A Visualization Tool for Exploratory Analysis of Cyclic Multivariate Data

MATTHEW O. WARD
Worcester Polytechnic Institute, 100 Institute Road, Worcester, MA 01609, USA

BENJAMIN N. LIPCHAK
Compaq Computer Corporation, 129 Parker Street, Maynard, MA 01754, USA

Abstract: This paper presents an interactive visualization tool for the qualitative exploration of multivariate data that may exhibit cyclic or periodic behavior. Glyphs are used to encode each multivariate data point, and linear, stacked, and spiral glyph layouts are employed to help convey both intra-cycle and inter-cycle relationships within the data. Users may interactively select glyph and layout types, modify cycle lengths and the number of cycles to display, and select the specific data dimensions to be included. We validate the usefulness of the system with case studies and describe our future plans for expanding the system's capabilities.

Key Words and Phrases: cyclic data; periodic data; time series; multi-dimensional data; multivariate data; visualization.

1 Introduction

Visualization is a common technique employed to gain a qualitative understanding of a data set prior to focussed computational analysis. By displaying a pictorial representation of the data and providing users with interactive tools for manipulating the data and views, visualization allows human pattern recognition skills to contribute to the analysis process. Visual identification and classification of outliers, redundant variables, clusters, patterns, and trends can facilitate the creation of accurate statistical models of the data (Cleveland, 1993, Huber, 1995).

Most commonly used statistical graphics techniques focus on univariate or bivariate data. Indeed, histograms, line plots, and scatterplots are found in most, if not all, statistical packages and are routinely used for tasks ranging from exploration to the confirmation of models to the presentation of results. Fewer methods, however, are available for visualizing data with more than two variables, as the logical mapping of data dimension to screen dimension can no longer be directly exploited. Still, many techniques have been developed and reported in the literature, including scatterplot matrices, parallel coordinates, glyphs, and embedded dimension (trellis) displays.

Besides dimensionality, another attribute of a data set is its implicit or explicit structure. For example, the data may be cyclic or periodic in nature, with a seasonal or otherwise predictable repetition that can be identified and isolated from the trend or noise components. Detecting this structure can be done in an analytic fashion (Bloomfield, 1994, Hwarng, 1995, Wang, 1993), but is susceptible to outliers, noise, and interference amongst the multiple variables. Even if the data is not cyclic in its own right, the viewer may wish to impose a cyclic structure on the data. For example, a company's quarterly results may not follow a cyclic pattern, but the viewer of this data might like to see how the data changes on a quarter-to-quarter basis as well as how the data changes on a year-to-year basis for the same quarter.

Our research focuses on developing visualization techniques for multivariate data that are customized to emphasize the structure as well as the content of the data. In particular, we are interested in methods of displaying cyclic data that permit users to quickly detect inter-cycle and intra-cycle patterns and anomalies, and use this information to enhance quantitative analysis. In this paper we first present an overview of multivariate and cyclic visualization techniques and describe some of their limitations. We then describe the system we have developed, named SpiralGlyphics, for exploring multivariate data sets with implicit or explicit cyclic behavior. We demonstrate the usefulness of the system through a case study, and summarize with a description of the limitations of and future directions for our work.

2 Visualization Techniques for Multivariate Data

There are four general approaches to visually presenting multivariate data:

1. Use multiple views of the data, each of which communicates a subset of the dimensions. Scatterplot matrices are an example of this

method (Fisherkeller, 1975), where all possible pairs or triplets of dimensions are used to form a grid of projections.

2. Reconfigure the axes so that they are no longer orthogonal and attempt to display data along all axes simultaneously. Examples of this method include parallel coordinates (Wegman, 1987, Inselberg, 1990) and glyphs/icons (Chernoff, 1973, Feinberg, 1979, Siegel, 1972).

3. Embed or combine two or more data dimensions to form composite spatial dimensions, such as found in dimensional stacking (LeBlanc, 1990) and trellis displays (Becker, 1996).

4. Reduce the dimensionality, as in statistical techniques such as principal component analysis (Jolliffe, 1986), factor analysis, dimensional clustering (Duda, 1973), Kohonen self-organizing feature maps (Kohonen, 1995), non-linear mapping, or multidimensional scaling (Kruskal, 1978), until the data is easily displayed using traditional graphics techniques.

