Shape coding of multidimensional data on a microcomputer display.

Jeff Beddow
Microsimulations Research
5321 Chateau Place
Minneapolis, MN 55417
E-Mail: Beddow@vz.acs.umn.edu

Abstract:

The visual representation of data from complex systems, whether databases, measured scientific data, or simulation output, holds the promise of discovering patterns in the data that will increase its management efficiency while revealing relationships invisible to numeric methods. In this paper we present a simple and flexible method of shape coding for higher dimensional data sets that allows the database operator or discipline scientist quick access to promising patterns within and among records or samples. The example used is a thirteen parameter set of solar wind, magnetosphere, and ground observation data collected hourly for twenty-one days in 1976.

Background.

There is currently a great deal of interest in understanding the topology of higher dimensional data sets. Applications for methods developed in this area encompass a broad range of computer-related fields, including management of extensive databases, scientific visualization, and complex process control. Visual methods in this field seem promising and are the subject of extensive research at this time[1] . One visual method involves mapping data values to a compound graphic object, usually called a "glyph" and occasionally called an "icon."[2] The precedent for this approach is to be found in advanced statistical methods. [3] ,[4] During the 60's and 70's several methods were proposed for projecting complex data sets onto graphic surfaces. Most of these efforts were ignored or sank into obscurity at the time because of the difficulty of creating and interpreting them. [5] Current interface and computing technology eliminates the drudgery of successive iterations of design, allowing the researcher to seek optimum display parameters in fairly short sessions. A greater commitment to the process is required, however, as perception of the salient features of complex glyph surfaces can usually require much more time than the inspection of histograms, draughtsman's arrays of scattergram's box and whisker charts, etc.

[1] McCormick, B., DeFanti, T. & Brown, M., "Visualization in Scientific Computing", Report to the National Science Foundation by the Panel on Graphics, Image Processing and Workstations. 1987

[2] Littlefield, Richard J., "Using the glyph concept to create user-definable display formats" Proceedings of the Fourth Annual Conference and Exposition of the National Computer Graphics Association.1983

[3] Wang, Peter, "Graphical Represenation of Multivariate Data", New York: Academic Press, 1978

[4] Everitt and Dunn "Advanced Methods of Data Exploration and Modelling.",London;Heinemann, 1983

[5] Conversation , Wesley Nicholsen, Batelle Labs, 1987

CH2913-2/90/0000/0238/$01.00 – 1990 IEEE

The system to be studied:

The solar wind and the Earth's magnetic field have an interesting relationship which is being studied from many perspectives. The plasma, or ionized gas outflow from the Sun, consists of both positively and negatively charged particles. When these particles encounter the magnetosphere of the Earth, it is similar to water molecules in a river encountering the pier of a bridge. A bow shock wave is formed, and behind that turbulence occurs in the transitional area. In these areas the behavior of the solar wind is similar to air or water currents. Because of the electrons that accompany the ions, however, there are enormous electrical currents in the plasma stream, and these interact further with the Earth's field, producing side effects as spectacular as the Auroras and as problematic as vast "radio storms" that interfere with radio communications over entire hemispheres. The bow shock always faces the Sun. Opposite is a tail that extends at least a million miles.

The fluid nature of the bowshock, magnetopause, and tail result in perturbations of the entire system when particularly active streams of particles from the sun arrive. The forces that generate the Earth's dipole magnetic field are fairly stable from the scale of a human lifetime. The interaction with the Sun's charged particles creates a vast, sensitive, electrically active caul around the planet that has significant impact on many people's daily activities.

Watching the system.

These behaviors are too complex to be completely predicted by mathematical models. An optical facility in Boulder Colorado keeps track of disturbances of the Sun's surface that will result, several days later, in perturbations of the Earths magnetosphere. These are reported to military and corporate clients whose activities depend on clear radio transmissions. Since the magnetosphere is not visible to the naked eye, its shape must actually be calculated from data about ion density, temperature and direction. Most of the interplanetary medium data is collected by satellites, either earth or solar orbiting. The quality and nature of the data from the satellite is affected by its orbit. An earth orbiting satellite is always in an elliptical orbit, with a near and far point that keep their alignment relative to the celestial sphere, but change their alignment relative to the earth-sun line as the earth moves in its orbit. Every six months the major axis will be wholly within the magnetotail. When it is in the magnetotail, it is not collecting accurate data about the solar wind. This leaves significant gaps in the data record, and lesser gaps of a daily scale as the satellite moves behind the earth in every orbit.

