Multivariate Glyphs for Multi-Object Clusters

Eleanor Boyle Chlan*
Johns Hopkins University
Whiting School of Engineering

Penny Rheingans†
University of Maryland,
Baltimore County

ABSTRACT

Aggregating items can simplify the display of huge quantities of data values at the cost of losing information about the attribute values of the individual items. We propose a distribution glyph, in both two- and three-dimensional forms, which specifically addresses the concept of how the aggregated data is distributed over the possible range of values. It is capable of displaying distribution, variability and extent information for up to four attributes at a time of multivariate, clustered data. User studies validate the concept, showing that both glyphs are just as good as raw data and the 3D glyph is better for answering some questions.

CR Categories: H.5.2 [Information Interfaces and Presentation]: User Interfaces—Evaluation/methodology; I.3.6 [Computer Graphics]: Methodology and Techniques—Interaction Techniques;

Keywords: information visualization, multivariate visualization, distribution, aggregated data

1 INTRODUCTION

Glyphs usually represent individual, multivariate items. For example, in the context of census data, we could imagine a glyph representing an individual, where the shape implied the type of paid employment, the color implied an age range, and size was associated salary. Supplementation with additional features would encode addition attributes of the data, but it still only reflects the one individual. It would not tell us the average salary for a group or range of ages of those working in a particular type of job. Also, with large quantities of data it is probable to have more data items than can be reasonably comprehended by the average viewer. Besides, the more items to be displayed, the greater the likelihood that significant subsets will be occluded. With large quantities of data, it is common to aggregate items to make the data more cognitively manageable. Clustering [8] is a popular method. However, an unfortunate side effect of showing aggregated data as a single entity is the loss of the underlying values associated with the individual data items. Simply displaying a glyph at the centroid of the clustered data fails to show the extent and variability of the clustered data items. Average data values are useful but not sufficient. A group composed only of teenagers and senior citizens may be indistinguishable from a group composed only of middle aged people. Therefore, we also wish to get a sense of whether the data is scattered across a wide range or focused on a narrow range of some attribute, if the distribution is uniform or skewed, and what the actual extent of the data is. Our goal is to facilitate aggregation by designing a compact glyph providing as much information as raw data.

Inspired by Tukey Box plots [14], we seek to restore some of the lost information of clustered data by displaying summary and distribution information about the cluster's contents. We have developed a distribution glyph which shows mean, standard deviation, and extent, and also implies distribution of the data in the cluster. Two user studies show that the distribution glyph is just as good as the raw data in terms of correctness and preference. The 3D version of the glyph was shown to be superior to raw data for answering questions about average values and distribution.

*e-mail: chlan@apl.jhu.edu
†e-mail:rheingan@csee.umbc.edu

IEEE Symposium on Information Visualization 2005
October 23-25, Minneapolis, MN, USA
0-7803-9464-X/05/$20.00 ©2005 IEEE.

2 RELATED WORK

There is a rich body of research on the topic of using glyphs to display information. Colin Ware's book [16] gives basic background on standard glyphs, where the emphasis is on finding useful methods to encode additional information into symbols.

Histograms and scatter-plots are examples of very simple glyphs, which are useful for showing smaller amounts of data with few attributes [3]. There are numerous glyph types which represent individual data items [2, 5, 9, 10, 12].

There has been less research on methods to show composite data. Hendley et al. have developed a scheme called Narcissus [7] which casts a translucent isosurface around the cluster, visually converting the cluster into a simpler item. This does not solve the cognitive management problem for huge data sets since the individual glyphs are still displayed. In a similar vein, Sprenger et al. [13] cast an isosurface around the clusters created with a hierarchical, agglomerative partitioning algorithm called H-BLOB. Fua, et al. [6] use a variation of parallel coordinates to display aggregated, clustered data. Tukey box plots [14] show extent and a limited sense of variability for two-dimensional data. See the example in Figure 1. They display a box representing the second and third quartile of the data with the median marked. The box is supplemented with lines that show the extent of the first and fourth quartiles. Outliers may be shown as well. The size of the box is visually compared to the overall range of the data to obtain an impression of the distribution of the data.

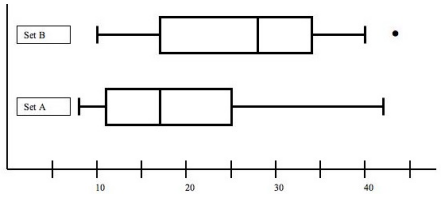

Figure 1: Tukey Box Plot Example

Other than Tukey box plots, which illustrate 2-dimensional data, glyphs do not show extent and variability for an attribute of interest. The glyphs discussed here represent a good cross-section of the kinds of multivariate glyphs available. Typical problems range from being overly abstract to being difficult to compare attributes across objects. In addition, none of these can handle aggregated data and nor show variability and extent of the data.

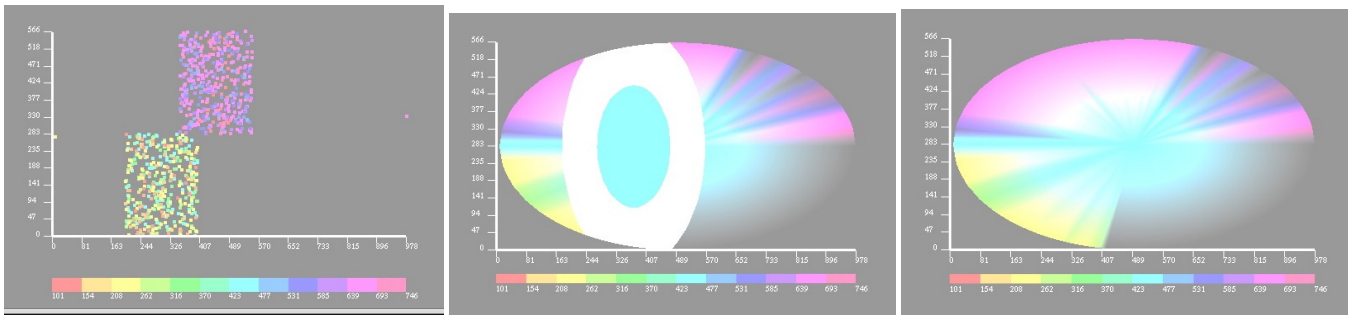

Figure 2: Test Images with Age, Education and Salary Mapped to X, Y and Color (a) Data Only - Shown as Dots (b) Combined Image with All Shells (c) Extent Shell

3 Approach

Since we want to imply the range of all the data values in the set without explicitly drawing them, we need to show the extent and distribution of the data. The distribution glyph has been inspired by Tukey box plots and is explicitly designed to show the average, standard deviation, and distribution of three attributes of the data set.

