Multivariate Visualization Using Metric Scaling

Pak Chung Wong* R. Daniel Bergeron†

Department of Computer Science‡
University of New Hampshire

Abstract

We present an efficient visualization approach to support multivariate data exploration through a simple but effective low dimensional data overview based on metric scaling. A multivariate dataset is first transformed into a set of dissimilarities between all pairs of data records. A graph configuration algorithm based on principal components is then used to determine the display coordinates of the data records in the low dimensional data overview. This overview provides a graphical summary of the multivariate data with reduced data dimensions, reduced data size, and additional data semantics. It can be used to enhance multidimensional data brushing, or to arrange the layout of other conventional multivariate visualization techniques. Real life data is used to demonstrate our approach.

1 Introduction

We present an efficient visualization approach to support multivariate data exploration through a simple but effective low dimensional data overview based on *metric scaling* [4, 5]. This overview can be used to enhance multidimensional data brushing, or arrange the layout of other conventional multivariate visualization techniques. Some of the underlying design concepts have been applied in various visualization tools. Keim et al. [11, 12] define a *distance function* as a metric to show the *relevance factor* of individual variates of a dataset in *VisDB*. Ward and Bentley [2] use multidimensional scaling to *animate* multidimensional datasets *in Mavis*. Hurley et al. use principal components to analyze data with *motion graphics* in *Data Viewer* [6] and *XGobi* [3].

We demonstrate our approach using a publicly accessible automobile dataset[1], which contains information about thirty eight 1978-79 model automobiles including miles per gallon, weight, drive ratio, horsepower, displacement, and number of cylinders. An example using a larger dataset with 329 records is described in Section 3.2. Suppose the data has *dissimilarities*, δ_{rs}, measured between all pairs of automobiles in the 6 dimensional *variate* space. The dissimilarity between two station wagons is expected to be much smaller than the one between a compact and a full size sedan. A graph configuration of 38 vertices, in which the r^{th} vertex represents the r^{th} automobile, is sought in a d dimensional display space such that the distances, d_{rs}, between all pairs of vertices *match* the corresponding dissimilarities, δ_{rs}, in variate space. This configuration is called a low dimensional data overview. Figure 1 shows an example of such a graph configuration of the automobile dataset in a two dimensional space. A quick look at the figure reveals that the full size sedans and wagons are located at the right hand side, the compacts and subcompacts are at the left hand side, the upscale medium size sedans are at the top, and the rest are scattered around the middle of the display. The graph successfully highlights all the major clusters, which reflect the dissimilarities among the six variates of the data.

This example is effective partly because common knowledge is sufficient to identify the nature of the clusters. When there is no obvious meaning coming out of the data overview, further analysis of the original data through other means is necessary. The data overview can then be used to guide the exploration of the local details.

This paper describes the construction of such a low dimensional display of multivariate data through the use of principal components [10]. The strengths and weakness of the approach as compared to other multivariate visualization techniques such as scatterplot matrix are also investigated. An effective visualization environment can be achieved by combining the strengths of the different techniques.

2 Metric Scaling

In this paper, we narrow our data choices to *quantitative* data which includes both *interval* and *ratio* data. Suppose we have a set of n records with v variates and dissimilarities, δ_{rs}, measured between all pairs of records in a v dimensional space. We want to configure a graph of n vertices in a d dimensional display space such that each vertex represents one record and the distances, d_{rs}, measured between all pairs of vertices in display space match δ_{rs} in variate space *as closely as possible*. This graph configuration problem falls into the broad research area of *metric scaling* studied mostly by mathematical psychologists. The goal is to determine the dissimilarities between all pairs of records by distances in v space, and then use them to compute the coordinates of the n vertices in the d dimensional display space. Figure 1 shows a low dimensional data overview of the 38 automobiles in two dimensional Euclidean space.

2.1 Data Dissimilarity Measurement

The first step is to determine the dissimilarities between all pairs of input records. Euclidean distance in v space is the most commonly used metric, but Cox and Cox [4] suggest nine other metrics including weighted Euclidean, Minkowski, and Manhattan (a.k.a. city blocks) that can be used to measure data dissimilarity of quantitative datasets. Usir_ Euclidean distance, the dissimilarity, δ_{rs}, between records r and s is given by

$$\delta_{rs} = \sqrt{\sum_{i=1}^{v} (x_{ri} - x_{si})^2}.$$

A dataset with n records generates an $n \times n$ real symmetric dissimilarity matrix. Each element of this matrix contains the dissimilarity, δ_{rs}, between records r and s of the original data. For example, given a dataset with five variates and six records as shown in Figure 2a, the Euclidean metric is applied to the data and the result is a 6×6 dissimilarity table, such as shown in Figure 2b.

*pcw@cs.unh.edu, http://www.cs.unh.edu/~pcw

†rdb@cs.unh.edu, http://www.cs.unh.edu/~rdb

‡Department of Computer Science, Kingsbury Hall, University of New Hampshire, Durham, New Hampshire 03824, USA. Phone: (603) 862-3778. Fax: (603) 862-3493.

