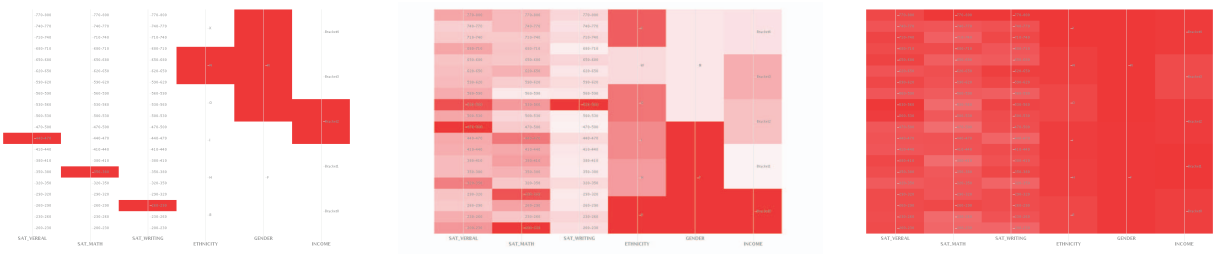
Visualization of Diversity in Large Multivariate Data Sets

Tuan Pham, Rob Hess, Crystal Ju, Eugene Zhang, and Ronald Metoyer



(a) Very low diversity (b) Medium diversity (c) Very high diversity

Fig. 1. 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.

Abstract—Understanding the diversity of a set of multivariate objects is an important problem in many domains, including ecology, college admissions, investing, machine learning, and others. However, to date, very little work has been done to help users achieve this kind of understanding. Visual representation is especially appealing for this task because it offers the potential to allow users to efficiently observe the objects of interest in a direct and holistic way. Thus, in this paper, we attempt to formalize the problem of visualizing the diversity of a large (more than 1000 objects), multivariate (more than 5 attributes) data set as one worth deeper investigation by the information visualization community. In doing so, we contribute a precise definition of diversity, 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. Our primary contribution, however, is a visual representation, called the *Diversity Map*, for visualizing diversity. 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.

Index Terms—Information visualization, diversity, categorical data, multivariate data, evaluation.

1 Introduction

The concept of diversity presents itself in many domains. 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 interplay between the physical characteristics of an environment (e.g. water levels, temperature, elevation, rainfall, etc.) and the diversity of species present there [23]. In both of these cases, many variables may be considered, and a general starting point is to examine the overall distribution of samples (i.e. applicants or species) over the attributes of interest. Many other domains 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 (i.e. they wish to understand the diversity of their training examples), and chemists are interested in assessing the similarity/diversity of a collection of molecular models in exploring the multitude of designs generated by simulations [1].

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 5 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 [30, 37, 18], 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 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 [10]. A visual encoding of the data offers the potential to allow this kind of direct observation, but this approach will only be useful if a representation is available in which the diversity of the data is readily apparent. While visually understanding the diversity of objects over a single attribute is fairly straightforward and is supported by representations such as histograms or Tukey box plots, very little work has been done to develop representations that specifically emphasize the diversity of a multi-dimensional data set.

In this paper, we attempt to formalize the problem of visualizing the diversity of a large, multivariate data set as one that warrants deeper attention by the information visualization community. Our primary contribution is a visual representation called the Diversity Map (Fig. 1), which is specifically intended to help users understand the diversity of a large set of multivariate objects. The Diversity Map 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 biological diversity, a set of requirements for diversity visualizations based on this definition, and a

• *Tuan Pham, Rob Hess, Crystal Ju, Eugene Zhang, and Ronald Metoyer are with the School of Electrical Engineering and Computer Science, Oregon State University, E-mail: {pham, hess, juji, zhange, metoyer}@eecs.oregonstate.edu*

Manuscript received 31 March 2010; accepted 1 August 2010; posted online 24 October 2010; mailed on 16 October 2010.
For information on obtaining reprints of this article, please send email to: tvcg@computer.org .

design for a formal user study intended to understand the effectiveness of a visual representation in communicating diversity information. We evaluate the Diversity Map by using this study design to compare it to Pearlman *et al.*'s visualization [25], the only other representation specifically designed to visualize diversity. In comparing user performance between Pearlman *et al.*'s representation and the Diversity Map, we show that users can consistently and as or more accurately judge elements of diversity using the Diversity Map.

