampling over shorter ranges (*e.g.* 1 foot intervals or less) may be required to capture a distance-decay relationship in such a highly variable habitat.

Differences in depth have also been shown to influence bacterial communities. Kuske *et al.* (2002) reported significant differences in communities at 10, 20, and 30 cm depths in Colorado grasslands utilizing T-RFLPs and HCA or *t*-tests. Griffiths *et al.* (2003) had similar results using DGGE and either PCA or HCA in a UK temperate upland grassland at depths of 5, 15, and 20 cm. The aforementioned depths were similar to the shallower ones investigated in the current study (*ca*. 2 – 10’’); however, the results differed. ∫-LIBSHUFF indicated all depths from surface – 2” were statistically different while the 5” and 10” samples did not differ. These were not consistent with the relative abundance charts (Appendix 6 Figure 74) however, as all samples were visually similar. UniFrac showed the surface and 1” as well as 5” and 10” depths to not differ, while the 2” and deeper depths did. NMDS configurations (Figures 44 and 45) and HCA

dendrograms (Figure 46 and Appendix 12 Figures 111 – 115) tended to support UniFrac findings. Kuske *et al.* (2002) and Griffiths *et al.* (2003) suggested that differences in soil chemistry influence bacterial populations as depth changes. They concluded moisture content and elemental abundances are the main drivers for the differences, which could be the reasons for the results found here. Additionally, both sets of authors noted that bacterial diversity decreased as depth increased. The exact opposite was found in this study, where the lowest depth had the highest number of classes identified. This relationship could not be supported by the literature as such depths have not been previously tested. The relatively lower diversity of the shallow depths could result from the chemical treatment of this yard, as noted above. Additionally, these findings could result from the use of next-generation sequencing rather than T-RFLPs or DGGE. The 454 platform develops considerably larger amounts of information than the other techniques, as well as a deeper exploration of bacterial communities. Bacterial sequences can be analyzed at very discriminate levels of classification (*i.e.* genus and species) allowing better differentiation of populations as opposed to T-RFLPs and DGGE. So, if samples all contained similar bacterial classes, which those in this study did, they could still be individualized because of the higher resolution of the members in each, rather than relying on slight differences in abundances. It is feasible that since the older techniques are comparatively limited in their detection ability, the differences seen are real.

Bacterial communities of similar habitats have the potential to differ if environmental conditions are conducive to increased diversity. Strickland *et al.* (2009) reported the chemical complexity of a habitat influenced bacterial communities, where the least complex allowed for the most diversity. They studied three habitats using clone libraries and ANOVA or Tukey's HSD, and showed that a grass field with lowest carbon complexity had the most diversity, while

harsher environments tended to drive bacterial communities to be more phylogenetically similar. Lower complexity could be the reason the ten yards in this study were mostly differentiated from each other with all statistics. Both pairwise comparisons revealed all but two of the yards as being statistically different, though those two did not associate closely in NMDS configurations (Figures 53 and 54), HCA dendrograms (Figure 55 and Appendix 16 Figures 128 – 132), or relative abundance charts (Appendix 6 Figure 76). Final configurations of the ten yards mostly agreed with each other, though samples tended to plot closer using SDC than BCDI. Dendrograms were variable among the linkage methods used, particularly those developed from SDC values. These differences could be a result of the small range of dissimilarity values calculated with SDC (most were around 0.7). Additionally, the SDC complete linkage and all BCDI dendrograms showed similar relationships with the NMDS final configurations, though cluster membership differed slightly among them. Plotting or clustering this many samples, especially if their dissimilarity values do not vary much, reduces the resolving power of these techniques because clusters are not well defined (discussed further below).

The inability to differentiate two of the yards (Michigan State University east and Lisa) with both pairwise comparisons suggests they are biologically similar, though the distance separating them and nearly significant values for UniFrac indicate they are probably not. A direct comparison of ∫-LIBSHUFF and UniFrac by Schloss (2008) on simulated bacterial communities showed UniFrac had high and low p-values dependent on if the populations were similar or not, respectively. On the other hand, ∫-LIBSHUFF routinely calculated low p-values regardless of whether or not the communities were different, making these results more inconclusive. The low value might just be a limitation of the technique rather than an indication bacterial abundances are significantly different. It is also possible that because of the high number of comparisons

made for both statistics, these findings are demonstrating type II error, *i.e.* samples are found to not differ significantly when they actually do.

The ten habitats in the diverse habitat study were mostly separated with the multivariate techniques. Lenz and Foran (2010) using T-RFLPs and NMDS reported similar differentiation of some, but not all habitats investigated. A major limitation of NMDS is when multiple dissimilar samples are plotted together, highly different ones can force less dissimilar ones to be close in multidimensional space. This was seen in the BCDI NMDS configuration (Figure 56) for the ten habitats, where the dirt road caused the other habitats to cluster tightly. The dirt road sample was extremely different in the relative abundance charts compared to the other samples, and had the lowest number of classes identified for this entire work. The road is treated with calcium chloride twice a year (Shiawassee County Road Commission personal communication) and high salinity in soil has been shown to reduce bacterial diversity as a result of the extreme environment (Hollister *et al.*, 2010). The most abundant bacterial classes in the dirt road were Acidobacteria, similar to other samples, but also Flavobacteria, Bacilli, and Clostridia, which contain halophilic members that have been shown to thrive in a highly saline environment (Amoozegar *et al.*, 2005; Quesada *et al.*, 1983; Ragab, 1993). When projecting all samples into two dimensions, differences between the remaining nine habitats were underrepresented even though the majority were highly dissimilar from each other (*e.g.* values greater than 0.9). Therefore, care must be taken when analyzing configurations so erroneous conclusions are not made because a greatly dissimilar data point is artificially forcing other samples to be closely associated. The final configuration without the dirt road sample had a larger spread to the samples in two dimensions, though the high dissimilarity among all points could be influencing random clustering of some of them. It was also useful to analyze all ten samples with a different combination of dimensions.

The one versus four dimension plot (Figure 57) showed similar relationships to the configuration when the dirt road sample was removed. Analyzing all ten samples in 1 – 8 dimensions resulted in erroneous Scree diagrams (similar to Figure 12) after 4 dimensions, so pairwise combinations of all dimensions were not considered. It could be worthwhile to investigate additional combinations of dimensions to elucidate relationships when they are believed to be underrepresented, like with the BCDI plot of all ten samples, but careful consideration of Scree and Shepard diagrams is needed for confidence in the placement of samples in multidimensional space.

The diverse habitats sampled were, however, easily differentiated with both pairwise statistics, which was reflected in the relative abundance charts (Figure 15). This is not unexpected in light of the findings for the much less diverse yard samples, which were likewise differentiated. If the chemical complexity of soil is an indicator of the bacterial community structure as found by Strickland *et al.* (2009), the ten habitats would reflect this in differences of bacterial abundance seen with ∫-LIBSHUFF. Lauber *et al*. (2009) showed pH directly influenced diversity of bacterial communities in diverse soils analyzed with UniFrac. These two factors, as well as environmental conditions, habitat heterogeneity, plant diversity, water content, *etc.* could be causing bacterial communities to be very different in the ten diverse habitats of this study. Forensically, it is important to be able to reliably differentiate habitats. Visually, the relative abundance charts indicated these habitats were distinct from each other and it would call into question the forensic utility of any statistical technique that could not differentiate them.

Smit *et al*. (2001) found that seasonal changes influence the structure of bacterial communities evaluated with DGGE, clone libraries, NMDS, HCA, and a permutation analysis. The authors reported that total bacterial abundance changed significantly season-to-season for

wheat field samples collected seasonally over the course of a year. Additionally, Lipson and Schmidt (2004) found significant changes in abundance and number of bacterial classes in alpine soils of the Colorado Rockies from winter to summer when evaluated with clone libraries and ANOVA or phylogenetic tree permutation analysis. These findings are weakly supported by pairwise comparisons in the current research. Pairwise evaluation of time series samples (Figures 51 and 52) resulted in statistical differences for both within habitat and between habitat samples. The number of samples that were different was similar between the two techniques, though the actual samples often were not.

Change in bacterial populations within a habitat due to seasonal variations has the potential to make the identification of evidentiary soil difficult. The possible convergence of populations due to extreme or unusual environmental conditions could make habitats more difficult to differentiate, limiting the forensic utility of bacterial profiling. The only statistics to completely discriminate the marsh edge, yard, and deciduous woods time series samples were HCA with UPGMA or single clustering (Figure 50 and Appendix 14 Figure 122) and NMDS with SDC (Figure 49), both of which had three clusters consisting of samples from individual habitats. Forensically, this is useful in discriminating among three distinct habitats while also associating samples separated by large amounts of time. The remaining configuration and dendrograms showed intermingling of deciduous woods and yard time points, while the marsh edge samples were consistently removed from the rest. The deciduous woods and yard were frequently found to be similar with the multivariate statistics throughout these studies and it is possible the marsh edge samples were influencing their clustering in configurations and dendrograms.

The pairwise comparisons showed the most conflicting results for time series samples. Fifty percent or more of all within habitat comparisons were significantly different for both ∫-LIBSHUFF and UniFrac. However, no month was consistently different from any others from the same habitat. Further, 50 – 100% of the between habitat comparisons were significantly different. Similar to the within habitat comparisons, no month was consistently different from the others when assessed against the other habitats, and results were different between ∫-LIBSHUFF and UniFrac. There were 12 comparisons that did not differ significantly between habitats when considering both pairwise statistics, three for ∫-LIBSHUFF and nine for UniFrac. Of those, only one was shared: the deciduous woods May vs yard November. The remaining comparisons showed no obvious correlation between the month in which a habitat was sampled and if it was not significantly different from other time points, seeming almost random. The discrepancies between the two pairwise statistics could largely be explained by the interpretation of the p-values, where only a single significant value for ∫-LIBSHUFF was required for the samples to be deemed different. This affected seven of the eleven differences between ∫-LIBSHUFF and UniFrac (the twelfth was shared already), as those seven all had one significant and one non-significant ∫-LIBSHUFF p-value. Forensically, significant differences occurring within a habitat or lacking between habitats over time are troubling and limit the potential of associating known and evidentiary soils when large amounts of time have passed. It is very possible using pairwise statistics for evaluation of these samples focused interpretation too much on whether samples were different or not, where the more pressing question is how different can they be while still associating with the habitat they originated. The multivariate statistics showed the same samples can still associate correctly, even while being statistically different.

Interestingly, UniFrac was the only statistic to not differentiate all the marsh edge and yard time points, where three, apparently random, comparisons were not significantly different. Further, NMDS and HCA always had clusters of only marsh edge samples that were easily differentiated from the deciduous woods and yard time points. Again the pairwise comparisons indicate how limited these types of statistics may be for forensic use, but also highlight that how bacterial communities are changing over time is still not well understood, and there might not be any predictability to how samples are going to differ based on the season in which they were collected.

The influence sample processing had on association of libraries was assessed with the comparison of the yard mix August and yard August samples. The former was the three replicate soils mixed in equal ratios and extracted together, while the latter was a compilation of their three sequencing libraries. Overall, both methods seemed to improve results, in that fewer significant differences existed between them and the other yard time points. Likewise, some yard to marsh edge or deciduous woods comparisons that were significantly different using single samples, were not using the mixed or compiled methods. The two methods did not always generate the same results however, as they differed in three between habitat comparisons using ∫-LIBSHUFF. Sequencing samples separately will require more libraries to be produced; however, as costs for sequencing decreases this becomes more feasible. If soils are mixed before extractions, it is possible rare bacterial communities could be underrepresented or lost based on stochastic sampling. Given these results, further investigation into how samples are extracted and sequenced is needed.

Regardless of how samples are processed, multiple ones are needed for forensic comparison purposes in order to overcome microhabitat heterogeneities, given that some

replicate samples in this study differed significantly. It would also be advisable to take multiple samples from discrete locations to increase sample size for statistical analysis (*e.g.* many samples are required for an adequate training set). Also, if samples are sequenced separately, potential outliers could be identified if pairwise comparisons are conducted. A significantly different library would indicate a microhabitat was potentially sampled and it could be excluded from a training set or from an analysis with a statistic like NMDS, where highly dissimilar samples will influence cluster membership. Currently, how many samples to collect from a habitat is unclear, but since it is now possible to cheaply sequence many more samples than what was even possible in this work, the more libraries produced the more robust the analysis will be.

The high amount of diversity in the least abundant bacterial classes of the biological replicates indicates this region would be a poor choice for statistical analysis. It is possible that differences among samples resulted from loss of rare classes during subsampling (noted above). These samples could then cluster poorly as it is probable both BCDI and SDC will calculate higher dissimilarities. Also, they would be expected to differ significantly with pairwise statistics. Finally, highly dissimilar training set samples can cause KNN to misclassify unknowns. Validation accuracy could be low since class members are very different from each other, and classification of unknowns would be suspect in light of known samples already being poorly classified. To overcome these potentialities, multiple rounds of subsampling could be conducted and libraries from each event merged. By doing this, the rare populations would still be represented in the final libraries and the frequency at which they were lost during subsampling could be determined.

If the exact origin of evidentiary soil is unknown, samples could be collected through the entirety of a crime scene to both increase the number of comparisons and potentially overcome

environmental heterogeneity. This would generate an extensive number of samples, so further studies need to address how many samples should be taken to accurately represent habitat-wide heterogeneity, and how accurately 'evidence' can be identified. Results from the within habitat distance study showed that for habitats that are fairly uniform, *e.g.* a yard, samples tend to differ more as distances between them increase, indicating the sampling used captured an expected distance-decay relationship. This would be forensically useful because known samples taken close to where evidentiary soil came from could show less difference than those further away. The exact location would not have to be sampled, and based on these data could be up to ten or twenty feet distant if analyzing libraries with UniFrac. Conversely, the deciduous woods environment did not show a distance-decay relationship, which poses a considerable problem forensically. Known samples would need to be collected very close, perhaps within a foot, to where evidentiary soil originated to show no statistical differences. This could be difficult if not impossible when an impression, *e.g.* shoeprint or tire track, is not discovered. The high variability within this environment could require extensive sampling to ensure known soils are taken close to where evidence could have originated.

Two additional variables that had a substantial effect on bacterial communities were time and depth. The influence of time between when a crime was committed (*i.e.* when evidentiary soil could have been deposited) and when soil samples are collected is not trivial. It would be advisable to collect samples from a crime scene as soon as it is discovered and also when environmental conditions are similar to when a crime occurred, but it is still unclear how these factors influence bacterial communities. Also, the burial of human remains or objects of forensic relevance presents a unique challenge to the investigation of soil evidence. During the act of burial the soil stratification is disrupted, potentially homogenizing or rearranging it. How this

influences bacterial communities is still unknown, though in this work, communities from previously undisturbed soil were shown to differ as depth increased.

If a suspect claims evidentiary soil came from a specific location other than where a crime is thought to have occurred, having the ability to differentiate habitats is crucial and of great forensic relevance. Both pairwise comparisons and multivariate statistics were able to differentiate all the diverse habitats and all but two of the similar habitats (yards). Previous studies using older molecular techniques have been limited in differentiating even diverse habitats, or in understanding how variables like time and depth are influencing bacterial communities. Next-generation sequencing produced close to a million raw sequences when all studies were combined, allowing investigation at the unique sequence level rather than at higher taxonomic levels common for the older techniques. Deeper investigation into bacterial communities was possible, and diverse statistics could be applied to better determine how they were different or changing. Overall, these findings show how powerful the analysis of next-generation sequencing coupled with different statistics is for forensic soil analysis.

The statistics evaluated in this research helped identify advantages and disadvantages of both multivariate techniques and pairwise comparisons. Throughout the studies, NMDS and HCA varied in the relationships they presented among samples, as did the pairwise comparisons. In no study did all statistical outputs completely agree. To understand how bacterial communities are different or changing, these statistics are helpful, but possibly forensically limited. Inconsistent results could lead to alternative interpretations of them (*e.g.* the prosecution and defense presenting conflicting conclusions from the same sequence libraries). This situation would lower the value of evidence in court and demonstrates the need for a single statistical

technique meeting *Daubert* considerations. Determining which technique to use is paramount for the acceptability of soil analysis.

Procedures that can utilize known samples for the purpose of identifying questioned ones would have the most forensic potential. KNN was used in the current study and had a high correct classification rate from a training set of samples taken a year or more removed from the unknown samples, though misclassifications did occur. It is possible the amount of time between when the training set and unknown samples were collected influenced how class membership was assigned to misclassified samples, but not knowing the reason why samples were classifying as they were, either correctly or incorrectly, makes the utility of KNN limited. Also, KNN is a hard classifier, forcing a class membership to every unknown sample. This was demonstrated with the four marsh edge samples that were misclassified as treated yard. There was no way to know these samples were incorrectly classified other than the *a priori* knowledge of where they were collected. This information will obviously not be available in an actual case and evidentiary soils will be classified regardless of whether or not the habitat they originated from is present in the training set. If different origins of evidentiary soil are proposed by both the prosecution and defense, a training set could be developed encompassing all of them. Classification of unknown samples to crime scene ones would support the prosecution's case, though the defense can easily call into question the accuracy of the technique because no confidence or measure of how well those samples are classified is possible with KNN. Supervised classification techniques that can classify within a confidence interval or not assign class membership at all, would be more appropriate forensically, especially when the possible location where evidentiary soil originated is unknown.

The research presented was designed to examine variables influencing bacterial communities in soil; however, additional work is needed to further the investigation. It would be valuable forensically to better understand how bacterial communities are changing over time. Ongoing studies in the laboratory are sampling soils at shorter and longer time intervals to address this, with a goal of elucidating how the number of bacterial species and their abundance change. Also, the results of the depth study leave many questions unanswered. Would sampling a more 'natural' environment show different relationships among samples or patterns as depth increases? Additionally, how does the disruption of soil stratification influence bacterial communities? A study to examine this could include the collection of samples from undisturbed soil near a simulated burial site. Samples could also be collected in intervals from the shovel used and the buried item to allow for the direct comparison of disturbed and undisturbed bacterial communities, as well as association to known samples. Finally, the treated and untreated yards showed a distance-decay relationship using most of the statistics investigated. It would be interesting to see if similar types of habitats exhibit cognate distance-decay relationships, which could lead to standardized collection schemes for the type of habitat being sampled. Further investigation into the spatial change of communities in highly variable habitats, like the deciduous woods, is needed to determine if a distance-decay relationship exists and the feasibility of developing a standard collection procedure to capture it.

A great deal remains unknown regarding what influences soil bacterial populations. The research conducted in this thesis was designed to investigate a few of these variables in order to evaluate the utility of soil as evidence. Using next-generation sequencing coupled with an array of statistics, molecular analysis of soil in forensics is becoming more viable and the findings in this work build upon others to make its utilization in casework possible.

APPENDICES

APPENDIX 1. PHOTOGRAPHS OF SAMPLING LOCATIONS.

Figure 61. The deciduous woods (W) sampling site at Fenner Nature Center in Lansing, MI.

Figure 62. The marsh edge (M) sampling site at Fenner Nature Center in Lansing, MI.

Figure 63. The yard (Y) sampling site at Fenner Nature Center in Lansing, MI.

Figure 64. The field (F) sampling site at Fenner Nature Center in Lansing, MI.

Figure 65. The Lake Lansing beach (LL) sampling site in Haslett, MI.

Figure 66. The corn agricultural field (CAF) sampling site in East Lansing, MI.

Figure 67. The roadside (RS) sampling site in Lansing, MI.

Figure 68. The fallow agricultural field (FAF) sampling site in Perry, MI.

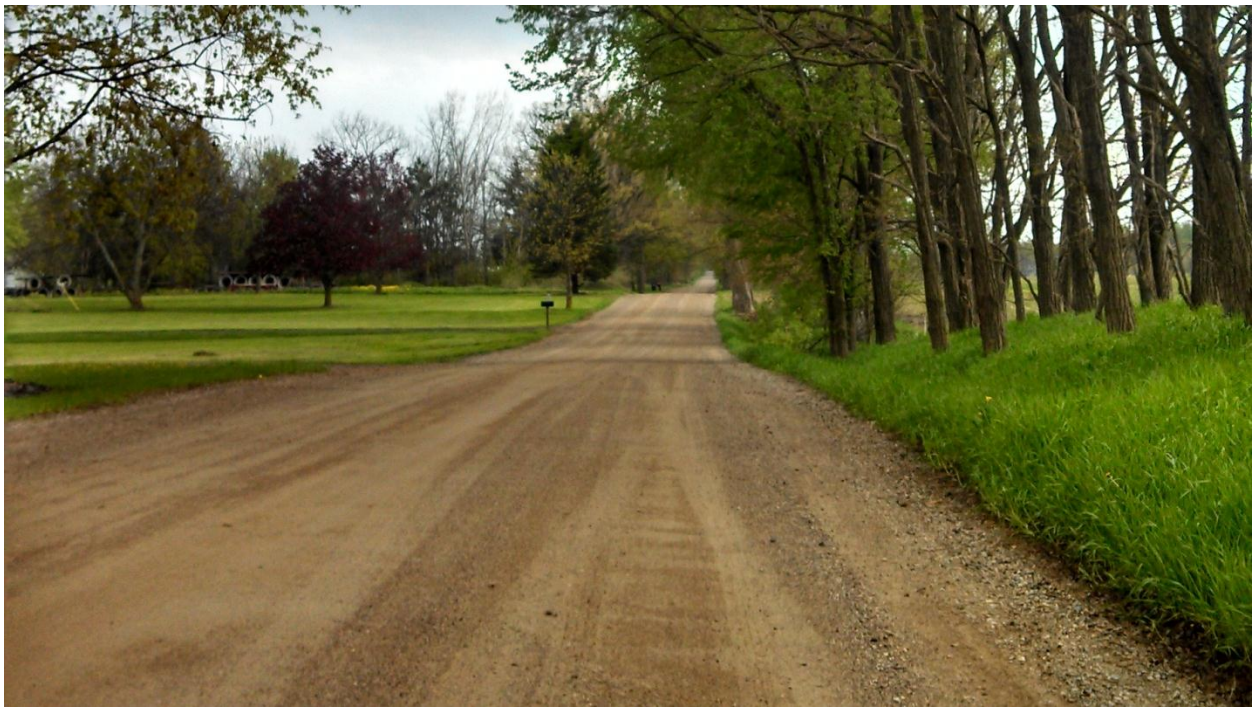

Figure 69. The dirt road (DR) sampling site in Perry, MI.

Figure 70. The coniferous forest (CF) sampling site at Woldumar Nature Center in Lansing, MI.

APPENDIX 2. INSTRUCTIONS FOR SEQUENCE PROCESSING WITH EXAMPLE FILES.

The file type outputted by the 454 sequencer is .sff, which contains the raw sequence data. For ease of explanation an example file (ABC.sff) will be 'processed' to illuminate the progression of the mothur steps. The bracketed sections are the codes inputted into mothur.

First, the sffinfo command is used to extract files from the .sff file [mothur>sffinfo(sff=ABC.sff)]. Three files are produced: a .fasta which contains the sequences, a .qual which contains the information about the quality of the base calls, and a .flows that can be used for additional analysis with the shhh.flows command. Next, the summary.seqs command is used [mothur>summary.seqs(fasta=ABC.fasta)]. This command can be applied at any point in the analysis to assess the number of sequences left and their sizes.

The trim.seqs command is then used, which instructs mothur to remove sequences from the fasta file based on specific characteristics, *e.g.* ambiguous base calls and quality base scores [mothur>trim.seqs(fasta=ABC.fasta, oligos=ABC.oligos, qfile=ABC.qual, maxambig=0, maxhomop=8, bdiffs=1, pdiffs=2, qwindowaverage=35, qwindowsize=50, filp=T)]. The oligos file in this step is designed specifically for the primers used during sequencing, including the primer sequences, their barcodes, and sample names affiliated with the primers. Additionally, using an oligos file breaks sequences into groups (.groups file) based on barcode and associated name in the oligos files. The maxambig command denotes any sequence with an ambiguous base call, which is removed, while the maxhomop removes any sequence with more than eight of the same bases in row. The bdiffs command allows one base difference between the inputted barcodes from the oligos file and the sequences and the pdiffs command accepts sequences with up to two base pair differences from the primer sequence. The qwindowaverage and qwindowsize commands establish the window size (50 base pairs) and the average quality score

(35) when screening sequences. Once the average quality score drops below 35 the sequence is trimmed at that location. Finally, the flip=T command reverses the sequence since the barcode is on the reverse primer. Again, the summary.seqs command can be used to evaluate the sequences left after the trim.seqs command is used [mothur>summary.seqs(fasta=ABC.trim.fasta)]. (Note each command adds an additional file extension that needs to be incorporated into the next command.)

Next, sequences are removed based on minimum length with the screen.seqs command [mothur>screen.seqs(fasta=ABC.trim.fasta, group=ABC.groups, minlength=250)]. The .groups file is generated when the trim.seqs command is used. The screen.seqs command removes any sequences that are under 250 base pairs. Along with the summary.seqs command, the count.groups command can be used to evaluate the number of sequences in each group. Additionally, if multiple files are being processed at the same time and they need to be combined, the merge.files command can be included [mothur>merge.files(input=fileA-fileB, output=fileAB)]. This command works with .fasta or .groups files but not with .sff files requiring some processing before files can be merged.

The number of sequences remaining after the trim.seqs and screen.seqs commands are executed can be further reduced with the sub.sample command. This command randomly samples the remaining sequences by group [mothur>sub.sample(fasta=ABC.trim.good.fasta, group=ABC.good.groups, size=6000, persample=T)]. The size parameter indicates that of the remaining sequences 6000 will be randomly chosen while persample=T commands that each sample will be subsampled to 6000 sequences. This command is at the discretion of the analyst and can be effectively used to speed up analysis based on the processing power of the computer utilized.

Redundant sequences can also be removed to speed up processing using the unique.seqs command [mothur>unique.seqs(fasta.ABC.trim.good.subsample.fasta)]. This command removes identical sequences from the .fasta file and returns a screened .fasta file and a .names file that documents which sequences were identical. Though the sequences are removed the number of them that were present is maintained through downstream processing. Again, the summary.seqs and count.groups commands can be used to assess the sequences remaining [mothur>summary.seqs(fasta=ABC.trim.good.subsample.unique.fasta, name=ABC.trim.good.subsample.names)].

Further removal of very similar sequences that are likely different due to sequencing error can be accomplished with the pre.cluster command [mothur>pre.cluster(fasta=ABC.trim.good.subsample.unique.good.filter.fasta, name=ABC.trim.good.subsample.names, diffs=1)]. This command ranks the sequences in order of abundance, then looks for rare sequences within a certain threshold of the abundant sequences that are likely due to sequencing error and removes them. In this case sequences that are within one base difference of abundant sequences are removed as being erroneous.

The final steps in processing sequences are the creation of a distance matrix and then the clustering of those sequences. To create the distance matrix the dist.seqs command is used [mothur>dist.seqs(fasta=ABC.trim.good.subsample.unique.good.filter.precluster.fasta, cutoff=0.30)]. The cutoff parameter can be used to reduce the number of distances being saved. In this case, distances larger than 0.30 are not considered. Distance refers to a measure of how far apart the sequences are from each other, similar to a Euclidian distance. This cutoff is meant to preserve hard drive space when saving these extremely large files and allows the process to be accomplished on personal computers. The remaining sequences are clustered using the cluster

command [mothur>cluster(column=ABC.trim.good.subsample.unique.good.filter.precluster.dist, name=ABC.trim.good.subsample.unique.good.filter.precluster.names, method=average)]. This step clusters the sequences based on the average neighbor method and outputs a file containing sequences binned into OTUs. This file cannot be viewed in excel, so a .shared file needs to be generated with the make.shared command [mothur>make.shared(list=ABC.trim.good.subsample.unique.good.filter.precluster.an.list, group=ABC.good.subsample.good.groups)]. A .an.shared file is generated which, when viewed in excel, indicates how many sequences are in each bin for each cutoff level at which they clustered for every group.

APPENDIX 3. INSTRUCTIONS FOR CLASSIFYING SEQUENCES.

Sequences are classified using the classify.seqs command [mothur>classify.seqs(fasta= ABC.trim.good.subsample.unique.good.filter.precluster.fasta, template=silva.bacteria.fasta, taxonomy=silva.bacteria.silva.taxonomy, group=ABC.good.subsample.good.groups)]. This command outputs a taxonomy file that has all sequences classified to the lowest possible taxonomy in the previously assigned groups. To classify the OTUs the classify.otu command is used

[mothur>classify.otu(taxonomy=ABC.trim.good.subsample.unique.good.filter.precluster.silva.taxonomy, list=ABC.trim.good.subsample.unique.good.filter.precluster.an.list)]. This command outputs a file that contains each OTU classified to lowest possible taxonomic level.

APPENDIX 4. INSTRUCTIONS FOR CALCULATION OF BCDI AND SDC.

Dissimilarities were calculated in mothur using the Bray-Curtis dissimilarity index and Sørensen-Dice coefficient using the summary.shared command [mothur>summary.shared(shared=ABC.trim.good.subsample.unique.good.filter.an.shared, calc=braycurtis-sorclass)]. This outputs a .shared file that has all the pairwise comparisons for each sample in the file analyzed. A square matrix was then developed from these comparisons.

APPENDIX 5. INSTRUCTIONS FOR RUNNING ∫-LIBSHUFF AND UNIFRAC COMPARISONS.

Before LIBSHUFF can be run, sequences need to be redistanced into a phylogeny inference package (phylip) formatted square matrix. This is done with the following modification from the previous code [mothur>dist.seqs(fasta=ABC.trim.good.subsample.unique.good.filter.precluster.fasta, output=square)]. This command will output a new .square.dist file. With this new .dist file LIBSHUFF can be run with the following input [mothur>libshuff(phylip=ABC.trim.good.subsample.unique.good.filter.precluster.square.dist, group=ABC.good.subsample.good.groups)]. Depending on the number of pairwise comparisons being made how many iterations that are run can be increased for p-values smaller than 0.0001, default is 10,000.

To run UniFrac a phylogenetic tree is first developed with the following input [mothur>clearcut(phylip=ABC.trim.good.subsample.unique.good.filter.precluster.square.dist)]. This will output a .tre file used in the next step. A comparison of all samples can first be run within the developed tree to determine if any of the groups are statistically different. If a nonsignificant value is determined in this step UniFrac can be stopped since pairwise comparisons would find again that no comparisons would be significant. To run the comparison of all samples input the following code [mothur>unifrac.unweighted(tree= ABC.trim.good.subsample.unique.good.filter.precluster.square.tre, group=ABC.good.subsample.good.groups, random=t)]. If statistical significance is found, pairwise comparisons for all groups can be completed with a slightly modified code [mothur>unifrac.unweighted(tree= ABC.trim.good.subsample.unique.good.filter.precluster.square.tre,

group=ABC.good.subsample.good.groups, random=t, groups=all)]. Like LIBSHUFF the number of iterations can be increased for smaller p-values, default is 1,000.

APPENDIX 6. RELATIVE ABUNDANCE CHARTS FOR REMAINING STUDIES AND ASSOCIATED LEGENDS.

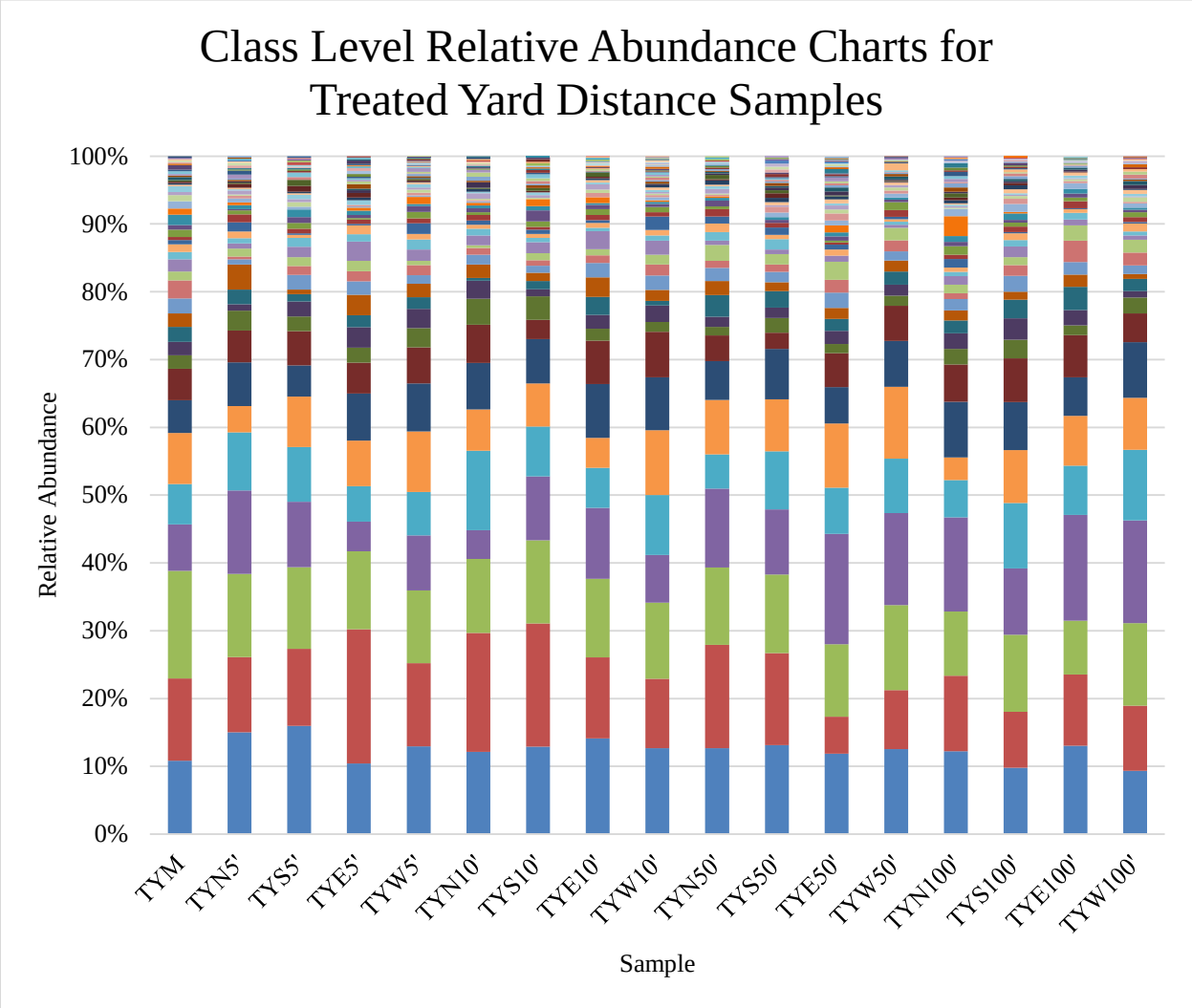

Figure 71. Class level relative abundance charts for treated yard distance samples representing 83 bacterial classes. Samples share a majority of bacterial classes up to 95% total relative abundance though the individual abundance of each class is variable sample to sample. No sample is noticeably different from the rest in general abundance of bacterial classes. See Table 3 for site names corresponding to abbreviations and Appendix 6 Figure 77 for legend of bacterial classes.

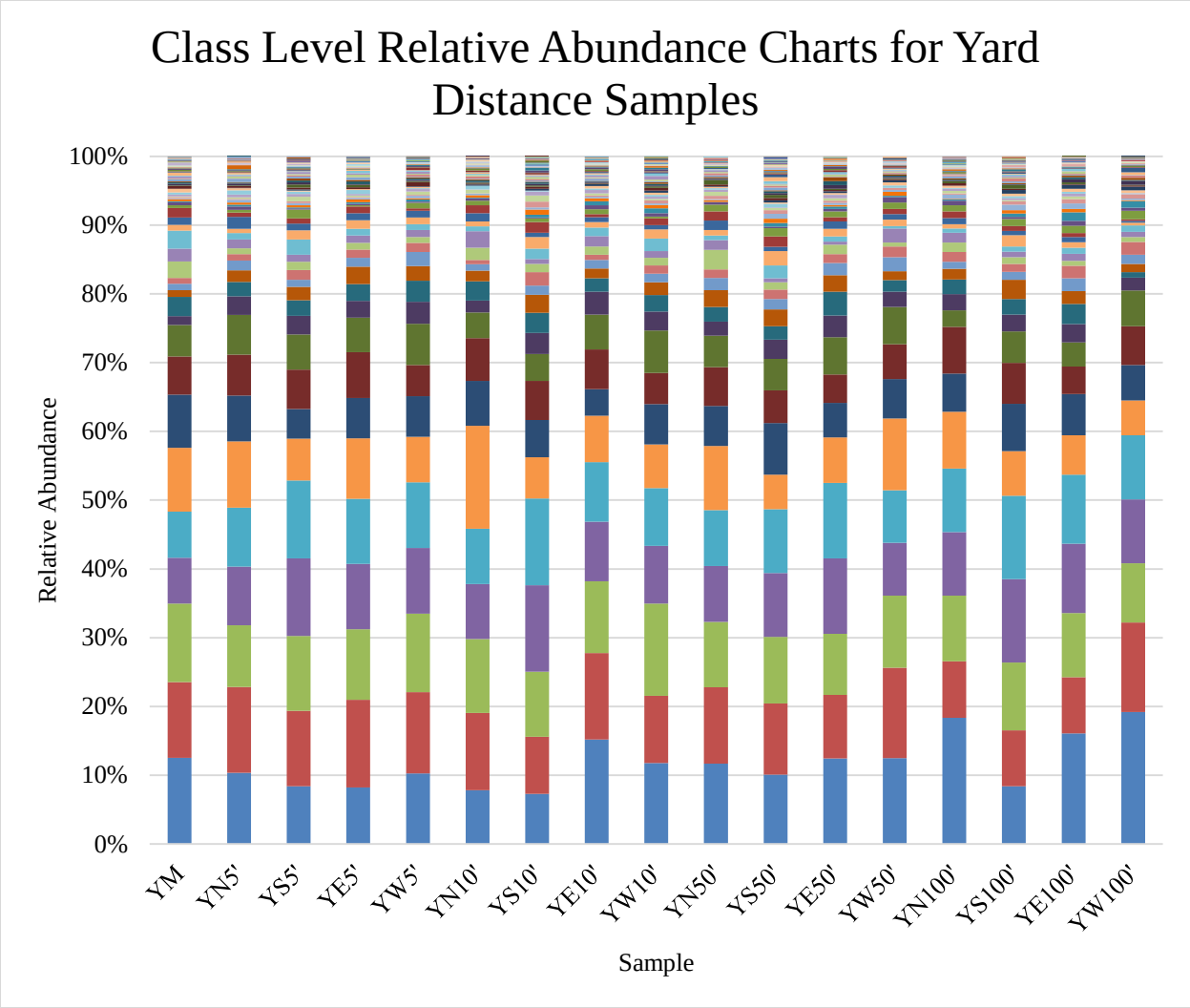

Figure 72. Class level relative abundance charts for yard distance samples representing 89 bacterial classes. Samples share a majority of bacterial classes up to 95% total relative abundance though the individual abundance of each class is variable sample to sample. No sample is noticeably different from the rest in general abundance of bacterial classes. See Table 3 for site names corresponding to abbreviations and Appendix 6 Figure 77 for legend of bacterial classes.

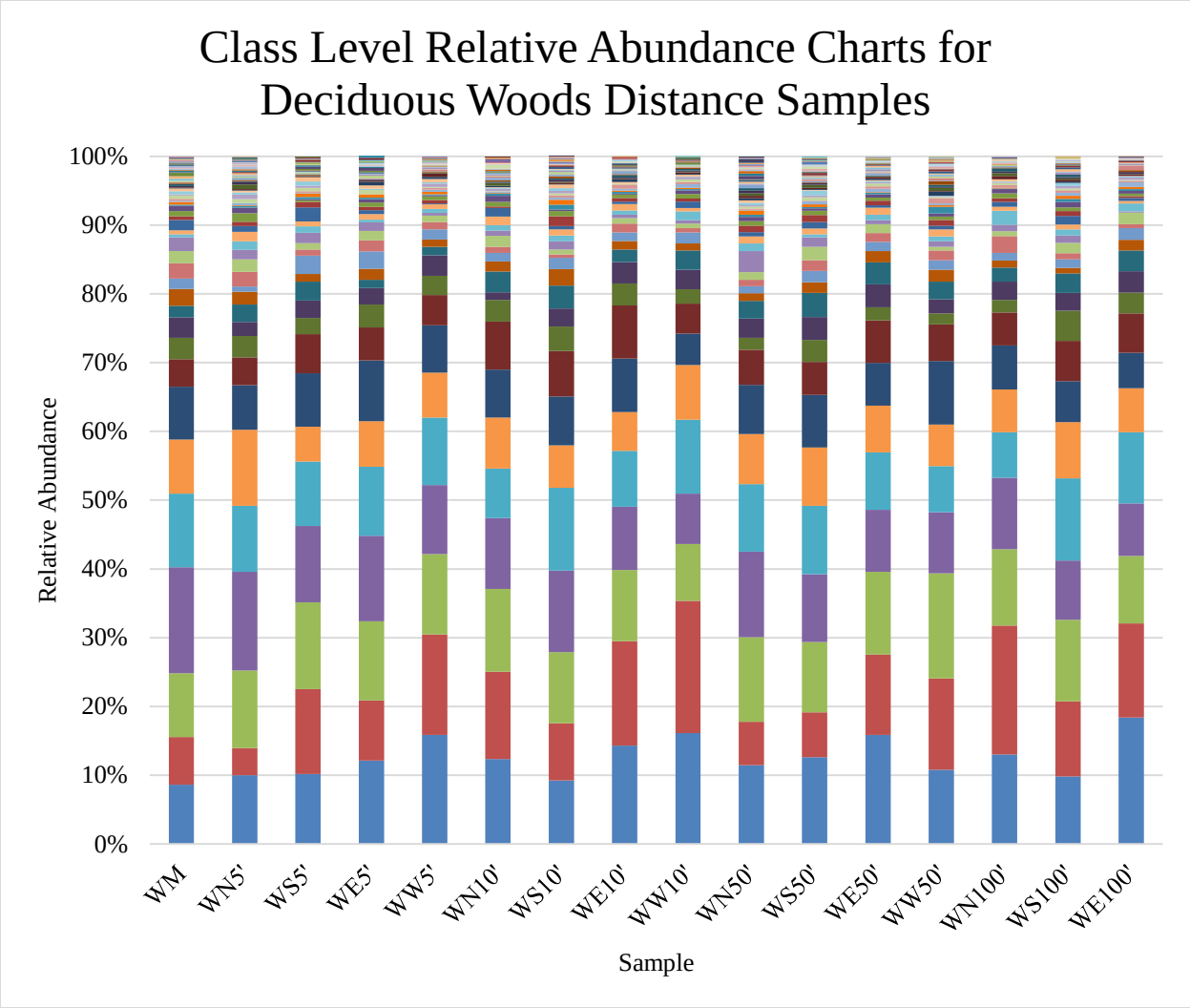

Figure 73. Class level relative abundance charts for deciduous woods distance samples representing 88 bacterial classes. Samples share a majority of bacterial classes up to 95% total relative abundance though the individual abundance of each class is variable sample to sample. No sample is noticeably different from the rest in general abundance of bacterial classes. See Table 3 for site names corresponding to abbreviations and Appendix 6 Figure 77 for legend of bacterial classes.

