

STAT 6560
Graphical Methods
Spring Semester 2011

Dr. Jürgen Symanzik

Utah State University

Department of Mathematics and Statistics

3900 Old Main Hill

Logan, UT 84322-3900

Tel.: (435) 797-0696

FAX: (435) 797-1822

e-mail: symanzik@math.usu.edu

Web: <http://www.math.usu.edu/~symanzik/>

Contents

- Acknowledgements** **1**

- 1 Introduction: A Couple of Good and Bad Examples** **1**
 - 1.1 Motivation 1
 - 1.2 Why Graphics !?!? 2
 - 1.3 How to Display Data Badly 3
 - 1.3.1 Don't show much data 4
 - 1.3.2 Show the data inaccurately 9
 - 1.3.3 Obfuscate the data 16
 - 1.4 Bad Graphics are Everywhere — In Space and in Time 32
 - 1.5 Rules for Good Data Displays 43
 - 1.6 Further Reading 47

- 2 History of Statistical Graphics: Plots, People, and Events** **49**
 - 2.1 General History 49
 - 2.1.1 Milestones in the History of Data Visualization (According to Friendly) 50
 - 2.2 Selected People 51
 - 2.3 Statistical Graphics and Events in History 64
 - 2.3.1 John Snow and the Cholera Epidemic in London, 1854 64
 - 2.3.2 The Challenger Disaster, 1986 68
 - 2.4 Further Reading 72

- 3 Color and Cognition** **73**
 - 3.1 Color-Deficiency and Color-Blindness 73
 - 3.2 Various Color Spaces 78
 - 3.2.1 The HSL and HSV Color Spaces 78
 - 3.2.2 The RGB Color Space 80
 - 3.2.3 The HCL Color Space 82

3.3	Suggestions for Color Selections	83
3.4	Good Color Choices	88
3.4.1	Work by Cindy Brewer and Collaborators	90
3.4.2	Work by Zeileis, Hornik, and Murrell	93
3.5	Change Blindness	94
3.6	Further Reading	97
4	Statistical Maps	98
4.1	Choropleth Maps	98
4.1.1	Choropleth Maps in R	102
4.2	Linked Micromaps	107
4.2.1	Template for LM Plots	108
4.2.2	Micromaps vs. Choropleth Maps	110
4.2.3	Additional Linked Micromap Examples	115
4.2.4	Web-based Applications of LM Plots	121
4.2.5	Linked Micromaps in Java	124
4.2.6	Linked Micromaps in R	125
4.3	Conditioned Micromaps	129
4.4	Comparative Micromaps	132
4.5	Further Reading	135
5	Categorical Plots	136
5.1	Which Plot Type to Choose?	136
5.2	Categorical Plots in R	140
5.2.1	Pie Charts	142
5.2.2	Bar Charts	144
5.2.3	Dot Charts	145
5.2.4	Mosaic Plots	145
5.2.5	Spine Plots and Spinograms	146

5.2.6	Four Fold Plots	146
5.3	Categorical Plots in Mondrian	148
5.3.1	Installation	148
5.3.2	The Titanic Data in Mondrian	149
5.4	Further Reading	151
5.5	R Code and Output	152
5.5.1	Example 1: UCBA admissions	153
5.5.2	Example 2: Titanic	165
5.5.3	Example 3: HairEyeColor	168
6	Univariate Plots	176
6.1	Histograms	176
6.2	Averaged Shifted Histograms	184
6.3	Stem-and-Leaf Plots	187
6.4	Boxplots (or Box-and-Whisker Plots)	188
6.5	Dot Charts for Univariate Data	189
6.6	Kernel Density Plots for Univariate Data (with Rug Plot)	191
6.7	Quantile-Quantile Plots (Q-Q Plots)	193
6.8	Empirical Cumulative Distribution Functions (ECDFs)	196
6.9	Graphics and Small Sample Sizes	198
6.10	Further Reading	202
7	Bivariate Plots	203
7.1	Scatterplots	203
7.2	Hexagon Binning	206
7.3	Bivariate Histograms	207
8	Trivariate Plots	209
8.1	Scatterplot Matrix	209
8.2	3D Scatterplots	210
8.3	Co-Plots	211
8.4	Trivariate Density Estimation	213

9	“Hypervariate” (High–Dimensional) Plots	221
9.1	Scatterplot Matrix (for $n \geq 4$)	221
9.2	Parallel Coordinate Plots	222
9.3	Faces, Star Plots, and other Glyph Representations	224
9.4	Andrews Plots	226
9.5	Data Images	229
9.6	From Tables to Plots	231
10	Interactive and Dynamic Graphics	237
10.1	Software for Interactive and Dynamic Graphics	238
10.1.1	REGARD, MANET, and Mondrian	238
10.1.2	HyperVision, ExplorN, and CrystalVision	241
10.1.3	Data Viewer, XGobi, and GGobi	243
10.2	Concepts of Interactive and Dynamic Graphics	248
10.2.1	Scatterplots and Scatterplot Matrices	248
10.2.2	Brushing and Linked Brushing/Linked Views	249
10.2.3	Focusing, Zooming, Panning, Slicing, Rescaling, and Reformatting	250
10.2.4	Rotations and Projections	252
10.2.5	Grand Tour	252
10.3	Spatial Data Analysis in the Linked ArcView 2.1 and XGobi Environment	254
10.3.1	Basic XGobi Layout	257
10.3.2	Graphical Methods	258
10.3.3	Linking Between More Complicated Objects for Data Analysis	265
10.4	Interactive and Dynamic Graphics in R	269
10.4.1	R Package <i>rggobi</i>	269
10.4.2	R Package <i>tourr</i>	269
10.4.3	R Package <i>iplots</i>	270
10.4.4	R Package <i>animation</i>	271
11	Graphics Galleries and Sources on the Web	273

Appendix	274
Homework Assignments	275
Homework Assignment 1	1
Homework Assignment 2	1
Homework Assignment 3	1
References	17

1 Introduction: A Couple of Good and Bad Examples

(Based on Wainer (1997), Chapter 1 & Tufte (1983), Chapter 2)

1.1 Motivation

Statistical graphics and data visualization are critical elements of modern data analysis and presentation. From initial exploration of a data set to the final presentation of results to the end user, statistical graphics play a vital role in shaping our understanding of our data. Through proper use of graphics, we can make critical discoveries, and communicate them clearly. Conversely, poor use or misuse of graphics can seriously mislead (by accident or design).

In this course, we will start with presentation graphics, including discussion of both tools and principles which lead to clear communication and those which serve only to confuse or mislead. We will spend most of the semester in exploratory graphics and data analysis, including data mining. This will be broken down largely by the dimension of the applicable data. One- and two-dimensional datasets require and allow far different methods than those of more than three dimensions. Categorical and regression data call for their own specialized methods.

Even more than most aspects of statistics, graphics and visualization involve art as well as science. In most cases, there are many reasonable approaches. Only an understanding of the options available and the underlying principles will lead to a successful analysis and presentation.

Via graphics, otherwise boring statistical information can become really exciting and entertaining. Hans Rosling (http://en.wikipedia.org/wiki/Hans_Rosling), a Swedish medical doctor and statistician, “sells” statistics extremely well. Enjoy his talk *Debunking Third-World Myths with the Best Stats You’ve Ever Seen*, accessible at http://www.ted.com/index.php/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen.html. The <http://TED.com> Web page is featured in Time, April 27, 2009, p. 44. In another presentation, Rosling shows an animation of 200 years of health and wealth data of 200 countries in just four minutes: <http://www.flixxy.com/200-countries-200-years-4-minutes>.

htm. The underlying software is accessible at <http://www.gapminder.org/> and it is further described in Rosling & Johansson (2009). Google has acquired the software, now called Google Public Data Explorer (<http://www.google.com/publicdata/home>), and they are planning to further extend it. We will revisit this software once we have discussed interactive and dynamic graphics.

1.2 Why Graphics !?!

Why do we need graphics at all. Aren't summary statistics sufficient?

Start R and load the Anscombe (1973) data set. Just type `anscombe` to check whether these data are available — if not, you may have to load the data via:

```
require(stats)
data(anscombe)
```

Then calculate some summary statistics (separately for the four columns of X's and Y's): mean of the X's, mean of the Y's, standard deviation of the X's, standard deviation of the Y's, correlation coefficient, slope and intercept of the regression line, rms error.

```
http://www.math.usu.edu/~symanzik/teaching/2011\_stat6560/RDataAndScripts/Anscombe.R
```

So, the four pairs of X/Y columns basically are identical !?!

But, didn't we forget to **plot** the data !!!

```
http://www.math.usu.edu/~symanzik/teaching/2011\_stat6560/RDataAndScripts/Anscombe2.R
```

See here for additional references:

```
http://en.wikipedia.org/wiki/Anscombe's\_quartet
```

```
http://pbil.univ-lyon1.fr/library/base/html/anscombe.html
```

Tufte (1983), p. 13, concludes:

“Graphics *reveal* data. Indeed graphics can be more precise and revealing than conventional statistical computations. Consider Anscombe's quartet: all four of these data sets are described by exactly the same linear model (at least until the residuals are examined).”

The Anscombe data show up in numerous textbooks, as early as in Tufte (1974), pp. 131–134, and as recent as in Moore et al. (2012), p. 120 (Exercise 2.73). In fact, this data set should be shown in every undergraduate class as well as in every regression class to demonstrate what might happen when blindly performing any statistical calculations without plotting the data first.

1.3 How to Display Data Badly

Wainer (1997), p. 12, states:

“The aim of good data graphics is to display data accurately and clearly.

[...]

Thus, if we wish to display data badly, we have three avenues to follow.

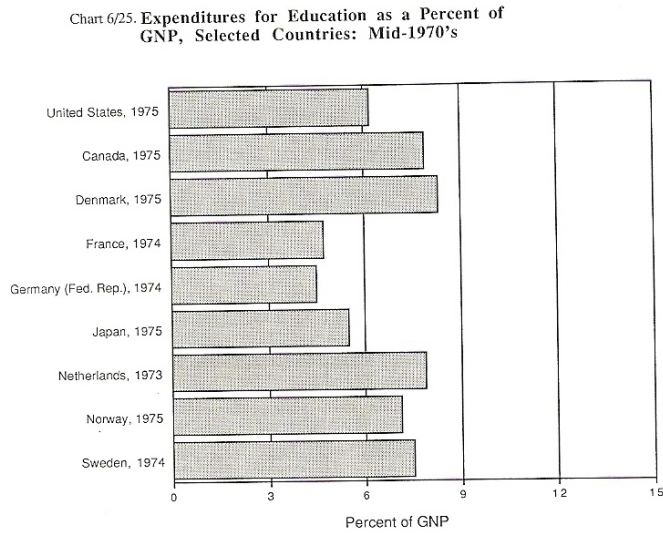
- A. Don’t show much data.
- B. Show the data inaccurately.
- C. Obfuscate the data.”[†]

Let us follow these strategies:

[†]Show the data unclearly.

1.3.1 Don't show much data

Rule 1: Show as little data as possible (minimize the data density).



Data Density = 9 numbers /63 sq. ins. = .14

FIGURE 2. Chart 6/25 from *Social Indicators III* showing expenditures for education for nine countries as a function of GNP.

Figure 2: Wainer (1997), p. 13, Figure 2.

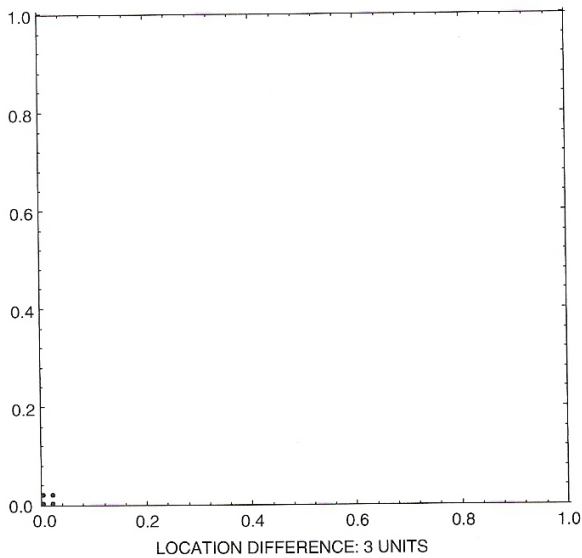


FIGURE 3. A graph of obviously low data density.

Figure 3: Wainer (1997), p. 13, Figure 3.

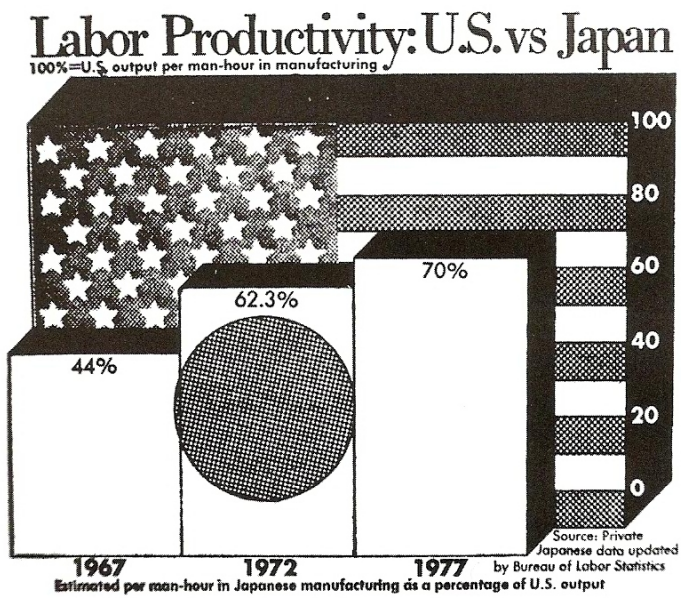


FIGURE 6. A graph with low data density filled in with chartjunk from the *Washington Post*, 1978.

Figure 4: Wainer (1997), p. 16, Figure 6.

Rule 2: Hide what data you do show (minimize the data/ink ratio).

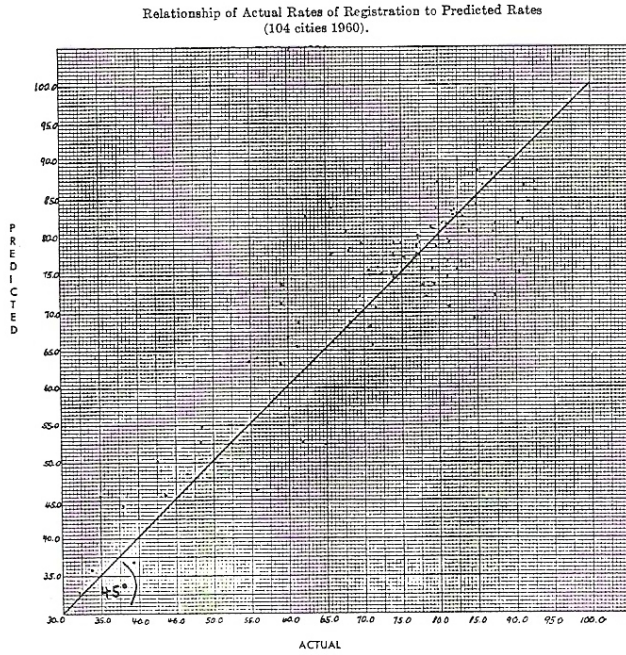


FIGURE 8. Hiding the data in the grid.

Figure 5: Wainer (1997), p. 17, Figure 8: Hiding the data in the grid.

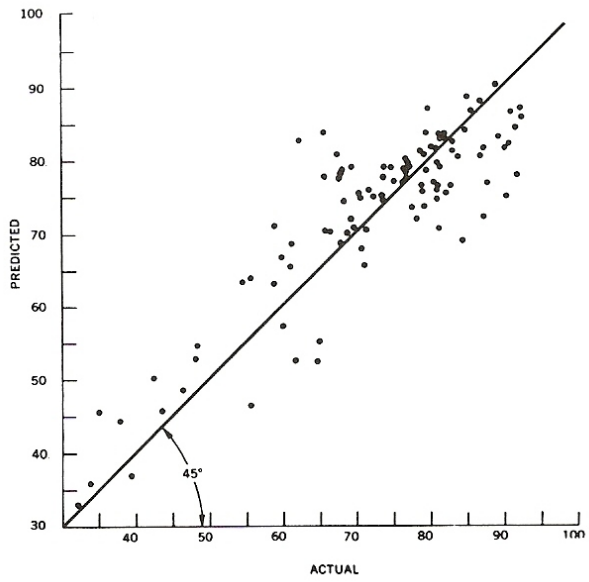
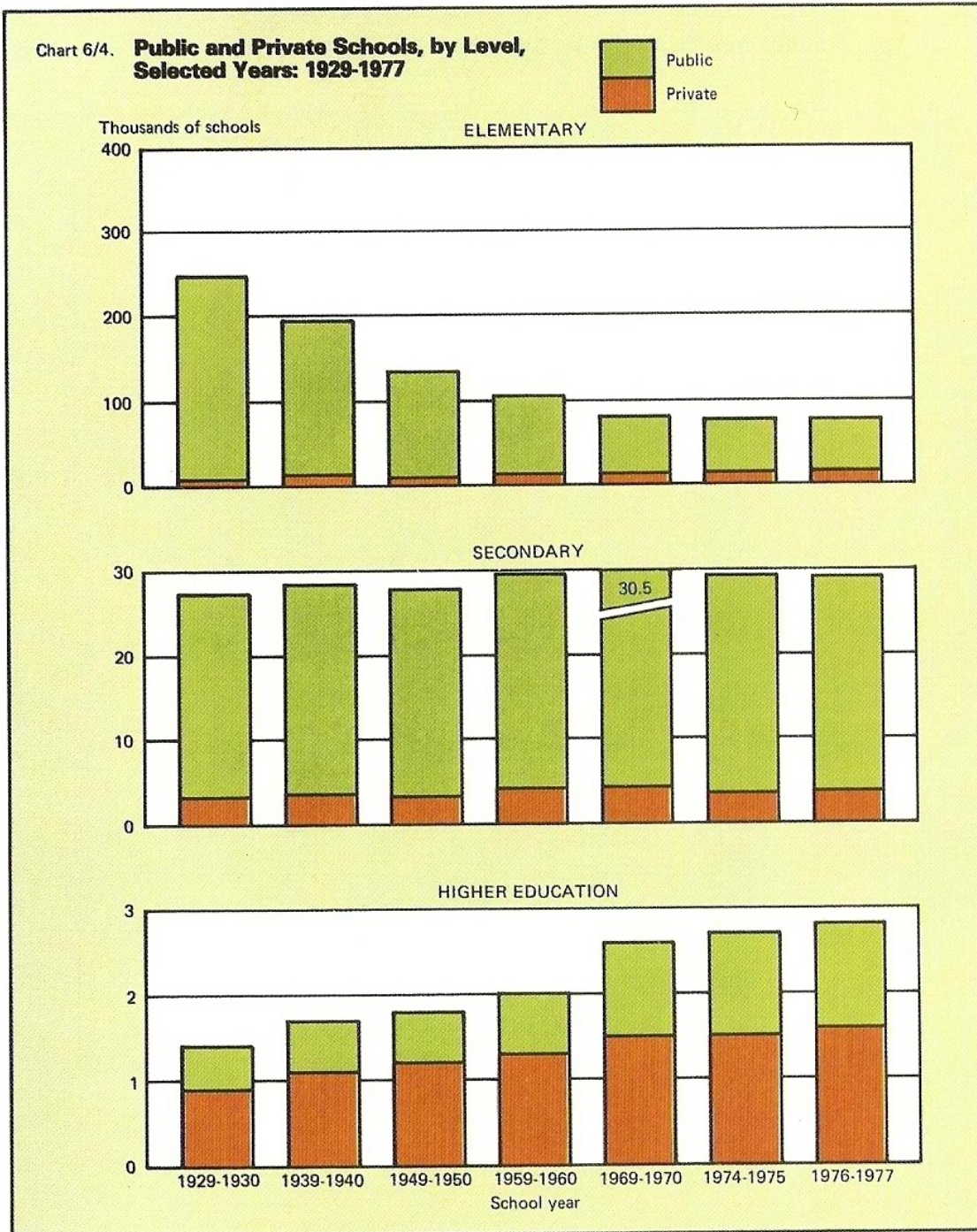


FIGURE 10. A redone example of the data from figure 8.

Figure 6: Wainer (1997), p. 18, Figure 10: Wainer (1997), p. 17, Figure 8, improved.



CHAPTER 1, FIGURE 11. Hiding the data in the scale.

Figure 7: Wainer (1997), p. 20A, Figure 11: Hiding the data in the scale.

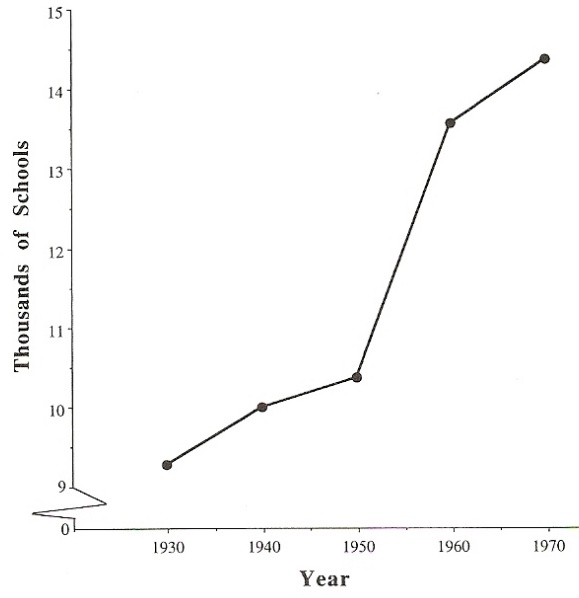


FIGURE 12. Expanding the scale and showing the data for the number of private elementary schools from figure 11.

Figure 8: Wainer (1997), p. 20, Figure 12: Wainer (1997), p. 20A, Figure 11, improved.

1.3.2 Show the data inaccurately

Rule 3: Ignore the visual metaphor altogether.

FIGURE 13. Ignoring the visual metaphor by letting a longer bar segment represent a smaller amount of coal (from the *New York Times*, 1978).

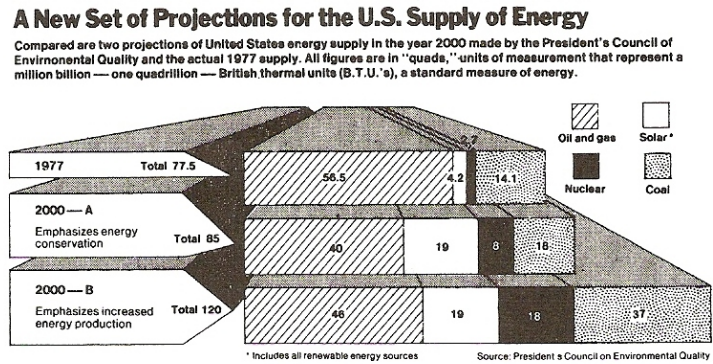


Figure 9: Wainer (1997), p. 20, Figure 13.

U.S. trade with China and Taiwan

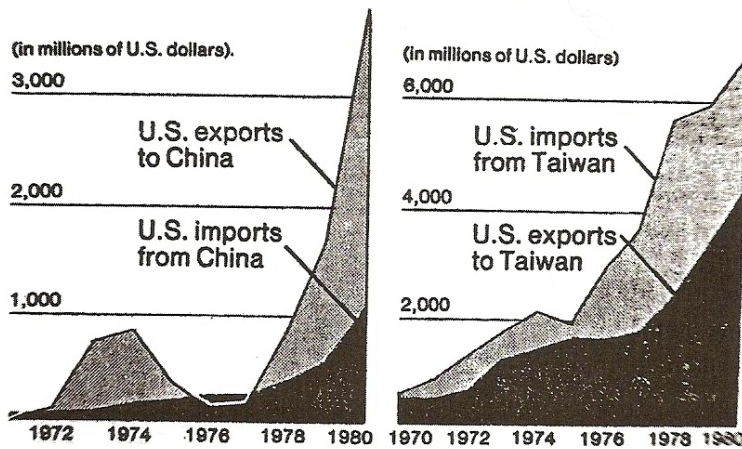


FIGURE 14. Reversing the metaphor in mid-graph while changing scales on both axes (from the *New York Times*, June 14, 1981).

Figure 10: Wainer (1997), p. 21, Figure 14.

FIGURE 15. Figure 14 redone with a consistent scale and visual metaphor.

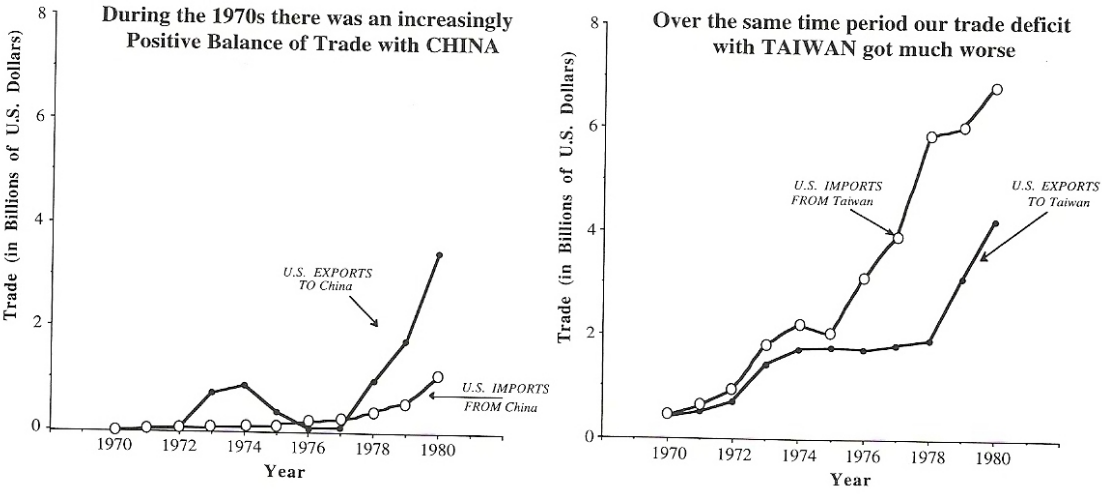


Figure 11: Wainer (1997), p. 21, Figure 15: Wainer (1997), p. 21, Figure 14, improved.

Rule 4: Only order matters.

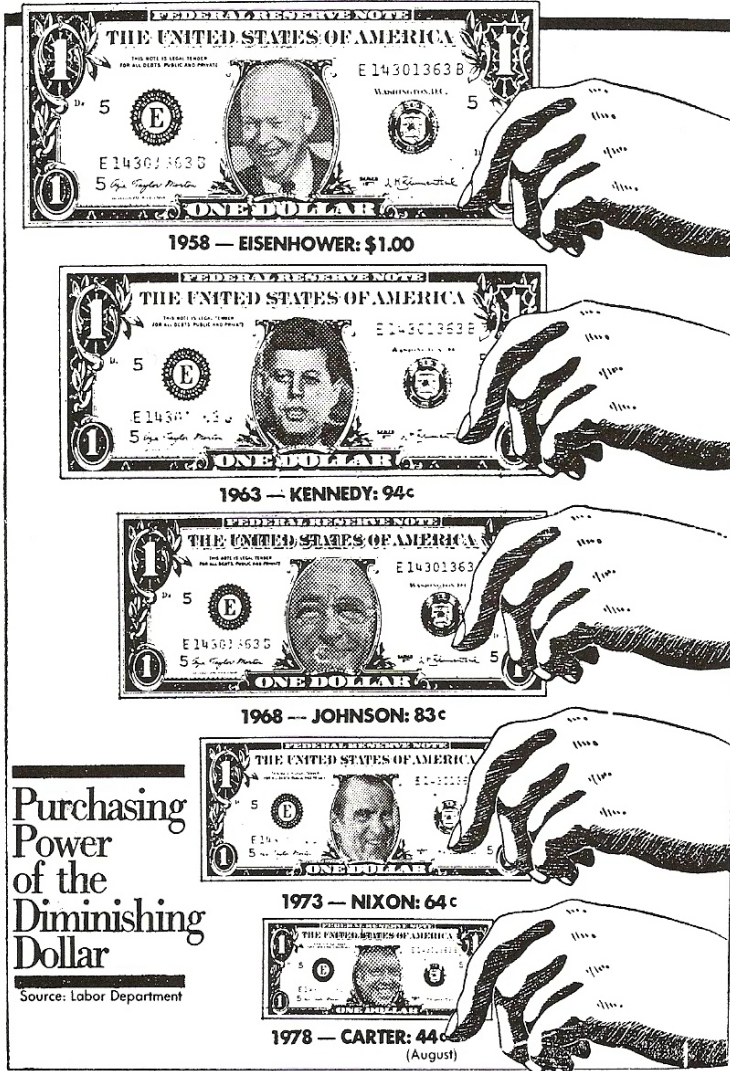


FIGURE 17. An example of how to goose up the effect by squaring the eyeball.