The rest of this paper is organized as follows. We begin in Section 2 with a precise definition of diversity derived from the definition of species diversity used by ecologists, and we lay out a set of requirements for diversity visualizations based on this definition. We then discuss related work in Section 3 in the context of the presented definition and requirements. In Section 4, we present and discuss the Diversity Map representation in the context of what is known about human perception. In Section 5, we describe a formal study designed to understand the effectiveness of a visual representation in communicating diversity information, and in Section 6, we evaluate the Diversity Map representation using this study design. Finally, we discuss the merits and shortcomings of the Diversity Map representation, suggest directions for future work, and draw our conclusions.

2 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 samples 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"* [23].

Two measures of diversity are used in Ecology: *richness*, which is simply the number of species in the unit of study represented out of all possible species; and *evenness*, which describes the variability in species abundances [23]. 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.

Beyond our definition of diversity, we also borrow several conventions from the study of biodiversity. Specifically, we adopt individual objects as our unit of measure, and, as in the study of biodiversity, 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.

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 [35], we understand that this is difficult for large attribute spaces).

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, as defined earlier. We focus only on representation methods and organize our review based on the taxonomy proposed by Keim *et al.* [16].

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 [29]. 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 small multiples may solve some of these problems. For example, scatter plot matrices may provide useful representations of diversity, especially for high and low diversity cases, but intermediate values may be difficult to disambiguate due to data overlap. While jittering techniques 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). Small multiples in matrix form, however, require screen space that grows with the square of the number of attributes. Small 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 4, is in fact a small multiple histogram representation with an alternative encoding that facilitates communication of overall diversity.

Alternatively, rank/abundance—or Whittaker—plots [37] 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.

3.2 Geometrically Transformed Displays

Geometrically transformed displays map one object to a set of points and lines in 2D or 3D space [16]. This category includes graph visualizations and coordinate-based visualizations. While graph-based visualizations are important in many domains, 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 [12, 11] and related variants [9, 17], and star coordinates [14].

Parallel coordinates [12, 11] are well-suited to visualizing various types of multivariate data (quantitative, ordinal, or nominal) and revealing data correlation between attributes. However, visual clutter becomes a limitation as the number of objects increases. Refinements to parallel coordinates have attempted to address visual clutter with brushing [9], clustering [3, 7], and dimension reordering [26].

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 [9, 17].

In one variant, a histogram is overlaid lengthwise on each parallel axis [9], 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.

The Parallel Sets [17] 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 [14] 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.

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 [4], 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[15], 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) [20, 31, 13]. 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 [2, 34, 38]. While these approaches are popular in many domains 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.

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 [25], 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). 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, and we revisit it in Section 6 where we formally compare its ability to communicate diversity information to that of our Diversity Map representation.

4 The Diversity Map Representation

To address the shortcomings of previous approaches, we developed a novel representation called the *Diversity Map* for visualizing the diversity of a set of objects. In this representation, depicted in Fig. 1, 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 the Diversity Map by placing a semi-transparent 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 *buckets*.

To satisfy the requirement from Section 2 that all objects are treated equally, each semi-transparent rectangle contributes an equal, fractional amount of opacity to the bucket 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 buckets corresponding to attribute values not represented in the visualized data set are fully transparent (i.e. $\alpha = 0$ in RGBA color space), and the bucket(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 bucket is calculated based on the ratio of the number of objects in that bucket to the number of objects in the bucket corresponding to the most abundant attribute value. We have empirically found that using the square-root of the number of objects per bucket in this calculation helps to make buckets corresponding to attribute values with low abundance more recognizable. Specifically, the opacity of each bucket x is calculated as $\alpha(x) = \sqrt{|x|/|x_{MAX}|}$, where $|x|$ denotes the number of objects in bucket x, and x_{MAX} is the bucket with the most objects for the attribute in question. Figure 2 illustrates the process of visualizing a single attribute using the Diversity Map.

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. 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.