Each method has its strengths and weaknesses in terms of the data characteristics (dimension, size, distribution), task (detection, classification, measurement), and focus (clusters, trends, outliers) to which it is applied (Ward, 1994). They also vary in the amount of information loss and distortion; e.g., scatterplot matrices show mainly pairwise relationships and are affected by occlusion, while dimensionality reduction techniques do not directly convey the original numeric values.

We chose to concentrate on the use of glyph representations for our research, as glyphs can convey the original data accurately and without occlusion. Conveniently, glyph placement can be used to express cyclic structure, leaving other glyph attributes to represent other data dimensions. Clearly, glyphs require the largest amount of display space per datum of all the common visualization techniques, which limits the size of the data set that can be effectively displayed in its entirety. However, we have found that providing interactive controls for the subsets being displayed, in conjunction with real-time updates, helps alleviate this limitation. Glyphs are also capable of encoding moderate numbers of data dimensions, while techniques such as scatterplot matrices work best with data of low dimensionality. A significant limitation to the effective use of glyphs is the sensitivity of the technique to the order of the dimensions; correlations between dimensions that map to adjacent glyph attributes are often easier to detect than those mapped to non-adjacent attributes. Thus the ability to reorder the dimensions, either manually or automatically, is an essential component in visual exploration using glyphs.

3 Visualization Techniques for Cyclic Data

Traditionally, data sets containing cyclic or periodic behavior, either univariate or multivariate, have been visualized with little specific regard to their cyclic structure. Most rely on the user to visually scan across data plots to identify repeated patterns, with minimal support for reorganizing the displays to help the user explore variations between cycles. In this section, we outline several layout strategies for plots that have been employed for cyclic data analysis. We present some of the strengths and weaknesses of each, and argue that no single layout method is useful for all visual analysis tasks.

3.1 Linear Techniques

Linear data visualization methods rely heavily on the human vision system to detect cyclic patterns from plots. Such plots often have time on the horizontal axis and use the vertical axis to encode one or more variables. Nelson Max et. al. (Max, 1993), describe a four-dimensional visualization with this property. The viewer must scan the plot to decide if the data appears cyclic, and then scan the adjacent cycles to determine what change occurred on a cycle-by-cycle basis. While the human vision system is generally well suited to this task, it is limited in scope. If the data set is lengthy, this system breaks down, as one can only distinguish change over proximal cycles rather than all cycles. Multivariate data further impedes this process. Cave Plots (Becker, 1994) allow the simultaneous display of two univariate time series data sets by vertically flipping one of the data plots and using the top of the display as a second base line. This technique could be used for bivariate cyclic data, but would not extend gracefully to higher numbers of dimensions.

Using animation to convey cyclic or periodic data, such as described by Rhyne et al. (Rhyne, 1993), also has limitations. While a useful tool for visualizing overall trends and anomalies, for the purpose of studying cyclic structure this might be considered a regression from the static linear visualization. Our vision system can detect periodicity in the animation if it exists, but we have no way of comparing in any detail the data of one cycle with those of previous or subsequent cycles.

3.2 Circular Techniques

Circular histograms, as described by Fisher (Fisher, 1993), are useful for visualizing distributions of data that have some circular component. Like a conventional histogram, circular histograms map a data set's population density to height. Instead of the height of bars on a line, this technique varies the height of bars on a circle. This technique is used to visually categorize data whose range is circular, such as angle, time of day, or month of the year.

While circular histograms effectively summarize the intra-cycle relationships in the data, the cycle number is not considered, so inter-cycle relationships are not apparent. Also, this method does not scale to many dimensions.

3.3 Cyclic Techniques

Jacques Bertin provides three side-by-side visualizations of cyclic univariate data (see (Tufte, 1983), pg. 169). The first is a traditional linear plot. This method offers intra-cycle coherence, but lacks inter-cycle coherence for reasons described above. The second is a stack of cycles. Data falling at the same position within each cycle line up vertically. This method provides inter-cycle coherence, but lacks intra-cycle coherence due to the discontinuity where one cycle ends and another begins. The third of Bertin's techniques makes use of a spiral to provide both intercycle and intra-cycle coherence. By wrapping his linear plot around a spiral, he maintains the continuity of the data, while at the same time lining up successive cycles radially.

