ADOPTION OF NEXT-GENERATION 16S BACTERIAL SEQUENCING PRACTICES FOR THE FORENSIC ANALYSIS OF SOIL

By

Ellen Marie Jesmok

A THESIS

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

Forensic Science—Master of Science

2015

ABSTRACT

ADOPTION OF NEXT-GENERATION 16S BACTERIAL SEQUENCNG PRACTICES FOR THE FORENSIC ANALYSIS OF SOIL

By

Ellen Marie Jesmok

Soil has the potential to link items or individuals to a crime scene; however, the nature of traditional forensic soil analysis does not typically allow for its individualization. Rather than develop an entirely new methodology, it is more logical to adopt next-generation sequencing of the bacterial 16S ribosomal RNA gene, a well-established, scientifically validated technique in microbiology, for the forensic analysis of soil. Next-generation sequencing was employed to generate bacterial profiles from soils collected in ten diverse and nine similar habitats, across time and space within three habitats, and from various evidentiary items. Bacterial abundance charts, nonmetric multidimensional scaling, and a supervised classification technique were used to analyze profiles. Soils from diverse and similar habitats were largely differentiated from one another, with bacterial profiles separating in multidimensional space and/or being correctly assigned to their location of origin. Time and space within a habitat affected bacterial profile similarity; however, not enough to prevent the traceability of soils. Bacterial profiles from evidentiary items were correctly classified to their location of origin over a full year of storage, regardless of evidentiary type or storage temperature. These studies highlight the utility of next-generation sequencing and a combination of robust bacterial profile statistical methods for forensic soil analysis while also emphasizing the adoption of established techniques from other scientific disciplines to strengthen forensic science as a whole.

ACKNOWLEDGMENTS

I would like to thank Dr. David Foran for his guidance not only as an advisor, but also as a co-author and co-presenter. Together we disseminated this thesis research in the hopes of greatly improving forensic soil analysis, and I could not have done so without his expertise and understanding. I would also like to thank my other thesis committee members, Dr. Eric Benbow and Dr. Steven Dow, as well as the many individuals who had a part in guiding my research, including: James "Mac" Hopkins who introduced me to the soil sampling and data analysis methods, Dr. Jeff Landgraf who taught me the many steps involved in MiSeq sample preparation, and Dr. Ruth Smith who provided knowledge and resources allowing me to apply effective statistical methods. I am also grateful to the National Institute of Justice, which funded portions of this work, as well as allowing Dr. Foran and me to present at their annual research and development meeting. Finally, I would like to thank the other members of the Michigan State Forensic Biology program (Timothy Antinick, Rebecca Ray, Kaitlyn Germain, and my fellow soil researcher, Alyssa Badgley) as well as my friends and family, especially my parents, John Jesmok and Annette Samson, without whom this thesis would not have been possible.

PREFACE

This research was funded in part by The National Institute of Justice Grant Number 2013-R2-CX-K010, awarded to Dr. David Foran, and portions of this document are included in the published grant report. Material in this thesis is also included in a Journal of Forensic Sciences manuscript entitled “Next-Generation Sequencing of the Bacterial 16S rRNA Gene for Forensic Soil Comparison: A Feasibility Study” (Jesmok et al., accepted for publication in July 2016).

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

Soil can become evidence in a criminal investigation when a crime occurs in an outdoor location or it is collected from clothing, shoes, tires, or other items connected with a crime. The association of such evidentiary soil samples with a location of origin has the potential to link a suspect or victim to the scene (Dawson and Hillier, 2010). Soil has been involved in criminal investigations as far back as the 1800s, when a visual comparison of sand was used to link a barrel that had once been filled with silver to a specific train station on the Prussian railroad, resulting in a conviction (Scientific American, 1856). Traditional forensic techniques for soil analysis, similar to those used in the silver theft case, involve the examination of class characteristics such as grain size, color, pH, and moisture content (Murray and Solebello, 2002; Saferstein, 2002; Ruffell, 2010). Soil containing very rare attributes may be highly probative; however, the comparison of soils using these techniques is time consuming and often does not result in definitive associations. Additionally, many such examinations are visual comparisons, with similarity being subjectively judged. Given this, the need for a more objective, robust forensic soil analysis method that allows for stronger sample association and potentially location of origin assignment is evident.

Ideal Elements of a Forensic Evidence Association Technique

A National Academy of Sciences (NAS) committee published a report in 2009 assessing the strength of current forensic methods and technologies (National Research Council). The main message was one of great concern; many forensic disciplines were criticized for their inability to statistically link or individualize evidentiary items in an objective manner. Pattern evidence was regarded as especially weak in the report, as associations between evidentiary items were based

largely on human interpretation. The lack of established, objective scientific methods is contrary to the 1993 *Daubert v. Merrell Dow Pharmaceuticals, Inc.* ruling (509 U.S. 579), in which an expert's "principles and methodology" must be based upon "scientific validity" to be acceptable in court.

Forensic methodologies should possess several elements in addition to objectivity, many of which were mentioned in the NAS report. One of these attributes, important to all scientific methods, is reproducibility. Multiple scientists may analyze the same evidence in a criminal investigation and their conclusions might differ if a technique is not reproducible, lowering its probative value and chances of gaining court acceptance. Associated with reproducibility is the availability and standardization of methodologies. This includes access to and acceptance of analysis and statistical procedures. Not all crime laboratories have the resources to purchase expensive analytical equipment. Additionally, a new analysis process has to be validated, technicians have to be trained, and in some cases databases need to be created before any data for criminal investigations are produced. Again, different experts may analyze the same piece of evidence, and all of these processes must be standardized to ensure the same conclusions are reached.

Another important quality of a scientific analysis technique specific to forensic science is layperson interpretability. Following data production and analysis, a forensic scientist might be called to testify on their results. This presents a challenge if the analysis process is overly complicated, as a judge or jury may not be able to easily understand how association or discrimination between samples was achieved, nor the strength of a scientist's conclusions. The ability to display data in a way that intuitively demonstrates how a conclusion was reached or

what the results mean for the particular case can prove advantageous for the non-scientist's comprehension in a court of law.

Current forensic methods to analyze soil possess some of the above qualities, but not all. A few traditional techniques are objective (e.g., pH measurements and elemental analysis); however, these measurements only provide class characteristics and can be time consuming, limiting their overall value. Such techniques are easily interpreted by laypersons, but definitive association between two soil samples is unlikely. There are three possible avenues that could be taken to improve forensic soil analysis: expand upon the traditional techniques, develop an entirely new methodology, or adopt already established techniques from another scientific discipline. The latter represents an interesting option because it could allow for a departure from the largely subjective traditional methods for the forensic examination of soil while avoiding the long process of developing a novel methodology. Additionally, there are several disciplines that currently analyze different aspects of soils, offering multiple options for adoption into forensic science. Experts in the fields of microbiology, agricultural biology, ecology, and geology have different soil analysis goals than forensic science; however, their methodologies are well-studied and might allow for the development of stronger associations between known and questioned soil samples than traditional forensic analysis. Microbiologists in particular have established techniques for the comparison of soils, and the adoption of their practices offers a possible approach to satisfy the recommendations of the NAS report, making soil evidence more valuable.

Microbiological Techniques for Soil Analysis

Microbiologists characterize soil types by analyzing the living organisms within, including plants, small eukaryotes, fungi, and bacteria. The latter is the most numerous in soil; one gram contains between 4×10^7 and 2×10^9 bacteria, which vary widely in diversity and species abundance (Daniel, 2005), providing massive amounts of information from small amounts of soil. Woese and Fox (1977) pioneered the use of 16S ribosomal RNA (rRNA) gene analysis for constructing bacterial phylogenies. This marker is conserved across bacteria and archaea but contains variable regions (Figure 1) that can be used to identify bacteria down to species. Two decades later, Liu et al. (1997) described terminal restriction fragment length polymorphism (T-RFLP) analysis, which can be employed to generate a snapshot of the microbial community within a sample (a microbial profile) via DNA amplification, restriction enzyme digestion, and electrophoresis. The resultant electropherograms or autoradiographs are compared to estimate microbial similarities among samples. Since then, T-RFLP analysis of the 16S rRNA gene has been widely used in both the microbiological (e.g., Fierer and Jackson, 2006) and forensic fields (e.g., Horswell et al., 2002; Meyers and Foran, 2008; Lenz and Foran, 2010) to generate soil bacterial profiles. Unfortunately, the massive number of bacterial species in soil and the limited resolving power of T-RFLP makes its forensic utility limited. Several different types of bacteria may share the same restriction site, resulting in an underrepresentation of diversity in a given profile. Additionally, background noise due to drop-in, which can occur when DNA is only partially digested, poses difficulties for microbial profile comparison (Egert and Friedrich, 2003). Given these problems, it is apparent that forensic analysis of soils based on their bacterial makeup requires the production of more complete and higher resolution data.

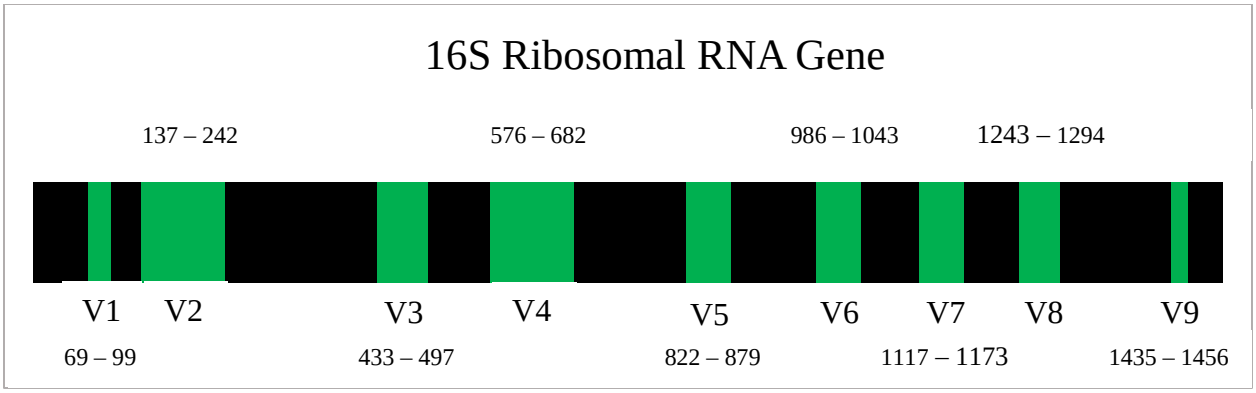

Figure 1—The 1500 base pair 16S rRNA gene showing conserved regions (black segments) and variable regions (green segments). Base pair ranges (Yarza et al., 2014) are shown above or below the respective variable region.

Next-generation sequencing of the 16S rRNA gene, first described by Jonasson et al. (2002), represents an extremely robust methodology for bacterial identification that does not suffer from the resolution and reproducibility weaknesses of past techniques. It is now used extensively by microbiologists for bacterial community analysis (e.g., Kravchenko et al., 2014; Luo et al., 2014), and large bacterial sequence reference databases have been created (e.g., Cole et al., 2004; Quast et al., 2013), providing classification of bacteria at taxonomic levels from phylum to species. Classification based on the entire 16S rRNA gene is ideal; however, many next-generation technologies do not allow for full 16S sequencing, and the choice of which variable region(s) to sequence depends on the capacities of the platform used, ranging from 100 to 900 base pairs (reviewed by Liu et al., 2012; Quail et al., 2012). The nine variable regions of the 16S gene confer different amounts of bacterial classification coverage at the sequence and/or taxonomic levels based on their levels of sequence diversity. Studies measuring the taxonomic classification accuracy of each variable region or combinations of regions (e.g., Yu and Morrison, 2004; Chakravorty et al., 2007) have shown V3 to be the most informative, having higher sequence diversity than other regions. Additionally, this region can be easily sequenced

due to its moderate size (Figure 1), making it an excellent target for forensic science. Mizrahi-Man et al. (2013) sequenced V1, V3 – V7, and V9 to determine which provided the highest level of bacterial differentiability, finding regions V3 and V4 conferred the greatest classification coverage, and paired sequencing of both regions was suggested as the most effective method for bacterial profile generation.

Several next-generation sequencing platforms exist, including ion-semiconductor sequencing (Life Technologies, Carlsbad, CA), pyrosequencing (e.g., Roche, San Francisco, CA), or Illumina sequencing by synthesis (San Diego, CA). Loman et al. (2012) compared the performance of these three platforms in sequencing an *E.coli* isolate and found the latter produced the lowest error rate and highest throughput. Illumina sequencing by synthesis has since been used to produce bacterial profiles from several different media (e.g., Bokulich et al., 2012; Maughan et al., 2012; Ward et al., 2013). Caporaso et al. (2012) used an Illumina HiSeq2000 and MiSeq to generate bacterial profiles from 24 different environments including the human gut, skin, and soil, but its forensic value for soil characterization was not examined. The reason behind the sequencing efficiency of Illumina technology may come from its cluster generation and sequencing by synthesis technology. Rather than beginning DNA sequencing immediately after a sample is placed in the machine, the Illumina platform first amplifies each DNA strand, forming clusters of identical sequences. Millions of these DNA clusters are then simultaneously sequenced, producing massive datasets from each sequencing run. The advantages of this method include the ability to sequence more than 95 samples concurrently and produce read lengths of up to 500 base pairs (www.Illumina.com), which is sufficient to sequence the informative V3 and V4 regions of the bacterial 16S rRNA gene.

Past Forensic Soil Research

Microbiologists have already developed robust, well-studied methodologies for soil analysis via generation and interpretation of bacterial profiles. These techniques have the potential to increase the value of forensic soil evidence; however, factors that influence such profiles must first be considered. The main goal of forensic soil analysis is to associate or discriminate samples from two sources: an evidentiary item and a crime scene. While variability among locations (of both diverse and similar habitat type) is necessary to distinguish forensic soils, variability within a habitat may result in the generation of differing bacterial profiles and potentially prevent association of evidentiary soil with its location of origin. Known soil samples will not be collected at the same time or in the exact same place as an evidentiary soil, making knowledge of temporal and spatial bacterial variation within a location important. Additionally, exposure to various items (e.g., tools, clothing, or weapons) and/or time in storage could affect soil bacterial profiles, resulting in false exclusions when soils originated from the same location.

Many profile generation techniques have been utilized to assess bacterial variation in past forensic research, detecting differences across habitats as well as over time and space. T-RFLP analysis is the most prevalent technique for bacterial profiling in the forensic soil literature. Horswell et al. (2002) exposed shoes and clothing to soil at a hypothetical crime scene and used T-RFLP analysis to generate bacterial profiles from soil on the evidentiary items, from the scene, and from three different locations of similar habitat type immediately following exposure and again after eight months. Similarity values were calculated based on shared T-RFLP peak presence or absence. Peaks were defined as over 150 relative fluorescence units (RFUs) and were regarded as identical if they differed by one base or less. They found that soil originating from the same location produced bacterial profiles more similar to one another than profiles from

other locations; however, samples collected after eight months in these same locations produced dissimilar profiles, showing temporal bacterial variation exists. Temporal changes could prevent the forensic scientist from making associations between two soil samples if the knowns are collected long after a crime occurs. Heath and Saunders (2006) generated bacterial T-RFLP profiles from soil collected at five points within three distinct habitats to assess spatial variation. Each sample was divided into three subsamples and DNA was extracted separately to determine the reproducibility of T-RFLP analysis. Again, the shared presence or absence of peaks (over 25 RFUs) were used to calculate similarities between profiles; however, only peaks that differed by less than 0.5 bases were considered shared, and DNA quantity within each T-RFLP profile was standardized based on the smallest total relative fluorescence across replicate subsamples before analysis. Additionally, hierarchical cluster analysis (Ward, 1963) was employed to group profiles based on relative similarity. The five soil samples collected in each location were more similar to one another than soils across habitats, but the reproducibility of T-RFLP bacterial profiles among replicate subsamples was only 67.6%, indicating this method may not be effective in forensic science where results must be reproducible. Meyers and Foran (2008) employed T-RFLP analysis to generate bacterial profiles from five diverse habitats over one year and from soil 10 feet in the cardinal directions within each habitat every three months to assess temporal and spatial variation. T-RFLP profiles were compared by calculating similarity based on normalized profiles, including only shared peaks over 50 RFUs. Their results mirrored those of past studies: the greatest similarity values were between soils collected from the same location, but the bacterial content of the soil varied across time. Spatial factors also lowered profile similarity slightly; however, soils within a habitat were usually more similar than soils across habitats. The reliability of T-RFLP for forensic analysis was also questioned in Meyers and Foran's research,

as the normalization of the data or number of peaks included in similarity calculations can affect associations. Although T-RFLP analysis may lack the resolution and repeatability desired for a forensic technique, these studies demonstrate the potential of using microbial populations within the soil to associate two samples.

The advent of next-generation sequencing techniques and their application in the microbiological world has also influenced forensic soil analysis research. Young et al. (2014) used next-generation sequencing in their assessment of which soil microbes confer the highest distinguishability of soils and therefore have the highest forensic potential. They compared the reproducibility and resolution of multiple molecular markers in distinguishing soils collected from two habitats. Portions of the bacterial 16S rRNA gene, eukaryotic 18S rRNA gene, a chloroplast tRNA gene intron (*trnL)*, and a fungal spacer between two RNA subunits (internal transcribed spacer [ITS]) were amplified and sequenced using an Ion Torrent Personal Genome Machine™ (Life Technologies). All markers allowed for discrimination of the two sites, however the ITS marker exhibited the highest profile reproducibility among both replicates from the same collection and replicates collected 1 meter apart within a habitat. ITS profiles also contained the lowest diversity; however, it and the eukaryotic 18S rRNA gene provided the most accurate sample discrimination. Additionally, analysis of the 16S rRNA gene was recommended for use in the discrimination of locations on a small scale, such as at short distances within habitats, due to the heterogeneous nature of soil bacteria. However, only 150 to 250 base pairs of each marker were sequenced, limiting their potential discriminatory power for soil analysis. Additionally, no assessment of microbial profiles across time and larger spaces was carried out, factors that may influence each group of microorganisms differently.

Soil will most often be recovered from evidentiary items in a forensic scenario, and it is crucial to determine whether such samples will produce microbial profiles that can be traced to a location of origin. Young et al. (2015) sequenced 200 base pairs of the eukaryotic 18S rRNA gene in an investigation of soil evidence traceability. A mock crime scenario was developed in which a body had been left on a roadside. A hypothetical suspect was identified, claiming to have never been at the crime scene; however, a soil covered shovel and pair of shoes were found in the trunk of his car 6 weeks after the body was discovered. Two soil samples were collected from the shoes, three were collected from the shovel, and one was collected from the trunk of the vehicle. Soil was also collected at the roadside where the body was located as well as three samples 5 meters from the center site. Additionally, soil was collected in triplicate from six other locations, three of similar and three of different habitat type. Resulting eukaryotic profiles were ordinated in non-constrained multidimensional scaling plots, and statistically compared through pairwise comparisons and analysis of similarities (Clarke, 1993). Evidentiary soil samples clustered near the crime scene samples in multidimensional space and were the most similar as a group to that site; however, some profile variation was obvious in those samples.

These studies highlight the potential of next-generation sequencing for forensic soil analysis, while also stressing the importance of considering various factors that might affect microbial profile composition in soil (e.g., time, space, and storage). However, these researchers did not measure the full potential of bacterial profiling for forensic soil analysis. Higher levels of distinguishability may be achieved if longer stretches of DNA are sequenced, allowing for differentiation of very similar taxa. Additionally, further studies assessing profile change after storage of soil on evidentiary items for longer periods of time and in various conditions are necessary to understand how bacteria will behave in forensic scenarios.

Statistical comparisons of microbial profiles differed in the above studies, with several requiring subjective data interpretation (e.g., multidimensional scaling) and others providing more objective statistical discrimination among groups of samples (e.g., analysis of similarities). However, no authors specifically mention if any of these statistical methods would be useful for forensic analysis. Hopkins (2014) aimed to identify next-generation sequencing data analysis techniques found in the microbiological literature that have potential to meet the demands of forensic bacterial profile comparison. He generated bacterial profiles from replicate soil samples taken at the surface and at various depths within a habitat, and at a central location within three habitats over one year. Additionally, soil samples were collected 5 to 100 feet from the central location in each of the cardinal directions. Statistical analysis methods used to compare bacterial profiles included: bacterial abundance charts (Whittaker, 1965), hierarchical cluster analysis (Ward, 1963), two types of pairwise comparisons (UniFrac; Lozupone and Knight, 2005 and ∫-LIBSHUFF; Schloss et al., 2004), nonmetric multidimensional scaling (NMDS; Kruskal, 1964), and *k*-Nearest Neighbor (*k*-NN; Cover and Hart, 1967). Three of the analysis techniques stood out as potential forensic options, as they possessed at least one of the traits desired in a scientifically robust forensic analysis method. Ideally, bacterial profile analysis would combine many statistical qualities, allowing for objective association of soil samples while also being easily interpreted by a lay audience. Abundance charts, NMDS, and *k*-NN were employed in the current research for bacterial profile analysis based on their abilities to associate soils collected in the same location and produce easily interpretable data depictions.

Soil Bacterial Profile Analysis Methods

Abundance charts (e.g., Figure 2) are generated from taxonomic data, producing a simple visualization of bacterial groups present in a given soil from most to least abundant. The charts can be created at any taxonomic level; however, if too many groups exist, such as when considering genera or species, the charts are largely uninterpretable; therefore, most microbiologists build abundance charts at the phylum or class level (e.g., Meadow and Zabinski, 2012; Jansson and Tas, 2014). Researchers have used bacterial abundance charts to examine, for instance, the influence of environmental stressors such as repeated wetting and drying (Barnard et al., 2013) or diesel fuel contamination (Sutton et al., 2013) on the bacterial makeup of soil. Such charts have also been used to show the changing bacterial abundance on and within decomposing bodies (Hyde et al., 2013). In court, abundance charts could provide the expert witness with a useful visualization tool for a jury. However, they do not provide a statistical measure of relative similarities among bacterial profiles; therefore, additional analysis methods are necessary.

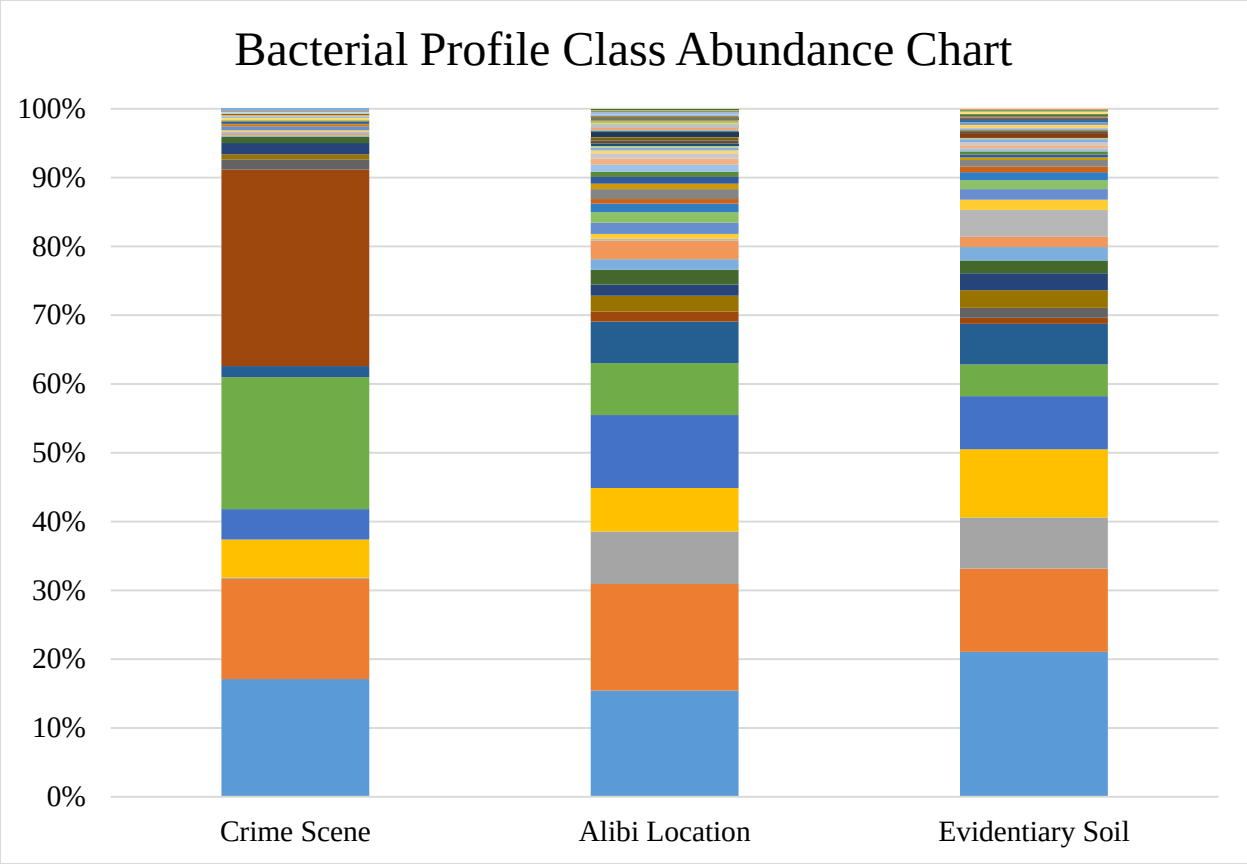

Figure 2—Exemplary abundance chart displaying bacterial profiles generated from soils collected at a hypothetical crime scene and alibi location, and from an evidentiary item. The chart was generated from taxonomic class data and are in ascending order from most overall abundant to least overall abundant class. The evidentiary profile appears more similar to the alibi location than the crime scene. Very different profiles are fairly noticeable in abundance charts; however, similar soil types are often difficult to visually discriminate.

Further analysis of bacterial profiles requires the calculation of dis/similarity values between profiles. More than fifty similarity and/or distance measures exist for the comparison of two samples containing complex data, calculated based on the shared presence/absence of specific groups and/or abundance differences within shared groups (Choi et al., 2010). The most widely used of these indices in microbiological research are the Bray-Curtis index (Bray and Curtis, 1957), the Jaccard index (Jaccard, 1901), and Sørensen-Dice coefficients (Dice, 1945; Sørensen, 1948), each of which are calculated in a different way. Bray-Curtis measures both

which sequences are shared between profiles as well as any difference in shared sequence abundance. Alternatively, Jaccard and Sørensen-Dice ignore abundance differences between profiles, measuring only the presence of shared sequences, with Sørensen-Dice placing twice the weight on such commonalities (see equation below). The choice of which statistic to employ can affect the outcome of downstream results. Daliresefat et al. (2009) compared several indices and concluded that for the analysis of specific DNA regions (such as the 16S rRNA gene), indices that are calculated based only on the shared presence of sequences, such as Jaccard or Sørensen-Dice, are more accurate. Further, with massive datasets like those produced via next-generation sequencing, it is intuitive that some fluctuation in abundance of specific sequences between two soil bacterial profiles are bound to exist, which will affect statistics like Bray-Curtis, making dissimilarity values artificially large. The Sørensen-Dice coefficient was used in the current research due to its wide application in microbiological studies as well as its success in characterizing soil bacterial profiles in past research (reviewed by Hopkins, 2014).

The equation for the Sørensen-Dice dissimilarity coefficient is

$$1 - \frac{2Z}{X+Y}$$

where X is the number of unique sequences in one profile, Y is the number of unique sequences in a second profile, and Z is the number of shared sequences between the two profiles. A dissimilarity matrix is developed from these calculations comparing each sample to the others in a pairwise fashion. This matrix can then be used for a variety of ordination and classification techniques (Chahouki, 2011).

NMDS is an ordination technique that provides a visualization of relative association among multiple samples in a dataset. It can be used in soil analysis for orienting bacterial profiles in multidimensional space (e.g., Figure 3) based on calculated similarity or dissimilarity

values. More similar bacterial profiles plot closer together in space, forming clusters (Ramette, 2007). Increasing the number of dimensions can tease out subtle differences among samples; however, two-dimensional plots are the easiest to interpret. A goodness of fit measure in the form of stress is generated with NMDS plots, providing the user with information on how well a plot is reflecting the inputted dissimilarity values (Holland, 2008). High stress indicates that not all pairwise relationships are accurately represented in the plot, potentially posing a problem for forming conclusions on relative similarity. Stress levels are typically higher when more samples are ordinated together or when a very dissimilar sample is included in the dataset, as such a sample will force others together (Kenkel and Orlóci, 1986). Thus, it may be more beneficial to ordinate a subset of samples from a given dataset to accurately reflect their dissimilarity. NMDS plots, like abundance charts, have a subjective component, as there is no standard mechanism for defining a cluster. Despite this, the data depiction that NMDS plots provide may have value for jury comprehension of soil bacterial profiling. The location of a forensic unknown (evidentiary) sample in an NMDS plot can impart important information on which known profiles it is most similar to, helping to form associations among samples.

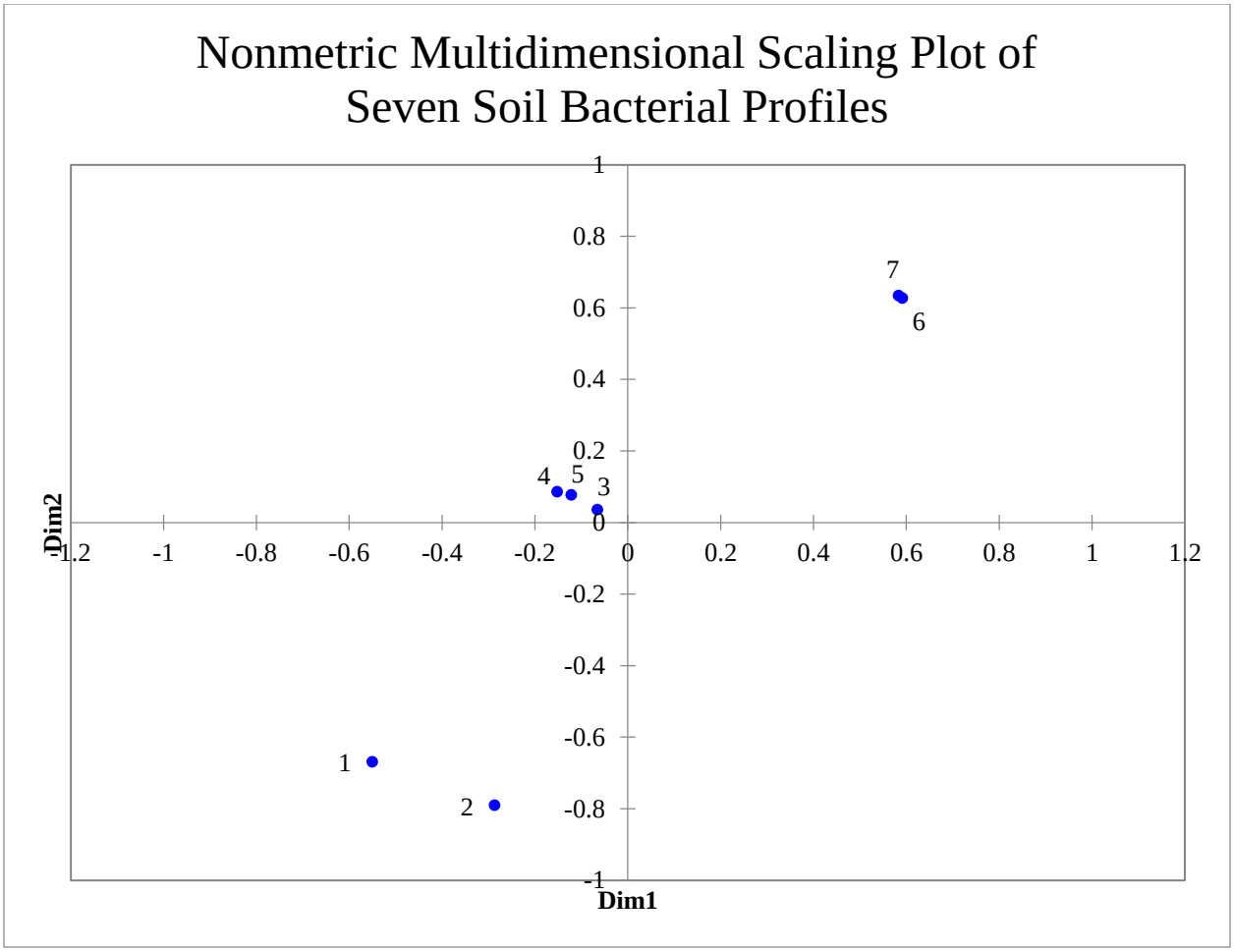

Figure 3—Ordination of hypothetical soil bacterial profiles in multidimensional space. Profiles 1 and 2 come from a hypothetical crime scene, while 4 – 7 represent two alibi locations. The bacterial profile produced from evidentiary soil (3) plots closest to an alibi location, indicating it is more similar to this location than the crime scene.

Supervised classification techniques allow for an objective assignment of bacterial profiles to a location of origin (Mohri, 2012). These techniques build models derived from groups of known samples collectively called training sets. Unknowns are compared to the training sets, either simultaneously or after training set validation, and assigned to the closest group or, depending on the technique, to no group at all. Yang et al. (2006) used supervised classification to assign soil microbial profiles to their location of origin with approximately 90% accuracy, based on length differences in 16S rRNA variable regions 1, 2, 3, and 9. This

methodology does not hold the resolving power of next-generation sequencing, data from which may classify with higher accuracy; however, it does highlight the potential utility of supervised classifiers for bacterial profile analysis.

k-NN is one of the simplest supervised classification techniques and is widely used in microbiology research (e.g., Gaus et al., 2006; Diaz et al., 2009). *k*-NN utilizes a majority-vote method in which the closest known samples in multivariate space determine the assignment of an unknown sample (Coomans and Massart, 1982). Figure 4 is a representation of how *k*-NN classifies unknowns. A training set is built from known samples representing different groups. A test set is also created, made up of evidentiary samples. The training and test sets are run together, forming a model while simultaneously identifying the unknowns' nearest neighbors and classifying them to a known group. One benefit of *k*-NN analysis is that it first measures how closely related the knowns in a training set are via a jackknife method (Tukey, 1958), where one sample is compared to the remaining knowns in a round robin fashion. This training set validation offers the user an opportunity to identify any problem samples (outliers), which can be removed, allowing for the construction of strong, representative known groups. A downfall of *k*-NN is that it is a hard classifier, meaning unknowns are assigned to a group even if they do not belong with any of the known groups. Misclassifications can often be identified through interpretation of threshold values in both the training set validation and classification of unknowns. These values are similar to a t-test statistic and are based on a comparison of the distance an unknown sample falls from its single nearest neighbor in the group it classified to and the smallest intra-point distances of the known samples within that group. The threshold value itself is the number of standard deviations away an unknown can fall and still be

considered correctly classified. It acts as a goodness-of-fit measure, telling the user how well a data point is classifying to a known group (Pirouette user guide, version 4.0).

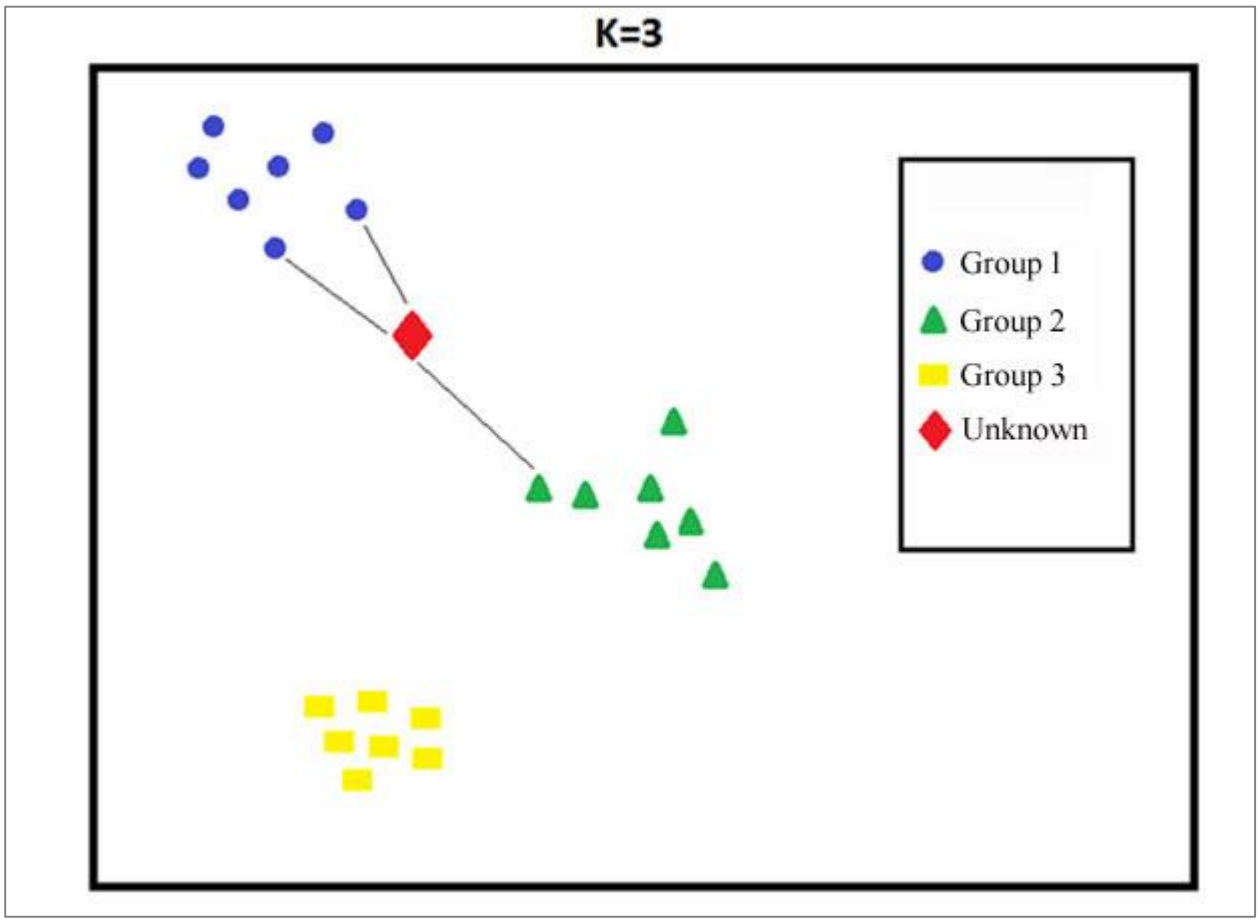

Figure 4—Visual representation of *k*-Nearest Neighbor analysis. In a hypothetical forensic scenario, three groups of known bacterial profiles (blue circles, green triangles, and yellow rectangles), representing three different soil sampling locations, make up the training set. An evidentiary bacterial profile (red diamond) is also analyzed, and its three nearest neighbors are identified. The majority-vote rule classifies the evidence as a member of group 1, meaning that it is most similar to bacterial profiles from that location.

Calculation methods for threshold values and the overall validity of *k*-NN analysis are disputed in the literature. Beebe et al. (1998) recommended calculating threshold values as described above, given that this strategy might help account for natural bacterial variation in a group of knowns, while allowing differences in threshold values based on the amount of

variation within a given group. Lavine and Davidson (2006) argued the opposite, saying *k*-NN cannot accurately measure how well an unknown is classifying, but rather acts as a benchmark test to show the potential of classification techniques in analyzing a particular type of data. This disagreement has led to the general consensus that *k*-NN should be regarded as a baseline technique (e.g., Zhang et al., 2006; Bubeck and Luxberg, 2009), where classification accuracy reflects how more advanced supervised classifiers are likely to perform with the same type of data. *k*-NN uses a relatively simple classification algorithm and only a small number of samples are needed to build training and test sets, providing faster run times and easier analysis, making it an ideal baseline technique.

Goals of This Thesis Research

The overall goal of the research presented in this thesis was to examine whether next-generation sequencing techniques developed by microbiologists could be successfully adopted for the forensic analysis of soil. Traditional forensic soil analysis techniques are often subjective and do not allow for statistical association between samples. The generation of microbial profiles using methods described in the microbiology literature offers the potential to form more definitive associations, as unique characteristics can be identified. Although several molecular methods exist for profile generation, next-generation sequencing of the 16S rRNA gene to assay the bacteria within soil has the most promise, producing massive datasets with high resolution. Sensabaugh (2009) outlined three criteria that should be satisfied for any new microbial technique to gain footing in the forensic sciences. First, it should allow for differentiation of samples collected from different locations. Second, it should have high discriminatory power while also remaining repeatable and robust. Finally, statistical methods used to analyze microbial

profiles should be objective. Attempts to satisfy the first two recommendations using next-generation sequencing were undertaken in this thesis research through comparison of soil bacterial profiles originating from diverse and similar habitats, as well as across time and space within habitats. Additionally, bacterial profiles were generated from soils on mock forensic items to examine whether evidentiary soil could be traced to a single location of origin after being stored up to 1 year. Sensabaugh's third condition was addressed in combination with the examination of what statistical analysis techniques can satisfy the many demands of forensic science. Bacterial profile comparison using abundance charts, NMDS plots, and supervised classification has shown forensic potential (Hopkins, 2014); however, these methods had not yet been used on a large scale to compare bacterial profiles generated from many different soil samples.

MATERIALS AND METHODS

Soil Sampling

Habitat types, GPS coordinates, the study or studies that soils were included in, and the number of collections at each location are displayed in Table 1. Soil samples (except those in the vertical space study) were gathered in the same manner: approximately 100 g of surface soil was collected with a garden trowel rinsed with RO water between collections. Three scoops of soil from a 1×1 foot area at each site were homogenized in an 18-oz Whirl-Pak® bag (Nasco, Fort Atkinson, WI). Soil samples were stored at -20°C until DNA extraction and were always extracted within 1 month of sampling. The bags were kept on ice if they could not be frozen within 1 hour of collection. Photographs of collection sites and evidentiary items can be found in Appendix A. The ground was covered with over a foot of snow and a layer of ice in February 2014, and soil was collected by digging through the snow and chiseling at the ground with a hammer and a screw driver. Chips of soil were collected in sampling bags and the screw driver was washed with RO water between collections. At least an inch of snow was present from December through March; however, chiseling through ice was only necessary in February.

Soil Collection from Diverse Habitats

Soil samples were collected from 10 diverse habitats in the Greater Lansing area every 3 months for 1 year in 2013 and 2014. A map of the sampling locations is displayed in Figure 5. Four of the habitats were located in the Fenner Nature Center, a 134-acre park, shown in the box in Figure 5.

Table 1—Summary of soils collected for all studies.

Sampling Location	GPS Coordinates	Study	n*
Marsh†	42°42'32.0"N 84°30'53.4"W	Diverse Habitat	5
Fallow Agricultural Field	42°45'06.4"N 84°39'42.8"W	Diverse Habitat	5
Beach	42°45'13.9"N 84°24'16.5"W	Diverse Habitat	5
Coniferous Forest	42°41'11.9"N 84°38'05.1"W	Diverse Habitat	5
Field†	42°42'38.9"N 84°31'15.4"W	Diverse Habitat	5
Corn Agricultural Field	42°42'33.5"N 84°28'17.5"W	Diverse Habitat	5
Dirt Roadside	42°48'17.2"N 84°09'33.5"W	Diverse Habitat	5‡
Roadside	42°48'03.4"N 84°11'10.1"W	Diverse Habitat	5
Deciduous Woodlot†	42°42'33.7"N 84°31'01.3"W	Diverse Habitat, Temporal, Horizontal and Vertical Space	51
Yard†	42°42'39.0"N 84°30'53.5"W	Diverse Habitat, Temporal, Horizontal and Vertical Space	51
Treated Yard 1	42°43'26.6"N 84°28'02.5"W	Temporal, Horizontal Space	44
Treated Yard 2	42°43'44.0"N 84°28'23.4"W	Vertical Space	7
Deciduous Woodlot 1	42°42'33.7"N 84°31'00.6"W	Similar Habitat, Evidentiary	5
Deciduous Woodlot 2	42°44'28.2"N 84°27'09.8"W	Similar Habitat, Evidentiary	5
Deciduous Woodlot 3	42°41'03.3"N 84°31'26.1"W	Similar Habitat, Evidentiary	5
Deciduous Woodlot 4	42°40'57.2"N 84°28'05.6"W	Similar Habitat, Evidentiary	5
Deciduous Woodlot 5	42°43'38.9"N 84°30'08.8"W	Similar Habitat, Evidentiary	5
Deciduous Woodlot 6	42°44'38.9"N 84°28'57.9"W	Similar Habitat, Evidentiary	5
Deciduous Woodlot 7	42°42'50.8"N 84°28'38.5"W	Similar Habitat, Evidentiary	5
Deciduous Woodlot 8	42°42'00.8"N 84°31'35.0"W	Similar Habitat, Evidentiary	5
Deciduous Woodlot 9	42°41'25.6"N 84°27'41.2"W	Similar Habitat, Evidentiary	5
Tire	-	Preliminary Evidentiary	5
Shoe	-	Preliminary Evidentiary	5
Sock	-	Preliminary Evidentiary	5
Shirt	-	Preliminary Evidentiary	5
Shovel	-	Preliminary Evidentiary	5
T-shirts (24°C)	-	T-Shirt Evidentiary	44
T-shirts (4°C)	-	T-Shirt Evidentiary	44

*Number of collections

†Habitats within the Fenner Nature Center

‡One soil sample from this set produced less than 3000 sequence reads and was excluded from further processing.

