Evaluation of Multivariate Visualizations: A Case Study of Refinements and User Experience

Mark A. Livingston* and Jonathan W. Decker
Naval Research Laboratory, 4555 Overlook Ave SW, M/S 5580, Washington, DC, USA 20375

ABSTRACT

Multivariate visualization (MVV) aims to provide insight into complex data sets with many variables. The analyst's goal may be to understand how one variable interacts with another, to identify potential correlations between variables, or to understand patterns of a variable's behavior over the domain. Summary statistics and spatially abstracted plots of statistical measures or analyses are unlikely to yield insights into spatial patterns. Thus we focus our efforts on MVVs, which we hope will express key properties of the data within the original data domain. Further narrowing the problem space, we consider how these techniques may be applied to continuous data variables.

One difficulty of MVVs is that the number of perceptual channels may be exceeded. We embarked on a series of evaluations of MVVs in an effort to understand the limitations of attributes that are used in MVVs. In a follow-up study to previously published results, we attempted to use our past results to inform refinements to the design of the MVVs and the study itself. Some changes improved performance, whereas others degraded performance. We report results from the follow-up study and a comparison of data collected from subjects who participated in both studies. On the positive end, we saw improved performance with Attribute Blocks, a MVV newly introduced to our on-going evaluation, relative to Dimensional Stacking, a technique we were examining previously. On the other hand, our refinement to Data-driven Spots resulted in greater errors on the task. Users' previous exposure to the MVVs enabled them to complete the task significantly faster (but not more accurately). Previous exposure also yielded lower ratings of subjective workload. We discuss these intuitive and counter-intuitive results and the implications for MVV design.

Keywords: User study, multivariate visualization, visual analytics

1. INTRODUCTION

One challenge created by the ever-increasing ability to collect data is the issue of how to display large and diverse data sets. Meta-data, such as uncertainty about measurements, can add another layer of complexity to data. Automated techniques can perform standard numerical analysis routines such as correlation, dimension reduction, and clustering; these summary statistics or aggregated variables provide some basic insights to the data. However, more complex relationships in the data may remain hidden during these simple analysis techniques. Human intelligence and pattern recognition capabilities are still needed to discover unexpected relationships in data. One response to the explosion of variables has been the creation of multivariate visualization (MVV) techniques (Figure 1), in which several variables are integrated into a single visualization, which distinguishes the variables through a variety of graphical parameters. The goal is to enable understanding of the details of several individual variables as well as the pattern of relationships between the variables.

Figure 1: Five multivariate visualization (MVV) techniques evaluated in the experiment described in this paper.

*mark.livingston@nrl.navy.mil; phone 1 202 767-0380; fax 1 202 404-1192;

Research on MVVs benefits from a user-centered approach, in which both perceptual and cognitive studies of human capabilities to discern subtle differences in graphical patterns and larger patterns of data within visualizations could lead to guidelines on how MVV techniques could improve human performance on a wide variety of data-intensive tasks. One long-term goal in our work is to determine how many variables can be presented in a visualization and still enable users to discover new insights. We hope that these relationships need not be explicitly represented, thus avoiding the need to fully understand the data in order to present it in the best manner for a task that may or may not be itself fully understood a priori. One may conceive of user-centered design of MVVs to be an optimization problem in a very large parameter space. MVVs have numerous parameters in the design space which may affect user performance on a visual task. We report results of a follow-up study to our previous exploration of MVVs. The results were surprising both in some numerical measures and in the behavior of subjects.

2. PREVIOUS WORK

A number of MVV techniques have been devised to display such data. However, evaluation of the utility of such techniques for general data sets remains difficult. Thus most techniques are studied on only one data set and task. Another criticism that may be offered for some evaluations of MVVs is that the task doesn't require the presence of multiple variables; the analyst would be best served by focusing only on one variable at a time. Further, since most MVVs have parameters that may be adjusted for a particular visualization, a comparison between techniques could be criticized as comparing only the specific instance of a technique and not the technique in general. Our previous studies[1,2] suffered from several of these difficulties. We briefly review techniques and evaluations, with a focus on the drawbacks noticed in our previous work.

2.1 Multivariate Visualizations

Brush strokes[3] compose a texture inspired by impressionist paintings; attributes of length, width, orientation, intensity, and hue enable five variables to be encoded in this MVV. Strokes are placed randomly over the surface. One difficulty of this technique is that the parameters do not have the same resolution. Intensity and hue will have more perceptible output levels than length and width of strokes, due to properties of display hardware and human perception. We further observed that wide strokes can appear to be blurred. Earlier glyph-based representations include *stick figures*, which used a torso-and-limb structure to encode data values with relative angles of the torso to the display and limbs to the torso[4].

Data-driven spots (DDS)[5] encode each data layer with Gaussian kernels on a randomly jittered grid. The layers are differentiated by the size and hue of the spots, while intensity encodes the data value. Animation of the layers over the surface further separates layers perceptually and synthesizes resolution beyond that created by the size and separation of the spots, albeit perhaps by raising a conflict with the jitter pattern. *Color weaving* similarly works on the same concept of separating colored elements so that multiple, overlapping features can be identified in the same spatial region[6].

The *oriented slivers*[7] technique encodes each data layer with short, grayscale lines on a randomly jittered grid. The orientation differentiates the data layers; the intensity encodes the data values. Similar to DDS and brush strokes, sliver density affects the frequency of the underlying data which may be reliably understood. Further, high sliver density, great length, or great width may prevent the user from distinguishing slivers. Still, the technique has the advantage of using few perceptually significant features, allowing the potential for many data layers to be visualized.

Early multivariate visualizations used simple shapes or shape patterns, such as small, adjacent blocks or sectors of a circle. Each shape represents one value in the data; a cluster of such shapes or a complete shape represents a multi-valued sample point. Individual values were typically depicted through the intensity (in a grayscale implementation) or hue (in a color implementation). One critical design decision is how to convey the resolution of the data. In *dimensional stacking*, the data range was divided into bins, with a finite set of gray levels or hues assigned to the bins. Given the limited resolution of human perception of intensity and color, this technique may be more valuable for showing gross differences than for fine details. This technique exhibits a problem in that a finite region is required to show the multiple data values at a single point of the domain.

Attribute blocks[8] build on early visualizations that use a cluster of shapes or a divided shape to represent multi-valued sample points[9-11]. Each attribute may be visualized with a continuous variable, such as color or intensity; variables are separate by their location within the cluster or shape. Dynamically changing the array's configuration and the size and origin of the individual components enable synthesizing higher resolution than the initial sampling of multivariate data.