Figure 12: Wainer (1997), p. 23, Figure 17.

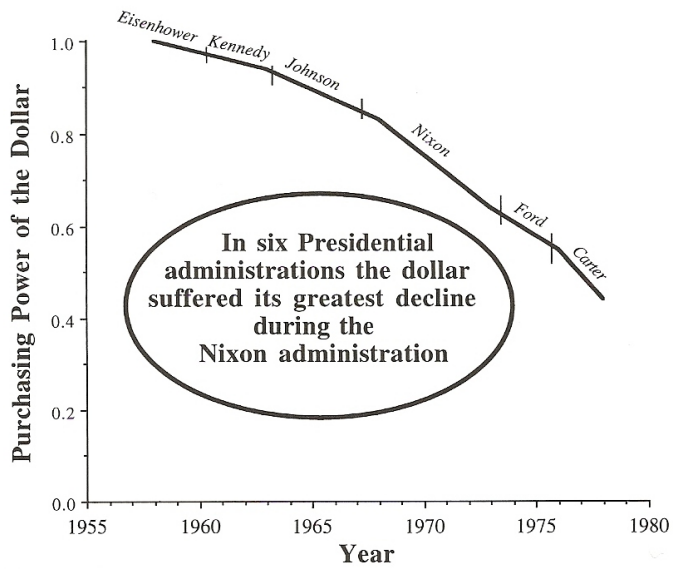


FIGURE 18. The data in figure 17 as an unadorned line chart (from Wainer, 1980).

Figure 13: Wainer (1997), p. 24, Figure 18: Wainer (1997), p. 23, Figure 17, improved.

FIGURE 19. Cubing the visual effect and choosing the origin to yield a near record lie factor of over 131,000% (from the *Washington Post*).

U.S. Beer Sales and Schlitz's Share

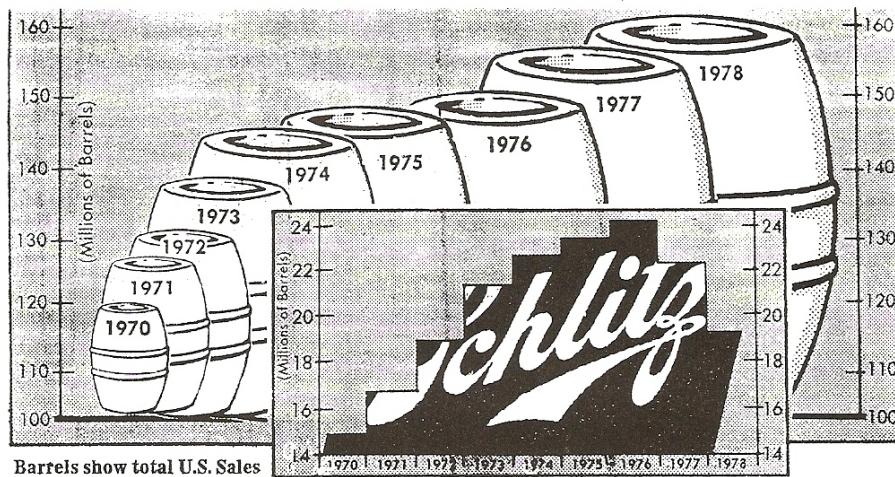


Figure 14: Wainer (1997), p. 24, Figure 19.

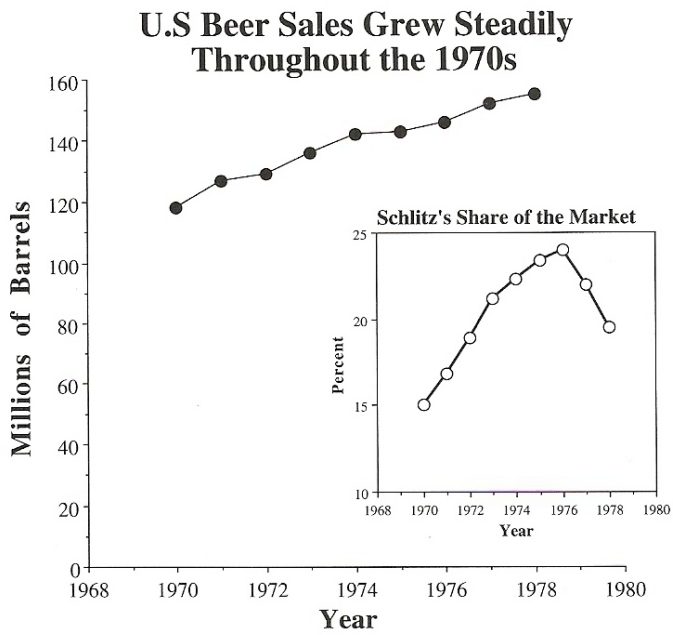


FIGURE 20. Data from figure 19 redone without tricks (from Wainer, 1980).

Figure 15: Wainer (1997), p. 25, Figure 20: Wainer (1997), p. 24, Figure 19, improved.

Rule 5: Graph data out of context.

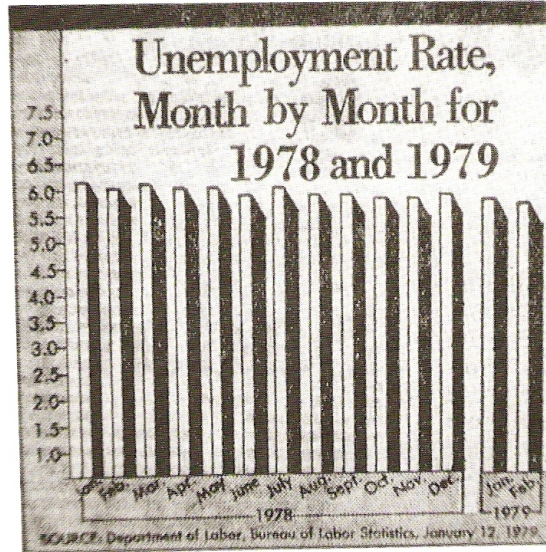


FIGURE 21. Hiding the effect by the careful choice of scale and origin (from the *Washington Post*).

Figure 16: Wainer (1997), p. 26, Figure 21.



FIGURE 22. Regraph of data from figure 21 with expanded scale, different starting point, and previous year's average added for context (from Wainer, 1980).

Figure 17: Wainer (1997), p. 26, Figure 22: Wainer (1997), p. 26, Figure 21, improved.

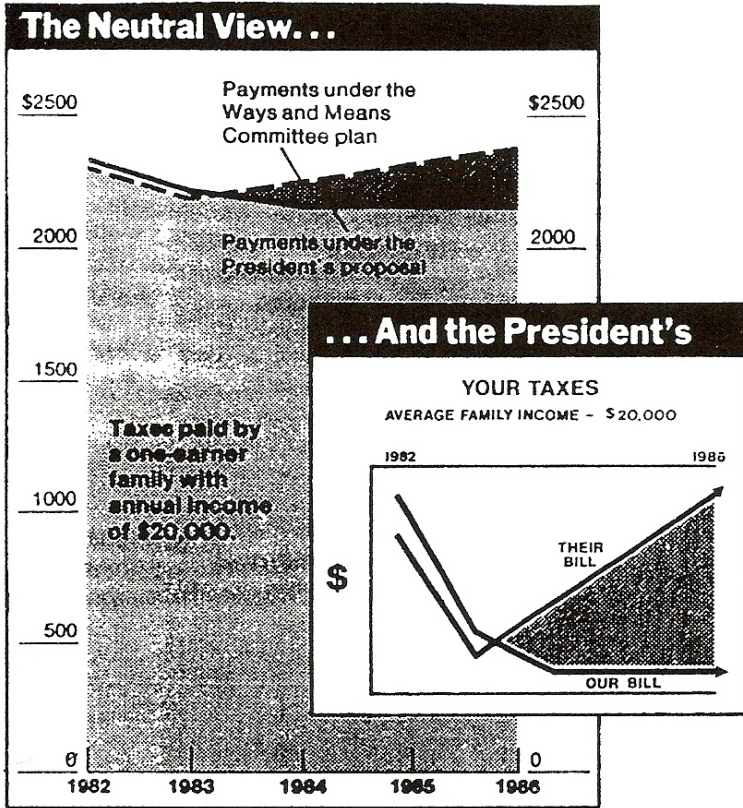


FIGURE 23. *New York Times* graphs showing how lack of context changes our perceptions about alternative tax bills.

Figure 18: Wainer (1997), p. 27, Figure 23.

1.3.3 Obfuscate the data

Rule 6: Change scales in mid-axis.

FIGURE 24. Changing the scale in mid-axis to make large differences seem small (from the *New York Post*, May 12, 1981).

The soaraway Post — the daily paper New Yorkers trust

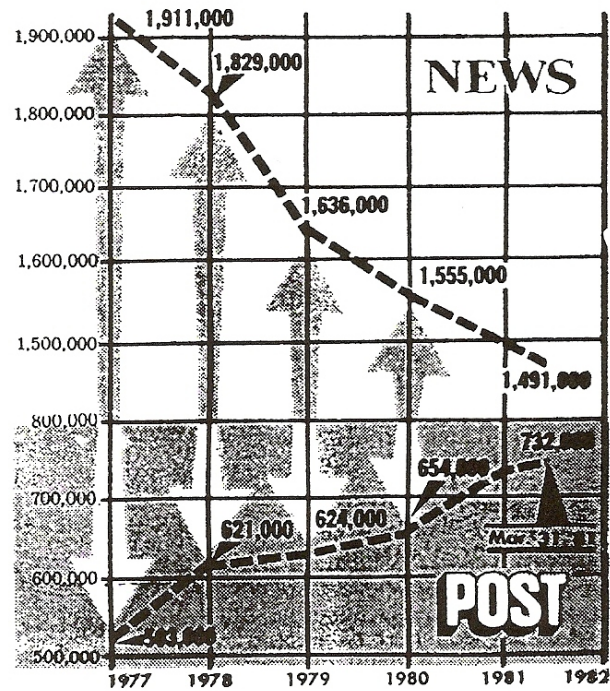


Figure 19: Wainer (1997), p. 28, Figure 24.

FIGURE 25. Changing scale in mid-axis to make exponential growth linear (from the *Washington Post*, Jan. 11, 1979, in an article titled "Pay, Practices of Doctors on Examining Table" by Victor Cohn and Peter Milius).

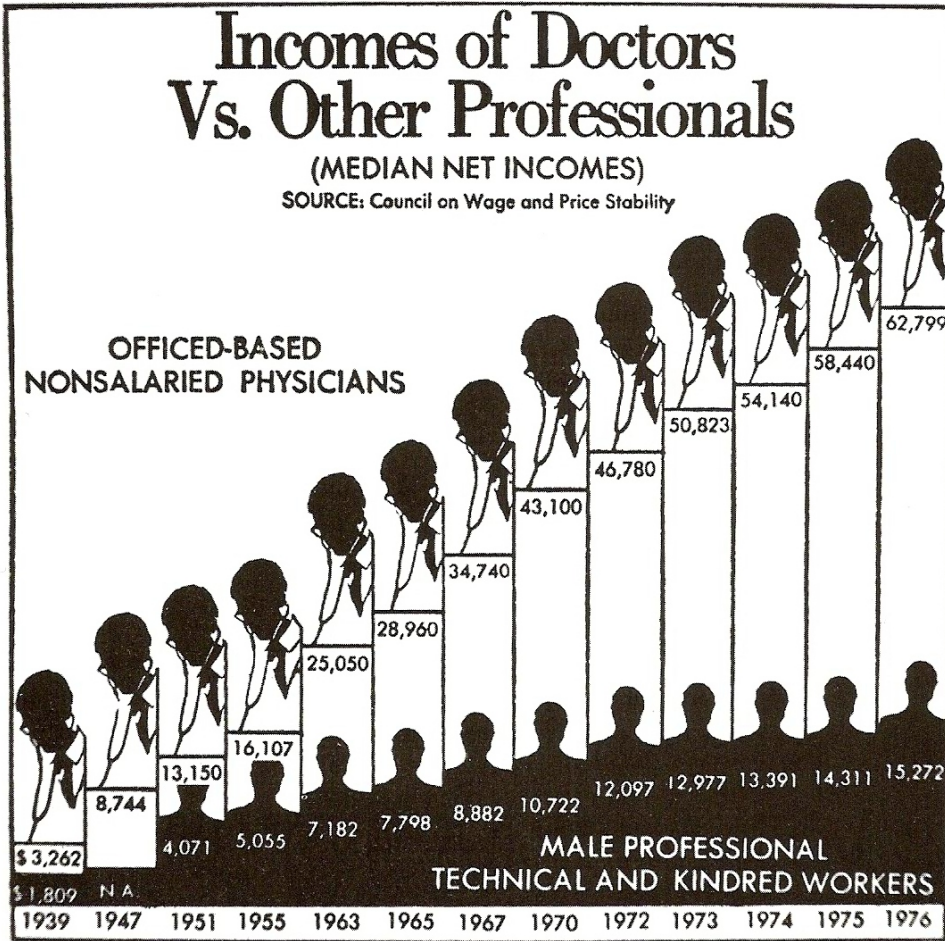


Figure 20: Wainer (1997), p. 29, Figure 25.

FIGURE 26. Data from figure 25 redone with a linear scale (from Wainer, 1980).

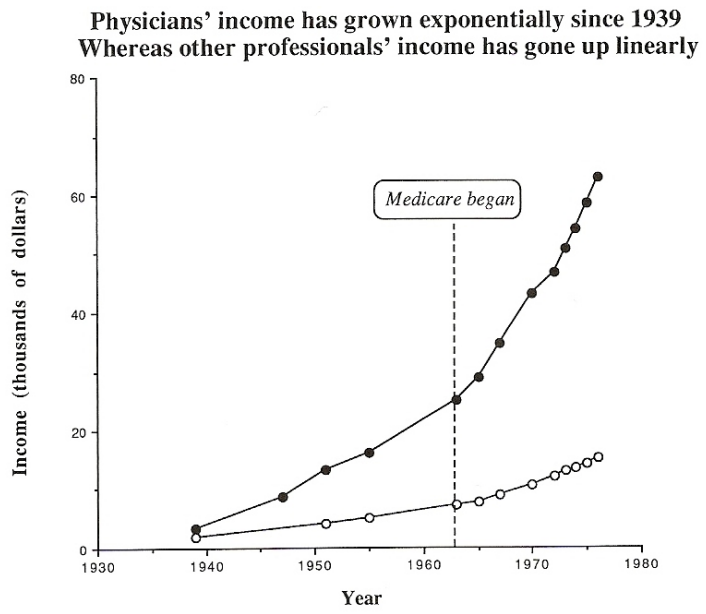
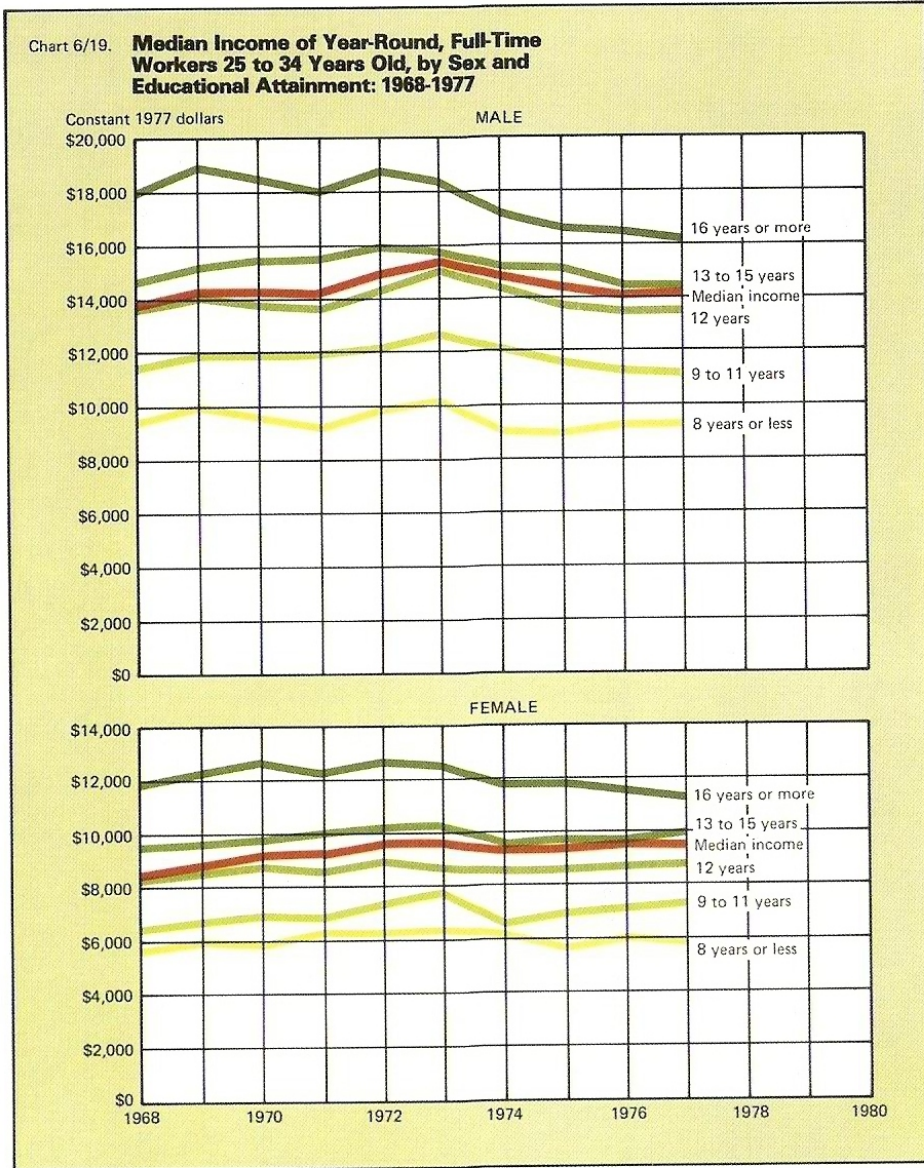


Figure 21: Wainer (1997), p. 30, Figure 26: Wainer (1997), p. 29, Figure 25, improved.

Rule 7: Emphasize the trivial (ignore the important).



CHAPTER 1, FIGURE 27. Emphasizing the trivial: Hiding the main effect of sex differences in income through the vertical placement of plots.

Figure 22: Wainer (1997), p. 20A, Figure 27.

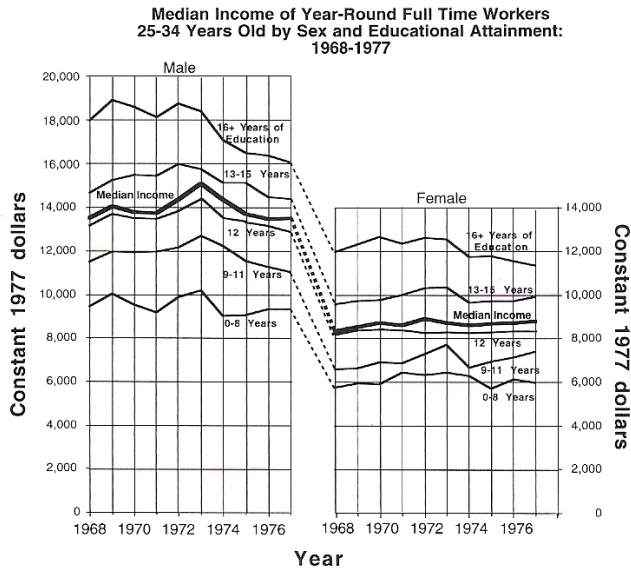


FIGURE 28. Figure 27 redone with the two plots horizontally opposed, showing the size of sex differences more clearly.

Figure 23: Wainer (1997), p. 31, Figure 28: Wainer (1997), p. 20A, Figure 27, improved.

**In the period 1968-1977 women's salaries
were about 60% of men's at comparable
levels of education**

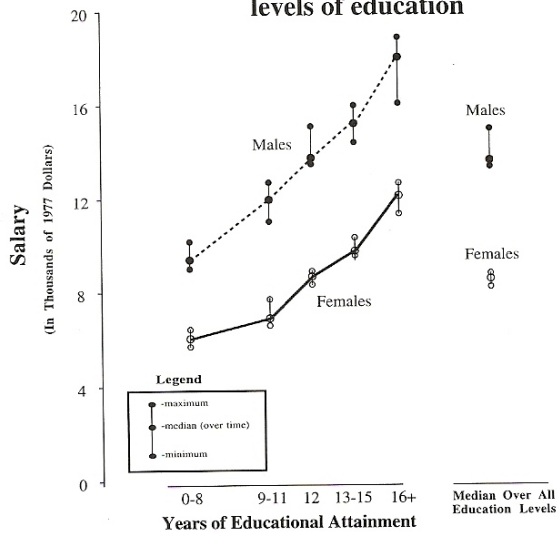
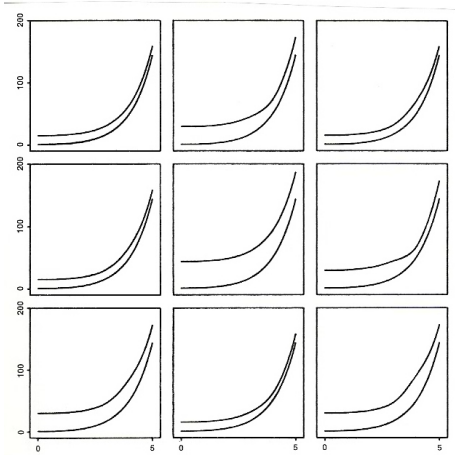


FIGURE 29. Figure 28 redone with the large effects of sex and education emphasized and the small-time trend suppressed.

Figure 24: Wainer (1997), p. 32, Figure 29: Wainer (1997), p. 31, Figure 28, further improved.

Rule 8: Jiggle the baseline.

FIGURE 30. A graphical experiment (from Cleveland and McGill, 1984). Without looking at the corresponding right panel, try to determine the difference between the two curves in the left panel.



*Sorry, these plots
got scrambled...*

Figure 25: Wainer (1997), p. 33, Figure 30.

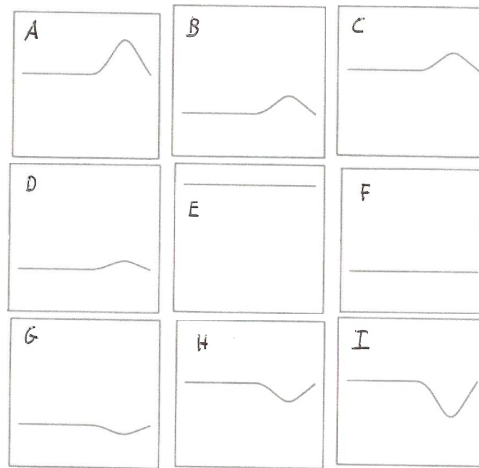
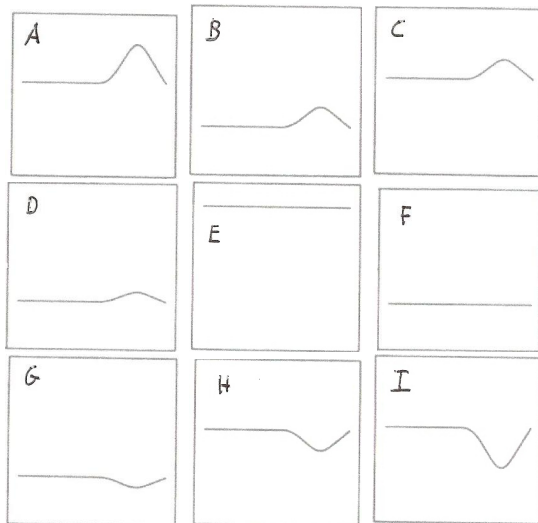
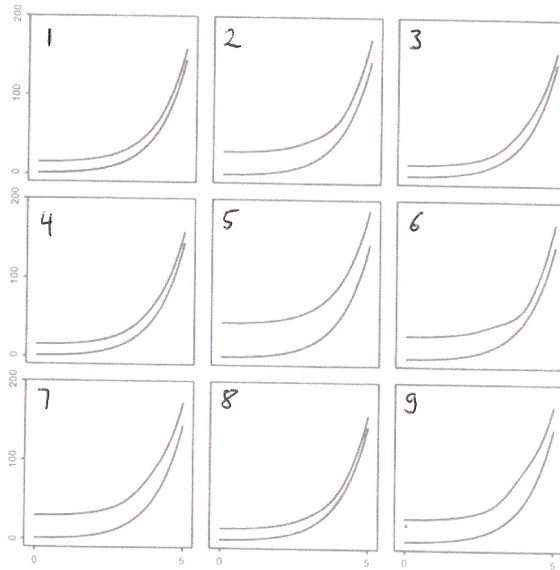


Figure 26: Wainer (1997), p. 33, Figure 30: Scrambled differences. The horizontal axis covers the interval 0 to 5, the vertical axis covers the interval 0 to 50.

Worksheet

Your Name: _____



Task: Match each original (labeled 1 to 9) with the plot (labeled A to I) that shows the difference between upper and lower line in the original plot.

Answer:

- 1: _____ 2: _____ 3: _____
 4: _____ 5: _____ 6: _____
 7: _____ 8: _____ 9: _____

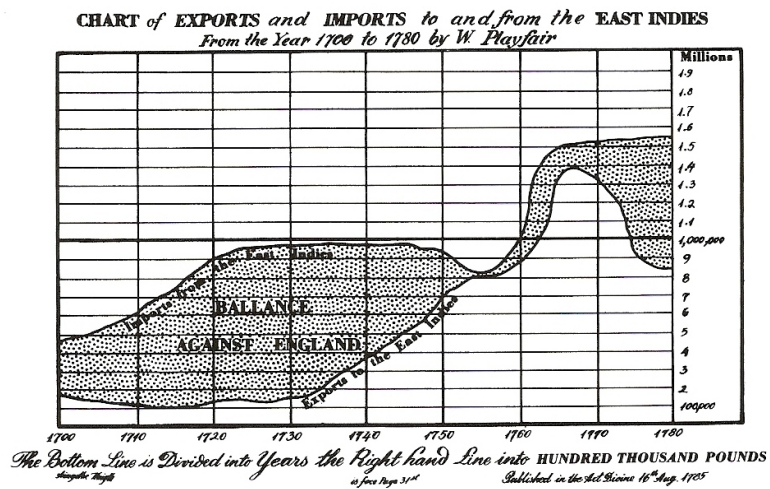


FIGURE 31. William Playfair's eighteenth-century graph of England's imports and exports with the East Indies (from Cleveland and McGill, 1984).

Figure 27: Wainer (1997), p. 34, Figure 31: One of William Playfair's few mistakes.

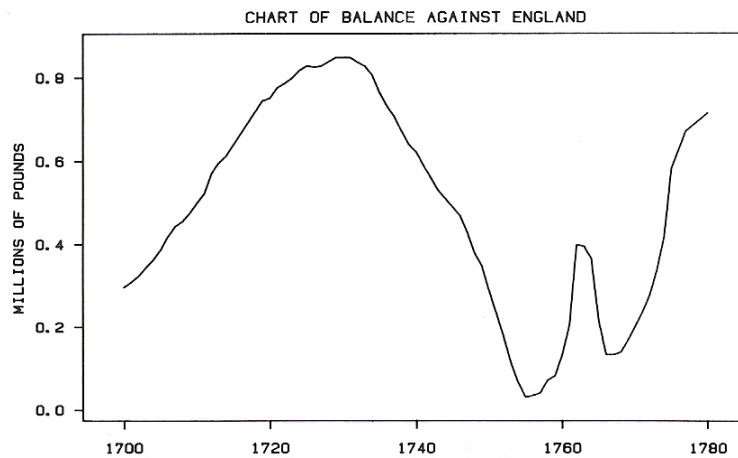


FIGURE 32. A graph of the difference between West Indies imports and exports showing explicitly the previously invisible jump in the 1760s (from Cleveland and McGill, 1984).