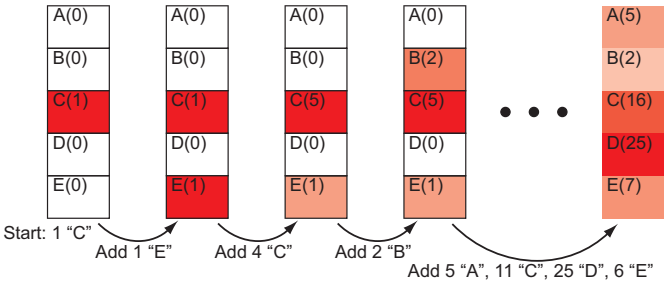


Fig. 2. The process of visualizing a single attribute using the Diversity Map. The depicted attribute has five possible values (A, B, C, D, and E). The visualization begins with a single object with attribute value “C,” and objects with other attribute values are added in subsequent steps. At each step, the number of objects in each bucket is shown in parentheses next to the bucket's label and the opacity (α-value) of each bucket 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 Fig. 1(b).

4.1 Design Considerations

As indicated earlier, our primary goal in designing the Diversity Map 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 bucket, 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 bucket 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 literature as the most effective encoding for quantitative information [22, 5], it is easy to justify its use in our design. However, several other quantitative encodings, such as length, angle, slope, and area, rank higher than opacity in [22] and [5]. 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 the Diversity Map representation, richness of variety is expressed by the number of buckets with non-zero opacity, and evenness of abundance is expressed by the uniformity of the color distribution across the buckets 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 buckets it will yield, and the more even is the abundance across the data set, the more uniform will be the colors of the buckets.

The overall diversity of a given data set—that is, the combined diversity of all its attributes—is communicated by the Diversity Map as the overall color density of the entire visual region: as the visualized data set becomes more richly various and more evenly abundant, more buckets 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 Figs. 1 (a) and (c).

Finally, we note that the Diversity Map is specifically designed to provide a holistic overview of the population sample of interest. As Shneiderman notes, [32], 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 [28]. 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 the Diversity Map is designed to serve as just such a gateway.

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.

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 [30, 37, 18]. 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 biodiversity, X corresponds to species; in the more general case, it could correspond to any single attribute) is

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

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 biodiversity, 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 2 that all attributes are considered as equal, we have adapted a variant of the Shannon index known as the *evenness measure* [27], which normalizes the value of $H(X)$ by its maximum possible value:

$$H_{max}(X) = -\sum_{i=1}^{S} \frac{1}{S}\log\frac{1}{S} = \log S. \tag{2}$$

Thus, the evenness of attribute X is

$$H_E(X) = \frac{H(X)}{H_{max}(X)} = -\frac{1}{\log S}\sum_{i=1}^{S} p(x_i)\log p(x_i). \quad (3)$$

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 instantiated 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.

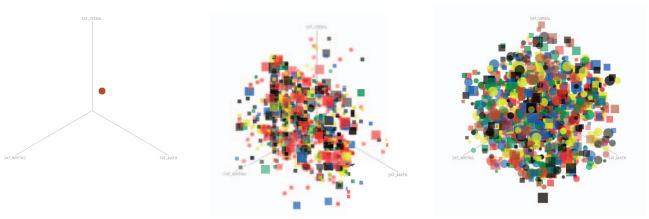
5.2 User Study Questions

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 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.



(a) (b) (c)

Fig. 3. 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 Fig. 1.

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.

6 Evaluation of the Diversity Map Representation

In this section, we use the formal user study described in the previous section to evaluate the effectiveness of the Diversity Map representation (DM; Fig. 1) at conveying diversity information by comparing it to the Glyph Hybrid representation [25] (GH; Fig. 3) discussed in Section 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 small multiples, such as multiple histograms.

Here we describe the specific implementation of the user study outlined in Section 5 that we used to compare the DM and GH representations, and we analyze and discuss the results of this study.

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 domain because it is one of the three domains examined as a case study by Pearlman *et al.* in [25] and because we believed it would be a domain with which our participants, who were all university students, would be familiar.

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

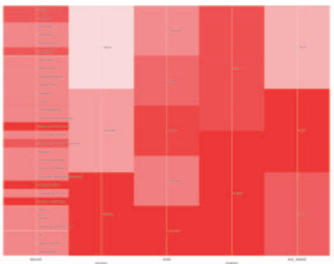


Fig. 4. Participants of the user study visualized using Diversity Map. The visualized attributes, from left to right, are major, degree, year in school, gender, and age-range. The participants represented a diverse range of majors, degrees, and ages.

Table 1. Allocation of 40 participants across 4 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)

posted around the campus. They represented a diverse range of majors, degrees, and ages (Fig. 4), 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.

