AN ABSTRACT OF THE DISSERTATION OF

Tuan N.T. Pham for the degree of Doctor of Philosophy in Computer Science presented on June 4, 2013.

Title: Interactive Visualization of Diversity in Multivariate Data Sets Unified across Fields of Study

Abstract approved: ______________________________

Ronald A. Metoyer

The study of the diversity of multivariate objects shares common characteristics across disciplines, including ecology and organizational management. Nevertheless, experts in these two disciplines have adopted somewhat separate diversity concepts and analysis techniques, limiting the ability of potentially sharing and cross comparing these concerns. Moreover, while complex diversity data may benefit from exploratory data analysis, most of the existing techniques emphasize confirmatory analysis based on statistical metrics and models. To bridge these gaps, interactive visualization is especially appealing because of its potential to allow users to explore diversity data in a direct and holistic way, prior to further statistical analysis.

This dissertation addresses the problem of designing multivariate visualizations that support exploration and communication of diversity patterns and processes in multivariate data. To this aim, the dissertation presents design considerations as well as implementation and evaluation of interactive visualizations targeting diversity analysis. The contributing visualization techniques and tools include (1) *Diversity Map*—a novel multivariate space-filling representation emphasizing diversity patterns in separate attributes; (2) *Ecological Distributions and Trends Explorer (EcoDATE)*—a web-based visual-analysis tool that is built upon the Diversity Map and facilitates the exploratory analysis of long-term ecological data with an emphasis on distribution patterns and temporal trends; and (3) *HIST*—a visual representation for communicating team diversity faultlines across multiple attributes that is based on multiple linked, stacked histograms. Further, drawing upon lessons from these designs, this dissertation cross compares analyses of species diversity (ecology), microbial diversity (microbiology), and workgroup diversity (organizational management) and introduces a *unified taxonomy of analytical tasks* to guide the creation and evaluation of future diversity visualizations. The design considerations, visualization techniques, tools, and task taxonomy are evaluated and refined in empirical user studies involving human participants and subject-matter experts.

Interactive Visualization of Diversity in Multivariate Data Sets Unified across Fields of Study

by

Tuan N.T. Pham

A DISSERTATION

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Doctor of Philosophy

Presented June 4, 2013
Commencement June 2014

Doctor of Philosophy dissertation of Tuan N.T. Pham presented on June 4, 2013.

APPROVED:

Major Professor, representing Computer Science

Director of the School of Electrical Engineering and Computer Science

Dean of the Graduate School

I understand that my dissertation will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my dissertation to any reader upon request.

Tuan N.T. Pham, Author

ACKNOWLEDGEMENTS

If you want to go quickly, go alone. If you want to go far, go together.

—African proverb

This dissertation is the result of a truly cross-disciplinary collaboration. I could not have gone this far and finished it without the support and collaboration of many people.

First, I would like to express my deep gratitude to my three mentors, Professor Ron Metoyer, Professor Julia Jones, and Professor Toshimi Minoura, for their generous support and guidance over the years. Professor Metoyer, my supportive major professor, has overseen this research from the beginning with great enthusiasm and vision. He is truly my role model of a great researcher, teacher, leader, in addition to a man of family. Professor Jones has brought to this work deep insights and knowledge about ecology and the scientific world. She has been incredibly generous in giving her precious time to help with the project. Lastly, Professor Minoura, my former advisor, has given me the opportunity to start my journey as a graduate student in Computer Science. He was the first to train me to write and think critically as a researcher.

I am grateful to other OSU professors who have served on my PhD committee: Professor Margaret Burnett, Professor Carlos Jensen, and Professor James Coakley (GCR) for their encouragement and input. This dissertation has greatly benefited from their constructive feedback on the evaluation and concrete outcomes of the work. In addition, Professor Eugene Zhang has served as both a committee member and collaborator on the publication of the Diversity Map representation. Many difficult aspects of my life as an international student have gotten easier thanks to his thoughtful advice. My gratitude is also extended to Professor Christopher Scaffidi for the feedback on my program of study; to other visualization researchers, who served as panelists at the doctoral colloquium at the VisWeek conference in 2010, for constructive feedback on my early thesis proposal.

This research started as an InfoVis course project and I was fortunate to work with two wonderful teammates and friends: Rob Hess and Crystal Ju, who contributed to the development and evaluation of the Diversity Map representation. I would also like to thank Onyekwere Ogba and Nicholas Hubbert, the two summer DREU interns, for their development efforts.

I am also fortunate to get involved with the H.J. Andrews LTER (HJA) community whose nature-loving members and activities have made my graduate school experience more rich and meaningful. Many thanks to Professor Jeff Miller, Steven Highland, and Don Henshaw for their contribution to the visualization of the moth data set; to Professor Frederick Swanson and Robert Pabst for their contribution to the evaluation of the EcoDATE tool. I want to extend my thanks to Lina DiGregorio, Mark Schulze, Suzanne Remillard, and Theresa Valentine for offering me the resources in conducting this project. Our work on visualization of biodiversity data is generously funded by the H.J. Andrews Experimental Forest Graduate Student GRA Support Award.

In addition to ecology, this work intersects with microbiology and organizational management. I genuinely appreciate Professor Rick Colwell and his students for their valuable and informative discussions on visualization of microbial diversity data. I owe Professor Katerina Bezrukova and Professor Chester Spell for contributing to the design and evaluation of our visualization of team faultlines. The enthusiastic and insightful feedback from Professor Bezrukova has deepened my understanding of workgroup diversity. I am also thankful for the input from the participants and other domain experts on our studies.

Besides this dissertation, my involvement with other groups and projects has been invaluable. I am proud to be a member of the Information Visualization Research Group at OSU among other fellow students including Karl Smeltzer, Nels Oscar, Josie Hunter, Islam Almusaly, Sheena Ellenburg, Andrew Atkinson, Catharina Vijay, and many others. It has also been my pleasure and honor to be a part of the Personal Understanding of Life and Social Experiences (PULSE) project under the considerate guidance of Professor Karen Hooker and Professor Ronald Metoyer and the enjoyable teamwork with Shannon Mejía and Soyoung Choun. I also value the opportunity to contribute my development efforts to the Take Charge programs, led and run by the Energize Corvallis Program Director, Carly Lettero and her volunteers. In addition, my understanding of software development in industry was greatly enhanced during my internship at the SAP Labs. My special thanks to Andreas Vogel and the Raptor team for creating a welcoming work environment and for widening my perspective.

Considering myself an introvert, I am lucky enough to be among circles of amazing friends and relatives. My wife and I owe our first and best experiences in the U.S. to the Ormans, our host family, for their constant support and care. Thanks to our friends in

Oregon (VNOR) and in Vietnam, whose names would constitute an additional page, for making our stay here more enjoyable and at the same time closer to our home country. Special thanks to the family of An Nguyen, for their natural committment to hospitality and delicious foods; to Nguyen Quan and Phuong Nguyen for making Thanksgiving celebration something we look forward to every year; to Larry Geer, Phuong Tran, and Hoa Nguyen for making our visits in the Bay area fun and unforgettable; to our aunt and cousins for making us feel like home every time we visit Seattle.

I dedicate this dissertation to my family, to whom I owe every good thing in my life. It is my Mom with her unconditional love and care, who is willing to sacrifice everything for her children's future; my Dad who is watching us from above and would be smiling seeing our success in life; my twin brother who takes care of the family in Vietnam as a breadwinner so I can pursue advanced degrees; my sister and brother-in-law who encouraged me to follow an exciting career in computer science as well as provided me with valuable work experience. Finally, the completion of this dissertation would be meaningless without my lovely and witty wife, who has shared with me all the ups and downs of this journey with her tender love.

TABLE OF CONTENTS

Page

TABLE OF CONTENTS (Continued)

Page

LIST OF FIGURES

LIST OF FIGURES (Continued)

LIST OF FIGURES (Continued)

LIST OF FIGURES (Continued)

LIST OF FIGURES (Continued)

LIST OF TABLES

LIST OF TABLES (Continued)

Chapter 1: Introduction

Over the years, researchers in various domains[1] have agreed that diversity is a key determinant of many system functions. This agreement has led to a desire for ecosystems, learning systems, and organizational systems, for example, that are diverse and thus more sustainable, persistent, and resilient in the long run (e.g., [90, 51, 52, 55]).

Ecologists study *diversity patterns* of species (e.g., plants and animals) and their *processes*—how the patterns drive ecosystem functions (e.g., productivity, sustainability, resilience) or how the patterns respond to environmental or human factors (e.g., location, climate, land use) (e.g., [51, 101]). The problem is important because the conservation of species, especially rare ones, may depend on the conservation of associated environmental resources. In a similar vein but at different scales, microbiologists and microbial ecologists study diversity of microorganisms and their functionality in various environments, including deep ocean, soils, and human bodies (e.g., [112, 4, 45]). Many other fields share similar characteristics as well. For example, supervised machine learning researchers are often interested in knowing how well their training examples span the space of features—that is, they wish to understand the diversity of their training examples; chemists are interested in assessing the similarity/diversity of a collection of molecular models in exploring the multitude of designs generated by simulations [77]; scholars study language diversity in order to understand societies [110]. In social sciences such as psychology and organizational management, scholars are interested in how the diversity of work team members across multiple characteristics (e.g., age, gender, race, functional background) affects team performance and outcomes (e.g., productivity, creativity, collaboration) (e.g., [90, 55, 14]). This area is interesting because the data set sizes may range from small teams of people to complete organizations. These are just a few examples of fields where an understanding of the diversity of a set of objects characterized by multiple attributes is desired.

To understand phenomena concerning diversity—for instance, diversity patterns and

[1] In this dissertation, *domain* refers to a field of study or a discipline, as opposed to a taxonomic subdivision in biological classification, unless otherwise stated.

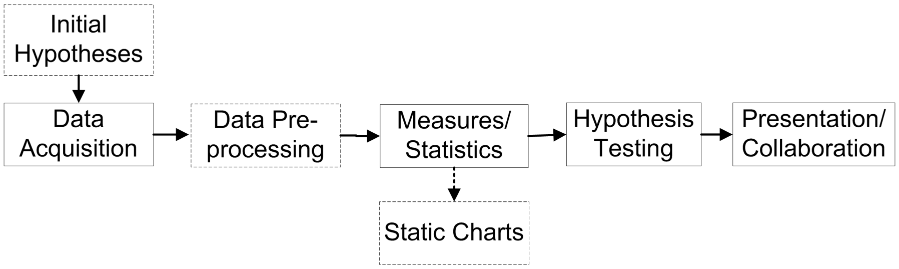


Figure 1.1: Common and traditional approach to diversity analysis. Each rectangle represents a subprocess and each arrow represents a direction the analyst can take to go through the process. The approach emphasizes *hypothesis-driven* or *confirmatory analysis* that relies heavily on diversity metrics and statistical tests of hypotheses. Static charts play a little role in the process.

processes [101]—subject-matter experts[2] typically undertake scientific studies. While different fields may adopt different perspectives on the conceptualization of diversity and hypotheses surrounding it, a common and traditional analysis approach follows three main steps [101, 48, 55, 14] (see Figure 1.1): (1) collect data on objects of interest and other related factors—for example, species and environmental variables for studying species diversity, demographics of team members and team performance for studying workgroup diversity, then (2) plot and observe data variables in typical charts such as histograms, scatter plots, and boxplots, and finally (3) use diversity metrics, such as the Shannon Index [139], and statistical tests to measure diversity and quantify relationships between diversity patterns and system processes. Ultimately, the goals are to characterize diversity patterns and processes across multiple variables and data subsets; if time and space are involved, how these characterizations vary over time and space. Depending on how diversity is conceptualized in different domains, these characterizations may manifest themselves as distributions, clusters, hierarchies, and correlations in the data.

This common approach to data analysis may work well when the number of variables is small and interesting hypotheses can be preconceived. However, when the number of variables is large (> 5), multiple subsets of data are involved, and/or hypotheses are not well pre-established, moving between static charts and statistical tests—often in different software packages—can become unwieldy, slow, and a limiting method of data

[2] We refer to scientists or researchers who study diversity as *subject-matter experts*, *domain experts*, or simply, *experts*.

exploration. Further, diversity metrics conceal tremendous amounts of information in the process [72]. Usually these measures use summary statistics to characterize and compare diversity patterns, where the summary statistics assume theoretical—not empirical—multimodal distributions of data. In addition to the paucity of approaches, domain-specific terminologies and measures preclude the understanding of how diversity functions and how it could be characterized similarly across disciplines.

Interactive visualizations of the data, when combined with traditional analysis approaches, offer the potential to overcome the aforementioned issues, provided that the representations of data coupled with interaction features properly support the analytical needs of experts and are well suited to characteristics of diversity data. Such an interactive visualization would serve as an *effective user interface* for subject experts to explore data directly, formulate and refine hypotheses iteratively, and discuss their findings with others, prior to further statistical analysis (Figure 1.2). By *data exploration*, we mean getting acquainted with data, detecting and describing patterns, trends, and relationships in data while incorporating the user's knowledge and intuition [154, 8]. Furthermore, not only is interactive visualization a potentially powerful analysis tool, but it could prove to be a very useful means for *communicating* diversity information (Figure 1.2). Such a tool, for example, could be used by experts to teach students about diversity analysis or to present scientific findings to the general public. Nevertheless, while typical static univariate or bivariate charts such as histograms, scatter plots, and boxplots have been used by experts to explore and communicate patterns of diversity and its related hypotheses, little work has focused on abstracting the concept of diversity from various fields to potentially unified design considerations and visualization

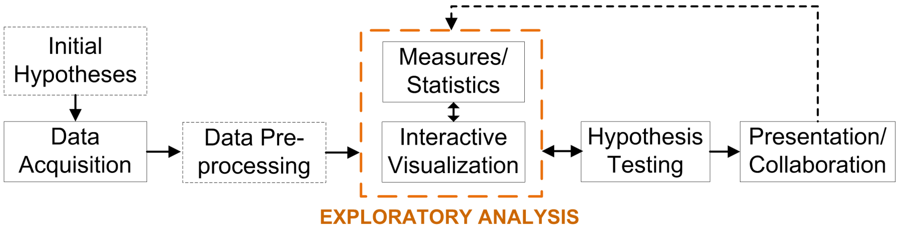


Figure 1.2: Proposed visual-analysis process of exploring diversity data. This dissertation concentrates on the exploration stage (the orange rectangle), as distinguished from to other stages such as data acquisition, data pre-processing, and hypothesis testing.

techniques that illuminate all facets of diversity in multivariate data sets.

1.1 Thesis Statement and Research Questions

To be more specific about the needs in analyzing diversity data, consider the common questions that experts seek to answer. First, there are descriptive questions of *diversity patterns*. For example, how many species are there in this location? how are they distributed? what is the structure of this work team with respect to member demographics? Second, following the first type of questions, there are questions of *diversity processes*. For example, what factors promote such diversity patterns? How are diversity patterns related to socio-ecological process and performance outcome? In other words, do diversity patterns of species, of people, or of institutions promote the capacity of ecosystems, social systems, and their inter-dependent interactions? While this dissertation does not propose answers to these questions, it does suggest that tools are necessary to help experts better approach them.

To answer these questions, experts need ways of exploring complex diversity data directly and holistically. By leveraging the human visual system, visual analytics, "the science of analytical reasoning facilitated by interactive visual interfaces" [151], provides a solid foundation for diversity data analysis. However, from the visualization standpoint, visualizing diversity is a difficult problem because diversity data sets are usually large and multivariate with varying types (i.e., quantitative, nominal, and ordinal); temporal and spatial aspects may also be involved. Furthermore, the dominant use of many diversity measures precludes the understanding of how diversity functions similarly in different domains such as ecology and organizational management. This work intends to align the concept of diversity across these domains and to operationalize the concept in visualizations for exploring and communicating diversity.

This dissertation addresses the following thesis statement:

> *Interactive visualizations of multivariate diversity data, when combined with existing analysis approaches, serve as an effective user interface for subject-matter experts in ecology and organizational management to explore and communicate diversity information directly, prior to further statistical analyses.*

In doing so, the dissertation answers the following three key research questions:

- **RQ1**: How is diversity conceptualized across the multiple fields that study it and what are the fundamental scientific questions/hypotheses of interest regarding diversity?
- **RQ2**: Which existing or novel multivariate representation and interaction techniques are particularly useful in exploring and communicating diversity data?
- **RQ3**: What is the role of interactive visualization in the real-world analysis process in which diversity is a key element?

1.2 Dissertation Contributions and Outline

This dissertation contributes design considerations and systems of interactive visualizations that enable exploratory data analysis of diversity unified across fields. The contributions, which represent joint work with my collaborators, can be grouped into three areas:

1. The design and evaluation of visual representations that communicate diversity patterns in multivariate data (**RQ1**and **RQ2**).
   (a) Design and evaluation of *Diversity Map*—a multivariate space-filling representation emphasizing diversity in separate attributes (Chapter 3). We introduce a precise definition of diversity adopted from the field of ecology, a set of requirements for diversity visualizations based on this definition, and a formal user study design intended to evaluate the capacity of a visual representation for communicating diversity information. An evaluation of the Diversity Map using our study design shows that users can judge elements of diversity consistently and as or more accurately than when using the only other representation specifically designed to visualize diversity. Furthermore, we illustrate the value of the Diversity Map visualization by several example scenarios of ecologists exploring diversity patterns in the moth data sets.
   (b) Design and evaluation of *HIST*—a multivariate visual representation for communicating team faultlines, a conceptualization of diversity in organizational management that shares many characteristics with clustering in computation (Chapter 5). The proposed technique is based on multiple linked, stacked

histograms in a parallel axis layout. We evaluate the effectiveness of the technique in a controlled user study, comparing it to the two other common cluster representations: parallel coordinates and a scatter plot matrix. While we chose the faultline-related tasks based on the requirements by domain experts, the study findings can be generalized to representations and tasks involving distributions of clusters in mixed-type data. Furthermore, inspired by geological faultlines, we propose several visual enhancements to stacked histograms to further facilitate the task of identifying faultlines within work teams.

2. Interaction techniques and interface components that support data exploration where diversity is a key element of the real-world analysis process (**RQ3**).

    (a) Design and evaluation of *Ecological Distributions and Trends Explorer (EcoDATE)*, a web-based visual-analysis tool that facilitates exploratory analysis of long-term ecological data with an emphasis on diversity/distribution patterns and temporal trends (Chapter 4). The tool, which is publicly available online, was created and refined through a user-centered design process, in which our team of ecologists and visualization researchers collaborated closely. Our collaboration resulted in (1) a set of visual representation and interaction techniques well suited to communicating distribution patterns and temporal trends in ecological data sets, and (2) an understanding of processes ecologists use to explore data, generate and test hypotheses. We present three case studies to demonstrate the utility of EcoDATE and the exploratory analysis processes using long-term data on cone production, stream chemistry, and forest structure collected as part of the H.J. Andrews Experimental Forest (HJA), Long Term Ecological Research (LTER), and US Forest Service Pacific Northwest Research Station programs. We also present results from a survey of 15 participants of a working group at the 2012 LTER All Scientists Meeting that showed that users appreciated the tool for its ease of use, holistic access to large data sets, and interactivity.

3. Design considerations and analytical tasks targeting diversity analysis unified across fields of study (**RQ1**and **RQ2**).

(a) *Unified visualization design considerations* and *a taxonomy of common analytical tasks* for exploratory analysis of diversity (Chapter 6). In developing the taxonomy, we cross compare the literature of species diversity (ecology), microbial diversity (microbial ecology), and workgroup diversity (organizational management) and we introduce a framework of diversity concerns aligned across the three areas. The alignment framework is validated and refined by feedback from subject-matter experts.

In addition to the chapters presenting the dissertation contributions, Chapter 2 covers background on information visualization and an overview of diversity concepts. Finally, Chapter 7 summarizes the contributions of this dissertation and suggests possible directions for future work.

Chapter 2: Background

This chapter covers background on the field of information visualization and the concept of diversity. The aim is to set up a shared understanding between subject-matter experts and visualization researchers in terms of common technical vocabulary and abstractions. These concepts and terminology will be used and extended in subsequent chapters.

2.1 Information Visualization

2.1.1 Basic Concepts and Terminology

Visualization systems provide visual representations of datasets intended to help people carry out some task more effectively [23, 146]. In doing so, such systems help people form a mental model of the data and gain insights into data for analysis, communication, and decision making. Visualization systems typically allow manipulation of the data views using a computer [23, 146].

Visualization can be loosely divided into two subfields of *information visualization (InfoVis)* and *scientific visualization (SciVis)*. While there is not always a clear boundary between the two, they differ in the characteristics of the data analyzed and the corresponding data representations. InfoVis tends to deal with interactive displays of abstract data without a direct physical correspondence, such as counts of insects, cone production, or vegetation cover collected over time [146]. SciVis concerns data that has a natural mappings to 2D or 3D space and the visualizations usually involves the physical properties of the data, such as rendering of multiple layers of trees in a forest from LiDAR data [23, 146]. Recently, intersecting InfoVis and SciVis, *visual analytics* emerges as a branch of science focusing on “analytical reasoning facilitated by interactive visual interfaces” [151]. This work adopt the techniques from information visualization and visual analytics, instead of scientific visualization.

A *data set* consists of one or many related data tables. A *data table* is a structured format typically organized as rows and columns. A spreadsheet is a typical example of a data table. A *column* may be referred to as a field, a dimension, an attribute, or a variable, and a row is an object, a tuple, a data case, a data point, a data item, a data observation, or a record [38]. These terms are used interchangeably in this dissertation.

Metadata are descriptive information about the data set, such as name, data type, and description for attributes. An attribute can be in one of three types: ordinal, nominal, or quantitative (or numerical). Categorical refers to both ordinal and nominal types. In some case, for instance, when judging diversity from a data set, we are also interested in the metadata of possible unique values in one attribute (e.g., all possible insect species, all possible ethnicities of team members).

A *multivariate data set* is a data set with more than three dimensions. Common numbers are 4 to 20 dimensions, even though higher-dimensional data sets are increasingly common [38]. The boundary of three corresponds to three dimensions in real world and therefore to human perceptual capability. However, 2D representations of data have been more widely used than 3D because computer displays are thus far two-dimensional. 3D representations still exist but are empirically proven to be more difficult for most users [136]. This dissertation studies 2D representations of data.

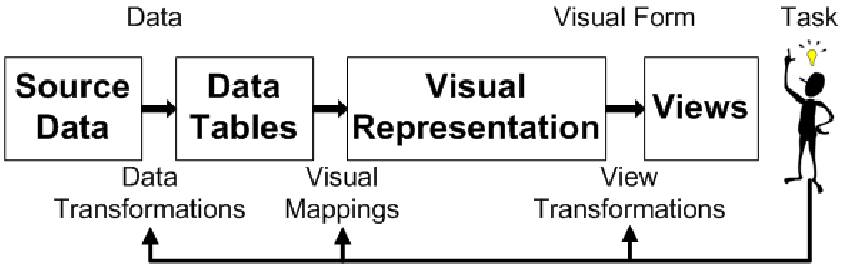


Figure 2.1: Information Visualization Reference Model [28, 23] illustrating the steps involved in building an interactive visualization. Image redrawn from Card et al. [23].

A common approach to designing a visualization system follows the widely-accepted information visualization reference model. The model—which was originally introduced by Chi [28] and later refined by Card et al. [23]—models the visualization process as discrete steps from inputting the source data and transforming them to appropriate formats to mapping data to visual representations and ultimately supporting view transformation via user interactions (Figure 2.1). The outcome of the process is an interactive visualization that helps users complete their tasks and/or gain additional insights into their data. In this dissertation, we adopt this model when designing visualization techniques and tools.

2.1.2 Taxonomies of Visualization Techniques

Visualization researchers have attempted to classify visualization techniques in general and multivariate data visualization techniques in particular [38]. We believe that possible solutions to the diversity visualization problem may come from existing visualization techniques. Therefore, here we briefly discuss different classifications of techniques. Subsequent chapters assess the applicability of specific techniques to diversity analysis.

Card et al. [23] categorized visualization techniques based on the type of *Visual Structures* they adopt. The Visual Structures concept indicates how space is used to encode information or the dimensionality of the data representations. Common types of Visual Structures are Physical, 1D, 2D, 3D, Multi-dimensional, and Tree and Network. Although this taxonomy draws a big picture of techniques, its focus is not on techniques for multivariate data.

On another taxonomy, Seo and Shneiderman [136] focused on 2D representation techniques. They distinguished the three categories of 2D representations by how axes are composed:

1. *Non axis-parallel projection* refers to methods that map a combination of two or more attributes to one axis of the 2D projection plane. In other words, the number of attributes may be processed and reduced before projection. For example, principal component analysis (PCA) [6] and multidimensional scaling (MDS) [152] are well-known techniques in this category.

2. *Axis-parallel projection methods* map attributes as axes of the projection plane. One attribute is selected as the horizontal axis, and another as the vertical axis, to make a familiar and comprehensible representation. Other dimensions can sometimes be mapped as color, size, length, shape, angle. Standard techniques such as scatter plot, histograms are typical examples of this group.

3. *Novel methods* use axes that are not directly derived from any combination of attributes, and axes are not necessarily orthogonal. For example, the parallel

coordinates technique is a new concept in which each attribute is represented by an attribute axis and attribute axes are aligned in parallel [75, 73, 74].

Although this classification is fairly complete for multivariate data, it is not fine-grained enough to differentiate techniques, especially from novel group.

Thus far, one of the most widely-used taxonomies was proposed by Keim [82, 83], who classified techniques according to three different criteria (Figure 2.2): (1) the data type to be visualized, (2) the visualization technique used, and (3) the interaction and distortion techniques. In this respect, each criterion forms a taxonomy of techniques.

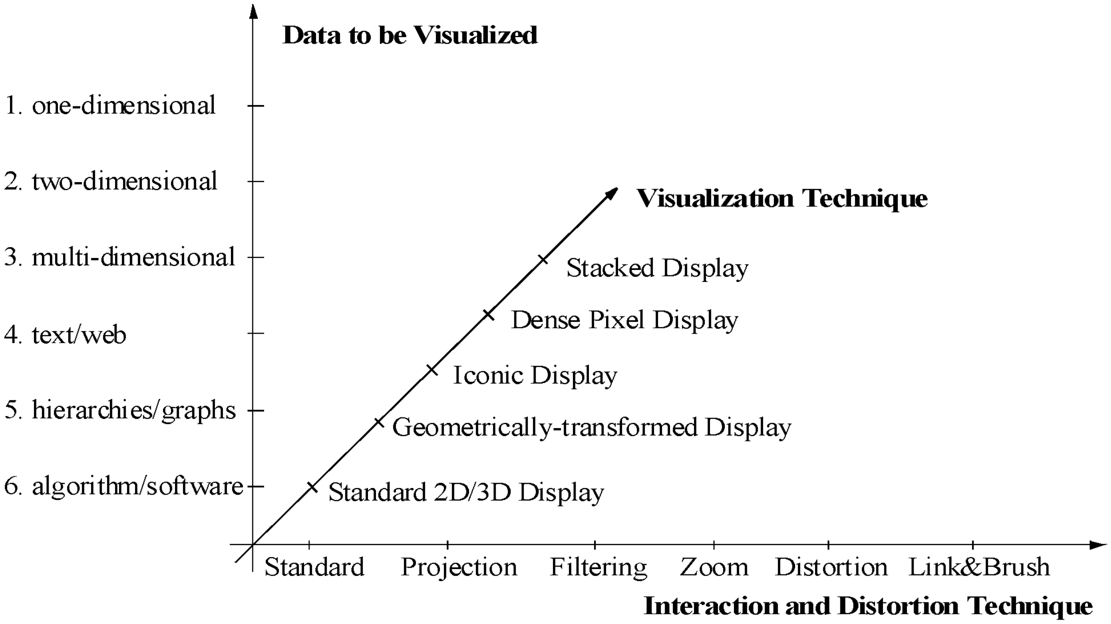


Figure 2.2: Classification of information visualization techniques according to three criteria: (1) the data type to be visualized, (2) the visualization technique used, and (3) the interaction and distortion techniques. Image reused with permission from Keim [83]. Copyright ©2011 IEEE.

Following the axis of visualization techniques used, Keim [82, 83] classified techniques into six groups:

- **Standard 2D/3D displays.** This group refers to visualizations that encode information by positioning marks in a plane with two or three orthogonal coordinate axes. A typical example is the scatter plot, in which two data attributes are projected along the x and y axes of a Cartesian coordinate system. Other examples include 1D scatter plot, box plot, bar chart, and histogram. These techniques effectively support such tasks as finding outliers, gaps, clusters, or correlation between two or three attributes [136].

- **Geometrically transformed displays.** This group refers to both coordinate-based visualizations and graph visualizations [82, 83]. *Coordinate-based visualizations* are an extension from standard 2D/3D displays, which performs geometric transformations and projections of data on coordinate axes. Usually, all data attributes are preserved and treated equally. However, the order of attribute axes may affect what can be perceived. Techniques in this subgroup can handle moderately high-dimensional data sets. Typical examples include scatter plot matrices [7, 10], parallel coordinates [75, 74] and its variants (e.g., [59, 86]), and star coordinates [81]. Beside coordinate-based visualizations, geometrically transformed displays also include *graph visualizations.* This subgroup refers to node-link representations for hierarchy data sets.

- **Data Preprocessing Techniques.** This group refers to statistics techniques typically used for data processing. Keim [83] originally listed these techniques under geometrically transformed displays. For the sake of clarity, we break data preprocessing techniques into a separate group. Generally, these techniques preprocess data to reduce the number of dimensions and/or the number of objects. For example, principal component analysis (PCA) [6] and multidimensional scaling (MDS) [152] are well-known in the group of dimensionality reduction techniques. On the other hand, to reduce data, subsetting techniques may use sampling to find a representative subset of the original data set; aggregate techniques aggregate data objects based on attribute values. These pre-processing techniques are usually required as a starting point of a data-mining project. They can handle very large and/or very high-dimensional data sets. However, users may have difficulty in interpreting data from only subsets or 2D/3D projections whose axes are combinations of several dimensions [136]. With regard to the diversity visualization problem, we assume diversity data are already pre-processed before being mapped to appropriate visual representations (Figure 2.1). Therefore, although data pre-processing techniques play an important role in the data analysis process, they are outside the scope of this dissertation.

- **Icon-based displays.** This group maps each data item to an icon (or glyph) whose visual features vary depending on the data values. Visual features include location, color, shape, size, and opacity. Common representative techniques include scatter plot with additional color encoding, Chernoff face [27] and star plot (or star glyphs) [30]. Within this group, dimensions may be treated differently, as some visual features of the icons (e.g. color) may draw more attention than others (e.g. opacity). Therefore, interpretation is not straightforward and may require legend and training. These techniques can handle a few thousand data items because icons tend to occupy several pixels on a screen. Data overlap possibly occurs if data attributes are mapped to the icon's display location.

- **Dense pixel displays.** In dense pixel displays, a colored pixel is used to represent an attribute value. Pixels belonging to each dimension are grouped into adjacent areas. How pixels are arranged may produce different visualizations, which may be categorized into recursive pattern, axes technique, spiral technique, and circle segments [82]. The techniques in this group are suitable for very large data sets on high-resolution displays (millions of records). Their visualizations may provide insight into information on local correlations, dependencies, and hot spots.

- **Stacked displays.** The group of stacked displays refers to techniques that represent data in a hierarchical fashion. They are usually space-filling techniques in which hierarchy is nested (or stacked). Typical examples include Dimension Stacking [93], mosaic plots [58], and treemaps [78, 140]. These techniques can effectively handle small to medium-sized data sets. They are suitable for handling data sets of low to medium dimensionality. Attributes are treated differently based on the order of being nested and may produce different views of the data. Therefore, interpretation of resulting visualizations may require training.

- **Hybrid techniques.** The last group in the taxonomy, hybrid techniques, simply integrates multiple visualization techniques, either in one or multiple windows, to enhance the expressiveness of the visualizations. Within the scope of this work, we focus on visualization techniques that integrated into one window. The main idea behind integration of techniques is to take advantage of strengths from techniques and to overcome weaknesses.

In summary, this classification by Keim [82, 83] focuses on representation techniques, which is in line with our primary purpose of designing visual representations that convey diversity information in multivariate data (**RQ2**). Therefore, in this dissertation, we assess multivariate data visualization techniques based on this taxonomy (Chapters 3 and 5).

2.2 Overview of Diversity Concepts

Here we briefly describe fundamental concepts of diversity studies in terms of data characteristics, diversity patterns, and diversity processes. To demonstrate the alignability of these concepts across domains, our description cross compares species diversity in ecology and workgroup diversity in organizational management.

2.2.1 Data Characteristics

Diversity data are samples of independent *observations* collected from the population of interest within one or multiple *units of study*. For example, in workgroup diversity, a work team represents a typical unit of study while an individual person is a unit of observation (or measurement) [55, 24]. Comparatively, in species diversity, a typical unit of observation is an individual of known species collected in a community or assemblage, which are typical units of study [101]. A community in ecology refers to group of species found at a given place and time. An assemblage refers to a community in which species are taxonomically related. Each unit of observation may be characterized by multiple mix-typed and hierarchical characteristics (*attributes*) necessary for gauging diversity of the corresponding unit of study—for example, demographics of team members or biological classification of species. In addition, observations can be collected in space and time and associated with additional process factors (e.g. team performance or ecosystem processes). In essence, diversity data sets are mix-typed, multivariate, and in many cases, spatiotemporal and large (thousands of records/observations).

2.2.2 Diversity Patterns and Processes

As mentioned in the first chapter, experts analyzing diversity data typically consider two types of scientific question: (1) questions on description of diversity (*patterns*) (e.g., which species are present in this ecological community? Between two work teams, which one is more diverse?) and (2) questions on causes and consequences of diversity (*processes*) (e.g., how do temperature gradients affect species richness? how are ethnic differences of team members related to conflict within team?).

Diversity pattern is an overarching term. From an analysis point of view, experts may investigate diversity patterns either *in separate attributes* (i.e., one by one) or *over multiple attributes simultaneously*; if time and space are involved, experts may be interested in how the patterns vary over time and space [101, 55].

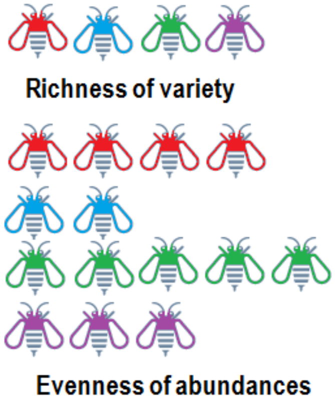
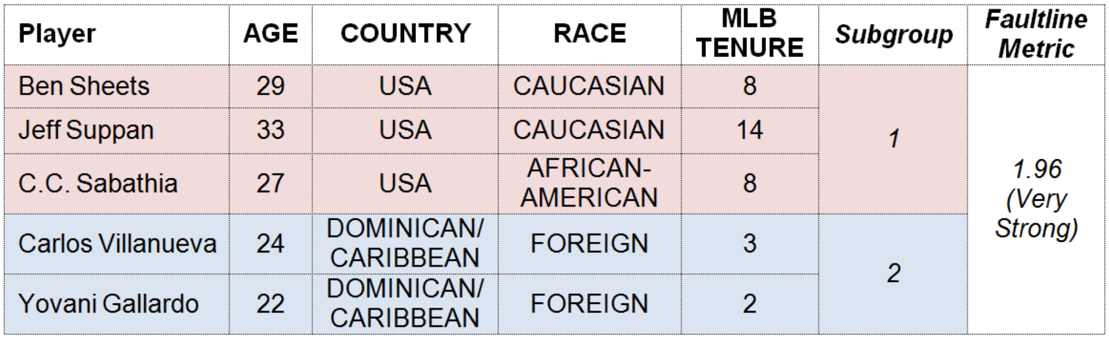


Figure 2.3: Illustration of species richness and evenness. Each icon represents an individual of a known species (e.g., insects). Species richness refers to the number of different species represented in a unit of study and species evenness concerns how close in abundances each species in a unit of study is.

When diversity patterns are considered in separate attributes, different fields adopt slightly different definitions of diversity. Nevertheless, these definitions are centered on the *distribution of data observations.* In ecology, species diversity is defined as "the variety and abundance of species in a defined unit of study" [101]. The definition emphasizes the two main diversity components and corresponding measures of *richness of variety* and *evenness of abundance* of species (Figure 2.3). Likewise, experts studying human organizations describe diversity as "the distribution of differences among the members of a unit with respect to a common attribute, X, such as tenure, ethnicity, conscientiousness, task attitude, or pay" [55]. Further, depending on the attributes under investigation, management experts categorize diversity not only as *variety* (e.g., skill sets) but also as *separation* and *disparity* (e.g., difference in pay among members may create disparity in a team; different cultural values of members represent team separation) [55]. These types of diversity differ in the shape of distribution for maximum diversity and minimum diversity as well as desired diversity, the level of diversity empirically associated with optimal outcomes for the examined unit of study. We present a study of diversity patterns in separate attributes and the design of corresponding multivariate visualization techniques in Chapters 3 and 4.

When diversity patterns are investigated as *interactions among multiple attributes simultaneously*, the concept is further overloaded. Ecologists recognize *functional diversity* as variety of roles played by different species based on their multiple *functional traits* (e.g., rooting depth and maximum growth rate of plants) [116]. In other words, composition of these traits cluster different species present in a unit of study into different *functional groups*. Moreover, since species is inherently hierarchical, traits under investigation could be extended to taxonomic organization such as genus and family, resulting in *taxonomic diversity* (diversity across taxa). There exist parallel components in organizational management. The *faultlines* concept, which is also adopted from multivariate clustering, concerns *subgroups* or *clusters* formed in a work team based on composition of multiple demographic characteristics of members [90]. Figure 2.4 depicts an example of how the faultlines concept is applied to a work team. Chapter 5 concentrates on diversity patterns across multiple attributes and presents a case study of visualizing diversity faultlines in work teams.

Diversity patterns are strongly associated with the functioning of the systems under investigation (*diversity processes*) [101]. Across ecology and organizational management, we can find parallels in the roles of diversity that are of interest to domain experts. For example, ecologists refer to positive effects of diversity such as sustainability and resilience in an ecological system while organizational management experts seek productivity, innovation, and flexibility, just to name a few [51, 55]. By exploring diversity processes, experts look for the causal relationships between the diversity patterns and system processes, which are typically characterized by *correlations* among values of corresponding data variables.



Player	AGE	COUNTRY	RACE	MLB TENURE	*Subgroup*	*Faultline Metric*
Ben Sheets	29	USA	CAUCASIAN	8	*1*	*1.96 (Very Strong)*
Jeff Suppan	33	USA	CAUCASIAN	14		
C.C. Sabathia	27	USA	AFRICAN-AMERICAN	8		
Carlos Villanueva	24	DOMINICAN/ CARIBBEAN	FOREIGN	3	*2*	
Yovani Gallardo	22	DOMINICAN/ CARIBBEAN	FOREIGN	2		

Figure 2.4: An example of how a faultline metric [14] is used to cluster a group of starting pitchers of the Major League Baseball (MLB) team Brewers in 2008 into two subgroups (subgroup 1 and subgroup 2) based on the similarity of group members across the attributes of interest: AGE, COUNTRY (of origin), RACE, and MLB TENURE (in years). Data courtesy Katerina Bezrukova and Chester Spell.

In summary, a variety of diversity components exist and based on research questions of interest, it becomes very important that experts choose the correct conceptualization (e.g. diversity components) and apply the appropriate operationalization (e.g. statistical measures). While the operationalization in terms of statistics is outside the scope of this work, the choice of analytical tasks and visualizations that fulfill those tasks are certainly operationalizations and must be chosen carefully. Chapter 6 aims to align the diversity concept across disciplines into a framework and to identify a unified taxonomy of analytical tasks for exploratory analysis of diversity.

Chapter 3: Visualization of Diversity in Separate Attributes [1]

3.1 Introduction

When exploration of diversity data involves multiple attributes, a general starting point is to examine the *overall distribution of observations* over the attributes of interest. For example, in selecting an incoming freshman class, college admissions officials may wish to consider how diverse a particular population of applicants is with respect to attributes such as GPA, gender, home state, and ethnicity. Similarly, in analyzing species diversity data, ecologists may wish to understand the diversity patterns of species present there in relation to other environmental or human factors (e.g., temperature, elevation, rainfall, land use) [101].

In most cases, determining the overall diversity of a set of objects can be decomposed into an examination of diversity in *each of a number of separate attributes.* Unfortunately, as the number of attributes and objects to be examined both increase (for example, beyond five and 1000 respectively), the number of values that must be considered in gauging diversity increases. This can make a text- or table-based assessment of the diversity of a large data set with many attributes extremely difficult and tedious. While metrics, such as the Shannon Index [139, 158], are intended to provide a measure of diversity, these generally reduce diversity to a *single* number, throwing away a large amount of information in the process. Moreover, these metrics can typically be applied to measure the diversity of only a single attribute. Experts in ecology have argued that metrics like the Shannon Index are not always useful and that scientists should rely on a more direct observation of the data to gauge its diversity [158, 133, 72].

In this chapter, we attempt to formalize the problem of visualizing *diversity patterns* as *distributions of separate attributes* in large, multivariate data sets. Our primary contribution is a visual representation called the Diversity Map (DM), which is specifically intended to help users understand the diversity patterns of a large set of multivariate

[1] The material in this chapter was previously published with co-authors Rob Hess, Crystal Ju, Eugene Zhang, and Ronald Metoyer in [119] and co-authors Steven Highland, Ronald Metoyer, Donald Henshaw, Jeff Miller, and Julia Jones in [120].

objects (Figure 3.1a). DM is designed to be efficiently perceived to give an accurate initial impression of a data set's *overall diversity*, while also allowing the user to explore relationships and interrogate the raw data using an overview as the interface.

We also contribute a precise definition of diversity based on the one used by ecologists in discussing species diversity, a set of requirements for diversity visualizations based on this definition, and a design for a formal user study intended to understand the effectiveness of a visual representation in communicating diversity information. We evaluate DM by using this study design to compare it to Pearlman et al.'s Glyph Hybrid visualization [114], the only other representation specifically designed to visualize diversity (Figure 3.1b). In comparing user performance between Pearlman et al.'s representation and DM, we show that users can as or more accurately judge elements of diversity using DM. The results across task questions are also more consistent for DM.

Finally, we deployed an interactive version of DM for use by ecologists. In the Oregon State University H.J. Andrews Experimental Forest, researchers have collected data on diversity and abundance of moth species [107, 106]. The data are of particular interest to ecologists because moths are indicators of broader biological diversity in plant types and physical environments. We illustrate the value of the DM visualization by several example scenarios of ecologists exploring the moth data sets and we discuss what we have learned from our interdisciplinary collaboration.

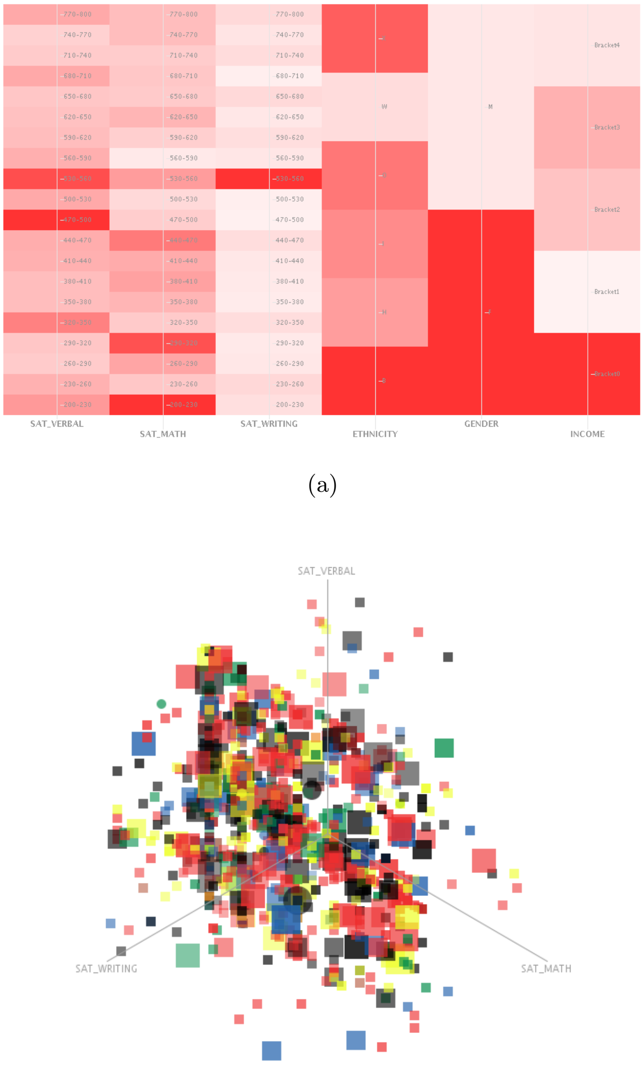
3.2 A Definition of Diversity and Design Requirements

3.2.1 Defining Diversity

Before discussing its visualization further, we must first establish a more thorough definition of diversity. With this in place, the requirements for a successful diversity visualization will become more clear.

The data sets in which we are interested represent observations of populations of objects (e.g. students, moths, stocks, etc.) that are described by multiple variables, or attributes (e.g. GPA, ethnicity, gender, etc.). To define the diversity of such a set, we borrow from the established field of Ecology, where biological diversity is defined as *"the variety and abundance of species in a defined unit of study"* [101].

Two measures of diversity are used in Ecology: *richness*, which is simply the number



(a)

(b)

Figure 3.1: A synthetic data set of medium-diversity visualized using the Diversity Map (a) and the Glyph Hybrid [114] (b). The data set contains 1000 objects and 6 attributes (SAT verbal, SAT math, SAT writing, ethnicity, gender, and income level). Visual mappings of the attributes in the Glyph Hybrid are described in Figure 3.9.

of species in the unit of study represented out of all possible species; and *evenness*, which describes the variability in species abundances [101]. Generalizing from Ecology, we say that a population sample is diverse with respect to a specific attribute if it exhibits a rich variety of values of that attribute and if each of those values is evenly abundant. In other words, high diversity corresponds to a *uniform distribution* of objects across all possible values of an attribute. We extend the definition of diversity to sets of arbitrary objects described by many different attributes by simply defining *overall diversity* as the aggregated diversity over all attributes being considered.

As an example of how this definition is applied, consider analyzing the diversity of a university's potential incoming freshman class. In particular, if we are considering the diversity of different populations of applicants with respect to their income levels, then a very diverse population will contain a similar number of applicants (i.e. even abundances) in each of many possible income brackets (i.e. a rich variety). In contrast, a very non-diverse population might contain applicants in only a single income bracket (i.e. no variety) or mostly applicants in a single income bracket with very few applicants in each of the others (i.e. very uneven abundances). The diversity of other attributes, such as GPA, ethnicity, gender, etc., would also contribute to the overall diversity of a particular population of applicants.

3.2.2 Design Requirements

Beyond our definition of diversity, we also borrow several conventions from the study of species diversity. Specifically, we adopt individual objects as our unit of measure, and, as in the study of species diversity, we treat all possible values of an attribute and all individuals in a population sample as equal. Additionally, since we have extended the definition to account for diversity over many attributes, we adopt the added convention that all attributes are treated as equal (i.e., equally perceived by the user).

In order to adequately convey diversity as defined above, a visualization should possess the following properties:

- Communicates the attributes of interest, the richness in variety of the values of each attribute, and the evenness of abundance of the population sample of interest over the values of each attribute while considering all attributes and objects equally.

- Scales well to large multivariate data sets, i.e. ones containing many objects (> 1000) and many attributes (> 5).
- Enables users to make judgments about diversity with little effort through an efficient perceptual encoding (while ideally, the visualization should be designed so that the user perceives diversity preattentively, i.e. without focused attention [153], we understand that this is difficult for large attribute spaces).

3.3 Related Work

In this section, we review a subset of existing multivariate visualization techniques, emphasizing those that apply to the problem of exploring the diversity of a set of objects in separate attributes, as defined earlier. We focus only on representation methods and organize our review based on the taxonomy proposed by Keim [83].

3.3.1 Standard 2D/3D Displays

Techniques such as scatter plots, box plots, bar charts, and histograms effectively support tasks such as finding outliers, gaps, clusters, and correlations over a small number of attributes [136]. However, while the box plot is well suited to displaying evenness of abundance, it fails in communicating richness of variety and is not applicable to categorical data. Likewise, without additional encoding, the scatter plot may lead to ambiguous communication of evenness of abundance due to occlusions caused by data overlap. A rectangular heatmap can be viewed as a special case of the scatter plot where a value is plotted for every combination of the two mapped attribute values and a point is replaced by a colored square. Like the scatter plot, heatmaps are limited to displaying diversity over only the two attributes being mapped. However, occlusion is no longer a problem. The histogram, in particular, is well suited to showing richness in variety and the evenness of distribution of objects over a single attribute. As noted, all of these approaches typically display only one or two attributes of interest.

The use of multiples may solve some of these problems. For example, scatter plot matrices may provide useful representations of diversity (Figure 3.2), especially for high and low diversity cases, but intermediate values may be difficult to disambiguate due to data overlap. While jittering techniques [29] may help alleviate this problem, they may give the misleading appearance of evenness when it is not actually present. A matrix of heatmaps would avoid the data overlap issue and could be an interesting approach to viewing diversity (both richness and evenness). Multiples in matrix form, however, require screen space that grows with the square of the number of attributes. Multiples of histograms could be a powerful method for diversity visualization, since these appear capable of conveying both richness of variety and evenness of abundance. However, it is not clear how well *overall* diversity is communicated by multiple spatially separated histograms. The Diversity Map representation, described in Section 3.4, is in fact a multiple histogram representation with an alternative encoding that facilitates communication of overall diversity.