3.1 Two Dimensional Distribution Glyph

The two dimensional (2D) distribution glyph consists of three nested shells. The innermost or "average" shell is an ellipse with the size of the axes of the ellipse tied to the standard deviation of the data attributes mapped to the *X* and *Y* axes. It is centered on the average value of each of the two attributes. The middle or "variability" shell is an ellipse with the size of the axes tied to twice the standard deviation of the attributes mapped to the *X* and *Y* axes. The outer or "distribution" shell is an ellipse with the size of the axes tied to the minimum and maximum values of the attributes mapped to the *X* and *Y* axes.

A simulated test data set is shown in Figure 2(a). The full glyph corresponding to this data is shown in (b), while just the distribution shell is shown in (c). Note that the distribution shell corresponds to the limits of the data in both X and Y, in this case to the positions of the outliers, which are about halfway up the Y axis, near 0 and near 978 in X. Since the shell is elliptical, it is normal for points to occur outside the "corners" of the ellipse. The shell is not trying to show the exact area covered by the data, but is trying to imply its extent. In the middle variability shell, we see the data is reasonably tightly distributed in the X direction but much more broadly distributed in the Y direction. Note that the variability shell has been clipped by the distribution shell. This supports the notion that the distribution shell implies the limits of where data will be found. Finally, we see that the inner or average shell shows that the average is about 365 in X and about 283 in Y. The standard deviation is about 90 in X and about 170 in Y. The average in X and Y is clearly offset from the center of the distribution ellipse, implying skew or an asymmetric bell curve.

A third attribute is mapped to color. Since we are not able to use physical size to imply average, variability, and distribution, as we do for the attributes mapped to X and Y, we map the color with different methods for each shell. The inner shell still represents the average, the middle shell still implies variability, and the outer shell still implies the distribution or extent by giving us an idea of where the data points actually lie in the color spectrum.

For the inner average shell, the attribute associated with color is averaged over all the data points. The average is then interpreted using the HSV color model. For the average shell in Figure 2(b), the data points range in color from 423 to 477.

Variability is implied in the middle variability shell through color saturation. The shell is colored by interpreting the attribute of each data item associated with color in terms of HSV. The essentially cylindrical HSV colors are converted to rectangular coordinates and averaged. This method of averaging the color tends to desaturate the shell color when it is generated by widely varying values. If the data points all have very similar values, the shell color will tend to be highly saturated. Looking at the test data in Figure 2(c), we see that the color of the variability shell is essentially white, implying a broad distribution of color values in the data set.

The outer distribution shell implies different concentrations of data values in different regions of the data set. The shell is colored by taking a weighted average of the colors of the data points located in wedge shaped windows around the ellipse. The user specifies the size in degrees of the wedge originating from the center of the ellipse. The color calculation finds all of the data points in the wedge and weights each point inversely by its distance from a target point, which is centered in the wedge on the edge of the ellipse. Points closer to the edge contribute more to the color than points near the center or significantly outside the edge. The calculated color is assigned to the target point on the edge of the ellipse. The average value colors the point at the center of the ellipse. The color between the center and the edge is automatically interpolated. Sectors of the ellipse with few points are more transparent than sectors with more points. The distribution shell is shown at full opacity when the number of points in the sector exceeds half of the number of points possible if they are distributed perfectly uniformly. The number of points in the window is divided by this threshold to get a density value that controls the transparency of the color.

For the distribution shell in Figure 2(c), the wedge is +/- two degrees. Transparency starts to occur when the number of points in the window falls below five. The average of this data for attribute is between 370 and 423. The upper half is dominated by values in the 639 to 693 range yet has sub-concentrations on the 531 to 585 range. The lower left is dominated by data with low (yellow) values of the attribute, again with some sub-concentrations around 423. The lower right is transparent, indicating few or no points in that region. This implies most of the data is in the left half and there are some outliers to the extreme right. The broad value distribution implied by the middle variability shell is substantiated by the distribution shell which clearly shows color from yellow through magenta. In looking at (a) we can see there are also orange valued points which are too few to contribute to the coloring of the distribution shell.

Figure 3 shows the relationship of several nodes in a hierarchically clustered set of census data [15]. In Figure 4, we see a series of examples of the two-dimensional distribution glyph using this data. The data in question has been clustered using the standard algorithm K-means. The K-means algorithm was applied repeatedly to the results to impose a hierarchy on the data set. All the figures have