Figure 28: Wainer (1997), p. 34, Figure 32: Wainer (1997), p. 34, Figure 31, improved.

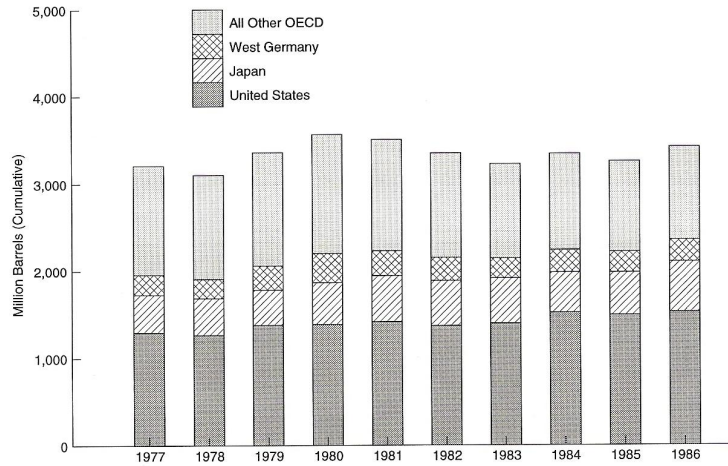


FIGURE 33. From the U.S. Department of Energy's *Annual Energy Review, 1986*, showing the changes in primary stocks of petroleum in OECD countries.

Figure 29: Wainer (1997), p. 35, Figure 33.

OECD PETROLEUM STOCKS HAVE STABILIZED
But Not All Countries Are Pulling Their Own Weight

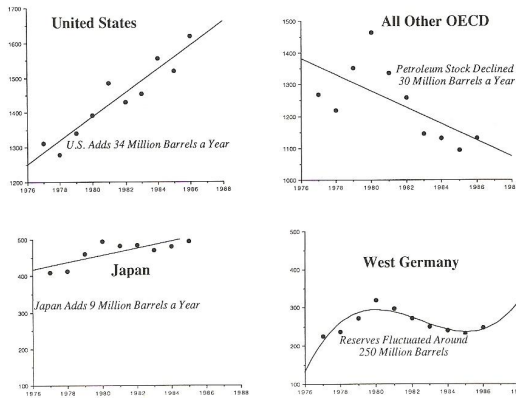
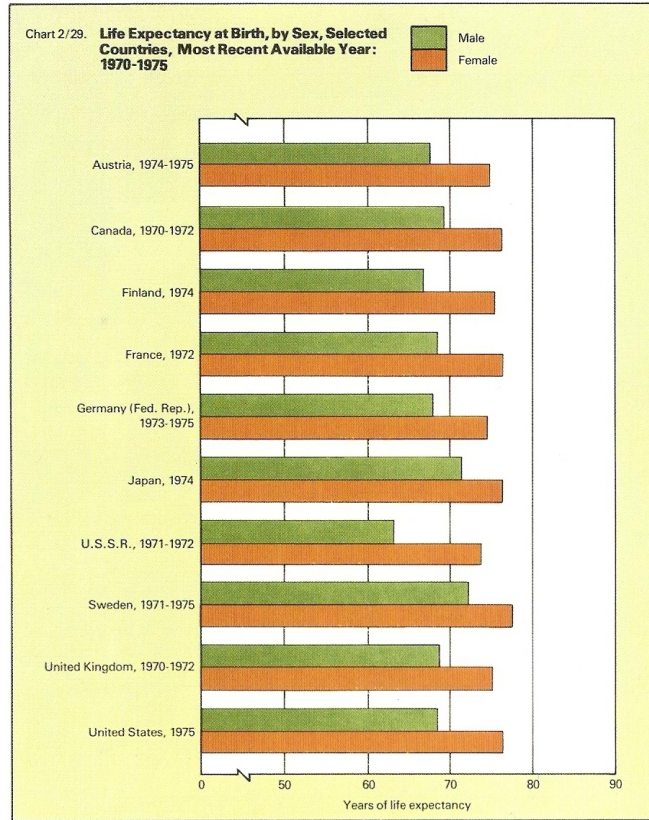


FIGURE 34. Regraphing of the data from figure 33 in which each country's data are shown relative to a straight line.

Figure 30: Wainer (1997), p. 36, Figure 34: Wainer (1997), p. 35, Figure 33, improved.

Rule 9: Alabama first!



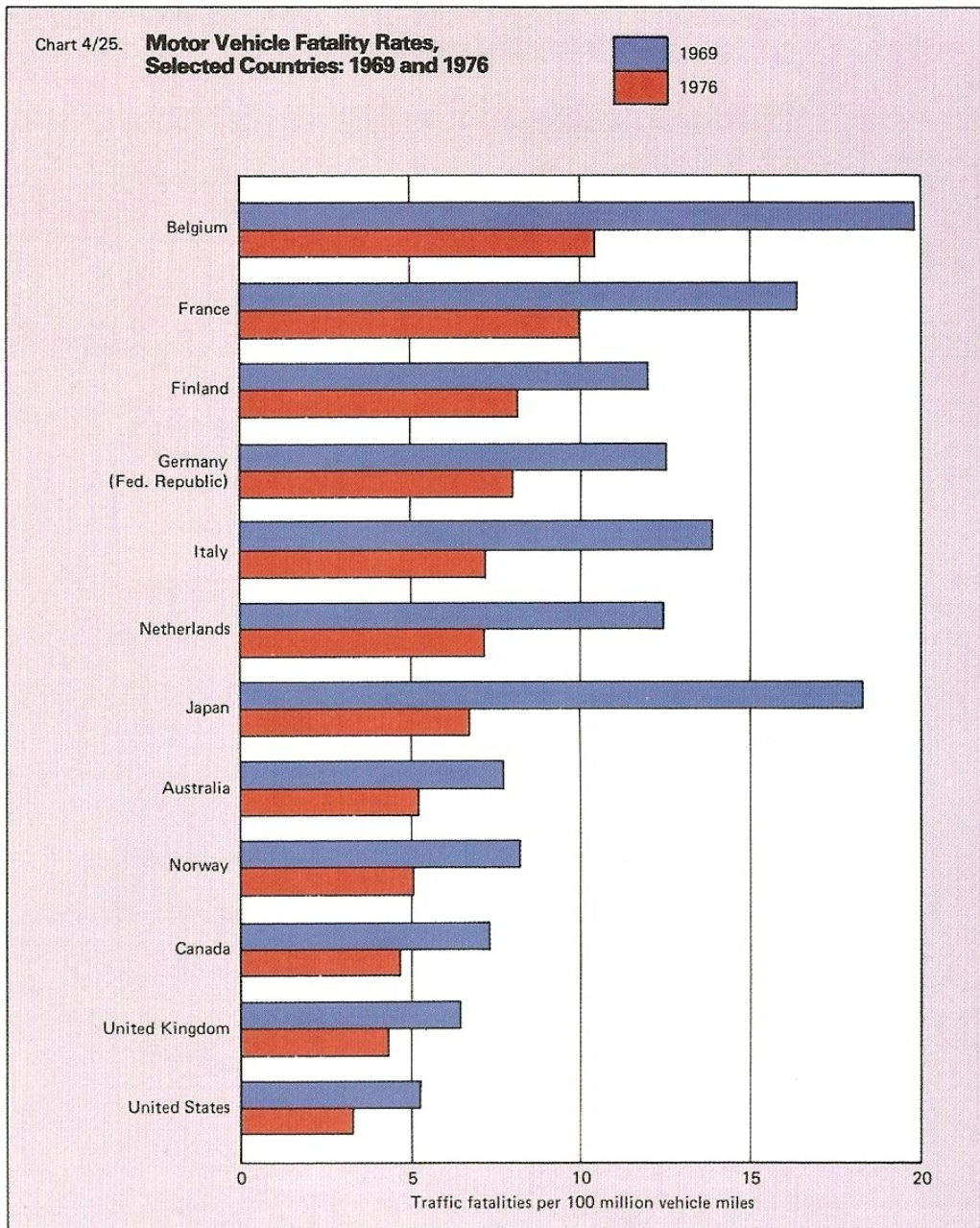
CHAPTER 1, FIGURE 35. Austria first! Obscuring the data structure in some life expectancy data by alphabetizing the plot.

Figure 31: Wainer (1997), p. 20B, Figure 35.



FIGURE 36. Ordering and spacing the data from figure 35 as a stem-and-leaf diagram provides insights previously invisible.

Figure 32: Wainer (1997), p. 37, Figure 36: Wainer (1997), p. 20B, Figure 35, improved.



CHAPTER 1, FIGURE 37. Ordering the bar chart by the data tells the tale a bit more clearly.

Figure 33: Wainer (1997), p. 20B, Figure 37: Layout similar to Wainer (1997), p. 20B, Figure 35, but improved due to ordering.

Rule 10: Label: (a) illegibly, (b) incompletely, (c) incorrectly, and (d) ambiguously.

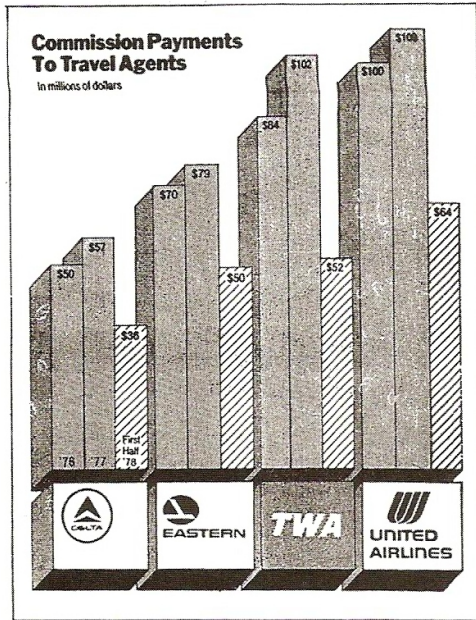


FIGURE 38. Mixing a changed metaphor with a tiny label reverses the meaning of the data.

Figure 34: Wainer (1997), p. 39, Figure 38.

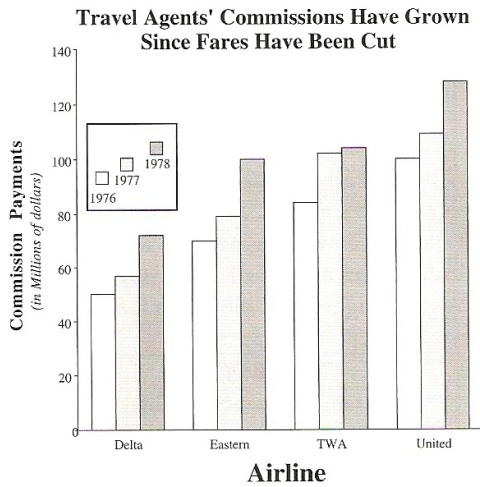


FIGURE 39. Figure 38 redrawn with 1978 data placed on a comparable basis shows that the fare cuts have been a boon to travel agents.

Figure 35: Wainer (1997), p. 39, Figure 39: Wainer (1997), p. 39, Figure 38, improved.

Rule 11: More is murkier: (a) more decimal places and (b) more dimensions.

Country	Male	Female
Argentina	56.90	61.40
Brazil	39.30	45.50
Canada	67.61	72.92
Iceland	66.10	70.30
Japan	65.37	70.26
Mexico	37.92	39.79
Netherlands	71.40	74.80
New Zealand	68.20	73.00
Norway	71.11	74.70
Spain	58.76	63.50

Figure 36: Wainer (1997), p. 40, Table 1.

Country	Male	Female
Netherlands	71	75
Norway	71	75
New Zealand	68	73
Canada	68	73
Iceland	66	70
Japan	65	70
Spain	59	64
Argentina	57	61
Brazil	39	46
Mexico	38	40

Figure 37: Wainer (1997), p. 40, Table 2: Wainer (1997), p. 40, Table 1, improved.

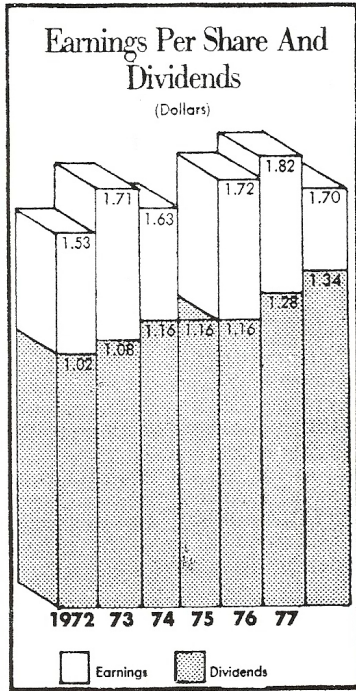


FIGURE 41. An extra dimension on earnings and dividends confuses even the grapher (from the *Washington Post*, 1979).

Figure 38: Wainer (1997), p. 41, Figure 41.

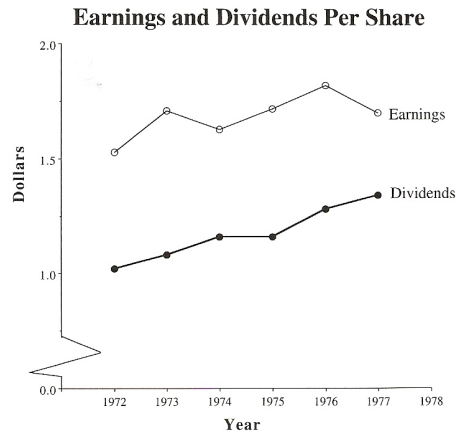
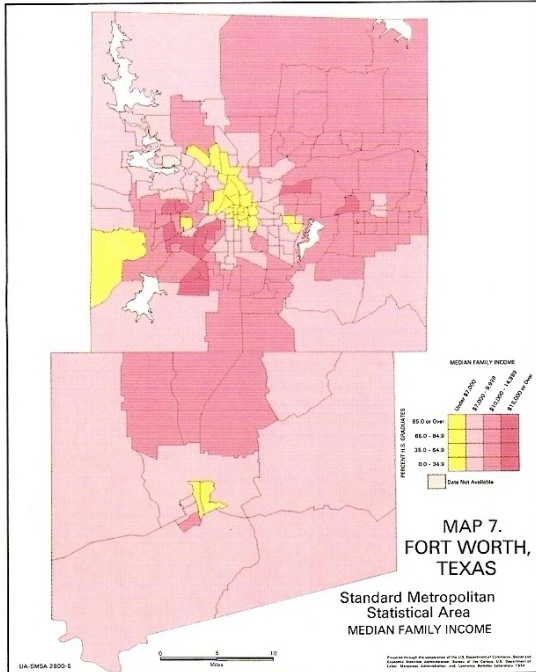


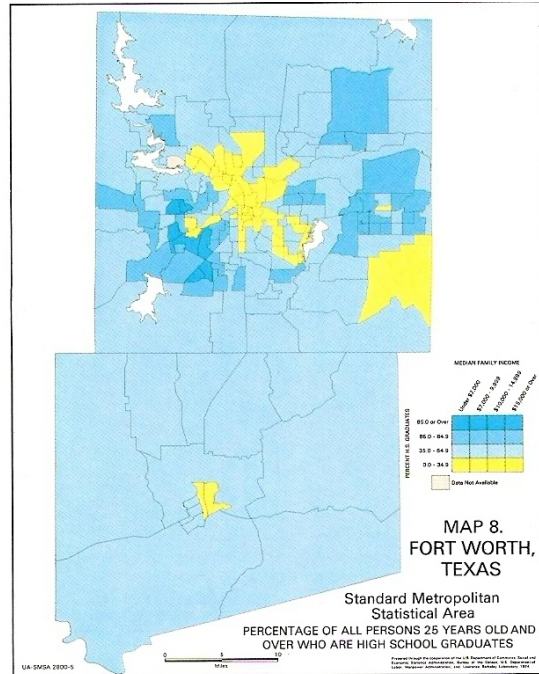
FIGURE 42. Data from figure 41 redrawn simply.

Figure 39: Wainer (1997), p. 42, Figure 42: Wainer (1997), p. 41, Figure 41, improved.

Rule 12: If it has been done well in the past, think of a new way to do it.

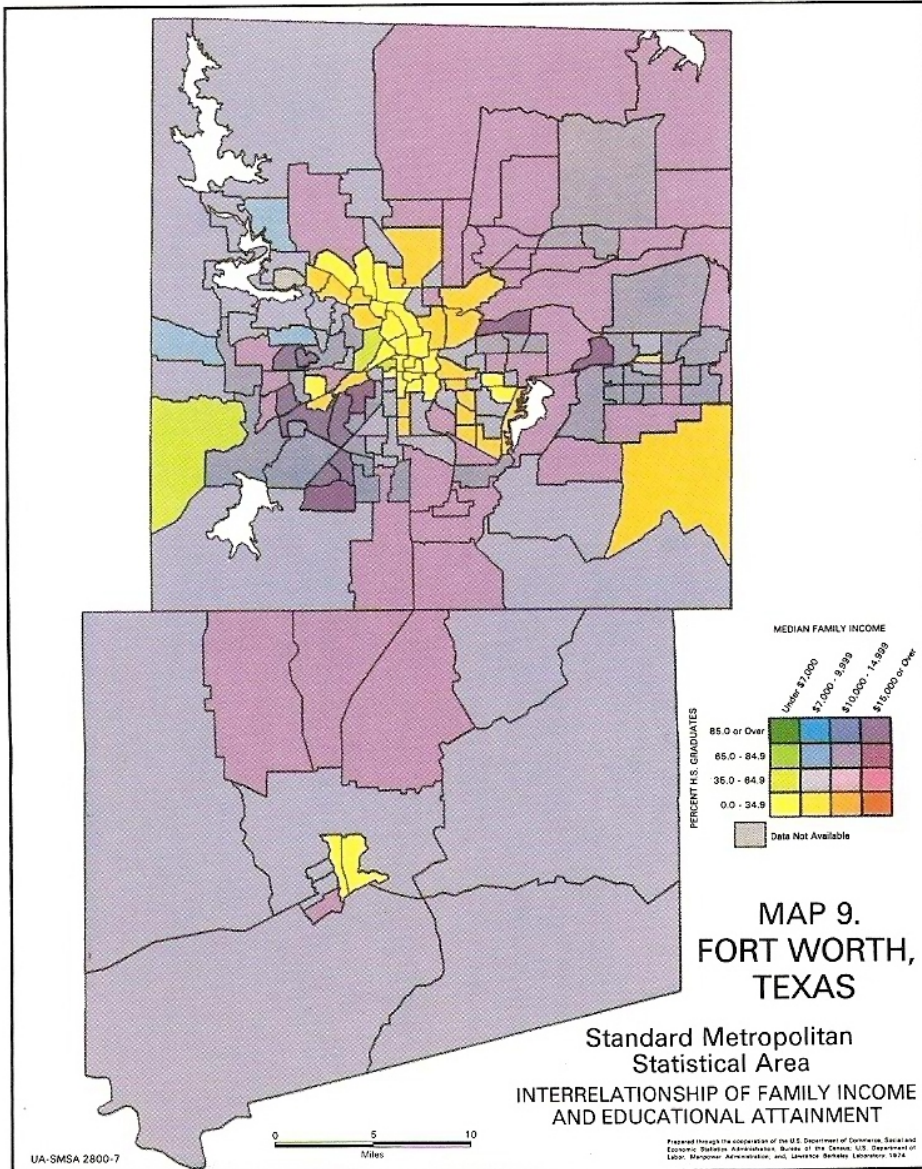


CHAPTER 1, FIGURE 44. The geographic distribution of median family income in Fort Worth, Texas, in 1974.



CHAPTER 1, FIGURE 45. The geographic distribution of percentage of high-school graduates in Fort Worth, Texas, in 1974.

Figure 40: Wainer (1997), p. 20C, Figures 44 & 45: Traditional maps.



CHAPTER 1, FIGURE 46. The geographic distribution of both median family income and percentage of high-school graduates in Fort Worth, Texas, in 1974, shown as a two-variable color map.

Figure 41: Wainer (1997), p. 20C, Figure 46: Wainer (1997), p. 20C, Figures 44 & 45, modified but **not** improved.

1.4 Bad Graphics are Everywhere — In Space and in Time

Example 1: Zion National Park, UT, Shuttle Parking Lot, December 28, 2002

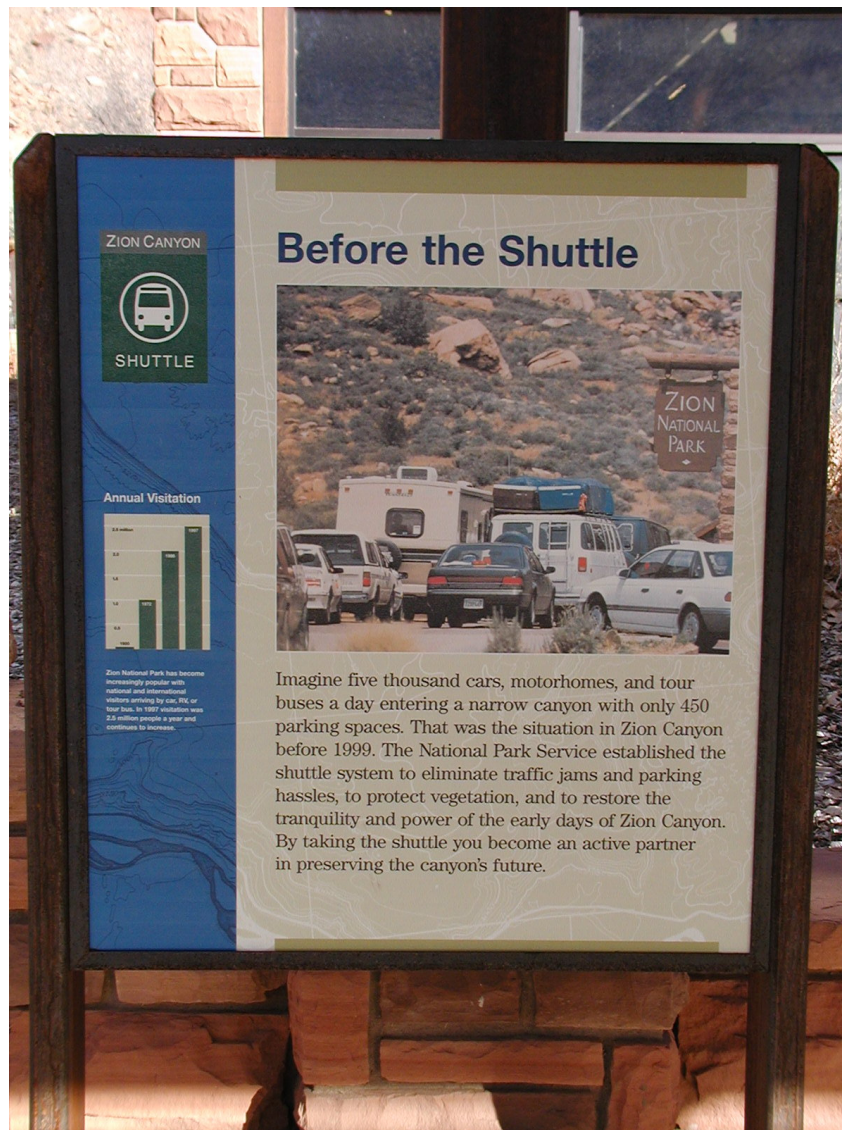


Figure 42: Personal Photograph: From the distance, the annual visitation appears to increase linearly, . . .

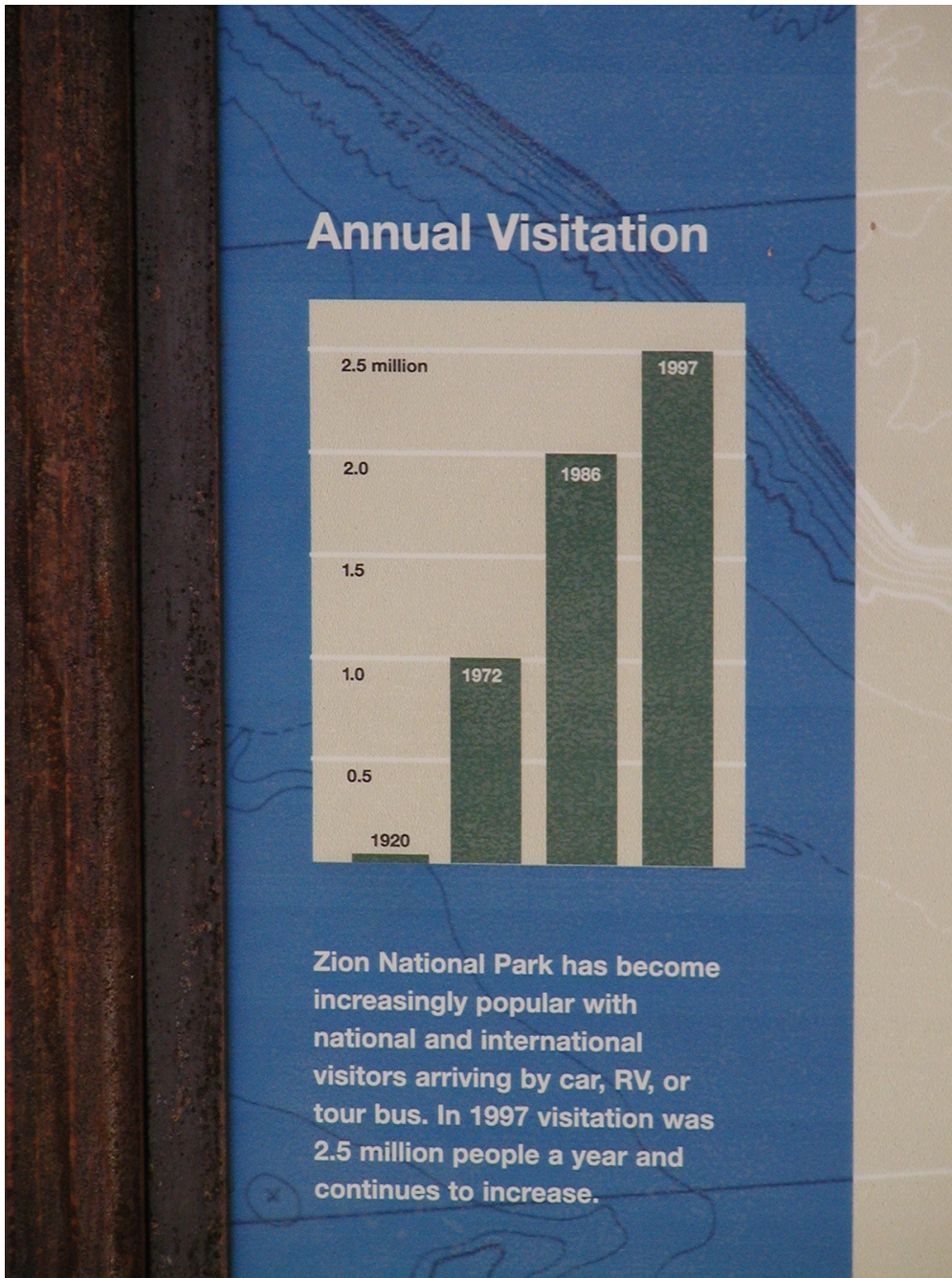


Figure 43: Personal Photograph: ... but at a closer view, this is certainly not the case.

Rules followed (to make this a bad graphic):

- Rule 6: Change scales in mid-axis.

Years on the horizontal axis are 1920, 1972, 1986, and 1997, i.e., the gaps are 52, 14, and 11 years. However, the same spacing has been used.

- Rule 10: Label: (a) illegibly, (b) incompletely, (c) incorrectly, and (d) ambiguously.

No axis label on vertical axis; what are 2.5 millions, etc. — visitors or cars? Also, no label on the horizontal axis. Moreover, listing the year near the top of each of the bars could be confusing as this might be interpreted as the actual number of visitors (in millions or so).

- Rule 1: Show as little data as possible (minimize the data density).

There are only 4 data points, but the figure is considerably filled with ink used for the bars.

- Rule 5: Graph data out of context.

Data are shown for only 4 years. However, the data range is 77 years. Why have these 4 years been chosen — and not any others (and in particular, why not more years)?

Improved Version:

http://www.math.usu.edu/~symanzik/teaching/2011_stat6560/RDataAndScripts/Zion.R

Example 2: Berlin, Germany, August 20, 2006



Figure 44: Personal Photograph: Exhibit at the 1936 Berlin Olympic Site, related to the history of the Olympic area from 1909 to 1936 to 2006.