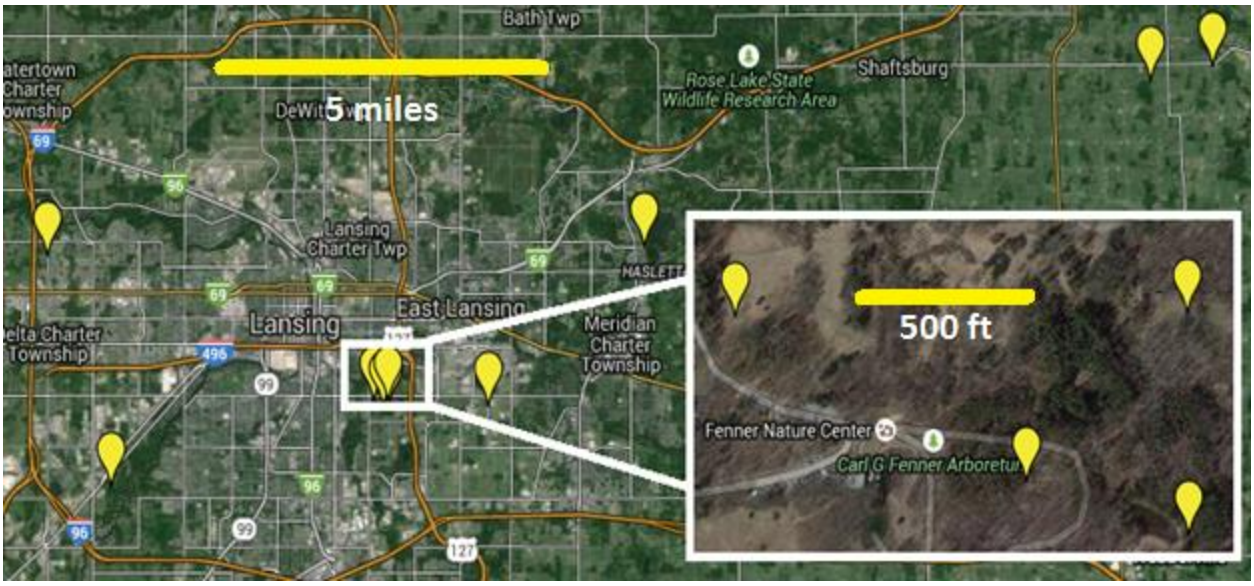

Figure 5—Map of soil sampling locations for the diverse habitat study. The four collection sites at the Fenner Nature Center are shown in the inset.

Soil Collection from Similar Habitats

Soil samples were collected from nine deciduous woodlots in the Greater Lansing area once every 2 weeks for 8 weeks starting in May 2014. Collection locations are shown in Figure 6.

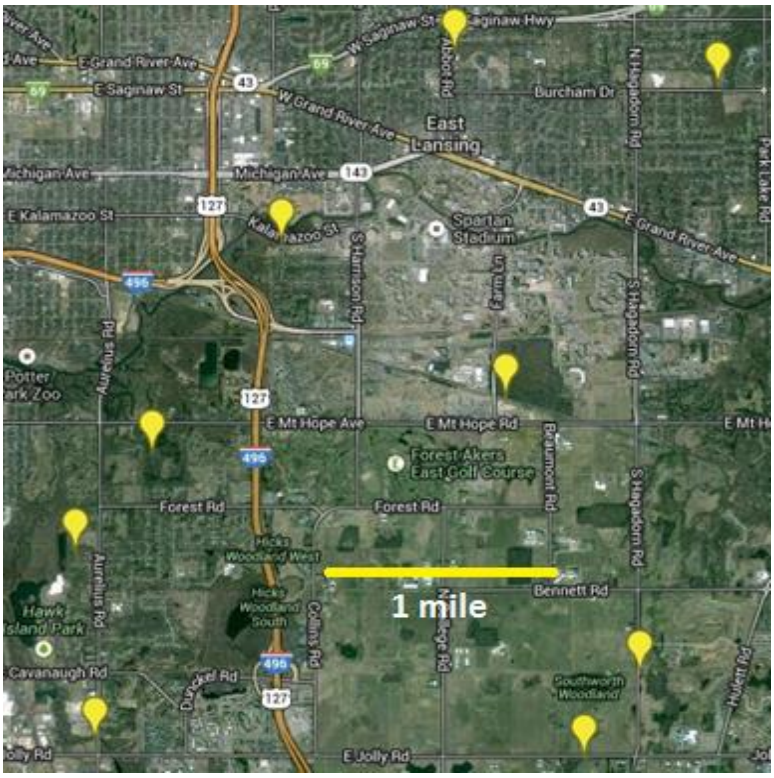

Figure 6—Map of soil sampling locations for the similar habitat study. Locations were within 6 miles of one another in the Greater Lansing area.

Soil Collection from Three Habitats over Time

Surface soil samples were collected at a central point in three habitats, a yard and a deciduous woodlot at the Fenner Nature Center and a yard treated with pesticides and fertilizer on the Michigan State University campus, once a day for 4 days, once a week for 2 months, and once a month for the remainder of the year starting in August 2013.

Soil Collection from Three Habitats over Horizontal Space

Surface soil samples were collected in three habitats, a yard and a deciduous woodlot at the Fenner Nature Center and a treated yard on the Michigan State University campus, in March 2014. A central soil sample and 16 additional samples 5, 10, 50, and 100 feet in each cardinal direction were collected, resulting in 17 samples per location.

Soil Collection from Three Habitats over Vertical Space

Soils samples were collected at a yard and a deciduous woodlot at the Fenner Nature Center and a second treated yard on the Michigan State University campus in October 2013[1] and April 2014 using a soil corer and mud auger (AMS, Inc. American Falls, ID) that were rinsed with RO water between samplings. After a surface soil sample was taken, the mud auger was driven into the ground, removed, and soil samples at 1, 2, 5, and 10 inches below the surface were collected. A soil corer was then driven into the ground in the same location, removed, and soil samples were collected at 20 and 60 inches below the surface. A maximum depth of 25 inches was reached in the treated yard due to obstruction. Soil samples at depths of 20 inches or more were lighter brown in color and clay-like in all habitats.

Soil Collection from Evidentiary Items

Soil was collected for a preliminary mock evidence study from deciduous woodlot 1 and deposited on a shoe, steel shovel blade, cotton polo shirt, and cotton sock while in the woodlot, which were placed in brown paper bags. Additional soil was collected from the site, transported to a storage room near the forensic science laboratory in Giltner Hall on the Michigan State University campus, and deposited on a tire. Items were stored in the room at ambient temperature for 1 year. Soil was collected from three different areas on each evidentiary item as well as one sample of homogenized soil after 6 months of storage. Soil was again collected from three locations on each evidentiary item and homogenized into one sample after the full year. Bacterial profiles were not generated from the items prior to soil exposure.

A second evidentiary study was conducted in which eight new white cotton T-shirts (Hanes®, Winston Salem, NC) were exposed to soil in a 2×2 foot area of deciduous woodlot 1

[1]The treated yard vertical space collections in October 2013 were processed via 454 pyrosequencing by Hopkins (2014). Therefore, the analysis of these bacterial profiles is not included in this research.

by placing soil on a shirt, folding the sides of the shirt around the soil, and rubbing it into the fabric while wearing vinyl exam gloves. The shirts were placed in numbered brown paper bags. T-shirts 1 – 4 were stored in an incubator (24°C), while T-shirts 5 – 8 were stored in a laboratory refrigerator (4°C) for the duration of the study. On day zero and once a week for 8 weeks, small (ca. 1 cm^2) soil-covered portions were cut and collected from each shirt. Additional portions were collected once a month for 2 months following the 8-week period. A cutting was taken from a clean shirt on the initial sampling day. On day 0 and every 2 weeks for 8 weeks, soil was collected from the deciduous woodlot of origin (woodlot 1) and from the eight other deciduous woodlots described in the similar habitat study.

DNA Extraction from Soil

A Spectrolinker XL-1500 UV Crosslinker (Spectronic Corporation, Lincoln, NE) was used to UV irradiate pipette tips, pipettes, tubes, and scissors for 5 min (~ 2.5 J/cm^2). DNA was extracted from soil samples with a PowerSoil® DNA Isolation Kit (MoBio, Carlsbad, CA) following the manufacturer's protocol with two changes: spin filters containing bound DNA were washed with 500 μL of 70% ethanol and centrifuged for 30 s at 10,000 × g immediately following protocol step 17. Additionally, solution C6 was heated in a 55°C incubator prior to its addition to the spin filter. The soil covered cuttings and the clean cutting collected from the T-shirts were placed directly into extraction tubes. Reagent blanks were processed with every extraction.

PCR Amplification of 16S rRNA Variable Regions 3 and 4

Bacterial 16S rRNA gene variable regions 3 and 4 were amplified using a forward primer (357F [Haas et al., 2011]) and a reverse primer (806R [Caporaso et al., 2010]) that contained one of 96 barcodes, allowing for identification of which soil sample each sequence originated from in downstream analysis. Fifteen microliter PCR reactions contained final concentrations of 1X AmpliTaq Gold buffer (Life Technologies, Carlsbad, CA), 2.5 mM $MgCl_2$, 0.2 mM nucleotide triphosphates, 0.4 µg/µL bovine serum albumin, 1U AmpliTaq Gold (Life Technologies), 1 µL of template DNA, and 1 µL of 10 µM premixed forward and reverse primers. DNAs were denatured on an Applied Biosystems® 2720 thermal cycler (Life Technologies) for 10 min at 94°C, followed by 35 cycles of 94°C for 30 s, 60°C for 45 s, and 72°C for 60 s, and a final extension of 10 min at 70°C. Four microliters of the PCR product were electrophoresed on a 1.5% agarose gel followed by ethidium bromide staining and UV visualization.

DNA Quantification and Equimolar Pooling

PCR products were quantified using a Quant-iT™ PicoGreen® dsDNA Assay Kit (Life Technologies) following the manufacturer's protocol, and pooled such that each bacterial sample was at equal concentration (~6 ng/µL). Pooled DNAs were purified using Agencourt® AMPure® XP (Beckman Coulter, Brea, CA) beads. The beads were vortexed and added in a 0.6:1 ratio to the tube containing pooled DNAs. This solution was vortexed for 15 s and allowed to incubate for 15 min at room temperature. The tube was placed in a MagnaRack™ (Life Technologies) for 5 min. The supernatant was aspirated and discarded. Five hundred microliters of 70% ethanol was used to wash the bound beads. The supernatant was aspirated away after 30 s and the ethanol washing process was repeated. The tube, still in the MagnaRack™, was placed in

a 37°C incubator and allowed to dry for 30 min. It was removed from the rack and 100 µL of 10 mM Tris at pH 8 was used to elute the DNA by vortexing for 10 s. The tube was placed back in the MagnaRack™ for 5 min, and the supernatant was transferred to a 1.5 mL microcentrifuge tube.

Sequencing of Purified PCR Products

The pooled bacterial DNAs were sequenced on an Illumina MiSeq (Illumina, San Diego, CA) following the manufacturer's protocol using a paired end 250 bp v2 Reagent Kit (Illumina). Base calling was performed with Real Time Analysis software v1.18.54 (Illumina), and the output was demultiplexed and converted to FastQ files with Bcl2fastq Conversion Software v1.8.4 (Illumina).

Next-Generation Sequencing Data Processing

Sequencing data were processed using open-source mothur software following the MiSeq sequence processing standard operating procedures on the mothur webpage (Schloss et al., 2009; www.mothur.org). Sequence processing commands for mothur are given in Appendix B. Bacterial profiles were subsampled to 3000 sequences per soil sample[2]. Sequences were organized into operational taxonomic units (OTUs) at a 97% similarity cutoff. The number of OTUs in each profile was documented and considered to represent sequence diversity.

[2] Subsampling is a necessary step in sequence processing due to the computational limits when handling massive amounts of data such as those produced via next-generation sequencing. The effects of subsampling on sequence libraries were examined by subsampling the diverse habitat soil samples (n=49) down to 3000 sequences four times and assessing congruity. Each subsampling resulted in approximately equivalent measures of dissimilarity and orientation in NMDS plots, demonstrating that subsampling had little effect on profile analysis.

Next-Generation Sequencing Data Analysis Procedures

OTUs were used to calculate Sørensen-Dice coefficients within mothur, and the resulting square symmetric dissimilarity matrices were used as the input for NMDS, which was run in XLSTAT Pro (Addinsoft, New York, NY), and *k*-NN, which was run in Pirouette 4.0 (Infometrix, Inc. ©, Bothell, WA). A jackknife resampling method was employed for *k*-NN analysis of diverse and similar habitat soil bacterial profiles, in which each soil bacterial profile was tested against the other four collected from the same site resulting in a calibration accuracy. Training and test sets for *k*-NN analysis are described in Table 2. The accuracy of *k*-NN was measured by its ability to classify soil bacterial profiles to their location of origin. Threshold values were recorded at a 95% confidence level. Bacterial profile OTUs were also classified using the SILVA bacterial reference alignment (Quast et al., 2013), and abundance charts were created at the taxonomic class level in Excel (Microsoft, Redmond, WA). The number of bacterial classes in each profile was documented and representative of taxonomic class diversity.

Table 2—Training and test sets for *k*-NN analysis of soil bacterial profiles.

Study	Training Set	Test Set
Diverse Habitats*	N=4 per habitat	N=1
Similar Habitats*	N=4 per woodlot	N=1
Temporal	First seven bacterial profiles (August and September)	All other collections over 1 year
Within-Habitat Horizontal Space	Center, 5 ft N, 5 ft S, 5 ft W, 5 ft E	All other distance soil bacterial profiles
	Center, 5 ft E, 10 ft N, 50 ft W, and 100 ft S†	All other distance soil bacterial profiles
	Center, 100 ft N, 100 ft S, 100 ft W, 100 ft E	All other distance soil bacterial profiles
Within-Habitat Vertical Space	Surface, 2 in,10 in, and 60 in	All other depth soil bacterial profiles
Preliminary Evidentiary	Woodlot soil bacterial profiles over 8-wk period	Various evidentiary soil bacterial profiles after 6 months and 1 year
T-Shirt Evidentiary	Woodlot soil bacterial profiles over 8-wk period	T-shirt evidentiary soil bacterial profiles over 4 months

*Analyzed via the jackknife resampling method (Tukey, 1958) in which each of the five soil bacterial profiles was systematically left out and tested against the other four profiles.
†Three additional spiral training sets were also tested in this manner with different 5 ft starting points.

RESULTS

Soil Collection, Extraction, and Amplification

Soil samples collected for the temporal study from the Fenner deciduous woodlot, the Fenner yard, and the Michigan State University treated yard in December – March and for the diverse habitat study in February contained ice and snow. These soils were very slushy and sometimes completely submerged when thawed for extraction. Water could not be avoided during soil weighing and the mass of the soil was exaggerated by the moisture, therefore its amount in the extraction tube was likely less than other collections.

DNA from the soil-covered T-shirt cuttings did not amplify as well as samples containing only soil; however, quantification revealed enough amplified DNA was present for sequencing in all cases. DNA from the clean T-shirt cutting also did not amplify well and had to be re-sequenced in order to obtain more than 3000 sequences. The resulting bacterial profile contained 256 OTUs representing 30 bacterial classes, all of which were present in at least one profile generated from the deciduous woodlot in which t-shirts were exposed to soil. The bacterial profile generated from the clean shirt forced others together in NMDS plots (data not shown) and did not classify as the woodlot of origin, nor was it under the threshold value for any other deciduous woodlot in *k*-NN. All other bacterial DNAs amplified well, with an average post-PCR quantification of approximately 20 ng/µL; however, no particular location possessed soil that consistently had the highest DNA quantifications.

Illumina MiSeq Sequencing Efficiency

MiSeq sequencing resulted in datasets of approximately 150,000 sequence reads per soil sample. Processing of bacterial sequence libraries in mothur led to the removal of 94 to 97% of

sequences, primarily during the subsampling portion of sequence analysis. Only the dirt road collection from February did not produce the requisite number of sequences and it was excluded from analysis.

Soil Bacterial Profile OTU Diversity

Soil bacterial profiles contained between 242 and 1231 OTUs (Tables C1 – C7, Appendix C)—each representing 97% similar sequences—with the dirt road soils having the lowest average (276) and the treated yard soils having the highest average (1033). Fewer OTUs were generated from soils collected in late February from the Fenner deciduous woodlot (483) and in late February and March from the Fenner yard (739 and 764, respectively) than from deciduous woodlot and yard soils collected in the other months of the year (average of 1012 and 1071, respectively). Additionally, the yard soil sample collected 5 feet east of the center sampling site contained fewer OTUs (781) than the rest of the yard surface samples (average of 1000). Soil samples collected at different depths contained similar numbers of OTUs with the exception of the 60 inch sample from the deciduous woodlot in October which contained substantially fewer (242 OTUs). The number of OTUs in the April 60 inch deciduous woodlot sample was also relatively low (647). Fewer OTUs were present in preliminary evidentiary soil profiles after storage for 6 months (average of 765) and 1 year (average of 895) than soil samples collected from the deciduous woodlot of origin (average of 1050). Soil samples from evidentiary T-shirts initially contained similar numbers of OTUs as woodlot of origin soils (Figure 7); however, the number of OTUs decreased over the 4-month period at both storage temperatures.

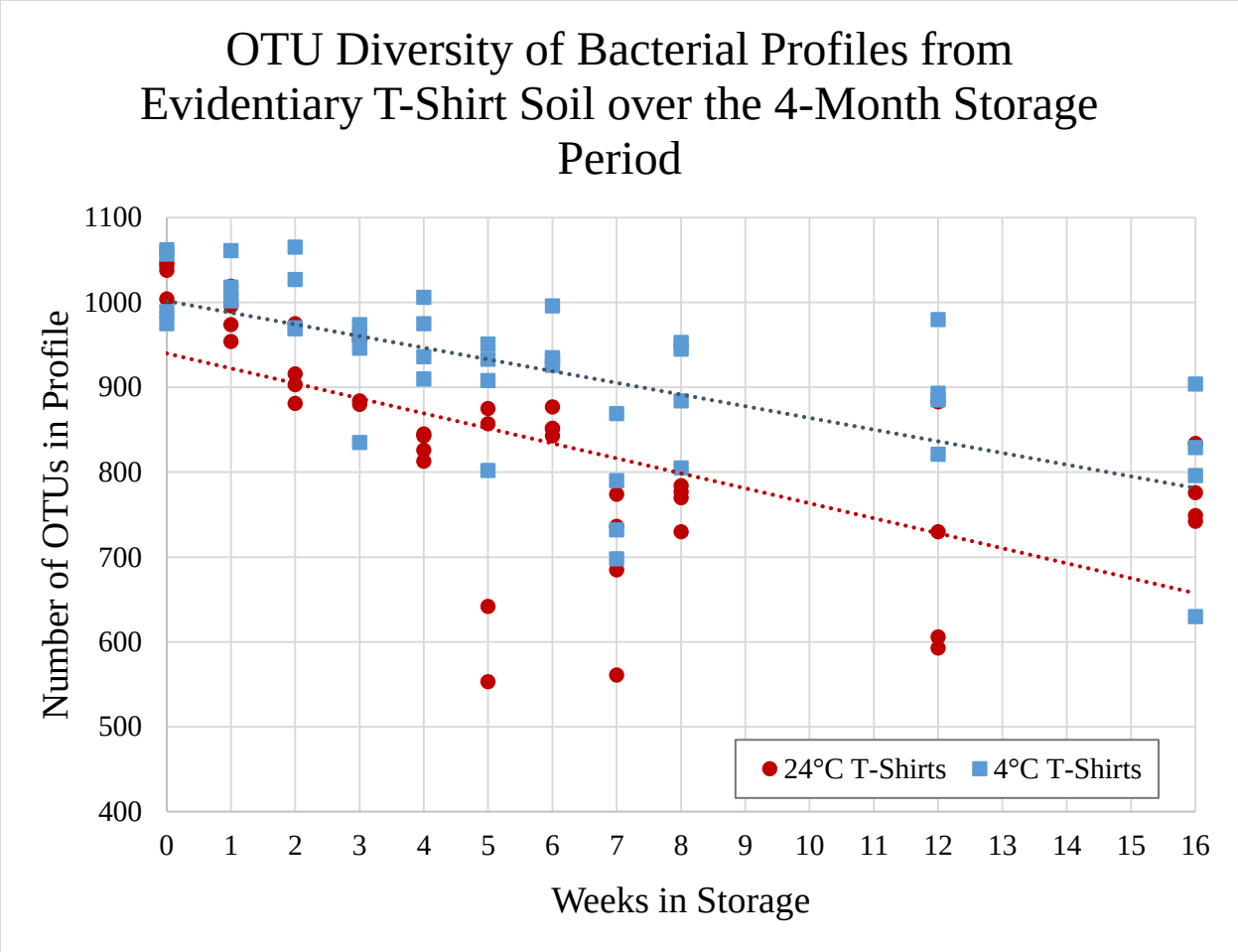

Figure 7—Number of OTUs in bacterial profiles generated from evidentiary T-shirts over the 4-month storage period. Trend lines showing the decrease of OTUs over time at each storage temperature are displayed. The quantity of OTUs varied among the replicate shirts at each sampling time; however, soils generally contained fewer OTUs the longer they were stored. This decrease occurred at a slightly faster rate in soil on T-shirts stored at 24°C.

General Soil Bacterial Profile Analysis Results

Each soil bacterial profile contained approximately 50 bacterial taxonomic classes (range: 41 to 58), with the exception of profiles generated from dirt road soils, which had lower class diversity (see below) and the profiles generated from soil collected in February from the Fenner deciduous woodlot, which contained 22 classes. All other bacterial profiles shared classes that

made up a large fraction of their abundance; differences among soil profiles tended to be in the least abundant classes[3].

The Scree diagrams produced with NMDS plots showed decreasing stress as dimensions increased, with a characteristic elbow at two dimensions (e.g., Figure 8). Additionally, Shepard diagrams produced with the plots exhibited close association between distances and disparities, confirming the low stress at two dimensions in the Scree diagram (e.g. Figure 9).

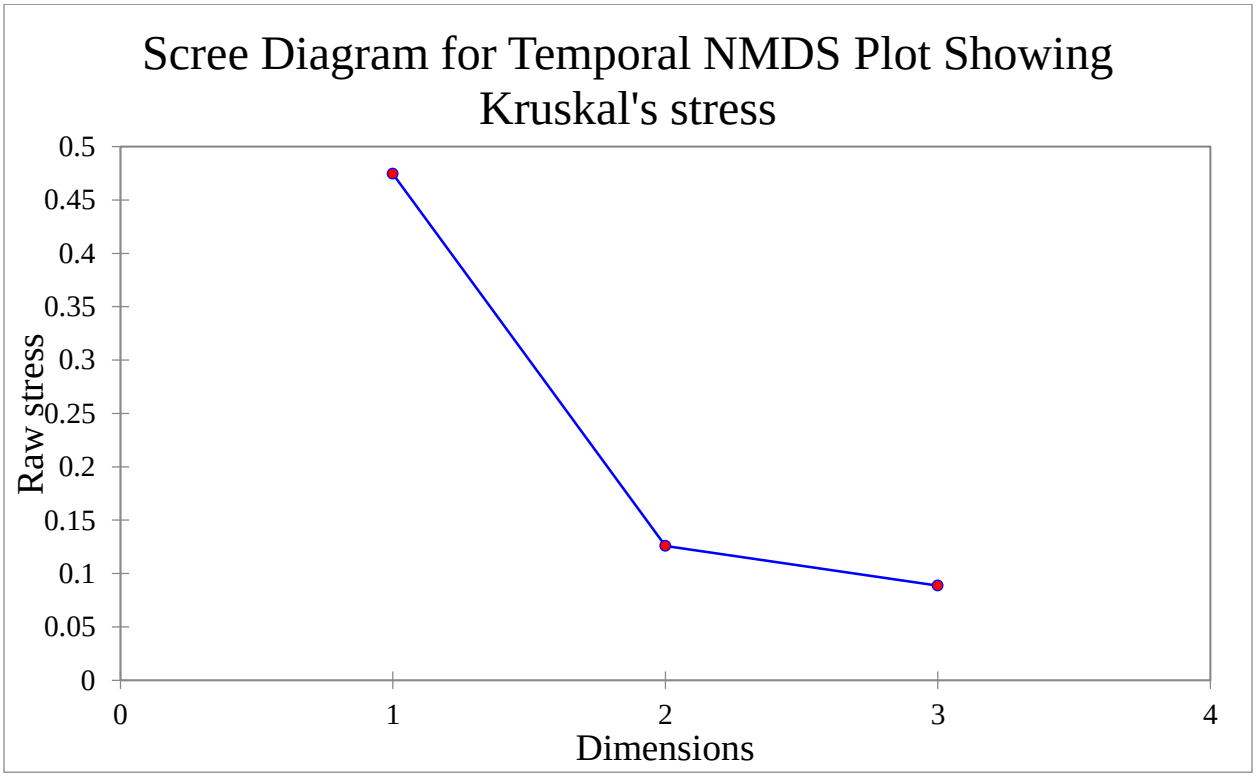

Figure 8—Scree diagram generated from the ordination of temporal soil bacterial profiles showing an elbow signifying the substantial decrease in stress from one to two dimensions, and a general leveling off with additional dimensions. All Scree diagrams were similar.

[3] Note that in the abundance charts that follow, bacterial classes are graphed in order of abundance for a given sample set, thus class order can differ among charts.

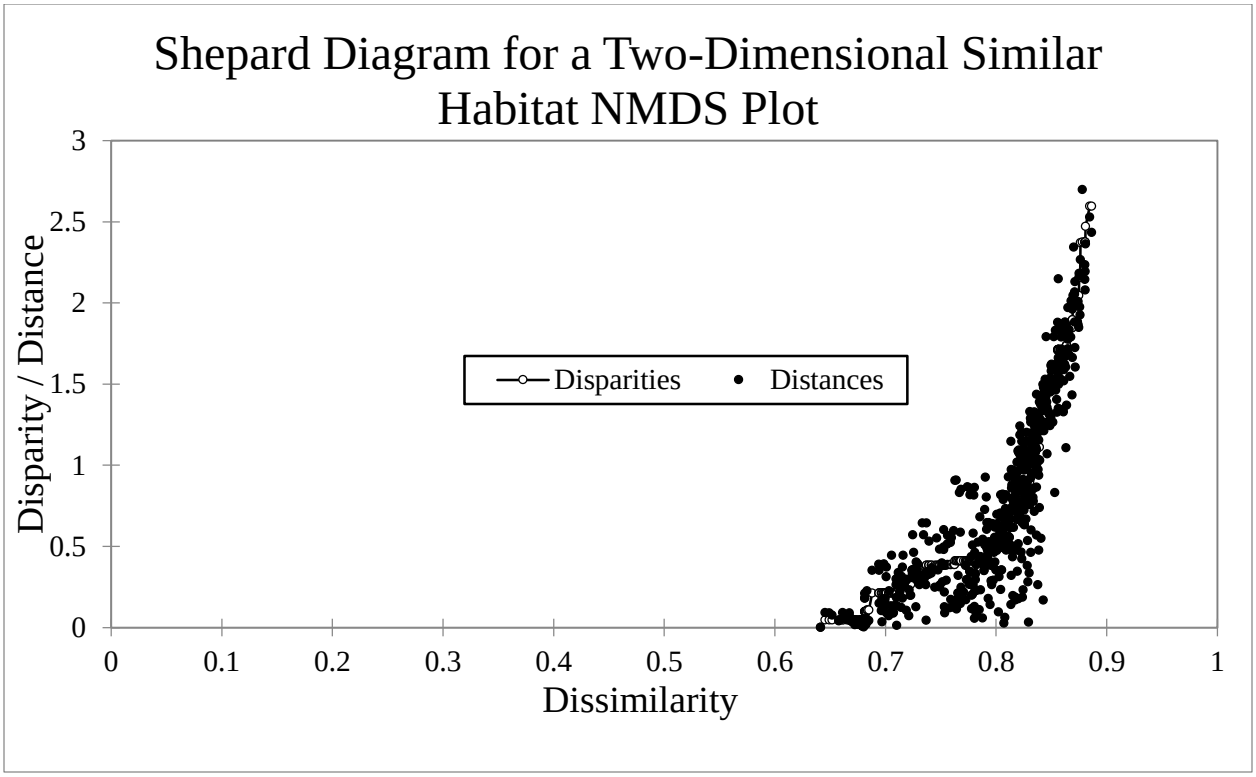

Figure 9—Shepard diagram generated with similar habitat NMDS plot. Distances fell close to corresponding disparities, indicating good correlation between the two metrics in the accompanying NMDS plot. All Shepard diagrams were similar.

Intermingling of soil sampling location clusters was common when many profiles were ordinated together in NMDS plots; however, these clusters were resolved when the locations containing overlapping members were ordinated in pairs or triads. Additionally, NMDS plots developed with all bacterial profiles for a given study always had higher stress than when a subset of those profiles were ordinated together (data not shown).

Table 3 summarizes the *k*-NN classification results for all bacterial profiles, specifying the training sets used, classification accuracy, and which profiles were misclassified. Training sets consisted of five or more members, with the exception of one vertical space training set, made up of only four. *k*-NN analysis resulted in accurate classification of soils to their location of origin for 97.6% of the bacterial profiles generated in this research (percentage based on the

training set that produced the most accurate classification in each study, as it did change slightly depending on the bacterial profiles acting as knowns). Table 4 summarizes *k*-NN threshold value results, specifying the number of profiles correctly classified, the percentage of profiles correctly classified and under the threshold values, and which profiles were correctly classified but over the threshold value.

Table 3—*k*-NN classification results. Multiple training sets were examined for all soil sets with the exception of the temporal and evidentiary studies.

Study	Training Set	Number of Profiles Analyzed	Classification Accuracy	Misclassified Profiles
Diverse Habitats	All habitat bacterial profiles*	49	88%	August Marsh, Fallow Ag† Field, Deciduous Woodlot, and Yard
	Marsh and Fallow Ag† Field profiles*	10	100%	-
	Deciduous Woodlot and Yard profiles*	10	100%	-
Similar Habitats	All location bacterial profiles*	45	87.5%	All from Deciduous Woodlot 8 and one from Deciduous Woodlot 9‡
	Deciduous Woodlots 1 – 7 profiles*	35	100%	-
Temporal	First seven profiles in each habitat (August and September)	48	93.8%	February Deciduous Woodlot and Yard, March Yard‡

Table 3 (cont'd)

Horizontal Space	Profiles from the center, 5 ft N, 5 ft S, 5 ft W, 5 ft E	36	94.4%	Yard 100 ft N, Deciduous Woodlot 100 ft S‡
	Profiles from the center, 5 ft E, 10 ft N, 50 ft W, and 100 ft S§	36	97.2%	Deciduous Woodlot 100 ft N‡
	Profiles from the center, 100 ft N, 100 ft S, 100 ft W, 100 ft E	36	94.4%	Yard 5 ft E, Yard 10 ft N‡
Vertical Space	Profiles from the surface, 2 in, 10 in, 60 in	15	100%	-
	Profiles from the surface, 1 in, 2 in, 5 in, 10 in	10	80%	Deciduous Woodlot 60 in (October and April) ‡
Preliminary Evidentiary	Deciduous Woodlots 1 – 9 profiles	25	100%	-
T-Shirt Evidentiary	Deciduous Woodlots 1 – 9 profiles	88	100%	-

* Analyzed via jackknife method (Tukey, 1958)
†Ag=Agricultural
‡Pairwise *k*-NN analysis did not resolve bacterial profiles from these locations
§Other training sets developed using this spiral method produced similar results (see below)

Table 4—*k*-NN threshold value results. Threshold values were evaluated only if profiles were correctly classified.

Study	Training Set	Number of Profiles Correctly Classified	Percent Correctly Classified and under Threshold Value	Profiles Correctly Classified and over Threshold
Diverse Habitats	All habitat bacterial profiles*	43	90.2%	February Coniferous Forest and Roadside, May Beach and Dirt Road
	Marsh and Fallow Ag† Field profiles*	10	90.0%	February Fallow Ag Field
	Deciduous Woodlot and Yard profiles*	10	100.0%	-
Similar Habitats	All location bacterial profiles*	39	93.3%	June Woodlot 4 and 5, July Woodlot 1
	Deciduous Woodlots 1 – 7 profiles*	35	91.4%	May Woodlot 7, June Woodlot 6, and July Woodlot 4
Temporal	First seven profiles in each habitat (August and September)	38	73.7%	January Treated Yard, February all habitats, March Woodlot, April Woodlot and Yard, May Woodlot, June Yard, December Woodlot
Horizontal Space	Profiles from the center, 5 ft N, 5 ft S, 5 ft W, 5 ft E	36	100.0%	-

Table 4 (cont'd)

Horizontal Space	Profiles from the center, 5 ft E, 10 ft N, 50 ft W, and 100 ft S	36	100.0%	-
	Profiles from the center, 100 ft N, 100 ft S, 100 ft W, 100 ft E	36	100.0%	-
Vertical Space	Profiles from the surface, 2 in, 10 in, 60 in	23	100.0%	-
	Profiles from the surface, 1 in, 2 in, 5 in, 10 in	20	100.0%	-
Preliminary Evidentiary	Deciduous Woodlots 1 – 9 profiles	25	0.0%	All
T-Shirt Evidentiary	Deciduous Woodlots 1 – 9 profiles	88	(See Figure X)	(See Figure X)

*Analyzed via jackknife method (Tukey, 1958)
†Ag=Agricultural

Analysis of Soils from Diverse Habitats

Bacterial Abundance Charts

Averaged diverse habitat profiles (Figure 10) appeared similar, with the exception of the dirt road, which had lower class diversity (range: 24 to 29) and yielded substantially lower levels of *Acidobacteria* and *Betaproteobacteria*, and higher levels of *Flavobacteria*,

Gammaproteobacteria, *Clostridia*, and *Bacilli* (denoted by arrows) relative to the other habitats.

Abundance charts of soils collected at each sampling time can be found in Appendix D.

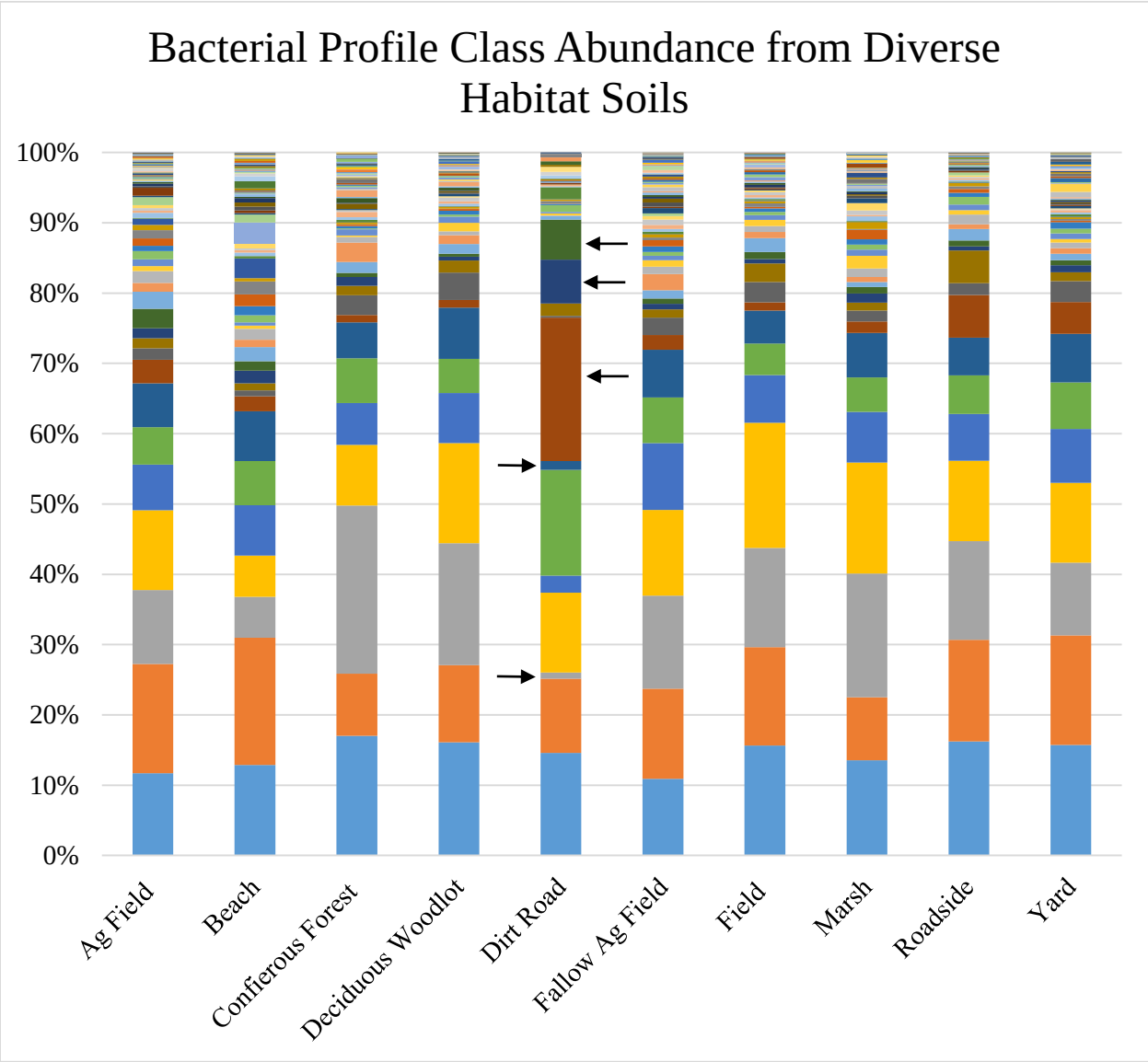

Figure 10—Average bacterial class abundance of five soil samples from ten diverse habitats. The dirt road soil clearly differed from the other habitats, containing higher levels of *Flavobacteria*, *Clostridia*, and *Bacilli* (denoted by arrows in ascending order on the right), along with lower levels of *Acidobacteria* and *Betaproteobacteria* (denoted by arrows in ascending order on the left). Ag=Agricultural.

Nonmetric Multidimensional Scaling

Bacterial profiles generated from soil collected within a habitat clustered together in NMDS plots (Figure 11), but some intermingling occurred among the 10 habitats. Clusters of bacterial profiles generated from soil collected in the deciduous woodlot, yard, and field at the Fenner Nature Center overlapped, the marsh and fallow agricultural field clusters intermingled slightly, and profiles from the agricultural field, roadside, and beach clusters intermixed. Removal of the dirt road profiles, which clustered the farthest away in the plot, did not resolve these intermingled clusters (data not shown). When two or three habitats were oriented at a time, clusters separated in all cases (e.g., Figure 12); however, profiles from the same location did not cluster as tightly as they did when all habitats were oriented together.

k-Nearest Neighbor

k-NN accurately classified diverse habitat soil bacterial profiles 88% of the time when all habitats were analyzed together (Table 3). Misclassifications occurred between the marsh and fallow agricultural field soil bacterial profiles and between the Fenner deciduous woodlot and yard profiles. These profiles were correctly classified to their habitat when analyzed as pairs in a *k*-NN model.

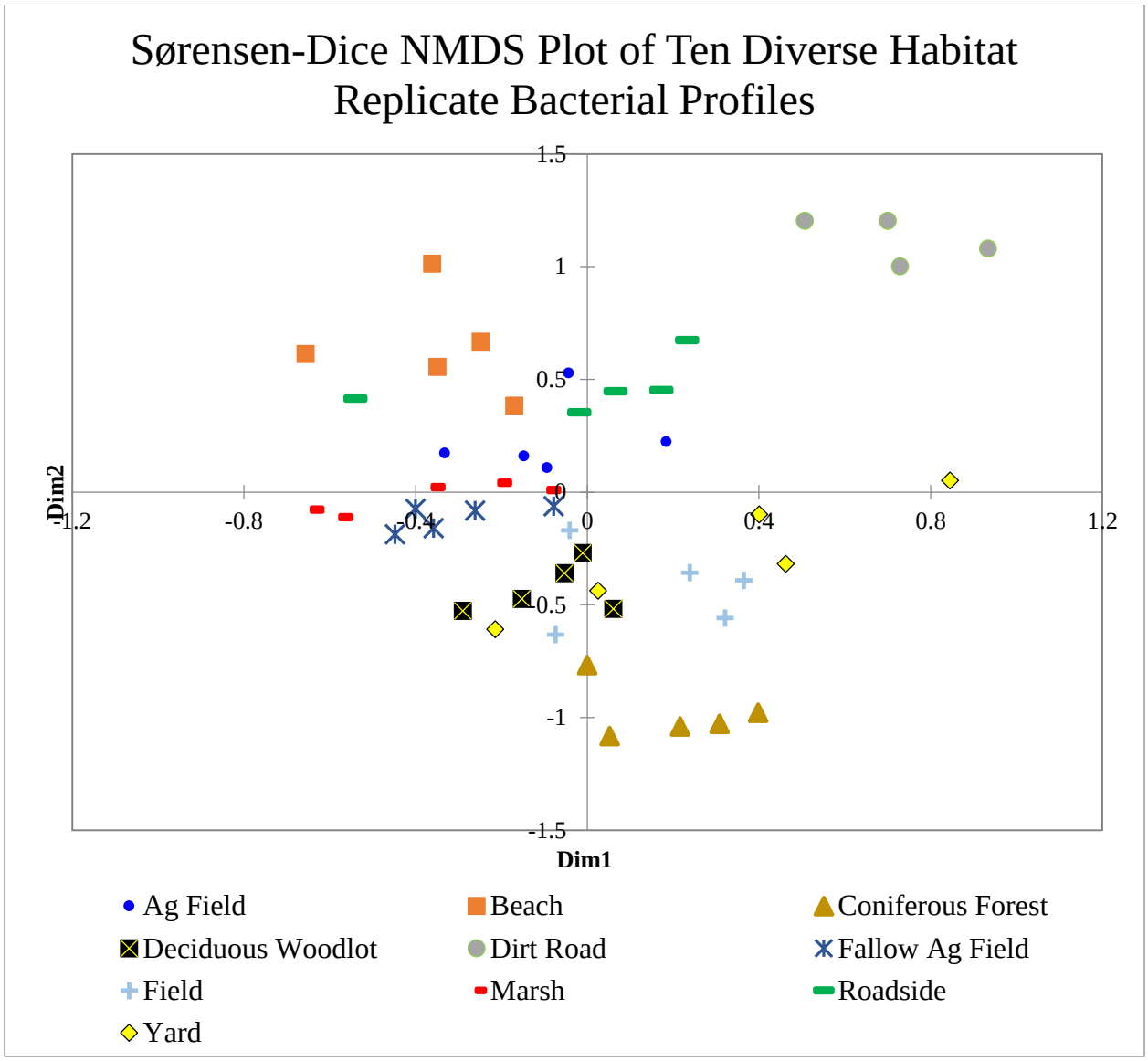

Figure 11—NMDS plot ordinating soil bacterial profiles from the 10 diverse habitats. Replicate profiles from the same habitat formed clusters, but intermingling occurred among some of the habitats. Ag=Agricultural.