A spiral visualization similar to Bertin's, attributed to Antonio Gabaglio (1888), inspired the Glyph Spiral technique described below (see (Tufte, 1983), pg. 72). Low-dimensional multivariate data is displayed on a series of concentric circles (the annual rings of a tree), with time spiraling out from the center, the period equal to one year. Months of the year are aligned radially, so comparison can be made from month to month within the same year, or from year to year for the same month. While this visualization's glyphs accommodate only two data dimensions, it has transcended the two fundamental obstacles of cyclic visualization: the simultaneous preservation of both intra-cycle and inter-cycle relationships in the data.

4 SpiralGlyphics

An interactive tool for exploring cyclic multivariate data sets has been designed and implemented. SpiralGlyphics, a cross-platform Java application (an applet version is also available), combines a variety of glyph types with several glyph layout methods designed to accentuate the cyclic characteristics of the data. A simple plug-in interface allows for extensibility of both glyph types and layouts.

Common across all layouts is a set of controls for navigating the data set. These controls allow the viewer to pan and zoom by altering the focus and scope of currently displayed data. In addition to these generic controls, there is also a control for specifying the length of each cycle. By interactively varying the hypothetical cycle length, or period, the viewer may perform exploratory analysis of the data set's cyclic structure. The display is updated immediately after any control is changed.

To enhance the exploration of relationships between data dimensions, controls are provided for enabling and disabling dimensions as well as reordering the dimensions. Thus dimensions determined to be redundant or otherwise uninteresting can be turned off. Similarly, relationships between specific dimensions can be emphasized by interactively modifying the order of dimensions such that dimensions of interest are adjacent to each other. In most glyph designs, reordering can significantly affect the detection of relationships or anomalies.

Color can be used in two different ways in SpiralGlyphics. First, it can be used to distinguish between different data dimensions within each glyph. This helps especially when the glyphs are encoding many dimensions and the user wishes to identify individual dimensions for analysis. Alternatively, color may be used as a mechanism to detect inter-cycle trends and as a visual cue to distinguish between glyphs as they migrate within a layout due to interactive scrolling. In this case, a single color is assigned to the entire glyph based upon its position within a cycle. Glyphs from different cycles that are at the same position within the cycle will share the same color. In both cases, colors are automatically chosen based on equal differences in hue, using an HSV (hue, saturation, value) color space. This method has proven to be effective in distinguishing at least 25 dimensions or cycle positions, though an alternate strategy would be needed for larger numbers.

4.1 Layout Types

SpiralGlyphics uses the layout of glyphs to enhance the cyclic structure of multivariate data. The layout is independent of the type of glyph used. Three layouts are currently provided. The user may switch among these to take advantage of the strengths of each layout.

The linear layout is familiar, but provides the least enhancement of cyclic structure. Glyphs are placed directly on the axis of the variable exhibiting or being imposed with cyclic structure. Above the glyphs, a more useful line plot is drawn with a different color for each dimension. The boundaries between cycles are marked by guidelines in the plot. While maintaining intra-cycle coherence and in fact full continuity, inter-cycle coherence is absent. The viewer must study the data to find these relationships. To assist the comparison of points at the same location within cycles, the user can click on a point and have all others at that location in other cycles highlighted. However, this layout is most useful for finding trends and anomalies rather than cyclic behavior.

The stacked cycle layout takes the axis of the linear layout above, and breaks it into segments each equal to the cycle length. These sub-axes are stacked vertically such that common positions within each cycle are aligned. Inter-cycle coherence is preserved, as the viewer can scan vertically to find relationships between cycles. Unfortunately, the discontinuities of the axis distract the intra-cycle relationships when scanning horizontally from one sub-axis to another. This limitation has been solved interactively, as the user can scroll such that data originally on the boundary between two sub-axes may later lie in the middle of one sub-axis.

The spiral layout attempts to enhance cyclic structure by providing continuity within and between cycles (Lipchak, 1997). The glyphs are positioned on a spiral axis, extending outward from the center of the spiral. The spiral is segmented such that consecutive cycles are aligned radially. This method preserves both intra-cycle and inter-cycle coherence. By maintaining the continuity of the axis, we preserve intra-cycle relationships. By segmenting the spiral such that data occurring at the same location within two adjacent cycles line up radially, we preserve inter-cycle relationships. Using a spiral of Archimedes, we can ensure that the glyphs positioned near the outside of the spiral are not size-biased when compared to those near the center of the spiral. The viewer may alternate the spiral direction in which the axis increases between spiraling outward and spiraling inward. For regularly sampled data, the viewer may prefer the dependent variable to increase toward the center where the closer proximity facilitates glyph comparison.