Figure 45: Personal Photograph: Historical graphic (from the late 1920ies), dedicated to the development of women's gymnastics as part of the *Deutsche Turnerschaft* (the governing body of German gymnastics).

Rules followed (to make this a bad graphic):

- Rule 6: Change scales in mid-axis.

Years on the horizontal axis are 1897, 1900, 1904, 1907, 1914, 1919, 1921, 1924, and 1927, i.e., the gaps are 3, 4, 3, 7, 5, 2, 3, and 3 years. However, the same spacing has been used.

- Rule 4: Only order matters.

The size of the figures is not proportional to the numbers presented. Moreover, which part of the figure represents a value? (Look at the raised arms in 1904 and 1924.)

- Rule 3: Ignore the visual metaphor altogether.

The figure for 1919 is smaller than that for 1897, 1900, 1904, and 1907 — although the value is bigger for 1919. Moreover, two different scales are used for the vertical axis in the upper and in the lower part of the figure.

Improved Version:

http://www.math.usu.edu/~symanzik/teaching/2011_stat6560/RDataAndScripts/Berlin.R

Example 3: Wikipedia, 2009

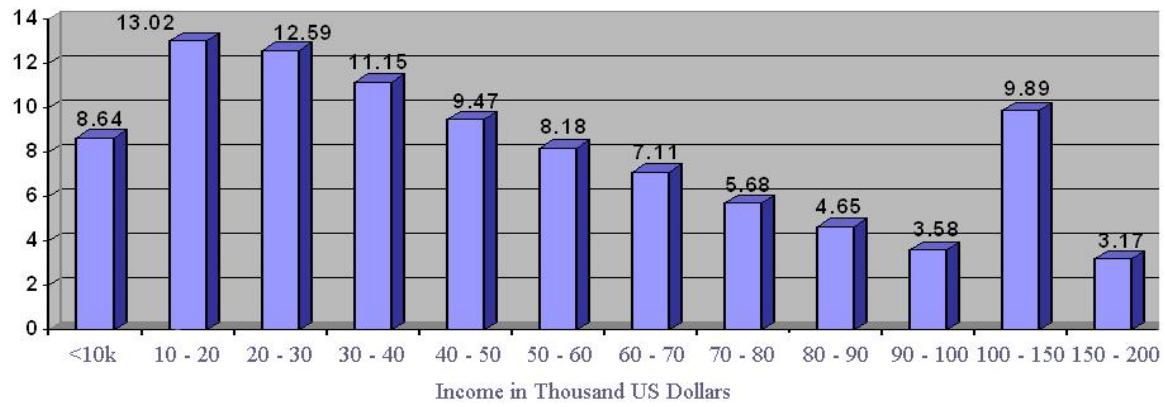


Figure 46: Figure taken from http://en.wikipedia.org/wiki/Household_income_in_the_United_States on 1/13/2009.

Rules followed (to make this a bad graphic):

- Incorrect plot type!!!

Income is quantitative and not categorical. We need a histogram here and not a bar chart.

- Rule 3: Ignore the visual metaphor altogether.

or: Rule 6: Change scales in mid-axis.

The 100–150 Thousand Dollars income interval seems to be the most outstanding interval, but this is due to the fact that this is a 50 Thousand Dollars wide interval. Most other intervals are only 10 Thousand Dollars wide. Histograms need to be drawn using the density scale if class intervals are differently wide, i.e., percentages have to be recalculated as percentage per unit.

- Rule 11: More is murkier: (a) more decimal places and (b) more dimensions.

No need for a third dimension for the bars. Also, no need to list two decimals for the percentages (e.g., 13.02).

- Rule 5: Graph data out of context.

Only incomes up to 200 Thousand Dollar are shown. But 2.87% of the incomes (not shown) are above 200 Thousand Dollars. Not showing these high incomes (and not even mentioning these incomes) is quite misleading.

- Rule 10: Label: (a) illegibly, (b) incompletely, (c) incorrectly, and (d) ambiguously.

Which year is this? The Wikipedia Web page deals with income data from 2004, 2005, and 2006 — so it is not clear which year is the basis for the data used in this figure.

Improved Version:

http://www.math.usu.edu/~symanzik/teaching/2011_stat6560/RDataAndScripts/Wiki.R

Example 4: Computational Statistics, 2002

430

age, surface and contour graphs. Various types of graphs created by KyPlot are shown in Figures 3 and 4.

Almost every component of each graph can be customized through dialog boxes. Double-clicking an axis of a graph brings up a dialog box through which one can change various settings for the axis interactively. The scales of the x- and y-axes of graphs can be individually set as either linear or logarithmic. Error bars can be attached to either x- or y-values, or both, and the attributes of individual error bars can also be customized. (For example, in Figure 3A, the error bars for two data points of a line graph have been partially suppressed to avoid overlapping.) A break along an axis can be set, over a specific range and at a specific location, to indicate that a range of values has been omitted (Figure 3B).

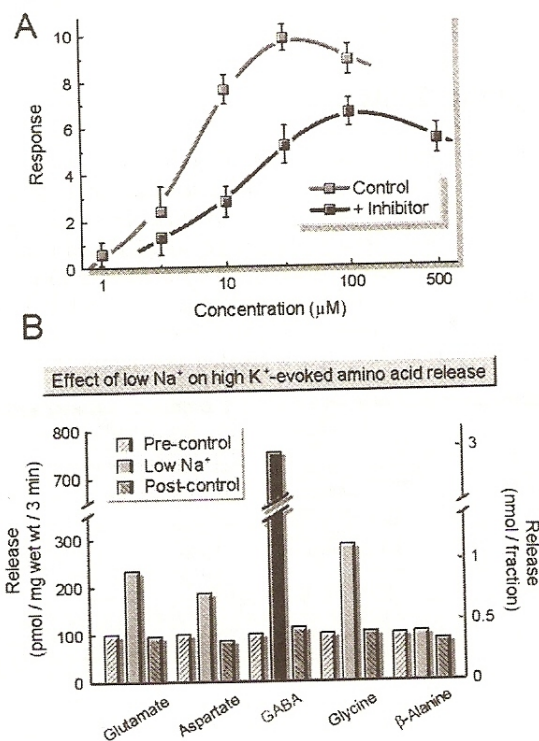


Figure 3: Line and bar graphs created with KyPlot

Figure 47: Yoshioka (2002), p. 430, Figure 3: Intended (!) features of the KyPlot software package for statistical data analysis and visualization.

Rules followed (to make this a bad graphic):

3A:

- Rule 10: Label: (a) illegibly, (b) incompletely, (c) incorrectly, and (d) ambiguously.

Concentration is drawn using a log₁₀-scale. This is not stated (and also not immediately clear with only one additional (unlabeled) ticmark. Moreover, just this additional ticmark (that is not labeled at all!) halfway between two labeled ticmarks makes it very difficult to read off concentration values. Reconstruction of these two missing labels as 3.2 and 32 requires some careful considerations.

- Rule 12: If it has been done well in the past, think of a new way to do it.

People have dealt with overplotting before. We can use different colors (or symbols) for example when parts of the data or information are being overplotted.

- Rule 3: Ignore the visual metaphor altogether.

Except for a concentration of 3.2, the error bars suggest an approximate symmetric (likely normal) distribution of the response. With the error bars partially suppressed, the distribution for the control seems to be skewed to the right and the distribution for the inhibitor seems to be skewed to the left for a concentration of 3.2. Otherwise, with error bars plotted for this concentration level as well, the likely message would be that there is no significant difference between control and inhibitor for a concentration of 3.2 (whereas there is a significant difference for the concentrations of 10, 32, and 100).

3B:

- Rule 6: Change scales in mid-axis.

There is a break in the vertical axis. 300 is followed by 700, but the distance between these two values is about the same as for differences of 100 elsewhere on the vertical axis.

- Rule 10: Label: (a) illegibly, (b) incompletely, (c) incorrectly, and (d) ambiguously.

There are two labeled vertical axes! Which of these axes/labels is used, and which is not used?

- Rule 9: Alabama first!

Even worse, there is no sorting at all here.

- Rule 7: Emphasize the trivial (ignore the important).

Five bars, i.e., some considerable amount of space, are used to display that the Pre-control is 100 for each of the five amino acids under investigation. Moreover, it takes a while to realize that the Post-control for all five amino acids is also close to 100 (with some small variation). The response under Low Na⁺ differs considerably, though, for the various amino acids.

- Rule 3: Ignore the visual metaphor altogether.

The “Low Na⁺” bar for “GABA” is shaded in black — while all other “Low Na⁺” bars are shaded in light gray. We can highlight a single observation (or a subset of observations), but then we should say so in the caption and indicate why these observations were highlighted.

Improved Version:

http://www.math.usu.edu/~symanzik/teaching/2011_stat6560/RDataAndScripts/Yoshioka.R

1.5 Rules for Good Data Displays

Wainer (1997), p. 46, suggests:

- “1. Examine the data carefully enough to know what they have to say, and then let them say it with a minimum of adornment.
2. In depicting scale, follow practices of “reasonable regularity.”
3. Label clearly and fully.”

Tufte (1983), p. 77, suggests:

“Graphical integrity is more likely to result if these six principles are followed:

- The representation of numbers, as physically measured on the surface of the graphic itself, should be directly proportional to the numerical quantities represented.
- Clear, detailed, and thorough labeling should be used to defeat graphical distortion and ambiguity. Write out explanations of the data on the graphic itself. Label important events in the data.
- Show data variation, not design variation.
- In time-series displays of money, deflated and standardized units of monetary measurements are nearly always better than nominal units.
- The number of information-carrying (variable) dimensions depicted should not exceed the number of dimensions in the data.
- Graphics must not quote data out of context.”

Robbins (2005), pp. 375–377, provides a “*Checklist of Possible Graph Defects*” in her Appendix A:

“Can the reader clearly see the graphical elements?”

- Do the data stand out? Are there superfluous elements?
- Are all graphical elements visually prominent?
- Are overlapping plotting symbols visually distinguishable?
- Can superposed data sets be readily visually assembled?
- Is the interior of the scale–line rectangle cluttered?
- Do data labels interfere with the quantitative data or clutter the graph?
- Is the data rectangle within the scale–line rectangle?
- Do tick marks interfere with the data?
- Do tick mark labels interfere with the data?
- Are axis labels legible?
- Are there too many tick marks?

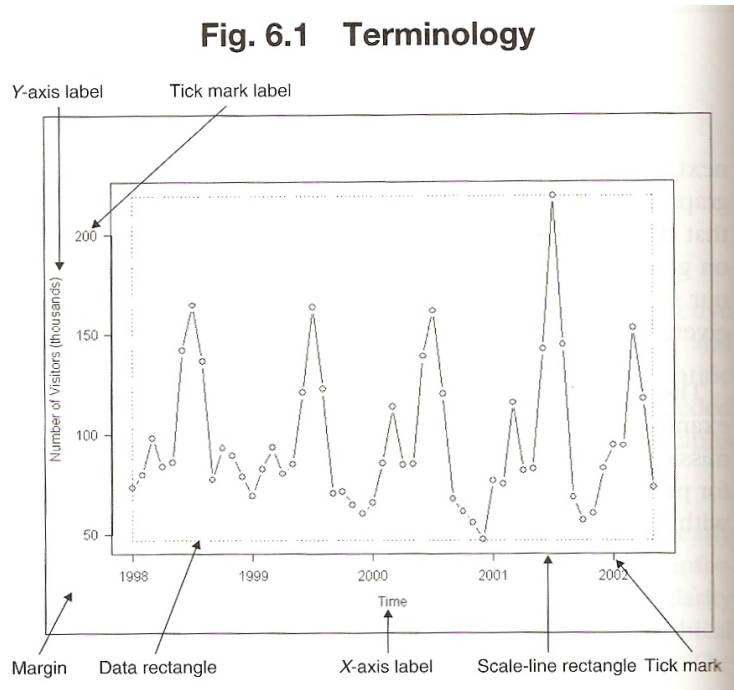


Figure 48: Robbins (2005), p. 156, Figure 6.1.

- Are there too many tick mark labels?
- Do the grid lines interfere with the data?
- Are there notes or keys inside the scale–line rectangle?
- Will visual clarity be preserved under reduction and reproduction?

Can the reader clearly understand the graph?

- Are the data drawn to scale?
- Is there an informative title?
- Is area or volume used to show changes in one dimension?
- Are there too many dimensions in the graph (more than in the data)?
- Are common baselines used wherever possible?
- Are all labels associated with the correct graphical elements?
- Is the reader required to make calculations?
- Are groups of charts drawn consistently?

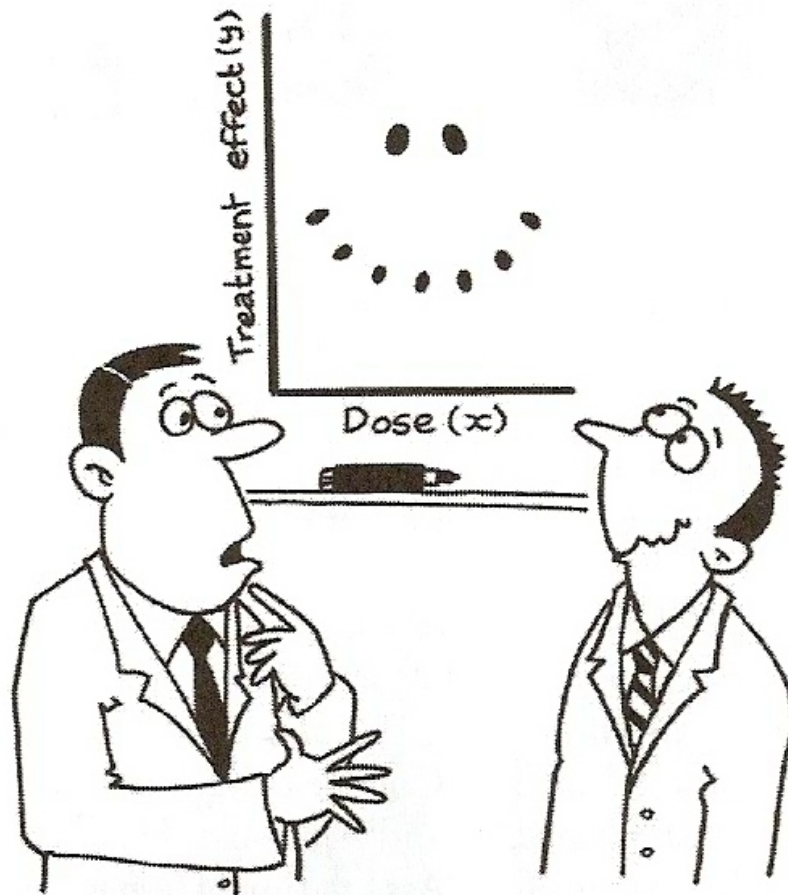
Are the scales well chosen and labeled?

- Is zero included for all bar graphs?
- Are there any unnecessary scale breaks?
- Is there a forceful indication of a scale break?
- Are there numerical values on two sides of a scale break that are connected?
- Does the aspect ratio allow the reader to see variations in the data?
- Are scales included for all axes?
- Are the scales labeled?
- Are tick marks at sensible values?
- Do the axes increase in the conventional direction?
- Does the data rectangle fill as much of the scale–line rectangle as possible?
- Are uneven time intervals handled correctly?
- Are the scales appropriate when different panels are compared?”

1.6 Further Reading

In addition to Wainer (1997), Tufte (1983), and Robbins (2005), cited so far in this chapter, many other sources exist that compare bad graphics with good graphics. Some of these additional sources are:

- Bertin (1977) and Bertin (2005) (first published in 1967)
- Henry (1995)
- Holmes (1991): check the author credentials and then decide whether this book is a source for good or bad graphics
- Huff & Geis (1954)
- Jones (2000)
- Kosslyn (1994) and Kosslyn (2006)
- Krämer (1991)
- Wainer (2005)
- Wainer (2007)
- Wallgren et al. (1996)
- Zelazny (2001)



"It's a non-linear pattern with outliers.....but for some reason I'm very happy with the data."

A CAUSE-commissioned cartoon that is part of the CAUSEweb collection and available for free noncommercial use by statistics teachers. Cartoon by John Landers ©. Provided by permission.

Figure 49: Amstat News, January 2009, p. 25, Cartoon.

2 History of Statistical Graphics: Plots, People, and Events

2.1 General History

- Michael Friendly's Web page: <http://www.math.yorku.ca/SCS/Gallery/milestone/>
- "*Milestones in the History of Data Visualization*" from the original "*Milestones in the History of Thematic Cartography, Statistical Graphics, and Data Visualization*". An illustrated chronology of innovations by Michael Friendly and Daniel J. Denis, York University, Canada. Organization by Mario Kanno (no longer active: www.infografe.com.br; current: <http://kanno-infografia.blogspot.com/>). http://www.math.yorku.ca/SCS/Gallery/milestone/Visualization_Milestones.pdf
- "*Milestones in the History of Thematic Cartography, Statistical Graphics, and Data Visualization*". Michael Friendly, August 24, 2009. <http://www.math.yorku.ca/SCS/Gallery/milestone/milestone.pdf>

2.1.1 Milestones in the History of Data Visualization (According to Friendly)

Pre-17th Century: Early Maps and Diagrams

1600–1699: Measurement and Theory

1700–1799: New Graphic Forms

1800–1849: Beginnings of Modern Data Graphics

1850–1899: Golden Age of Data Graphics

1900–1949: Modern Dark Ages

1950–1974: Re-birth of Data Visualization

1975–present: High-D Data Visualization

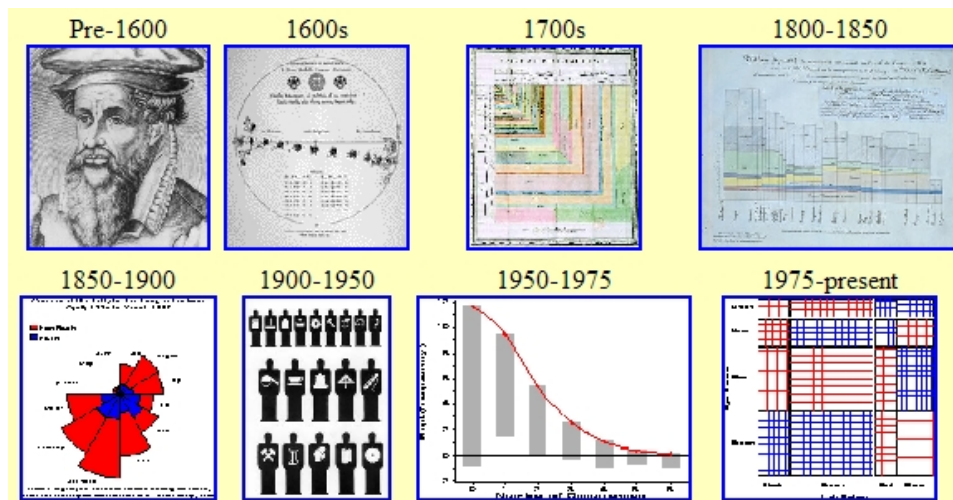


Figure 50: Screenshot taken from <http://www.math.yorku.ca/SCS/Gallery/milestone/> on 1/25/2009.

2.2 Selected People

Below are some of the individuals listed in Michael Friendly's "*Milestones in the History of Thematic Cartography, Statistical Graphics, and Data Visualization*". Heyde & Seneta (2001) present biographies of 103 important statisticians born between 1601 to 1900. Christiaan Huygens, William Playfair, Florence Nightingale, and Francis Galton are listed in Heyde & Seneta (2001) as well as in Friendly's milestones overview.

Christiaan Huygens: (1629–1695), Netherlands

1669: First graph of a continuous distribution function, a graph of Gaunt's life table, and a demonstration of how to find the median remaining lifetime for a person of given age.

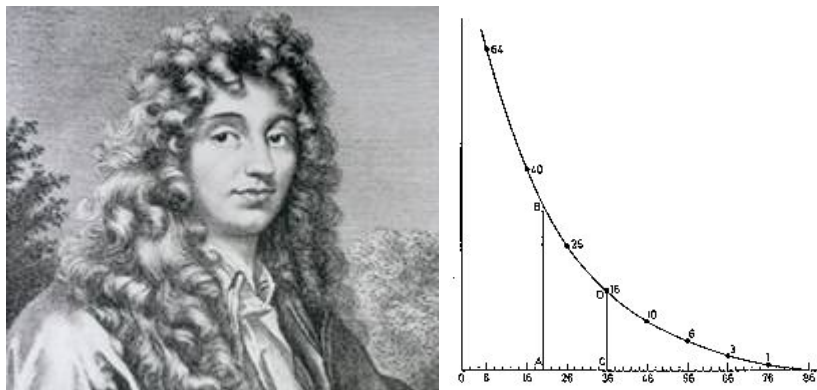


Figure 51: Figures taken from <http://www.math.yorku.ca/SCS/Gallery/images/portraits/huygens.jpg> and <http://www.math.yorku.ca/SCS/Gallery/images/huygens-graph.gif> on 1/27/2009.

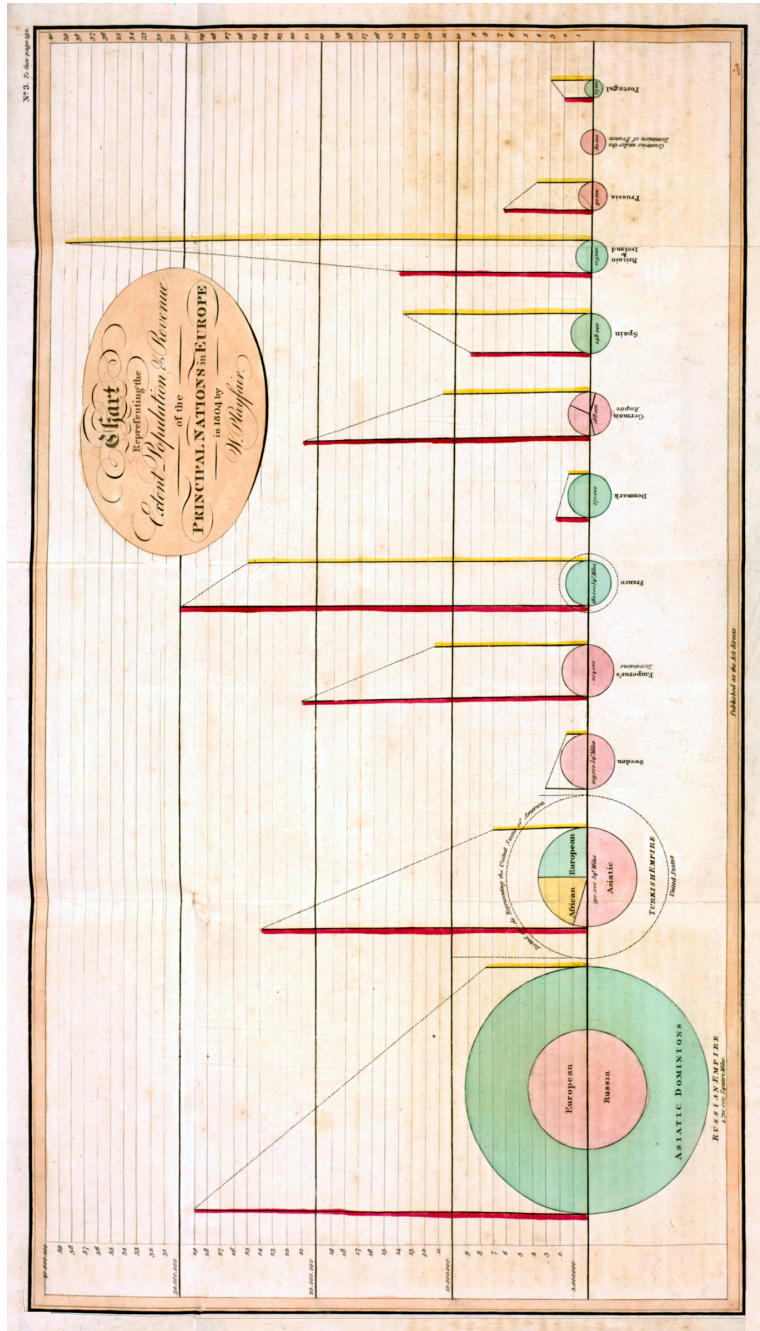


Figure 53: Symanzik et al. (2009), p. 553, Figure 1: *Chart Representing the Extent, Population & Revenue of the Principal Nations in Europe in 1804*. Plate 2 (labeled No. 3) from Playfair (1805). Courtesy of the Thomas Fisher Rare Book Library, University of Toronto.

Charles Joseph Minard: (1781–1870), France

1844: “Tableau-graphique” showing transportation of commercial traffic by variable-width (distance), divided bars (height \sim amount), area \sim cost of transport [An early form of the mosaic plot.]

1851: Map incorporating statistical diagrams: circles proportional to coal production (published in 1861).

1869: Minard’s flow map graphic of Napoleon’s Russian Campaign; often called “the best statistical graphic ever drawn” (Tufte 1983, p.40).

Further Reading:

- Tufte (1983), pp. 40–41 and more
- Wainer (1997), pp. 63–65
- Robinson (1967)
- Hankins (1999)
- Friendly (2000a) (a pdf of the newsletter is available at <http://stat-computing.org/newsletter/v111.pdf>)
- Tufte (2006), pp. 122–139

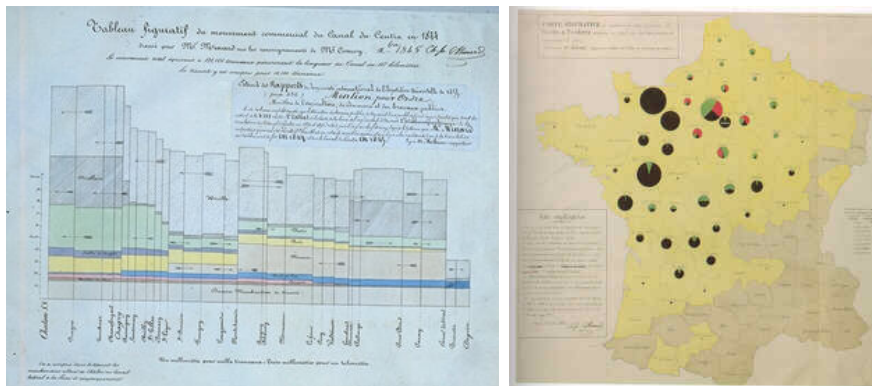


Figure 54: Figures taken from <http://www.math.yorku.ca/SCS/Gallery/images/enpc/img09a.jpg> and <http://www.math.yorku.ca/SCS/Gallery/images/Robinson/viandes.jpg> on 2/1/2009.

Florence Nightingale: (1820–1910), England

1857: Polar area charts, known as “coxcombs” (used in a campaign to improve sanitary conditions of army) or as “Nightingale’s Rose”.

Additional details and an animation of her coxcombs can be found at http://www.sciencenews.org/view/generic/id/38937/title/Math_Trek__Florence_Nightingale_The_passionate_statistician.

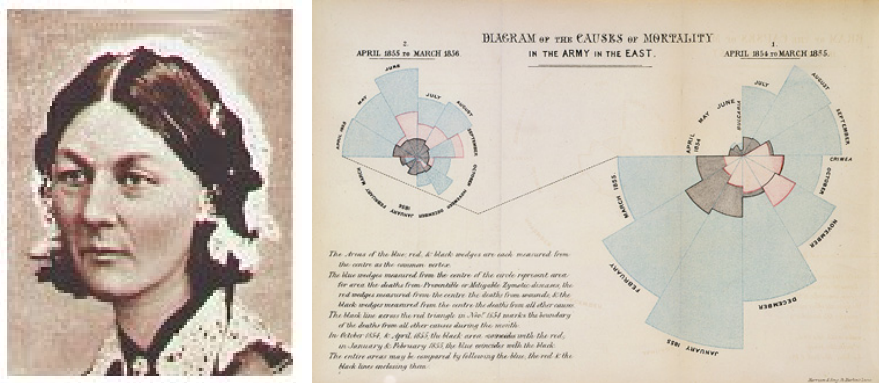


Figure 56: Figures taken from <http://www.math.yorku.ca/SCS/Gallery/images/portraits/nightingale.jpg> and <http://en.wikipedia.org/wiki/File:Nightingale-mortality.jpg> on 1/27/2009.