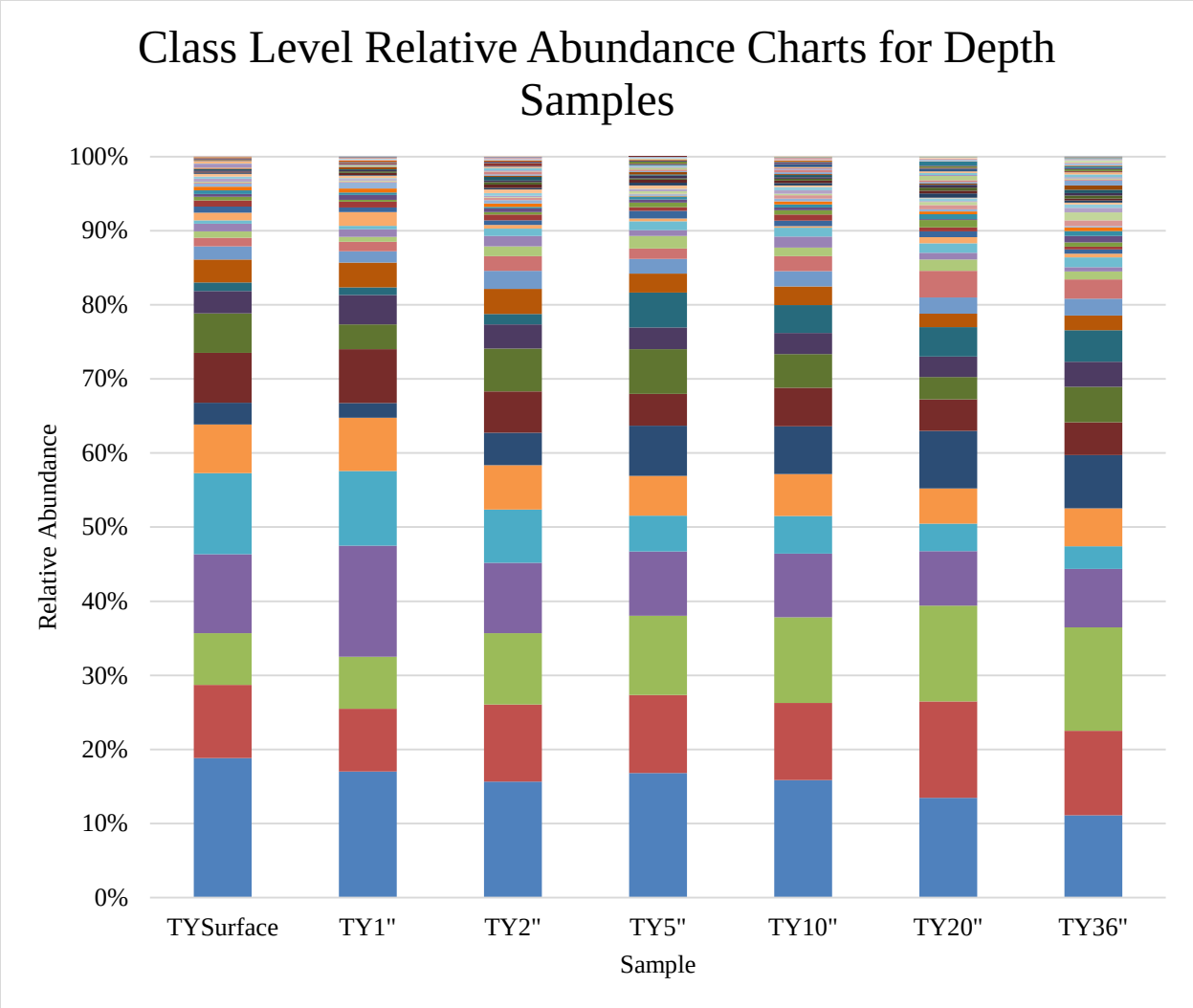

Figure 74. Class level relative abundance charts for depth samples representing 62 bacterial classes. Samples share a majority of bacterial classes up to 95% total relative abundance though the individual abundance of each class is variable sample to sample. No sample is noticeably different from the rest in general abundance of bacterial classes. See Table 4 for site names corresponding to abbreviations and Appendix 6 Figure 77 for legend of bacterial classes.

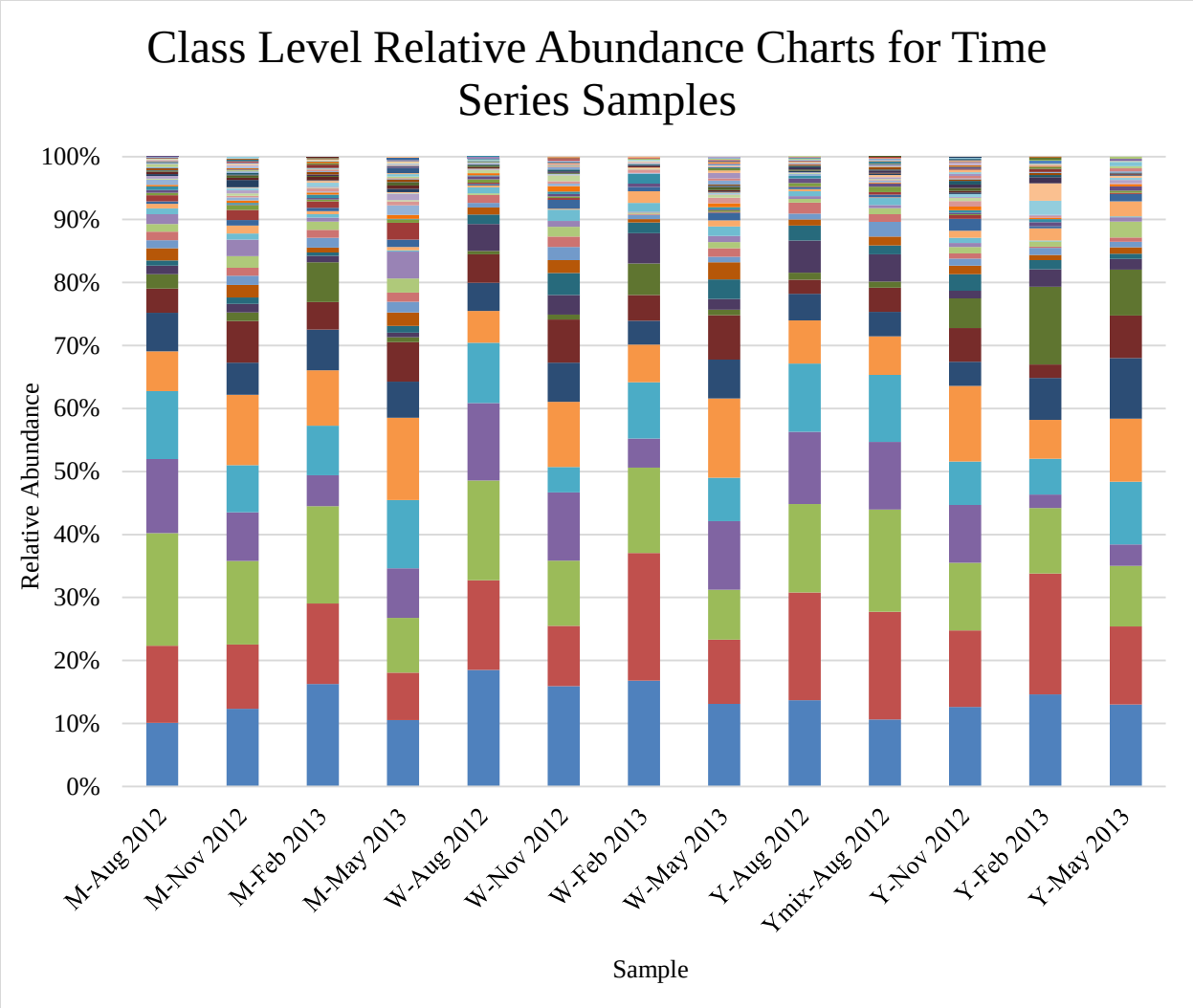

Figure 75. Class level relative abundance charts for time series samples representing 87 bacterial classes. Samples share a majority of bacterial classes up to 95% total relative abundance though the individual abundance of each class is variable sample to sample. No sample is noticeably different from the rest in general abundance of bacterial classes. See Table 2 for site names corresponding to abbreviations and Appendix 6 Figure 77 for legend of bacterial classes.

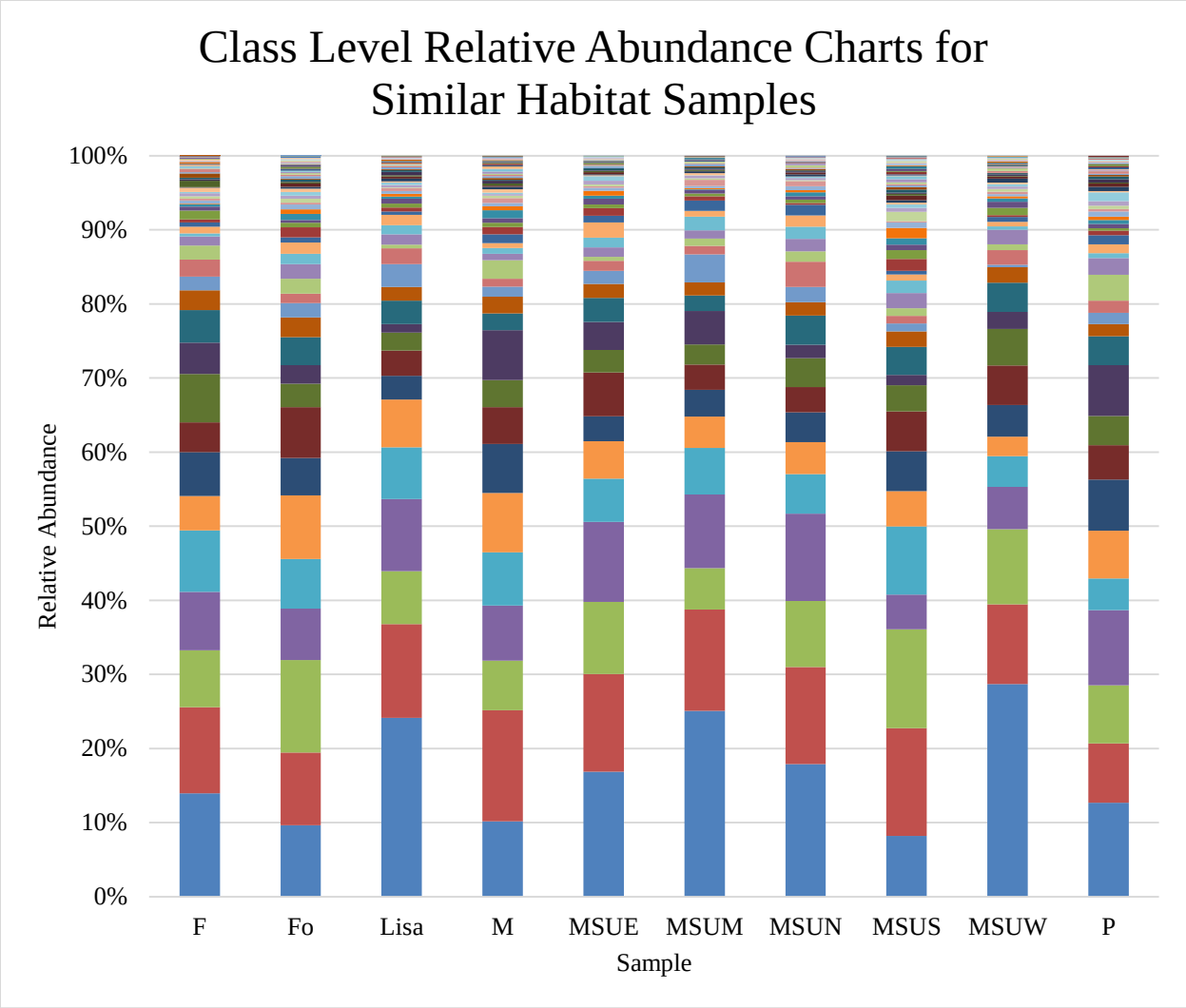

Figure 76. Class level relative abundance charts for similar habitat samples representing 67 bacterial classes. Samples share a majority of bacterial classes up to 95% total relative abundance though the individual abundance of each class is variable sample to sample. No sample is noticeably different from the rest in general abundance of bacterial classes. See Table 5 for site names corresponding to abbreviations and Appendix 6 Figure 77 for legend of bacterial classes.

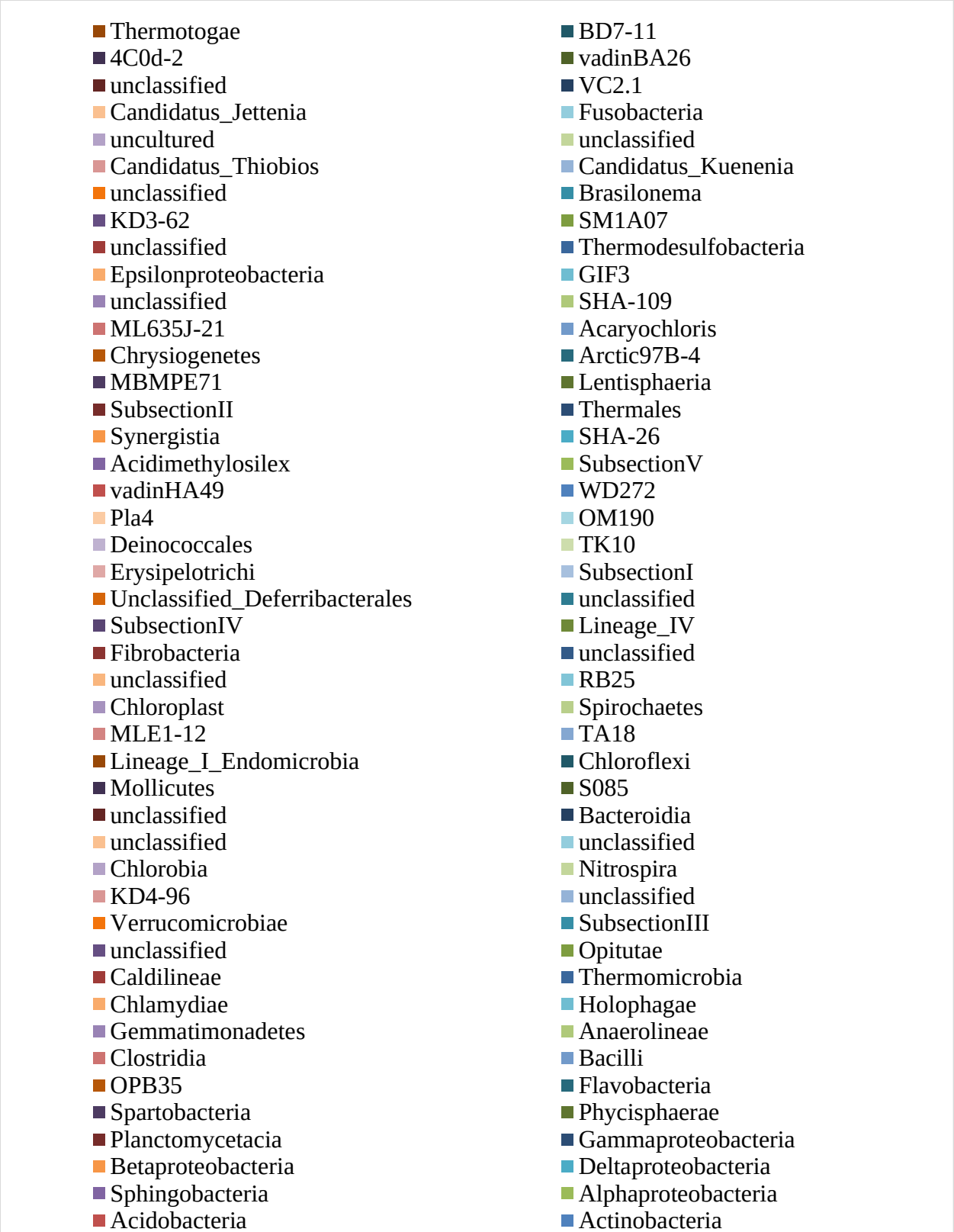

Figure 77. Legend of bacterial classes for total relative abundance charts.

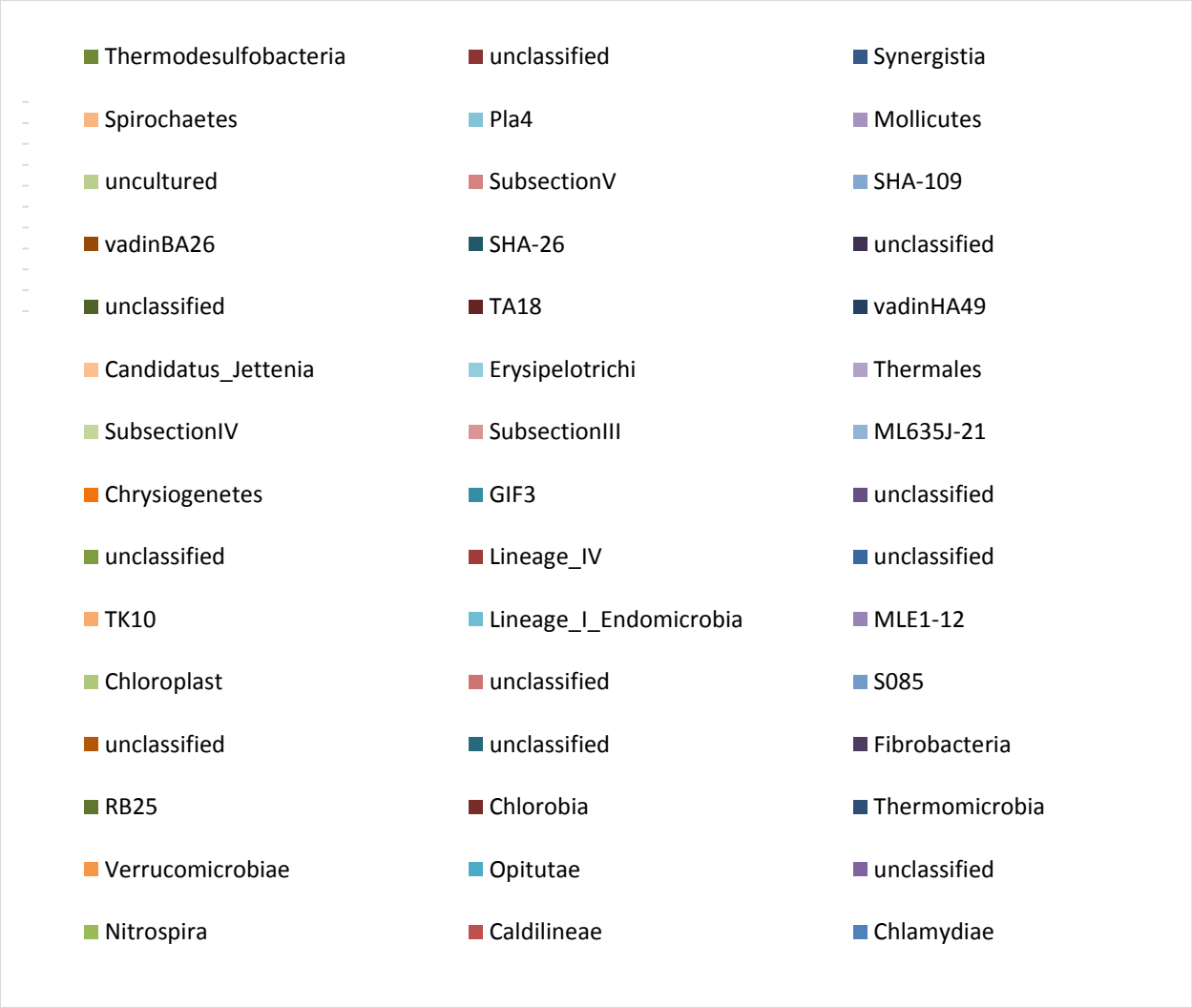

Figure 78. Legend of least abundant bacterial classes.

APPENDIX 7. ADDITIONAL NMDS DIAGRAMS FOR BIOLOGICAL REPLICATE SAMPLES AND SDC.

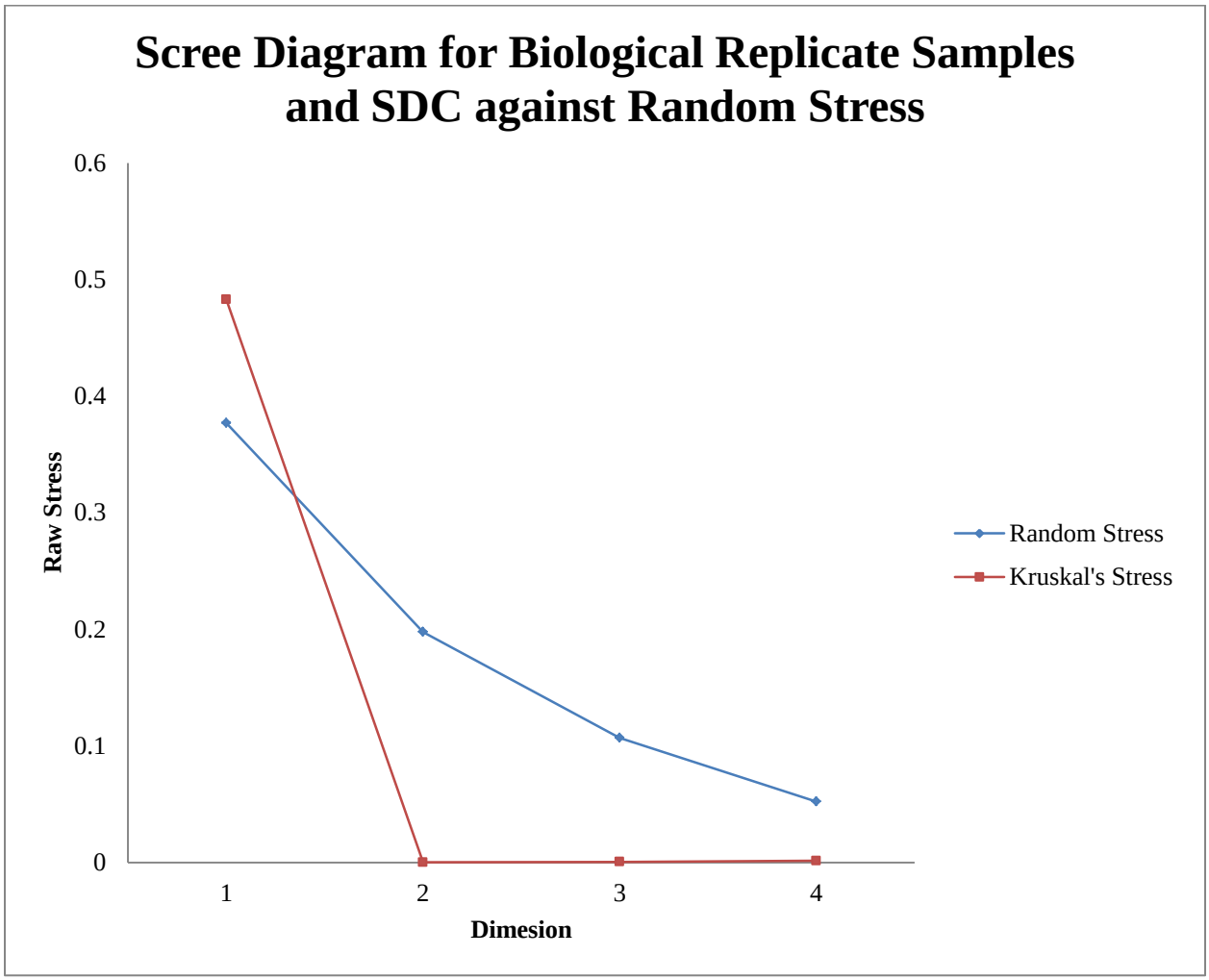

Figure 79. Scree diagram developed by NMDS for the final configurations of the ten biological replicate soil samples and SDC over four dimensions. Random stress was used as a threshold for the acceptance of final configurations. The one dimension configuration was the only to exceed the threshold and was rejected. The elbow in the curve is at two dimensions.

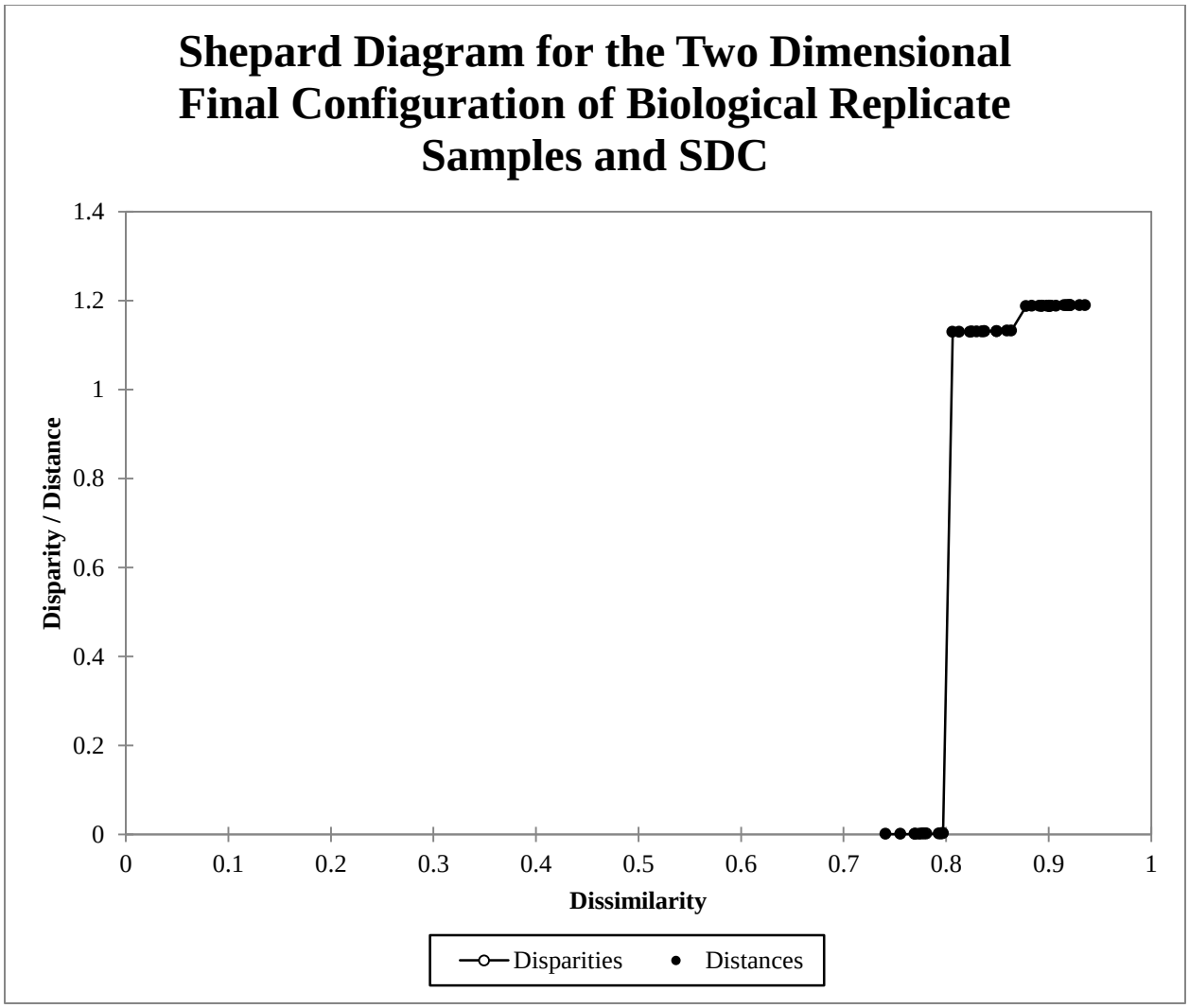

Figure 80. Shepard diagram for the two dimensional final configuration developed from SDC of biological replicate soil samples. All distances fall nearly on top of their corresponding disparities indicating good correlation of the two in the configuration.

APPENDIX 8. ADDITIONAL DENDROGRAMS OF BIOLOGICAL REPLICATE SAMPLES AND SDC.

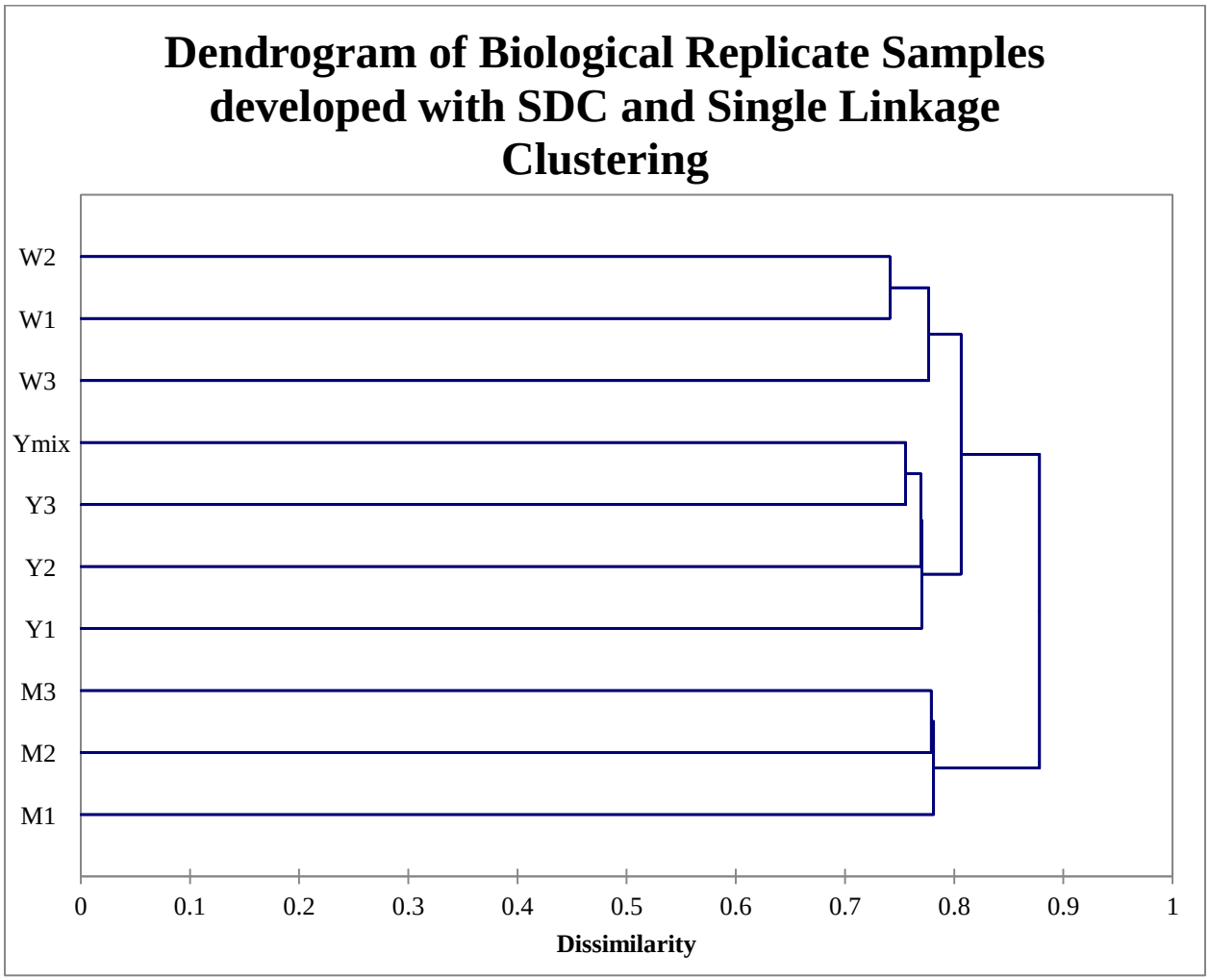

Figure 81. Dendrogram of biological replicate soil samples developed with SDC and single linkage clustering. Three clusters are formed by habitat with replicate samples more similar to each other than to other habitats. The yard samples cluster at 0.777 while the deciduous woods group at 0.770. The marsh edge samples are most dissimilar from the other two habitats clustering at 0.781 and forming a cluster with the others at 0.878. See Table 2 for site names corresponding to abbreviations.

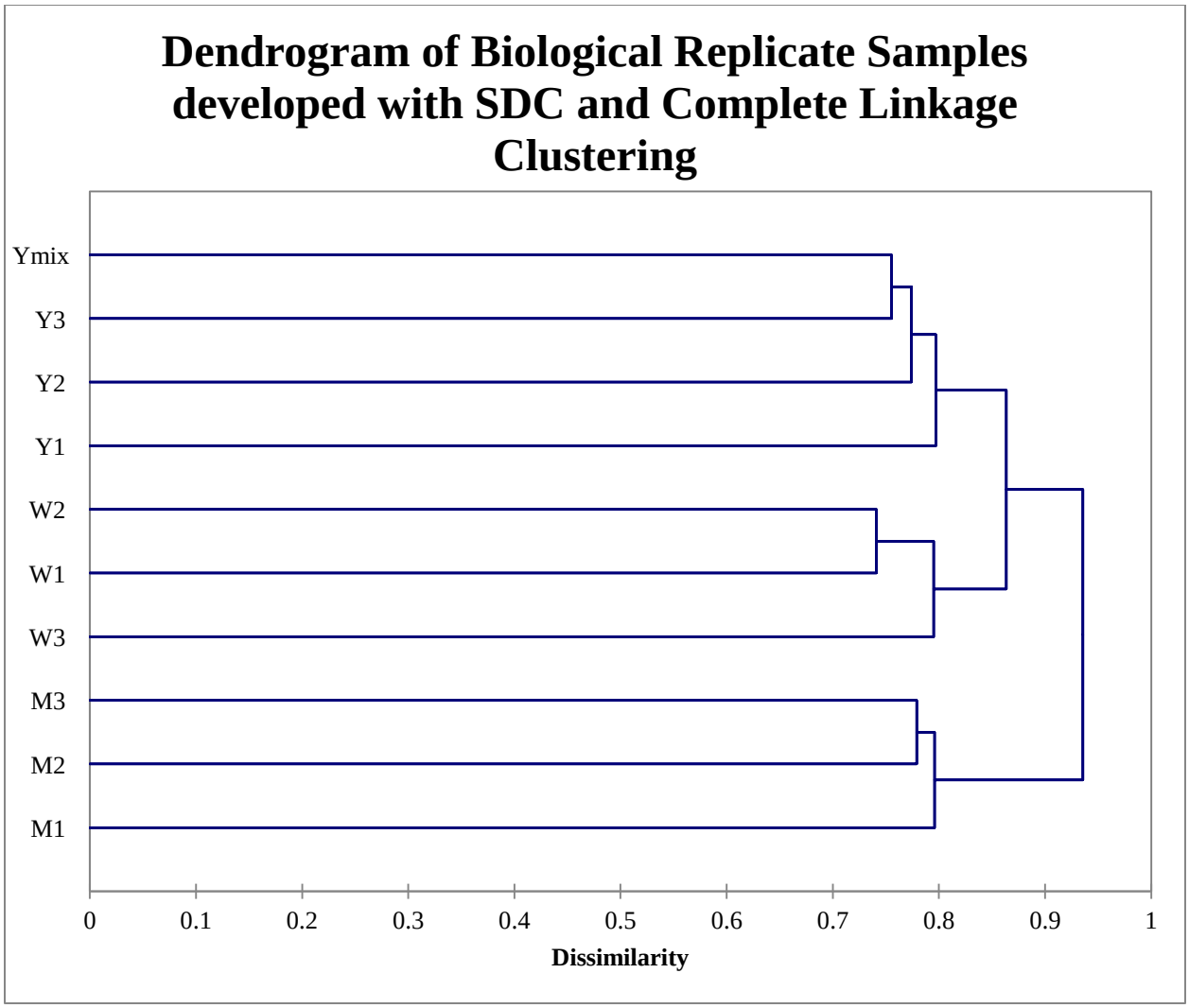

Figure 82. Dendrogram of biological replicate soil samples developed with SDC and complete linkage clustering. Three clusters are formed by habitat with replicate samples more similar to each other than to other habitats. Yard and deciduous woods samples cluster independently at 0.797 and 0.795, respectively. The marsh edge samples group at 0.796 and are most dissimilar from the other two habitats forming a cluster with them at 0.936. See Table 2 for site names corresponding to abbreviations.

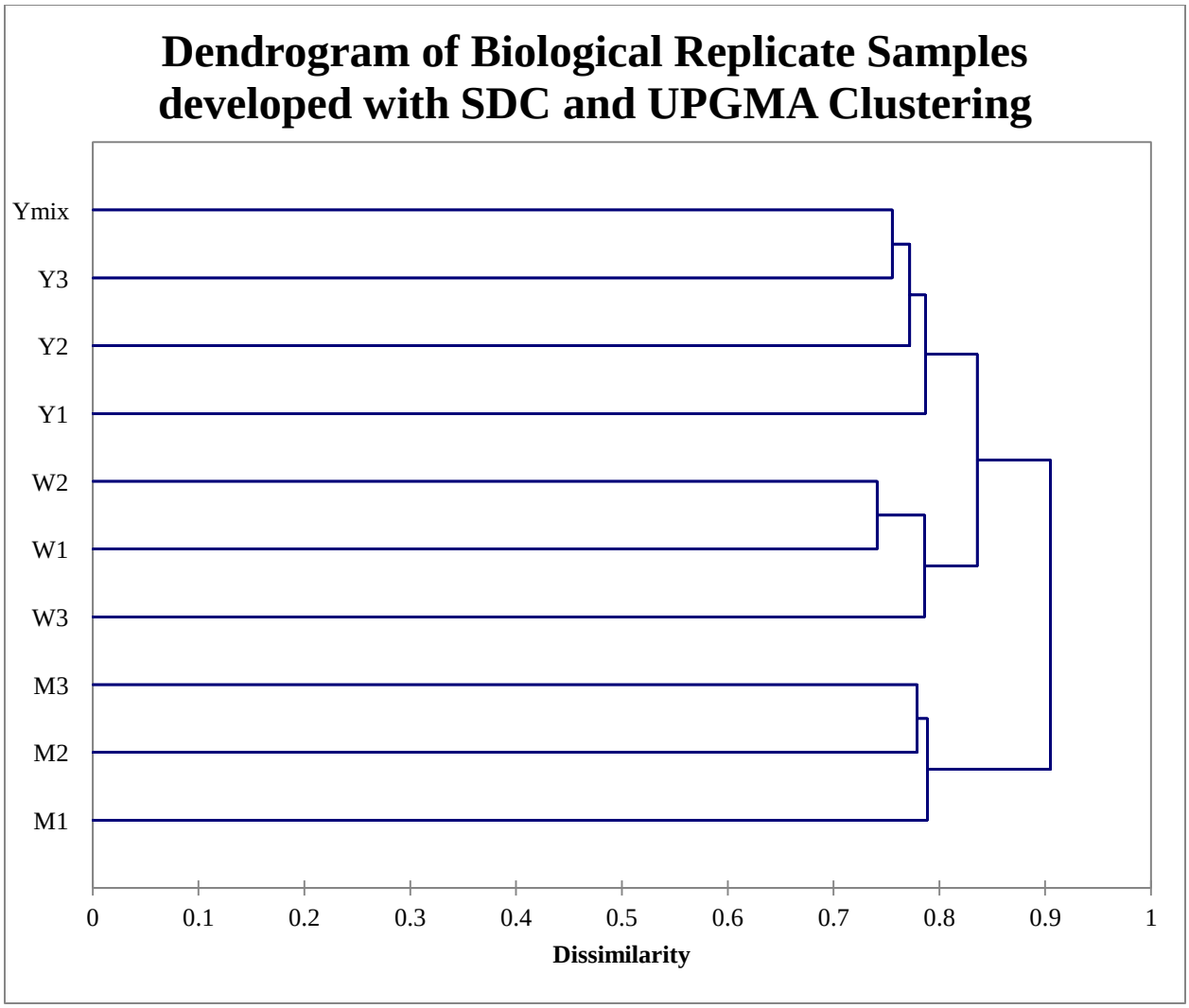

Figure 83. Dendrogram of biological replicate soil samples developed with SDC and UPGMA clustering. Three clusters are formed by habitat with replicate samples more similar to each other than to other habitats. Yard and deciduous woods samples cluster independently at 0.787 and 0.786, respectively. The marsh edge samples group at 0.788 and are most dissimilar from the other two habitats forming a cluster with them at 0.905. See Table 2 for site names corresponding to abbreviations.

APPENDIX 9. ADDITIONAL NMDS DIAGRAMS FOR HABITAT DISTANCE SAMPLES.

APPENDIX 9-1. SCREE AND SHEPARD DIAGRAMS FOR TREATED YARD DISTANCE SAMPLES.

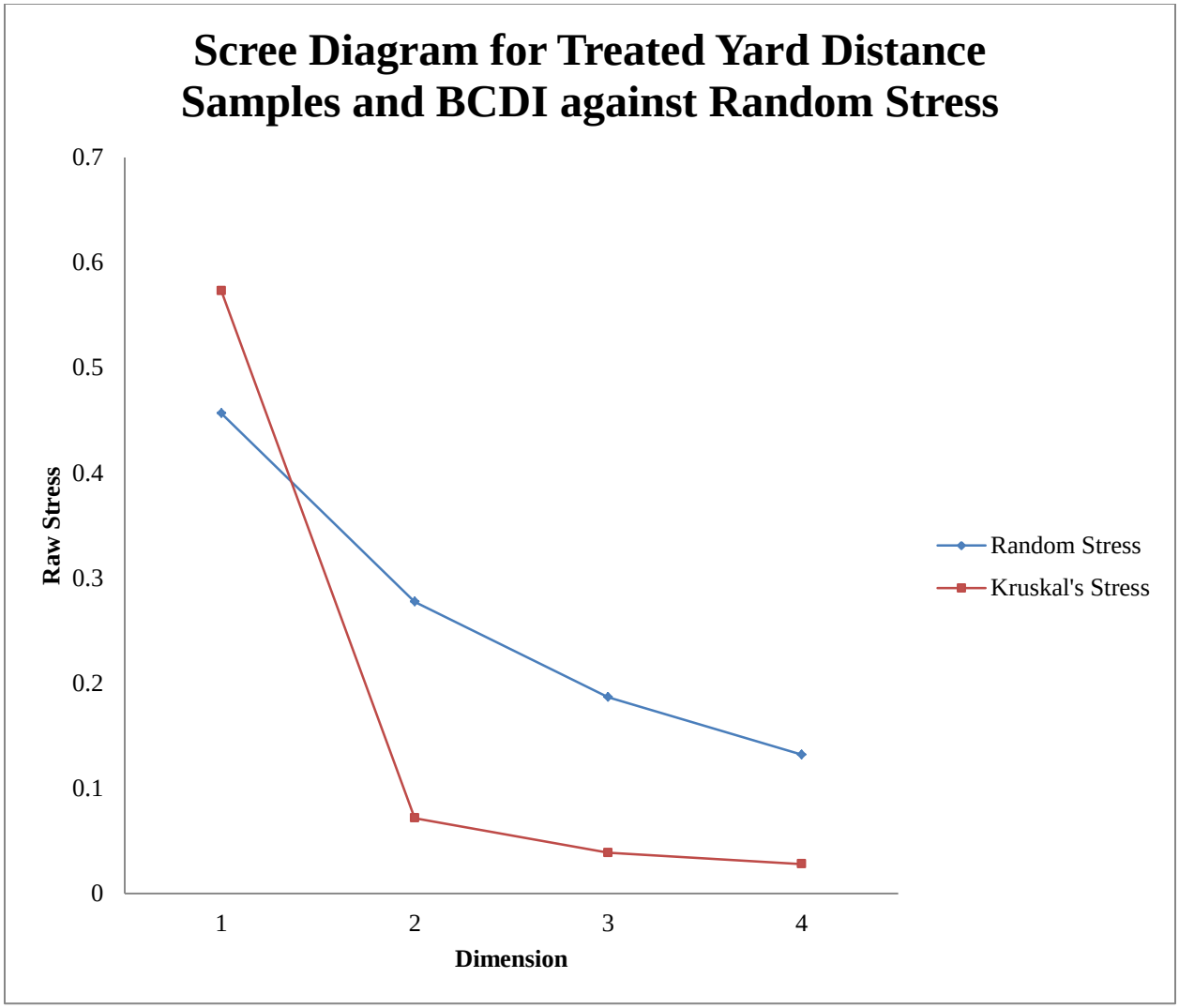

Figure 84. Scree diagram developed by NMDS for the final configurations of treated yard distance soil samples and BCDI over four dimensions. Random stress was used as a threshold for the acceptance of final configurations. The one dimension configuration was the only to exceed the threshold and was rejected. The elbow in the curve is at two dimensions.