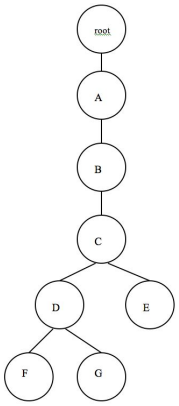

Figure 3: Relationship of Nodes in Clustered Census Data

the education level mapped to the x-axis, the marital status mapped to the Y-axis, and age mapped to color. Similarly, marital status starts at zero to represent three categories of "Married" and continues through separated, divorced and widowed to never married. The node in Figure 4(a) is Node A from Figure 3. You can see that, while the data in this part of the tree is broadly distributed, most is located in a smaller region centered in the lower right quadrant with some outliers. The white color of the variability shell shows a wide distribution of age. The average age is in the 35 to 41 range. Node B in Figure 4(b) is the child of the node in (a). The range of education and marital status is similar to the parent node but the age is a little more narrowly distributed, the variability shell having a bluer cast than the variability shell of the glyph in (a). The average age is also noticeably higher. Nodes D and E are shown in Figure 4(c) and (d). Node E seems to have the highest average age of all the nodes in this section of the tree. Although it has a smaller range of education and marital status, there are fewer outliers. As you descend though the hierarchy, the nodes reflect smaller and smaller ranges of the associated attributes. This is reflected by smaller sizes in X and Y and more saturated color in the variability shell. In Nodes D and F (Figure 4(c) and (e)), we see the node has become sufficiently cohesive in age that we can no longer distinguish the variability and average shells from each other. At the highest level in Node A, the distribution shell is almost white, telling us clearly that age is not big factor in the high level clustering of this data. For comparison, the Node A is redrawn in Figure 5(a) with the color attribute now mapped to hours worked per week. Clearly, the hours worked per week is a major factor in the clustering. Figure 5(b), shows the corresponding raw data, illustrating a common problem that occurs when representing a high dimensional set in a reduced number of dimensions. There are more than 3000 points in (a) but hundreds of points end up as collocated in (b) since each attribute has only a limited number of possible values.

3.2 Three Dimensional Distribution Glyph

A three-dimensional (3D) version of the glyph allows an additional attribute to be summarized for the data set and more appropriately fits into existing visualizations in 3D. The three dimensional Distribution Glyph is implemented like the two dimensional version except that a third attribute is mapped to the **Z** axis. The ellipses become ellipsoids. The size of the average shell represents standard deviation, the size of the variability shell represents twice the standard deviation, and the size of the distribution shell represents the minimum and maximum values. However, simply adding a dimension to the 2D model is not adequate. When properly shaded and lighted, all the viewer would see would be the distribution shell. It's

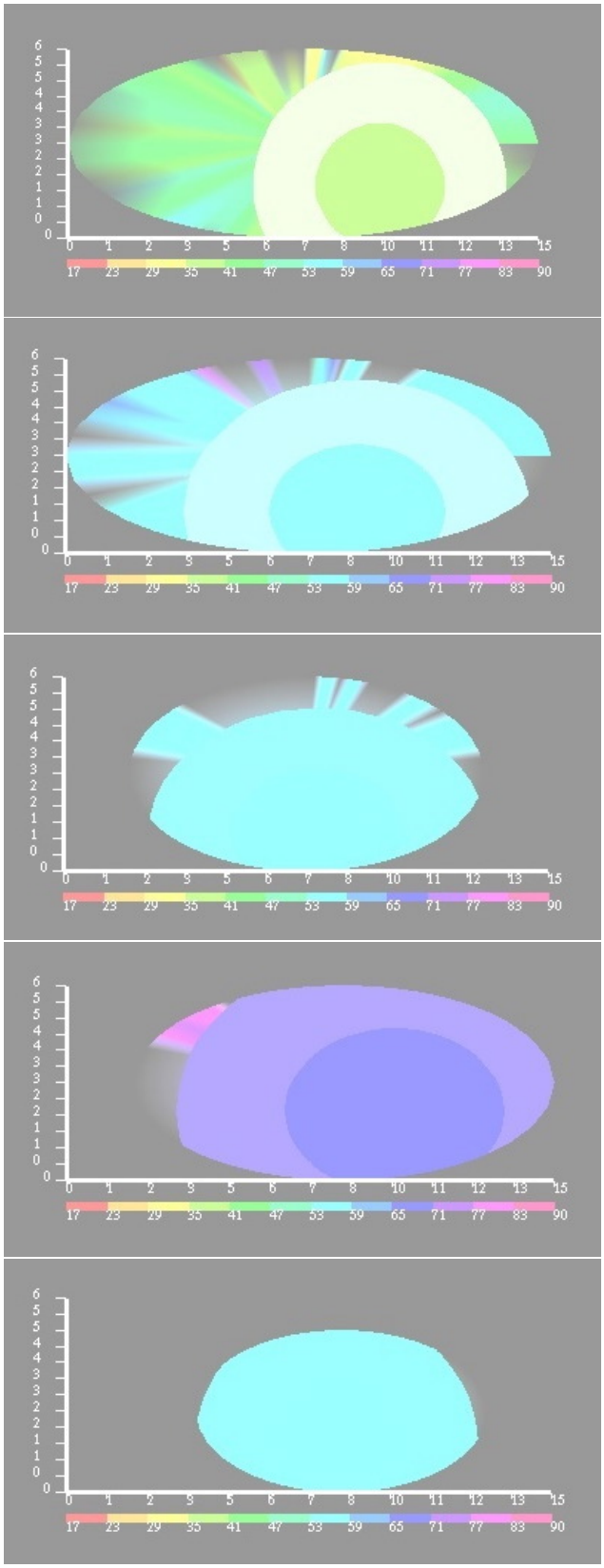

Figure 4: Related Nodes in Clustered Census Data with Education, Marital Status,and Age Mapped to X, Y, and Color (a) Node A (b) Node B (c) Node D (d) Node E (e) Node F