Francis Galton: (1822–1911), England

1861: The modern weather map, a chart showing area of similar air pressure and barometric changes by means of glyphs displayed on a map. These led to the discovery of the anti-cyclonic movement of wind around low-pressure areas.

c. 1874: Galton's first semi-graphic scatterplot and correlation diagram, of head size and height, from his notebook on *Special Peculiarities*.

1875: Galton's first illustration of the idea of correlation, using sizes of the seeds of mother and daughter plants.

1885: Normal correlation surface and regression, the idea that in a bivariate normal distribution, contours of equal frequency formed concentric ellipses, with the regression line connecting points of vertical tangents.

1899: Idea for “log-square” paper, ruled so that normal probability curve appears as a straight line.

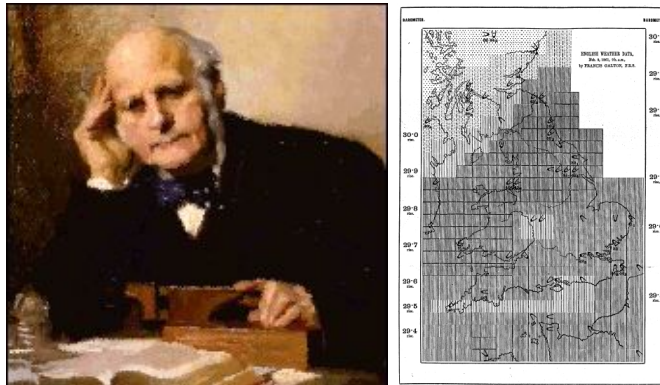


Figure 57: Figures taken from <http://www.math.yorku.ca/SCS/Gallery/images/portraits/galton-furse.gif> and <http://galton.org/essays/1860-1869/galton-1861-charts.pdf> on 1/27/2009.

John W. Tukey: (1915–2000), USA

1965: Beginnings of Exploratory Data Analysis (EDA): improvements on histogram in analysis of counts, tail values (hanging rootogram).

1969: Graphical innovations for exploratory data analysis (stem-and-leaf, graphical lists, box-and-whisker plots, two-way and extended-fit plots, hanging and suspended rootograms).

1974: Start of true interactive graphics in statistics; PRIM-9, the first system in statistics with 3-D data rotations provided dynamic tools for projecting, rotating, isolating and masking multidimensional data in up to nine dimensions — M. A. Fishkeller, Jerome H. Friedman and John W. Tukey.

1981: The “draftsman display” for three-variables (leading soon to the “scatter-plot matrix”) and initial ideas for conditional plots and sectioning (leading later to “coplots” and “trellis displays”) — John W. Tukey and Paul A. Tukey (a fifth cousin).

1990: Textured dot strips to display empirical distributions — Paul A. Tukey and John W. Tukey.

Further Reading:

- The PRIM-9 video, available at <http://stat-graphics.org/movies/>.
- Tukey (1977)
- Brillinger (2002) (preprint available at <http://www.stat.berkeley.edu/~brill/Papers/life.pdf>)

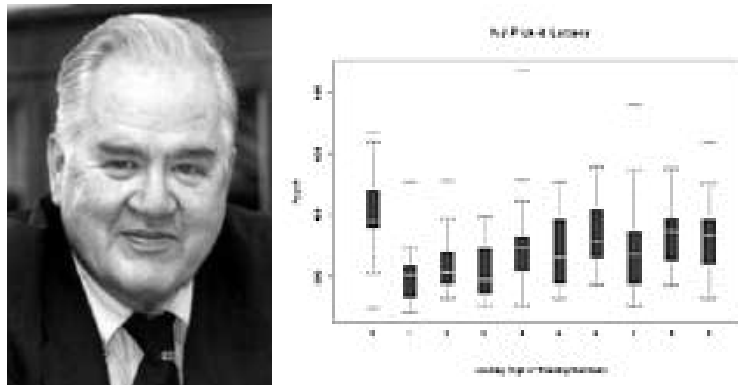


Figure 58: Figures taken from <http://www.math.yorku.ca/SCS/Gallery/images/portraits/tukey2.jpg> and <http://www.math.yorku.ca/SCS/Gallery/icons/NJPick-it.gif> on 2/1/2009.

Jacques Bertin: (1918–), France

1967: Comprehensive theory of graphical symbols and modes of graphics representation.

Among other things, Bertin introduced the idea of reordering qualitative variables in graphical displays to make relations more apparent, the reorderable matrix.

Further Reading:

- Bertin (1977)
- Bertin (2005)

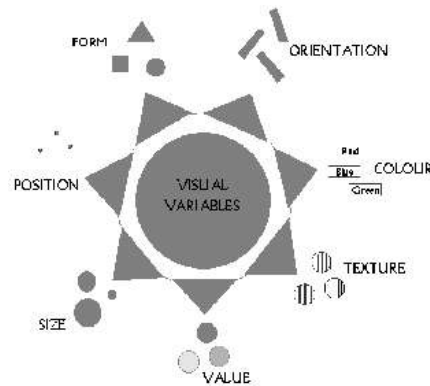
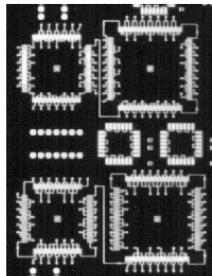


Figure 59: Figures taken from <http://www.math.yorku.ca/SCS/Gallery/images/portraits/jbertin.jpg> and <http://www.math.yorku.ca/SCS/Gallery/images/bertin-ve.jpg> on 2/1/2009.

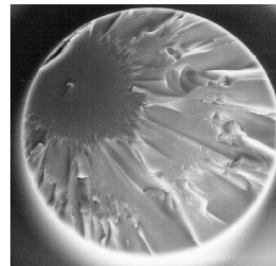
John W. Chambers and Collaborators: USA

Personal Home Page: <http://stat.stanford.edu/~jmc4/>

1978: S, a language and environment for statistical computation and graphics. S (later sold as a commercial package, S-Plus; more recently, a public-domain implementation, R is widely available), would become a *lingua franca* for statistical computation and graphics — Richard A. Becker and John M. Chambers.



Wafer Solder Image



Optical Fiber Preform

Figure 61: Figures taken from <http://stat.stanford.edu/~jmc4/> and <http://stat.stanford.edu/~jmc4/papers/93.1.ps> on 2/1/2009.

Antony Unwin and Collaborators: Ireland, England, Germany

Personal Home Page: <http://stats.math.uni-augsburg.de/~unwin/>

1988: Interactive graphics for multiple time series with direct manipulation (zoom, rescale, overlaying, etc.) — Antony Unwin and Graham Wills.

1989: Statistical graphics interactively linked to map displays — Graham Wills, J. Haslett, Antony Unwin and P. Craig.

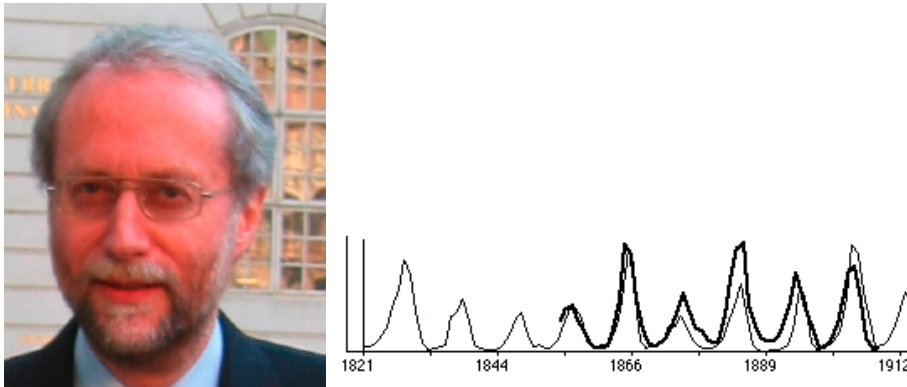


Figure 62: Figures taken from <http://stats.math.uni-augsburg.de/~unwin/> and <http://www.math.yorku.ca/SCS/Gallery/images/DiamondFast.jpg> on 2/1/2009.

Edward J. Wegman: USA

Personal Home Page: http://statistics.gmu.edu/people_pages/wegman.html

See here for a summary of his impressive vita — but also read about his denial of global warming:

<http://www.nationalpost.com/story.html?id=22003a0d-37cc-4399-8bcc-39cd20bed2f6&k=0>

1990: Statistical theory and methods for parallel coordinates plots.

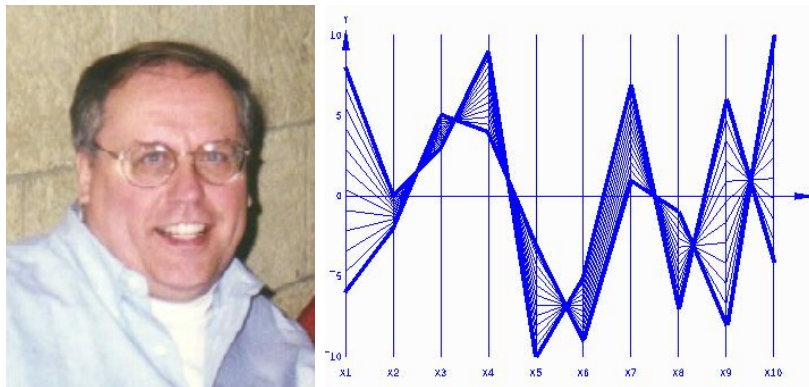


Figure 63: Figures taken from <http://www.math.yorku.ca/SCS/Gallery/people/EdWegman.jpg> and <http://www.math.yorku.ca/SCS/Gallery/images/parallel-coords.gif> on 2/1/2009.

2.3 Statistical Graphics and Events in History

2.3.1 John Snow and the Cholera Epidemic in London, 1854

(Based on a Student Project by William L. Welbourn in Spring 2009)

- John Snow (1813 – 1858): British anesthesiologist
- His investigation of the 1854 cholera outbreak in London, pioneered the field of epidemiology
- Some consider him the “*father of epidemiology*” (Vachon 2005)
- Cholera (Centers for Disease Control and Prevention 2010):
 - Acute diarrheal illness caused by intestinal infection by the bacterium *Vibrio cholerae*
 - Leads to rapid loss of body fluids and ultimately to dehydration and shock
 - Without treatment, death can occur within hours
- 1854 London Cholera Outbreak:
 - Six week period, beginning August 19, 1854
 - More than 575 deaths
 - “... Mortality in this limited area probably equals any that was ever caused in this country, even by the plague.” (Snow 1936)
- Snow’s Hypotheses:
 - Cholera is transmitted from person to person via fecal-oral route
 - Incubation period is 24 to 48 hours
 - The drinking water of the Broad Street Pump was the cause of the cholera outbreak



Figure 64: John Snow's rendition of the 1850 C.F. Cheffins Company Map, taken from http://www.ph.ucla.edu/epi/snow/snowmap1_1854_lge.htm. The white section of the map shows the area of particular interest to John Snow. The area encompassing the Broad Street Pump is circled (blue).

- Snow's Action to Control the Epidemic:
 - Utilized his map and empirical evidence to convince the Board of Guardians to remove the handle of the Broad Street Pump
 - A mere 48 fatal attacks occurred, following the removal of the handle of the Broad Street Pump, indicative that the water feeding the Broad Street Pump could be the source of the cholera epidemic

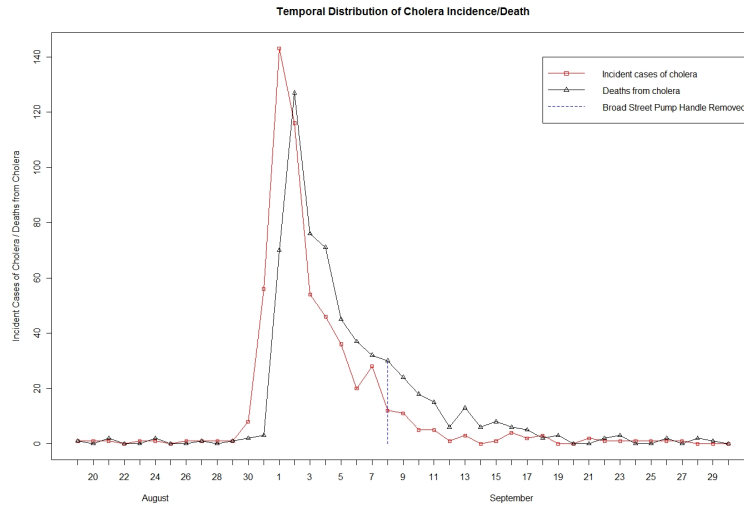


Figure 65: Time-series plot of the incident cases (red line) of cholera and deaths (black line) from cholera, for the time period, August 19, 1854 to September 30, 1854. The handle of the Broad Street Pump was removed September 8, 1854. (By William L. Welbourn)

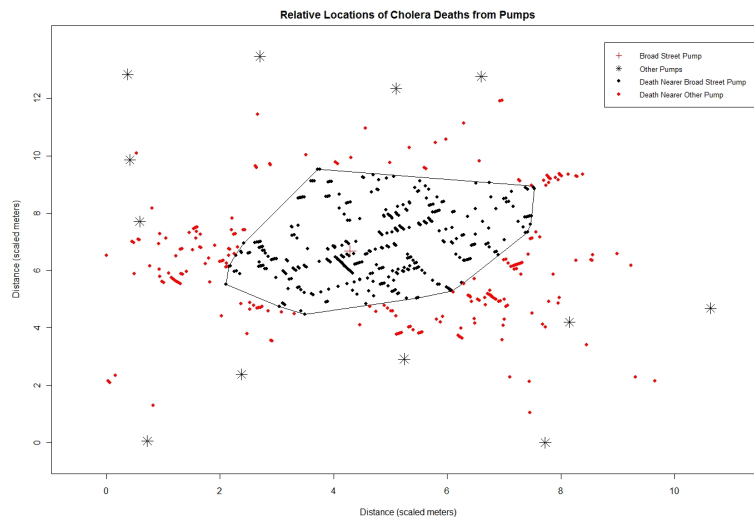


Figure 66: Relative positions of the 578 deaths arising from cholera and the thirteen pumps. 359 (62%) of these deaths occurred (black dots within the polygon) at a distance closer to the Broad Street Pump than to any other pump location. (By William L. Welbourn)

Further Reading:

- *John Snow — A Historical Giant in Epidemiology*, accessible at <http://www.ph.ucla.edu/epi/snow.html>
- Snow (1936), pp. 36–55
- Tufte (1997), pp. 27–37
- Wainer (1997), pp. 60–62
- Carvalho et al. (2004)
- R code to reproduce Figures 65 and 66:
http://www.math.usu.edu/~symanzik/teaching/2009_stat6560/RDataAndScripts/welbourn_william_project1_cholera.R

2.3.2 The Challenger Disaster, 1986

(Based on a Student Project by Abbass Sharif in Spring 2009)

- Background:
 - Lunch Time: January 28, 1986
 - The temperature was 31°F
 - Exploded after 73 seconds from its lunch leading to the death of its seven crew members
 - The Shuttle consisted of:
 - * The orbiter: Housed crew and controls
 - * An external fuel tank
 - * Two solid-rocket booster motors
 - * Each rocket-booster was shipped in 4 pieces
 - * Each rocket-booster has three joints called *O-rings* (6 total)

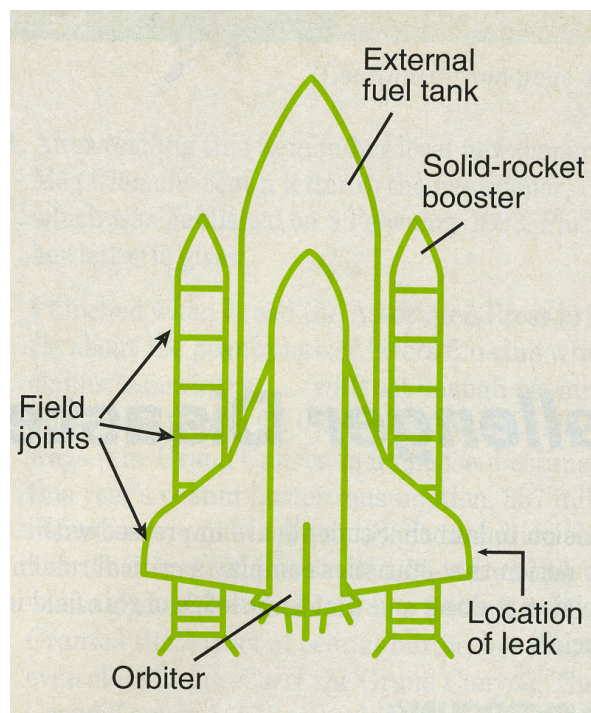


Figure 67: Tappin (1994), p. 424, Figure 1.

- Challenger Pre-launch Discussion:

- The night before the scheduled launch, discussions occurred as to whether the launch should be postponed because of the low temperature
- It was believed, by some of the people who were involved in the decision, that low temperatures might harden the O-ring seals, and thus leading to a potentially dangerous combustion-gas leak (Wainer 1997)
- The O-rings had been designated as a “Criticality 1” component
- The engineers and manufacturers of the rocket motors believed that they should abort the flight
- They presented several (hand written) tabulated numbers to show their point

BLOW BY HISTORY	HISTORY OF O-RING TEMPERATURES (DEGREES - F)				
	MOTOR	MGT	AMB	O-RING	WIND
SRM-15 WORST BLOW-BY					
o 2 CASE JOINTS (80°), (110°) ARC	DM-4	68	36	47	10 MPH
o MUCH WORSE VISUALLY THAN SRM-22	DM-2	76	45	52	10 MPH
SRM 22 BLOW-BY	QM-3	72.5	40	48	10 MPH
o 2 CASE JOINTS (30-40°)	QM-4	76	48	51	10 MPH
	SRM-15	52	64	53	10 MPH
	SRM-22	77	78	75	10 MPH
SRM-13A, 15, 16A, 18, 23A 24A	SRM-25	55	26	29	10 MPH
o NOZZLE BLOW-BY				27	25 MPH

Figure 68: Tufte (1997), p. 42, Figure.

- These tables were unconvincing to their managers because:
 - * NASA’s pressure
 - * Did not show clearly the relationship between temperature and the number of O-rings failing to operate

- Challenger Post-launch Discussion:

- Subsequent to the explosion, commission staff members tried to graphically replicate the flaws in the pre-launch reasoning process
- They plotted the data from previous twenty three space shuttle launching where there was at least one O-ring failure
- The dataset consisted of two variables: the launching temperature and number of damaged O-rings

- The temperature has an average around 63°F and standard deviation equal to 8°F
- The conclusion was: there is no effect of temperature

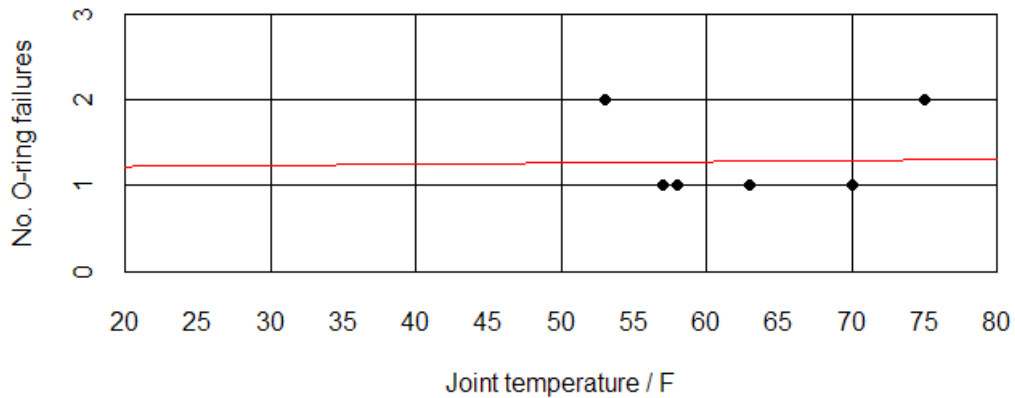


Figure 69: Reconstruction of Tufte (1997), p. 46, Figure (Right), with a fitted (almost) horizontal straight line. Such a figure was created after the accident to explain the poor reasoning in the pre-launch debate. (By Abbass Sharif)

- What went wrong?
 - **The data were graphed and analyzed out of context**
 - The dataset of O-ring damages was very small (7 cases only)
- What could have been done?
 - Use the complete dataset, and don't include only O-ring failure cases
 - Extend the “Number of Incidents” axis limit to 6
 - Fit a non-linear curve

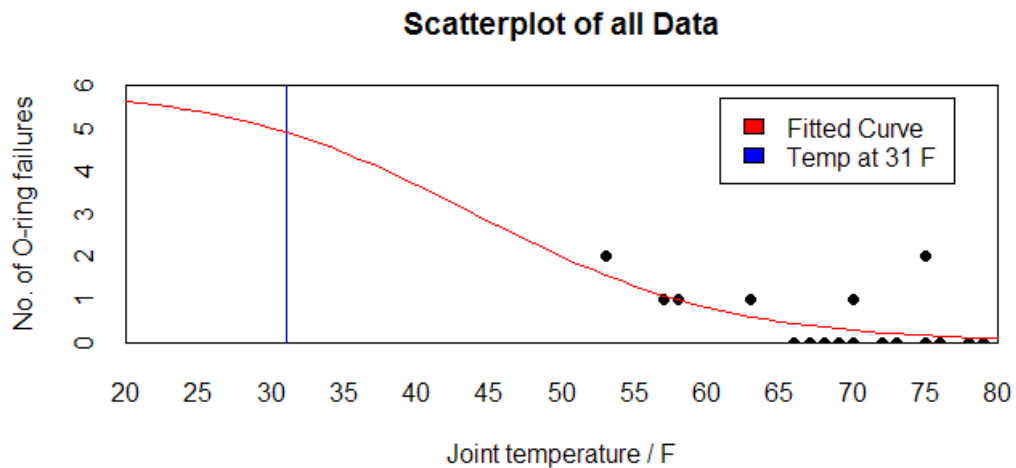


Figure 70: Reconstruction of a figure similar to Tufte (1997), p. 45, with a fitted non-linear curve. (By Abbass Sharif)

Further Reading:

- “Challenger Disaster Live on CNN”, Video posted at <http://youtube.com: http://www.youtube.com/watch?v=j4J0jcDFtBE>
- Tappin (1994)
- Tufte (1997), pp. 27 & 38–53
- Wainer (1997), pp. 51–53
- R code to reproduce Figures 69 and 70:
http://www.math.usu.edu/~symanzik/teaching/2009_stat6560/RDataAndScripts/sharif_abbass_project1_challenger.R

2.4 Further Reading

Additional sources for the history of statistical graphics, selected people, and events in history are:

- Friendly (2005) (3/15/2006 preprint available at <http://www.math.yorku.ca/SCS/Papers/gfkl.pdf>)
- Friendly (2008) (3/21/2006 preprint available at <http://www.math.yorku.ca/SCS/Papers/hbook.pdf>)
- Wainer (2009*a*)

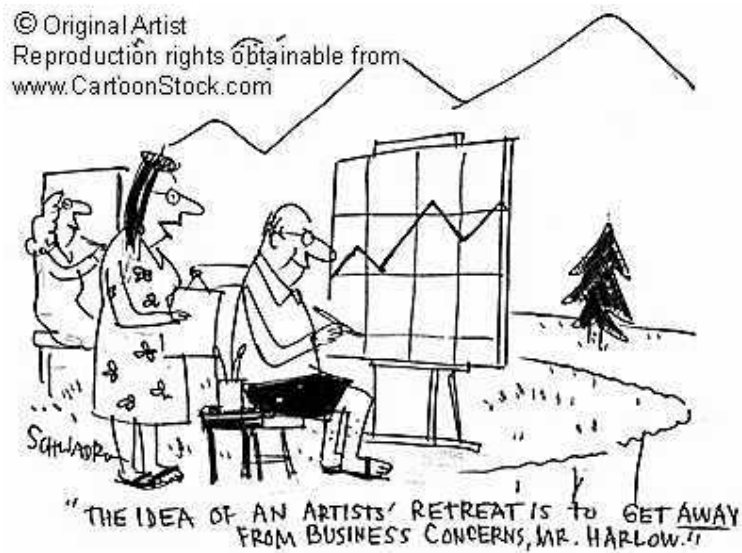


Figure 71: http://www.cartoonstock.com/blowup_stock.asp?imageref=hsc3714&artist=Schwadron,+Harley&topic=statistics+, Cartoon.

3 Color and Cognition

3.1 Color-Deficiency and Color-Blindness

Tufte (1983), p. 183, states:

“There are many specific differences between friendly and unfriendly graphics. [...]

Friendly: colors, if used, are chosen so that the color-deficient and color-blind (5 to 10 percent of viewers) can make sense of the graphic (blue can be distinguished from other colors by most color-deficient people). [...]

Unfriendly: design insensitive to color-deficient viewers; red and green used for essential contrasts. [...]

Question: Which numbers and letters can you see hidden in each figure?



#ADAM

Figure 72: Illustration of various tests for color blindness. Figure taken from <http://www.nlm.nih.gov/medlineplus/colorblindness.html> on 2/3/2009.

Definitions:

“**Color blindness**, a color vision deficiency, is the inability to perceive differences between some of the colors that others can distinguish. It is most often of genetic nature, but may also occur because of eye, nerve, or brain damage, or due to exposure to certain chemicals. The English chemist John Dalton published the first scientific paper on the subject in 1798, “Extraordinary facts relating to the vision of colours”, after the realization of his own color blindness; because of Dalton’s work, the condition is sometimes called Daltonism, although this term is now used for a type of color blindness called deuteranopia. [...]

Dichromacy: Protanopes, deuteranopes, and tritanopes are dichromats; that is, they can match any color they see with some mixture of just two spectral lights (whereas normally humans are trichromats and require three lights). These individuals normally know they have a color vision problem and it can affect their lives on a daily basis. Protanopes and deuteranopes see no perceptible difference between red, orange, yellow, and green. All these colors that seem so different to the normal viewer appear to be the same color for this two percent of the population. [...]

Protanopia (1% of males): Lacking the long-wavelength sensitive retinal cones, those with this condition are unable to distinguish between colors in the green-yellow-red section of the spectrum. [...]

Deuteranopia (1% of males): Lacking the medium-wavelength cones, those affected are again unable to distinguish between colors in the green-yellow-red section of the spectrum. [...]

Tritanopia (less than 1% of males and females): Lacking the short-wavelength cones, those affected are unable to distinguish between the colors in the blue-yellow section of the spectrum. [...]

Anomalous trichromacy: Those with protanomaly, deuteranomaly, or tritanomaly are trichromats, but the color matches they make differ from the normal. They are called anomalous trichromats. [...]

Protanomaly (1% of males, 0.01% of females): Having a mutated form of the long-wavelength (red) pigment, whose peak sensitivity is at a shorter wavelength than in the normal retina, protanomalous individuals are less sensitive to red light than normal. [...] This causes reds to reduce in intensity to the point where they can be mistaken for black. [...]

Deuteranomaly (most common — 6% of males, 0.4% of females): Having a mutated form of the medium-wavelength (green) pigment. [...] This is the most common form of color blindness, making up about 6% of the male population. The deuteranomalous person is considered “green weak”. For example, in the evening, dark green cars appear to be black to Deuteranomalous people. Similar to the protanomates, deuteranomates are poor at discriminating small differences in hues in the red, orange, yellow, green region of the spectrum.[...]

Tritanomaly (equally rare for males and females): Having a mutated form of the short-wavelength (blue) pigment. [...]

(Definitions taken from http://en.wikipedia.org/wiki/Color_blindness on 2/3/2009.)

See <http://www.nlm.nih.gov/medlineplus/colorblindness.html> for additional information and links.

Task: Start with the R code at http://www.math.usu.edu/~symanzik/teaching/2011_stat6560/RDataAndScripts/Ch3_TestColors.R and modify the colors. Save your figure as a jpg and run it through the different simulations for color vision deficiencies accessible at <http://www.vischeck.com/vischeck/vischeckImage.php>. After having tested a few options, can you suggest what may work well for all viewers (and what may not work well for some viewers)?

3.2 Various Color Spaces

(Based on Kosslyn (1994), Chapter 7, Kosslyn (2006), Chapter 7 & Few (2004), Chapter 2)

3.2.1 The HSL and HSV Color Spaces

Color is not a single entity, but it can be broken down into three components:

Hue (H): This is what we usually mean by color. It is the qualitative aspect which depends on the wavelength of the light (from long at the red end of the spectrum to short at the violet end).