In terms of relative effectiveness, the stacked layout has the best screen utilization, followed by the spiral and linear layouts. While the spiral layout preserves both intra- and inter-cycle relationships, the visual comparison of vertically or horizontally offset neighbors (as in the stacked and linear layouts) is easier than circular or radial offset glyphs. This is especially true when the cycle length and number of visible cycles is small. Thus we believe that each of the currently implemented layouts has strengths and weaknesses, and have found an effective strategy for data exploration is to regularly switch between the different options for alternate views.

Because of the plug-in design of the system, it is relatively straightforward to develop new layout schemes for the glyphs. For example, the line plots used in the linear layout could be used in conjunction with the stacked layout to show rows of line plots corresponding to cycles. Similarly, the plots of the individual dimensions could be separated to avoid congestion in the display. Other variations could include the simultaneous display of two or more data sets, using various interleaving strategies for the layout. For example, rows of glyphs in the stacked layout could alternate between two data sources, or two spirals could be placed in a nested fashion or adjacent to each other. The code is structured specifically for this sort of user customization and extensibility.

4.2 Glyph Types

Three glyph types are currently provided by SpiralGlyphics (see Fig. 1). While layout of the glyphs remains the same regardless of glyph type, each is effective at enhancing different properties of the data. The user may switch among these, employing the glyphs best suited to her purpose.

The bar glyph (also known as a profile glyph) is simply a mapping of each data dimension, discounting the dimension used to derive or impose cyclic structure, to the height of a bar. The bars are stacked side-by-side with a common baseline, much the same as a histogram or bar chart. While this glyph type is readily familiar, it becomes ineffective when displaying many dimensions. Because the bars vary in height along parallel axes, one cannot easily detect trends or anomalies within individual variables. The use of one color per dimension helps relieve this limitation.

The star glyph is a mapping of dimensions to the lengths of line segments resting on equally spaced rays originating at the center of the glyph (Siegel, 1972). The endpoints of adjacent dimensions are then connected, forming a star shape. The shape is thus influenced in a different direction by each dimension, eliminating the parallel axes inherent to bar glyphs. The human vision system is adept at recognizing and distinguishing between shapes.

The pie glyph is a mapping of dimensions to slices of pie in a conventional pie chart. The size of a dimensions slice is based on the magnitude of the dimension relative to each of the other dimensions. Of course this relationship is most interesting when the dimensions share a common unit.

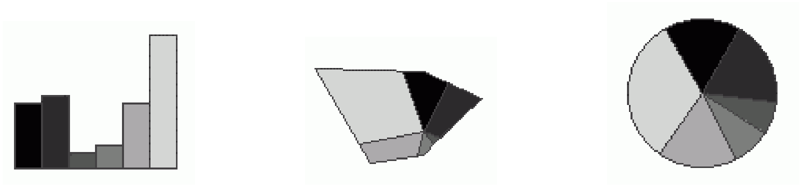

Figure 1: Glyph types in SpiralGlyphics: bar, star, and pie.

Each of the glyph types currently implemented portrays the multivariate data in a distinct manner, and thus are best suited for different tasks. The bar glyphs allow for the most accurate measurement of absolute value, as research has shown our ability to gauge length using a common baseline is very accurate (Cleveland, 1993). Star glyphs allow easier isolation of individual dimensions than bar glyphs, especially when the number of data dimensions is large. They also draw on our ability to perceive both length and shape, providing an alternate means of interpretation. The size of pie glyphs are automatically normalized, which means they are best for comparing data which represents percentages of a total rather than data fields with different units.

Other glyph types can be added to SpiralGlyphics through its plug-in interface. For example, by implementing a few interface routines describing how to draw facial attributes, a user could add Chernoff faces to the repertoire of available glyph types. Similarly, variations on existing glyphs could also be easily incorporated.