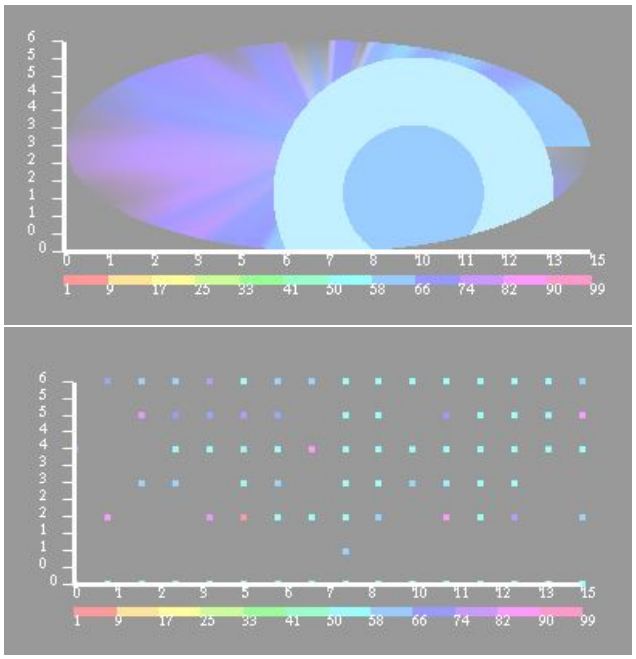

Figure 5: Census Data Example Node A Colored by Hours worked Per Week (a) 2D Glyph (b) Raw data

important to be able see the inner and middle shells with the colors accurately rendered. Some viewers may also find it more difficult to accurately perceive the shapes and relationships in a static 3D image.

In order to see all three ellipsoids, the distribution shell is shown as a solid with every other row of triangles omitted. The variability shell is also shown as a solid, but with every other sector of longitude omitted. The overlay of the variability and distribution shells creates a basket weave though which the average shell can be seen. This methods allows all three shells to be seen without the misleading color changes that would occur if the variability and distribution shells were sufficiently transparent to allow the average shell to be seen [11]. Showing the density of points via transparency is omitted.

Like the 2D model, the color of the average shell represents the average value of the fourth attribute which is then mapped to color. The color of the variability shell represents the average of the colors after the attribute values are converted to color values. The color of the distribution shell is calculated by taking the weighted average of the color of all the points in a three dimensional wedge centered on a point on the surface of the ellipsoid.

A simulated test data set is shown in Figure 6(a). The average shell corresponding to this data is shown in (b). Note the average color is 394 to 453.

The variability shell is shown with the average shell in Figure 6(c). The color of the variability shell is essentially white, implying a broad distribution of color values in the data set. This is substantiated by the raw data which clearly shows colors from yellow through magenta. In looking at the raw data in Figure 6(a), we can see there are also orange valued points which are too few to contribute to the coloring of the distribution shell.

The distribution shell is shown in Figure 6(d) and the complete glyph is shown in (e). The imposition of the distribution shell over the variability shell creates the basket weave effect noted above but still allows the average shell to be seen. To make it easier to see the real extent of the data in 3D, the 3D glyph has been supplemented with the convex hull of the points in the dataset. The hull is calcu-

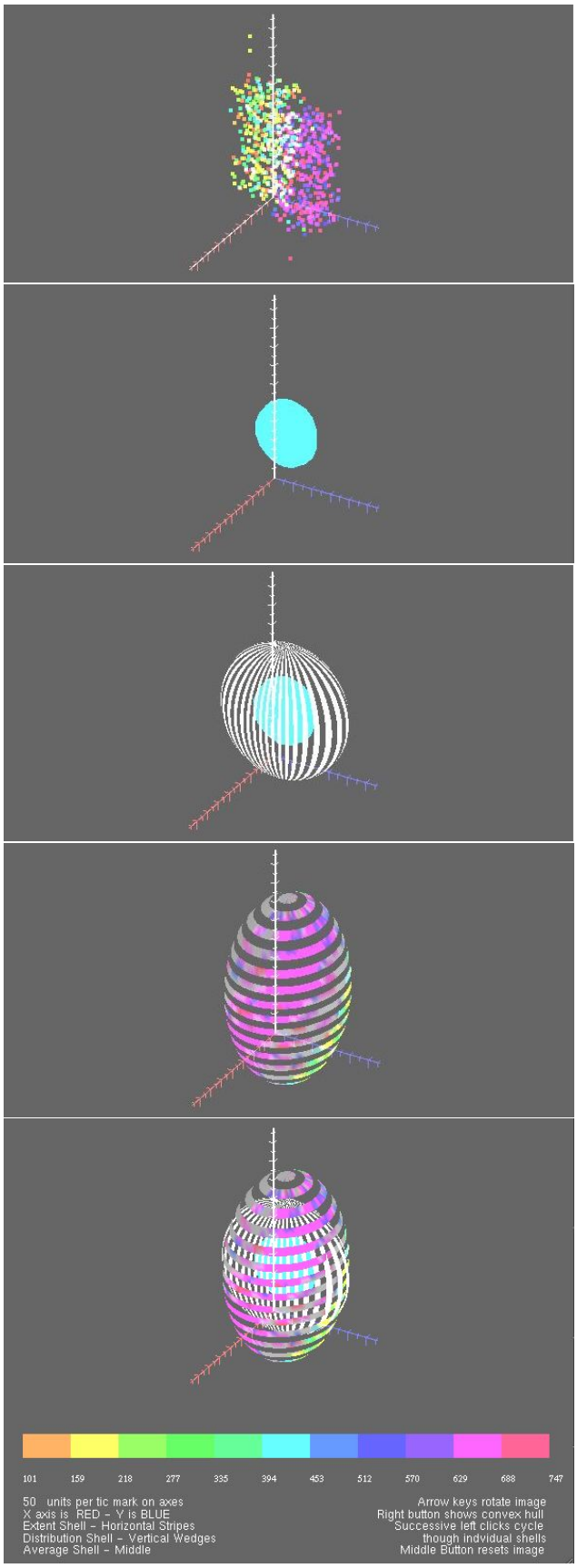

Figure 6: Test Image with Age, Education, Hours Worked and Salary Mapped to X, Y, Z and Color (a) Raw Data (b) Average Shell (c) Average and Variability Shells (d) Distribution Shell (e) Complete Glyph

