TECHNIQUES FOR VISUALIZING MULTIDIMENSIONAL DATA

W T Hewitt, S Larkin, A J Grant
Manchester Visualization Centre, Manchester Computing, University of Manchester, UK

Abstract
This paper will review methods for visualization of multidimensional and multivariate data, covering techniques such as scatter plots, Chernoff faces, Andrews plots and parallel coordinates.

1 WHAT IS MULTIDIMENSIONAL DATA?

1.1 Introduction

Usually we use the acronym mDv for multidimensional data, with the m indicating that there are m dimensions or in a physical experimental setup m independent variables, and v corresponds to the number of dependent variables of at each point in m dimensional space. We can assume that both m and v are significantly bigger than 3 otherwise the problem would be straightforward. But what is the problem? Typically such data sets are in excess of 100Mbytes, and may as large as Terabytes, from which we extract sufficient information to enable us to gain and understanding of what is going on and subsequently present some visual information to enable the viewer to make some deduction or inference from the picture.

By implication there is some complex relationships embodied in this vast amount of data.

1.2 Examples

Some examples of the sources of such data are national census data sets, stock exchange data, and two specific examples of “small” data sets are:

The authors have been working with National Power, a UK electricity generating company. They supply the UK national grid with electricity, and to do this they offer the electricity in 30 minute units. But of course they are in competition with a number of other companies, and given that there are 13 variable controlling the cost, how do you provide timely analysis?

Another example is from a researcher in the Department of Sociology, who has spent a long time collecting data about people who held office in Medieval times. It contains a number of variables, including name, year and position in the community, and is far from complete. He is trying to analyze job movement, promotion/demotion, and kinship/nepotism.

Of course mDv datasets arise in many of the scientific disciplines, and computational fluid dynamics, and high energy physics are examples of such data.

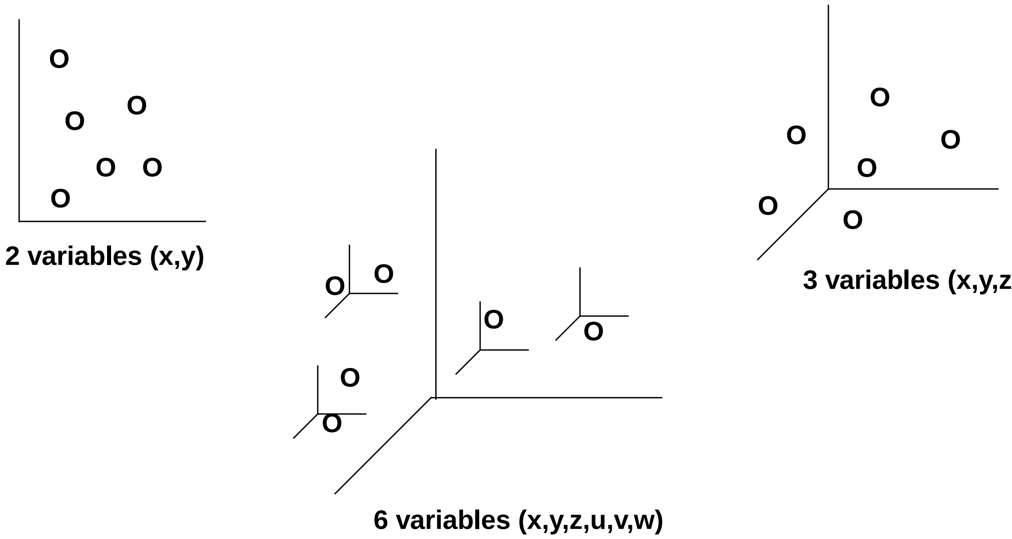
1.3 Why not use existing techniques?

We could apply the known techniques for scalar and vector, 2D and 3D data on subsets of these large data sets, and whilst it might give some useful results, its difficult to decide how to subset, and of course it fails to convey an overall impression of the data. Correlations could be made by stacking or overlaying results, though careful use is needed as they can produce cluttered and incomprehensible results.

In reality the techniques used are based upon the existing methods. There are two major problems with mDv visualization: the techniques are application specific, there are few generic methods, and most methods require some selection, by the user, of which portion of the data to view. The correct selection of data may be critical to the success or failure of a particular method.