Color blending is perhaps the oldest and conceptually simplest MVV. Each variable is assigned a particular color; the value of each pixel is computed to be the weighted sum of the colors, with the weights derived from the data values. Thus the dominant hue of a pixel or region in visualization indicates the greatest component value among the data values at that location. Each pixel is a visual sample, so the spatial resolution is equal to that of the display device, but the color resolution of modern display hardware limits the effectiveness of this technique for more than three variables.

2.2 Evaluations of Multivariate Visualizations

A growing number of studies have explored the qualitative and quantitative evidence of the usefulness of MVV for a variety of tasks. Several user studies have examined the perceptual properties of individual techniques. Height and density of vertical bars over a 2D domain were easily identified, but certain combinations with background elements (such as salience or regularity of samples in a dense field) made it hard to understand the data[12]. Brush strokes (using color, texture, and feature hierarchies among luminance, hue, and texture) enabled verification[13] that perceptual guidelines for visualization[14] apply to non-photorealistic visualizations as well. Oriented slivers[7] enabled users to perceptually separate layers within a data set. To get the best performance on identifying the presence of a constant rectangular target in a constant background field required a minimum separation of 15° between layers.

A key type of evaluation is testing task performance with a MVV. DDS enabled users to discern boundaries amongst as many as eight layers of data[5]. Art-inspired techniques such as pointillism, speed lines, opacity, silhouettes, and boundary enhancement enabled users to track a feature over time more accurately and with a subjective preference[15]. Adding colors and altering texture properties such as line thickness or orientation in line-integral convolution created effective visualizations for multiple flow fields, as assessed by domain experts[6]. Ellipsoid glyphs were effective at showing tensor structure in diffusion tensor images, whereas layered brush strokes encoded field values and enabled users to understand relationships between layers, albeit with a potential for cluttered images[16]. This was not a serious problem because the application displayed a number of dependent variables (data layers).

Other studies have compared multiple, diverse visualization techniques. Users performed better when a visualization explicitly represented a feature sought[17], e.g. showed the sign of vectors in the field, represented integral curves, and showed critical point locations. Experts and non-experts did not show significant differences. Maintaining the source data colors, as in color weaving, outperformed color blending[18]; this difference increased with the number of data layers. Color selection for the various scales was a critical issue in the blending methods. Multi-layer texture synthesis enabled users to perform with no significant difference from brush strokes for weather data visualization[19]. In a previous study, we found[1] that the parameterized patterns of DDS and oriented slivers helped users perform critical point (maximum) detection more accurately and faster than glyph representations of brush strokes and stick figures. We also found that some techniques were sensitive to monitor settings (brightness and contrast) and room lighting conditions. On a trend detection task, we found[2] that DDS and separate grayscale visualizations outperformed the other techniques mentioned for accuracy, but not for response time. We noted several possible improvements to the visual representations and the interface, which prompted the revised study described in Section 3.

3. REVISED STUDY DESIGN

3.1 Implementations of Multivariate Visualizations

The following section summarizes how we applied a technique to our data and gives the descriptions and hints on the trend localization task that subjects read in our study. The task was to find the county (region) exhibiting the greatest increase or greatest decrease within a five-year time span. On each trial, a technique legend (if applicable) was provided for the subjects' reference (Figure 2). Refer to the description and key to the cropped images provided in Figure 1 in addition to the figures in this section. All MVV images were generated by offline rendering and displayed at 1024x1024 pixels on a Dell 3008WFP monitor with factory default settings and resolution set to 2560x1600. The juxtaposed maps were each 512x512 pixels. The study took place in an office with standard fluorescent lighting.

Juxtaposed Maps

Our baseline technique used a series of grayscale maps, each of which corresponded to a single variable. To localize the trend, subjects had to scan the maps. The intensity (brightness) indicated the data value; brighter pixels indicated higher values. So if the county's brightness increased across the maps, the trend was increasing. If the county's brightness was decreasing, the trend was decreasing. In the original study, subjects selected their answers by clicking only on an empty map at the lower right of the map array (Figure 2, right). Their selection was highlighted by a blue border. For our

follow-up study, we added the ability to select a county from any map. We hypothesized that more direct interaction would help the users respond faster.

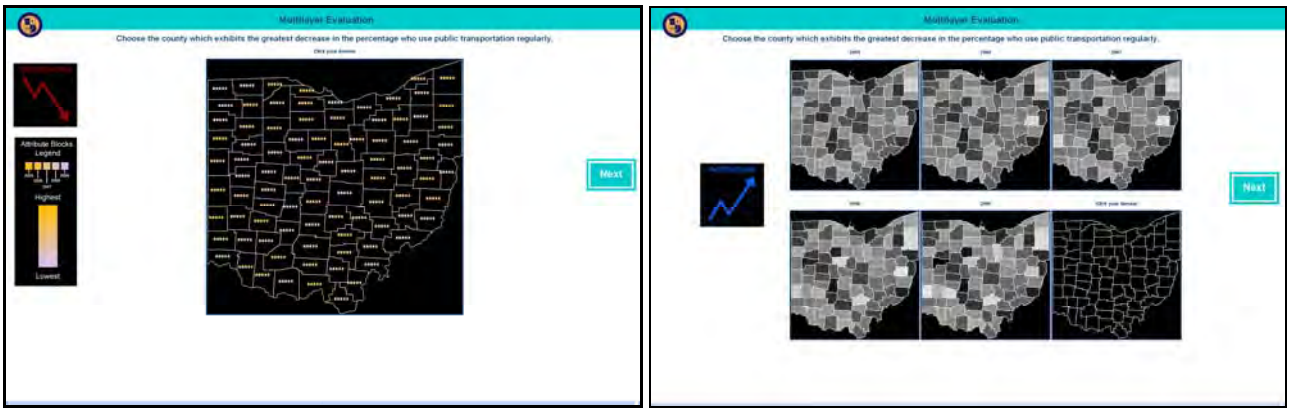

Figure 2 *Left:* Data trial layout for integrated MVVs, showing an icon indicating trial type (red arrow, decreasing) and a technique legend at the left, the MVV in the center, and a trial submit button at the right. *Right:* Data trial layout for juxtaposed maps, showing the trial type indicator (blue arrow, increasing), the map array, and the submit button.

Brush Strokes

The chief design issue for brush strokes was mapping parameters to variables. Our previous work showed that the final values were most helpful, so we mapped intensity and hue to the fourth and fifth data values, respectively. These two were exchanged in the mapping (Figure 3) from our previous study. Length and width were more subtle than hue and intensity; the range of (horizontal) stroke width was six to twelve pixels (0.51°-1.02°) and of stroke length, 25 to 49 pixels (2.13°-4.16°). Thus we hypothesized that users would find the initial value harder to interpret than the final; this could be exacerbated by county boundaries cutting strokes. Stroke orientation (third value) was horizontal for zero; a stroke tilted 135° clockwise from horizontal (slanting down to the left) represented the maximum value for any county in the current map. Strokes that were dim and blue, but long and wide, indicated decreasing trends; bright-and-gold but short-and-narrow strokes indicated increasing trends. Since a value could start in the middle, the target and other trends could have a more subtle pattern. (See Figure 1, left.)