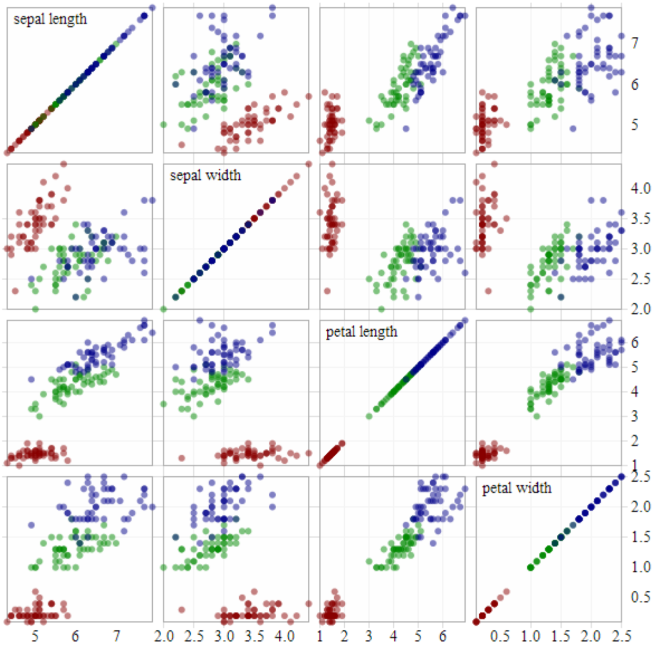


Figure 3.2: Scatterplot matrix representing Edgar Anderson's Iris data set [39]. The example is created with the D3 toolkit [16].

Alternatively, rank/abundance—or Whittaker—plots [158] are commonly used by ecologists to visualize species abundance distribution. The representation is a variation of the scatter plot in which species are ranked from most to least abundant and then plotted along the x axis, while the y axis shows the relative abundance of species. The shape of the resulting curve provides insight into species evenness (or dominance). Although this approach is specific to species abundance, it and the other standard approaches serve as a starting point for exploring techniques for visualizing distributions of data over many dimensions.

3.3.2 Geometrically Transformed Displays

Geometrically transformed displays map one object to a set of points and lines in 2D or 3D space [83]. This category includes graph visualizations and coordinate-based visualizations. While graph-based visualizations are important in many fields, we do not discuss them because we assume that limited (or no) explicit relationship information is present in the data sets we consider.

Coordinate-based visualizations extend standard 2D/3D displays by performing geometric transformations and projections of data onto coordinate axes. Data attributes are typically preserved and treated equally during this process. These techniques are generally applicable to multivariate data sets and offer potential solutions to the diversity visualization problem. Examples include parallel coordinates [75, 73] and related variants [59, 86], and star coordinates [81].

Parallel coordinates [75, 73] are well-suited to visualizing various types of multivariate data (quantitative, ordinal, or nominal) and revealing data correlation between attributes (Figure 3.3). However, visual clutter becomes a limitation as the number of objects increases. Refinements to parallel coordinates have attempted to address visual clutter with brushing [59], clustering [44, 9], and dimension reordering [115].

Despite these improvements, accurately judging richness of variety and evenness of abundance may still be difficult using parallel coordinates, especially for larger data sets. However, several variants of parallel coordinates overcome this limitation by providing information on the distribution of values for each attribute [59, 86].

In one variant, a histogram is overlaid lengthwise on each parallel axis [59], and bin intervals are created for quantitative attributes by partitioning them into ranges. Each histogram communicates both the richness of variety and the evenness of abundance of the values of the corresponding attribute. However, the (necessary) spatial separation of the histograms in this approach may likely affect the user's ability to interpret overall diversity without significant effort.

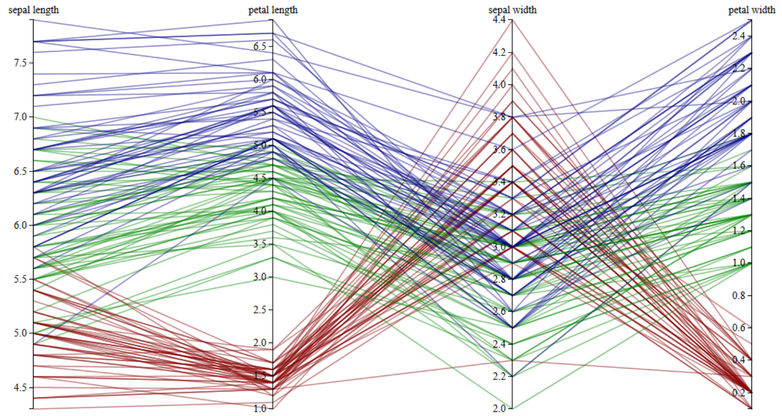


Figure 3.3: Parallel coordinates representing Edgar Anderson's Iris data set [39]. The example is created with the D3 toolkit [16].

The Parallel Sets [86] technique is another variant of parallel coordinates that targets categorical data in particular. This representation adopts the layout of parallel coordinates and uses a box to represent each possible value of a categorical attribute. Box size is scaled lengthwise along the axis in proportion to the frequency of the value in the data set. Connections between values of two different attributes are also scaled according to their frequency values. Parallel Sets convey the distribution of objects over the values of an attribute (i.e. the evenness of abundance for an attribute), as well as relationships between the distributions of values across multiple attributes. However, while this approach scales to large data sets, the number of possible attribute values it can display is limited due to space restrictions. In addition, the boxes corresponding to outliers, i.e. attribute values exhibited by very few objects, can become imperceptibly small. Moreover, this method does not display attribute values not represented in a particular data set. When combined, these limitations make it very difficult to accurately perceive richness of variety using Parallel Sets.

Star coordinates [81] is well-suited to visualizing the overall distribution of a set of objects. Unfortunately, the mapping between a data point and its location in star coordinates is many-to-one. That is, several different data points with equal vector sums will end up in the same location. This ambiguity makes it difficult to discern richness of variety and evenness of abundance over the attribute space.

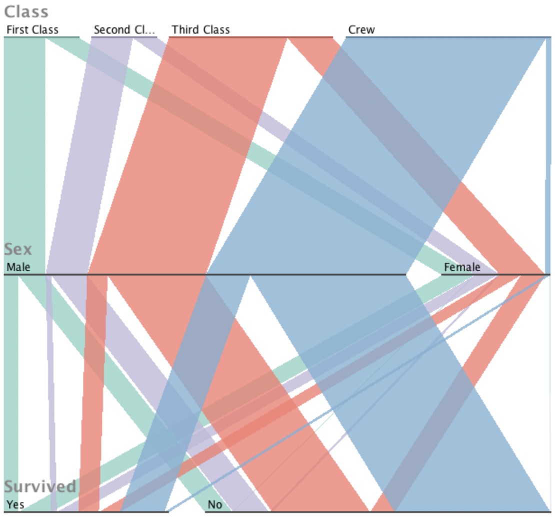


Figure 3.4: A Parallel Sets visualization showing the Titanic data set with three attributes: Class, Sex, and Survived. Image reused with permission from Kosara et al. [86]. Copyright ©2006 IEEE.

3.3.3 Icons, Dense Pixels and Stacked Displays

Several other classes of multivariate visualization techniques have been developed that are not well suited to diversity visualization. Icon-based displays, such as Chernoff faces [27], typically treat attributes differently and as a result, some visual features of the icons (e.g. color) may draw more attention than others, thus violating our requirement of equal consideration for all attributes. Star glyphs, on the other hand, give equal treatment to attributes, however this approach will not scale well with a large number of objects due to occlusion. While dense pixel displays scale well with the number of objects [82], they do not necessarily display all possible values (only the ones represented in the data set), making it difficult to gauge richness of variety. Stacked display techniques represent data in a hierarchical fashion and are often space-filling approaches where a hierarchy is nested (or stacked) [93, 78, 140]. Since we are not specifically concerned with hierarchical data, these techniques are not considered further.

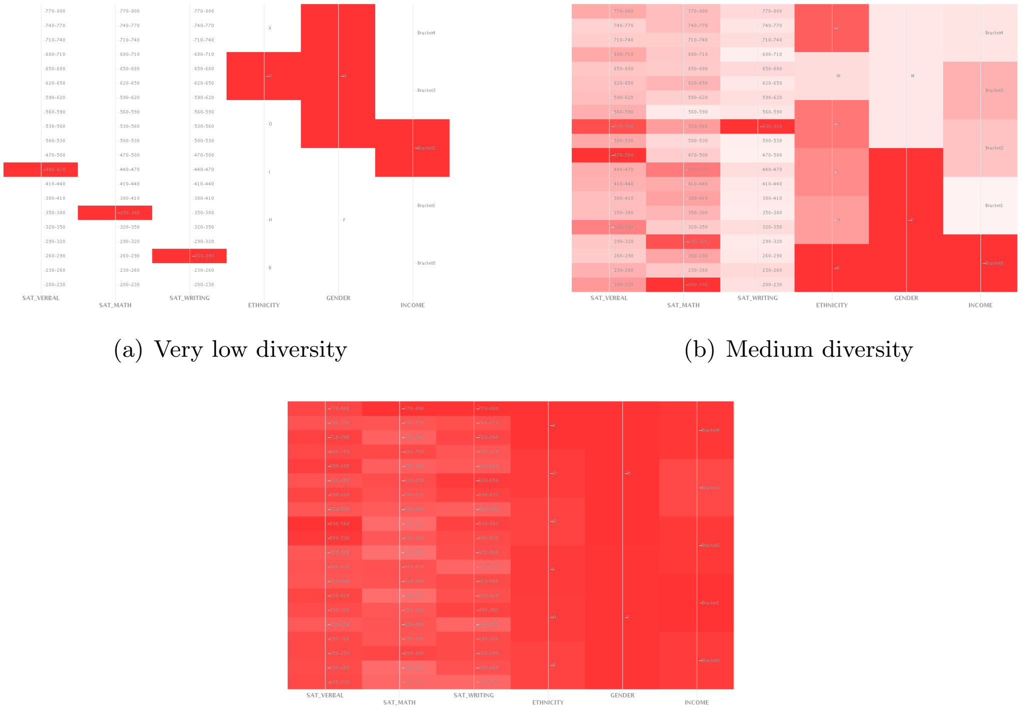
Finally, there is a large group of approaches that fall into the category of data preprocessing techniques that generally manipulate the data to reduce the number of dimensions and/or the number of objects [152, 6, 162]. While these approaches are popular in many fields as a starting point for exploring data, they typically result in a loss of information and sometimes yield results that are reduced to a non-intuitive space and are thus difficult for users to interpret, especially with respect to the richness of variety. Thus, we do not consider these techniques further in this chapter.

3.3.4 Hybrid Techniques

Hybrid techniques integrate multiple visual representations in one or more windows. The most relevant technique in this class is Pearlman et al.'s glyph-based approach [114], the only proposed technique to explicitly address the problem of visualizing the diversity of a set of objects. Pearlman et al. focus on communicating both diversity, loosely defined as the distribution of attribute values across a set, as well as depth, defined as the attribute values of individual members of the set. This technique represents objects as glyphs in a coordinate frame, where three attributes (of possibly many) are used to map objects to the 2D space of the frame in much the same way as multi-dimensional objects are mapped to 2D space using star coordinates (See Figure 3.9). Other glyph properties, such as shape, size, opacity and color are used to represent additional attributes and are typically described in an accompanying legend. Unfortunately, the number of attributes that can be successfully encoded using this technique is limited by the perceptual and cognitive loads placed on the user by icon-based approaches. Moreover, the number of objects that can be successfully visualized using this technique is limited by occlusion. Nonetheless, this representation is important, since it is the first to explicitly address the problem of visualizing diversity in multivariate data, and we revisit it in Section 3.6 where we formally compare its ability to communicate diversity information to that of our Diversity Map representation.

3.4 The Diversity Map Representation

To address the shortcomings of previous approaches, we developed a novel representation called the *Diversity Map* (DM) for visualizing the diversity of a set of objects. In this representation, depicted in Figure 3.5, each attribute is represented as one of a set of parallel axes, similar to the parallel coordinate layout. Unlike traditional parallel coordinates, however, each object is represented in DM by placing a semi-transparent



(a) Very low diversity

(b) Medium diversity

(c) Very high diversity

Figure 3.5: Synthetic data sets of (a) very low-, (b) medium-, and (c) very high-diversity visualized using the Diversity Map representation. Each visualized data set contains 1000 objects and 6 attributes (columns from left to right: SAT verbal, SAT math, SAT writing, ethnicity, gender, and income level). The very high-diversity data set is 6 times more diverse than the very low-diversity one.

rectangle on each attribute axis at the locations corresponding to the object's attribute values. In other words, for a data set containing N attributes, each object is represented by placing one semi-transparent rectangle on each of N parallel axes. Note that in our approach, we discretize continuous numerical attributes. We refer to the distinct locations along the attribute axes corresponding to discrete attribute values as *bins*.

To satisfy the requirement from Section 3.2 that all objects are treated equally, each semi-transparent rectangle contributes an equal, fractional amount of opacity to the bin in which it is placed. To satisfy the requirement that all attributes are treated equally, we normalize the opacity values on a per-attribute basis so that bins corresponding to

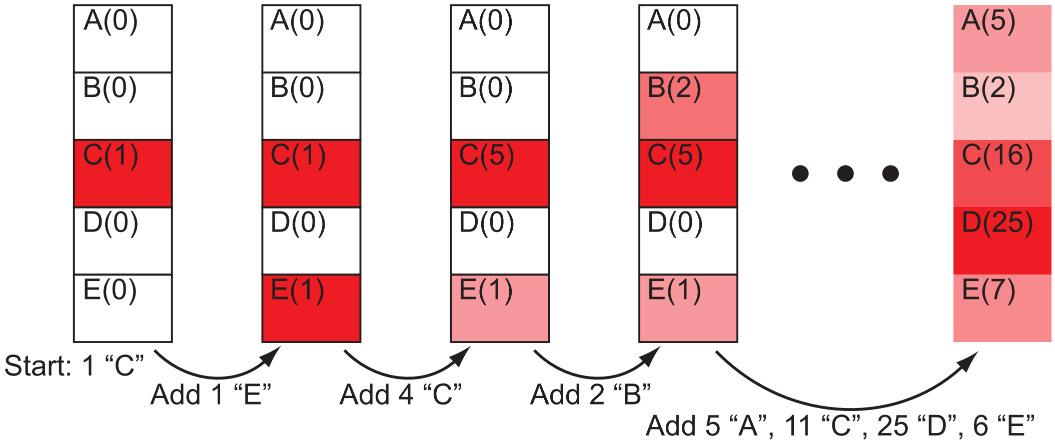


Figure 3.6: The process of visualizing a single attribute using DM. The depicted attribute has five possible values (A, B, C, D, and E). The visualization begins with a single object with attribute value “C,” and objects with other attribute values are added in subsequent steps. At each step, the number of objects in each bin is shown in parentheses next to the bin’s label and the opacity (α-value) of each bin is calculated as described in the text. Note that, while it is instructive to illustrate the process step-wise, as above, our implementation simply aggregates object counts and computes opacity values in a single step. Also, note that for a multivariate data set, every object would contribute to each of the parallel attribute axes in the same way as depicted above, resulting in a visualization as depicted in Figure 3.5(b).

attribute values not represented in the visualized data set are fully transparent (i.e. $\alpha = 0$ in RGBA color space), and the bin(s) corresponding to the most abundant attribute value(s) in the data set are fully opaque (i.e. $\alpha = 1$). The opacity of every remaining bin is calculated based on the ratio of the number of objects in that bin to the number of objects in the bin corresponding to the most abundant attribute value. We have empirically found that using the square-root of the number of objects per bin in this calculation helps to make bins corresponding to attribute values with low abundance more recognizable. Specifically, the opacity of each bin x is calculated as $\alpha(x) = \sqrt{|x|/|x_{MAX}|}$, where $|x|$ denotes the number of objects in bin x, and x_{MAX} is the bin with the most objects for the attribute in question. Figure 3.6 illustrates the process of visualizing a single attribute using DM.

An alternative way to view our design is to imagine each attribute axis as a histogram over the values of that attribute constructed in 3D space by stacking semi-transparent tiles on top of each other, as illustrated in Figure 3.7. When viewed from above, the taller stacks of tiles appear darker, while the shorter stacks appear lighter, according to the total combined contribution of the tiles in each stack to that stack's opacity.

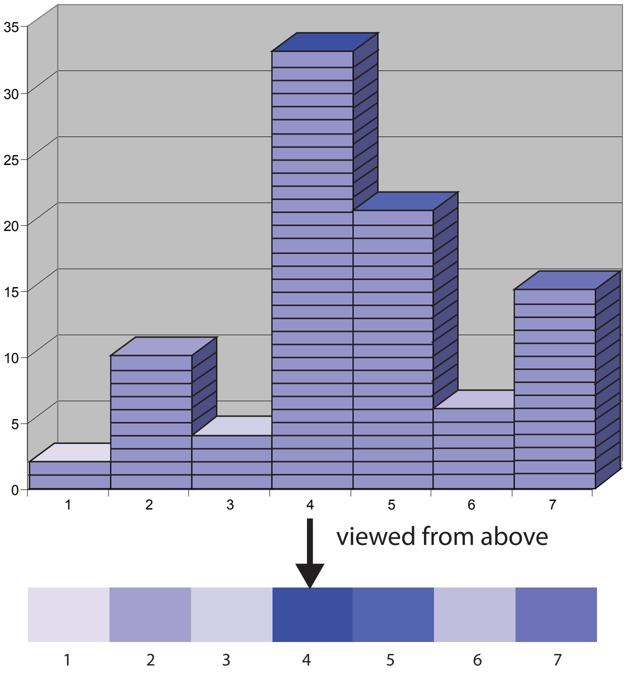


Figure 3.7: Each attribute axis of DM can be viewed as a histogram over the values of that attribute constructed in 3D space by stacking semi-transparent tiles on top of each other.

3.4.1 Design Considerations

As indicated earlier, our primary goal in designing DM was to make easily apparent the richness of variety and the evenness of abundance of the attribute values exhibited in the data set being visualized. While we do not explicitly calculate or assign values for richness and evenness, we consider them to be quantitative features of the data, in that we can think of one data set as being more or less rich or even than another. For this reason, we have chosen visual encodings that are known to be effective for conveying quantitative information.

Specifically, we encode variety using spatial position by assigning a distinct 2D location, or bin, to each of the possible discrete attribute values that can be taken by objects in the visualized data set, and we encode abundance using opacity, with each semi-transparent rectangle representation of an object's attribute value contributing a constant, fractional amount of opacity to the bin in which it is placed. Under this encoding, more abundant regions of attribute space are indicated by visual regions of higher opacity.

Because spatial position ranks in the visualization literature as the most effective encoding for quantitative information [29, 99], it is easy to justify its use in our design. However, several other quantitative encodings, such as angle, slope, and area, rank higher than opacity [29, 99]. Unfortunately, these encodings appear to conflict with our chosen spatial encoding. In contrast, we found that opacity serves as a natural complement to the spatial encoding and allows us to elegantly convey both the richness of variety and the evenness of abundance of the visualized data. In particular, under this combination of encodings, "occlusions" in the 2D visual plane that result from one or more objects sharing a certain attribute value serve simply to increase the opacity of that visual region, thereby indicating increased abundance.

In DM, richness of variety is expressed by the number of bins with non-zero opacity, and evenness of abundance is expressed by the uniformity of the color distribution across the bins of a single attribute, as well as over the entire visualization. In other words, the more rich is the variety of a given data set, the more non-transparent bins it will yield, and the more even is the abundance across the data set, the more uniform will be the colors of the bins.

The overall diversity of a given data set—that is, the combined diversity of all its attributes—is communicated by DM as the overall color density of the entire visual region: as the visualized data set becomes more richly various and more evenly abundant, more bins will exhibit a similar non-transparent color. In the limit of "perfect" diversity, where all possible values of each attribute are represented equally, the entire visual region will be a solid, completely opaque color. Conversely, a set with little diversity will produce a visualization with regions of very high contrast. As examples of these phenomena, consider the synthetic data sets with zero and near-perfect diversity visualized using a Diversity Map in Figures 3.5 (a) and (c).

Finally, we note that DM is specifically designed to provide a holistic overview of the population sample of interest. As Shneiderman notes, [141], providing an overview of the data is an important part of a visualization system, as overviews help the user build a mental model of how the data covers the attribute space. This model in turn helps the user formulate actions such as queries [126]. Indeed, a good overview representation should serve as a gateway by allowing the user to interact with the visualization in order to investigate the data based on the mental model he or she has formed. While we reserve deeper investigation of this matter for future work, we simply note that DM is designed to serve as just such a gateway.

3.4.2 Application to Real-world Applicants Data

We also explore the application of Diversity Map to real-world data. As an example, we applied Diversity Maps to a real data set containing 2550 applicants (one year worth) to a particular university. Each applicant is characterized by ten attributes. Interestingly, this real data set was preprocessed using an existing proprietary software package designed to recommend a set of applicants using a holistic evaluation process intended to produce a diverse incoming class[2]. The DM visualizations of this data set are shown in Figure 3.8. These results are promising in that they agree with the output of the holistic evaluation software.

3.5 User Study Design

In this section, we describe a formal user study designed to measure a given visualization's ability to communicate diversity information. In particular, the study is a controlled user study intended to be conducted in a laboratory setting, and it is designed to compare the visualization of interest against a given baseline visualization. There are two important components to this design: (1) a method for generating synthetic data sets with controllable, varying levels of diversity and (2) a set of questions, each of which is meant to assess a study participant's ability to comprehend a particular aspect of diversity using each of the visualizations under comparison. We describe these components next.

[2] http://www.applicationsquest.com/

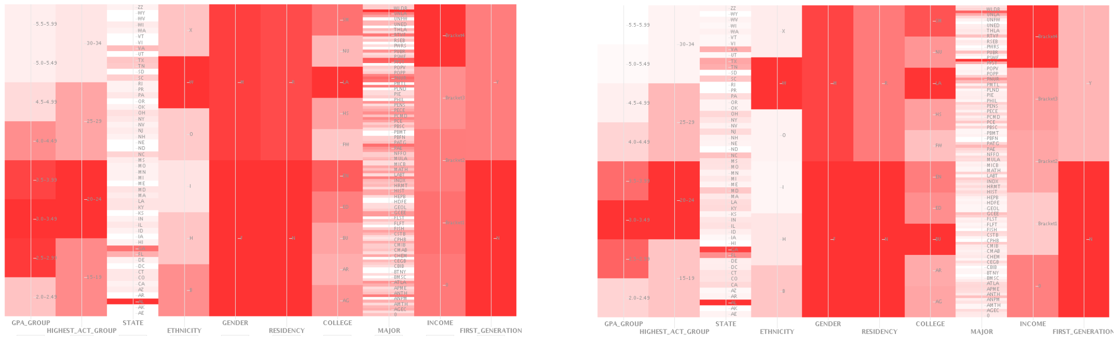


Figure 3.8: A real data set of 2550 college applicants with 10 attributes visualized using DM. Left: the subset of students recommended for acceptance based on a holistic admissions process implemented by a proprietary software package and designed to produce a diverse incoming class. Right: the subset of rejected students. The recommended students yield a visualization with a more even distribution of opacity, especially in attributes like GPA, ethnicity, residency, and major (columns 1, 4, 6, and 8 respectively). This suggests that the recommended applicants are more diverse than the rejected ones.

3.5.1 Synthetic Data Generation

While, ideally, we would use a real data set to evaluate a visualization, we require data with specific distributions of values over attributes. Since it is difficult to find data sets that can accommodate this requirement, we developed a technique for creating synthetic data sets of controllable, varying diversity over a set of independent attributes. In particular, our procedure generates synthetic sets of objects over a manually defined set of attributes and attribute values, where the richness of variety and evenness of abundance over each attribute is controlled and measured.

Our data generation procedure is based on the Shannon index, or Shannon entropy, a measure of diversity that is widely used in ecology [139, 158, 87]. Shannon entropy is also used in other fields, such as information theory, where it is used to measure the amount of information contained in a coded message. In its general form, the entropy of a single random variable, X (in species diversity, X corresponds to species; in the more general case, it could correspond to any single attribute) is

$$H(X) \;=\; -\sum_{i=1}^{S} p(x_i) \log p(x_i), \tag{3.1}$$

where $\{x_1, \ldots, x_S\}$ is the set of possible values of X and $p(x_i)$ is the probability that X takes value x_i. In species diversity, for example, $x_1, \ldots, x_S$ represent the possible species and $p(x_i)$ represents the probability of observing one particular species x_i. In practice, we compute $p(x_i)$ as the ratio of the number n_i of instances of value x_i to the total number N of individuals in the set, i.e. $p(x_i) = \frac{n_i}{N}$. In other words, $p(x_i)$ represents the relative abundance of value x_i in the total set.

$H(X)$ is directly proportional to the level of diversity within a single attribute, in that higher values of $H(X)$ correspond to richer variety and more even abundances. Unfortunately, it is difficult to compare values of $H(X)$ across attributes, since it is scaled to the number of possible values of the attribute being measured. This implies that an attribute with many possible attribute values (e.g. the home state of a student) may be considered more diverse under entropy than an attribute with few possible values (e.g. the gender of a student), even if it is not.

In order to meet our requirement from Section 3.2 that all attributes are considered as equal, we have adapted a variant of the Shannon index known as the *evenness measure* [124], which normalizes the value of $H(X)$ by its maximum possible value:

$$H_{max}(X) \;=\; -\sum_{i=1}^{S} \frac{1}{S} \log \frac{1}{S} \;=\; \log S. \tag{3.2}$$

Thus, the evenness of attribute X is

$$H_E(X) \;=\; \frac{H(X)}{H_{max}(X)} \;=\; -\frac{1}{\log S} \sum_{i=1}^{S} p(x_i) \log p(x_i). \tag{3.3}$$

Note that, despite its name, this measure captures both the richness and evenness of attribute X. In particular, richness, which measures the number of values of represented out of all possible values of X, is indicated by the number of values x_i with non-zero probability. The more of these that are present for attribute X, the higher the value of $H_E(X)$. Likewise, evenness is indicated by the uniformity of the probabilities $p(x_i)$, and $H_E(X)$ is maximized when each attribute value x_i occurs with the same probability. An important property of this measure is that it always takes a value between 0 (zero diversity) and 1 (full diversity).

In our setting, we have one variable X^k corresponding to each attribute, and we hand-specify the possible values $\{x_i^k\}_{i=1}^{S^k}$ for each attribute X^k. We model the distribution $p(x_i^k)$ over the possible values of each attribute as multinomial. In other words, associated with each possible attribute value x_i^k is a weight w_i^k, where $w_i^k \geq 0$ for $i = 1, \ldots, S_k$ and $\sum_i w_i^k = 1$, and the attribute values in a given set are distributed in proportion to those weights.

To rigorously test visualization methods, we wish to be able to generate data that achieves a pre-specified target value $H_E^*(X^k)$ of the evenness measure for each attribute X^k. We model this as a set of separate non-linear optimization problems, one for each attribute. The objective for each problem is to find the set of weights $\{w_i^k\}_{i=1}^{S^k}$ that minimizes the squared error between the resulting evenness $H_E(X^k)$ and the target evenness $H_E^*(X^k)$. We solve for these weights using a gradient-based quasi-Newton method.

Once the distribution $p(x_i^k)$ is specified with weights $\{w_i^k\}_{i=1}^{S^k}$ for each attribute X^k, we generate synthetic data by simply drawing samples from each of these distributions and using the j^{th} sample for each attribute as the corresponding attribute value of the j^{th} object in the data set. Then we use $H = \sum_k H_E(X^k)$ as a measure of the overall diversity of a particular data set.

3.5.2 User Study Questions

Our user study contains four types of questions. Each type is designed to assess a different aspect of the user's ability to perceive diversity using a particular visualization. We outline each question type here.

Q1: *Between two visualizations generated with the same method, which picture represents a more diverse set of objects?* (possible answers: picture A or picture B) The primary goal of this question type is to determine if a visualization technique is discriminative enough to allow a user to distinguish and compare the levels of overall diversity depicted in two visualizations generated with the same technique. The difficulty of each question of this type can be determined by the difference in the overall diversity values H between two visualized data sets. The bigger this difference, the easier the question is.

Q2: *How diverse is the data set represented in this picture?* (possible answers: very low diversity, low diversity, medium diversity, high diversity, very high diversity) This question type is intended to identify how well a user can interpret and assign a diversity value to a visualization given baseline examples of zero and full diversity (which we provide to users in tutorials; see Section 3.6). The level of diversity of a data set is determined based on its overall diversity value H.

Q3: *What is the most/least diverse attribute in the data set represented in this picture?* (possible answers: the possible attributes) This question type is designed to understand the participant's ability to identify relative differences in diversity among attributes that may have different levels of richness of variety or evenness of abundance. The difficulty of each question of this type can be determined by the difference between the values of the evenness measurements $H_E(X^k)$ of the most/least and second-most/least diverse attributes. The bigger this difference, the easier the question is.

Q4: *Which value of attribute X contains the most/least objects?* (possible answers: possible values of attribute X) The last question type is designed to determine the participant's ability to isolate attribute values with high and low relative abundance of objects, given a particular attribute to inspect (e.g. ethnicity). The difficulty of each question of this type can be determined by the difference between the number of objects exhibiting the most/least abundant attribute value and the number exhibiting the second-most/least abundant attribute value. The bigger this difference, the easier the question is.

In a study, questions of each question type are the same in terms of wording. However, they can be asked multiple times on different data sets to vary the difficulty (Q1, Q3, Q4) or the level of diversity (Q2). For each of these question types, ground-truth answers are based on the distribution of objects and the evenness measure values obtained using our synthetic data generation method.

3.6 User Study Implementation and Results

In this section, we use the formal user study described in the previous section to evaluate the effectiveness of the Diversity Map representation (DM; Figure 3.5) at conveying diversity information by comparing it to the Glyph Hybrid representation [114] (GH; Figure 3.9) discussed in Section 3.3. We chose GH as the baseline for this comparison because it is the only previous method developed specifically to visualize diversity. Nevertheless, in the future work, it will be informative to compare DM with other traditional multiples, such as multiple histograms.

Here we describe the specific implementation of the user study outlined in Section 3.5 that we used to compare the DM and GH representations, and we analyze and discuss the results of this study.

3.6.1 User Study Implementation

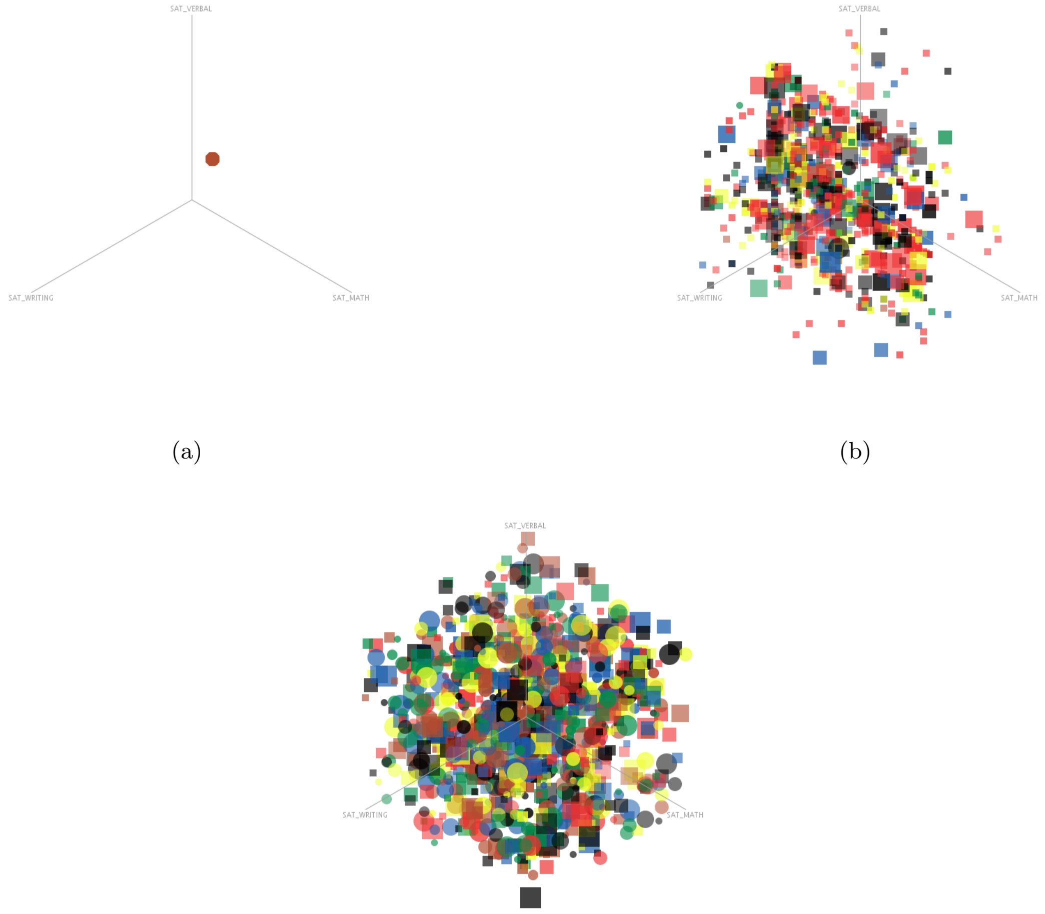
Data. The synthetic data sets we used in our user study simulated college applicant pools where the objects are applicants characterized by the following six attributes:

- SAT Verbal Score: 200-800, discretized by steps of 30
- SAT Math Score: 200-800, discretized by steps of 30
- SAT Writing Score: 200-800, discretized by steps of 30
- Ethnicity: B, H, I, O, W, or X
- Gender: F or M
- Income: Bracket 0, Bracket 1, Bracket 2, Bracket 3, or Bracket 4

We chose the college applicant application because it is one of the three applications examined as a case study by Pearlman et al. [114] and because we believed it would be an application with which our participants, who were all university students, would be familiar. We used single-letter labels as values of categorical attributes (e.g., B, H, I, . . . for Ethnicity) to prevent participants from associating their own knowledge of demographics (e.g. ethnic differences) into their answers.

Participants and Protocol. The participants in our study were 40 students at our university, all with normal color vision. All of the participants volunteered to participate in our study in response to fliers posted around the campus. They represented a diverse range of majors, degrees, and ages (Figure 3.10), and, although their participation in our study might indicate interest in diversity visualization, most of the participants were unfamiliar with the field of information visualization.

After the signing of an informed consent document required by our university's Institutional Review Board, each participant was randomly assigned to different experimental conditions as described below. Participants were encouraged to ask any questions they might have at any time during the course of the study.



(a)

(b)

(c)

Figure 3.9: Synthetic data sets of (a) very low-, (b) medium-, and (c) very high-diversity visualized using the Glyph Hybrid (GH) representation (the accompanying legend is not shown). Each visualized data set contains 1000 objects and 6 attributes (SAT verbal, SAT math, SAT writing, ethnicity, gender, income level). The SAT attributes are mapped to the 3 coordinate axes. Ethnicity, gender, and income are mapped to color, shape, and size of the glyphs respectively. Additionally, opacity encodes composite SAT scores (as in Pearlman's implementation) to remedy ambiguity caused by the many-to-one mapping. The very high diversity data set is 6 times more diverse than the very low data set. The data sets are identical to the ones in Figure 3.5.

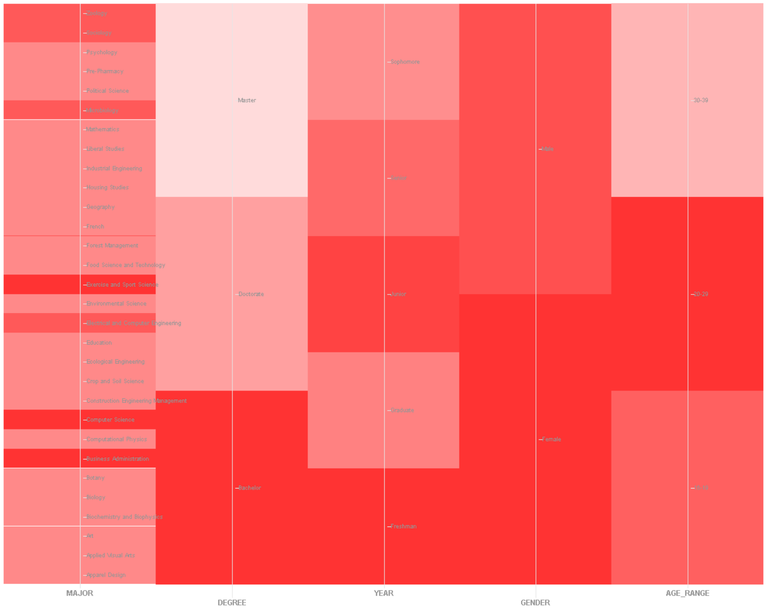


Figure 3.10: Participants of the user study visualized using the Diversity Map. The visualized attributes, from left to right, are major, degree, year in school, gender, and age-range. The participants represented a diverse range of majors, degrees, and ages.

Experiment Design. We followed a two-phase crossover experiment design and used two collections of synthetic data sets, collection A and collection B, for the two phases to avoid learning effect when participants moved from one visualization method to the other. Note that both data set collections are considered equivalent in all respects. They were simply generated with separate runs of the data generation algorithm described in Section 3.5.1.

Each participant's session was divided into two phases. In the first phase, the participant answered questions about visualizations of one collection of data sets created with one visualization method. In the second phase, the participant answered the same questions about visualizations of the other collection of data sets using the other visualization method. The order of visualization methods and data set collections was counter-balanced across participants (see Table 3.1).

In each phase, the participant first completed a short tutorial that explained the visualization method involved in the phase and included several example images generated

Table 3.1: Allocation of 40 participants across four treatments. E.g., 10 (DM, A)–(GH, B) indicates 10 participants answered questions on collection A with DM in phase 1 then on collection B with GH in phase 2.

10 (DM, A)–(GH, B)	10 (DM, B)–(GH, A)
10 (GH, B)–(DM, A)	10 (GH, A)–(DM, B)

using that method. After completing the tutorial, the participant answered several questions of each of the types described in Section 3.5. Note that participants were supplied with a hard copy of each tutorial to consult while answering these questions. Note also that the questions of one type are the same, but each one is asked about visualizations of different data sets. The ordering of question types was randomized across two phases and across participants, but all questions of the same type were asked as a block.

Each participant answered six questions of type Q1. A secondary goal for this question type was to determine whether data set size affected participants' ability to judge and compare overall diversity levels. Thus, each participant was asked Q1 questions of three levels of difficulty (easy, medium, hard) for each of two different data set sizes (100 and 1000 objects). Half of the participants answered questions using the smaller data sets first and the larger ones second, and the other half answered questions using the larger data sets first and the smaller ones second. The order of the three difficulty levels was randomized within each data set size for each participant. This ordering convention was chosen to avoid ordering effects among participants.

Each participant answered three questions of type Q2, and six questions each of types Q3 and Q4. To avoid ordering effects for these questions, we used a counterbalancing/randomization approach similar to the one used with Q1 questions. For all of these questions, we used data sets with only 100 objects. Though our goal is to develop visualizations that can handle data sets with more that 1000 objects, we believed that GH would suffer with larger data sets because of occlusion/clutter. To compare the capabilities of the respective methods to effectively communicate information about diversity, we used data sets with only 100 objects so as not to disadvantage GH.

We collected answers to these questions not only to measure absolute correctness but also to identify how far each participant's response was from the correct answer. We accomplished this by assigning an error distance to each response. For questions of type

Q1, correct responses were assigned an error distance of 0, while incorrect responses were assigned an error distance of 1. For questions of type Q2, Q3, and Q4, the error distance of each response was computed as the rank order of the participant's selected response in relation to the correct answer. In particular, the best (correct) answer was assigned an error distance of 0, the second-best answer was assigned an error distance of 1, the third-best answer was assigned an error distance of 2, and so on. As an example, consider a question of type Q2 whose correct answer was "low diversity." For this question, a response of "very low diversity" would be assigned an error distance of 1, as would a response of "medium diversity," while a response of "high diversity" would be assigned an error distance of 2, and a response of "very high diversity" would be assigned an error distance of 3. We used a similar system to assign error distances for questions of type Q3 and Q4 based on the diversity ordering of the attributes and the cardinality ordering of the attribute values, respectively.

We also collected response times in addition to error distances. Participants were given a time limit of two minutes to answer each question. If the participant did not answer the question in the allotted time, the system timed out and sent the participant to the next question. The participant was assigned the maximum possible error distance for the question type for any question on which he or she timed out.

In addition to the questions of type Q1-Q4, at the end of each phase, the participants answered a short questionnaire about their experience with each method. This questionnaire contained both Likert-style questions as well as open-ended questions. We discuss these questions in more detail in our analysis of the study results below.

The entire study was administered through a web-based interface that collected demographic information, presented tutorials and questions, collected user answers, computed error distances and response times, and stored these in a database for analysis. The resolution of the monitor used was the standard 1920×1200 pixels. The resolution of each image produced by the DM visualization was 900×537 pixels, and the resolution of each image produced by the GH visualization was 640×640 pixels. Each question for the GH method required a legend image of 200×524 pixels. When the legend is taken into account, visualizations of both methods are roughly the same size.

3.6.2 Results and Analysis

Here, we analyze the results obtained from the user study. Our initial hypothesis about these results was that, for each question type, DM would outperform GH, both in terms of accuracy and response time. In particular, we believed that GH would suffer for some questions due to the fact that it does not treat all attributes as equal. Specifically, we expected users to have difficulty accurately judging diversity for the attributes mapped to GH's three spatial axes, due to the ambiguous many-to-one mapping these axes produce. We also expected GH to suffer in terms of time and/or accuracy due to the need for users to consult the legend to remember the mappings.

For each question type, we did not analyze individual answers but computed the sum of error distances and the sum of response times across the questions of that type for each participant and compared the distributions of these aggregated values using statistical hypothesis testing. While we initially planned to use ANOVA (Analysis of variance) and repeated-measures ANOVA directly for this comparison, we found that the response data did not meet these methods' normality requirements. We therefore first applied a rank transformation [33] to the response data before using these techniques.

Our primary focus in analyzing the results of the study is on error distance, since we believe this is the most important performance measure for a given representation. However, we still pay close attention to response time, as well. In all cases, our null hypothesis is that no difference exists between the distributions of corresponding performance measures across the methods DM and GH.

We chose a two-phase crossover experiment design in order to reduce the number of participants and to keep individual subject variability low. However, the design also required us to account for additional within-subjects factors, namely, phase of method (first or second) and collection of data sets (A or B). While we did not expect either of these factors to have a statistically significant effect on our results, this was not the case. Our preliminary analysis showed that the collection of data sets had a statistically significant effect on the error distance for method GH. This effect was not statistically significant for DM. As a result of this effect we analyze error distance separately for each collection. In addition, our preliminary analysis showed statistically significant evidence of an effect of phase of method on response time for DM. Specifically, participants performed slightly faster with DM in the second phase of the study than in the first phase.

Interestingly, there was no significant evidence for this effect for GH. Regardless, due to this effect, we analyze response time using only data collected during the first phase of participants' sessions. Tables 3.2 and 3.3 respectively summarize the error distance and response time results.

Analysis of Results for Q1. *Between two visualizations generated with the same method, which picture represents a more diverse set of objects?* As Table 3.2 indicates, participants answered Q1 questions more accurately with DM than with GH, particularly for collection B. In fact, there is convincing statistical evidence for an effect of visualization method on error distance with collection B, $F(1, 38) = 7.53$, $p = 0.009$. However, with collection A, there is no evidence of such an effect, $F(1, 38) = 0.21$, $p = 0.65$. These results hold consistent when analyzing data separately over 100 and 1000 object data sets, suggesting no effect of data set size on accuracy for questions of this type. With regard to response time, though Table 3.2 suggests that participants performed slightly faster using GH in phase 1, the evidence for this effect is not statistically significant, $F(1, 38) = 1.20$, $p = 0.28$.

While these results do not support our initial hypothesis that users would perform more quickly when using DM than when using GH, they do substantiate our hypothesis that users would be able to more accurately compare the diversity of two data sets when using DM than when using GH. Examining these results more closely, we found that much of the difference in performance between collections A and B for participants using GH was accounted for by the fact that many participants (13 out of 20) incorrectly answered one particular question of medium difficulty from collection B using GH. In this question, the data set with lower overall diversity contained a very diverse Ethnicity attribute, while the data set with higher overall diversity contained a very diverse Gender attribute but a much less diverse Ethnicity attribute. In GH, the Ethnicity attribute is mapped to glyph color and the Gender attribute is mapped to glyph shape. We believe that in answering this question, participants placed more weight on the distribution of color in the visualization than on the distribution of shape, misleading them into an incorrect judgment of overall diversity. If this explanation is correct, it points to an interesting consequence of GH's unequal treatment of attributes. DM, on the other hand, does not seem to suffer from this consequence because it treats all attributes as equal.

Table 3.2: Mean sum of error distances for each question type as a function of visualization method (DM or GH), collection of data sets (A or B), and phase (P1 or P2). Standard deviations are shown in parentheses. The table structure is split by collections of data sets because our preliminary analysis showed that the collection of data sets had a statistically significant effect on the error distance for method GH.

Question	Method	Collection A			Collection B		
		P1	P2	P1&2	P1	P2	P1&2
Q1	GH	0.50 (0.71)	0.40 (0.52)	0.45 (0.60)	0.90 (0.74)	1.40 (0.70)	1.15 (0.75)
	DM	0.60 (1.07)	0.30 (0.48)	0.45 (0.83)	0.50 (0.53)	0.60 (0.70)	0.55 (0.60)
Q2	GH	2.70 (1.25)	3.70 (1.06)	3.20 (1.24)	2.00 (0.94)	2.40 (0.97)	2.20 (0.95)
	DM	2.10 (1.66)	1.70 (0.67)	1.90 (1.25)	1.70 (0.67)	2.10 (0.74)	1.90 (0.72)
Q3	GH	16.10 (2.02)	15.70 (3.09)	15.90 (2.55)	9.10 (3.31)	9.30 (2.98)	9.20 (3.07)
	DM	3.60 (4.53)	3.50 (4.88)	3.55 (4.58)	5.50 (3.27)	4.70 (4.03)	5.10 (3.60)
Q4	GH	2.20 (1.40)	1.20 (1.40)	1.70 (1.45)	2.90 (1.60)	3.10 (2.02)	3.00 (1.78)
	DM	0.50 (1.27)	1.90 (3.38)	1.20 (2.59)	0.70 (1.89)	3.30 (8.27)	2.00 (5.99)

Table 3.3: Mean sum of response times (in seconds) for each question type as a function of visualization method (DM or GH), phase (Phase 1 or Phase 2), and collection of data sets (A or B). Standard deviations are shown in parentheses. The table structure is split by phases because our preliminary analysis showed statistically significant evidence of an effect of phase of method on response time for DM.

Question	Method	Phase 1			Phase 2		
		A	B	A&B	A	B	A&B
Q1	GH	114.40 (53.74)	120.50 (66.44)	117.50 (58.90)	105.70 (53.22)	93.40 (37.66)	99.55 (45.31)
	DM	151.60 (88.32)	121.90 (44.42)	136.80 (69.73)	79.40 (37.89)	91.20 (24.05)	85.30 (31.48)
Q2	GH	53.90 (37.73)	56.00 (21.29)	54.95 (29.84)	41.20 (23.19)	50.00 (23.75)	45.60 (23.29)
	DM	66.90 (26.54)	53.60 (15.21)	60.25 (22.13)	38.70 (31.73)	43.20 (17.85)	40.95 (25.16)
Q3	GH	179.70 (61.18)	208.40 (85.49)	194.10 (73.84)	216.50 (61.67)	180.80 (46.88)	198.70 (56.37)
	DM	143.60 (43.96)	153.30 (43.21)	148.40 (42.72)	98.50 (34.95)	105.30 (29.61)	101.90 (31.72)
Q4	GH	130.50 (44.94)	118.00 (34.91)	124.20 (39.69)	97.10 (18.88)	108.10 (55.89)	102.60 (40.99)
	DM	93.40 (42.86)	120.90 (76.02)	107.20 (61.70)	86.10 (43.18)	53.40 (23.33)	69.75 (37.71)

Analysis of Results for Q2. *How diverse is the data set represented in this picture?* The results for Q2 were similar to Q1's, with participants tending to judge absolute levels of overall diversity more accurately with DM. Again, with GH, users' performance depended heavily on data set collection: participants using GH performed worse on collection A than on collection B. In fact, for collection A, there was convincing evidence for an effect of visualization method on error distance, $F(1, 38) = 15.02$, $p = 0.0004$. For collection B, there was not statistically significant evidence for this effect, $F(1, 38) = 1.56$, $p = 0.22$. Again for Q2, there was no evidence for an effect of method on response time in phase 1, $F(1, 38) = 1.91$, $p = 0.18$.

These results, too, do not support our initial hypothesis that users would perform more quickly when using DM than when using GH, but they do sustain our hypothesis that users would be able to more accurately assign an absolute diversity value to a given data set when using DM than when using GH. Again, more closely examining these results, we found that the three data sets used for Q2 questions from collection A (low, medium, and very high diversity) tended to be more diverse than the corresponding data sets from collection B (very low, medium, and high diversity). With this in mind, we suspect that participants may have been more hesitant to choose a higher diversity response when using GH than when using DM, perhaps because, while it is clear what very low overall diversity looks like under GH (very few spatial locations, colors, shapes, etc.; see Figure 3.9(a)), what very high overall diversity looks like under GH is much more ambiguous (evenly "spread out" glyphs with evenly distributed colors, shapes, etc.; see Figure 3.9(c)). On the other hand, using DM, it was likely much easier for participants to understand exactly how very low and very high diversity appear visually (very low and very high total color density of the entire visual region, respectively; see Figs. 3.5 (a) and (c)), and we believe this led them to be more confident in choosing responses at both ends of the diversity spectrum when using DM for Q2 questions.

Analysis of Results for Q3. *What is the most/least diverse attribute in the data set represented in this picture?* The results for Q3 very much favored DM. There was convincing evidence for an effect of visualization method on error distance for both collections of data sets, A and B, $F(1, 38) = 75.54$, $p = 1.45 \times 10^{-10}$ and $F(1, 38) = 13.565$, $p = 0.0007$, respectively. In addition, there was suggestive but inconclusive evidence for an effect of visualization method on response time in phase 1, $F(1, 38) = 3.50$, $p = 0.07$. These results appear to confirm our initial hypothesis that users would perform better—

both in terms of error distance and response time—when making judgments about the diversity of a single attribute when using DM than when using GH.