1.4 Coping with greater than 3D

One simple way would be to extend the 2D and 3D scatterplots illustrated in Fig. 1 to higher dimensions as shown, by introducing local-axes for the extra dimensions, emphasizing that it becomes hard to perceive, navigate, relate and compare values. It is difficult to choose the principle axes. The proper visualization of 3D objects on 2D screens is a topic in its own right, though will not be covered in this paper.



.

Fig. 1 Greater than 3D

2 SOME TECHNIQUES

The main techniques to be reviewed, in this paper, are Glyphs [1], [2], [4], [5], [6], [16], Textures [3], [9], [13], [17] Tables and Stacked Plots [2], [12], Scatterplots [2], [7], Andrews curves [10], Permutation Matrix [8] (not covered in this paper), Parallel coordinates , Data Sonification [18], [20] and Virtual Reality [22] (not covered in this paper)

2.1 What are you looking for?

The techniques generally produce results which appear to be very cluttered and the viewer must spend sometime gaining experience in the use of all the techniques described here. Usually the methods , require the viewer to look for: a) unexpected results or anomalies (spotting a stranger), b) grouping or clusters, or c) patterns or trends and correlations

2.2 Glyphs

Graphical icons (glyphs) are attributed to Edgar Anderson in 1957, who used instead of the local axes of Fig. 1 symbols or glyphs at each of the points in 2D. He used circular icons with rays (Fig. 2)

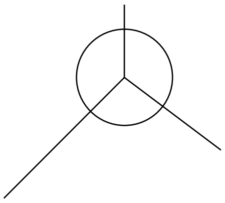


Fig. 2 Anderson Glyph

where the values beyond the 2nd dimension were encoded via the radius of the circle and the length of the rays. In 1966 Pickett White used a triangle with sides and orientation related to different variables, and in

2.2.1 Chernoff Faces

In 1973 Chernoff whose name is usually associated with this technique used a traditional 2D scatterplot with facial characteristics to represent 3,4,5,...,22 variables. Variations are normally grouped into distinct classes:

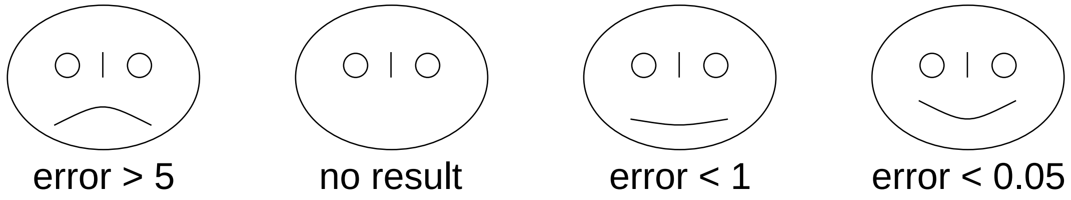


Fig. 3 Encoding error as variation of the mouth; one component of a Chernoff Face

The methods allows the viewer to spot trends or strangers (Fig. 4), as it relies on the fact we are good at recognizing faces

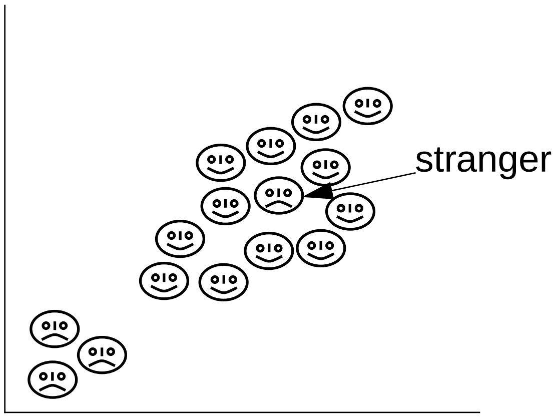


Fig. 4 An example Chernoff Face Plot

Clearly other symbols could be used, for example oil tankers, for data from the petroleum industry.

2.2.2 Star Glyphs

An extension of the above method is the Star glyph each dimension in the dataset is represented as a “prong” in the star, [19]

For each datapoint a star is drawn with the size of the “prongs” representing the value in each dimension for that particular point:

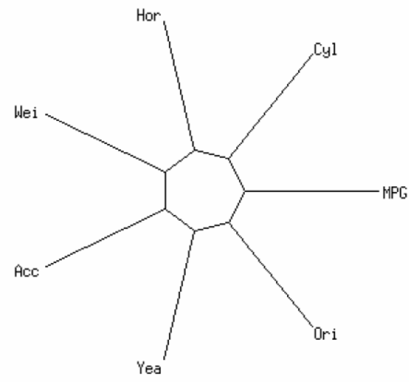


Fig. 5 Example of Star Glyphs

2.2.3 Haber Glyphs

The Haber glyph [5] is used to visualize the stress-strain in a tensor in an engineering mechanics application. The stress tensor can be split into the sum of a symmetric and anti-symmetric parts. The glyph is a cylinder and an ellipse (Fig. 6), the cylinder axis direction shows major principal direction, of the stress, the ellipse axes showing the other two directions. The cylinder and axis lengths show stretching in each axis.