5 Case Studies

The first case study (Fig. 2) explores a four-dimensional data set consisting of Dow Jones average, retail sales, Standard and Poor's 500 Index and unemployment over a five year period (glyphs for January emanate straight up from the center). As would be expected, sales is highest in December, and unemployment peaks in June and July. A few anomalies are clearly visible, however. Unemployment during September and October of year 2 is unusually low for that time of year. Likewise, the first part of year 4 is flatter than other years (low Dow Jones and S+P levels), and is followed by a larger than normal level of unemployment. Clearly, these and other patterns in the image are worth examining in more detail.

In the second example (Fig. 3), we look at several years of data showing the relative levels of industrial (dark grey) and utilities (light grey) indices from the archives of the New York Stock Exchange (monthly averages). Early years show rapid shifts surrounded by periods of relative constancy, while changes in more recent years are much slower.

In the final example (Fig. 4), we examine currency exchange rates over a 10 year period, comparing the U.S. dollar against the currencies of (from left to right) Italy, U.K., Canada, Sweden, France, Belgium, Netherlands, Switzerland, Germany, and Japan. We reordered the dimensions with SpiralGlyphics so that a monotonically increasing level was achieved across the ten countries for the first month of information. For the first couple years, levels and relative ordering remained fairly constant. However, later years showed a number of significant changes in both the relative and absolute magnitudes, including inverse changes for Canadian and Japanese currencies. Other examples of common and inverted variations can be perceived on further inspection.

6 Summary

In this paper we have presented a visualization tool, SpiralGlyphics, for the interactive exploration of multivariate data sets, with a focus on emphasizing the cyclic or periodic characteristics of one or more variables. The program supports several styles of glyphs for encoding a multivariate data element, as well as several glyph layout strategies to help users detect both inter- and intra-cycle patterns and anomalies. An intuitive, yet powerful, user interface provides users with instantaneous control of cycle properties and data configuration, along with several customization options to increase the effectiveness of the exploration process. The program is implemented in Java and is freely available at the WWW site http://davis.wpi.edu/ matt/projects/SpiralGlyphics. Also at that site are more details on the functionality and design of the system, including an applet version and numerous screen dumps highlighting its features.

The current limitations of the system are predominantly those resulting from the use of glyphs for data representation, i.e., only a small number of data points (generally less than 200) may be effectively visualized on a single image. Similarly there is a bias imposed by the ordering of the dimensions which makes detection of patterns or anomalies between some dimensions harder than others. Finally, no comprehensive usability study has been performed on the system to assess the usefulness and power compared to traditional cyclic data visualization and analysis methods.

Our major goal for the future is to integrate the package with analytic techniques and expand the visualizations to permit simultaneous display of both the original data and data points resulting from the development of an analytic model. With tight integration, this combination of visualization and modeling could provide data analysts a powerful tool for multivariate data analysis. Other activities planned include performing usability studies, experimenting with alternate glyphs and layout strategies (including 3D methods), and attempting to address the data set size limitation by integrating clustering for large sets.

References

Becker R, Cleveland W, Shyu M (1996). The design and control of trellis displays. Journal of Computational and Graphical Statistics 5:123-155

Becker R, Clark L, Lambert D (1994). Cave Plots: a graphical technique for comparing time series. Journal of Computational and Graphical Statistics 3:277-283

Bloomfield P, Hurd H, Lund R (1994). Periodic correlation in stratospheric ozone data. Journal of Time Series Analysis 15:127-150

Chernoff H (1973). The use of faces to represent points in k-dimensional space graphically. Journal of the American Statistical Association 76:361-368

Cleveland W (1993). Visualizing data. Hobart Press

Duda R, Hart P (1973). Pattern classification and scene analysis. John Wiley and Sons

Feinberg S (1979). Graphical methods in statistics. American Statistician 33:165-178

Fisher N (1993). Statistical analysis of circular data. Cambridge University Press

Fisherkeller M, Friedman J, Tukey J (1975). PRIM-9, an interactive multidimensional data display and analysis system. In Dynamic Graphics for Statistics (Cleveland W and McGill M, eds.), Wadsworth and Brooks/Cole Publishers, Pacific Grove, CA 91-109

Huber T, Nagel M (1995). Look first, then model: exploring the data using the interactive graphics of PC-ISP. Computational Statistics 10:29-36

Hwarng H (1995). Proper and effective training of a pattern recognizer for cyclic data. IIE Transactions 27:746-756