[1] http://lib.stat.cmu.edu/DASL/Stories/ClusteringCars.html

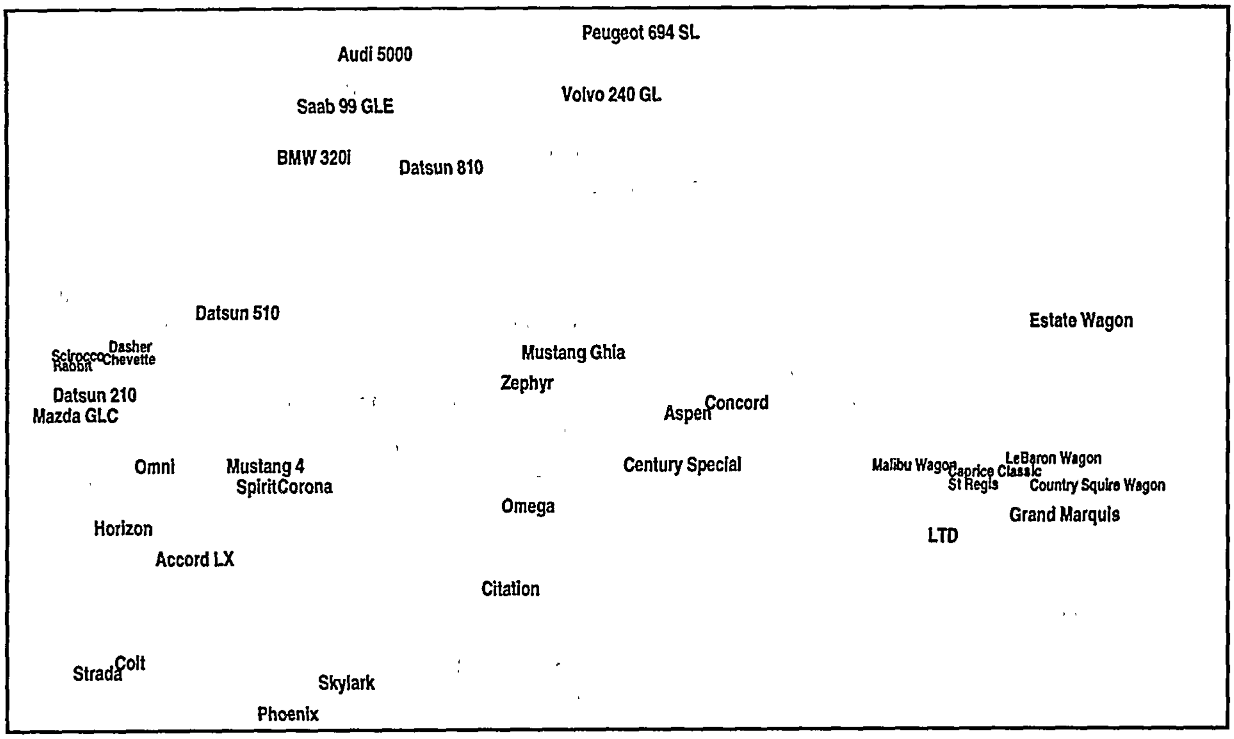
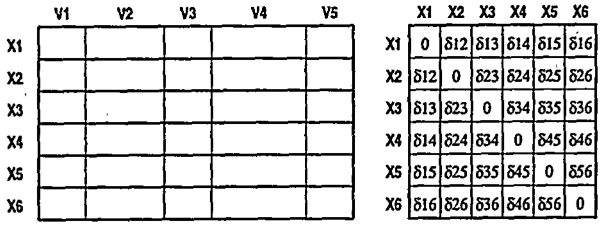


Figure 1: A two dimensional display of the six variate automobile dataset.



(a)

	V1	V2	V3	V4	V5
X1					
X2					
X3					
X4					
X5					
X6					

(b)

	X1	X2	X3	X4	X5	X6
X1	0	$\delta12$	$\delta13$	$\delta14$	$\delta15$	$\delta16$
X2	$\delta12$	0	$\delta23$	$\delta24$	$\delta25$	$\delta26$
X3	$\delta13$	$\delta23$	0	$\delta34$	$\delta35$	$\delta36$
X4	$\delta14$	$\delta24$	$\delta34$	0	$\delta45$	$\delta46$
X5	$\delta15$	$\delta25$	$\delta35$	$\delta45$	0	$\delta56$
X6	$\delta16$	$\delta26$	$\delta36$	$\delta46$	$\delta56$	0

Figure 2: a) A dataset with five variates and six records. b) A dissimilarity matrix of the dataset.

2.2 Recovery of Coordinates

Using the approach of principal components, we can represent the data as points in a new p dimensional space where $p \leq n$. We create an inner product matrix from the dissimilarities δ_{rs} in variate space, and find its non-negative eigenvalues $\lambda_1, \ldots, \lambda_p$ and the corresponding eigenvectors to yield the Euclidean coordinates of the n vertices in the p dimensional space. (See Appendix A for more details.)

If the eigenvalues are sorted in descending order, i.e., $\lambda_1 > \lambda_2 > \ldots > \lambda_p$, the first principal component associated with λ_1 is more important than the second component, which in turn is more important than the third and so on. (See [10] for details.) The distance, Δ_{rs}, between vertices r and s is given by

$$\Delta_{rs}^2 = \sum_{i=1}^{p} \lambda_i (x_{ri} - x_{si})^2$$

where x_r and x_s are the distance vectors associated with points r and s respectively (i.e., columns/rows of the matrix in Figure 2b). A smaller eigenvalue contributes much less weight to the distance d_{rs}, so these smaller eigenvalues can be truncated with less error. In many cases, the first two to three components can approximate the data very well. Suppose we select the d most significant eigenvalues to display the data overview, the degree of accuracy of the approximation can then be measured by

$$\frac{\sum_{i=1}^{d} \lambda_i}{\sum_{i=i}^{p} \lambda_i}.$$

The two dimensional graph of the automobile dataset depicted in Figure 1 has a degree of accuracy of 94.61%. It reaches 97.37% if the third principal component is included.

It is important to realize that the eigenvalues based principal components are not the only approach to do scaling. Other popular methods include Monte Carlo scaling and Least Squares scaling [4].

3 Strengths and Weakness

We investigate the strengths and weakness of the low dimensional data overview, and compare it to other multivariate visualization techniques including scatterplot matrix and parallel coordinates. All the figures of scatterplot matrix and parallel coordinates were generated by an enhanced version of *XmdvTool* [13, 15, 18].