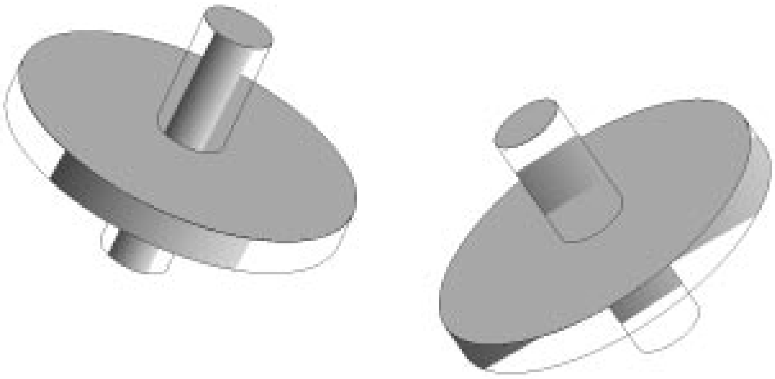


Fig. 6 The Haber Glyph

2.2.4 de Leeuw and van Wijk glyphs[4]

de Leeuw and van Wijk extended this method to visualize the tensor field in the context of the associated velocity field, for steady state flows. This and the Haber glyphs are best used as probes or a small number distributed throughout the data. The method is to construct a local coordinate axis as with Haber glyphs, and then decompose the velocity field tensor into *parallel* and *perpendicular* components. Components for display (Fig. 7) are then extracted from these:

- **acceleration, shear, curvature (parallel)**
- **torsion, divergence (perpendicular)**

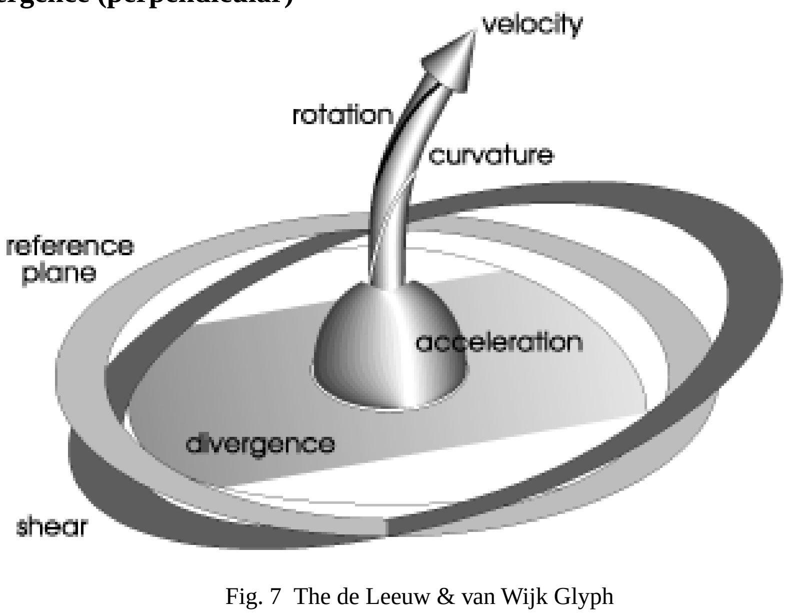


Fig. 7 The de Leeuw & van Wijk Glyph

2.3 Textures

On 3D surface plots information such as height above a plane is encoded through colour, in addition texture (bump mapping) may be added to encode yet another dimension.The bump map is a collection of bumps (texture) used to add additional information to a graphical primitive. For example the surface may appear rough in regions of for example high wind, and smooth in regions of low wind.