Interestingly, participants using GH appeared to perform worse on Q3 questions where the correct answer was an attribute assigned to a spatial axis, likely due to GH's ambiguous many-to-one spatial mapping. In contrast, participants using DM did not appear to favor any single attribute for questions of this type. Again, this suggests that DM's treatment of all attributes as equal is one of its strengths.

Analysis of Results for Q4. *Which value of attribute X contains the most/least objects?* As with Q3, the results for Q4 very much favored DM. For questions of this type, there was convincing evidence for an effect of visualization method on error distance for both collections of data sets A and B, $F(1, 38) = 7.58$, $p = 0.009$ and $F(1, 38) = 25.18$, $p = 1.26 \times 10^{-5}$, respectively, and there was suggestive but inconclusive evidence for an effect of visualization method on response time in phase 1, $F(1, 38) = 2.61$, $p = 0.11$. Again, these results support our initial hypothesis that users would be able to more quickly and more accurately make judgments about relative abundances within a single attribute when using DM than when using GH.

Summary. The results across Q1–Q4 consistently supported our hypothesis that users would be able to make more accurate judgments about various aspects of the diversity of data when using DM than when using GH. While we found some evidence suggesting that users performed more quickly with DM than with GH, these results were not conclusive. Similarly, we found no conclusive evidence that size of data set had an effect on user performance for questions of type Q1.

3.6.3 Subjective Evaluation

After each participant answered all of the questions of types Q1–Q4 for a particular method, he or she also completed a short questionnaire on that method. The questionnaire, whose form we adopted from [147], consisted of nine Likert-style statements, where participants were asked to indicate their level of agreement on a scale of 1 (strongly disagree) to 5 (strongly agree), and three open-ended questions.

Table 3.4 lists each of the Likert-style questions along with the participants' mean responses for both GH and DM. Participants slightly favored DM over GH in making judgments of diversity components and this is consistent with their performance in the objective portion of the study. Participants also slightly favored DM over GH in terms of applicability, ease of understanding, and affinity.

Table 3.4: Mean responses to each of nine Likert-style statements presented to participants immediately after using each visualization method. These responses are based on a scale of 1 (strongly disagree) to 5 (strongly agree). Standard deviations are shown in parentheses.

Statement	GH	DM
L1) I was able to compare the diversity of two data sets using this method.	3.75 (0.81)	3.93 (0.92)
L2) I was able to judge the diversity of a single data set using this method.	3.63 (0.90)	4.25 (0.84)
L3) I was able to determine the most/least diverse attributes in a data set using this method.	3.58 (0.96)	4.15 (0.86)
L4) I was able to determine the ethnicity with the most/least objects using this method.	4.05 (0.88)	4.28 (0.82)
L5) After the initial training session, I knew how to use this method well.	3.33 (0.83)	3.55 (0.99)
L6) After answering all of the questions, I knew how to use this method well.	3.74 (0.88)	3.88 (0.91)
L7) There are definitely times that I would like to use this method.	3.20 (1.04)	3.75 (0.93)
L8) I found this method to be confusing.	3.38 (1.21)	2.77 (1.13)
L9) I liked using this method.	2.95 (0.96)	3.50 (1.01)

In addition to the Likert-style statements, the questionnaires included the following three open-ended questions:

O1) What aspect(s) of this method did you like most?

O2) What aspect(s) of this method did you dislike most?

O3) If possible, how would you change this method to improve it?

Many participants indicated an affinity for GH because it was intuitive, in that, as the diversity of the underlying data increased, so too did the diversity of the visual properties (color, shape, size, etc.) of the generated visualization. On the other hand, many participants expressed concern about GH's ambiguous spatial layout, which they found confusing.

Participants indicated that they liked the "clean layout" of DM; the simplicity of comparing color opacity under DM; and its ability to easily handle different data set sizes. On the other hand, some participants disliked comparing the diversity of an attribute with several bins (e.g. ethnicity) to that of an attribute with only a few bins (e.g. gender). Interestingly, though this appears to be an issue with GH as well, participants did not seem to notice it when using GH.

Finally, most participants (29 out of 40) preferred DM to GH. In general, participants tended to feel GH would be best suited for judging the overall diversity of a data set, especially to determine if the set is not diverse. Interestingly, this is in direct contradiction to their performance in questions Q1 and Q2 which favored DM. In contrast, participants generally believed DM would be useful for investigating the data more deeply and examining the diversity of separate attributes.

3.7 Formative Evaluation with Ecologists

We also deployed an interactive version of DM for use by ecologists at Oregon State University. Initial findings indicate that the visualization is valuable for ecologists in the early stages of data exploration, prior to further statistical analysis. Here we demonstrate how the tool can help scientists gain insights into spatial and temporal patterns of diversity and abundance in ecological data. The use of this tool is illustrated with data on moth diversity and abundance from the H.J. Andrews Experimental Forest. The data set has been difficult to analyze because the data set is large ($> 69,000$ individual moths), many species (> 500) are present, common species are widespread, and most species are rare (see Figure 6.4).

3.7.1 Moth Diversity and Abundance Data

Ecologists have sampled moths in the $64-km^2$ H.J. Andrews Experimental Forest (HJA) and Long Term Ecological Research (LTER) site within the Willamette National Forest, Lane County, Oregon (Figure 3.12) for five years. Moths were sampled at 20 sites every two weeks from May-October from 2004 to 2008, using UV light traps. Moth abundance refers to the number of individuals caught in a single trap in a single night, or the total number of individuals in any aggregated assemblage of trapping events. Host plants for

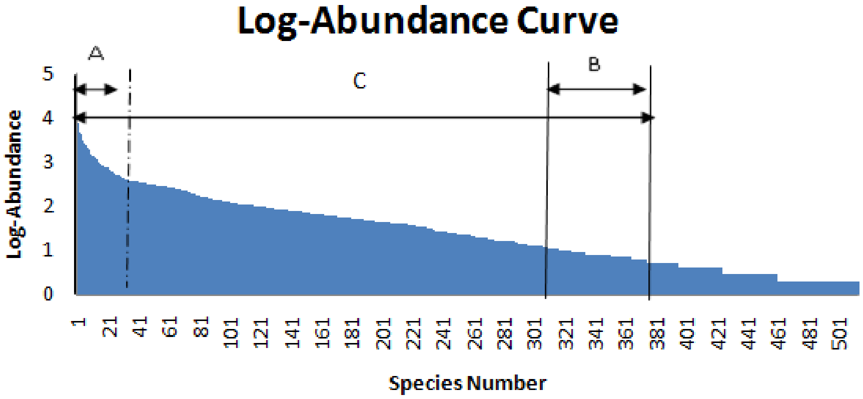


Figure 3.11: Rank abundance curve (with logarithmic scale) showing the evenness of moth species in the moth dataset [106]. A shows the common moths, B shows the rare moths, and C shows the common through rare moths.

moths, if known, were based on Miller and Hammond [107]. Additionally, the following environmental variables were used to explain the distributional patterns of moths: calendar day (sampling period), temperature (accumulated heat-units), vegetation type, watershed, and elevation. Values of vegetation type, watershed, and elevation are determined based on trap sites and values of temperature are based on sampling periods. The structure of the data set is described in Table 3.5.

In summary, a total of 69,168 individual moths from 514 species were captured (Figure 6.4). Species richness was high, but most species were rare, producing highly varied patterns of diversity (Figure 6.4). Fifty-four (10%) of the 514 moth species were represented by only 1 individual, and 46 (9%) were represented by 2 individuals.

We used two subsets of the entire moth dataset in the analyses: 26 common moth species and 66 rare moth species. We define common moth species ($n = 26$) as those for which 500 or more individuals were captured over the entire five-year sampling period. We define rare moth species ($n = 66$) as those for which a total of 5-10 individuals were captured over the five year sampling period. Note that we do not include moths with 1-4 individuals as part of the rare moths because we assume that an average abundance of at least one per year will provide enough information to identify the moth's spatial and temporal associations. Moth species with 1-4 individuals will not provide the level of detail needed to sufficiently identify the environmental associations of the moth species.

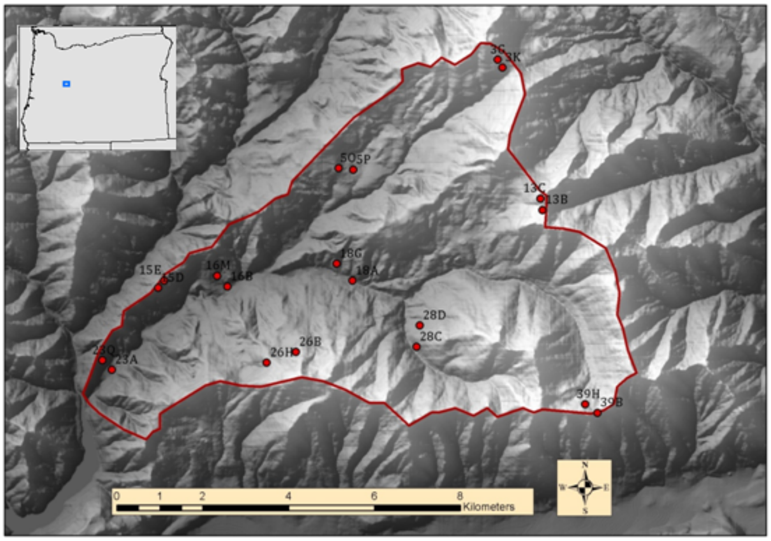


Figure 3.12: Map showing the location of the Andrews Forest in the central western Cascades, Oregon with 20 moth trap sites (red dots). The red line is the boundary of the forest.

For example, singletons and doubletons are very difficult to understand because they do not occur often enough to analyze statistically.

The 26 most common moth species (A in Figure 6.4) accounted for 41,889 individuals (60.6% of the total abundance). The 66 moth species considered as rare (B in Figure 6.4) accounted for 467 individuals (0.7% of the total abundance).

3.7.2 Visualization of the Moth Data

Figures 3.13 and 3.14 depict the two subsets of common moth species and rare moth species visualized with DM. In this DM implementation, we scale the number of moth individuals in each bin according to the total abundance of all individuals in the visualization. Thus, the opacity of each bin x is calculated as $f(x) = |x|/|total|$, where $|x|$ denotes the number of individuals in bin x and $|total|$ is the total number of individuals from the visualized data set. Although we use linear scaling in our implementation, the method can accommodate other forms of scaling, such as logarithmic, for species whose abundances span multiple orders of magnitude [101]. We choose white as the

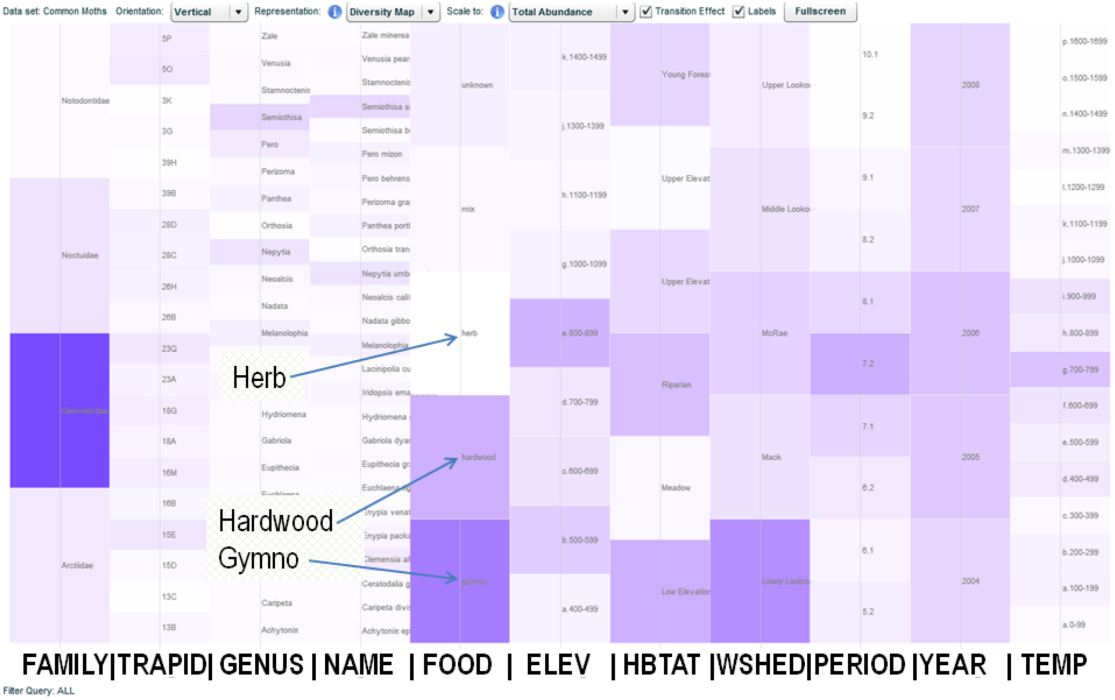
Table 3.5: Structure of the moth data set. Each row represents a moth species with non-zero individual abundance (NO_INDIV) collected at a trap site on a sampling date. Numerical attributes are discretized based on convenient divisions of the data.

Attribute Name	Type	Description
LEP_NAME	categorical	Lepidoptera (moth) scientific name; includes genus and species
LEP_FAMILY	categorical	Lepidoptera taxonomic family
LEP_GENUS	categorical	Lepidoptera taxonomic genus
FOOD_PLANT	categorical	Host functional feeding group
TRAP_ID	categorical	Identifier for a trap site
ELEVATION	numerical	Elevation. Discretized by 100m band.
HABITAT	categorical	Habitat
WATERSHED	categorical	Watershed
COLLECT_PERIOD	categorical	2-week collect period. E.g., 7.2 represents the second half of July
COLLECT_YEAR	categorical	Collect year
TEMPERATURE	numerical	Temperature (Heat unit). Discretized by 100 unit band.
NO_INDIV	numerical	Number of individuals

background color and blue as the foreground color, because the human eye is known to be more sensitive to changes in blue than in other colors [98]. We map opacity values to values in the CIELAB color space [156], which is perceptually uniform, meaning that a visual difference in color opacity is equally perceptible across the range of that color. We then convert CIELAB values to RGB values for representation on a computer screen. In addition to the DM representation (opacity encoding), the visualization tool also allows users to switch to a multiple histogram representation (bar length encoding) (Figure 3.15).

In addition to the representation, we also equipped the DM visualization with interactive features. These features allow the transformation of the view to alternative views so that users can interact with and explore their data. In particular, these features can be used to query the data (e.g., filtering) and to change the representation of the data (e.g., switch between DM and multiple histogram representations or sort the bins within an attribute).

Notably, data filtering extends the static DM to facilitate subsetting of data. For

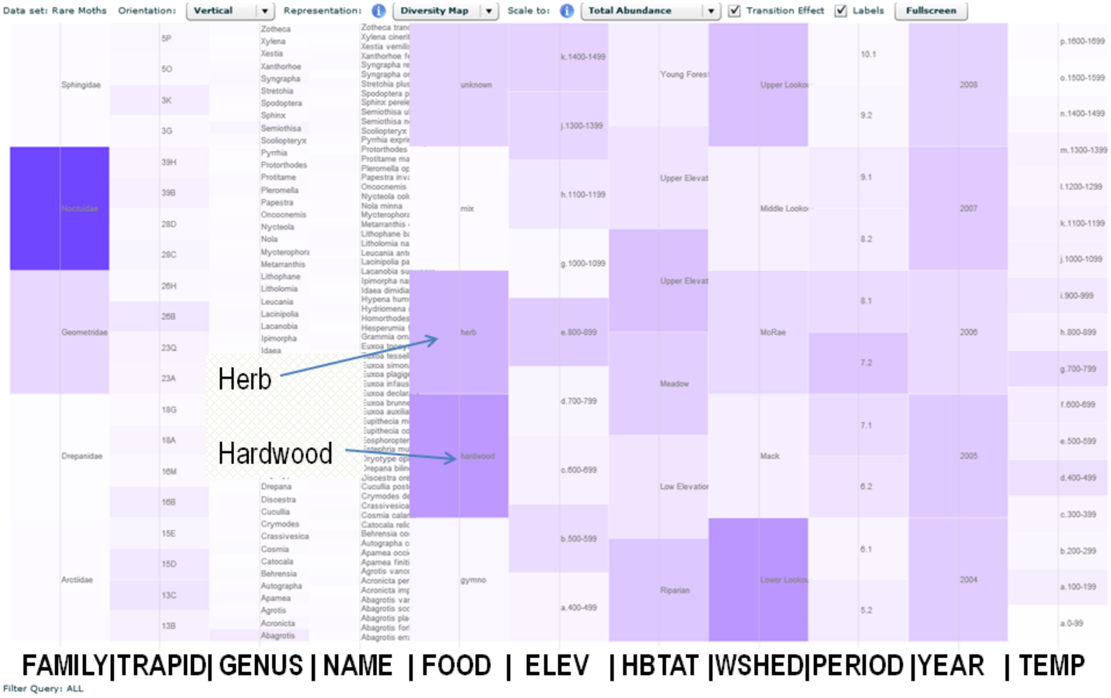


Data set: Common Moths Orientation: Vertical ▾ Representation: Diversity Map ▾ Scale to: Total Abundance ▾ ☑ Transition Effect ☑ Labels Fullscreen

Filter Query: ALL

Figure 3.13: The DM representation of common moths. The data set contains 41,889 individual moths and 11 attributes whose labels are enlarged (columns from left to right: LEP_FAMILY, TRAP_ID, LEP_GENUS, LEP_NAME, FOOD_PLANT, ELEVATION, HABITAT, WATERSHED, COLLECT_PERIOD, COLLECT_YEAR, TEMPERATURE). The view shows that common moths are associated with common habitats (conifer forests in the HJA) and therefore, mostly conifer-feeders (Hardwood and Gymno): 'Gymno' and 'Hardwood' are the most opaque bins within the FOOD_PLANT axis while 'Herb' is the most transparent bin.

example, a user can constrain, or "filter," a single attribute or multiple attributes to one or more particular values (bins) (e.g. show all moths that were sampled at TRAP_ID X and in COLLECT_YEAR Y). The remaining attributes then display the distribution of only those individuals that fall within the specified range of the filtered attribute values. Filtering facilitates direct comparison of the attributes of a subset of specific samples as well as comparisons of subsets of data. To construct a complex filtering query consisting of multiple bins (or attribute values), we follow a simple and commonly used rule articulated by ecologists: bins within an attribute are connected by the "OR"



Data set: Rare Moths Orientation: Vertical Representation: Diversity Map Scale to: Total Abundance ☑ Transition Effect ☑ Labels Fullscreen

Filter Query: ALL

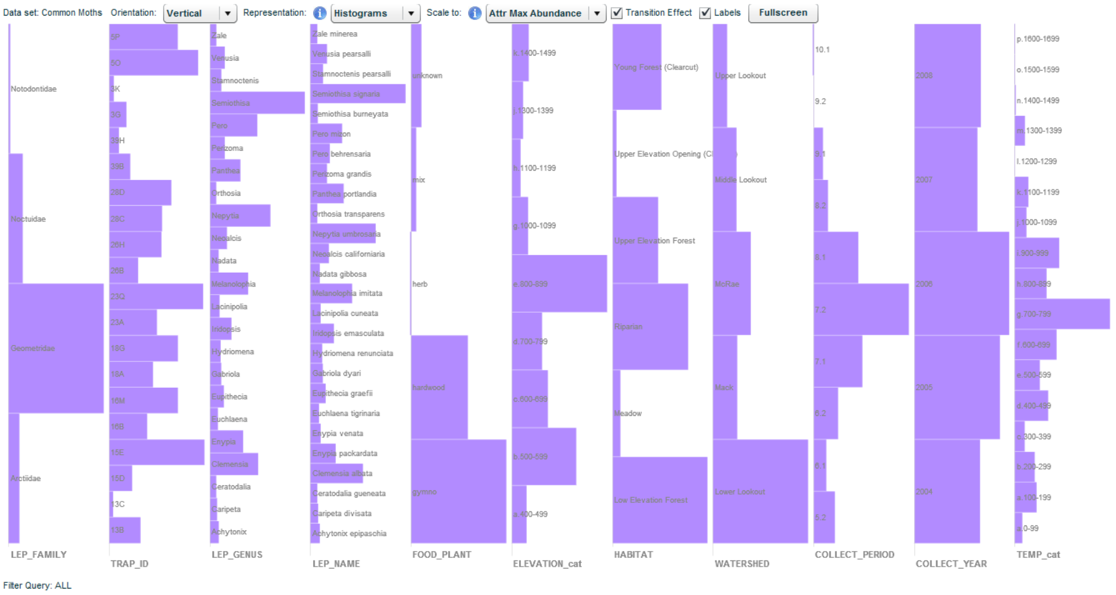
Figure 3.14: The DM representation of rare moths. The data set contains 467 individual moths and 11 attributes (whose labels are enlarged) ordered as in Figure 3.13. The view shows that rare moths are associated with rare habitats (meadows in the HJA) and therefore, herb and grass-feeders.

condition, whereas groups of filtered bins across attributes are connected by the "AND" condition.

3.7.3 Exploration of the Moth Data - Example Scenarios

Here we illustrate the value of the DM visualization by several example scenarios of ecologists exploring the moth data sets. The ecological findings presented in this section are primarily for demonstrating the utility of the visualization. Ecology readers are encouraged to refer to Highland [68] for a more detailed analysis of these findings.

First, without requiring any interactions from users, the overview of moths (Figures 3.13 and 3.14) quickly suggests that common moths are associated with common habitats (conifer forests in the HJA) and rare moths are associated with rare habitats



Data set: Common Moths Orientation: Vertical Representation: Histograms Scale to: Attr Max Abundance ☑ Transition Effect ☑ Labels Fullscreen

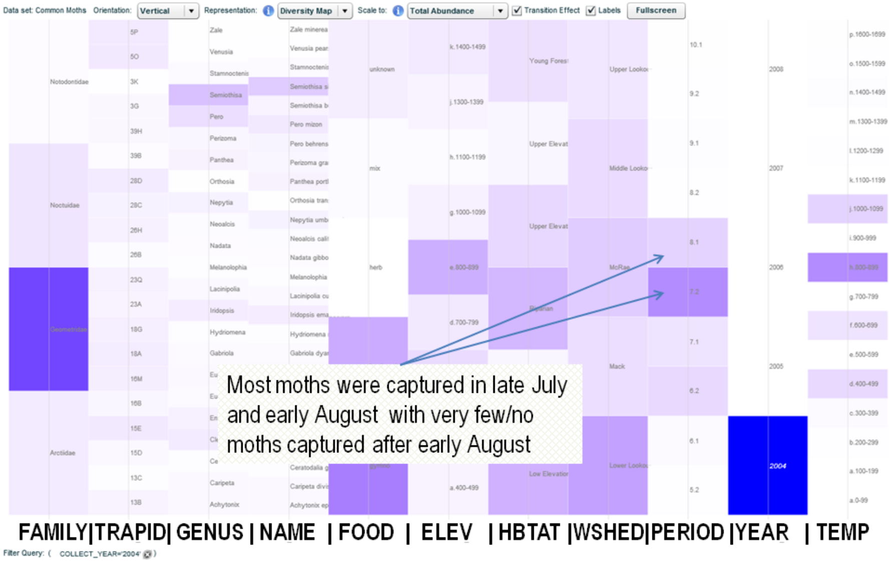
Filter Query: ALL

Figure 3.15: The multiple histogram representation of common moths. The visualized attributes are ordered as in Figure 3.13. Users can select their preferred representation in the drop-down list located on the control bar at the top. In each of the histograms, the bars are pointing to the right (in contrast to the familiar upward-pointing display). The length of each bar is scaled according to $l(x) = |x|/|x_{MAX}$ where $|x|$ denotes the number of observations in bin x, and x_{MAX} is the bin with the most observations for the variable in question.

(meadows in the HJA). In addition, the visualization shows that common moths are mostly conifer-feeders and rare moths are mostly hardwood, herb, and grass-feeders. That is, the view of common moths (Figure 3.13) shows 'gymno' is the most opaque bin within FOOD_PLANT axis and the view of rare moths (Figure 3.14) shows 'herb' and 'hardwood' are the most opaque bins within the same axis.

Second, consider this example, which demonstrates how interactions facilitate the investigation of temporal relationships in the moth data sets. Because moth development is temperature dependent, ecologists hypothesize that adult moths emerge earlier in warm years and later in colder years. According to the temperature records, while 2004 was a warm year, 2008 was a much colder year. Ecologists can filter the moth records by COLLECT_YEAR and/or COLLECT_PERIOD to observe temporal trends. The views help verify that the peak in common moth abundance occurred earlier in 2004 (and

2006) than in 2008 (Figure 3.16 top and bottom). Note that they show moth capture by 2-week sampling period (8th column) and by degree days (last column). In 2004, most

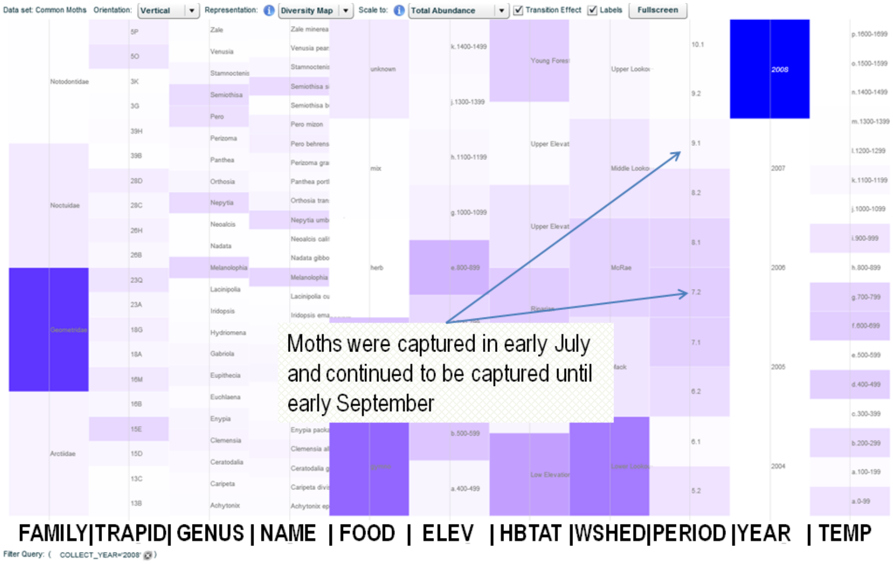


Data set: Common Moths Orientation: Vertical Representation: Diversity Map Scale to: Total Abundance ☑ Transition Effect ☑ Labels Fullscreen

Most moths were captured in late July and early August with very few/no moths captured after early August

FAMILY|TRAPID| GENUS | NAME | FOOD | ELEV | HBTAT |WSHED|PERIOD |YEAR | TEMP

Filter Query: (COLLECT_YEAR='2004')



Data set: Common Moths Orientation: Vertical Representation: Diversity Map Scale to: Total Abundance ☑ Transition Effect ☑ Labels Fullscreen

Moths were captured in early July and continued to be captured until early September

FAMILY|TRAPID| GENUS | NAME | FOOD | ELEV | HBTAT |WSHED|PERIOD |YEAR | TEMP

Filter Query: (COLLECT_YEAR='2008')

Figure 3.16: The DM representations of common moths sampled in COLLECT_YEAR of '2004' (top) and of '2008' (bottom). The views help verify that the peak in common moth abundance occurred earlier in 2004 (warm year) than in 2008 (cold year).

moths were captured in sampling periods 7.2 and 8.1 with very few/no moths captured after 8.1, whereas in 2008, moths were captured in sampling periods 7.1 to 8.1 and continued to be captured until 9.1. Common moths were initially captured in a much more concentrated time span in 2004 than 2008, with many more moths initially captured later in the year in 2008 than in 2004. In this example, while ecologists need to observe only three attributes (COLLECT_YEAR, COLLECT_PERIOD, and TEMPERATURE) to answer their question, they can potentially look at other attributes for additional insights. For example, they may initially pre-define the ordering of moth species in LEP_NAME attribute (e.g., by abundance) and then quickly verify whether the ordering pattern remains consistent over these two years.

3.7.4 User-Centered Design with Ecologists

A close collaborative effort between ecologists and visualization researchers was required to understand the analysis process for integration of the DM into active research. We employed a user-centered, participatory design approach (Figure 3.17) [134, 132] where the ecologists were included as part of the design team from the beginning of the collaborative effort. The initial prototype of the DM served as the starting point for this particular collaboration.

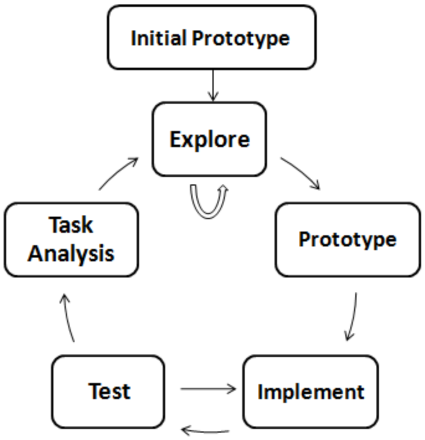


Figure 3.17: The collaboration between ecologists and visualization researchers taking an iterative user-centered, participatory design approach

The initial prototype was initially developed for a small subset of the data, and it proved invaluable as a means for stimulating discussion and identifying design alternatives. In early meetings, the prototype served as a way to introduce the ecologists to the visual representation in the particular context of their data set. Subsequent meetings followed a very informative and dynamic process. In particular, each session generally started with the visualization team running the visualization, projecting the view onto a large screen for the entire team to view. The ecologists would then begin to explore the data set in an iterative fashion, asking questions and modifying views to answer those questions, and repeating. The process was typically very fast-paced and very collaborative with team members posing questions to each other and devising views together to answer those questions. When a question could not be answered using the provided representation and interactions, the entire team would break from the exploration cycle to discuss how the system could be modified to further enhance the application. In the weeks following each meeting, the visualization team would integrate the design modifications into the system in preparation for the next design meeting. As the design matured, the work centered more on dedicated exploration and analysis of the data set.

3.8 Discussion and Future Work

We have presented (1) an infrastructure for studying the problem of diversity visualization, (2) a novel representation for visualizing diversity patterns in separate attributes of a large set of multivariate objects, and (3) a rigorous evaluation of the effectiveness of the proposed technique. The infrastructure includes a precise definition of diversity that takes both richness and evenness into account, a method for generating synthetic data of controllable levels of diversity, and a formal user study design for evaluating diversity visualization representations. Based on this definition and study design, we developed and evaluated our approach to diversity visualization, the Diversity Map, which is based loosely on ideas from both parallel coordinates and multiple histograms. In a formal user study, we show that DM allows users to as or more accurately judge elements of diversity than the only other existing method designed to visualize diversity. The results across task questions are also more consistent for DM. The DM representation was further refined into an interactive tool for ecologists to explore diversity patterns and processes in ecological diversity data. While we believe we have taken a positive step in understanding diversity visualization, there are several issues left to address as well as lessons learned.

3.8.1 Study Design Issues

First, while our study design focuses on static visualizations only, both DM and GH are interactive visualizations. We avoided interactive features to limit the scope of our study to first understand the merits and shortcomings of DM and GH as *representations*. As is well known, evaluating the ability of a fully interactive visualization to facilitate insight is a difficult and open problem that remains for the information visualization community. This is evidenced by the recent conference focused solely on methods to perform this kind of evaluation [13]. Future work will address the interactive capabilities of DM (see Chapter 4).

Additionally, implementing GH required us to choose a mapping of attributes to the various visual properties of the representation (the three spatial axes, color, size, shape, etc.). While we based our mapping on the one used by Pearlman et al. [114], our choices here nonetheless represent a possible threat to construct validity.

Finally, our study does not include a specific question to determine the richness of variety of an attribute. At first glance, it would appear that richness of variety was obvious in both methods. However, while richness is clearly communicated in DM and in the non-spatial attributes of GH (e.g. color, shape, size), it is not clear how well richness is communicated in the spatial axes of GH (e.g. the richness of SAT scores in Figure 3.9 is ambiguous). The evaluation study would benefit from explicit attention to the ability to communicate richness.

3.8.2 Limitations of Diversity Map

The DM representation itself is also not without limitations. First, DM is currently designed to visualize only categorical data, requiring a discretization of quantitative attributes. Second, the static visualization provides limited insight into the relationships between attributes. However, variations on the interactive version of DM can address these problems. For example, traditional parallel coordinates poly-lines can be selectively displayed over DM to allow the user to view the actual quantitative attribute values. These poly-lines also allow the user to see and select individual objects, which are currently not visible in the static DM visualization as presented. Filtering techniques are also implemented in the interactive version of DM to allow users to perform queries. For example, as demonstrated in the formative evaluation of ecologists, the user can constrain a single attribute to one or more particular values (bins) using the mouse. The remaining attributes then display the diversity of only those objects that fall within the specified range of the filtered attribute. With filtering, users can answer questions regarding the relationship between two attributes such as “In what bin in attribute X are objects most/least diverse in attribute Y?”.

While the DM representation scales well with the number of objects to be visualized, like many multivariate visualization methods, scaling with an increase in attributes is limited by screen space. Likewise, the number of bins for any one attribute is also similarly limited, and it is not clear how “small” a bin can be made before the representation becomes ineffective. Studies to understand these limitations are left for future work.

The DM representation, like many others, requires initial training for users to be effective in reading the visualization. Indeed, many pilot users of the visualization found the representation counter-intuitive. They assumed that if a representation is to convey diversity, high diversity should be shown with an image in which all of the objects look different, however, in our implementation, high diversity results in a uniform image (see Figure 3.5(c)). This confusion stems from the users associating each box with an individual object to be visualized. Once users understood that DM did not display individual objects, but rather their distribution over the attribute space, they were much more receptive and able to interpret the visualizations consistently as shown in the study. We believe that this representation, where the space filling effect denotes diversity, helps the user establish a baseline for high diversity. In fact, this is supported by the results of our formal study. While participants tended to underestimate diversity when using the GH method, where more diversity implies more dissimilar symbols, they were able to more accurately assign absolute diversity values as well as compare diversity between two data sets using DM.

Note that initially we empirically chose white as the background color and red as the foreground color in DM. However, we have since found studies indicating that blue may be a more appropriate foreground color, since our eyes are known to be more sensitive to changes of saturation (or opacity) in blue than in red, given a fixed hue and

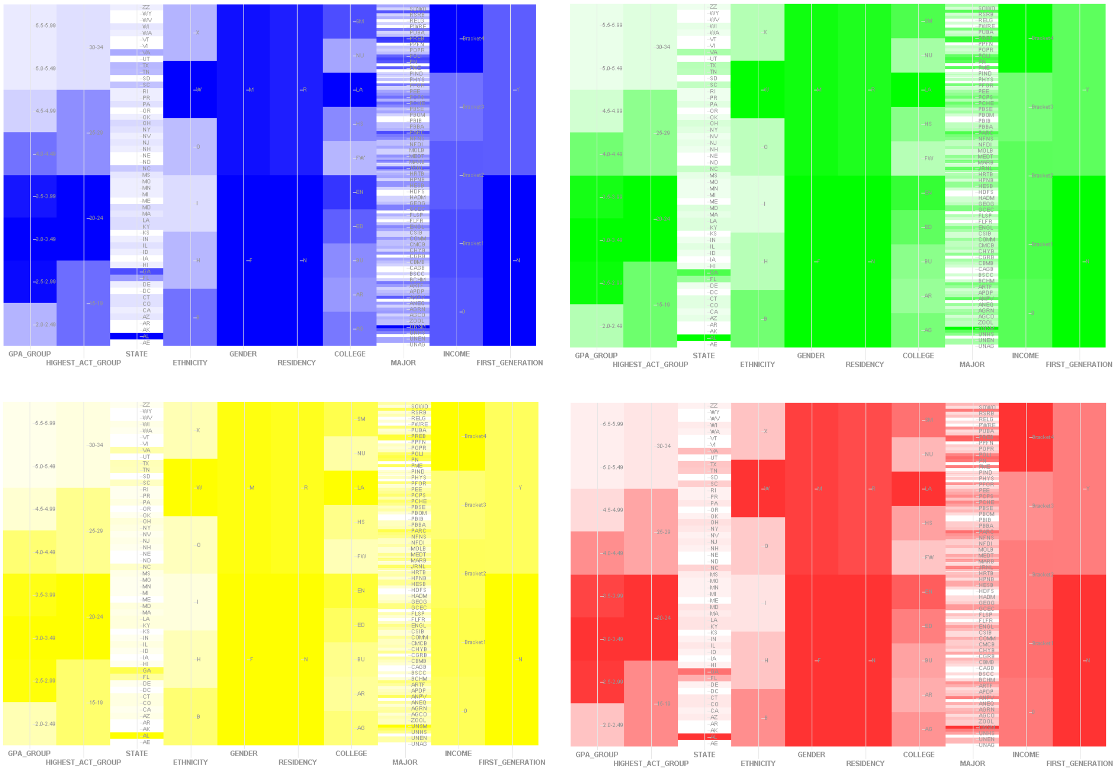


Figure 3.18: The subset of college applicants recommended for acceptance visualized using DM in different foreground colors. Opacity (or saturation) values are mapped to values in the CIELAB color space before converted to RGB values for representation on a computer screen. That is, given a fixed hue and a fixed brightness, increasing saturation adds intensity to colors, changing color from white to saturated colors. According to the CIELAB color space, the distance from no saturation to full saturation is the longest for blue [98]. Therefore, changes of saturation levels in blue (top-left view) appear the most noticeable to our eyes—followed by red, green, and yellow.

a fixed brightness [98]. Figure 3.18 demonstrates this phenomenom. Additionally, we empirically used the square root-based normalization in determining the color opacity (α-value) of bins to help make bins corresponding to attribute values with low abundance more recognizable. Nevertheless, this ad-hoc scaling factor is not necessarily a preferred choice by all users. In fact, ecologists may prefer a log transformation to accommodate species whose abundances span multiple orders of magnitude [101]. Moreover, we could employ an alternative to the RGBA color space, such as *CIELAB* or *CIELUV*, which are perceptually uniform color spaces and may be more appropriate for representing quantitative abundances [156]. All of these issues have been addressed in the interactive version of DM as described in the formative evaluation with ecologists.

3.8.3 Limitations of Our Definition of Diversity

Our definition of diversity generalizes the one used in the field of Ecology to the case of arbitrary multivariate data. As a consequence our definition looks at the diversity of each attribute *independently* and does not take into account the interaction between attributes. Consider an example of two teams of employees that have four members each (see Table 3.6). While it is obvious that Team 2 is divided into more subgroups, the current definition concludes that both teams are at the same level of overall diversity with respect to gender and age–that is, in each of the two teams, members are uniformly distributed in both gender and age. In Chapter 5, we will investigate a definition to account for this interaction. The area of business management provides useful insights as researchers in that field discuss diversity across multiple attributes [90, 55].

Table 3.6: Employee Diversity Example. Each of the two teams has four members.

Team 1	
Female, over 50	Male, under 50
Female, over 50	Male, under 50

Team 2	
Female, over 50	Male, over 50
Female, under 50	Male, under 50

3.8.4 Lessons Learned from the Evaluation with Ecologists

The user-centered design process was important in reaching a design that truly met the needs of the target users (ecologists). An initial prototype was a key component in starting the ‘discussion’ between ecologists and visualization researchers and helping the design team to understand the exploration process (Figure 3.17). Although the prototype may not be the final design, some means for rapidly exploring the data allows the team members to begin to understand the typical process and types of questions they can and would like to ask of the data.

Characteristics/Process. Given interactive tools, ecologists were able to quickly and iteratively explore data that was originally in a very inaccessible format. The visualization provided an environment in which ecologists could rapidly answer questions and visually verify expected relationships. The process was typically iterative with several cycles of starting with a question, taking an exploration path, getting insight, and then starting over with a different path through the data. In some cases, ecologists felt the need to explore two paths simultaneously to observe the differences in the outcome. This multiple path exploration capability is a fundamental requirement of creativity tools [142]. Data analysis through visualization must support the creative process of hypothesis generation.

Data Queries. In this particular collaborative effort, the visualization served as a means for rapid high-level exploration of complex data that was then followed with detailed statistical analyses. Data exploration tools, such as the DM, which overview the data, should provide mechanisms for exporting subsets of data associated with the current view so that scientists can conduct appropriate statistical analyses.

Communication. On several occasions an ecologist sought to explain a particular insight or finding by walking the team through the necessary interactions to produce a specific view. Exploration tools must provide mechanisms for storing and retrieving history in order to help users tell their stories. In addition, the tools need to permit users to mark and recreate paths of exploration in order to explain ideas to one another.

Context of Collaboration. Our meetings were typically held in a conference room in the computer science building. On several occasions, the team would have benefited from being located in the context of the ecologist so that the team could refer to or use artifacts that are typically at their disposal—such as topographic maps. A more contextual design process that included, for example, sessions in the office of an ecologist or visits to field sites, might have revealed additional useful views/tools that would provide powerful insight capabilities when combined with the visual representation.

Educational Outreach. Education and outreach are key components of the H.J. Andrews Experimental Forest and LTER. We believe that visualization tools are promising in this setting, because they provide a mechanism for clearly communicating complex ideas and data through images, which are often more easily explained than data sets and scientific findings. We plan to make the tool publicly available to a broader audience, including scientists, students, and educators (see Chapter 4). The tool will allow users to explore existing HJA data sets or upload and explore their own data sets.

Role of Diversity in Real-world Analysis. While diversity is a key element, it may not necessarily be the only element in the real-world analysis process. Built upon the DM representation, we have deployed a visual analysis tool to ecologists that targets ecological long-term data sets with an emphasis on diversity/distribution patterns and temporal trends. We describe the tool and the processes whereby an ecologist explores data, generates and test hypotheses in Chapter 4.

3.9 Conclusions

The Diversity Map represents a first attempt to design a representation with the specific goal of visualizing diversity in separate attributes as we defined in this chapter. Subsequent chapters extend the representation in two directions. First, the representation is developed into a fully functional interactive tool for ecologists exploring long-term ecological data. The goal is to understand the role of interactive visualization in the real-world analysis process in which diversity is a key element (Chapter 4). Second, in Chapter 5, we will investigate a definition of diversity to account for the interaction among attributes. In doing so, we will present a study exploring the design space for graphical representation of team diversity faultlines, a fundamental construct in organizational management.

Chapter 4: EcoDATE–Exploratory Analysis of Distribution Patterns and Temporal Trends in Long-Term Ecological Data [1]

4.1 Introduction

In recent decades, scientists have witnessed the proliferation of complex and large data sets within many fields of study. In ecology, observations of long-term change are the key to understanding ecosystem function and environmental change (e.g., [85, 17, 50, 138, 100, 25, 60, 91, 129, 19]). In ecosystem and community ecology, long-term trends in stream water nutrient concentrations and fluxes from watersheds are used to examine ecosystem dynamics, such as retention and flux of nutrients and atmospheric pollutants [96]. Similarly, long-term data on plant succession are used to analyze temporal changes in community composition, structure, biomass and nutrients (e.g., [40]).

Long-term ecological studies commonly involve a variety of data sets and hypotheses, but the analysis usually follows three main steps: (1) collect ecological and—hopefully—relevant environmental data; (2) plot and observe overall distributions, temporal trends, and correlation of variables in typical charts such as static histograms, line charts, and scatter plots; and (3) use statistical tests to confirm or refute the initial hypotheses. This approach may work well when the number of variables is small and interesting hypotheses can be preconceived. When data sets span many decades, it is likely that the hypotheses and objectives, under which a study began, evolve as a result of unforeseen trends as well as changes in the knowledge and perceptions of the scientists who work with or inherit the data and experiments. Thus, exploration of new or alternative hypotheses is an inherent part of long-term studies. The exploration process usually involves hypothesis generation as opposed to hypothesis testing, decision-making, scientific modeling, or theory development [154, 8].

Interactive visualizations of data, when combined with traditional analysis approaches, offer the potential to facilitate exploratory data analysis (Figure 4.1), provided that the

[1] The material in this chapter was previously published with co-authors Julia Jones, Ronald Metoyer, Frederick Swanson, and Robert Pabst in [121].

charts and interactivity fulfill the analytical needs of ecologists and are well suited to characteristics of long-term data. Nevertheless, while typical static charts such as histograms, scatter plots, and line charts have been used by scientists to explore distribution patterns and temporal trends in individual variables, little work has been done to develop interactive visual-analysis tools that support rapid exploration of large, multivariate, and long-term data. The paucity of tools also hinders understanding of potentially different strategies and processes whereby scientists gain knowledge and generate hypotheses from long-term data.

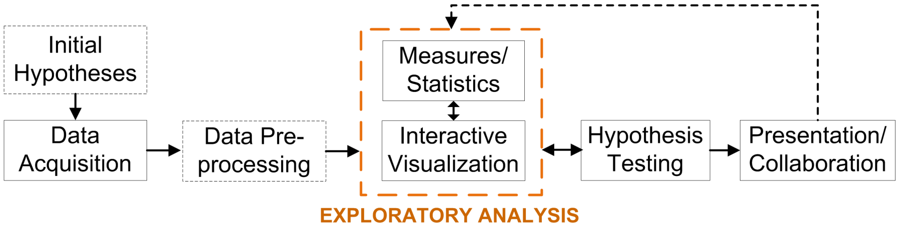
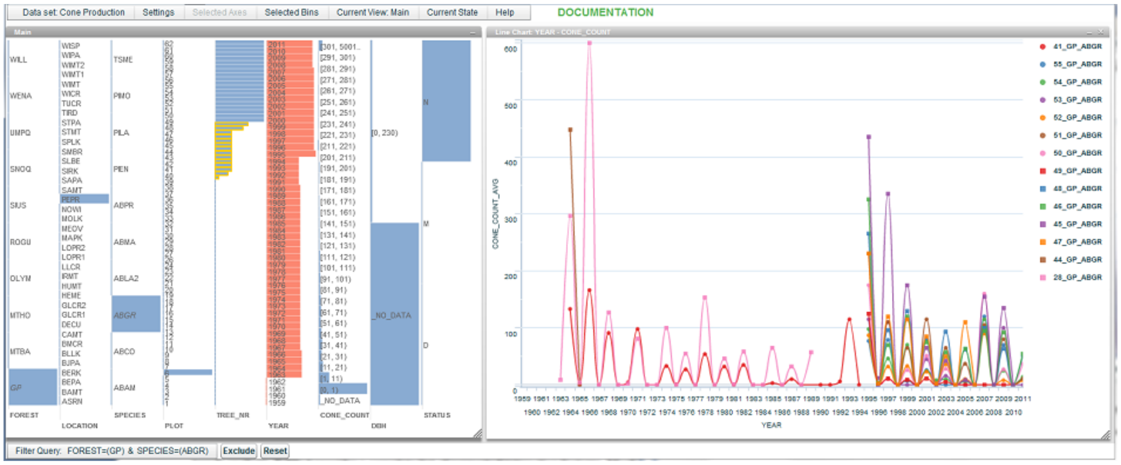


Figure 4.1: The visualization driven exploratory analysis process the EcoDATE tool aims to support. Each rectangle represents a subprocess and each arrow represents a direction the user can take to go through the process.

We have developed the *Ecological Distributions and Trends Explorer* (EcoDATE), a web-based visual-analysis tool that facilitates the collaborative visual inspection of the distribution patterns and temporal trends of ecological long-term data (Figure 4.2). It was refined and evaluated using the user-centered design approach [143, 132] in which ecologists worked closely with visualization researchers during all stages of the development process from assessing analytical needs to testing. The tool, which is readily available at `http://purl.oclc.org/ecodate`, supports multiple chart views and a wide range of interaction features involving collaboration of multiple users.

This chapter describes the development and initial application of the tool to three large, long-term data sets: cone production [80], stream chemistry [79], and forest structure [54] collected as part of the H.J. Andrews Experimental Forest (HJA), Long Term Ecological Research (LTER), and US Forest Service Pacific Northwest Research Station programs (`http://andrewsforest.oregonstate.edu/`). We describe how ecologists have used this tool to overview these datasets, examine and compare distributions and temporal trends, and generate and share hypotheses with others (Figure 4.1). We also



Filter Query: FOREST=(GP) & SPECIES=(ABGR)

Figure 4.2: The EcoDATE interface for the cone production data set opened in a browser window. On the left is the multiple histogram view of a data subset. On the right is the time-series line chart showing the temporal trends of cone production among 14 individual trees.

describe an evaluation of the tool in a working group at the 2012 LTER All Scientists Meeting (`http://asm2012.lternet.edu/`).

4.2 Problem Characterization

Here we characterize the analytical needs of ecologists approaching long-term ecological data. These needs are prerequisites for understanding if and how visual analysis can enable insight and discovery.

4.2.1 Long-term Ecological Research and Data

Our study was structured around the central research questions of the HJA LTER program (`http://andrewsforest.oregonstate.edu/`): (1) how do land use, natural disturbances, and climate affect three key ecosystem properties: carbon and nutrient dynamics, biodiversity, and hydrology? and (2) how do these relationships change over time and space? The focus of this work is not to answer these questions but rather to develop a visual-analysis tool to help ecologists better approach these questions. To demonstrate the utility of the tool and the data exploration process, we selected three long-term data sets that represent the three major ecological components of biodiversity, carbon, and hydrology.

Cone Production Data. Conifer trees commonly dominate the forests in which they occur. Seed production by conifers is not only critical to tree reproduction, but also a vital food resource for many organisms. Since readily-observed cone production is an index of seed production, the history of cone crops gives clues to roles of endogenous (physiological) versus exogenous (climate) factors regulating cone and seed production. For instance, cone production is known to be cyclical as well as responsive to climate and local environmental conditions [41].

In the Cascade Range of Oregon and Washington (USA), ecologists have collected data on cone production of upper-slope conifers at 37 locations across 10 national forests every year over a period of 53 years (from 1959 to 2011) [41, 80]. The data set has been difficult to analyze because it is large (45,704 observations) and contains many sampled trees (934 distinct trees of 9 species), some of which died or could not be found again, and others were added to replace those lost (Table 4.1).

Table 4.1: Structure of the cone production data set [80]. Each record described by the following variables represents a cone count observation of a particular tree sampled at a particular plot in a particular year. Each plot falls within a location which is situated in a national forest.