3.1 Data Clustering

Both scatterplot matrix and parallel coordinates are very flexible multivariate visualization techniques which perform well in a wide variety of visualization situations [19]. A scatterplot matrix consists of an array of panels which present pairwise adjacent scatterplots of a multivariate dataset. Parallel coordinates [7, 8, 9] use parallel axes to plot a multivariate dataset. Examples of scatterplot matrix and parallel coordinates are shown in Figures 3b and 3c. With the addition of high dimensional brushing [15, 13], both techniques enable

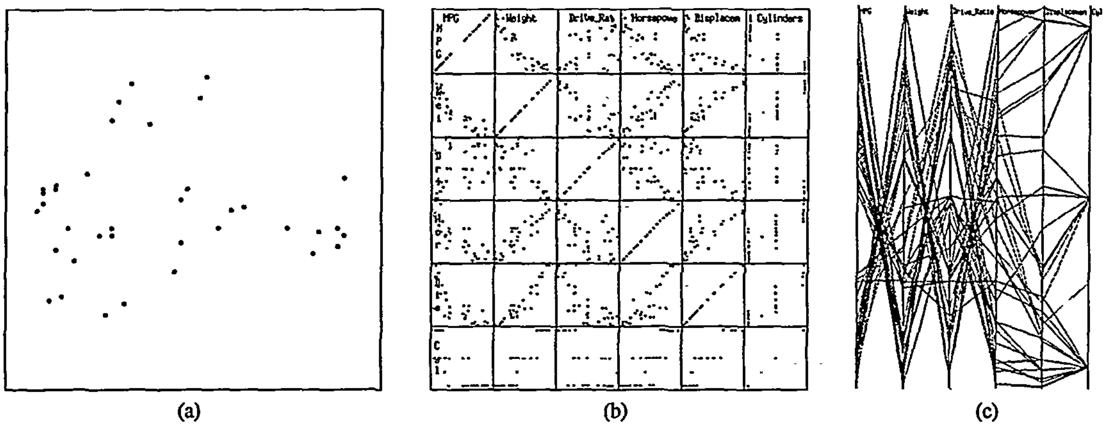


Figure 3: The cluster of the upscale medium size cars is displayed in red in the graphs. See also Color Plate 1.

clustering analysis in a limited form. Figure 3 (Color Plate 1) depicts the same automobile dataset shown in Figure 1 using the low dimensional data overview, scatterplot matrix, and parallel coordinates. The data in red represents the information for the six upscale medium size cars. We notice that the red dots in Figure 3a are all by themselves without any blue dots close by. The same data cluster is not so obvious in Figure 3c where the red polylines are embedded among the blue polylines. It is very difficult to spot the data cluster without brushing, and it is not clear how easy it would be to figure out how to determine the right brush to use. The situation is worse in Figure 3b, where the red dots are scattered all over most of the display tiles. We would not be able to identify this cluster from the scatterplot display alone.

3.2 Display Density

A second advantage of the low dimensional data overview is its relatively low display density. Because of the data reduction during the scaling process, the low dimensional data overview handles larger datasets (especially with a high number of variates) much better than the other two techniques. Figure 4 depicts a dataset[2] of US cities with 10 variates and 329 records that measure various quality of living parameters. As we can see, the scatterplot matrix in Figure 4b suffers from the high number of data variates, which requires 81 display tiles to cover the data completely. The problem is even more obvious with the parallel coordinate display in Figure 4c in which valuable information and data clusters are mostly hidden behind the polylines.

The large number of data points in Figures 4b and 4c also makes it very difficult to brush the data accurately. That is because a majority of data points are plotted very close to each other while some of the others simply vanish due to overlappings. It is also important to realize that even though two data points are displayed in the same neighborhood in certain tiles in Figure 4b (or some axes in Figure 4c), the two data records may still be very far away when all variates are considered.

The data points in the low dimensional overview in Figure 4a have a significantly lower density than those in Figure 4b and there is much less overlap. This makes it much easier to brush neighboring sets of data values than is possible with either of the other displays. In addition, proximity in Figure 4a is a measure of the similarity between data records (i.e., cities) whereas proximity in Figures 4b and 4c only indicate similarity of variate values.

3.3 Outlier Detections

By brushing the two isolated data points at the far left side of Figure 4a, we discover that they represent two major cities (New York and San Francisco) which have higher living costs, more recreation facilities such as theaters, museums, and zoos, and more physicians and hospitals per capita than the other cities included in the database. The same process which allows us to identify this outlying data also enables us to detect anomalous outliers.

The document which comes with the automobile dataset indicates that the Buick Estate Wagon record is an outlier. This is because the data was collected on a test track and the car was operated with a higher than recommended tire inflation pressure, while the rest of the data were collected by EPA under standard test conditions. Since the track condition and the tire pressure are not included in the dataset, the miles per gallon value of the Buick is unexpectedly better than the other cars in the same category. The outlier shows up in the data overview in Figure 1 as the Estate Wagon is located away from the full size sedan/wagon cluster. However, the other six variates of the Buick record keep the car fairly close to its peers. Once again, it is much easier to spot the outlying data (shown in red in Figure 5 and Color Plate 2) in the low dimensional data overview than in the other two visualization techniques depicted in Figures 5b and 5c.