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 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 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 using that method. After completing the tutorial, the participant answered several questions of each of the types described in Section 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 for all studies 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.

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 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 [6] to the response data before using these techniques.

Our primary focus in analyzing the results of the study is on error

Table 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. 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)

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 2 and 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 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 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.

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 Fig. 3(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 Fig. 3(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. 1 (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.

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 [33], 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 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.

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.

Table 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)

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 buckets (e.g. ethnicity) to that of an attribute with only a few buckets (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 individual attributes.

7 Discussion and Future Work

We have presented 1) an infrastructure for studying the problem of diversity visualization and 2) a novel representation for visualizing the diversity of a large set of multivariate objects. 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 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 small multiple histograms. We show that the Diversity Map allows users to consistently and as or more accurately judge elements of diversity than the only other existing method designed to visualize diversity. While we believe we have taken a positive step in understanding diversity visualization, there are several issues left to address.

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*. Future work will address the interactive capabilities of DM.

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.* [25], 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 is ambiguous). The evaluation study would benefit from explicit attention to the ability to communicate richness.

Limitations of DM. The Diversity Map representation (DM) 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, the user can constrain a single attribute to one or more particular values (buckets) 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 bucket in attribute X are objects most/least diverse in attribute Y?".

In future work, we will explore other interaction features including user-defined orderings for nominal-valued attributes, user-defined ordering of the attribute axes themselves, and user specification of attribute value ranges (over multiple buckets) of interest for interactive filtering. In addition, we plan to investigate mechanisms for constructing sets of objects of a desired diversity. Finally, we also plan to explore the advantages of small multiples of DMs for trend analysis.

While the Diversity Map 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 buckets for any one attribute is also similarly limited, and it is not clear how "small" a bucket can be made before the representation becomes ineffective. Studies to understand these limitations are left for future work.

The Diversity Map 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 Fig. 1(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 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 in blue than in red [21]. Additionally, we empirically used the square root-based normalization in determining the color opacity (α-value) of buckets to help make buckets 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 [23]. 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 [36].

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. In future work, we will investigate a definition to account for this interaction. The area of business management may provide useful insights as researchers in that field discuss diversity across multiple attributes [19, 8].

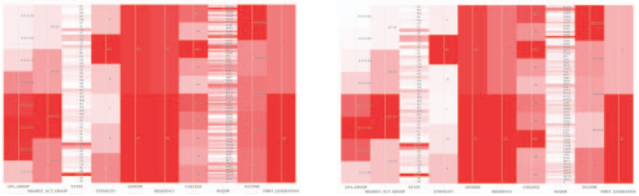


Fig. 5. A real data set of 2550 college applicants with 10 attributes visualized using the Diversity Map. Left: the subset of students recommended for acceptance based on a holistic admissions process implemented by a proprietary software package and designed to produce a diverse incoming class. Right: 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.

Application to Real Data. We also intend to explore the application of Diversity Maps 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. The DM visualizations of this data set are shown in Fig. 5. These results are promising in that they agree with the output of the holistic evaluation software.

We are also in the process of deploying an interactive version of the Diversity Map for use by ecologists at our university. Ecologists commonly collect species inventory data and analyze it to better understand the interactions between the environment and the species under study. For example, species distribution modeling is used to relate environmental covariates to each species. Data sets of this type are often challenging in many ways. In the Oregon State University H.J. Andrews Experimental Forest, researchers have collected data on 606 moth species by sampling $\sim$200 sites across a $\sim$100 km^2 region of study every summer week over a period of 23 years [24]. This data is interesting because moths are indicators of broader biological diversity in plant types and physical environments. Diversity measures such as the Shannon Index provide little insight into relationships that may be present in this data between the species and the environment because, in reducing diversity to a single number, they conceal a tremendous amount of information.

8 Conclusions

The Diversity Map represents a first attempt to design a representation with the specific goal of visualizing diversity as we have defined it in this paper. While, to date, little attention has been paid to this problem, we hope that this work will serve to provide a foundation for future studies into the design and evaluation of visualization methods for exploring the increasingly important area of diversity.

Acknowledgments

The authors wish to thank Onyekwere Ogba and Nicholas Hubbert for their development efforts supported in part by the CRA-W/CDC Distributed Research Experiences for Undergraduates (DREU) program. In addition we would like to thank Dr. Margaret Burnett, Dr. Juan Gilbert, Dr. Julia Jones, Dr. Jeff Miller, and Steven Highland for numerous discussions regarding our user study, application to admissions data, and application to moth diversity data. Finally we would like to thank our user study participants for their time and input. This work was supported in part by NSF IIS-0546881.