Saturation (S): This is the deepness of the color (which can be varied by the amount of white that is added).

Lightness (L), Value (V), Intensity (I), or Brightness (B): This is the amount of light that is reflected (if shown on a printed page) or that is emitted (if the display is projected from a slide or shown on a computer screen). In either case, intensity can be varied by the amount of gray that is added.

Color encoded via hue, saturation, and lightness is called a HSL color space. Equivalent color spaces are based on hue, saturation, and intensity (HSI) or on hue, saturation, and brightness (HSB). Still using the idea of adding gray as the third component, but using a different encoding results in the hue, saturation, and value (HSV) color space. Sometimes, the last two letters of the color space are swapped, so we may speak of a HLS (instead of a HSL) or a HVS (instead of HSV) color space.

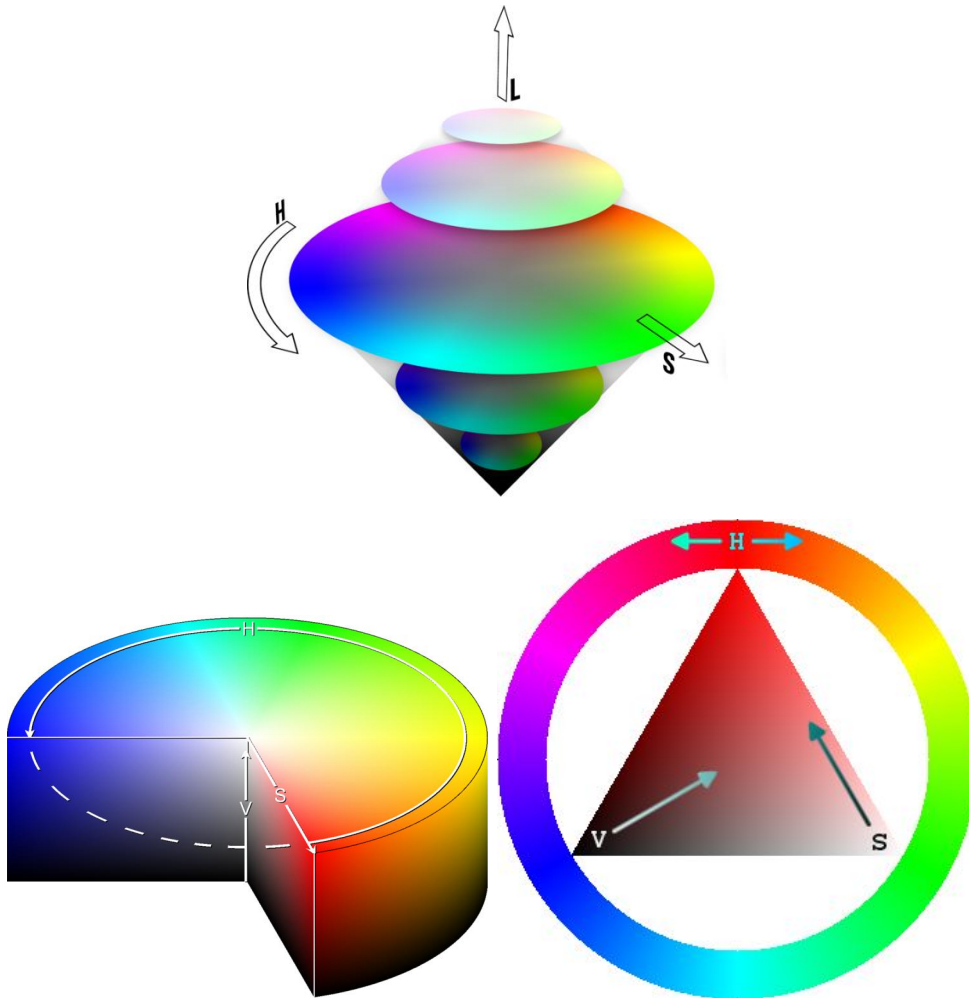


Figure 74: Illustration of the HSL (top) and HSV (bottom) color spaces. Figures taken from http://en.wikipedia.org/wiki/File:Color_cones.png, http://en.wikipedia.org/wiki/File:HSV_cylinder.png and http://en.wikipedia.org/wiki/File:Triangulo_HSV.png on 2/5/2009 (top) and 2/3/2009 (bottom).

3.2.2 The RGB Color Space

The **RGB color model** is an additive color model in which red, green, and blue light are added together in various ways to reproduce a broad array of colors. The name of the model comes from the initials of the three additive primary colors, red (R), green (G), and blue (B).

(Definition taken from http://en.wikipedia.org/wiki/RGB_color_model on 2/3/2009.)

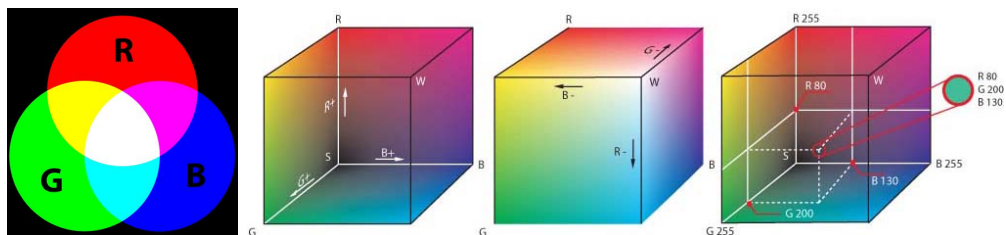


Figure 75: Illustration of the RGB color model and color space. Figures taken from <http://en.wikipedia.org/wiki/File:AdditiveColor.svg> and http://en.wikipedia.org/wiki/File:RGB_farbwuerfel.jpg on 2/3/2009.

In many programming languages including R, the RGB color space is being used as the primary (and easiest to use) color space. Values for each of the three components (R, G, and B) can originate from the discrete set $\{0, \dots, 255\}$ or the continuous interval $[0 \dots 1]$.

Some main color combinations in RGB are:

Color name	Red	Green	Blue	Hexadecimal
Black	0	0	0	#000000
White	255	255	255	#FFFFFF
Red	255	0	0	#FF0000
Green	0	255	0	#00FF00
Blue	0	0	255	#0000FF
Yellow	255	255	0	#FFFF00

Task: Start with the R code at http://www.math.usu.edu/~symanzik/teaching/2011_stat6560/RDataAndScripts/Ch3_RGBColors.R where we vary the red (R) component in a linear way. Can we clearly distinguish among the 10 different colors? If not, about how many visually distinct colors can you easily identify?

Repeat the same for the green (G) and the blue (B) component. How many visually distinct colors can you easily identify for these?

In contrast, repeat the same with a variation of gray levels where each gray level can be obtained by tripling the same value x , i.e., (x, x, x) . Is this better? Let's see whether we can obtain even some further improvement later on ...

3.2.3 The HCL Color Space

This color space consists of the following three components:

Hue (H): As before, this component describes the dominant wavelength.

Chroma (C): This component describes the colorfulness, i.e., the intensity of the color as compared to gray.

Luminance (L): This component relates to brightness, i.e., the amount of gray.

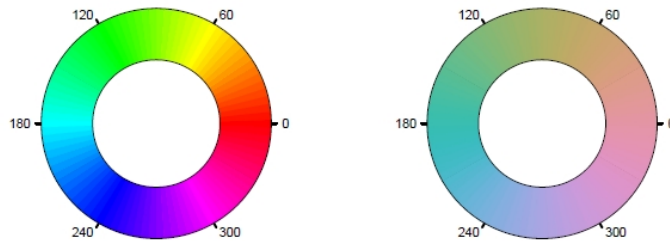


Figure 3: HSV-based and HCL-based color wheel.

Figure 76: Zeileis et al. (2009), p. 6, Figure 3.

3.3 Suggestions for Color Selections

(Based on Kosslyn (1994), Chapter 7 & Kosslyn (2006), Chapter 7)

Similar to our lists in Chapter 1 how to create bad and good graphics, there exist suggestions how to use color in graphics. Below the suggestions from Kosslyn (2006), Chapter 7:

- Use colors that are well separated in the spectrum.
- Make adjacent colors have different brightness.

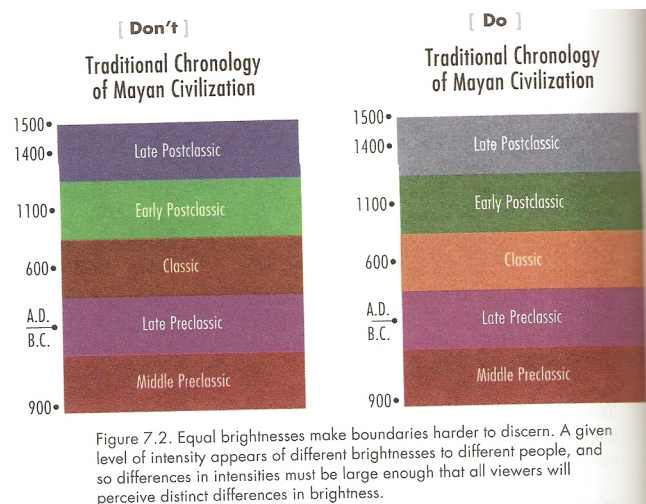


Figure 77: Kosslyn (2006), p. 160, Figure 7.2.

- Make the most important content element the most salient.

- Use warm colors to define a foreground.

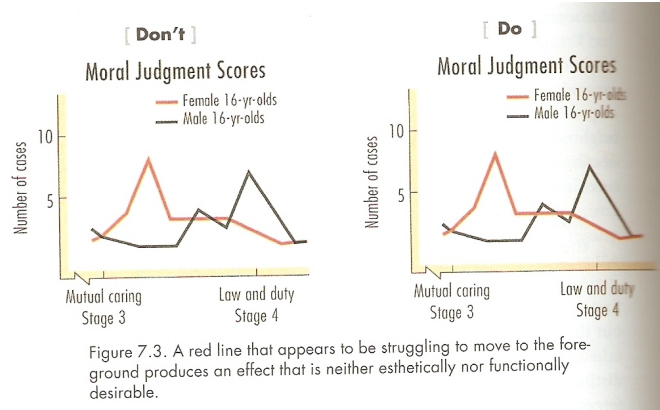


Figure 78: Kosslyn (2006), p. 162, Figure 7.3.

- Avoid using red and blue in adjacent regions.

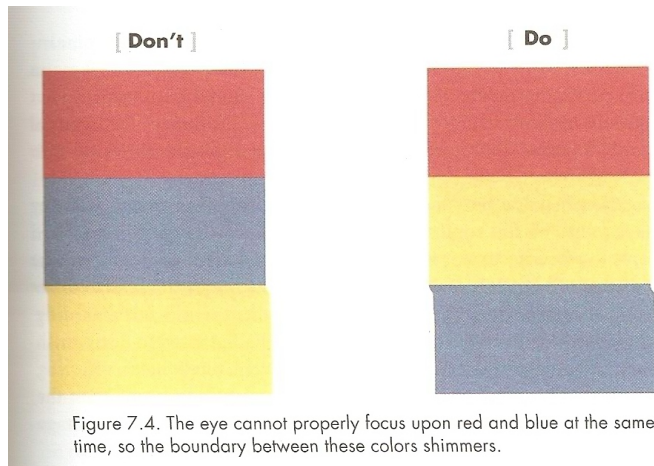


Figure 79: Kosslyn (2006), p. 163, Figure 7.4.

- Respect compatibility and conventions of color.

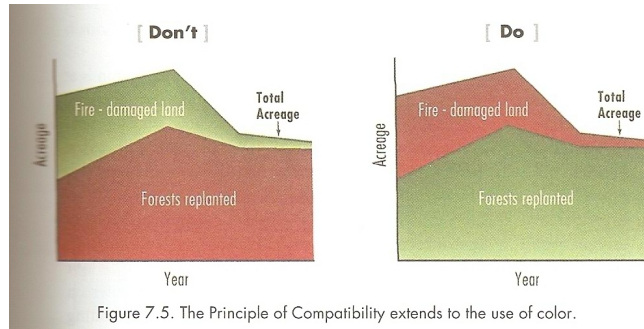


Figure 80: Kosslyn (2006), p. 163, Figure 7.5.

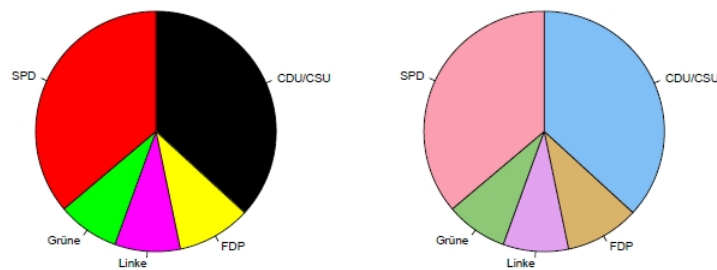


Figure 10: Seats in the German parliament.

Figure 81: Zeileis et al. (2009), p. 12, Figure 10: While HCL-based color palattes are a worthwhile alternative to standard RGB color choices, the right part of this figure clearly should be labeled “[Don’t]. The original colors (black, red, yellow, green, and purple) are so strongly associated with the German parties that they shouldn’t be changed. Imagine that someone would change the tradional red–blue US election map to pink–cyan, or, even worse, that someone would change the colors of traffic lights from red–yellow–green to something different because that is more appealing to the viewer — unthinkable.

- Use color to group elements.

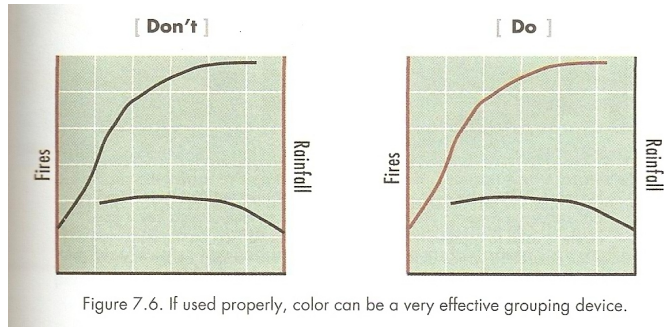


Figure 82: Kosslyn (2006), p. 165, Figure 7.6.

- Avoid using blue if the display is to be photocopied.
- Avoid using hue to represent quantitative information.
- Use deeper saturations and greater intensities for hues that indicate greater amounts.

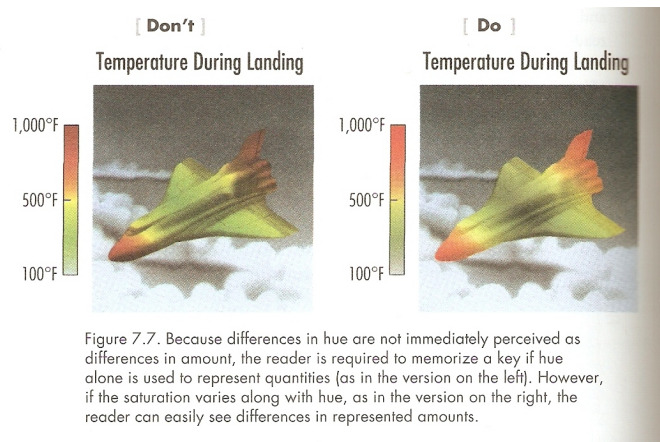


Figure 83: Kosslyn (2006), p. 166, Figure 7.7.

- Do not use hue, saturation, and intensity to specify different measurements.

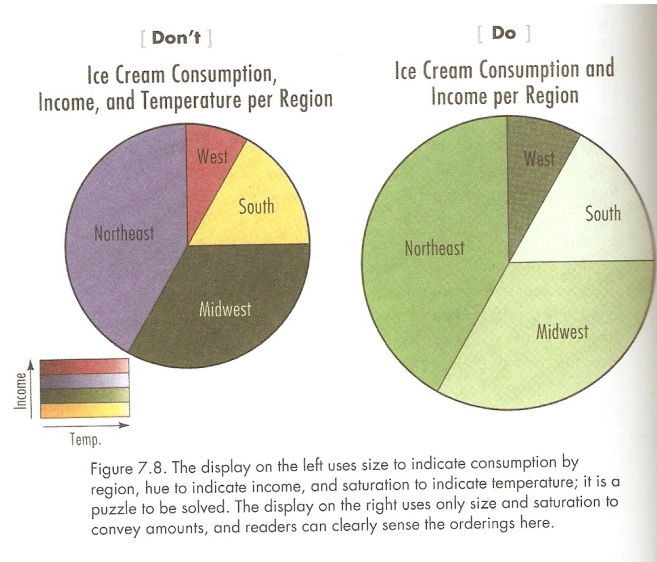


Figure 84: Kosslyn (2006), p. 168, Figure 7.8.

3.4 Good Color Choices

Compare Figures 85 and 86, taken from Tufte (1997). Both use 21 distinct colors to communicate altitude and ocean depth. Which is better? — Why?

76 VISUAL EXPLANATIONS

Showing the Japan Sea and the great trenches of the western Pacific, this classic map below makes extraordinary use of small and effective differences. The General Bathymetric Chart of the Oceans depicts depth (the blue, bathymetric tints) and altitude (tan, hypsometric tints) in 21 color gradations—with “the deeper or the higher, the darker the color” serving as the visual metaphor for the color scale. To indicate depth, the contour lines are labeled by numbers, a design that enhances accuracy of reading and nearly eliminates any need to refer back to the legend. Every color tint on the map signals four variables: latitude, longitude, sea or land, and depth or altitude measured in meters. Then, on a visual layer separated from the blue tints, thin gray lines trace out the routes of the oceanographic ships that measured the depth (outside of areas with detailed surveys, such as ports and coastlines).

These gray lines are a small miracle of information design. Floating on top of the ocean and coexisting with the blue tints and contours, the thin lines depict a distinct, second layer of data relevant to the depths below. There is sufficient visual space for the gray lines because the representation of depth does not use up all the informational possibilities of color in the map. And since the contours are directly

General Bathymetric Chart of the Oceans,
International Hydrographic Organization
(Ottawa, Canada, 5th edition, 1984). 5.06.

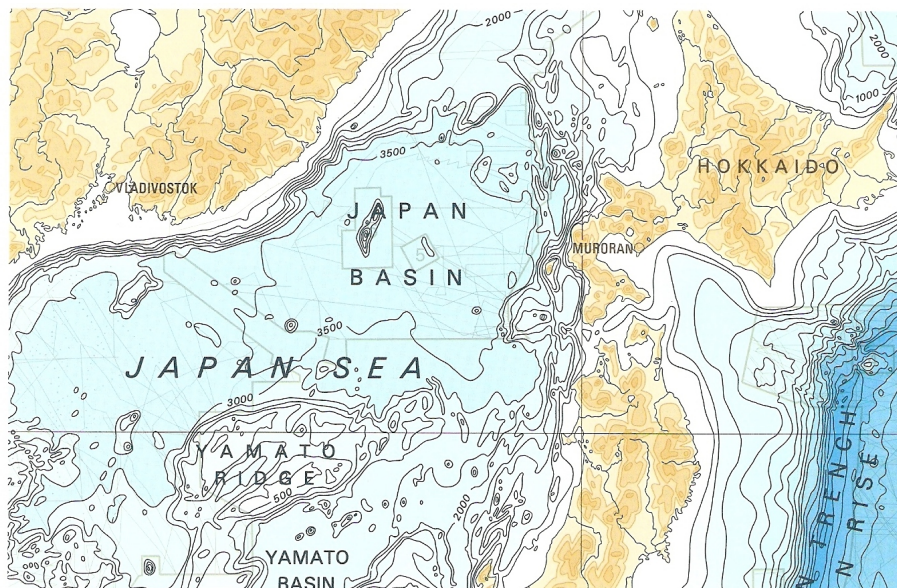
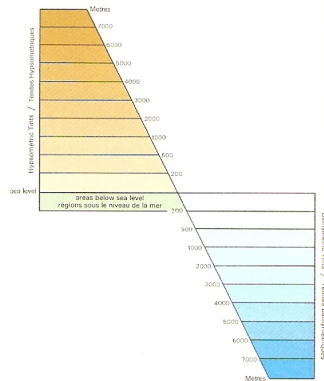


Figure 85: Tufte (1997), p. 76, Figure.

labeled with numbers, the fine distinctions in blue remain clear and readable. By indicating depth with visually minimal gradations in color, the cartographers were able to add an extra two-dimensional layer of gray-line data right on top of the ocean contours. Minimal differences allow more differences.

In ghastly contrast below, a rainbow encodes depth. Although often found in scientific publications, such a visually naive color-scale would be laughed right out of the field (or ocean) of cartography. These aggressive colors, so unnatural and unquantitative, render the map incoherent, with some of the original data now lost in the soup.

Minimal distinctions reduce visual clutter. Small contrasts work to enrich the overall visual signal by increasing the number of distinctions that can be made within a single image; thus design by means of small effective differences helps to increase the resolution of our images. In practice, the appropriate size of small contrasts will depend on the context, priority of particular elements in the overall visual story, number of differentiations made within an image, and characteristics of those viewing the image. Despite these local complications, the global principle of the smallest effective difference resolves many visual issues—serving perhaps even as an algorithm for automated design.

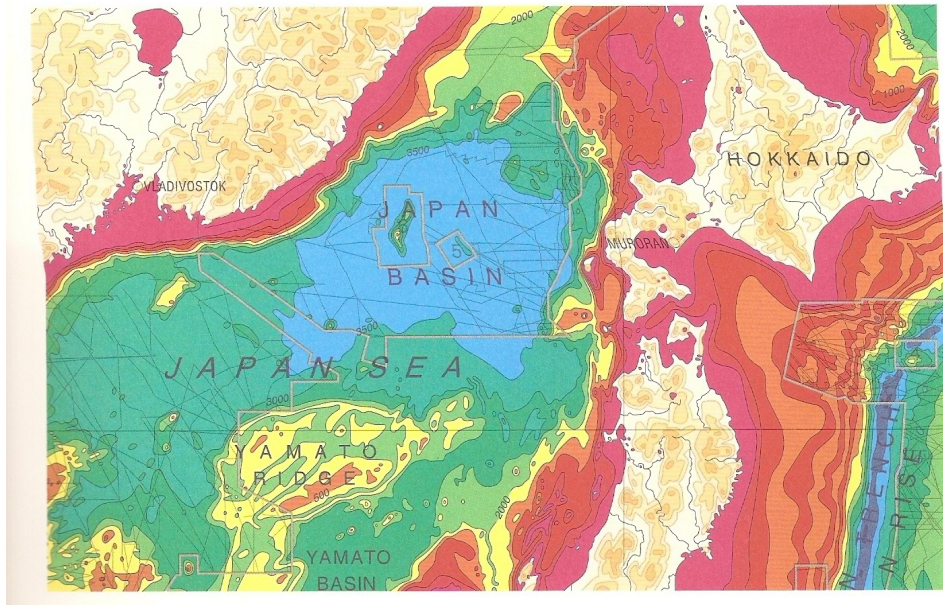
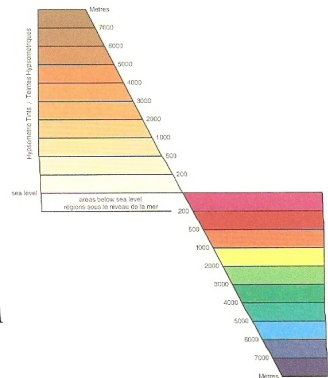


Figure 86: Tufte (1997), p. 77, Figure.

3.4.1 Work by Cindy Brewer and Collaborators

Extensive work regarding the use of color, in particular on maps, was done by Cindy Brewer and her collaborators. Examples of her work include:

- Brewer (1997)
- Brewer et al. (1997)
- Brewer (1999)
- Brewer & Pickle (2002)

A resulting software tool, *ColorBrewer* is described in:

- Leslie (2002), a brief independent review (online version available at <http://www.sciencemag.org/cgi/reprint/296/5567/435c>)
- Harrower & Brewer (2003)
- Brewer et al. (2003)
- Brewer (2003)

ColorBrewer is accessible at <http://colorbrewer2.org> and it is described as:

“ColorBrewer is an online tool designed to help people select good color schemes for maps and other graphics. It is free to use, although we’d appreciate it if you could cite us if you decide to use one of our color schemes.”

Main Color Schemes in <http://colorbrewer2.org> are:

sequential: best suited for ordered data that progress from low to high

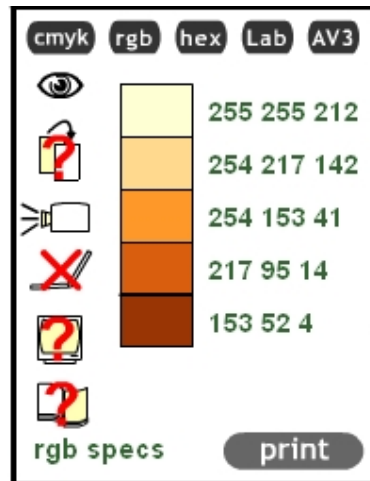


Figure 87: <http://ColorBrewer.org>: 5-class sequential YlOrBr.

diverging: equal emphasis on mid-range critical values and extremes at both ends of the data range

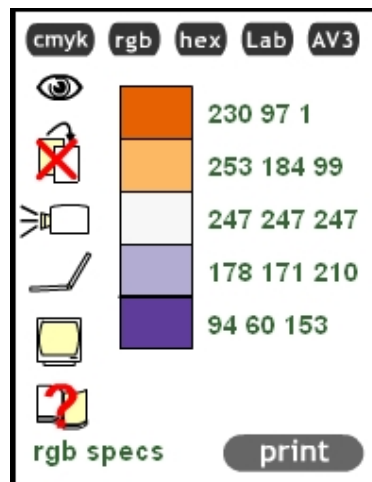


Figure 88: <http://ColorBrewer.org>: 5-class diverging PuOr.

qualitative: no difference in magnitude between legend classes; hues are used to create the primary visual differences; best suited for categorical data

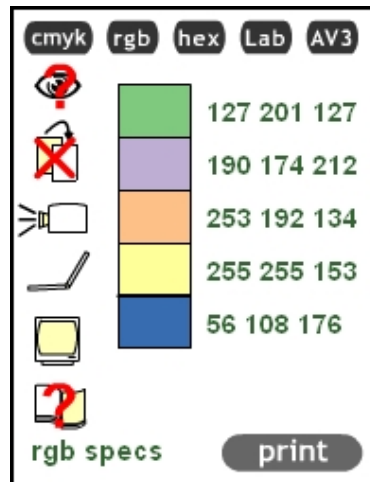


Figure 89: <http://ColorBrewer.org>: 5-class qualitative Accents.

The related R package *RColorBrewer* is documented at <http://cran.r-project.org/web/packages/RColorBrewer/index.html>. First run the examples on pages 3 and 4 of the reference manual (<http://cran.r-project.org/web/packages/RColorBrewer/RColorBrewer.pdf>), then further experiment with these color palettes.

Ultimately, when you work with color palettes from *ColorBrewer* or *RColorBrewer*, do not forget to cite the source of the color palettes you use. See http://www.personal.psu.edu/cab38/ColorBrewer/ColorBrewer_updates.html, 5. Citation, for details.

3.4.2 Work by Zeileis, Hornik, and Murrell

In a recent paper, Zeileis et al. (2009) suggest to work with HCL color palettes instead of HSV or RGB palettes.

They use the same distinction among color palettes as in <http://colorbrewer2.org>:

Qualitative Palettes:

Sequential Palettes:

Diverging Palettes:

The related R package *colorspace* is documented at <http://cran.r-project.org/web/packages/colorspace/index.html>. First run the examples on pages 14 and 15 in the reference manual (<http://cran.r-project.org/web/packages/colorspace/colorspace.pdf>) that are related to *rainbow_hcl*, then further experiment with these color palettes.

3.5 Change Blindness

(Based on a Student Project by Ying Jin in Spring 2009)

Simons & Rensink (2005), p. 16, state:

“The term ‘change blindness’ refers to the surprising difficulty observers have in noticing large changes to visual scenes.”