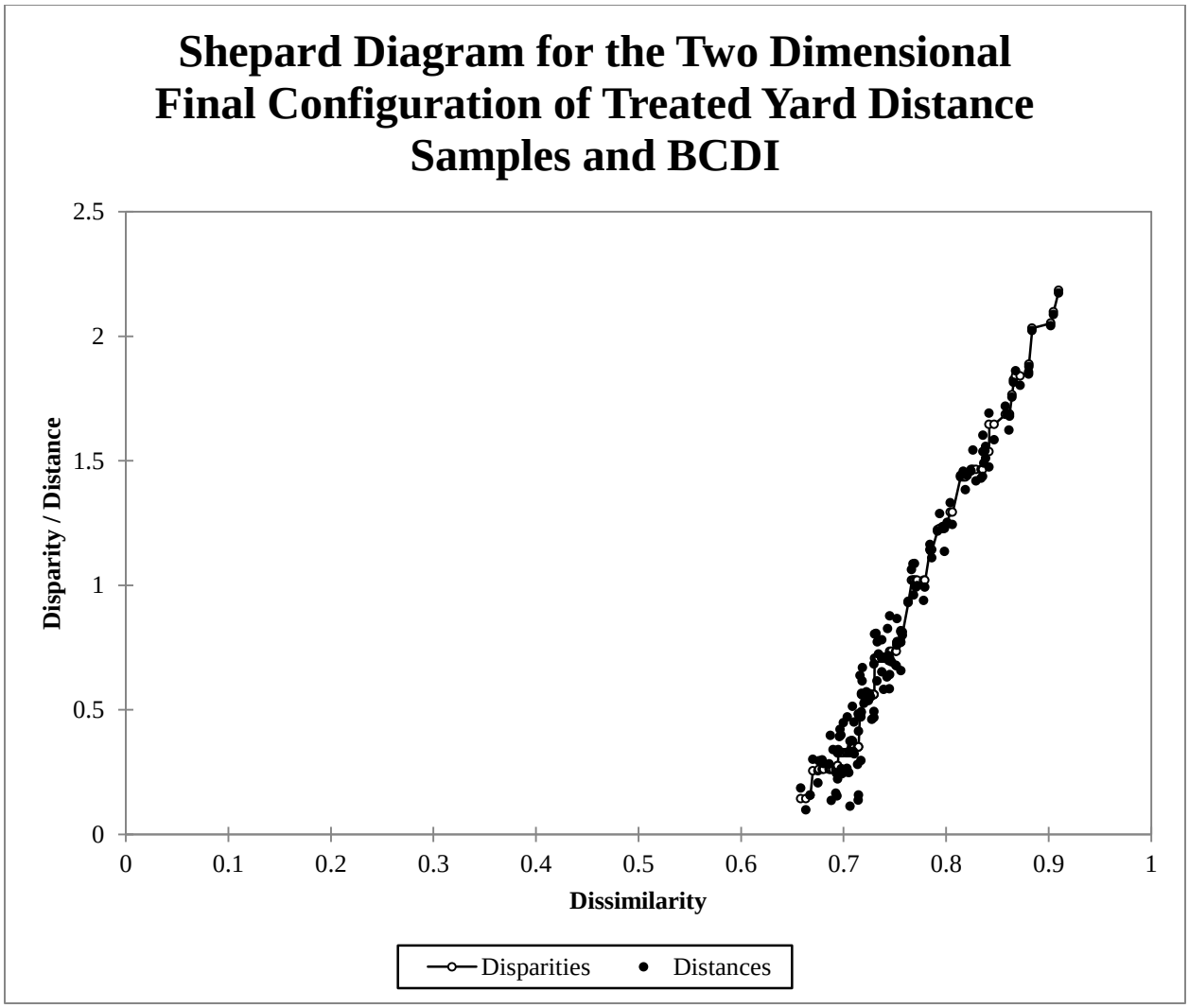

Figure 85. Shepard diagram for the two dimensional final configuration developed from BCDI of treated yard distance soil samples. All distances fall close to their corresponding disparities indicating good correlation of the two in the configuration.

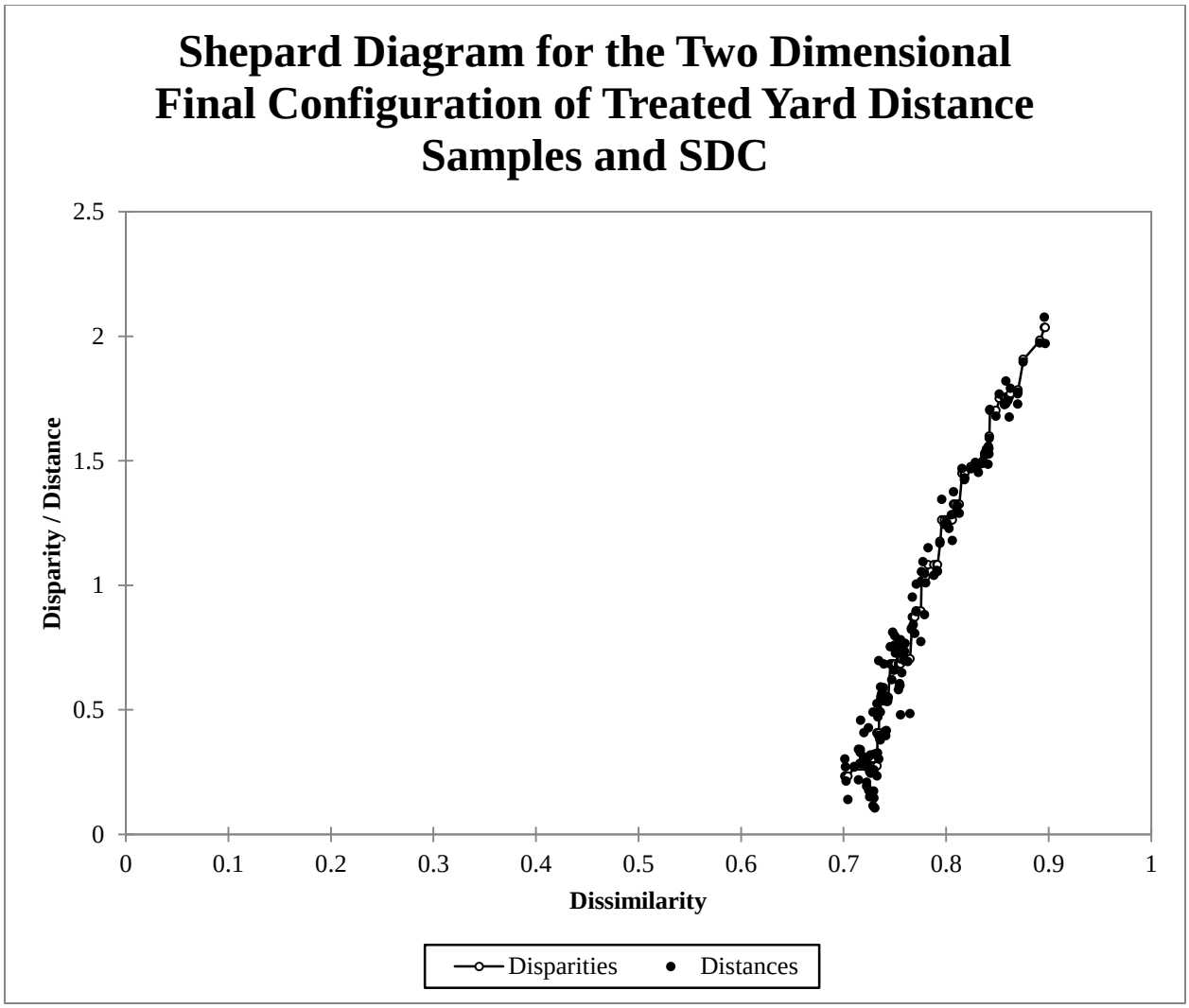

Figure 86. Shepard diagram for the two dimensional final configuration developed from SDC of treated yard distance soil samples. All distances fall close to their corresponding disparities indicating good correlation of the two in the configuration.

APPENDIX 9-2. SCREE AND SHEPARD DIAGRAMS FOR YARD DISTANCE SAMPLES.

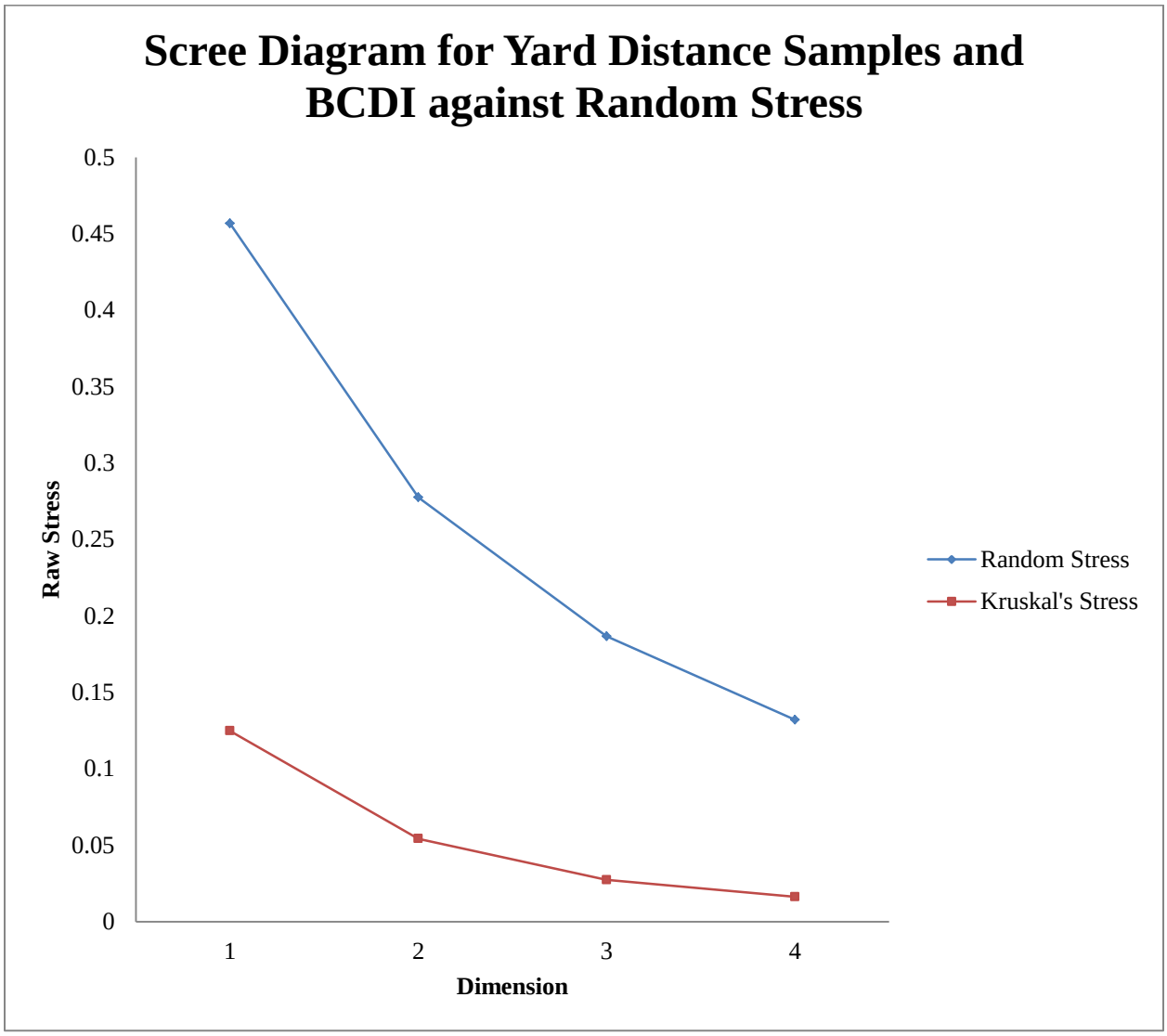

Figure 87. Scree diagram developed by NMDS for the final configurations of yard distance soil samples and BCDI over four dimensions. Random stress was used as a threshold for the acceptance of final configurations. No configuration exceeded the threshold though the stress for the one dimension plot was higher than the others. No elbow is noticeable at any dimension; however, for consistency two dimensional plots were chosen.

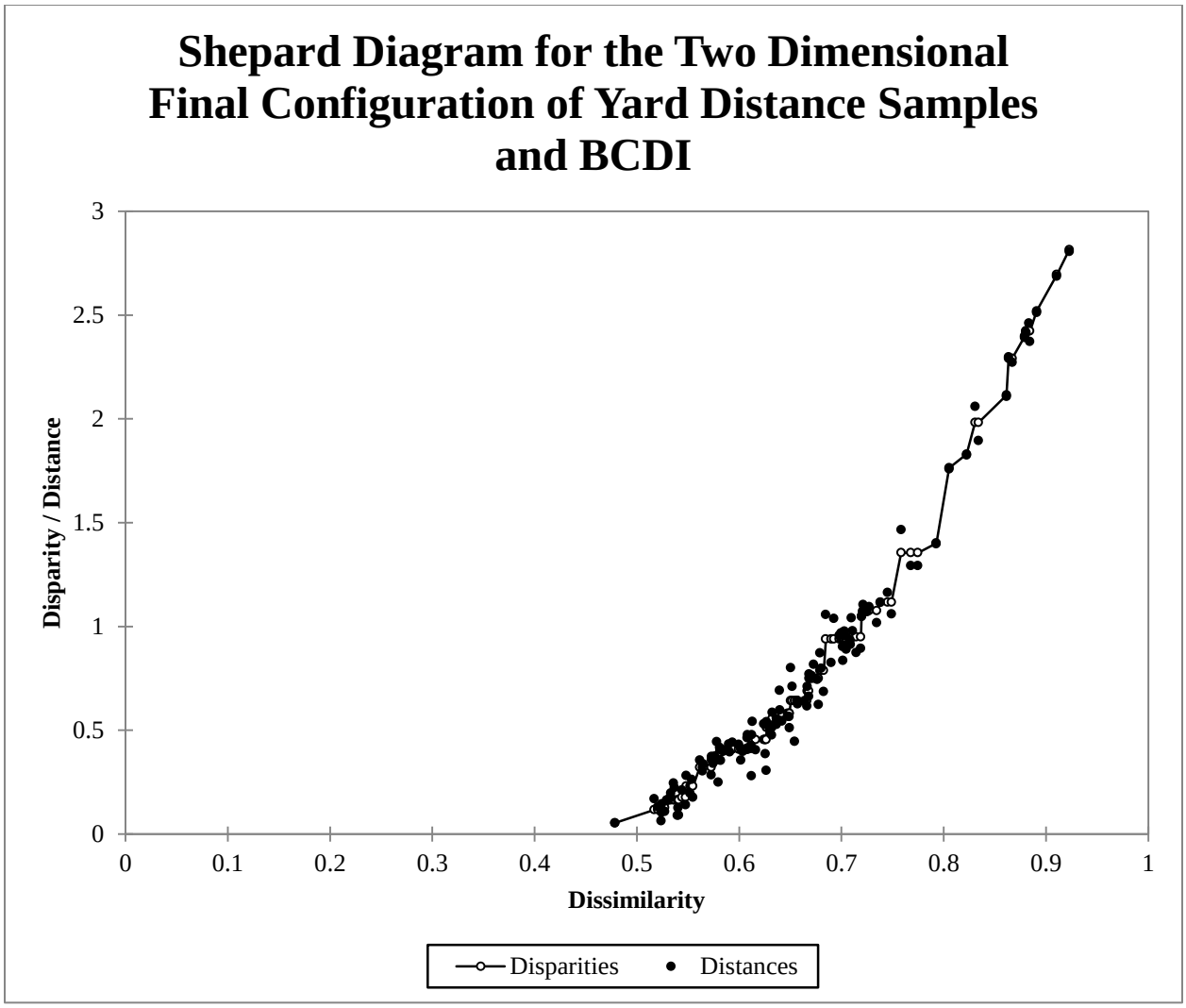

Figure 88. Shepard diagram for the two dimensional final configuration developed from BCDI of yard distance soil samples. All distances fall close to their corresponding disparities indicating good correlation of the two in the configuration.

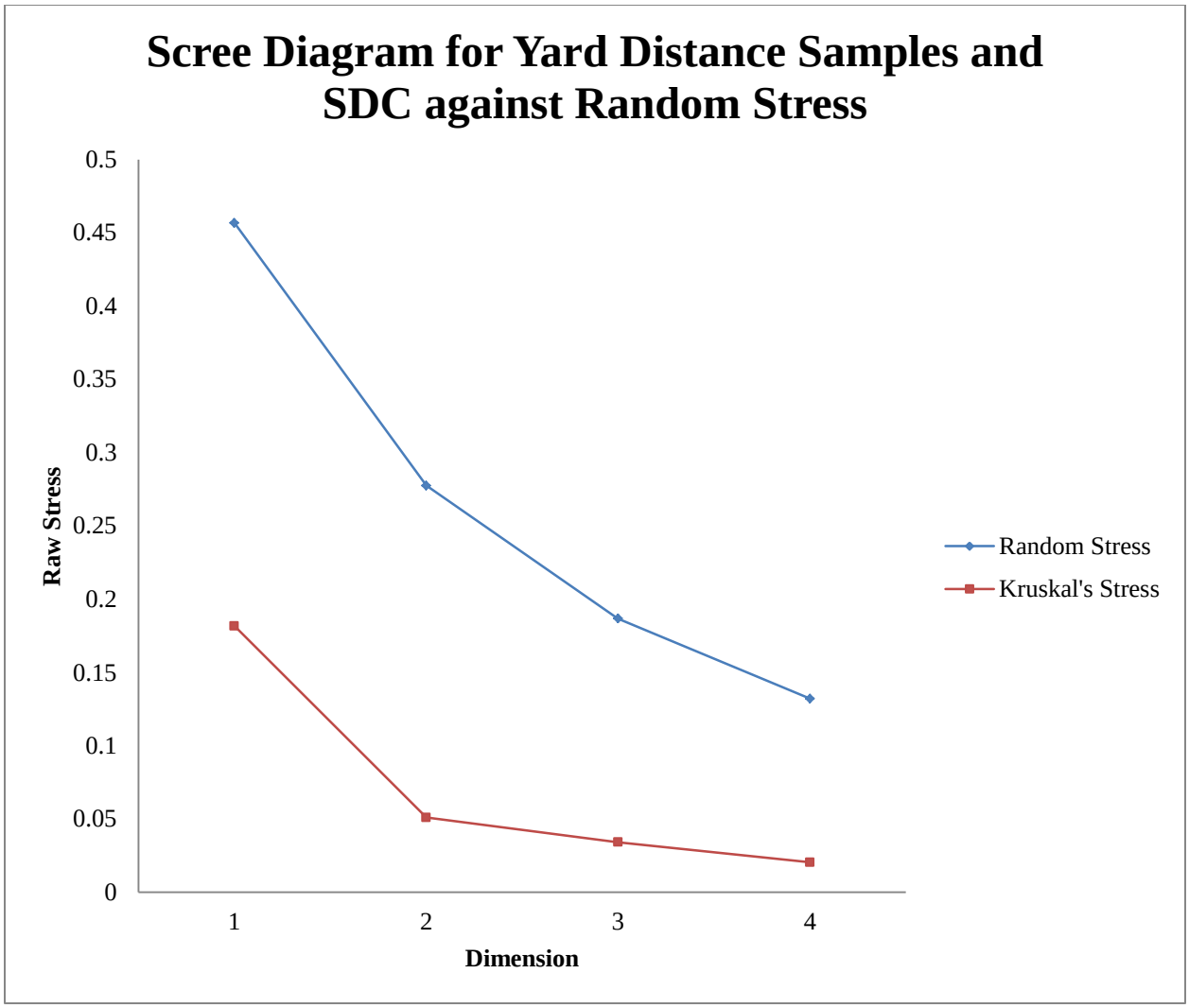

Figure 89. Scree diagram developed by NMDS for the final configurations of yard distance soil samples and SDC over four dimensions. Random stress was used as a threshold for the acceptance of final configurations. No configuration exceeded the threshold though the stress for the one dimension plot was higher than the others. The elbow in the curve is at two dimensions.

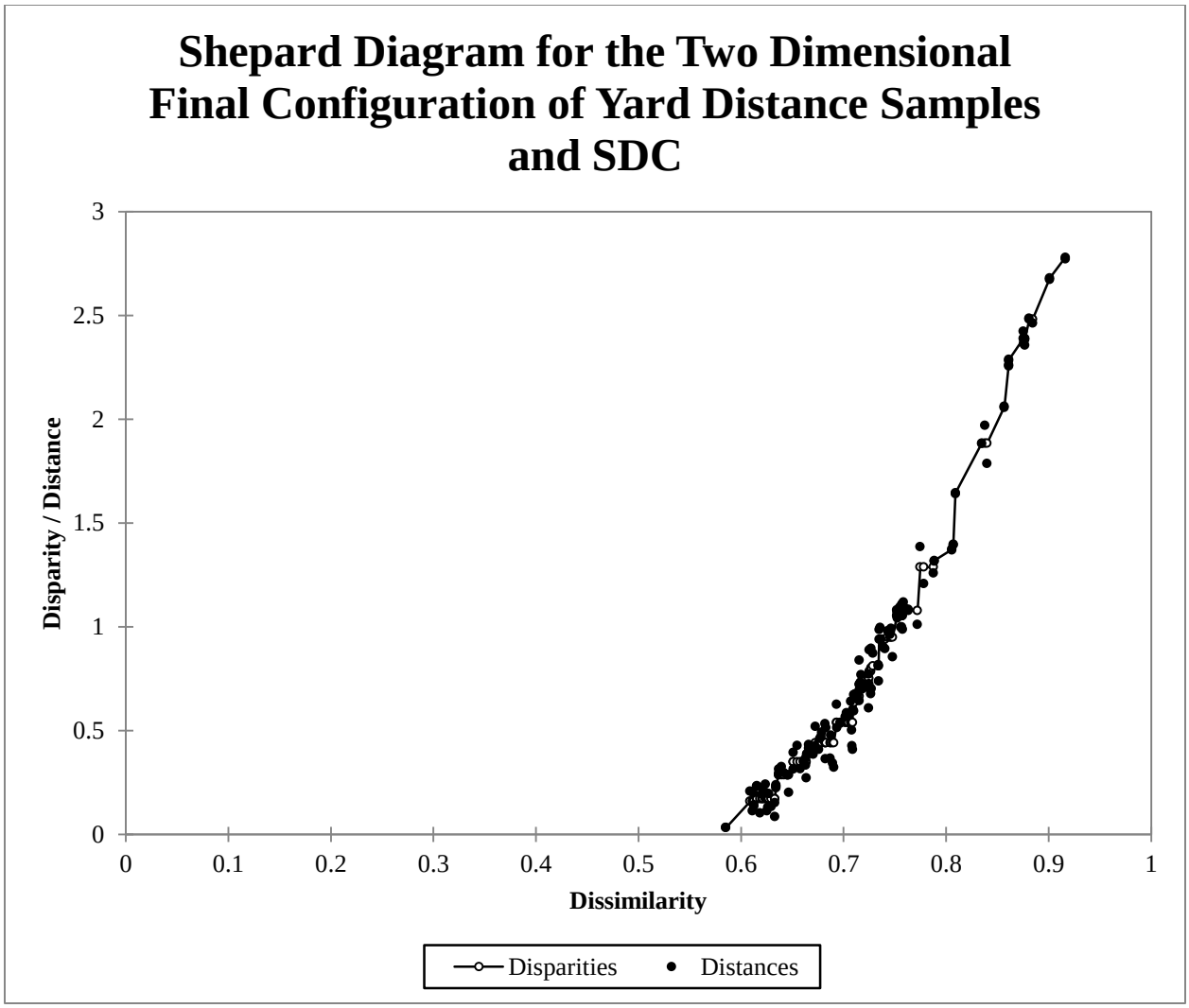

Figure 90. Shepard diagram for the two dimensional final configuration developed from SDC of yard distance soil samples. All distances fall close to their corresponding disparities indicating good correlation of the two in the final configuration**.**

APPENDIX 9-3. SCREE AND SHEPARD DIAGRAMS FOR DECIDUOUS WOODS DISTANCE SAMPLES.

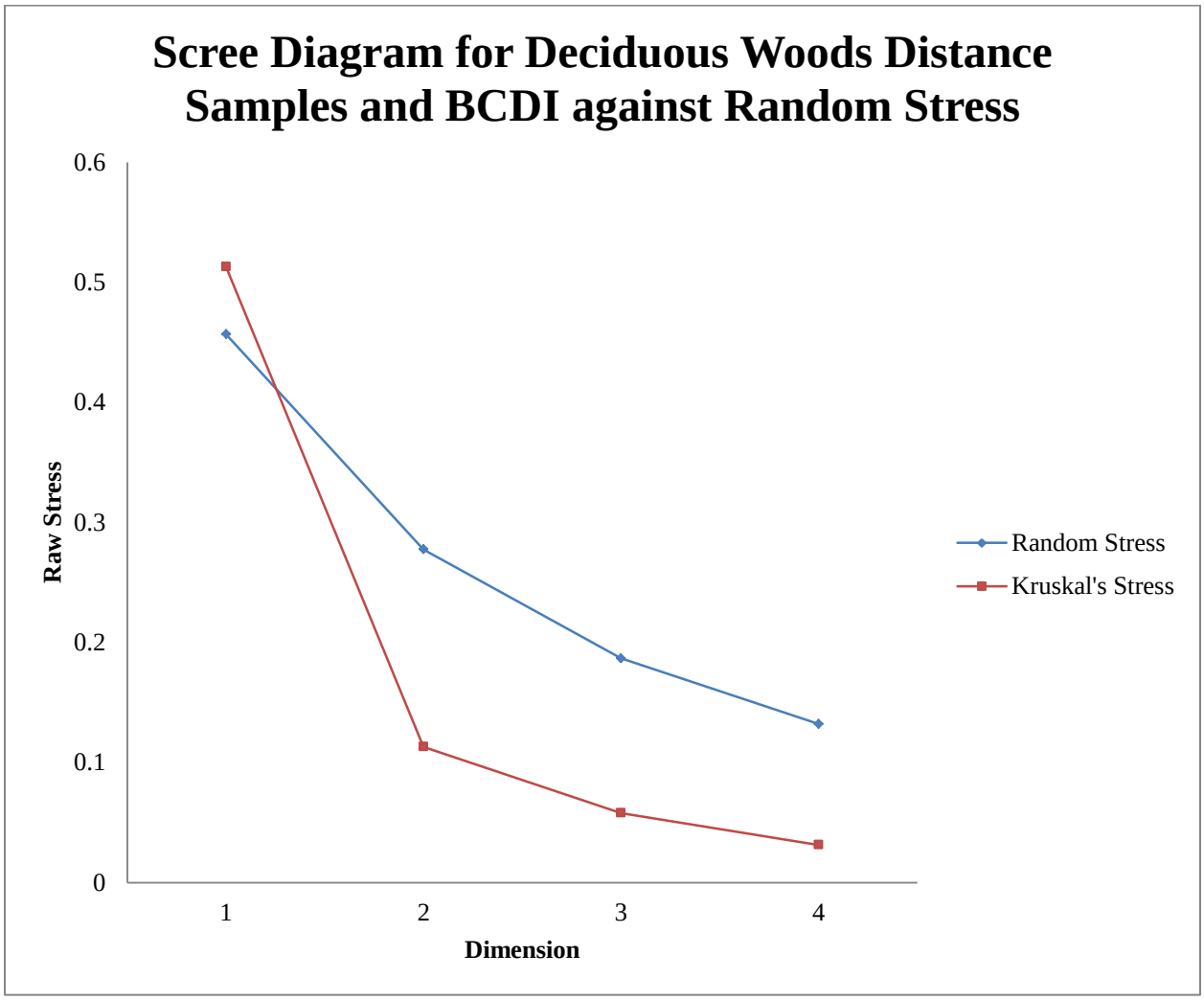

Figure 91. Scree diagram developed by NMDS for the final configurations of deciduous woods distance soil samples and BCDI over four dimensions. Random stress was used as a threshold for the acceptance of final configurations. The one dimension configuration was the only to exceed the threshold and was rejected. The elbow in the curve is at two dimensions.

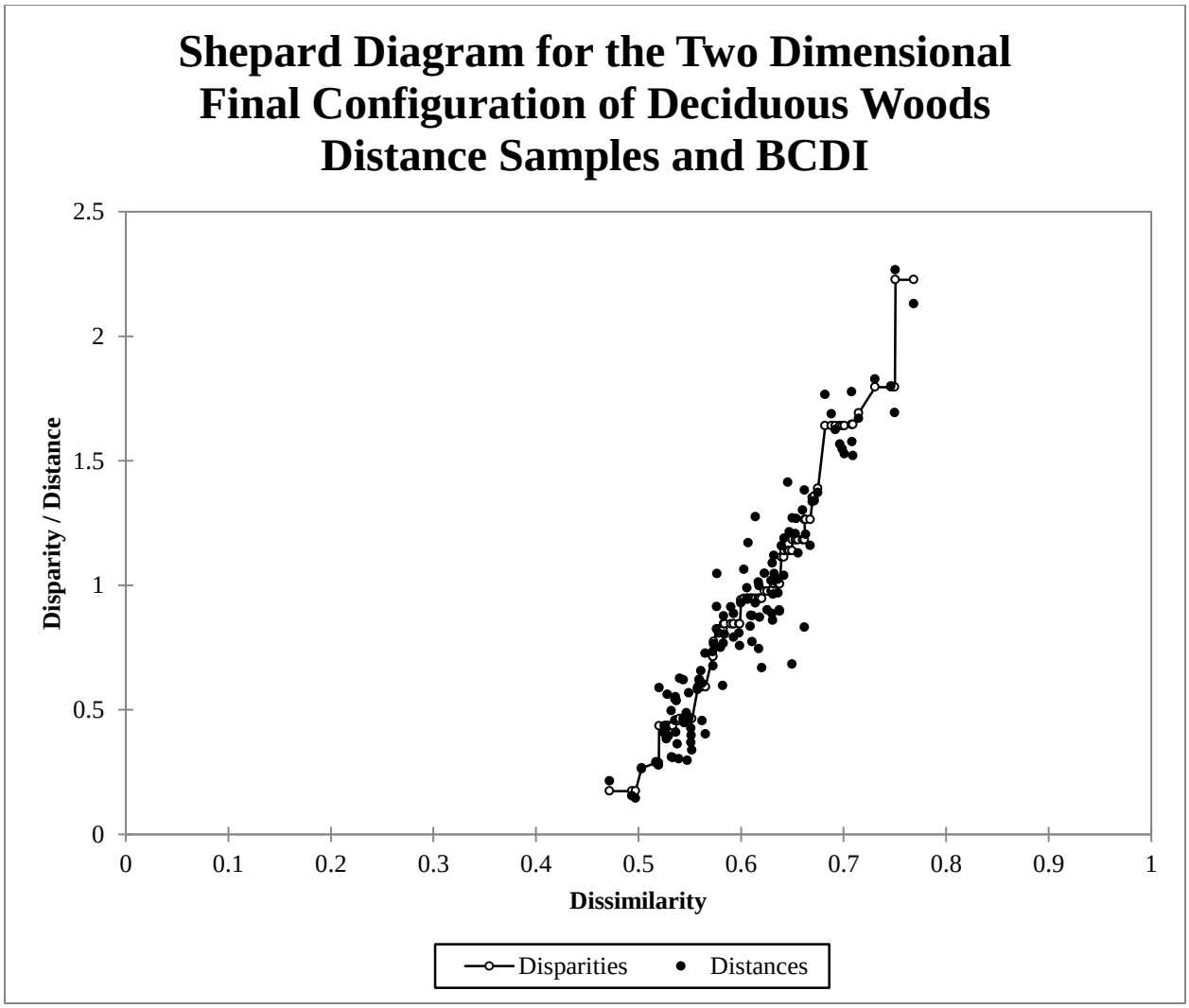

Figure 92. Shepard diagram for the two dimensional final configuration developed from BCDI of deciduous woods distance soil samples. Distances do not associate well with their corresponding disparities agreeing with the higher stress seen with the final configuration.

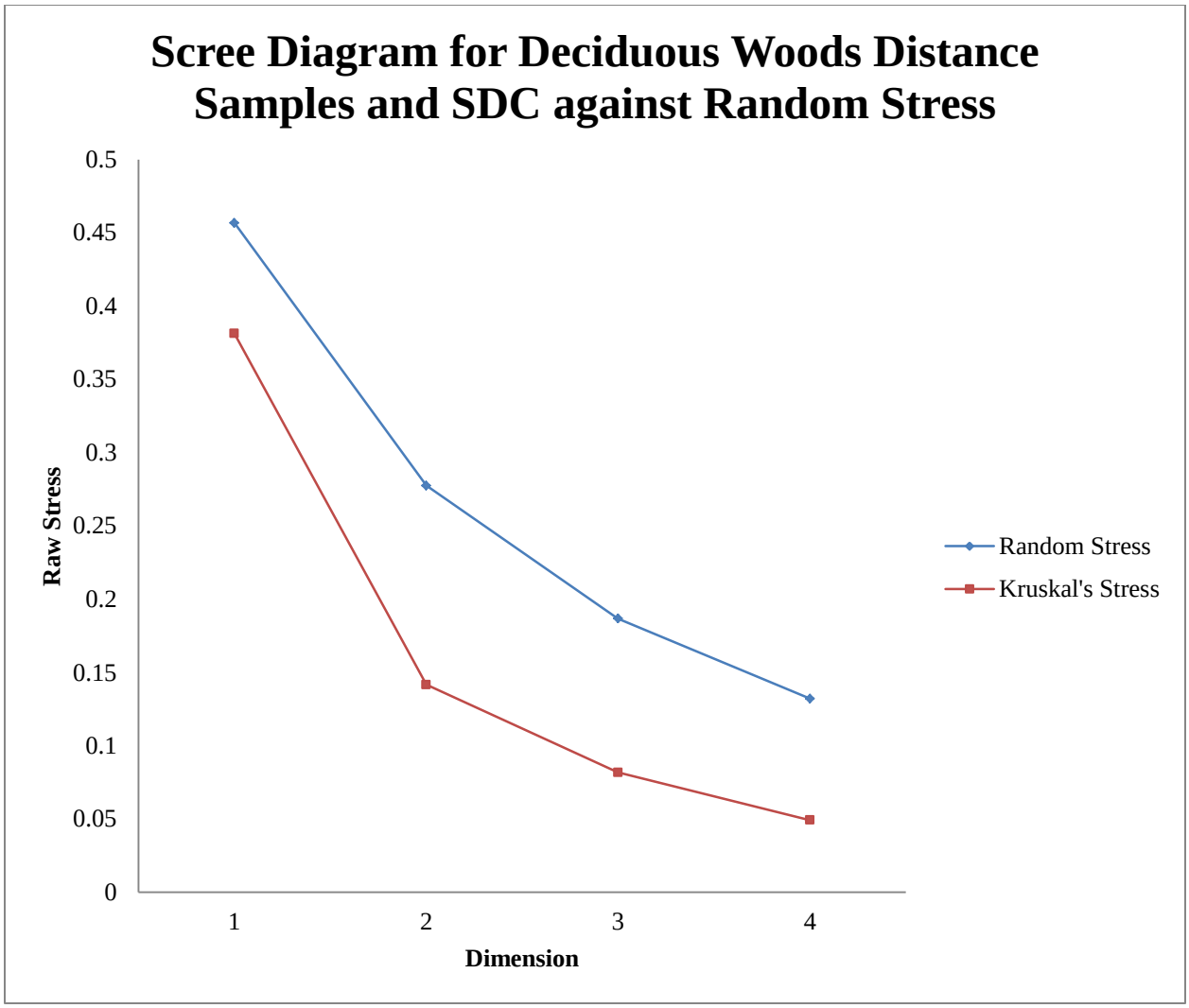

Figure 93. Scree diagram developed by NMDS for the final configurations of deciduous woods distance soil samples and SDC over four dimensions. Random stress was used as a threshold for the acceptance of final configurations. No configuration exceeded the threshold though the stress for the one dimension plot was much higher than the others. The elbow in the curve is at two dimensions.

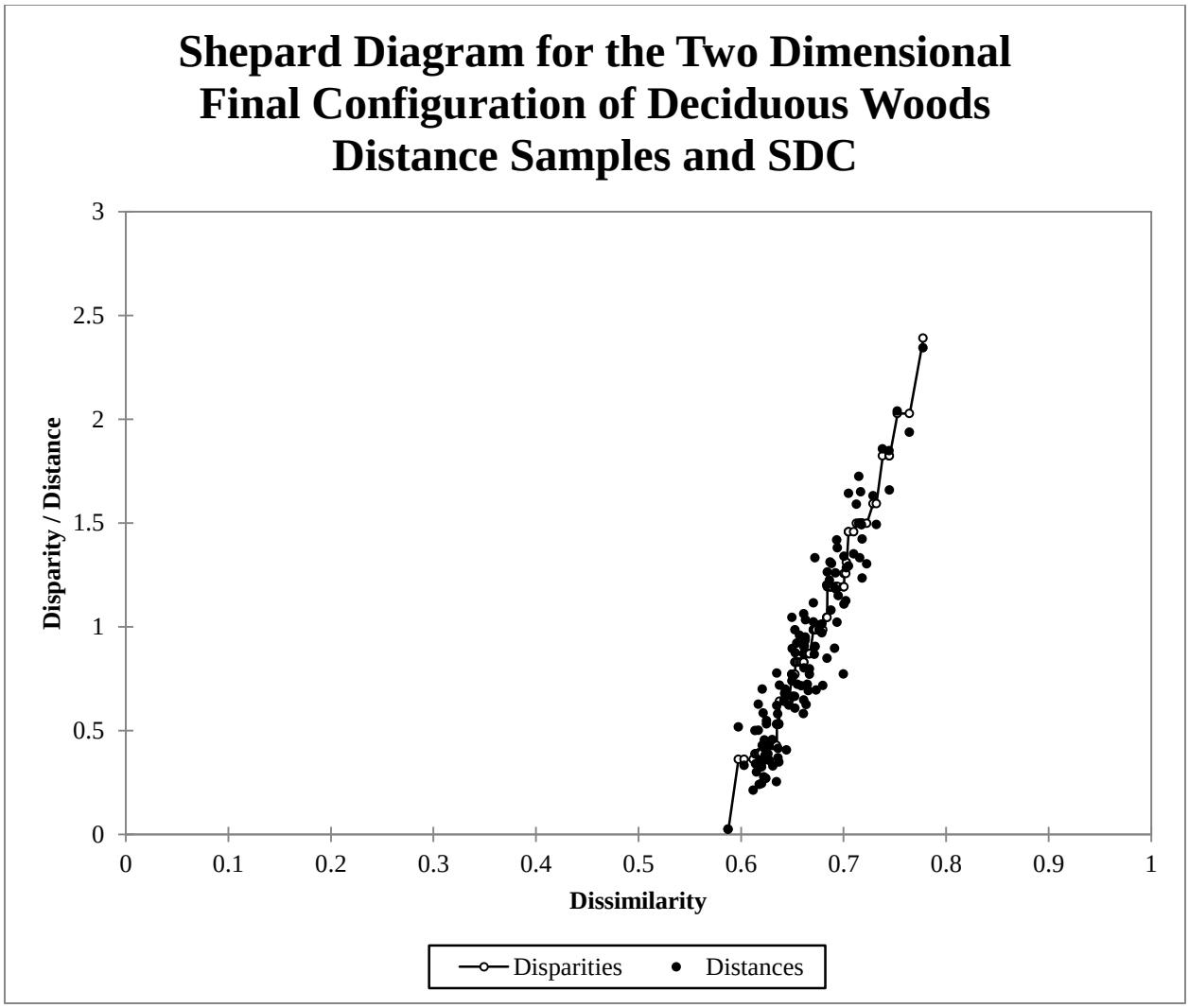

Figure 94. Shepard diagram for the two dimensional final configuration developed from SDC of deciduous woods distance soil samples. Distances do not associate well with their corresponding disparities agreeing with the higher stress seen with the final configuration.

APPENDIX 10. DENDROGRAMS FOR HABITAT DISTANCE SAMPLES.

APPENDIX 10-1. DENDROGRAMS OF TREATED YARD DISTANCE SAMPLES FOR BCDI AND SDC.