3.4 Multiresolution Visualization

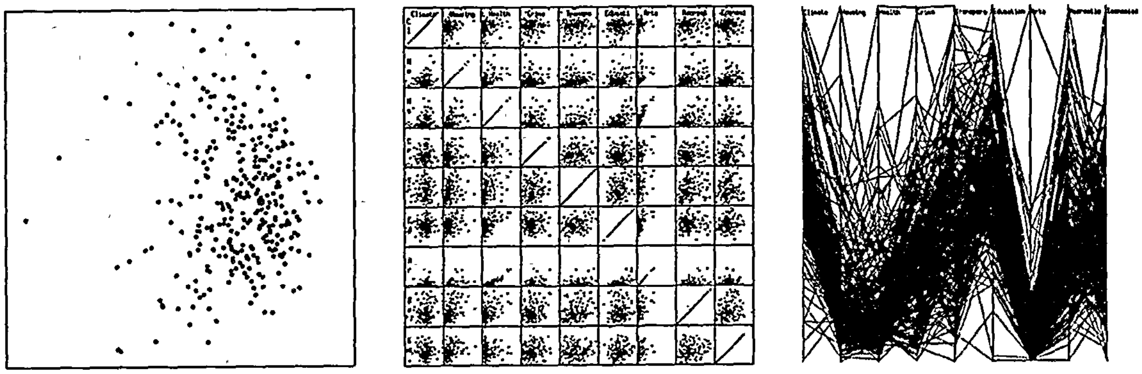
The visualization of large multivariate datasets continues to be one of the major challenges of visualization research. We investigate the use of wavelets to support visualization of data with a large number of records in [17, 18, 20, 16]. In this paper, we propose another *progressive refinement* solution to visualize datasets with a high number of *variates.*

As we mentioned in Section 2.2, the low dimensional data overview based on principal components is only an approximation of the data. Its degree of accuracy is determined by the number of principal components used for display. In many cases, the first two to three principal components of the data overview convey most of the important information of the data. Nonetheless, there are times when a higher number of dimensions is needed.

As an example, we use a 10-variate dataset[3] which contains information about protein consumption in Europe during the 70's. The first five data overviews, which use eigenvectors from the first one, two, three, four, and five principal components, achieve degrees of accuracy of 45.57%, 63.15%, 76.02%, 86.04%, and

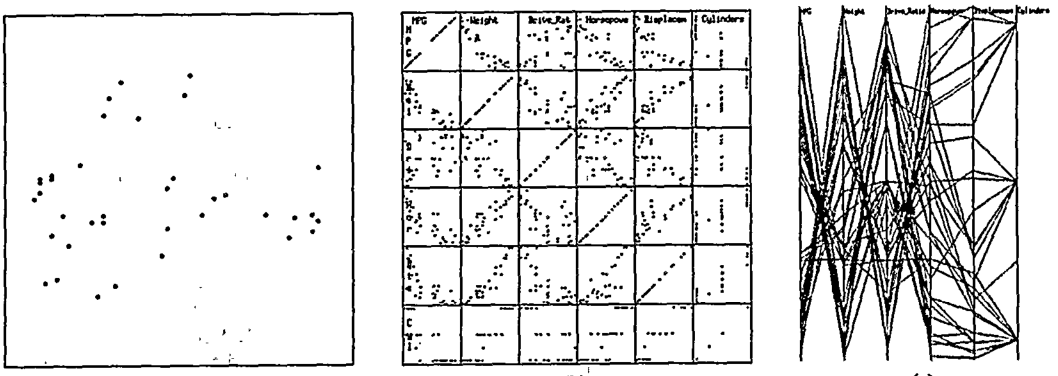
[2] http://lib.stat.cmu.edu/datasets/places.data

[3] http://lib.stat.cmu.edu/DASL/Datafiles/Protein.html



(a) (b) (c)

Figure 4: A 10 variate dataset with data records of 329 US cites.



(a) (b) (c)

Figure 5: The Buick Estate Wagon (an outlier) is displayed in red in the graphs. See also Color Plate 2.

91.26% respectively. After the first component, every unit increase in dimension improves the degree of accuracy by about 5–10%. If we consider the 80% size reduction rate, the two dimensional data overview (63.15%) is indeed a very practical result. Nevertheless, error information is always welcome for data visualization which involves data reduction.

We use the glyph technique [1, 14] to demonstrate the idea of multiresolution visualization. Bear in mind that our goal is to reduce the number of data dimensions, not the size of the data. Figure 6 depicts three coarse to fine data overviews of the protein data in two, three, and four dimensional spaces using the glyph representation, followed by an annotated graph to show the identities of the data points. The first two principal components are mapped to the two axes in all three overviews. In Figures 6b and 6c, the third principal component is mapped to the intensity. In Figure 6c, the fourth component is mapped to the diameter of the glyph, which is a circle. The multiresolution display is particularly important in this dataset because the first two principal components do not reflect a very high degree of accuracy, i.e., 63.15%. This shows up when we look at the data points at the lower left side of Figure 6c, where the glyphs include both intensity and size parameters whose values come from the third and fourth components respectively. Even though the first two major components put these data points close to each other, there are still differences among the neighbors which deserve attention.