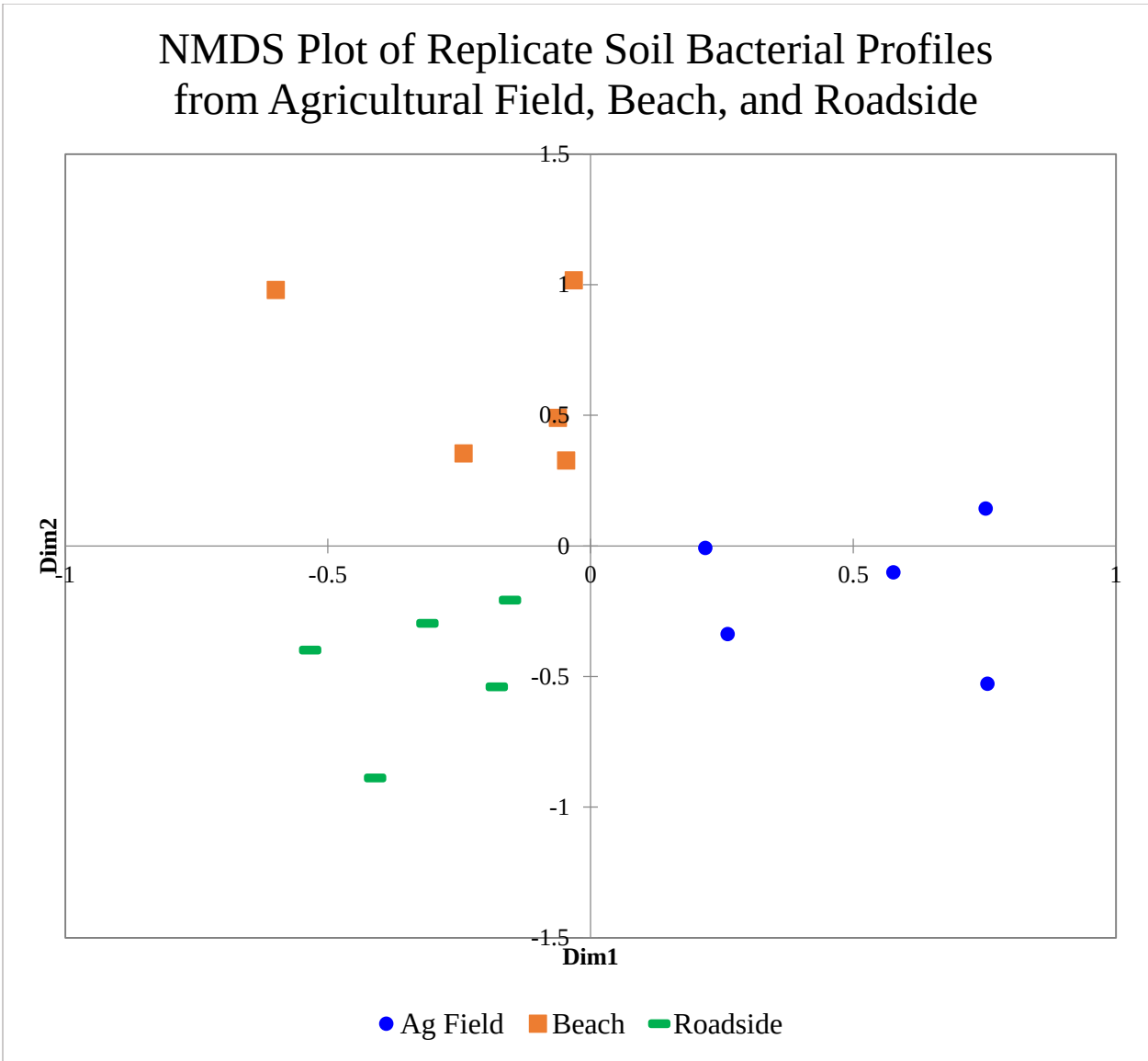

Figure 12—NMDS plot ordinating soil bacterial profiles from the agricultural (Ag) field, beach, and roadside. Profiles from these locations intermingled when all habitats were ordinated together, but were resolved when analyzed as pairs or triads in NMDS plots.

Analysis of Soils from Similar Habitats

Bacterial Abundance Charts

Averaged similar habitat profiles (Figure 13) appeared more similar than those obtained from the diverse habitats (Figure 10); however, class diversity differed among deciduous

woodlots, ranging from 40 to 59 bacterial classes per profile. Abundance charts of soils collected at each sampling time can be found in Appendix D.

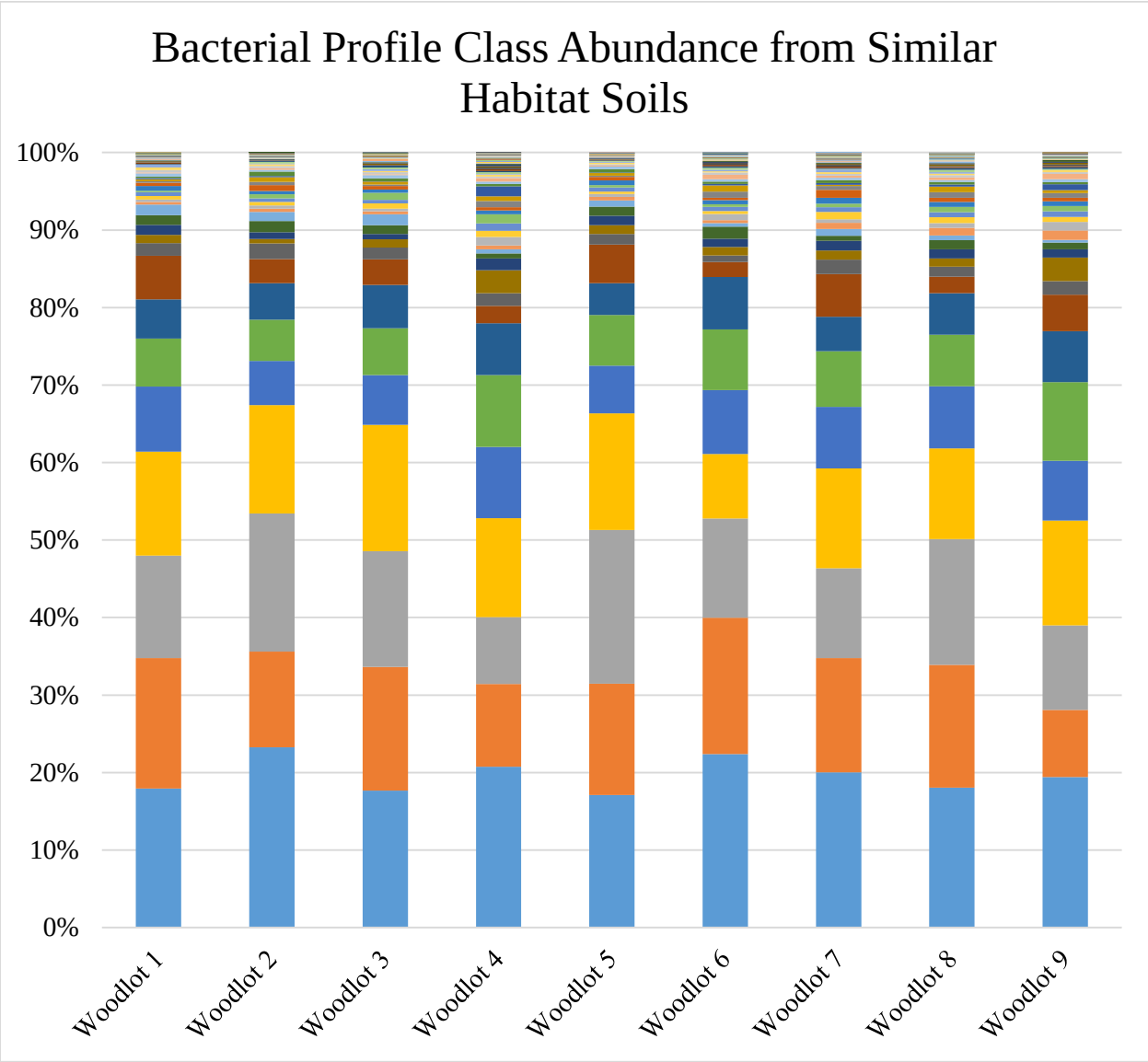

Figure 13—Average bacterial class abundance of five soil samples from nine deciduous woodlots. The profiles appeared very similar, sharing the most abundant bacterial classes.

Nonmetric Multidimensional Scaling

Soil bacterial profiles from a given deciduous woodlot clustered together in NMDS plots (Figure 14), but intermingling occurred among several of the clusters, such as woodlots 2 and 3, and one profile from woodlot 5. The most substantial overlap involved woodlot 8, whose bacterial profiles were interspersed among several other clusters. By ordinating profiles in pairs or triads, separation of deciduous woodlots occurred in all cases (e.g., Figure 15). NMDS plots showing the separation of woodlot 8 from others can be found in Appendix E.

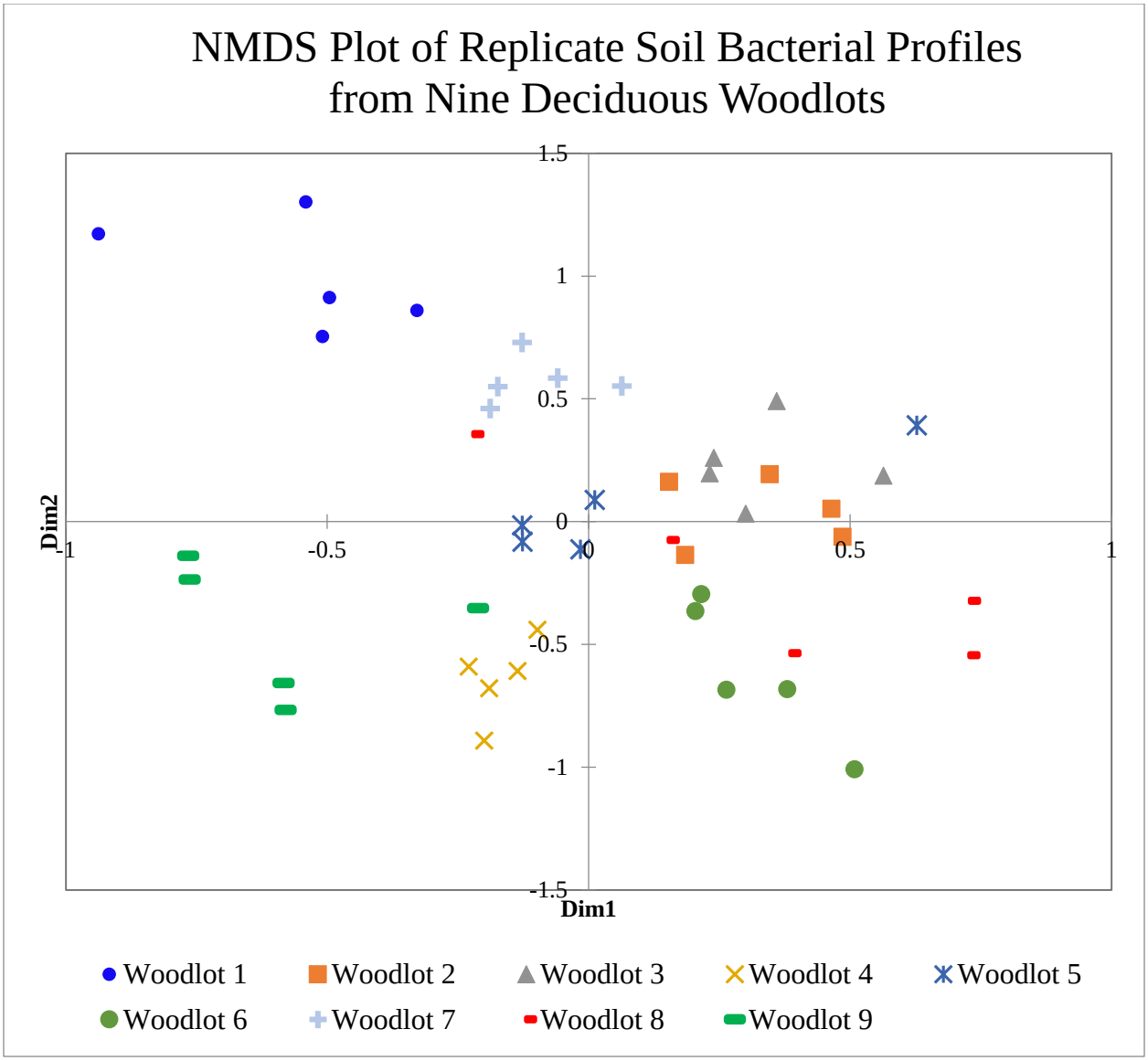

Figure 14—NMDS plot ordinating soil bacterial profiles from the nine deciduous woodlots. Profiles from the same location formed clusters, but intermingling occurred among some of the location clusters. Woodlot 8 replicate profiles clustered relatively poorly, intermingling with several other woodlot profiles; however, these clusters were resolved when fewer woodlots were ordinated together.

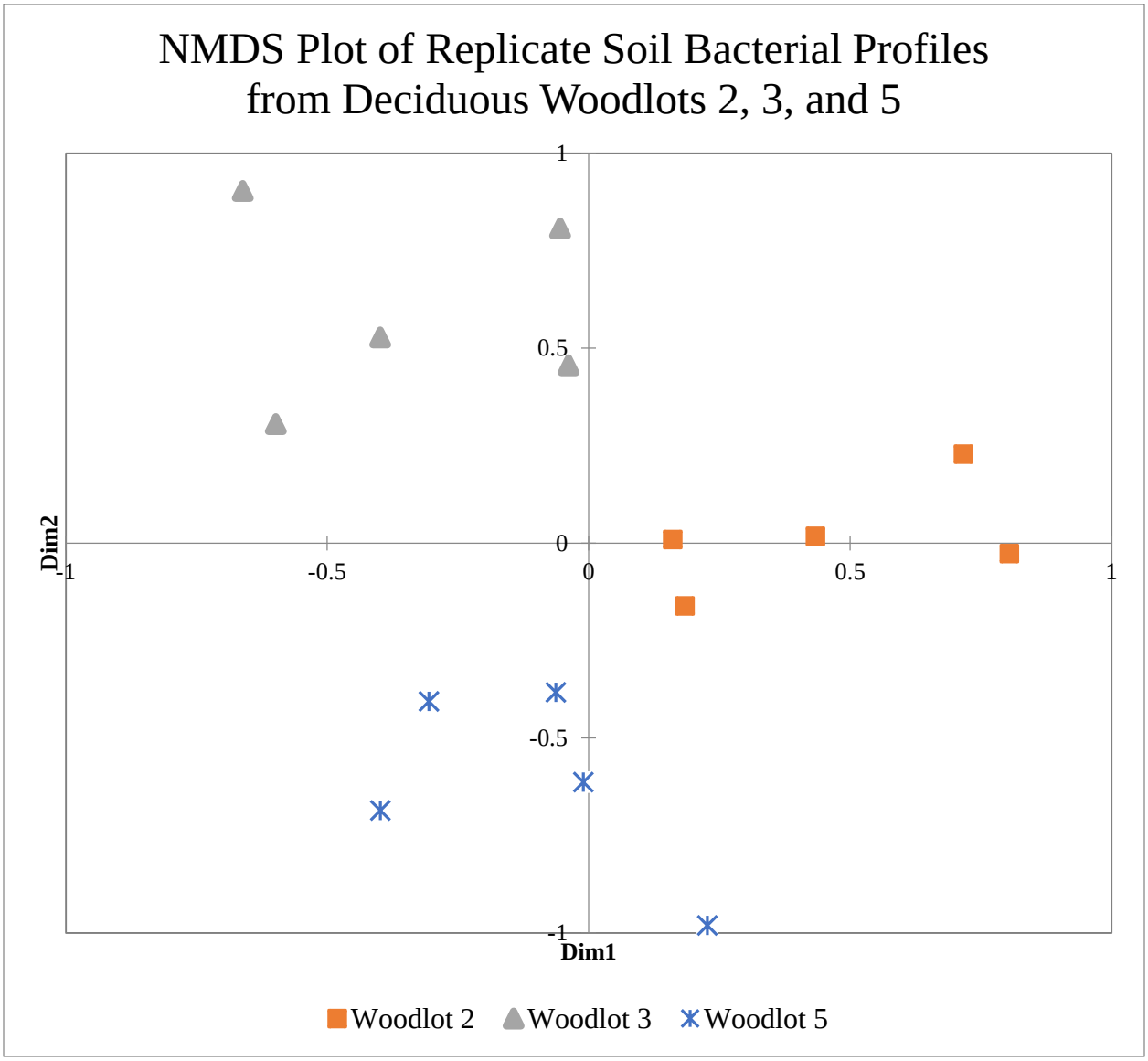

Figure 15—NMDS plot ordinating bacterial profiles generated from soil collected in deciduous woodlots 2, 3, and 5. These profiles were intermingled when all woodlots were ordinated together (Figure 14), but were resolved when they were analyzed alone using NMDS.

k-Nearest Neighbor

k-NN accurately classified deciduous woodlot bacterial profiles 87.5% of the time (Table 3). All soil bacterial profiles from woodlot 8 and one profile from woodlot 9 were misclassified,

even when run in pairs with the woodlots to which they classified. When these profiles were removed from the model, 100% classification accuracy was achieved.

Analysis of Soils from Three Habitats over Time

Bacterial Abundance Charts

Class abundance of soil bacterial profiles over time within the Fenner deciduous woodlot, yard, and treated yard appeared very similar (e.g. Figure 16). Temporal fluctuations were evident in all habitats, but were not as pronounced as differences across the 10 diverse habitat profiles. Abundance charts of temporal soil collections from the deciduous woodlot and treated yard can be found in Appendix D.

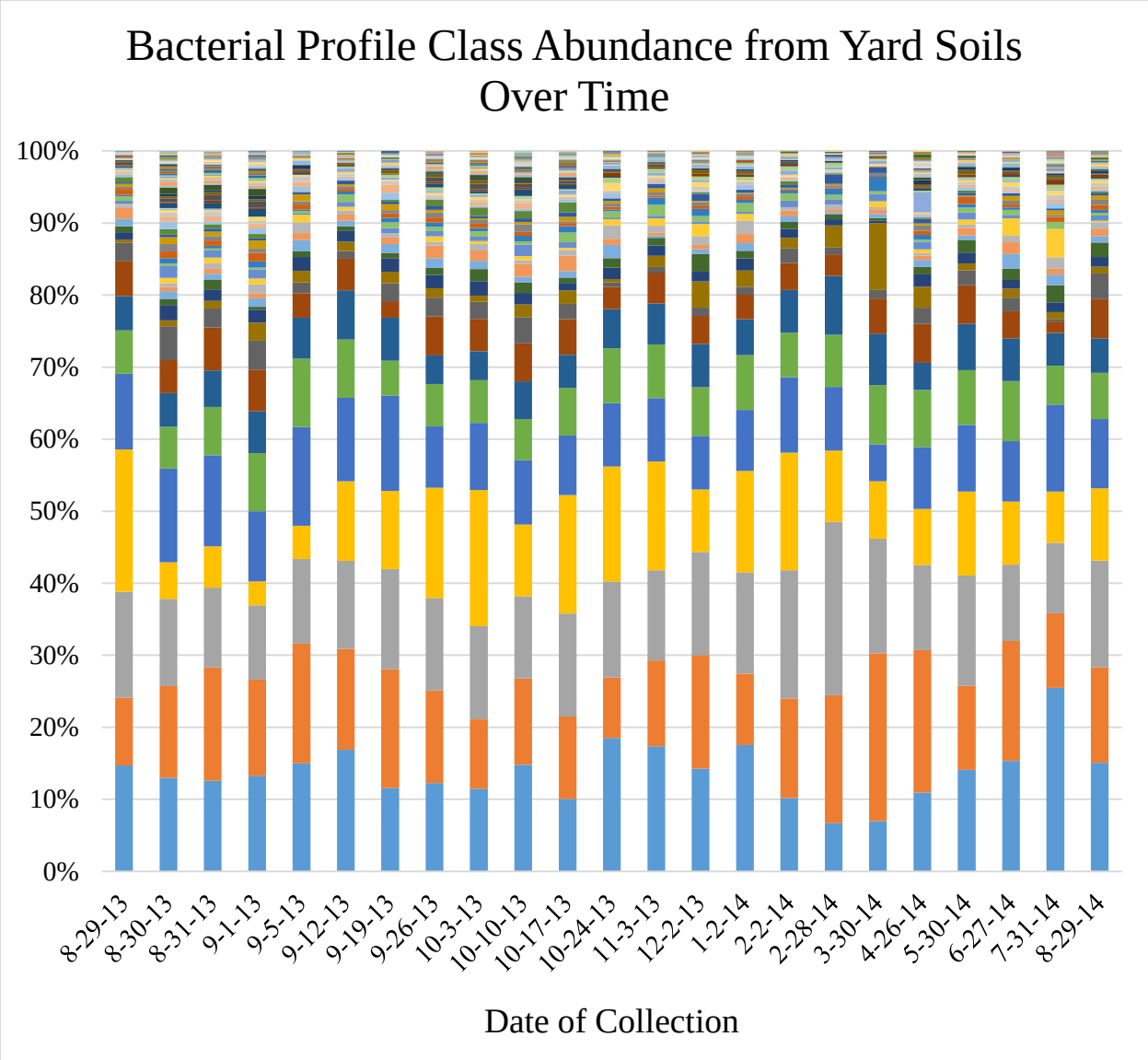

Figure 16—Bacterial class abundance of yard soils over 1 year (left to right). Soil was collected daily for 4 days, weekly for two months and monthly for the remainder of the year, so the chart is not evenly spaced in time. Slight fluctuations in abundance were evident, but soils shared the most abundant bacterial classes throughout the year.

Nonmetric Multidimensional Scaling

Soil bacterial profiles generated from soil collected over time formed clusters based on their habitat of origin in multidimensional space (Figure 17). February and March profiles from the Fenner deciduous woodlot and yard fell the farthest from the main habitat clusters. Habitat

clusters did not change substantially when February and March profiles were removed from the plot (data not shown).

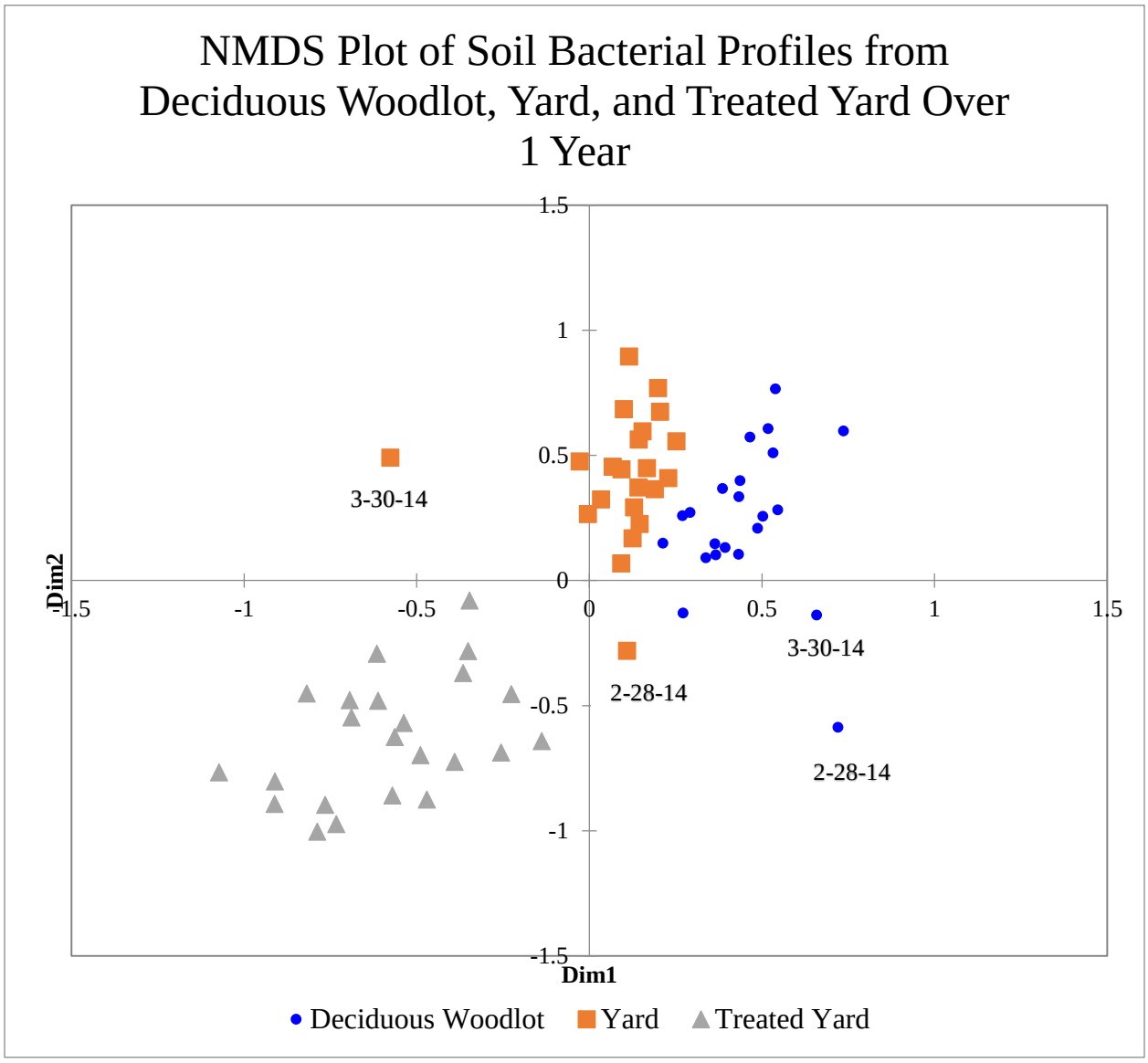

Figure 17—NMDS plot ordinating temporal soil bacterial profiles from three habitats. Profiles from each habitat formed distinct clusters. Bacterial profiles generated from Fenner deciduous woodlot and yard soils collected in late February and March fell the farthest away from their corresponding habitat cluster (labeled below point with date of collection).

k-Nearest Neighbor

k-NN accurately classified 93.8% of soil bacterial profiles to their site of origin over the full year (Table 3). Profiles from the Fenner deciduous woodlot in February and the yard in February and March were misclassified, all being assigned to the treated yard. Ten of the 38 correctly classified bacterial profiles were above the threshold value for their given habitat, six of which were generated from soil samples collected in December, January, February, or March.

Analysis of Soils from Three Habitats over Horizontal Space

Bacterial Abundance Charts

Bacterial profiles generated from soil samples collected across the surface of three habitats appeared similar, with shared taxonomic classes making up a large portion of each profile (e.g. Figure 18). Class abundance differences were evident in profiles generated from soils within a habitat, but were not pronounced. Abundance charts of horizontal soil collections from the yard and treated yard can be found in Appendix D.

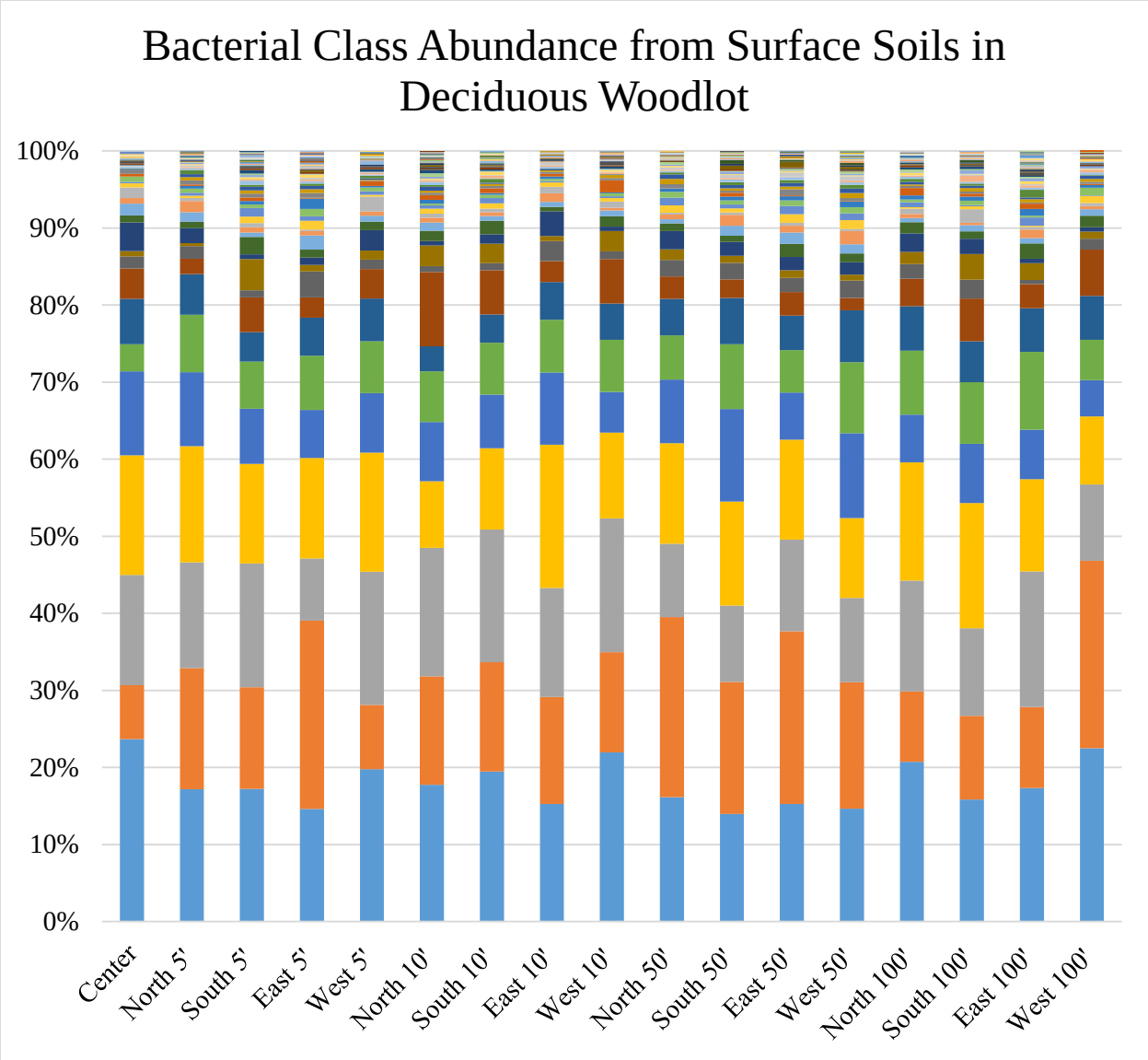

Figure 18—Bacterial class abundance of Fenner deciduous woodlot surface soils collected at a center point and 5, 10, 50, and 100 ft in the cardinal directions. Shared bacterial classes made up a large proportion of each bacterial profile, but slight differences in abundance were evident.

Nonmetric Multidimensional Scaling

Bacterial profiles generated from soil samples collected in the Fenner deciduous woodlot, yard, and treated yard clustered together based on habitat in NMDS plots (Figure 19), with profiles obtained 50 and 100 feet from the center sampling site generally being the farthest from the middle of the clusters. The bacterial profile generated from the yard soil sample collected 5

feet east of the center point plotted relatively far from the rest of the yard profiles. The treated yard cluster was completely separated from the deciduous woodlot and yard clusters, while the latter two intermingled slightly. Ordination of bacterial profiles from only the deciduous woodlot and yard did not resolve these clusters.

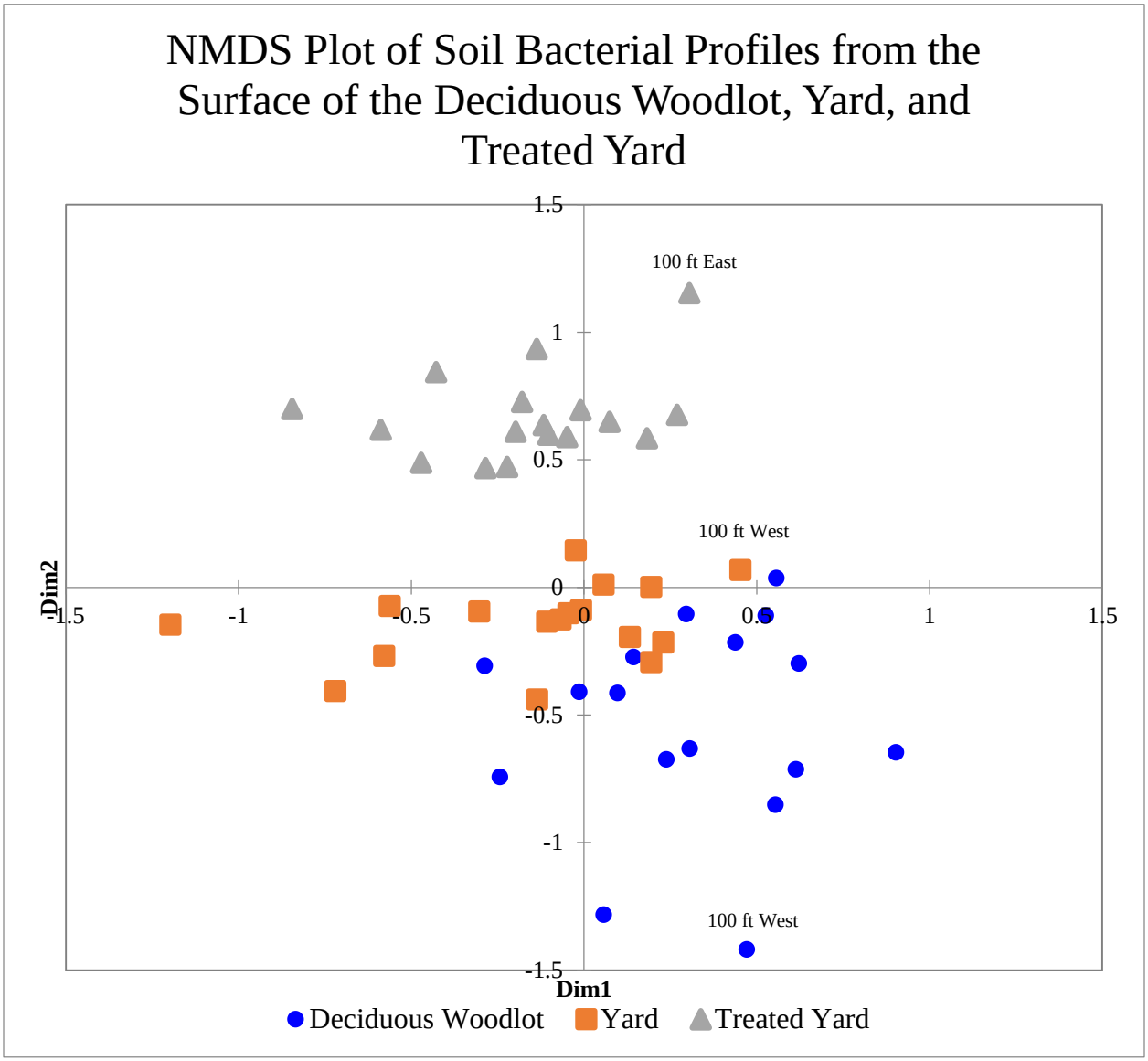

Figure 19—NMDS plot ordinating soil bacterial profiles generated from soil collected on the surface of three habitats. Profiles from each habitat formed clusters, but the deciduous woodlot and yard profiles intermingled. Profiles from soils collected the farthest from the center sampling site, plotted farther away in multidimensional space (one 100 ft profile from each habitat is labeled above the corresponding point). Additionally, the profile generated from the 5 ft east soil sample in the yard (far left square) plotted relatively far from the center of the cluster.

k-Nearest Neighbor

k-NN accurately classified bacterial profiles from across a habitat surface to their location of origin 94.4 to 97.2% of the time (Table 3), depending on the profiles used for the training set.

The most accurate classification occurred when using the center and one profile from 5, 10, 50, and 100 feet distances going in a counterclockwise spiral as the training set (four different training sets total). Each spiral training set resulted in a misclassification of a 100 feet deciduous woodlot profile to the yard and was always at least 90 feet distant from the nearest training profile. The majority (94.4%) of deciduous woodlot and yard profiles were under the threshold for both habitats using any training set.

Analysis of Soils from Three Habitats over Vertical Space

Bacterial Abundance Charts

Abundance charts generated from the vertical soil bacterial profiles revealed class differences with depth (e.g., Figure 20); however, no differences in the number of taxonomic classes (diversity) were evident (Figure 21). The most substantial class abundance differences in all habitats were higher amounts of *Clostridia*, *Nitrospira,* and *SHA-26* (Saale, Halle, library A) and lower amounts of *Spartobacteria* (denoted by arrows in Figure 20) as depth increased. Abundance charts of vertical soil collections from the yard in October and April and the treated yard in April can be found in Appendix D.

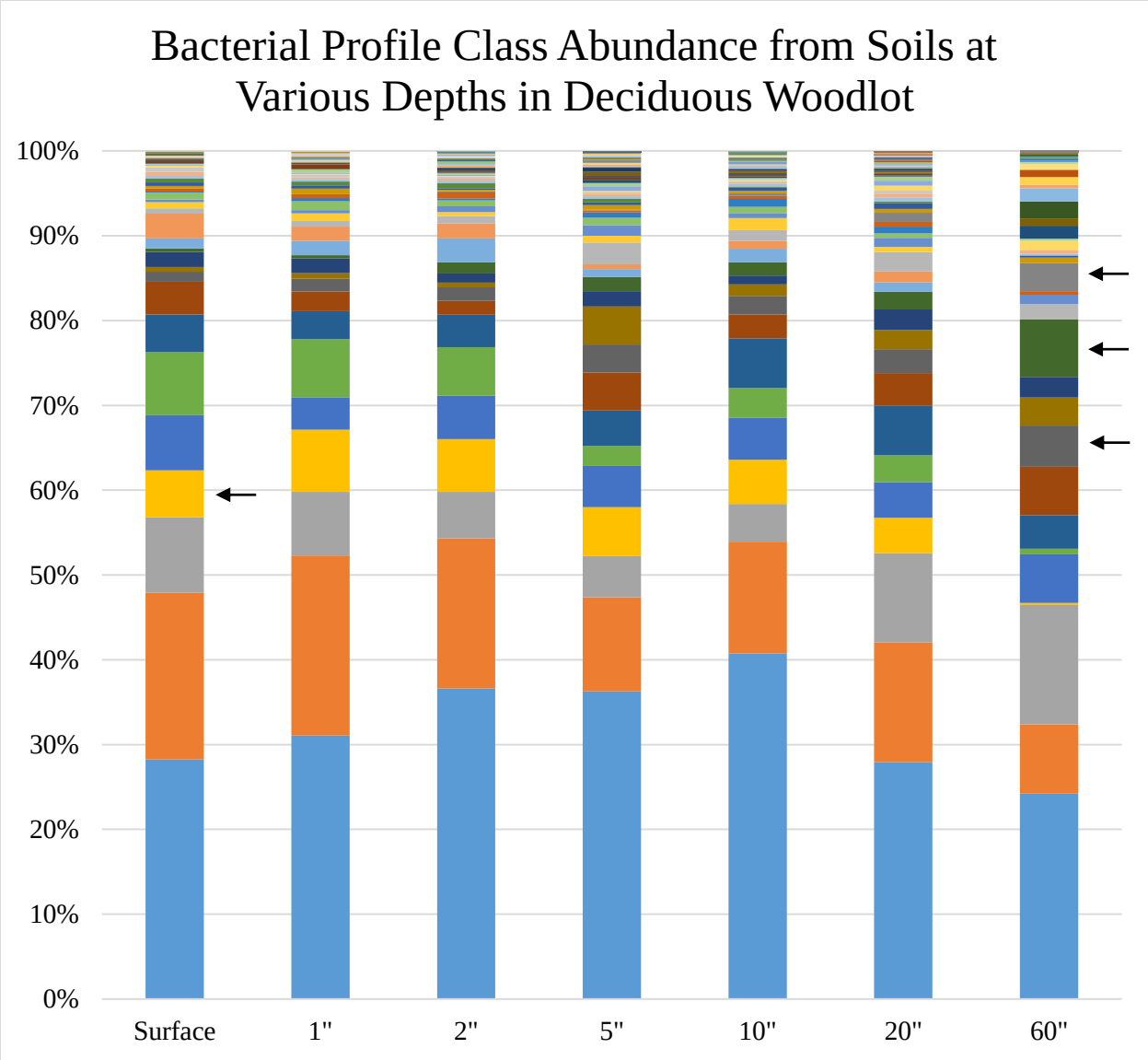

Figure 20—Bacterial class abundance of deciduous woodlot depth soils in October. As depth increased, substantial increases in *Clostridia*, *Nitrospira*, and *SHA-26* (denoted by arrows in ascending order on the right) and decreases in *Spartobacteria* (denoted by left arrow) existed in all habitats.

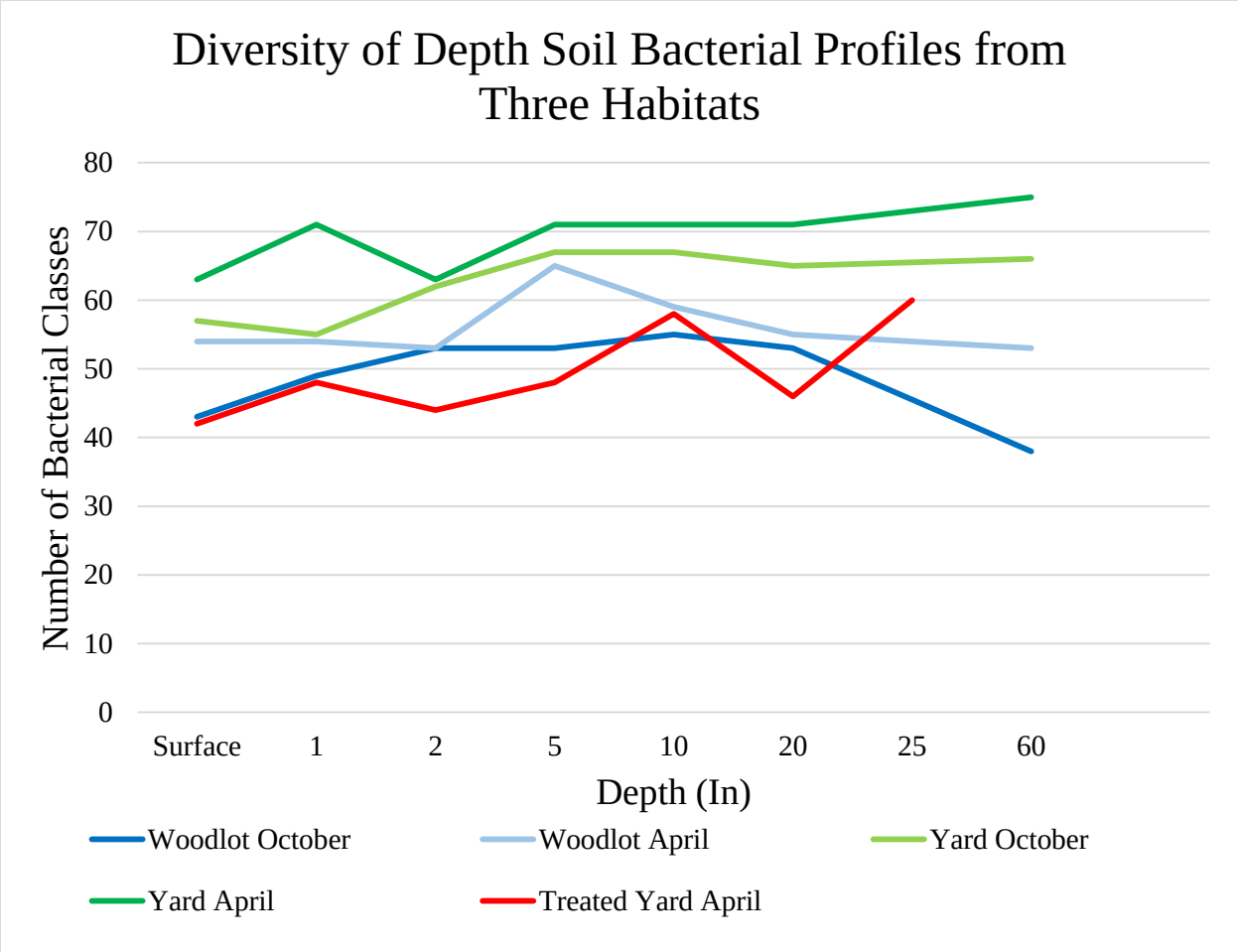

Figure 21—Levels of bacterial taxonomic class diversity in soils collected at various depths within the deciduous woodlot, yard, and treated yard. No diversity trends were evident across or within habitats.