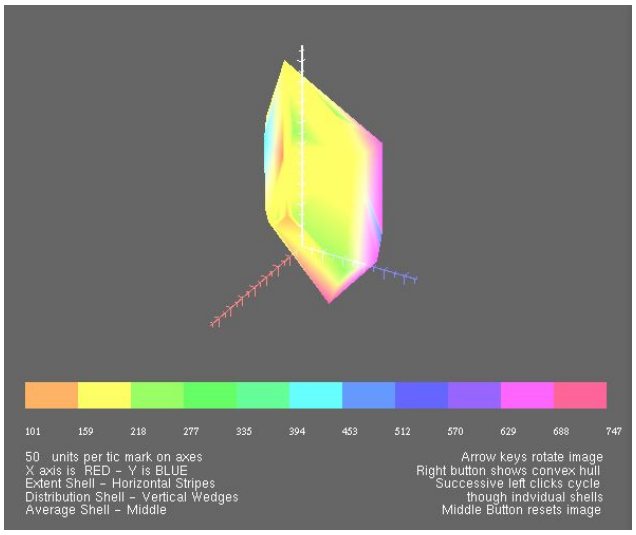

Figure 7: Test Image with Age, Education, Hours Worked and Salary Mapped to X, Y, Z and Color - Convex Hull

lated using the freeware Qhull [1]. Figure 7 shows the convex hull corresponding to Figure 6(d).

Previous research has shown that 3D can be more powerful and more easily understood than 2D when supplemented with rotation [4], so the image can be viewed from multiple positions. To that end, the 3D distribution glyph can be rotated in two directions using the arrow keys on the keyboard. In addition, shells can be shown individually. Therefore, the variability shell is no longer clipped by the distribution shell.

Examples of the three dimensional glyph are shown in Figures 8 and 9. The node in Figure 8(a) shows the root node of the data set in Figure 3. Age is shown in color and it is obvious from the color of the inner shell that the average age of the entire set is between 36 and 43. The white color of the middle shell indicates that the ages are fairly broadly distributed. The upper parts of the outer shell show concentrated regions of ages 23 to 56. Education level is shown along the red X axis where higher values indicate more education. The data set includes people with no education as well as all levels of education. The average education is higher than the midpoint. Education is relatively narrowly distributed. The ages of the people in the Node A are similar to the root but Node A represents a subset with a higher educational level. The education level of Node B is similar to Node A, however Node B clearly has a smaller range of ages, as shown by the more cyan color of the middle shell and the lack of yellows in the outer shell. The average age of this group is in the 50 to 56 range, but the presence of purple and tiny amounts of magenta in the outer shell indicate that the group includes elderly people. The average age of Node C is higher than Node B and is in the 56 to 63 years range. The more saturated color of the middle shell means the age distribution is narrower than in the parent. Marital status is shown along the white Z axis. Married is zero and never married is six with various categories of single in between. Marital status is broadly distributed with the average closer to married than to never married, offsetting the middle shell from the outer shell. This trend continues in the nodes in Figure 9. The average age of Node E is in the 63 to 70 range. The distribution of ages is quite narrow as indicated by the highly saturated color of the middle shell. The blue Y axis shows occupation categories. Clearly Node E has a broader range of occupation than its sibling, Node D. Node F clearly has a narrower, more highly educated group of people than its sibling in Node G.

Both versions of the glyph execute in linear time with respect to the size of the data set. The color calculation for the outer or distribution shell is quadratic.

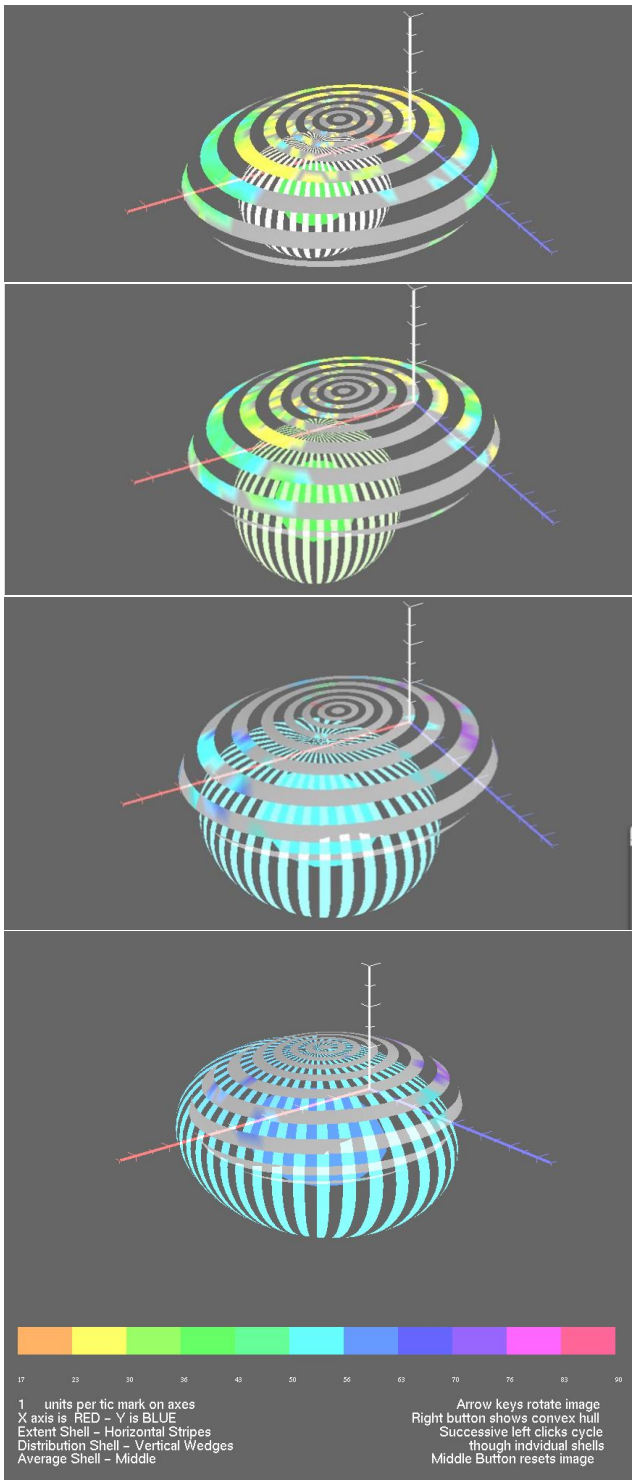

Figure 8: Related Nodes in Clustered Census Data - 3D Glyph Showing Age, Education, Occupation and Marital Status Mapped to Color, X, Y and Z (a) Root (b) Node A (c) Node B (d) Node C