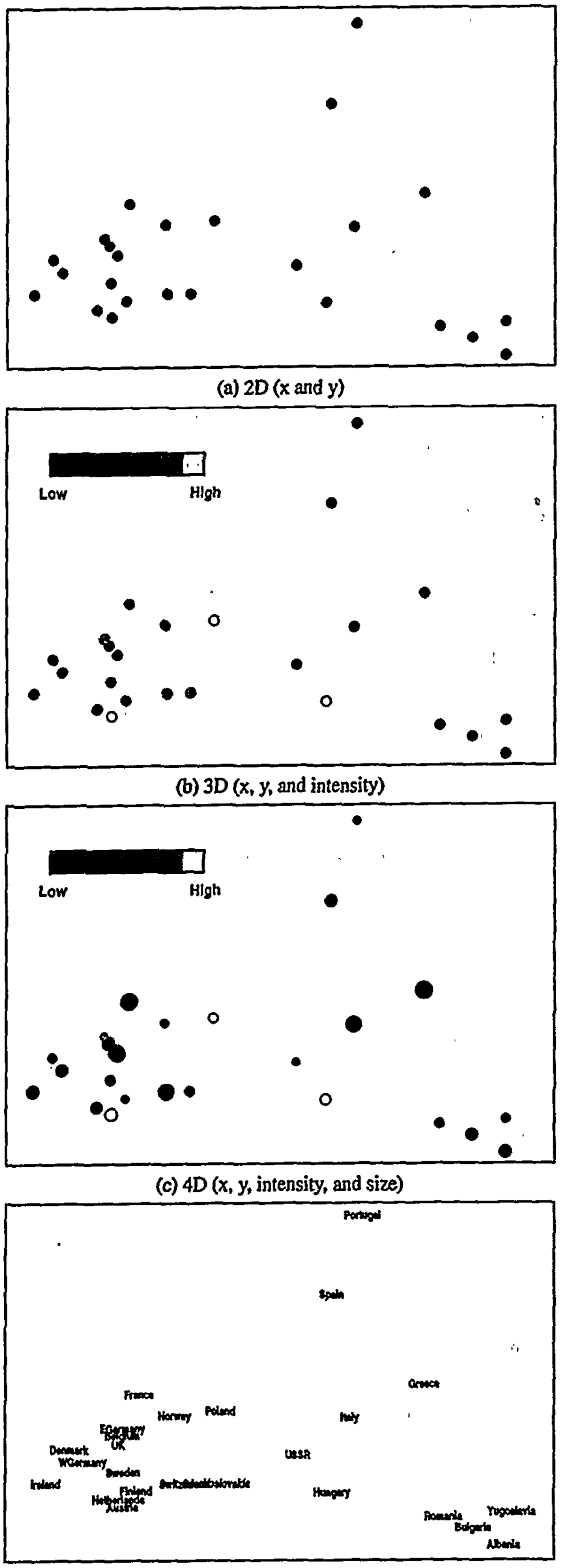
From Figure 6d, we see that the spatial locations of the glyphs resemble the geographical locations of the corresponding countries. The interpretation is that people who live in neighboring countries share similar diets, which determine the way they consume protein. By looking at the degrees of accuracy achieved by the additions of the third and fourth principal components, the intensities and the sizes of the glyphs are expected to play important roles in the interpretations. The display proximity and the similarity of the intensity of the two data vertices representing the two Germanys is strong evidence that they share very similar cultures, even though they belong to two different data groups, i.e., Eastern and Western Europe.

3.5 Shortcomings

It is not entirely fair to compare the effectiveness of visualization tools when we only look at one aspect of the results. The metric scaling technique is good at representing clusters of data points but the individual variate values are lost. For example, a car with more cylinders does not always imply more horse power than one with less. This can be spotted easily in Figure 3c, but not in Figure 3a. In the following section, we describe how to integrate the metric scaling technique with techniques that maintain more of the variate value information.

4 Integration of Techniques

From the previous examples, we have learned that the low dimensional data overview cannot be used to replace the other visualization tools for effective multivariate visualization. The goal is to combine these tools together easily and productively.



(a) 2D (x and y)

(b) 3D (x, y, and intensity)

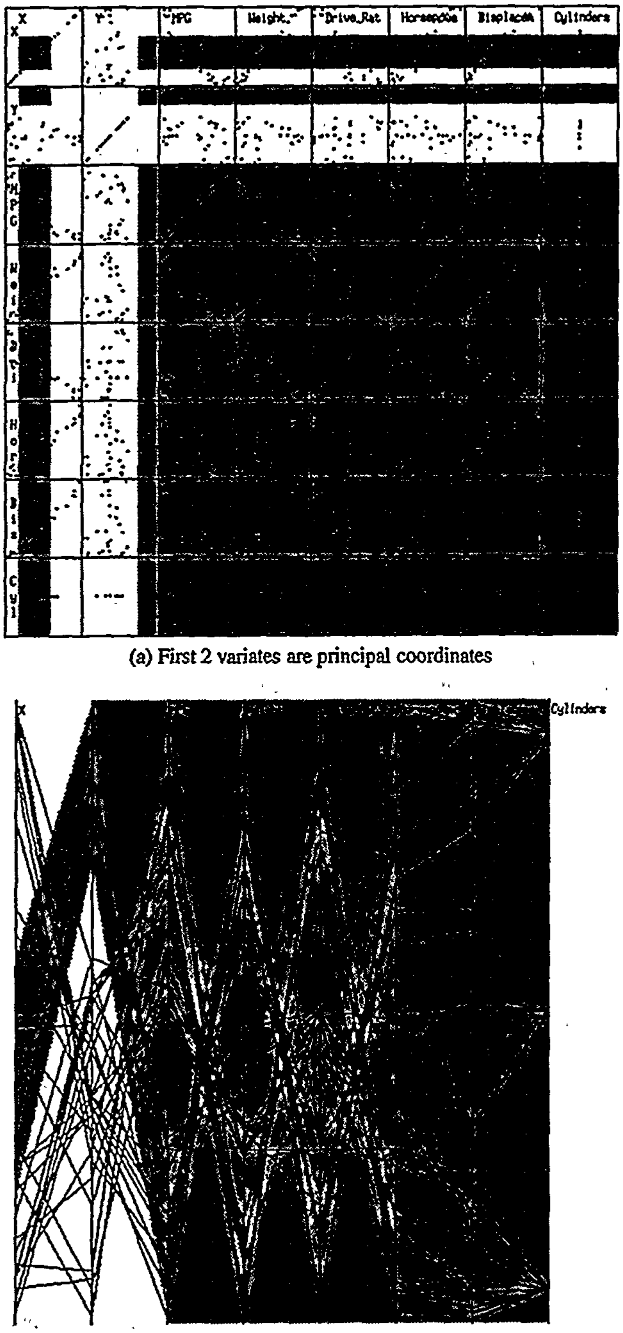
(c) 4D (x, y, intensity, and size)

(d) Annotated data

Figure 6: The glyph representation of three coarse to fine data overviews of the protein data, and an annotated graph to show the identity of the data.