Nonmetric Multidimensional Scaling

The treated yard depth profiles clustered separately in NMDS plots (Figure 22), while the Fenner deciduous woodlot and yard depth profiles intermingled. When deciduous woodlot and yard profiles were ordinated as a pair (Figure 23), they did not intermingle. Within all habitats, a trend existed wherein the soil bacterial profiles moved farther away from the surface profile in multidimensional space as depth increased. NMDS plots of bacterial profiles generated from soil

collected in the deciduous woodlot and yard in October and plots ordinating the same habitat in each sampling month can be found in Appendix E.

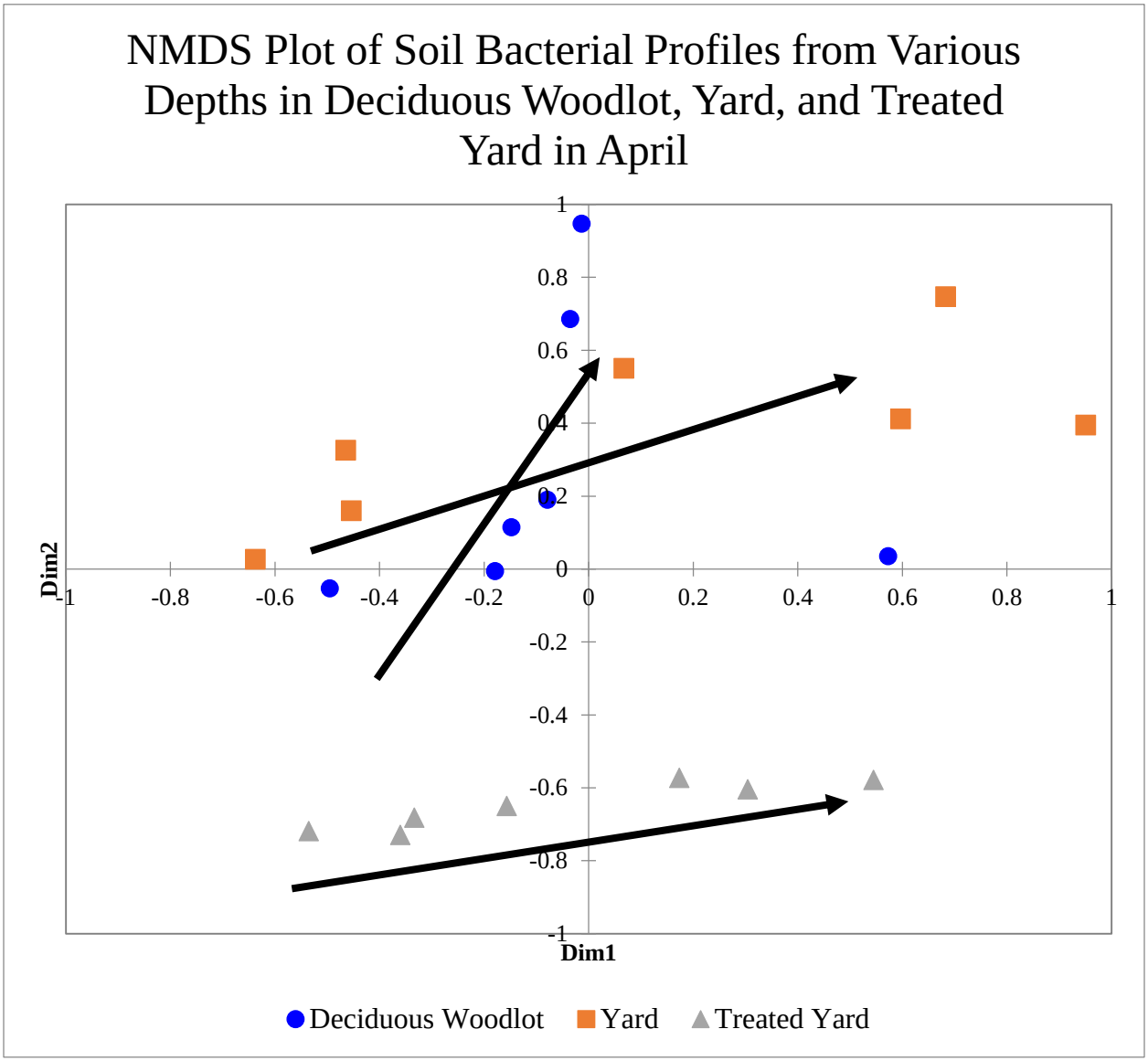

Figure 22—NMDS plot ordinating soil bacterial profiles generated from soil collected at various depths within three habitats in April. The treated yard profiles clustered separately, while the deciduous woodlot and yard profiles intermingled. A trend existed across all habitats, with the soil bacterial profiles moving away from the surface profile in multidimensional space as depth increased (arrows point in the direction of increasing depth).

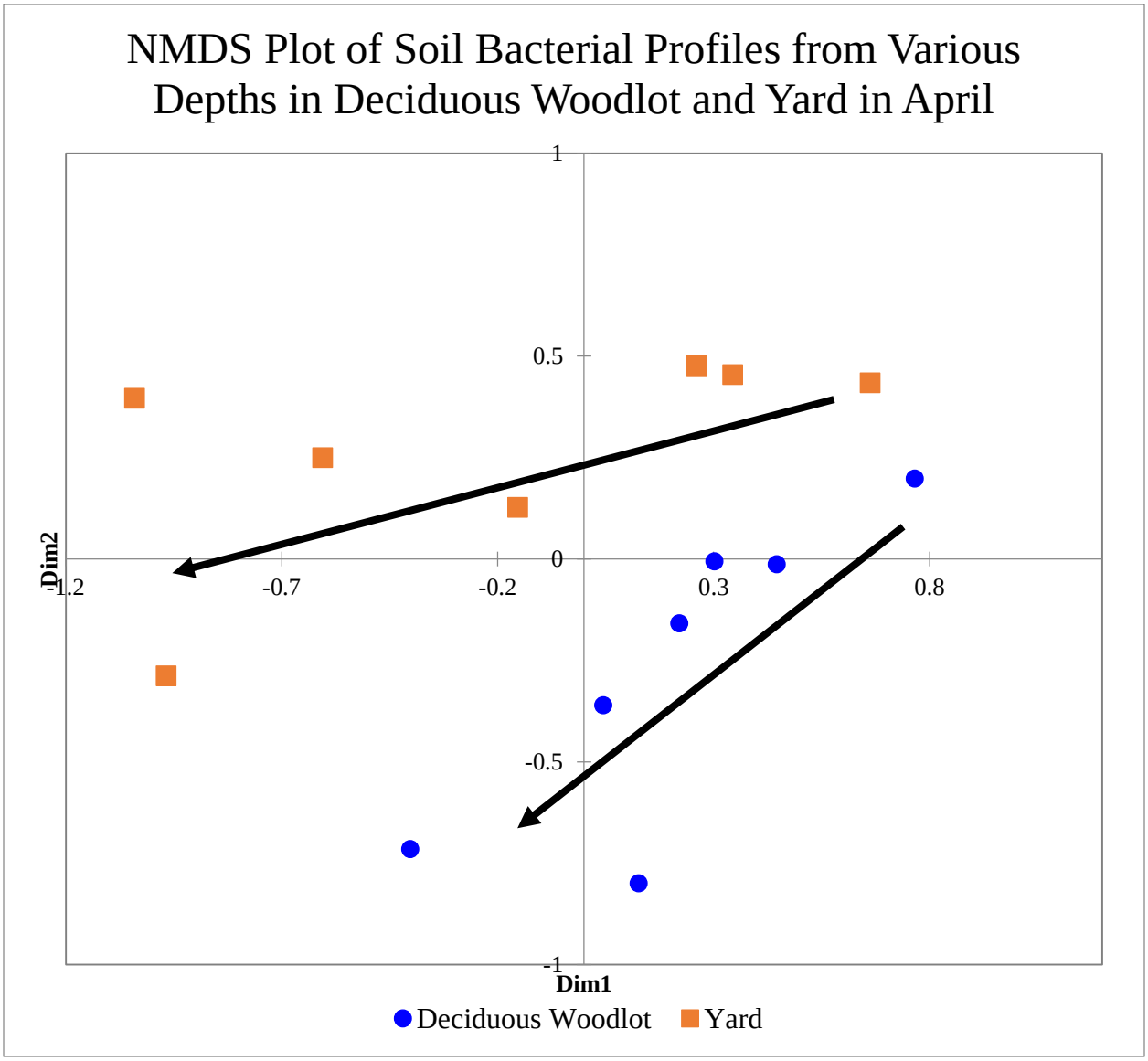

Figure 23—NMDS plot ordinating deciduous woodlot and yard depth profiles in April. Although intermingled when plotted with the treated yard profiles (Figure 22), deciduous woodlot and yard clusters separated when ordinated as a pair. Again, plots reflected the trend of soil bacterial profiles moving away from the surface profile in multidimensional space as depth increased (arrows point in the direction of increasing depth[4]).

[4] General orientation of bacterial profiles in NMDS plots is random, thus, the different direction of arrows across plots is not analytically relevant.

k-Nearest Neighbor

Bacterial profiles generated from soils collected at different depths were accurately classified 80% of the time when the shallowest five profiles made up the training set (Table 3). The only misclassifications were the 60 inch deciduous woodlot bacterial profiles in both months. Soil profiles were accurately classified (Table 3) and fell under the threshold values (Table 4) when the surface, 2, 10, and 60 inch profiles were used as the training set in *k*-NN.

Analysis of Preliminary Evidentiary Soils

Bacterial Abundance Charts

Bacterial profiles generated from the various evidence types exhibited abundance changes over time (e.g., Figure 24). Notably consistent change across all evidence types included an increase in *Actinobacteria* and *Bacilli* and a decrease in *Acidobacteria, Sphingobacteria, Betaproteobacteria,* and *Spartobacteria.* More change occurred within the first 6 months of storage, while less change happened between the 6 month and 1 year collections. Abundance charts of soil bacterial profiles from the other evidence types can be found in Appendix D. There were fewer bacterial classes present, on average, from the evidence items that had been stored for 6 months (41) than the deciduous woodlot of origin (56), however the average class diversity did not decrease after the full year (43).

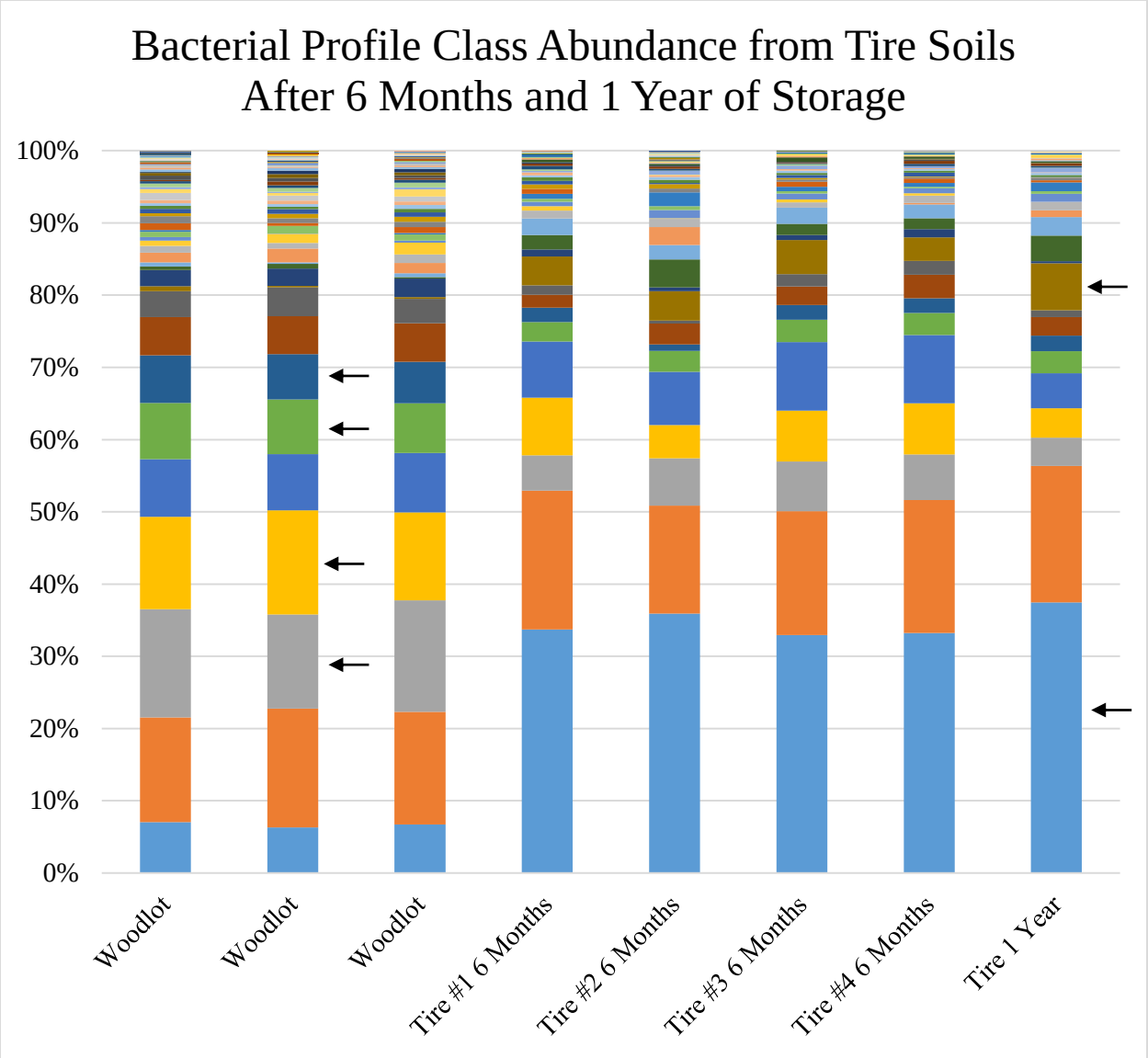

Figure 24—Bacterial class abundance of three replicate soil collections from the deciduous woodlot of origin (left) and soil collections from the tire after 6 months and 1 year of storage at room temperature. Evidentiary profiles exhibited an increase in *Actinobacteria* and *Bacilli* (denoted by arrows in ascending order on the right of the figure) and a decrease in *Acidobacteria*, *Sphingobacteria*, *Betaproteobacteria*, and *Spartobacteria* (denoted by arrows in ascending order on the left of the figure).

Nonmetric Multidimensional Scaling

Bacterial profiles generated from evidentiary item soil after 6 months and 1 year of storage clustered together in multidimensional space, away from all deciduous woodlots, but in

closest proximity to the woodlot of origin (Figure 25). Soil bacterial profiles generated from soils collected after 1 year of storage were slightly more distant from the woodlot of origin profiles than were those collected after 6 months.

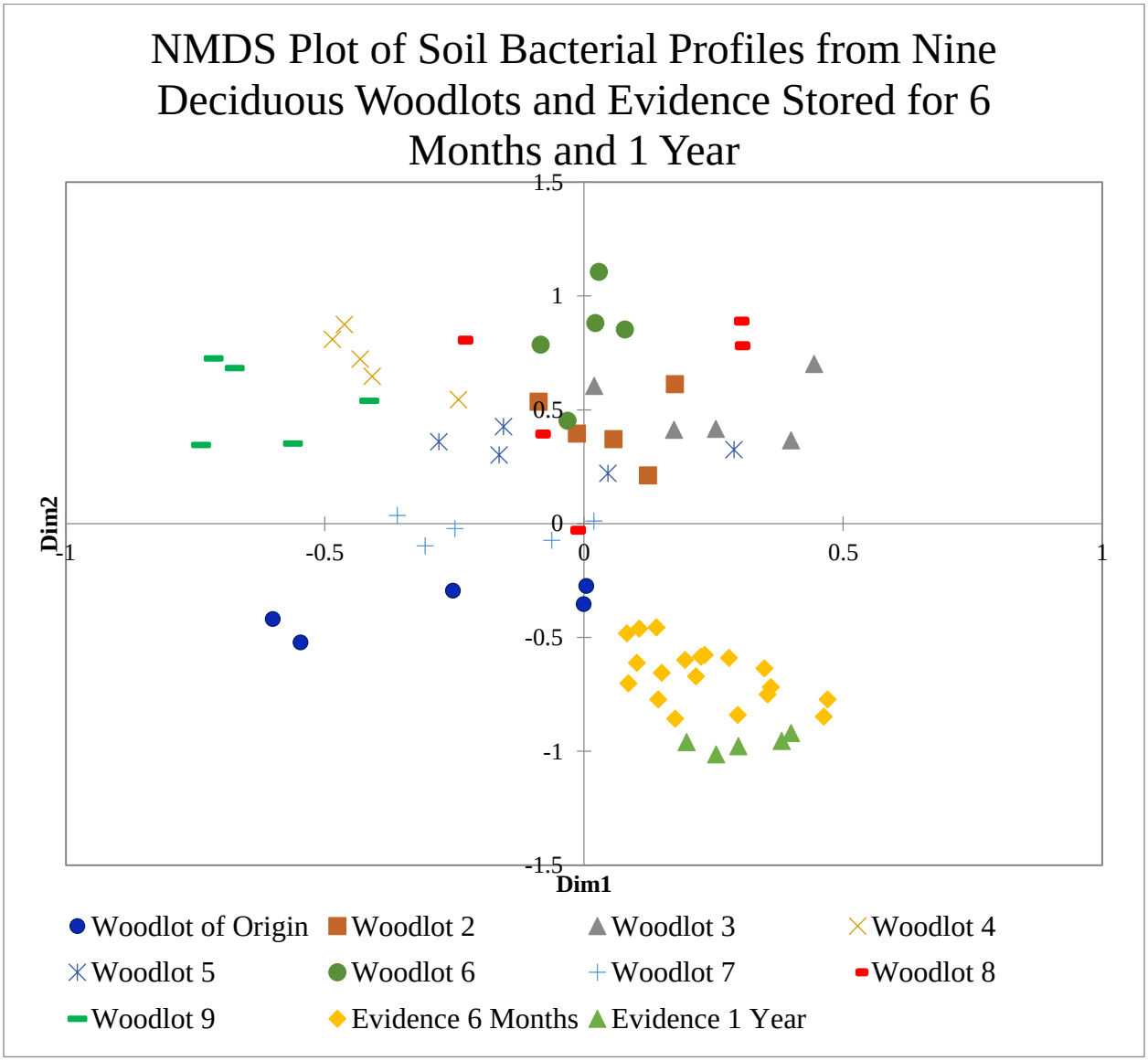

Figure 25—NMDS plot ordinating evidentiary and deciduous woodlot soil bacterial profiles. Evidence profiles after both 6 months and 1 year in storage clustered together, nearest the woodlot of origin, with the 1-year profiles plotting slightly farther away.

k-Nearest Neighbor

k-NN accurately classified soil bacterial profiles from evidentiary items to their location of origin 100% of the time after 6 months and 1 year of storage; however, no profiles were under the threshold value (Table 4).

Analysis of T-Shirt Evidentiary Soils

Bacterial Abundance Charts

Bacterial profile abundance changes were evident in soil collections from T-shirts over the 4-month period (e.g., Figures 26 and 27), regardless of storage temperature. The same changes that occurred on the other evidentiary items occurred on the T-shirts: an increase in *Actinobacteria* (Figure 28) and *Bacilli* and a decrease in *Acidobacteria, Sphingobacteria* (Figure 29), *Betaproteobacteria,* and *Spartobacteria.* These abundance changes began more slowly in the soils collected from T-shirts stored at 4°C.

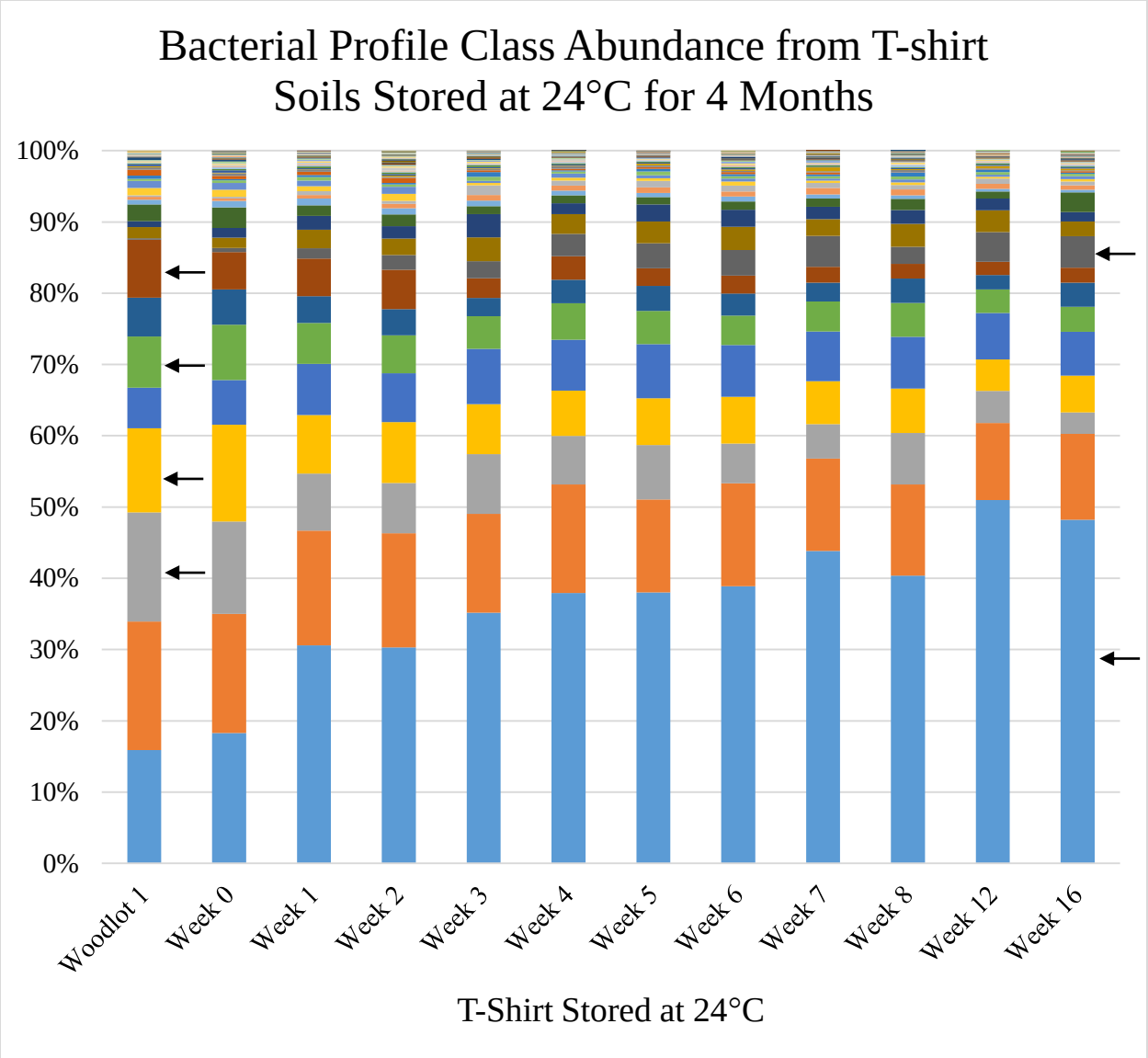

Figure 26—Bacterial class abundance of 24°C T-shirt soil collections over the 4-month sampling period compared to soil collected from the deciduous woodlot of origin. Evidentiary soil profiles exhibited increases in *Actinobacteria* and *Bacilli* (denoted by arrows in ascending order on the right of the figure) and decreases in *Acidobacteria, Sphingobacteria, Betaproteobacteria,* and *Spartobacteria* (denoted by arrows in ascending order on the left of the figure).

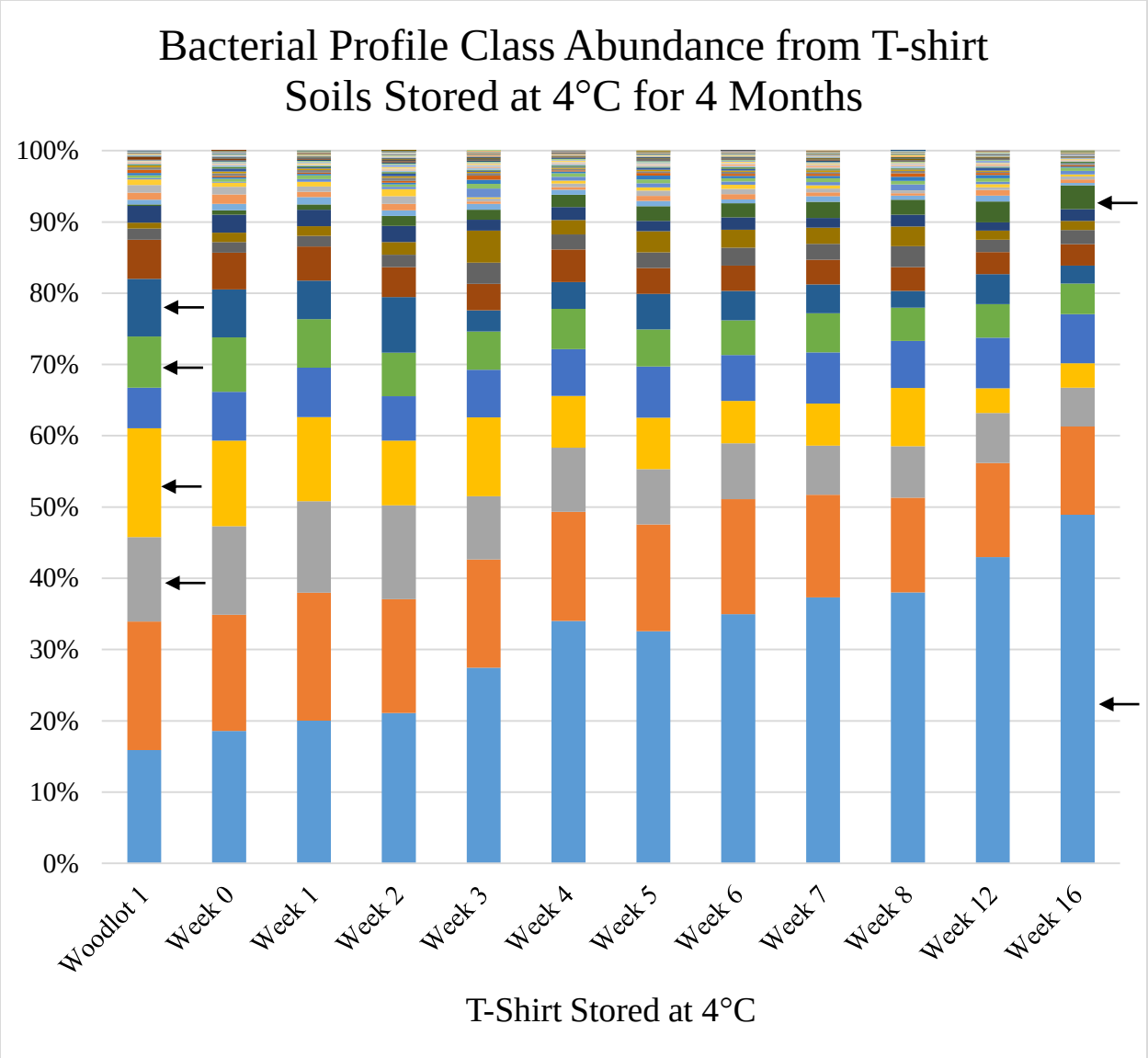

Figure 27—Bacterial class abundance of T-shirt soil profiles stored at 4°C and collected over the 4-month sampling period compared to one profile generated from deciduous woodlot of origin soil. Evidentiary soils exhibited notable increases in *Actinobacteria* and *Bacilli* (denoted by arrows in ascending order on the right of the figure) and decreases in *Sphingobacteria, Acidobacteria, Betaproteobacteria,* and *Spartobacteria* (denoted by arrows in ascending order on the left of the figure).

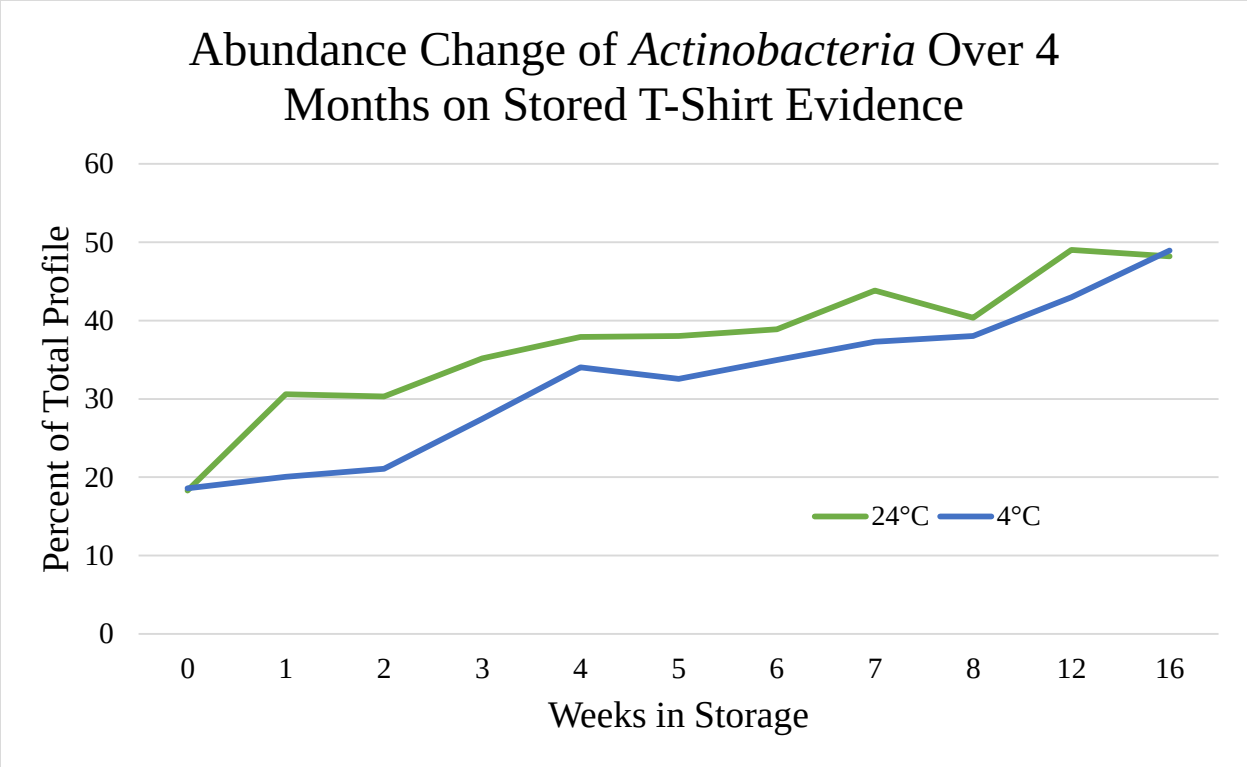

Figure 28—Average (n=4) *Actinobacteria* abundance in bacterial profiles generated from soil on T-shirts stored at 24°C and 4°C over a 4-month period. Members of this class increased in abundance over time in storage.

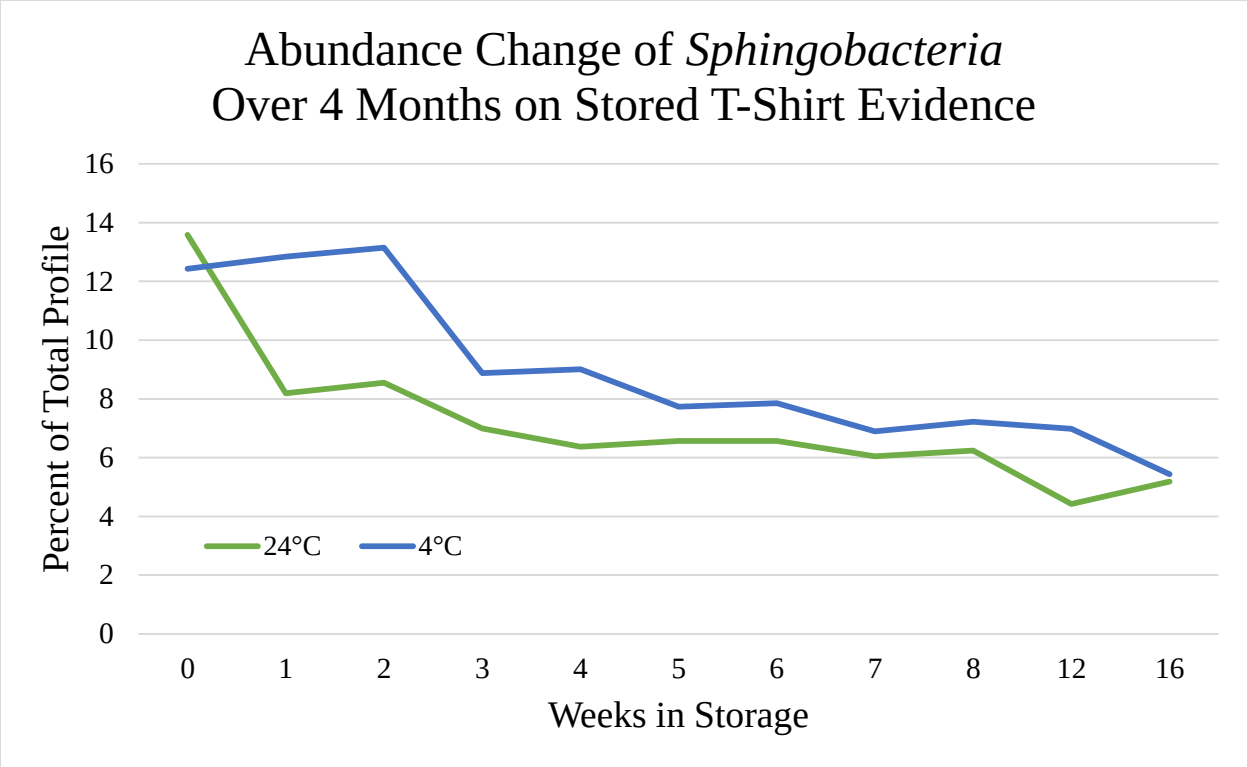

Figure 29—Average (n=4) *Sphingobacteria* abundance in bacterial profiles generated from soil on T-shirts stored at 24°C and 4°C over a 4-month period. Members of this class decreased in abundance over time in storage.

Additionally, the average taxonomic class diversity from evidentiary items (48) was initially slightly lower than the diversity of deciduous woodlot of origin soils (56); however, this diversity did not decrease substantially over the 4-month storage period (Figure 30).

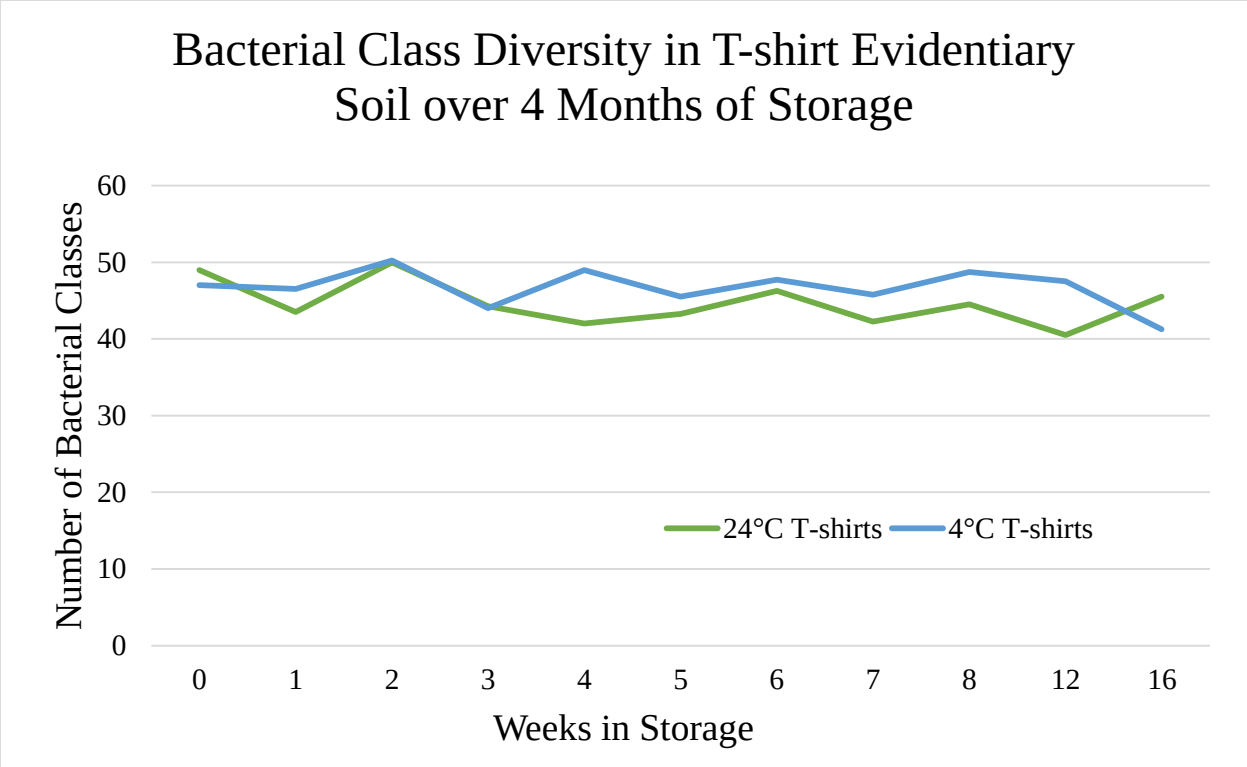

Figure 30—Number of bacterial classes in soils on evidentiary T-shirts over a 4-month storage period. T-shirt bacterial profiles had lower diversity than the deciduous woodlot of origin (which had an average of 56 bacterial classes); however, the diversity did not decrease markedly over the storage period in either temperature.

Nonmetric Multidimensional Scaling

T-shirt soil bacterial profiles initially clustered together near their deciduous woodlot of origin (Figure 31) and began to drift away from all woodlots in multidimensional space over the 4-month period (e.g., Figure 32 and Figure 33). An NMDS plot of all soil bacterial profiles generated from the T-shirts and nine deciduous woodlots can be found in Appendix E.

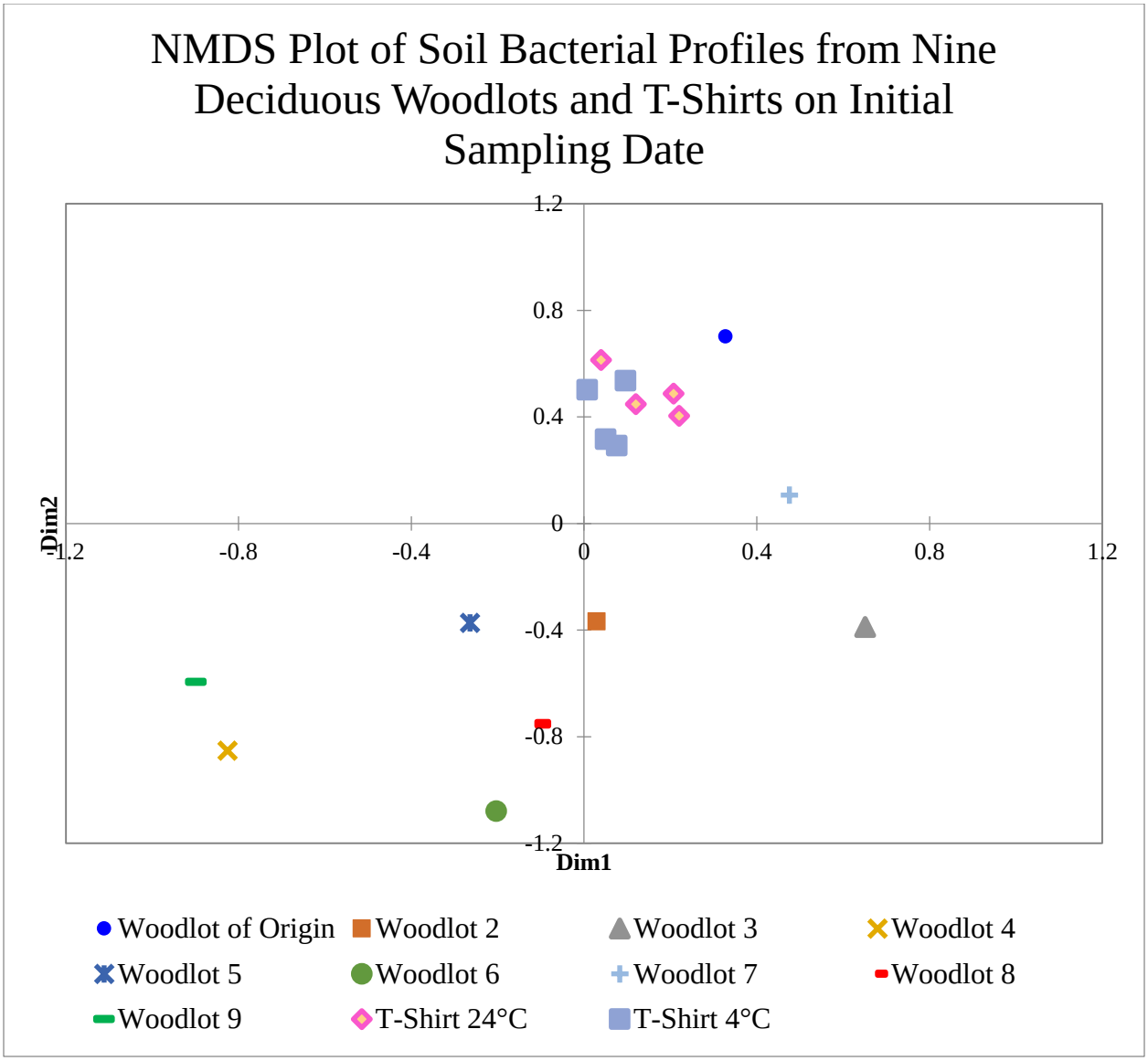

Figure 31—NMDS plot ordinating initial deciduous woodlot and T-shirt soil bacterial profiles. Evidentiary soil profiles clustered together nearest the deciduous woodlot of origin profile.