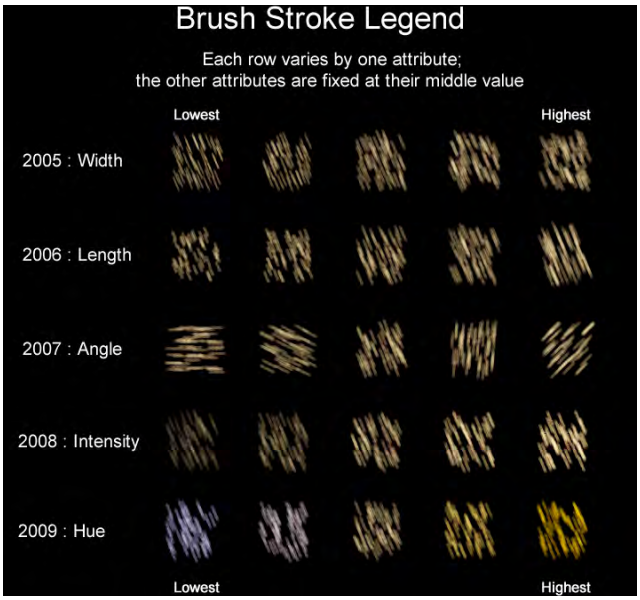

Figure 3: The legend for brush strokes show the mapping of graphical attributes and the range of visual parameters for each one. Hue and intensity swapped years between studies.

Data-Driven Spots

In our previous DDS implementation, we created a set of jittered grids that avoided the possibility of kernels from different layers overlapping each other. Our implementation in other respects was no more restrictive or permissive than the original DDS technique. The standard deviations of the Gaussian kernels were 25 (red, initial), 25, 17, 17, and 13 (green, final) pixels, for the five data layers (2.13°, 1.45°, and 1.11°, respectively). The technique legend (Figure 4, left) mapped the size and color of spots to years. The spots were dim if the value was low, while the spots were bright if the value was high. So if the red spots appear brightest, followed by the brown, purple, blue, and then green, the trend was decreasing. Conversely, if red spots were dimmest, but the brown, purple, blue, and green got progressively brighter, then the trend was increasing.

In revising the study, we proposed to study whether these grid-like layers were too distracting for the user; thus, we replaced this grid-based layout with random placement. We also decreased the overall sizes of the spots so that small regions were represented by spots that were more numerous and from all the layers. The new Gaussian kernel sizes were as follows: 8 (red, initial), 10, 12, 14, 16 (green, final) pixels. These sizes were slightly more uniformly distributed relative to one another than in the previous study, and increased in size rather than decreased in size over the data span.

Oriented Slivers

With oriented slivers, we did not prevent overlap of slivers from different data layers; thus, our implementation is neither more restrictive nor more permissive than the original guidelines for the technique. In this study, the initial year (2005) was specified by the vertical sliver. Each subsequent year was represented by a sliver rotated 36° clockwise from the previous layer's sliver (Figure 4, second from left), for angles of 90°, 54°, 18°, -18°, and -54°. The slivers were three pixels wide and twenty pixels long (0.26°x1.70°) in the vertical orientation; any change from this was due to anti-aliasing or county boundaries. If the slivers for a particular county got progressively brighter, then the county's value was increasing. If they got progressively dimmer, then the county's value was decreasing.

In the original study, the pixels where slivers overlapped were brighter, since we simply added the individual sliver layers together in the compositing of data layers. In this study, we instead computed the maximum intensity value at each pixel. This helped reduce misleading intensity within the images. Also, the previous study's images were not rendered with sufficient anti-aliasing, and these artifacts were distracting and gave the impression of brighter intensity than what was intended. We removed these artifacts by running a Gaussian filter on the final image.

Color Blending

The color blending scheme deployed for the original study used the same color set as DDS. The colors were blended by w_i v_i, where w={0.1111,0.1111,0.1111,0.2222,0.4445} and v={*data*}. The colors {red, orange, purple, blue, green} were specified in CIELAB by *L*=50, *b*=50, and *a*={-95,-45,0,45,95}. The weights w_i v_i were renormalized so that the *L* and *b* parameters were unchanged. In each region, the dominant color corresponded to the highest value at those pixels. Thus, if the tint were more towards the red and brown and less towards the blue and green, then the trend was decreasing. If the tint were more towards green and less towards red, then the trend was increasing.

In our follow-up study, we replaced the five distinct colors we used previously with a heat map derived from the line in RGB color space between {0,49,0} and {255,0,45}. We also replaced the weights defined above with the vector w_i v_i, where now w={0.30,0.16,0.08,0.16,0.30} and v={*data*}. The *L*, *a*, and *b* parameters, as specified in CIELAB, remained unchanged. We made these changes in an attempt to aid the subjects in reading the color as an indication of trend direction and magnitude. The new equation produces a color representative of the trend direction (green for increasing trends, red for decreasing). However, trend magnitude was still quite difficult to determine. To further benefit our subjects, we created a new key which showed a sample of the color result for varying trend magnitudes and directions (Figure 4, second from right). Notice that intensity gave no clear indication of trend magnitude, and in fact the subject was unable to differentiate a large trend from a small trend which started and stopped at a weighted end of the scale. At least with this legend, our subjects were aware of the potential ambiguity in the technique.

Dimensional Stacking, replaced by Attribute Blocks

In our previous study, we evaluated dimensional stacking; its poor performance motivated us to replace it with attribute blocks. Since the techniques look similar, we describe them in sequence. Our color version of dimensional stacking used five discrete values for the field values; thus we used the same color set as for DDS and color blending, although it indicated data value (bin) rather than data layer (year). Each bin contained 20% of the values over the five-valued data for a trial; red for the lowest values, then brown, purple, blue, and finally green for the highest values. Our blocks were 9x11 pixels (0.77°x0.94°), and the set was arranged horizontally. Color progression across the row indicated trends.

Our attribute blocks also consisted of a horizontal row of blocks which represented each one data layer. The main difference is that our attribute blocks used a blue-to-yellow heat map to indicate the field value (Figure 4, right). Here, if the sequential blocks changed from blue to yellow, we had an increasing trend; likewise, if they changed from yellow to blue, there was a decreasing trend. As for the magnitude of the trend, this maps to the range of color within the row of blocks. For example, if the colors changed from full blue to a full yellow, that represented a large increasing trend.