4.1 Linking

Since the size of each principal component of the data overview is the same as the number of records of the data, we can simply treat the components of the data overview as part of the data for visualization. *This approach is* demonstrated in Figure 7 (Color Plate 3)



(a) First 2 variates are principal coordinates

(b) First 2 axes are principal coordinates

Figure 7: Multidimensional brushing with a low dimensional data overview. See also Color Plate 3.

using the same automobile dataset presented in Section 1. The first two variates in the upper left corner of Figure 7a are the principal components of the data overview. The scatterplots generated by these two variates show the principal components view while the scatterplots along the first two rows/columns show how each variate correlates with each of the principal components. With the use of high dimensional data brushing, we can select data points in the principal component overview tiles and study the responses in the other display tiles at the same time. For example, we can brush the cluster of the upscale medium size sedans in tile (x, y) as shown in Figure 7a. Brushing such clusters is difficult without the support of the data overview.

The idea of linking a data overview to a visualization technique seems simple, but the implications reach far beyond just the point-and-click operations. By applying the Euclidean metric to different subsets of variates, we provide a very realistic overview which we can use to query the data with a far more powerful and flexible language than the conventional database query languages using aggregate functions. The data overview presents a graphic summary with reduced data dimensions, reduced data size, additional data semantics, and most important of all, better user-friendliness for the underlying visualization technique.

4.2 Merging

The second strategy is to merge the Euclidean coordinates of the data overview and the data into one visualization display. Conventional glyph representations such as the stick figure icon [14] and autoglyph [1] can be useful only when the icons are arranged in certain dimensions. The patterns may vanish altogether with even a slight change of the plotting axes. This limits the flexibility and functionality of the technique. The low dimensional data overview, however, brings new perspective to the conventional icon visualization. Our approach is to arrange the icons according to the Euclidean coordinates of the data overview determined by the principal components of the data.

An arbitrary glyph based on Beddow's autoglyph, which is defined in Figure 8, is used to visualize the protein data described in Section 3.4 using the same intensity scale as in Figures 6b and 6c. The glyph contains four layers and nine blocks. Each layer approximately represents one food category and each block represents one kind of food. The result is shown in Figure 9. As we can see, the glyphs representing the Balkans indicate that the countries including Yugoslavia, Romania, Bulgaria, and Albania consume a relatively small amount of meat, milk, fruits, vegetables, and fish, but a large amount of starch and cereals. Scandinavian countries such as Sweden and Finland consume much more meat, milk, and vegetables than starch and cereals. The Iberian countries including Spain and Portugal consume a lot of fruits, vegetables, and fish, and only a small amount of meat and milk. And the Mediterranean countries including Greece and Italy have relatively balanced diets. All this information, however, can also be revealed by the icons themselves. However, if we look closely at the spatial locality of the glyphs, countries with higher meat, egg, and milk consumptions (the top two layers) tend to be located at the lower left hand side; countries with higher starch and cereals consumptions (the bottom layer) tend to be located at the lower right hand side; countries with higher fruits, vegetables, and fish consumptions (the third layer) are at the top; and finally, countries with relatively balanced diets are around the middle of Figure 9. These explain why the Balkans are at the lower right hand side, Scandinavian and Western Europe countries are at the lower left hand side, Iberian countries are at the top, and Mediterranean countries are in the middle of Figure 9.

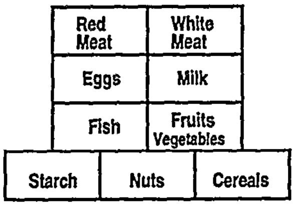


Figure 8: The definition of a nine variate glyph.

The visualization of the data and its principal components together offers a lot more than just the data itself. The principal coordinates of the data also provide new opportunities to create new shape and texture patterns with the conventional iconographic technology.

5 Conclusions and Future Research

We present an implementation of a powerful low dimensional data overview generated by metric scaling. Real life data is used to demonstrate the strengths and weakness of this data overview. It is shown that a combination of the low dimensional data overview with other multivariate data representations can greatly improve the exploration power of the underlying visualization techniques. Currently we are in the process of developing a non-metric scaling model based on *topology*. Our long term goal is to extend the current design to cover categorical data including ordinal and nominal data.

Acknowledgements

This work has been supported in part by the National Science Foundation under grant IRI-9117153. The authors wish to thank Professor Matthew Ward of Worcester Polytechnic Institute, who made his XmdvTool publicly available.

References

[1] Jeff Beddow. Shape coding of multidimensional data on a microcomputer display. In Arie Kaufman, editor, *Proceedings of IEEE Visualization '90*, pages 238–246, Los Alamitos, California, October 1990. IEEE Computer Society Press.

[2] Chris L. Bentley and Matthew O. Ward. Animating Multidimensional Scaling to Visualize N-Dimensional Data Sets. In Stuart Card, Stephen G. Eick, and Nahum Gershon, editors, *Proceedings of IEEE Information Visualization '96*, pages 72–73, Los Alamitos, California, Oct 1996. IEEE Computer Society Press.

[3] Dianne Cook, Andreas Buja, Javier Cabrera, and Catherine Hurley. Grand Tour and Projection Pursuit. *Journal of Computational and Graphical Statistis*, 4(3):155–172, 1995.

[4] Trevor F. Cox and Michael A. Cox. *Multidimensional Scaling*. Monographs on Statistics and Applied Probability. Chapman & Hall, London, 1994.

[5] Mark L. Davison. *Multidimensional Scaling*. Wiley series in Probability and Mathematical Statistics. Wiley, New York, 1983.

[6] Catherine Hurley and Andreas Buja. Analyzing High-Dimensional Data with Motion Graphics. *SIAM Journal on Scientific and Statistical Computing*, 11(6):1193–1211, 1990.

[7] A. Inselberg, M. Reif, and T. Chomut. Convexity algorithms in parallel coordinates. *Journal of ACM*, 34(4):765–801, October 1987.

[8] Alfred Inselberg and Bernard Dimsdale. Parallel coordinates for visualizing multi-dimensional geometry. In T. L. Kunii, editor, *Proceedings of Computer Graphics International '87*, Tokyo, 1987. Springer-Verlag.