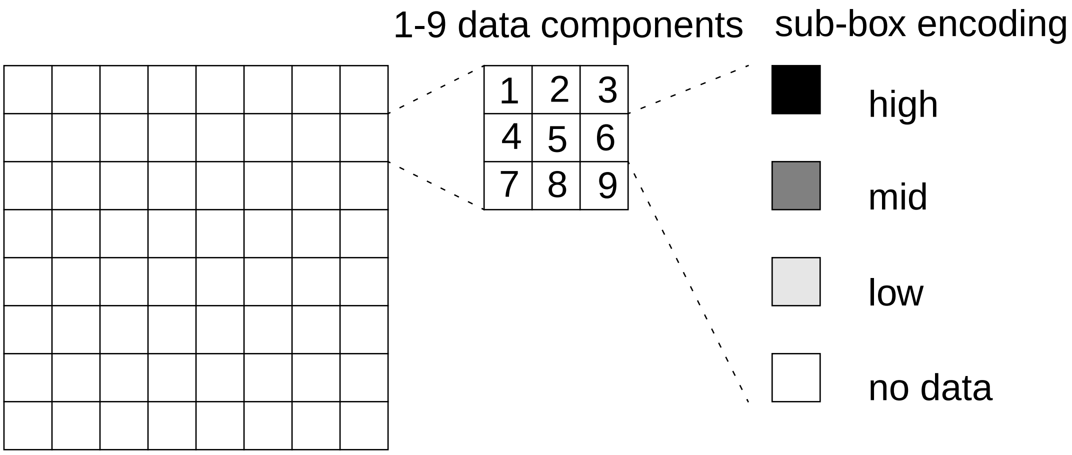
Interactive adjustment of parameters is desirable to obtain best results and careful use is needed as additions to an already rough surface can be distracting.

Texture maps can be used to represent more information about vectors and tensors than just magnitude. The Line Integral Convolution method [9], [17], produces through texture a continuous version of an ‘arrow’ or ‘vector’ plot. The output image is a one-one correspondence of a 1D convolution of a filter kernal and texture pixels along a local streamline in the vector field. More simply the texture is “smeared” in the direction of the vector field.

This work has led to the use of texture for the visualization of tensor fields [9] and looks a promising line of research.

2.4 Tables

Each point in the dataset is represented as a rectangle and the rectangle contains encodings for the value of the point in each particular dimension in the dataset [1]



We show an example of the 13 parameters of magnetosphere and solar wind readings taken every hour over a number of days from NASA Goddard Space Flight Center.

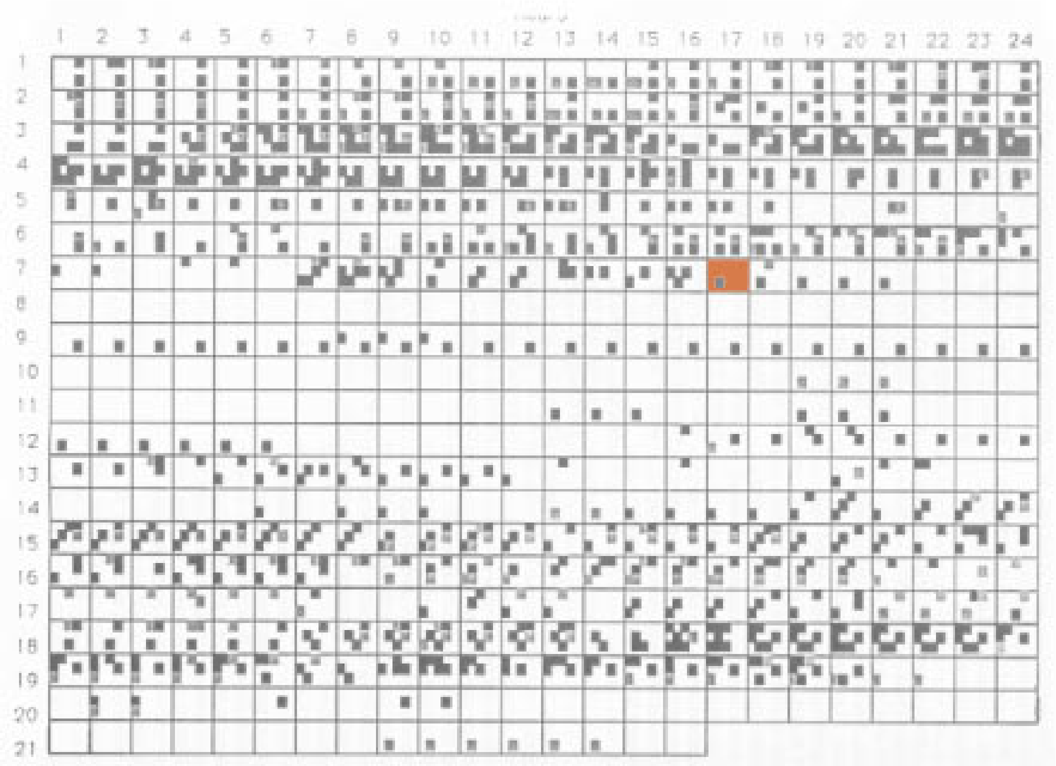


Fig. 8 Beddow J, Microsimulations Research, [1]