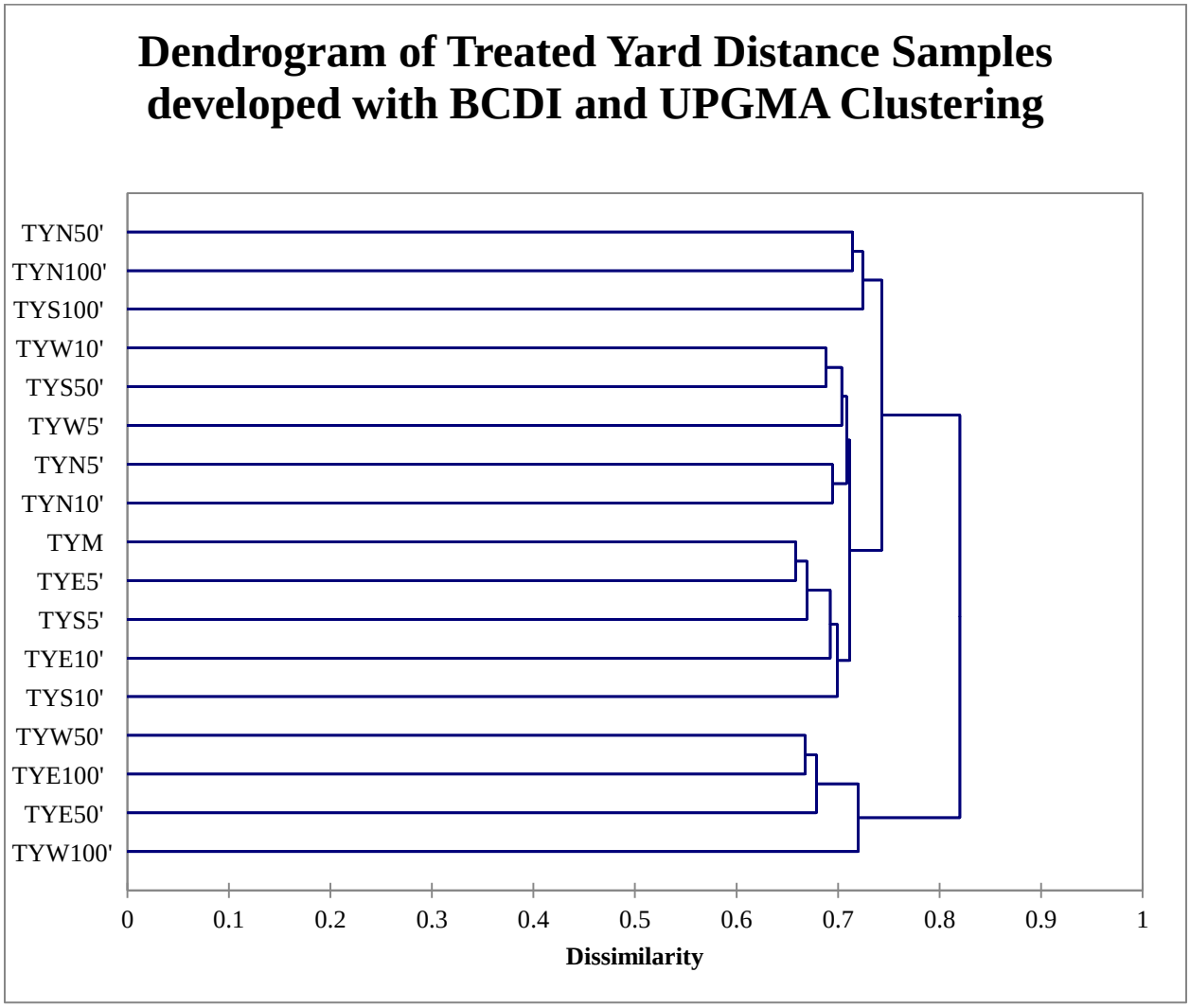

Figure 95. Dendrogram of treated yard distance soil samples developed with BCDI and UPGMA clustering. Three loose clusters are present. The first is made up of the north 50’ and 100’ samples followed by the south 100’ distance grouped at 0.724. The second is formed at 0.712 by two five member clusters and contain the main, all the 5’ and 10’ distances along with the south 50’ samples. The final cluster is composed of the east 50’ and 100’ samples as well as the west 50’ and 100’ distances and formed at 0.720. This cluster joins the others at 0.820. See Table 3 for site names corresponding to abbreviations.

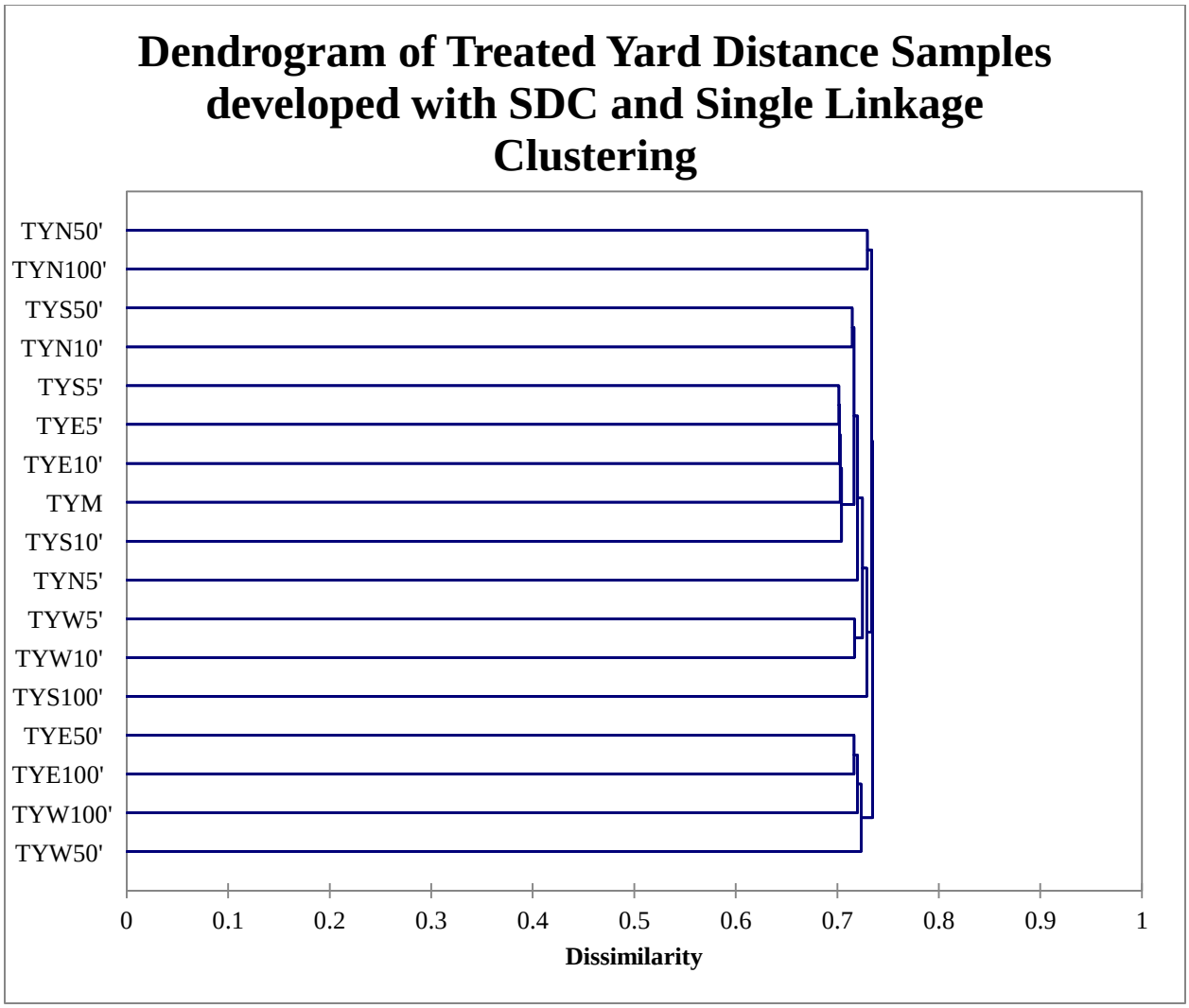

Figure 96. Dendrogram of treated yard distance soil samples developed with SDC and single linkage clustering. There are three possible clusters present; however, the largest of them shows chaining and was not analyzed. See Table 3 for site names corresponding to abbreviations.

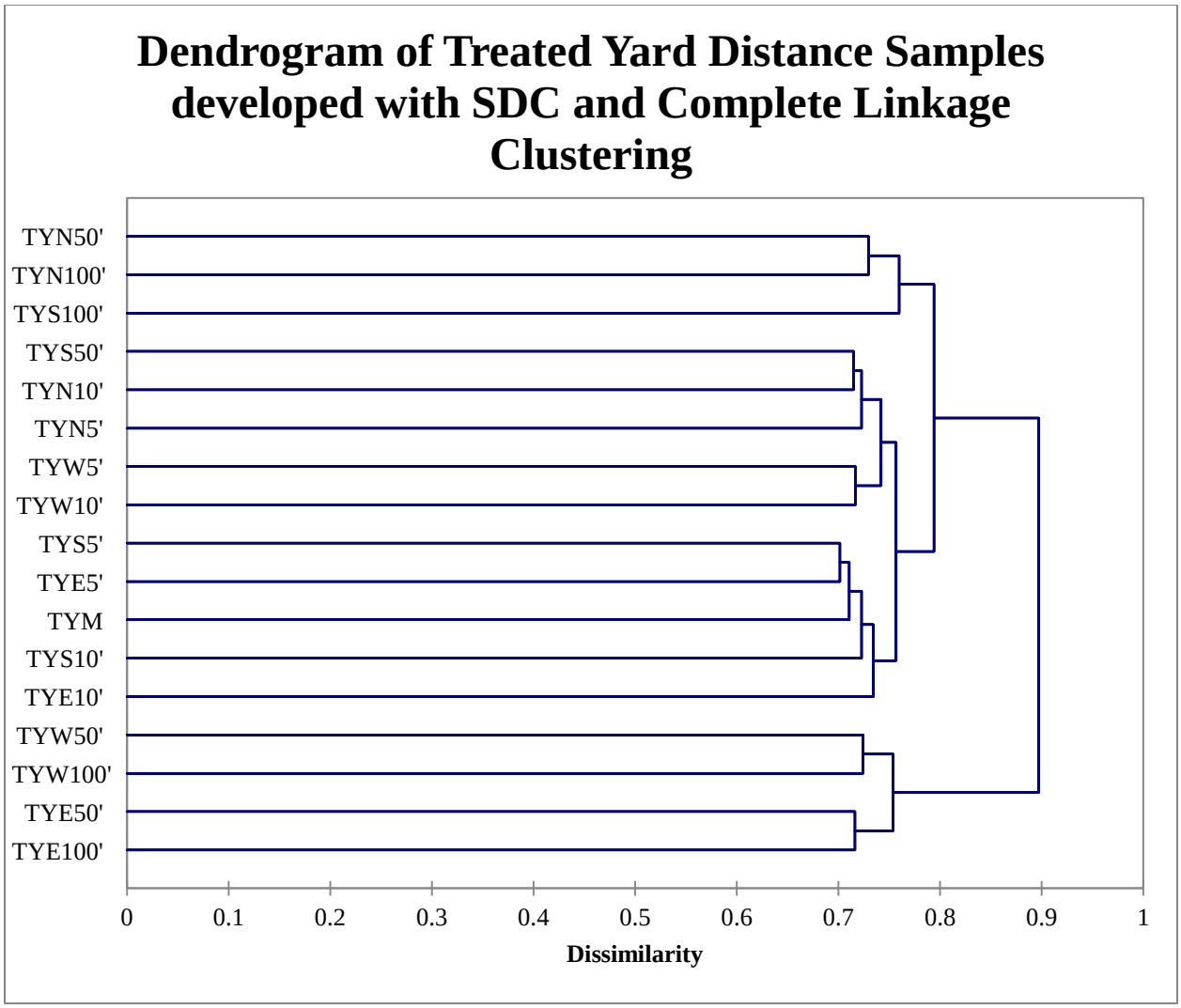

Figure 97. Dendrogram of treated yard distance soil samples developed with SDC and complete linkage clustering. Three loose clusters are present. The first is made up of the north 50' and 100' samples followed by the south 100' distance clustering at 0.760. The second is formed at 0.757 by two five member clusters and contain the main, all the 5' and 10' distances along with the south 50' samples. The final cluster is composed of the east 50' and 100' samples as well as the west 50' and 100' distances grouping at 0.734. This cluster joins the others at 0.900. See Table 3 for site names corresponding to abbreviations.

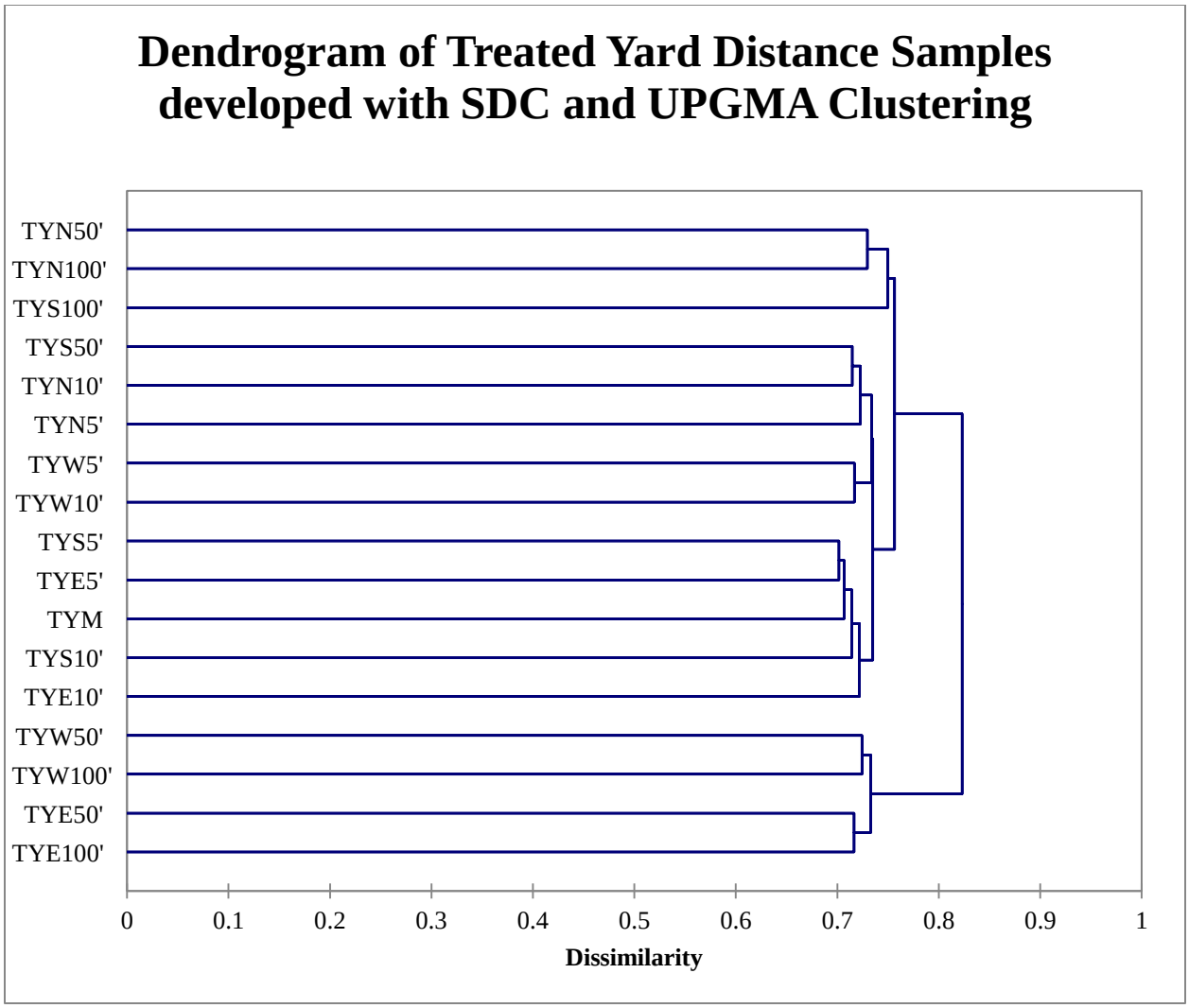

Figure 98. Dendrogram of treated yard distance soil samples developed with SDC and UPGMA clustering. Three loose clusters are present. The first is made up of the north 50’ and 100’ samples followed by the south 100’ distance clustering at 0.740. The second is formed at 0.735 by two five member clusters and contain the main, all the 5’ and 10’ distances along with the south 50’ samples. The final cluster is composed of the east 50’ and 100’ samples as well as the west 50’ and 100’ distances grouping at 0.733. This cluster joins the others at 0.823. See Table 3 for site names corresponding to abbreviations.

APPENDIX 10-2. DENDROGRAMS OF YARD DISTANCE SAMPLES FOR BCDI AND SDC.

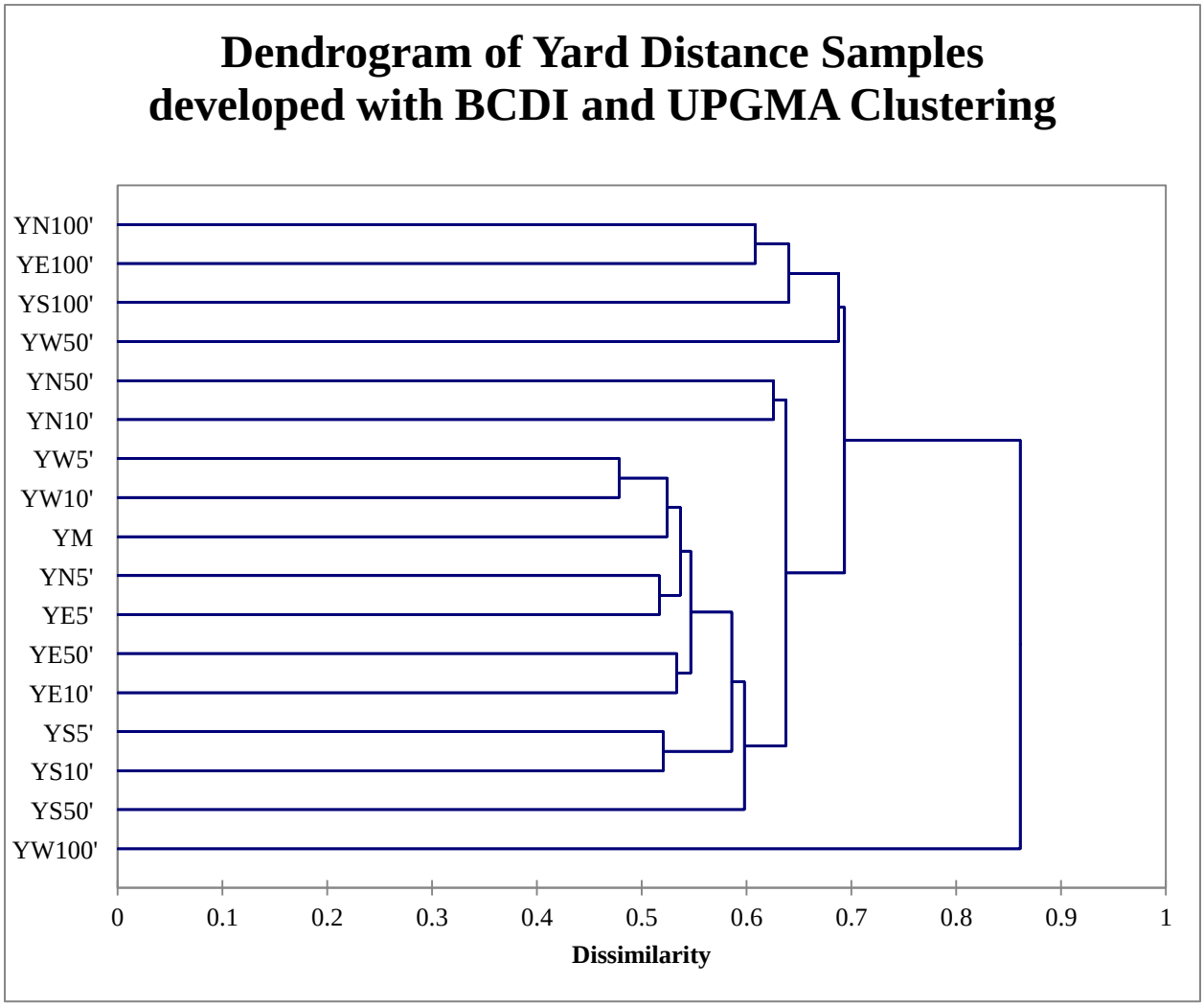

Figure 99. Dendrogram of yard distance soil samples developed from BCDI and UPGMA clustering. Two clusters are present. The first is comprised of the north 100', south 100', east 100', and west 50' distances clustering at 0.688. The second cluster is formed by smaller two sample groups and single samples at 0.638. The west 100' distance joins the two clusters at 0.861. See Table 3 for site names corresponding to abbreviations.

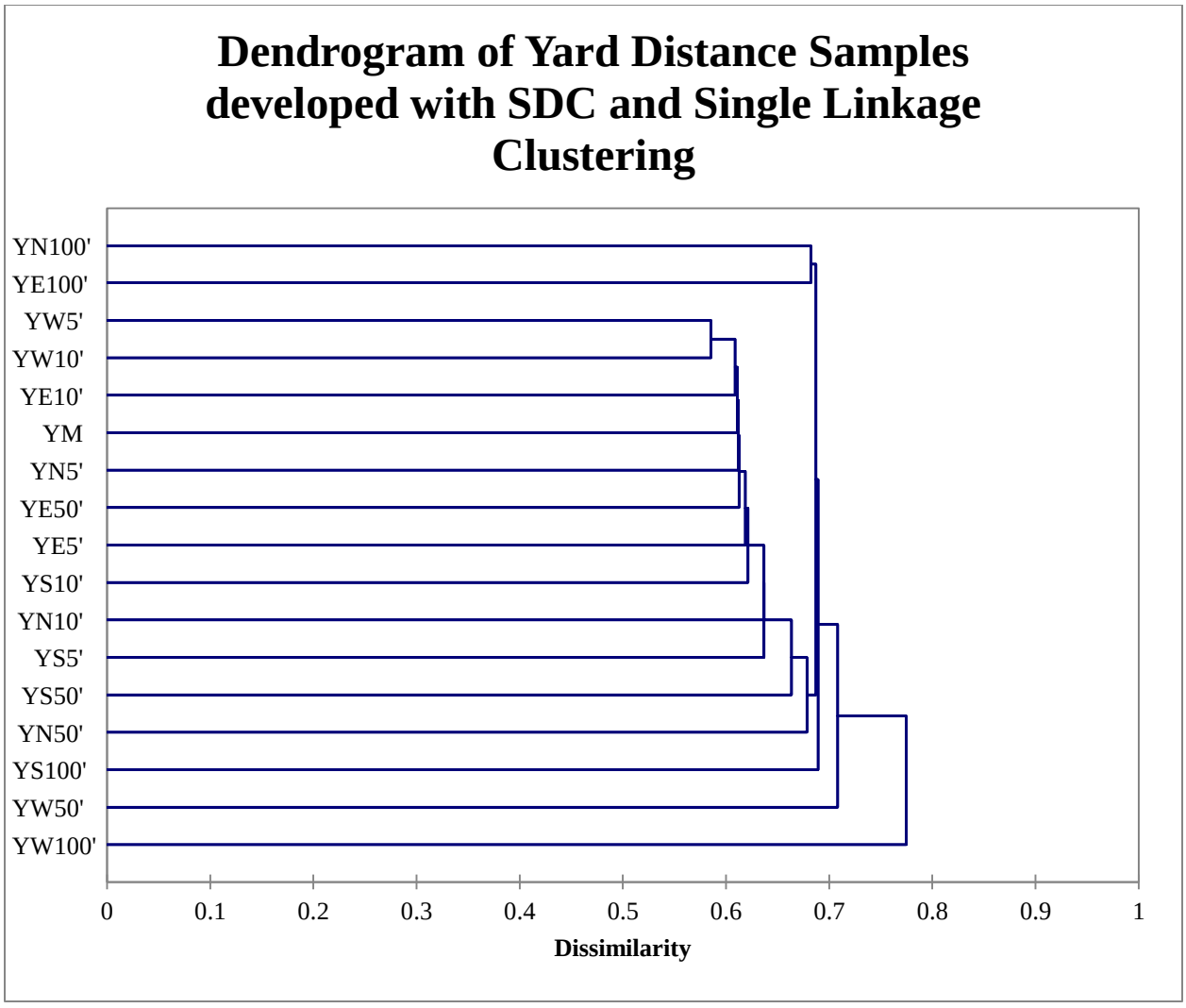

Figure 100. Dendrogram of yard distance soil samples developed with SDC and single linkage clustering. There are possible clusters present; however, extensive chaining of samples makes interpretation difficult. This dendrogram was not analyzed. See Table 3 for site names corresponding to abbreviations.

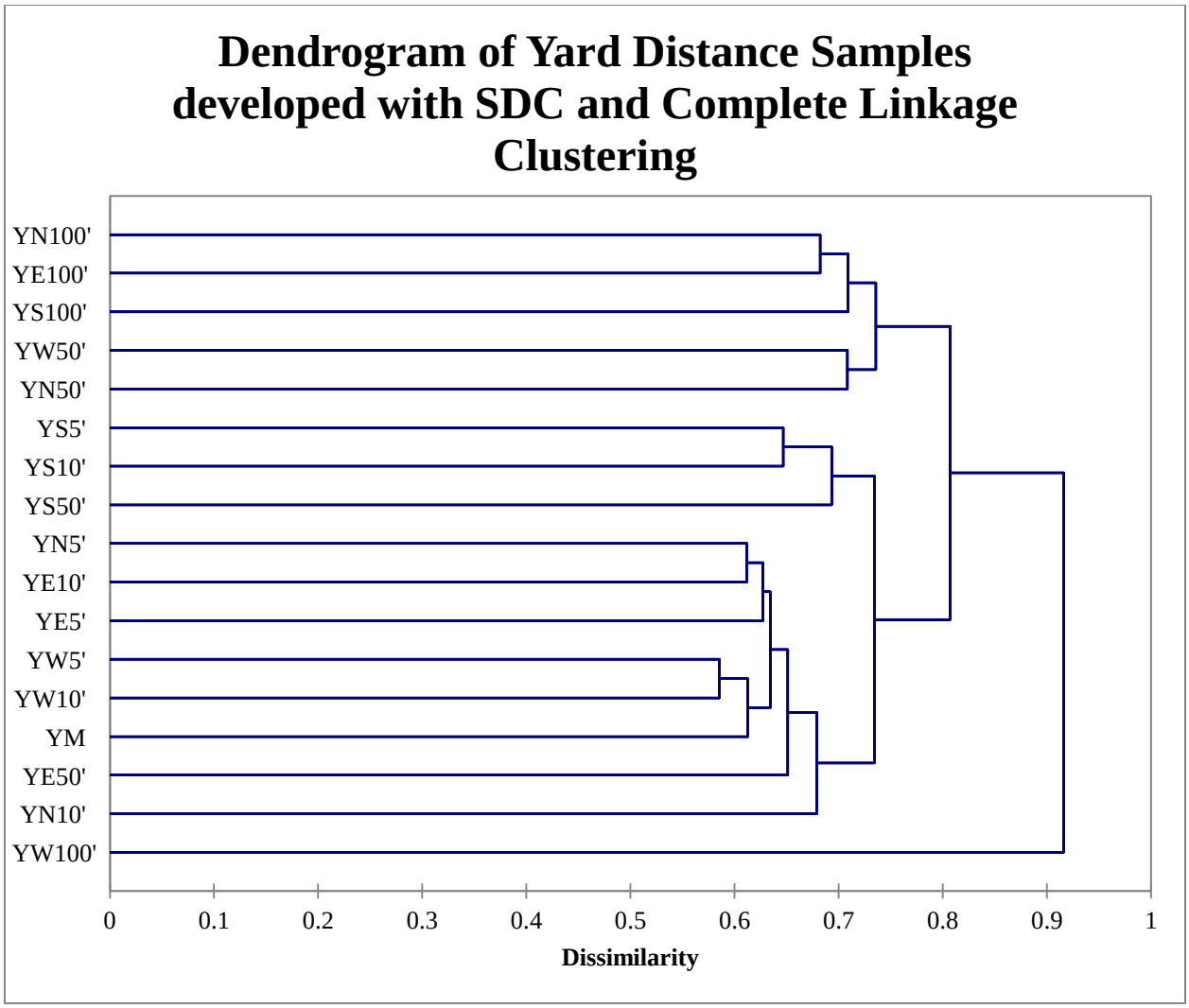

Figure 101. Dendrogram of yard distance soil samples developed with SDC and complete linkage clustering. Two clusters are formed. The first is a joining of a two and three member group that contain the north 50’ and west 50’ and north 100’, south 100’, and east 100’ samples, respectively, at 0.736. The second cluster is also a grouping of two smaller clusters. The three member one contains the south 5’, 10’, and 50’ samples while the other has the main, the remaining 5’ distances, east 10’ and 50’, and west 10’ samples and join at 0.734. The west 100’ distance groups with the two clusters at 0.916. See Table 3 for site names corresponding to abbreviations.

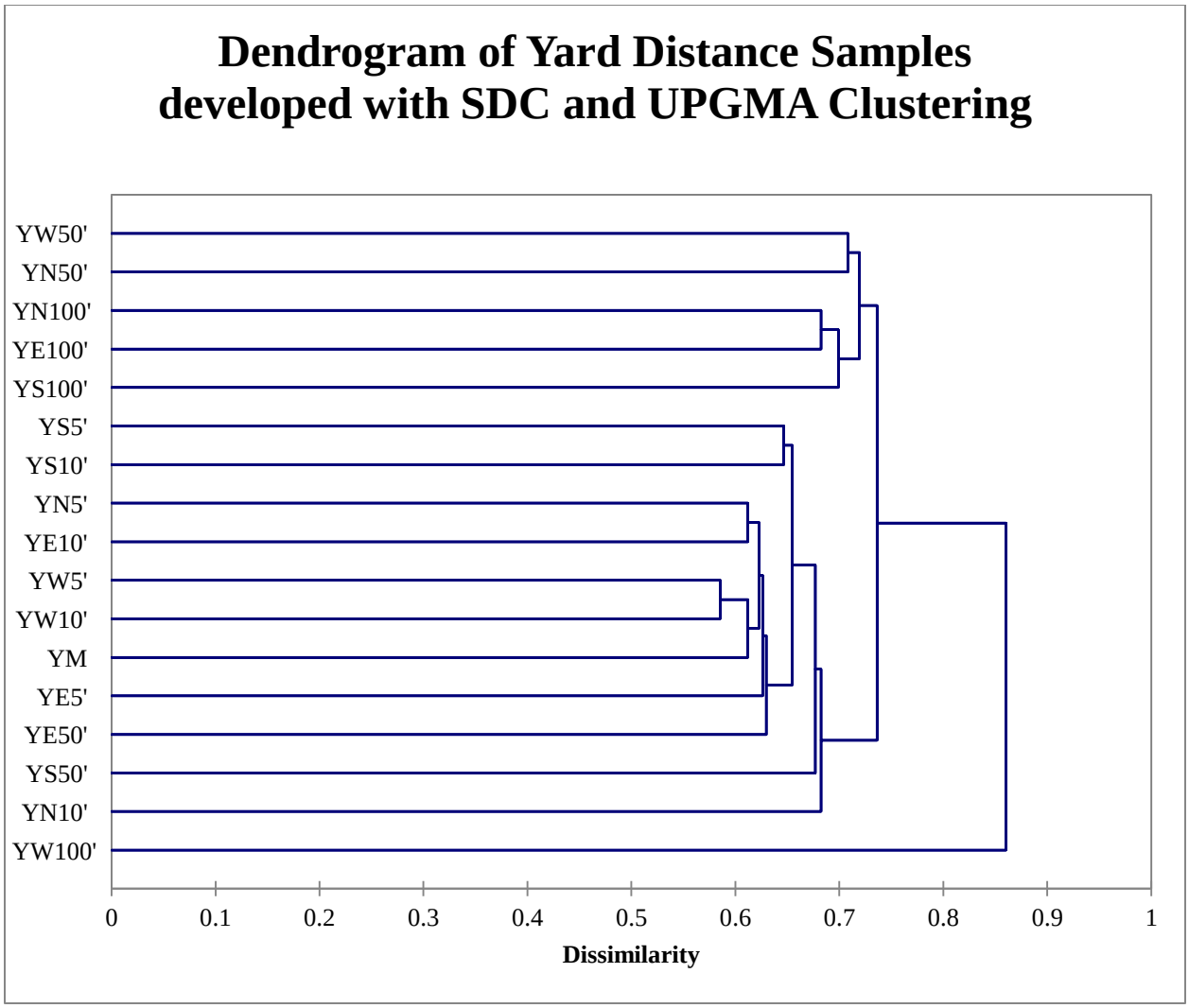

Figure 102. Dendrogram of yard distance soil samples developed with SDC and UPGMA clustering. Two clusters are formed at 0.719 and 0.682, respectively. The first is comprised of a two and three member group that contain the north 50' and west 50' and north 100', south 100', and east 100' samples, respectively. The second cluster is formed by smaller two sample groups and single samples. The west 100' distance joins the two clusters at 0.860. See Table 3 for site names corresponding to abbreviations.

APPENDIX 10-3. DENDROGRAMS OF DECIDUOUS WOODS DISTANCE SAMPLES FOR BCDI AND SDC.

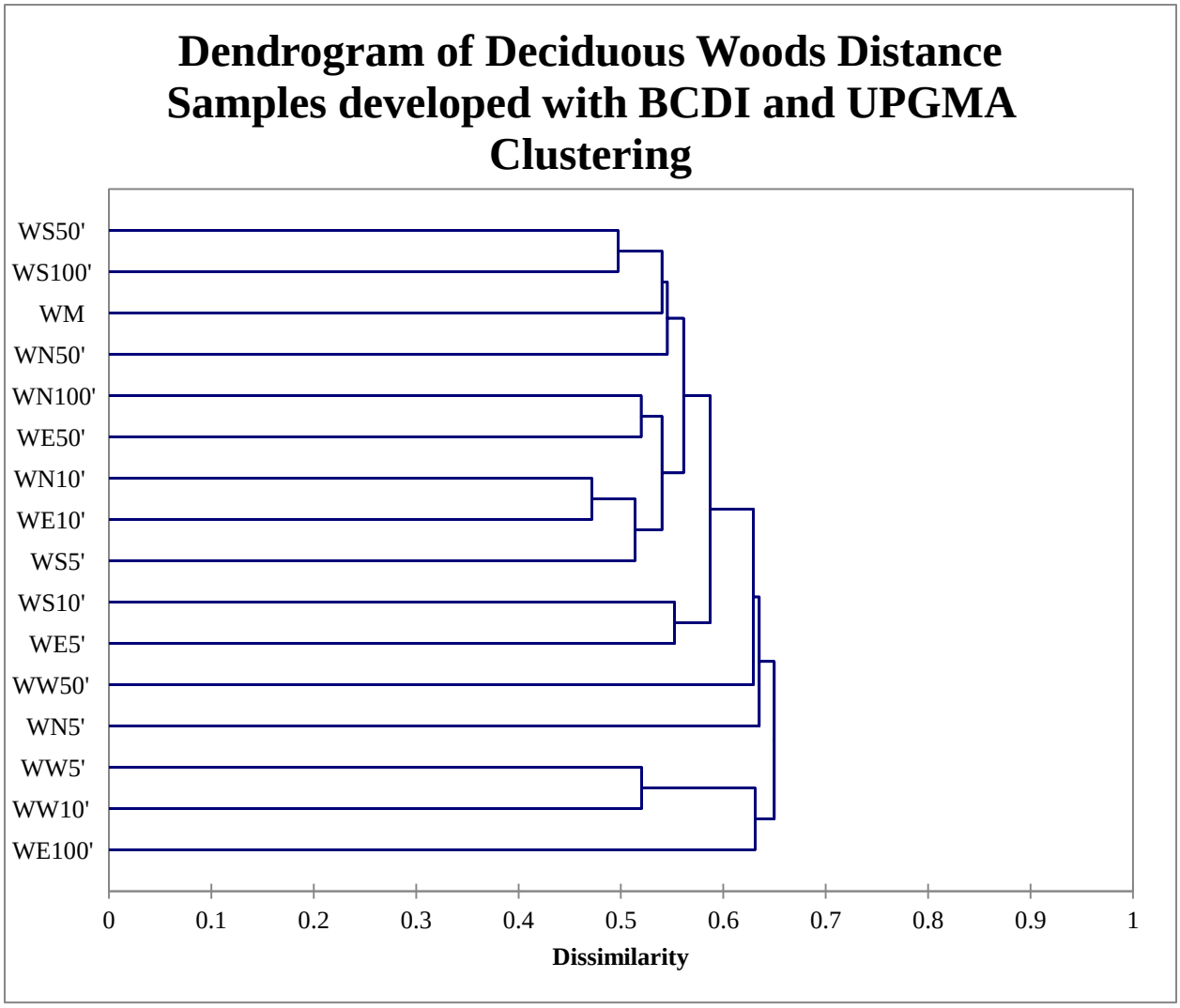

Figure 103. Dendrogram of deciduous woods distance soil samples developed with BCDI and UPGMA clustering. No distinct clusters are noticeable. Small groupings of samples are present; however, a pattern to their associations is not evident. See Table 3 for site names corresponding to abbreviations.

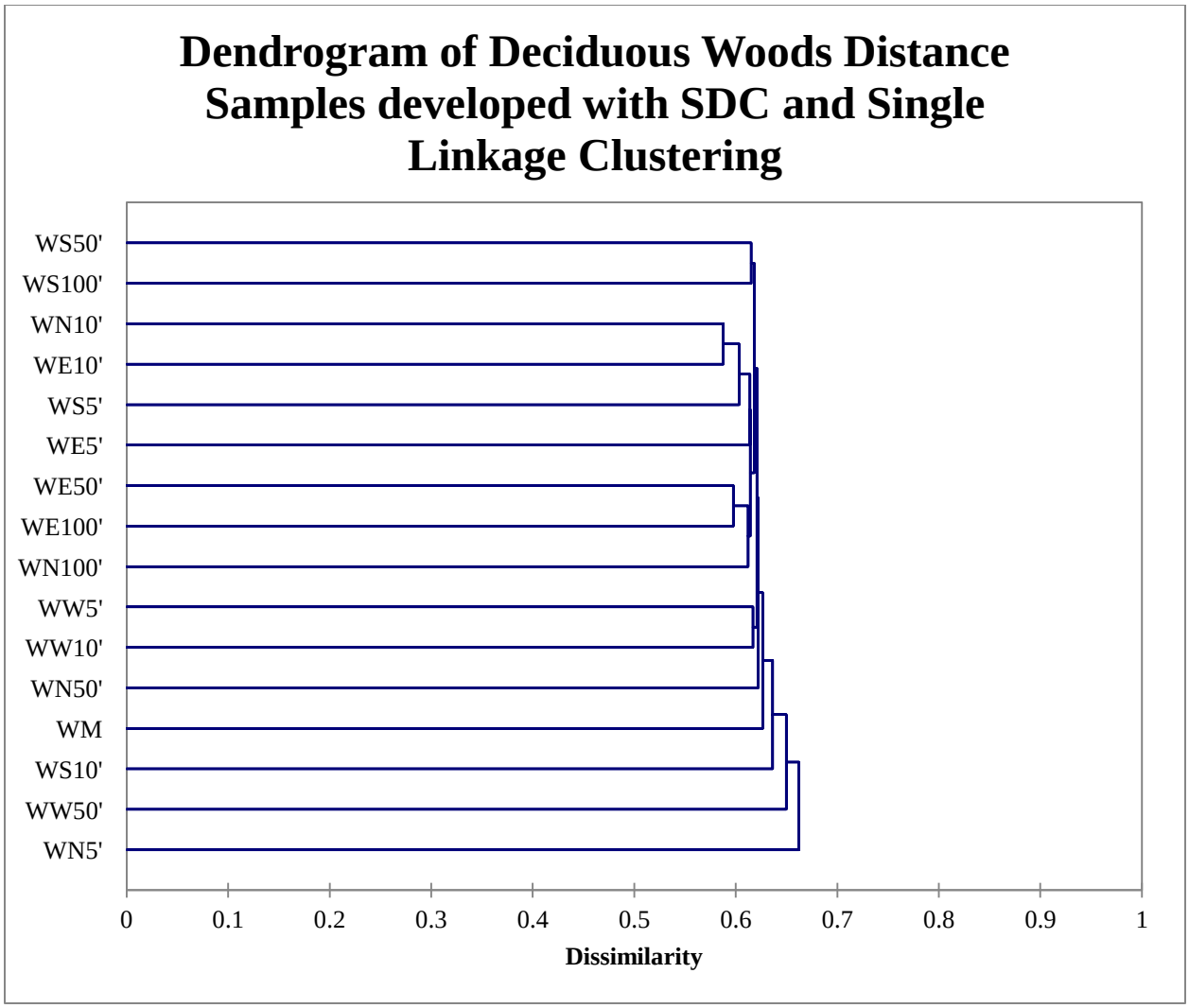

Figure 104. Dendrogram of deciduous woods distance soil samples developed with SDC and single linkage clustering. There are possible clusters present; however, extensive chaining of samples makes interpretation difficult. This dendrogram was not analyzed. See Table 3 for site names corresponding to abbreviations.

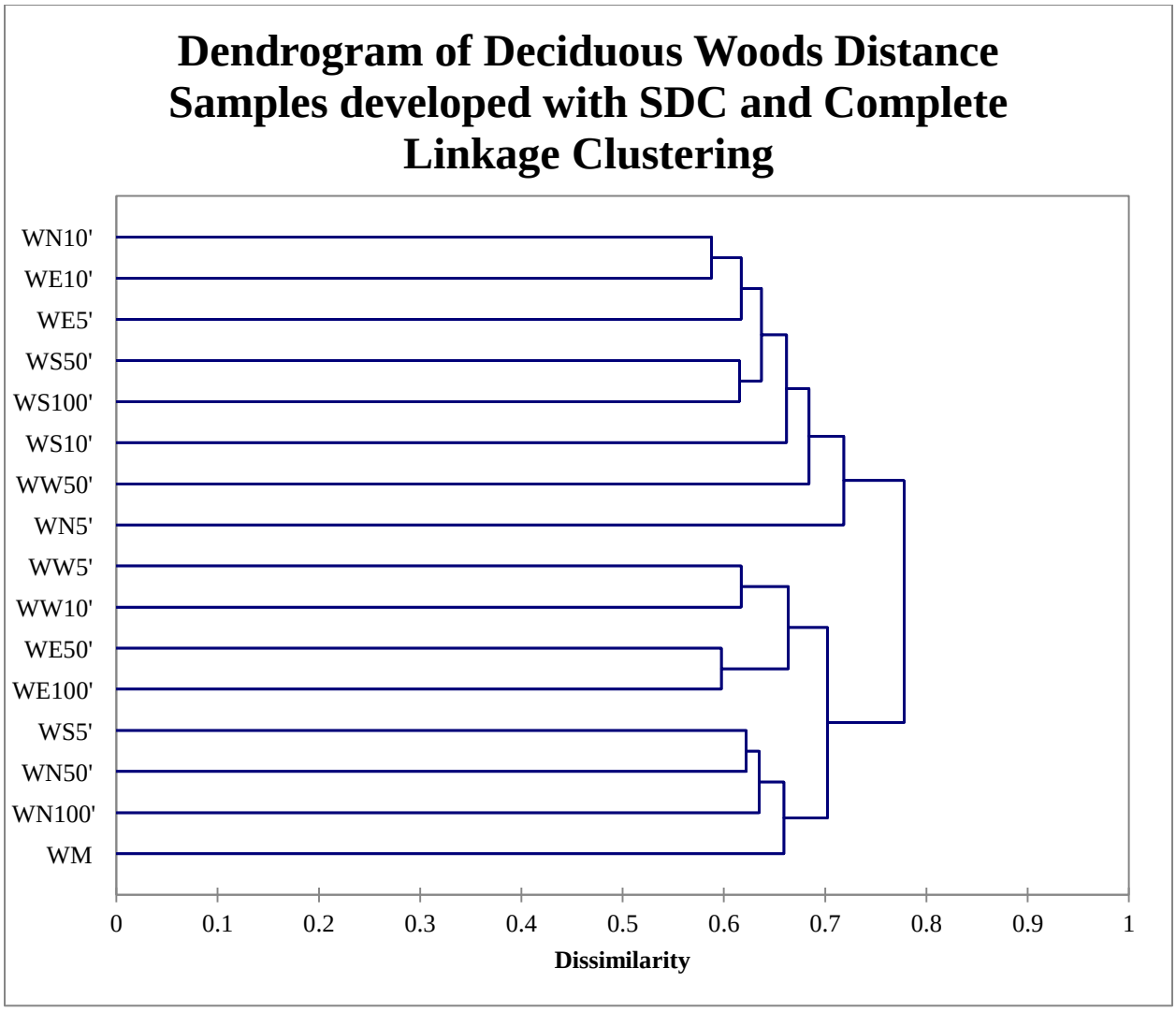

Figure 105. Dendrogram of deciduous woods distance soil samples developed with SDC and complete linkage clustering. Two distinct clusters are present. The first is comprised of the north 5' and 10', south 10', 50', and 100', east 5' and 10', and west 50' distances grouping at 0.718. The other is formed at 0.702 by the remaining samples as two smaller clusters. The two larger clusters group together at 0.778. See Table 3 for site names corresponding to abbreviations.