The data set.

NASA' s Goddard Space Flight Center has collected the interplanetary medium and magnetic activity data over the last twenty years and packaged it for general scientific uses on the NODIS system as the OMNI data set. This set consists of some 35 parameters which can be downloaded by the investigator in ascii format. I chose the following thirteen for this exercise:

1. Field magnitude average: this is the average strength of the magnetic field, ignoring any shifts in vector orientation.

2. Average magnetic vector strength: This average includes direction, which would cancel out two 50-

gamma readings from opposite directions, for example.

3. Vector Latitude: An indication of magnetic field direction based on a spherical coordinate system, as is #4.

4. Vector Longitude

5. Plasma Temperature: A measure of kinetic energy in the plasma particles, not comparable to measurement of atmospheric temperature, for example

6. Ion Density: Density of ions in plasma flow, not directly correlated locally to electron density.

7. Flow Speed:Solar wind, outside of magnetopause, as are #8 & #9.

8. Flow Longitude: An indication of solar wind direction.

9. Flow Latitude: An indication of solar wind direction.

10. Kp*10: a measure of Earth's magnetic field strength and motion from the ground

11.C9: An index of activity in the 10 cm radio bandwidth.

12. r (sunspot)

13. year/day/hour

The parameters of the solar wind are taken from sensors outside the magnetopause. This ensemble was chose to depict, in broad strokes, concurrent states and levels of activity in the three main domains of the system: solar wind, magnetosphere, and ground level effects.

The display system and software:

The software system is a prototype developed by the author to demonstrate the glyph approach to depicting higher-dimensional data sets. It accepts data from the clipboard of a Macintosh personal computer with minimum format constraints, and automatically scales the display for physical size of glyph, location of glyph, and color mapping based on the mean and standard deviation of columns. If you copy a 625 row and 25 column table from the clipboard, the system will present the data as a major matrix of 20x25 cells and a minor matrix within each cell of 5x5 elements. It will calculate the mean and standard deviation for each column in the table, and assign colors to each of three subranges and zero or missing data values within the column value range. The default subranges are:

Low values, $<-1\sigma$

Middle values, between -1σ and $+1\sigma$

High values, $>+1\sigma$

Once the default display is constructed, the investigator has control over several of the mappings from data to graphic that take place:

1. Shape of display matrix

2. Number of standard deviations above and below the mean to establish color thresholds

3. Shape of elements

4. Non-standardized scaling of elements to emphasize certain variables, bringing them into the perceptual foreground, as it were.

5. Visibility and color coding of elements.

6. Alternate support geometry consisting of an n-dimensional Euclidian distance measure/scaling facility that allows the selection of any subset of variables to be grouped and computed as the y axis and any other subset of variables to be grouped and computed as the x axis of a scatter diagram. In this technique the entire glyph is presented instead of a single point, with or without labels.

7. Limited animation facility for moving the color threshold down from the top and up from the bottom of the range partition in regular steps, allowing the perception of the "knottedness" of variables in higher dimensional space,

as well as the general distribution of values within the entire data set.

The key features are:

▪Automated translation of data format

This increases the likelihood of using the program, and of submitting interesting or important data to it.

▪Automated setting of defaults for an initial display:

1. Size and position of glyph and its internal elements

2. Subranges within value ranges for color coding

3. Color assignment

This also increases the attractiveness of the program, and greatly improves the researcher's orientation to the data.

▪Interactive control over color map, shape·assignment and differential scaling of elements.

Once the general layout of the data has been assessed, many variations can be tried by the operator in establishing an optimum figure/ground relationship between high and low-priority variables or parameters.

The experiment:

The experiment was to depict all parameters simultaneously, and see if any global or local patterns emerged. Several difficulties attended the use of this data which should be noted. The occlusion of the sensor bearing satellites by the earths magnetotail results in dropouts in the data set, with several parameters missing for days at a time. In Fig. 2 the missing data has been coded as solid black elements. (The last sample screen on the video.) The scale of the set, at approximately 6000 data points, was appropriate for the visual coding strategy of depicting each hour as a sub matrix of solid squares within the display matrix.

The Display.