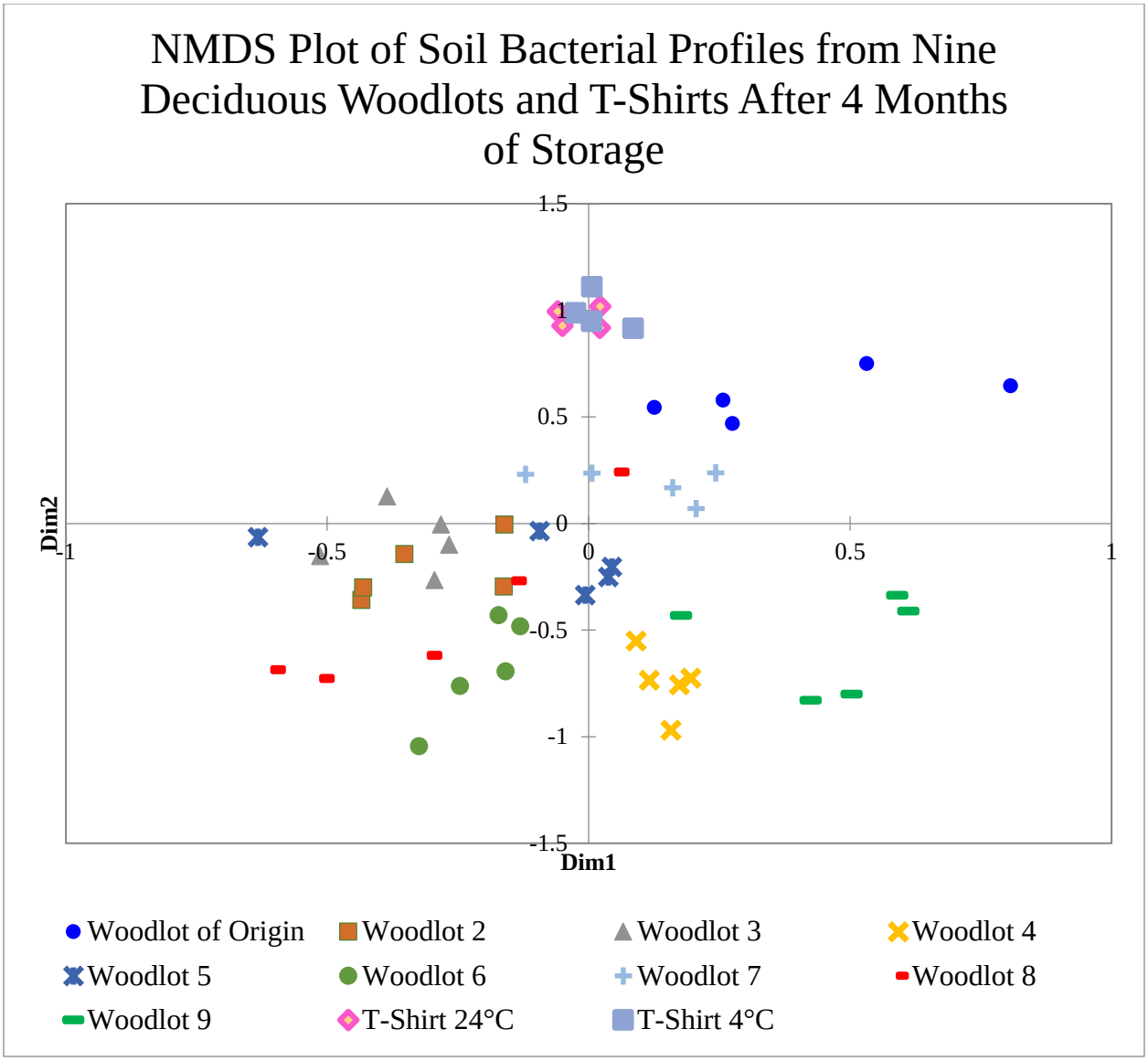

Figure 32—NMDS plot ordinating nine deciduous woodlots and T-shirt evidentiary bacterial profiles after 4 months of storage. Profiles generated from T-shirts kept at both storage temperatures clustered away from all woodlot profiles, remaining closest to the woodlot of origin cluster.

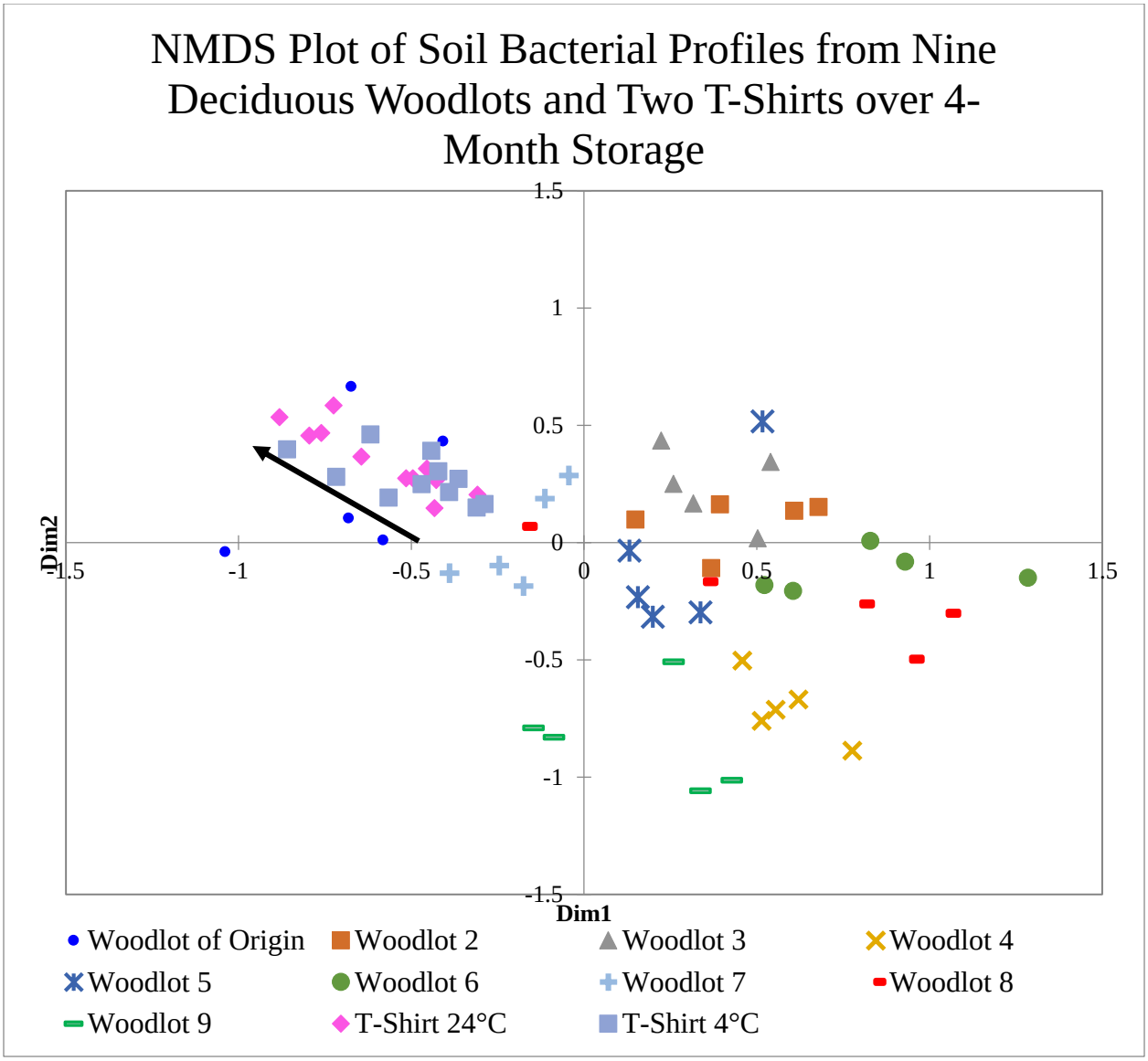

Figure 33—NMDS plot ordinating nine deciduous woodlots and soil bacterial profiles generated from one T-shirt at each storage temperature over 4 months. T-shirt soil profiles clustered together near the woodlot of origin cluster. Profiles drifted away from all woodlot profiles over time (in the direction of the arrow).

k-Nearest Neighbor

k-NN accurately classified soil bacterial profiles from T-shirt evidence to their deciduous woodlot of origin 100% of the time over the 4-month period (Table 3). Profiles from the four T-shirts stored at 4°C were under the *k*-NN threshold for the woodlot of origin more often than 24°C T-shirts (Figure 34); however, both fluctuated over the 4-month storage period. All eight

T-shirt profiles were under the threshold on the initial exposure date and after 1 week of storage. Week 3 stood out, as only one T-shirt profile from each storage temperature was under the threshold value for the woodlot of origin.

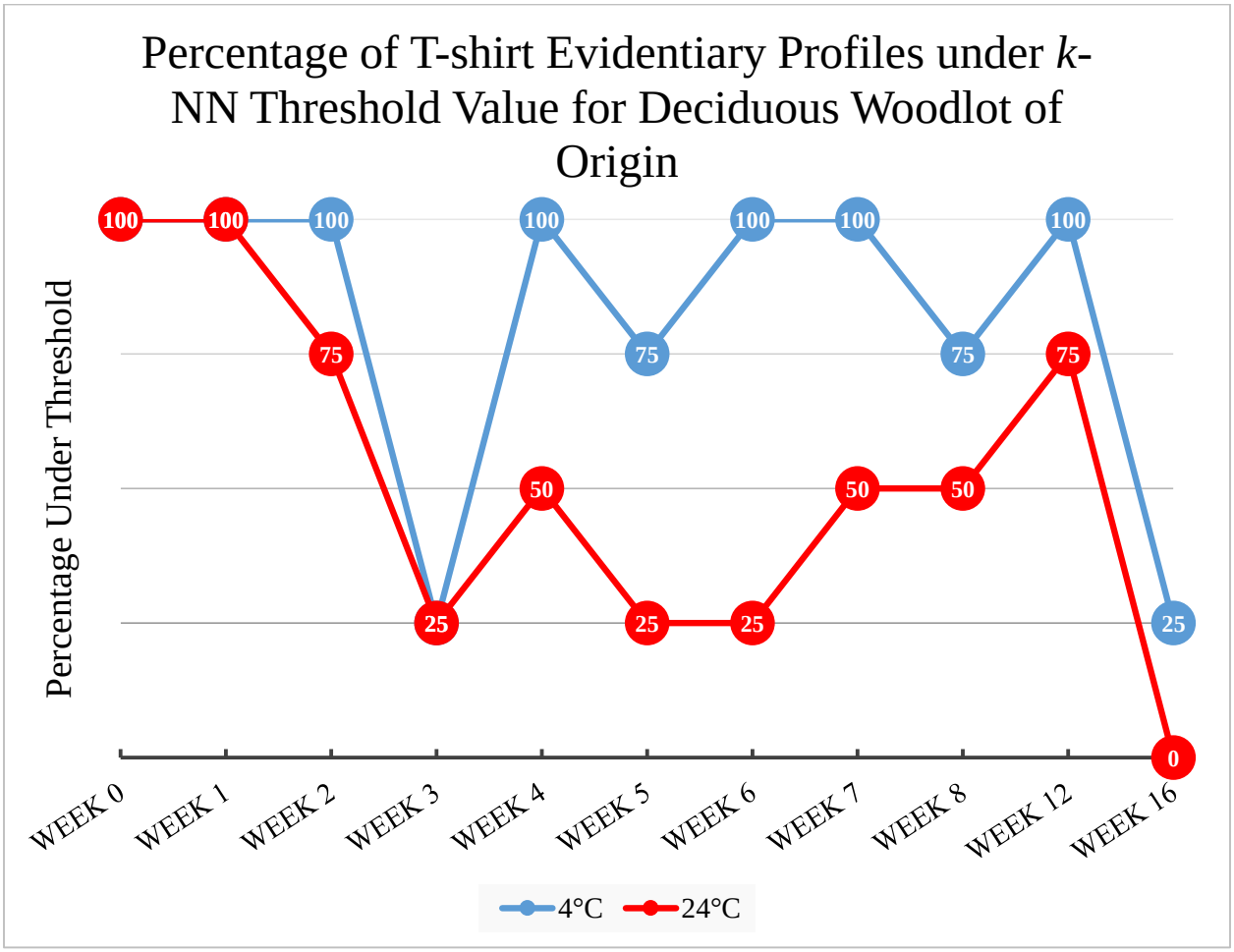

Figure 34—T-shirt evidentiary bacterial profiles under the threshold value for the deciduous woodlot of origin over a 4-month storage period at either 4°C (n=4) or 24°C (n=4). All T-shirt profiles were under the threshold initially and after 1 week of storage, but fluctuated for the rest of the sampling times. Soil profiles from T-shirts stored at 4°C were under the *k*-NN threshold value more often than profiles from T-shirts stored at 24°C.

DISCUSSION

The transfer of soil to an item or individual during the course of a crime offers the potential to link a suspect or victim to the scene. In these situations, forensic scientists aim to compare two or more soil samples and ultimately determine whether they originated from the same location. Traditional forensic soil analysis techniques involve the examination of class characteristics, which can be time consuming and often do not result in definitive association. The adoption of a more individualizing technique from another soil science discipline has the potential to strengthen forensic soil analysis. Microbiologists have been developing techniques to compare soil samples for decades (reviewed by van Elsas and Boersma, 2011) via the generation of soil bacterial profiles. Forensic comparison of such profiles may provide stronger discrimination or association than traditional soil analysis techniques due to the diversity and complexity of bacterial communities within soil. However, adoption and implementation of microbiological profiling methods into crime laboratories requires multiple steps, the first of which is verification that the technique will meet the needs of forensic soil analysis. Sensabaugh (2009) described three conditions that any microbial technique must satisfy to gain footing in forensic science—differentiability, reproducibility, and objectivity—all of which were generally satisfied through the research presented in this thesis. Next-generation sequencing of the 16S rRNA gene in soil bacteria allowed for the differentiation of soils from diverse and similar habitats over both time and space. Additionally, the combination of abundance charts, NMDS plots, and *k*-NN offered visual and statistical comparisons of soil bacterial profiles.

Illumina next-generation sequencing both surpasses the resolution of older microbial profile generation methods (e.g., T-RFLP [Cao et al., 2013]) and produces a greater number of sequence reads than other next-generation platforms (e.g., pyrosequencing [Will et al., 2010;

Sato et al., 2013; Hopkins, 2014]). Larger amounts of data provide both advantages and disadvantages to forensic soil analysis. High resolution sequence data allow for the identification of subtle variances (e.g., rare or low abundance bacteria), which can help differentiate soil from two or more very similar locations. The same differentiability can be detrimental in profile comparisons however, as slight bacterial variation within an area can result in higher dissimilarity between samples and potentially a false exclusion. This did not seem to be a problem in the research presented here, as bacterial profiles from the same location generally clustered or classified together. Large datasets also pose a problem for forensic soil analysis, requiring powerful computers to process and analyze sequences that may not be available in forensic laboratories. Subsampling reduces the amount of data to a more manageable level; however, this process results in data loss, potentially affecting association of profiles. Replicate subsampling performed using the diverse habitat samples in this research produced bacterial profiles that showed similar taxonomic class abundance and exhibited consistent clustering patterns in NMDS plots (data not shown), indicating that subsampling did not affect profile association. Given this, it was concluded that subsampling profiles to 3000 sequences (which had been necessary for all studies due to computational analysis capabilities) was adequate to accurately reflect the bacteria present in a soil sample, and the potential loss of rare sequences did not greatly affect the discrimination or association among bacterial profiles.

The next-generation sequencing data analysis methods utilized in this research allowed for the comparison of bacterial profiles both visually at the taxonomic class level and statistically at the sequence level; proving beneficial when used in combination. Abundance charts offered a clear visualization of taxonomic classes present within each bacterial profile, along with differences in their quantity and taxonomic diversity. Almost all of the soils contained the same

most abundant bacterial classes, although no two charts were identical. Bacterial abundance charts from diverse habitats were less alike than those from similar habitats, which in turn were less alike than those from soils within a habitat. An exception to this was the vertical space soils, which had clear class differences as depth increased (e.g., increasing levels of *Clostridia, Nitrospira,* and *SHA-26* and decreasing levels of *Spartobacteria).* Environmental factors in deep soils such as lower oxygen and nitrogen levels (Hinchee and Leeson, 1997; Schramm et al., 1999) offer a possible explanation for why *Clostridia* and *Nitrospira* were more abundant in deeper samples from all habitats, as their members thrive under such conditions (O'Brien and Morris, 1971). *SHA-26* has not been as thoroughly studied; however, members of this class have been found in deeper soils by Tsitko and Bomberg (2014), indicating its species can also thrive in the conditions below topsoil.

The dirt road soil samples represent another interesting example of bacterial variation detectable in abundance charts. Soils from this habitat exhibited differences in specific bacterial classes compared to other habitats, as well as lower class diversity. Upon further investigation, this likely resulted from treatment of the road with calcium chloride to reduce dust levels (Shiawassee County Road Commission, personal communication, 2015). Such chemical treatment increases soil salinity, which has been shown to lower the overall number of bacteria (Hollister et al., 2010), while favoring halophilic species (Quesada et al., 1983; Amoozegar et al., 2005), many of which exist in the bacterial classes that were unusually abundant (*Clostridia, Bacilli*, *Flavobacteria*, and *Gammaproteobacteria* [Oren, 1983; Ventosa et al., 1998; Albuquerque et al., 2008; Sorokin et al., 2010]).

Bacterial abundance charts also have forensic value beyond the identification of extremely different profiles. The relative abundance of taxonomic groups within known bacterial

profiles could be used to develop confidence intervals for a given location. Evidentiary profile relative abundance would then be compared to known profile confidence intervals and classified as being consistent or not. For example, the average ratio of *Actinobacteria* to *Spartobacteria* could be determined for a set of known profiles, measuring the variation within that location. An evidentiary sample with a very similar *Actinobacteria-Spartobacteria* ratio that falls within the confidence interval would be considered consistent with that location. The combination of many such ratio comparisons would increase the assurance of evidentiary and known profile association. Ratio calculations have been used in past soil microbiological research, but have not been studied for forensic application. Bossio et al. (1998) used phospholipid fatty acid profiles to calculate the ratio of fungal to bacterial biomass and determine how fertilizer affects microbial communities. Thomson et al. (2010) examined whether the removal of vegetation on agricultural land influenced bacteria based on relative T-RFLP peak height at specific restriction sites within the 16S gene. Ratio comparisons similar to these past studies, but based on the relative abundance of various bacterial groups (at any taxonomic level), would be especially helpful when analyzing forensic samples from very similar locations, where differences may not be readily visualized through abundance charts. However, such methods may be attacked in court, especially if associations were based on ratios involving rare bacterial classes. For example, it could be argued that two soils are inconsistent because the known profiles have a 1:500 ratio of a rare class to a common class, but the rare class is not present in the single evidentiary profile. Such differences may result from natural bacterial variation within a location, from subsampling, or from other sources of variation, such as long storage periods or the environmental conditions of storage (discussed below). Therefore, it may be more accurate to calculate ratios among only common groups at the respective taxonomic level to ensure slight bacterial variation does not

influence association or discrimination. Further research is necessary to determine which taxonomic level and bacterial groups provide the most discriminatory information; ideally differing among habitats but not within a single location.

NMDS can help differentiate bacterial profiles that appear very similar in abundance charts, and it too generates a visualization of the data, potentially providing the expert witness an easily explainable concept when presenting results in court. NMDS reflected differences apparent in bacterial abundance charts in this thesis research (e.g., the dirt road profiles plotted the farthest from other habitats, and the depth profiles plotted progressively farther from the surface profile), while also allowing for differentiation of profiles that were visually similar in abundance charts. However, plots ordinating many locations together had relatively high stress, and profile clusters often intermingled. Several locations may be in question as the source of an evidentiary soil sample in a criminal investigation, and such intermingling of known bacterial profiles could prevent the association of an evidentiary profile and a single location. Stress is typically lower when fewer samples are ordinated in NMDS plots (Holland, 2008), resulting in a better depiction of sample dissimilarity and, in a forensic scenario, a more accurate representation of which location an evidentiary soil profile is most similar to. Ordination of profiles from pairs or triads of locations in this research resulted in the resolution of intermingled clusters, allowing for discrimination of the most similar bacterial profiles from different locations in both the diverse and similar habitat studies. It should be noted that the intermingled profiles were not removed in this exercise, they were simply analyzed without the pressure of other, more dissimilar profiles forcing them together. The clustering characteristics of NMDS plots can also be used to exclude locations as a possible source of an evidentiary soil sample. The relative position of evidentiary soil profiles in a plot allows the user to identify the most

dissimilar known clusters. Distant clusters can then be excluded as a site of origin, reducing the number of potential soil transfer locations to be compared with more definitive analysis methods. However, while these benefits make NMDS a powerful analysis technique, the evaluation of stress and the identification of clusters is still somewhat subjective. It has been suggested that stress values over 0.2 should be interpreted with caution, and that lower stress is always a better representation of profile relationships (Clarke, 1993), but a universally accepted stress value is not defined in the microbiology literature. The individual interpreting an NMDS plot is also responsible for identifying clusters, as there is no defined proximity measure for whether profiles are close enough to be considered associated. However, once identified, clusters of profiles can be compared using statistical tests such as analysis of similarities (Clarke, 1993) or multivariate analysis of variance (Johnson and Wichern, 2002), which provide a p-value and greater discriminatory power than the visual interpretation of an NMDS plot alone. Although helpful for statistically discriminating profiles from two locations, such methods probably have limited forensic value, as they require multiple profiles within each group being compared, and there may only exist one evidentiary sample in a criminal investigation. Due to its subjectivity and inadequate statistical power, NMDS cannot stand alone as a soil bacterial profile analysis technique in forensic science, if the goal is to definitively determine where a soil sample originated.

Supervised classification techniques provide the objectivity necessary for forensic science, allowing for definitive assignment of a bacterial profile to a location of origin. In a forensic setting, training sets will be made up of known soil bacterial profiles representative of two or more potential locations of origin. Misclassification of a known profile can be identified in the training set validation step, and can be removed from analysis, or profiles from fewer

locations can be analyzed in a process analogous to what was done by analyzing pairs or triads of samples with NMDS. However, the former option is not ideal, as failure to classify in the training set validation could be indicative of variation within a location, and removal of the profile would make the training set less representative of the location of origin, and possibly result in a false exclusion of an evidentiary profile from a crime scene. Additionally, removal of profiles may be seen as data manipulation in court, potentially lowering the reputability of soil analysis results. The misclassification of marsh and fallow agricultural field profiles when all habitats were analyzed together in this research provides one example of how the comparison of fewer profiles at once, rather than the removal of profiles, resulted in accurate classification. Although there is no way to be sure what an unknown's neighbors were in *k*-NN, it is likely some nearest neighbors of the misclassified profiles were from a third location, reducing the number that could be divided between the habitats with which they were most similar. This represents an instance where NMDS could be used together with supervised classification to visualize potential nearest neighbors and determine if a single known profile is plotting near an evidentiary profile and skewing classification. Marsh profiles in this research clustered near both fallow agricultural and agricultural field profiles in NMDS plots, suggesting nearest neighbors were being divided among those three locations, forcing classification based on a smaller nearest neighbor majority. Analysis with training sets made up of only marsh and fallow agricultural field knowns allowed all nearest neighbors to be decided between them, providing classification to the most similar location without losing neighbors to less similar groups. Supervised classification may be better suited for forensic science if knowns from pairs of locations are compared to evidentiary profiles in this manner. In the event that several potential locations are involved in a criminal investigation, all pair combinations could be analyzed with the evidentiary

profile, narrowing down the location with which an unknown sample is most similar, while ensuring all nearest neighbors are decided between only two locations at once.

A training group assignment does not hold much forensic value without the interpretation of threshold values, as *k*-NN will force classification even if an unknown belongs to no training group. Although questioned as a statistically valid measure (Lavine and Davison, 2006), threshold values allow for an assessment of how well soil profiles are classifying, adding another layer of confidence. In the current research, dissimilarity of bacterial profiles in abundance charts or NMDS plots was usually reflected in *k*-NN threshold values. For example, evidentiary soil bacterial profiles were not consistently under the threshold value for their woodlot of origin as storage time increased, mirroring the divergence seen in both class abundance and NMDS plots.

Conversely, there were instances where bacterial profiles were similar in abundance charts and NMDS plots but were not under threshold values in *k*-NN. Bacterial profiles from the nine deciduous woodlots exhibited this discrepancy, appearing alike in abundance charts and clustering well when ordinated as pairs or triads in NMDS plots; however, not all profiles were under the threshold values of their location of origin when run with all woodlots or in pairs. The calculation of threshold values can explain why these profiles were poorly classified. If the four training profiles acting as knowns are extremely similar, their average intra-point standard deviation would be low, and even slight variation in the fifth profile would push it over the threshold value. Based on this research, collection of more than five samples may be better suited for forensic analysis, lowering the chance of having one extremely dissimilar profile while more accurately capturing within habitat variation (discussed below).

Other supervised classification techniques that can produce statistical values not affected by slight variations among bacterial profiles from the same location may perform better than *k*-

NN for forensic analysis. Two supervised classifiers that provide a more robust measure of association are soft independent modelling of class analogies (SIMCA) and decision trees, which assess the different attributes of data (e.g., OTU or taxonomic class similarities), rather than dissimilarity measures, to group samples. SIMCA statistically compares the residual variance between an unknown and the group it is classified to (Lavine and Davidson, 2006), outputting a group assignment and goodness of fit value. Unlike *k*-NN, it is a soft classifier, not forcing assignment if the unknown does not fit with any of the known groups in the training set. Alternatively, decision trees produce probabilities, measuring how likely an unknown sample belongs with a specific training group based on classification along a branching tree (Rokach and Maimon, 2008), where each branch is a different profile attribute (e.g., a specific taxonomic class). The variance and probability statistics generated by both of these classifiers, as well as several others, is much more definitive than *k*-NN thresholds, and these techniques may provide an even more objective assessment of the location from which soil originated.

Supervised classification via *k*-NN analysis correctly classified the vast majority of profiles to their location or origin in this research; however, one profile from deciduous woodlot 9 and all profiles from deciduous woodlot 8 failed to correctly classify, whether being compared with all similar habitat profiles or in a smaller training set. Complete cluster separation was achieved in NMDS when profiles from pairs or triads of locations were ordinated, but the bacterial variation among these profiles was too high for accurate classification. The dissimilar profile from woodlot 9 was collected on the final sampling date (July 14), when construction was occurring on a nearby road. Although abundance charts for all profiles from woodlot 9 appeared alike, chemicals or disturbances from the construction activity might have altered the bacterial makeup of the soil at the OTU level. A similar possibility exists for the intra-location variability

of profiles generated from woodlot 8. An area of land directly adjacent to this sampling site was previously a gravel pit, which was converted to a park in 1999 (Ingham County Website, 2015). Such anthropogenic soils (IUSS Working Group WRB, 2006) are often complex mixtures of different soil types and can possess extreme spatial heterogeneity (Fitzpatrick, 2013) that influences the microbial community, sometimes resulting in very different profiles from soils collected short distances apart. Soil samples were collected based on GPS coordinates, but it is possible that small distances resulted in dissimilar profiles if strong spatial variability existed. Other anthropogenic soils in this research (e.g., agricultural field and roadside) did not exhibit the same micro-spatial variation, potentially due to the uniform treatment of such habitats compared to a mining location where massive turnover of soil occurs. Extreme spatial heterogeneity would affect classification, especially if only a few soil samples are collected from an anthropogenic location. Variation among known profiles will result in training sets with high standard deviation and low training set validation success. Again, rather than collecting fewer known samples or removing dissimilar profiles to avoid variation, the collection and analysis of additional knowns is recommended based on these results. Highly variable soil can still produce tight clustering and strong training sets if profiles across a gradient of differences are represented. This means that two profiles within a training set can be fairly dissimilar; however, additional profiles at varying degrees of dissimilarity would also exist in the knowns, representing the location of origin while keeping standard deviation relatively low.

It may also be beneficial to combine bacterial profiling with traditional soil analysis techniques for human-influenced soils. The identification of extreme variation can act as a signal that the soil is anthropogenic and potentially possesses individualizing characteristics beyond microbes. The presence of rare elements has led to strong associations between evidentiary and

crime scene soil in both historical cases (e.g., the Eva Disch murder investigation [Murray and Tedrow, 1975]) and modern investigations (e.g., the Adelaide hills double murder [Fitzpatrick et al., 2007]). Elemental analysis (e.g., X-ray florescence) or microscopy may reveal these unique characteristics (Dawson and Hillier, 2010) and allow the forensic scientist to associate two soil samples with or without the use of bacterial profiling. Unique soil characteristics are extremely rare however (Dawson and Hillier, 2009), and traditional techniques are not as useful when analyzing similar soil types. Future studies on the variation present within anthropogenic locations would help in determining if bacterial profiling is a viable option or if combinations of techniques are better suited for the analysis of heterogeneous soils.

Another important consideration for the forensic use of bacterial profiling is how bacterial communities are changing over time. It is fundamentally impossible to collect known soil samples at the time a crime occurs; consequently, temporal changes in bacterial makeup must be examined. Past studies assessing change over time through T-RFLP analysis (Meyers and Foran, 2008) and pyrosequencing (Lauber et al., 2013; Hopkins, 2014) of the 16S locus have revealed substantial differences in bacterial profiles generated from soil collected in different months; however, no specific trends were evident. Some temporal differences in bacterial profiles were apparent in the current study, most of which were based on the months of collection rather than the time separating two collections. Six of the 10 profiles that classified correctly in *k*-NN but were over the threshold value came from soil samples collected December – March. Additionally, profiles from the deciduous woods and yard in February and the yard in March were misclassified and contained substantially fewer OTUs than profiles generated from soil in other months. These bacterial profile differences seem to be more seasonal than temporal, and several seasonal factors could have affected the bacteria, including the amount of daylight,

colder temperatures, snow, and ice. However, not all samples collected in winter months produced profiles that misclassified or did not cluster together in NMDS plots, suggesting that the seasonal conditions of winter are not the main cause of dissimilar profile generation.

Rather than being an exclusively seasonal or temporal affect, profile dissimilarity may have stemmed from the collection of water with the soil samples. Soils from the winter months contained snow and ice and were slushy when thawed for extraction, which potentially affected the bacterial diversity of the resulting profile due to the chemistries of the extraction kit. A PowerSoil® kit utilizes bead beating technology, where cells are lysed in the initial extraction step by the collision of beads moving in solution (PowerSoil® Instruction Manual). Extra water in the extraction tube both dilutes regents and spreads beads further apart, potentially reducing their lysing capabilities. Stronger celled bacteria (often gram positive) will not be lysed as readily in these situations and would thus not be represented in the final bacterial profile. Although there was no indication that such bacteria (e.g., members of *Actinobacteria* and *Bacilli*) were in lower abundance in profiles generated from wet samples, there may have been differences at other taxonomic levels. Extraction reagent and bead dilution could potentially be solved by pelleting a wet soil sample and removing excess water before addition of reagents; as is recommended in the troubleshooting section of the PowerSoil® kit instruction manual; however, if some bacteria do not pellet, the resulting profile would still be inaccurate. A second option is to dry the soils before weighing them for extraction. Drying would allow for the correct mass of soil to be added to extraction tubes and prevent both reagent and bead dilution. Young et al. (2015) found that air drying soil samples for 72 hours in a 25°C incubator did not affect the OTU diversity nor change similarity among eukaryotic profiles from dried and fresh soil samples; however, little research has been done on the effects of drying soil before extraction on

bacterial profiles. Soils that were dried during storage in the t-shirt evidentiary study presented here exhibited differences after several weeks, but sample collections up to 1 week of storage did not produce noticeably dissimilar profiles from the location of origin, suggesting short drying periods (e.g., overnight) do not affect representative profile generation. Despite the slushy soil samples producing dissimilar bacterial profiles, the vast majority (93.7%) of profiles correctly classified to their location of origin over the full year of sampling, indicating that temporal changes do not greatly influence the ability to associate soil profiles generated via next-generation sequencing.

It is also unlikely that known soil samples will be collected from the precise spot an item was exposed, but instead could be collected feet, yards, or greater distances away, stressing the importance of understanding spatial variability of bacterial communities within a location. Differences in bacterial profiles over small distances have been attributed to microenvironmental factors such as foliage, pH, and nutrient supply (Ettema and Wardle, 2002; Eichorst et al., 2007), although in reality, any number of factors could come into play. Bacterial profile variation or patchiness similar to that found in past studies (e.g., Meyers and Foran, 2008) was present in the current research, wherein profiles generated from soils collected across a habitat, either from the surface or at different depths, did not always cluster tightly in NMDS plots. Despite their loose clustering, *k*-NN analysis resulted in the accurate classification of both horizontal and vertical spatial soils when profiles from a range of distances and depths were used as the training set, highlighting the importance of using a variety of known samples to capture within-habitat bacterial variation. It is noteworthy that horizontally collected soil bacterial profiles that were misclassified stemmed from locations at least 90 feet distant from all or four of the five training samples, and were collected in either the deciduous woodlot or yard, which were less than 800

feet apart at the Fenner nature center. This result presents both a strength and weakness of bacterial profile classification. The generation of profiles via next-generation sequencing allowed for differentiation of these close-proximity sites; however, the edges of habitats began to blend, becoming more similar to the neighboring habitat. It may be beneficial to collect known soil samples across a habitat as well as a subset from the edges of habitats to capture these bacterial differences. The dissimilarity of resulting bacterial profiles could then be compared in NMDS plots to examine clustering behavior. If separate clusters representative of the habitat edge form, they can be analyzed in *k*-NN as separate known groups, representing different areas within a location. Classification to one of these groups would allow for the determination of whether soil transfer occurred in the center of a habitat or on the edge.

Soil bacterial profiles generated from samples collected across time and space in this research classified to their location of origin with high reliability; however, actual forensic scenarios will combine many such factors with the added influence of the material evidentiary soil is deposited on, which may already possess its own microbial community. Items in the preliminary evidentiary study were previously used and/or worn; however, they had been rinsed or wiped with water prior to soil exposure. A background bacterial profile may still have existed that either resembled woodlot of origin profiles or contained bacteria that were substantially different from woodlot soils, both potentially influencing traceability. Bacterial profiles generated from those items contained similar numbers of OTUs and bacterial classes, indicating background bacteria did not have a substantial effect on the soil bacterial profiles after exposure in the woodlot. A clean T-shirt from the secondary study produced a bacterial profile with far fewer OTUs and bacterial classes than woodlot soils. Additionally, the profile was dissimilar from all deciduous woodlots and T-shirt soil bacterial profiles, clustering far away in NMDS

plots (data not shown). The clean shirt profile was assigned to deciduous woodlot 2 in *k*-NN; however, it was far above the threshold for all woodlots; an average of 25 standard deviations away from each training group's intra-point distance. Realistic forensic scenarios will likely involve items that are not as pristine when first exposed to soil, such as a worn T-shirt or a used handkerchief, which could harbor very different bacteria than soil. The ubiquity of bacteria poses a potential challenge for forensic analysis in these cases, as the bacteria already present on such items could impact the soil bacterial profile that is generated and in turn, the ability to form associations. In these instances, it may be beneficial to extract DNA from a clean portion of an evidentiary item and subtract the OTUs or taxonomic groups from the resulting profiles, allowing for comparison of only soil bacteria. Subtraction should be done before subsampling, so as not to decrease the number of sequences within a profile more than necessary. There exists a risk of subtracting bacteria that are present in both the soil and on the clean item using this technique, and it is unclear how association would be affected. Therefore, more research is necessary to determine whether evidentiary soil profiles still classify to their location of origin after background bacterial subtraction.

Fewer OTUs were generated from soils deposited on evidentiary items the longer they were stored in this research. Soils removed from their location of origin are no longer exposed to the same environment, and it is likely that not all bacterial types will persist, so the loss of OTU diversity is somewhat intuitive. The differences in abundance of taxonomic classes also made sense in the context of the environment change. *Actinobacteria,* which increased over time, contains species that thrive in dry environments (Ghorbani-Nasrabadi et al., 2013), and *Bacilli*, which also increased, contains species that persist in changing environments due to spores that are much more resilient than those of other bacteria (Claus and Berkeley, 2009).

Sphingobacteria, which decreased over time, is a class whose members produce large lipid membranes that require moisture (Boone and Castenholz, 2001), which was not present in storage. Characteristics like these could explain why members of specific taxonomic classes did or did not persist in stored deciduous woodlot soils. Ongoing research at Michigan State University has shown these same bacterial class changes in stored soils from an agricultural field, dirt road, treated yard, and coniferous forest (Alyssa Badgley, personal communication, 2015), establishing that the abundance differences are not a phenomena specific to deciduous woodlot soils. Owing to these findings, it seems highly likely that the same bacterial class abundance changes will occur in most or all soil types when stored. Rather than being a hindrance for soil evidence investigation, predictable changes in bacterial profiles, at both the class and OTU level, could be used to develop a biological clock measuring how long soil has been removed from a location. Similar to how a post-mortem interval can be estimated based on insect presence and/or the stage of tissue decomposition (e.g., Catts and Goff, 1992; Nelson, 2000), the rate of change in both taxonomic class and OTU abundance in storage might allow estimation of a time period within which evidentiary soil transfer occurred. For example, *Actinobacteria*, which never made up more than 20% of the profiles generated from soils in this research, increased markedly in soil stored on evidentiary items. If an evidentiary item profile showed extremely high levels of *Actinobacteria* (e.g., threefold higher than any known profile generated), it can be assumed that the soil sample is not fresh and that transfer probably occurred more than 2 months previous (based on this research). The relative abundance of several bacterial taxonomic classes combined with a measure of how many OTUs are present within a profile would result in a more precise clock, allowing for smaller range estimations. However, it should be noted that such a clock may be influenced by storage conditions, such as temperature,

humidity, or nutrient availability. Not all of these factors were examined in the current research, but cooler temperatures slowed the onset of bacterial abundance change and OTU reduction in profiles generated from the T-shirts. Temperature-dependent changes will affect how biological clocks are calibrated (e.g., if a piece of evidence is stored in an outdoor shed during winter versus an indoor closet). Studies examining various storage conditions and how they affect soil bacteria will need to be performed to increase our knowledge on how evidentiary profiles might differ from their location of origin after soil transfer occurs.

The woodlot in which items were exposed to soil in the evidentiary studies was one of the most unique of the similar habitats, clustering the farthest from all other woodlot profiles in NMDS plots. This distinctiveness may have influenced the evidentiary soil traceability results, as evidentiary profiles were initially slightly different from all other locations. However, the profiles never developed characteristics of other woodlots over time, remaining most similar to the woodlot of origin even after 1 year of storage. Additionally, abundance charts of evidentiary profiles were not similar to any other location from the diverse habitat study, suggesting the same traceability results would have been obtained if the evidence were compared to those profiles. The ongoing research at Michigan State University mentioned above, assessing evidentiary profile traceability over time within four other habitat types has shown similar results: evidentiary soil on T-shirts classified to its location of origin in the months following exposure (Alyssa Badgley, personal communication, 2015). In combination with the research presented in this thesis, these results highlight the strong potential for bacterial profile traceability, as soil profiles correctly classified to their location of origin regardless of storage time or the material on which soil was stored.

The evidence traceability success in this research is promising; however, without proper collection of known soil samples, bacterial profile comparison and association is difficult or impossible. Crime scene investigators must be informed about where, when, and how to collect known soil samples as well as the number of samples to collect. A collection strategy can begin to be developed based on the results of this thesis research. Multiple known samples must be collected in order to develop location clusters in ordination techniques or training sets for supervised classification. Additionally, the collection of multiple known samples proved beneficial in capturing the bacterial variation present within habitats in this research, providing more accurate classification when a range of profiles generated from soil across a given location were used as the training set (e.g., the horizontal space soils). For these reasons, multiple samples over short distances across a habitat should be collected, which would allow for the development of strong training sets representative of habitat extremes, as well as a range of bacterial profiles within these extremes. Evidentiary profiles from a given location could then be discriminated or associated with a certain level of confidence, knowing that bacterial variation was represented within the training set.

The collection of soil samples weeks or months after a crime occurred will likely not produce profiles dissimilar to evidentiary soils; however, excessive wetness due to environmental phenomena such as snow, ice, or rain should be taken into account by crime scene investigators when collecting knowns. Avoiding water during collection is the most logical course of action, but this may not be possible, especially in icy conditions. Therefore, the individual collecting known soil samples must be mindful of excess water, handling these samples differently, either by drying the soils before packaging at the crime scene or marking the

sample in some way to ensure laboratory analysts are aware that water could be a factor when the samples are thawed for extraction.

Storage of known soil samples after collection is another facet of forensic analysis that must be considered. Bacterial change occurred at both evidentiary storage temperatures in this research (4°C and 24°C). The assessment of bacterial change in a subset of the known soil samples, which were stored at -20°C before DNA extraction, was recently performed (Alyssa Badgley, personal communication, 2015) in which four soil samples collected from the treated yard and stored for 1 year were re-extracted and their corresponding stored DNAs were re-amplified for comparison to the original profiles. Re-amplified DNAs produced bacterial profiles very similar to the original amplification; however, frozen soils generated slightly dissimilar profiles, plotting farther away from habitat clusters in NMDS. Additionally, the bacterial classes that changed in frozen soils were different than those that had changed on stored evidentiary items (*Acidobacteria* and *Actinobacteria* decreased while *Flavobacteria* increased), showing cold storage temperatures affect bacteria differently. Forensic laboratories may store samples for long periods if a backlog of cases exists, and it will be important that the bacteria within stored soil are not changing during this time. Even colder storage of knowns (e.g., at -80°C) is one possibility to maintain bacterial profiles representative of their origin; however, evidentiary soils that have been exposed to a warmer environment before discovery may produce dissimilar bacterial profiles compared to fresh samples from the origin location. A better option is to recreate the evidentiary soil storage conditions and expose known samples from all locations involved in the investigation to that environment. Storage under similar conditions would likely result in similar bacterial changes if a known and evidentiary soil sample came from the same location. Conversely, soil profiles originating from different locations could be differentiated

using this technique, as their starting bacteria would be different, and the resulting stored profiles would also be dissimilar.

A final consideration for the adoption of next-generation sequencing for forensic soil analysis is its integration into both crime laboratories and court, a process that involves multiple steps. First, sequencing technology must be available, and in most cases this means adding it to the laboratory, as forensic laboratories do not currently employ next-generation sequencing for casework. However, it is unlikely every crime laboratory will have the means or desire to purchase and maintain expensive equipment such as a next-generation sequencer, especially if only a handful of samples are run per year. In this regard, the technology could be made available through a central or regional laboratory, much like the FBI has done with some of its mtDNA testing laboratories (Forensic Science Communications, 2003). Such a laboratory would receive samples from across the country or region, requiring only one team of personnel to be trained on next-generation sequencing analysis, saving time and money.

The next step in the implementation process is the acceptance of the methods by the courts through satisfaction of Frye and Daubert standards. The methodology being presented by an expert witness must have been empirically tested, widely accepted in its appropriate field, published in the peer reviewed literature, have a known error rate, and possess standards for its operation (*Daubert v. Merrell Dow Pharmaceuticals, Inc.*, 1993). The generation of bacterial profiles through next-generation sequencing already largely meets these criteria because it has been successfully used by so many in microbiology and other fields (reviewed by Shokralla et al., 2012). Bacterial sequencing has been extensively tested and presented in the microbiology peer reviewed literature, and approximate error rates for base calling are known for many next-generation sequencing platforms (e.g., Loman et al., 2012; Liu et al., 2012; Quail et al., 2012).

However, the goals of forensic science differ from those of microbiology, where definitive association or discrimination is not necessary, as it is in forensic soil analysis. Forensic applications of soil bacterial profiling are not prevalent in the peer reviewed literature, but studies like those presented in this thesis represent the first steps toward validation of microbiological soil comparisons for use in forensic analysis.

Another challenge remaining for laboratory implementation of bacterial profile soil analysis is to develop universal standards for the technique. Differences already exist among countries on which markers to assay for human DNA analysis (e.g., the United States 13 core loci [Budowle et al., 1998] and the Extended European Standard Set [European Council, 2001]), making international profile comparison difficult. Microbiological soil analysis offers even more options than human DNA, as several taxa can be assayed, and it is unclear whether profiles of different organisms within soil provide similar results. Young et al. (2014) found differences in diversity and reproducibility of profiles when sequencing genetic markers of soil fungi, eukaryotes, plants, and bacteria; however, profiles from all four taxa allowed for discrimination of two sites, suggesting any of these taxa could be utilized for forensic analysis. The use of different sequence processing methods has the potential to produce conflicting results across laboratories. For example, the employment of the Greengenes reference database (DeSantis et al., 2006) instead of SILVA in this research may have provided higher discriminatory power among bacterial profiles, affecting soil sample association. It seems beneficial to standardize the entire forensic soil analysis process rather than attempt to meet Daubert standards for and achieve court acceptance of the many microbial profile generation methods and processing techniques that exist in the microbiology literature. Forensic soil analysis could be regulated by a standardization body such as the Scientific Working Group on Microbial Genetics and Forensics

(Budowle, 2003), which already has guidelines for the implementation of microbiology into crime laboratories, but does not specifically outline standards for soil evidence. The same marker in one soil taxon could be assayed by all forensic laboratories, expediting the court acceptance process. Standardization would also make the presentation of results in court more understandable, allowing laypersons to easily compare two experts' testimony, rather than focusing on how to compare data.

Despite the need for continued evidentiary soil research, crime laboratory implementation, and court acceptance, the research presented here shows the tremendous potential of next-generation sequencing to meet the main goal of forensic soil analysis: discrimination or association of an evidentiary sample and a crime scene. The development of soil bacterial profiles based on the 16S rRNA gene offers a promising avenue for such analysis, surpassing the value of class characteristics measured through time-consuming traditional methods. Many microbiological profiling techniques have shown potential for forensic analysis; however, this research demonstrates that next-generation sequencing provides the resolution and discriminatory power necessary for criminal investigations.

APPENDICES

APPENDIX A. Photographs of Sampling Sites and Evidence

Diverse Habitats

Figure A1—Agricultural field in East Lansing, MI.