3.2 User Study

In this section, we describe the design of the user study we conducted, with an emphasis on those aspects that were changed from our previous study[2].

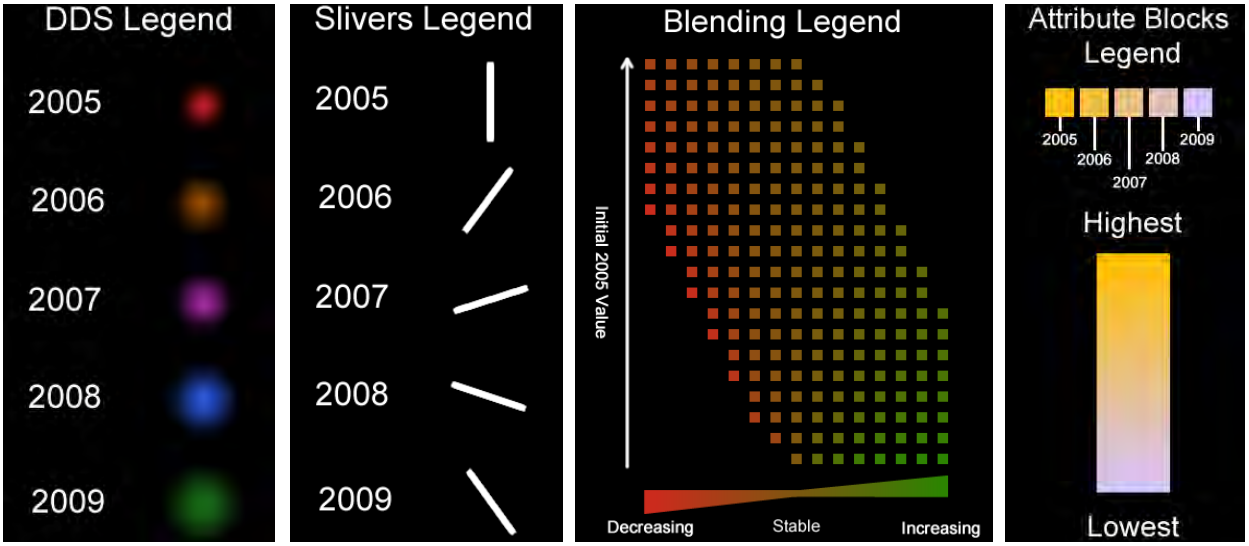

Figure 4: The legends for DDS, oriented slivers, color blending, and attribute blocks showed the mapping of graphical parameters to years in DDS and oriented slivers, and to values for color blending and attribute blocks. Note in particular the legend for color blending, in which we tried to capture the ambiguity inherent in the color mapping that weighted five colors for a three-color display.

Task

The task of localizing the greatest increase or greatest decrease was retained; users were shown synthetic data for Indiana or Ohio counties. These data sets were simulated to be five years of survey data on basic demographic questions and consumer electronics opinions. We used the same formulas to generate the data: initial year values were randomly selected in [0.3,0.7], a target was selected from among counties near the mean county area, and its final value was computed (initial value + 0.3 for greatest increase, initial value – 0.3 for greatest decrease). Distracter counties were selected (for both targets): up to five with trend size 0.2, up to ten with trend size 0.1, and up to 20 with trend size 0.05. There was a target for the greatest increase and one for the greatest decrease in every set, along with corresponding distracters. All other counties were assigned "noise" values: an increase or decrease in the range [-0.03,0.03].

Independent and Dependent Variables

The independent variables from our previous study were maintained: *visualization technique*, *trend type*, and the *distance* (by county adjacency) to the nearest distracter of size (including undefined if no distracters of size 0.2 were present). The values for visualization technique were changed slightly, as noted above: attribute blocks were substituted for dimensional stacking, and the parameters of other techniques were changed[20] in the hopes of improving performance. We added a new independent variable of *returning subject* (values: true, false). The dependent variables were identical: *error* (difference between target size of 0.3 and trend size of selected county), *response time*, and *subjective workload* for each technique (measured with NASA Task-load Index[21]).

Subject Procedures

Users were asked to identify the county with the greatest increase or greatest decrease over the five-year span of data, with a note that this may or may not be the county with (respectively) the maximum or minimum final value. We divided the trials by MVV technique so that subjective workload ratings could be measured. Within a block of trials for a technique, users saw a general tutorial with hints (summarized in Section 3.1) about how to identify trends with that technique, two tutorial questions which gave immediate feedback of the correct answer, and then two blocks of ten trials. In our previous study, the polarity (increasing or decreasing) of the question alternated between trials. Because several users in our previous study made mistakes on the polarity of the trial (picking the greatest trend in the wrong direction),

we divided the polarity by ten-question sets, alternating the order of (increasing, decreasing) with a Latin square. Unfortunately, this resulted in our first surprising "result" of the study: users made more mistakes of looking for the wrong trend type than in our previous study. Previously, we saw 37 such errors, or 1.7% of the trials. With the revised design and the addition of a large icon (256x256 on a screen of 2560x1600; show in Figure 2) to indicate the desired trend type, we assumed we would see fewer such errors. Instead, we saw a few subjects make this mistake for entire sets of ten questions, and an overall count of 83, or 3.8%. While a few users noted after their session that they had made this error, apparently our design did not encourage participants to pay careful attention to their task. We are confident that when they were looking for the correct type of trend, they were performing the task to the best of their ability, and due to their infrequency, these outliers were not removed from the data points used for analysis. Doing so would change the repeated measures design to a more permissive non-repeated measures design. (Such an analysis is planned as future work, for comparison.)

Eighteen subjects (thirteen male, five female) completed the experiment; with six visualizations, two trend types, and an effort (but not a requirement) to have one of each possible distance of a distracter of trend size 0.2 in {1..9,undefined}, that yielded 18x6x2x10=2160 data points. Subjects ranged in age from 27 to 66 (average = 42) and were drawn from the technical and clerical staff of our laboratory. Nine users were returning subjects from our previous study; nine were new to the visualization techniques. All reported being moderate to heavy computer users; none reported any color blindness.

Hypotheses

With the adjustments reported above, we expected to see decreased error with all of the techniques that were revised, and that attribute blocks would have lower error than dimensional stacking. We further expected the change in the interface for juxtaposed maps to results in faster response time for that technique. Regarding returning subjects, we expected their error to decrease and their response time to decrease, a standard learning effect.

4. RESULTS

We analyzed the accuracy and response times with a repeated measures ANOVA, using a 6 (visualization technique) × 2 (trend type) x 10 (distracter distance) design, with returning subject as a between-subjects variable. We also checked for interactions between the visualization techniques and trend type (increasing, decreasing). We analyzed the subjective workload data with a 6-way ANOVA. We report the F-statistic with its two degrees of freedom and p-values.