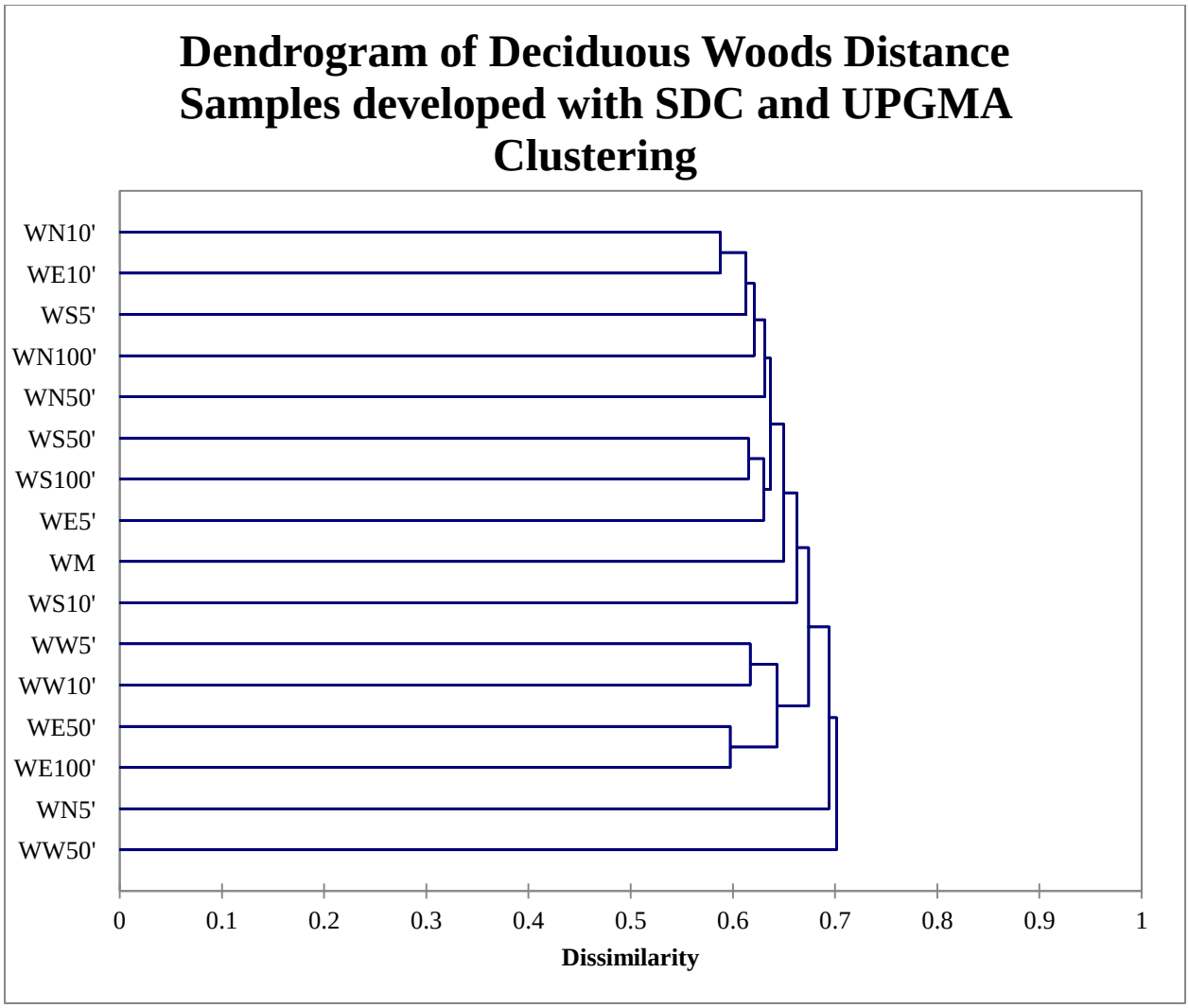

Figure 106. Dendrogram of deciduous woods distance soil samples developed with SDC and UPGMA clustering. No distinct clusters are noticeable. Small groupings of samples are present; however, a pattern to their associations is not evident. See Table 3 for site names corresponding to abbreviations.

APPENDIX 11. ADDITIONAL NMDS DIAGRAMS FOR DEPTH SAMPLES.

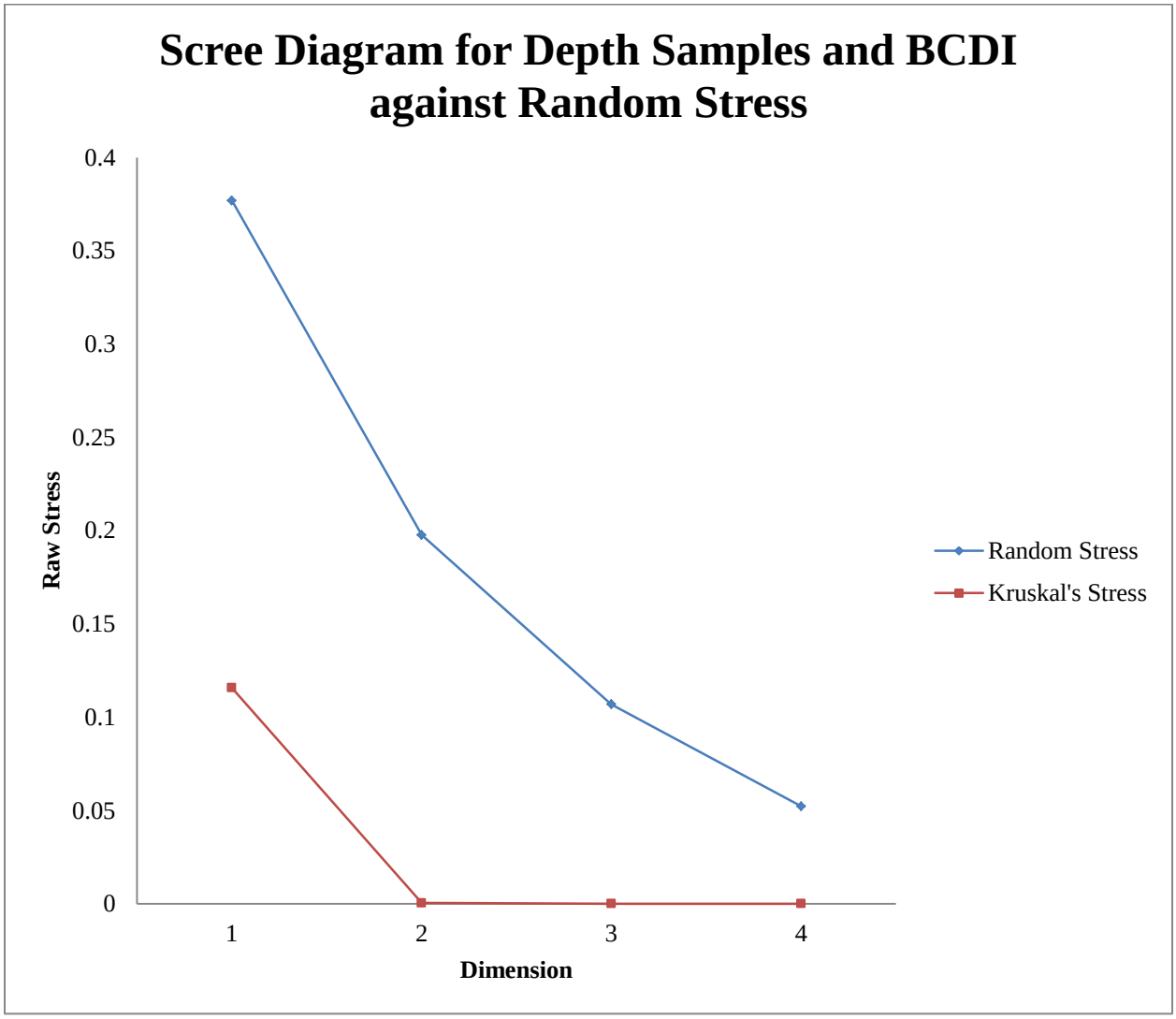

Figure 107. Scree diagram developed by NMDS for the final configurations of depth soil samples and BCDI over four dimensions. Random stress was used as a threshold for the acceptance of final configurations. No configuration exceeded the threshold though the stress for the one dimension plot was higher than the others. The elbow in the curve is at two dimensions.

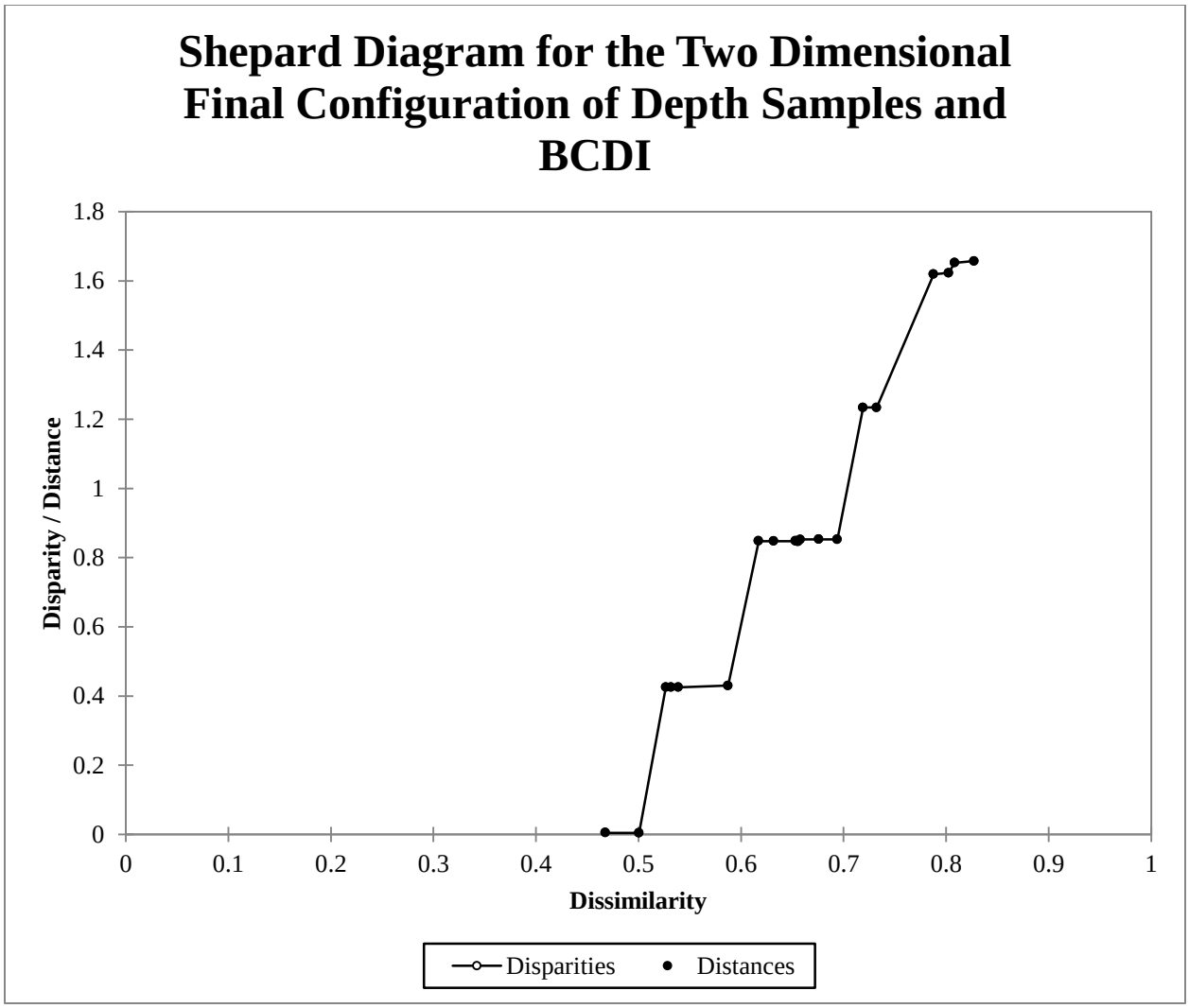

Figure 108. Shepard diagram for the two dimensional final configuration developed from BCDI of depth soil samples. All distances fall close to their corresponding disparities indicating good correlation of the two in the final configuration.

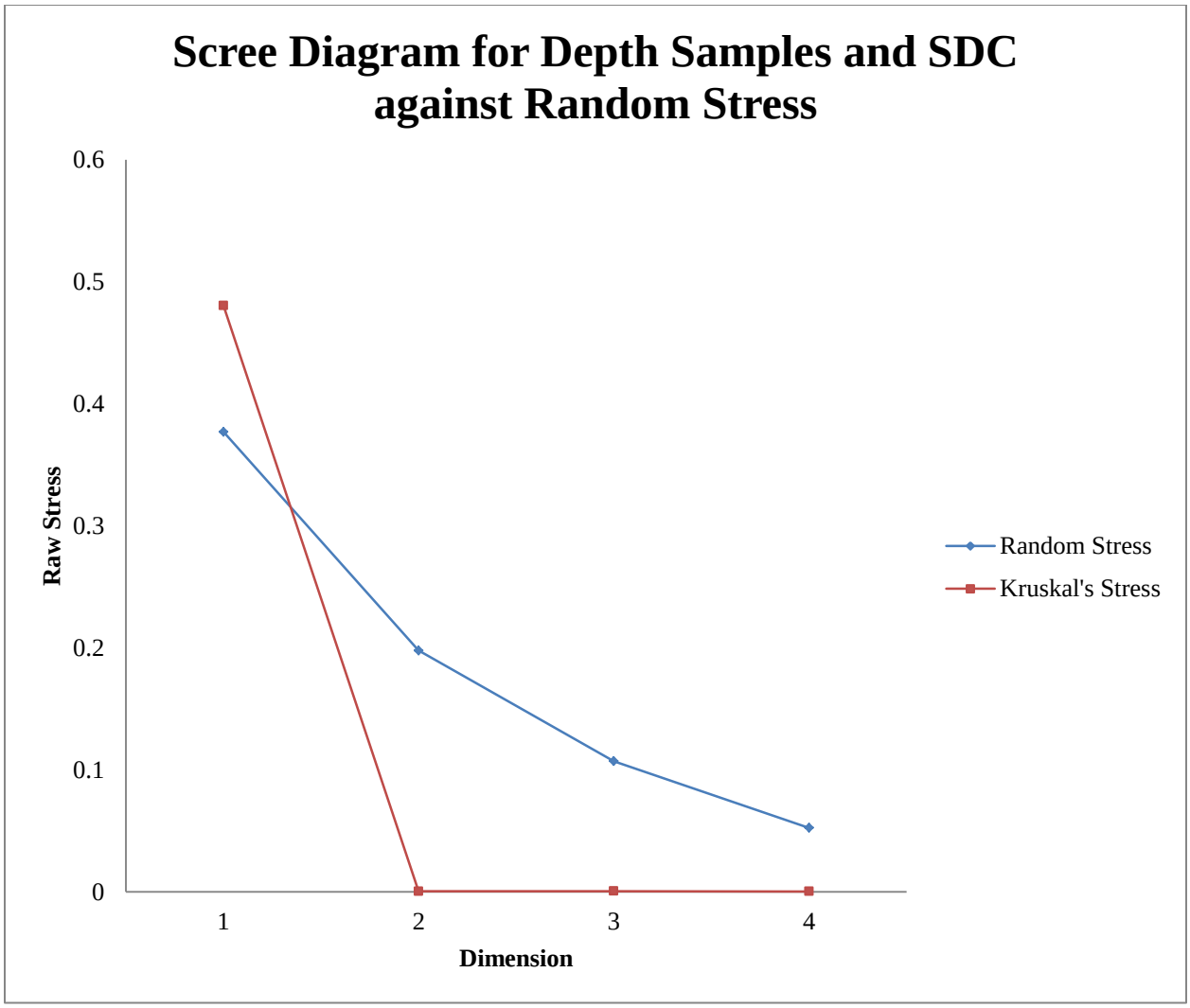

Figure 109. Scree diagram developed by NMDS for the final configurations of depth soil samples and SDC over four dimensions. Random stress was used as a threshold for the acceptance of final configurations. The one dimension configuration was the only to exceed the threshold and was rejected. The elbow in the curve is at two dimensions.

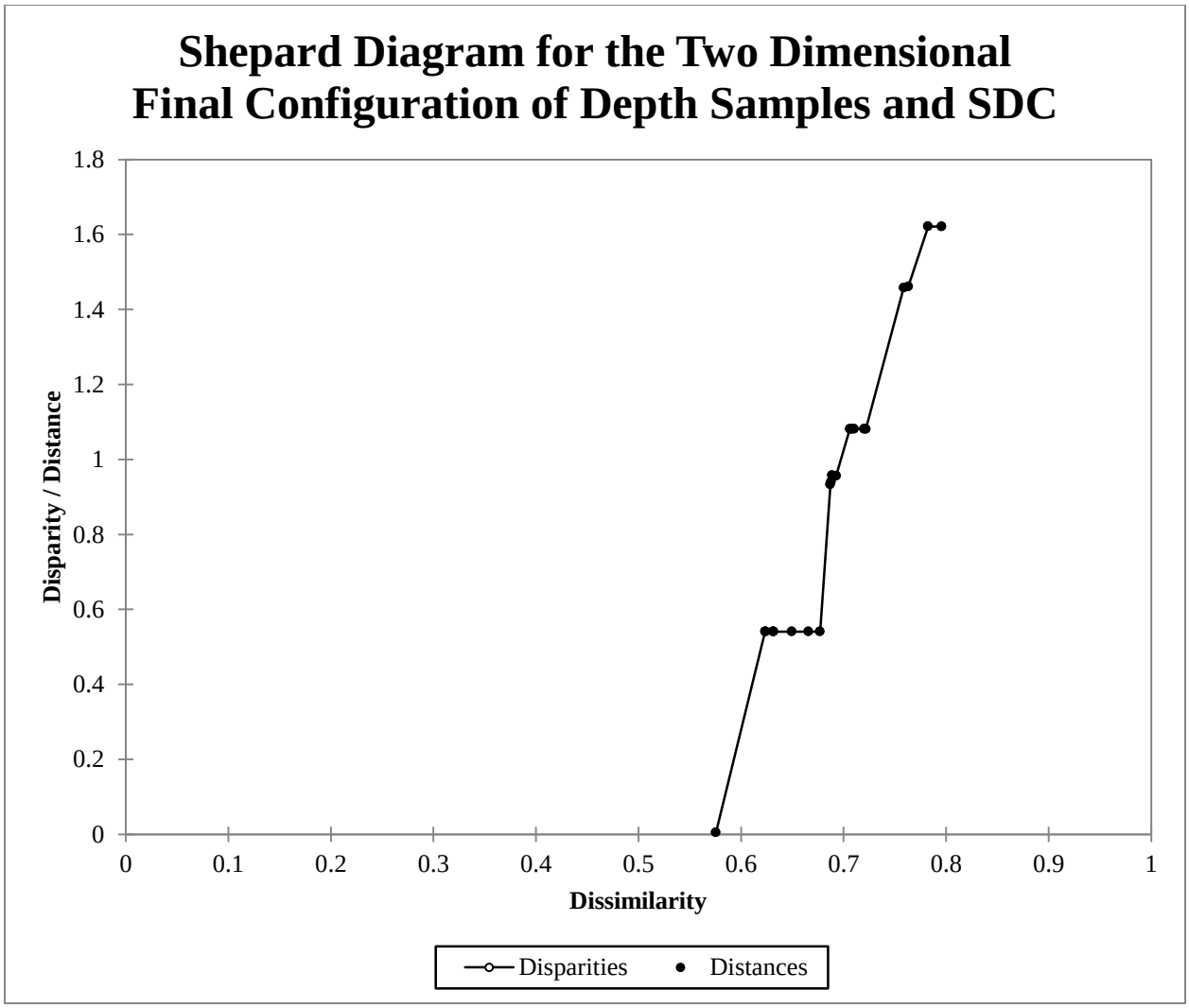

Figure 110. Shepard diagram for the two dimensional final configuration developed from SDC of depth soil samples. All distances fall close to their corresponding disparities indicating good correlation of the two in the final configuration.

APPENDIX 12. DENDROGRAMS OF DEPTH SAMPLES FOR BCDI AND SDC.

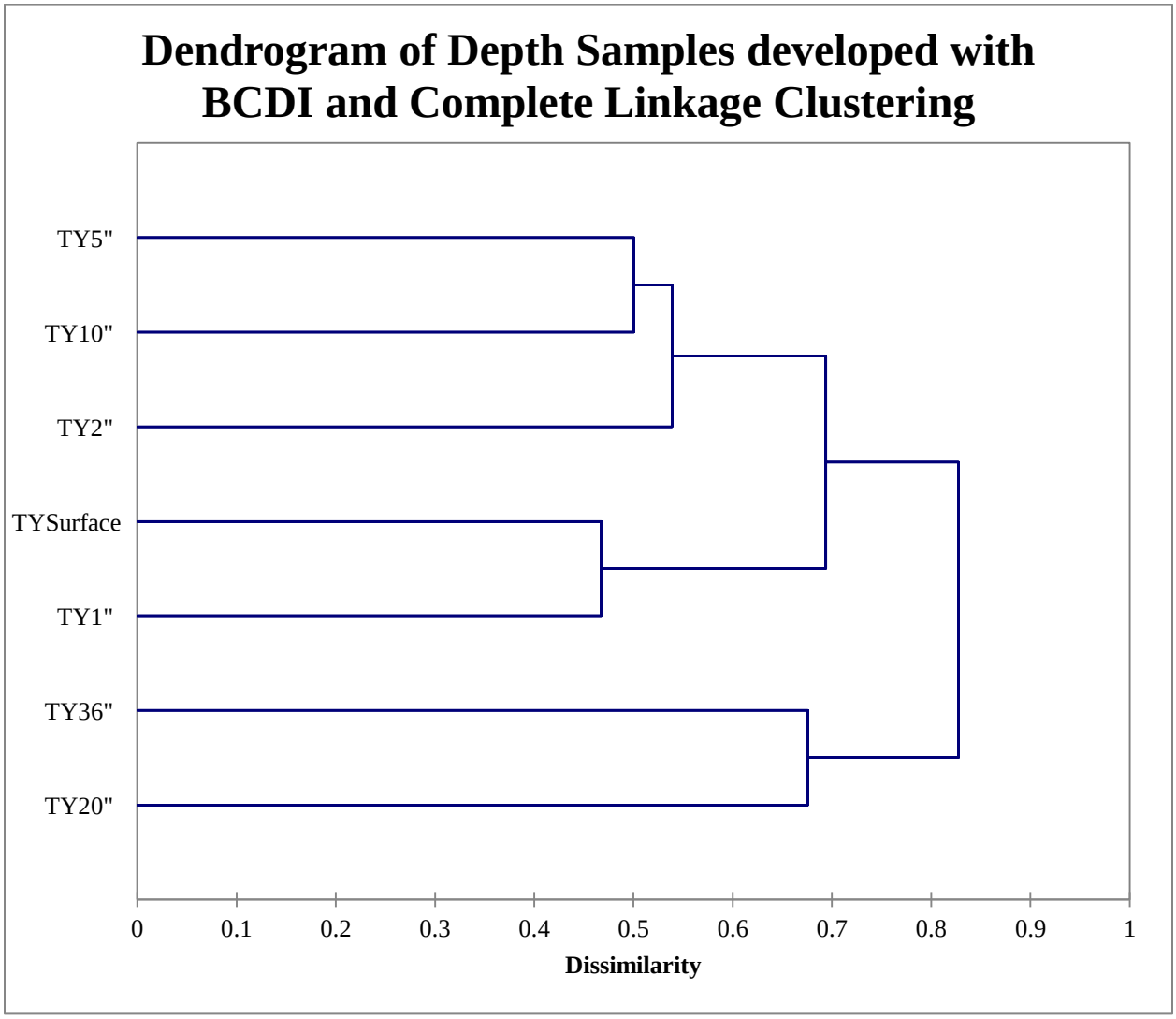

Figure 111. Dendrogram of depth soil samples developed with BCDI and complete linkage clustering. Three clusters are noticeable. The first is a two member group formed by the surface and 1" samples at 0.468. This cluster is joined by a three member group of the 5", 10", and 2" depths at 0.694. The final grouping is the 20" and 36" depths which cluster with the other two at 0.827. See Table 4 for site names corresponding to abbreviations.

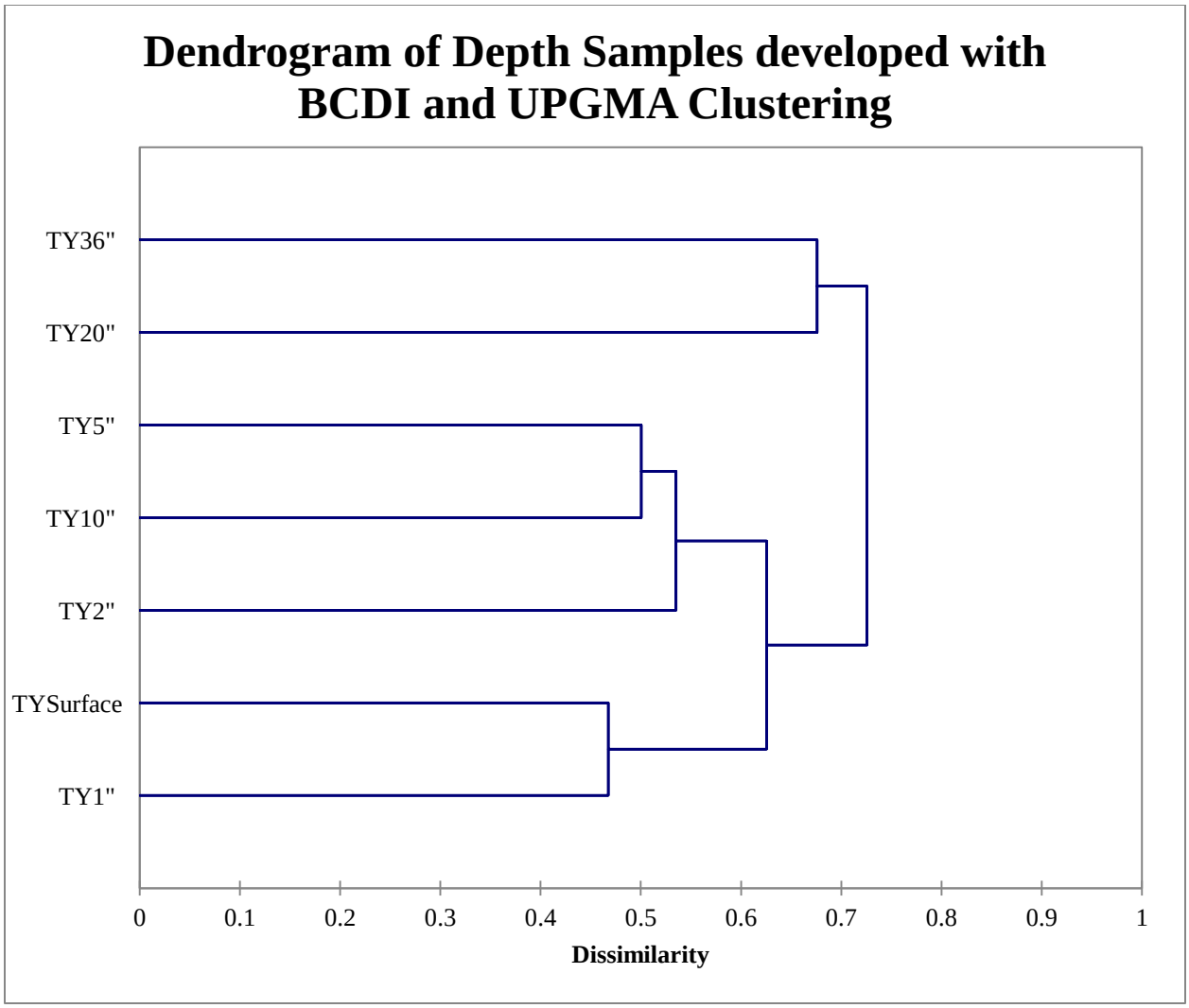

Figure 112. Dendrogram of depth soil samples developed with BCDI and UPGMA clustering. Three clusters are noticeable. The first is a two member group formed at 0.468 by the surface and 1" samples. This cluster is joined by a three member group of the 5", 10", and 2" depths at 0.626. The final grouping is the 20" and 36" depths which cluster with the other two at 0.726. See Table 4 for site names corresponding to abbreviations.

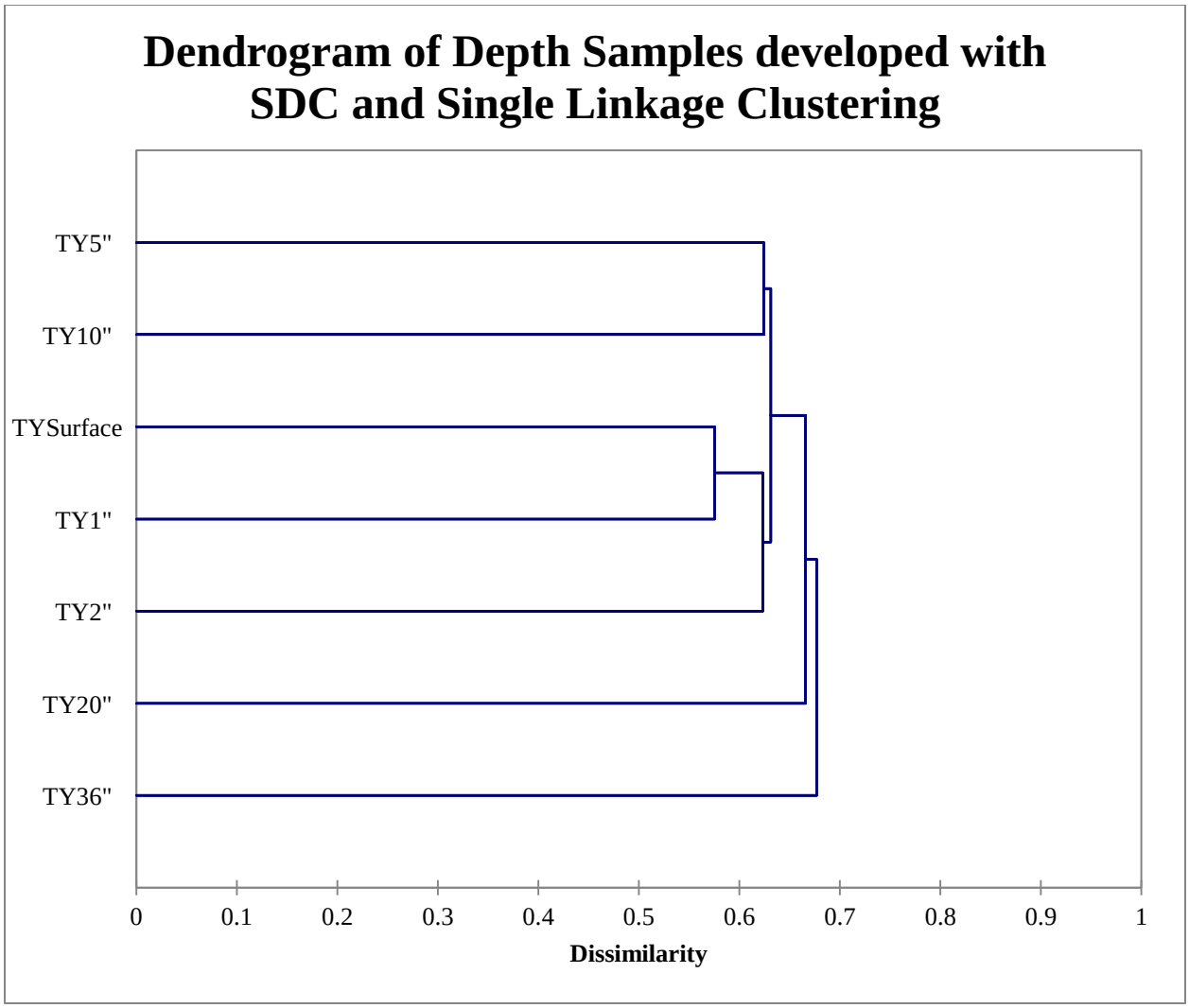

Figure 113. Dendrogram of depth soil samples developed with SDC and single linkage clustering. Two clusters are present. The first is a three member group formed first by the surface and 1" samples followed by the 2" depth at 0.624. This cluster is joined by a two member group of the 5" and 10" depths at 0.631. The 20" and 36" depths group with the rest at 0.665 or greater. See Table 4 for site names corresponding to abbreviations.

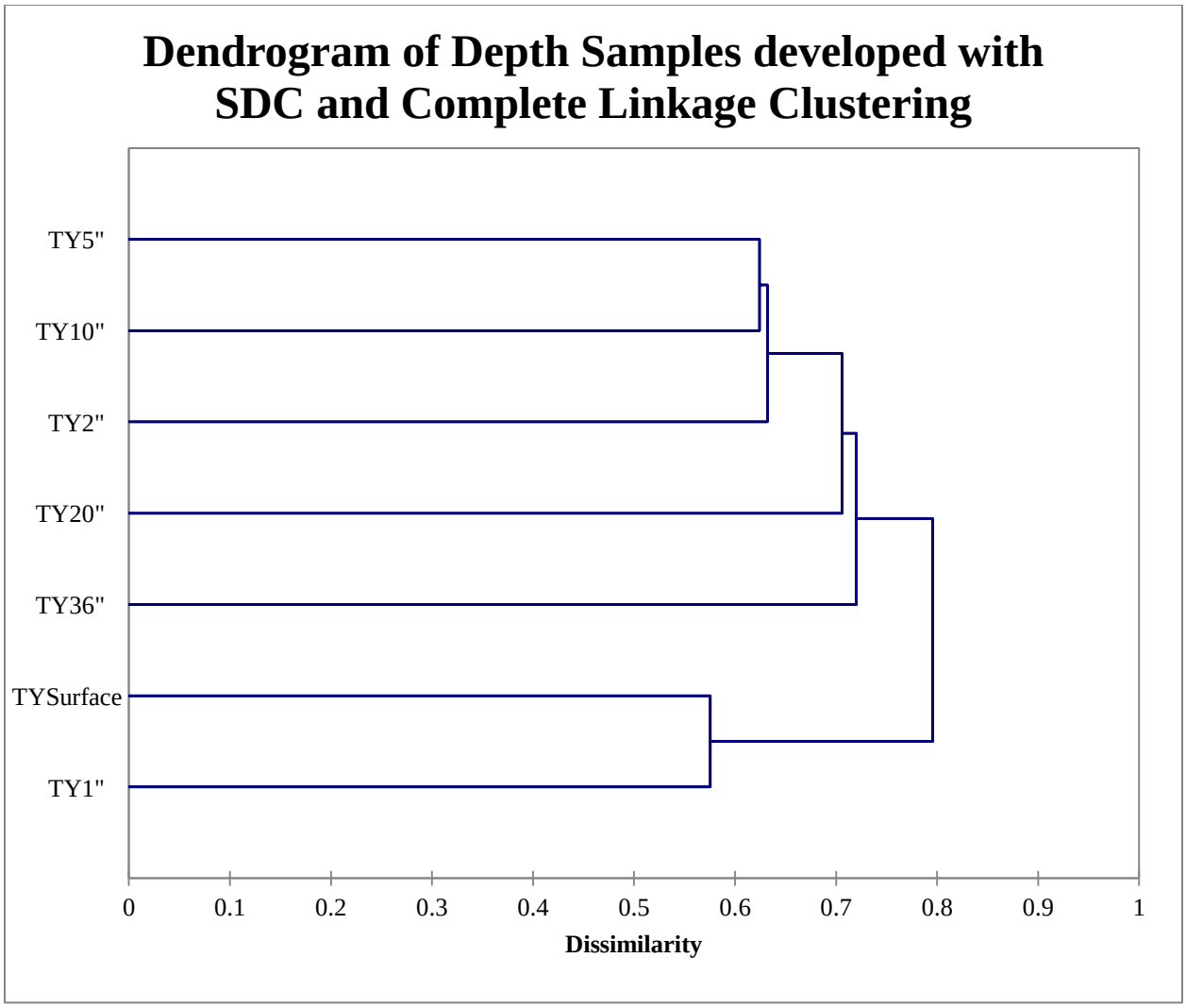

Figure 114. Dendrogram of depth soil samples developed with SDC and complete linkage clustering. Two possible clusters are present. The first is made up of the surface and 1" depths joining at 0.576. The other is formed initially by the 5" and 10" depths followed by the 2", 20", and 36" samples at 0.720. The two are grouped together at a dissimilarity of 0.796. See Table 4 for site names corresponding to abbreviations.

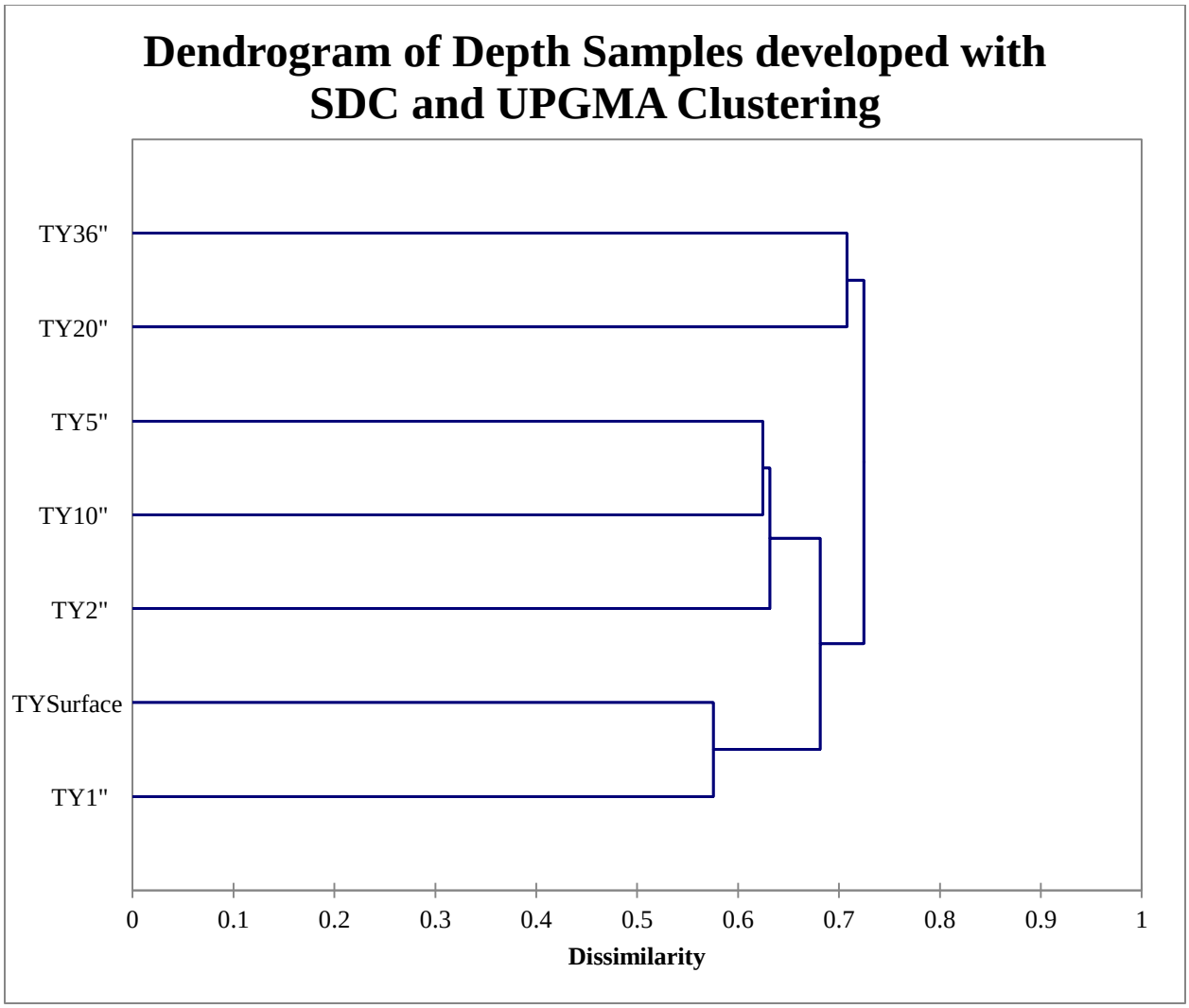

Figure 115. Dendrogram of depth soil samples developed with SDC and UPGMA clustering. Three clusters are formed. The first is a two member group formed by the surface and 1" samples at 0.576. This cluster is joined by a three member group of the 5", 10", and 2" depths at 0.681. The final grouping is the 20" and 36" depths which cluster with the other two at 0.724. See Table 4 for site names corresponding to abbreviations.

APPENDIX 13. ADDITIONAL NMDS DIAGRAMS FOR TIME SERIES SAMPLES.