2.5 Scatterplot Matrix

A scatter plot, more commonly called a graph of y versus x, shows the relationship of 2 variables and with the addition of colour can represent a 3rd variable. A scatterplot matrix of n variables is obtained by projection of the data onto n*(n-1) scatter plots, i.e., all possible combinations of scatter plots are drawn as illustrated in (Fig. 9 and Fig. 10) which is an example for pressure, temperature, and velocity (6 plots) data...

Pressure	PvT	PvV
TvP	Temperature	TvV
VvP	VvT	Velocity

Fig. 9 Scatter Plot Matrix, each gray rectangle is an y vs x plot

Notice the plot in the diagonal correspond to x vs x and are redundant, and each combination is plotted twice, one the transpose of the other. This redundancy may be useful.

2.6 Andrews Curves

The Andrews curve and parallel coordinates of the next section are methods for representing a multi-dimensional data point via a 2D curve or polyline respectively. In the Andrews covers each multidimensional point x (x_1,x_2,...,x_m) is mapped to a periodic functionG(t) via, for example a function like:

$$G(t) = \frac{F1}{\sqrt{2}} + F2\sin(t) + F3\cos(t) + F4\sin(2t) + F5\cos(2t) + \dots$$

.

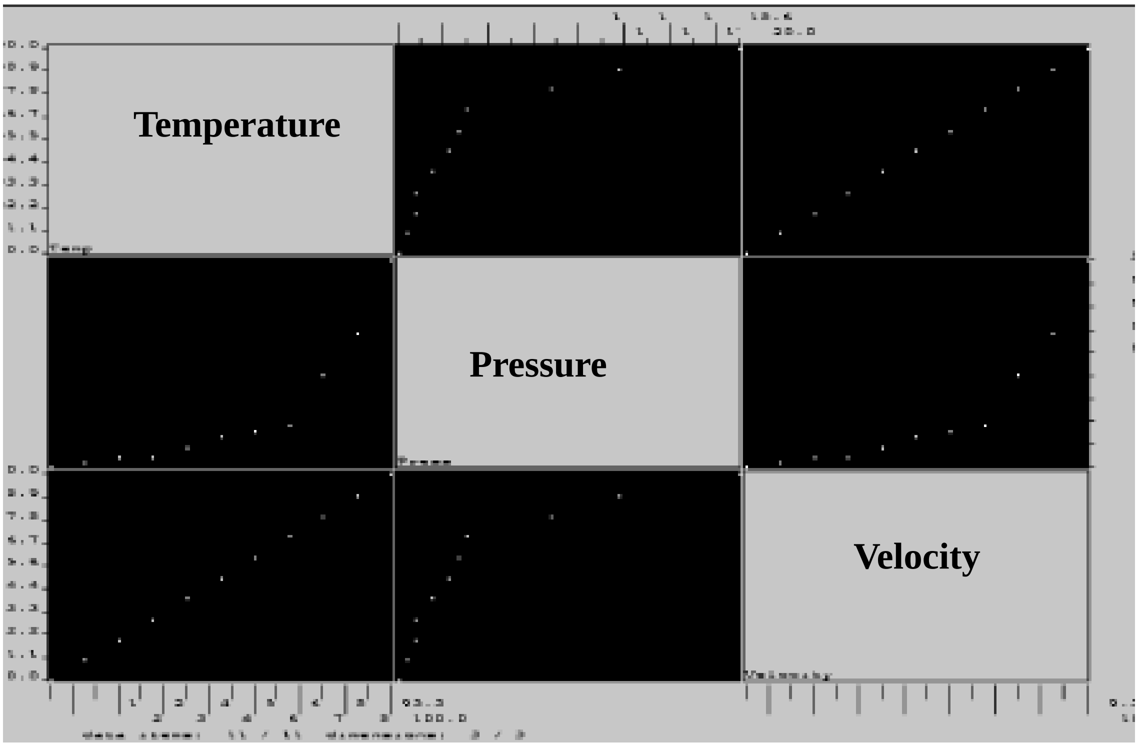


Fig. 10 The Scatter Plot of Pressure, Temperature and Velocity

The curves are plotted over the, range $-\pi < t < \pi$, and produces an iconic representation of each point through multidimensional space, and clusters of n-dimensional points map to similar shaped curves