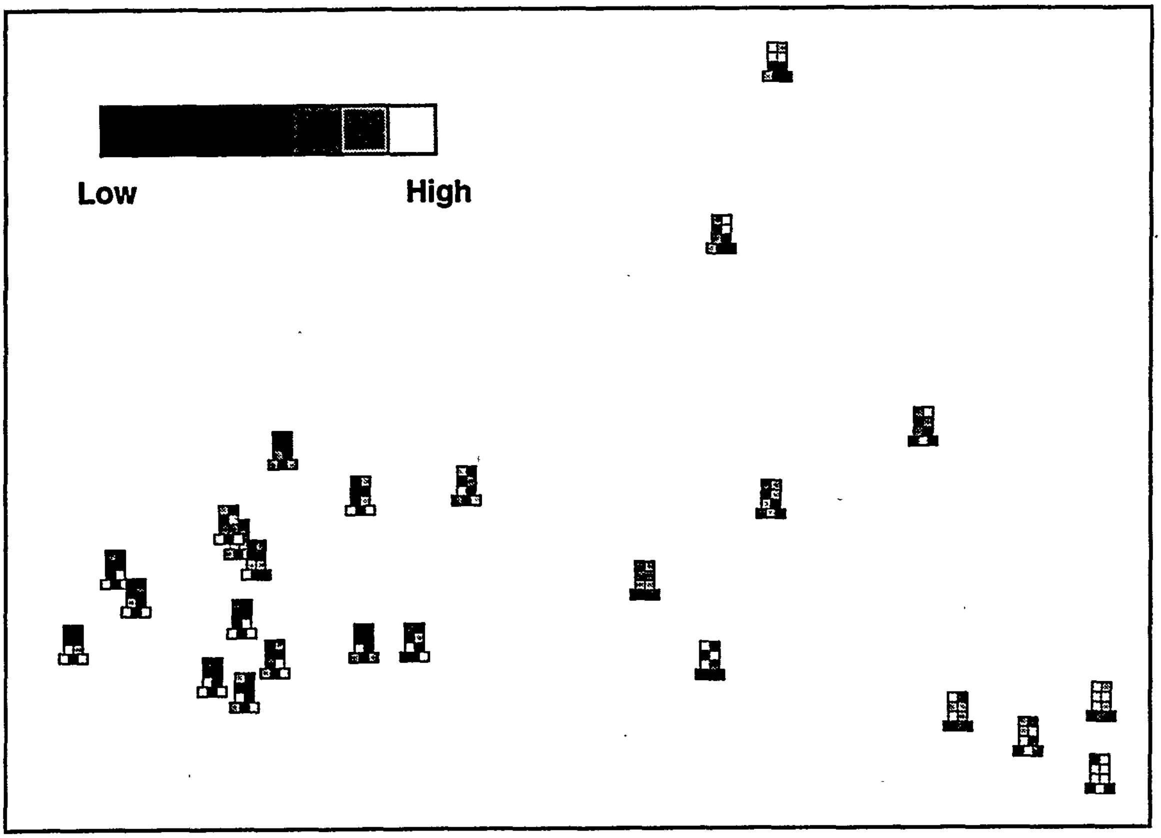


Figure 9: The nine variate protein consumption data represented by the glyph representation.

[9] Alfred Inselberg and Bernard Dimsdale. Parallel coordinates: A tool for visualizing multi-dimensional geometry. In Arie Kaufman, editor, *Proceedings of IEEE Visualization '90*, pages 361–375, Los Alamitos, California, October 1990. IEEE Computer Society Press.

[10] J. Edward Jackson. *A User's Guide to Principal Components.* Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, New York, 1991.

[11] Daniel A. Keim and Hans-Peter Kriegel. VisDB: Database exploration using multidimensional visualization. *IEEE Computer Graphics and Applications*, 14(5):40–49, September 1994.

[12] Daniel A. Keim, Hans-Peter Kriegel, and Thomas Seidl. Visual Feedback in Querying Large Database. In Gregory M. Nielson and R. Daniel Bergeron, editors, *Proceedings IEEE Visualization '93*, pages 158–165, Los Alamitos, California, October 1993. IEEE Computer Society Press.

[13] Allen R. Martin and Matthew O. Ward. High dimensional brushing for interactive exploration of multivariate data. In Gregory M. Nielson and Deborah Silver, editors, *Proceedings IEEE Visualization '95*, pages 271–278, Los Alamitos, California, October 1995. IEEE Computer Society Press.

[14] Ronald M. Pickett and Georges G. Grinstein. Iconographics displays for visualizing multidimensional data. In *Proceedings IEEE Conference on Systems, Man, and Cybernetics*, pages 514–519, Beijing and Shenyang, PRC, May 1988.

[15] Matthew O. Ward. XmdvTool: Integrating multiple methods for visualizing multivariate data. In R. Daniel Bergeron and Arie E. Kaufman, editors, *Proceedings IEEE Visualization '94*, pages 326–336, Los Alamitos, California, October 1994. IEEE Computer Society Press.

[16] Pak Chung Wong. *Adaptive Multiresolution Representation of Large Multidimensional Multivariate Scientific Datasets.* PhD thesis, Department of Computer Science, University of New Hampshire, Durham, New Hampshire, May 1997.

[17] Pak Chung Wong and R. Daniel Bergeron. Authenticity analysis of wavelet approximations in visualization. In Gregory M. Nielson and Deborah Silver, editors, *Proceedings of IEEE Visualization '95*, pages 184–191, Los Alamitos, California, October 1995. IEEE Computer Society Press.

[18] Pak Chung Wong and R. Daniel Bergeron. Multiresolution Multidimensional Wavelet Brushing. In Roni Yagel and Gregory M. Nielson, editors, *Proceedings of IEEE Visualization '96*, pages 141–148, New York, NY, October 1996. ACM Press.