REFERENCES

[1] A method for quantifying and visualizing the diversity of qsar models. *Journal of Molecular Graphics and Modelling*, 22(4):275 – 284, 2004.

[2] T. Anderson. *An introduction to multivariate statistical analysis*. Wiley New York, 1958.

[3] A. Artero, M. de Oliveira, and H. Levkowitz. Uncovering clusters in crowded parallel coordinates visualizations. In *Proceedings of the IEEE Symposium on Information Visualization*, pages 81–88. IEEE Computer Society, 2004.

[4] H. Chernoff. The use of faces to represent points in k-dimensional space graphically. *Journal of the American Statistical Association*, pages 361–368, 1973.

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

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

[7] Y. Fua, M. Ward, and E. Rundensteiner. Hierarchical parallel coordinates for exploration of large datasets. In *IEEE Visualization*, volume 99, pages 43–50, 1999.

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

[9] H. Hauser, F. Ledermann, and H. Doleisch. Angular brushing of extended parallel coordinates. In *Proceedings of the IEEE Symposium on Information Visualization*, pages 127–130, 2002.

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

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

[12] A. Inselberg and B. Dimsdale. Parallel coordinates: a tool for visualizing multi-dimensional geometry. In *Proceedings of the 1st conference on Visualization'90*, page 378. IEEE Computer Society Press, 1990.

[13] B. Johnson and B. 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 Los Alamitos, CA, USA, 1991.

[14] E. 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 New York, NY, USA, 2001.

[15] D. Keim. Visual Database Exploration Techniques. In *Proc. Tutorial Int. Conf. on Knowledge Discovery & Data Mining, Newport Beach, CA*, 1997.

[16] D. Keim. Information visualization and visual data mining. *IEEE transactions on Visualization and Computer Graphics*, pages 1–8, 2002.

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

[18] C. Krebs. *Ecological methodology*. Harper & Row New York, 1989.

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

[20] J. LeBlanc, M. Ward, and N. Wittels. Exploring n-dimensional databases. In *Proceedings of the 1st conference on Visualization'90*, page 237. IEEE Computer Society Press, 1990.

[21] D. MacAdam. Visual sensitivities to color differences in daylight. *J. Opt. Soc. Am*, 32:247–273, 1942.

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

[23] A. Magurran. *Measuring biological diversity*. Wiley-Blackwell, 2003.

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

[25] J. Pearlman, P. Rheingans, and M. des Jardins. Visualizing diversity and depth over a set of objects. *IEEE Computer Graphics and Applications*, pages 35–45, 2007.

[26] W. Peng, M. Ward, and E. Rundensteiner. Clutter reduction in multidimensional data visualization using dimension reordering. In *Proceedings of the IEEE Symposium on Information Visualization*, pages 89–96. IEEE Computer Society, 2004.

[27] E. Pielou. *Ecological diversity*. Wiley New York, 1975.

[28] C. Plaisant, B. Shneiderman, K. Doan, and T. Bruns. Interface and data architecture for query preview in networked information systems. *ACM Trans. Inf. Syst.*, 17(3):320–341, 1999.

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

[30] C. Shannon and W. Weaver. The mathematical theory of information. *Urbana: University of Illinois Press*, 97, 1949.

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

[32] B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *IEEE Visual Languages*, pages 336–343, 1996.

[33] J. Stasko, R. Catrambone, M. Guzdial, and K. McDonald. An evaluation of space-filling information visualizations for depicting hierarchical structures. *International Journal of Human Computer Studies*, 53(5):663–694, 2000.

[34] W. Torgerson. Multidimensional scaling: I. Theory and method. *Psychometrika*, 17(4):401–419, 1952.

[35] A. Treisman. Preattentive processing in vision. In *Papers from the second workshop Vol. 13 on Human and Machine Vision II*, page 334. Academic Press Professional, Inc., 1986.

[36] C. Ware. *Information visualization: perception for design*. Morgan Kaufmann, 2004.

[37] R. 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.

[38] F. Young and R. Hamer. *Multidimensional scaling: History, theory, and applications*. L. Erlbaum Associates Hillsdale, NJ, 1987.