We found a main effect of distracter distance [$F(9,144)=2.964$, $p=0.003$]; subjects were most accurate when there was no distracter (distance=undefined) and were more accurate when a distracter was near the target, but not monotonically so (Figure 5). Distracter distance also had a main effect on response time [$F(9,144)=2.658$, $p=0.007$]. Users were fastest when a distracter was far away, with the cases of no such distracter or an immediately adjacent distracter next fastest (Figure 5). We saw no main effect of returning [$F(1,16)=1.090$, $p=0.312$].

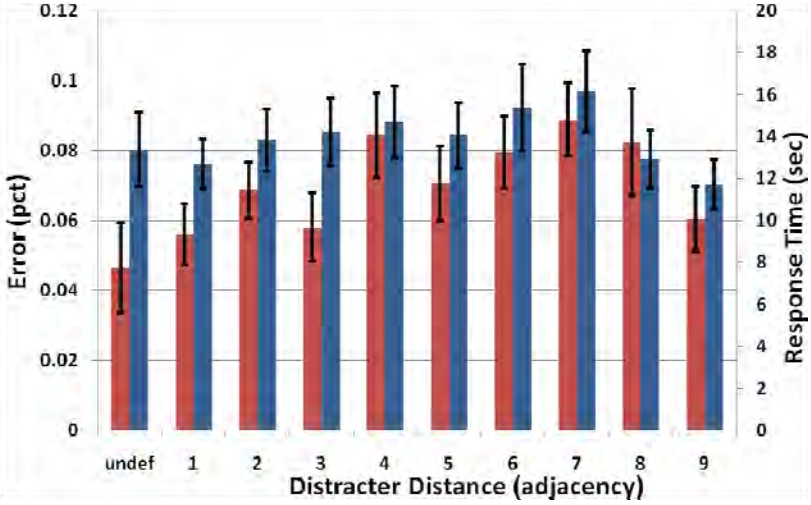

Figure 5: Distracter distance had a main effect on both error (light red bars) and response time (dark blue bars); users were more accurate when there was no distracter or it was close to the target, but fastest when a distracter was very distant. In all result graphs, error bars indicate the standard error, red hues indicate error, and blue hues indicate response time.

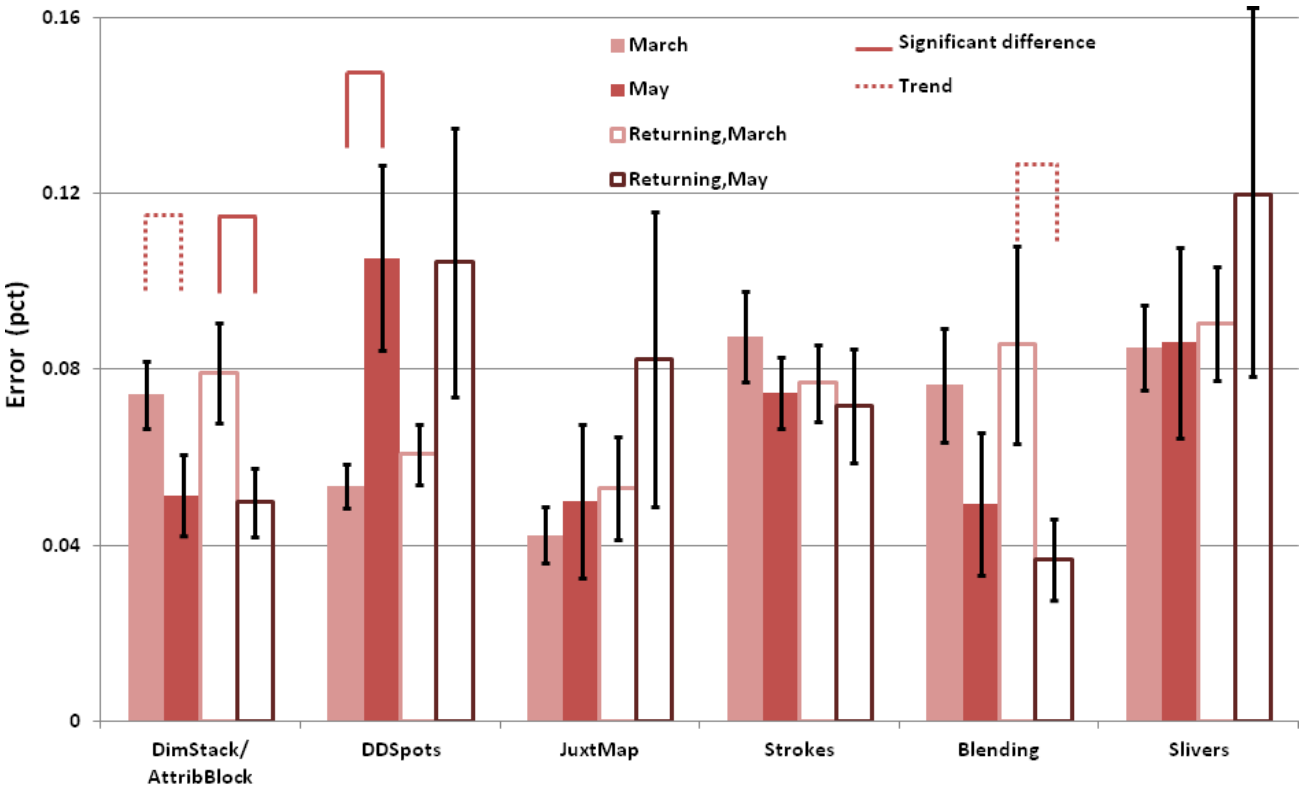

Figure 6: Performance changes with regard to error, separated by visualization technique, included improvement for attribute blocks over dimensional stacking and a loss of performance for DDS. Filled bars compare all data of the two studies; outlined bars compare performance of only subjects who participated in both. Coupling bars indicate significant differences (solid) or trends (dashed).

We found no main effect of visual technique [F(5,80)=1.332, p=0.259] or trend type [F(1,16)=1.383, p=0.257] on error. This was surprising, as it contrasted with our previous findings[2] that subjects were previously more accurate with juxtaposed maps and DDS than with the other techniques, and they had been more accurate with decreasing trends. We regard the lack of a significant main effect of trend type as reflecting improvements in the visualization techniques, since the cases of increasing and decreasing trends should be symmetric and have roughly equal error.