Table 1: 2D Accuracy Results Showing Mean Percentage Correct By Question for Ten Subjects Using T-test Paired Two Sample with alpha=.05

Questions	Mean Glyphs	Mean Dots	t-value	p-value
1) What is the average value of the X-attribute?	.73	.63	0.68	0.55
2) What is the range of the Y-attribute?	.65	.68	-0.23	0.84
3) What is the range of the color attribute?	.10	.43	-2.03	0.14
For each remaining question, if color was salary, X was age and Y was education level:				
4) Is salary narrowly distributed?	.83	.88	-0.77	0.50
5) Is age narrowly distributed?	.88	.93	-0.58	0.60
6) Is education level narrowly distributed?	.80	.83	-0.17	0.87
7) Is salary uniformly distributed across age and education level?	.58	.68	-0.56	0.61
8) Is there a correlation between salary and age?	.83	.93	-1.10	0.35
9) Is this distribution skewed or lopsided in age and education level?	.93	.70	3.00	0.06
10) Roughly, how many poor, well educated people, middle aged people are in this set?	.75	.63	0.78	0.49

4 User Studies

Our intention with the Distribution Glyph was to create a user-friendly glyph that made it feasible for information to be extracted from the aggregated data without resorting to the underlying data. To test this, we conducted a user study for each version of the distribution glyph, comparing the distribution glyph with the raw data. The primary hypothesis was that the Distribution Glyph was just as useful as raw data for answering basic statistical questions on the data. Timing, correctness and user-preference were measured. These are fully factorial, within-subject experiments.

4.1 Experimental Design

Ten subjects participated in the 2D study and 16 subject participated in the 3D study. Most were students, ranging in age from 18 to 30, and were pre-screened to have a minimum level of computer skills. All had normal or corrected-to-normal acuity and normal color vision. None had previously participated in a user study for this project. No subject participated in both studies.

The experiment started with subject training. The subject was shown a series of slides which explained the two visualizations. The raw data is displayed as dots. For the distribution glyph, each shell was explained and shown with examples. Samples are shown in Figure 2(b) and Figure 6(d). The subjects were allowed as much time as they wanted to review the training material and ask questions. In the 3D study, subjects were given the opportunity to practice the rotation and image manipulation options and to answer some sample questions as part of the training. For both studies, the user task was to answer the questions about the images. See Table 1 and Table 2 for the specific questions.

In the timed portion of the 2D study, a subject was shown a set of four representations of the same type and answered the same ten questions for each one. Then the subject was shown the same four data sets rendered in the other visualization and again the subject answered the same ten questions about each one. For the 3D study, there were two representations in each set and twelve questions. Subjects were encouraged to take as much time as needed to answer the questions before moving on, but were not permitted to make direct comparisons or go back. Half the subjects viewed the visualizations in one order and half in the other order. After the timed portion, the subjects answered a brief questionnaire. Subjects took anywhere from 30 minutes to 75 minutes to complete the studies, including the training time. Subjects were not informed of the accuracy of their responses.

4.2 2D Results

Correct answers to the statistical questions on each data set were counted and compared by question across the two visualization types. A standard paired t-test with two tails was used. Table 1 shows the mean number of correct answers for all subjects on each question for each visualization type. It also shows the t-value calculated by the test and gives the probability that the means would be the same by random chance. The standard cutoff is .05. By that standard, none of the glyph means are significantly different from the corresponding dots (raw data) mean. Question 9 comes very close to meeting that standard (p=0.06). This question asks the subject to estimate if there is skew in the distribution.

4.3 2D Discussion

Even based on a few subjects, it is easy to see that this glyph requires a substantial investment in experience. It is complex enough that training and practice is necessary to use it successfully. In fact, there is a clear decrease in the time needed to process the distribution glyph as subjects progress through the set. Analysis also shows that the interaction of the subjects with the data sets over time infers an advantage to the glyphs over the dots. Interestingly enough subjects with no or minimal experience in code development showed no clear disadvantage in terms of training times to review the visualizations or in correctness. Also, experience playing computer games, even in 3D, seemed to impart a minimal advantage to reviewing the dots visualizations but none to reviewing the glyphs. Although no statistically significant preferences were identified overall, some of the more experienced subjects preferred the distribution glyph and thought it was better for answering the questions. Although statistically significant differences could not be found, the order in which subjects viewed the visualizations seemed to influence the study. It is interesting to note that all the subjects, who viewed the glyphs first, answered more questions correctly about both sets of visualizations than did the subjects who viewed the dots first. This means that people learn something with the glyphs that they do not learn with the raw data. Overall, there was equal preference for the two types, however all subjects preferred the type they viewed first. The two sets of visualizations averaged similarly as easy or hard to understand, but subjects viewing the dots first found them easier to understand. Overall, subjects preferred the glyphs, preferred the training for the glyphs and felt they were more useful and easier for answering questions. Again, subject preference seemed to be tied to which set of visualizations the subject viewed first.