It is not possible to pinpoint single data components i.e., all the data components are combined into one function

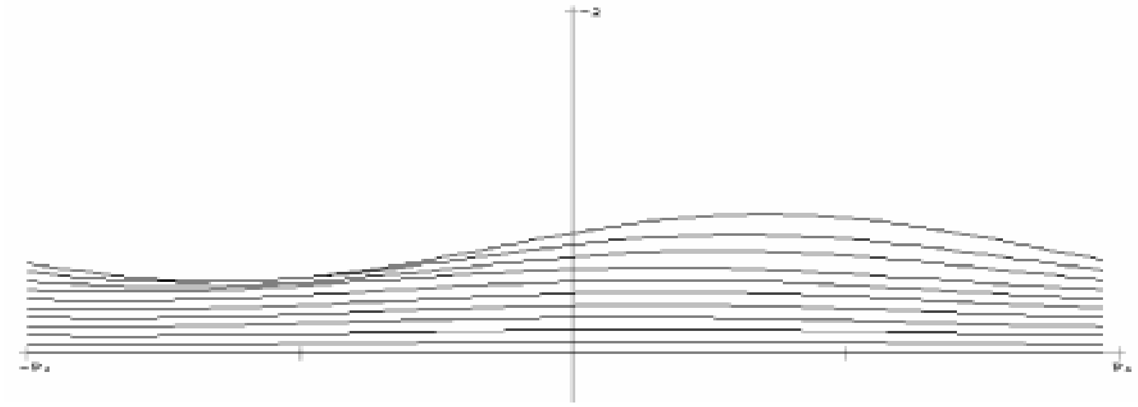


Fig. 11 Andrews curves - Simple Points through pressure, temperature, velocity

2.7 Parallel Coordinates

Due to Alfred Inselberg [11] parallel coordinates organize each axis vertically on a 2D and for each multidimensional point x $(x_1,x_2,...,x_m)$ mark the appropriate value on the axis, and join the marks with line segments.

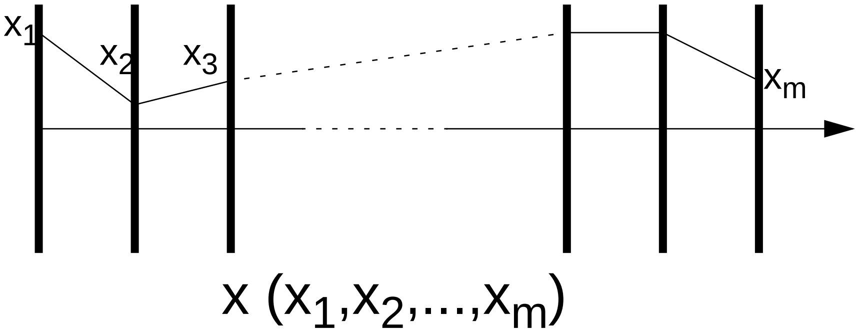


Fig. 12 Parallel Coordinates

Therefore a m dimensional point is represented as a line through m parallel vertical lines or axes. The results seem extremely cluttered, and systems which provide this technique allow interactive marking and highlighting of groups of lines. There are some patterns/shapes to look for, A proportional to B will give parallel lines, and A inversely proportional to B will give lines crossing over each other.

Fig. 13 show a picture of some data assoicated with car manufacture. MPG is theconsumption of the car in miules per gallon, Cylinders, Horsepower, Weight and acceleration are obvious, year is the year of manufacture, and Origin is one of Europe, North America or Japan. All Japanese cars are highlighted.

2.8 Data Sonification

2.8.1 Introduction

Data Sonification is the name associated with the use of sound to complement a graphical representation [18,20]. The assimilation to data is simple, or is it? Sounds has seven attributes which can be used to encode data:

2.8.2 Pitch

Logarithmic changes in frequency produce linear changes in pitch, and this is intuitive for relating to magnitude of a scalar component. There are similar problems as with colourmaps; adjacent values are difficult to distinguish

2.8.3 Loudness

Variations in amplitude are not linear as it is also affected by frequency and timbre changes

2.8.4 Timbre

Or wavering where different instruments play the same pitch/loudness can be used to differentiate between data components

2.8.5 Location

The physical location of the sound source, which is affected by acoustics of the surrounding environment, can provide locational cues to results

2.8.6 Rhythm

The music is organized around a periodic event rate or pulse and can be used to represent temporal separation between time stamped events or behavioral cycles.