Figure A2—Beach on Lake Lansing in Haslett, MI.

Figure A3—Coniferous forest at Woldumar Nature Center in Lansing, MI.

Figure A4—Deciduous woodlot at Fenner Nature Center in Lansing, MI.

Figure A5—Dirt road in Perry, MI.

Figure A6—Fallow agricultural field in Perry, MI.

Figure A7—Field at Fenner Nature Center in Lansing, MI.

Figure A8—Marsh edge at Fenner Nature Center in Lansing, MI.

Figure A9—Roadside in Lansing, MI.

Figure A10—Yard at Fenner Nature Center in Lansing, MI.

Similar Habitats

Figure A11—Deciduous Woodlot 1 at Fenner Nature Center in Lansing, MI.

Figure A12—Deciduous Woodlot 2 in East Lansing, MI.

Figure A13— Deciduous Woodlot 3 in East Lansing, MI.

Figure A14— Deciduous Woodlot 4 in Lansing, MI.

Figure A15— Deciduous Woodlot 5 in East Lansing, MI.

Figure A16— Deciduous Woodlot 6 in East Lansing, MI.

Figure A17— Deciduous Woodlot 7 in East Lansing, MI.

Figure A18— Deciduous Woodlot 8 in Lansing, MI.

Figure A19— Deciduous Woodlot 9 in Okemos, MI.

Spatial and Temporal Habitats

Figure A20—Deciduous woodlot at Fenner Nature Center in Lansing, MI.

Figure A21—Yard at Fenner Nature Center in Lansing, MI.

Figure A22—Treated Yard at Michigan State University in East Lansing, MI. Used for temporal and horizontal spatial studies.

Figure A23—Treated Yard at Michigan State University in East Lansing, MI. Used for depth study.

Evidentiary Items

Figure A24—Tire with soil collected from woodlot 1.

Figure A25—Shovel with soil collected from woodlot 1.

Figure A26—Shirt with soil collected from woodlot 1.

Figure A27—Shoes with soil collected from woodlot 1.

Figure A28—Sock with soil collected from woodlot 1.

Figure A29—T-shirt being exposed to soil in woodlot 1.

APPENDIX B. Sequence Processing Commands for Mothur Version 1.33.3

"Samplefile" is the file name used in the following commands. All files must be in the same location as the mothur file being used for processing. Refer to http://www.mothur.org/wiki/MiSeq_SOP for additional information.

1. mothur > make.contigs(file=samplefile.txt*, processors=8)
2. mothur > summary.seqs(fasta=samplefile.trim.contigs.fasta)
3. mothur > screen.seqs(fasta=samplefile.trim.contigs.fasta, group=samplefile.contigs.groups, maxambig=0, maxlength=475)
4. mothur > summary.seqs(fasta=samplefile.trim.contigs.good.fasta)
5. mothur > count.groups(group=samplefile.contigs.good.groups)
6. OPTIONAL MERGE FILES COMMAND. Must merge both the fasta files and the group files separately.
   mothur > merge.files(input=fileA-fileB-fileC, output=fileABC)
7. mothur > sub.sample(fasta=samplefile.trim.contigs.good.fasta, group=10locationsstability.contigs.good.groups, size=3000,persample=T)
8. mothur > unique.seqs(fasta=samplefile.trim.contigs.good.Subsample.fasta)
9. mothur > count.seqs(name=samplefile.trim.contigs.good.Subsample.names, group=10locationsstability.contigs.good.Subsample.groups)
10. mothur > summary.seqs(count=samplefile.trim.contigs.good.Subsample.count_table)
11. mothur > pcr.seqs(fasta=silva.bacteria.fasta†, start=11894, end=25319, keepdots=F, processors=8)
12. mothur > system(rename silva.bacteria.pcr.fasta silva.samplefile.fasta)
13. mothur > summary.seqs(fasta=silva.samplefile.fasta)
14. mothur > align.seqs(fasta=samplefile.trim.contigs.good.Subsample.unique.fasta, reference=silva.samplefile.fasta)
15. mothur > summary.seqs(fasta=samplefile.trim.contigs.good.Subsample.unique.align, count=samplefile.trim.contigs.good.Subsample.count_table)
16. mothur > screen.seqs(fasta=samplefile.trim.contigs.good.Subsample.unique.align, group=samplefile.contigs.good.Subsample.groups, name=samplefile.trim.contigs.good.Subsample.names, summary=samplefile.trim.contigs.good.Subsample.unique.summary, start=1968, end=11550, maxhomop=8,processors=8)
17. mothur > summary.seqs(fasta=samplefile.trim.contigs.good.subsample.unique.good.align, count=samplefile.trim.contigs.good.subsample.count_table)
18. mothur > filter.seqs(fasta=samplefile.trim.contigs.good.subsample.unique.good.align, vertical=T, trump=., processors=8)
19. mothur > unique.seqs(fasta=samplefile.trim.contigs.good.subsample.unique.good.filter.fasta, count=samplefile.trim.contigs.good.subsample.count_table)

20. mothur > pre.cluster(fasta=samplefile.trim.contigs.good.subsample.unique.good.filter.unique.fasta, count=samplefile.trim.contigs.good.subsample.unique.good.filter.count_table, diffs=2)
21. classify.seqs(fasta=samplefile.trim.contigs.good.subsample.unique.good.filter.unique.precluster.fasta, template=silva.samplefile.fasta, taxonomy=silva.bacteria.silva.tax†)
22. mothur > cluster.split(fasta=samplefile.trim.contigs.good.subsample.unique.good.filter.unique.precluster.fasta, count=samplefile.trim.contigs.good.subsample.unique.good.filter.unique.precluster.count_table, taxonomy=samplefile.trim.contigs.good.subsample.unique.good.filter.unique.precluster.silva.wang.taxonomy,cutoff=.15, splitmethod=classify, taxlevel=3,processors=8)
23. mothur > make.shared(list=samplefile.trim.contigs.good.subsample.unique.good.filter.unique.precluster.an.unique_list.list, count=samplefile.trim.contigs.good.subsample.unique.good.filter.unique.precluster.count_table, label=0.03)
24. mothur > summary.shared(shared=current,calc=braycurtis-sorclass)
25. mothur > classify.otu(list=current, count=current, taxonomy=current, label=0.03)

Summary files for Bray-Curtis/Sørensen-Dice calculation and taxonomy can be opened in excel. Final shared file contains operational taxonomic units of each sample.

*This is a stability file constructed in Microsoft Excel and saved as a txt file. The format is sample name, R1 file name, R2 file name for each sample.

†Bacterial reference files downloaded from SILVA data base at http://www.arb-silva.de/ (Quast et al., 2013)

APPENDIX C. Bacterial Profile OTU Diversity

Table C1—Number of OTUs in each bacterial profile generated from diverse habitat soils.

	Date of Collection				
Habitat	8/29/2013	11/24/2013	2/7/2014	5/16/2014	8/13/2014
Agricultural Field	857	1067	835	1031	648
Beach	894	799	766	743	980
Coniferous Forest	617	919	1062	829	803
Fenner Deciduous Woodlot	889	1176	1039	1106	1038
Dirt Road	259	282	-	285	278
Fallow Agricultural Field	915	1198	1126	1065	1083
Field	850	981	954	889	985
Marsh	911	1121	1193	1168	927
Roadside	584	859	581	709	738
Yard	990	1149	762	628	1143

Table C2—Number of OTUs in each bacterial profile generated from similar habitat soils.

	Date of Collection				
Deciduous Woodlot	5/16/2014	5/30/2014	6/13/2014	6/27/2014	7/14/2014
Woodlot 1	1056	1036	957	1044	952
Woodlot 2	983	1046	1005	935	919
Woodlot 3	1024	1099	1016	1050	1029
Woodlot 4	983	1020	955	898	888
Woodlot 5	1046	1027	904	1003	933
Woodlot 6	1036	1052	845	920	911
Woodlot 7	1067	1015	1004	1029	1036
Woodlot 8	1119	1227	1009	908	954
Woodlot 9	959	931	882	919	902

Table C3—Number of OTUs in each bacterial profile generated from three habitats over time.

	Date of Collection											
Habitat	8/29/13	8/30/13	8/31/13	9/1/13	9/5/13	9/12/13	9/19/13	9/16/13	10/3/13	10/10/13	10/17/13	10/24/13
Fenner Deciduous Woodlot	889	1130	1030	1092	1175	1019	1086	1070	1022	1056	1016	1032
Treated Yard	1160	1059	1042	1071	1067	1090	996	1084	962	1032	976	882
Yard	990	1159	1107	1052	1086	1082	1050	1158	1038	1041	1065	1051

	Date of Collection										
Habitat	11/3/13	12/2/13	1/2/14	2/2/14	2/28/14	3/30/14	4/26/14	5/30/14	6/27/14	7/31/14	8/29/14
Fenner Deciduous Woodlot	1007	1121	998	794	483	852	1029	956	943	996	960
Treated Yard	923	891	830	874	823	1063	985	942	1099	985	1055
Yard	1018	1095	1092	1120	739	764	954	1095	997	1183	1063

Table C4—Number of OTUs in each bacterial profile generated from soils across the surface of three habitats.

	Location of Collection								
Habitat	Center	5'North	5'South	5'East	5'West	10'North	10'South	10'East	10'West
Fenner Deciduous Woodlot	1023	1076	1141	977	1104	1170	1144	973	1172
Treated Yard	1231	1026	1049	1109	871	1127	1097	1022	1126
Yard	972	860	1043	781	836	945	1015	861	1061
	Location of Collection								
Habitat	50'North	50'South	50'East	50'West	100'North	100'South	100'East	100'West	
Fenner Deciduous Woodlot	940	1063	1001	1057	1057	1029	1228	900	
Treated Yard	994	1129	1155	1005	807	1118	1008	1124	
Yard	1014	1129	1069	1027	1011	1010	1108	1031	

Table C5—Number of OTUs in each bacterial profile generated from at different depths within three habitats.

	Depth of Collection						
Habitat	Surface	1"	2"	5"	10"	20"	60"
Fenner Deciduous Woodlot October	1020	943	939	963	1058	993	242
Fenner Deciduous Woodlot April	992	1203	1040	926	887	1197	647
Yard October	1159	1138	1174	1139	1058	968	1047
Yard April	1027	1057	1046	1141	1087	882	1010
Treated Yard April	866	964	1048	1144	948	613	1063*

*Soil collected at 25” due to obstruction.

Table C6—Number of OTUs in each bacterial profile generated from soil on stored evidentiary items.

	Time in Storage	
Evidentiary Item	6 Months*	1 Year
Shirt	605	853
Shoe	873	958
Shovel	714	699
Sock	660	1015
Tire	871	952

*OTU value presented represents homogenized soil profile.

Table C7—Average number of OTUs in bacterial profile generated from soil on stored T-shirts (n=4). The clean shirt bacterial profile contained 256 OTUs.

	Time in Storage										
Storage Temperature	0 Weeks	1 Week	2 Weeks	3 Weeks	4 Weeks	5 Weeks	6 Weeks	7 Weeks	8 Weeks	12 Weeks	16 Weeks
24°C	1033	986	919	732	919	832	856	765	689	703	775
4°C	1021	1022	1008	899	930	957	947	897	772	895	790

APPENDIX D. Additional Bacterial Abundance Charts

Diverse Habitat Soil Samples

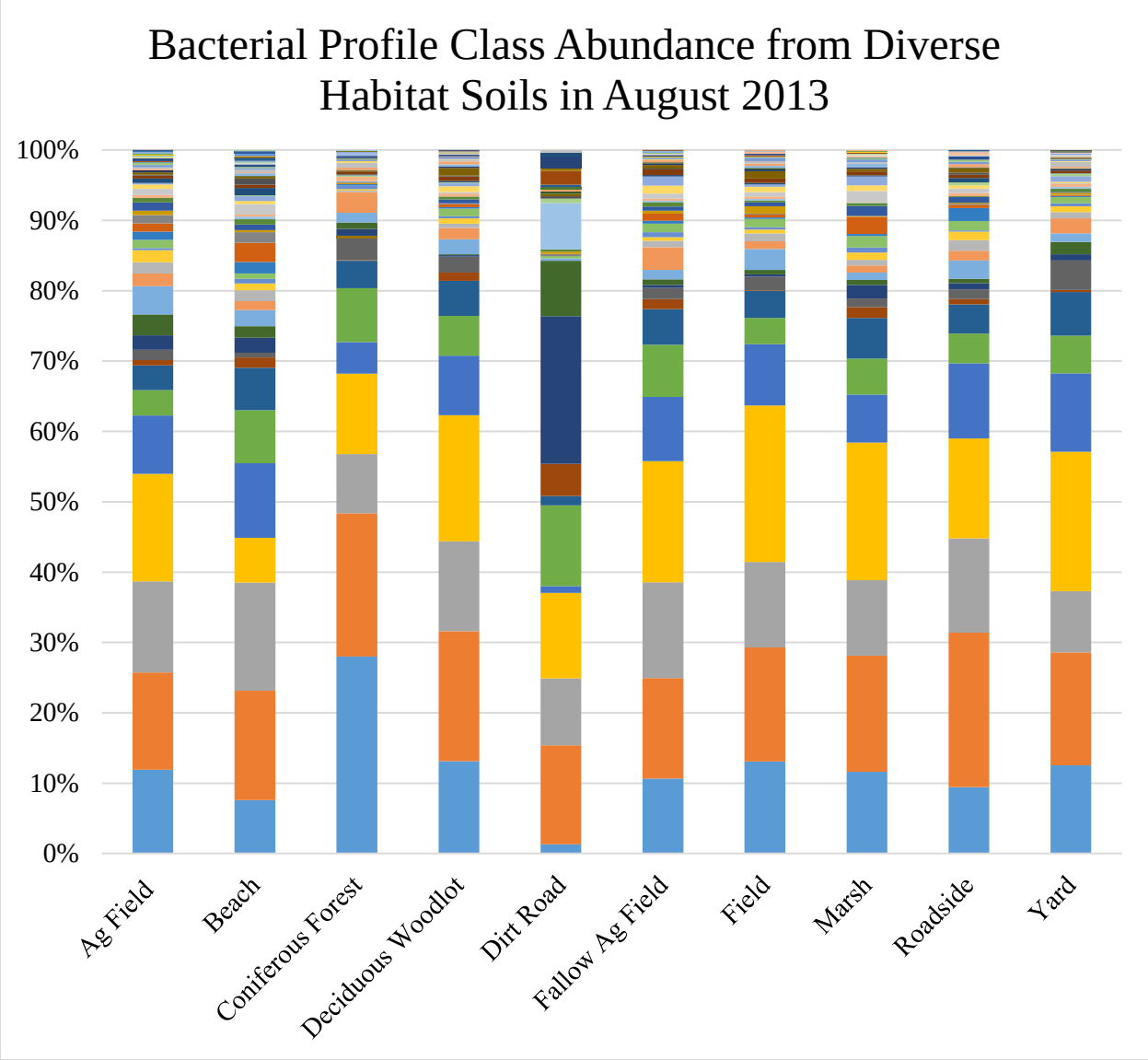

Figure D1—Bacterial class abundance of soil collections from the ten diverse habitats in August of 2013.

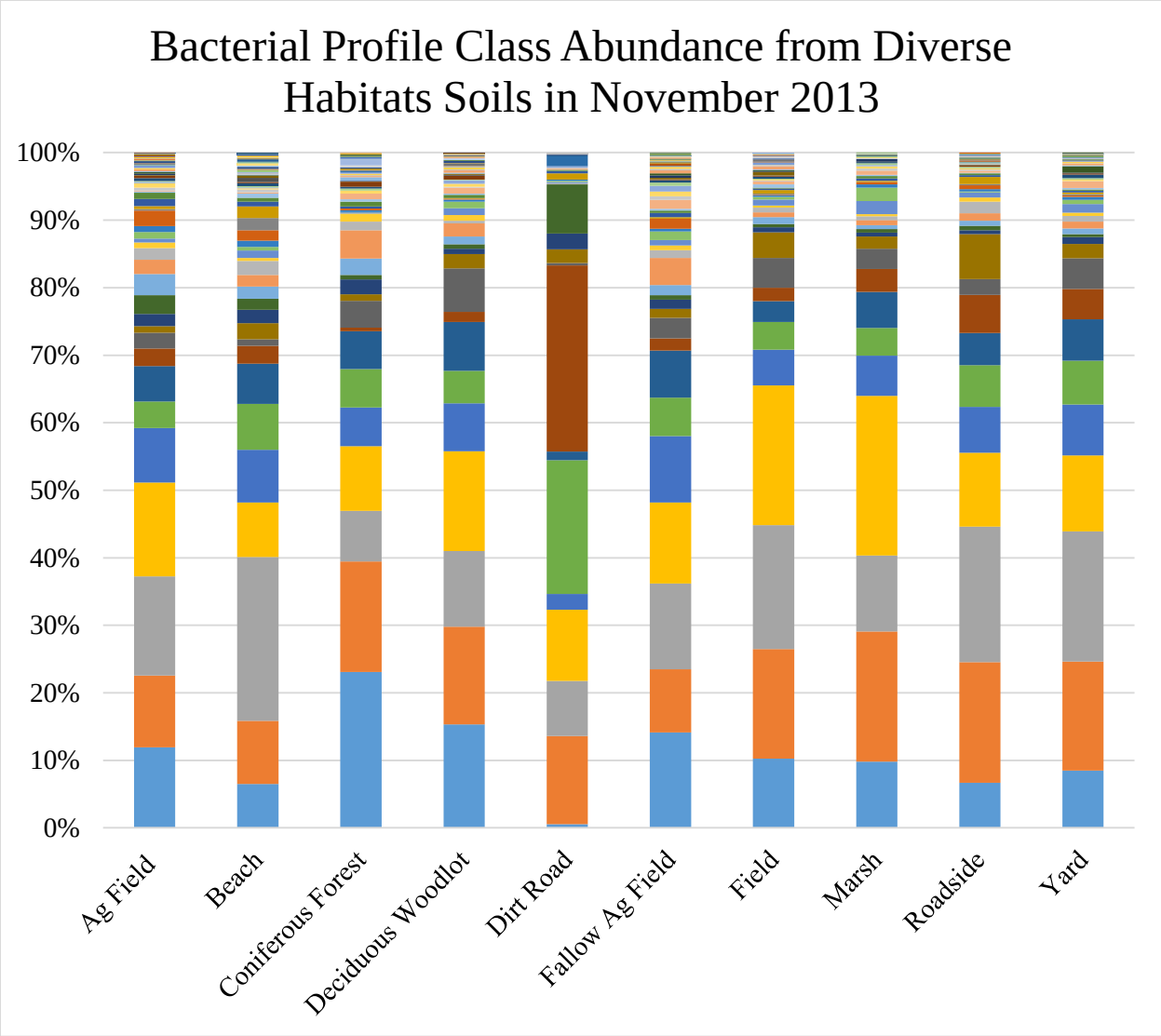

Figure D2—Bacterial class abundance of soil collections from the ten diverse habitats in November of 2013.

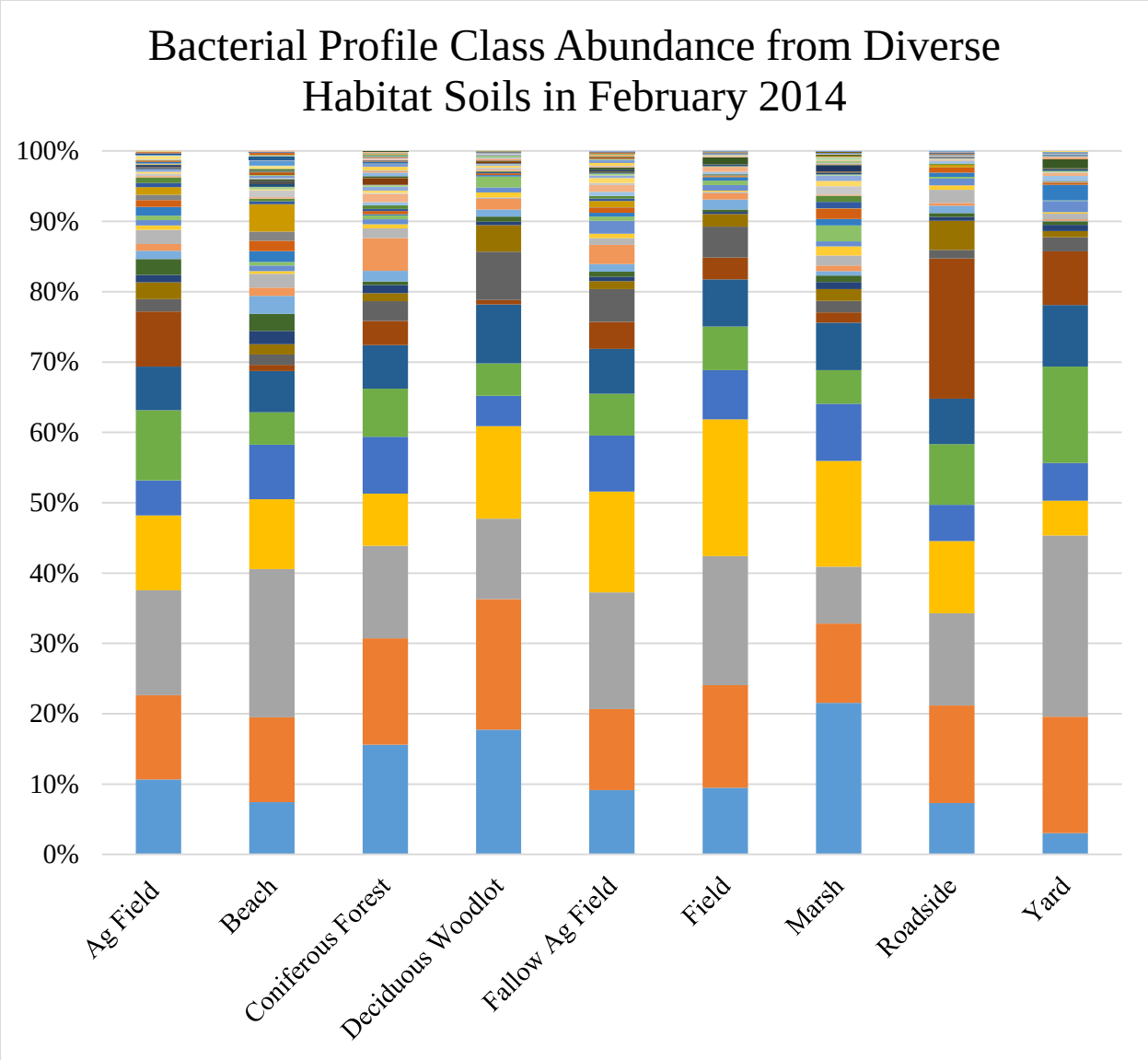

Figure D3—Bacterial class abundance of soil collections from the ten diverse habitats in February of 2014. Dirt road sample failed to produce 3000 sequences and was excluded from further processing.

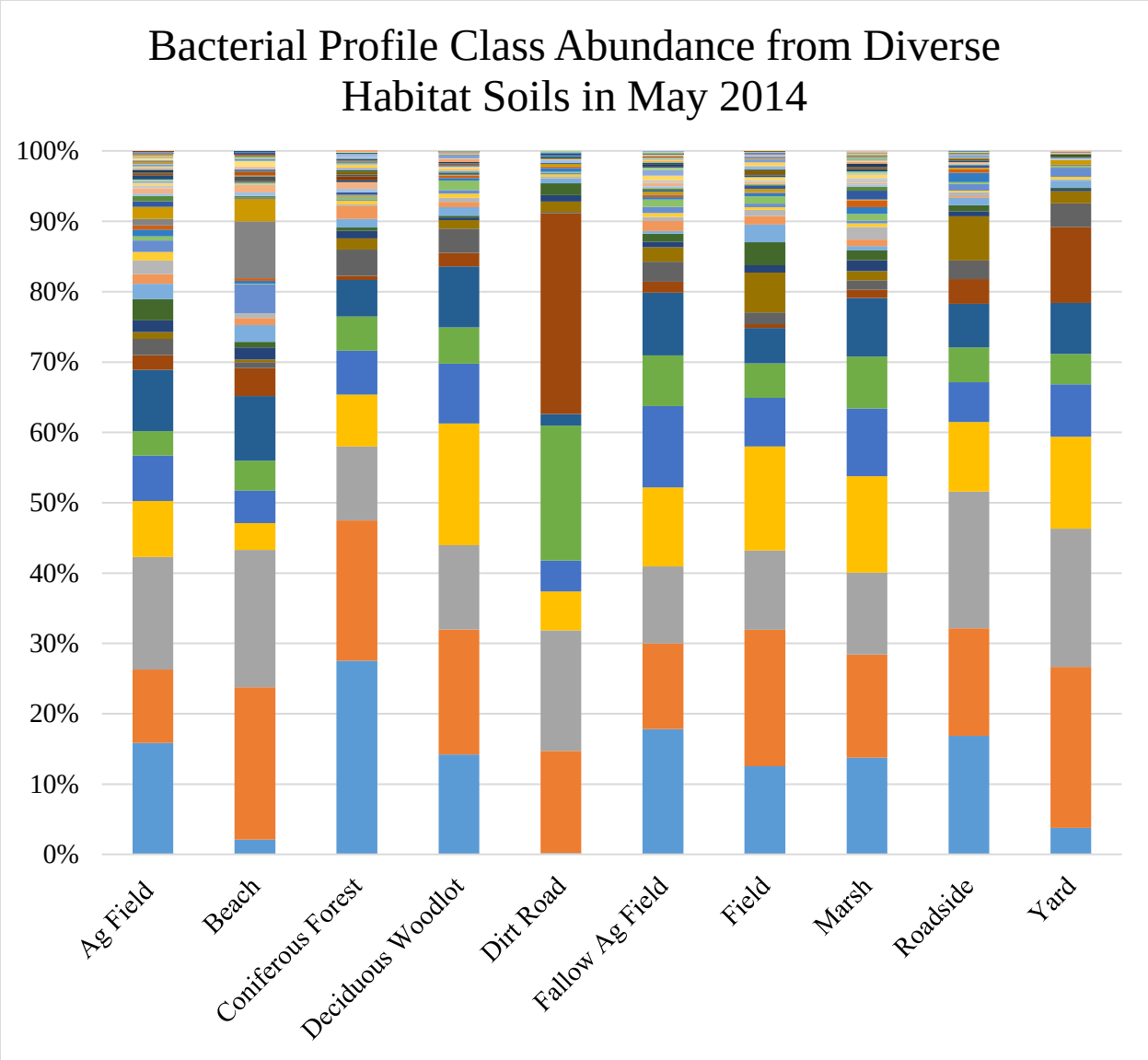

Figure D4—Bacterial class abundance of soil collections from the ten diverse habitats in May of 2014.

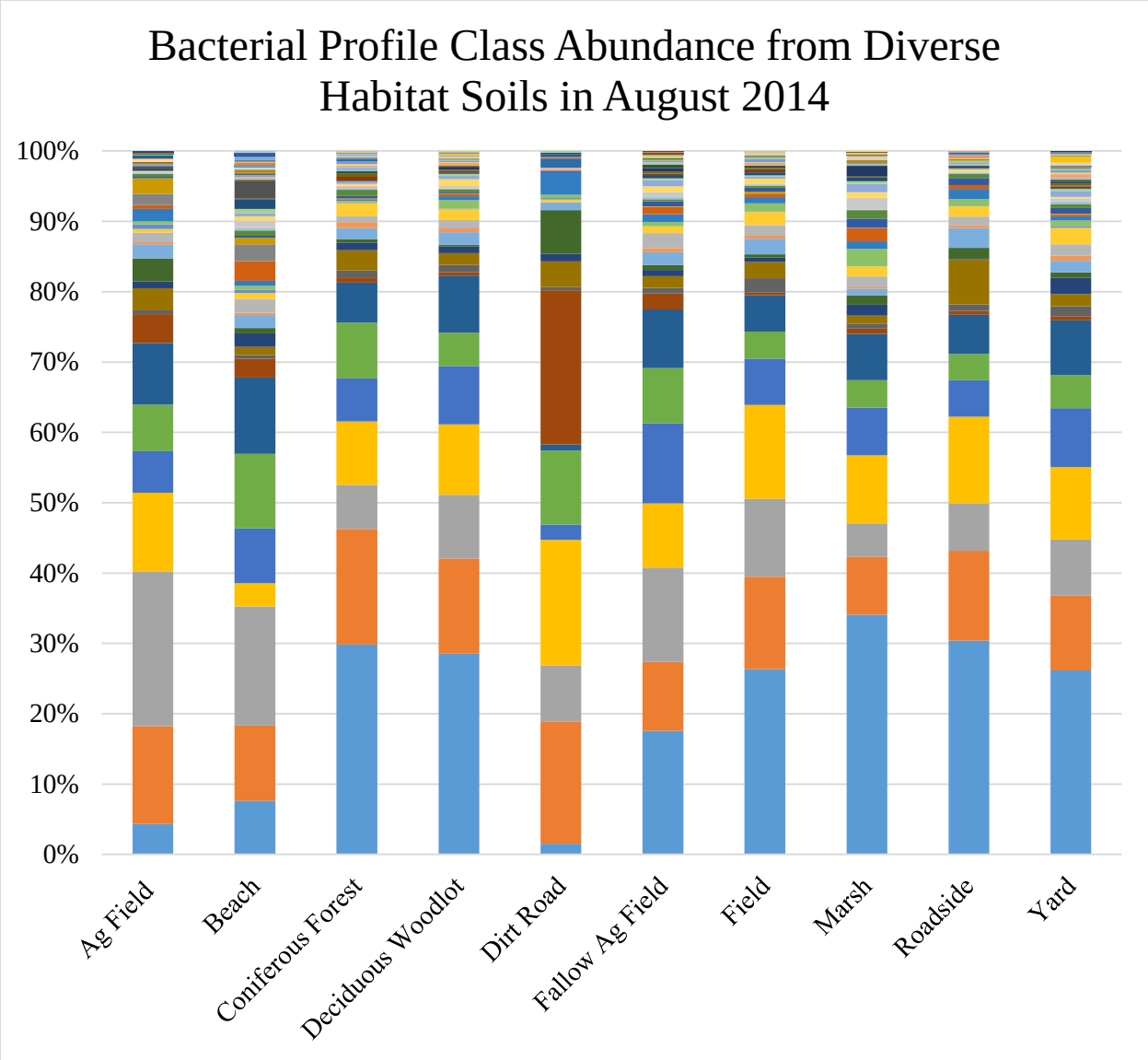

Figure D5—Bacterial class abundance of soil collections from the ten diverse habitats in August of 2014.

Similar Habitat Soil Collections

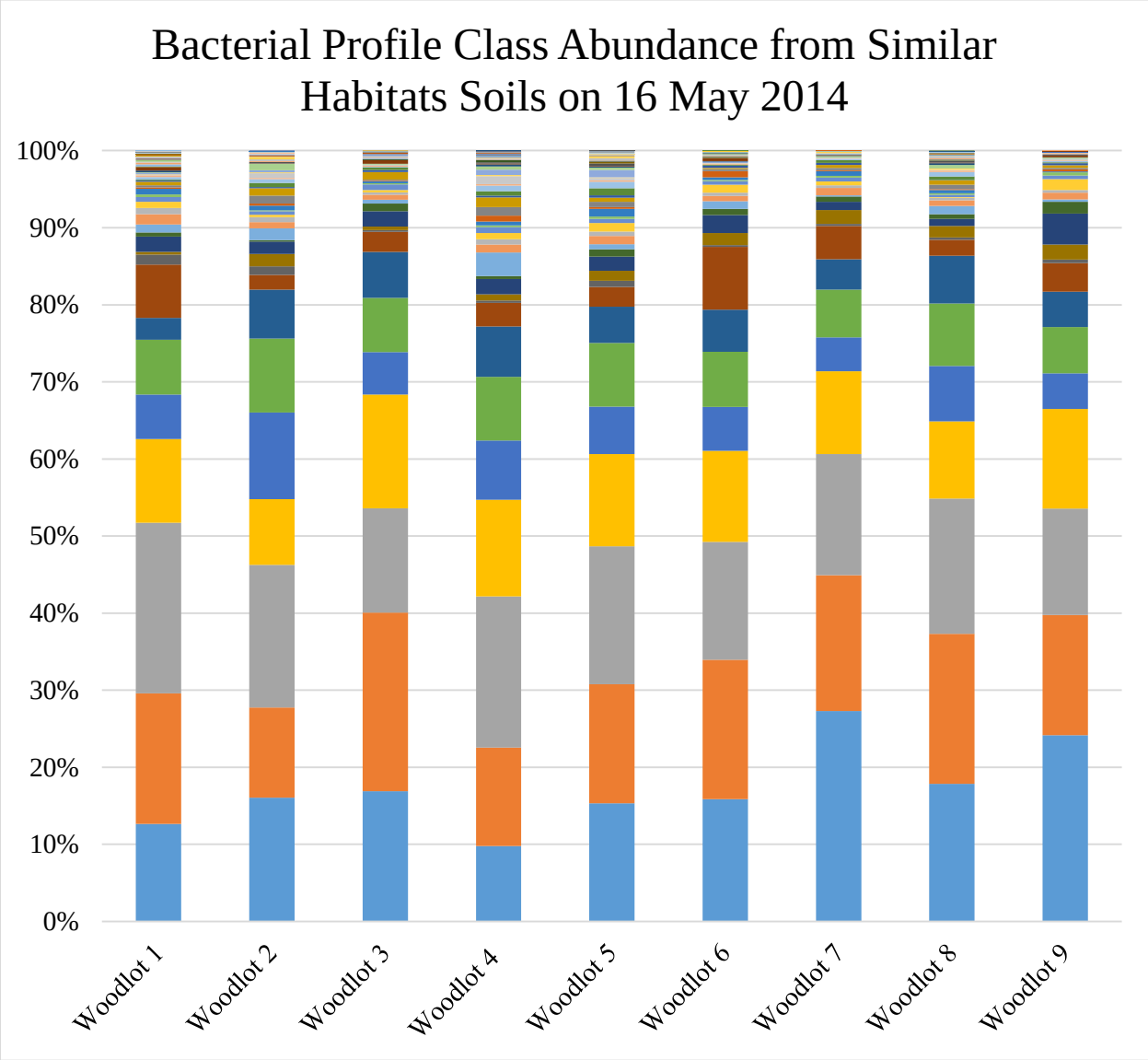

Figure D6—Bacterial class abundance of soil collections from the nine deciduous woodlots in May of 2014.

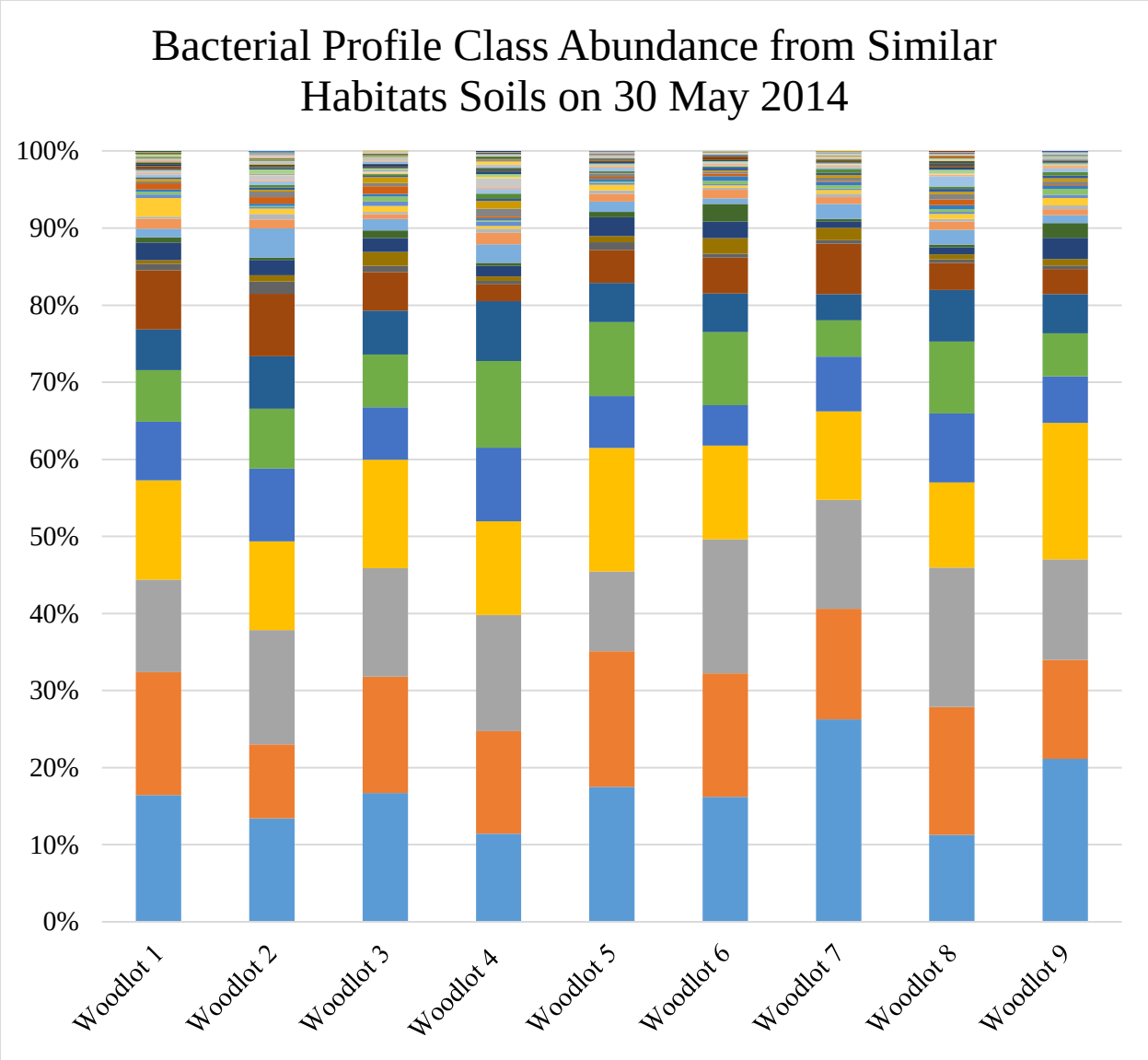

Figure D7—Bacterial class abundance of soil samples collected from the nine deciduous woodlots in May of 2014.

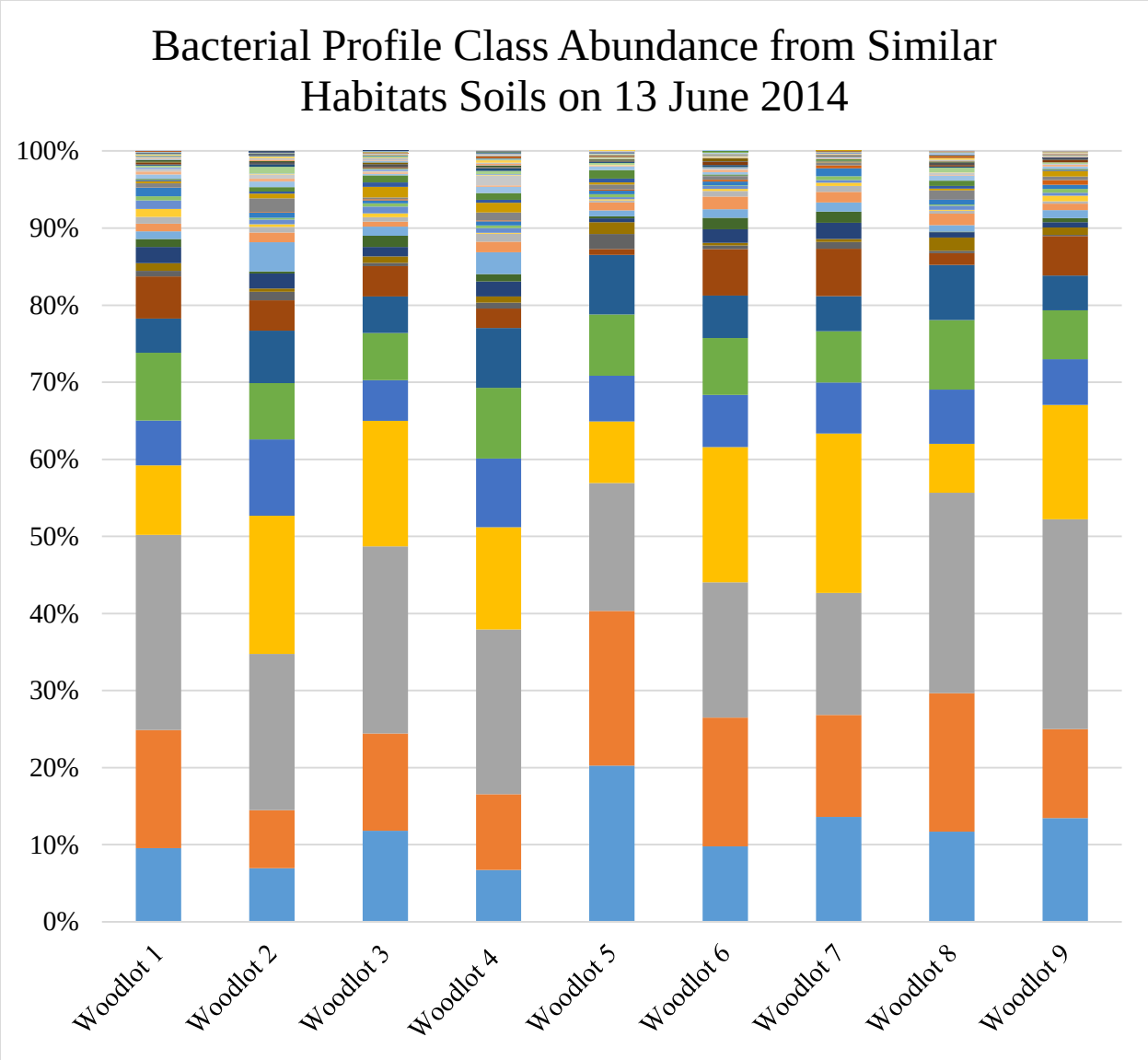

Figure D8—Bacterial class abundance of soil collections from the nine deciduous woodlots in June of 2014.

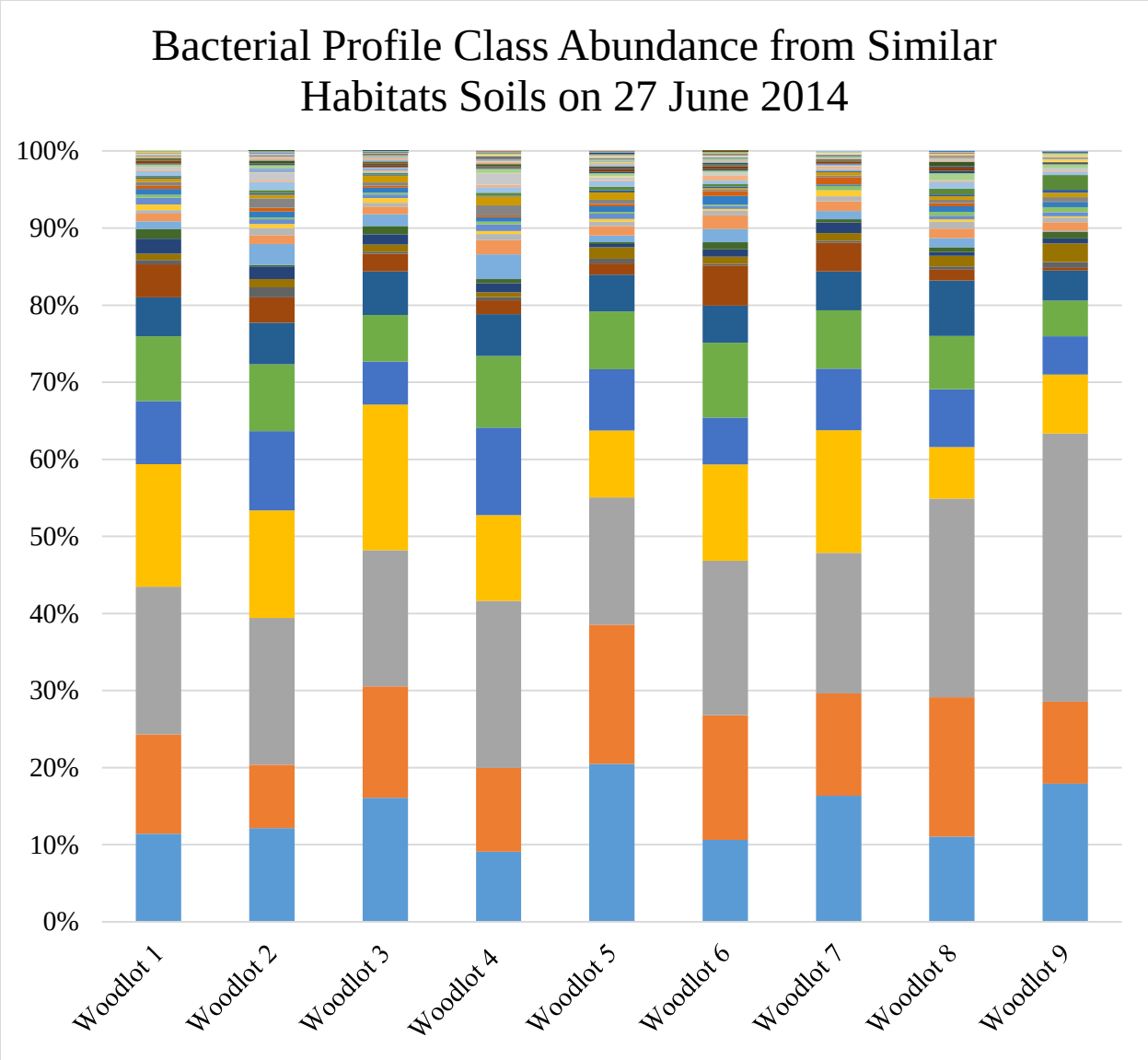

Figure D9—Bacterial class abundance of soil collections from the nine deciduous woodlots in June of 2014.