Regarding the individual techniques (Figure 6), we saw some statistically significant results using Welch's t-test. Most interesting for us was the comparison of attribute blocks in the new study against dimensional stacking in the previous study. We saw a trend for attribute blocks to cause less error [t(68)=1.910, p=0.060]. When restricting the analysis only to the returning subjects who saw both techniques, we did see a significant difference [t(30)=2.148, p=0.040], although to be fair, some of this improvement could be due to a learning effect. We saw a significant *increase* in error with DDS [t(39)=2.410, p=0.021], largely due to an increase in error on decreasing trends. Clearly, our adjustment of the technique was not only ineffective at improving it, it served to reduce performance. There were no significant differences between the accuracy with the old and new versions of oriented slivers, color blending, brush strokes, or juxtaposed maps.

Response time saw several other main effects. There was a main effect of Returning [F(1,16)=8.801, p=0.009]; users taking the study for the second time (ignoring the changes in the study design and visualization parameters) were approximately 40% faster (though not significantly different in error). Subjects were significantly faster with color blending [F(5,80)=17.392, p=0.000]. This result echoes a main effect found in our previous study, which also saw users complete the task most quickly with color blending. There was a significant interaction between visualization technique and returning; color blending was used at approximately the same speed by both returning and new participants, contrary to the general trend of returning participants to be faster (Figure 7). There was also a significant interaction between visualization technique and trend type for response time [F(5,80)=5.741, p=0.000]. A slightly different interaction occurred than in our previous study, however. Previously, only color blending was notably faster for decreasing trends than increasing trends. This was the case again for color blending, but it was also true for attribute blocks. Further,

juxtaposed maps were now slightly slower for decreasing trends (Figure 8), which is a reversal from our previous study, when juxtaposed maps were slightly slower for increasing trends.

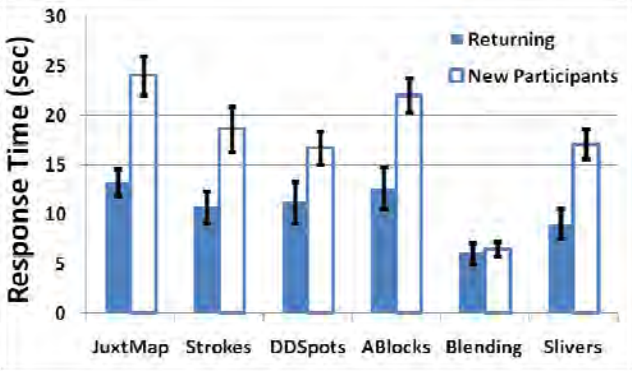

Figure 7: The main effect of visual technique on response time was largely due to the speed with color blending.

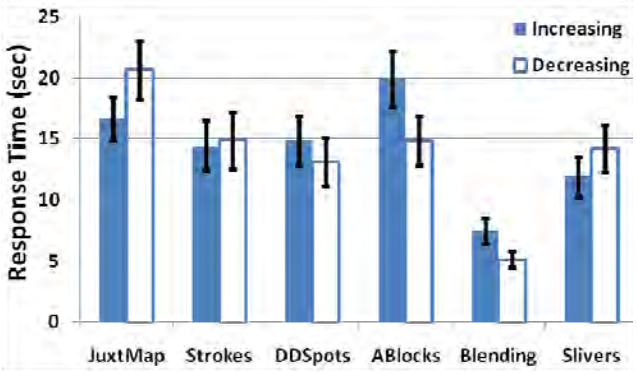

Figure 8: The interaction between visual technique and trend type for response time showed that color blending and attribute blocks were faster for decreasing trends, while juxtaposed maps were slower for decreasing trends.

Turning to individual techniques, we again saw some statistical results of our changes (Figure 9). Subjects required significantly less response time for attribute blocks than they had for dimensional stacking, in addition to their decreased error. This was true both for all subjects [t(42)=2.600, p=0.013] and for the subjects who took part in both experiments [t(29)=3.225, p=0.003]; presumably, the stronger result for the latter group includes a learning effect. Combined with the results on error showing (respectively) a trend and a significant decrease for these two groups, this gives clear evidence that attribute blocks were more usable than dimensional stacking. We saw a trend for a decrease in response time for juxtaposed maps when restricting the analysis to those subjects who participated in both experiments [t(26)=1.744, p=0.093]. Similarly, there was a trend for oriented slivers to require less response time for returning subjects in the second experiment than those subjects required in the first experiment [t(28)=1.984, p=0.057]. While these are hopefully both encouraging results for (respectively) the flexibility in the interface and the error-fixes in the rendering, we had expected to see stronger results.

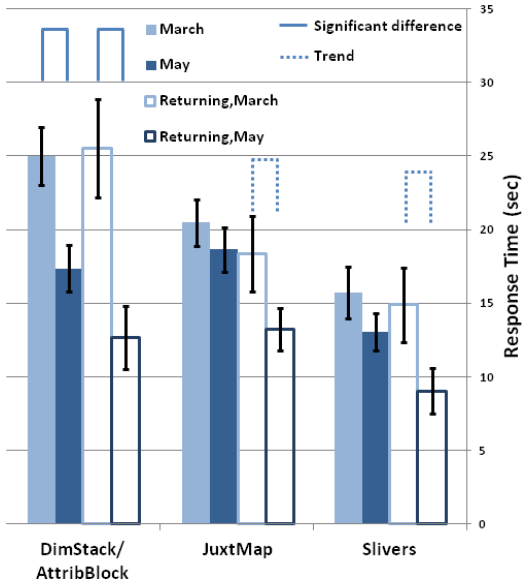

Figure 9: Comparison of response times between the two experiments showed a significant decrease in the response time needed for attribute blocks compared to dimensional stacking, and trends for decreased response times for juxtaposed maps and oriented slivers.

Contrary to our previous study, there was no main effect of visualization technique on subjective workload [F(5,80)=1.889, p=0.105]. There was a main effect of returning on workload [F(1,16)=5.782, p=0.029]. Returning participants gave significantly lower workload ratings than new participants. This was true across all visual techniques.

We also conducted an ANOVA using the data from the returning participants in the two studies. The independent variables for this were the visual technique, trend type, and session (values= {first, second}). Distracter distance was not considered in this analysis. Returning subjects showed an interaction between trend type and session for error

[F(1,8)=11.051, p=0.010]. Returning participants improved their performance between sessions on increasing trends, but declined in accuracy with decreasing trends. There was still a main effect of trend type on error [F(1,8)=11.331, p=0.010]. On the combined data set, returning participants were slightly better on decreasing trends (as had occurred in the previous study, but the difference was smaller). No further effects were found for returning participants for response time beyond those captured above. Returning participants lowered their subjective workload ratings [F(1,8)=6.155, p=0.038]. Considering only the returning subjects, but in both studies, we did see a main effect of visual technique on workload [F(5,40)=3.435, p=0.011]. It appears that the new subjects made this effect disappear in the second study.