2.8.7 Duration

It is hard to distinguish the duration of a sound unless it is exaggerated, and should not be used as quantitative measure but it is useful to identify outliers or activity lifetimes

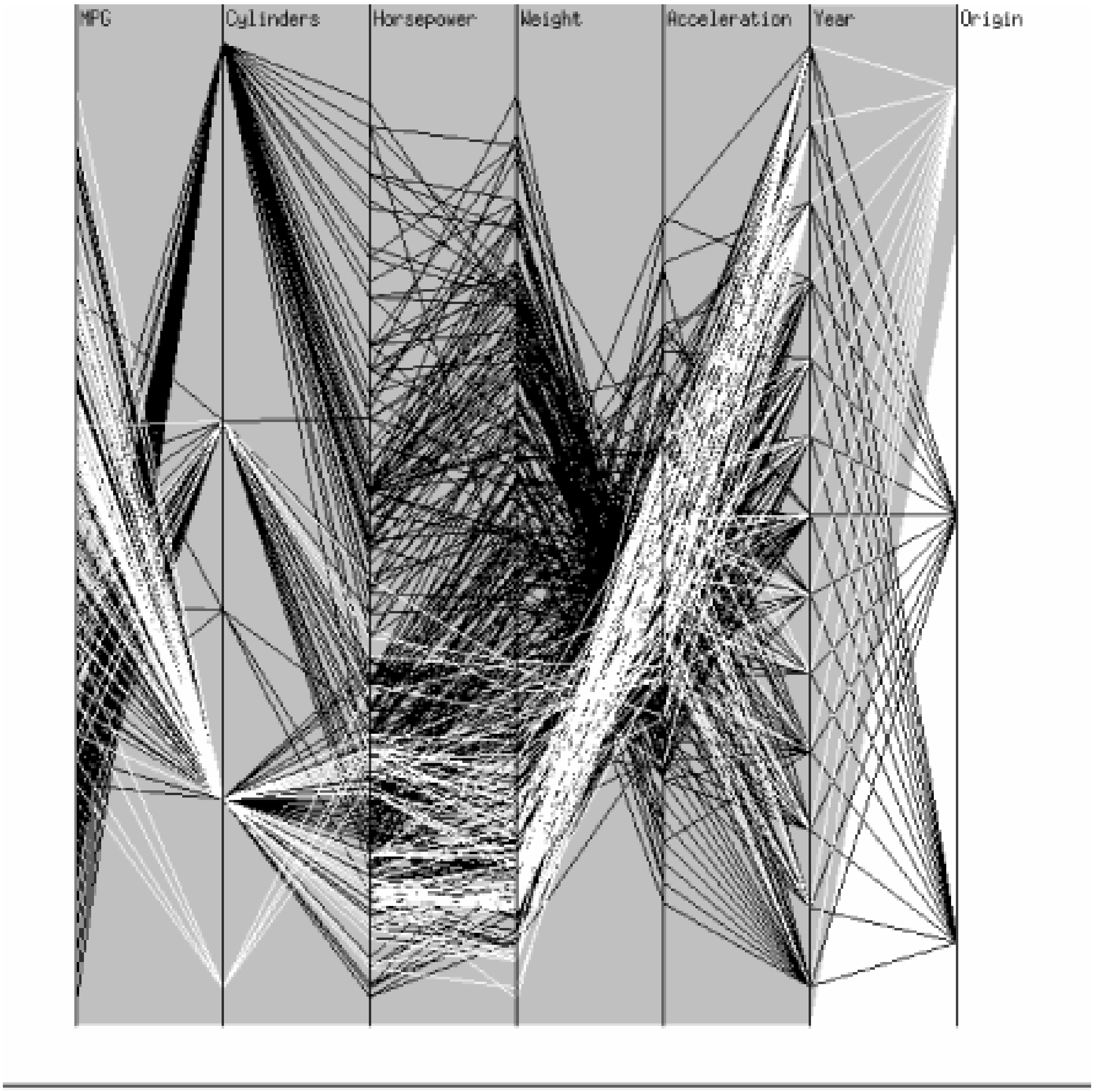


Fig. 13 Parallel Coordinate plot of car manufacture data

2.8.8 Melody

"The first thing remembered, the last thing forgotten". What constitutes a melody is the subject of considerable research but certain patterns of notes are more "melodic" than others; therefore the choice of scale or starting pitch is significant

2.8.9 Some real examples

This technique has been used in the analysis of climate data; a probe samples data components and assimilates them to sound: wind: varying the pitch of a siren, and rain: varying the amplitude (loudness) of the sound of "rain".