Variable Name	**Type**	**Description**
SPECIES	nominal	Species code
TREE_NR	nominal	Tree number, unique for plot
FOREST	nominal/spatial	National forest code
LOCATION	nominal/spatial	Location code (within forest)
PLOT	nominal	Unique plot number (within location)
YEAR	ordinal/time-based	Sampling year
CONE_COUNT	quantitative	Number of cones
DBH	quantitative	Diameter at breast height
STATUS	nominal	Status of tree (live, dead, missing)

Stream Chemistry Data. For the past 50 years, small watersheds have been a major setting for ecosystem studies based on long-term records of inputs and outputs [102, 103, 96]. Ecologists have assessed aspects of ecosystem dynamics, such as retention of nutrients and atmospheric pollutants in response to natural and management disturbances of vegetation, growth of vegetation, and chemical inputs to the ecosystem. Stream chemistry sampling and analysis was initiated in two small watersheds within HJA in 1968. Over time, sampling expanded to eight gauged watersheds. Water samples are collected automatically as a function of stage height and flow and composited at stream gauging sites. Analytes include dissolved and particulate nitrogen, phosphorus, carbon, as well as pH, conductivity, suspended sediment, and a full suite of cations and anions (Table 4.2).

Table 4.2: Structure of the stream chemistry data set [79]. Each record represents a monthly stream chemistry property collected and aggregated at a particular location in a particular month of a year.

Variable Name	**Type**	**Description**
SITE_CODE	nominal	Gaging station site code
WATERYEAR	ordinal/time-based	Water year (October-September)
YEAR	ordinal/time-based	Calendar year
MONTH	ordinal/time-based	Month
Q_AREA_MO	quantitative	Total monthly (TM) streamflow
ALK_OUT_MO	quantitative	TM alkalinity outflow as HCO3-C
SSED_OUT_MO	quantitative	TM suspended sediment outflow
SI_OUT_MO	quantitative	TM silica outflow
TDP_OUT_MO	quantitative	TM total dissolved phosphorus outflow
PO4P_OUT_MO	quantitative	TM ortho phosphorus (PO4-P) outflow
TDN_OUT_MO	quantitative	TM total dissolved nitrogen outflow
DON_OUT_MO	quantitative	TM dissolved organic N outflow
NO3N_OUT_MO	quantitative	TM nitrate-nitrogen (NO3-N) outflow
NA_OUT_MO	quantitative	TM sodium outflow
K_OUT_MO	quantitative	TM potassium outflow
CA_OUT_MO	quantitative	TM calcium outflow
MG_OUT_MO	quantitative	TM magnesium outflow
SO4S_OUT_MO	quantitative	TM sulfate-sulfur (SO4-S) outflow
CL_OUT_MO	quantitative	TM chloride outflow
DOC_OUT_MO	quantitative	TM dissolved organic carbon outflow

Forest Structure Data. In a study of long-term forest development, ecologists are studying temporal changes in the structure and composition of unmanaged Douglas-fir (*Pseudotsuga menziesii*) forests [54] that established after a stand-replacing wildfire disturbance. The analysis is based on records collected from 21 permanent plots at eight locations along the Pacific Coast and the Cascade Mountains in western Oregon and Washington. The plots were established between 1910 and 1940, when the forests ranged from 42 to 72 years of age, for the purpose of tracking growth and timber yield of young Douglas-fir forests; in the 1970s forest ecologists began to study forest succession in these plots. Of the 21 plots, 17 are still being measured at regular intervals, providing a data record of up to 100 years on rates of tree growth, trajectories of stand productivity, and the processes and patterns associated with tree mortality, growth, and regeneration. The plots are part of a larger network of long-term plots maintained through the Pacific Northwest Permanent Sample Plot program (PNW-PSP) [1] (Table 4.3).

Table 4.3: Structure of the forest structure data set [54]. Each record represents an observation of trees in terms of basal area, density, and biomass sampled at a particular location in a particular year.

Variable Name	**Type**	**Description**
STANDLOC	nominal	Stand location
STANDID	nominal	Stand identifier
AGE	ordinal	Stand age
SPP	nominal	Species code
ELEV_M	quantitative	Elevation (meters)
L_BAPH	quantitative	Basal area of live trees (m^2/ha)
L_TPH	quantitative	Density of live trees (trees/ha)

In summary, long-term ecological data sets are characterized by their large size (thousands of records) and their complexity in terms of the multiple biotic and abiotic variables (e.g., location, elevation, temperature, and rainfall) of varying types (e.g., quantitative, nominal, and ordinal) that are sampled through time. These characteristics—multivariate, geospatial, and connected through time—make them good candidates for visualization. In this chapter, we focus on observational and experimental data and exclude modeled or real-time ecological data (e.g., continuous stream data from sensors).

4.2.2 Visual Analytical Needs of Ecologists

From the information visualization perspective, each of the three long-term ecological data sets presents a challenging multivariate visualization problem. Employing the user-centered design approach—which we describe in Section 4.5—we have identified the general requirements for a visual-analysis tool targeting ecological long-term data with an emphasis on distributions and temporal trends. Specifically, the tool should enable users to do the following:

Requirement 1 (R1) - Distribution/Diversity Patterns. See and relate distributions of variables simultaneously and iteratively without making assumptions about their shapes. In doing so, the tool should also allow users to repetitively filter data to specific subsets and compare them. In addition, the tool should be able to handle large data sets (thousands of records).

Requirement 2 (R2) - Temporal Trends. See temporal trends of variables and compare these trends iteratively across space and species. For example, for the cone production data set, ecologists are interested in the patterns and relative strengths of synchronicity of cone production across time, space, and species. Therefore, in this example, the tool should enable ecologists to isolate time-series for different sets of trees of interest and to use an appropriate chart that supports time-oriented data to compare these series.

Requirement 3 (R3) - Collaboration. Keep track of findings at any stage of visualization, share findings with other users, and invite others to build on or modify the visualizations. Scientists and educators may also use the tool to teach students about data exploration in general, and their exploratory process in particular.

Requirement 4 (R4) - Usability. Learn to use the tool quickly and easily. From our experience, users of the tool may have varying levels of comfort with computer applications. Therefore, the tool should be simple and easy to use.

4.3 Existing Visualization Solutions

Design of the EcoDATE tool was informed by related work on visual representation techniques and visual-analysis tools, including those currently employed by ecologists. In this section, we assess their applicability to exploring long-term ecological data, with regards to the four design requirements (R1 - R4).

4.3.1 Visual Representations for Ecologists

A visual representation or chart type determines how data are represented or visualized. Along with interaction features, visual representation techniques serve as the primary components in visual analysis tools that we assess here. Ecologists typically employ standard 2D/3D displays as classified by Keim [83]. Examples include histograms, boxplots, and scatter plots. They effectively support tasks such as inspecting distributions, outliers, clusters, and correlations over one or two variables [136] (support of R1). Ecologists also use rank/abundance curves or Whittaker plots [158] to visualize species abundance and diversity (support of R1). Ecologists commonly represent time series data as a line chart in which time is presented as a linear, ordered x-axis and data cases are plotted by their time values [3] (support of R2). The EcoDATE tool incorporates existing standard displays commonly used by ecologists, such as multiple histograms and time-series line charts, into a simple interface and augments them with appropriate interaction features.

4.3.2 Visual Analysis Tools for Ecologists

A visual analysis tool facilitates data analysis with visual representations and interactive features. To the best of our knowledge, little work has been done to develop visual analysis tools specifically for analysis of distributions and temporal trends in long-term ecological data. Here we discuss the merits of four types of tools used by ecologists that contain visual analysis components: (1) widely used software packages such as spreadsheet programs and statistical software packages; (2) specific tools for particular calculations (e.g., estimates of species diversity, calculation of primary productivity); (3) data repositories or portals; and (4) workflow management systems (e.g., Kepler). O'Donoghue et al. [111] provides an overview on visualization of biological data.

Ecologists often use charting components in spreadsheets and statistical software packages for visual analysis prior to statistical analyses; these tools permit quick and simple visual inspection and they are easy to learn (support of R4). However, these tools lack interactive capacities. For instance, they do not readily permit iterative subsetting and replotting of data, which are essential steps in hypothesis formulation [8] (lack of interactivity for R1 and R2).

A second group of tools includes software designed for specific types of ecological data analysis, such as estimation of species diversity and abundance [32] or simulation of hydrologic models with input data [130]. These tools provide rigorous statistical tests and modeling techniques to answer specific scientific questions—for example, what is the species richness of dataset A?, or what data should be used to define parameters for hydrologic model B? However, these tools do not support exploration of distribution patterns and temporal trends with interactive charts (lack of R1 and R2). Therefore, we do not consider these tools further.

A third type of analysis tool is ecological data repositories or portals that support collection, archival, and synthesis of long-term data from multiple sites, for example, EcoTrends [137, 118] and Clim-DB/Hydro-DB [66]. These web-based portals are usually equipped with static visual representations such as line charts for simple and quick visual exploration of temporal trends in existing long-term data sets (partial support of R2). Although these tools may have limited capacity for subsetting, they are not designed to support distribution patterns in multiple attributes (lack of R1), interaction features (lack of interactivity for R1 and R2), or collaboration features (lack of R3).

A fourth class of software tools for visual analysis is designed to support "workflows," i.e., the analysis process of scientists (support of R3). Representative tools include Kepler [97] and VisTrails [20]. Although these tools are powerful and potentially useful to ecologists, they require customization and programming to fit the specific analytical needs of ecologists, especially with respect to visual representations and interaction features (lack of R4). Therefore, these tools may be more suitable for information managers who have expertise in managing data in repositories and who help ecologists with data pre-processing tasks such as data gathering and cleansing.

4.3.3 General Visualization Tools

In addition to tools developed by and for ecologists, a wide range of information visualization tools is available that, to some extent, meet the design requirements for ecologists [131, 61, 65]. For instance, software systems such as Tableau (`http://www.tableausoftware.com/`) and Spotfire (`http://spotfire.tibco.com/`) are dedicated visual analysis tools, as distinguished from charting components in spreadsheet or statistical tools. They provide pre-defined chart types and a variety of controls for interacting with data, for example, to subset data (support of R4). They also support multiple, coordinated views; and users can publish and share visualization dashboards as interactive Web pages (support of R3). However, these applications are not necessarily tailored to specific analytical needs of ecologists (lack of R1 and R2). For example, ecologists may want to discretize quantitative variables interactively to reveal different distribution features of the data (R1). Also, ecologists may want to repeatedly generate subsets of time-series data and plot them in a line chart in order to examine temporal trends (R2).

4.4 The EcoDATE Tool

A visual analysis tool consists of (1) representations (i.e., charts, graphs) and (2) interaction features (i.e., subsetting, bookmarking, etc.). The various types of interaction features can be described using a classification system for visual analysis tasks proposed by Heer and Shneiderman [65]. The classification consists of three high-level categories of task types: a user makes a set of decisions about types of charts and organization of data (*data view and specification*), how to manipulate the visualization views (*view manipulation*), and how to reproduce and share the visualizations (*process and provenance*). The representations and interaction features of EcoDATE are outlined following this classification system (Table 4.4) and discussed based on the four design requirements presented earlier (R1 - R4).

The EcoDATE interface (Figure 4.2) supports multiple views (or windows) each of which can be manipulated (select, drag and drop, resize, and close). While the interface is web-based, its look and feel is similar to a desktop interface that is familiar to users (support of R4).

4.4.1 Chart Types

The current version of EcoDATE (ver. 1.0) supports two widely used chart types: multiple histograms and a line chart. Coordination between views of these chart types loosely follows the master/slave relationship [131], in which the master views of multiple histograms are used to query/retrieve data and to generate line series for line charts. Other than that, views are independent from each other.

Table 4.4: Interaction techniques supported by the EcoDATE tool. Each of the techniques is designed to facilitate specific analytical needs of ecologists. Most of the techniques (if not explicitly noted) are applied to the multiple histogram views. The classification is adapted from Heer and Shneiderman [65].

High-level Category	**Task Type**	**EcoDATE's features**	**Specific analytical needs of target users**
Data and View Specification	Visualize	Choose among multiple histograms and line charts	Inspect distributions of variables with multiple histograms and temporal trends with time-series line charts
	Filter	Filter data based on selection of bins	Examine different data subsets or samples of observations
	Sort and Reorder	Sort bins within a variable by names or by abundances	Organize the data according to a familiar unit of analysis (e.g., rank species from rare to common)
		Reorder variable axes	Group axes by their common or user-defined characteristics (e.g., group of covariate/response variables)
	Derive	Discretize quantitative variables	Experiment with different discretization settings (e.g., isolate specific range of interest) to reveal different features of the data
		Group/ungroup bins within an variable	Group outliers or similar variable values to fit users' hypotheses (e.g., group species of the same genus or family)
		Scale (normalize) bins' abundances	Accommodate data sets with different distributions
View Manipulation	Select/ Highlight	Select or highlight a view, axes, bins, or line series	Select or highlight elements of interest for other operations, such as filter, sort, derive
	Navigate	Navigate and control views using the top menu bar and the bottom status bar	Know where and how to navigate views
	Coordinate	Duplicate multiple histogram views	Compare data subsets side-by-side
		Use multiple histograms as a query builder to construct series data for line charts	Construct multiple line series and compare them
	Organize	Open, close, resize, and layout views	Manage views for comparison or effective presentation
		Show/hide error bars in line charts	Access additional information on demand
Process and Provenance	Record	Log user interactions	Undo/redo actions, reproduce states step-by-step. These features are reserved for future work
	Annotate	Color axes and Label line series	Distinguish among axes or line series based on their common or user-defined characteristics
	Share	Bookmark visualization states	Revisit/share visualization states with others for collaborative and iterative exploration of data
		Export view data	Pursue further analysis with existing statistical tools.
	Guide	Display data tips for menu bars, axes, bins, and line series	Guide users through menu items and provide additional information on highlighted items

Multiple Histograms. The purpose of this representation is to show distributions of multiple variables, which permit the user to identify and interactively specify subsets of data (support of R1). Like previous work, this multiple histogram representation presents variables in a parallel axis layout [59, 120]. Histograms are placed vertically side-by-side, one histogram for each variable, as opposed to horizontally. In these views, the bars extend to the right (in contrast to the familiar upward-extending display). A vertical arrangement of histograms allows more variables to fit in wide-screen displays and facilitates the placement and reading of labels from left to right, as shown by an example of a subset of the cone production data set (Figure 4.2, left view). The ecologist user can duplicate multiple histogram views to compare data subsets side-by-side. Continuous numerical variables are discretized into bins to plot relative frequency. The length of each bar is scaled according to $l(x) = |x|/|x_{MAX}|$ where $|x|$ denotes the number of observations in bin x, and x_{MAX} is the bin with the most observations for the variable in question.

Line Chart. The purpose of this representation is (1) to show overall trends in a continuous, real-valued variable, such as cone production or tree density, over the sampling period of interest; and (2) to support comparison of values of the variable at different time points or intervals (support of R2) and across multiple samples. In line charts, ordinal variables such as time are presented as a linear ordered axis, X-axis, and values at each point in time are plotted along the Y-axis. For example, Figure 4.2, right view depicts multiple line series of average cone count over time for multiple sets of trees in the cone production data set. Optionally, users can display error bars as standard errors or standard deviations on the line series (Figure 4.3).

4.4.2 Interaction Features

The EcoDATE tool supports a wide range of interaction features (Table 4.4). We extend our description of a subset of prominent features here, emphasizing its utility in the context of the distributions and trends analysis.

Subset/Filter. Given an overview of data distribution in multiple histogram views, ecologists often want to shift their focus repetitively among different subsets or samples of observations, for example, to examine distributions of species at different locations. Ecologists also want to generate subsets of data for other representations, such as a

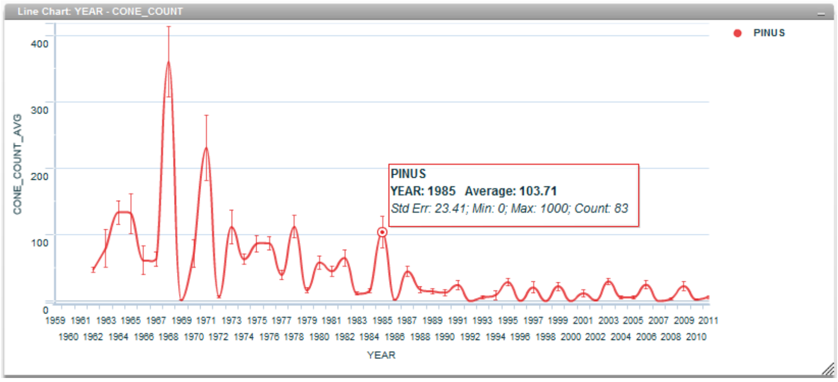
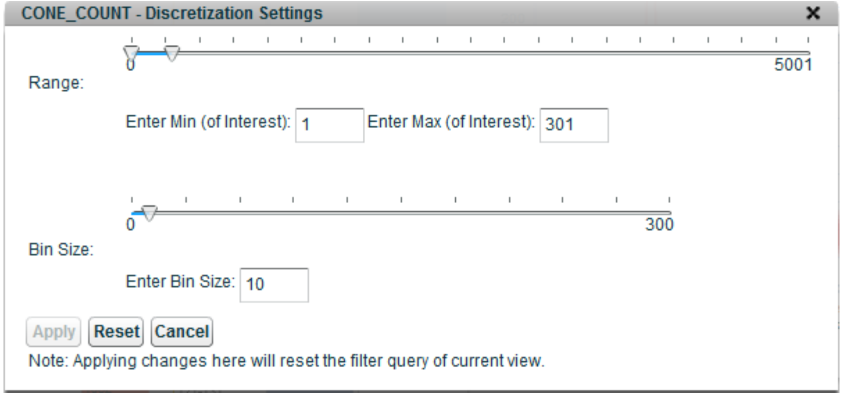


Figure 4.3: Line series of average cone production from *Pinus* species (pine trees) from 1962-2011 showing a declining trend. Users can select to display error bars as standard errors or standard deviations on the line series. Users can place the mouse pointer over the data points on the line series for additional information

line chart. Subsetting or filtering operates on selected bins. A filter 'status' bar at the bottom will show the filter query for the currently selected view (see Figure 4.2). To construct a complex filtering query consisting of multiple bins, we follow a simple and commonly used rule articulated by ecologists: bins within a variable are connected by the "OR" condition, whereas groups of filtered bins across variables are connected by the "AND" condition. For example, the left view of Figure 4.2 visualizes a subset of observations filtered by *Abies grandis* trees (grand fir) AND sampled at Peterson Prairie in the Gifford Pinchot National Forest, Washington, USA (see the bottom bar for the query). Users can also inverse (or exclude) the query to obtain the complement of a subset. To some extent, multiple histogram views can be used to quickly and visually construct a query (as opposed to typing a query command) (support of R1).

Sort/Reorder. Users can sort bins within a variable by names or by abundances (support of R1). The goal is to organize the data according to a familiar unit of analysis, for example, species ranked from rare to common. They can also reorder variable axes according to common or user-defined characteristics of variables. For example, they might group a set of covariate/response variables or create groups of nominal (e.g., species, habitat), ordinal (e.g., sampling month, year), or quantitative (e.g., cone count, basal area) variables.

Derive. In many cases, to examine different data distribution settings, ecologists wish to generate derived data such as discretized quantitative variables or groups of bins. While they can do so prior to importing data for visual analysis, moving between tools disrupts the flow of the iterative exploration process [36]. Using EcoDATE, ecologists can discretize quantitative variables, based on their knowledge of ecology, without leaving the application (support of R1). EcoDATE allows ecologists to flexibly experiment with discretization settings by specifying the range of interest and bin size (see Figure 4.4). In addition, for categorical variables, similar to discretization, ecologists can group or ungroup bins within a variable based on their hypotheses (support of R1). For example, they can group species based on their rarity or their functional groups. For example, ecologists exploring the forest structure data set may wish to select and group species such as western hemlock, western redcedar, Pacific yew, and bigleaf maple into a group of shade-tolerant species before comparing it to Douglas-fir.


CONE_COUNT - Discretization Settings

Range:

Enter Min (of Interest): 1 Enter Max (of Interest): 301

Bin Size:

Enter Bin Size: 10

Note: Applying changes here will reset the filter query of current view.

Figure 4.4: Discretization settings for variable CONE_COUNT. Ecologists can narrow the range of interest for this variable to [1, 301] and specify a bin size of 10. The result will automatically include two separate out-of-range bins for [0, 1) and [301, 5001) as shown in Figure 4.2 (left view, the CONE_COUNT histogram).

Share. Collaborators are often geographically dispersed with the physical distance and time differences making collaborative exploration difficult. Using the EcoDATE tool, ecologists can discuss and share findings with collaborators by bookmarking visualization states (e.g., Figure 4.2) as unique web URLs (support of R3). These bookmarks can be easily shared via email or embedded to the user's notes, serving as a common ground for discussions among collaborators.

The implementation of bookmarking in EcoDATE stores “snapshots” of visualization states (e.g., Figure 4.2) including aggregated static data (for bins and line series) as opposed to providing dynamic access to the most current data [64]. This implementation decision is based on the understanding of characteristics of large ecological data sets. The data are usually static and the analysis process involves inspecting distributions or trends of observations as aggregation of data as opposed to individual data points. Because EcoDATE stores only aggregated data, the storage cost is efficient and the state loading time is fast.

4.5 Design and Implementation

4.5.1 User-centered Design with Ecologists

A close collaboration between ecologists and visualization researchers was critical for design and integration of the EcoDATE tool into the ecologists' analysis process. We employed the user-centered and participatory design approach [143, 132] in which the ecologists were included as part of the design team. User-centered design is both a philosophy and a process in which the needs, desires, and limitations of the target users (e.g., scientists) are considered very closely at every stage of the design process (establishing requirements, design, implementation, evaluation). The process has involved three ecologists and two visualization researchers, who are co-authors of the work presented in this chapter.

Our participatory design process was iterative, required group design sessions over many weeks, and involved a variety of tools for assessment of user performance and tool usability such as observations, interviews, log books, and automated logging of user interactions [143]. In addition, the visualization researchers engaged with the ecologists to the point of becoming assistants in the process of data exploration. We used email communications to share and discuss visualization state bookmarks. We set up weekly one-hour meetings between a researcher and an ecologist in the ecologist's workplace for several months. Activities during these meetings varied. In early meetings, the visualization researcher learned about the data set and initial hypotheses of the ecologist and observed her using an Excel spreadsheet to create time-series line charts for data subsets. We followed up with a discussion of the ecologists' difficulties with Excel's charting component. In subsequent meetings, after implementing line charts in EcoDATE, we gave tutorials on how to use the chart, observed the ecologist exploring her data using the new representations, and discussed hypotheses and insights. In other meetings, the researcher observed the ecologist as she performed hypothesis testing using a statistical tool and discussed how she preferred to export data sets. In addition, we also discussed entries in the log book [143]. The ecologists were given a log book so they could record notes from using both the EcoDATE and their existing tools. They were encouraged to record not only successes but also any difficulties or frustrations. Because the log book was for both the EcoDATE and their existing tools, we refrained from implementing an online log book feature within the EcoDATE tool.

Finally, during the development process, we also collaborated with information managers, who manage the ecological data repository of the HJA LTER site. They helped clean data, explained the structure of data sets, and gave feedback on EcoDATE.

4.5.2 Implementation

The EcoDATE tool is a web-based database application implemented following the client-server architecture. In this section, we describe the client and server components of the tool and justify our choice of the architecture.

Client. The client side of EcoDATE is responsible for representing processed data from the server—that is, representing multiple histogram views and line charts, laying out views and menus, and communicating user interactions with the server. We developed the EcoDATE client interface with Flex 3 and the Degrafa graphics framework. Flex 3 is an open-source framework by Adobe for creating Flash rich internet applications. Degrafa is an open-source graphics framework that facilitates the process of creating pre-composed graphics in Flex 3. In particular, Degrafa helps create lightweight geometry building blocks such as rectangular bins and variable axes in the EcoDATE tool. We used Action Message Format (AMF), a binary format by Adobe, to serialize data and send messages between the client and server (remote service). Because Flash is web-based, no installation of the tool is required and it can be potentially available on any browser or device that supports Flash.

Server. Data sets are stored and managed with the MySQL database management system (DBMS). In addition, we rely on the programming languages of PHP (Hypertext

Preprocessor) and SQL (Structured Query Language) to handle requests from the client. Specifically, the server is responsible for all data-related logic and computation, such as retrieving and manipulating ecological data, building and maintaining data structures of visualization states, and logging interactions. Note that by design of the tool, the multiple histogram views and line charts require only aggregated data. By pushing logic such as data aggregation to the server side, we leverage computation and the database management capabilities of the server while keeping the workload on the client low. For example, we set up data indexes for all fields of interest in the data sets and we take advantage of caching in MySQL. This client server model was a natural choice considering that most of the ecological data repositories are structured and stored in a DBMS [67].

Metadata are another distinctive property of scientific data in general, and ecological data sets in particular. While generated to aid analysis, metadata present another challenge to data visualization. Specifically, the key variables described in Tables 4.1, 4.2, and 4.3 were supplemented with additional information about the variable such as descriptions of SPECIES or LOCATION. Technically, the metadata tables needs to be joined with the primary data table to form the data set for use in EcoDATE.

Our implementation approach scales well to large data sets. Feedback on performance from ecologists indicates that it is highly responsive for all three data sets of interest on a typical desktop PC. From our tests, heavy interactions such as filtering usually respond in a few seconds provided a high-speed internet connection.

4.6 Evaluation

One of the most effective ways of evaluating an information visualization tool is through long-term case studies of target users exploring real world data sets using the tool [143]. In this section, we evaluate EcoDATE by three case studies, one for each of the three data sets: cone production, stream chemistry, and forest structure. Further, we discuss the results from the evaluation of the tool during a working group meeting at the LTER All Scientists Meeting in 2012.

The objectives of the case studies are (1) to demonstrate the utility of EcoDATE for ecologists and (2) to describe how use of the tool reveals how scientists analyze data, both individually and collaboratively, and provides scientists with hypotheses that can be tested outside the tool (Figure 4.1). Each of the case studies involved multiple observations of ecologists in multiple work sessions in normal working environments (i.e., offices) during which they used the EcoDATE tool to explore the three data sets.

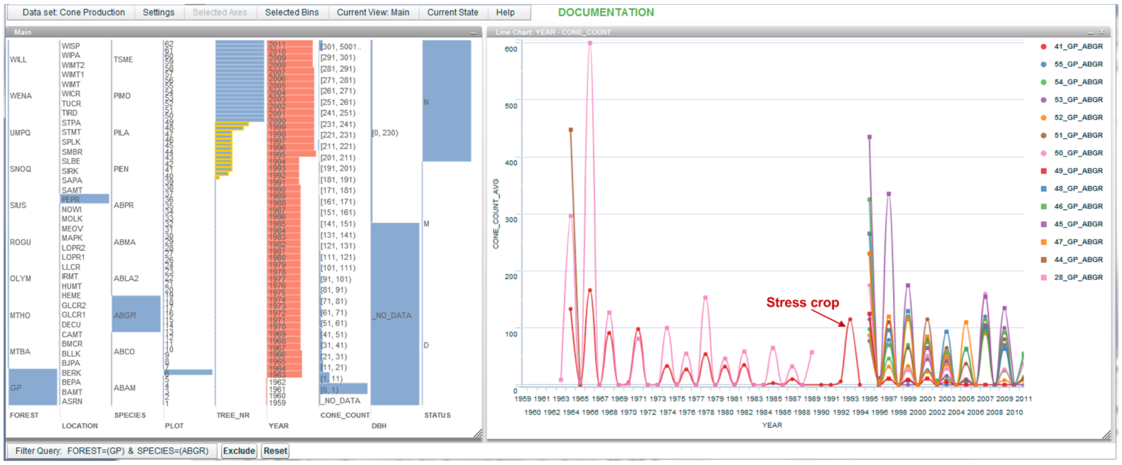
4.6.1 Cone Production Data Case Study

The primary objective of this case study is to demonstrate the utility of EcoDATE in terms of its supported visual representation and interaction techniques. The design of EcoDATE followed closely the Visual Information Seeking Mantra, the widely accepted visual design guideline introduced by Shneiderman [141]: "overview first, zoom and filter, then details on demand". This mantra suggests that when the user seeks information from a data set, a tool should allow the user to start first with an overview of the entire data set, then to subset the data (filtering and zooming), and ultimately to get additional fine details as needed.

Summary of information needs. According to the design requirements, the ecologist user was interested in two key aspects of the cone production data set. First, she wanted to see the overall distribution of samples in time and space (geographic and environmental) and to be able to relate multiple distributions simultaneously and iteratively. Second, she was interested in the patterns and relative strengths of synchronicity of cone production variation across time, space, and species.

Overview. The initial multiple histogram view helped the ecologist quickly assess the numbers of sampled trees by species and their distributions across locations and years. She also detected that the range for CONE_COUNT (number of cones per tree) was large (0-5000) and its distribution was positively skewed with very few high values. To examine the number of trees that produced no cones (observations with zero cone count), she was able to use the discretization settings (Figure 4.4) to derive (Table 4.4) new bins that displayed the numbers of trees with zero cones (Figure 4.5, left view, the CONE_COUNT axis).

Filtering/Subsetting. After inspecting the overview, the ecologist focused on a specific location, in this case, the Gifford Pinchot National Forest (GP), Washington, USA. This forest was of interest because of its complex topography and proximity to Mount St. Helens, whose 1980 eruption may have affected cone production history. First, she filtered the data by bin 'GP' in the 'FOREST' variable. While she could select and filter multiple bins at once, she preferred first to inspect the distribution of all cone production



Filter Query: FOREST=(GP) & SPECIES=(ABGR)

Figure 4.5: The EcoDATE interface for the cone production data set opened in a browser window. On the left is the multiple histogram view of observations of *Abies grandis* trees (grand fir) sampled at Peterson Prairie in the Gifford Pinchot National Forest, Washington, USA. Ecologists used this view (1) to inspect distribution of this sample with respect to the variables of interest and (2) to generate multiple line series of average cone count over time for multiple sets of trees. The time-series line chart (right view) shows the high degree of synchrony of cone production among 14 individuals of *Abies grandis* (grand fir). It suggests that cone production of *Abies grandis* occurs on a biennial cycle but skipped several years, for example, 1969-1970 and 1972-1973, perhaps due to climate control. Tree 41 (red line) shows very little cone production from 1973-1992, and then a stress crop in 1993, just before the tree died.

observations (i.e., all species) in the GP. Then she filtered the data to examine cone production for *Abies grandis* (grand fir) (ABGR) only. *Abies grandis* was of interest because it is a common species in mixed conifer forest communities. The view of the new subset helped the ecologist discover that the sampling process was not consistent over time: trees were sampled starting in 1963, but because of gradual, cumulative mortality, the sample size declined over time, so new trees were added in 1995 (Figure 4.5, left view, the YEAR axis). She was then able to further filter the data to examine cone production in individual trees with long-term cone production records, as well as to examine mortality at tree, plot, species, and regional scales.

Details on demand. While inspecting the distribution of the subset of interest (cone production in *Abies grandis* at the GP), the ecologist wanted to compare trends of cone production between trees that died and those that were added to replace them. Tree status (health) during the study period was important because tree health, morbidity, and mortality affect cone production. For example, stand-level cone production may depend on tree-level processes, including stress crops from dying individuals, insect attacks, and partial wind damage. Using EcoDATE, she was able to identify and plot the trees that were sampled for subsets of the record, which produced a visualization of cone production in trees that died and trees that were added to replace them. Specifically, to identify trees that were not sampled throughout the entire study period, the ecologist sorted trees (i.e., variable TREE_NR) by the numbers of years of observation. After sorting, she selected trees that had less than a certain number of observations (Figure 4.5, left view, the TREE_NR axis) and added data for each of the selected trees as a line series into the YEAR-CONE_COUNT line chart (Figure 4.5, right view).

The time-series line chart (Figure 4.5, right view) helped ecologists quickly formulate hypotheses that the trees of interest produced cones in synchrony (timing and magnitude) on a biennial cycle, but skipped several years, for example, 1969-1970 and 1972-1973, suggesting the hypothesis that some external factor or event may have disrupted the biennial cycle. The view also shows multiple trees that were added to the plot in 1995. In addition, the visualization allowed a discovery that trees that died sometimes produced "stress crops" just before dying: tree 41 (red line) shows very little cone production from 1973-1992, and then a stress crop in 1993, just before the tree died. In all, while the ecologist found no evidence of an effect of the 1980 eruption of Mount St. Helens on cone production history, her exploration led to other interesting scientific discoveries that she did not anticipate before using the visualization.

Note that while this sequence of actions demonstrates a single exploration path, the tool supports pursuit of multiple paths simultaneously and iteratively. For example, the ecologist repeated the process and retrieved the data subset for all 'PINUS' or pine trees (Figure 4.3). The time-series line chart for this subset revealed a declining trend of average cone production of *Pinus* spp. from 1962-2011, suggesting the hypothesis that tree aging, mortality, or expansion of influence of a pest/pathogen may be contributing to declining cone production.

Sharing and further analyses. Satisfied with her findings, the ecologist bookmarked the current state of the visualization (as shown in Figure 4.5) as a URL and emailed the link to her collaborators with a description of her findings. She also exported the data subsets and pursued further analysis using existing statistical tools (e.g., Pearson's correlation test to quantify the correlation between multiple line series with respect to sampling years).

Case Study Summary. Using EcoDATE, the ecologist became acquainted with the cone production data and the tool and developed a concrete analysis plan, which, to her, had been vague or possibly subconscious before. Specifically, EcoDATE provided a holistic overview of the observations interest and helped the ecologist build a mental model of how multiple variables were distributed in the entire data set. This model helped the ecologist to formulate actions such as filter/subset queries, and explore the data broadly and deeply.

4.6.2 Stream Chemistry Data Case Study

The objective of the next case study is to illustrate the process of using EcoDATE to gain insights into data and to generate hypotheses. Following her experience with EcoDATE and the analysis of the cone production data, the ecologist was more aware of the exploration paths that she would take. From our observations, the ecologist followed a hypothesis generation process that can be summarized as three main tasks (Figure 4.6): (1) *specify visualization views* (e.g., filter, sort, reorder, derive data), (2) *characterize views* (e.g., distribution patterns and temporal trends), and (3) *gain insights and generate hypotheses*. The process is highly iterative with multiple rounds of exploration, guided by discoveries in each round and the ecological knowledge of the user.

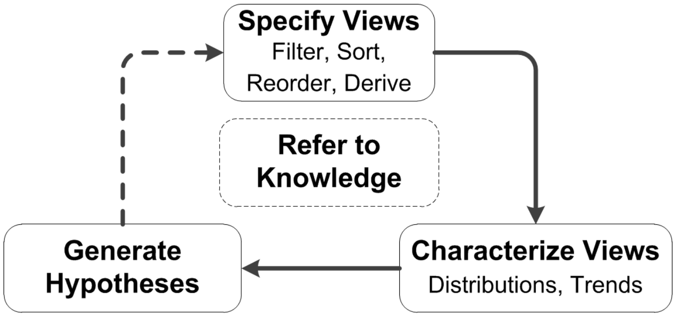


Figure 4.6: The iterative process of hypothesis generation supported by EcoDATE. Each rectangle represents a task and each arrow represents a transition from one task to another. The task of involving knowledge in the center supplements all other tasks.

Summary of information needs. Exploring the stream chemistry data set, the ecologist wanted to investigate distribution and temporal patterns of multiple chemical properties within and across locations (e.g., watersheds) over time, and ultimately to make inferences about ecological processes and events driving these patterns. Before the work sessions, she had examined temporal patterns of stream chemistry using statistical tools. Despite this prior knowledge, the multiple histogram views of the data facilitated by EcoDATE helped her generate additional hypotheses based on the shapes of distributions of different chemical constituents. The visualization of the stream chemistry data is available at `http://purl.oclc.org/ecodate/chemistry`.

Round 1 of hypothesis generation. Starting with the default specification of the multiple histogram view (*specify*), the ecologist quickly noticed (*characterize*) (1) differences in numbers of samples by year and location, (2) differences in the shapes of distributions of the chemical properties over the years and from one property to another, and (3) a relatively large number of extreme values. Using her knowledge, the ecologist related the difference in record lengths (characterization 1) to the hypothesis that sampling must have been turned off and on intentionally at some watersheds (*generate hypotheses*). To confirm this hypothesis, she planned to access the sampling logs for more information. Further, the characterizations (2) and (3) prompted the ecologist to pursue these paths further, as described next.

Round 2 of hypothesis generation. To compare the shapes of distributions of the chemical properties, the ecologist first used the discretization feature (Figure 4.4) to *specify* equal numbers of bins as well as equal numbers of observations in the extreme value bin (upper range bin) for each of the histograms of the corresponding chemical properties. She then found that distributions varied among properties in the degree of skew (*characterize*) (Figure 4.7). Specifically, the distributions for silica (SI) and discharge (Q_AREA_MO) were similar to one another and differed from the distributions for nitrate-nitrogen (NO3-N) and suspended sediment (SSED).

From this characterization of the data, the ecologist *referred to her knowledge* and *formulated several hypotheses*, e.g., (1) extreme suspended sediment output and nitrate-nitrogen output may occur under extreme storm events, when sediment and decomposed litter are entrained; (2) silica output is more dominated by chronic export, which is consistent with its origin from chemical weathering.

Additional rounds of hypothesis generation. Following up on the hypotheses gen-

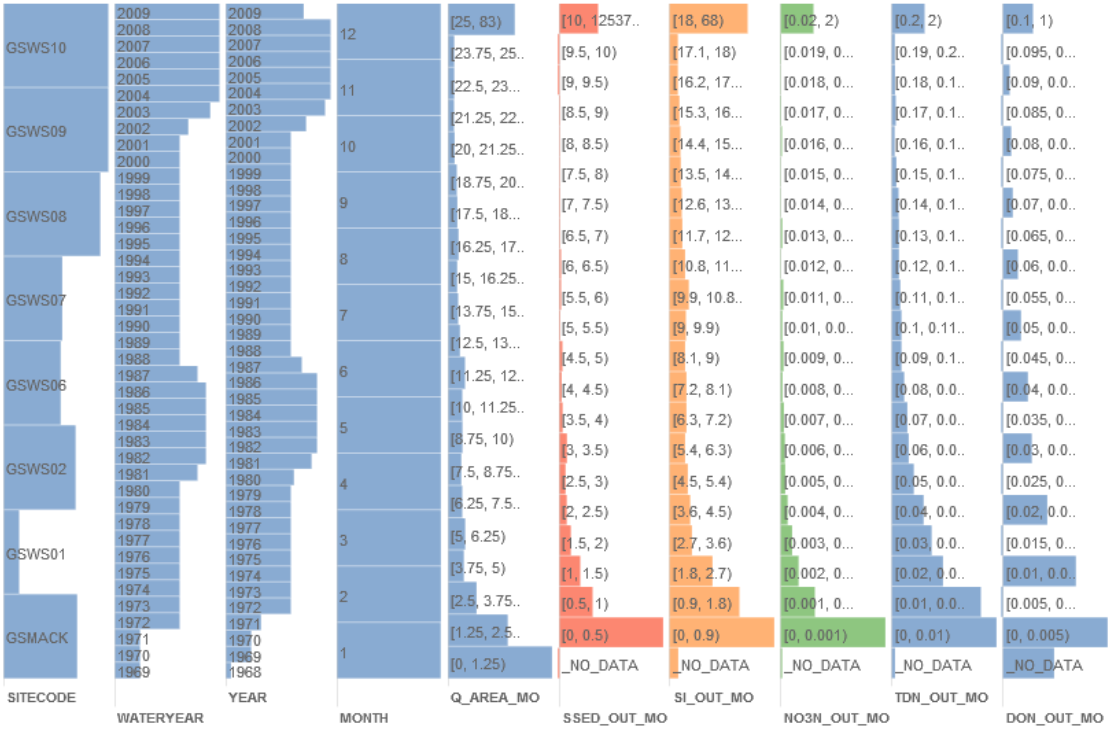


Figure 4.7: Multiple histogram view of observations in the stream chemistry data set. In this case, the ecologist was interested in the distribution patterns of total monthly streamflow (Q_AREA_MO, blue axis), total monthly suspended sediment outflow (SSED_OUT_MO, red axis), total monthly silica outflow (SI_OUT_MO, orange axis), and total monthly nitrate-nitrogen outflow (NO3N_OUT_MO, green axis).

erated in Round 2, the ecologist rapidly completed additional rounds of exploration. She specified the time-series line charts for the chemical properties of interest to investigate how the extreme values of the properties coincided over time. She subsetted the data to two specific locations (watersheds) and cross-compared their temporal trends of specific chemical properties. After each of the exploration rounds, the ecologist was able to bookmark the visualization state, take snapshots of the visualization views, and save them along with her notes. In summary, within four one-hour work sessions, the ecologist completed ten rounds of data exploration, generating hypotheses that could be statistically confirmed quickly as well as questions that prompted further analyses (inside or outside of the EcoDATE tool).

Case Study Summary. Even though the ecologist had prior knowledge of the stream chemistry data, EcoDATE nevertheless permitted in-depth analysis of the data that led to new insights, especially with respect to specification and characterization of multivariate distribution shapes using the interaction feature of discretization of bins. Although existing analysis tools such as spreadsheet programs also permit this kind of specification, the process would be cumbersome and time-consuming. We summarized the analysis strategy in this case study as an iterative three-step process of specifying visualization views, characterizing views, and gaining insights while incorporating ecological knowledge and intuition (Figure 4.6).

4.6.3 Forest Structure Data Case Study

While the stream chemistry case study aims to emphasize the hypothesis generation process supported by EcoDATE, this case study highlights how EcoDATE helped another ecologist prepare data to upload into EcoDATE, construct the line charts, and gain insights into the forest structure data set.

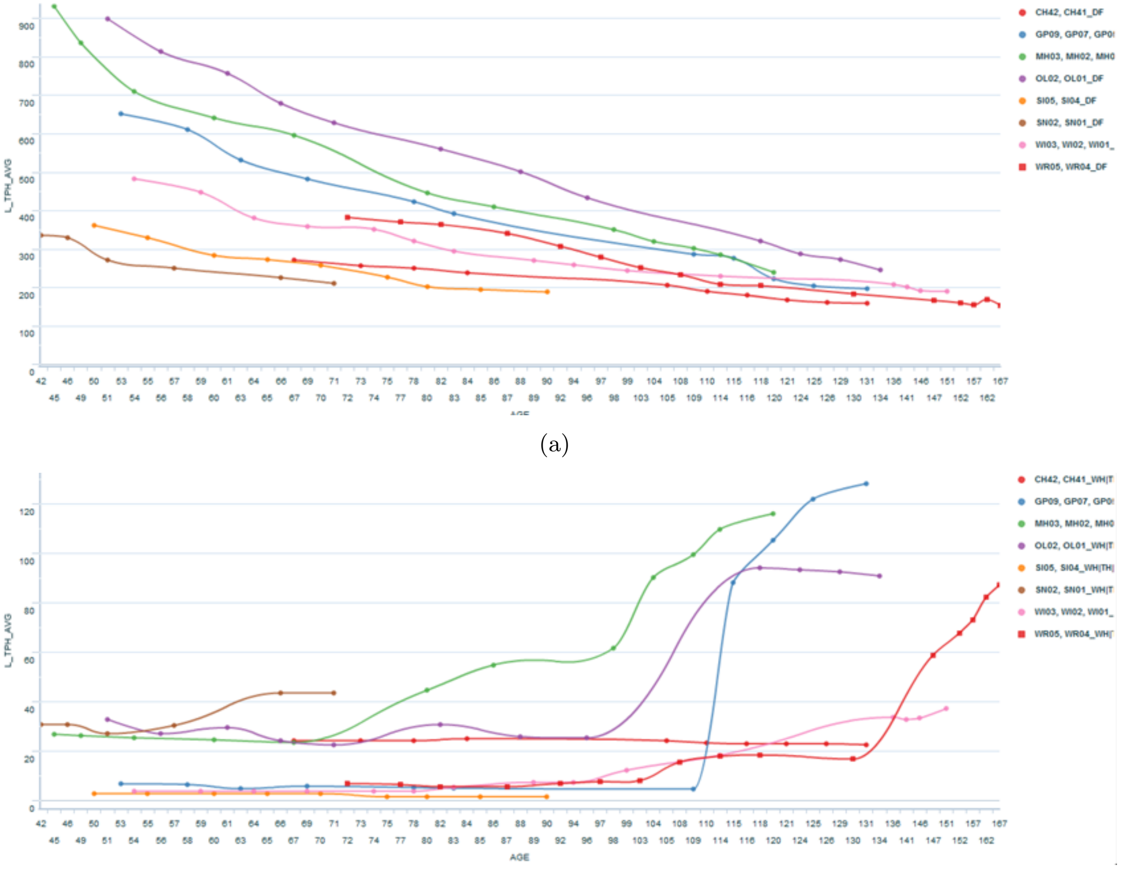
Summary of information needs. The ecologist user exploring the forest structure data was interested in temporal changes in species composition as Douglas-fir forests of the Pacific Northwest transitioned from early to mid-succession stages of development. Of particular interest were trends in density and basal area of shade-tolerant species such as western hemlock (*Tsuga heterophylla*), western redcedar (*Thuja plicata*), Pacific yew (*Taxus brevifolia*), and bigleaf maple (*Acer macrophyllum*) in relation to the dominant Douglas-fir trees across the eight study locations. Therefore, the time-series line chart played an important role for this data set. Nevertheless, the ecologist also benefited from the multiple histogram views of the data when preparing line charts.

Preparing the data. EcoDATE facilitated the preparation of the forest structure data in two ways. First, the tool allowed the ecologist to load and visualize the data quickly in three straightforward steps: (1) upload data (e.g., comma-separated values file format), (2) configure data structure (e.g., specify types for each of the variables of interest), (3) optionally, add additional metadata for each of the categorical variables (e.g., species common names to supplement species codes) (see the EcoDATE tutorials at `http://purl.oclc.org/ecodate/tutorials/`). In this case it was important for the user to be able to upload multiple successive versions of data to EcoDATE because data have been collected from many sites over many years and discoveries from the visualization may prompt the ecologist to re-consider, synthesize, and re-upload data. For example, the initial exploration only considered Douglas-fir and western hemlock. Subsequently, the ecologist expanded the data to include other shade-tolerant species and re-uploaded the data.

Second, in addition to discretization of quantitative variables, EcoDATE supports grouping of categories in nominal and ordinal data variables using the multiple histogram views of the data (Table 4.4). The ecologist grouped shade-tolerant species into a single functional group for comparison to Douglas-fir. The grouping process was exploratory in the sense that the ecologist was able to experiment iteratively with different combinations of species based on his ecological knowledge.

Constructing the line charts. EcoDATE supports creation of line charts for any ordinal variable (e.g., age, year) on the x-axis and any quantitative variable (e.g., tree density, basal area) on the y-axis. Note that based on the configuration of the data structure, EcoDATE can detect the temporal variables at different resolutions and derive new temporal variables based on their combinations (e.g., YEAR and MONTH variables combined creates YEAR-MONTH). For the forest structure data, the ecologist favored AGE over YEAR as the x-axis, which facilitated comparisons of successional trends across the eight study locations, where each location was an average of 2-5 plots (Figures 4.8 and 4.9). The ecologist followed the same process of constructing line series for each of the data subsets of interest as described in the cone production data case study.

Gaining insights. Findings from the visualization underscore the importance of long-term data in tracking the response of forests and other ecosystems to disturbance agents and changes in the environment. The view in Figure 4.8a helped the ecologist quickly assess the declining and converging trends in mean density of Douglas-fir across locations. Although this trend was not unexpected given knowledge of stand development [113, 42], the finding was interesting given the three-fold range in density (about 250 to over 800 trees/ha) when the stands were about 55 years of age. Equally interesting was the variability in the timing of increases in the mean density of shade-tolerant species (Figure 4.8b). Figure 4.9a displays recent declines in Douglas-fir basal area at several locations (GP, MH, OL, WI, WR). This prompted the ecologist to revisit the raw data on mortality assessments of individual trees at these locations. At two of the locations, GP and WR, the mortality data indicated that Douglas-fir bark beetles (*Dendroctonus pseudotsugae*)



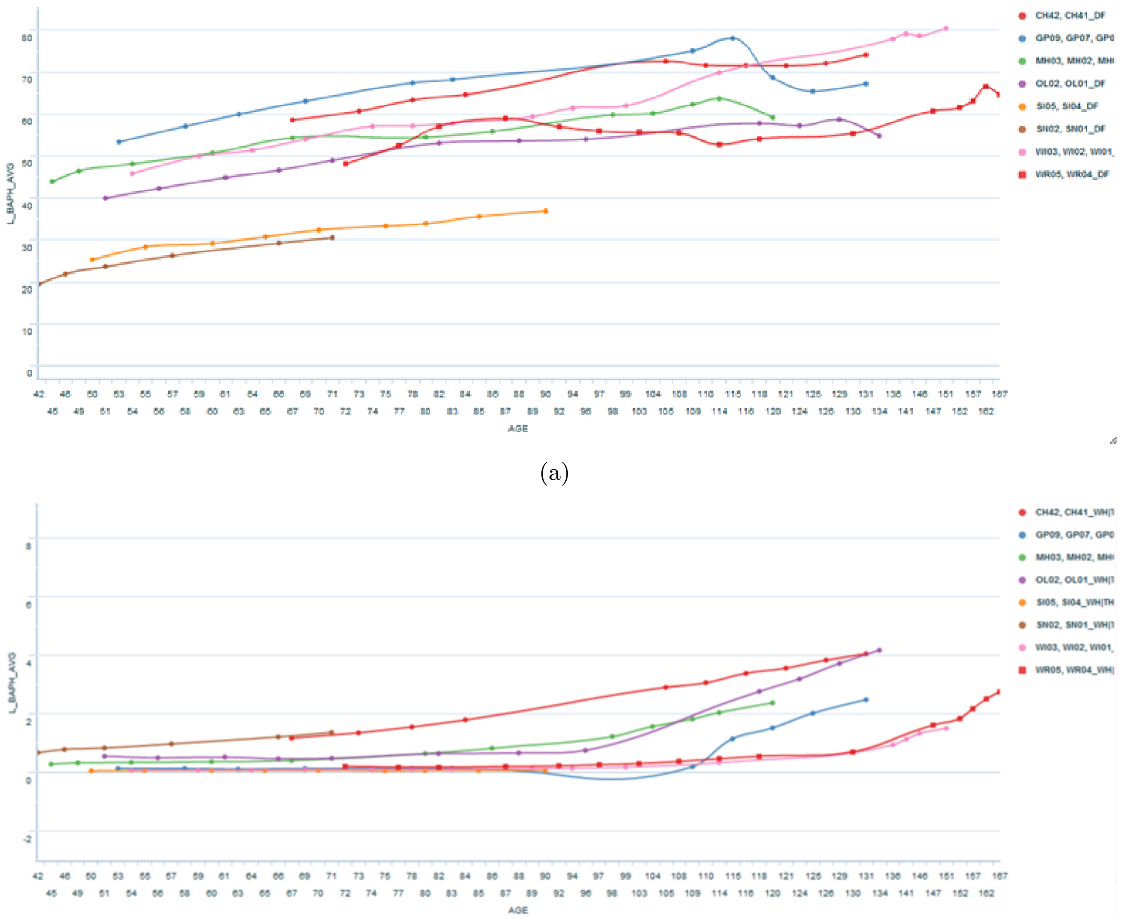
(a)

(b)

Figure 4.8: Long-term trends in density (trees/ha) of (a) Douglas-fir, and (b) shade-tolerant species in Douglas-fir-dominated permanent plots in Oregon and Washington ($n = 8$ locations, 2-5 plots per location). Note different scales of y-axes.

may have caused tree death. The beetle mortality occurred first at GP when the stand was 120 years old, and led to a pronounced but temporary decline in Douglas-fir basal area. The drop in Douglas-fir basal area there was accompanied by increases in both mean density and basal area of the shade-tolerant species, likely as a result of increased resources (e.g., light, nutrients, water) available to the understory trees. The ecologist also planned to share the visualization with an entomologist to gain insights on localized and regional outbreaks of Douglas-fir bark beetle.

