Barcodes, Generalized Pairs Plots, and Sparkmats

Jay Emerson, Walton Green, and John Hartigan, Yale University

June 17, 2006, UseR! Vienna, Austria

The barcode plot aids in comparing distributions. It shares some of the characteristics of side-by-side histograms or boxplots, and of rugs or stripplots. We have found it particularly useful with clumped data, when other methods obscure detail.

As an example, we display the 2003 Tour de France stage results. Flat stages (Nevers and Lyon) are easily identified by the tallest spikes (finishers in the peleton arriving together). In contrast, mountain stages (Gap and L'Alpe d'Huez) stretch the distribution of finishing times. In some mountain stages (Gap and Morzine), trailing clumps of slower (exhausted) riders finish together (in a group called the autobus), avoiding elimination.



Tour de France 2003

Traditional pairs plots (or scatterplot matrices) show relationships between quantitative variables, and package vcd provides similar plots for categorical data using mosaic tiles. Many data frames, however, include both quantitative and categorical variables, and the generalized pairs plot recognizes the fundamental importance of the relationships between different types of variables.

Characteristics of a sample of properties in New Haven, CT, USA, illustrate the generalized pairs plot. Not all the properties are residential; outliers are evident, dominating the distribution of the data (for example, one property has more than 80 acres of land), and transformations are clearly needed. Some properties seem to have negative values of depreciation!



Here, we limit our attention to residential properties (removing many of the outliers). The current property value and square footage are logged, while the square root of property acreage is used. This latter choice retains condominiums without land in the plots (which would have been log(0) = -Inf using the log transformation). RS1-zoned properties are larger single-family homes in more affluent areas; not surprisingly, these homes have more bedrooms. The barcode plots show clumping of depreciation (most noticeable in the barcode of depreciation versus air-conditioning). According to the City of New Haven, about half of all residential properties have exactly 18% or 28% depreciation.



Finally, we present a new graphical tool called a sparkmat for spatially distributed time series. It generalizes Edward Tufte's sparklines for continuous variables distributed in both space and time. In the example below, we can identify certain peculiarities of three 6-year-long monthly time series of climate data on a 24 by 24 raster covering an area of South and Central America. Note the difference in the amplitude of temperature cycles (black) at different latitudes (less variability and warmer temperatures near the equator). The data, provided by NASA for a JSM poster competition, also show unexplained change points and truncation in pressure measurements (red).



Black = Temperature, Red = Pressure, Green = Low clouds

Contact: Jay Emerson (john.emerson@yale.edu) barcode() and gpairs() source, documentation, and examples:

http://www.stat.yale.edu/~jay/R/ViennaUseR/