5. DISCUSSION

We expected Figure 6 and Figure 9 to provide the most interesting results of the current study by comparing the "improved" MVV techniques against the versions we used in our previous study. Clearly, some of our efforts to improve the techniques were unsuccessful, while others did in fact lead to better performance as measured by error and/or response time. We hypothesized that attribute blocks would improve over dimensional stacking; the discrete representation of the field value in dimensional stacking clearly has a disadvantage compared to the (digital approximation to) continuous representation used by attribute blocks. While we do not have sufficient statistical power to assert potential equality, it appears that attribute blocks are competitive with our previous implementation of DDS and with juxtaposed maps. Further data collection should confirm whether this is statistically likely to occur.

Perhaps the most surprising result in the *loss* in accuracy with DDS. Previously, it was among the best techniques; now, it appears to be the weakest with regard to error. The changes we made (a more random jitter pattern, smaller dots, and differentiating the dots by size) clearly hurt the technique. Of these changes, we note that the first two had the effect of reducing the overall intensity (Figure 10, lower row). We believe this explains why the technique was hurt moderately on increasing trends but significantly on decreasing trends (Figure 10, upper left). Ware[14] concludes that high-contrast (e.g. black) boundaries around colored regions can help make them distinct; thus the space between the spots should have improved performance. We interpret our results as indicating that overall intensity was more important than the contrast. Intensity was already low for decreasing trends; our changes made them even darker. Ware also cautions about the relative nature of lightness and contrast effects. This is another potential limitation on the usefulness of the boundaries.

We are also surprised that our implementation of attribute blocks so outperformed the revised DDS implementation. DDS uses color to identify separate variables in a jittered pattern of Gaussian kernels; if these kernels are jittered just right, the result can be something that looks strikingly like arrangements proposed for implementations of attribute blocks (Figure 10, upper right). On the other hand, it may be that the resulting colors are not perceptually balanced for the change in intensity which DDS uses to denote value to appear perceptually the same for all variables. The identical heat maps we used for our attribute blocks would allow us to test this hypothesis. Another possibility is that our DDS implementation resulted in multiple randomly jittered kernels within each county that corresponded to a single variable. By contrast, our attribute blocks contained one linear array of blocks per county (one for each variable). Perhaps the repetition in DDS confused users.

Another surprise was the failure to improve performance with oriented slivers. We thought the change to maximum intensity projection for the combining of layers of slivers would improve the technique by removing the bright spots that appeared at points of overlap between slivers of difference orientations. While the artifact was successfully removed, we did not find a significant improvement in the performance (although again we lack sufficient statistical power to assert that there was in fact no difference). Ware advises that small objects should have high saturation; this is not a property of oriented slivers in low-value regions. One could hypothesize for future implementations that a suitable minimum value for the mapping of intensity to oriented slivers should be identified.

Similarly, we did not expect to see any change in the error with juxtaposed maps; we expected only faster response time. We merely enabled users to click on any of the component maps to select a county without changing the visual representation. While returning subjects did get faster, new subjects were barely faster than the response times in our earlier study. It appears that the interface didn't matter as much as experience with the technique. We also note that the mean error of returning subjects increased (though not by a statistically significant amount). Perhaps the change was simply too disruptive, or users concentrated less. The change in the weighting scheme we made for color blending did appear to have its intended effect, as the overall mean error was lower (though not by a statistically significant amount), and there was a trend for returning subjects to improve their performance with the technique, which was a reflection of

the statistically significant improvement returning subjects showed on increasing trends. It would appear that our new weighting scheme and/or new legend were helpful in improving returning subjects' performance on this trend type. However, new subjects had a similar disparity between trend types, so we interpret our data as demonstrating a learning effect. Again, further testing could enable us to verify this hypothesis.

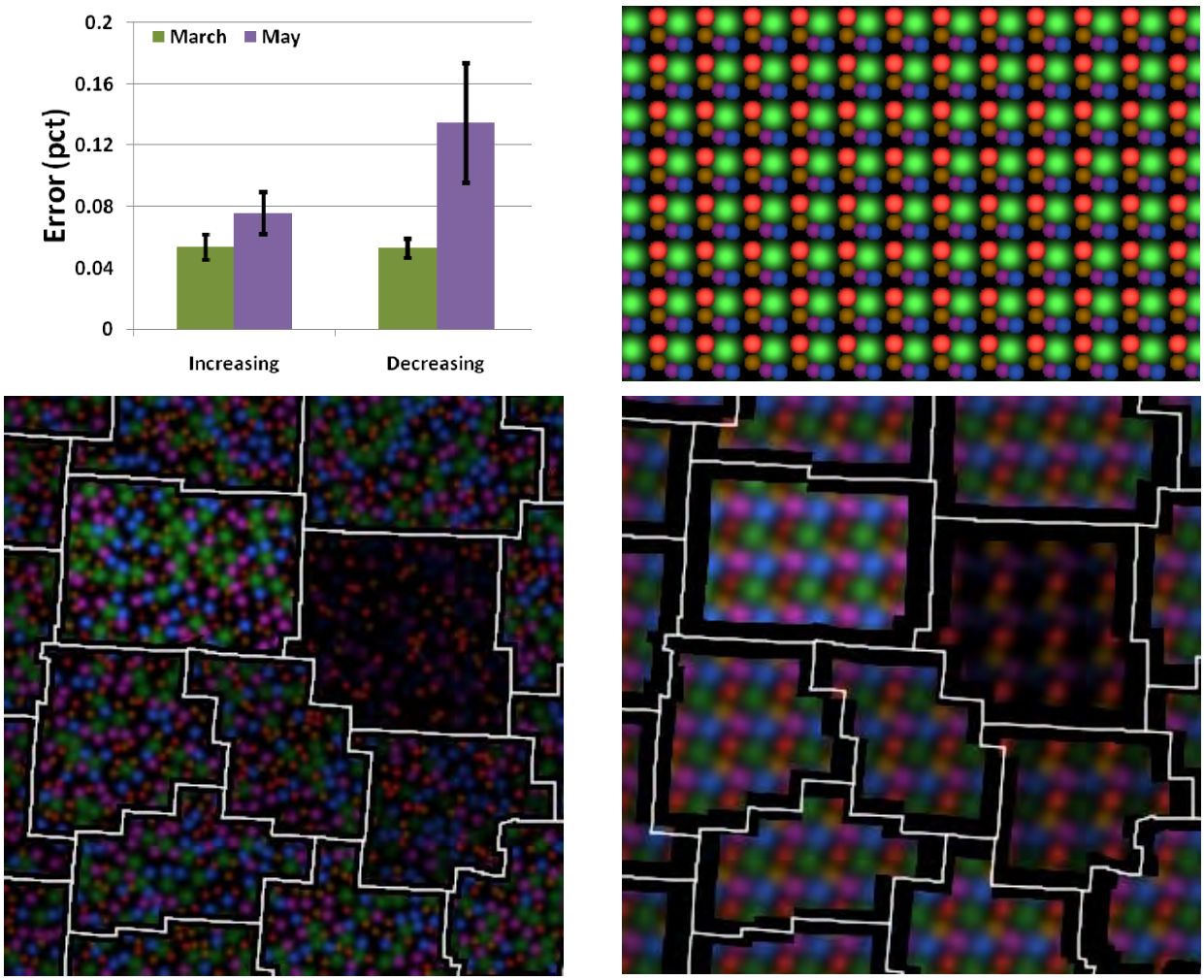

Figure 10: Exploring the behavior of DDS, which showed a significant increase in error when localizing decreasing trends (upper left). The new implementation (lower left) fills less of the space and is more randomized than the previous implementation (lower right), resulting in lower aggregate brightness. We had expected DDS to perform much like attribute blocks, since a DDS implementation can appear to look very similar to an implementation of attribute blocks (upper right).