4.4 3D Results

As with the 2D study, correct answers to the statistical questions on each data set were counted and compared by question across the two visualization types. A standard paired t-test with two tails was used. Table 2 shows the mean number of correct answers for all subjects on each question for each visualization type. It also shows

Table 2: 3D Accuracy Results Showing Mean Percentage Correct By Question for Sixteen Subjects Using T-test Paired Two Sample for Means with alpha=.05 if Color is Salary Per Week, X is Age in Months, Y is Education Level and Z is Hours Worked in a Four Month Test Period

Questions	Mean Glyphs	Mean Dots	t-value	p-value
1) What is the average value of salary?	0.76	0.50	-2.32	0.03
2) What is the minimum and maximum hours worked?	0.47	0.50	0.24	0.81
3) What is the minimum and maximum of salary?	0.24	0.59	3.52	0.00
4) The distribution of salary is Narrow In-between Broad?	0.71	0.71	0.00	1.00
5) The distribution of age is Narrow In-between Broad	0.71	0.68	-0.24	0.81
6) The distribution of education level is Narrow In-between Broad	0.68	0.79	1.00	0.33
7) Is salary uniformly distributed across hours worked?	0.71	0.41	-2.15	0.04
8) Is salary uniformly distributed across age and education level?	0.47	0.44	-0.23	0.82
9) Is there a correlation between salary and age?	0.71	0.88	1.98	0.06
10) Is this distribution skewed or lopsided in education level?	0.62	0.79	1.64	0.11
11) Is this distribution skewed or lopsided in salary?	0.44	0.29	-1.15	0.26
12) Roughly, how many poor, well educated older people, working very long hours are in this set?	0.71	0.79	0.83	0.41

the t-value and p-value calculated by the test. The p-value gives the probability that the means would be the same by random chance. A cutoff of .05 was used. Subjects gave significantly more accurate responses on question one (p = 0.03), asking for the average of an attribute, when reviewing the glyphs. We suspect that the ability to examine the shells individually influenced this result. Subjects also gave significantly more accurate responses for question seven (p = 0.04), which asked subjects to compare the distribution of one attribute against another. Question nine is a harder questions asking about correlation between attributes. The mean number correct using glyphs is noticeably better than the mean correct using the raw data even though it is not statistically significant (p = 0.06). Unfortunately, subjects did significantly better on question three using the dots (p = 0.00). The effect is primarily from people who viewed the dots first. This question asked the range of the color attribute.

The order in which subjects viewed the visualizations seemed to influence the study. Subject viewing the dots firsts did statistically significantly better on question three, and subjects viewing the glyphs first did statistically significantly better on question one, as noted above. These order effects are also statistically significant.

4.5 3D Discussion

Subject found both types equally easy to use to answer questions, but as a group they all thought the second set they saw was easier. This result is the reverse of what was observed on the 2D study, but is probably just a learning curve effect. The dots were easier to understand, as expected, especially for subjects viewing the glyphs first. Subjects viewing the dots first tended to think both types equally easy to understand. A learning curve is also seen in overall subject preference to like the second type that they viewed more and think it more useful. This was not statistically significant. In spite of this, subjects thought both visualization types equally easy to use to answer questions and were equally comfortable with them. Subject also thought the glyphs more useful and better for answering harder questions. As the questions become more difficult, the 3D glyphs rise to the challenge better than the raw data does.

We noted four subjects for whom language was potentially an issue. It had a noticeable but not statistically significant effect on times required to complete the study. These four subjects had a statistically significant lower correctness on question 3 for the distribution glyphs and a statistically significant higher correctness on question 9 for the raw data. Prior programming experience and prior 3D experience had an impact on a few questions but no effect across the entire study. Gender on the other hand seemed to play no role in correctness.

Question two asked for the range of the Z attribute. For question two the probability is very high that the answers to both the glyphs and raw data images came from the same population. In contrast, both the 2D and 3D versions of the glyph had problems with the questions asking range of the color attribute. We believe that the method of calculating the color of the distribution shell by averaging the data occurring in a wedge accounts for this. If a portion of the range is represented by only a few points then the method tends to hide that color.

Just as in the 2D study, this study showed that the 3D distribution glyph requires a substantial investment in experience. It is complex enough that training and practice is necessary to use it successfully. In fact, there is a clear decrease in the time needed to process the distribution glyph as subjects progress from the first visualization in the pair to the second. Analysis also shows that over time the glyphs accrue an advantage over the dots. Interestingly enough, gender, programming experience, and 3D gaming experience seemed to have no clear impact on times required to complete the study. On the other hand, language did seem to impact the times for reviewing the glyphs. All subjects took longer to review first glyph visualization than the second. However the four subjects identified as having potential English language issues took significantly longer to process the first glyph visualization.

5 Conclusions and Future Work

This tool is aimed towards technically oriented investigators who are willing to invest some time into learning to use a more powerful tool. Some subjects were more accurate with the dots and some were more accurate with the distribution glyph but overall subjects gave more correct answers with the glyphs. It also gave better results on certain types of questions.

This work shows that it is reasonable to aggregate data and still retain a sense of how the data is distributed. We have developed two and three dimensional versions of a glyph that show extent, average, variability, and distribution for up to four attributes. It is appropriate to use it for summarizing data sets or to display clusters of data items in extremely large clustered sets. This represents an important step in the visualization of large data sets because it allows the number of physical items to be reduced while minimizing the accompanying information loss. When combined with other visualization tools, the distribution glyph can simplify the display of very large, multivariate data sets by reducing the number of glyphs displayed, while still retaining useful information about the underlying data. This is further supported by the user studies which show that the glyphs are just as good as raw data for answering statistical