(Please refer to figures)

The display matrix presented one entire day's data on one row, with twenty-one rows. This arrangement leaves space at the bottom of the screen for the detail alphanumeric data, which is polled by the cursor. By allowing polling of alphanumeric detail in real time, the investigator is free to look for and interrogate any number of potential visual patterns without the interruptions of continual requests to display and discard display of alphanumeric identities or values.

The Figures.

Figure 1.

This is a laserwriter snapshot of a color screen which maps the value set as follows:

1. Missing values -> White
2. Low values -> Grey
3. Mid Values -> White
4. High Values -> Black

In the third and fourth rows from the top and bottom (the symmetry is coincidental) there is a dense clustering of high and low values which show both identifiable shape and shape development over time as a function of the topology of the higher dimensional space.

As an exploratory device this succeeds in indicating to the investigator some areas to follow up. One example starts to take shape in row three, column four, is established in column 7 and breaks up in column 12. The "z" shape that occurs consists of high readings in the field magnitude, average vector, ion density, kp, c9 and sunspot indices. A slight sub pattern of low value in flow latitude starts to appear. When the middle of the "z" drops out in column 12, the element to the left of center goes high, along with the element on the right edge of the center row, forming a broken donut.

An even stronger sequence starts in row 3, column 18 and continues through row 4, column 12, or further, depending upon the investigators sense of threshold for shape groups. An interesting set of diagonal triplets starts to take shape at the end of row 14, is quite coherent through the first eight hours of row 15, struggles to reshape itself through the remainder of row 15 before giving up at the end of the row. Finally, at the end of row 17 there is an echo of the morphology of row 3 and 4.

Figure 2.

This maps the values as;

1. Missing values -> Black
2. Low values -> White
3. Mid Values -> White
4. High Values -> White

It is not necessary, in this paper, to pin down the exact relationships that give rise to these shapes and validate or invalidate them statistically. The example is sufficient to show that even fairly simple pattern formative results enrich the perceptual response to the intrinsic data geometry, and supplement traditional computational techniques, while potentially depicting patterns that other methods would miss altogether.

Color Page.

Figures c1-c4

These figures show the actual color display as follows:

Figure c1.

This maps the values as;

1. Missing values -> Black
2. Low values -> Blue
3. Mid Values -> Red
4. High Values -> Green

The entire data set is displayed as it would appear to the operator on initial conditions.

Figure c2

This maps the values as;

1. Missing values -> White
2. Low values -> Cyan
3. Mid Values -> White
4. High Values -> Red

This allows the operator to perceive grouping of high and low values within individual cells, while discriminating between the high and low.

Figure c3

In this figure the shape and scaling assignments have been changed. One variable of interest has been assigned a outline circle shape, and its scaling has been increased to five times normal in order to exaggerate its variation while preserving information about activity among the other variables. A related variable has been assigned a filled circle shape, while retaining its normal scaling. This is a technique that establishes and exploits a "figure-ground" relationship among variables that is determined interactively, according to the operator's priority.

It is difficult to explain interactive visualization methods in print, but an analogy might be made with the stops on an organ, that affect timbre, voicing, etc. across the entire range of tones. In this respect, it can be understood that nonlinear quantitative operations can serve the interest of qualitative perception, which pays off in distinctions that are ultimately quantitative, but not predictable from the initial conditions or appearance.

Figure c4

In this figure the shape and color maps are the same as in Fig. c3, but the value subrange thresholds are changed consistently among all variables as follows:

Low values, $<-3\sigma$

Middle values, between -3σ and $+3\sigma$

High values, $>+3\sigma$

This serves two main purposes: as a decluttering strategy it simplifies the visual field. As a quantitative strategy it identifies distribution of extreme values within the data set.

Visual browsing of higher dimensional displays is a supplement to the numerical and statistical analysis of patterns, trends, and exceptions within the data. There are several methods for extracting some shape sense from higher dimensional data sets, such as principle component analysis, multi-dimensional scaling, cluster analysis, and variants on these techniques. [4] Most of these techniques are forced to make assumptions about the distribution of values within the data set. The visual inspection of the data patterns does not require that any assumptions be made.

Implications of the research and further directions.

This experiment proves that much more complex data can be presented for visual pattern extraction than standard methods allow. This has significant implications for traditional database management problems, where "similarity recognition" is more important than set-theoretic filtering and extraction.

Other methods of shape coding might involve more sophisticated visual techniques that exploit visual edge detection, negative space perception, and complementary value enhancement . A display feature that treated each glyph as one frame in an animated sequence would combine motion and shape coding for enhanced feature extraction.