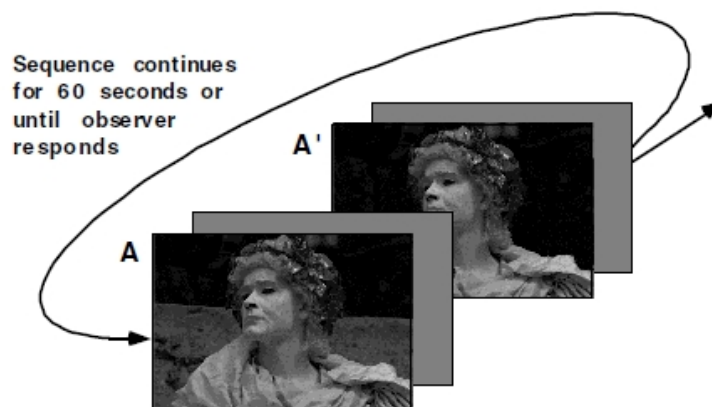


Figure 90: Rensink (2006), p. 2417, Figure 5: Change blindness. Original image A (statue with wall in background) and modified image A' (statue with wall gone) are displayed in the order A, A', A, A', ... with gray fields placed between successive images. Observers typically require several seconds to see such a change, even though it is large.

Worksheet

Your Name: _____

Viewable/Downloadable Examples of Change Blindness:

<http://www2.psych.ubc.ca/~rensink/flicker/download/>

You will be allowed to watch each of the following sequences for exactly 30 seconds. It is your task to describe in a few words what is different between the two images that are shown in each sequence.

1. Airplane: _____

2. Chopper & Truck: _____

3. Dinner: _____

4. Farm: _____

5. Harborside: _____

6. Market: _____

7. Money: _____

8. Sailboats: _____

9. Street Corner: _____

10. Tourists: _____

Causes of Change Blindness:

- The impaired localization of the motion signals that accompanies the change, which means the motion signals failed to draw attention (i.e., eye blinks, brief occlusions, or changes made gradually).
- Changes are not expected (i.e., inattentional blindness).
- The limited size for visual short-term memory. Normally, attention can only be distributed to 4 to 5 items at a time. Graphs containing large amounts of information would result in the failure of change detection.

Implications for Designs of Efficient Graphics:

The following are guidelines and principles to avoid change blindness and to create efficient graphics, adapted from Rensink (2006):

- Principle of limited information. Attention capacity is limited to 4 to 5 items at a time.
- Principle of maximal static representation. Motion signals are easy to lose.
- Principle of minimal motion. If motion is necessary, use motions as less as possible.
- Principle of single dynamic source. Use at most one dynamic motion source at a time.

3.6 Further Reading

Additional sources for the use of colors in statistical graphics are:

- Tufte (1990), Chapter 5



He then drew a number of smaller pie charts behind the bigger chart. That helped to put it into perspective.

Figure 91: http://www.cartoonstock.com/blowup_stock.asp?imageref=mban818&artist=Baldwin,+Mike&topic=statistics+, Cartoon.

4 Statistical Maps

4.1 Choropleth Maps

Symanzik & Carr (2008), p. 270, state:

“Choropleth maps use the color or shading of regions in a map to represent region values. Choropleth maps have proved very popular but have many problems and limitations as indicated by writers such as Robinson et al. (1978), Dent (1993), and Harris (1999). [. . .]

There are two kinds of choropleth maps, called unclassed and classed. Unclassed maps use a continuous color scale to encode continuous values (statistics). This is problematic because perception of color is relative to neighboring colors and because color has poor perceptual accuracy of extraction in a continuous context. Classed choropleth maps ameliorate this problem and dominate in the literature.

Classed choropleth maps use class intervals to convert continuous estimates into an ordered variable with a few values that can be represented using a few colors. When a few colors are easily discriminated and regions are sufficiently large for color perception, color identification problems are minimal. The color scheme also needs to convey the class ordering based on values. Brewer (1997) and Brewer et al. (1997) provided results evaluating different color schemes in a mapping context. The Web site <http://colorbrewer.org> (see Leslie 2002, for a short description) contains guidance on ordered color schemes and additional issues such as suitable schemes for people with color vision deficiencies and for different media. Perfect examples on how colors should be used in choropleth maps can be found in the 1996 “Atlas of United States Mortality” (Pickle et al. 1996).”

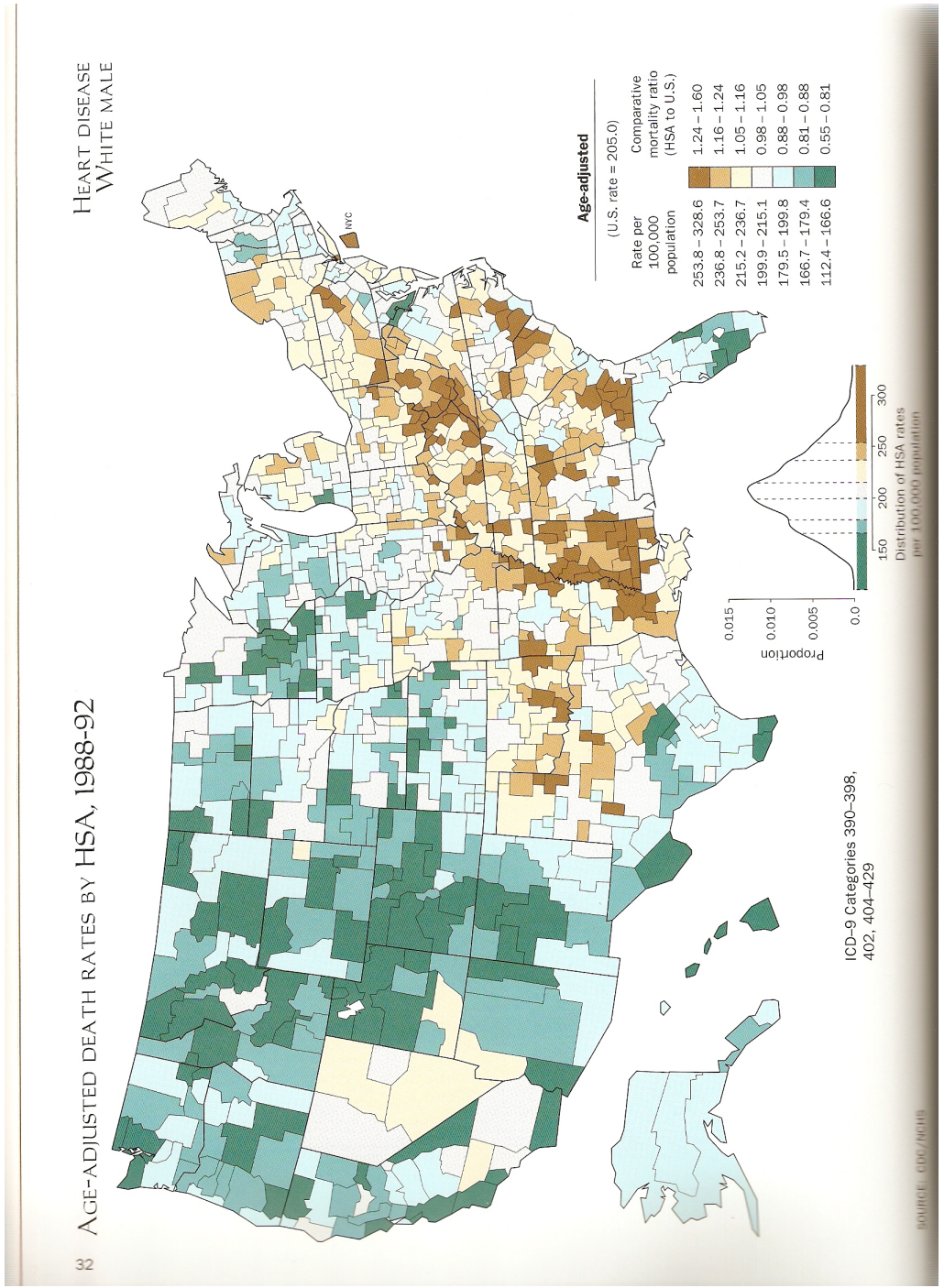


Figure 92: Pickle et al. (1996), p. 32, Figure, showing Heart Disease White Male by Health Service Area (HSA), 1988-92.

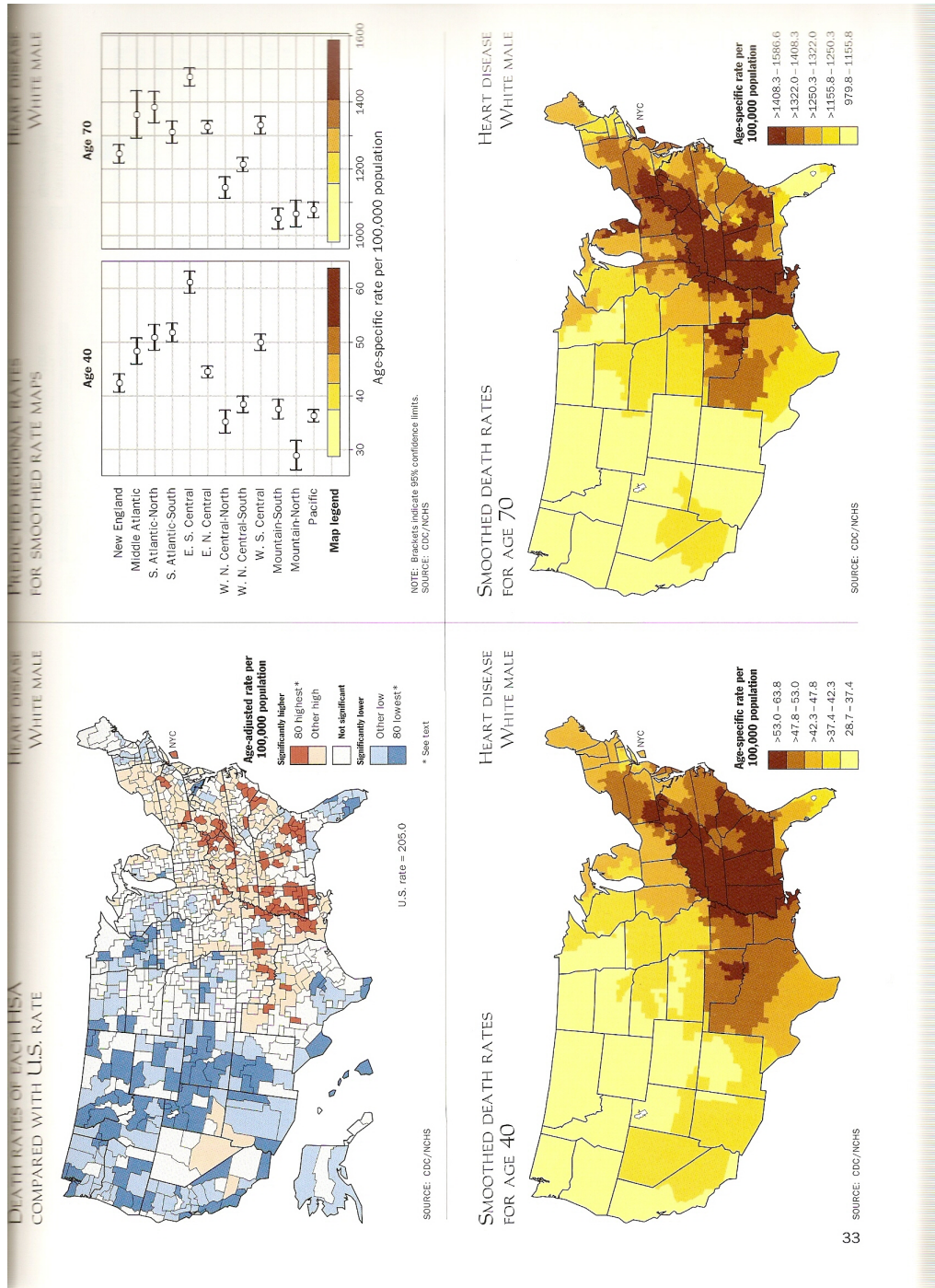


Figure 93: Pickle et al. (1996), p. 33, Additional Figures, showing Heart Disease White Male, 1988-92.

Choropleth Map Construction:

Choropleth Maps are easy to construct on paper. <http://www.teachingideas.co.uk/geography/chlormap.htm> teaches children age 5 to 11 “How to make a Choropleth Map”.

Weather Maps:

We are used to see various new choropleth maps almost daily, in newspapers, on TV, or on the Web. Probably the most popular choropleth maps are weather maps, such as provided by <http://www.usatoday.com/weather/default.htm>. Move your mouse over the various options (Radar, Satellite, Precip, Temps, Fronts) to see various weather-related choropleth maps.

Unfortunately, except for Temps, there are usually no exact values shown on choropleth maps. So, it is difficult to exactly compare different regions or cities.

4.1.1 Choropleth Maps in R

Various R packages support the creation of maps and choropleth maps in R, provide commands to read and write ESRI shapefiles, and link to various Geographic Information Systems (GIS). A more detailed overview can be found at <http://geodacenter.asu.edu/projects/rsp/content/map-packages-cr>.

Example 1:

Basic geographic maps, ranging from a world map to detailed county maps with labels and city names.

```
library(maps)

map() # World (default)

map("state") # US

map("state", c("Utah", "Colorado", "Idaho", "Wyoming", "Montana"))

data(state)
state.region
state.name[state.region == "Northeast"]
map("state", state.name[state.region == "Northeast"])

map("state")
map.axes() # add latitude (North/South) and longitude (East/West)

map("county") # US counties

map("county", "Utah") # Utah counties

map.text("state") # too many labels

map.text("state", state.name[state.region == "South"])

map("county", "Utah")
map.cities() # too many labels
```

```

map("county", "Utah", xlim = c(-113,-111), ylim = c(40,41))
map.cities()
map.axes()
points(c(-112.5, -111.5), c(40.2, 40.8), pch = c(17, 20), col = "red", cex = 1.5)
  ## fill in two additional points

map("state", state.name[state.region == "Northeast"], fill = T, col = 0:8)

library(RColorBrewer)

map("state", state.name[state.region == "Northeast"], fill = T,
  col = brewer.pal(9, "Set3"))

map("state", fill = T, col = brewer.pal(9, "Set3"))

map("state", fill = T, col = brewer.pal(5, "Blues"))

```

Example 2:

Choropleth maps.

```

library(maps)

data(state)

murder <- state.x77[,5]
range(murder)

breaks <- c(1, 4, 7, 10, 13, 16)
m.class <- cut(murder, breaks)
brewer.pal(5, "Blues")
m.col <- brewer.pal(5, "Blues")[m.class]

map.m.col <- m.col[match.map("state", state.name)]
map("state", fill = T, col = map.m.col)
legend("bottomright", legend = levels(m.class), fill = brewer.pal(5, "Blues"))

```

The next four examples incorporate google maps as background into an R graphics window. This can be done with the package *RgoogleMaps*. The main documentation for this package can be found at <http://cran.r-project.org/web/packages/RgoogleMaps/index.html>. The reference manual is located at <http://cran.r-project.org/web/packages/RgoogleMaps/RgoogleMaps.pdf> and additional examples can be found at <http://cran.r-project.org/web/packages/RgoogleMaps/vignettes/RgoogleMaps-intro.pdf>.

The following R packages are required to run the next four examples and should be first installed before running these examples: *RgoogleMaps*, *PBSmapping*, *maptools*, *ReadImages*, and *rgdal*.

The R code below originates from the Web sites above, but has been further modified in some cases. Here is an example how David Kahle, a Ph.D. Candidate in the Department of Statistics, Rice University, has used *RgoogleMaps* in combination with *ggplot2*: <https://github.com/hadley/ggplot2/wiki/Crime-in-Downtown-Houston%2C-Texas-%3A-Combining-ggplot2-and-Google-Maps>. Unfortunately, we cannot reproduce the maps as the data files are not provided.

Example 3:

Google map of Washington, D.C., as background.

```
library(RgoogleMaps)
library(PBSmapping)

source(url("http://www.math.usu.edu/~symanzik/teaching/2011_stat6560/RDataAndScripts/GetMap.R"))
# this is needed while RgoogleMaps is being updated on CRAN

shpFile <- paste(system.file(package = "RgoogleMaps"), "/shapes/bg11_d00.shp", sep = "")

shp = importShapefile(shpFile, projection = "LL")
bb <- qbbox(lat = shp[,"Y"], lon = shp[,"X"])
MyMap <- GetMap.bbox(bb$lonR, bb$latR, destfile = "DC.jpg")
PlotPolysOnStaticMap(MyMap, shp, lwd = .5, col = rgb(0.25, 0.25, 0.25, 0.025), add = F)

# add some random points to the map
x = runif(100, -0.1, 0.1) + 38.8933115
y = runif(100, -0.1, 0.1) - 77.0146475
PlotOnStaticMap(MyMap, x, y, FUN = points, col = "red", add = T)
```

Example 4:

Google satellite map of lower Manhattan, New York, NY.

```
library(RgoogleMaps)

source(url("http://www.math.usu.edu/~symanzik/teaching/2011_stat6560/RDataAndScripts/GetMap.R"))
# this is needed while RgoogleMaps is being updated on CRAN
```



```
bb <- qbbox(c(40.702147, 40.711614, 40.718217),c(-74.015794, -74.012318, -73.998284),
  TYPE = "all", margin = list(m=rep(5, 4), TYPE = c("perc", "abs")[1]))
MyMap <- GetMap.bbox(bb$lonR, bb$latR, destfile = "MyTile3.png", matype = "satellite")

PlotOnStaticMap(MyMap)
```

Example 5:

Google street map of lower Manhattan, New York, NY.

```
library(RgoogleMaps)

source(url("http://www.math.usu.edu/~symanzik/teaching/2011_stat6560/RDataAndScripts/GetMap.R"))
# this is needed while RgoogleMaps is being updated on CRAN

bb <- qbbox(c(40.702147, 40.711614, 40.718217),c(-74.015794, -74.012318, -73.998284),
  TYPE = "all", margin = list(m = rep(5, 4), TYPE = c("perc", "abs")[1]))
MyMap <- GetMap.bbox(bb$lonR, bb$latR, destfile = "MyTile3.png")

PlotOnStaticMap(MyMap)
```

Example 6:

Google street map of Washington, D.C., with markers and lines.

```
library(RgoogleMaps)

source(url("http://www.math.usu.edu/~symanzik/teaching/2011_stat6560/RDataAndScripts/GetMap.R"))
# this is needed while RgoogleMaps is being updated on CRAN

#Define the markers:
mymarkers <- cbind.data.frame(lat = c(38.898648, 38.889112, 38.880940),
  lon = c(-77.037692, -77.050273, -77.03660), size = c('small', 'small', 'small'),
  col = c('blue', 'green', 'red'), char = c('', '', ''))

#get the bounding box:
bb <- qbbox(lat = mymarkers[, "lat"], lon = mymarkers[, "lon"])

#download the map:
MyMap <- GetMap.bbox(bb$lonR, bb$latR, destfile = "DC.png", GRAYSCALE = T,
  markers = mymarkers)

#determine the max zoom, so that all points fit on the plot
# (not necessary in this case):
#zoom <- min(MaxZoom(latrange=bb$latR, lonrange=bb$lonR))

#plot:
library(RColorBrewer)
pal <- brewer.pal(3, "Set2")

tmp <- PlotOnStaticMap(MyMap, lat = mymarkers[, "lat"], lon = mymarkers[, "lon"],
  cex=1.5, pch = 20, col = pal, add=F)
tmp <- PlotOnStaticMap(MyMap, lat = mymarkers[, "lat"], lon = mymarkers[, "lon"],
  col = c('purple'), add = T, FUN = lines, lwd = 2)
```

4.2 Linked Micromaps

Symanzik & Carr (2008), p. 268, state:

“Over the last decade, researchers have developed many improvements to make statistical graphics more accessible to the general public. These improvements include making statistical summaries more visual and providing more information at the same time. Research in this area involved converting statistical tables into plots (Carr 1994, Carr & Nusser 1995), new ways of displaying geographically referenced data (Carr et al. 1992), and, in particular, the development of linked micromap (LM) plots, often simply called micromaps (Carr & Pierson 1996, Carr et al. 1998, Carr, Olsen, Pierson & Courbois 2000). LM plots, initially called map row plots as well as linked map–attribute graphics, were first presented in a poster session sponsored by the American Statistical Association (ASA) Section on Statistical Graphics at the 1996 Joint Statistical Meetings (Olsen et al. 1996). More details on the history of LM plots and their connection to other research can be found in these early references on micromaps. More recent references on LM plots (Carr, Wallin & Carr 2000, Carr 2001) focused on their use for communicating summary data from health and environmental studies.”

A comprehensive source for micromaps is Carr & Pickle (2010). This book deals with (ordinary) linked micromaps, conditioned micromaps, and comparative micromaps. Three accompanying Web pages at <http://www.crcpress.com/product/isbn/9781420075731>, <http://www.statnetconsulting.com/micromaps.html>, and <http://mason.gmu.edu/~dcarr/Micromaps/> provide access to numerous data sets, shape files, and R code used in the book.

4.2.1 Template for LM Plots

Gebreab et al. (2008), pp. 112–113, state:

“A typical template of a LM plot consists of four key features (Carr & Pierson 1996). Figure 1 shows a hypothetical LM plot. The first feature is **three or more sequence panels in parallel linked by location**. In the hypothetical case, Figure 1 shows five parallel sequences of panels. The first (leftmost) sequence of panels is the micromap panel itself that typically contains small caricatures of map outlines of a region. The caricature map maintains the shape and neighborhood relationship while making the small subregions more visible. The second (from the left) sequence of panels is the label panel that provides the names of the geographical subregions (here, Region 1 through Region 10). The third through the fifth (from the left) sequence of panels display the statistical summaries. These panels may represent many forms of statistical summaries including box-plots, dot-plots (as shown in Figure 1), time series plots, confidence intervals, etc. **Sorting the geographic subregions based on the statistical variable(s) of interest** is the second feature. Sorting improves perception between consecutive panels from the top to the bottom of the display. The third feature is the **partitioning of the regions into perceptual groups of size five or less to allow the viewer’s attention to focus on explicit areas at a time**. The fourth feature is **color and location that links corresponding elements within the parallel sequence panels**, i.e., the color red in the topmost panels relates to the geographic subregion in the northeast of the map, the subregion name (Region 5), and a red dot in each of the three statistical panels. The color red is reused in the next consecutive set of panels for Region 2, but there is no relationship between Region 5 and Region 2 as one might at first assume. Simply, there do not exist enough distinguishable colors to populate an entire display (with, say, 50 different subregions) such that colors have to be reused in different panels.

In the hypothetical Figure 1, the rows are sorted by decreasing values with respect to the statistical panel 2. The statistical data displayed in the statistical panel 1 and 2 show a strong positive association (the correlation r calculated as 0.99), expressed in the almost parallel behavior of the dots and lines representing the values for these two variables. In contrast, the statistical data in panel 3 and 1 (as well as 3 and 2) show a strong negative

association (the correlation r calculated as $\sqrt{0.94}$ for 3 and 1 and as $\sqrt{0.92}$ for 3 and 2). This negative association is seen in the movement of the dots and lines in opposite directions for these variables. Moreover, the data in panel 3 shows an unusual outlier, the value for Region 1. It is this outlier that considerably reduces the almost perfect negative association otherwise present in this data. Just a simple numerical calculation of r might not be able to reveal the influence of a single subregion on the overall relationship.

The map panels of the LM plot in Figure 1 exhibit a strong geographic pattern: Highest occurrences with respect to the statistical panel 1 and 2 can be found in the north and in the east; lowest occurrences can be found in the west and in the south. Additional features of LM plots exist and are described in more details in Symanzik & Carr (2008).”

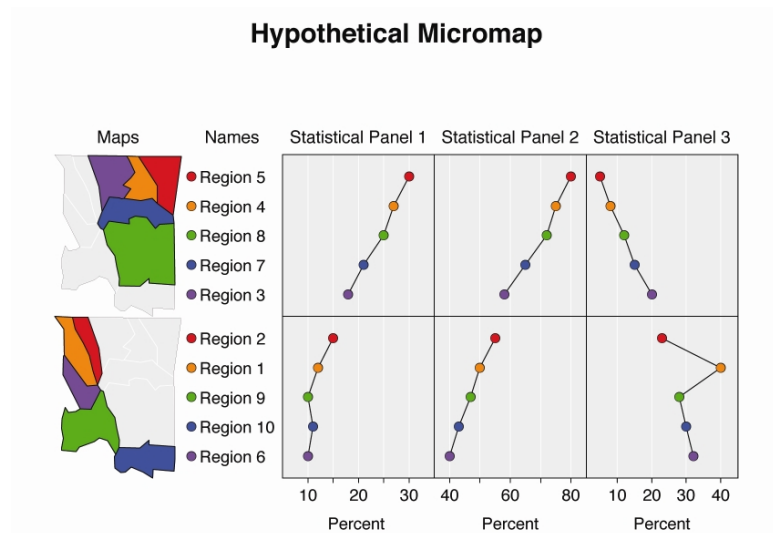


Figure 94: Gebreab et al. (2008), p. 113, Figure 1: Hypothetical LM plot illustrating the main features of such plots: the leftmost sequence of map panels, the second (from the left) sequence of label panels, and the third through the fifth (from the left) sequence of statistical panels.

4.2.2 Micromaps vs. Choropleth Maps

Symanzik & Carr (2008), pp. 270–271, state:

“LM plots often provide a good alternative to displaying statistical information using choropleth maps. Choropleth maps use the color or shading of regions in a map to represent region values. Choropleth maps have proved very popular but have many problems and limitations as indicated by writers such as Robinson et al. (1978), Dent (1993), and Harris (1999). Reviewing these problems helps to indicate why LM plots are a good alternative. [...]

Even with a good color scheme, three key problems remain for classed choropleth maps. **The first problem relates to region area. As suggested above, some map regions can be too small to effectively show color.** Examples include Washington, D.C., on a map of the United States (U.S.) and Luxembourg on a European map. Map caricatures, such as Monmonier’s state visibility map (Monmonier 1993), can address this problem, by enlarging small regions in a way that maintains region identifiability and shows each region touching the actual neighboring regions. Another facet of the area problem is that large areas have a strong visual impact while in many situations, such as in the mapping of mortality rates, the interpretation should be weighted by the region population. Dorling (1995) addressed this problem by constructing cartograms that changed region shapes to make areas proportional to population. Issues related to this approach are region identifiability, and map instability over time as their shapes change with changing populations. Area related problems persist in choropleth maps.

A second key problem is that converting a continuous variable into a variable with a few ordered values results in an immediate loss of information. This loss includes the relative ranks of regions whose distinct values become encoded with the same value. The task of controlling conversion loss has spawned numerous papers about proposing methods for defining class intervals. Today, guidance is available based on usability studies. Brewer & Pickle (2002) indicated that quintile classes (roughly 20% of regions in each class) tend to perform better than other class interval methods when evaluated across three different map reading tasks. Still, the best of the class interval selection approaches loses information.

The third key problem is that it is difficult to show more than one variable in a choropleth map. MacEachren et al. (1995) and MacEachren et al. (1998) were able to clearly communicate values of a second binary variable (and indicator of estimate reliability) by plotting black and white stripped texture on regions with uncertain estimates. However, more general attempts such as using bivariate colors schemes have been less successful (Wainer & Francolini 1980). Thus, choropleth maps are not suitable for showing estimate standard errors and confidence bounds that result from the application of sound statistical sampling or description. It is possible to use more than one map to show additional variables. However, Monmonier (1996, page 154) observed that when plotting choropleth maps side by side it can easily happen that “similarity among large areas can distort visual estimates of correlation by masking significant dissimilarity among small areas.” The change blindness (Palmer 1999, page 538) that occurs as the human eyes jump from one map to another map makes it difficult to become aware of all the differences that exist in multiple choropleth maps and hard to mentally integrate information in a multivariate context.”

Symanzik & Carr (2008), pp. 272–274, provide the following motivational example:

“Fig. 95 shows two variables, the soybean yield and acreage from the 1997 Census of Agriculture for the United States, displayed in two choropleth maps. Five equal size class intervals were chosen for each of the maps. [...]

The two choropleth maps in Fig. 95 indicate that highest yields and highest acreages for soybeans occur in the Midwest. There seems to be some spatial trend, i.e., some steady decrease for both variables from the Midwest to the Southeast. Overall, there appears to be a positive correlation between these two variables since high yields/high acreages and low yields/low acreages seem to appear in the same geographic regions. The correlation coefficient between yield and acreage is only 0.64, suggesting departures from linearity that would be better revealed using scatterplots or LM plots. [...]