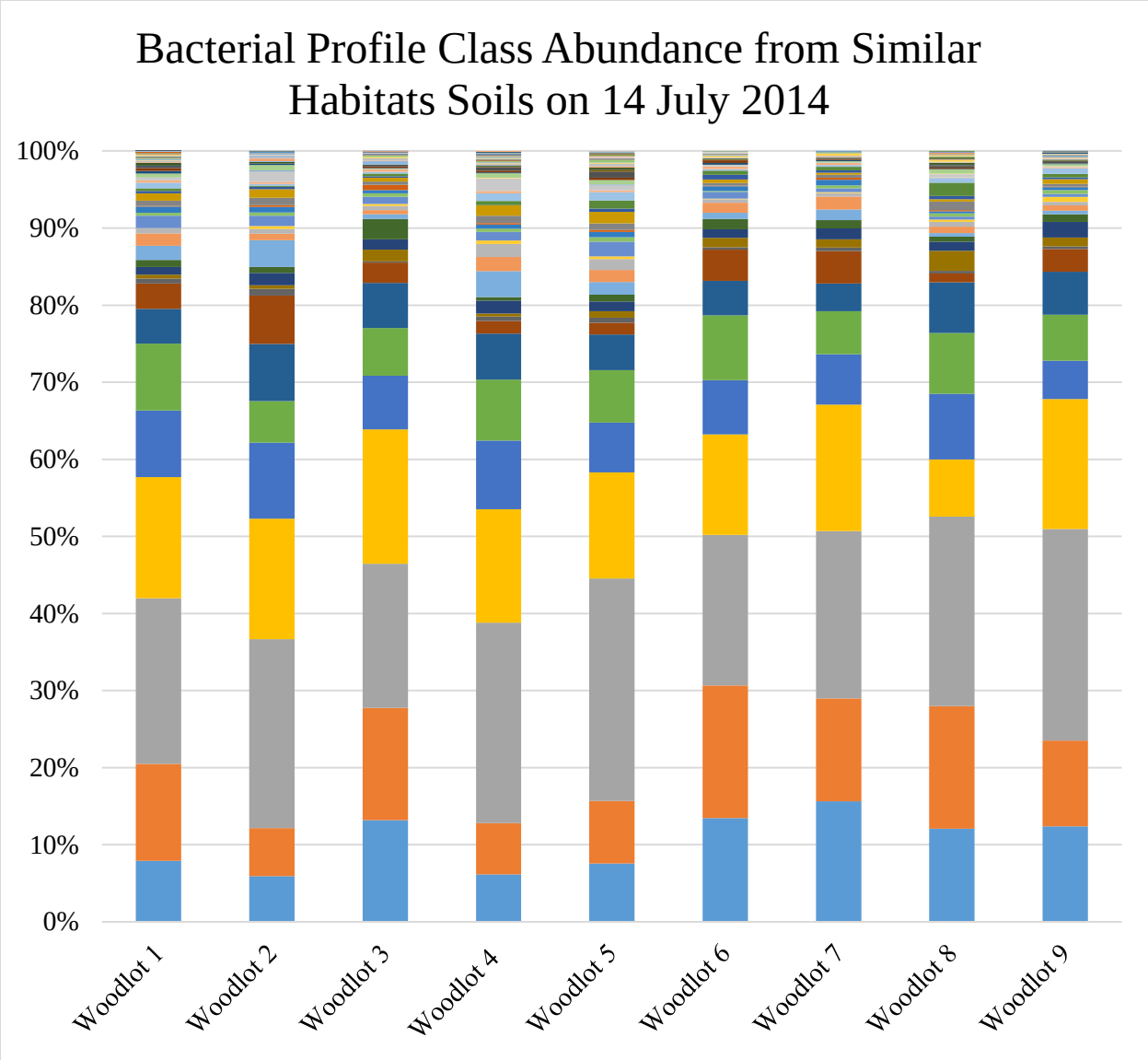

Figure D10—Bacterial class abundance of soil collections from the nine deciduous woodlots in July of 2014.

Temporal Soil Samples

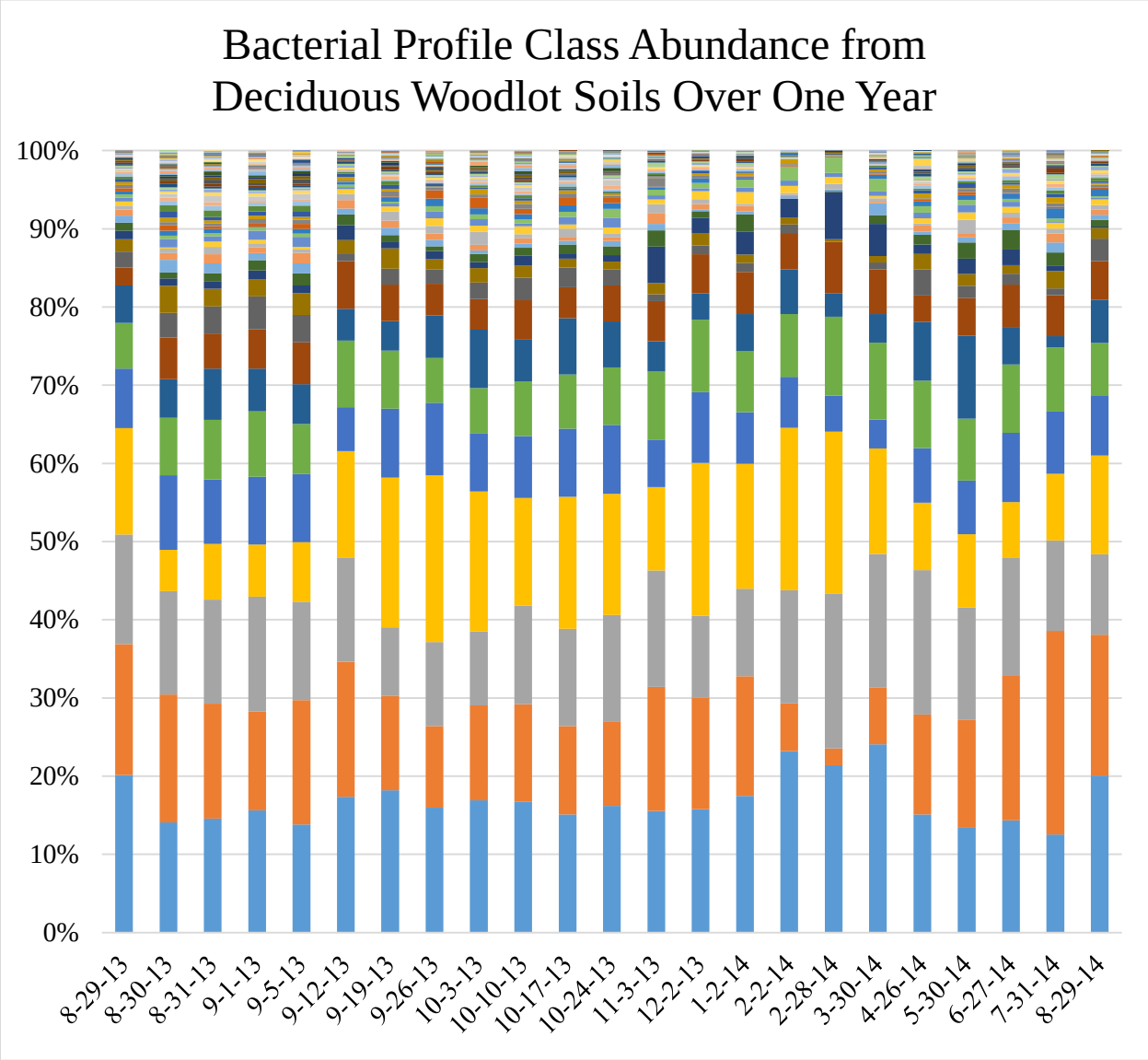

Figure D11—Bacterial class abundance of soil collections from the same location within a deciduous woodlot from August 2013 – August 2014. Soils were collected daily for 4 days, weekly for 2 months, and monthly for the remainder of the year so chart is not evenly spaced over time.

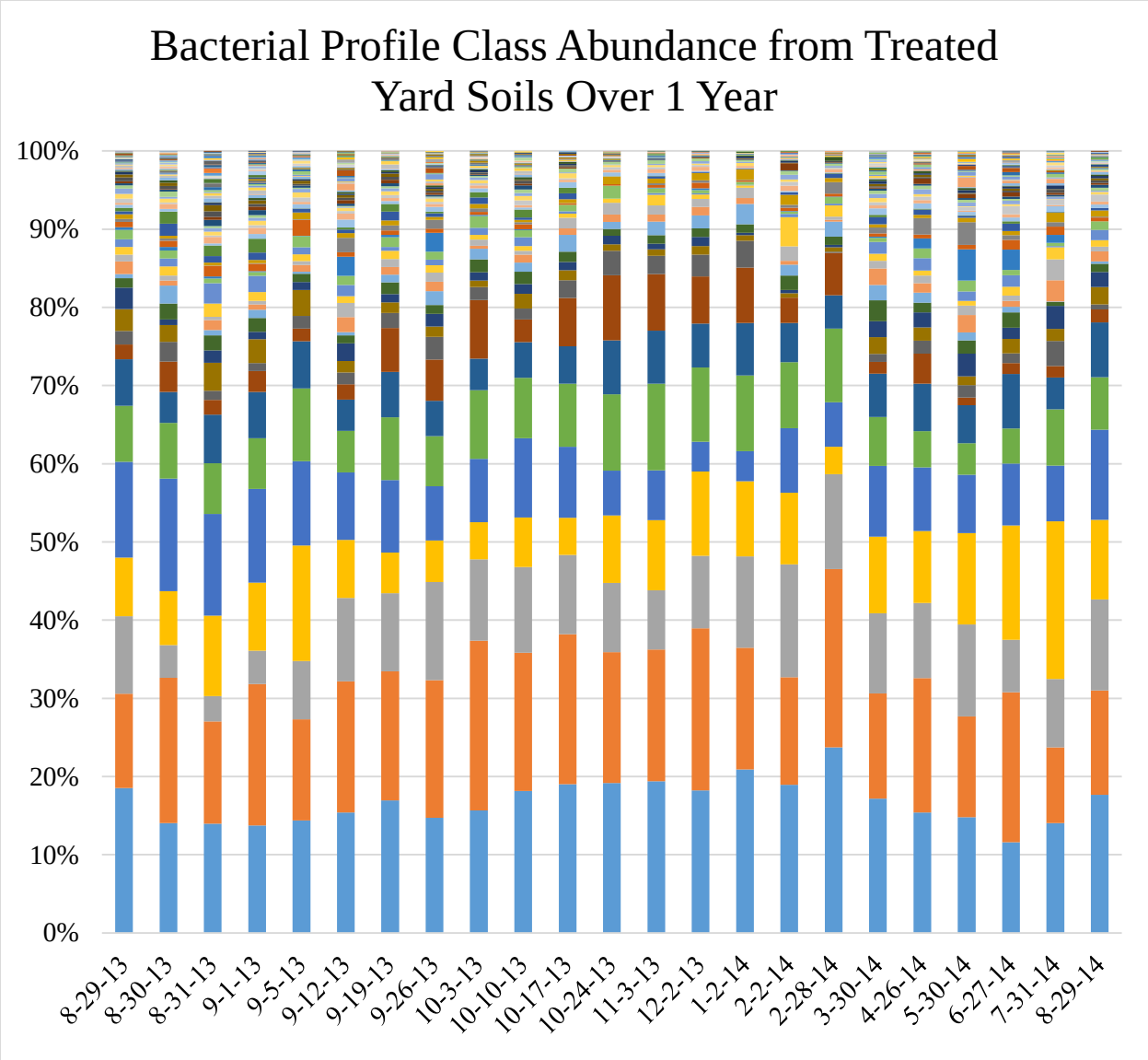

Figure D12—Bacterial class abundance of soil collections from the same location within a treated yard from August 2013 – August 2014. Soils were collected daily for 4 days, weekly for 2 months, and monthly for the remainder of the year so chart is not evenly spaced over time.

Spatial Soil Samples

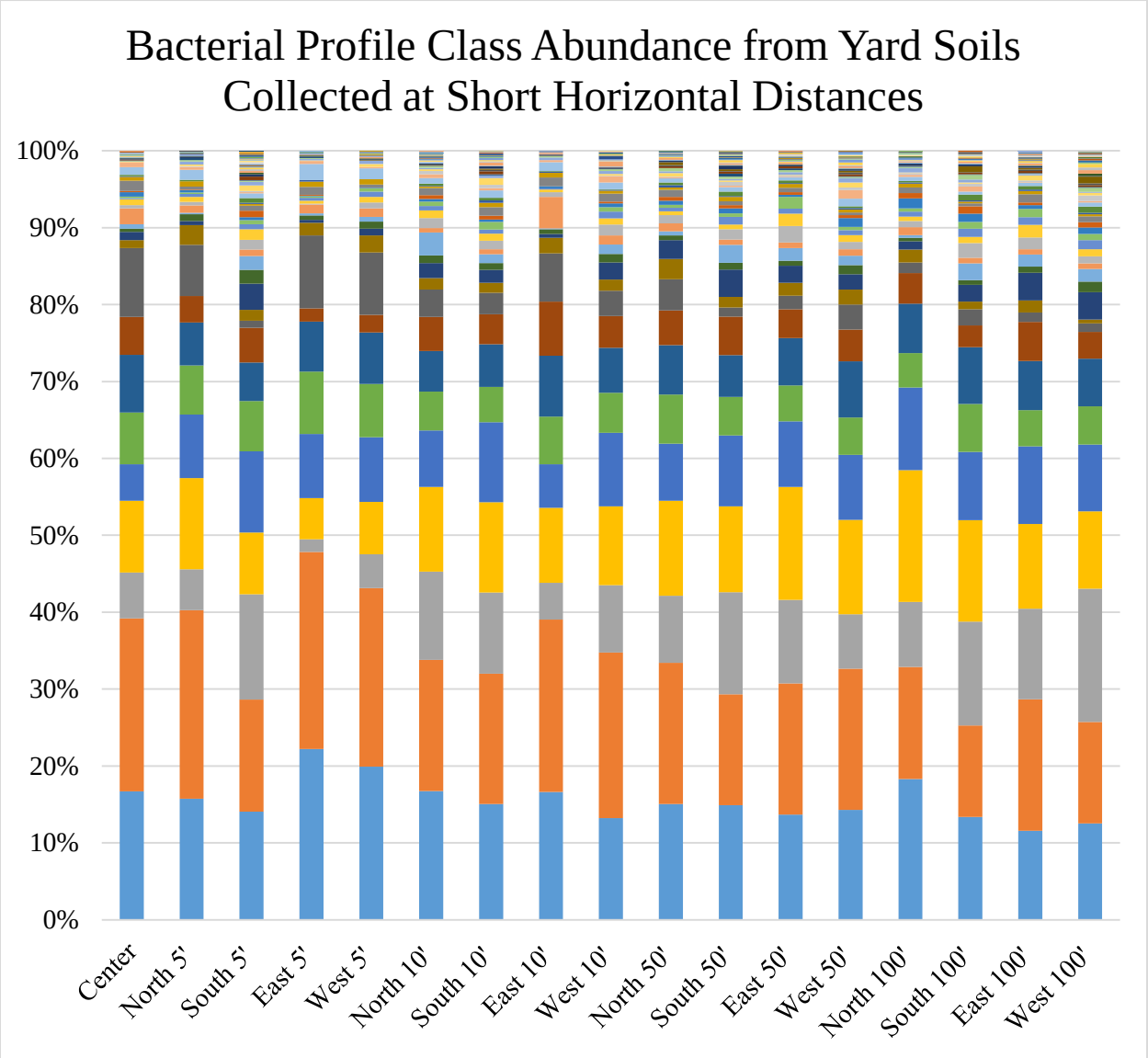

Figure D13—Bacterial class abundance of soil collections across the surface of a yard in March.

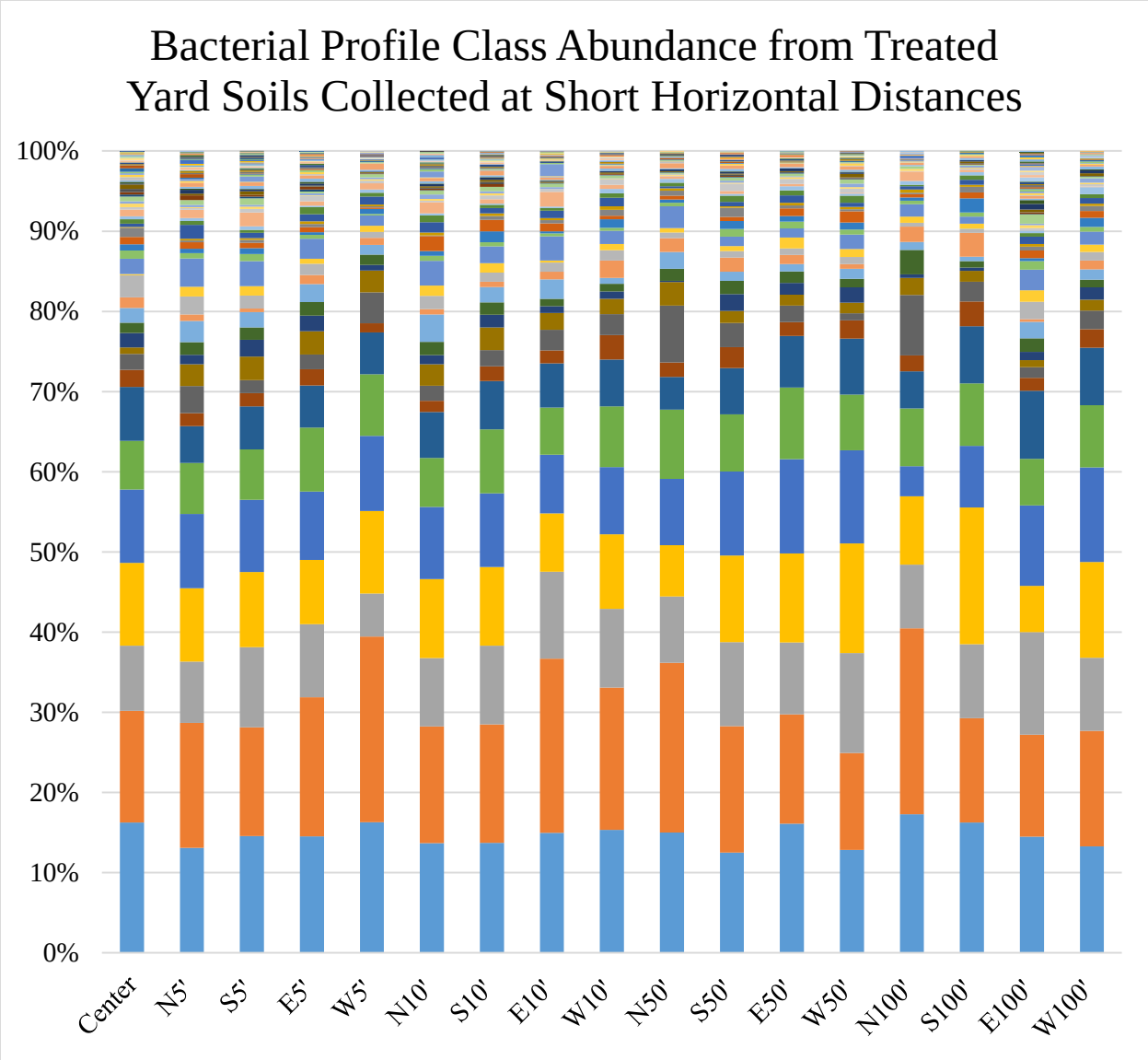

Figure D14—Bacterial class abundance of soil collections across the surface of a treated yard in March.

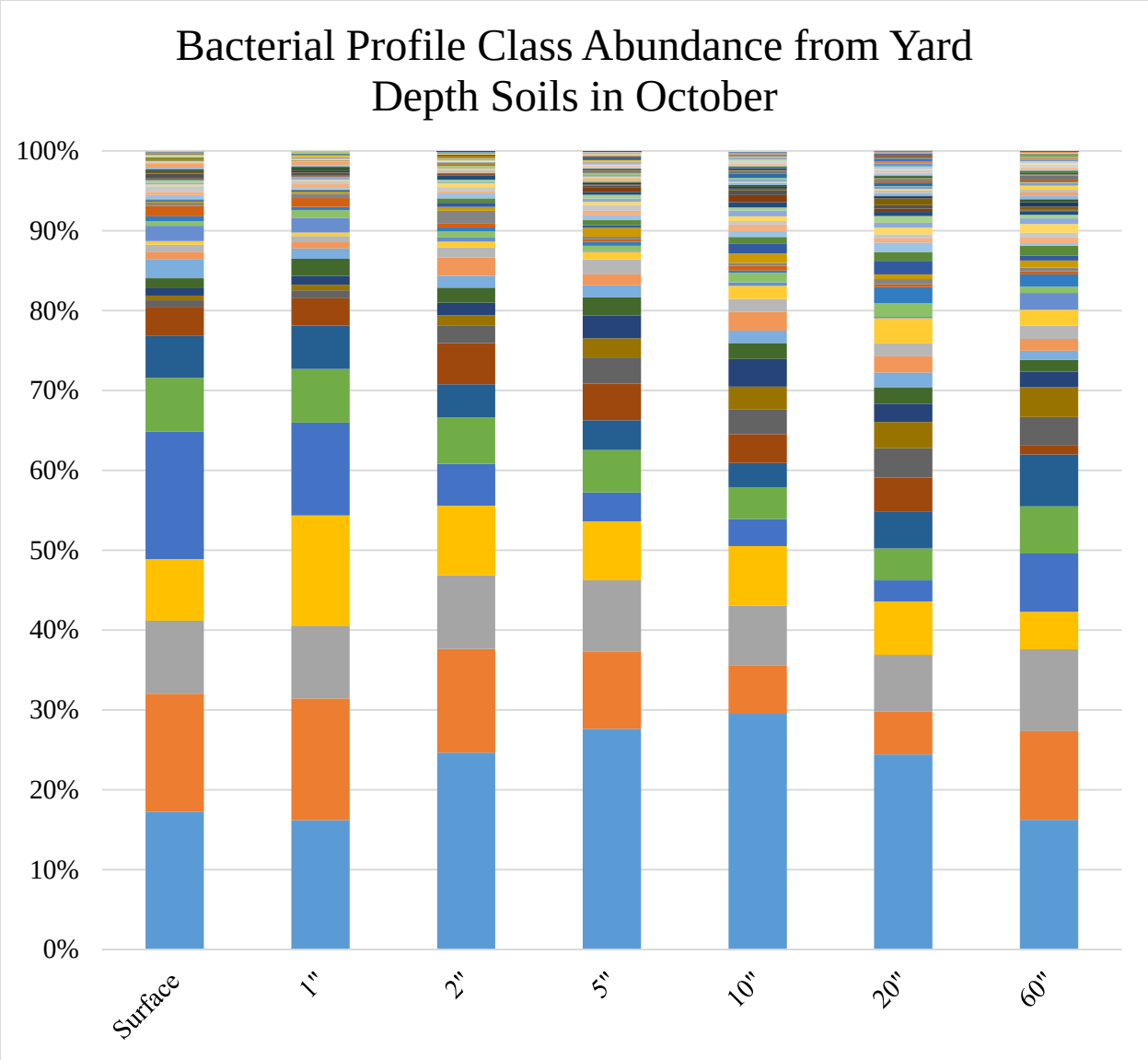

Figure D15—Bacterial class abundance of soil samples collected at different depths within a yard in October.

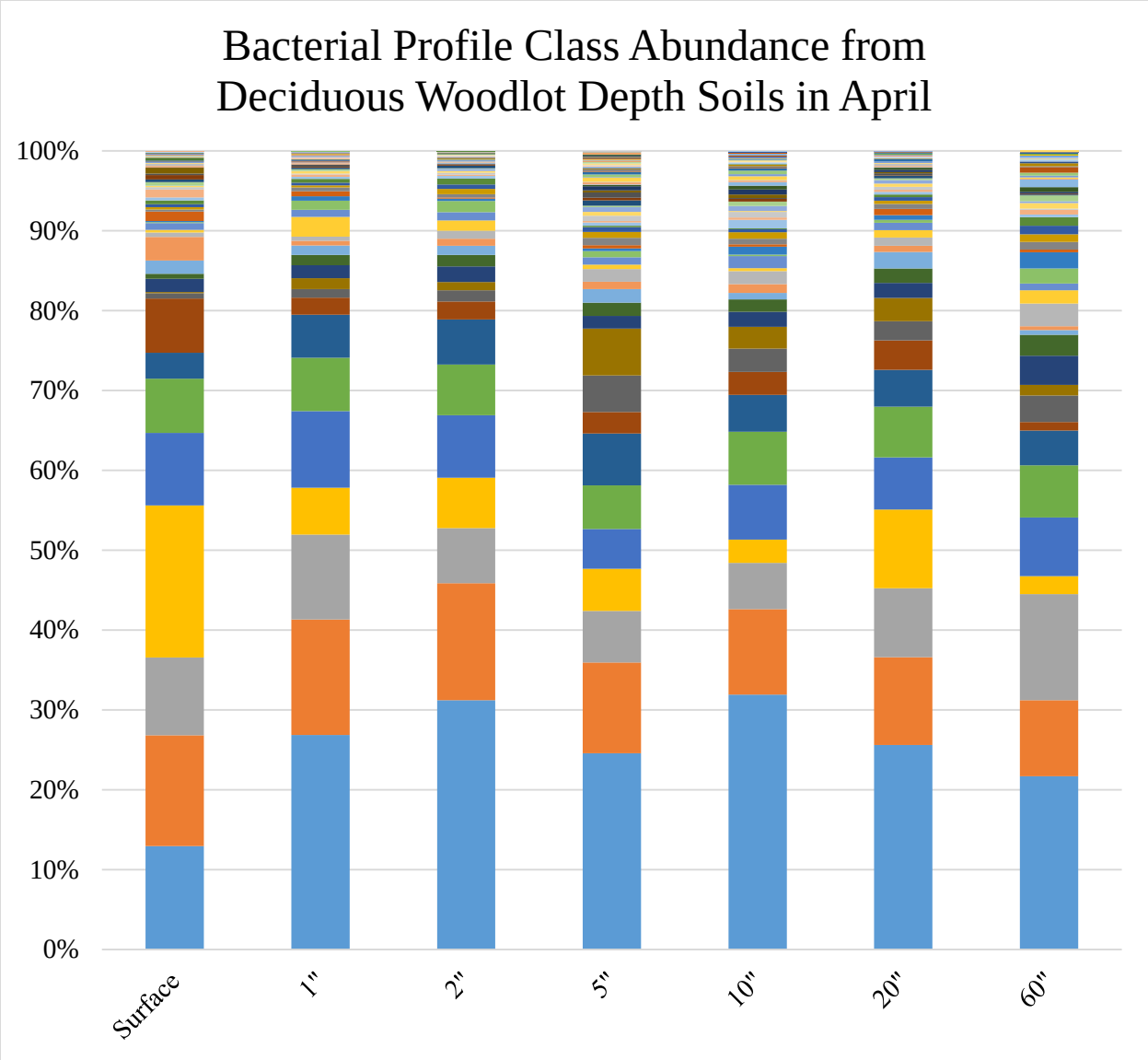

Figure D16—Bacterial class abundance of soil collections at different depths within a deciduous woodlot in April.

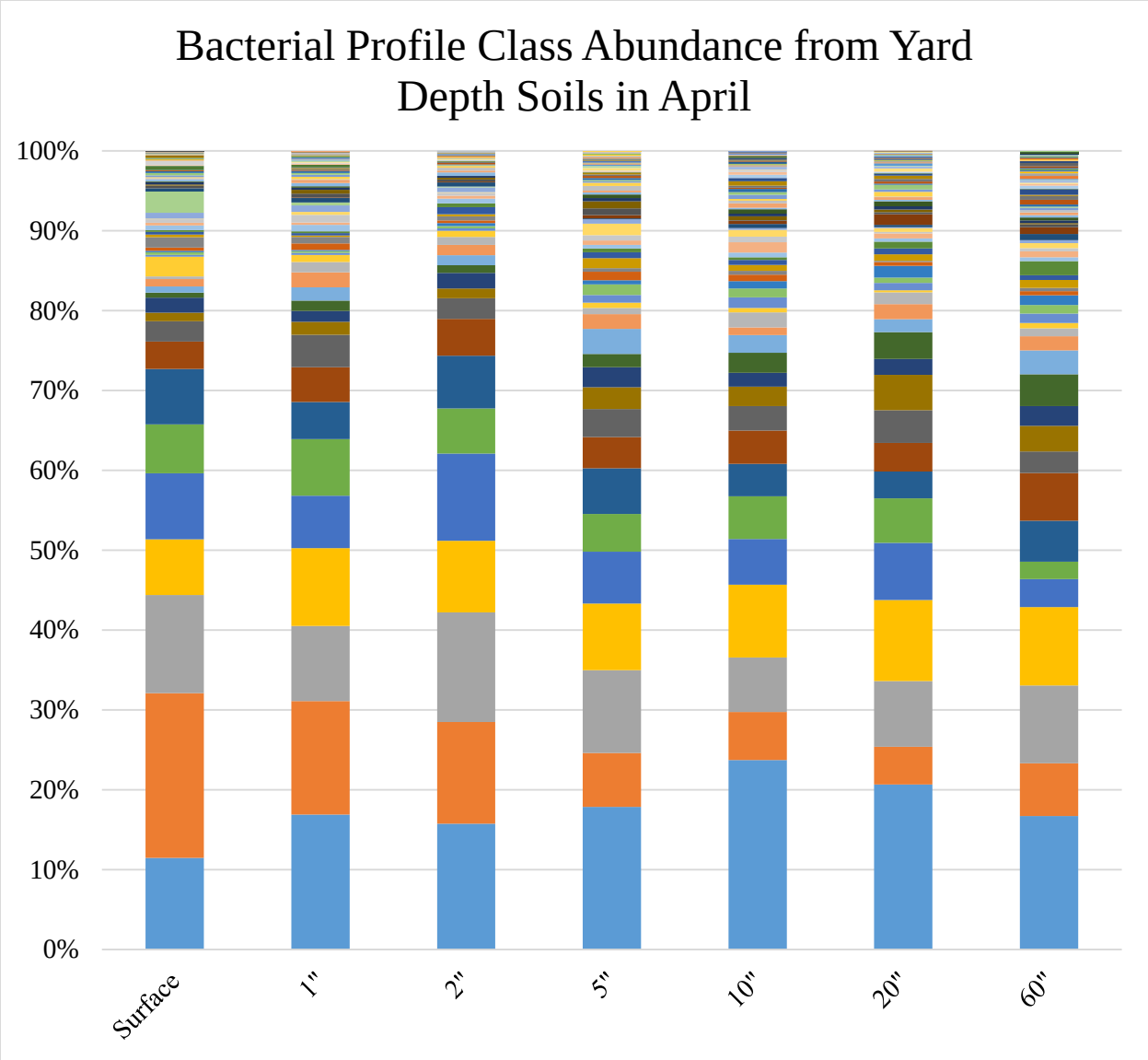

Figure D17—Bacterial class abundance of soil collections at different depths within a yard in April.

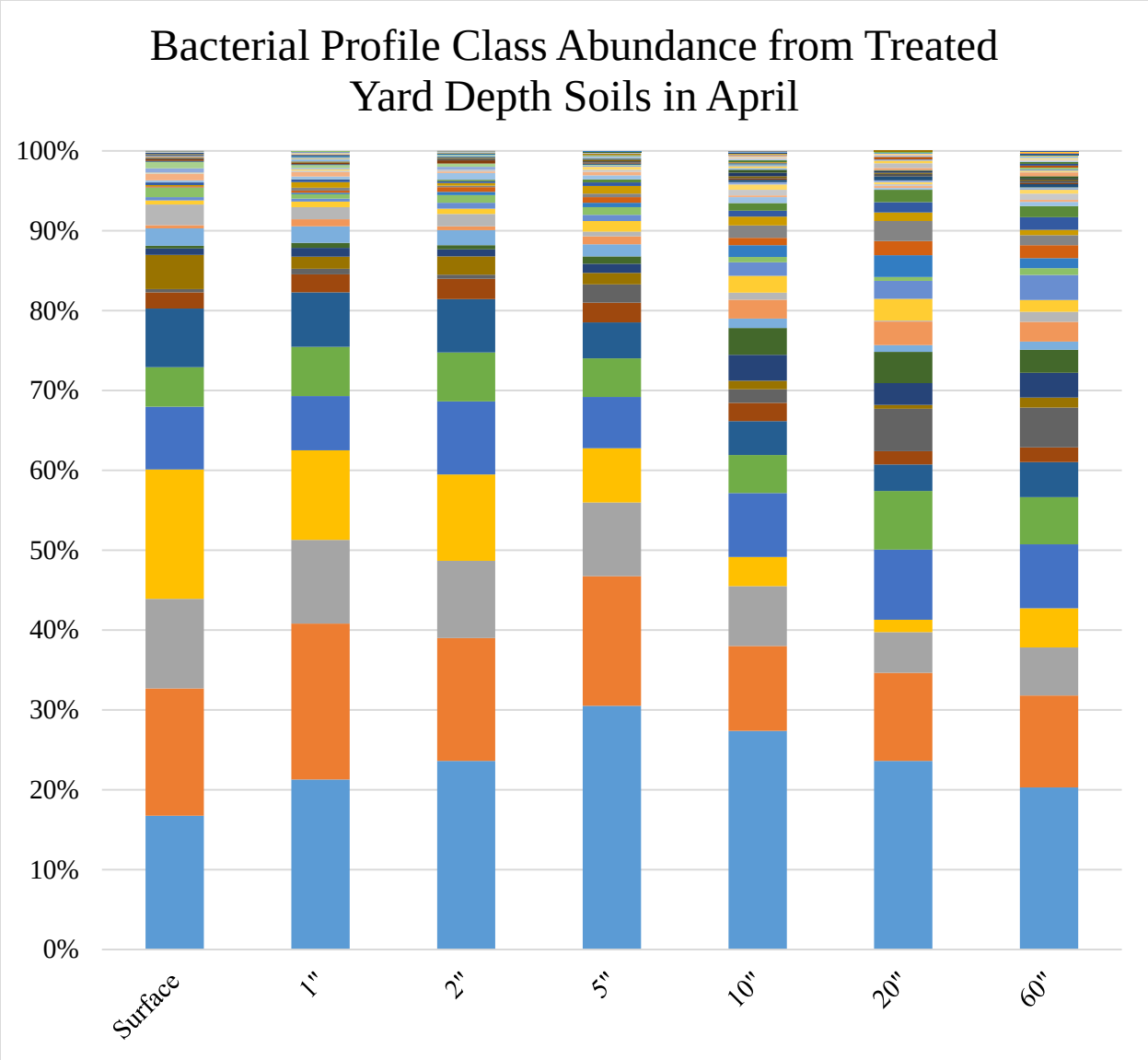

Figure D18—Bacterial class abundance of soil collections at different depths within a treated yard in April.

Evidentiary Soil Samples

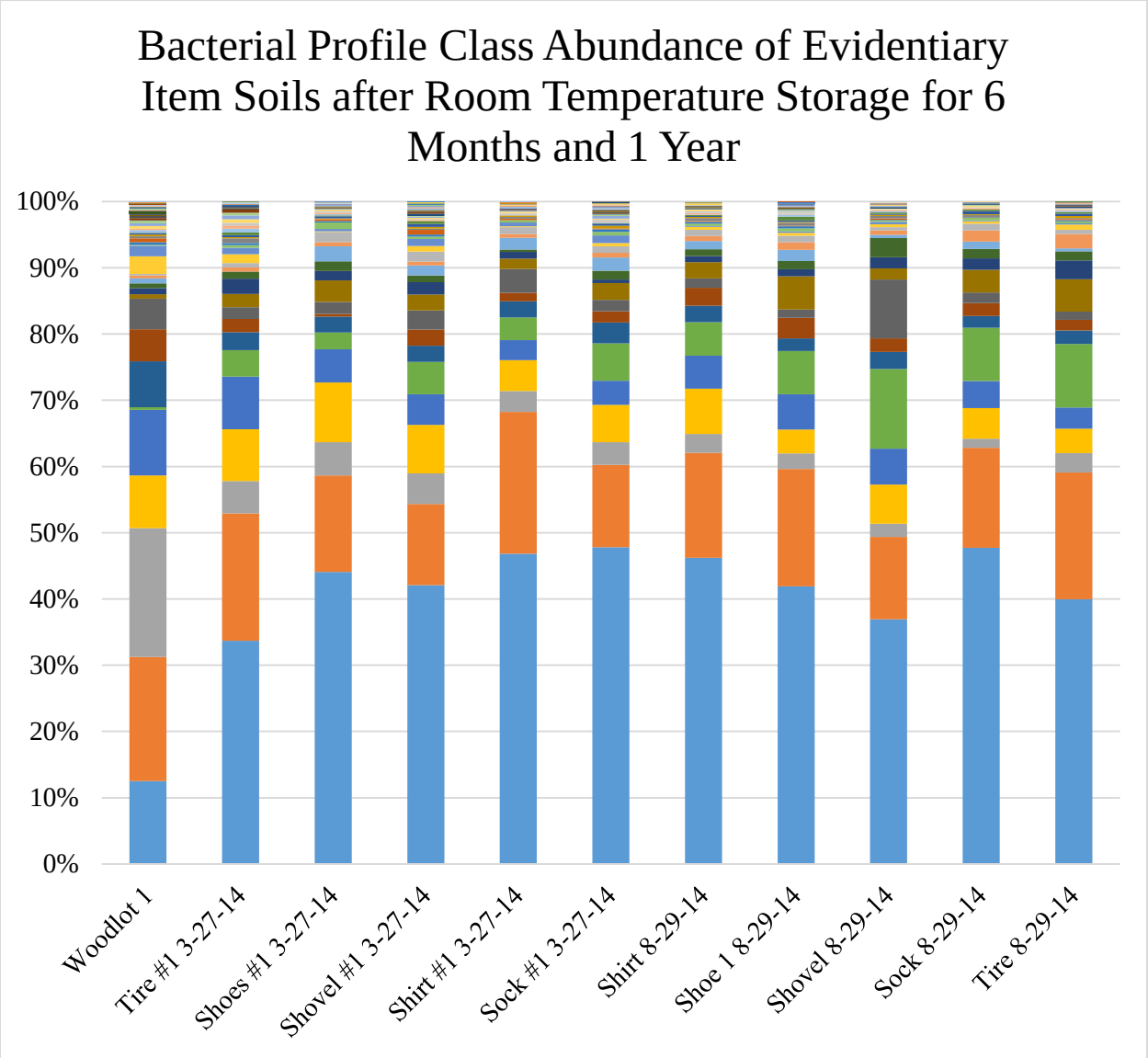

Figure D19—Bacterial class abundance of soil samples collected off of evidentiary items that had been stored at room temperature for 6 months (3-27-14) and 1 year (8-29-14).

APPENDIX E. Additional Nonmetric Multidimensional Scaling Plots

Similar Habitat Soil Samples

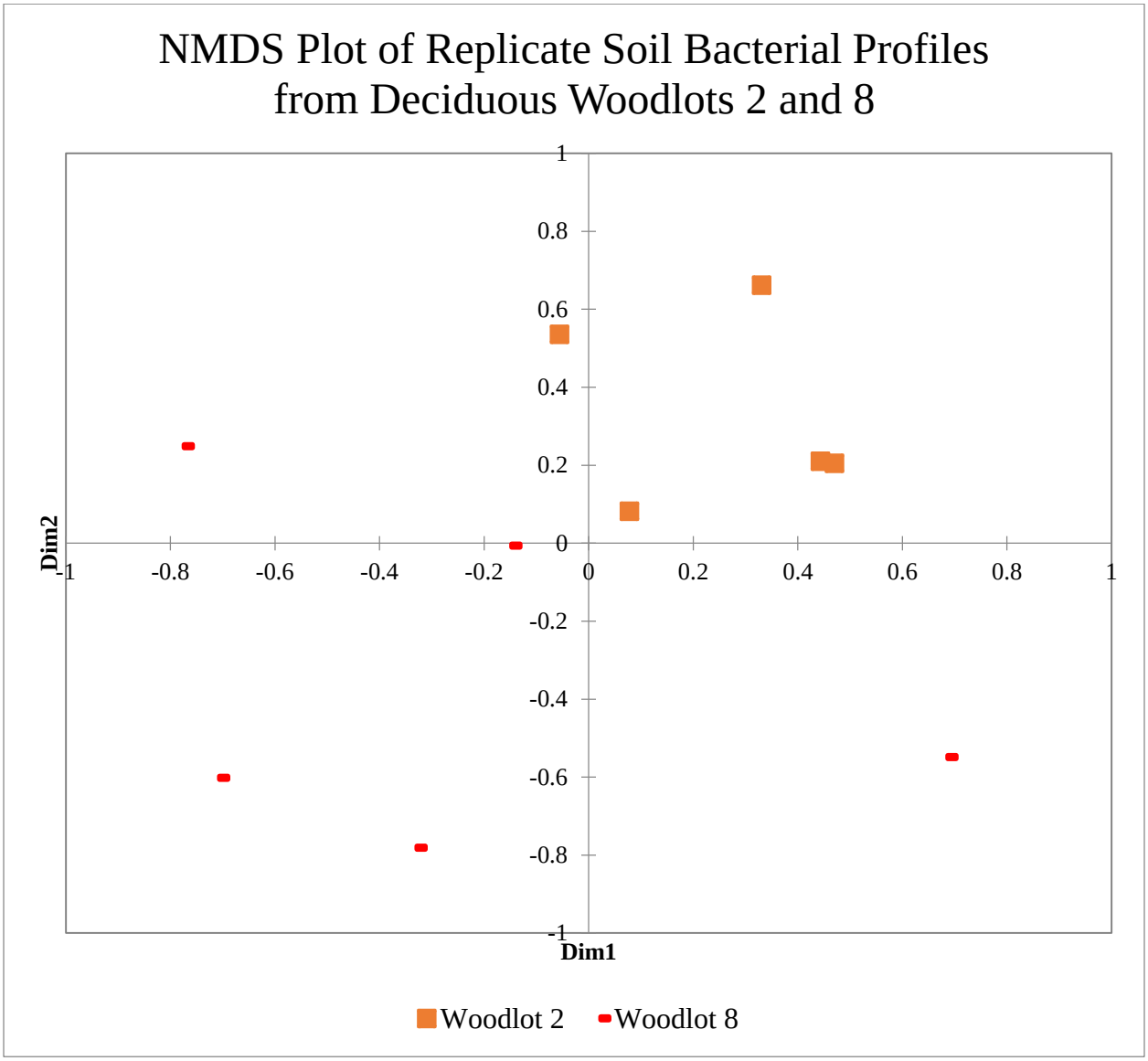

Figure E1—NMDS plot ordinating soil bacterial profiles from deciduous woodlots 2 and 8 over an 8-week period. These profiles were intermingled when all woodlots were ordinated together (Figure 14), but were resolved when they were analyzed alone in a NMDS plot.