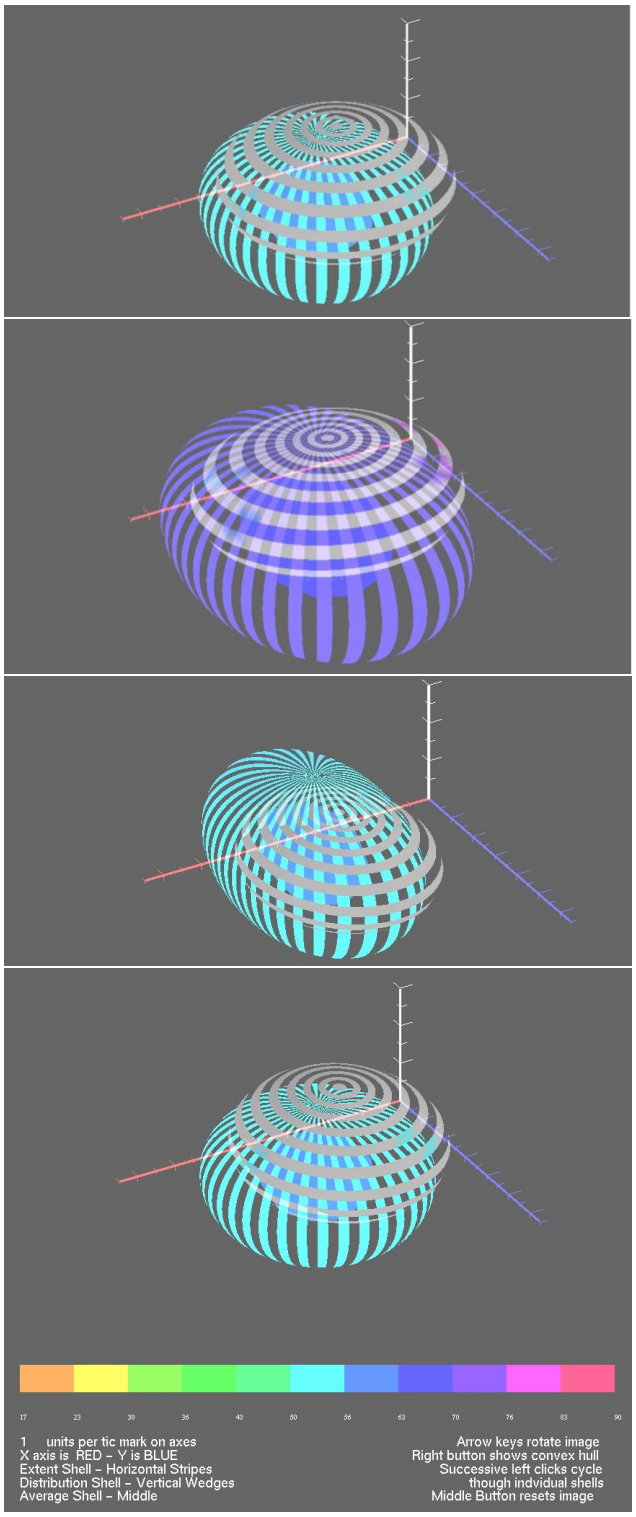

Figure 9: Related Nodes in Clustered Census Data - 3D Glyph Showing Age, Education, Occupation and Marital Status Mapped to Color, X, Y and Z (a) Node D (b) Node E (c) Node F (d) Node G

questions. It further shows that for some questions it is even better.

A user study comparing the glyphs to data represented by centroids should provide further insight into the usefulness of the glyphs. A feature to display multiple, related nodes in a data set would allow more comparison across the data sets.

6 Acknowledgments

This work supported in part by the Department of Defense (CADIP), the National Science Foundation (0121288) and the AT&T Foundation.

References

[1] C. B. Barber, D. P. Dobkin, and H. T. Huddanpaa. The quickhull algorithm for convex hulls. *ACM Transactions on Mathematical Software*, 22(4):469–483, December 1996.

[2] Mei C. Chuah and Stephen G. Eick. Information rich glyphs for software management. *IEEE Computer Graphics and Applications*, pages 2–7, July–August 1998.

[3] William S. Cleveland. *Visualizing Data*. Hobart Press, Summit, NJ, 1993.

[4] Andy Cockburn and Bruce McKenzie. 3d or not 3d? evaluating the effect of the third dimension in a document management system. In *Proceedings of CHI'01*, pages 434–441, New York, NY, May 2001. Addison-Wesley Publishing Co.

[5] David Ebert, James Kukla, Christopher Shaw, Amen Zwa, Ian Soboroff, and D. Aaron Roberts. Automatic shape interpolation for glyph-based information visualization. *IEEE Visualization*, October 1997.

[6] Ying-Huey Fua, Matthew O. Ward, and Elke A. Rundensteiner. Hierarchical parallel coordinates for exploration of large datasets. In *Proceedings of the conference on Visualization '99*, pages 43–50. IEEE Computer Society Press, 1999.

[7] R. J. Hendley, N. S. Drew, A. M. Wood, and R. Beale. Narcissus: Visualizing information. In Nahum D. Gershon and Steve Eick, editors, *Proceedings Symposium on Information Visualization*, pages 90–96. IEEE Computer Society Press, 1995.

[8] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: A review. *ACM Computing Surveys*, 31(3):264–323, 1999.

[9] Eser Kandogan. Visualizing multi-dimensional clusters, trends, and outliers using star coordinates. In *Proceedings of KDD-2001*, San Francisco, 2001.

[10] Martin Kraus and Thomas Ertl. Interactive data exploration with customized glyphs. In *Proceedings of WSCG'01*, pages 20–23, 2001.

[11] Penny Rheingans. Opacity-modulating triangular textures for irregular surfaces. In *Proceedings of IEEE Visualization '96*, pages 219–225. IEEE Computer Society Press, October 1996.

[12] Randall M. Rohrer, John L. Sibert, and David S. Ebert. A shape-based visual interface for text retrieval. *IEEE Computer Graphics and Applications*, pages 2–8, September/October 1999.

[13] T. C. Sprenger, R. Brunella, and M. H. Gross. H-blob: A hierarchical visual clustering method using implicit surfaces. In *Proceedings Symposium on Information Visualization*, pages 61–68. IEEE Computer Society Press, October 2000.

[14] John W. Tukey. *Exploratory Data Analysis*. Series in Behavioral Science: Quantitative Methods. Addison-Wesley, 1977. page 39.

[15] University of California, Irvine. *UCI Machine Learning Repository*, 1999.

[16] Colin Ware. *Information Visualization Perception for Design*. Morgan Kaufmann Publishers, San Francisco, CA, 2000.