Product graphics

Heike Hofmann and Hadley Wickham

May 13, 2009

1 Introduction

The Grammar of Graphics [reference] is a grammar describing the components of statistical graphics and how they can be combined together. These components are very flexible and can be used to describe any type of plot. However, the book (Grammar with a capital G) does not present many examples of graphics for categorical data, and in those examples the use of the grammar gives little insight into plot. This paper describes common categorical plots, and presents a specialisation of the Grammar of Graphics specifically designed to describe these types of plots and give insight into the types of relationships that they can show.

Area plots [reference] display categorical data with a rectangle for each combination of the factors of interest, with the area of the rectangle proportional to the number of observations in that combination. The layout of the rectangles give rise to the difference categorical plots that we are familiar with: bar charts, spine charts, mosaic plots, equal-bin-size plots, and fluctuation diagrams. Trellis graphics are another type of display that uses categorical variables to create small multiples for different subsets of the data. Visually, trellis graphics look like equal-bin-size plots with another plot drawn inside each bin. Figure 1 illustrates some of these plots.

Figure 1: Area plot examples.

Area plots are the graphical equivalent to contingency tables. And like contingency tables, they can be used to display raw counts, or to display proportions or percentages. Similarly, it may be useful to augment area plots with marginal displays as we augment contingency tables with marginal sums.

Area plots different to other types of graphics, because must be able to flexible change level of aggregation. Values are typically counts, but can also be sums of observation level weights.

These techniques can also be used with continuous data, if the data is binned to create a categorical variable. This gives rise to the histogram and spinogram, analogues of the bar chart and spine chart respectively. A long standing tradition is that no gaps are displayed between adjacent rectangles when used for originally continuous data. Two examples are shown in Figure 2.

These techniques are also closely related to TreeMaps and dimensional stacking (LeBlanc et al., 1990)

Figure 2: Area plots of originally continuous data

What are the common components of all these plots, and how can be describe them in a succinct way that also allows us to describe what tasks each display is good for? We propose that there is one key feature that makes these plots special. That is there their coordinate system, which is hierarchical.

In most coordinate systems that we are used to, such as Cartesian or polar coordinates, are symmetric in the sense that you can locate a point by starting from either dimension. For example, in Cartesian coordinates you can identify either the x- or y-coordinate first, and the other next. This also implies a distance symmetry when projected onto 2D Euclidean geometry. When you vary one coordinate by a small amount (holding all others constant) points in the projective geometry will be close.

Neither of these properties hold for hierarchical coordinate systems.

We will discuss hierachical coordinate systems in general, and then present a specific hierachical coordinate system which gives rise to the area plots described previously. We also present an R implementation, in the R package `ggplot`.

Related work:

(Vliegen et al., 2006) (Ahlberg, 1996)

(Hartigan and Kleiner, 1984, 1981) (Friendly, 1999, 1994) (Theus and Lauer, 1999) (Hofmann, 2000, 2001, 2003)

Goals:

- Put all area plots (bar charts, spine plots, mosaic plots, fluctuation diagrams) on a common footing
- Continue to build link between graphics and log-linear models
- Provide tools for "residuals and re-iteration"

Basic idea is factorising table of probabilities:

- as products of marginal and conditional distributions
- as areas in plots
- as sums of parameters in log-linear models

2 Data

To illustrate the techniques in this paper, we'll use a small sample from the general social survey (GSS) focussing on happiness. 51 020 observations from a yearly survey from 1972 to 2006. Happy, age, sex, marital status, terminal education level, relative financial status and health. Also contains a weighting variable wtssall, which is a probability weight from the survey.

2.1 Factorising probabilities

Let $f(x, y, z)$ be the 2d pmf function generated. Can write this pmf as product of marginals and conditionals

- $f(x, y, z) = f(x, y|z)f(z)$
- $f(x, y, z) = f(x|y, z)f(y, z)$
- $f(x, y, z) = f(x|y, z)f(y|z)f(z)$

Compare to area of mosaic plot.

One other special 2d case is the equal bin size plot - which all 1d x 1d primitive plots collapse down to when counts are equal. The important difference is the formatting - a grid rather than nested boxes.

2.2 Weighting

We have assumed the probabilities represent counts, but WLOG we can use any set of non-negative additive weights instead. Some of the applications of such weighted plots are described in Unwin et al. (In progress).

2.3 Continuous data

Binning.

Convention when using binned continuous data is to remove the gap between regions.