More time needs to be spent in lab conditions comparing subject response to the display to computed response. Other methods of interactively changing display characteristics in real time need to be developed on more powerful microcomputers. Studies need to be made of the cross-over threshold where the "gestalts" of each glyph become mere texture components of a large field display.

Interactive methods of marking groups to submit for further analysis are necessary. Neural network methods could be used to characterize and recognize distinctions among data sets, and qualitative infrastructures within the data set.

More research needs to be done in which traditional knowledge of competent graphic artists is combined with "transparent" interface strategies to facilitate browsing and rapid re-mapping of display parameters.

Acknowledgements:

The author would especially like to acknowledge Lloyd Treinish, then of Goddard, now of IBM, and H. Kent Hills, of Goddard, for their patient assistance in acquiring and interpreting the data set used in this paper. Any mistakes in interpretation are entirely my own.

Follow-up.

"The Blackboard", a newsletter on this and related topics, is currently published by the author of this paper.

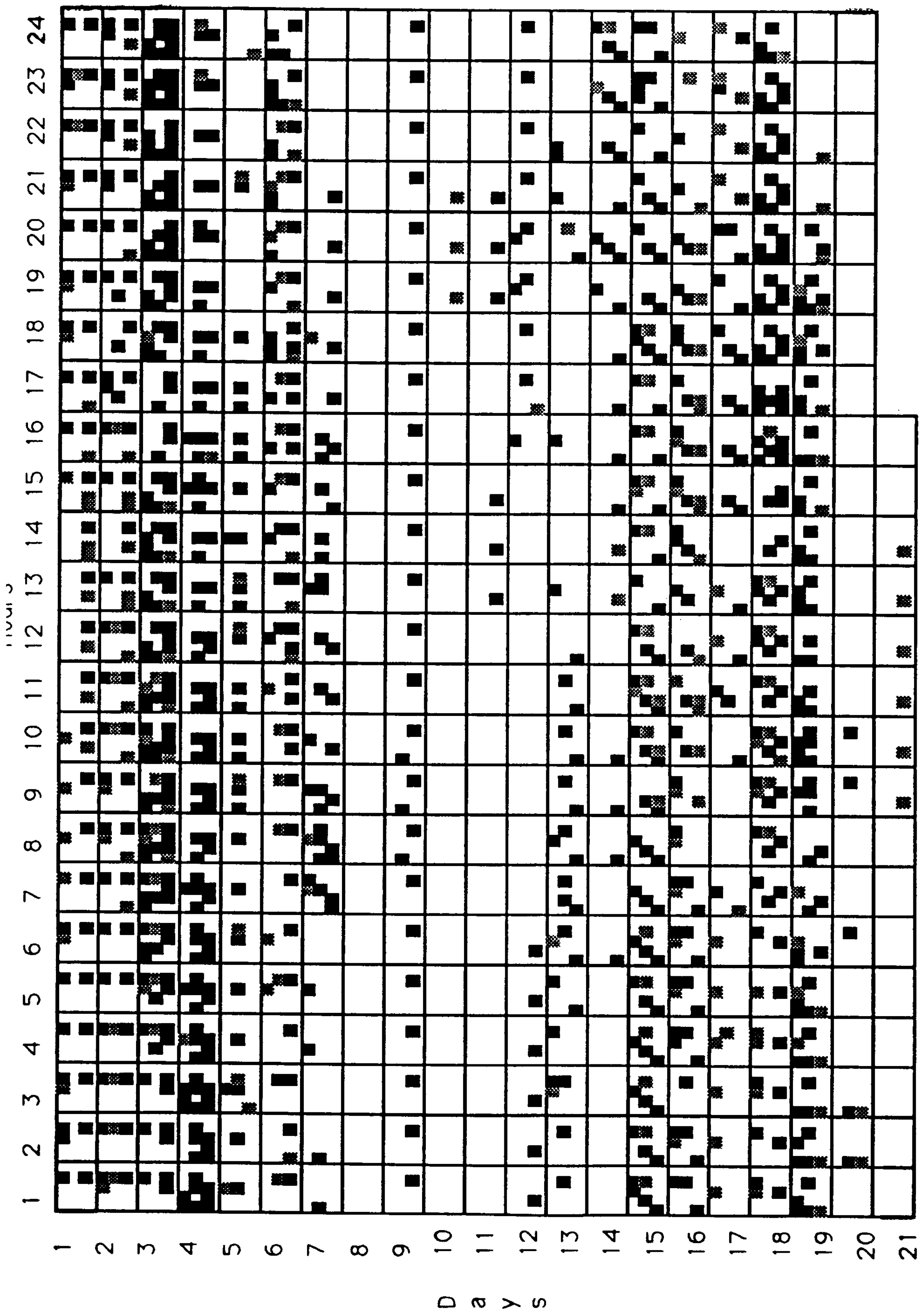

Figure 1 :
Day by Hour: Thirteen Parameters of Magnetosphere and Solar Wind Data

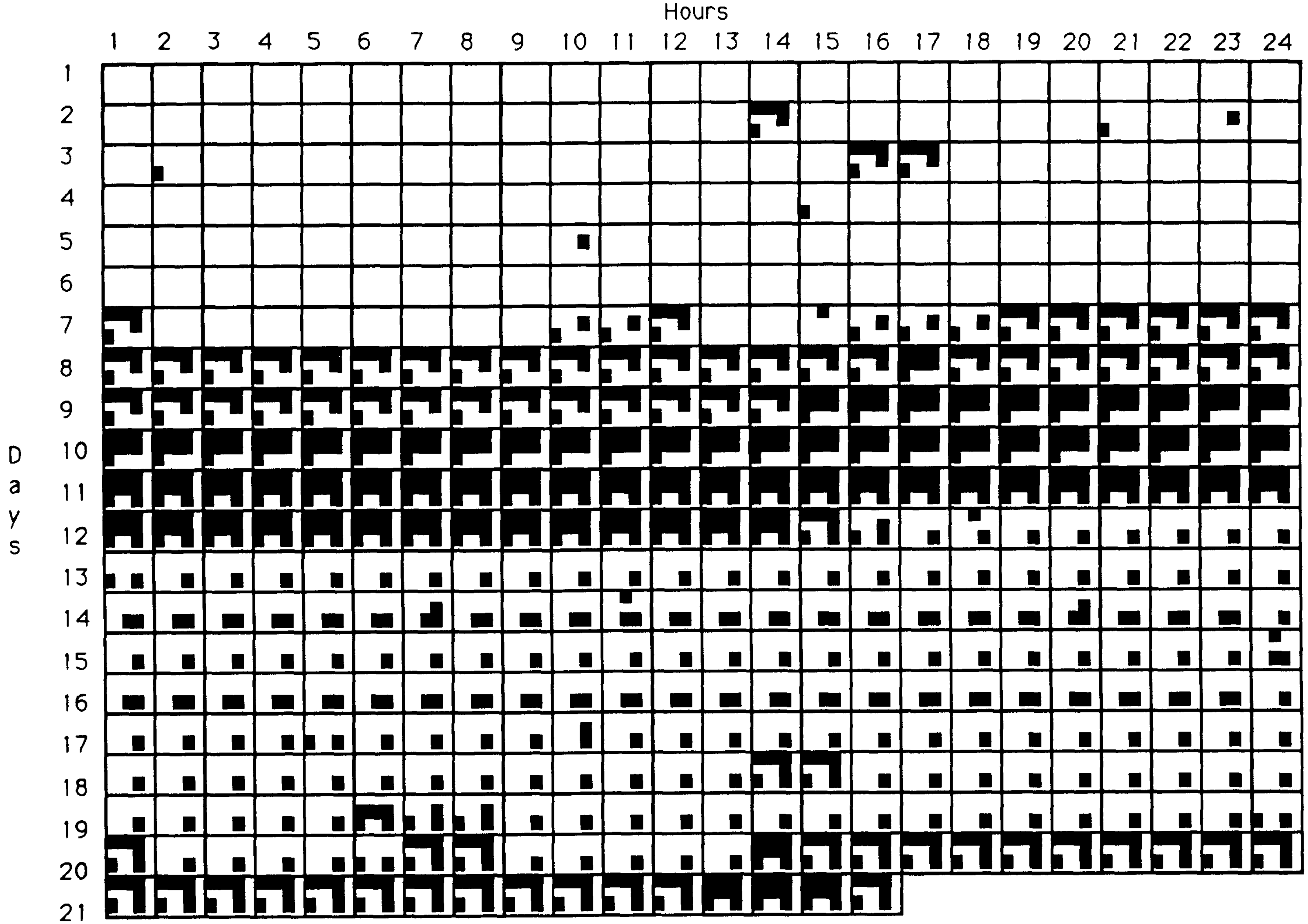

Figure 2:
Day by Hour: Thirteen Parameters of Magnetosphere and Solar Wind Data

Shape Coding of Multidimensional Data on a Microcomputer Display

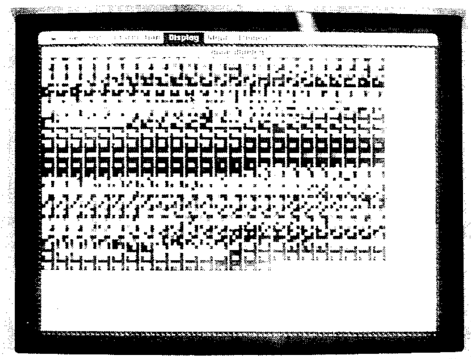

Figure c1. *(Color Plate 109, page 478)*

This maps the values as;

1. Missing values —> Black
2. Low values —> Blue
3. Mid Values —> Red
4. High Values —> Green

The entire data set is displayed as it would appear to the operator on initial conditions.

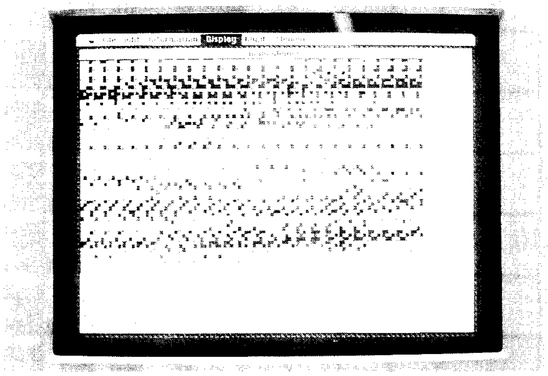

Figure c2 *(Color Plate 110, page 478)*

This maps the values as;

1. Missing values —> White
2. Low values —> Cyan
3. Mid Values —> White
4. High Values —> Red

This allows the operator to perceive grouping of high and low values within individual cells, while discriminating between the high and low.