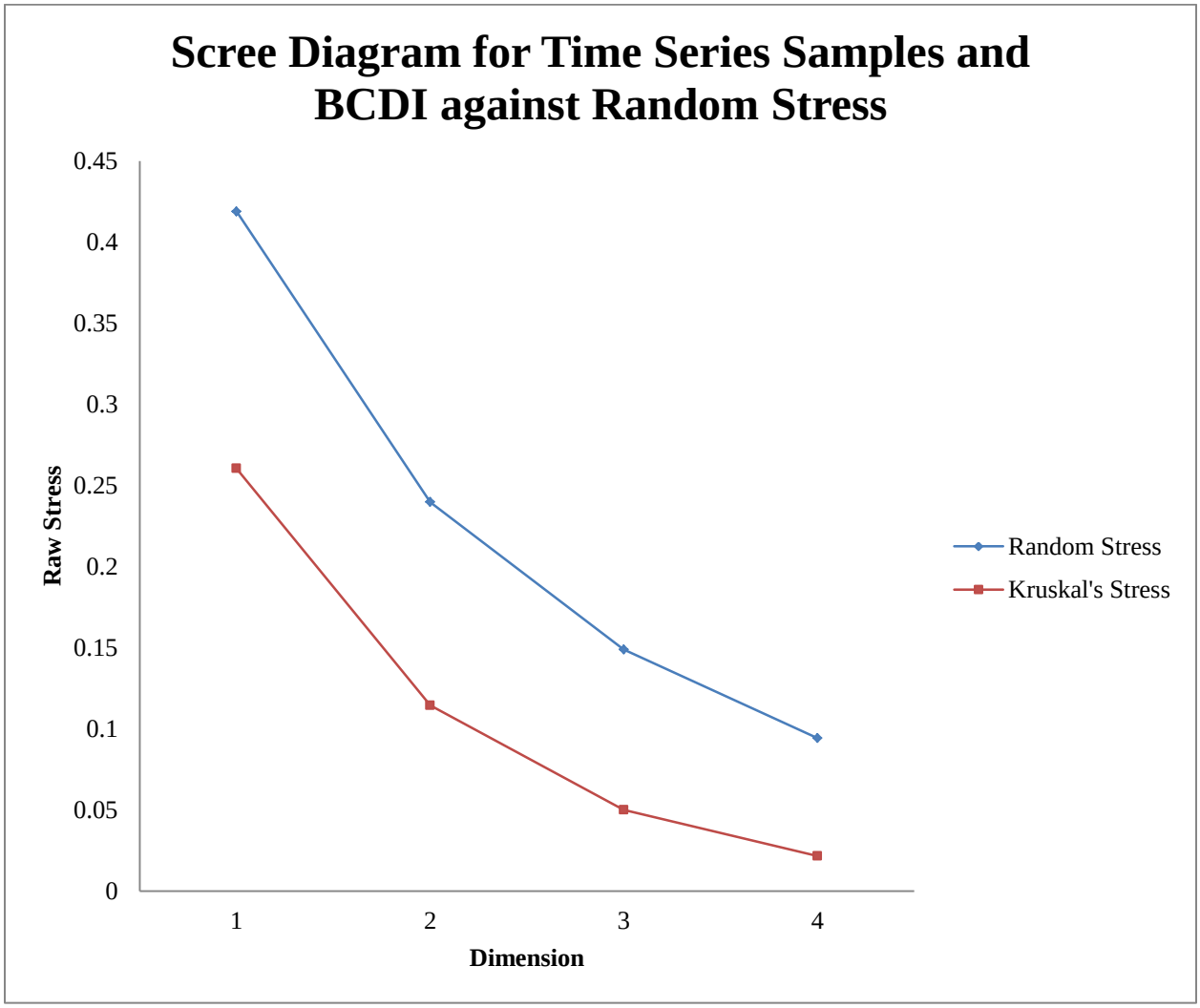

Figure 116. Scree diagram developed by NMDS for the final configurations of time series soil samples and BCDI over four dimensions. Random stress was used as a threshold for the acceptance of final configurations. No configuration exceeded the threshold though the stress for the one dimension plot was much higher than the others. There is no elbow at any dimension; however, for consistency two dimensional plots were chosen.

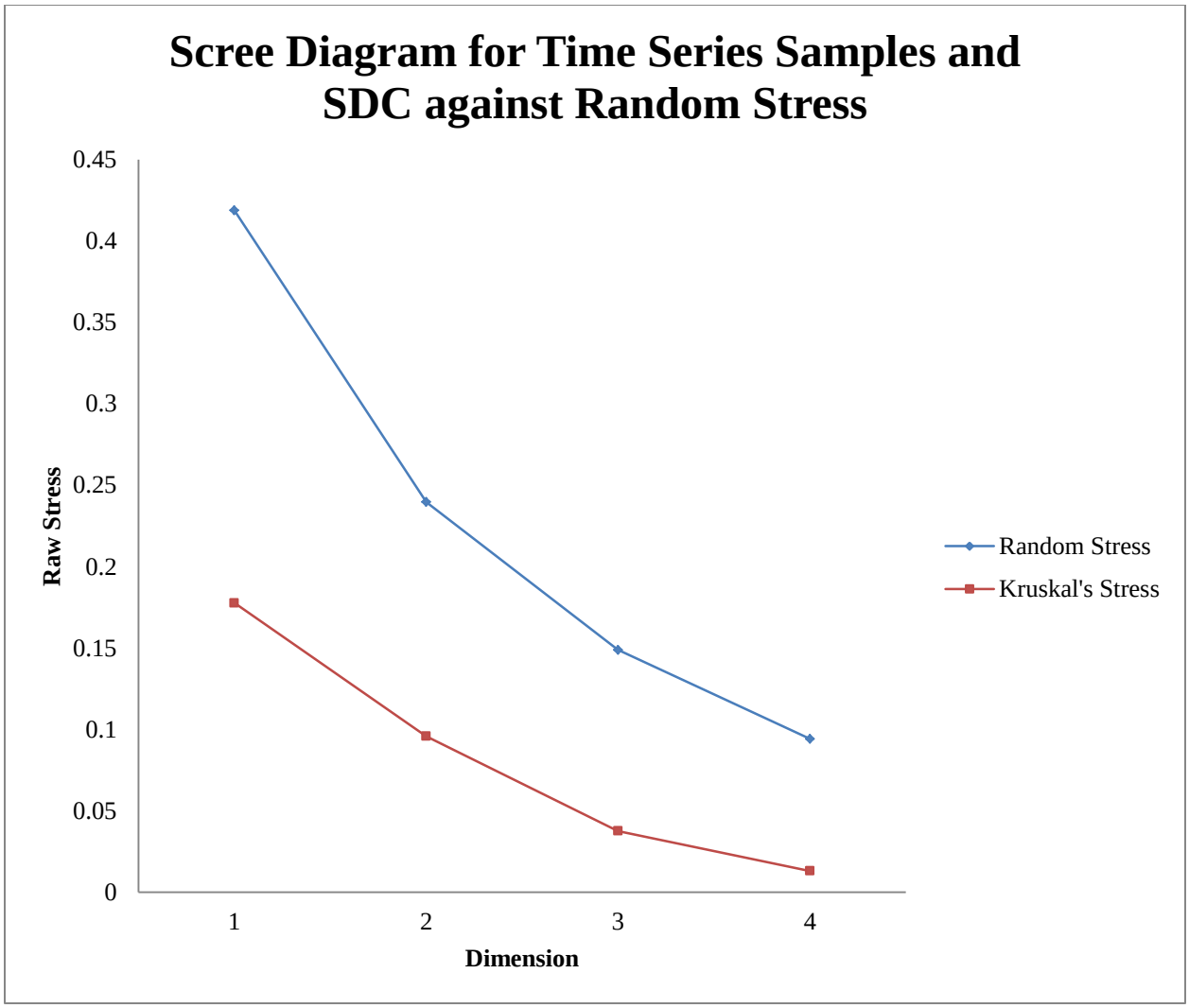

Figure 117. Scree diagram developed by NMDS for the final configurations of time series soil samples and SDC over four dimensions. Random stress was used as a threshold for the acceptance of final configurations. No configuration exceeded the threshold though the stress for the one dimension plot was higher than the others. No elbow is noticeable at any dimension; however, for consistency two dimensional plots were chosen.

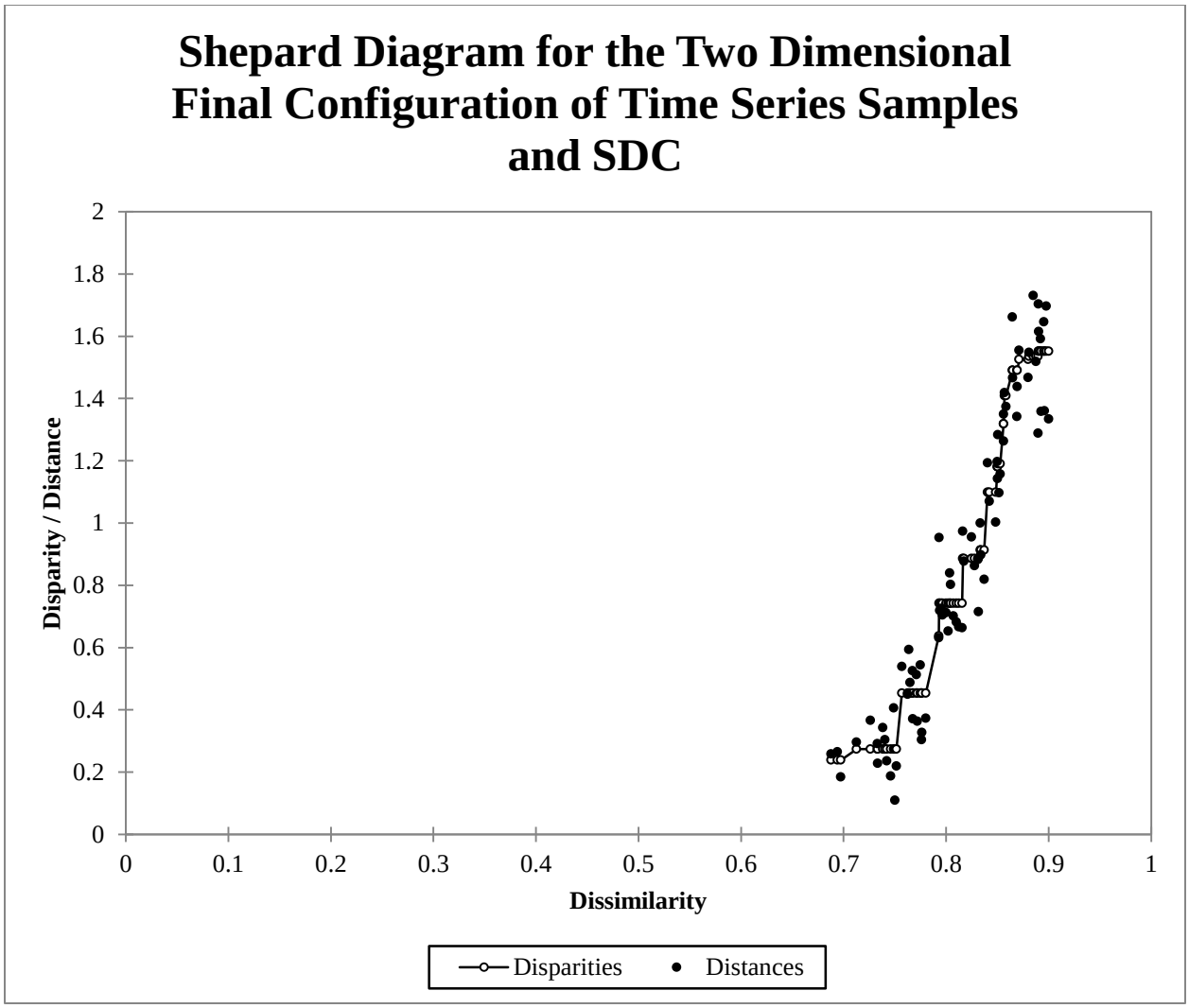

Figure 118. Shepard diagram for the two dimensional final configuration developed from SDC of time series soil samples. Distances do not associate well with their corresponding disparities agreeing with the higher stress seen with the final configuration.

APPENDIX 14. DENDROGRAMS OF TIME SERIES SAMPLES FOR BCDI AND SDC.

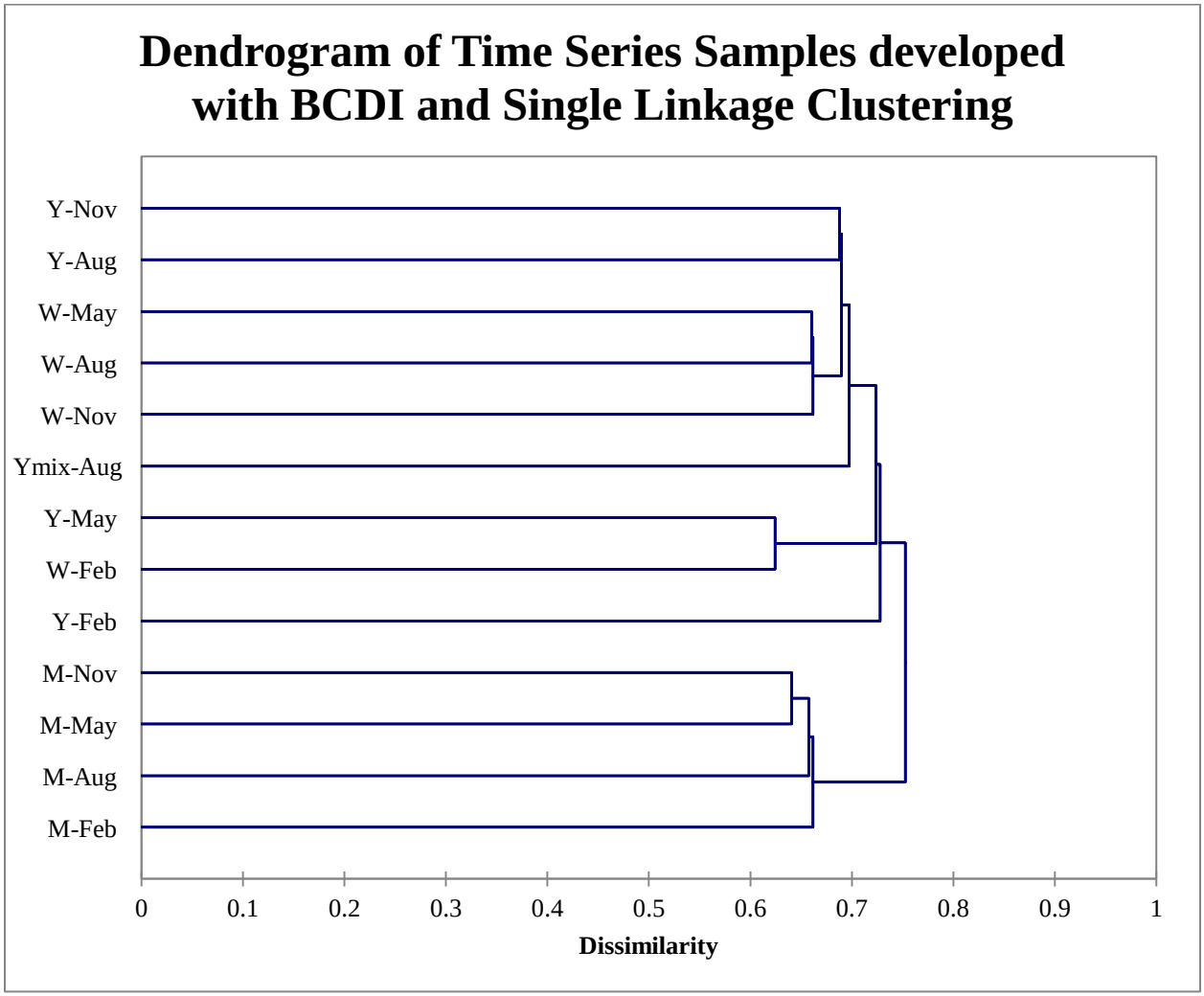

Figure 119. Dendrogram of time series soil samples developed with BCDI and single linkage clustering. Four clusters are present at 0.690, 0.662, 0.624, and 0.662. The first is a two member group of the yard August and November samples while the second is composed of the deciduous woods May and August samples followed by the November time point. The third cluster is another two member grouping of the yard May and deciduous woods February samples. The final cluster is comprised of all the marsh edge samples where the November and May samples are the most similar followed by the August then February ones. The Ymix August sample clusters with the first two groups at 0.697 while the yard February sample clusters with the first three groups at 0.728. The marsh edge cluster groups with the rest of the samples at 0.753. See Table 2 for site names corresponding to abbreviations.

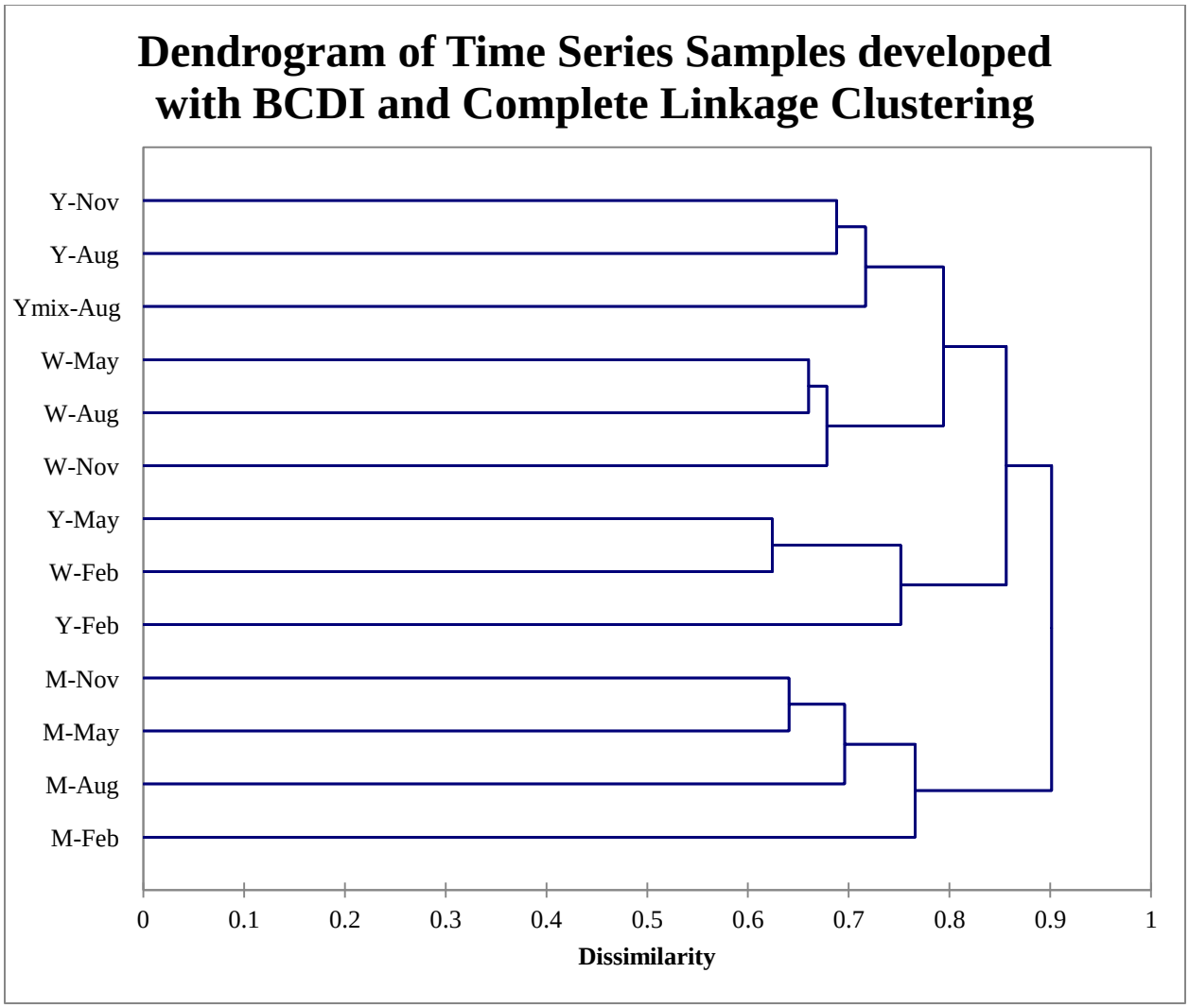

Figure 120. Dendrogram of time series soil samples developed with BCDI and complete linkage clustering. Four noticeable clusters are present. The first consists of the yard November and August samples followed by the Ymix August sample grouping at 0.717. The second cluster has the deciduous woods May and August samples being most similar then the November sample at 0.678. The third grouping is comprised of the yard May and deciduous woods February time points followed by the yard February one joining at 0.752. The final cluster formed at 0.766 is comprised of all the marsh edge samples where the November and May samples are the most similar followed by the August then February ones. All clusters are joined at a dissimilarity of 0.794 or greater. See Table 2 for site names corresponding to abbreviations.

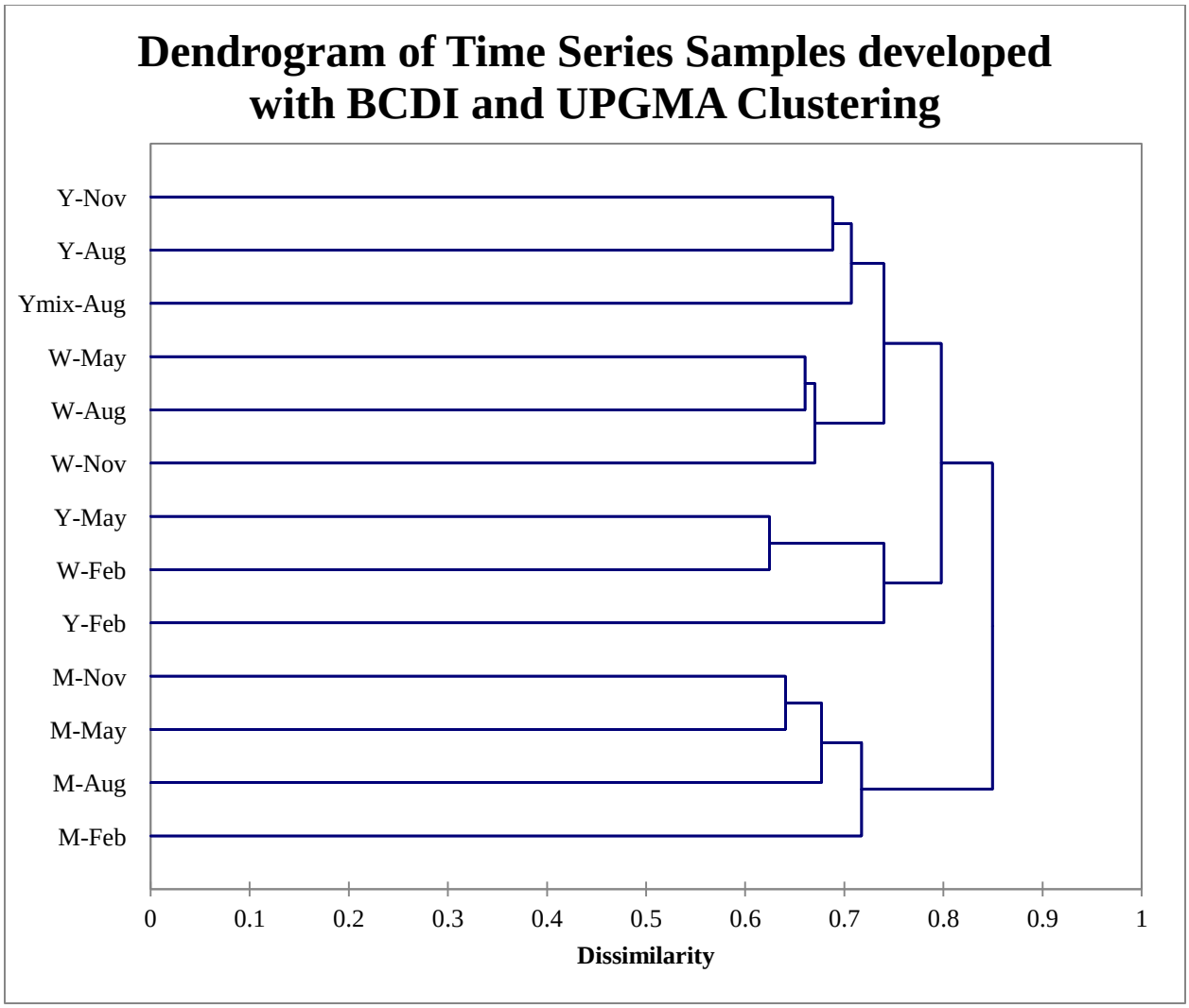

Figure 121. Dendrogram of time series soil samples developed with BCDI and UPGMA clustering. Four noticeable clusters are present. The first consists of the yard November and August samples followed by the Ymix August sample grouping at 0.707. The second cluster has the deciduous woods May and August samples being most similar then the November sample joining at 0.670. The third grouping is comprised of the yard May and deciduous woods February time points followed by the yard February one at 0.740. The final cluster formed at 0.717 is comprised of all the marsh edge samples where the November and May samples are the most similar followed by the August then February ones. All clusters are joined at a dissimilarity of 0.740 or greater. See Table 2 for site names corresponding to abbreviations.

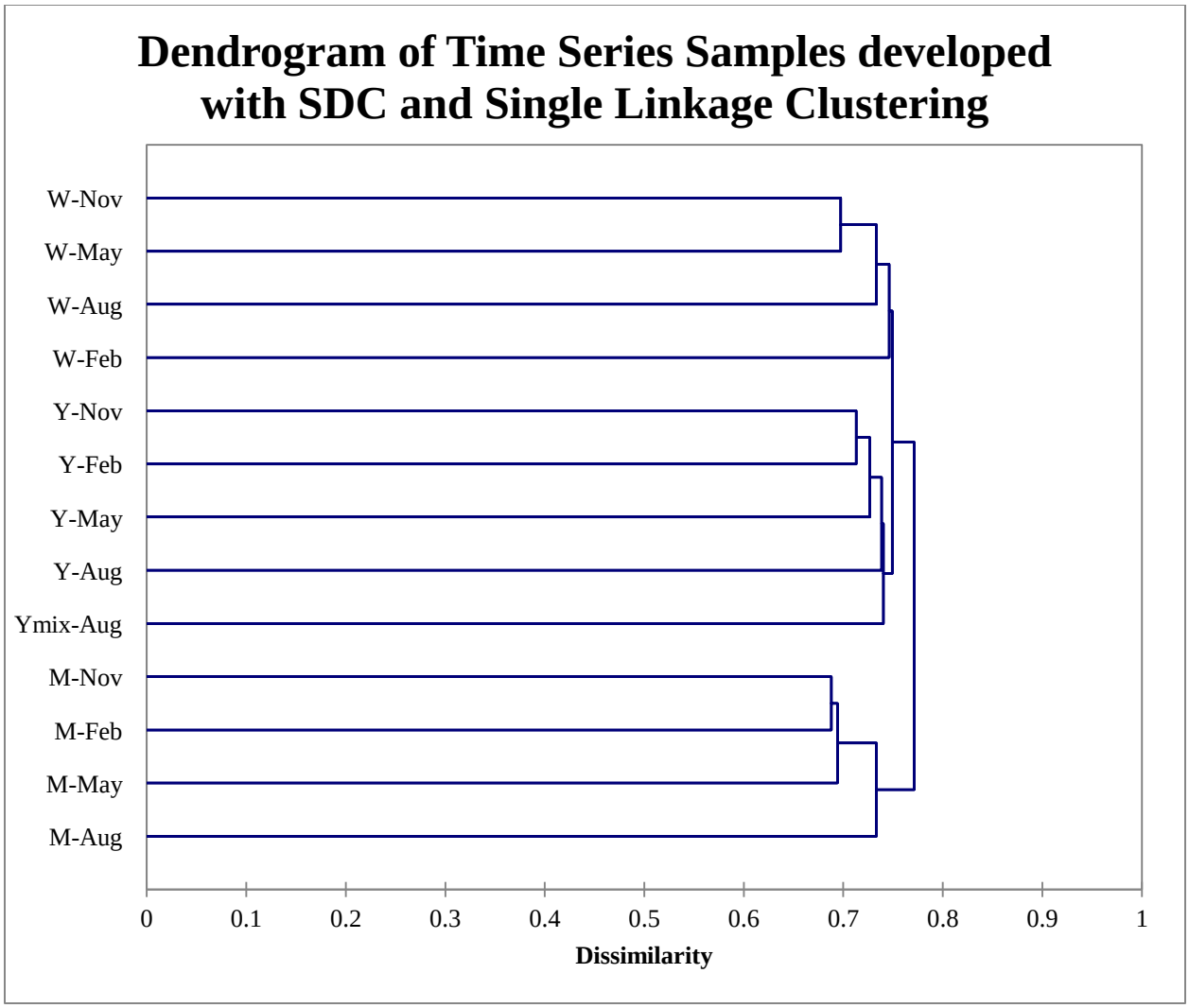

Figure 122. Dendrogram of time series soil samples developed with SDC and single linkage clustering. Three clusters are formed by habitat, two at 0.749 and the other at 0.733. The first contains the deciduous woods November and May samples clustering first followed by the August and February ones. The second cluster has the yard November and February samples being most similar preceded by the May, August, and Ymix August time points. The third is comprised of the marsh edge November and February samples followed by the May then August ones. The first and second clusters group around 0.749 with the third grouping around 0.771. See Table 2 for site names corresponding to abbreviations.

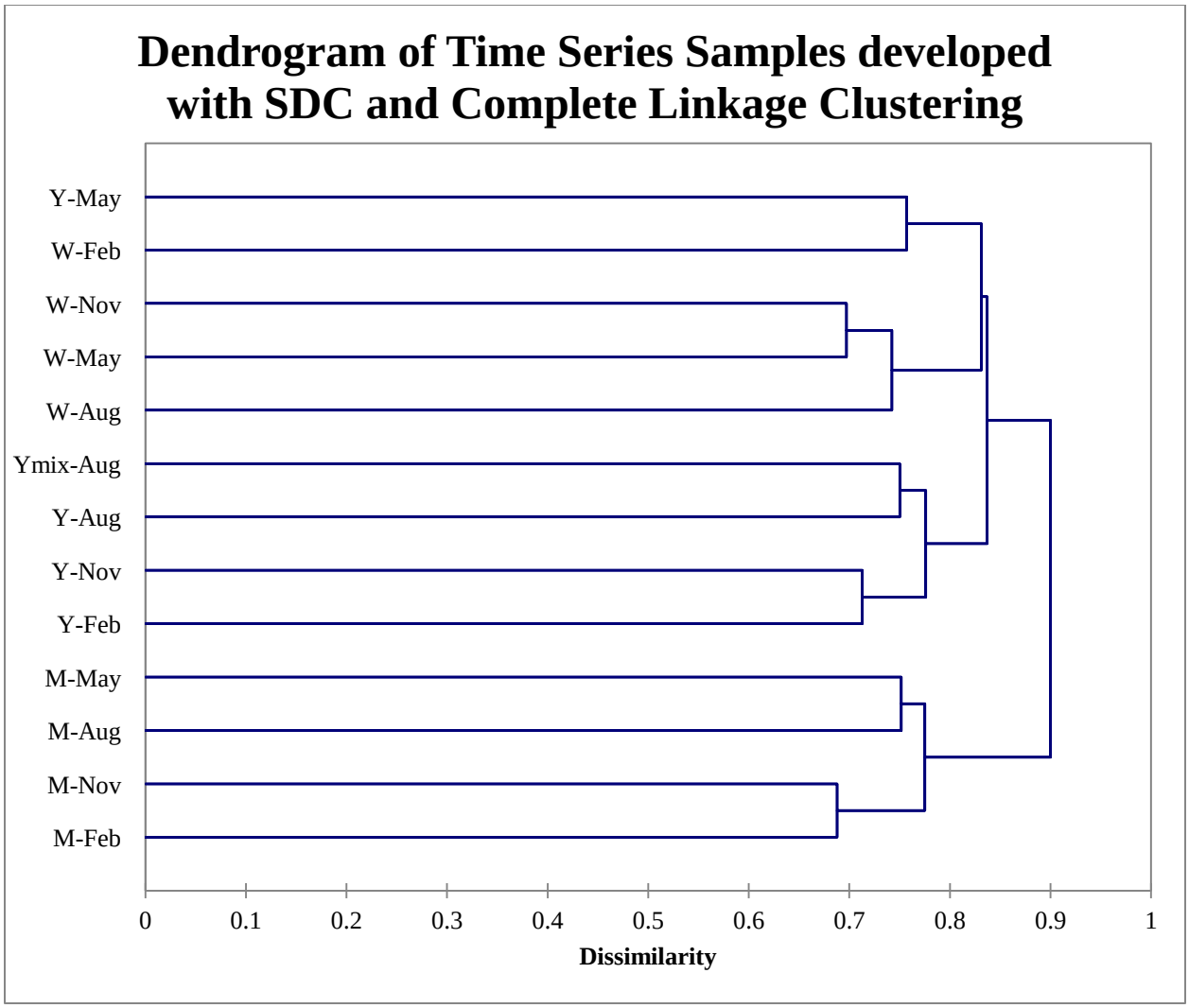

Figure 123. Dendrogram of time series soil samples developed with SDC and complete linkage clustering. There are four clusters present, three of which contain samples from only one habitat. The first cluster is a two member group composed of the yard May and deciduous woods February samples at 0.757. The second is formed at 0.743 first by the deciduous woods November and May time points, followed by the August sample. The first and second clusters join at a dissimilarity of 0.837. The third cluster contains two two member groups, the yard August and Ymix August samples and the yard November and February samples. These two join together at 0.776 and then cluster with the first and second groups at 0.837. The final cluster is another formed of two two member groups, the marsh edge May and August samples and the marsh edge November and February samples. They join together at 0.775 then with the others at 0.900. See Table 2 for site names corresponding to abbreviations.

APPENDIX 15. ADDITIONAL NMDS DIAGRAMS FOR SIMILAR HABITAT SAMPLES.

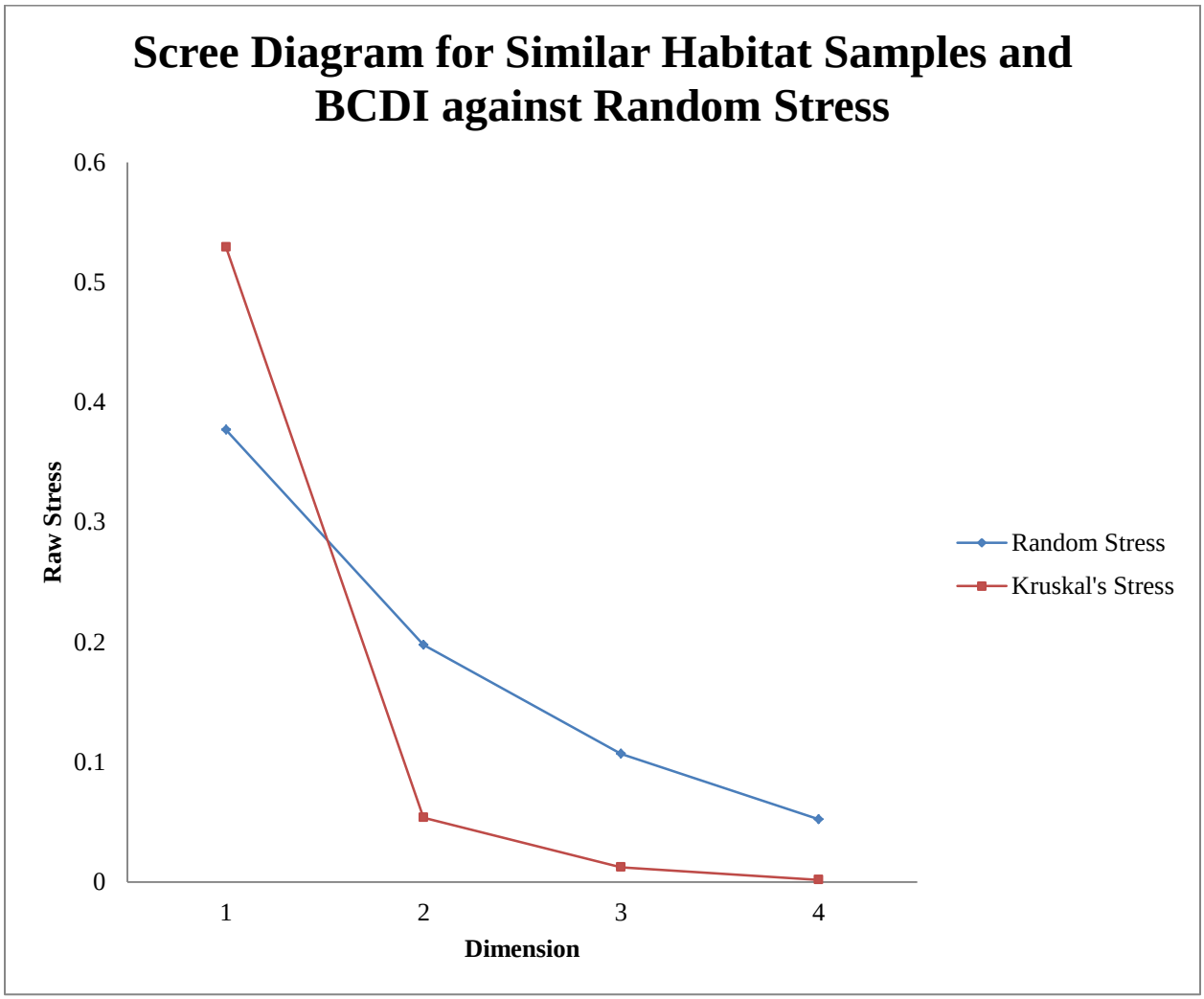

Figure 124. Scree diagram developed by NMDS for the final configurations of similar habitat soil samples and BCDI over four dimensions. Random stress was used as a threshold for the acceptance of final configurations. The one dimension configuration was the only to exceed the threshold and was rejected. The elbow in the curve is at two dimensions.

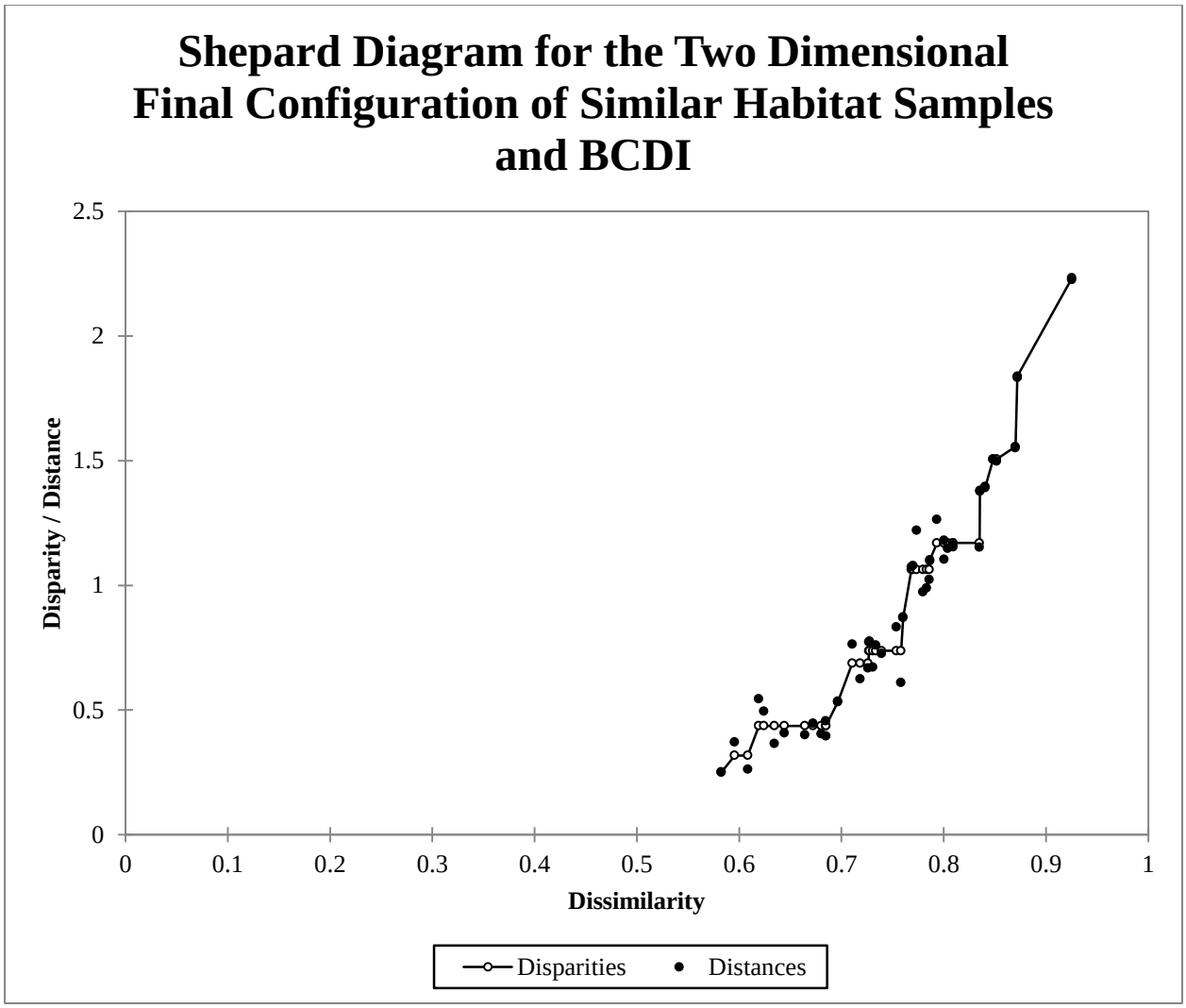

Figure 125. Shepard diagram for the two dimensional final configuration developed from BCDI of similar habitat soil samples. Most distances fall close to their corresponding disparities indicating good correlation of the two in the final configuration.