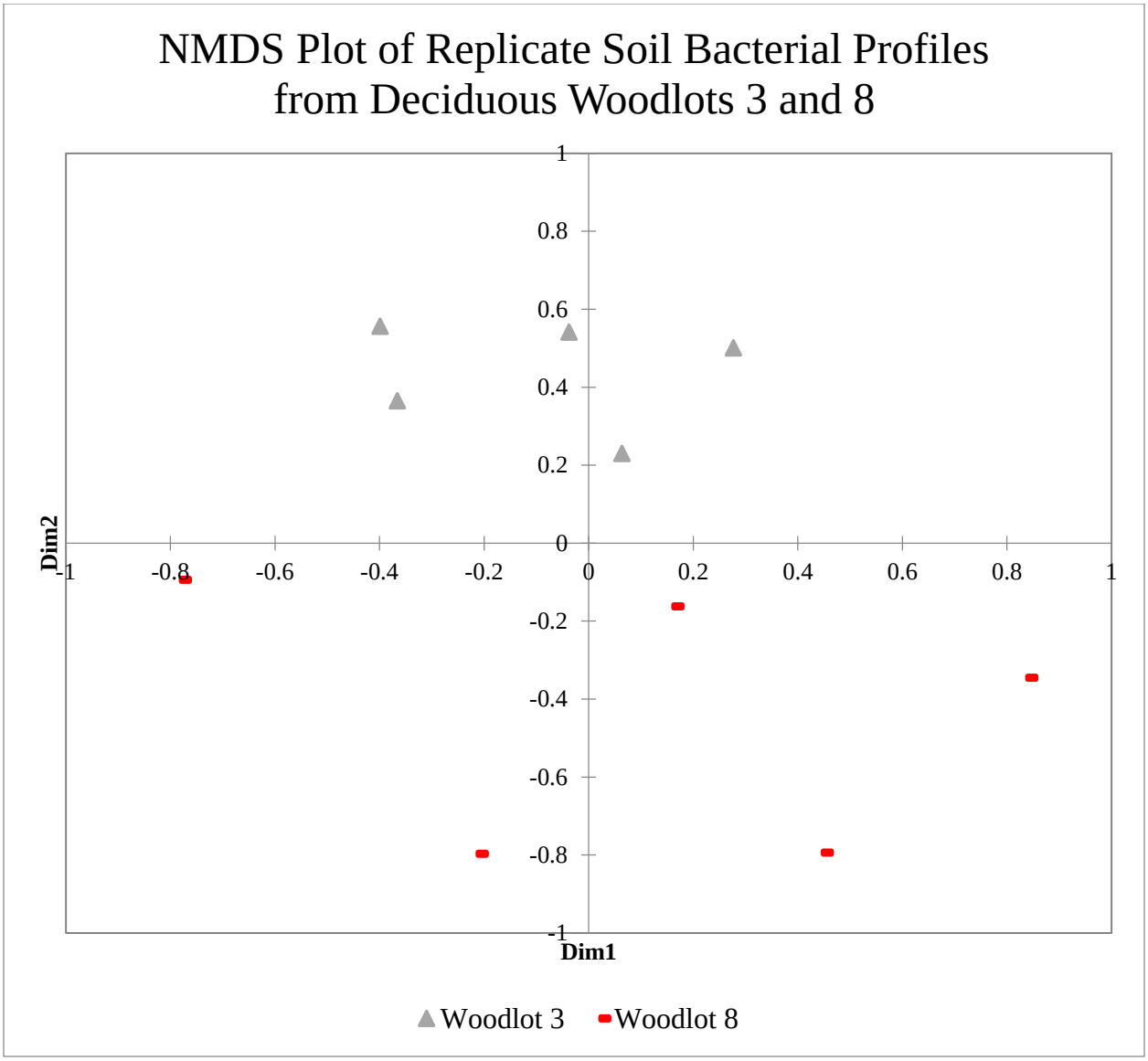

Figure E2—NMDS plot ordinating soil bacterial profiles from deciduous woodlots 3 and 8 over an 8-week period. These profiles were intermingled when all woodlots were ordinated together (Figure 14), but were resolved when they were analyzed alone in a NMDS plot.

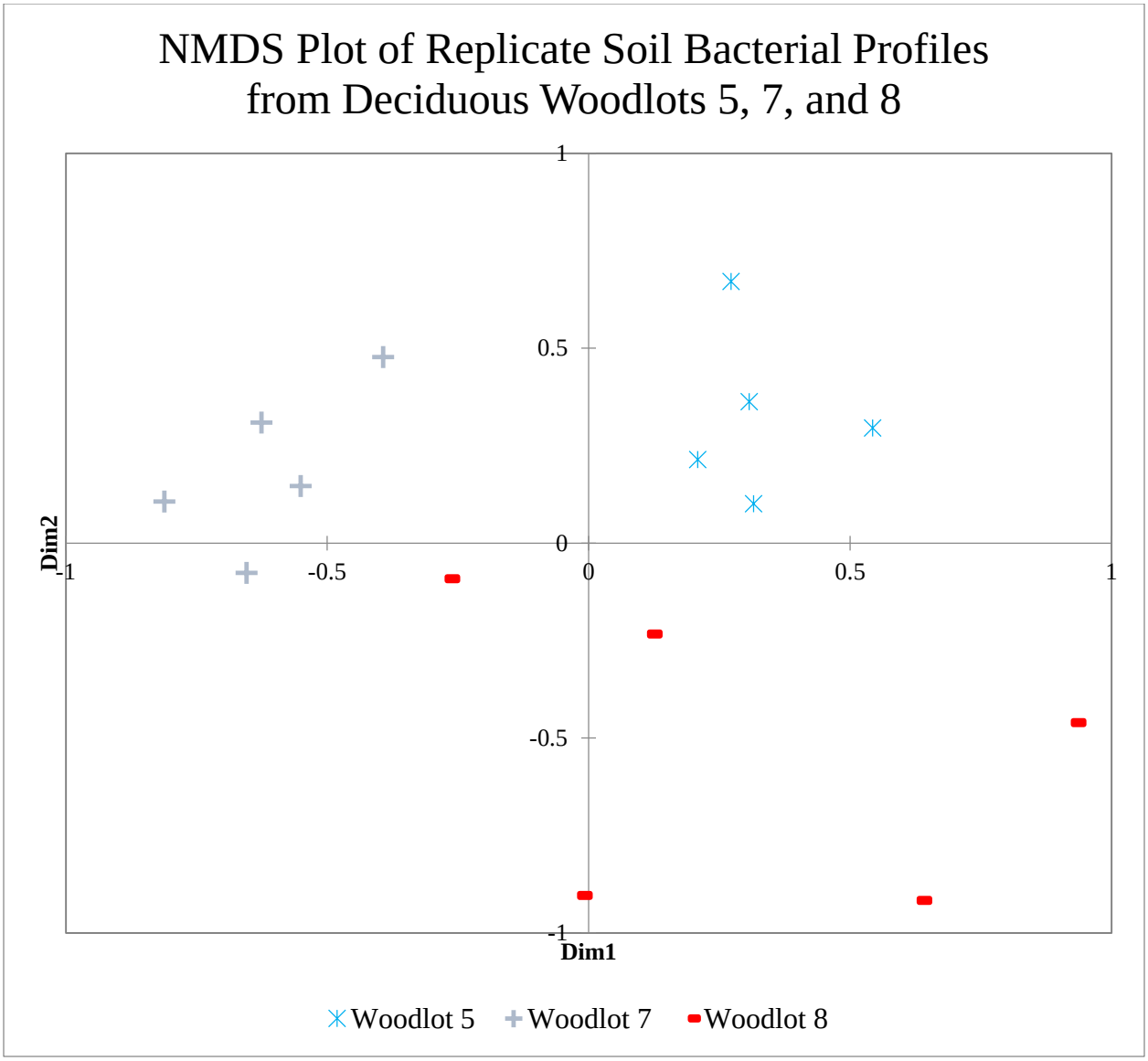

Figure E3—NMDS plot ordinating soil bacterial profiles from deciduous woodlots 5, 7, and 8 over an 8-week period. These profiles were intermingled when all woodlots were ordinated together (Figure 14) but were resolved when they were analyzed alone in a NMDS plot.

Spatial Soil Samples

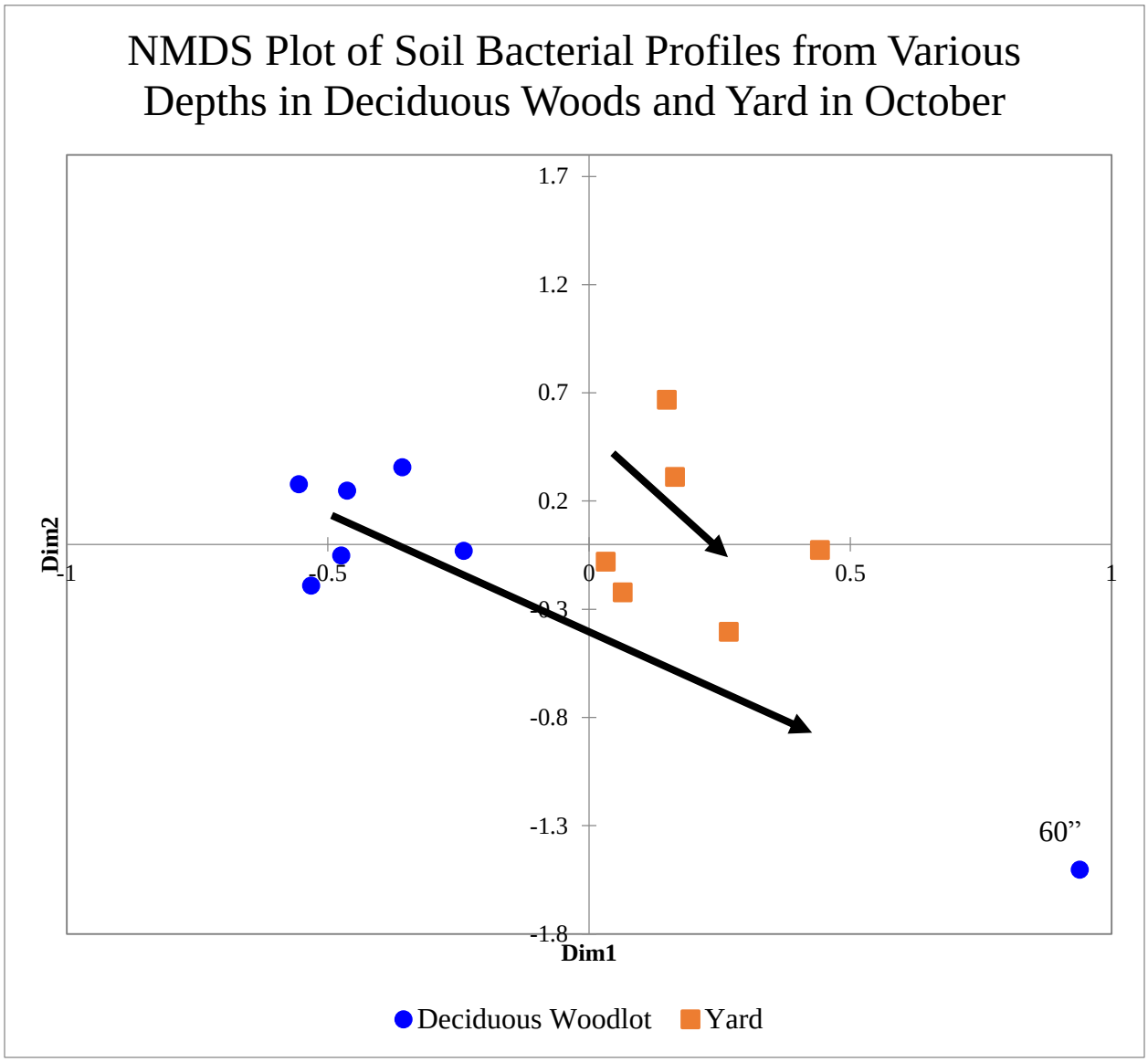

Figure E4—Ordination of soil bacterial profiles from soils at various depths within a deciduous woodlot and yard in October. Profiles from each habitat formed clusters with the exception of the 60 inch collection in the deciduous woodlot (far right, labeled). A trend existed in both habitats where the soil bacterial profiles moved away from the surface profile in multidimensional space as depth increased (arrows point in the direction of increasing depth).

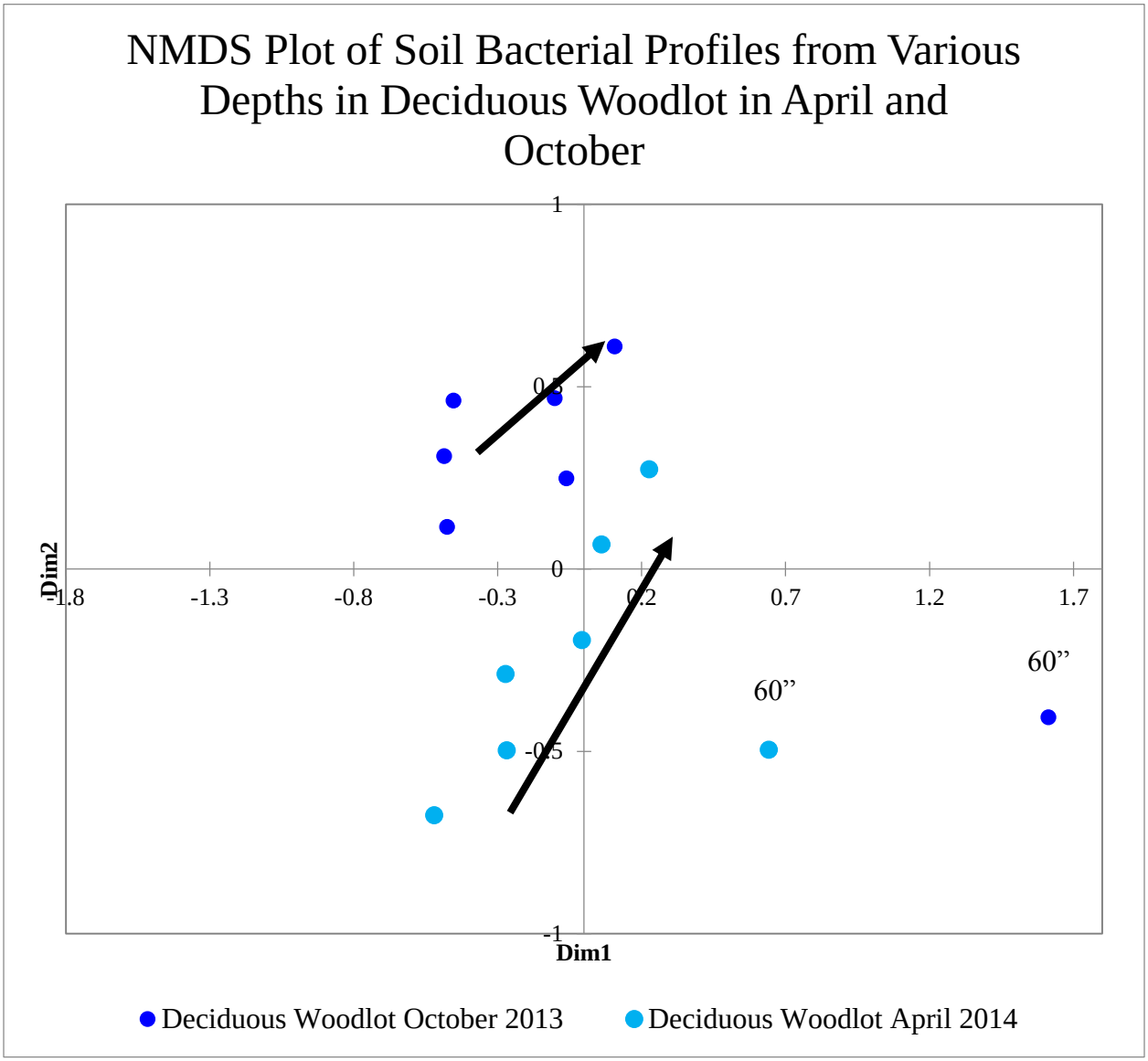

Figure E5—Ordination of soil bacterial profiles from soil collected at various depths within a deciduous woodlot in October 2013 and April 2014. Profiles from each sampling time formed clusters with the exception of the 60 inch profiles (far right, labeled). A trend existed in both months where the soil bacterial profiles moved away from the surface profile in multidimensional space as depth increased (arrows point in the direction of increasing depth).

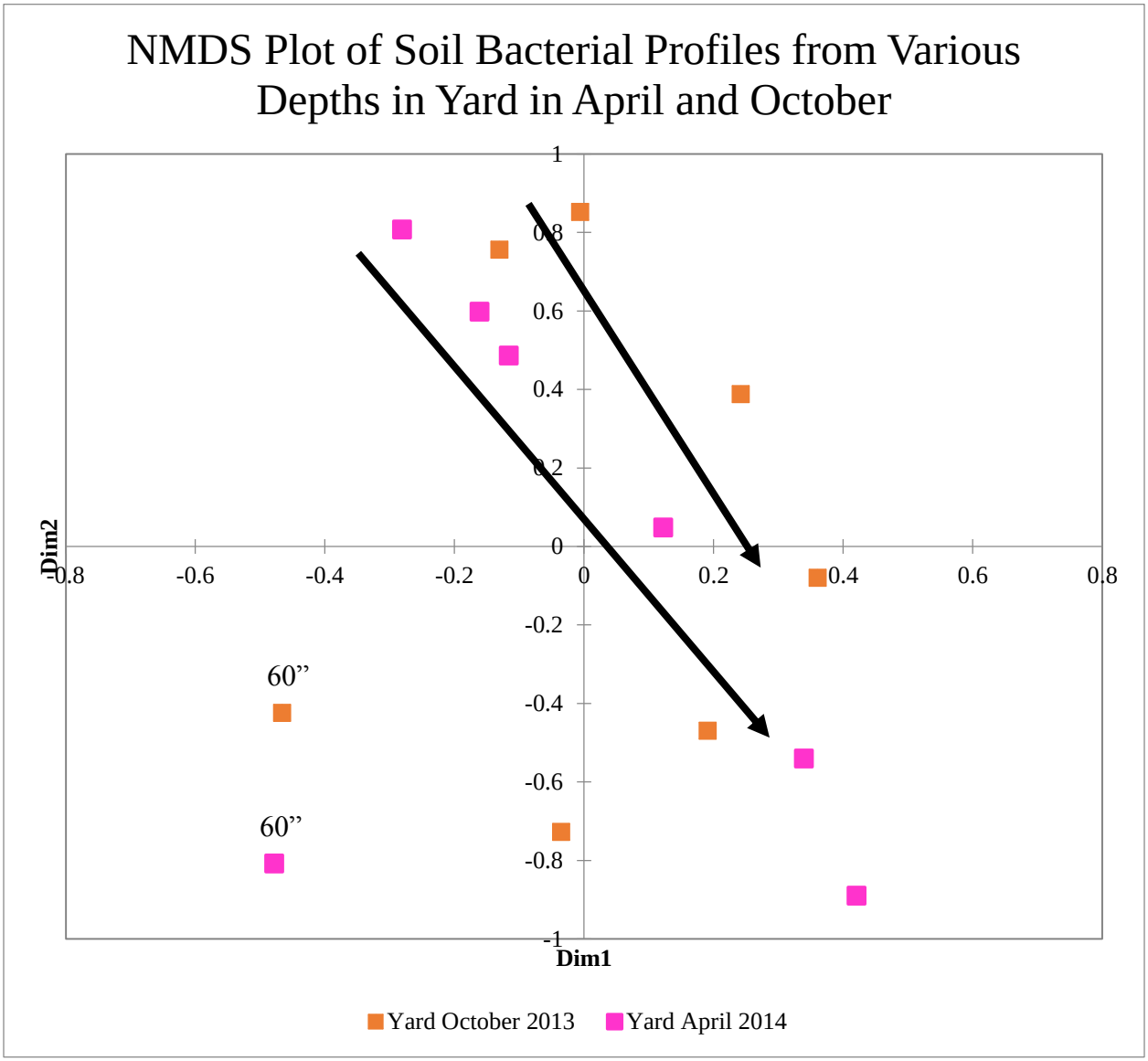

Figure E6—Ordination of soil bacterial profiles from soil collected at various depths within the yard in October and April. Profiles from each sampling time intermingled. A trend existed in both months where the soil bacterial profiles moved away from the surface profile in multidimensional space as depth increased (arrows point in the direction of increasing depth).

T-Shirt Soil Samples

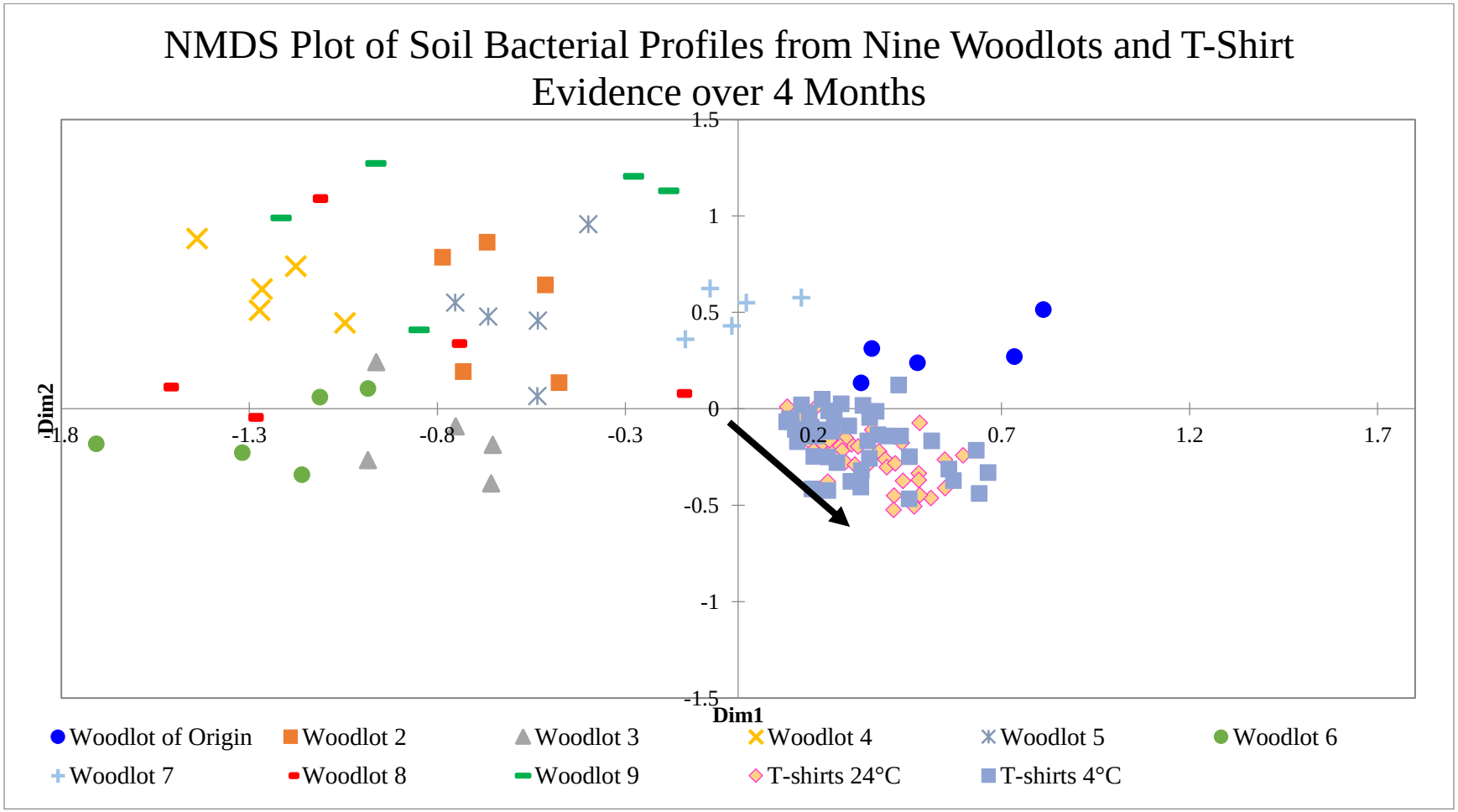

Figure E7—NMDS plot of nine deciduous woodlots and soil profiles generated from t-shirts over four months. T-shirt profiles cluster together near the woodlot of origin cluster. Profiles drifted away from all woodlots over time (in the direction of the arrow).

REFERENCES

REFERENCES

Albuquerque L, Tiago I, Taborda M, Nobre MF, Verissimo A, da Costa MS. Bacillus isabeliae sp. nov., a halophilic bacterium isolated from a sea salt evaporation pond. International Journal of Systematic and Evolutionary Microbiology 2008; 58 (Pt 1): 226 – 30.

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.

Barnard RL, Osborne CA, Firestone MK. Responses of soil bacterial and fungal communities to extreme desiccation and rewetting. The International Society for Microbial Ecology Journal 2013; 7: 2229 – 2241.

Beebe KR, Pell RJ, Seasholtz MB. Chemometrics: A Practical Guide. John Wiley & Sons 1998; 4: 56 – 182.

Bokulich MA, Joseph CL, Allen G, Benson AK, Mills DA. Next-generation sequencing reveals significant backterial diversity of botrytized wine. PLoS ONE 2012; 7 (5): e36357.

Boone DR and Castenholz RW, eds. Bergey's Manual of Systematic Bacteriology, 2nd ed., vol.1. Springer Verlag, New York, 2001.

Bossio DA, Scow KM, Gunapala N, Graham KJ. Determinants of Soil Microbial Communities: Effects of Agricultural Management, Season, and Soil Type on Phospholipid Fatty Acid Profiles. Microbial Ecology 1998; 36: 1 – 12.

Bray JR and Curtis JT. An Ordination of the Upland Forest Communities of Southern Wisconsin. Ecological Monographs 1957; 27(4): 325 – 49.

Bubeck S and von Luxburg U. Nearest Neighbor Clustering: A Baseline Method for Consistent Clustering with Arbitrary Objective Functions. Journal of Machine Learning Research 2009; 10: 657 – 98.

Budowle B, Moretti TR, Niezgoda SJ, Brown BL. "CODIS and PCR-Based Short Tandem Repeat Loci: Law Enforcement Tools". Second European Symposium on Human Identification. Promega Corporation, Madison, Wisconsin 1998; 1998: 73 – 88.

Budowle B. Quality Assurance Guidelines for Laboratories Performing Microbial Forensic Work Produced by the Members of the Scientific Working Group on Microbial Genetics and Forensics (SWGMGF). Forensic Science Communications 2003; 5 (4).

Cao Y, Van De Werfhorst LC, Dubinsky EA, Badgley BD, Sadowsky MJ, Andersen GL, Griffith JF, Holden PA. Evaluation of molecular community analysis methods for discerning fecal sources and human waste. Water research 2013; 47 (18): 6862 – 72.

Caporaso GJ, Lauber CL, Walters WA., Berg-Lyons D, Huntley J, Fierer N, Owens SM, Betley J, Fraser L, Bauer M, Gormley N, Gilbert JA, Smith G, Knight R. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. The ISME Journal 2012; 6: 1621 – 4.

Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Lozupone CA, Turnbaugh PJ, Fierer N, Knight R. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. Proceedings of the National Academy of Sciences USA 2010; 108 (Suppl 1): 4516 – 22.

Catts EP, Goff ML. Forensic Entomology in Criminal Investigations. Annual Review of Entomology 1992; 37: 253 – 72.

Chahouki, MAZ. Dr. Imran Ahmad Dar (Ed.). Multivariate Analysis Techniques in Environmental Science. Earth and Environmental Sciences 2011. ISBN: 978-953-307-468 – 9, InTech, Available from: http://www.intechopen.com/books/earth-and-environmental-sciences/multivariate-analysis-techniques-inenvironmental-science

Chakravorty S, Helb D, Burday M, Connell N, Alland D. A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. Journal of Microbiological Methods 2007; 69 (2): 330 – 9.

Choi SS, Sung-Hyuk C, Tappert CC. A survey of binary similarity and distance measures. Journal of Systemics, Cybernetics and Informatics 2010; 8.1: 43 – 8.

Clarke KR. Non-parametric multivariate analyses of changes in community structure. Australian Journal of Ecology 1993; 18: 117 – 43.

Claus D and Berkeley RCW eds. Bergey's Manual of Systematic Bacteriology, Second Edition, Volume 2. Springer Dordrecht Heidenberg London, New York 2009; 1105 – 39.

Cole JR, Chai B, Farris RJ, Wang Q, Kulam SA, McGarrell DM, Garrity GM, Tiedje JM. The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis. Nucleic Acids Research 2004; 33 (Suppl 1): D294 – 6.

Coomans, D, Massart, DL. Alternative k-Nearest Neighbour Rules in Supervised Pattern Recognition. Analytical Chimica Acta 1982; 136: 15 – 27.

Cover T and Hart P. Nearest neighbor pattern classification. Information Theory, IEEE Transactions 1967; 13 (1): 21 – 7.

Daliresefat SB, Meyer AS, Mirhoseini SZ. Comparison of Similarity Coefficients used for Cluster Analysis with Amplified Fragment Length Polymorphism Markers in the Silkworm, *Bombyx mori.* Journal of Insect Science 2009; 9 (71): 1 – 8.

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).

Dawson L and Hillier S. Measurement of soil characteristics for forensic applications. Surface and Interface Analysis 2010; 42 (5): 363 – 77.

DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P, Andersen GL. Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB. Applied and Environmental Microbiology 2006; 72 (7): 5069 – 72.

Diaz NN, Krause L, Goesmann A, Niehaus K, Nattkemper TW. TACOA—Taxonomic classification of environmental genomic fragments using a kernelized nearest neighbor approach. BMC Bioinformatics 2009; 10: 56.

Dice LR. Measures of the Amount of Ecologic Association between Species. Ecology 1945; 26 (3): 297 – 302.

Egert M and Friedrich MW. Formation of Pseudo-Terminal Restriction Fragments a PCR-Related Bias Affecting Terminal Restriction Fragment Length Polymorphism Analysis of Microbial Community Structure. Applied and Environmental Microbiology 2003; 69 (5): 2555 – 62.

Eichorst SA, Breznak JA, Schmidt TM. Isolation and Characterization of Soil Bacteria That Define Terriglobus gen. nov., in the Phylum Acidobacteria. Applied and Environmental Microbiology 2007; 73 (8): 2708 – 17.

Ettema CH and Wardle DA. Spatial soil ecology. Trends in Ecology and Evolution 2002; 17 (4): 177 – 83.

European Council. Resolution of 25 June 2001 on the exchange of DNA analysis results, 2001.

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

Fitzpatrick RW, Raven MD, Heath M, Rinder G. How soil evidence helped solve a double murder case: A display. 2nd International Conference on Criminal and Enviornmental Soil Forensics Book of Abstracts 2007. Edinburgh, UK. http://www.macaulay.ac.uk/geoforensic/RFitzpatrick.pdf. Accessed 8-3-15.

Fitzpatrick RW. Soil: Forensic Analysis. Wiley Encyclopedia of Forensic Science, 2009. 1 – 14.

Forensic Science Communications. "Regional mtDNA Laboratory Program". FBI 2003; 5 (2).

Gaus K, Rosch P, Petry R, Peschke KD, Ronneberger O, Burkhardt H, Baumann K, Popp J. Classification of lactic acid bacteria with UV-resonance Raman spectroscopy. Biomedical Systems and Microorganisms 2006; 82 (4): 286 – 90.

Ghorbani-Nasrabadi R, Greiner R, Alikhani HA, Hamedi J, Yakhchali B. Distribution of actinomycetes in different soil ecosystems and effect of media composition on extracellular phosphatase activity. Journal of Soil Science and Plant Nutrition 2013; 13 (1): 223 – 36.

Haas BJ, Gevers D, Earl AM, Feldgardent M, Ward DV, Giannoukos G, Ciulla D, Tabbaa D, Highlander SK, Sodergren E, Methe B, DeSantis TZ, The Human Microbiome Consoritum, Petrosino JF, Knight R, Birren BW. Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. Genome Research 2011; 21(3): 494 – 504.

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

Hinchee RE and Leeson A. *Soil Bioinventing: Principles and Practices.* CRC Press 1996.

Holland SM. Non-Metric Multidimensional Scaling (MDS). Department of Geology, University of Georgia, Athens, GA 30602 – 2501, 2008. R Software Tutorial. http://strata.uga.edu/software/pdf/mdsTutorial.pdf

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. The ISME Journal 2010; 4 (6): 829 – 38.

Hopkins JM. Forensic soil bacterial profiling using 16S rRNA gene sequencing and diverse statistics. Published Master's Thesis. Michigan State University, 2014.

Horswell J, Cordiner SJ, Maas EW, Martin TM, Sutherland KB, Speir TW, Nogales B, Osborn AM. Forensic comparison of soils by bacterial community DNA profiling. Journal of Forensic Sciences 2002; 47 (2): 350 – 3.

Hyde ER, Haarmann DP, Lynne AM, Bucheli SR, Petrosino JF. The Living Dead: Bacterial Community Structure of a Cadaver at the End of the Bloat Stage of Decomposition. PLoS One 2013; DOI: 10.1371/journal.pone.0077733.

Ingham County Website. Hawk Island History, 2015. http://pk.ingham.org/ParksTrails/HawkIsland/History.aspx Accessed 6-2-15.

IUSS Working Group WRB. World reference base for soil resources 2006. World Soil Resources Reports 2006; No. 103 FAO, Rome.

Jaccard P. Étude comparative de la distribution florale dans une portion des Alpes et des Jura. Bulletin de la Société Vaudoise des Sciences Naturelles 1901; 37: 547 – 79.

Jansson JK and Tas N. The microbial ecology of permafrost. Nature Reviews Microbiology 2014; 12: 414 – 25.

Johnson RA, Winchern DW. Applied Multivariate Statistical Analysis. 5th ed. Upper Saddle River: Prentice-Hall, 2002.

Jonasson J, Olofsson M, Monstein HJ. Classification, identification and subtyping of bacteria based on pyrosequencing and signature matching of 16S rDNA fragments. Apmis 2002; 110 (3): 263 – 72.

Kenkal NC and Orlóci L. Applying Metric and Nonmetric Multidimensional Scaling to Ecological Studies: Some New Results. Ecology 1986; 67 (4): 919 – 928.

Kravchenko AN, Negassa WC, Guber AK, Hildebrandt B, Marsh TL, Rivers, ML. Intra-aggregate pore structure influences phylogenetic composition of bacterial community in macroaggregates. Soil Science Society of America Journal 2014; 78 (6): 1924 – 39.

Kruskal, JB. Nonmetric Multidimensional Scaling: A Numerical Method. Psychometrika 1964; 29 (2): 115 – 29.

Lauber CL, Ramirez KS, Aanderud Z, Lennon J, Fierer N. Temporal variability in soil microbial communities across land-use types. The ISME Journal 2013; 1 – 10.

Lavine BK and Davidson CE. Gemperline, P ed. Practical Guide to Chemometrics 2nd Edition. CRC Press 2006; 9: 339 – 78.

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

Liu L, Li Y, Li S, Hi N, He Y, Pong R, Lin D, Lu L, Law M. Comparison of Next-Generation Sequencing Systems. Journal of Biomedicine and Biotechnology 2012; 2012: Doi:10.1155/2012/251364.

Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Saheer E. Performance comparison of benchtop high-throughput sequencing platforms. Nature Biotechnology 2012; 30.5: 434 – 9.

Lozupone C and Knight R. UniFrac: A New Phylogenetic Method for Comparing Microbial Communities. Applied and Environmental Microbiology 2005; 71 (12): 8228 – 35.

Luo C, Rodriguez RLM, Johnston E, Wu L, Cheng L, Xue K, Tu Q, Deng Y, He Z, Shi Z, Yuan M, Rebecca S, Li D, Luo Y, Schuur EAG, Chain P, Tiedje J, Zhou J, Konstantinidis K.

Soil microbial community responses to a decade of warming as revealed by comparative Metagenomics. Applied and Environmental Microbiology 2014; 80: 1777 – 86.

Maughan H, Wang PW, Caballero JD, Fung P, Gong Y, Donaldson SL, Yuan L, Keshavjee S, Zhang Y, Yau YCW, Waters VJ, Tullis E, Hwang DM, Guttman DS. Analysis of the Cystic Fibrosis Lung Microbiota via Serial Illumina Sequencing of Bacterial 16S rRNA Hypervariable Regions. PLoS one 2012; 7 (10): e45791.

Meadow JF and Zabinski CA. Spatial heterogeneity of eukaryotic microbial communities in an unstudied geothermal diatomaceous biological soil crust: Yellowstone National Park, WY, USA. FEMS Microbiology Ecology 2012; 82: 182 – 91.

Meyers MS and Foran DR. Spatial and Temporal Influences on Bacterial Profiling of Forensic Soil Samples. Journal of Forensic Sciences 2008; 53 (3): 652 – 60.

Mizrahi-Man O, Davenport ER, Gilad Y. Taxonomic Classification of Bacterial 16S rRNA Genese Using Short Reads: Evaluation of Effective Study Designs. PLoS ONE 2013; 8 (1): e53608. Doi:10.1371/journal.pone.0053608.

Mohri M, Rostamizadeh A, Talwalkar A. *Foundations of Machine Learning.* The MIT Press, 2012. ISBN 9780262018258.

Murray RC and Solebello LP. Saferstein, R ed. *Forensic Science Handbook, Volume 1, Second Edition.* Chapter 11: Forensic Examination of Soil. Pearson Education Inc, 2002. 11: 615 – 33.

Murray RC and Tedrow JC. Forensic Geology—Earth Sciences and Criminal Investigation. National Criminal Justice Reference Service, 1975: NCJ 017479.

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

Nelson EL. Estimation of short-term postmortem interval utilizing core body temperature: a new algorithm. Forensic Science International 2000; 109: 31 – 8.

O'Brien RW and Morris JG. Oxygen and the Growth and Metabolism of *Clostridium acetobutylicum.* Microbiology 1971; 68 (3): 307 – 18.

Oren A. Clostridium lortetii sp. nov., a halophilic obligatory anaerobic bacterium producing endospores with attached gas vacuoles. Archives of Microbiology 1983; 136: 42 – 8.

Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, Bertoni A, Swerdlow HP, Gu Y. A tale of three next-generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BioMed Central Genomics 2012; 13: 341.

Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Research 2013; 41 (D1): D590 – 6.

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.

Ramette A. Multivariate analyses in microbial ecology. FEMS Microbiology Ecology 2007; 62 (2): 142 – 60.

Rokach L and Maimon O. *Data Mining with Decision Trees: Theory and Applications*. World Science Pub Co Inc., 2008; Hackensack, NJ.

Ruffell A. Forensic pedology, forensic geology, forensic geoscience, geoforensics and soil forensics. Forensic Science International 2010; 202 (1 – 3): 9 – 12.

Saferstein RE. *Forensic Science Handbook*. Prentice Hall, 2002; New Jersey, 784 pp.

Sato Y, Willis BL, Bourne DG. Pyrosequencing-based profiling of archaeal and bacterial 16S rRNA genes identifies a novel archaeon associated with black band disease in corals. Environmental Microbiology 2013; 15 (11): 2994 – 3007.

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

Schloss PT, Larget BR, Handelsman J. Integration of microbial ecology and statistics: a test to compare gene libraries. Applied and Environmental Microbiology 2004; 70 (9): 5485 – 92.

Schramm A, Beer D, Heuvel JC, Ottengraf S, Amann R. Microscale Distribution of Populations and Activities of *Nitrosospira* and *Nitrospira* spp. Along a Macroscale Gradient in a Nitrifying Bioreactor: Quantification by In Situ Hybridization and the use of Microsensors. Applied and Environ Microbiology 1999; 65 (8): 3690 – 6.

Scientific American. Curious use of the microscope. Science and Art 1856; 11: 240.

Sensabaugh GF. Microbial Community Profiling for the Characterization of Soil Evidence: Forensic Considerations. Criminal and Environmental Soil Forensics. Springer Netherlands 2009; 49 – 60.

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

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.

Sorokin, DY, Kovaleva OL, Tourova TP, Muyzer G. *Thiohalobacter thiocyanaticus* gen. nov., sp. nov., a moderately halophilic, sulfur-oxidizing gammaproteobacterium from hypersaline lakes, that utilizes thiocyanate. International Journal of Systematic and Evolutionary Microbiology 2010; 60: 444 – 50.

Sutton NB, Maphosa F, Morillo JA, Al-Soud WA, Langenhoff AAM, Grotenhuis T, Rijnaarts HHM, Smidt H. Impact of Long-Term Diesel Contamination on Soil Microbial Community Structure. Applied and Environ Microbiology 2013; 79 (2): 619 – 30.

Thomson BC, Ostle N, McNamara N, Bailey MJ, Whiteley AS, Griffiths RI. Vegetation Affects the Relative Abundances of Dominant Soil Bacterial Taxa and Soil Respiration Rates in an Upland Grassland Soil. Microbial Ecology 2010; 59 (2): 335 – 43.

Tsitko I and Bomberg M. The Diversity of Bacterial Community in Lastensuo Bog Characterized by DNA-Based 16S rRNA-gene targeted 454 Pyrosequencing. POSIVA Working Report 2014; 2014 – 34.

Tukey JW. Bias and confidence in not-quite large samples. The Annals of Mathematical Statistics 1958; 29: 614 – 5.

van Elsas JD and Boersma FGH. A review of molecular methods to study the microbiota of soil and the mycosphere. European Journal of Soil Biology 2011; 47 (2): 77 – 87.

Ventosa A, Nieto JJ, Oren A. Biology of Moderately Halophilic Aerobic Bacteria. Microbiology and Molecular Biology Reviews 1998; 62 (2): 504 – 44.

Ward JH. Hierarchical Grouping to Optimize an Objective Function. Journal of the American Statistical Association 1963; 58 (301): 236 – 44.

Ward TL, Hosid S, Ioshikhes I, Altosaar I. Human mild metagenom: a functional capacity analysis. BMC Microbiology 2013; 13 (1): 116.

Will C, Thurmer A, Wollherr A, Nacke H, Herold N, Schrumpf M, Gutknecht J, Wubet T, Buscot F, Daniel R. Horizon-Specific Bacterial Community Composition of German Grassland Soils, as Revealed by Pyrosequencing-Based Analysis of 16S rRNA Genes. Applied and Environmental Microbiology 2010; 76 (20): 6751 – 9.

Wittaker RH. Dominance and Diversity in Land Plant Communities: Numerical relations of species express the importance of competition in community function and evolution. Science 1965; 147 (3655): 250 – 260.

Woese CR and Fox GE. Phylogenetic structure of the prokaryotic domain: The primary kingdoms. Proceedings of the National Academy of Sciences USA 1977; 74 (11): 5088 – 90.

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 Microbial Methods 2006; 65 (1): 49 – 62.

Yarza P, Yilmaz P, Pruesse E, Glockner FO, Ludwig W, Schleifer KH, Whitman WB, Euzeby J, Amann R, Rossello-Mora R. Uniting the classificaiton of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. Nature Reviews Microbiology 2014; 12: 635 – 45.

Young JM, Weyrich LS, Breen J, Macdonald LM, Cooper A. Predicting the origin of soil evidence: High throughput eukaryote sequencing and MIR spectroscopy applied to a crime scene scenario. Forensic Science International 2015; 251: 22 – 31.

Young JM, Weyrich LS, Cooper A. Forensic soil DNA analysis using high-throughput sequencing: A comparison of four molecular markers. Forensic Science International: Genetics 2014; 13: 176 – 84.

Yu Z and Morrison M. Comparisons of Different Hypervariable Regions of rrs Genes for Use in Fingerprinting of Microbial Communities by PCR-Denaturing Gradient Gel Electrophoresis. Applied and Environ Microbiology 2004; 70 (8): 4800 – 6.

Zhang H, Berg AC, Marie M, Malik J. SVM-KNN: Discriminative nearest neighbor classification for visual category recognition. Computer Vision and Pattern Recognition 2006; 2: 2126 – 36.