It has been used in the Stanford ParalleL Applications for SHared memory benchmark suite (SPLASH) where the type of process (system, network, application) is mapped to pitch, the process's time quantum is mapped to duration, and the processor to an instrument.

2.8.10 Conclusions

Sound is as complex a medium as other more traditional ones for visualization e.g., colour, There are many pitfalls, yet it is an interesting and exciting research area. You have to be aware of the "tone deaf" equivalent of a "colour blind" user

3 CONCLUSIONS

This paper has presented a number of techniques for visualizing complex data. The key features are that the techniques tend to be application specific and not generic, the viewer must spend time learning to use the technique.

4 ACKNOWLEDGMENTS

The authors wish to thank all members of the Manchester Visualization Centre for the help.

5 REFERENCES

[1] Beddow J, "Shape Encoding of Multidimensional Data", Proceedings of IEEE Vis '90, pages 238-246

[2] Tufte E R, "Envisioning Information", Graphics Press, 1990

[3] Cabral B, Leedom L C, "Imaging Vector Fields using Line Integral Convolution", SIGGRAPH '93 Proceedings, pages 263-272

[4] de Leeuw W C, van Wijk J J, "A Probe for Local Flow Field Visualization", Proceedings of IEEE Visualization '93, pages 39-48

[5] Haber R B, "Visualization Techniques for Engineering Mechanics", Computing Systems in Engineering I, 1990, pages 37-55

[6] Ellson R, Cox D, "Visualization of Plastic Injection Moulding", Simulation 51, 5, 1988, pages 184-188

[7] Cleveland M, "Elements of Graphing Data", Wadsworth, 1985

[8] Bertin J, "Semiologie graphique", Editions Gauthier-Villars, 1967

[9] Delmarcelle T, Hesselink L, "The topology of 2nd order tensor fields", Proceedings of IEEE Visualization '94, pages 140-148

[10] Andrews D, "Plots of Higher Dimensional Data", Biometrics, March 1972, pages 125-136

[11] Inselberg A, "Parallel Coordinates - A Tool for visualizing multi-dimensional geometry", Proceedings of IEEE Visualization '90, pages 361-390

[12] Eick S G, Steffen J L, "Visualizing Code Profiling Line Oriented Statistics", Proceedings of IEEE Visualization '92, pages 210-217

[13] Crawfis R A, Allison M J, "A Scientific Visualization Synthesiser", Proceedings of IEEE Visualization '91

[14] Gardiner V L, Lazarus R B, Stein P R, "Solutions of Diophante Equation $x^3 + y^3 = z^3 - d$", Math Comp 18, 1964, pages 408-413

[15] Nielson G M, "Modelling and Representing Multivariate Data", Course Notes on Advanced Techniques for Scientific Visualization, SIGGRAPH '94, Orlando Florida, July 1994.

[16] Chernoff H, "The use of faces to represent points in k-dimensional space graphically",

Journal of American Statistical Association 76, June 1973, pages 361-368

[17] Forsell L K, “Visualizing Flow over curvilinear grid surfaces using Line Integral Convolution” Proceedings of IEEE Visualization ‘94, pages 240-247

[18] Scaletti C, Craig A B, “Using sound to extract meaning from complex data”, `http://www.ncsa.uiuc.edu/VR/VR/Papers/sound.ps`

[19] Siegel J H, Farrell E J, Goldwyn R M, Friedman H P, “The surgical implications of physiological patterns in myocardial infractions shock”, Surgery Volume 72, pages 126-141, 1972

[20] Madhyastha T M, Reed D A, “Data Sonification: Do you see what I hear?”, `http://bugle.cs.uiuc/edu/Papers/IEEEsound.ps`

[21] Ward M O, “XmdvTool: Integrating Multiple Methods for Visualizing Multivariate Data”, Proceedings of IEEE Visualiztion ‘94, pages 326-336

[22] Bryson S, Levit C, “The virtual windtunnel: an environment for the exploration of 3D