Case Study Summary. This case study emphasizes the reusability of EcoDATE (i.e.,



(a)

(b)

Figure 4.9: Long-term trends in basal area (m^2/ha) of (a) Douglas-fir, and (b) shade-tolerant species, in Douglas-fir-dominated permanent plots in Oregon and Washington ($n = 8$ locations, 2-5 plots per location). Note different scales of y-axes.

data upload and configuration) and how it aids the scientist in adapting to the pre-defined structure of the data. In this example, EcoDATE supported the process of constructing and deriving visualization views—for example, automatic combinations of ordinal variables (e.g., month and year) derived new ordinal variables (e.g., month-year) for line charts. These features prove important to analysis of long-term ecological data since the data may get updated periodically over time and there exist multiple levels of data aggregation by various factors such as time (e.g., day, month, year), space (e.g., plot and stand), and species groups.

To summarize, the three case studies serve a primary purpose of assessing the utility the EcoDATE tool in the context of its target users, three ecologists in this case, exploring real world data sets in their normal working environment. The qualitative results show how use led to refinement of the tool and helped ecologists gain insights into their data and formulate new research questions. Our next step was to deploy the tool to a broader pool of ecologist users, starting with a working group at the 2012 LTER All Scientists Meeting as we describe next.

4.6.4 Working Group at the LTER ASM 2012

We further evaluated an early version of the EcoDATE tool in a working group at the 2012 LTER All Scientists Meeting, a network-wide meeting of over 750 scientists and students for scientific discussions, plenary talks, working groups, and scientific posters (http://asm2012.lternet.edu/). The EcoDATE working group was an information exchange session focused on (1) how ecologists approach analysis of long-term ecological data, (2) how interactive visualization may help with the data exploration process, and (3) the pros and cons of the proposed EcoDATE tool. During the session, we demonstrated the application of EcoDATE using several long-term data sets, invited participants to experiment with the visualizations in focus-group settings, and obtained feedback via a survey. Fifteen participants experimented with the tool and completed the survey: one professor, four LTER site managers/information managers, five post-docs, and five graduate students.

The evaluation survey consisted of five Likert-style statements, in which participants were asked to indicate their level of agreement on a scale of one (Strongly Disagree) to five (Strongly Agree), and three open-ended questions. In spite of relatively short usage time (around 30 minutes), most of the participants agreed that the tool is easy to use (L1 and L2) and they strongly liked using it (L4 and L5) (Figure 4.10).

In addition to the Likert-style statements, the survey included the following three open-ended questions: (1) what aspect(s) of the tool did you like most? (2) what aspect(s) of the tool did you dislike most? and (3) if possible, how would you change the tool to improve it? Overall, many participants praised the tool for its interactivity,

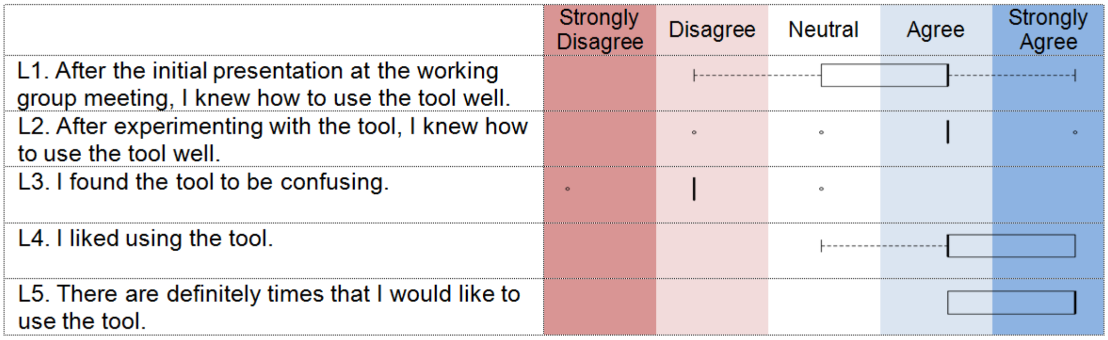


Figure 4.10: Boxplot of responses to each of the five Likert-style statements. The participants were asked to indicate their level of agreement on a scale of 1 (Strongly Disagree) to 5 (Strongly Agree).

holistic view of multivariate data/histograms, and ability to share visualizations with others. Among interactive features, participants highly favored data subsetting/filtering (nine out of 15 participants). However, some found it difficult to compare the temporal trends across variables (i.e., align the time axes across different line chart views), and they suggested superimposed line chart with two y-axes [76], which is a feature we want to study for future work. Participants also expressed the wish to use the tool with their data. We responded to that request and equipped the current version of the tool with the data upload feature.

4.7 Discussion

Although long-term ecological studies are essential for detecting changes in the environment, understanding of these changes is limited by capacity for data analysis (Figure 4.1). Data often accumulate faster than ecologists can analyze them, creating a bottleneck. Over time, hypotheses that guided establishment of a study may become irrelevant, and new hypotheses and new factors may emerge. Therefore, long-term studies require exploratory analysis to deal with growing data and changing scientific questions. Tools such as machine learning and statistics, which aim to simplify and automate data analysis, are of limited value for analysis of long-term ecological data because they assume well-defined and confirmatory tasks and hypotheses, such as computing the correlation between two variables or predicting the occurrence of some specific ecological event. In this chapter, we argue that interactive visualization provides a visual gateway to long-term ecological data, allowing users to explore data directly and complementing further analyses using statistics or machine learning.

Development and evaluation of the EcoDATE tool reveals several different strategies used by ecologists to explore long-term ecological data. Target users of interactive visualization for long-term ecological data occupy a spectrum ranging from scientists who are interested in general ecological phenomena and may have little specific knowledge of the data to scientists who have collected the data and studied them intensively. Therefore, an interactive visualization tool must permit overview of data as well as exploration of a priori hypotheses and generation of new ones. The EcoDATE tool supports a "breadth-first" exploration approach as demonstrated in the cone production data case study, in which the ecologist analyzed the data for the first time. In that case, the main analysis strategy followed the visual information-seeking mantra "overview first, zoom and filter, then details on demand" [141]. On the other hand, the tool also facilitates "depth-first" analysis, as demonstrated in the stream chemistry data and forest structure case studies, in which the ecologists had prior knowledge of the data. In these cases, the analysis followed a three-step process of specifying visualization views, characterizing views, and gaining insights (Figure 4.6). A visualization tool that facilitates open-ended exploration is essential to accommodate the varied analysis strategies used with long-term ecological data. A visualization tool is only part of a larger analysis process (Figure 4.1).

Development and evaluation also suggests the potential for integration of the EcoDATE tool with other tools and archived data sets. We envision that visual-analysis tools such as EcoDATE could become an add-on module in a workflow system or could take advantage of that framework for managing provenance or history of interactions (e.g., visualization states). Current workflow systems such as Kepler [97] lack support for interactive visualizations and usability. Also, web-based interactive visualization tools could support more complex on-site data exploration within existing data repositories or portals such as EcoTrends [137, 118] and Clim-DB/Hydro-DB [66] where large collections of long-term ecological datasets are archived. EcoDATE, a web-based application, could be easily integrated into these portals. In addition, sample visualizations could be presented to visitors to promote data analysis.

Design and use of EcoDATE also provides an insight into the evolution of long-term ecological data collection and analysis. The three data sets of interest were initiated several decades ago, and involve capital- or labor-intensive data collection at a limited number of pre-defined locations and times. Nowadays, long-term ecological data are increasingly being collected at fine temporal and spatial scales, at many sites, and possibly even at moving sites (e.g., tagged organisms) (e.g., [127]). For these data, visual analysis tools will need to accommodate combinations of time, space, and multiple variables. As an example, while filtering supported by EcoDATE is limited to values of bins, we intend to investigate more expressive filtering based on natural language used by ecologists or on a structured query language [62]. Further, the visualization community has shown interest in techniques for spatio-temporal visualization or geovisualization [8].

EcoDATE is now available to the public at `http://purl.oclc.org/ecodate`. We hope it will be utilized by ecologists, who will bring a variety of data sets and provide feedback and suggestions for improvements to the tool. In addition, we will analyze log data to identify dominant usage patterns and features and to understand how EcoDATE may play a role in shaping the scientists hypothesis generation strategies in the context of long-term ecological data.

4.8 Conclusions

In this chapter, we describe the design, implementation, deployment, and evaluation of EcoDATE, an interactive web-based visual-analysis tool designed for the analysis of long-term ecological data with a focus on distribution patterns and temporal trends. The tool combines information visualization techniques with chart types commonly used in ecology. EcoDATE was developed through a process of user-centered design in collaboration with long-term ecological research. Application of the EcoDATE tool to long-term ecological data sets on cone crop production, stream chemistry, and forest structure reveals that it facilitates overview, initial hypothesis testing, and hypothesis formulation in an open-ended framework. Ecologists' initial formative evaluation of EcoDATE indicates that interactive visualizations promote discovery in ecology and reveal several alternative pathways ecologists pursue for analysis of long-term ecological data. This study demonstrates that collaboration between ecologists and visualization researchers can potentially provide powerful tools for identifying important ecological patterns and trends while supporting scientific collaboration. Visual analysis collaboration between visualization researchers and ecologists underscores a promising direction likely to benefit ecology as a discipline.

Chapter 5: Visualization of Diversity across Multiple Attributes: A Case Study of Diversity Faultlines in Work Teams [1]

5.1 Introduction

Effective management of work teams is widely regarded as critical to the success of organizations. Therefore, leveraging the benefits of teamwork while reducing negative outcomes associated with groups has been a central focus of organizational research [90, 14, 150, 24]. For example, researchers study how the demographic diversity of team members (e.g., age, gender, ethnicity, functional background) affects performance and outcome processes (e.g., productivity, collaboration, conflict). They investigate diversity not only as a distribution along one employee attribute (e.g., group ethnic diversity) but also as *a complex composition of multiple attributes* that results in *diversity faultlines* [90]. For instance, faultlines may split a diverse project team into the two subgroups of two senior male software engineers and two junior female QA testers.

A common approach to understanding faultlines within a team relies on faultline metrics [149, 14, 150], which measure the extent to which the given team is divided into *relatively homogeneous subgroups* across the attributes of interest, and tabular data of subgroup structure (see Table 5.1 for an example). Unfortunately, as the number of attributes and team members to be examined both increase, table-based assessment of faultlines and subgroup structure becomes difficult, time-consuming, and tedious. To our knowledge, very little work has been done to develop visual representations that reveal faultlines across multiple attributes. In fact, this lack of tools is considered a challenge in management research that hinders the development of the faultline theory to a more applicable and useful level [150, 24].

We envision that a visualization that leverages faultline metrics with appropriate representation and interaction techniques would complement the faultline approach. Specifically, such a visualization would allow researchers to explore faultlines within teams

[1] The material in this chapter represents joint work with Ronald Metoyer, Katerina Bezrukova, and Chester Spell [122, 123].

quickly and iteratively. Managers or human resources departments could use the visualization to inspect team dynamics based on faultlines and potentially reassign members in hopes of improving performance. Such visualizations could also prove useful in understanding the dynamics of online volunteer teams (e.g., open source software development teams) [108, 26].

In this chapter, we formalize the faultlines visualization problem and provide three contributions. First, we propose a representation that aims to reveal faultlines and subgroup structure of diverse teams across multiple attributes. The proposed representation, HIST, is based on multiple linked, stacked histograms in a parallel axis layout [75, 74]. To our knowledge, while these techniques separately are well-known, as a whole, their application to representing clusters in general and team diversity faultlines in particular is novel and it is a first attempt to explore the design space for the problem. Moreover, the novelty of HIST is in the approach of attribute visibility (or object distribution) [146] to representing clusters as opposed to object visibility studied in previous work [70, 135].

Second, we contribute results of a controlled user study to compare HIST to the parallel coordinate plot (PCP) [75, 74] and the scatter plot matrix (SPLOM) [29], the two other common techniques for representing clusters of multivariate objects [70]. With respect to user performance, the results show that users can judge faultlines using HIST as or more consistent and accurate as the other methods. Furthermore, the findings can be generalized to representations and tasks involving distributions of clusters/subgroups in mixed-type data, extending the previous work [70].

Finally, we incorporate computational analysis into HIST to assist users in detecting faultlines or subgroup separation. Specifically, inspired by the physical form of geological faults, we propose visual enhancements as connecting dashed lines across attribute axes to represent "cracks" within a team, as depicted in Figure 5.2. In our algorithm, we cluster attribute values by subgroups using Bertin Classification Criterion [125] and we introduce a metric, Total Separation Criterion, to automatically detect attributes with separable subgroups.

5.2 Diversity Faultines Background and Design Requirements

5.2.1 Diversity Faultlines Concept

Faultlines are described as *hypothetical dividing lines* that may split a team or workgroup into relatively homogeneous subgroups based on one or more attributes [90]. Measuring faultlines of a team is adopted from multivariate clustering—that is, the measure assigns team members into subgroups (or clusters) according to their similarity across the attributes of interest (e.g. demographics). Clusters (or subgroups) have maximum internal homogeneity or between-cluster heterogeneity.

Team data represent team members characterized by multiple demographic attributes of varying types (e.g., numeric, ordinal, and nominal). As an example, consider two teams as shown in Table 5.1: Teams 1 and 2 consist of five and seven members, respectively. We computed team faultlines along three characteristics of AGE, ETHNICITY, and EDUCATION (degree) using a widely accepted measure proposed by Thatcher et al. [149]. For each team, the measure identifies the subgroups (*Subgroup* column) corresponding

Team	**AGE**	**ETHNICITY**	**EDU**	***Subgroup***	***Fau***
1	21	T	E	*1*	*1.00*
1	23	T	E	*1*	
1	20	T	E	*1*	
1	50	Y	A	*2*	
1	52	Y	A	*2*	
2	21	W	E	*1*	*0.56*
2	23	W	A	*1*	
2	22	U	B	*2*	
2	26	X	B	*2*	
2	21	Z	D	*3*	
2	23	Z	C	*3*	
2	22	Z	B	*3*	

Table 5.1: Synthetic data of the two work teams. Faultline measure [149] clusters each of the two teams into subgroups (*Subgroup*) and identifies the team faultline strength (*Fau*).

to *the strongest group partitioning* following the formula:

$$Fau_g = \left[\frac{\sum_{j=1}^{p} \sum_{k=1}^{n_g} n_k^g (\bar{x}_{.jk} - \bar{x}_{.j.})^2}{\sum_{j=1}^{p} \sum_{k=1}^{n_g} \sum_{i=1}^{n_k^g} (x_{ijk} - \bar{x}_{.j.})^2} \right] \quad g = 1, 2, ..S, \tag{5.1}$$

where p is the number of attributes of interest, n_g is the number of subgroups in the partition g, n_k^g is the number of members in subgroup k of partition g, $\bar{x}_{.jk}$ is the mean value of attribute j in subgroup k, $\bar{x}_{.j.}$ is the overall mean value of attribute j, and x_{ijk} is the value of attribute j of member i in subgroup k. Since Fau_g takes numeric values, each categorical attribute must be recoded into a series of dummy variables and rescaled across the attributes [149]. For example, a five year difference in AGE is equivalent to a difference in ETHNICITY and a difference in EDUCATION based on a given data sample.

The variable faultline strength Fau, which always takes a value between zero and one, is the maximum over all $\{Fau_g\}_{g=1}^{S}$. The larger the faultline strength value, the stronger the separation between subgroups or equivalently, the more attributes in which the subgroups are separable. The concept is inspired by *geological faults* whose strength increases with the number of layers it cuts through [90]. Since Fau is based on brute-force search, it is suited only for small teams. Thatcher et al. [150] and Meyer and Glenz [104] present thorough surveys of existing faultline measures.

5.2.2 Design Requirements

Table 5.1 does not clearly show where the separation (or “cracks”) occur in a team and this is precisely the problem that we address with our visual representation. Here we discuss design requirements as validated by our collaborators, who are experts in management research and also co-authors on this chapter. Moreover, these requirements are empirically associated with team outcomes in faultlines literature. Specifically, a faultline representation of a given team should allow users to explore efficiently:

- **R1.** Faultline value (e.g., faultline strength *Fau*). Such numeric quantification of a faultline can be used to compare different teams quickly or to predict the effects of faultlines on outcome processes [14, 150].
- **R2.** Faultlines themselves, or where do the “cracks” occur in the team? A “crack”

or total separation occurs within an attribute when members of different subgroups fall into different subset of values in the attribute space.

- **R3.** The inner structure of subgroups in the team including the number of subgroups, evenness of subgroups, and subgroup diversity or distribution [55]. These important constructs are associated with distribution of power, resources, and abilities in the team [90, 24].

In addition, the representation should scale well to the number of members in a team and number of attributes of interest. Management researchers have typically studied small teams of up to 16 members that may potentially split into up to seven subgroups, depending on team size and the number of attributes [89, 24], yet they are also interested in teams of larger sizes (e.g., online volunteer groups [108, 26]).

Finally, while conventional cluster analysis usually concerns object visibility and separation in attribute space of quantitative attributes [135], we note that faultlines analysis emphasizes distribution or alignment of objects across multiple attributes of varying types. Furthermore, a faultlines visualization requires a faultline measure or a clustering algorithm as an external data pre-processing step to pre-assign team members to subgroups, as opposed to letting users identify potential subgroups or implicit clusters from representations of raw data [70].

5.3 Related Work

Design and evaluation of our proposed technique was informed by related work on visual representations and user studies of cluster representations, which we discuss here.

5.3.1 Representing Clusters

Here we review a subset of existing representation techniques that are potentially applicable to clustering and team faultlines. More general surveys of visual representations can be found in [83, 38].

Scatter plots are probably the most common technique to represent clusters of objects [70]. However, without additional encoding, possible data overlap/occlusion may lead to ambiguous interpretation of the abundance of objects, especially among categorical attributes. The histogram, on the other hand, takes advantage of data overlap to show the distribution of objects over a single attribute. Our proposed technique, which is based on histograms, aims to convey object distribution instead of object visibility. As noted, these techniques display only one or two attributes of interest.

The dimensionality problem may be solved by using *multiples.* For example, the scatter plot matrix (SPLOM) [29] extends scatter plots to represent clusters of multivariate objects, although multiple pairwise projections of the data attributes require more screen space and potentially cognitive load placed on the user. On the other hand, multiple histograms could be useful for representing distribution of multiple attributes in a parallel axis layout [59]. Furthermore, histograms have been proved effective in communicating diversity information in separate attributes in previous work [120, 119] (see Chapters 3 and 4). Our proposed representation of diversity faultlines is in fact multiple histograms augmented with histogram stacking and color encoding.

The parallel coordinates plot (PCP) [75, 74] is another common approach to representing clusters of high-dimensional objects [70]. Similar to SPLOM, PCP may suffer from occlusions caused by data overlap as the number of objects increases and many categorical attributes exist, as in the case of demographic data. Several variants of PCP such as Parallel Sets [86] and Diversity Map [119] overcome this limitation by providing information on the distribution of values for each attribute. However, it is not clear how multiple clusters are embedded into these techniques.

Star coordinates [81] may be suited to representing the overall structure of a set of objects over multiple attributes. Additional encoding such as colors may be used to reveal explicit clusters in the data. Unfortunately, the mapping between a data point and its location in star coordinates is not one-to-one. Consequently, several different data points may end up in the same location if they have equal vector sums.

Among stacked displays [83, 161], the mosaic plot [58] could be used for showing subgroup structure since subgroups are stacked within a team. While in theory, the stacking process may be repeated multiple times, in practice space constraint limits the number of attributes as well as number of possible values in an attribute. Therefore, mosaic plots can be useful only when the number of attributes is relatively small. In our proposed histogram-based technique, we apply the stacking process to histogram bars only once.

Finally, there are hybrid approaches that integrate multiple representations in one view. The most relevant technique is DICON [21], a treemap- and icon-based technique designed to visualize structure of clusters. Unfortunately, the technique supports only quantitative attributes.

5.3.2 Evaluating Cluster Representations

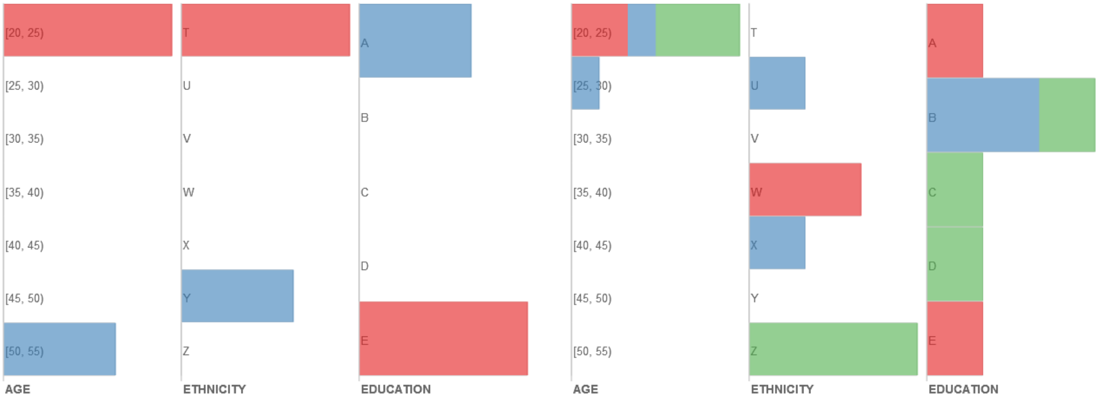
The closest exemplar to our user study is that of Holten and van Wijk [70]. They evaluated cluster identification performance of nine PCP variants, two of which are the standard PCP and a variant with embedded scatterplots (SP). Nevertheless, unlike our scenario involving explicit clusters in demographic data, their study used simulated quantitative data with no pre-computation of clusters. The most interesting finding from their study is that despite the apparently valid improvements of the PCP variants, scatterplots are more effective than PCPs with respect to PCP-based cluster identification tasks. Furthermore, participants favored SP as the least difficult variation. Following the result, the authors called for further evaluation of techniques that explicitly highlight pre-computed clusters, for example, with unique colors. We respond to that call in our user study by augmenting standard PCP and SPLOM—the two controlled methods—with color encoding of explicit clusters. We also extend the study to include other tasks appropriate for faultlines/cluster analysis.

In another related study, Li et al. found that scatter plots are more effective than PCP's in supporting visual correlation analysis [95]. We suspect that scatter plots are also more effective than histograms in conveying correlation information because histograms represent each of the attributes independently. Therefore, while management researchers may be interested in correlations among demographic attributes (e.g., age and experience), we do not consider correlation-related tasks in our user study.

5.4 Visualization Design

5.4.1 Design Considerations and Prototype

A histogram is well suited to showing the diversity or distribution of objects within an attribute (requirement **R3**). According to Mackinlay [99], position and length are ranked highly for encoding nominal and numeric values such as variety of attribute values and abundance of objects, respectively. In addition, previous work suggests that the parallel



(a) (b)

Figure 5.1: Synthetic data (Table 5.1) of (a) Team 1 and (b) Team 2 visualized using HIST. Distinct colors are used to differentiate the subgroups: subgroup 1, subgroup 2, and subgroup 3. While the two subgroups of Team 1 are totally separated in all three attributes of AGE, ETHNICITY, and EDUCATION, the three subgroups of Team 2 are totally separated in ETHNICITY only (column 2).

axis layout [75, 74] of multiple distributions is capable of conveying a *holistic* object distribution over multiple attributes [59, 119]. However, the previous work does not consider how distributions of multiple subgroups align over multiple attributes. Since subgroups are nested within a team, to maintain bar length encoding, a natural solution to encoding subgroups is to stack bars within each bin (Figures 5.1 and 5.2). We then use distinct color hues on a white background to differentiate stacked subgroups. Our choice of qualitative colors provided by ColorBrewer [56] meets the requirement of encoding up to seven subgroups. On another note, the length of each bar is scaled according to $l(x) = |x|/|x_{MAX}|$, where $|x|$ denotes the number of objects in bin x, and x_{MAX} is the bin with the most objects for the attribute in question. We also discretized numeric attributes into bins based on their rescaled factors (equation 5.1).

Following this design, a total separation or “crack” occurs at a *nominal* attribute when distinct subgroups (or distinct colors) occupy distinct positions along the vertical axis (requirement **R2**). Total separation at a *numeric* or *ordinal* attribute further requires that these distinct positions—including ones without objects (zero-length bars)—are contiguous (e.g., AGE and MLB_TENURE histograms in Figure 5.2).

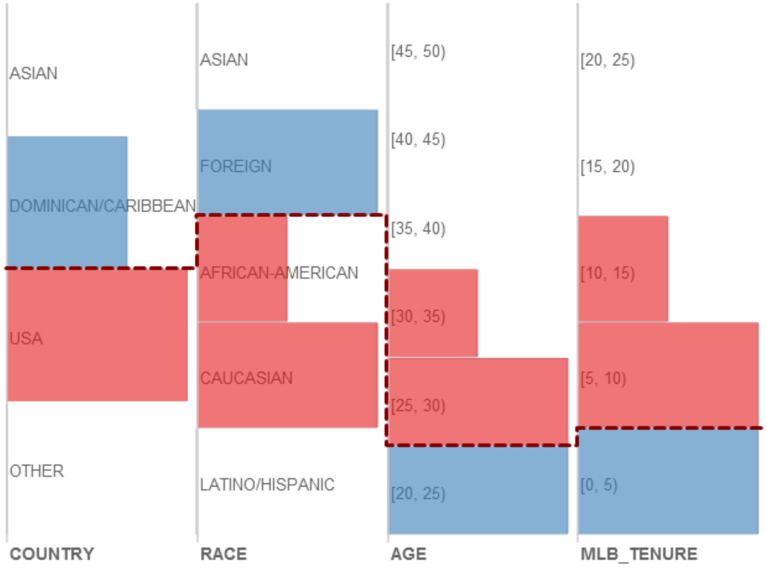


Figure 5.2: Group of starting pitchers of the MLB team Brewers in 2008 visualized using HIST. The two subgroups are totally divided in all four attributes of COUNTRY, RACE, AGE, and MLB TENURE. The connecting dashed lines, which are described in detail in Section 5.7, are overlaid to represent the holistic "cracks" between the two subgroups.

The HIST representation communicates the overall degree of separation of subgroups in a given team (i.e., faultline strength) as the combined separation of all demographic attributes under investigation (requirement **R1**). In the limit of *perfectly strong* faultlines, where different subgroups occupy different subset of attribute values across all the attributes, all the bars of the histograms will have solid colors, as depicted in Figure 5.1(a). On the contrary, a team with *very weak* faultlines will produce a visualization with most of the bars stacked with at least two colors like the AGE histogram in Figure 5.1(b). Moreover, while the chosen Fau measure (equation 5.1) [149] does not consider how far apart the subgroups are, especially on quantitative attributes (i.e., *faultline distance* [14]), we note that stacked histograms of quantitative axes are able to reveal the potential gaps or distances between subgroups. For instance, the AGE histogram in Figure 5.1a shows a big "generation gap" between the two subgroups.

5.4.2 Informal Evaluation with Management Researchers

A close collaboration between management and visualization researchers was critical for the design of HIST. The management researchers help validate the design requirements and evaluate the design iterations and prototypes. Thus far, we have applied the prototypes to two real-world data sets: Major League Baseball (MLB) teams (Figure 5.2) and an empirical faultlines study [15] (Figure 5.3). The domain experts found the representation helpful in inspecting subgroup structure of different teams and in developing a sense of where the separations are likely to occur following their configuration of the faultline measures.

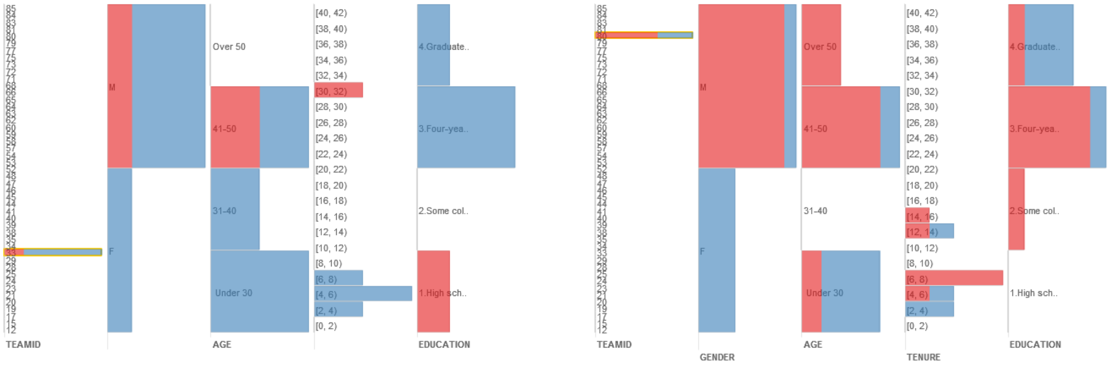


Figure 5.3: HIST representation of the subgroup structure of a team with strong faultlines (left view, Team 33) and a team with weak faultlines (right view, Team 80) from the faultlines study data set [15]. Columns from left to right are Team ID, gender, age, company tenure, and education.

5.4.3 Motivation for Futher Evaluation

While the qualitative results from our design study with domain experts are encouraging, they have limitations. First, the study represents an informal evaluation based on observations [88]. It lacks controlled visualization techniques (control groups) as well as various data sets with controllable characteristics serving as ground truth answers. Second, our two management researcher collaborators represent only a small set of potential users of the visualization. The proposed faultlines visualization HIST could potentially support a wide range of target users: (1) management researchers who study faultlines and subgroups theories [24], (2) human resources departments who manage current employees and recruiting new employees [49], and (3) managers and officials from many areas concerning work teams such as education, sports, and entertainments to name a few. Third, thus far, the management researchers limited the use of the faultline visualizations to data exploration only, accompanied by further statistical analysis. The design of HIST targets both data exploration (e.g., data analysis) and communication (e.g., charts in a publication or training). Finally, while HIST is designed based on the requirements of faultlines and subgroup structures in work teams, it can be potentially utilized to communicate distributions of clusters/subgroups in mixed-type data, for example, compare structures of functional groups in ecological and microbiological data [116, 117, 37].

To overcome these limitations and make the evaluation results generalizable, in the next section, we extend our design evaluation with a controlled user study designed to understand the effectiveness of a visual representation in a broader context of communicating information on faultines as well as on distributions of clusters/subgroups in mixed-type data.

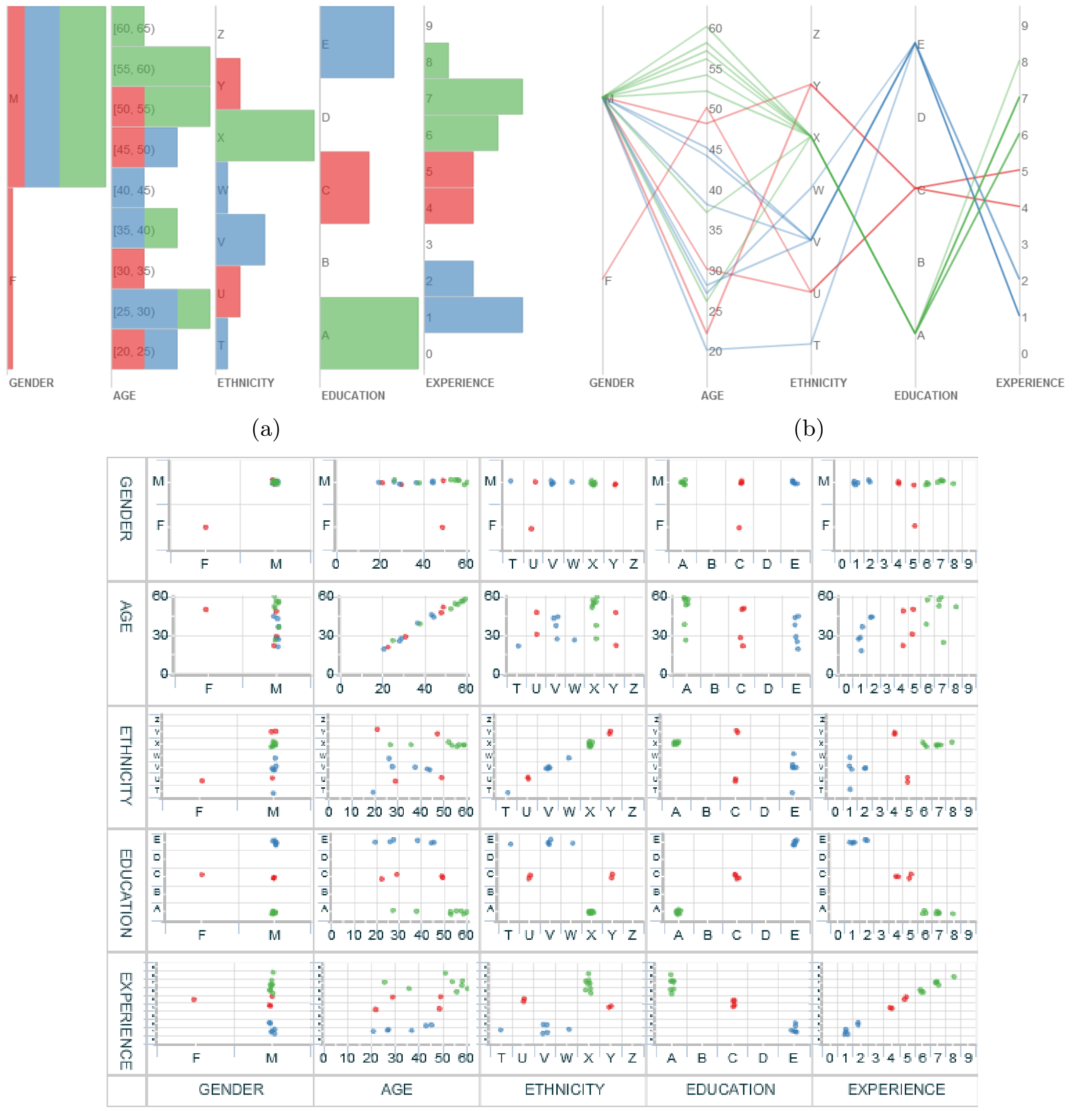
5.5 User Study Design and Implementation

In this section, we describe the design and implementation of a formal user study intended to evaluate the effectiveness of HIST at communicating faultlines information in teams. Specifically, we compare HIST to PCP [75, 74] and SPLOM [29], the two common techniques for representing clusters of multivariate objects. In fact, a previous study has shown that scatterplots are the most effective among variants of PCP for cluster identification tasks [70]. Figure 5.4 depicts examples of the three techniques.

5.5.1 Task Design and Implementation

The task design includes three important components: (1) a set of task-oriented questions, (2) a procedure for generating synthetic team data, and (3) design of the three visualization techniques under comparison.

User Study Task Questions. The study contains six types of questions intended to assess the capability of a particular visual representation in conveying different aspects



(a)

(b)

(c)

Figure 5.4: Example team of size 18 visualized using (a) HIST, (b) PCP, and (c) SPLOM. Distinct colors are used to differentiate the three subgroups: subgroup 1, subgroup 2, and subgroup 3. While subgroup 3 is the biggest, subgroup 1 is the smallest. The three subgroups are totally separated along ETHNICITY, EDUCATION, and EXPERIENCE because different subgroups occupy different subsets of values along these attributes. The three subgroups overlap in GENDER and AGE because there exist values of these attributes shared by different subgroups. The faultline level is MEDIUM considering that the subgroups are totally separated in three out of five attributes.

of team faultlines (requirements R1 - R3). Note that in accordance with the previous cluster identification study [70], we design the tasks to be relevant to both faultlines and general cluster representations of mixed-type data and not tied to users with specialized demographics knowledge.

Q1: *How many subgroups are there in the given team?* (possible answers: 1 to 7). This question type is designed to determine if a representation technique supports users in identifying the number of subgroups/clusters in a team/data set (requirement R3). This type is equivalent to the only cluster identification task in the previous study [70].

Q2a/b: *Among the existing subgroups in the given team, which one is the biggest/smallest?* (possible answers: Subgroup 1 to 7). These two types are intended to measure the user's ability to determine evenness of subgroups or equivalently, isolate subgroups/clusters that contain most and least members/objects using a representation (requirement R3).

Q3: *In which attributes are the subgroups totally separated?* (possible answers: the attributes under investigation). The goal of this question type is to test if a representation technique supports users in isolating the attributes that totally separate subgroups/clusters and result in faultlines (or "cracks") within a team (requirement R2).

Q4: *To what extent are the subgroups separated across all attributes?* (possible answers: Very Weak, Somewhat Weak, Medium, Somewhat Strong, Very Strong). This question type is intended to gauge how well a user can interpret and assign a faultline level to a team using a visual representation (requirement R1). Within the scope of this study, the faultline level of a team is determined by the number of attributes in which the subgroups are totally separated. While this assessment does not consider attributes with partial separation of subgroups as the way the Fau measure (equation 5.1) quantifies separation of subgroups, it makes answering this task question more straightforward to participants.

Q5: *Between two different teams, which team has stronger separation of subgroups?* (possible answers: Team A or Team B). This last question type is intended to determine if a representation technique is discriminative enough to allow a user to compare the faultline levels of two teams depicted in two visualizations of the same technique (requirement R1).

In our user study, each of the question types was asked multiple times on different teams/data sets. We identified the best answers to the questions based on the distribution of members across subgroups and the attributes in which subgroups are separable. These constructs are achieved using our synthetic data generation procedure, which is described next.

Synthetic Team Data Generation. For the study, we create teams formed from automatically generated data sets. Technically, our method generates *pre-clustered* teams over a manually defined set of mixed-type demographic attributes, where team size, number of subgroups, evenness of subgroups, and separation of subgroups are controlled. The aim was to simulate teams with realistic distributions of members while controlling the faultlines and subgroup structure.

In our setting, we have one variable X for each attribute, and we hand-specify the categorical values or range of values for X. To generate a team, we specify its input parameters including the number of subgroups k, subgroup sizes $\{n_i\}_{i=1}^{k}$, and the set of attributes in which the subgroups are totally separated $\{X_s\}$. Note that $n = \sum_i^k n_i$ denotes the size of the entire team. For each X_s, we randomly partition its attribute space into k distinct subsets of values and we draw randomly n_i samples from each subset for each subgroup i. This guarantees that the subgroups are totally separated in these attributes $\{X_s\}$. For rest of the attributes $\{X_{ns}\}$, we model the distribution over its possible values either as *uniform* or *skewed* distribution and we draw randomly n samples from each of these distributions. We choose these specific distributions based on the realistic distributions of the team demographics in management literature: *uniform* distribution corresponds to diversity as *variety* and *skewed* (or relatively *homogeneous*) distribution corresponds to diversity as *disparity* [55]. For example, while both genders may be uniformly represented in some teams (e.g., student body), either male or female gender may be dominant in other teams (e.g., organizational groups). Once the samples are created for each of the attributes, we use the j^{th} sample for each attribute as the corresponding attribute value of the j^{th} member in the generated team. Finally, since the team is already clustered into subgroups, we simply use the *Fau* formula (equation 5.1) to calculate the faultline strength value for the team.

In our generated teams, team members or objects are characterized by the following five independent demographic attributes. We chose these attributes because they are the most commonly used in faultline literature [150].

- GENDER: F or M
- AGE: 20-60, discretized by steps of 5 corresponding to the rescale factor
- ETHNICITY: T, U, V, W, X, Y, or Z

- EDUCATION (degree): A, B, C, D, or E
- EXPERIENCE (level): 0-9

While we believed the study participants would be familiar with these attributes, we used single-letter labels as values of categorical attributes (e.g., T, U, V, . . . for ETHNICITY) to prevent participants from associating their own knowledge of demographics (e.g. ethnic differences) into their answers.

We generated teams whose sizes range from four to 50 members and number of subgroups range from two to seven. The teams with at most 16 members were considered *small teams*. The teams with more than 16 and less than 50 members are considered *large teams* to simulate other workgroup settings such as online volunteer groups [108, 26].

Visual Representations. Figure 5.4 presents examples of stimulus materials of the three techniques under comparison. The design of HIST (without visual enhancements) was described in Section 5.4. In our design of PCP and SPLOM, we also use distinct color hues to differentiate subgroups. To prevent total occlusion due to data overlap, both PCP polylines and SPLOM dots are drawn at a constant opacity of 40% and 60%, respectively. The opacity encoding matches our PCP implementation with that of Holten and Van Wijk [70]. Furthermore, we employed a jittering technique [29] to alleviate data overlap issues in SPLOM.

The resolution of each image produced by the three techniques was 630×430 pixels. Each visualization image was accompanied by a subgroup color legend of 80×270 pixels. We chose these resolutions to ensure that visualization images would fit into a standard 1024×768 pixel screen without requiring any scrolling (the usable screen space for a web page is approximately 960×600 pixels).

5.5.2 Experiment Design and Implementation

Participants. Participants were recruited from Amazon's Mechanical Turk (mTurk), a popular crowdsourcing Internet marketplace which has been shown to be a viable platform for graphical representation experiments [63]. The marketplace allows requesters to post jobs (also called Human Intelligence Tasks or HITs) for a large pool of users (also called workers or turkers) to consider and complete. Since mTurk is a world-wide marketplace, we targeted our participants specifically to those registered in the US with normal vision, at least 95% “approval” rating, and at least 100 tasks approved. After passing the color blindness qualification test hosted on the mTurk website, each participant visited our external study website, read an explanation of the research study (in lieu of a signed consent form), and was randomly assigned to a visualization technique. The qualification test, which is based on the Ishihara Color Test [53], was to detect and exclude interested individuals with color blindness.

In total, 57 participants completed the study (19 for each visualization technique). They represented a diverse range of majors/occupations, gender, and ages (Figure 5.5). Most of them were unfamiliar with the field of InfoVis. In addition to the 57 participants, we excluded the other 10 participants who stopped at the beginning or in the

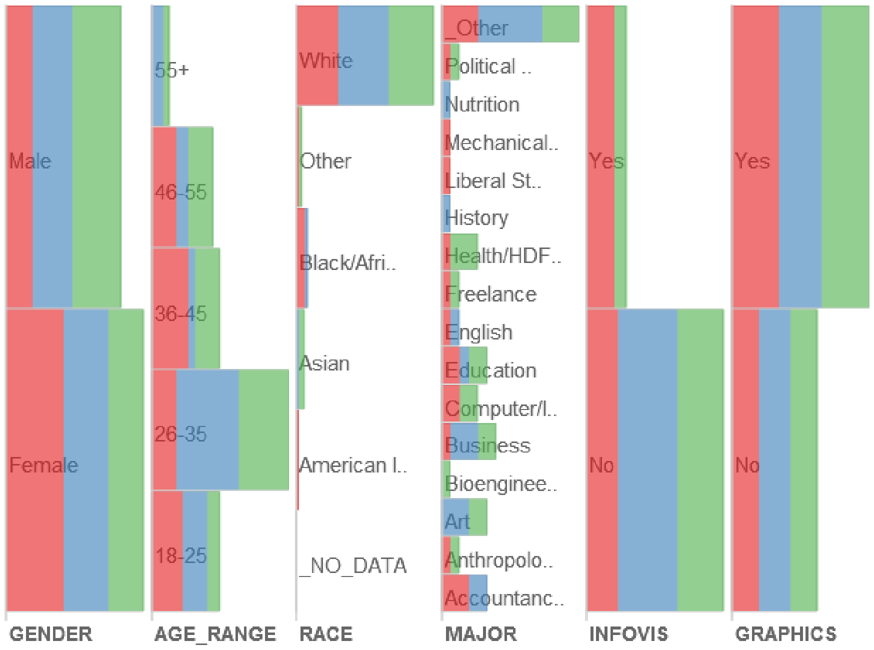


Figure 5.5: Participants of the user study visualized using multiple stacked histograms. The visualized attributes, from left to right, are gender, age range, race, major/occupation, familiarity with InfoVis (yes or no), and familiarity with computer graphics (yes or no). Participants of the three techniques are differentiated by distinct colors: HIST, PCP, and SPLOM. While the three groups of participants were mixed in most of the attributes, they collectively represented a diverse range of majors/occupations, genders, and ages.

middle of the study. These withdrawn participants are evenly distributed across the three techniques (4 for HIST, 3 for PCP, and 3 for SPLOM).

Experiment Design and Procedure. We followed a randomized between subjects study design where the primary factor consisted of three levels (HIST, PCP, SPLOM). Each of the techniques was randomly assigned to each of the participants. We used a common collection of synthetic team data sets for each of the three visualization techniques.

The participant first completed a short tutorial that explained the technique. The tutorial included several baseline visualization examples of very strong, weak, and medium faultline levels. The participant then answered six task questions of each of the types described earlier: three for smaller teams and three for larger teams. During a question, the participant could access a visualization example with annotations highlighting various aspects of faultlines. Note that the questions of one type are the same, but each one is asked about visualizations of different data sets. The ordering of question types was randomized across participants, but all questions of the same type were asked as a block. The ordering of questions in each type was also randomized to avoid ordering effects (e.g., primacy and recency effects) among participants. In total, the number of task questions was 36.

Following the data collection approach in our previous work on visualization of diversity in separate attributes [119] (see Chapter 3), we assigned an error distance to each participant's response to measure how far each response was from the correct answer. We identified the correct answers from the distribution of members across subgroups and the attributes in which subgroups are separated. These constructs are achieved using our data generation procedure described earlier. We also collected the total time participants spent on each response.

In addition to the questions of type Q1-Q5, at the end of the study, the participants answered a short questionnaire about their experience with each technique. This questionnaire contained both Likert-style questions as well as open-ended questions. We discuss the results of these questions in the next section.

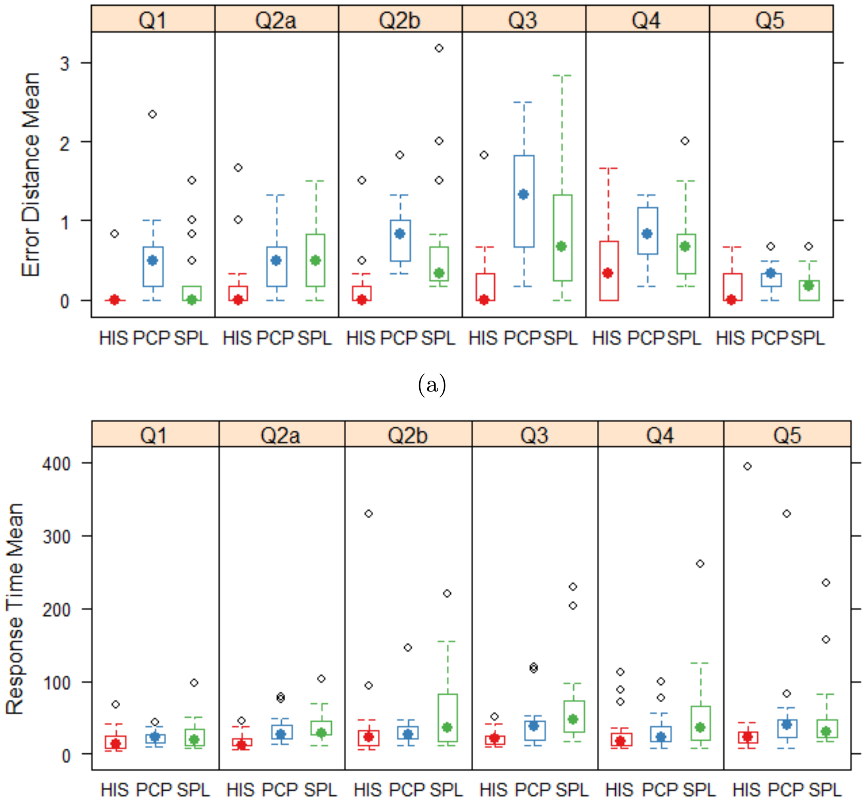
5.6 User Study Results

Initially, we hypothesized that for each type of question, HIST would outperform PCP and SPLOM, both in terms of accuracy and response time. Specifically, we expected users would have difficulty accurately identifying evenness of subgroups (Q2) and separation attributes (Q3) using PCP or SPLOM due to occlusion and visual clutter that may occur with increasing number of objects (i.e., large teams). A secondary factor of the study was to determine whether data set/team size affected participants' ability to judge information on diversity faultlines using a visualization technique.