Inselberg A, Dimsdale B (1990). Parallel coordinates: a tool for visualizing multi-dimensional geometry. Proceedings of IEEE Conference on Visualization 361-378

Jolliffe I (1986). Principle component analysis, Springer Series in Statistics. Springer, Berlin

Kohonen T (1995). Self-organizing maps, Springer Series in Information Sciences, Vol. 30. Springer, Berlin

Kruskal J, Wish M (1978). Multidimensional scaling. Sage Publications

LeBlanc J, Ward M, Wittels N (1990). Exploring n-dimensional databases. Proceedings of IEEE Conference on Visualization 230-237

Lipchak B, Ward M (1997). Visualization of cyclic multivariate data. Proceedings of IEEE Conference on Visualization, Late Breaking Hot Topics 61-64

Max N, Crawfis R, Williams D (1993). Visualization for climate modeling. IEEE Computer Graphics & Applications 13:4:34-40

Rhyne T, Bolstad M, Rheingans P, Petterson L, Shackelford, W (1993). Visualizing environmental data at the EPA. IEEE Computer Graphics & Applications 13:2:34-38

Siegel J, Farrell E, Goldwyn R, Friedman H (1972). The surgical implication of physiologic patterns in myocardial infarction shock. Surgery 72:126-141

Tufte E (1983). The visual display of quantitative information. Graphics Press

Wang H-X, de Paola R, Norwood W (1993). Analysis of intermittent periodic modes within complex data. Physical Review Letters 71:3039-3042

Ward M (1994). XmdvTool: integrating multiple methods for visualizing multivariate data. Proceedings of IEEE Conference on Visualization 326-333

Wegman E (1987). A parallel coordinate approach to statistical graphics. Proceedings of National Computer Graphics Association Conference 574-580

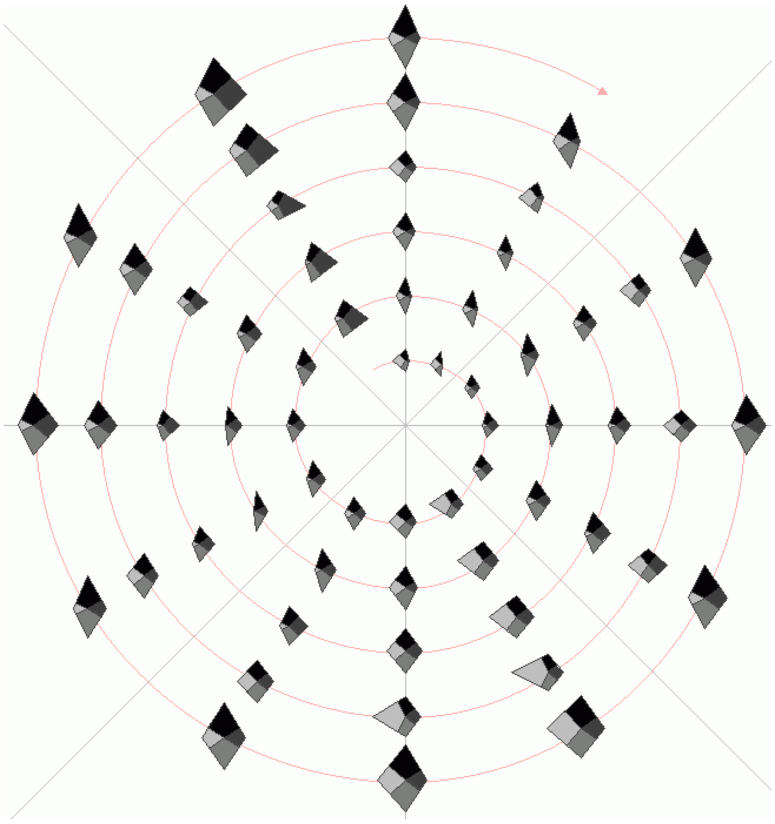

Figure 2: Financial, sales, and unemployment data over five years. Top is Dow Jones average, bottom is Standard and Poor's 500 index, right is retail sales, and left is unemployment.