These conflicting observations lower our confidence in the interpretation of our results. We take the following lessons from the two studies discussed in this paper as we proceed. First, based largely on the demonstration of ambiguous representation that can result from color blending, evidenced by the legend in Figure 4 (second from right), we feel that we should abandon exploration of this technique. This is despite the apparent adaptation made by returning users in our most recent study. Second, the success of attribute blocks in the current study, of DDS in the previous study, and juxtaposed maps in both studies (to varying extents) would argue for the continued exploration of these techniques. In particular, we should try to understand the potential overlap discussed above between attribute blocks and DDS. This may help us explore which parameters are truly the most useful for this type of task. Brush strokes continue to lag in performance behind other techniques, and appear to have success in this largely due to the frequency with which the extreme value (maximum or minimum) coincided with the greatest trend. Taking this with our study of critical point

detection[1], this technique appears to be less promising. Oriented slivers performed well on critical point detection, but has not fared as well on the trend localization task. It would also appear to be a candidate for dismissing from future exploration. The conclusion that one could make from these collected results is that the more tightly integrated the representation of each individual variable is, the lower the performance of users with the representation appears to be. This will make for an interesting hypothesis for future studies.

REFERENCES

[1] Livingston, M.A., Decker, J.W., and Ai, Z. "An Evaluation of Methods for Encoding Multiple, 2D Spatial Data," Proceedings of Visualization and Data Analysis, SPIE Vol. 7868, 2011

[2] Livingston, M. A. and Decker, J. W., "Evaluation of Trend Localization with Multivariate Visualizations," IEEE Transactions on Visualization and Computer Graphics 17(12):2053-2062, December 2011.

[3] Kirby, R. M., Marmanis, H., and Laidlaw, D. H., "Visualizing Multivalued Data from 2D Incompressible Flows using Concepts from Painting." Proceedings of IEEE Visualization, pgs. 333-340, October 1999

[4] Pickett, R. M., and Grinstein, G. G., "Iconographic displays for visualizing multidimensional data," Proceedings of the 1988 IEEE International. Conf. on Systems, Man, and Cybernetics, pgs. 514–519, August 1988

[5] Bokinsky, A. A., "Multivariate Data Visualization with Data-driven Spots," Ph.D. Dissertation, Department of Computer Science, University of North Carolina at Chapel Hill, 2003

[6] Urness, T., Interrante, V., Longmire, E., Marusic, I., O'Neill, S., and Jones, T. W. "Strategies for the Visualization of Multiple 2D Vector Fields," IEEE Computer Graphics & Applications 26(4):74-82, July/August 2006

[7] Weigle, C., Emigh, W., Liu, G., Taylor II, R. M., Enns, J. T., and Healey, C. G., "Effectively Visualizing Multi-Valued Flow Data using Color and Texture," Proceedings of Graphics Interface, pgs. 153-162, 2000

[8] Miller, J.R., "Attribute Blocks: Visualizing Multiple Continuously Defined Attributes," IEEE Computer Graphics & Applications 27(3):57-69, May/June 2007

[9] Beddow, J., "Shape Coding of Multidimensional Data on a Microcomputer Display," Proceedings of IEEE Visualization, pgs. 238-246, October 1990

[10] LeBlanc, J., Ward, M. O., and Wittels, N., "Exploring N-Dimensional Databases," Proceedings of IEEE Visualization, pgs. 230-237, October 1990

[11] Levkowitz, H., "Color Icons: Merging Color and texture Perception for Integrated Visualization of Multiple Parameters," Proceedings of IEEE Visualization, pgs. 164-170, 420, October 1991

[12] Healey, C. G. and Enns, J. T., "Building Perceptual textures to visualize multidimensional datasets," Proceedings of IEEE Visualization, pgs. 111-118, October 1998

[13] Healey, C. G., Tateosian, L., Enns, J. T., and Remple, M., "Perceptually based Brush Strokes for Nonphotorealistic Visualization," ACM Transactions on Graphics 23(1):64-96, January 2004

[14] Ware, C., "Information Visualization: Perception by Design (2nd ed.)," Morgan Kaufmann, 2004

[15] Joshi, A., "Art-inspired techniques for visualizing time-varying data," Ph.D. Dissertation, Department of Computer Science, University of Maryland-Baltimore County, 2007

[16] Laidlaw, D. H., Ahrens, E. T., Kremers, D., Avalos, M. J., Jacobs, R. E., and Readhead, C., "Visualizing Diffusion Tensor Images of the Mouse Spinal Cord," Proceedings of IEEE Visualization, pgs. 127-134, October 1998

[17] Laidlaw, D. H., Kirby, R. M., Jackson, C. D., Davidson, J. S., Miller, T. S., da Silva, M., Warren, W. H.. and Tarr, M. J., "Comparing 2D Vector Field Visualization Methods: A User Study," IEEE Transactions on Visualization and Computer Graphics 11(1):59-70, January 2005

[18] Hagh-Shenas, H., Interrante, V., Healey, C., and Kim, S., "Weaving versus blending: a quantitative assessment of the information carrying capacities of two alternative methods for conveying multivariate data with color," Proceedings of the 3rd Symposium on Applied Perception in Graphics and Visualization, pg. 164, 2006

[19] Tang, Y., Qu, H., Wu, Y., and Zhou, H. "Natural Textures for Weather Data Visualization," Tenth International Conference on Information Visualization, pgs. 741-750, July 2006

[20] Decker, J. and Livingston, M.A. "Poster: An Interactive, Visual Composite Tuner for Multi-Layer Spatial Data Sets," IEEE Visualization Poster Session, October 2010

[21] Hart, S. G. and Staveland, L. E., "Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research," In *Human Mental Workload*, pgs. 239-250, Elsevier Science Publishers, 1988