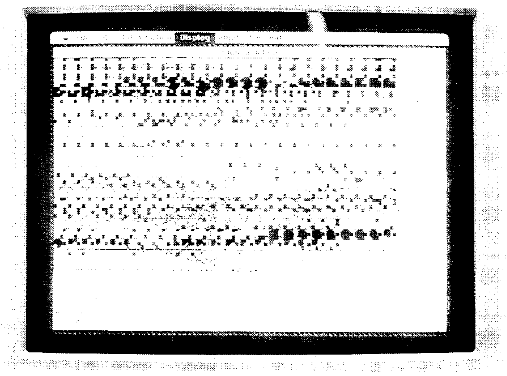

Figure c3 *(Color Plate 111, page 478)*

In this figure the shape and scaling assignments have been changed. One variable of interest has been assigned a outline circle shape, and its scaling has been increased to five times normal in order to exaggerate its variation. A related variable has been assigned a filled circle shape, while retaining its normal scaling. This is a technique that extablishes and exploits a "figure-ground" relationship among variables that is determined interactively, according to the operator's priorities.

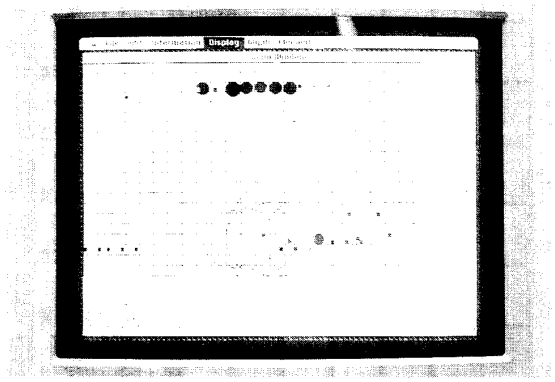

Figure c4 *(Color Plate 112, page 478)*

In this figure the shape and color maps are the same as in Fig. c3, but the value subrange thresholds are changed consistently among all variables as folows:

Low values, <-3σ

Middle values, between -3σ and +3σ

High values, >+3σ

This serves two main purposes: as a decluttering strategy it simplifies the visual field. As a quantitative strategy it identifies distribution of extreme values within the data set.