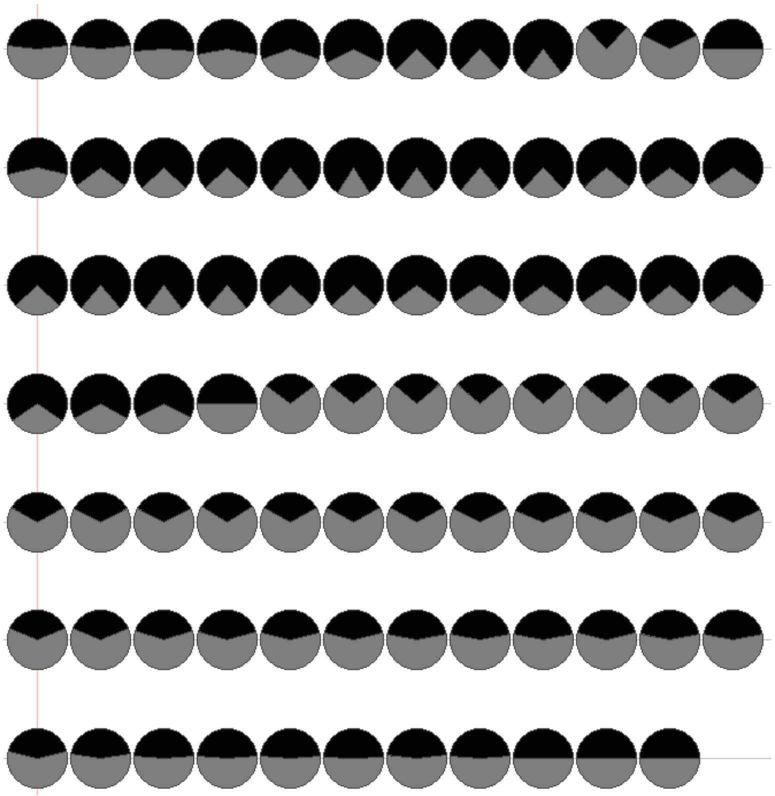

Figure 3: Relations between industrial (top segment of pie) and utilities indices over a seven year period.

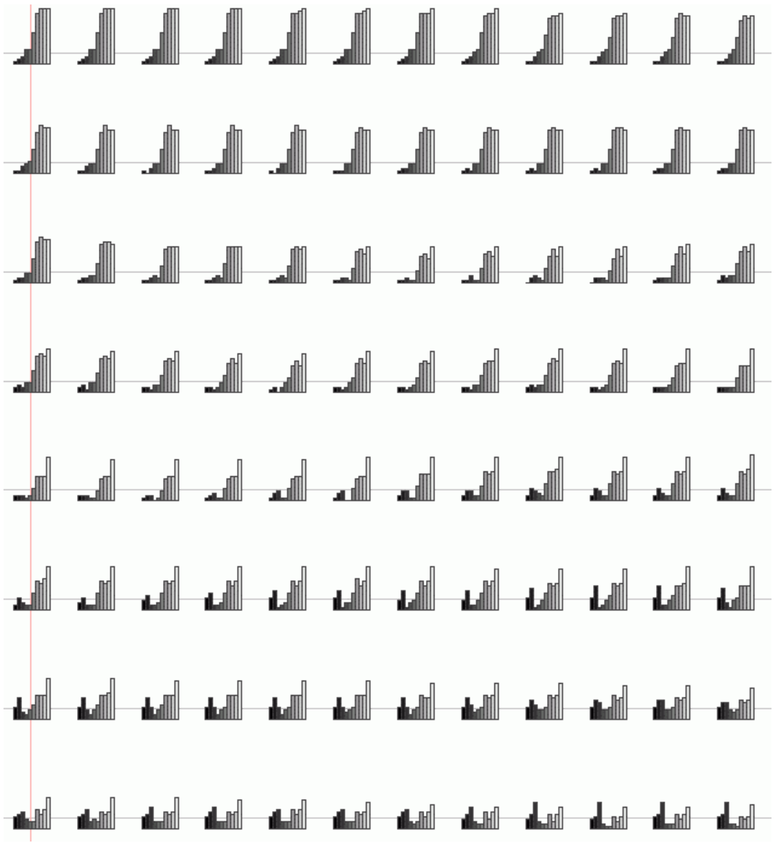

Figure 4: Exchange rates of ten currencies against the U.S. dollar over eight years, sorted by ascending value at the start of the data set . Note that after more than two years of relative stability, several dramatic changes occur.