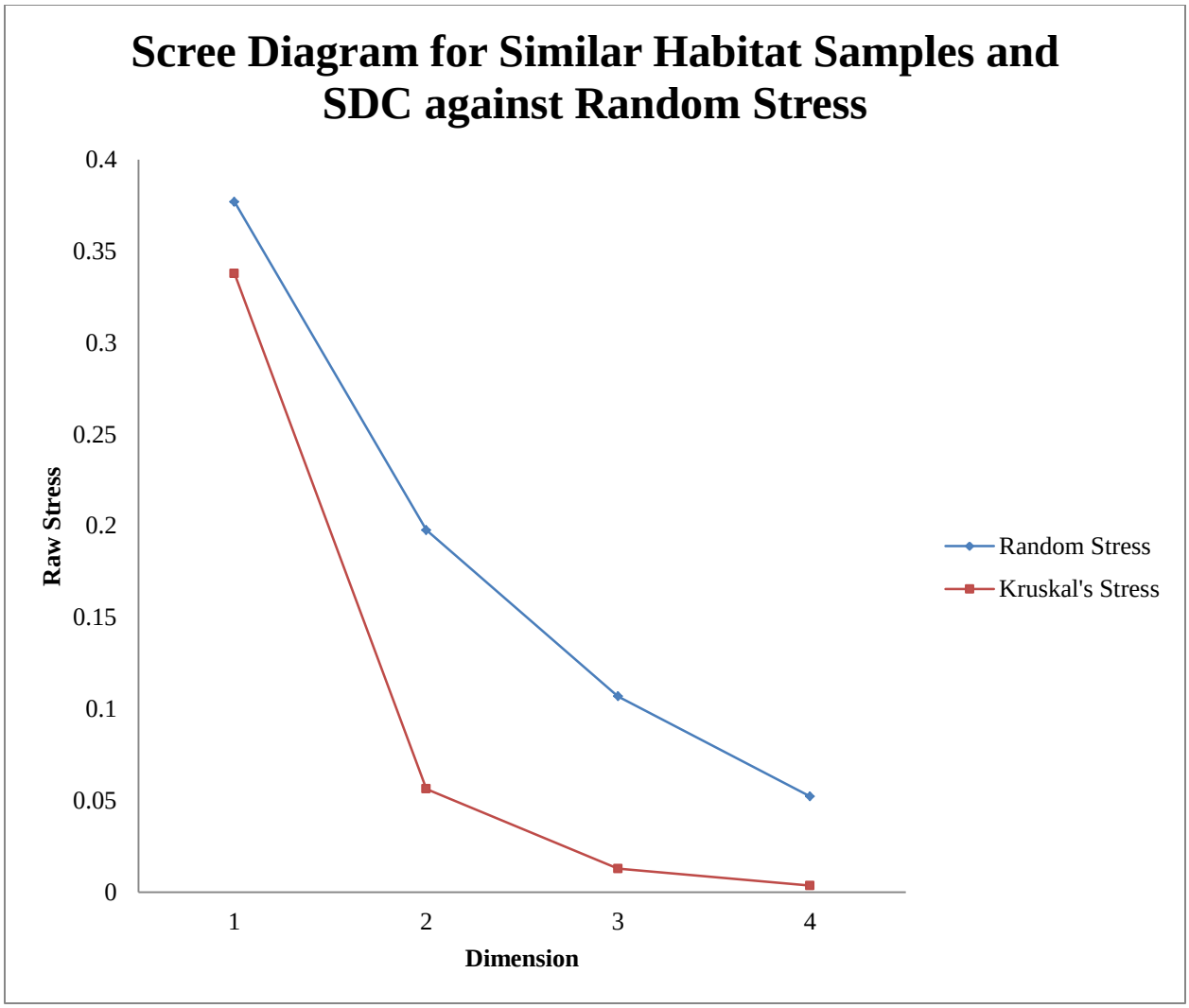

Figure 126. Scree diagram developed by NMDS for the final configurations of similar habitat soil samples and SDC over four dimensions. Random stress was used as a threshold for the acceptance of final configurations. No configuration exceeded the threshold though the stress for the one dimension plot was higher than the others. The elbow in the curve is at two dimensions.

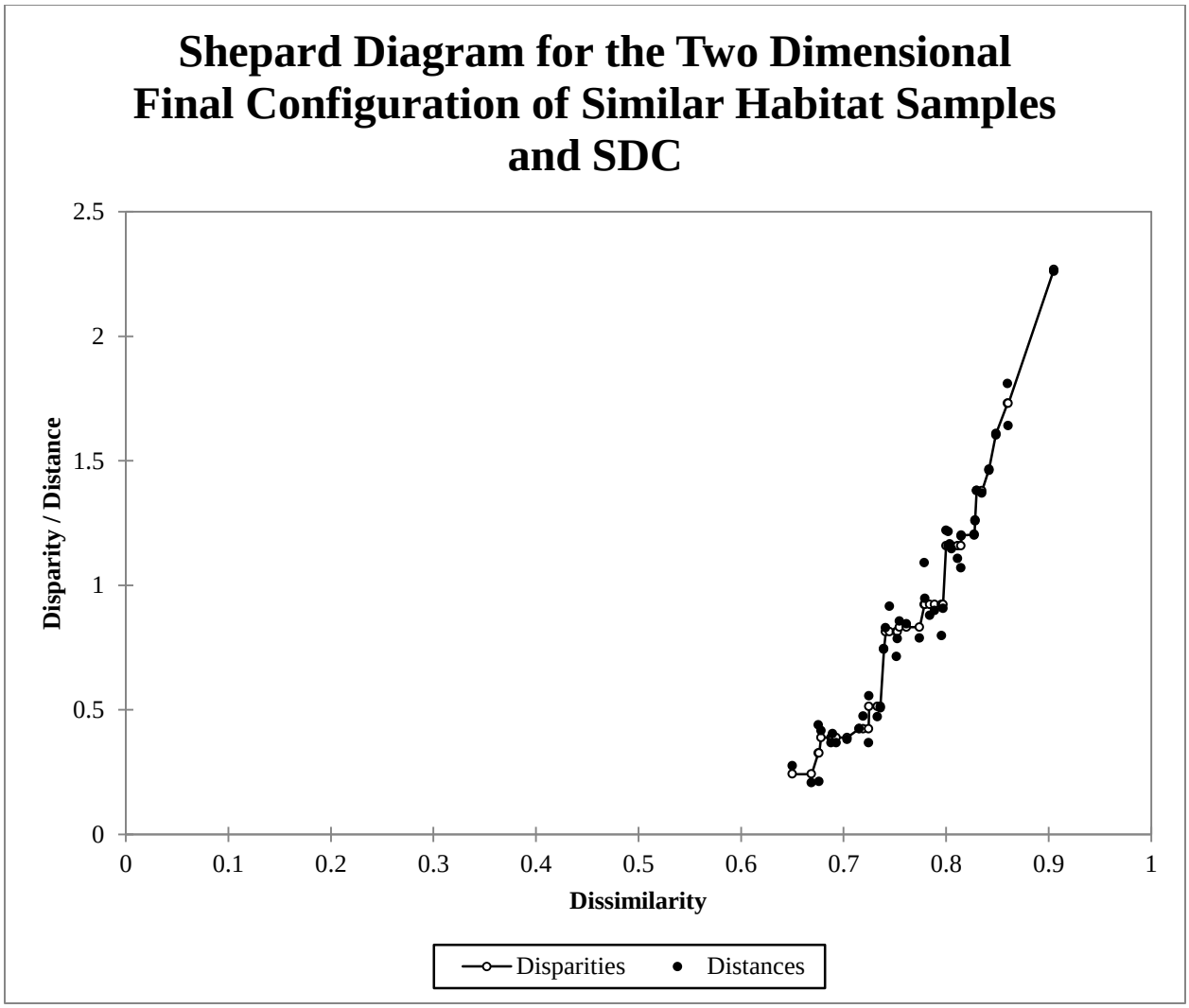

Figure 127. Shepard diagram for the two dimensional final configuration developed from SDC of similar habitat soil samples. Most distances fall close to their corresponding disparities indicating good correlation of the two in the final configuration.

APPENDIX 16. DENDROGRAMS OF SIMILAR HABITAT SAMPLES FOR BCDI AND SDC.

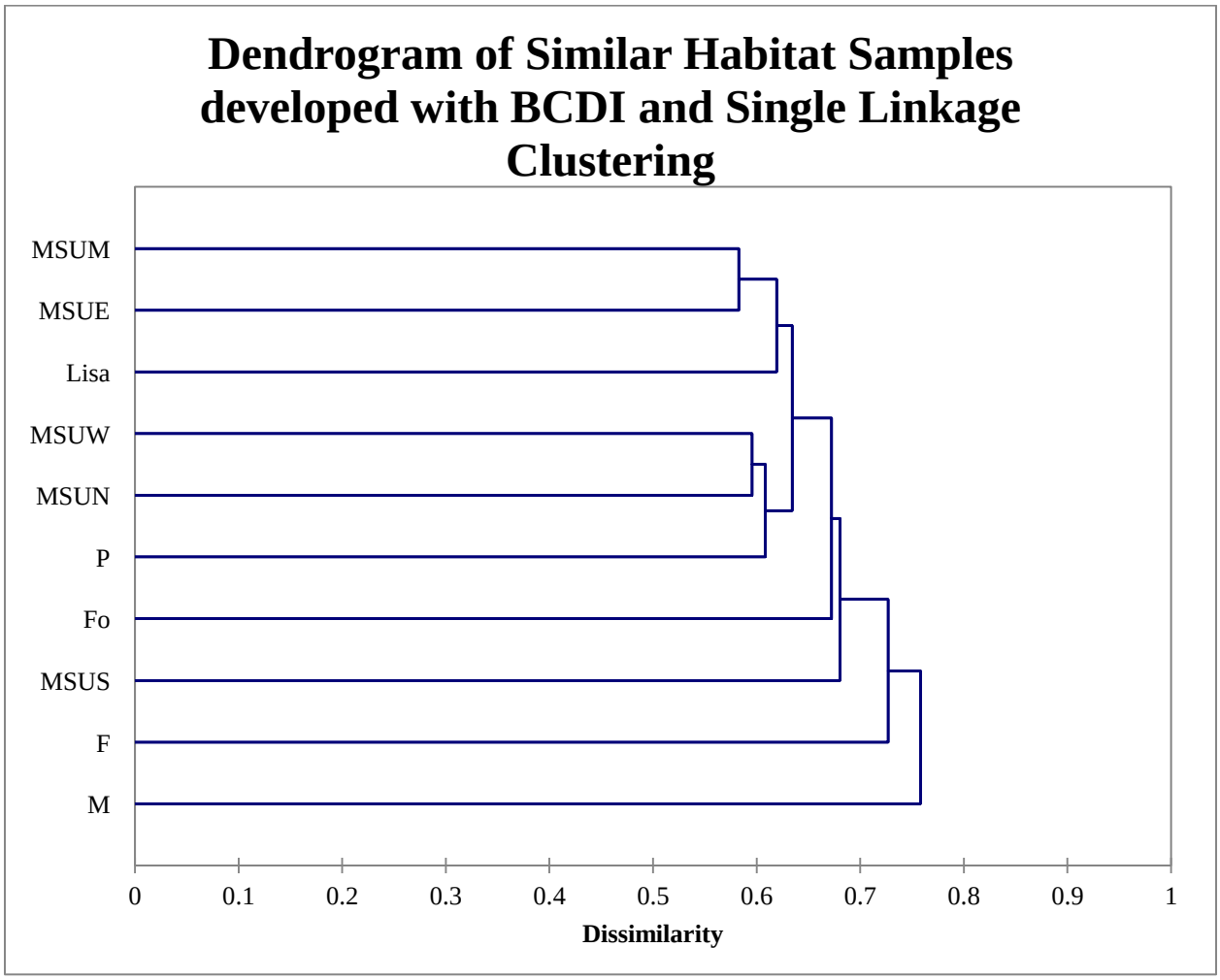

Figure 128. Dendrogram of similar habitat soil samples developed with BCDI and single linkage clustering. Two clusters are evident at 0.619 and 0.608. The first has the Michigan State University main and east samples being the most similar followed by the Lisa yard. The second is comprised of the Michigan State University west and north samples clustering first then the Perry yard. The remaining samples cluster with the two groups greater than 0.634 dissimilarity with the final sample, the Michelle yard, clustering under 0.758. See Table 5 for site names corresponding to abbreviations.

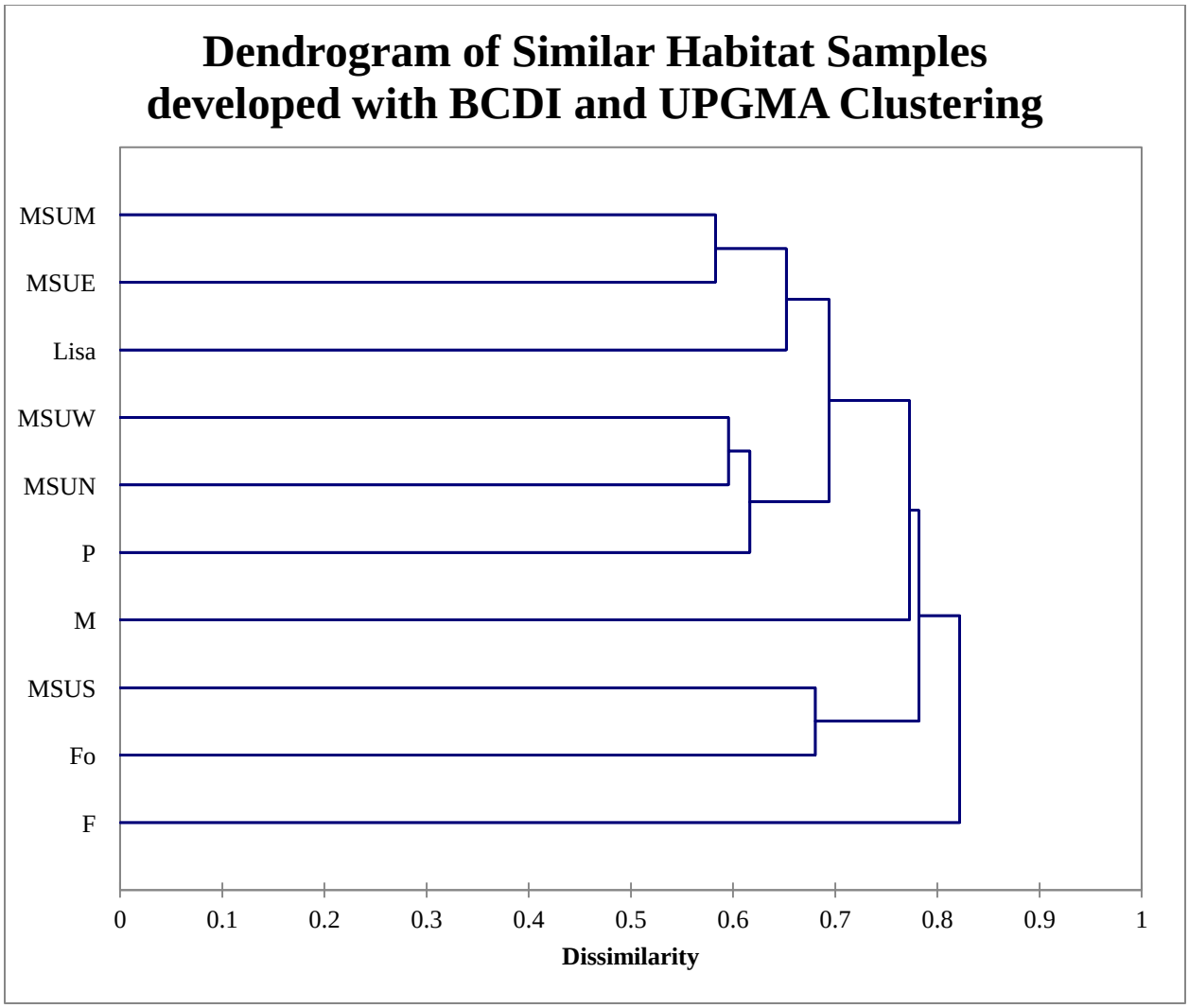

Figure 129. Dendrogram of similar habitat soil samples developed with BCDI and UPGMA clustering. Three clusters are formed at 0.652, 0.616, and 0.680. The first has the Michigan State University main and east samples being the most similar followed by the Lisa yard. The second is comprised of the Michigan State University west and north samples clustering first then the Perry yard. The final cluster contains the Michigan State University and Foran yards. The Michelle yard groups with the first two clusters at 0.773, followed by the third cluster at 0.782. The Fenner yard clusters with the rest at 0.822. See Table 5 for site names corresponding to abbreviations.

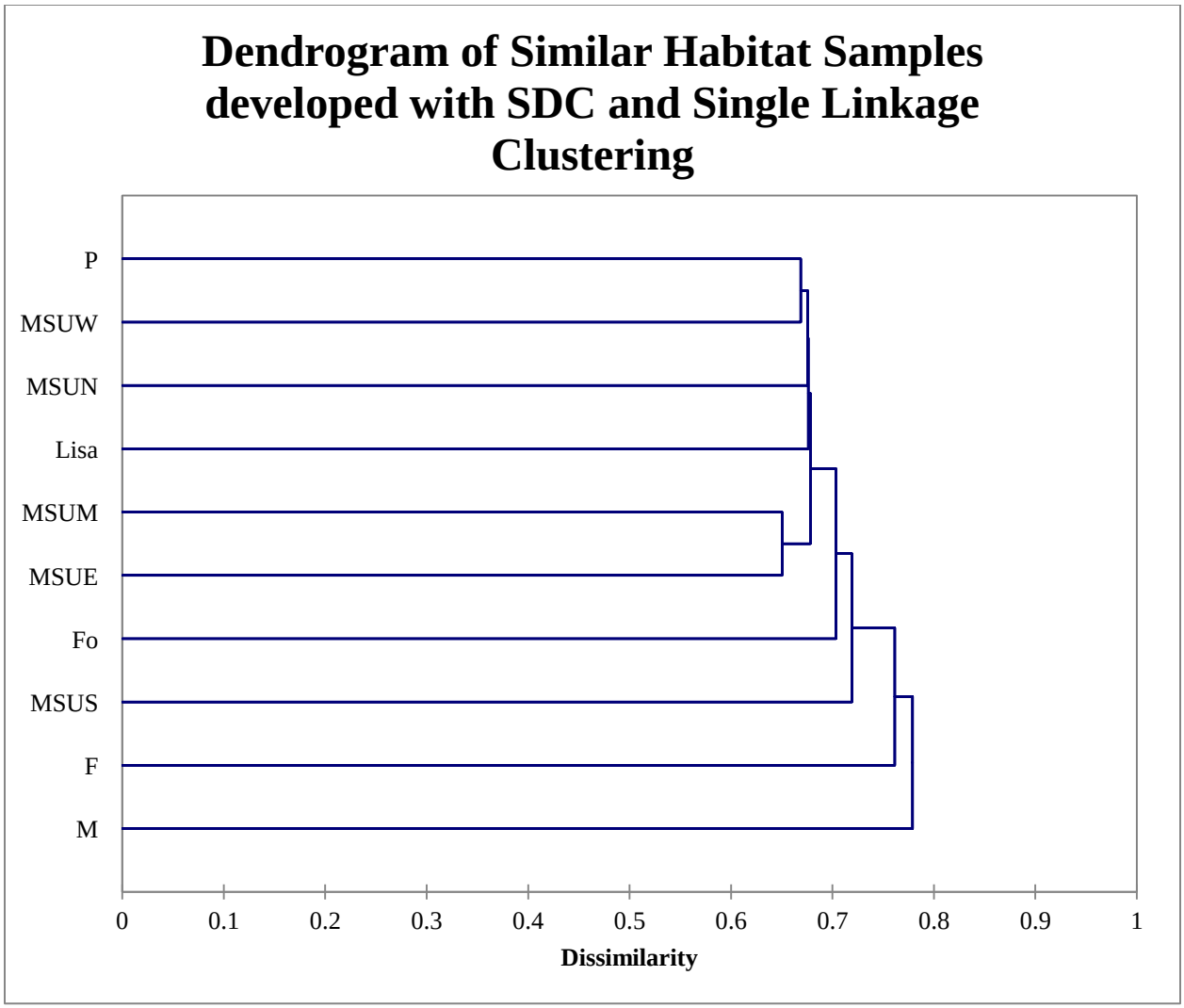

Figure 130. Dendrogram of similar habitat soil samples developed with SDC and single linkage clustering. Two clusters are evident at 0.676 and 0.650. The first has the Michigan State University main and east samples. The second is comprised of the Michigan State University west and Perry yards. This group is followed closely by the Michigan State University north and Lisa yards then the first cluster. The remaining four samples cluster with the others at 0.704 or greater, with the final sample, the Michelle yard, clustering at 0.779. See Table 5 for site names corresponding to abbreviations.

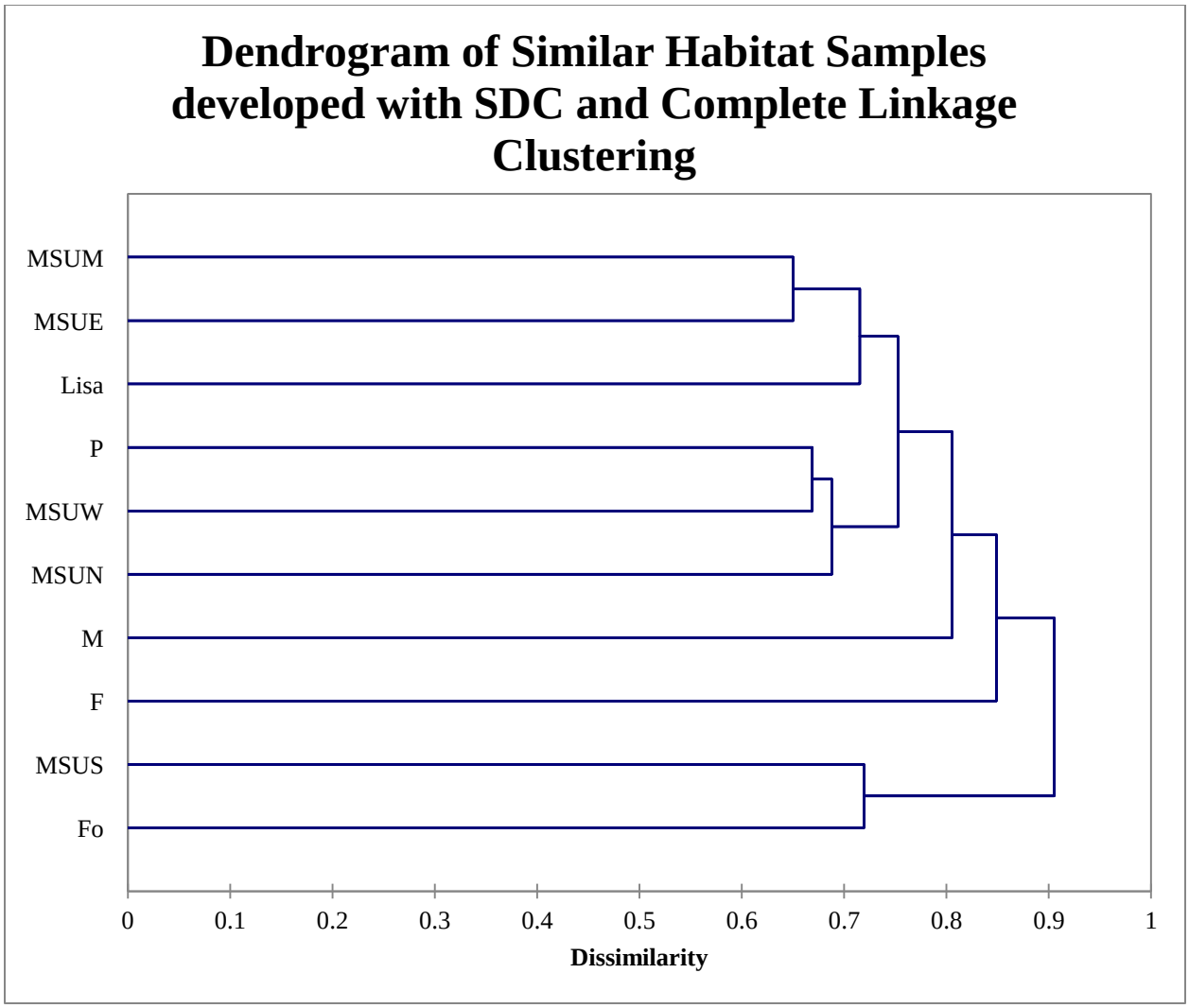

Figure 131. Dendrogram of similar habitat soil samples developed with SDC and complete linkage clustering. Three clusters are present. The first has the Michigan State University main and east samples being the most similar followed by the Lisa yard clustering at 0.715. The second is comprised of the Michigan State University west and Perry samples clustering first then the Michigan State University north yard at 0.688. The final cluster contains the Michigan State University and Foran yards joining at 0.719. The two remaining samples cluster with the first two groups at 0.805 or greater. The third cluster groups with the rest at 0.905. See Table 5 for site names corresponding to abbreviations.

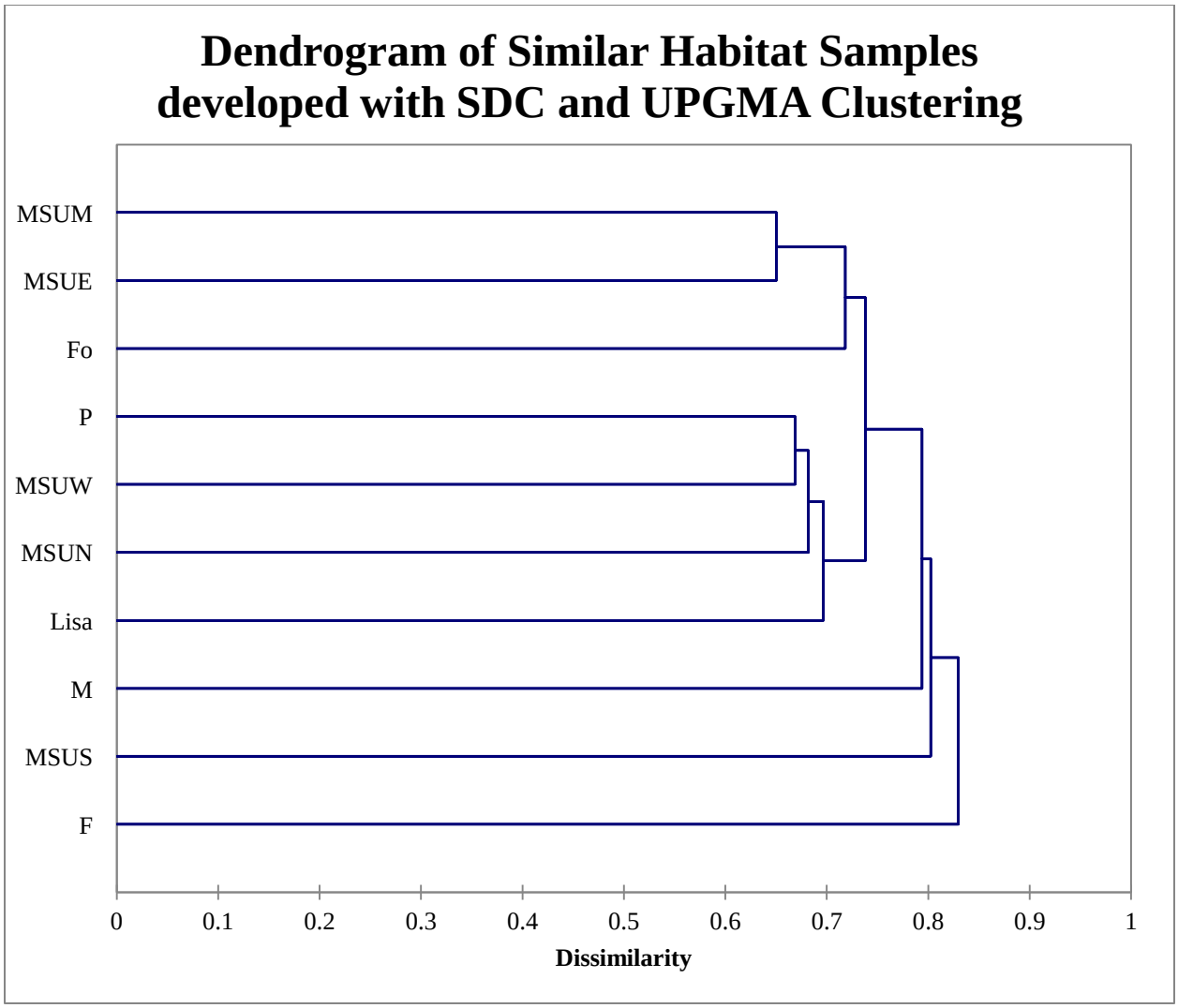

Figure 132. Dendrogram of similar habitat soil samples developed with SDC and UPGMA clustering. Two clusters are present. The first has the Michigan State University main and east samples being the most similar followed by the Foran yard clustering at 0.718. The second is comprised of the Michigan State University west and Perry yards clustering first then the Michigan State University north and Lisa samples at 0.697. The three remaining samples cluster with the rest at 0.794 or greater. See Table 5 for site names corresponding to abbreviations.

APPENDIX 17. ADDITIONAL NMDS DIAGRAMS FOR DIVERSE HABITAT SAMPLES.

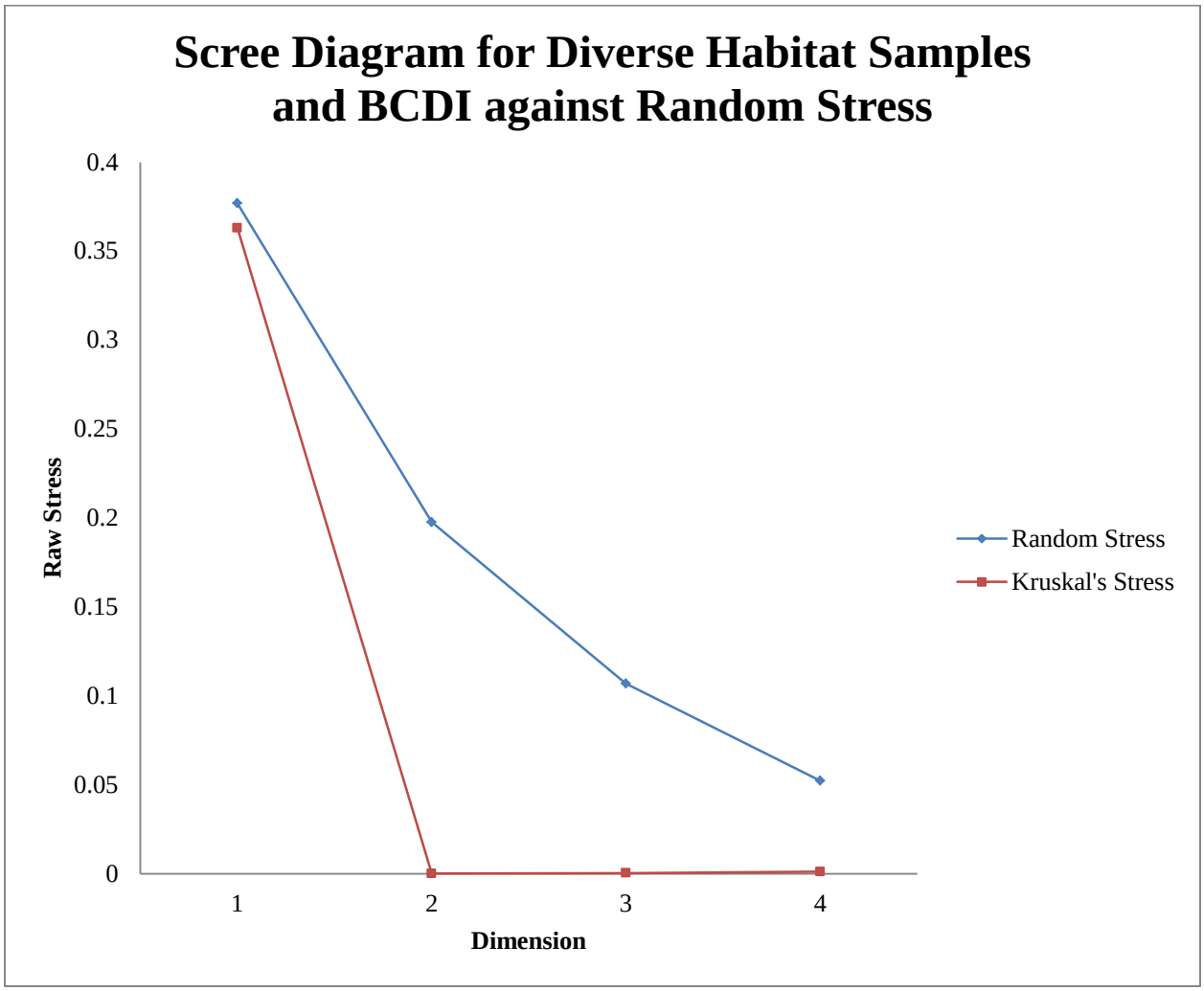

Figure 133. Scree diagram developed by NMDS for the final configurations of diverse habitat soil samples and BCDI over four dimensions. Random stress was used as a threshold for the acceptance of final configurations. No configuration exceeded the threshold though the stress for the one dimension plot was much higher than the others. The elbow in the curve is at two dimensions.

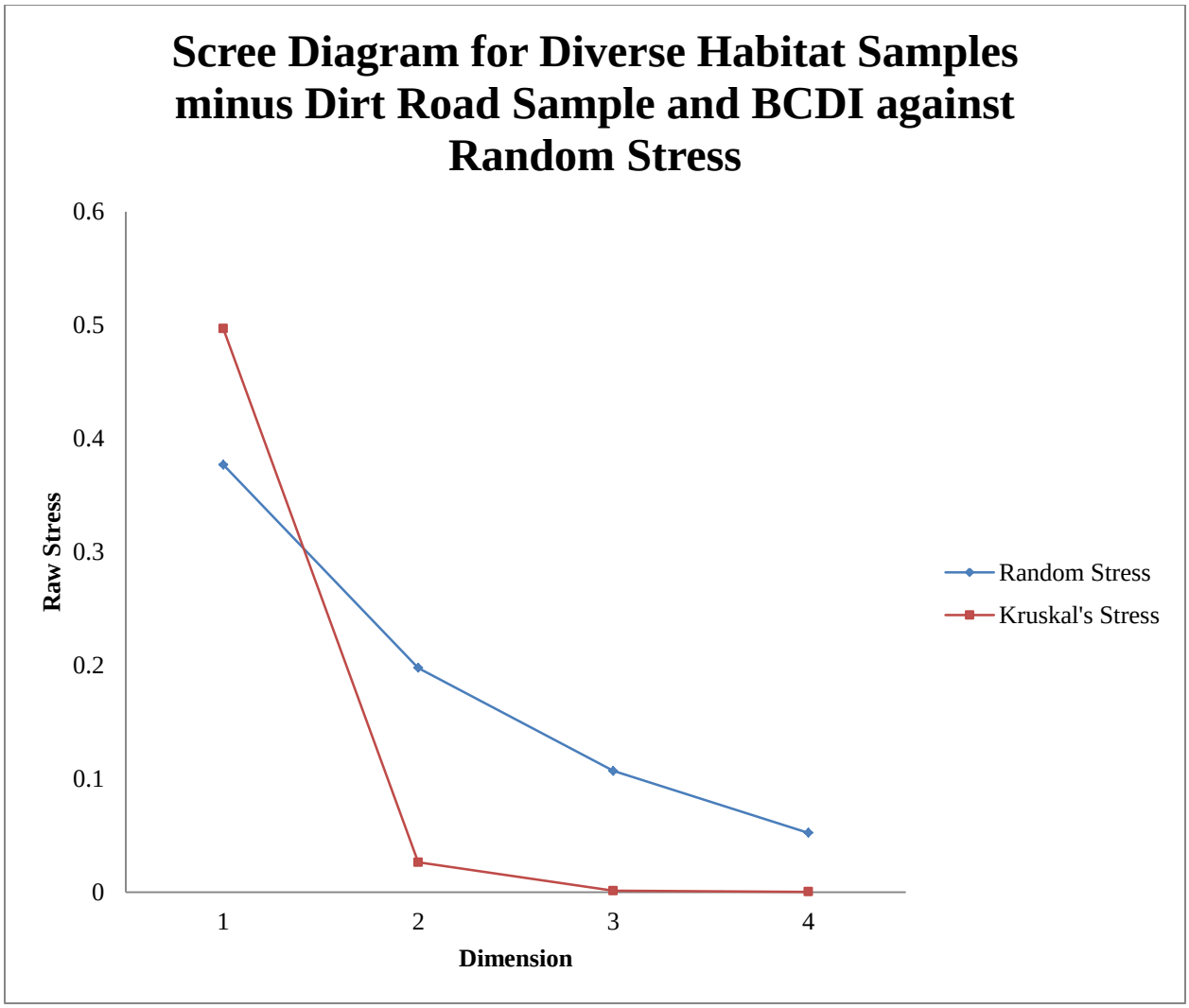

Figure 134. Scree diagram developed by NMDS for the final configurations of diverse habitat soil samples, minus the dirt road sample, and BCDI over four dimensions. Random stress was used as a threshold for the acceptance of final configurations. The one dimension configuration was the only to exceed the threshold and was rejected. The elbow in the curve is at two dimensions.

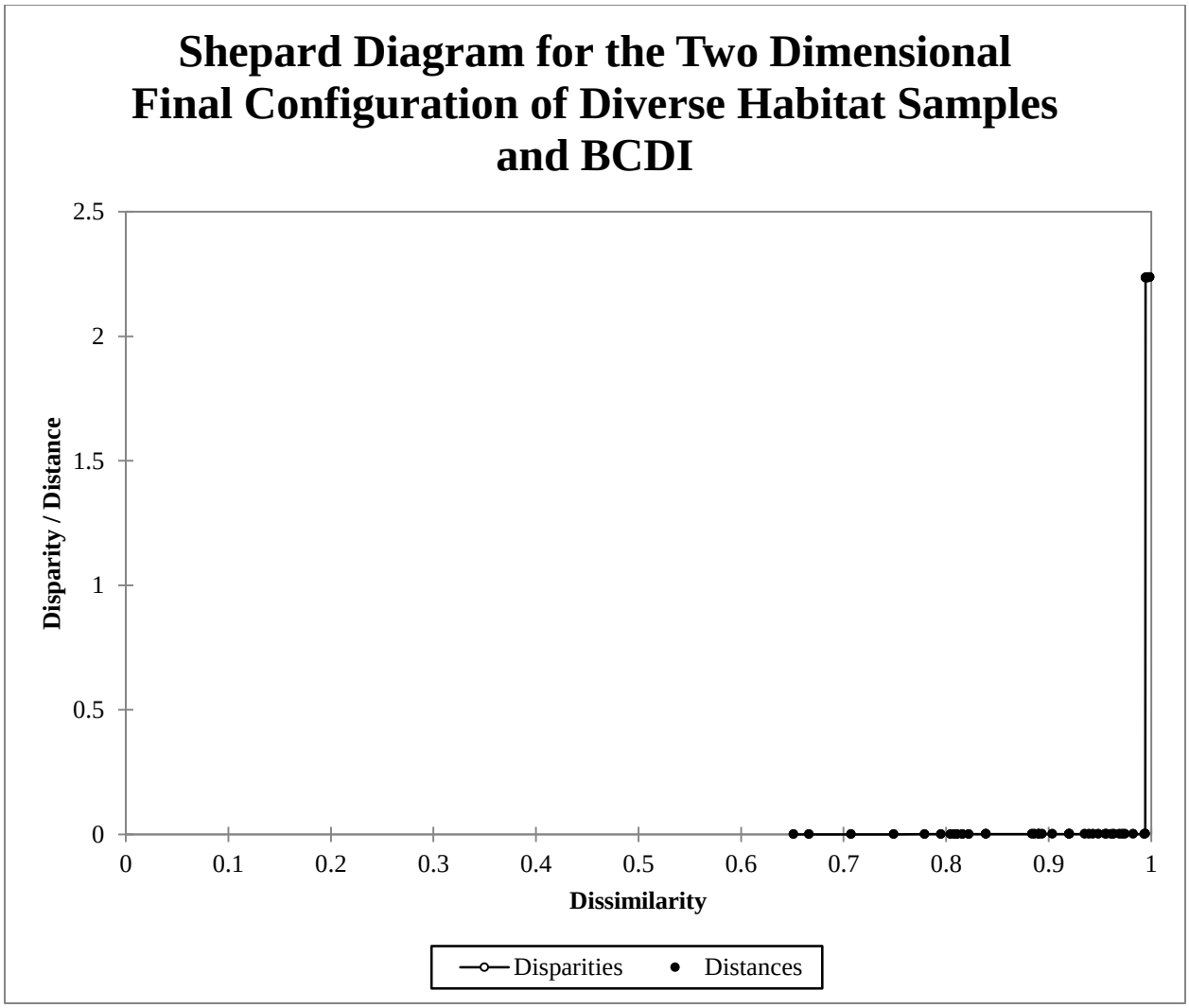

Figure 135. Shepard diagram for the two dimensional final configuration developed from BCDI of diverse habitat soil samples. All distances fall nearly on top of their corresponding disparities indicating good correlation of the two in the final configuration.

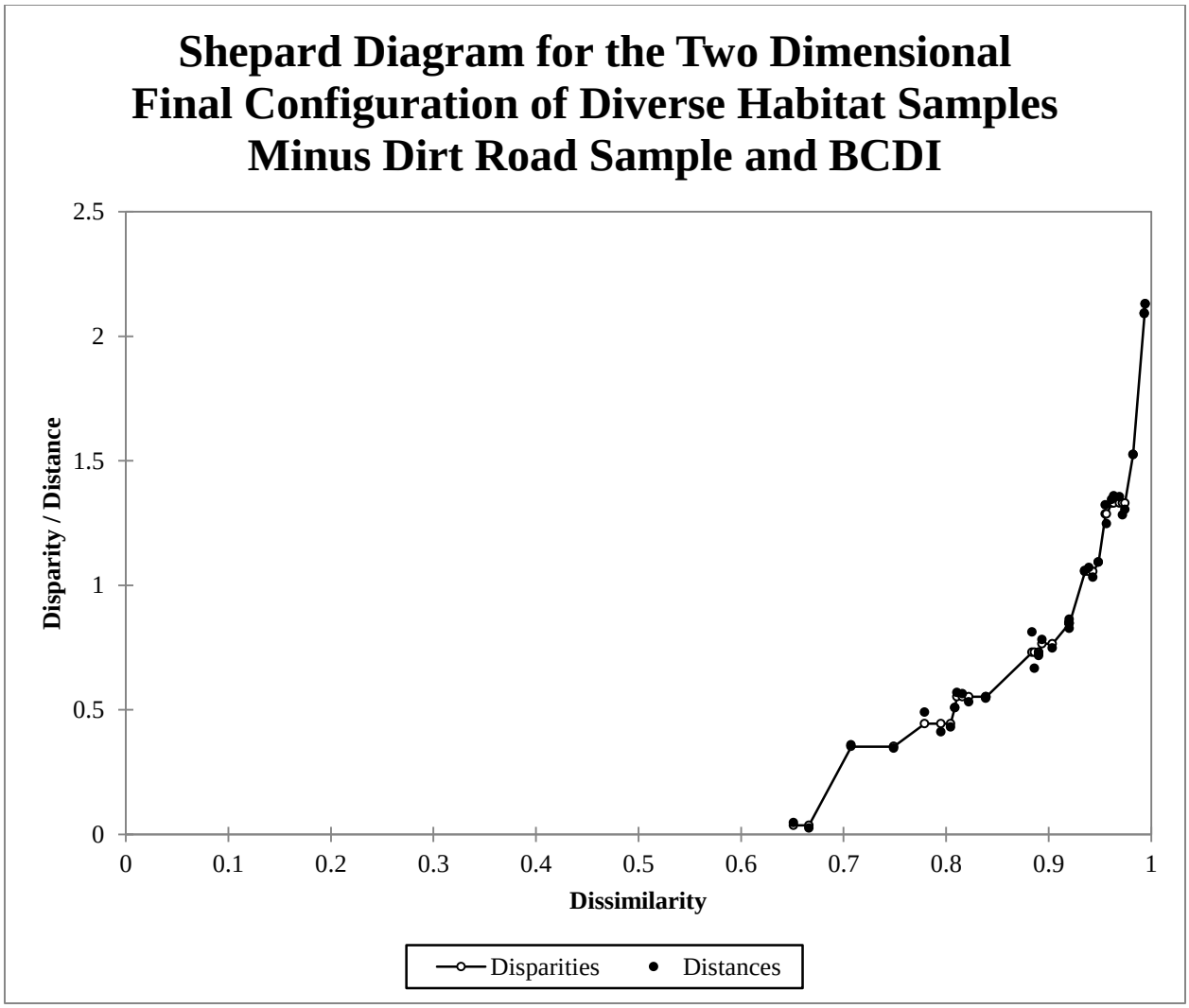

Figure 136. Shepard diagram for the two dimensional final configuration developed from BCDI of the diverse habitat soil samples minus the dirt road sample. All distances fall nearly on top of their corresponding disparities indicating good correlation of the two in the final configuration.