2.4 Nested data

3 Display

The three plots above are special cases of $a(i, j) = w(i, j)^x * h(i, j)^y$. $x + y = 1$.

Perceptual constraints. Best at comparing area when:

- Simple shape (i.e. square vs polygon)
- Aspect ratios similar
- Areas are large
- Areas are close
- One border is constant.

Impossible to simultaneously satisfy all constraints. Every product plot requires the creator to trade off between these.

Restrictions on partitioning

- Area should be proportional to weight

- Containment
- Rectangular
- Non-overlapping

Pseudo 2d, where you take a long 1d structure and wrap it into 2d.
Pseduo 1d, where you make a new variable from multiple variables.

3.1 1d primitives

Space-filling vs. not space-filling. Read length on common scale, vs. put more information in the display.

Basic principle of all area plots is area should be proportional to probability. For the plots we are interested in, the shape used to represent the count is a rectangle, so that $a(i,j) = w(i,j) * h(i,j)$. Each plot type constrains this relationship in a different way:

- **bar**: height is proportional to value, width equally divides space. Bars allow comparison between absolute numbers.
- **spine**: width is proportional to value, height equally divides space. Spines not useful at top level, but allow comparison of proportions at next level (ie. conditioned on top level values)
- **treemap**: no restrictions on height and width.

3.2 2d primitives

- **fluct**: width and height proportional to square root of value. Flucts allow comparisons of (approximately) absolute numbers in two directions.

3.3 Plot templates

- **Stacked** barchart: 1 bar + $n-1$ spines in opposite direction.
- **Nested** barchart: n bars.
- **Mosaic** plot: spines in alternating directions. The 1d case has a special name, the spineplot.
- **Double-decker** plot: $n-1$ spines + 1 spine in opposite direction.
- **Fluctuation** diagram:

3.4 Violation of constraints

3.4.1 Small values and zeros

3.4.2 Cascading

The cascaded treemaps of Lü and Fogarty (2008) is an idea that illustrates how the violation of containment can be productive. In the cascaded treemap, each level is slightly offset from the one above to create a pseudo-3d perspective.

3.4.3 Non-rectangular partitions

Pie charts fall out naturally, as bars in polar coordinates, with angle and radius instead of height and width. And to ensure that counts stay proportional to areas, the square-root is taken of the y-axis. This is related to the infoslices of Andrews and Heidegger (1998), which only use half of the disk. Figure 3 shows some examples of product plots in polar coordinates.

- Pie charts
- Fourfold displays (**?**)
- Bullseye chart

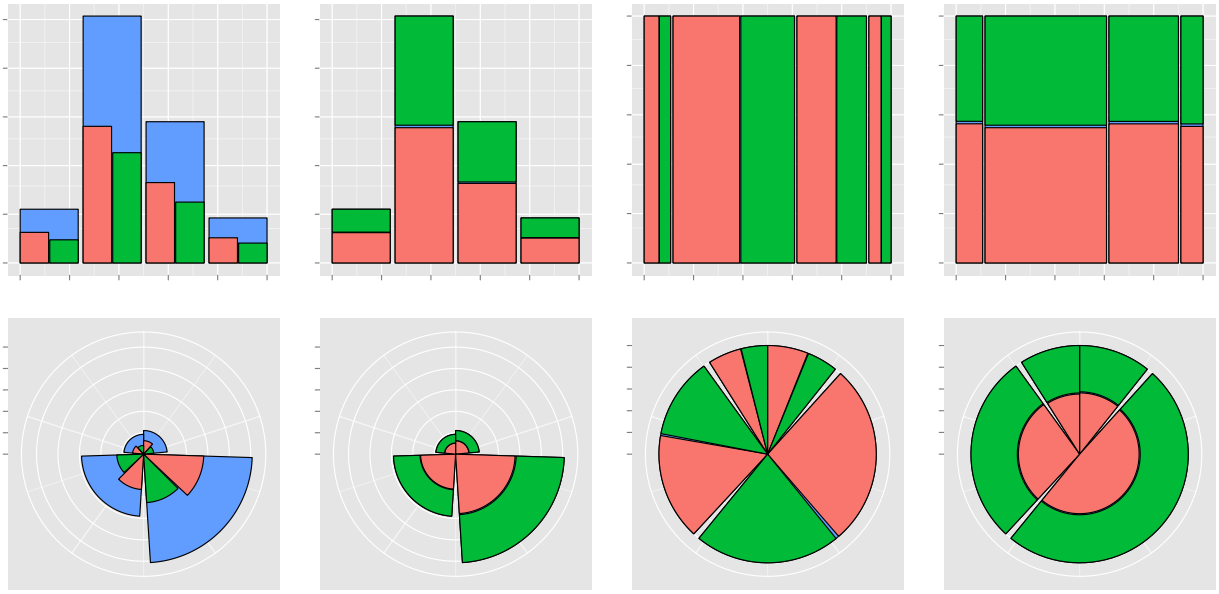

Figure 3:

Circular treemaps of Wetzel (2008) which use circles instead of rectangles, but is not space filling because you can not tile a circular region with circles. The radial displays of Stasko and Zhang (2000) deliberately keep the radius constant in order to relatively highlight the lower levels. Similarly so does the fanlens of Lou et al. (2007)

Other non-rectangular treemaps have been proposed, such as the space-filling curve approach of (Wattenberg, 2005), or the voronoi treemaps of (Wattenberg, 2005), but because of the perceptual difficultly of comparing areas of arbitrary polygons, these approaches tend to be attractive rather than useful.

4 Connection with log-linear models

which imply the obvious mapping between d(i, j) and a(i, j). For ¿2d, the definitions generalise in the obvious, but the constraints become more complicated.

$log(d(i,j)) = 1 + d_{i\cdot} + d_{j|i}$ $log(d(i,j)) = 1 + d_{\cdot j} + d_{i|j}$

5 Labelling

Labelling these plots is particularly challenging.

- Additional info
  - colour (map of the market)
  - texture (sieve plots)
  - embedded plots (time series in lab escape)
  - photographs
  - text (tables)
- Display of hierarchy
  - Spacing / borders
  - Shading
  - Cascading
  - Labelling

References

C. Ahlberg. Cocktailmaps: A space-filling visualization method for complex communicating systems. In *AVI '96*, 1996.

K. Andrews and H. Heidegger. Information slices: Visualising and exploring large hierarchies using cascading, semi-circular discs. In *IEEE Symposium on Information Visualization (InfoVis'98)*, 1998.

M. Friendly. Mosaic displays for multi-way contingency tables. *Journal of the American Statistical Association*, 89(425):190–200, 1994.

M. Friendly. Extending mosaic displays: Marginal, conditional, and partial views of categorical data. *Journal of Computational and Graphical Statistics*, 8(3):373–395, 1999.

J. A. Hartigan and B. Kleiner. Mosaics for contingency tables. In *Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface*, pages 268–273, Fairfax Station, VA, 1981. Interface Foundation of North America, Inc.

J. A. Hartigan and B. Kleiner. A mosaic of television ratings. *The American Statistician*, 38(1): 32–35, 1984.

H. Hofmann. Exploring categorical data: Interactive mosaic plots. *Metrika*, 51(1):11–26, 2000.

H. Hofmann. Generalized odds ratios for visual modeling. *Journal of Computational and Graphical Statistics*, 10(4):628–640, 2001.

H. Hofmann. Constructing and reading mosaicplots. *Computational Statistics and Data Analysis*, 43 (4):565–580, 2003.

J. LeBlanc, M. Ward, and N. Wittels. Exploring n-dimensional databases. In *Proceedings of Visualization '90*, pages 230–237, 1990.

X. Lou, S. Liu, and T. Wang. Fanlens: Dynamic hierarchical exploration of tabular data. In *Infovis 2007 (poster)*, 2007.

H. Lü and J. Fogarty. Cascaded treemaps: examining the visibility and stability of structure in treemaps. In *Proceedings of Graphics Interface 2008*, pages 259–266. Canadian Information Processing Society, 2008.

J. Stasko and E. Zhang. Focus+context display and navigation techniques for enhancing radial, space-filling hierarchy visualizations. In *Information Visualization 2000*, 2000.

M. Theus and S. R. W. Lauer. Visualizing loglinear models. *Journal of Computational and Graphical Statistics*, 8(0):396–412, 1999.

A. Unwin, H. Hofmann, and H. Wickham. Weight and see. In progress. Graphical techniques for weighted data.

R. Vliegen, J. J. van Wikj, and E.-J. van der Linden. Visualizing business data with generalized treemaps. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):789–796, 2006.

M. Wattenberg. A note on space-filling visualizations and space-filling curves. In *Proceedings of Infovis 05*, 2005.

K. Wetzel, 2008. URL `http://lip.sourceforge.net/ctreemap.html`.