For each question type, we computed the mean of error distances and the mean of response times across the questions of that type for each participant and compared these aggregated values using hypothesis testing. Since the response data were not normally distributed, we first applied a rank transformation [33] to the data before using ANOVA for statistical tests. Figures 5.6(a) and (b) summarize the error distance and response time results. We pay more attention to error distance when analyzing the results because it is the most important performance measure for a given representation.

Results for Q1. *How many subgroups are there in the given team?* As Figure 5.6 indicates, participants answered Q1 questions more accurately with HIST and SPLOM than with PCP. In fact, there was convincing evidence for an effect of visualization technique on error distance, $F(2, 54) = 10.01$, $p = 0$. Post-hoc analysis using Tukey's HSD (honestly significant difference) revealed convincing evidence for an error distance difference between HIST and PCP ($p_{HIST-PCP} = 0$) but no evidence for such a difference between HIST and SPLOM ($p_{HIST-SPLOM} = 0.255$). Interestingly, when analyzing data separately over small and large teams, we could not find evidence for such a difference between HIST and PCP for small teams ($p_{HIST:small-PCP:small} = 0.277$). In addition, there was no evidence of the effect of visualization on response time.

The results for Q1 suggest that users can identify the number of subgroups existing in a team equally well using both HIST and SPLOM, and PCP for only small teams. We suspect that encoding subgroups with unique colors make identifying the number of subgroups or clusters straightforward. However, PCP performance decreases when data size increases. We suspect crowded and overlapping poly lines may hinder participants from determining the correct number of subgroups in a team. Our results agree with the previous study [70] that SPLOM performs better than PCP on cluster number



(a)

(b)

Figure 5.6: Boxplots of mean of error distances (a) and of response times (b) for each question type as a function of visualization technique (HIST, PCP, and SPLOM).

identification tasks, both for implicit and explicit clusters.

Results for Q2a/b. *Among the existing subgroups in the given team, which one is the biggest/smallest?* The results for Q2 very much favored HIST (Figure 5.6). For Q2a—which involves the *biggest* subgroup—there was convincing evidence for an effect of visualization on both error distance, $F(2, 54) = 9.809$, $p = 0$ and response time $F(2, 54) = 10.87$, $p = 0$. Tukey's HSD multiple comparison tests showed statistically significant differences between HIST and PCP as well as between HIST and SPLOM in terms of error distance ($p_{HIST-PCP} = 0$; $p_{HIST-SPLOM} = 0.001$) and response time ($p_{HIST-PCP} = 0.002$; $p_{HIST-SPLOM} = 0$). The results for error distance held consistent when small and large teams were analyzed separately. The results for Q2b were similar to

Q2a's, with participants tending to identify the *smallest* subgroup more accurately with HIST. With respect to error distance, Tukey's HSD tests revealed convincing evidence for the difference in the two pairs of techniques ($p_{HIST-PCP} = 0$; $p_{HIST-SPLOM} = 0$). Interestingly, when we analyzed error distance data separately over small and large teams, the results held true for large teams only. With small teams, while we found a statistically significant difference between HIST and PCP ($p_{HIST:small-PCP:small} = 0.014$), there was no such evidence when comparing HIST and SPLOM ($p_{HIST:small-SPLOM:small} = 0.890$).

The results confirm our hypothesis that users would make better judgments about subgroup evenness with HIST than with SPLOM or PCP. Again, PCP is the least favorable choice for this task perhaps due to both occlusion caused by data overlap and visual clutter caused by large data sets. As the results suggest, data overlap also hurts SPLOM's performance, especially when the task involved identifying the biggest subgroup in large teams. In contrast, participants using HIST produced consistent answers for both smallest and biggest subgroups and independent of the data set size.

Results for Q3. *In which attributes are the subgroups totally separated?* The results also favored HIST (Figure 5.6). We found statistically significant effects of visualization on both error distance, $F(2, 54) = 17.58$, $p = 0$ and response time $F(2, 54) = 12.15$, $p = 0$. Tukey's HSD tests yielded significant differences between HIST and PCP as well as HIST and SPLOM on both error distance ($p_{HIST-PCP} = 0$; $p_{HIST-SPLOM} = 0.001$) and response time ($p_{HIST-PCP} = 0.013$; $p_{HIST-SPLOM} = 0$). The results held true when we analyzed error distances for small and large teams separately.

These results confirm our initial hypothesis that HIST is the most effective in supporting users in determining attributes in which subgroups/clusters are totally separated, followed by SPLOM and PCP. *This finding is important considering that to the best of our knowledge, no previous work has explored the use of stacked histograms to show the separation of clusters in separate attributes.*

Results for Q4. *To what extent are the subgroups separated across all attributes?* The results somewhat favored HIST, which showed a statistically significant effect of visualization technique on error distance, $F(2, 54) = 4.047$, $p = 0.023$. Tukey's HSD multiple comparison tests reveal convincing evidence of the error distance differences between HIST and PCP as well as suggestive but inconclusive evidence of the error distance differences between HIST and SPLOM ($p_{HIST-PCP} = 0.019$; $p_{HIST-SPLOM} = 0.161$). When error distance data are analyzed separately over small and large teams, the results hold true for small teams only. These results suggest that users would be able to assign a faultline level to a given team at least as accurately using HIST as using PCP or SPLOM.

Results for Q5. *Between two different teams, which team has stronger separation of subgroups?* While there was convincing evidence for an effect of visualization technique on response time, $F(2, 54) = 3.554$, $p = 0.036$, evidence for an effect of visualization on error distance was suggestive but inconclusive, $F(2, 54) = 2.566$, $p = 0.086$. Post-hoc analysis reveals that users answered this question the most quickly using HIST ($p_{HIST-PCP} = 0.053$; $p_{HIST-SPLOM} = 0.076$). In addition, it is suggestive that response accuracy favored HIST over PCP ($p_{HIST-PCP} = 0.092$) but not SPLOM ($p_{HIST-SPLOM} = 0.909$). While these results do not support our initial hypothesis that users would perform more accurately with HIST than with SPLOM, they do substantiate our hypothesis that users would be able to compare the faultline level of two teams the most quickly when using HIST.

Result Summary. The results across Q1–Q5 consistently supported our hypothesis that among the three techniques under investigation, HIST—followed by SPLOM and PCP—is the most effective representation in supporting users investigating faultlines (requirement R2) and inner structure of subgroups (requirement R3) in a given team. For the task involving assigning a faultline level to the team (requirement R1), HIST is at least as effective as SPLOM and PCP. Moreover, users can identify the number of subgroups existing in a team equally well using both HIST and SPLOM. Conversely, PCP performs the worst consistently across the tasks.

These results are complementary to the findings from the previous diversity visualization studies by Pham et al. [120, 119], which showed that multiple histograms are well-suited to communicating the diversity or distribution of objects over multiple attributes separately (see Chapters 3 and 4). Within our study, we could conclude that the multiple linked stacked histograms technique, which takes the approach of attribute visibility (or object distribution) as opposed to object visibility, is well-suited to communicating diversity faultlines and composition distribution in teams.

5.6.1 Subjective Evaluation

After answering the task questions, participants also completed a short questionnaire requesting their thoughts on the visualization technique and their study experience. The questionnaire consisted of 10 Likert-style statements, four NASA TLX questions [57], and three open-ended questions:

- **L1.** I was able to identify the number of subgroups in a team using the chart.
- **L2.** I was able to identify the biggest/smallest subgroup in a team using the chart.
- **L3.** I was able to identify attributes in which the subgroups were totally separated using the chart.
- **L4.** I was able to judge the overall degree of separation (faultline strength) in the team using the chart.
- **L5.** I was able to identify between two different teams, which team had stronger separation of subgroups.
- **L6.** After the initial tutorial session, I knew how to use the chart well.
- **L7.** After answering all of the questions, I knew how to use the chart well.
- **L8.** There are definitely times that I would like to use the chart.
- **L9.** I found the chart to be confusing.
- **L10.** I liked using the chart.
- **O1.** What aspect(s) of the chart did you like most?
- **O2.** What aspect(s) of the chart did you dislike most?
- **O3.** If possible, how would you change the chart to improve it?
- **TLX1.** Mental Demand: How mentally demanding were the task questions?
- **TLX2.** Physical Demand: How physically demanding were the task questions?
- **TLX3.** Temporal Demand: How hurried or rushed was the pace of the task questions?
- **TLX4.** Frustration: How insecure, discouraged, irritated, stressed, and annoyed were you?

We first discuss the results of the Likert-style statements and NASA TLX questions. Figure 5.7 presents the responses to each of the Likert-style statements from the samples of HIST, PC, and SPLOM participants. Overall, the level of agreement from participants was slightly higher for HIST than for PCP and SPLOM regarding making judgments of diversity faultlines components (L01–L05). This evaluation is consistent with participant performance during the task questions. Notably, we found statistically significant difference in level of agreement among the three groups of participants when it comes to identification of attributes with total subgroup separation (**L3**)–the primary task to judge faultlines in a team–$F(2, 54) = 5.14$, $p = 0$. Tukey's HSD tests show significant differences between HIST and PCP as well as between HIST and SPLOM ($p_{HIST-PCP} = 0.02$; $p_{HIST-SPLOM} = 0.02$). The participants also slightly favored HIST over PCP and SPLOM in terms of applicability, ease of understanding, and affinity (L06–L10). These results are supported by the NASA TLX questions (Figure 5.8), which showed significant differences on mental demand (TLX1) and frustration (TLX4) among the three methods, $p = 0.016$ and $p = 0.02$ respectively.

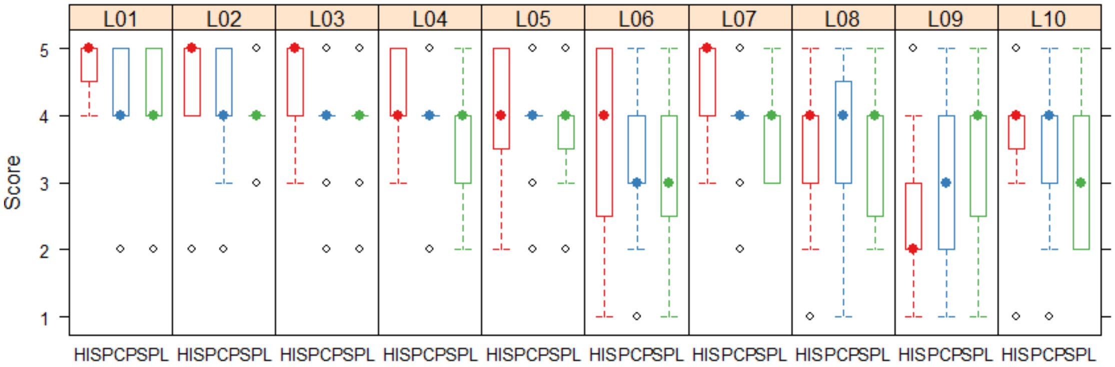


Figure 5.7: Boxplot of responses to each of 10 Likert-style statements as a function of visualization method (HIST, PCP, and SPLOM). The participants were asked to indicate their level of agreement on a scale of 1 (strongly disagree) to 5 (strongly agree).

In addition to quantitative analysis, we also performed qualitative analysis of the three open-ended questions. Overall, many participants praised HIST for its effectiveness and ease of use, especially the use of qualitative colors for encoding subgroups. However, some found it difficult to compare the small differences among bar lengths. As an improvement, they suggested that we selectively attach numbers in the bars. This suggestion is interesting considering that despite the stacking of multiple subgroups, HIST still has screen space to accommodate more information. Regarding PCP, several

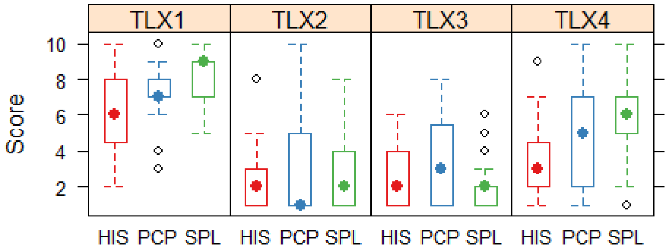


Figure 5.8: Boxplot of responses to each of four NASA TLX questions as a function of visualization method (HIST, PCP, and SPLOM). The participants were asked to indicate the level on a scale of 1 (very low) to 10 (very high).

participants liked its layout, which is novel to them and is able to represent multiple attributes in a single view. Nevertheless, many participants expressed concern about transparency of polylines, which are difficult to discern especially when they are of similar colors (e.g., red and orange). Participants also mentioned that the charts become extremely overwhelming for large data sets. Commenting on SPLOM, several participants liked the technique for its familiarity and ease of understanding. However, similar to PCP, many participants disliked the similar colors among dots. Additionally, many participants requested bigger charts or the zoom-in ability. This confirmed our initial assessement that without interaction techniques [35], the matrix form space requirement of SPLOM is a limitation.

5.7 Faultlines Visualization Enhancement

To further facilitate the faultlines identification tasks, we incorporate computational analysis into HIST. Inspired by the physical layered form of *geological faultlines*, we augment the representation with connecting dashed lines to indicate the *holistic* boundaries of existing separation among the subgroups across the attributes of interest (Figures 5.2 and 5.9). To our knowlegde, this visual enhancement is novel considering that while measures exist to detect separable clusters of quantitative data in 2D scatterplots [144, 135], measures and enhancements for mixed type data in stacked histograms are non-existent.

Technically, the augmentation requires three main computation procedures: (1) reordering values in attribute space, (2) identifying attributes with total subgroup separation, and (3) drawing the lines.

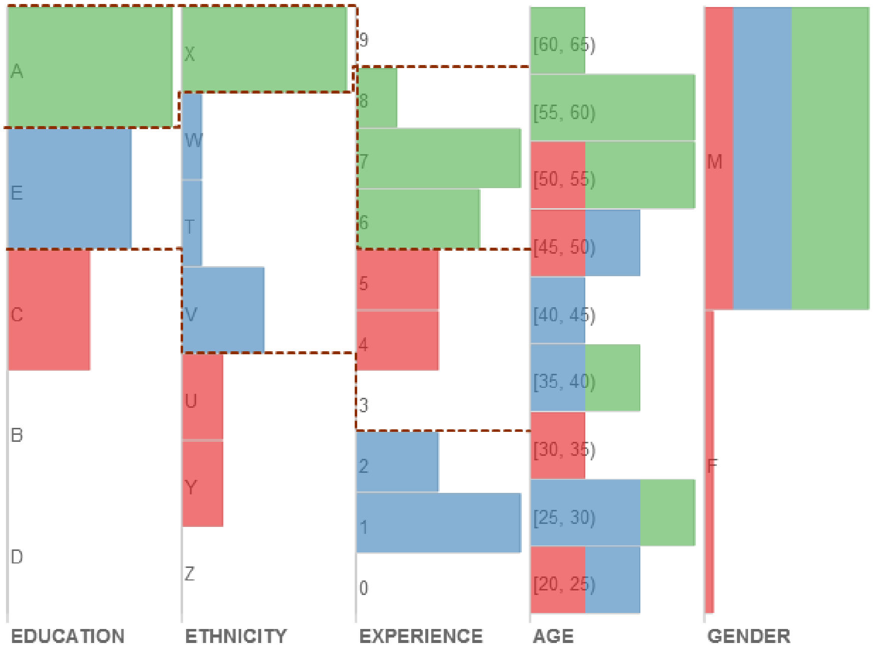


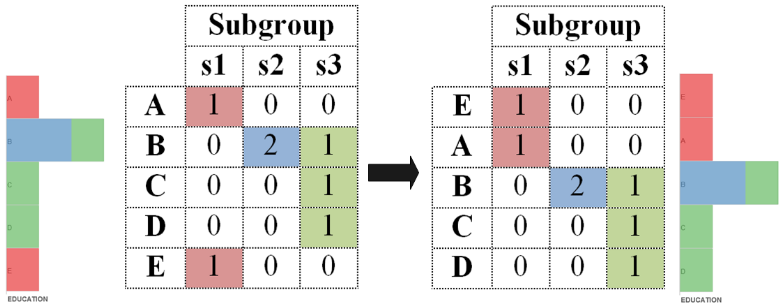
Figure 5.9: The HIST representation of the example team (Figure 5.4(a)) enhanced with connecting dashed lines to indicate the boundaries of separation among subgroups across the attributes. Within each of the nominal attributes, categories are clustered using Bertin Classification Criterion. Attribute axes are sorted using the Total Separation Criterion. The lines show that the three subgroups are totally separated along EDUCATION, ETHNICITY, and EXPERIENCE.

5.7.1 Reordering of Attribute Values

The first step is to reorder values within nominal attributes to reveal meaningful boundaries among subgroups along the corresponding axes. For each attribute X, we first construct the corresponding contingency table (or matrix), A, by subgroups. Second, we reorder attribute values or matrix rows by optimizing the Bertin Classification Criterion (BCC), as illustrated in Figure 5.10. The criterion, which is proposed by Pilhöfer et al. [125] and related to Kendall's τ [84], is an implementation of Bertin's idea that reordering of data would improve the understandability of graphical displays [12]. The goal is to minimize

$$BCC(X) = \sum_{i>i', j<j'} A_{ij} A_{i'j'} \tag{5.2}$$

where A_{ij} denotes the entry value at row i and column j and similarly, $A_{i'j'}$ the entry value at row i' and column j'. Note that optimization of BCC does not indicate whether total separation of subgroups/clusters occurs within an attribute. Also note that since we want to preserve the stacking order of subgroups (i.e., subgroup 1–red followed by subgroup 2–blue and subgroup 3–green as in Figure 5.9), this first step optimizes BCC by re-arranging attribute values only, instead of both subgroups and attribute values.



	Subgroup		
	s1	**s2**	**s3**
A	1	0	0
B	0	2	1
C	0	0	1
D	0	0	1
E	1	0	0

	Subgroup		
	s1	**s2**	**s3**
E	1	0	0
A	1	0	0
B	0	2	1
C	0	0	1
D	0	0	1

Figure 5.10: Reordering of categories in attribute EDUCATION of synthetic Team 2 (Table 5.1 and Figure 5.1) by optimizing BCC. The goal is to arrange the matrix rows to get close to a pseudo-diagonal form [125] and to reveal the boundary among subgroups.

5.7.2 Total Separation Criterion

The second step is to determine if total separation of subgroups/clusters occurs within an attribute X. Technically, if X is a *nominal* attribute, total separation occurs when each row (attribute value) of the matrix is fully contained in exactly one column (subgroup). In other words, different subgroups share no common attribute values, or

$$R(X) = \sum_{i=i', j\neq j'} A_{ij} A_{i'j'} = 0 \tag{5.3}$$

If X is a *numeric* or *ordinal* attribute, total separation of subgroups further requires that rows fully contained in one specific column must be contiguous, or $BCC(X) = 0$, assuming the ordering of subgroups (or matrix columns) are optimized (i.e., pseudo-diagonal form of the matrix [125]) . Combining the two requirements, total separation of subgroups occurs within an attribute X when

$$TSC(X) = R(X) + min(BCC(X)) = 0 \tag{5.4}$$

We refer to TSC as *Total Separation Criterion.* Its values are also used to reorder attribute axes in ascending order from left to right (Figure 5.9) before executing the faultlines drawing algorithm. Note that for the purpose of computing TSC, this step simply calculates BCC with different permutations of subgroups (or matrix columns), as opposed to actual re-arrangement of attribute values as in the first step.

5.7.3 Faultlines Drawing Algorithm

For each of the attributes with total separation of subgroups, since its values are already in optimal ordering after the first two steps, our algorithm simply traverses the values and marks the boundary between two adjacent subgroups. The traversal also *wraps around* the values to include the boundary between the two subgroups occupying the top and bottom values along the attribute axis. Finally, we draw a dashed polyline along the boundaries of the two specific subgroups across the attributes with total separation of subgroups. Note that values of nominal attributes without objects (zero-length bars) can be selectively excluded to adjoin boundaries among subgroups. We also apply a jittering technique to alleviate the possible overlap of vertical line segments (Figure 5.9). Our informal test indicates that real-time computation of the lines is reasonably fast on a typical desktop PC.

5.8 Discussion and Future Work

In this chapter, we propose, design, and evaluate visualization solutions to a new and worthwhile domain-specific problem concerning diversity faultlines in work teams. Like most studies, ours has limitations that we discuss here along with suggested directions for future work.

5.8.1 Study Design Issues

First, our study evaluated static visualizations only to first understand the merits and shortcomings of HIST, PCP, and SPLOM as *standalone representations.* Since the faultline concept is still new to end-users (e.g., managers) and no visualization solution exists, we must begin by understanding representation approaches that are linked to generic clustering. This decision was also made to keep the study implementation feasible in the online setting of mTurk. Future work will address the interactive capabilities of HIST. For example, interaction features can potentially allow users to configure their faultline requirements, such as faultline measures, attributes of interest, and rescale factors for each of the examined attributes.

Second, while we collected response time, we did not set a time limit for each question considering that the online setting of the study may be associated with more interruptions than in a lab setting. This design decision resulted in several unexpected outliers as shown in Figure 5.6(b). Nevertheless, these outliers were counter-balanced among the three visualization techniques and we applied a rank transformation [33] to the data before performing statistical tests.

Third, faultlines visualization enhancement (i.e., reordering of attribute values and drawing of connecting dashed lines) also requires formal evaluation. Early feedback from our management researcher collaborators were highly positive—they praised the enhancement for its simplicity and usefulness. However, an interesting point was suggested regarding reordering of attribute values not only in nominal attributes—as currently implemented—but also in ordinal and discretized quantitative attributes. The aim would be to make the separation of subgroups along the dashed lines more clear-cut (i.e., no crossing lines) but at the expense of losing the information on the possible distance/gaps among subgroups in ordinal and quantitative attributes. A follow-up user study of such trade-off in design choices with target users such as managers would be a potential direction for future work.

5.8.2 Limitations of HIST

Multiple histograms also have limitations. First, the technique requires a discretization of quantitative attributes. Second, since the technique treats each attribute independently, it provides limited insight into the correlation between attributes, at least with the static representation. On the other hand, PCP is well-suited to showing correlation between two neighboring attributes. To enable correlation analysis in HIST, we envision that PCP poly-lines can be selectively overlaid to allow the user to inspect the relationship among attributes as well as individual objects. Alternatively, it would be informative to consider approaches that decouple the primary faultlines/subgroups view from a relationship view where correlations are shown, for example, in a scatter plot matrix.

On a related note, implementing an interactive faultlines visualization would require efficient faultline measures as an external data clustering step. Nevertheless, to our knowledge, there are currently no well-established measures that would be scalable to large teams with multiple subgroups [150]. We suspect that modern cluster algorithms from the field of data mining such as Affinity Propagation [43] deserve further investigation for the faultline measurement challenge.

5.9 Conclusion

We present the first study exploring the design space for graphical representation of team faultlines, a fundamental construct in management that shares many characteristics with clustering in computation. In doing so, we contribute (1) the novel application and refinement of existing stacked histograms technique to the faultlines visualization, (2) a rigorous evaluation of the effectiveness of the proposed technique, (3) additional visual enhancements and metrics to further facilitate the faultlines identification tasks. To visualization researchers, the findings from our study suggest the need for revisiting cluster representations in general and investigating techniques for the important problem of faultlines in particular. To management researchers, our proposed visualization provides a useful means to conceptualize visually the output of faultlines measures, a requirement which is extremely difficult to achieve with a table-based assessment. We also hope the visualization will help bring the benefits of studying faultlines to more end-users such as managers or human resources departments.

Chapter 6: Toward Exploratory Analysis of Diversity Unified Across Fields of Study [1]

6.1 Introduction

Understanding diversity *patterns* and their causes and consequences (*processes*) is one of the greatest challenges in ecology, both at the scales of species such as plants and animals and of microorganisms (e.g., [51, 101, 112, 45]). Although this problem is shared by other disciplines, ecologists might not be fully aware of the potential improvements that could be gained by a formalized understanding of diversity studies in many arenas. For instance, while researchers and managers of human organizations may use different vocabulary, they are also concerned with diversity (e.g., [90, 55, 14]).

A common approach to understanding diversity patterns and processes is *hypothesis-driven* or *confirmatory* analysis that relies on rigorous statistical metrics and tests of data observations [101, 48, 55, 150]. These techniques may work well when the hypotheses exist, are falsifiable and testable with reasonable metrics and tests. Otherwise, the utility of the current approach diminishes quickly when the number of diversity attributes under investigation is large, multiple subsets of data are involved, and/or hypotheses are not pre-established. Still, indices of diversity have greatly dominated over more direct exploration of diversity in studies of ecology and human organizations.

Decades ago, ecology experts such as Whittaker [158], Sanders [133], and Hurlbert [72] suggested that in addition to diversity indices, ecologists should gauge diversity patterns by direct observation of data. Following this advice, visual representations of data such as histograms and rank-abundance plots [158] have been employed to communicate species variety and abundance. Nevertheless, these techniques supported limited number of variables, no interaction, and thus limited exploration capacities—perhaps due to a lack of computational interfaces and tools at that time. Experts who study human organizations have also suggested that configurations of work team structure are important and have direct consequences on team outcome processes [24]. Yet no tools exist

[1] The material in this chapter represents joint work with Julia Jones and Ronald Metoyer.

to enable direct investigation of team structure, besides text- or table-based assessment of data. In addition to the paucity of tools, discipline-specific terminologies and metrics preclude the understanding of how diversity functions and how it could be characterized similarly across disciplines.

Recently, visual analytics, "the science of analytical reasoning facilitated by interactive visual interfaces" [151], offers a new, and powerful aid to the analytical reasoning of diversity patterns and processes in complex data. By leveraging the human visual system, visual analytics—a subfield of data visualization—provides a visual gateway to the data, complementing existing diversity metrics and allowing users to explore data directly and iteratively prior to further statistical analysis (Figure 6.1). As demonstrated in previous chapters, data exploration facilitates the generation of hypotheses and insights into the data [154, 8].

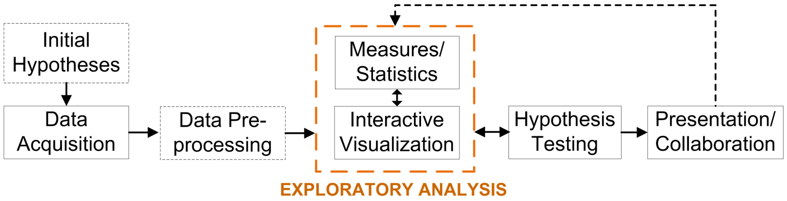


Figure 6.1: Proposed visual-analysis process of exploring diversity data. Each rectangle represents a subprocess and each arrow indicates a direction the analyst can take to go through the process. This work focuses on exploratory analysis tasks (the orange rectangle), as distinguished from data pre-processing or hypothesis testing tasks.

The visualization community has shown considerable interest in interactive visualization tools for exploring diversity in ecology and its subfield—microbial ecology. Notably, there are tools designed to facilitate understanding of (1) patterns of species distributions in separate attributes (e.g., the EcoDATE tool [121]), (2) structures of microbial populations (e.g., the MicrobiVis tool [37]), and (3) taxonomic classification and structure (e.g., the TaxonTree tool [94]). Unfortunately, these tools serve specific subsets of information needs that are somewhat separated and not transferrable from one to another. To our understanding, very little work has focused on abstracting diversity analyses from various fields to *unified analytical tasks* that target all facets of diversity in multivariate data sets. By analytical task, we mean one or a series of actions carried out by the target users on the data to fulfill an information need. Analytical tasks serve as prerequisites for designing visual-analysis tools that in turn support those tasks (Figure 6.2).

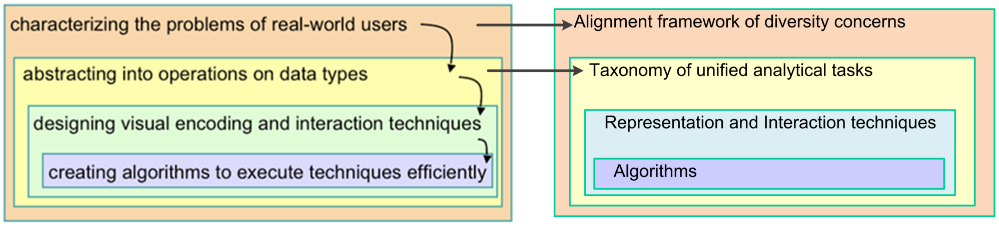

characterizing the problems of real-world users
abstracting into operations on data types
designing visual encoding and interaction techniques
creating algorithms to execute techniques efficiently
Alignment framework of diversity concerns
Taxonomy of unified analytical tasks
Representation and Interaction techniques
Algorithms

Figure 6.2: A model of visualization creation with four nested layers introduced by Munzner [109] (left) and its instantiation in the context of diversity analysis (right). This chapter emphasizes the two outer layers: (1) characterize the problem in terms of diversity concerns and information needs ("the framework") and (2) abstract the concerns into analytical tasks ("the taxonomy") that can be accomplished with visual-analysis tools.

This chapter draws upon lessons from the design of diversity visualizations in previous chapters to identify a taxonomy of analytical tasks for exploratory analysis of diversity potentially unified across fields of study (the yellow layer in Figure 6.2). In doing so, we first characterize the problem (the orange outermost layer in Figure 6.2). Specifically, we review, cross compare, and align diversity concerns across the three areas of *species diversity* (ecology), *microbial diversity* (microbiology), and *workgroup diversity* (organizational management). By *concerns*, we mean elements of diversity that can be conceptualized in a manner that transcends the three disciplines and the type of question being asked. We also illustrate these concerns with several examples of commonly used visualization techniques. The aim of the alignment framework is to set up a shared understanding between subject-matter experts and visualization researchers in terms of common diversity-related vocabulary and design considerations. We then translate these concerns into analytical tasks that are well defined by existing generic task taxonomies in visual analytics (e.g., [5, 8]). Simply put, while the diversity concerns are the vocabulary of subject-matter experts that represent their information needs and transcend across fields, the analytical tasks are the vocabulary of computer science, or more specifically, of visual analytics that represent user requirements that can be met by design of visual-analysis tools.

Our results aim to benefit various users. Subject-matter experts can cross compare diversity concerns and scientific findings as well as adopt analytical tasks and visualization techniques. Further, visualization designers and researchers have common vocabulary and abstractions for designing and evaluating different diversity visual-analysis tools. Finally, we are aware that the proposed framework and taxonomy are by no means comprehensive considering the complexity of ecological and human systems and their interactions. Therefore, we expect this work will stimulate further discussions regarding validation and improvement to both the framework and the taxonomy.

6.2 Alignment of Diversity Concerns

Here we discuss a framework for aligning diversity concerns ("the framework") across the analyses of species diversity (ecology), microbial/genomic diversity (microbial ecology), and workgroup diversity (organizational management). By framework, we mean a set of thoughts, theories, and approaches that are accepted by subject-matter experts as the guiding principles for characterizing the problem. The concerns of interest include (1) characteristics of diversity data, (2) description of diversity patterns, and (3) hypotheses regarding the causes and consequences of diversity (processes). The framework is summarized in Table 6.1.

6.2.1 Data Characteristics

Ecologists typically make a distinction between two types of phenomenon concerning diversity: (1) the description of diversity (*diversity patterns*) and (2) the causes and consequences of diversity (*diversity processes*) [101]. To understand these phenomena, a common approach is to undertake scientific studies. Specifically, experts collect data and make inferences about the underlying phenomena based on *data behaviors* (or data patterns). Data behavior is defined as a set of inherent features specific to a (sub)set of data observations considered *as a whole* as opposed to individual observations [8]. For instance, a data behavior may manifest itself as notions of distributions, clusters, or trends. Simply put, analysts characterize behaviors of their collected data to understand underlying phenomena. This approach of analysis follows inductive reasoning, as distinguished from deductive reasoning such as forming hypotheses based on theoretical models [48], which is outside the scope of this work.

Table 6.1: Alignment of diversity concerns across the analyses of species diversity (ecology), microbial diversity (microbial ecology and microbiology), and workgroup diversity (organizational management). Note that cells marked with "–" indicate missing concerns that may not yet be studied in the corresponding fields. The last column suggests how the data behavior for each of the concerns (if applicable) should be characterized.

	Species Diversity	**Microbial/Genomic Diversity**	**Workgroup Diversity**	**Data Behavior Characterization**
Typical Unit of Study	Community (α-diversity)	Microbe Sample (α-diversity)	Work team	N/A
Typical Unit of Observation	Individual of known species or biomass	OTUs with abundance (classified from microbe sample)	Individual person	N/A
Diversity Components concerning Separate Attributes	Variety and Abundance	Variety and Abundance	Variety	Distributions Metrics
	Niche Separation	–	Separation	
	Dominance/Rarity	Dominance/Rarity	Disparity	
Diversity Components concerning Interactions among Attributes	Functional Diversity	Functional Diversity	Faultines/Subgroups	Distributions Clusters Metrics
	Taxonomic Diversity	Taxonomic Diversity	–	Distributions Clusters Hierarchies Metrics
Diversity in Space and Time	β-diversity or turnover; γ-diversity	β-diversity or turnover	Between-unit diversity; Macro-faultlines	Spatial & Temporal Characterization Metrics
Diversity as Responder (Cause of Diversity)	Landscape patterns (Climate, Disturbance, Land Use)	Environmental patterns or Biological patterns (Human body)	Organizational factors (e.g., culture, recruitment)	Correlations/ Regressions Metrics
Diversity as Driver or Moderator (Effect of Diversity)	Ecosystem functions and processes	Eco or Human system functions and processes	Workgroup functions and outcomes	

Diversity data are samples of independent *observations* collected from the population of interest within one or multiple *units of study* (Table 6.1). For example, in workgroup diversity, a work team represents a typical unit of study while an individual person represents a unit of observation (or measurement) [55, 150]. Comparatively, in species diversity, a typical unit of observation is an individual of known species such as animals and plants collected in a community or assemblage [101]. On the other hand, a typical unit of study of microbial community diversity is a biological sample (i.e., biological specimen) that can be classified into various Operational Taxonomic Units (OTUs), a close approximation to microbial species (as opposed to plant or animal species) with corresponding abundances [112, 45]. The identification of OTUs is performed by extracting and sequencing DNA from the biological sample [45].

Each unit of observation may be characterized by multiple mix-typed and, in some cases, hierarchical characteristics (*attributes*) necessary for gauging diversity of the corresponding unit of study and its role in the examined ecological or human system. For instance, a team member may be characterized by multiple demographics and non-demographics attributes; an individual of a known species may be described by multiple known characteristics (e.g. size, food type) and hierarchical levels of Linnaean taxonomy (e.g., family, genus, and species). In addition, observations can be collected in space and time (*independent variables*) and associated with additional process (cause and effect) factors (e.g. team performance or ecosystem functions). In essence, diversity data sets are mix-typed, multivariate, and in many cases, hierarchical, spatiotemporal, and large (thousands of records/observations).

6.2.2 Diversity Patterns

Diversity patterns are an overarching concept that includes various and related components adopted by the three areas of interest but usually under slightly different terms, especially between species/microbial diversity and workgroup diversity. The components can be loosely classified based on the ideas that (1) diversity is *attribute-specific*—that is, attributes are not treated as equal and (2) one or multiple diversity attributes can be investigated either *separately* (i.e., one by one) or *simultaneously* (Table 6.1, Diversity Components concerning Separate Attributes vs. Diversity Components concerning Interactions among Attributes) [90, 101, 55]. Here we aim to (1) briefly describe the common components, (2) demonstrate how they could be aligned across the three examined areas, and (3) more importantly, characterize each of the components with corresponding data behaviors of interest to analysts.

6.2.2.1 Diversity Patterns concerning Separate Attributes

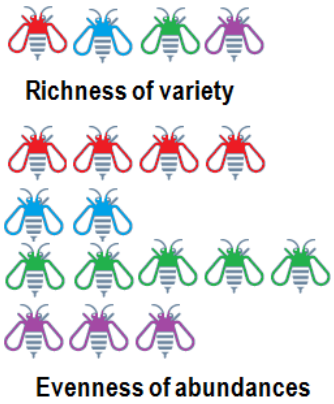


Figure 6.3: Illustration of species richness and evenness. Each icon represents an individual of a known species (e.g., insects). Species richness refers to the number of different species represented in a unit of study and species evenness concerns how close in abundances each species in a unit of study is.

First, consider the diversity patterns in separate attributes–for example, investigation of biodiversity at species level only. In this regard, species diversity (or α-diversity) is "the variety and abundance of species in a defined unit of study", as defined by Magurran [101]. The definition emphasizes the two main components and corresponding metrics of *richness of variety* and *evenness of abundance* of species (Figure 6.3). Similarly but at genomic level, microbial community diversity also concerns variety and abundance of microorganisms in a community. With respect to data behaviors, in addition to diversity metrics, richness of variety and evenness of abundance are typically characterized by *shapes of distribution*, as depicted by a rank-abundance curve (Figure 6.4).

When space and time are involved, ecologists introduce additional diversity components and related metrics [159]. α-diversity refers to species diversity within a particular community or site as defined earlier [101]. β-, and γ-diversity correspond to variation in

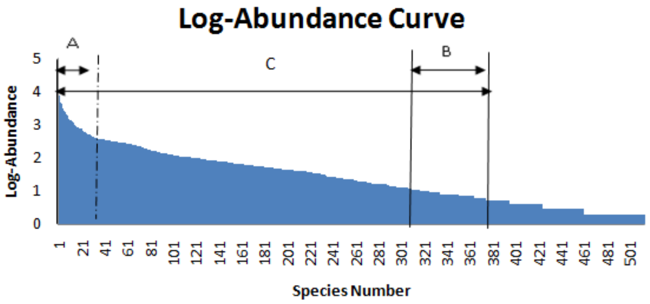


Figure 6.4: Rank abundance curve (with logarithmic scale) showing the evenness of moth species in the moth dataset [106]. 'A' shows the common moths, 'B' shows the rare moths, and 'C' shows the common through rare moths. The technique is a variation of the histogram in which species are ranked from most to least abundant and then plotted along the x-axis. The technique is limited to a single attribute. Image taken from Pham et al. [120].

species composition from one site to the others or from time to time (i.e., turnover) and to the diversity at the landscape scale, respectively [159].

On the other hand, experts studying human organizations describe workgroup diversity in separate attributes as "the distribution of differences among the members of a unit with respect to a common attribute, X, such as tenure, ethnicity, conscientiousness, task attitude, or pay" [55]. Similar to the definition of species diversity, this definition is also centered on the generalization of *diversity as distributions.*

Besides, workgroup diversity is explicitly *attribute-specific.* Depending on the attributes under investigation, the experts conceptualize diversity not only as *variety* but also as *separation* and *disparity*, as introduced by Harrison and Klein [55]. Variety represents differences in kind or category (e.g., different skill sets) and reflects information in the unit. Separation represents differences in position or opinion and is considered a horizontal difference between members of a unit. For instance, different cultural values of members represent team separation. Disparity represents differences in concentration of valued social assets or resources and is considered a vertical difference between members of a unit. For example, difference in pay among members may create disparity in a team. Disparity thus reflects differences in possession. Note that several characteristics, specifically the demographic attributes of age, gender, race, and tenure, can represent more than one type of diversity. For instance, 'age' may indicate *variety* in one case (age comes with experience) and *separation* in other cases (age represents generation gaps).

While these diversity types have different names conceptually, from an analysis point of view, their patterns differ only in the *shapes of the distribution* of interest for minimum, moderate, and maximum diversity (Figure 6.5) [55]. Consider maximum diversity, for example. Maximum variety occurs when each of the possible values of an attribute (e.g., skill sets) is evenly represented; maximum separation is depicted by a bimodal distribution with half of the unit at each of the extremes of the attributes of interest (e.g., cultural values); maximum disparity, on the other hand, is depicted by a skewed distribution with one member at the high end of the vertical scale and all other members at the low end (e.g., pay differences). These shapes of distributions are in turn empirically associated with different outcomes for the examined unit of study [55].

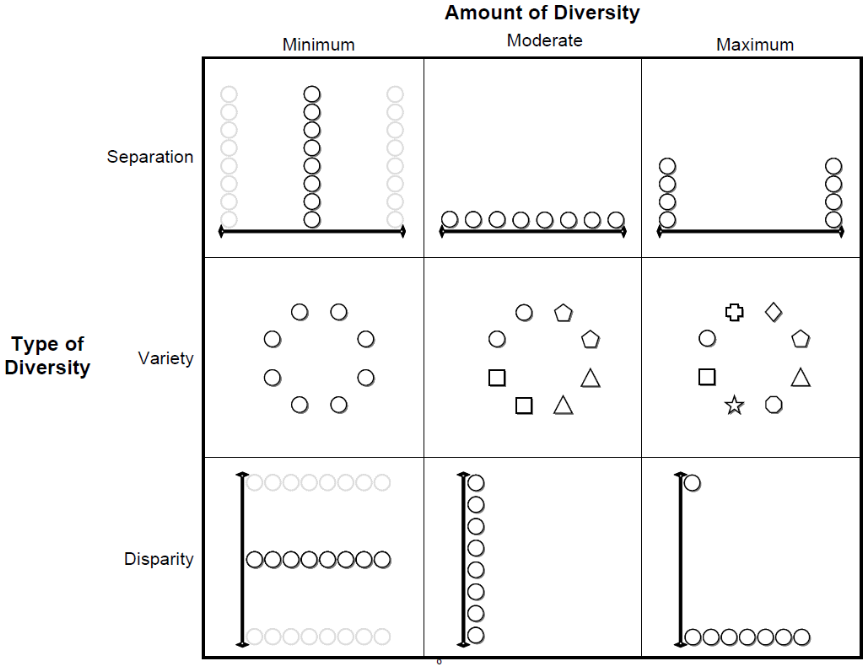


Figure 6.5: Illustration of the three types of diversity within work teams and the corresponding shapes of distributions for the three levels of diversity: minimum, moderate, and maximum. Each of the icons represents a team member. Image reused with permission from Harrison and Klein [55]. ©2007, Academy of Management.

While ecologists do not explicitly discuss separation and disparity, they do discuss *species dominance* and *niche separation*, which correspond well to disparity and separation distributions in management, if these components are considered in separate attributes. Specifically, species dominance refers to the degree to which a species is more numerous than others are or makes up (or possesses) more of the biomass, thus representing a vertical difference in makeup (as in disparity) [11, 47, 46]. Niche separation is the process of naturally partitioning competing species into different patterns of resource use or different niches so that they do not out-compete each other [92]. As an example, food type of animals could be considered as a separation attribute—carnivore (meat eater) and herbivore (plant eater) may represent two extreme ends of the food type spectrum. To some extent, this concept is comparable to separation in organizational management, which is a horizontal difference in makeup.

In all, we argue that when diversity is considered in separate attributes, the concept of species diversity matches well with that of workgroup diversity in which team members equate to individuals of species (or their equivalents such as OTUs). These components are centered on the generalization of *diversity as distributions.* Furthermore, it is important that the analysts choose the correct conceptualization (e.g. type of diversity) and apply the appropriate data characterization (e.g. statistical metrics or shapes of distribution). To summarize, we propose the following consideration for characterizing data behavior of diversity patterns in separate attributes:

> ***Data Behavior Characterization - Consideration 1.*** *From an analysis point of view, when diversity patterns are considered in separate attributes, depending on the types of diversity under consideration (e.g., variety, separation, and disparity), the corresponding data behaviors are typically characterized by the* ***shapes of distributions*** *of observations in separate attributes, in addition to summary statistics such as diversity metrics. If time and space are involved, the data behavior should also consider how the distribution patterns and summary statistics vary over time and space.*

To demonstrate how this consideration may benefit design of interactive visualization techniques, consider the example visualization in Figure 6.6. The figure depicts the multiple histogram representation of the moth diversity and abundance data set supported by the EcoDATE tool [121]. Consideration 1 emphasizes the characterization of data behavior as shapes of distributions in separate attributes. According to information visualization design principles [99], a histogram is well suited to showing the distribution of objects within an attribute. Further, placing histograms vertically side-by-side in parallel [74] aims to convey a holistic object distribution over multiple attributes. Finally, the characterization of distributions (Consideration 1) is further assisted by interaction features. For instance, users can sort bins within a histogram by abundances to form the rank-abundance curve (e.g., green histogram LEP_NAME); annotate histograms with different colors to distinguish attributes of different diversity types (i.e., variety, separation, and disparity); subset data by time (COLLECT_YEAR) or space (TRAP_ID) to see how distribution patterns vary over time and space.

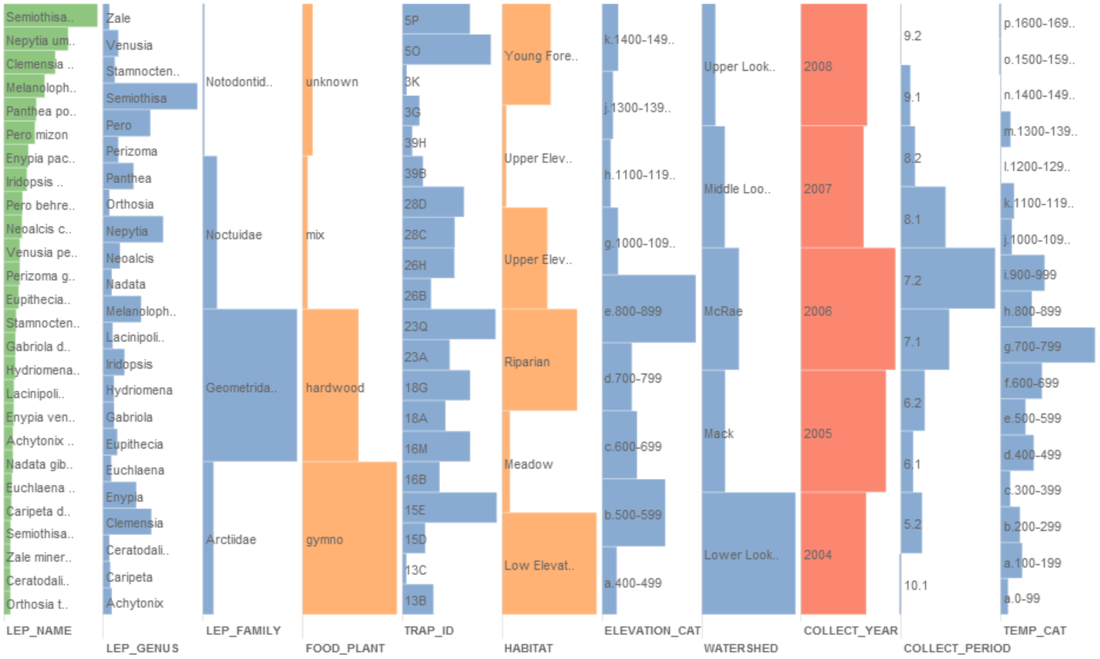


Figure 6.6: The multiple histogram representation of common moths. The visualized attributes from left to right are LEP NAME (moth scientific name including genus and species), LEP GENUS, LEP FAMILY, FOOD PLANT, TRAP ID, HABITAT, ELEVATION, WATERSHED, COLLECT YEAR, COLLECT PERIOD, TEMPERATURE. Note that LEP is short for *Lepidoptera* (moth). In each of the histograms, the bars are pointing to the right (in contrast to the familiar upward-pointing display). The structure of the moth data set is described in [120]. The interactive version of the visualization is available at `http://purl.oclc.org/ecodate/commonmoth`.

6.2.2.2 Diversity Patterns concerning Interactions among Multiple Attributes

Diversity definitions that look at the diversity of each attribute separately have a limitation. They do not take into account the *interaction among attributes.* Consider an example of two teams of employees that have four members each in Table 6.2. While it is obvious that Team 2 is divided into more subgroups, the current definition concludes that both teams are at the same level of overall diversity with respect to gender and age—that is, in each of the two teams, members are uniformly distributed in both gender and age. To address this limitation, here we discuss diversity patterns that consider interactions among multiple attributes. In this regard, we also find parallel components across the three areas (Table 6.1).

Table 6.2: Employee Diversity Example. Each of the two teams has four members.

Team 1	
Female, over 50	Male, under 50
Female, over 50	Male, under 50

Team 2	
Female, over 50	Male, over 50
Female, under 50	Male, under 50

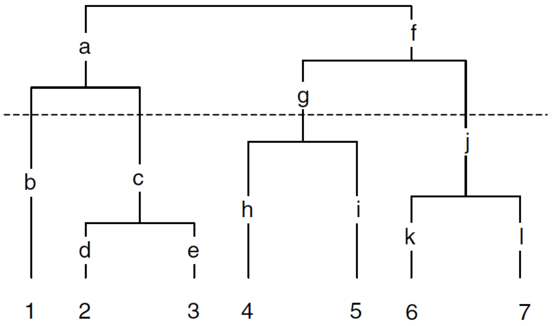


Figure 6.7: A dendrogram representation demonstrating how seven species 1-7 are assigned to four functional groups based on hierarchical clustering of the species across multiple functional traits. The four functional groups include {1}, {2, 3}, {4, 5}, {6, 7}. The dashed line indicates an arbitrary stopping condition for the clustering process. Image reused with permission from Petchey and Gaston [116]. Copyright ©2002, John Wiley and Sons.

On the one hand, ecologists and microbiologists recognize *functional diversity* as variety of roles played by different species (or their equivalents) based on their composition of multiple functional traits (e.g., rooting depth and maximum growth rate of plants) [116]. Technically, composition of these traits can be used to *cluster* different species (or their equivalents) present in a unit of study into different *functional groups* and to derive, for example, the functional diversity (FD) metric [116] (Figure 6.7). There are potentially many ways to compute similarity or dissimilarity among objects or among variables. Therefore, it is important to choose appropriate and ecologically meaningful clustering algorithms or to allow analysts to iteratively explore them (Figure 6.1). Ramette [128] provides an in-depth review of cluster analysis techniques for microbial diversity data.

Furthermore, species and OTUs are inherently *hierarchical*—that is, species are grouped into taxa. Therefore, traits or attributes under investigation might be extended to taxonomic organization such as *species*, *genus*, and *family*, resulting in *taxonomic diversity* (diversity across taxa) and corresponding metrics such as taxonomic distinctness [157]. Figure 6.8 illustrates an example of two hypothetical units of study whose diversity levels are determined by not only species level but also as composition of higher taxa. From an analysis perspective, *hierarchy* of different species present is the primary data behavior of interest for taxonomic diversity.