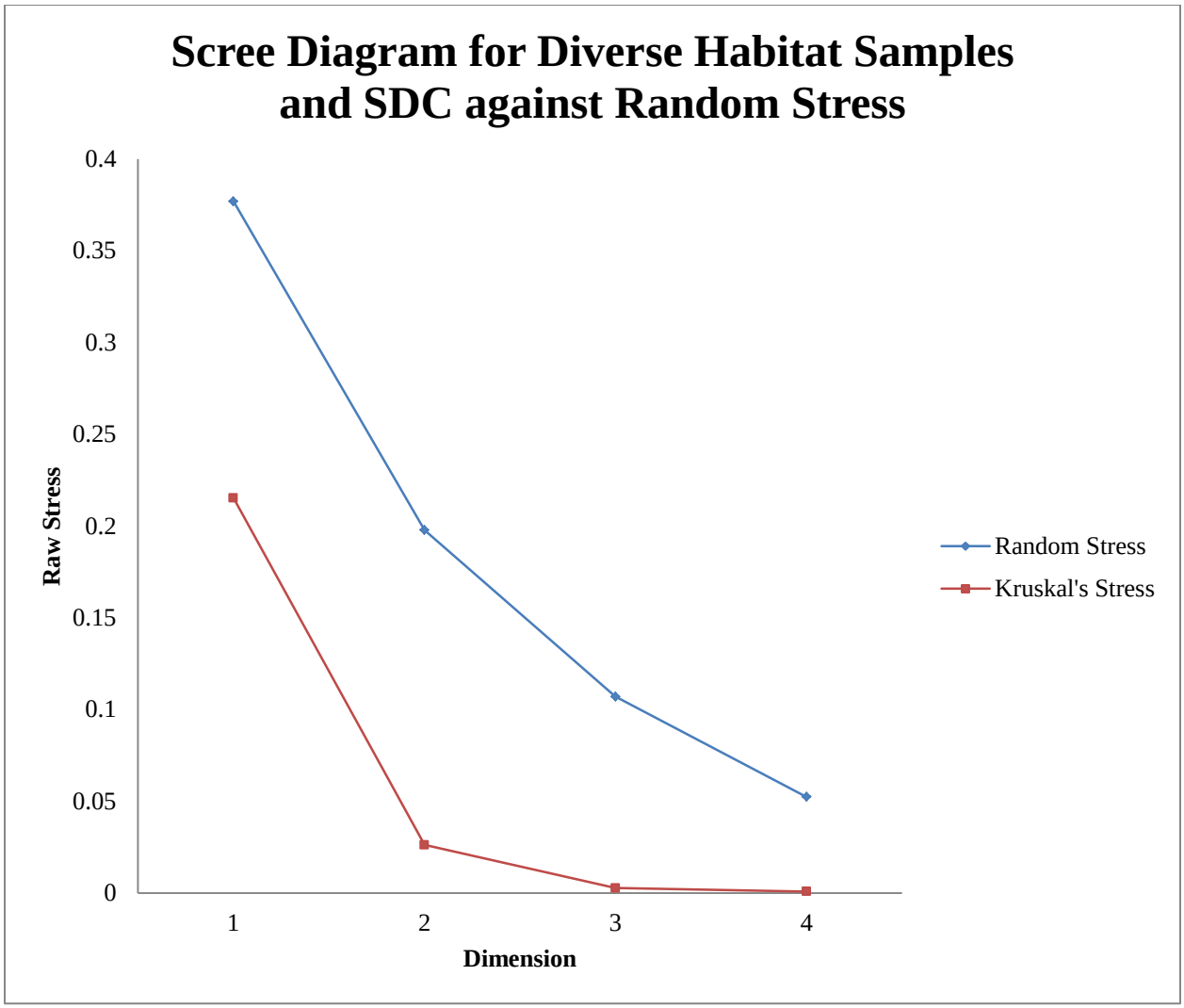

Figure 137. Scree diagram developed by NMDS for the final configurations of diverse habitat soil samples and SDC over four dimensions. Random stress was used as a threshold for the acceptance of final configurations. No configuration exceeded the threshold though the stress for the one dimension plot was much higher than the others. The elbow in the curve is at two dimensions.

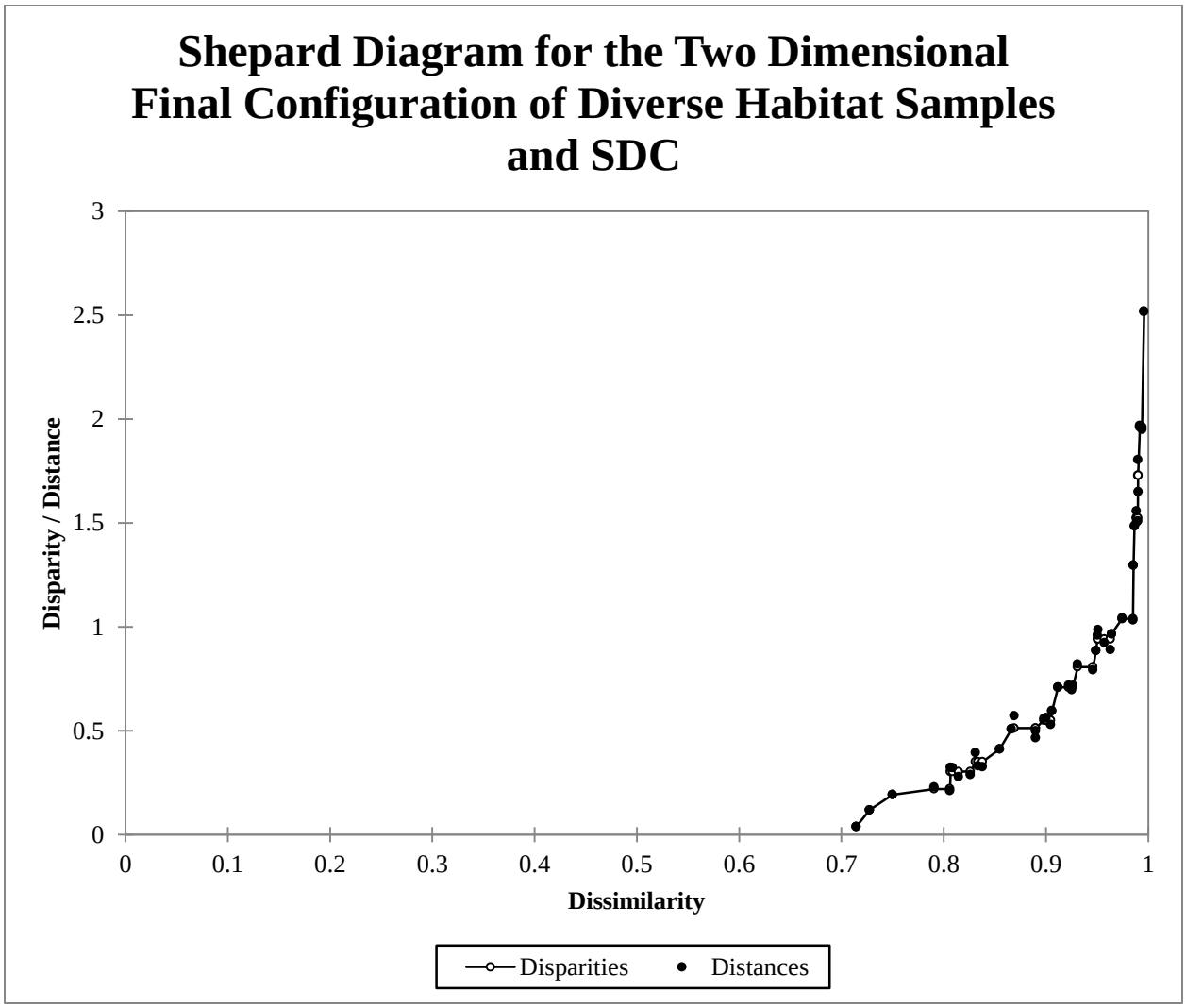

Figure 138. Shepard diagram for the two dimensional final configuration developed from SDC of diverse habitat soil samples. All distances fall close to their corresponding disparities indicating good correlation of the two in the final configuration.

APPENDIX 18. DENDROGRAMS OF DIVERSE HABITAT SAMPLES FOR BCDI AND SDC.

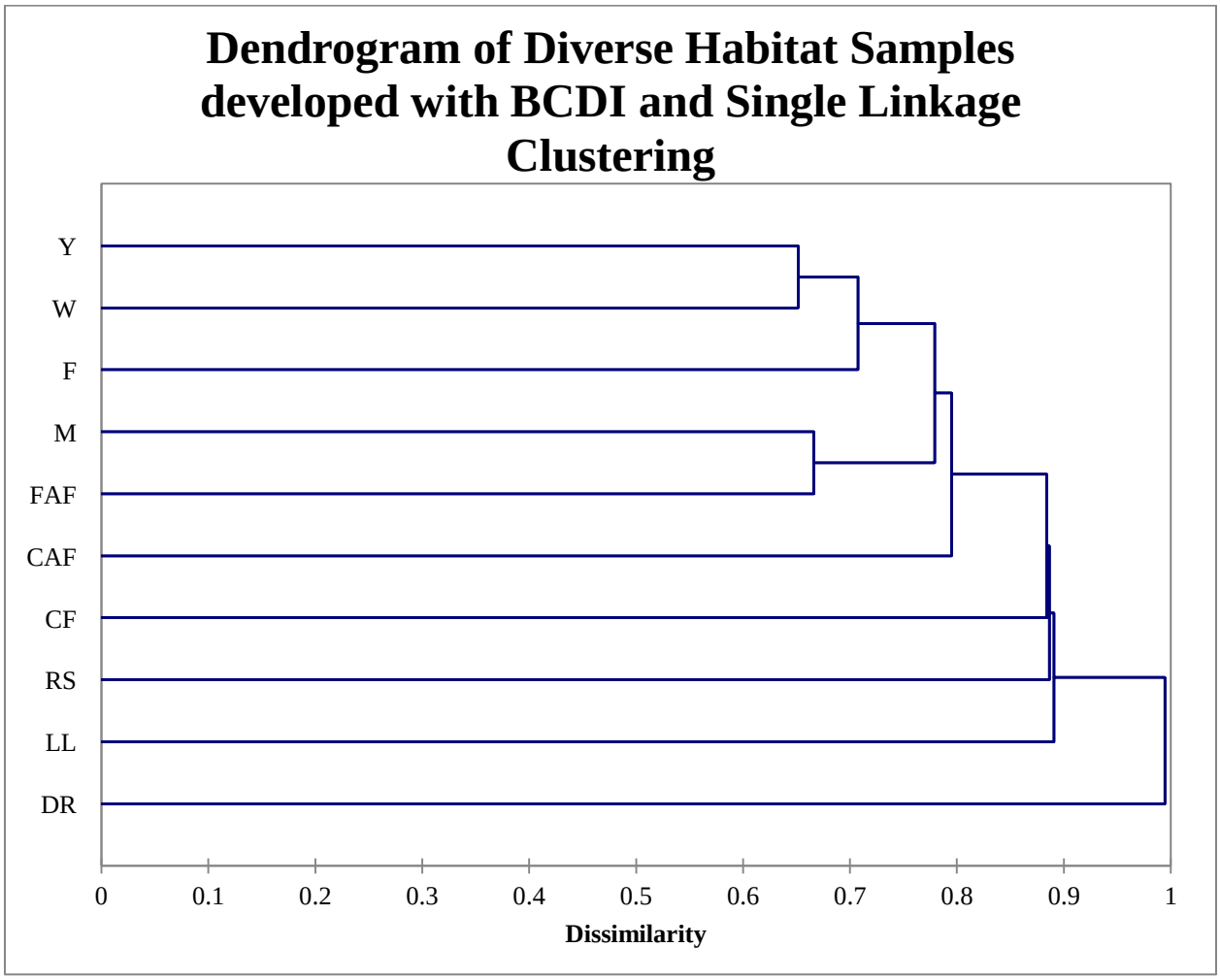

Figure 139. Dendrogram of diverse habitat soil samples developed with BCDI and single linkage clustering. Two clusters are evident; the first has the yard and deciduous woods samples being the most similar followed by the field grouping at 0.708. The second cluster has the marsh edge and fallow agricultural field clustering at 0.666. The remaining samples were a dissimilarity of 0.795 or greater from the two clusters with the dirt road being the most dissimilar from all other samples. See Table 6 for site names corresponding to abbreviations.

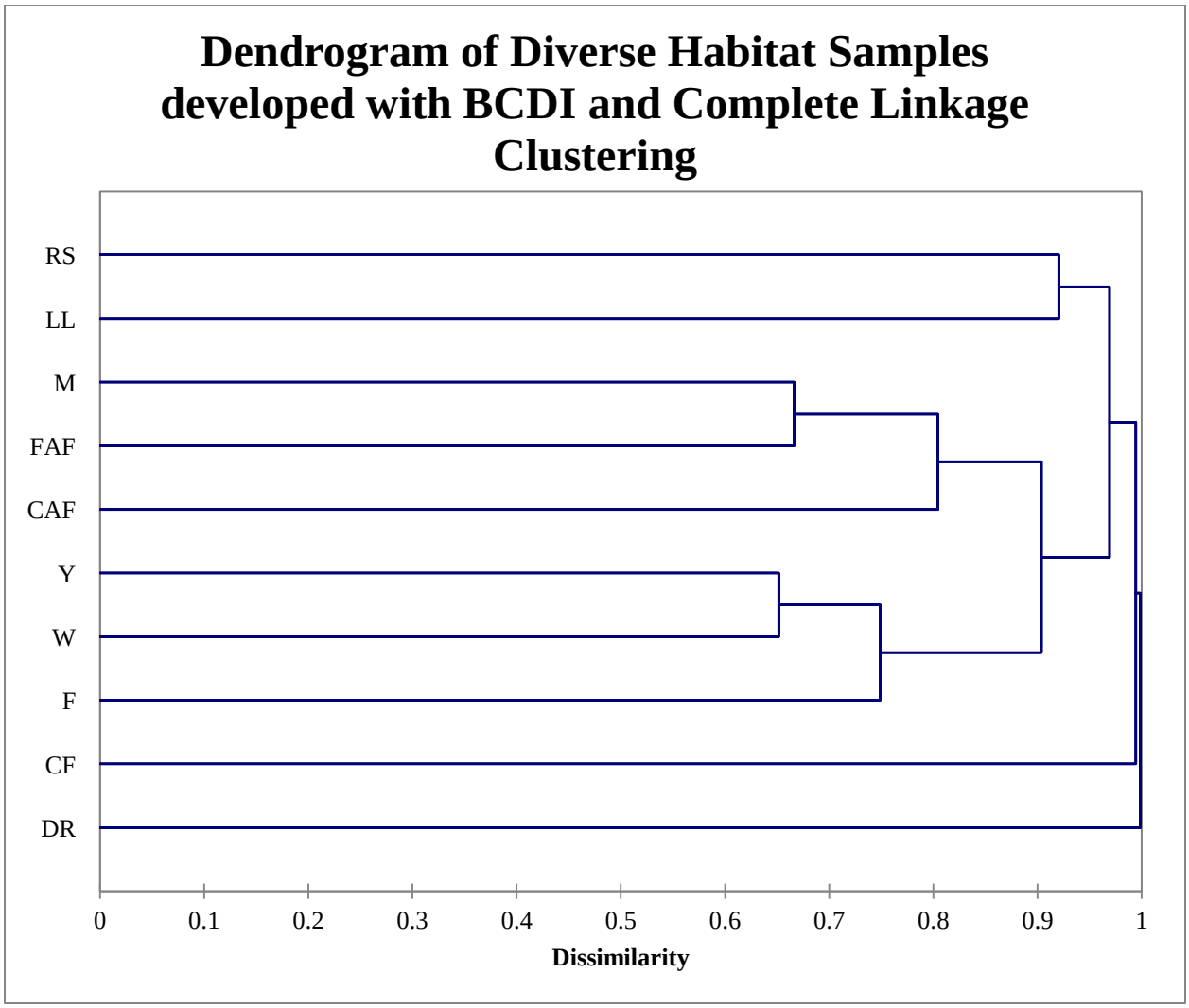

Figure 140. Dendrogram of diverse habitat soil samples developed with BCDI and complete linkage clustering. Three clusters are present. The first has the yard and deciduous woods samples being most similar followed by the field grouping at 0.749. The second has the marsh edge and fallow agricultural field clustering followed by the corn agricultural field at 0.804. The third cluster contains the Lake Lansing beach and roadside samples joining at 0.920. The coniferous forest and dirt road samples are the most dissimilar from the rest. See Table 6 for site names corresponding to abbreviations.

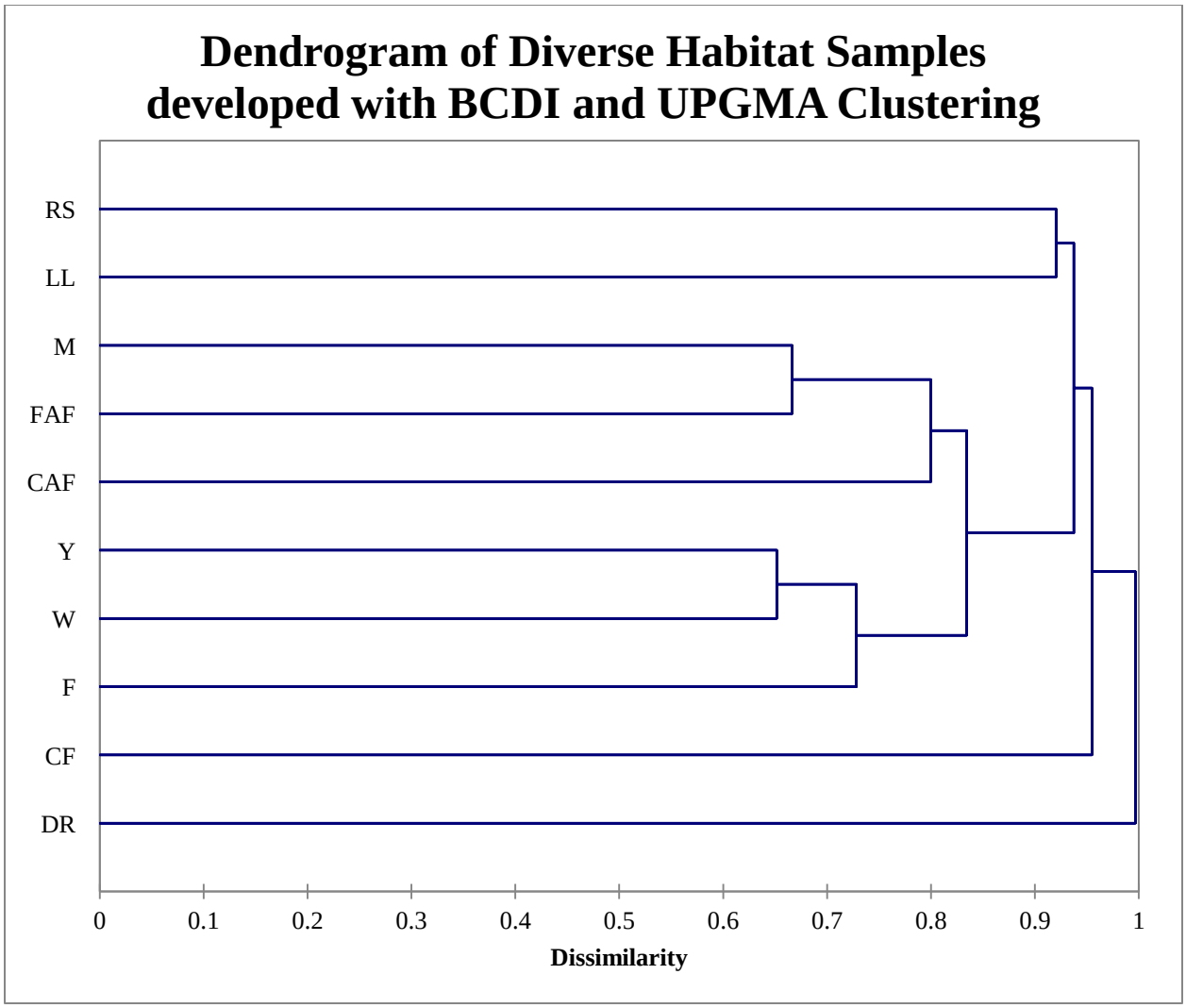

Figure 141. Dendrogram of diverse habitat soil samples developed with BCDI and UPGMA clustering. Three clusters are present. The first has the yard and deciduous woods samples being most similar followed by the field joining at 0.728. The second has the marsh edge and fallow agricultural field clustering followed by the corn agricultural field at 0.800. The third cluster contains the Lake Lansing beach and roadside samples grouping at 0.923. The coniferous forest and dirt road samples are the most dissimilar from the rest. See Table 6 for site names corresponding to abbreviations.

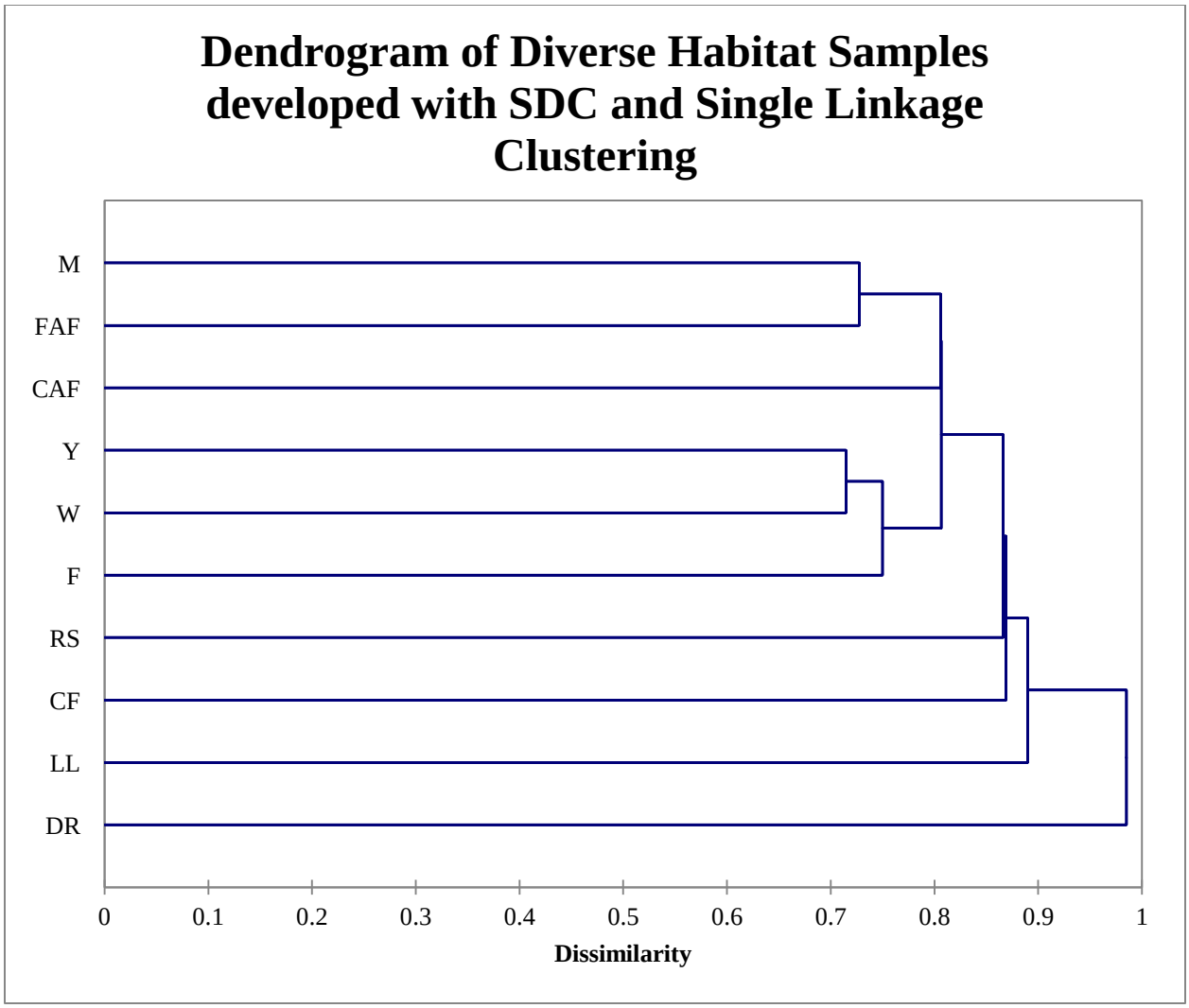

Figure 142. Dendrogram of diverse habitat soil samples developed with SDC and single linkage clustering. Two clusters are evident; the first has the yard and deciduous woods samples being the most similar followed by the field at 0.750. The second cluster has the marsh edge and fallow agricultural field clustering at 0.728. The remaining samples are a distance of 0.807 or greater from the two clusters with the dirt road being the most dissimilar from all other samples. See Table 6 for site names corresponding to abbreviations.

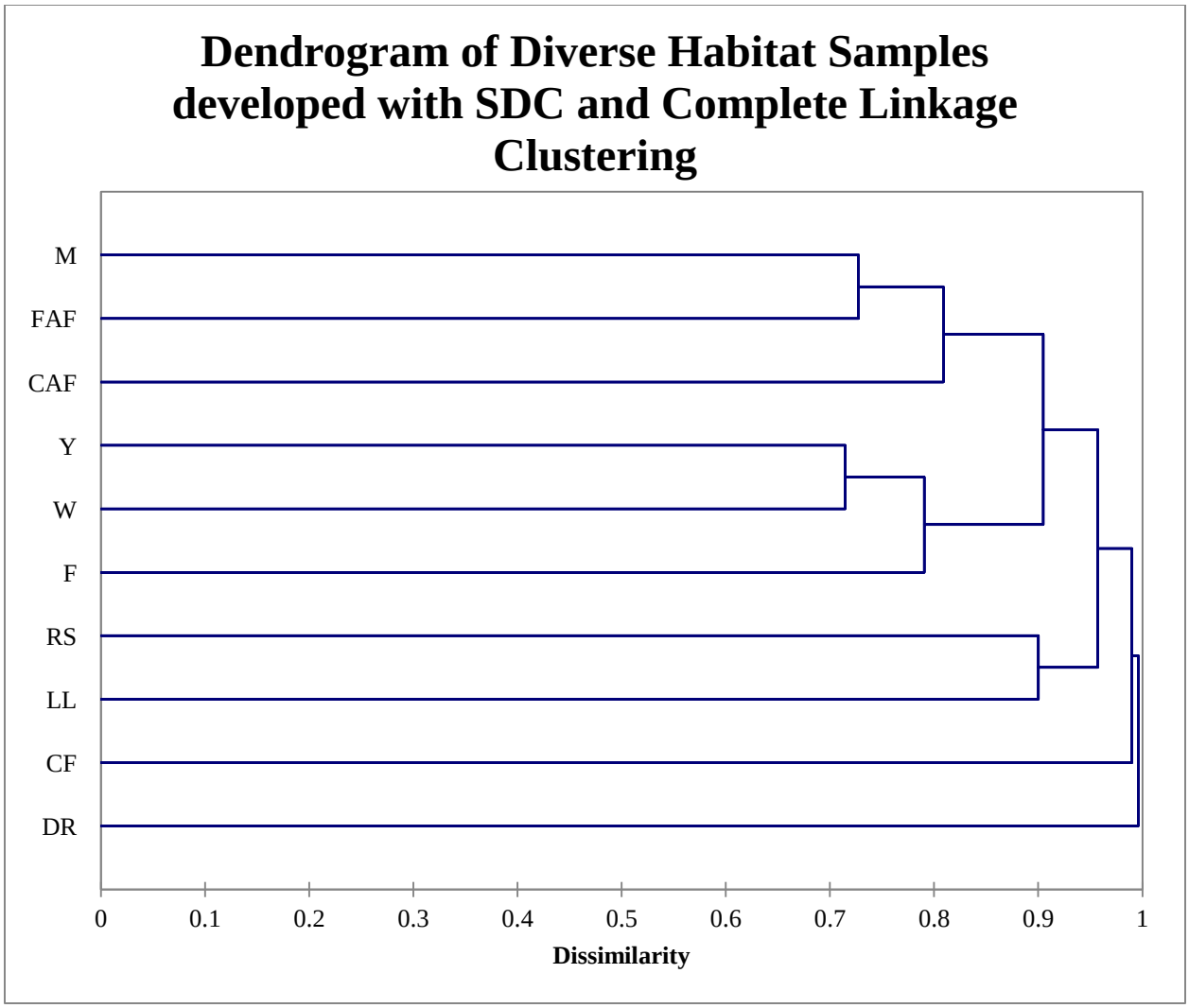

Figure 143. Dendrogram of diverse habitat soil samples developed with SDC and complete linkage clustering. Three clusters are present. The first has the yard and deciduous woods samples being most similar followed by the field grouping at 0.791. The second has the marsh edge and fallow agricultural field clustering followed by the corn agricultural field at 0.809. The third cluster contains the Lake Lansing beach and roadside samples joining at 0.900. The coniferous forest and dirt road samples are the most dissimilar from the rest. See Table 6 for site names corresponding to abbreviations.

REFERENCES

REFERENCES

Alden A. Sherlock Holmes, Forensic Geologist. http://geology.about.com/od/bookreviews/a/sherlock.htm

Amoozegar MA, Hamedi J, Dadashipour M, Shariatpanahi S. Effects of salinity on the tolerance to toxic metals and oxyanions in native moderately halophilic spore-forming bacilli. World Journal of Microbiology and Biotechnology 2005;21(6-7):1237-43.

Armougom F, Raoult D. Exploring microbial diversity using 16s rRNA high-throughput methods. Journal of Computer Science & Systems Biology 2009;2(1):74-92.

Beebe KR, Pell RJ, Seasholtz MB. Chemometrics a practical guide. New York: John Wiley & Sons, Inc., 1998.

Bergslien E. An Introduction to Forensic Geoscience. New Jersey: John Wiley & Sons, Ltd., 2012.

Bittman S, Forge TA, Kowalenko CG. Responses of the bacterial and fungal biomass in grassland soil to multi-year applications of dairy manure slurry and fertilizer. Soil Biology and Biochemistry 2005;37(4):613-23.

Bray JR, Curtis JT. An ordination of the upland forest communities of southern Wisconsin. Ecological Monographs 1957;27(4):325-349.

Bressan D. C.S.I Geology: Forensic geosciences and its application to criminal investigations. <http://historyofgeology.fieldofscience.com/2010/07/csi-geology-forensic-geosciences and.html>. Accessed 2014 March 3.

Borg I, Groenen PJF, Mair P. Applied multidimensional scaling. Berlin: Springer, 2013.

Borg, I.; Groenen, P. J. F. Modern multidimensional scaling: theory and applications. New York: Springer, 2005.

Brons JK, van Elsas JD. Analysis of bacterial communities in soil by use of denaturing gradient gel electrophoresis and clone libraries as influenced by different reverse primers. Applied and Environmental Microbiology 2008;74(9):2717-27.

Buckley DH, Huangyutitham V, Nelson TA, Rumberger A, Thies JE. Diversity of *Planctomycetes* in soil in relation to soil history and environmental heterogeneity. Applied and Environmental Microbiology 2006;72(7):4522-31.

Carson JK, Gonzalez-Quiñones V, Murphy DV, Hinz C, Shaw JA, Gleeson DB. Low pore connectivity increases bacterial diversity in soil. Applied and Environmental Microbiology 2010;76(12):3936-42.

Chau JF, Bagtzoglou AC, Willig MR. The effects of soil texture on richness and diversity of bacterial communities. Environmental Forensics 2011;12:333-41.

Coyle HM, editor. Nonhuman DNA Typing. Boca Raton: Taylor & Francis Group, 2008.

Cox, T. Multidimensional Scaling. Boca Raton: Chapman and Hall, 2001.

Cui H, Yang K, Pagaling E, Yan T. Spatial and temporal variation in Enterococcal abundance and its relationship to the microbial community in Havaii beach sand and water. Applied and Environmental Microbiology 2013;79(12):3601-9.

Daniel R. The Metagenomics of Soil. Nature Reviews Microbiology 2005;3:470-8.

Daubert v. Merrell Dow Pharmaceuticals (92-102), 509 U.S. 579 (1993).

Daugherty G. Pattern recognition and classification. New York: Springer Science + Business Media, 2013.

Dice, L. Measures of the amount of ecologic association between species. Ecology 1945;26(3):297-302.

Engelen B, Meinken K, von Wintzingerode F, Heuer H, Malkomes HP, Backhaus H. Monitoring impact of a pesticide treatment on bacterial coil communities by metabolic and genetic fingerprinting in addition to conventional testing procedures. Applied and Environmental Microbiology 1998;64(8):2814-21.

Everitt BS, Landau S, Lesse M, Stahl D. Cluster Analysis, 5th edition. Published Online: John Wiley & Sons, Ltd, 2011.

Fierer N, Bradford MA, Jackson RB. Toward and ecological classification of soil bacteria. Ecology 2007;88:1354-64.

Fierer N, Jackson RB. The diversity and biogeography of soil bacterial communities. Proceedings of the National Academy of Sciences 2006;103(3):626-31.

Filkins LM, Hampton TH, Gifford AH, Gross MJ, Gross MJ, Hogan DA *et al.* Prevalence of Streptococci and increased polymicrobial diversity associated with cystic fibrosis patient stability. Journal of Bacteriology 2012;194(17):4709-17.

Garbeva P, van Veen JA, van Elsas JD. Microbial diversity in soil: selection of microbial populations by play and soil type and implications for disease suppressiveness. Annual Review of Phytopathology 2004;42:243-70.

Ge Y, He J, Zhu Y, Zhang J, Xu Z, Zhang L, *et al.* Differences in soil bacterial diversity: driven by contemporary disturbances or historical contingencies? ISME 2008;2:254-64.

Gemperline P. Practical guide to chemometrics. Boca Raton: Taylor & Francis Group, 2006.

Griffiths RI, Whiteley AS, O'Donnell AG, Bailey MJ> Influence of depth and sampling time on bacterial community structure in an upland grassland soil. FEMS Microbiology Ecology 2003;43:35-3.

Griffiths RI, Thomson BC, James P, Bell T, Bailey M, Whiteley AS. The bacterial biogeography of British soils. Environmental Microbiology 2011;13(6):1642-54.

Grundmann GL, Debouzie D. Geostatistical analysis of the distribution of NH_4^+ and NO_2^- - oxidizing bacteria and serotypes at the millimeter scale along a soil transect. FEMS Microbiology Ecology 2000;34(1):57-62.

Heath LE, Saunders VA. Assessing the Potential of Bacterial DNA Profiling for Forensic Soil Comparisons. Journal of Forensic Sciences 2006;51(5):1062-8.

Hollister EB, Engledow AS, Hammett AJM, Provin TL, Wilkinson HH, Gentry TJ. Shifts in microbial community structure along an ecological gradient of hypersaline soils and sediments. ISME Microbial Ecology 2010;4:829-38.

Horner-Devine MC, Lage M, Hughes JB, Bohannan BJM. A taxa-area relationship for bacteria. Nature 2004;432:750-3.

Horswall J, Cordiner SJ, Maas EW, Martin TM, Sutherland KBW, Speir TW *et al.* Forensic comparison of soils by bacterial community DNA profiling. Journal of Forensic Sciences 2002;47(2):350-3.

Infometrix, Inc. Multivariate Data Analysis [Pirouette user guide]. Version 4.0.

Janssen PH. Identifying the dominant soil bacterial taxa in libraries of 16S rRNA and 16S rRNA genes. Applied and Environmental Microbiology 2006;72(3):1719-28.

Kaltenbach HM. A concise guide to statistics. New York: Springer, 2012.

Kruskal, J. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika 1964;29(1):1-27.

Kuske CR, Ticknor LO, Miller ME, Dunbar JM, Davis JA, Barns SM, *et al.* Comparison of soil bacterial communities in rhizospheres of three plant species and the interspaces in arid grassland. Applied and Environmental Microbiology 2002;68(4):1854-63.

Lauber CL, Hamady M, Knight R, Fierer N. Pyrosequencing-based assessment of soil pH as a predictor of soil bacterial community structure at the continental scale. Applied and Environmental Microbiology 2009;75(15):5111-20.

Legendre P, Legendre L. Numerical ecology. Kidlington: Elsevier Ltd, 2012.

Lerner A, Shor Y, Vinokurov A, Okon Y, Jurkevitch E. Can denaturing gradient electrophoresis (DGGE) analysis of amplified 16s rDNA of soil bacterial populations be used in forensic investigations? Soil Biology & Biochemsitry 2006;38(6):1188-92.

Lenz EJ, Foran DR. Bacterial Profiling of Soil Using Genus-Specific Markers and Multidimensional Scaling. Journal of Forensic Sciences 2010;55(6):1437-42.

Lipson DA, Schmidt SK. Seasonal changes in alpine soil bacterial community in the Colorado Rocky Mountains. Applied and Environmental Microbiology 2004;70(5):2867-79.

Liu WT, Marsh TL, Cheng H, Forney LJ. Characterization of microbial diversity by determining terminal restriction fragment length polymorphisms of genes encoding 16S rRNA. Applied and Environmental Microbiology 1997;63(11):4516-22.

Macdonald CA, Ang R, Cordiner SJ, Horswell J. Discrimination of Soils at Regional and Local Levels Using Bacterial and Fungal T-RFLP Profiling. Journal of Forensic Sciences 2011;56(1):61-9.

MacLean D, Jones JDG, Studholme DJ. Application of 'next-generation' sequencing technologies to microbial genetics. Microbiology 2009;7:287-96.

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA *et al.* Genome sequencing in microfabricated high-density picolitre reactors. Nature 2005;437(15):376-80.

Martiny JBH, Bohannan BJM, Brown JH, Colwell RK, Fuhrman JA, Green JL, *et al.* Microbial biogeography: putting microorganisms on the map. Nature Review Microbiology 2006;4:102-112.

Metzker ML. Sequencing technologies — the next generation. Nature Reviews January 2010;11(1):31-46.

Meyers MS, Foran DR. Spatial and temporal influences on bacterial profiling of forensic soil samples. Journal of Forensic Sciences 2008;53(3):652-9.

Moreno LI, Mills DK, Entry J, Sautter RT, Mathee K. Microbial metagenome profiling using amplicon length heterogeneity-polymerase chain reaction proves more effective than elemental analysis in discriminating soil specimens. Journal of Forensic Sciences 2006;51(6):1-8.

Muyzer G, Smalla K. Application of denaturing gradient gel electrophoresis (DGGE) and temperature gradient gel electrophoresis (TGGE) in microbial ecology. Antonie van Leeuwenhoek 1998;73(1):127-41.

Nakatsu CH, Torsvik V, Øvreäs L. Soil community analysis using DGGE of 16S rRNA polymerase chain reaction products. Soil Science Society of America Journal 2000;64:1382-8.

Niemi RM, Heiskanen I, Wallenius K, Lindström. Extraction and purification of DNA in rhizosphere soil samples for PCR-DGGE analysis of bacterial consortia. Journal of Microbial Methods 2001;45(3):155-65.

National Research Council. Strengthening Forensic Science in the United States: A Path Forward. Washington DC: The National Academies Press, 2009.

Pye K. Geological and Soil Evidence. Boca Raton: Taylor & Francis Group, 2007.

Quesada E, Ventosa A, Rodriguez-Valera F, Megias L, Ramos-Cormenzana A. Numerical taxonomy of moderately halophilic gram-negative bacteria from hypersaline soils. Journal of General Microbiology 1983;129:2649-57.

Ragab M. Towards the rational use of high salinity tolerant plants. Tasks for Vegetation Science 1993;27:467-72.

Ramette A, Tiedje JM. Biogeography: an emerging cornerstone for understanding prokaryotic diversity, ecology, and evolution. Microbial Ecology 2005;53:197-207.

Roesch LFW, Fulthorpe RR, Riva A, Casella F, Hadwin AKM, Kent AD, *et al.* Pyrosequencing enumerates and contrasts soil microbial diversity. ISME Journal 2007;1:283-90.

Saferstein R, editor. Forensic Science Handbook. Volume 1, 2nd edition. Upper Saddle River: Prentice Hall, 2002.

Sarathchandra SU, Ghani A, Yeates GW, Burch F, Cox NR. Effect of nitrogen and phosphate fertilisers on microbial and nematode diversity in pasture soils. Soil Biology and Biochemistry 2001;33(7 – 8):953-64.

Schloss PD. Evaluating different approaches that test whether microbial communities have the same structure. ISME Journal 2008;2:265-75.

Schloss PD, Westcott SI, Ryabin T, Hall JR, Hartmann M, Hollister EB *et al.* Introducing mothur: Open-source, platfrom-independent, community-supported software for describing and comparing microbial communities. Applied and Environmental Microbiology 2009;75(23):7537-41.

Schütte UM, Abdo Z, Bent SJ, Shyu C, Williams CJ, Pierson JD, *et al.* Advances in the use of terminal restriction fragment length polymorphism (T-RFLP) analysis of 16S rRNA genes to characterize microbial communities. Applied Microbial Biotechnology 2008;80(3):365-80.

Shokralla S, Spall JL, Gibson JF, Hajibabaei M. Next-generation sequencing technologies for environmental DNA research. Molecular Ecology 2012;21(8):1794-805.

Singleton DR, Furlong MA, Rathbun SL, Whitman WB. Quantitative comparisons of 16S rRNA gene sequence libraries from environmental samples. Applied and Environmental Microbiology 2001;67(9):4372-6.

Smit E, Leeflang P, Gommans S, van den Broek J, van Mil S, Wernars K. Diversity and seasonal fluctuations of the dominant members of the bacterial soil community in a wheat field as determined by cultivation and molecular methods. Applied and Environmental Microbiology 2001;67(5):2284-91.

Spence, I. A simple approximation for random rankings stress values. Multivariate Behavioral Research 1979;14:355-365.

Sørensen, T. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. Kongelige Danske Videnskabernes Selskab 1948;5(4):1–34.

Strickland MS, Lauber C, Fierer N, Bradford MA. Testing the functional significance of microbial community composition. Ecology 2009;90(2):441-51.

Sullivan TS, McBride MB, Thies JE. Soil bacterial and archaeal community composition reflects high spatial heterogeneity of pH, bioavailable Zn, and Cu in a metalliferous peat soil. Soil Biology & Biochemistry 2013;66:102-9.

Whittaker, R.. Vegetation of the Siskiyou mountains, Oregon and California. Ecological Monographs 1960;30(3):279-338.

Yang C, Mills D, Mathee K, Wang Y, Jayachandran K, Sikaroodi M *et al.* An ecoinformatics tool for microbial community studies: Supervised classification of amplicon length heterogeneity (ALH) profiles of 16S rRNA. Journal of Microbiological Methods 2006;65(1):49-62.