[19] Pak Chung Wong and R. Daniel Bergeron. 30 Years of Multidimensional Multivariate Visualization. In Gregory M. Nielson, Heinrich Mueller, and Hans Hagen, editors, *Scientific Visualization: Overviews, Methodologies & Techniques*, Los Alamitos, California, 1997. IEEE Computer Society Press. In press.

[20] Pak Chung Wong, Andrew H. Crabb, and R. Daniel Bergeron. Dual Multiresolution HyperSlice for Multivariate Data Visualization. In Nahum Gershon and Steven Eick, editors, *Pro-*

ceedings of IEEE Information Visualization '96, pages 74–75, Los Alamitos, California, October 1996. IEEE Computer Society Press.

A Principal Coordinates

This appendix describes the details of generating an inner product matrix from the dissimilarities matrix described in Section 2.1, and solving the Euclidean coordinates of the vertices in n dimensional space from the inner product matrix by the method of principal components.

Let the Euclidean coordinates of the n vertices in the n dimensional Euclidean space be a matrix $X = [x_1, x_2, \ldots, x_n]$ such that $x_r = [x_{r1}, \ldots, x_{rn}]^T$ where $r = 1, \ldots, n$. The Euclidean distance between vertices r and s is given by

$$\begin{aligned} d_{rs}^2 &= [x_r - x_s]^T [x_r - x_s] \\ &= x_r^T x_r + x_s^T x_s - 2 x_r^T x_s. \qquad (1) \end{aligned}$$

Hence

$$\begin{aligned} \frac{1}{n}\sum_{r=1}^{n} d_{rs}^2 &= \frac{1}{n}\sum_{r=1}^{n} x_r^T x_r + \frac{1}{n}\sum_{r=1}^{n} x_s^T x_s \\ &\quad -\frac{2}{n}\sum_{r=1}^{n} x_r^T x_s. \qquad (2) \end{aligned}$$

If we standardize the data to have zero mean and unit variance, the center of mass of the vertices is the origin. So we have

$$\frac{1}{n}\sum_{r=1}^{n} x_r^T x_s = 0.$$

Also, since

$$\frac{1}{n}\sum_{r=1}^{n} x_s^T x_s = x_s^T x_s,$$

equation (2) becomes

$$\frac{1}{n}\sum_{r=1}^{n} d_{rs}^2 = \frac{1}{n}\sum_{r=1}^{n} x_r^T x_r + x_s^T x_s. \qquad (3)$$

Similarly,

$$\frac{1}{n}\sum_{s=1}^{n} d_{rs}^2 = \frac{1}{n}\sum_{s=1}^{n} x_s^T x_s + x_r^T x_r. \qquad (4)$$

Furthermore, from (4),

$$\begin{aligned} \frac{1}{n^2}\sum_{r=1}^{n}\sum_{s=1}^{n} d_{rs}^2 &= \frac{1}{n^2}\sum_{r=1}^{n}\sum_{s=1}^{n} x_s^T x_s + \frac{1}{n}\sum_{r=1}^{n} x_r^T x_r \\ &= \frac{1}{n}\sum_{s=1}^{n} x_s^T x_s + \frac{1}{n}\sum_{r=1}^{n} x_r^T x_r \\ &= \frac{2}{n}\sum_{r=1}^{n} x_r^T x_r \qquad (5) \end{aligned}$$

Defining an inner product matrix B such that

$$[B]_{rs} = b_{rs} = x_r^T x_s,$$

and substituting (3), (4), and (5) into (1) gives the inner product matrix B in terms of d_{rs},

$$\begin{aligned} b_{rs} &= x_r^T x_s \\ &= \frac{1}{2}\left(x_r^T x_r + x_s^T x_s - d_{rs}^2\right) \\ &= \frac{1}{2}\left(\frac{1}{n}\sum_{s=1}^{n} d_{rs}^2 - \frac{1}{n}\sum_{s=1}^{n} x_s^T x_s + \frac{1}{n}\sum_{r=1}^{n} d_{rs}^2 \right. \\ &\quad \left. -\frac{1}{n}\sum_{r=1}^{n} x_r^T x_r - d_{rs}^2\right) \\ &= \frac{1}{2}\left(\frac{1}{n}\sum_{r=1}^{n} d_{rs}^2 + \frac{1}{n}\sum_{s=1}^{n} d_{rs}^2 - \frac{1}{n^2}\sum_{r=1}^{n}\sum_{s=1}^{n} d_{rs}^2 - d_{rs}^2\right). \end{aligned}$$

The next step involves the use of principal components to recover the Euclidean coordinates of the n dimensional space denoted by the matrix X from B. We seek a different coordinate system in the d dimensional display space such that the distances between points in the display space can be approximated by measuring the distances only along some subset of the axes of this new coordinate system. By the definition of an inner product matrix, B can be expressed as

$$B = XX^T. \qquad (6)$$

Since B is symmetric and positive semi-definite (i.e., only some of the eigenvalues are positive), it has p positive eigenvalues. Let Λ be the eigenvalue matrix in which the diagonal is the sorted eigenvalues $\lambda_1, \ldots, \lambda_p, \ldots, \lambda_n$. Let the corresponding normalized eigenvector of Λ be V. By the definition of eigenvectors, matrix B can be described as

$$B = V\Lambda V^T.$$

Since there are only p positive eigenvalues, B can be expressed as

$$\begin{aligned} B &= V_1 \Lambda_1 V_1^T \\ &= V_1 \Lambda_1^{\frac{1}{2}} \Lambda_1^{\frac{1}{2}} V_1^T \qquad (7) \end{aligned}$$

where Λ_1 is the eigenvalue matrix in which the diagonal is the eigenvalues $\lambda_1, \ldots, \lambda_p$ and V_1 is the corresponding eigenvector of Λ_1. From (6) and (7), therefore

$$X = V_1 \Lambda_1^{\frac{1}{2}}.$$