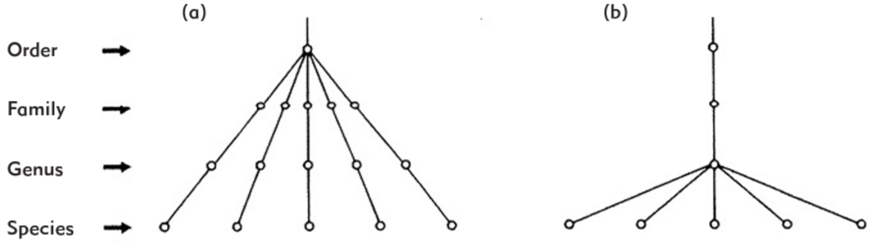


Figure 6.8: A node-link diagram (tree) representation of two hypothetical units of study (e.g., assemblages) with the same level of species richness (i.e., five species represented) but different levels of taxonomic diversity when higher taxa such as genus and family are considered; unit of study (a) is more diverse than unit of study (b). Image reused with permission from Magurran [101]. Copyright ©2003, John Wiley and Sons.

It is important to note that just as diversity patterns in separate attributes, functional diversity and taxonomic diversity also concern richness of variety and evenness of abundances within or between clusters (e.g., functional groups) [117], making the generalization of *diversity as distributions* still applicable. As an example, while the dendrogram alone in Figure 6.7 does not consider evenness of observations in each of the four functional groups, a heatmap is commonly used along with dendrogram to communicate evenness of abundances (Figure 6.9).

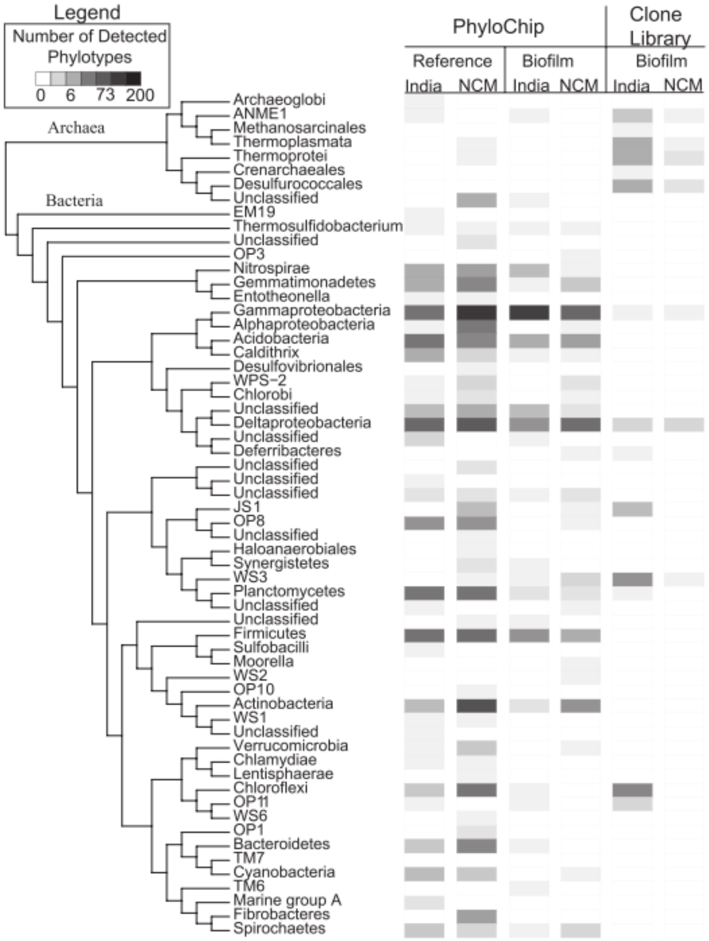
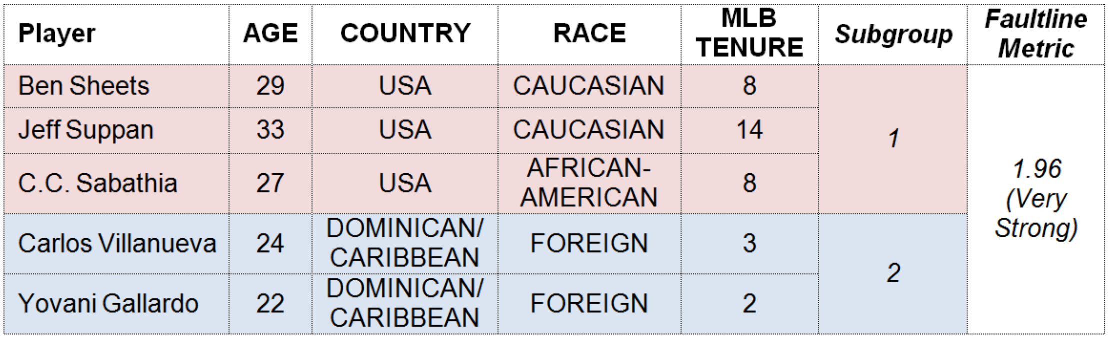


Figure 6.9: A hybrid representation of dendrogram and heatmap used to depict the taxonomic diversity of archael and bacteria phyla along with corresponding abundances detected in several samples of a microbial diversity study. Image reused with permission from Briggs et al. [18]. Copyright ©2011, American Society for Microbiology.

On the other hand, in organizational management, the *diversity faultlines* concept, which is also adopted from *multivariate clustering*, concerns *subgroups* or *clusters* formed in a work team based on alignment (or composition) of multiple demographic or non-demographic characteristics of members, as first introduced by Lau and Murnighan [90]. Figure 6.10 depicts an example of how the faultlines concept is applied to a work team. Just as ecologists studying functional diversity, management experts are also interested in (1) structure of subgroups with respect to the number of subgroups, evenness of subgroups, and subgroup variety and abundance; and (2) faultlines or attributes in which subgroups are separable or far apart from each other [14, 24, 123]. Note that subgroups in workgroup diversity, functional groups in functional diversity, and clusters in computation are now considered equivalent in the framework. Surveys of various faultline concerns and metrics can be found in [150] and [104].



Player	AGE	COUNTRY	RACE	MLB TENURE	*Subgroup*	*Faultline Metric*
Ben Sheets	29	USA	CAUCASIAN	8	*1*	*1.96 (Very Strong)*
Jeff Suppan	33	USA	CAUCASIAN	14		
C.C. Sabathia	27	USA	AFRICAN-AMERICAN	8		
Carlos Villanueva	24	DOMINICAN/ CARIBBEAN	FOREIGN	3	*2*	
Yovani Gallardo	22	DOMINICAN/ CARIBBEAN	FOREIGN	2		

Figure 6.10: An example of how a faultline metric [14] is used to cluster a group of starting pitchers of the MLB team Brewers in 2008 into two subgroups (subgroup 1 and subgroup 2) based on the similarity of group members across the attributes of interest: AGE, COUNTRY (of origin), RACE, and MLB TENURE (in years). The table does not clearly show how the subgroups (or clusters) are separable or far apart across the attributes under investigation. Figure 6.11 depicts a multivariate visualization technique that addresses this issue. Data courtesy Katerina Bezrukova and Chester Spell.

In all, we argue that the concept of faultlines in organizational management could be matched with that of functional diversity in ecology from an analysis perspective. Both are adopted from *multivariate cluster analysis*. Therefore, appropriate operationalizations of the concepts in terms of diversity metrics or data behaviors of interest could potentially be exchangeable. We summarize a consideration for characterizing data behavior of diversity patterns concerning interactions among multiple attributes as follows:

> ***Data Behavior Characterization - Consideration 2.*** *From an analysis point of view, when diversity patterns involve interaction among multiple attributes simultaneously, the corresponding data behaviors are typically characterized by the* ***shapes of distributions*** *of observations that are grouped into* ***clusters*** *across multiple attributes, in addition to summary statistics such as diversity metrics. Clusters may represent functional groups in an ecological unit of study (e.g., communities) or subgroups in an organizational unit of study (e.g., work teams); clusters may also represent different units of study under comparison. In addition, in some cases, the data behavior of interest is the* ***hierarchical relationships*** *if the patterns of interest concern taxonomic organization (e.g., taxonomic diversity) or hierarchical clustering. Further, if time and space are involved, the data behavior should also consider how these distributions, clusters, and/or hierarchies as well as corresponding summary statistics vary over time and space.*

Figure 6.11 and Figure 6.12 show examples of visual representation design of diversity faultlines that followed closely Consideration 2 [122, 123]. First, to convey distributions of observations across attributes, the design reuses multiple histograms (Figure 6.6). Then, since subgroups (or clusters) are nested within a team, to maintain bar length encoding, a natural solution to encoding subgroups is to stack bars within each bin (Figure 6.12). Distinct color hues on a white background are used to differentiate stacked subgroups. Following this design, structure of each of the subgroups is conveyed across attributes. In addition, a total separation of subgroups occurs at a nominal attribute when distinct subgroups (or distinct colors) occupy distinct positions along the vertical axis. Total separation at a numeric or ordinal attribute further requires that these distinct positions—including ones without objects (zero-length bars)—are contiguous. The visual representation in Figure 6.11 makes it obvious that the two subgroups formed in a group of baseball players are totally separated apart in all four attributes under investigation; the team visualized in Figure 6.12 represents a less extreme example of faultline separation in a work team: the three subgroups are separated along several attributes but the members overlap in other attributes.

To further show that the diversity faultlines concept in management corresponds well to functional diversity in ecology, we also apply the multiple linked stacked histograms

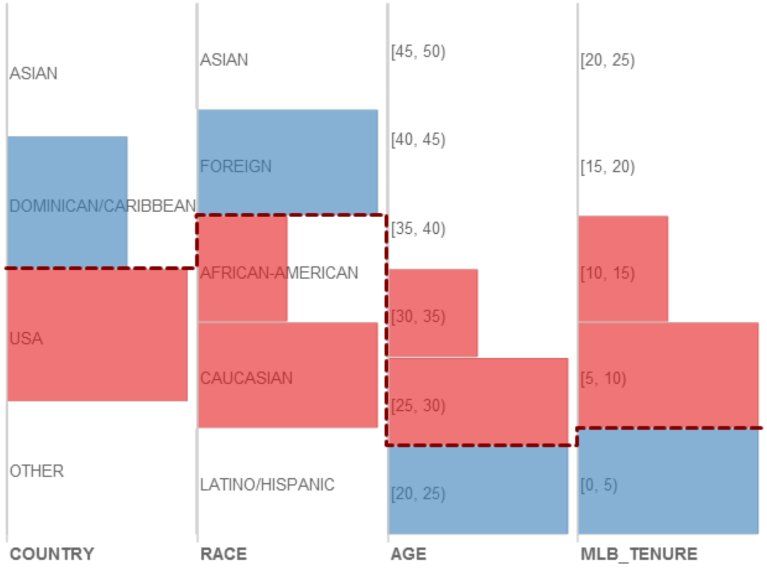


Figure 6.11: Group of starting pitchers of the MLB team Brewers in 2008 (Figure 6.10) visualized using multiple linked stacked histograms (HIST). The two subgroups (subgroup 1 and subgroup 2) are totally divided in all four attributes of COUNTRY, RACE, AGE, and MLB TENURE. The connecting dashed lines are overlaid to represent the holistic separation between the two subgroups.

representation (HIST) to visualizing the two groups of common moths and rare moths from the moth data set (Figure 6.13). To some extent, the two groups are equivalent to functional groups in functional diversity. The results are encouraging. The visualization provides insights into the separation between the two groups with respect to species, genus, and family as well as food plant: common moths are mostly conifer-feeders and rare moths are mostly hardwood, herb, and grass-feeders (Figure 6.13).

6.2.3 Diversity Processes

Thus far, we have focused on diversity patterns, however these patterns are causally associated with other phenomena in the system under investigation. Across the three fields, we can find parallels in the roles of diversity as *responder* (cause), *driver* (effect), or *moderator* (effect). For instance, ecologists refer to positive effects of diversity such as sustainability and resilience in an ecological system (e.g., [51, 11]) while organizational

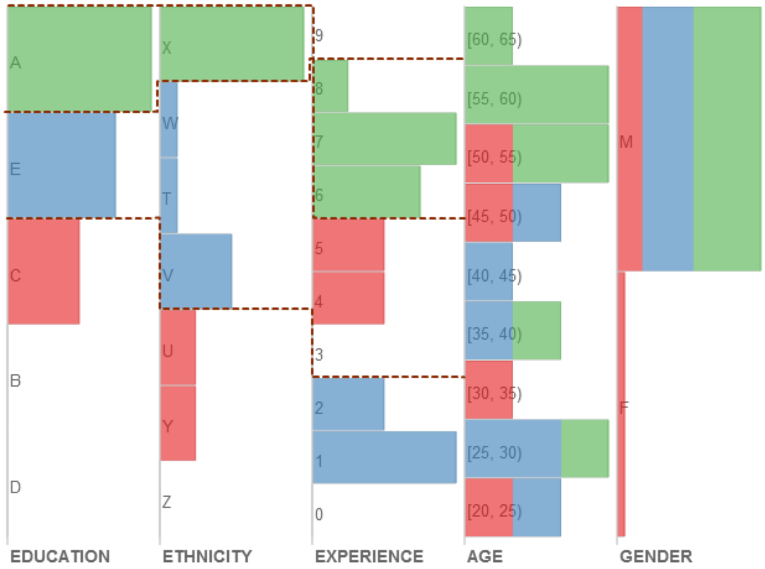


Figure 6.12: Example team of size 18 visualized using HIST. Distinct colors are used to differentiate the three subgroups: subgroup 1, subgroup 2, and subgroup 3. While subgroup 3 is the biggest, subgroup 1 is the smallest. The three subgroups are totally separated along ETHNICITY, EDUCATION, and EXPERIENCE because different subgroups occupy different subsets of values along these attributes. The three subgroups overlap in GENDER and AGE because there exist values of these attributes shared by different subgroups.

management experts seek innovation and flexibility, just to name a few (e.g., [55, 155, 105]). The causes of diversity in ecology are related to climate, disturbance, and land use while in organizational management they are organizational factors such as culture or recruitment.

During the data exploration process involving the analyst's knowledge, the analyst may be able to make inferences about the diversity processes from direct observation of diversity patterns considering that the causal links are well understood [8, 37, 120]. For example, ecologists found that the richness of species tends to be higher in lower latitudes than in higher latitudes [69]. In another example, it is widely accepted in organizational management that high variety of expertise in a team results in greater creativity and innovation [55]. Nevertheless, based on information needs of users and availability of process data (e.g., environmental factors or performance), visual analysis tools may

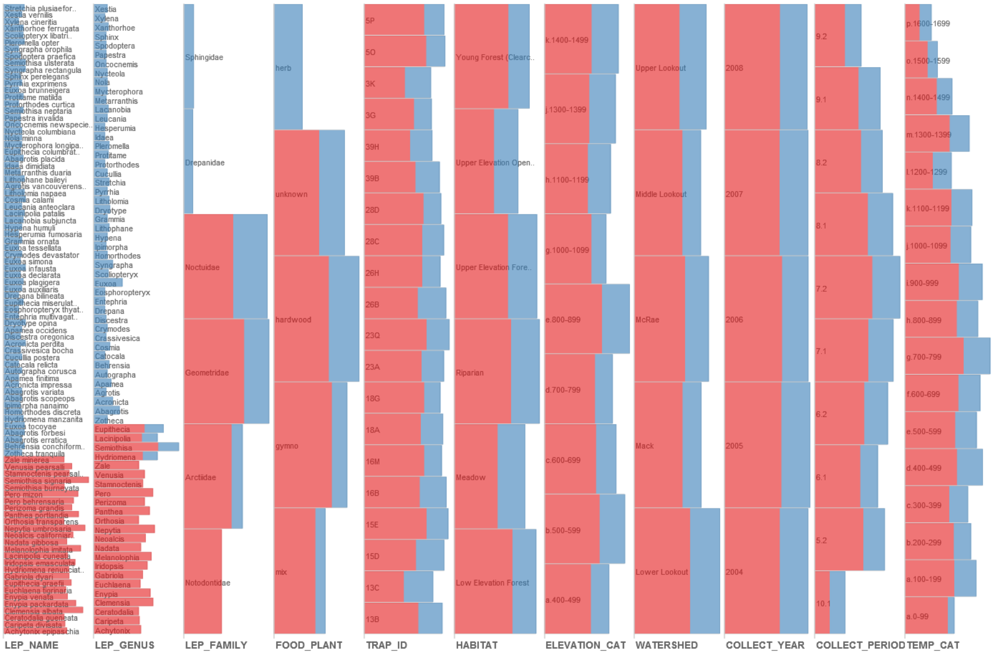


Figure 6.13: Two groups (or clusters) of common moths and rare moths visualized using HIST. Since the common moths are much more abundant than the rare moths, the length of each bar is scaled according the logarithm with base 10. The view suggests the two groups are far apart with respect to species, genus, and family as well as food plant—attribute axes 1-4 from left to right. However, the two groups overlap in the other attributes. The structure of the moth data set is described in [120].

support users in examining the associations between observed diversity patterns and system processes directly via *correlation and regression analyses.* Note that regression and correlation indicate only how or to what extent data variables are associated with each other. To make conclusions about the causal relationships, analysists may need to involve their domain knowledge. We introduce another consideration for characterizing data behavior of diversity processes as follows:

> ***Data Behavior Characterization - Consideration 3.*** *From an analysis point of view, the data behaviors of diversity processes are typically characterized by how diversity patterns and system processes are* ***correlated****, if the observed diversity patterns are investigated as a driver or responder; or by*

how diversity patterns ***moderate*** *correlations between system processes, if the observed diversity patterns are investigated as a moderator. If time and space are involved, the data behavior should also consider how these correlations/regressions vary over time and space. Note that following Consideration 1 and 2, diversity patterns and system processes may be characterized by corresponding data behaviors (e.g., distributions, clusters, hierarchies) or summary statistics (e.g., diversity metrics).*

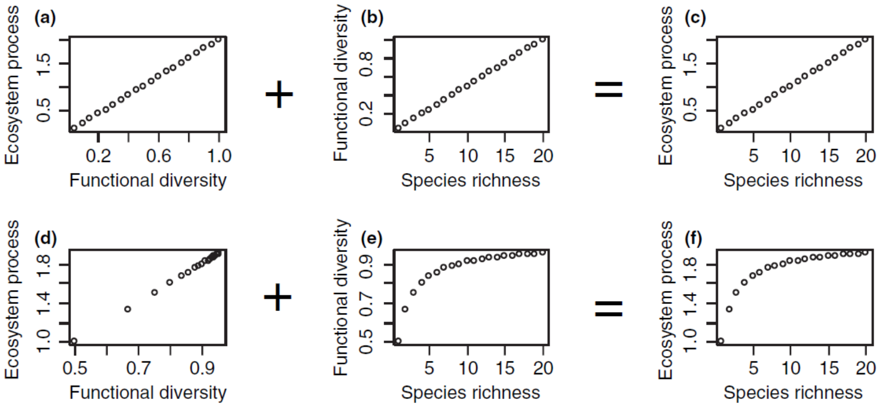


Figure 6.14: An example of scatter plots used to depict possible relationships between species richness, functional diversity metric, and ecosystem processes. Data points may represent unique units of study or a unit of study repeatedly measured over time. Image reused with permission from Petchey and Gaston [117]. Copyright ©2006, John Wiley and Sons.

To demonstrate the relevance of this consideration to designing visual representations, we first present two examples from ecology and organizational management. According to information visualization guidelines, scatter plot and line chart are effective for communicating relationships between two variables [136]. In Figure 6.14, scatter plots are used to demonstrate possible *correlations* between measures of species richness, functional diversity, and ecosystem processes (e.g., retention of nitrogen, total aboveground biomass). In Figure 6.15, a line chart is used to depict the role of diversity faultlines as “moderator”. These two static examples, which are taken from research papers, serve the primary purpose of *explaining* the correlations and regressions found from hypothesis testing. Nevertheless, the techniques, if equipped with appropriate interaction features

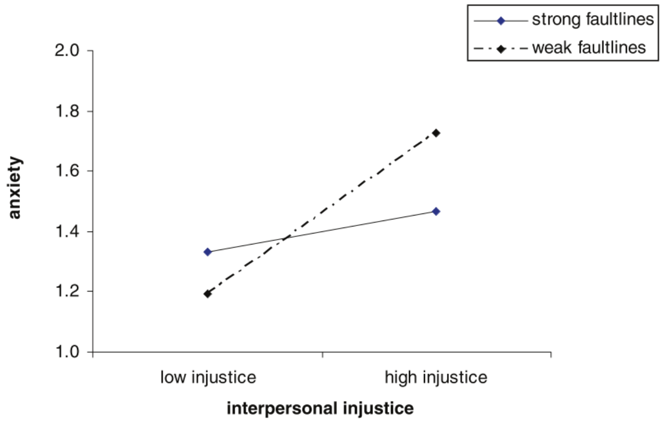


Figure 6.15: An example of a line chart used to depict the role of diversity faultline as "moderator": psychological distress of team members was positively related to their perceived injustice in the team (the dashed line); strong group faultlines weakened that positive relationship (the solid line). Image reused with permission from Bezrukova et al. [15]. Copyright ©2010 Wiley Periodicals, Inc.

such as highlight/select and filter/subset [65], would be still applicable for enabling *exploration* of the correlations. On a related note, in both examples, the examined diversity patterns and system processes are quantified by summary statistics, as opposed to more descriptive data behaviors such as distributions or clusters, which we demonstrate in the next example.

Reusing the moth data set, the next example shows how investigation of the relationships among distribution patterns (Consideration 3) is facilitated by multiple linked histograms augmented with interaction features, especially data filtering (Figure 6.16). Ecologists interact with the visualization to inspect the effect of temperature on the emergence patterns of common moth (diversity as responder). By filtering the moth records by COLLECT_YEAR, ecologists observed that common moths were captured in a much more concentrated time span in a warm year such as 2004 than in a cold year such as 2008. In this example, while ecologists need to observe only three attributes (COLLECT YEAR, COLLECT PERIOD, and TEMPERATURE) to discover their relationship, they can potentially relate other attributes for additional insights. For instance, they may initially pre-define the ordering of moth species in LEP_NAME attribute (e.g., by abundance) and then quickly verify whether the ordering pattern remains consistent

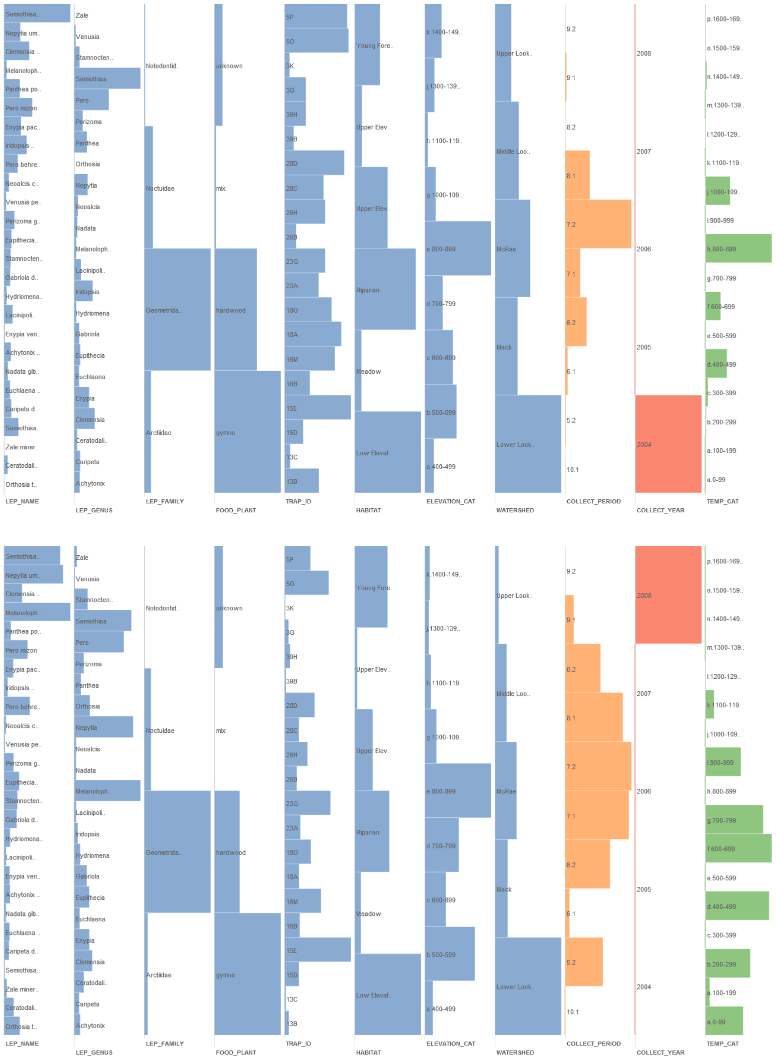


Figure 6.16: The multiple histogram representation of common moths sampled in COLLECT YEAR of ‘2004’ (top) and of ‘2008’ (bottom). The views allow exploration of the effect of TEMPERATURE on the sampling distribution of common moth (COLLECT PERIOD): common moths were captured in a much more narrow time span in 2004 (warm year) than in 2008 (cold year)

over these two years in response to temperature patterns.

6.2.4 Summary of the Diversity Concerns

We describe a variety of diversity concerns that represent information needs aligned across the three areas. In summary, exploratory analysis of diversity patterns aims to reveal the structure of multivariate objects of interest (e.g., species individuals, team members) in units of study of interest (e.g., communities, work teams). Such structure may manifest itself in the observed data as distributions, clusters, and/or hierarchies (Considerations 1 and 2). Exploration of diversity processes concerns the existence of the causal relationships between the diversity patterns and system processes. Such relationships are typically characterized by correlations and regressions among values of corresponding data variables (Consideration 3).

Based on research questions of interest, it becomes very important that experts choose the correct conceptualization, such as diversity concerns, and apply the appropriate operationalization such as statistical metrics or visual representations of data behaviors. Therefore, we accompany each of the diversity concerns, if applicable, with several examples of appropriate visualizations. By discussing these examples, we wish to emphasize how visualization design should be guided by the information needs of users that can be abstracted into corresponding data behaviors.

Motivations for the taxonomy of analytical tasks. Thus far, the information needs are described mostly in the vocabulary of experts and do not yet illuminate the *analytical tasks* or *processes* that fulfill those needs. For instance, investigation of diversity patterns typically precedes that of diversity processes. Understanding these tasks and processes is extremely important to designers of visualization tools. The process could be iterative and exploratory considering that multiple configurations of diversity may be involved and affect the results of the investigation (Figure 6.1). For example, ecologists may wish to explore taxonomic hierarchies of species observations after finding the two units of study are of the same level of diversity in terms of species richness [101] (see Figure 6.8). Similarly, ecologists may wish to experiment with different combinations of functional traits when investigating functional diversity. In another example, management researchers may wish to conceptualize the ‘age’ attribute as variety in one case and as separation in other cases. While the operationalization in terms of statistics is outside the scope of this work, the choice of analytical tasks that could be accomplished with corresponding visual representation and interaction techniques are certainly operationalizations and must be chosen carefully. Next, we review existing generic taxonomies of analytical tasks and introduce a specific task taxonomy for diversity analysis unified across ecology and organizational management.

6.3 Assessment of Existing Generic Taxonomies of Analytical Tasks

Design of our taxonomy of analytical tasks was informed by existing generic task taxonomies in the field of visual analytics. In this section, we assess applicability of a subset of relevant taxonomies to diversity analysis. More thorough reviews of existing task taxonomies can be found in [5] and [8].

To guide the design of information visualization tools, Shneiderman [141] proposed the now well-known visual information seeking mantra "overview first, zoom and filter, then details on demand" followed by a classification of corresponding analytical tasks. The mantra and tasks are potentially useful to guide analysis strategies. Nevertheless, the tasks are somewhat driven by the tool capabilities (e.g., support of zoom and filter features) and there are no explicit mappings between the tasks and specific information needs in the context of diversity studies (e.g., what is the purpose of overview or filter?).

Following a different approach based on user analytical activities when using visualization tools, Amar et al. [5] introduced a taxonomy of ten *low-level* tasks (see Table 6.3). Applied to diversity analysis, these tasks, while not necessarily comprehensive, are relevant as building blocks since they aim to capture primitive analytical operations (e.g., filter/subset, sort, characterize distribution). Nevertheless, to be more useful, the low-level operations need to be coupled with specific high-level information needs (e.g., which components of diversity require characterizing distributions, hierarchies, and/or clusters?).

Table 6.3: Ten low-level analytical tasks by Amar et al. [5] followed by the three additional tasks of *Characterize Hierarchy*, *Annotate*, and *Fit Models/Metrics*. The tasks are described in the context of diversity analysis.

Task	**Description**	**Example**
Retrieve Value	Given a set of observations, find attributes of those observations	What is the tenure of a given player in a given baseball team (Figure 6.10)?
Filter/Subset	Given some concrete conditions on attribute values, find observations satisfying those conditions.	What are the moth observations collected in HJA Forest in 2008 (Figure 6.16)?
Compute Derived Value/Metric	Given a set of observations, compute an aggregate numeric representation of those observation.	What is the faultline level of a given team (Figure 6.10)?
Find Extremum	Find data observations possessing an extreme value of an attribute over its range within the data set.	What is the moth species with highest abundance (Figure 6.4)?
Sort	Given a set of observations, rank them according to some ordinal metric.	Sort the moth species observations by abundances (Figure 6.4).
Determine Range	Given a set of observations and an attribute of interest, find the span of values within the set.	What is the age range of members in a given team (Figure 6.11)?
Characterize Distribution	Given a set of observations and an attribute of interest, describe the distribution of that attributes values over the set.	What is the tenure distribution of members in given team (Figure 6.11)?
Find Anomalies	Identify any anomalies within a given set of observations with respect to a given relationship or expectation, e.g., statistical outliers	Are there any rare moth species (Figure 6.4)?
Characterize Clusters	Given a set of observations and multiple attributes of interest, find clusters of similar attribute values.	Are there functional groups of trees with similar traits (Figure 6.7)?
Correlate	Given a set of observations and two attributes, determine useful relationships between the values of those attributes.	Is there a correlation between species richness, functional diversity, and ecosystem processes (Figure 6.14)?
Characterize Hierarchy	Given a set of data observations and hierarchy-based attributes, describe the hierarchical classification of the set over the attributes	What is the hierarchy of species in a unit of study (Figure 6.8)?
Annotate	Note or distinguish among attributes or observations based on their common or user-defined characteristics	Annotate ’age’ attribute as variety or as separation.
Fit Models/Metrics	Given a set of observations, fit a statistical or computational model to those observations—usually in the forms of visual indicators such as lines or colors.	Fit a specific distribution curve to the data (i.e., dash line on the data histogram)

Our proposed taxonomy is further motivated by the work of Andrienko and Andrienko [8], who introduced a classification of tasks strongly based on information needs of analysts and tied closely to spatiotemporal data. In their framework, a task is defined as a *query* to find the unknown information (*target*) corresponding to the specified or known information (one or more *constraints*). The target represents data behavior of interest such as distributions, clusters, correlations fulfilled by one or many constraints (e.g., population, space, time). The classification, whose general outline is illustrated in Figure 6.17, makes a distinction between the two classes of task: *elementary tasks*—which concern individual elements of data (e.g., what is the height of a given tree measured in a given date?), and *synoptic tasks*—which involves data behaviors in a set or subset of data *as a whole* (e.g., what is the shape of distribution of moth species caught in a given date and location?). Synoptic tasks are further classified into *descriptive* (e.g., characterize distributions) and *connectional* (e.g., characterize correlation) tasks. Since the classification presents high-level analytical tasks, it can potentially serve as a generic framework for building a task taxonomy for field-specific needs like diversity analysis.

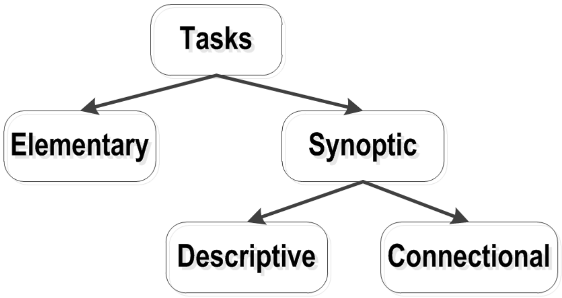


Figure 6.17: General outline of the classification of analytical tasks proposed by Andrienko and Andrienko [8]. Image redrawn from [8].

6.4 A Unified Task Taxonomy for Exploratory Diversity Analysis

While the generic task taxonomies do not necessarily consider or readily support specific tasks in diversity analysis, they serve as a framework and primitive building blocks for our proposed unified task taxonomy. In fact, our taxonomy offers an instantiation, combination, and extension of the taxonomy of data-centric queries by Andrienko and Andrienko [8] and the analytic low-level operations by Amar et al. [5] in the context of a specific analysis. Figure 6.18 depicts the outline of our proposed taxonomy.

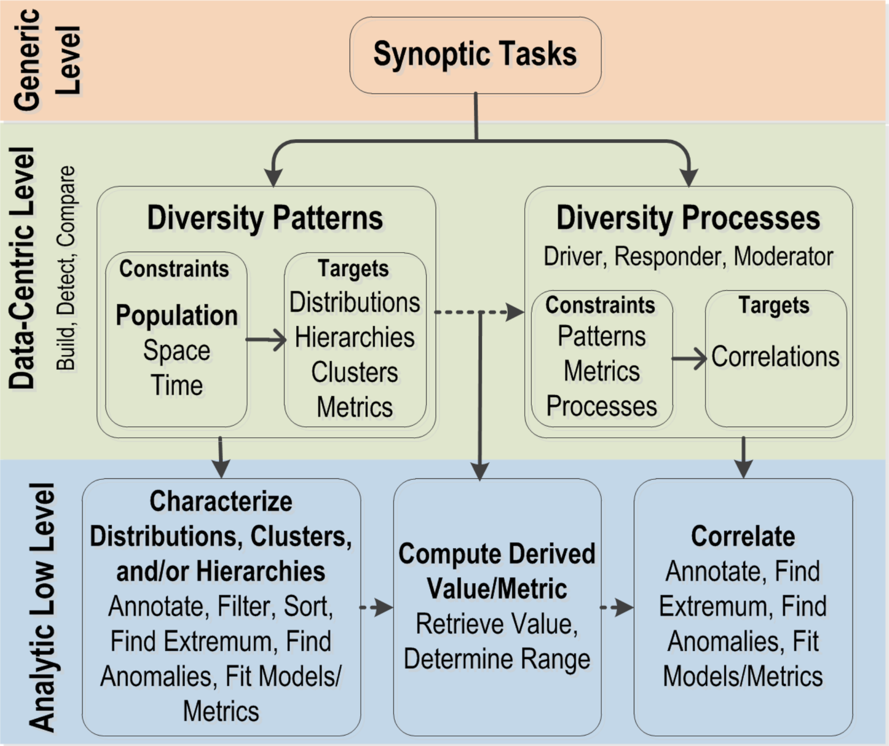

Generic Level
Synoptic Tasks
Data-Centric Level
Build, Detect, Compare
Diversity Patterns
Constraints
Population
Space
Time
Targets
Distributions
Hierarchies
Clusters
Metrics
Diversity Processes
Driver, Responder, Moderator
Constraints
Patterns
Metrics
Processes
Targets
Correlations
Analytic Low Level
Characterize Distributions, Clusters, and/or Hierarchies
Annotate, Filter, Sort, Find Extremum, Find Anomalies, Fit Models/ Metrics
Compute Derived Value/Metric
Retrieve Value, Determine Range
Correlate
Annotate, Find Extremum, Find Anomalies, Fit Models/Metrics

Figure 6.18: Proposed task taxonomy for exploratory analysis of diversity organized at three levels of abstractions: (1) Generic Tasks, (2) Data-centric Queries, and (3) Low-level Analytical Operations. Vertical solid arrows represent how the tasks in an upper level can be mapped to one or many tasks in a lower level. Horizontal dashed arrows suggest the workflow between tasks within the same level.

In essence, the taxonomy can be viewed at three levels of organization (or abstraction), representing the reasoning process of how analytical tasks transform information needs into knowledge and insights. Specifically, an information need starts in a very abstract form of synopsis (*Generic Level*), then is realized with specific queries on diversity patterns and processes in the analyst's mind (*Data-Centric Level*), and finally can be achieved with low-level operations on appropriate analysis tools (*Analytic Low Level*). The following subsections describe each of the three levels.

6.4.1 Generic-level Tasks

At the generic and also highest level, the taxonomy considers only synoptic tasks (Figure 6.18) as opposed to both elementary and synoptic tasks as in Andrienko and Andrienko's framework [8] (Figure 6.17). We made that decision based on the understanding that diversity patterns and processes concern behaviors of (sub)sets of observations *as a whole* as opposed to individual data elements (Table 6.1). While specific individual observations (e.g., rare or extreme observations) may be of interest to researchers—especially to certain microbiologists who have access to only a small number of samples due to sampling challenges (e.g., subsurface environments)—these observations are usually assessed in relation to other (sub)set of observations and are still considered as a whole. On a related note, the value of a visualization typically lies in its capacity to uncover patterns or behaviors in data as a whole. Investigation of individual observations may be better served by raw tables coupled with database queries.

6.4.2 Data-centric Queries

Data-centric queries (Figure 6.18, middle level), which are morphed from synoptic tasks, encompass specific information needs regarding *building, detecting, or comparing diversity patterns and processes* as presented in the alignment framework (Table 6.1). To some extent, patterns and processes match the descriptive and connectional tasks in Andrienko and Andrienko's framework [8] (Figure 6.17). At this level, we also adopt their definition of task as query, which consists of two parts: one target (unknown information) and one or many constraints (known information).

Diversity Patterns. These queries aim to gain knowledge into diversity patterns. Based on the constraints, the main objective is to characterize data behaviors (targets) as distributions, hierarchies, clusters, or summary statistics, following the three considerations. The primary constraint is *population*, which is represented by collected samples of independent observations characterized by multiple attributes. In addition, data samples could be collected in the context of *space* and *time*, two additional secondary constraints. For example, information needs regarding functional diversity in ecology as well as faultlines in organizational management may involve building, detecting, or comparing distributions of clusters of data observations (targets) collected from a specific population—and in some cases—in space and time (constraints) (Consideration 2).

Diversity Processes. These queries examine the scientifically meaningful relationships between diversity patterns and system processes. Diversity patterns can play multiple roles in such causal relationships: diversity as driver, as responder, and/or as moderator (Table 6.1). These roles are characterized by the correlation between data behaviors of diversity patterns/metrics—acquired from the diversity pattern queries—and of system processes. As an example, in ecology, as the name suggests, functional diversity, which is often characterized by statistical metrics or distribution of functional groups, is directly correlated with various ecosystem processes [116]. In a synthesis study, Díaz and Cabido [34] reported a strong correlation between different functional groups of plants such as nitrogen-fixing legumes, warm-season bunchgrasses, or rosette forbs and ecosystem processes such as retention of nitrogen and total aboveground biomass. During exploratory analysis, the queries on diversity patterns usually serve as prerequisites for understanding diversity processes. This kind of "workflow" is represented as horizontal dashed arrows in the task taxonomy (Figure 6.18).

6.4.3 Low-level Analytical Operations

Data-centric queries in the analyst's mind are finally decomposed to low-level operations to be fulfilled by visual-analysis tools (Figure 6.18, bottom level). All Amar et al.'s operations [5] described in Table 6.3 are relevant to diversity analysis. Note that secondary operations can be combined to accomplish a primary operation, which is denoted as bold texts in Figure 6.18. For instance, characterizing distribution of a data subset may require filtering data first, and then sorting the data.

The ten original operations [5] are not comprehensive. Guided by the alignment framework of diversity concerns, we introduce three additional low-level operations: *Characterize Hierarchy*, *Annotate*, and *Fit Models/Metrics*. Hierarchy characterization is required when users inspect taxonomic diversity of species (or their equivalents) (Consideration 2). Annotation is useful when analysts wish to note or distinguish among attributes or observations based on their common or user-defined characteristics (Consideration 1). For example, management researchers may wish to annotate the 'age' attribute as *variety* in one case and as *separation* in other cases. Fitting Models/Metrics represents a scenario in which analysts, given a set of observations, may want to fit a statistical or computational model to those observations—usually in the forms of visual indicators. For example, they may want to fit a straight line to a set of observations in a scatter plot to represent linear correlation (Figure 6.15); they may want to see some visual indicator to represent separation among clusters of observations (Figure 6.11 and Figure 6.12).

In summary, guided by the alignment framework and existing generic task taxonomies, our proposed taxonomy aims to capture all possible queries and operations in the process of exploring diversity data. The reasoning process of the analyst may start with high-level queries on scientific phenomena such as “what are the functional diversity patterns?”, followed by low-level analytical operations such as “characterize clusters of the observed data”. The data behavior characterization in turn enables the analyst to understand and make inferences about the underlying scientific phenomena. Understanding of the reasoning process as well as the specific queries and operations on diversity data is a critical task for designers in designing visual-analysis tools.

6.5 Discussion

This work presents the first cross-disciplinary synthesis study targeting exploratory analysis of diversity. Our study provides two contributions: (1) understanding of the diversity concerns aligned across the analyses of species, microbial, and workgroup diversity and (2) a unified taxonomy of analytical tasks guiding the design of visual-analysis tools to address these concerns. Here we extend our discussion on (1) validation and refinement of the alignment framework with subject-matter experts, (2) limitations and future work, and (3) implications for diversity studies and design of visualizations.

6.5.1 Feedback from Subject-Matter Experts

Feedback from experts is critical to ensure the alignment framework fulfills its intended purpose of characterizing the diversity analysis problem. Our validation of the framework consists of two phases: (1) a pilot phase with our domain expert collaborators and (2) a survey study with other external experts. To stimulate our discussion, we adopt the following feedback criteria [2] (Table 6.4): comprehensiveness (of the framework), ease of use, precision, usefulness, discoverability, and alignability.

Table 6.4: Criteria and corresponding questions for validation and refinement of the alignment framework of diversity concerns. The criteria are adopted from Ahn et al. [2].

Feedback Criterion	**Question**
Comprehensiveness	Are any concerns missing from the framework?
Ease of Use	Is the framework easy to understand?
Precision	Does the framework describe precisely the concerns and the corresponding data behaviors?
Usefulness	Can the framework be used by experts to organize and cross compare their studies?
Discoverability	Does the framework help experts discover new concerns they had not thought of?
Alignability	Would the experts think concerns could be aligned across the three fields of interest?

Pilot Feedback. In developing the alignment framework, we have set up multiple face-to-face discussion sessions and follow-up email correspondence between two visualization researchers and two domain expert collaborators: one ecologist and one microbiologist/microbial ecologist. The aims are to understand their information needs and to collect feedback on early thoughts on the framework before a full survey study.

The ecologist was instrumental in helping refine the vocabulary and presentation of the framework as well as validate the concerns of species diversity. Discussing the framework's comprehensiveness, she pointed out niche separation and dominance in ecology as potentially parallel concepts to separation and disparity in organizational management, respectively. With regards to the usefulness criterion, she requested compelling examples to demonstrate the operationalizations of diversity concepts as well as their alignment across the three areas. We responded with examples of visualization and introduced the three considerations for data behavior characterization. Interestingly, after seeing how multiple stacked histograms are used to communicate subgroups and faultlines in organizational management (Figure 6.11), the ecologist immediately requested the same chart for comparison of the structure of different groups of moths from the moth data set (Figure 6.13). We take that request as a positive sign that the framework helped the ecologist discover new diversity concerns she had not thought of, such as separation between clusters of observations or functional groups.

Our discussion with the microbiologist/microbial ecologist suggested that between microbial diversity and species diversity, while there are some parallels in analysis approach, there also exist distinctions in information needs and characteristics of diversity data. Specifically, microbial diversity analysis emphasizes exploration of genomic information to identify previously unknown microorganisms and ultimately, to understand their functionality. The classification is usually performed in the data pre-processing stage (Figure 6.1), using DNA extracting and sequencing [145, 45]. Genomic information is rich in terms of OTUs but is somewhat limited in terms of the number of biological samples and the number of attributes (e.g., spatial and temporal) because in many cases (e.g., subsurface environment), data sampling remains a challenge. However, surface microbial communities (e.g., in soils, waters, humans, animals) can be sampled much more frequently as in the MicrobiVis example [37]. On the other hand, species diversity deals with already known species and their well-understood characteristics (e.g., taxonomic classification, food types, habitats) so its analysis emphasis is really on the diversity patterns of multiple observations and their causes and consequences, providing that ecologists have access to larger number of observations and other environmental factors.

In all, the microbiologist assessed that the alignment framework was useful for cross-comparison of diversity studies. It helped him discover new diversity concerns such as diversity faultlines and corresponding techniques such as multiple stacked histograms. It is also encouraging to hear his comment that the future of microbiology would benefit from a similar species diversity analysis, and essentially from the alignment framework, providing that microorganisms are well classified and more data replicates are available. He also recommended related work on microbial diversity that we reference in this work.

Survey Study with Other Experts. After the pilot phase, we further evaluated the framework in a survey study involving nine domain experts whose expertise was in species diversity (four), both species and microbial diversity (one), and workgroup diversity (four). All of them, who authored published research work cited in this chapter, volunteered to participate in the survey in response to our emails solicitating their feedback. They answered the survey after reading a technical report presenting the framework.

The evaluation survey consisted of six Likert-style statements (Table 6.4), in which the experts were asked to indicate their level of agreement on a scale of one (Strongly

Disagree) to five (Strongly Agree), and two open-ended questions: (1) if you disagree with any of the above statements, please explain your reason and (2) please comment on any aspects concerning the framework or the technical report.

The survey results were encouraging (Figure 6.19). Most of the experts strongly agreed or agreed on the framework's comprehensiveness (seven (out of nine)), ease of use (six), precision (seven), usefulness (seven), discoverability (six), and alignability (seven). Several experts expressed their enthusiasm for the work, especially its novelty, necessity, and timeliness: "I really like how you bring together three so different disciplines in the first cross-disciplinary synthesis study about diversity! Your report represents a tremendous amount of work."; "It's [the framework] looking great! I'm so happy that you're tackling this challenge–it's sorely needed"; and "Overall, I found the text very interesting and well written. The subject is also very timely."

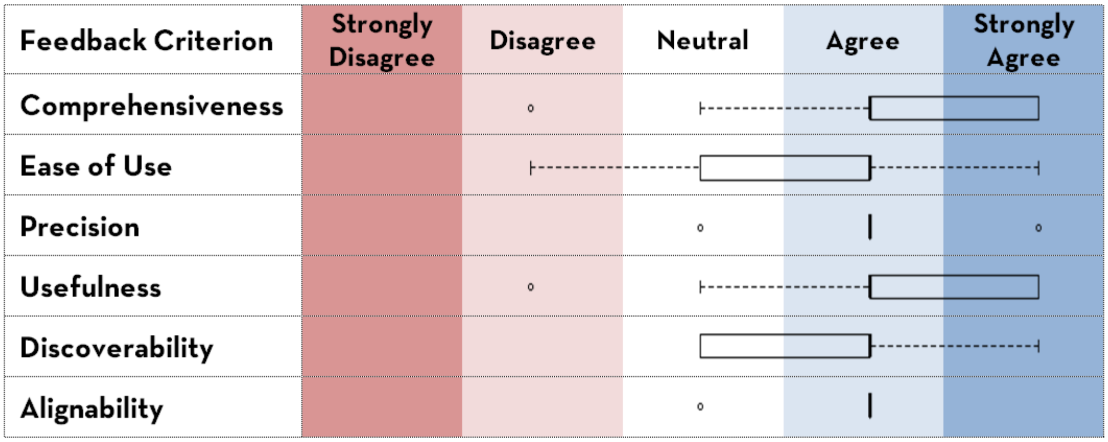


Figure 6.19: Boxplot of responses from n domain-experts to each of the six Likert-style feedback criteria/statements (Table 6.4). The experts were asked to indicate their level of agreement on a scale of 1 (Strongly Disagree) to 5 (Strongly Agree).

Nevertheless, some experts also pointed out several limitations of the work. With respect to comprehensiveness of the framework, one mentioned the lack of diversity components concerning data acquisition and pre-processing, which we elaborate in the next subsection. Commenting on the role of visual exploration in diversity analysis, one expressed concern about the issue with post-hoc analysis (i.e., use visualization to look for patterns that were not specified a priori or discriminate actual patterns from noise). We respond to that comment that visual exploration, which is only part of a larger analysis process (Figure 6.1), may prompt further statistical tests (that take into account post hoc analysis) or graphical inference tests [160], additional data collection, and experiments. We also argue that while traditional statistical tests are well studied, they may not be able to uncover unexpected data behaviors such as shapes of distribution, outliers, separation of clusters of interest to diversity analysis. Complementing statistics, visualizations are particularly effective for those tasks.

In addition to critical comments, the experts also offered suggestions for improvement. One pointed out related domains that may share common diversity analysis such as community detection among social networks and additional analysis techniques such as Bayesian approaches for workgroup data. Several of them suggested minor changes for the terms used in Table 6.1 (e.g., correlation vs. regression, taxonomy vs. ontology vs. typology) and other related work. We considered them carefully, followed up with the corresponding experts via email if necessary, and incorporated them into the framework.

6.5.2 Limitations and Future Work

This work emphasizes the exploration stage of the analysis process (i.e., hypothesis generation), following data acquisition and pre-processing stages and preceding further hypothesis testing, as illustrated in Figure 6.1. Other stages may involve additional diversity concerns and corresponding analytical tasks. For example, microbe samples could be pre-classified into OTUs at different taxonomic levels using the Ribosomal Database Project (RDP) [31] and the process could benefit from dedicated analytical tasks such as dimensionality reduction using principal component analysis (PCA) and Nonmetric multidimensional scaling (NMDS) [128]. In another example, data acquisition (or sampling) plays a critical role because it affects diversity patterns and processes. Species richness, for instance, tends to increase when the number of samples increases [101]. Such analysis of assessing species richness from the results of sampling is supported by dedicated analytical tasks such as constructing and comparing species accumulation curves or rarefaction curves [101]. Extending this work beyond the exploration stage deserves deeper investigation in future work.

To keep our proposed taxonomy concise, we excluded analytical tasks necessary for collaborative exploration. For instance, analysts may wish to keep track of their findings and share their findings with other users [121]. These tasks are generic and relevant to almost all scientific analysis workflows [61].

Our literature review examines only three areas of interest. Diversity represents itself in many other fields. For example, chemists consider the similarity/diversity of molecular models in exploring the multitude of designs generated by simulation [77]; scholars study language diversity in order to understand societies [110]. All of these fields are advancing and new findings and analysis techniques may prompt revision of the framework and the taxonomy. Alternatively, we may have to create new ones for specific fields.

6.5.3 Implications for Diversity Studies

The alignment framework aims to support experts in adopting new diversity concerns within their own field of expertise or across fields. In addition to the examples presented in the Alignment Framework Section, we discuss several other usage scenarios here.

The first scenario demonstrates how the three types of diversity as variety, separation, and disparity in separate attributes could be extended to interaction among multiple attributes (Table 6.1). In fact, depending on the types of attribute under investigation, experts studying workgroups already conceptualize variety-based, separation-based, and disparity-based faultlines and subgroups, as introduced by Carton and Cummings [24]. For example, composition of disparity attributes such as pay, rank, and decision power may form disparity-based faultlines and subgroups in a team [24]. The same conceptualization might be applied to functional diversity in ecology, depending on the types of examined functional traits. For example, composition of resource-based functional traits for plants such as nutrient consumption, tree density, body size could create disparity-based functional groups in the examined unit of study.

The second usage scenario extends our discussion on the alignability between diversity faultlines in organizational management and functional diversity in ecology. Across the two areas, it would be informative to cross compare statistical metrics [116, 117, 150] and visual representations. For example, while the faultline metric used in the baseball data (Figure 6.10) does not involve a hierarchy of clusters [14], hierarchical clustering algorithms such as the FD metric in ecology [116] could potentially be adopted and vice versa. Other modern cluster algorithms from the field of data mining such as Affinity Propagation [43] could be potentially utilized. Further, configuration of attribute weighting is another unique feature of diversity faultlines potentially applicable to ecological functional diversity. For example, management researchers studying the impact of faultlines in workgroups may ask how many years of age difference between team members should be considered as equally important as a difference in gender or ethnicity [149]. Ecologists studying functional diversity may adopt similar configuration of the relative importance of functional traits, depending on the corresponding system processes of interest.

The third usage scenario discusses the missing component of taxonomic diversity in workgroup diversity (Table 6.1). To our understanding, experts studying workgroups have not yet examined hierarchical classification of attributes yet. That missing link may suggest a potential research direction. For example, functional expertise of team members are potentially hierarchical (e.g., ecology and microbiology majors are closely related since they are classified under life sciences) and the hierarchical information can be taken into account during investigation of faultlines and subgroups.

6.5.4 Implications for the Design of Visualization

Use of the alignment framework and the task taxonomy also has implications for the design of visualizations. Specifically, it provides visualization designers and researchers with a common vocabulary and considerations for designing and evaluating different visual-analysis tools targeting diversity data. We expect a set of base visualization techniques and tools for illuminating various components of diversity and providing new ways of looking at data across fields.

Typical Visual Representations. Following the three considerations and examples of visualization presented in the Alignment Framework Section, Table 6.5 attempts to make a list of typical visual representations that are well-suited to communicating the data behaviors of interest concerning diversity. The techniques, which by no means represent an exhaustive list, are suggested based on the understanding of their pros and cons from the field of information visualization [119, 120, 123]. This classification would serve as a useful reference for visualization designers targeting specific diversity concerns. A thorough survey of various existing visualization techniques for general purpose can be found in [83].

Table 6.5 could be extended to include techniques targeting diversity in space and

Table 6.5: Data behaviors of interest to diversity analysis and corresponding typical visual representations.

Data Behavior	**Examples of Diversity Concern**	**Data Characteristics**	**Typical visual representations (with example citations)**
Distributions	Variety and abundance in separate attributes	Univariate	Boxplot [154]
			Histogram [101]
			Stacked Bar Chart [22]
			Rank-abundance Curve [158, 101]
			Cumulative Frequency Curve [101]
		Multivariate	2D Scatter plot and its variants (2D Heatmap, Fluctuation Diagram)
			Multiples of univariate representations (e.g., Boxplot [154], Histogram, Scatter plot matrix [29] , Diversity Map [120])
Distributions + Clusters	Functional diversity; Subgroups/ Faultlines	Bivariate	Scatter Plot [135]
			Mosaic Plot [58]
		Multivariate	Multiple Stacked Histograms [123]
			Scatter Plot Matrix [123]
Distributions + Hierarchies	Taxonomic Diversity (Richness+Evenness)	Multivariate	Treemap [140, 71]
			Sunburst, Icicle [148]
Hierarchies	Taxonomic Diversity (Richness)	Multivariate	Node-link diagram and its variants (e.g., Tree [94], Dendrogram [18])
Correlations	Processes of Diversity	Bivariate	Scatter plot or Line Chart [117, 15]
		Multivariate	Scatter Plot Matrix [29]
			Parallel Coordinates [74, 37]
			Parallel Sets [86]

time. Recall that the three considerations suggest that if time and space are involved, the techniques should support users to explore how the data behaviors of interest (e.g., summary statistics, distributions, clusters, and/or hierarchies) vary over time and space. To communicate spatial distributions or clusters in univariate data, a geographical map with an additional encoding (e.g., a heat map) is widely used. However, visualizing data behaviors of multivariate data on a map remains a challenge. Potential solutions include overlaying other representations on a geographical map or alternatively, presenting geographical maps and other representations in separate windows connected by interactions [8]. On the other hand, to convey how the data behaviors of interest vary over time, one possible solution is to employ multiple snapshots of visual representations—for example, multiple histograms—one for each time point. Alternatively, animation of visualization states over time may potentially be useful. Andrienko and Andrienko [8] present a thorough investigation of exploratory analysis of spatial and temporal data in their book.

Assessment of existing visual-analysis tools. In addition to guiding the invention of future visualizations, the three considerations could be used to assess existing techniques and tools. For example, consider the MicrobiVis tool [37], which employed parallel coordinate plot (PCP)—among other techniques—to convey the separation between two groups of microbial samples across multiple OTUs (Figure 6.20). PCP is well suited to make the data observations visible as well as to convey the correlation between two neighboring attribute axes [74] (Table 6.5). However, we argue that the choice of PCP does not support Consideration 2—PCP is not effective in supporting users in comparing the distributions and separation of clusters across multiple attributes [123]. Figure 6.21 presents an alternative design in which stacked histograms are selectively overlaid along the axes to convey the distribution of clusters as well as separation among clusters across the attributes of interest.

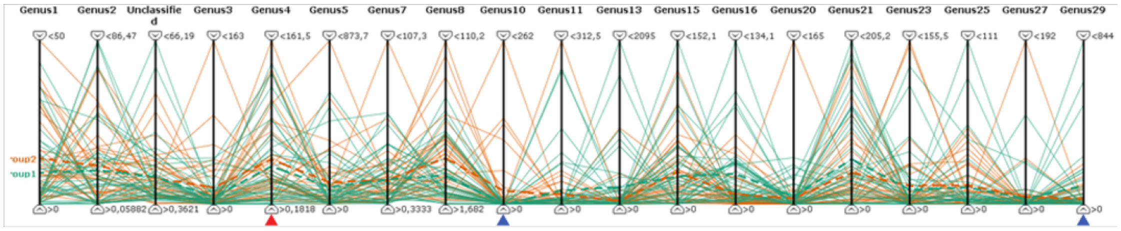


Figure 6.20: Samples from two oral bacterial populations visualized using Parallel Coordinate Plot (PCP) supported by the MicrobiVis tool [37]. Vertical axes represent a set of Genus OTUs of interest to the analyst; each of the polylines represents a sample that intersects each genus axis at the value corresponding to the abundance of the genus detected in the sample. Distinct colors are used to differentiate the two populations: group 1 and group 2. Red and blue arrows indicate some interesting genera identified by the analyst. For example, the first blue arrow from the left marks Genus 4 in which group 1 and group 2 are separated with respect to abundance. Image reused with permission from Fernstad et al. [37]. Copyright ©2011 IEEE.

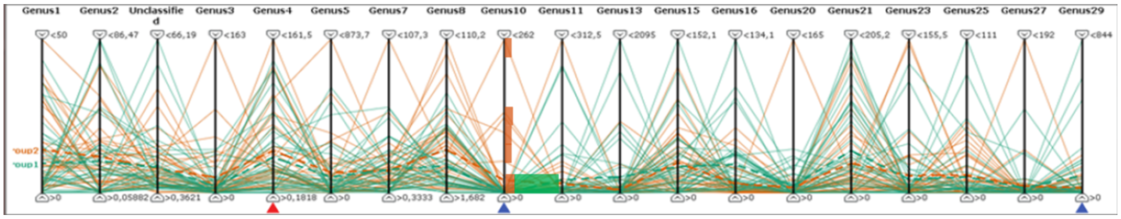


Figure 6.21: An alternative design to the PCP in Figure 6.20 in which stacked histograms are selectively overlaid along the axes to convey the distribution of clusters as well as separation among clusters across the attributes of interest. For demonstration purpose, Genus 4 axis—marked by the first blue arrow from the left—is re-drawn with stacked histograms. We argue that the stacked histograms make separation between the two groups of microbe samples within Genus 4 stand out: while all samples of group 1 contain low abundance of Genus 4, most of the samples of group 2 contain higher abundance of Genus 4. We adapt the Figure from Fernstad et al. [37]. Original figure ©2011 IEEE.

6.6 Conclusions

Ecologists are increasingly concerned about changes in diversity patterns of species communities and how they influence ecosystem functioning. However, ecologists may not be aware of analysis techniques (e.g., visualization and statistics) in other fields, such as organizational management, that may help improve their own understanding. Aiming to connect that missing link, this interdisciplinary work abstracts diversity concerns across the three areas of species diversity, microbial diversity, and workgroup diversity in an alignment framework and offers an operationalization of these concerns in terms of data behaviors of interest and common analytical tasks. Subject-matter experts and tool designers may take advantage of this work to find a common ground for the diversity analysis problem. We expect this work will help guide the evaluation and refinement of existing visualization techniques as well as the invention of future ones. We also anticipate further discussions regarding validation and amendment to both the alignment framework and the unified task taxonomy.

Chapter 7: Conclusion

As current technological advances continue to drive the generation of tremendous amounts of data, appropriate tools for data analysis become increasingly important. In many application domains, there is a need for visualizations that enable exploration and communication of the diversity aspect of the multivariate data. Ecologists collect and analyze species diversity data to conserve biodiversity and understand the interactions, for example, between communities of plant and animal species and the environment. Likewise, organizations strive to build diverse teams of effective team players and problem solvers, admissions officials seek to build diverse incoming classes, and scholars study language diversity to understand societal development. Nonetheless, there are few interactive visualization techniques and tools designed specifically to serve those purposes. This research work aimed to understand the issues involved in designing diversity visualizations. In doing so, we provide answers to the three key research questions summarized in Table 7.1.

This dissertation utilized a variety of scientific methods. In collaboration with domain experts in ecology, microbiology, and organizational management, we characterized the diversity analysis problem, investigated existing visualization techniques, and developed new ones. Subsequently, we designed and carried out controlled experiments to understand the effectiveness of the techniques. We also employed a user-centered design process to integrate the effective techniques into the analysis process of ecologists and management researchers as they analyze real-world diversity data. We futher generalized the results to design considerations and a taxonomy of analytical tasks, guiding the creation of future visualizations that target diversity data.

Table 7.1: Summary of the dissertation contributions grouped by the three research questions.

Research Question	**Contributions**	**Evaluation of Contributions**
RQ1. How is diversity conceptualized across the multiple fields that study it and what are the fundamental scientific questions/hypotheses of interest regarding diversity?	A framework of diversity concerns aligned across the three fields of ecology (species diversity), microbiology (microbial diversity), and organizational management (workgroup diversity) (Chapter 6).	Cross-comparison of diversity literature across the three fields and feedback from subject-matter experts.
RQ2. Which existing or novel multivariate representation and interaction techniques are particularly useful in exploring and communicating diversity data?	Diversity Map–Multivariate visual representation of diversity patterns in separate attrbutes (Chapter 3)	Controlled user study: Diversity Map vs. Glyph-based Technique [114] User-centered design study with ecologists exploring the moth data sets
	HIST–Multivariate visual representation of diversity patterns across multiple attributes: A case study of team faultlines (Chapter 5).	Design study with management researchers analyzing real-world workgroup data sets (e.g. MLB data) Controlled user study: HIST vs. PCP [75, 74] vs. SPLOM [29]
	A unified taxonomy of analytical tasks for exploratory analysis of diversity (Chapter 6).	Visual analytics design principles and feedback from subject-matter experts.
RQ3. What is the role of interactive visualization in the real-world analysis process in which diversity is a key element?	EcoDATE (Ecological Distributions and Trends Explorer)–web-based visual-analysis tool that facilitates exploratory analysis of long-term ecological data (Chapter 4).	User-centered design study with ecologists exploring the three long-term ecological data sets of cone production, stream chemistry, and forest structure. Working group at the LTER All Scientists Meetings 2012.

7.1 Review of Dissertation Contributions

After the first two chapters introducing the diversity analysis problem and providing an overview of our contributing solutions, Chapter 3 presented (1) a novel representation, the Diversity Map, for visualizing diversity patterns in separate attributes of a large set of multivariate objects and (2) a rigorous evaluation of the effectiveness of the proposed technique. Our design considerations and user study design centered on a precise definition of diversity adopted from the field of ecology that takes both richness of variety of attribute values and evenness of relative abundances of objects into account [101]. The design of the Diversity Map built loosely upon ideas from both parallel coordinates [75, 74] and multiple histograms. In a formal user study, we found that the Diversity Map allows users to as or more accurately judge elements of diversity than the only other existing multivariate method [114] designed to visualize diversity.

The Diversity Map representation was further refined into an interactive tool, EcoDATE (Ecological Distributions and Trends Explorer), for ecologists to explore diversity patterns and temporal trends in long-term ecological data (Chapter 4). EcoDATE was developed through a process of user-centered design in collaboration with long-term ecological research. Application of the EcoDATE tool to long-term ecological data sets on cone crop production, stream chemistry, and forest structure reveals that it facilitates overview, initial hypothesis testing, and hypothesis formulation in an open-ended framework. Further, the use of EcoDATE underscored an understanding of the potentially different pathways to gaining insights into data and generating hypotheses. The tool is readily available at `http://purl.oclc.org/ecodate`.

While Chapter 3 and Chapter 4 emphasized diversity patterns generalized as distributions of objects in separate attributes, Chapter 5 explored the design space for graphical representation of diversity patterns conceptualized as distributions of clusters of objects across mixed-type attributes. We instantiated the problem in the context of diversity team faultlines, a fundamental construct in management that was derived from multivariate cluster analysis. In collaboration with management researchers, we contributed a set of requirements for faultline visualizations. Then we designed and evaluated our proposed technique, HIST, which is based on multiple linked, stacked histograms in a parallel axis layout. The controlled user study results show that HIST supports users in inspecting elements of team faultlines as or more effectively than the other two common cluster representation methods of parallel coordinate plot (PCP) [75, 74] and scatter plot matrix (SPLOM) [29]. Finally, we augmented HIST with visual elements to further facilitate the faultlines identification tasks.

Finally, building on and extending the results from the previous chapters, Chapter 6 presented the first cross-disciplinary study that aims (1) to align understanding of diversity concerns across fields and (2) to propose a unified taxonomy of analytical tasks guiding the design of visualizations addressing these concerns. The three fields under investigation include species diversity in ecology, microbial diversity in microbiology, and workgroup diversity in organizational management. The concerns of interest cover (1) characteristics of diversity data, (2) description of diversity patterns, and (3) hypotheses regarding the causes and consequences of diversity (processes), and (4) most importantly, characterization of each of the concerns (if applicable) with corresponding data behaviors of interest to analysts. The proposed task taxonomy offers an instantiation of existing task taxonomies from the field of visual analytics (e.g., [5, 8]) in the context of a specific diversity analysis. In essence, the taxonomy captures the reasoning process of how analytical tasks transform information needs into knowledge and insights, facilitated by interactive visualization. The results from our synthesis study aim to provide domain experts and visualization designers with common vocabulary and considerations for designing and evaluating different visualizations targeting diversity analysis.

7.2 Directions for Future Work

While we believe this dissertation represents a positive step in understanding interactive diversity visualization of multivariate data, it also has limitations that we discussed in specific chapters. Here we elaborate several broader directions in which the work could be extended.

7.2.1 Diversity Visualization in Other Endeavors and Domains

Not only is interactive visualization a potentially powerful analysis tool, but it could prove to be a very useful means for constructing sets of multivariate objects that possess the desired types and levels of diversity. Such a tool lends itself to organizational management where the task at hand often involves building a team or an organization that exhibits particular diversity profiles and that ideally will produce positive outcomes. Future work would investigate tools that allow users to explore what-if scenarios by observing the effects of adding/removing particular objects on diversity patterns *on the fly*.

We are also intrigued by the possibility of investigating diversity in online collaborations and other social contexts such as political science. There have been studies of the effects of group diversity on productivity as well as member withdrawal behaviors among Wikipedia projects [26], however, the effects of attributes are studied one at a time. In future work, it will be informative to re-visit the problem and investigate the effects of multiple attributes simultaneously with diversity faultlines as the primary measure. In addition, political science, which studies demographic diversity and how diversity relates to voting patterns and election results, opens another area that may benefit from faultline-based visualization.

While this dissertation considers only three domains of ecology, microbiology, and organizational management, diversity is of interest in others such as chemical diversity [77] and language diversity [110]. These domains are advancing and new findings and analysis techniques may require extension and refinement to our proposed alignment framework of diversity concerns or the unified taxonomy of analytical tasks.

7.2.2 Integration of Exploration with Other Analysis Stages

While this dissertation emphasizes the exploration stage of the analysis process, other stages of data acquisition, data pre-processing, and hypothesis testing are crucial and demand additional analytical tasks. For example, in microbial diversity analysis, pre-classification of microbe samples into taxonomic units is facilitated by ribosomal databases (e.g., [31]) and by dedicated dimensionality reduction or cluster analysis techniques [128]. In another example, sampling of data plays an important role in the diversity investigation and such data acquisition stage is supported by specific visualization techniques such as species accumulation curves or rarefaction curves [101]. This work could be extended to consider other stages of domain-specific analyses, in addition to exploration.

7.2.3 Strategies to Exploratory Data Analysis

From our collaboration with experts across domains, we have observed different approaches to insight and hypothesis generation from exploratory analyses—in particular—regarding diversity. As demonstrated in Chapter 4, in some cases, the visual information seeking mantra "overview first, zoom and filter, and details on demand" [141] serves as the main guidance. In other cases, analysts may adopt the iterative in-depth three-step process of specifying visualization views, characterizing views, and gaining insights. In all, exploratory data analysis is a very open-ended process and could benefit from future research that investigates varying analysis strategies and the facilitation of a visual-analysis tool in such an open-ended exploration.

On a related note, while existing statistical tools (e.g., R) and workflow systems (e.g., Kepler [97]) serve as potentially powerful frameworks for structuring and managing domain-specific analysis processes, they still lack support for interactive visualizations and usability (perhaps due to a steep learning curve of their default command-line interfaces). We envision that visual-analysis techniques and tools such as HIST (Chapter 5) and EcoDATE (Chapter 4) could be integrated into such frameworks and leverage their capacities of provenance management and statistical computation.

7.3 Concluding Remarks

Arguably the most challenging questions in disciplines that study diversity center on gauging how diversity patterns of multivariate objects of interest are structured and explaining how those patterns can be related to ecosystem or human system processes. Answering such questions, which are inherently data-driven, are now facilitated by advancements in data collection and analysis techniques. Answering such questions may also require a collaboration not only between disciplines that concern diversity but also with disciplines that specialize in data analysis and knowledge discovery such as statistics and data visualization. This research work shows that collaboration between domain experts and visualization researchers can potentially produce powerful techniques, tools, and visualization design guidelines that enable us to answer critical scientific questions regarding diversity that advance and sustain our world.

Bibliography

[1] Steven A Acker, W Arthur McKee, Mark E Harmon, and Jerry F Franklin. Long-term research on forest dynamics in the Pacific Northwest: a network of permanent forest plots. *Man and the Biosphere Series*, 21:93–106, 1998.

[2] Jae-wook Ahn, Catherine Plaisant, and Ben Shneiderman. A task taxonomy of network evolution analysis. *University of Maryland, Human-Computer Interaction Lab Tech Report HCIL-2012-09*, 2012.

[3] Wolfgang Aigner, Silvia Miksch, Wolfgang Müller, Heidrun Schumann, and Christian Tominski. Visualizing time-oriented data—a systematic view. *Computers & Graphics*, 31(3):401–409, 2007.

[4] Eric E Allen and Jillian F Banfield. Community genomics in microbial ecology and evolution. *Nature Reviews Microbiology*, 3(6):489–498, 2005.

[5] Robert Amar, James Eagan, and John Stasko. Low-level components of analytic activity in information visualization. In *IEEE Symposium on Information Visualization*, pages 111–117. IEEE, 2005.

[6] Theodore W Anderson. *An introduction to multivariate statistical analysis.* Wiley New York, 1958.

[7] David F Andrews. Plots of high-dimensional data. *Biometrics*, pages 125–136, 1972.

[8] Natalia Andrienko and Gennady Andrienko. *Exploratory analysis of spatial and temporal data.* Springer Berlin, Germany, 2006.

[9] Almir Olivette Artero, Maria Cristina Ferreira de Oliveira, and Haim Levkowitz. Uncovering clusters in crowded parallel coordinates visualizations. In *IEEE Symposium on Information Visualization*, pages 81–88. IEEE Computer Society, 2004.

[10] Richard A Becker and William S Cleveland. Brushing scatterplots. *Technometrics*, 29(2):127–142, 1987.

[11] Michael Begon, Colin R Townsend, and John L Harper. *Ecology: from individuals to ecosystems.* Wiley-Blackwell, 2005.

[12] Jacques Bertin. *Semiology of graphics: diagrams, networks, maps.* University of Wisconsin press, 1983.

[13] Enrico Bertini, Adam Perer, Catherine Plaisant, and Giuseppe Santucci. BELIV'08: Beyond time and errors: novel evaluation methods for information visualization. In *CHI'08 extended abstracts on Human factors in computing systems*, pages 3913–3916. ACM, 2008.

[14] Katerina Bezrukova, Karen A Jehn, Elaine L Zanutto, and Sherry MB Thatcher. Do workgroup faultlines help or hurt? A moderated model of faultlines, team identification, and group performance. *Organization Science*, 20(1):35–50, 2009.

[15] Katerina Bezrukova, Chester S Spell, and Jamie L Perry. Violent splits or healthy divides? Coping with injustice through faultlines. *Personnel Psychology*, 63(3):719–751, 2010.

[16] Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. D^3 data-driven documents. *Visualization and Computer Graphics, IEEE Transactions on*, 17(12):2301–2309, 2011.

[17] William D Bowman and Timothy R Seastedt. *Structure and function of an alpine ecosystem: Niwot Ridge, Colorado.* Oxford University Press, USA, 2001.

[18] BR Briggs, JW Pohlman, M Torres, M Riedel, EL Brodie, and FS Colwell. Macroscopic biofilms in fracture-dominated sediment that anaerobically oxidize methane. *Applied and environmental microbiology*, 77(19):6780–6787, 2011.

[19] Nicholas Brokaw, Todd Crowl, Ariel Lugo, William McDowell, Frederick Scatena, Robert Waide, and Michael Willig. *A Caribbean Forest Tapestry: The Multidimensional Nature of Disturbance and Response.* Oxford University Press, USA, 2012.

[20] Steven P Callahan, Juliana Freire, Emanuele Santos, Carlos E Scheidegger, Cláudio T Silva, and Huy T Vo. Vistrails: visualization meets data management. In *Proceedings of the 2006 ACM SIGMOD international conference on Management of data*, pages 745–747. ACM, 2006.

[21] Nan Cao, David Gotz, Jimeng Sun, and Huamin Qu. DICON: Interactive visual analysis of multidimensional clusters. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2581–2590, 2011.

[22] J Gregory Caporaso, Justin Kuczynski, Jesse Stombaugh, Kyle Bittinger, Frederic D Bushman, Elizabeth K Costello, Noah Fierer, Antonio Gonzalez Pena, Ju-

lia K Goodrich, Jeffrey I Gordon, et al. QIIME allows analysis of high-throughput community sequencing data. *Nature methods*, 7(5):335–336, 2010.

[23] Stuart K Card, Jock D Mackinlay, and Ben Shneiderman. *Readings in information visualization: using vision to think*. Morgan Kaufmann, 1999.

[24] Andrew M Carton and Jonathon N Cummings. A theory of subgroups in work teams. *Academy of Management Review*, 37(3):441–470, 2012.

[25] F Stuart Chapin, Mark W Oswood, Keith Van Cleve, Leslie A Viereck, and David L Verbyla. *Alaska's changing boreal forest*. Oxford University Press, USA, 2006.

[26] Jilin Chen, Yuqing Ren, and John Riedl. The effects of diversity on group productivity and member withdrawal in online volunteer groups. In *Proceedings of the 28th international conference on Human factors in computing systems*, pages 821–830. ACM, 2010.

[27] Herman Chernoff. The use of faces to represent points in k-dimensional space graphically. *Journal of the American Statistical Association*, 68(342):361–368, 1973.

[28] Ed Huai-hsin Chi. *A Framework for Information Visualization Spreadsheets*. PhD thesis, University of Minnesota, 1999.

[29] William S Cleveland and Robert McGill. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association*, 79(387):531–554, 1984.

[30] JA Coekin. A versatile presentation of parameters for rapid recognition of total state. In *Proceedings of the IEEE International Symposium on Man-Machine Systems*, 1969.

[31] JR Cole, Q Wang, E Cardenas, J Fish, B Chai, RJ Farris, AS Kulam-Syed-Mohideen, DM McGarrell, T Marsh, GM Garrity, et al. The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic acids research*, 37(suppl 1):D141–D145, 2009.

[32] Robert K Colwell. EstimateS Ver. 8.2: Statistical estimation of species richness and shared species from samples. Freeware published at `http://viceroy.eeb.uconn.edu/EstimateS`, 2010.

[33] WJ Conover and Ronald L Iman. Rank transformations as a bridge between parametric and nonparametric statistics. *The American Statistician*, 35(3):124–129, 1981.

[34] Sandra Dıaz and Marcelo Cabido. Vive la difference: plant functional diversity matters to ecosystem processes. *Trends in Ecology & Evolution*, 16(11):646–655, 2001.

[35] Niklas Elmqvist, Pierre Dragicevic, and Jean-Daniel Fekete. Rolling the dice: Multidimensional visual exploration using scatterplot matrix navigation. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1539–1148, 2008.

[36] Niklas Elmqvist, Andrew Vande Moere, Hans-Christian Jetter, Daniel Cernea, Harald Reiterer, and TJ Jankun-Kelly. Fluid interaction for information visualization. *Information Visualization*, 10(4):327–340, 2011.

[37] Sara Johansson Fernstad, Jimmy Johansson, Suzi Adams, Jane Shaw, and David Taylor. Visual exploration of microbial populations. In *IEEE Symposium on Biological Data Visualization (BioVis)*, pages 127–134. IEEE, 2011.

[38] Maria C Ferreira de Oliveira and Haim Levkowitz. From visual data exploration to visual data mining: A survey. *IEEE Transactions on Visualization and Computer Graphics*, 9(3):378–394, 2003.

[39] Ronald A Fisher. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188, 1936.

[40] David R Foster and John D Aber. *Forests in Time: The Environmental Consequences of 1,000 Years of Change in New England*. Yale University Press, 2006.

[41] Jerry F Franklin. *Cone production by upper-slope conifers*. Pacific Northwest Forest and Range Experiment Station, US Department of Agriculture, 1968.

[42] Jerry F Franklin, Thomas A Spies, Robert Van Pelt, Andrew B Carey, Dale A Thornburgh, Dean Rae Berg, David B Lindenmayer, Mark E Harmon, William S Keeton, David C Shaw, et al. Disturbances and structural development of natural forest ecosystems with silvicultural implications, using douglas-fir forests as an example. *Forest Ecology and Management*, 155(1):399–423, 2002.

[43] Brendan J Frey and Delbert Dueck. Clustering by passing messages between data points. *Science*, 315(5814):972–976, 2007.

[44] Ying-Huey Fua, Matthew O Ward, and Elke A Rundensteiner. Hierarchical parallel coordinates for exploration of large datasets. In *IEEE Visualization*, volume 99, pages 43–50, 1999.

[45] Jed A Fuhrman. Microbial community structure and its functional implications. *Nature*, 459(7244):193–199, 2009.

[46] Angélique Gobet, Simone I Böer, Susan M Huse, Justus EE van Beusekom, Christopher Quince, Mitchell L Sogin, Antje Boetius, and Alban Ramette. Diversity and dynamics of rare and of resident bacterial populations in coastal sands. *The ISME journal*, 6(3):542–553, 2011.

[47] Angélique Gobet, Christopher Quince, and Alban Ramette. Multivariate cutoff level analysis (multicola) of large community data sets. *Nucleic acids research*, 38(15):e155–e155, 2010.

[48] NG Gotelli and AM Ellison. *A Primer of ecological statistics*. Sinauer Associates, 2004.

[49] Lynda Gratton, Andreas Voigt, and Tamara J Erickson. Bridging faultlines in diverse teams. *MIT Sloan management review*, 48(4):22–29, 2007.

[50] David Greenland, Douglas G Goodin, and Raymond C Smith. *Climate variability and ecosystem response at long-term ecological research sites*. Oxford University Press, USA, 2003.

[51] Lance H Gunderson. Ecological resilience–in theory and application. *Annual review of ecology and systematics*, pages 425–439, 2000.

[52] Patricia Gurin, Eric L Dey, Sylvia Hurtado, and Gerald Gurin. Diversity and higher education: Theory and impact on educational outcomes. *Harvard Educational Review*, 72(3):330–367, 2002.

[53] LeGrand H Hardy, Gertrude Rand, and M Catherine Rittler. Tests for the detection and analysis of color-blindness. *Journal of the Optical Society of America*, 35(4):268–271, 1945.

[54] Mark Harmon and Jerry Franklin. Tree growth and mortality measurements in long-term permanent vegetation plots in the Pacific Northwest (LTER Reference Stands). `http://andrewsforest.oregonstate.edu/data/abstract.cfm?dbcode=TV010`, 2012.

[55] David A Harrison and Katherine J Klein. What's the difference? Diversity constructs as separation, variety, or disparity in organizations. *Academy of Management Review*, 32(4):1199, 2007.

[56] Mark Harrower and Cynthia Brewer. Colorbrewer. org: an online tool for selecting colour schemes for maps. *The Cartographic Journal*, 40(1):27–37, 2003.

[57] Sandra G Hart. NASA-task load index (NASA-TLX); 20 years later. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting Annual Meeting*, volume 50, pages 904–908. SAGE Publications, 2006.

[58] John A Hartigan and Beat Kleiner. Mosaics for contingency tables. In *Proceedings of Interface 1981*, pages 268–273, 1981.

[59] Helwig Hauser, Florian Ledermann, and Helmut Doleisch. Angular brushing of extended parallel coordinates. In *Proceedings of IEEE Symposium on InfoVis*, pages 127–130, 2002.

[60] Kris M Havstad, Laura F Huenneke, and William H Schlesinger. *Structure and function of a Chihuahuan desert ecosystem: the Jornada Basin long-term ecological research site*. Oxford University Press, USA, 2006.

[61] Jeffrey Heer and Maneesh Agrawala. Design considerations for collaborative visual analytics. *Information visualization*, 7(1):49–62, 2008.

[62] Jeffrey Heer, Maneesh Agrawala, and Wesley Willett. Generalized selection via interactive query relaxation. In *Proceedings of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, pages 959–968. ACM, 2008.

[63] Jeffrey Heer and Michael Bostock. Crowdsourcing graphical perception: using mechanical turk to assess visualization design. In *Proceedings of the 28th international conference on Human factors in computing systems*, pages 203–212. ACM, 2010.

[64] Jeffrey Heer, Jock Mackinlay, Chris Stolte, and Maneesh Agrawala. Graphical histories for visualization: Supporting analysis, communication, and evaluation. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1189–1196, 2008.

[65] Jeffrey Heer and Ben Shneiderman. Interactive dynamics for visual analysis. *ACM Queue*, 10(2):30, 2012.

[66] Donald L Henshaw, Wade M Sheldon, Suzanne M Remillard, and Kyle Kotwica. ClimDB/HydroDB: a web harvester and data warehouse approach to building a cross-site climate and hydrology database. In *Proceedings of the 7th International Conference on Hydroscience and Engineering*, 2006.

[67] Donald L Henshaw and Gody Spycher. Evolution of ecological metadata structures at the HJ Andrews Experimental Forest Long-Term Ecological Research (LTER) site. In *North American science symposium: toward a unified framework for inventorying and monitoring forest ecosystem resources*, pages 2–6, 1998.

[68] Steven A Highland. *The Historic and Contemporary Ecology of Western Cascade Meadows— Archeology, Vegetation, and Macromoth Ecology.* Ph.D. Dissertation, Oregon State University, 2011.

[69] Helmut Hillebrand. On the generality of the latitudinal diversity gradient. *The American Naturalist*, 163(2):192–211, 2004.

[70] Danny Holten and Jarke J Van Wijk. Evaluation of cluster identification performance for different PCP variants. *Computer Graphics Forum*, 2010.

[71] Michael S Horn, Matthew Tobiasz, and Chia Shen. Visualizing biodiversity with voronoi treemaps. In *Sixth International Symposium on Voronoi Diagrams (ISVD'09)*, pages 265–270. IEEE, 2009.

[72] Stuart H Hurlbert. The nonconcept of species diversity: a critique and alternative parameters. *Ecology*, 52(4):577–586, 1971.

[73] Alfred Inselberg. Multidimensional detective. In *IEEE Symposium on Information Visualization*, pages 100–107, 1997.

[74] Alfred Inselberg. *Parallel coordinates: visual multidimensional geometry and its applications*. Springer, 2009.

[75] Alfred Inselberg and Bernard Dimsdale. Parallel coordinates: a tool for visualizing multi-dimensional geometry. In *Proceedings of Visualization'90*, pages 361–378. IEEE, 1990.

[76] Petra Isenberg, Anastasia Bezerianos, Pierre Dragicevic, and Jean-Daniel Fekete. A study on dual-scale data charts. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2469–2478, 2011.

[77] Sergei Izrailev and Dimitris K Agrafiotis. A method for quantifying and visualizing the diversity of QSAR models. *Journal of Molecular Graphics and Modelling*, 22(4):275–284, 2004.

[78] Brian Johnson and Ben Shneiderman. Tree-maps: A space-filling approach to the visualization of hierarchical information structures. In *Proceedings of the 2nd conference on Visualization'91*, pages 284–291. IEEE Computer Society Press, 1991.

[79] Sherri Johnson and Richard Fredriksen. Stream chemistry concentrations and fluxes using proportional sampling in the Andrews Experimental Forest, 1968 to present. `http://andrewsforest.oregonstate.edu/data/abstract.cfm?dbcode=CF002`, 2012.

[80] Julia Jones and Jerry Franklin. Cone production of upper slope conifers in the Cascade Range of Oregon and Washington. Long-Term Ecological Research. `http://andrewsforest.oregonstate.edu/data/abstract.cfm?dbcode=TV019`, 2012.

[81] Eser Kandogan. Visualizing multi-dimensional clusters, trends, and outliers using star coordinates. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 107–116. ACM, 2001.

[82] Daniel A Keim. Visual Database Exploration Techniques. In *Proceedings of Knowledge Discovery & Data Mining*, 1997.

[83] Daniel A. Keim. Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics*, 8(1):1–8, 2002.

[84] David G Kendall. Abundance matrices and seriation in archaeology. *Probability Theory and Related Fields*, 17(2):104–112, 1971.

[85] Alan K Knapp, John M Briggs, David C Hartnett, and Scott L Collins. *Grassland dynamics: long-term ecological research in tallgrass prairie.* Oxford University Press New York, NY, USA, 1998.

[86] Robert Kosara, Fabian Bendix, and Helwig Hauser. Parallel sets: Interactive exploration and visual analysis of categorical data. *IEEE Transactions on Visualization and Computer Graphics*, 12(4):558–568, 2006.

[87] Charles J Krebs. *Ecological methodology.* Harper & Row New York, 1989.

[88] Heidi Lam, Enrico Bertini, Petra Isenberg, Catherine Plaisant, and Sheelagh Carpendale. Empirical studies in information visualization: Seven scenarios. *IEEE Transactions on Visualization and Computer Graphics*, 18(9):1520–1536, 2012.

[89] James R Larson Jr. *In search of synergy in small group performance.* Psychology Press, 2010.

[90] Dora C Lau and J Keith Murnighan. Demographic diversity and faultlines: The compositional dynamics of organizational groups. *Academy of Management Review*, 23(2):325–340, 1998.

[91] William K Lauenroth and Ingrid C Burke. *Ecology of the shortgrass steppe: a long-term perspective.* Oxford University Press, USA, 2008.

[92] Lawrence R Lawlor. Overlap, similarity, and competition coefficients. *Ecology*, pages 245–251, 1980.

[93] Jeffrey LeBlanc, Matthew O Ward, and Norman Wittels. Exploring n-dimensional databases. In *Proceedings of the 1st conference on Visualization'90*, pages 230–237. IEEE Computer Society Press, 1990.

[94] Bongshin Lee, Cynthia Sims Parr, Dana Campbell, and Benjamin B Bederson. How users interact with biodiversity information using taxontree. In *Proceedings of the working conference on Advanced visual interfaces*, pages 320–327. ACM, 2004.

[95] Jing Li, Jean-Bernard Martens, and Jarke J Van Wijk. Judging correlation from scatterplots and parallel coordinate plots. *Information Visualization*, 9(1):13–30, 2010.

[96] Gene E Likens and F Herbert Bormann. *Biogeochemistry of a forested ecosystem.* Number Ed. 2. Springer-Verlag New York Inc., 1995.

[97] Bertram Ludäscher, Ilkay Altintas, Chad Berkley, Dan Higgins, Efrat Jaeger, Matthew Jones, Edward A Lee, Jing Tao, and Yang Zhao. Scientific workflow management and the Kepler system. *Concurrency and Computation: Practice and Experience*, 18(10):1039–1065, 2006.

[98] David L MacAdam. Visual sensitivities to color differences in daylight. *Journal of the Optical Society of America*, 32(5):247–273, 1942.

[99] Jock Mackinlay. Automating the design of graphical presentations of relational information. *ACM Transactions on Graphics (TOG)*, 5(2):110–141, 1986.

[100] John J Magnuson, Timothy K Kratz, and Barbara J Benson. *Long-term dynamics of lakes in the landscape: long-term ecological research on North Temperate lakes.* Oxford University Press, USA, 2005.

[101] Anne E Magurran. *Measuring biological diversity.* Wiley-Blackwell, 2003.

[102] C Wayne Martin and R Dennis Harr. Precipitation and streamwater chemistry from undisturbed watersheds in the cascade mountains of oregon. *Water, Air, and Soil Pollution*, 42(1-2):203–219, 1988.

[103] C Wayne Martin and R Dennis Harr. Logging of mature douglas-fir in western oregon has little effect on nutrient output budgets. *Canadian Journal of Forest Research*, 19(1):35–43, 1989.

[104] Bertolt Meyer and Andreas Glenz. Team faultline measures: A computational comparison and a new approach to multiple subgroups. *Organizational Research Methods*, In press.

[105] Bertolt Meyer, Meir Shemla, and Carsten C Schermuly. Social category salience moderates the effect of diversity faultlines on information elaboration. *Small Group Research*, 42(3):257–282, 2011.

[106] Jeffrey Miller. Spatial and temporal distribution and abundance of moths in the Andrews Experimental Forest. `http://andrewsforest.oregonstate.edu/data/abstract.cfm?dbcode=SA015`, 2005.

[107] Jeffrey C Miller and Paul C Hammond. Lepidoptera of the Pacific Northwest: caterpillars and adults. *Lepidoptera of the Pacific Northwest: caterpillars and adults*, 2003.

[108] Audris Mockus, Roy T Fielding, and James D Herbsleb. Two case studies of open source software development: Apache and Mozilla. *ACM Transactions on Software Engineering and Methodology (TOSEM)*, 11(3):309–346, 2002.

[109] Tamara Munzner. A nested model for visualization design and validation. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):921–928, 2009.

[110] Daniel Nettle. Explaining global patterns of language diversity. *Journal of anthropological archaeology*, 17(4):354–374, 1998.

[111] Seán I O'Donoghue, Anne-Claude Gavin, Nils Gehlenborg, David S Goodsell, Jean-Karim Hériché, Cydney B Nielsen, Chris North, Arthur J Olson, James B Procter, David W Shattuck, et al. Visualizing biological datanow and in the future. *Nature methods*, 7:S2–S4, 2010.

[112] Oladele Ogunseitan. *Microbial diversity: form and function in prokaryotes.* Blackwell Publishing, 2005.

[113] Chadwick Dearing Oliver, Bruce C Larson, et al. *Forest stand dynamics.* McGraw-Hill, Inc., 1990.

[114] Jason Pearlman, Penny Rheingans, and Marie des Jardins. Visualizing diversity and depth over a set of objects. *IEEE Computer Graphics and Applications*, 27(5):35–45, 2007.

[115] Wei Peng, Matthew O Ward, and Elke A Rundensteiner. Clutter reduction in multi-dimensional data visualization using dimension reordering. In *Proceedings of the IEEE Symposium on Information Visualization*, pages 89–96. IEEE Computer Society, 2004.

[116] Owen L Petchey and Kevin J Gaston. Functional diversity (FD), species richness and community composition. *Ecology Letters*, 5(3):402–411, 2002.

[117] Owen L Petchey and Kevin J Gaston. Functional diversity: back to basics and looking forward. *Ecology letters*, 9(6):741–758, 2006.

[118] DPC Peters, CM Laney, AE Lugo, SL Collins, CT Driscoll, PM Groffman, J Morgan Grove, AK Knapp, TK Kratz, MD Ohman, et al. Long-term trends in ecological systems: a basis for understanding responses to global change. *US Department of Agriculture, Agricultural Research Service*, 2011.

[119] Tuan Pham, Rob Hess, Crystal Ju, Eugene Zhang, and Ronald Metoyer. Visualization of diversity in large multivariate data sets. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1053–1062, 2010.

[120] Tuan Pham, Steven Highland, Ronald Metoyer, Donald Henshaw, Jeffrey Miller, and Julia Jones. Interactive visualization of spatial and temporal patterns of diversity and abundance in ecological data. In *Proceedings of Environmental Information Management*. Publisher of University of California, 2011.

[121] Tuan Pham, Julia Jones, Ronald Metoyer, Frederick Swanson, and Robert Pabst. Interactive visual analysis promotes exploration of long-term ecological data. *EcoSphere*, 2013 (In Press).

[122] Tuan Pham, Ronald Metoyer, Katerina Bezrukova, and Chester Spell. "Show Me the Cracks in Our Teams": Visual Representations of Demographic Diversity Faultlines. In *IEEE Information Visualization Poster Abstract*, 2012.

[123] Tuan Pham, Ronald Metoyer, Katerina Bezrukova, and Chester Spell. "Show Me the Cracks in Our Teams": Visualization of Team Faultlines. Technical Report (February 2013), 2013.

[124] Evelyn C Pielou. *Ecological diversity*. Wiley New York, 1975.

[125] Alexander Pilhofer, Alexander Gribov, and Antony Unwin. Comparing clusterings using Bertin's idea. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2506–2515, 2012.

[126] Catherine Plaisant, Ben Shneiderman, Khoa Doan, and Tom Bruns. Interface and data architecture for query preview in networked information systems. *ACM Transactions on Information Systems (TOIS)*, 17(3):320–341, 1999.

[127] John Porter, Peter Arzberger, Hans-Werner Braun, Pablo Bryant, Stuart Gage, Todd Hansen, Paul Hanson, Chau-Chin Lin, Fang-Pang Lin, Timothy Kratz, et al. Wireless sensor networks for ecology. *BioScience*, 55(7):561–572, 2005.

[128] Alban Ramette. Multivariate analyses in microbial ecology. *FEMS Microbiology Ecology*, 62(2):142–160, 2007.

[129] Charles Redman and David R Foster. *Agrarian Landscapes in Transition: Comparisons of Long-Term Ecological & Cultural Change.* Oxford University Press, USA, 2008.

[130] Karsten Rink, Thomas Fischer, Benny Selle, and Olaf Kolditz. A data exploration framework for validation and setup of hydrological models. *Environmental Earth Sciences*, pages 1–9, 2013.

[131] Jonathan C Roberts. State of the art: Coordinated & multiple views in exploratory visualization. In *Fifth International Conference on Coordinated and Multiple Views in Exploratory Visualization (CMV'07)*, pages 61–71. IEEE, 2007.

[132] Yvonne Rogers, Helen Sharp, and Jenny Preece. Interaction design: Beyond human computer interaction. 2007.

[133] Howard L Sanders. Marine benthic diversity: a comparative study. *American Naturalist*, pages 243–282, 1968.

[134] Douglas Schuler and Aki Namioka. *Participatory design: Principles and practices.* CRC, 1993.

[135] Michael Sedlmair, Andrada Tatu, Tamara Munzner, and Melanie Tory. A taxonomy of visual cluster separation factors. In *Computer Graphics Forum*, volume 31, pages 1335–1344. Wiley Online Library, 2012.

[136] Jinwook Seo and Ben Shneiderman. A rank-by-feature framework for interactive exploration of multidimensional data. *Information Visualization*, 4(2):96–113, 2005.

[137] Mark Servilla, Duane Costa, Christine Laney, Inigo San Gil, and James Brunt. The EcoTrends web portal: an architecture for data discovery and exploration. In *Proceedings of the Environmental Information Management Conference*, pages 10–11, 2008.

[138] Moshe Shachak, James R Gosz, Stewart TA Pickett, and Avi Perevolotsky. *Biodiversity in drylands: toward a unified framework.* Oxford University Press, USA, 2004.

[139] Claude E Shannon and Warren Weaver. The mathematical theory of information. *Urbana: University of Illinois Press*, 97, 1949.

[140] Ben Shneiderman. Tree visualization with tree-maps: 2D space-filling approach. *ACM Transactions on graphics (TOG)*, 11(1):92–99, 1992.

[141] Ben Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings of Symposium on Visual Languages*, pages 336–343. IEEE, 1996.

[142] Ben Shneiderman. Creativity support tools: accelerating discovery and innovation. *Communications of the ACM*, 50(12):20–32, 2007.

[143] Ben Shneiderman and Catherine Plaisant. Strategies for evaluating information visualization tools: multi-dimensional in-depth long-term case studies. In *Proceedings of the 2006 AVI workshop on BEyond time and errors: novel evaluation methods for information visualization*, pages 1–7. ACM, 2006.

[144] Mike Sips, Boris Neubert, John P Lewis, and Pat Hanrahan. Selecting good views of high-dimensional data using class consistency. In *Computer Graphics Forum*, volume 28, pages 831–838. Wiley Online Library, 2009.

[145] Mitchell L Sogin, Hilary G Morrison, Julie A Huber, David Mark Welch, Susan M Huse, Phillip R Neal, Jesus M Arrieta, and Gerhard J Herndl. Microbial diversity in the deep sea and the underexplored rare biosphere. *Proceedings of the National Academy of Sciences*, 103(32):12115–12120, 2006.

[146] Robert Spence. *Information Visualization: Design for Interaction.* Prentice Hall, 2007.

[147] John Stasko, Richard Catrambone, Mark Guzdial, and Kevin McDonald. An evaluation of space-filling information visualizations for depicting hierarchical structures. *International Journal of Human-Computer Studies*, 53(5):663–694, 2000.

[148] John Stasko and Eugene Zhang. Focus+context display and navigation techniques for enhancing radial, space-filling hierarchy visualizations. In *IEEE Symposium on Information Visualization (InfoVis 2000)*, pages 57–65. IEEE, 2000.

[149] Sherry MB Thatcher, Karen A Jehn, and Elaine Zanutto. Cracks in diversity research: The effects of diversity faultlines on conflict and performance. *Group Decision and Negotiation*, 12(3):217–241, 2003.

[150] Sherry MB Thatcher and Pankaj C Patel. Group faultlines a review, integration, and guide to future research. *Journal of Management*, 38(4):969–1009, 2012.

[151] James J Thomas and Kristin A Cook. A visual analytics agenda. *IEEE Computer Graphics and Applications*, 26(1):10–13, 2006.

[152] Warren S Torgerson. Multidimensional scaling: I. theory and method. *Psychometrika*, 17(4):401–419, 1952.

[153] Anne Treisman. Preattentive processing in vision. *Computer vision, graphics, and image processing*, 31(2):156–177, 1985.

[154] John W Tukey. *Exploratory Data Analysis*. Addison-Wesley, 1977.

[155] Daan Van Knippenberg, Carsten KW De Dreu, Astrid C Homan, et al. Work group diversity and group performance: An integrative model and research agenda. *Journal of applied psychology*, 89(6):1008–1022, 2004.

[156] Colin Ware. *Information visualization: perception for design*. Morgan Kaufmann, 2004.

[157] RM Warwick and KR Clarke. Taxonomic distinctness and environmental assessment. *Journal of Applied Ecology*, 35(4):532–543, 1998.

[158] Robert H Whittaker. Dominance and Diversity in Land Plant Communities: Numerical relations of species express the importance of competition in community function and evolution. *Science*, 147(3655):250, 1965.

[159] Robert H Whittaker. Evolution and measurement of species diversity. *Taxon*, pages 213–251, 1972.

[160] Hadley Wickham, Dianne Cook, Heike Hofmann, and Andreas Buja. Graphical inference for infovis. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):973–979, 2010.

[161] Hadley Wickham and Heike Hofmann. Product plots. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2223–2230, 2011.

[162] Forrest W Young and Robert M Hamer. *Multidimensional scaling: History, theory, and applications*. L. Erlbaum Associates Hillsdale, NJ, 1987.