In fact, Fig. 96 shows the LM plots of the same two variables as Fig. 95, plus a third statistical panel for the variable production. Data is available for 31 of the 50 U.S. states only. An identical color links all of the descriptors for a region. Successive perceptual groups use the same set of distinct colors. In Fig. 96, the sorting is done (from largest to smallest) by soybean yield in

those 31 U.S. states where soybeans were planted. Here, the points within a panel are connected to guide the viewer's eyes and not to imply that interpolation is permitted. The connecting lines are a design option and can be omitted to avoid controversy or suit personal preferences. The list of 31 U.S. states is not evenly divisible by five. Two perceptual groups at the top and two groups at the bottom contain four states, while three perceptual groups in the middle contain five states. The middle groups require the use of a fifth linking color. Using distinct hues for linking colors works best in full color plots. For grey-level plots, colors need to be distinct in terms of grey-level. Fig. 96 shows different shades of green and is suitable for production as a grey-level plot. Readers new to LM plots sometimes try to compare regions with the same color across the different perceptual groups, but quickly learn the linkage is meaningful only within a perceptual group."

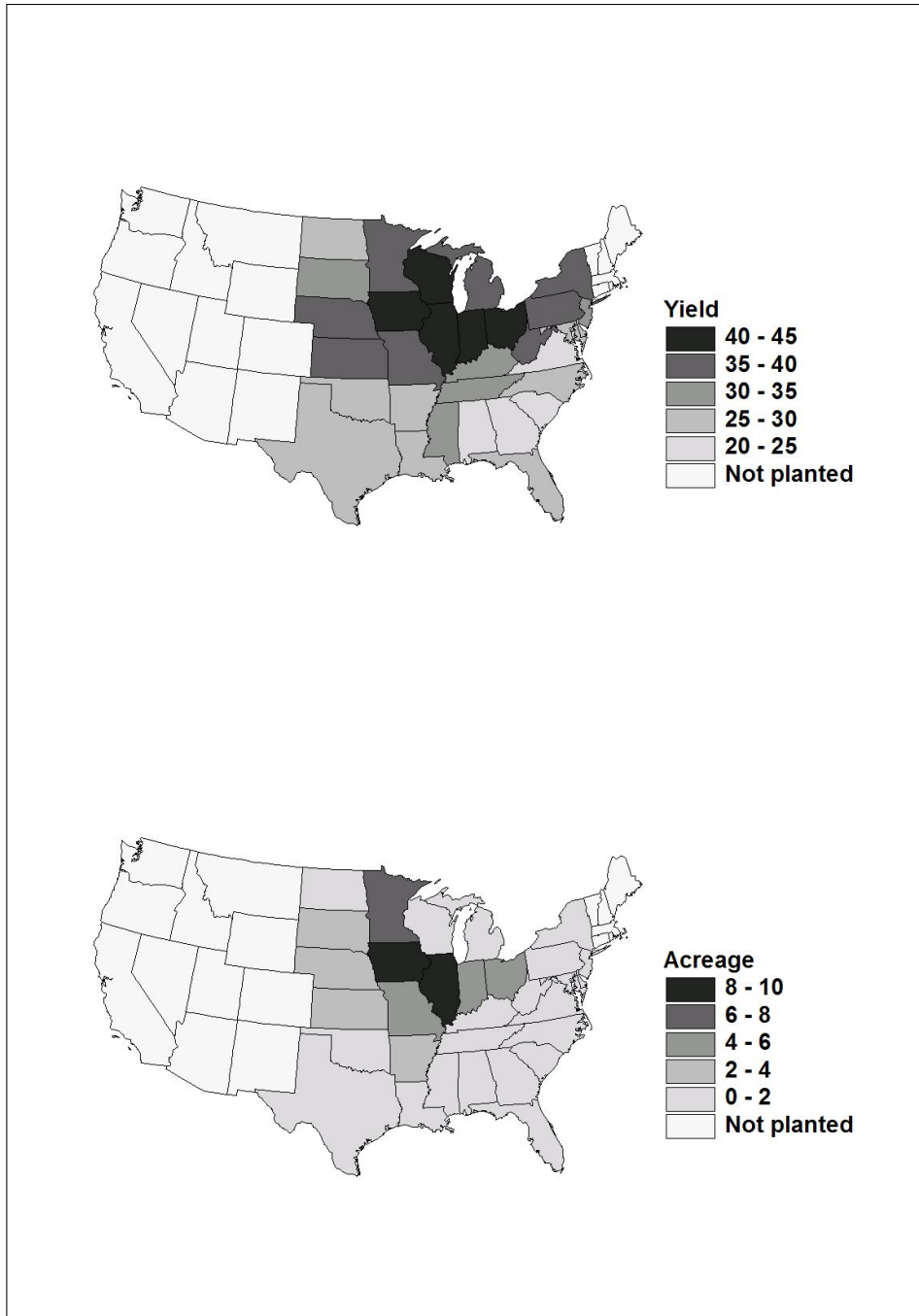


Figure 95: Symanzik & Carr (2008), p. 273, Figure 1.2: Choropleth maps of the 1997 Census of Agriculture, showing the variables soybean yield (in bushels per acre) and acreage (in millions of acres) by state. The data represent the 31 U.S. states where soybeans were planted.

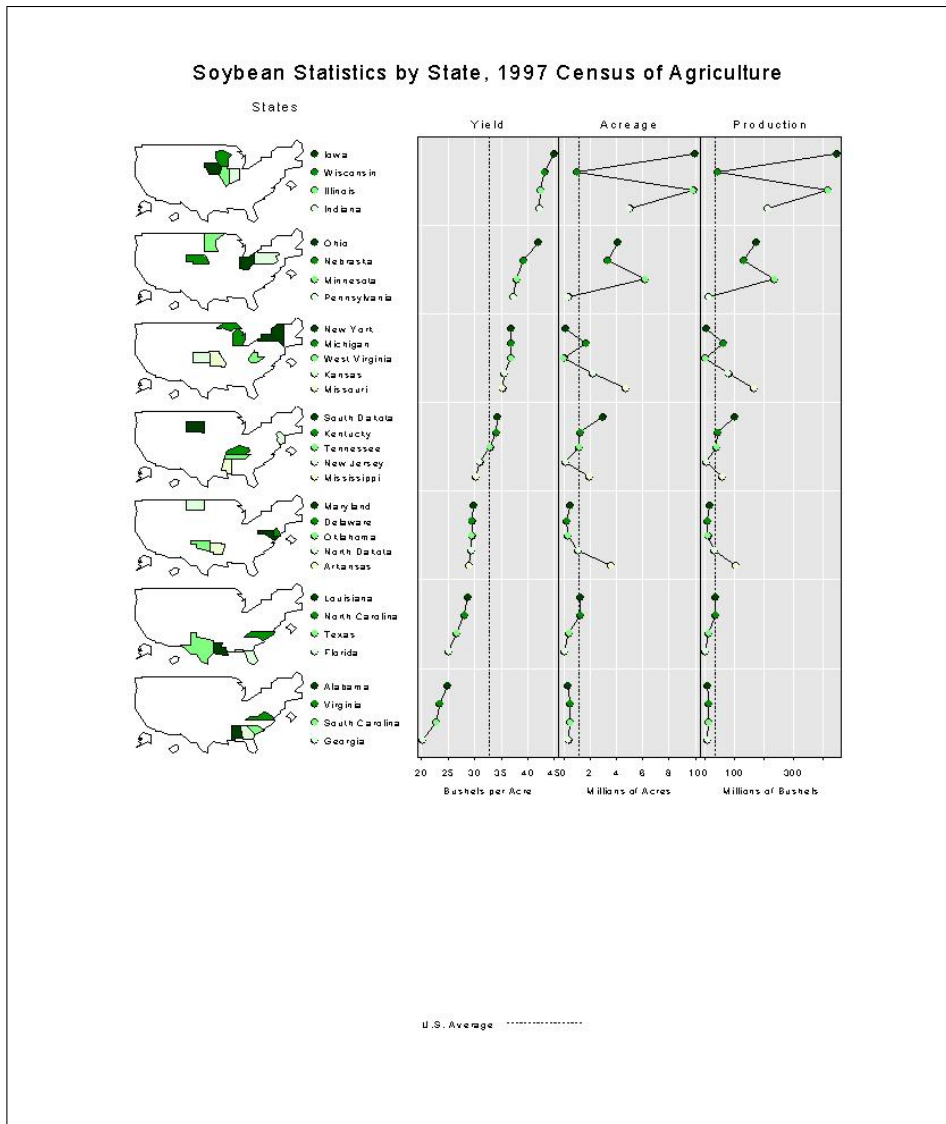


Figure 96: Symanzik & Carr (2008), p. 275, Figure 1.3: LM plots of the 1997 Census of Agriculture, showing soybean yield (in bushels per acre), acreage (in millions of acres), and production (in millions of bushels) by state. The data is sorted by yield and shows the 31 U.S. states where soybeans were planted. The “U.S. Average” represents the median, i.e., the value that splits the data in half such that one half of the states has values below the median and the other half of the states has values above the median. For example, Tennessee is the state with the median yield. This figure has been republished from <http://www.nass.usda.gov/research/gmsoyyap.htm> without any modifications (and ideally should contain much less white space in the lower part).

4.2.3 Additional Linked Micromap Examples

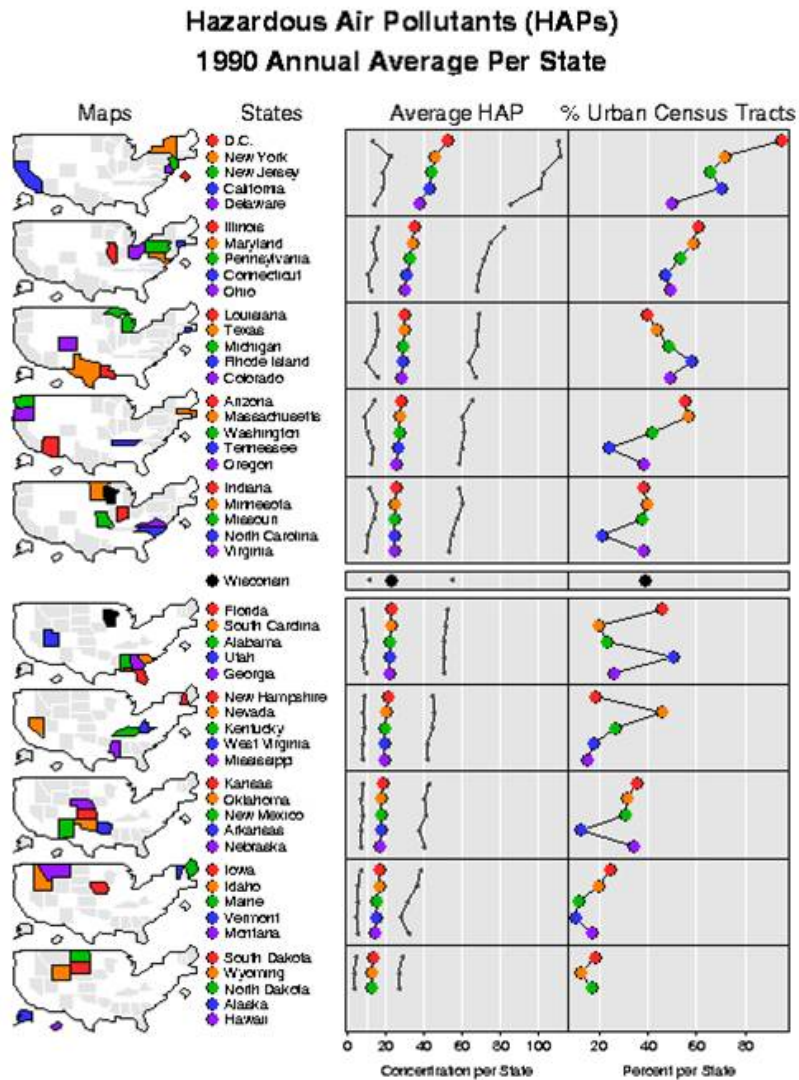


Figure 97: Hazardous Air Pollutants (HAPs) for the US from work with the EPA.

Oral Cleft Occurrence by State 1998–2002

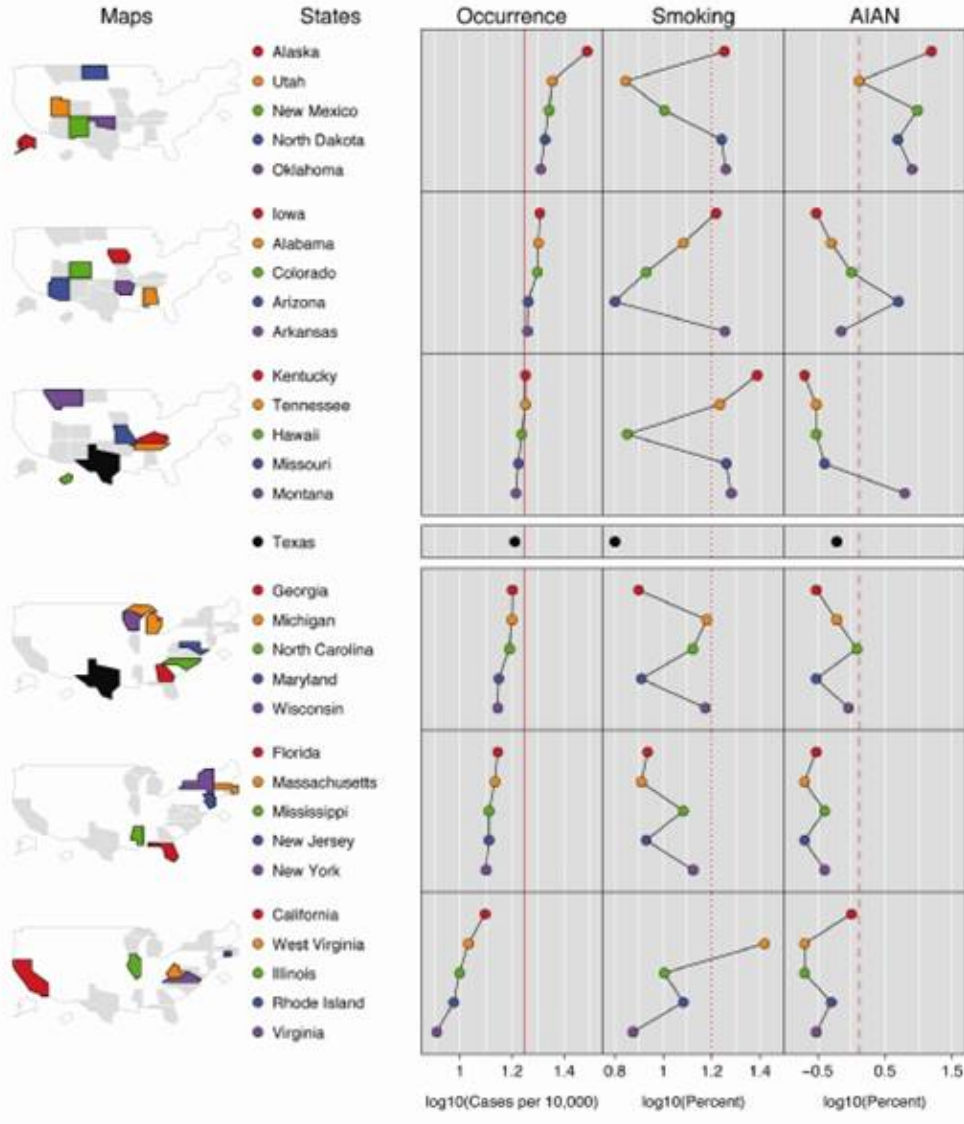


Figure 99: Gebreab et al. (2008), p. 114, Figure 2: LM plot showing oral cleft occurrence by state for the period of 1998–2002. Only oral cleft occurrence for 31 out of the 50 US states was available and displayed here. Smoking rates for California were not available. The red lines show the national average (i.e., mean) of oral cleft occurrence (17.7 per 10,000), smoking rate 16%, and AINA proportion of 1.3%. Note that Texas had the median oral cleft occurrence among the 31 states for which data were available.

West Nile Virus 2003 Lab-Positive Human Cases

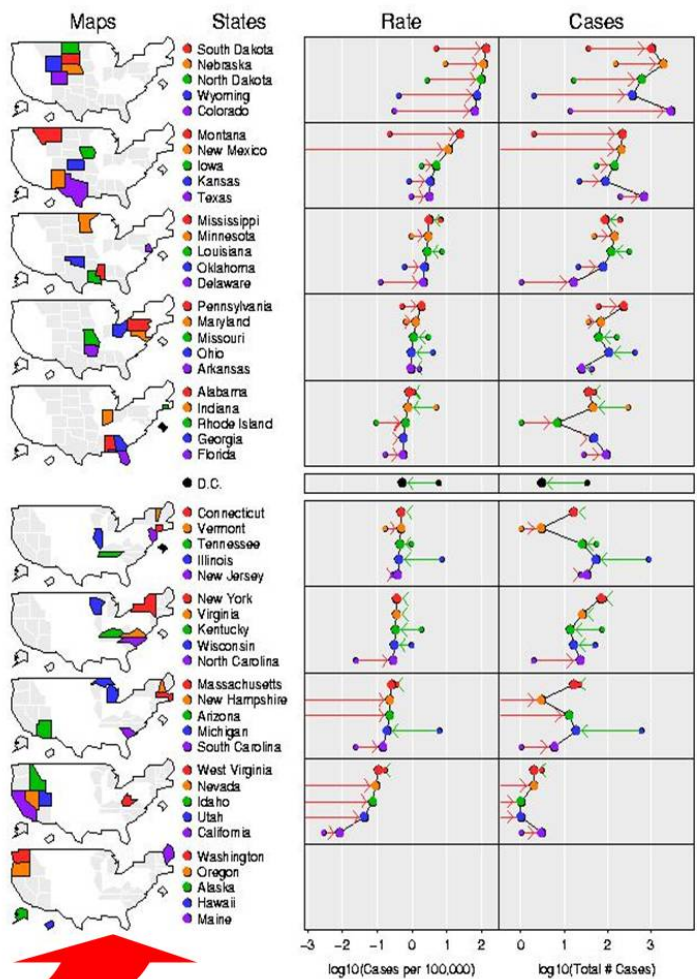


Figure 100: Spread of the West Nile Virus (WNV) across the US, 2002 vs 2003.

Annual CO2 Emissions From Energy Use

Units = Tons Per Person

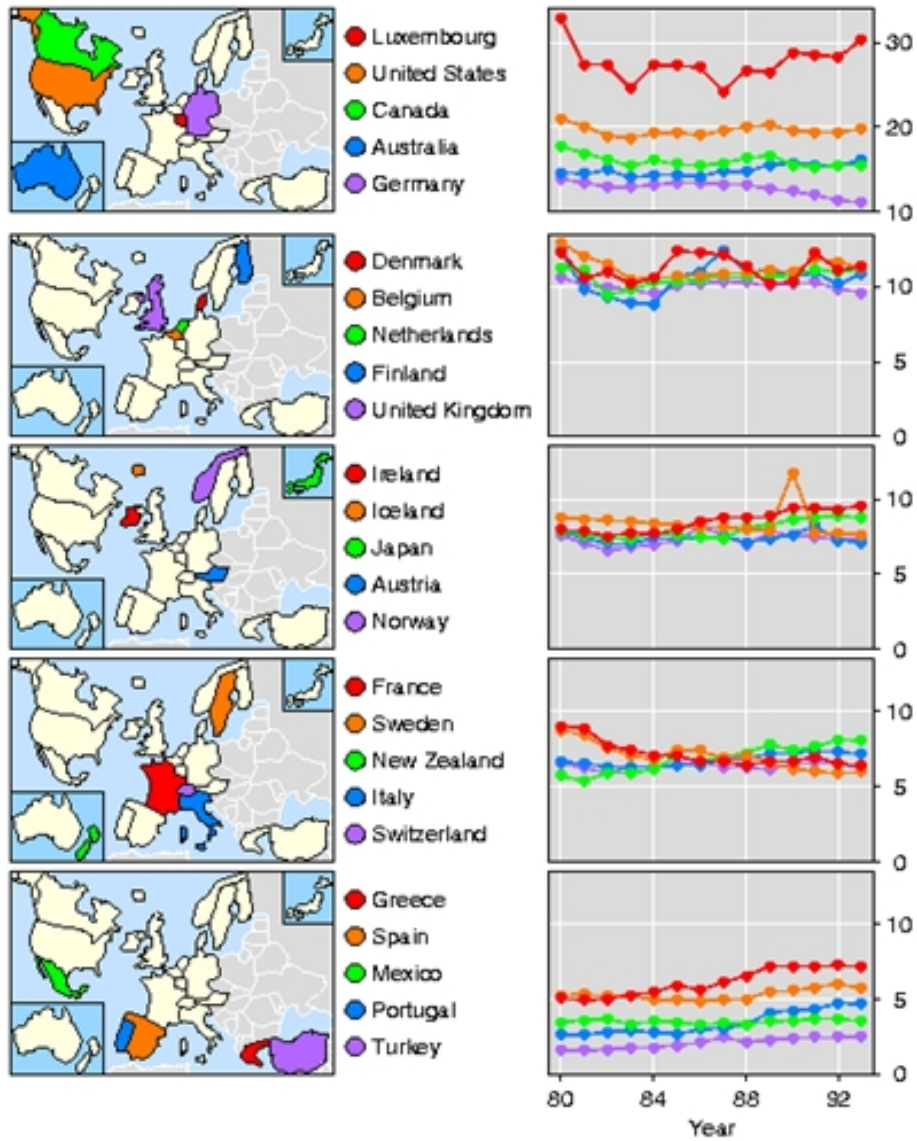


Figure 101: Example from Dan Carr, showing CO2 emissions over time.

White Male Lung Cancer Mortality Rates

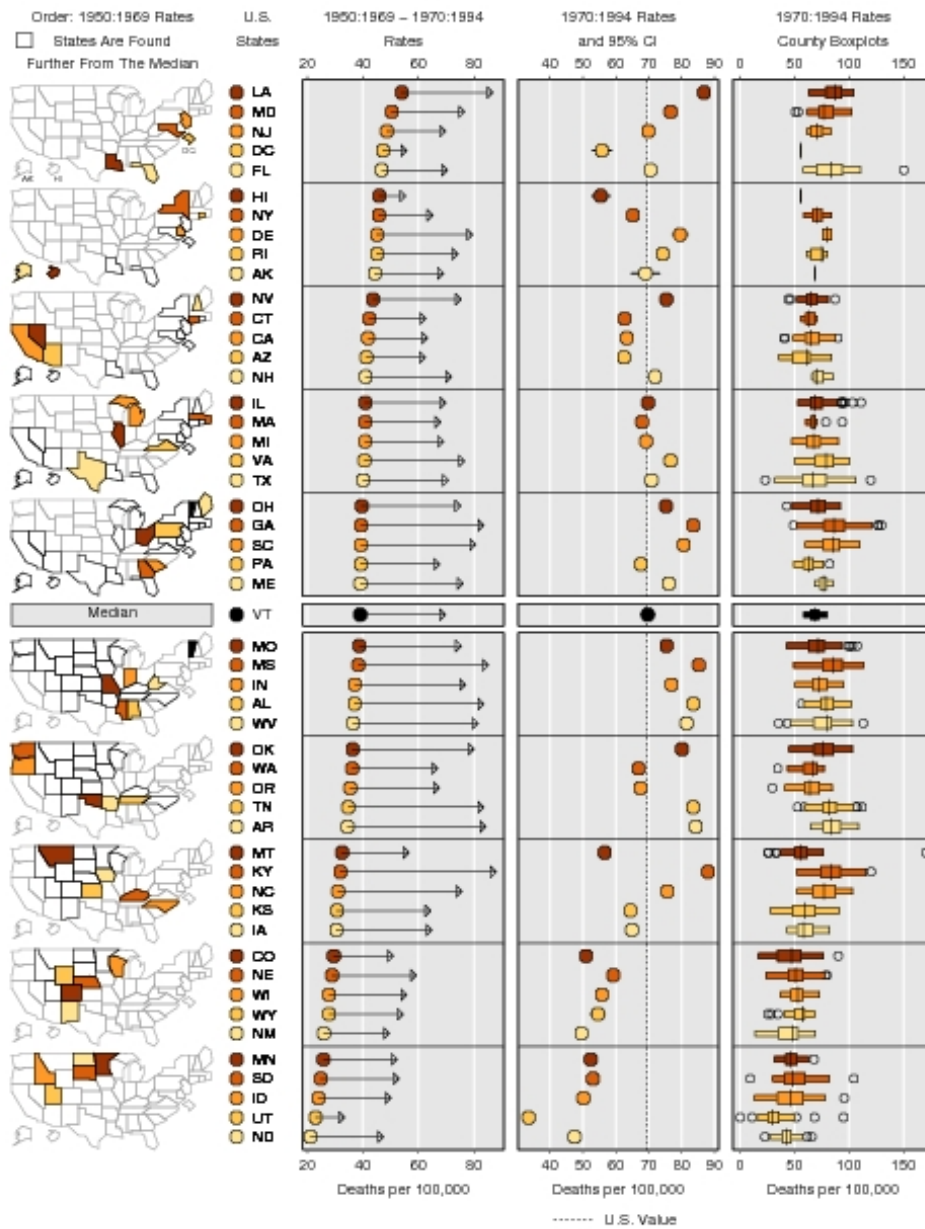


Figure 102: Symanzik & Carr (2008), p. 286, Figure 1.6: LM plots, based on data from the NCI Web page, showing summary values for the years 1950 to 1969 and for the years 1970 to 1994 in the left data panel, rates and 95% confidence intervals in the middle data panel, and boxplots for each of the counties of each state in the right data panel.

4.2.4 Web-based Applications of LM Plots

Symanzik & Carr (2008), pp. 276–282, state:

“Over the last decade, U.S. Federal Agencies and other institutions increasingly focused attention on distributing large amounts of geographically referenced statistical data, either in print or through the Web. The Web-based distribution of data is aimed at replacing printed tabular displays and at providing access to current data quickly. Several approaches have been developed that provide a user-friendly Web-based interface to tabular and graphical displays of Federal data. The user can interactively and dynamically query and sort the data, compare different geographic regions, and look at the data at different spatial resolutions, e.g., at the state or the county level. Carr & Olsen (1996) provide examples on the visual appearance of patterns in data when properly sorted.

The direction of LM plots development shifted from static LM plots towards interactive micromap displays for the Web. Work done for the EPA CEP Web site (Symanzik, Wong, Wang, Carr, Woodruff & Axelrad 2000) was the first in this regard. This project was soon followed by Web-based examples of micromaps produced by the USDA–NASS such as in Fig. 96.

The Digital Government (dg.o) initiative (<http://www.diggov.org>)[‡] is a major research initiative funded by the National Science Foundation (NSF) and several Federal Agencies such as the EPA, the USDA–NASS, the U.S. Census Bureau, the NCI, the U.S. Bureau of Labor Statistics (BLS), etc. This initiative addresses multiple aspects related to Federal data such as visualization, access, disclosure, security, etc. One of the proposals funded under dg.o was the Digital Government Quality Graphics (DGQG) project that included the development of LM plots (<http://www.geovista.psu.edu/grants/dg-qg/index.html>).

In the remainder of this section, we look at four main applications of interactive Web-based LM plots, three of them on Federal Web sites. A short overview of interactive micromaps, as well as a micromap of the “Places” data (Boyer & Savageau 1981), can be found in Symanzik (2004). However, additional details are given in this section.

[‡]Archived versions of these Web pages can be found at http://web.archive.org/web/*/http://www.diggov.org.

Micromaps at the EPA CEP Web Site

The idea of using micromaps on the Web was first considered for the EPA CEP Web site (previously accessible at <http://www.epa.gov/CumulativeExposure/>).[§] Initially, the EPA wanted to provide fast and convenient Web-based access to its hazardous air pollutant (HAP) data for 1990. In this data set, concentrations of 148 air pollutants were estimated for each of the 60,803 U.S. census tracts in the 48 contiguous U.S. states (Rosenbaum et al. 1999). The EPA Web site was designed to allow the user to easily move through the data set to find information on different air pollutants at different geographic locations and at different levels of geographic resolution (e.g., state, county, census tract) via interactive tables and micromaps. Unfortunately, no part of the interactive CEP Web site was ever published due to concerns that the 1990 data was outdated at the intended release date in 1998. Only a static version of the CEP Web site without tables and micromaps was accessible for several years. More details on the work related to the planned interactive CEP Web site can be found in Symanzik, Wong, Wang, Carr, Woodruff & Axelrad (2000), Symanzik, Axelrad, Carr, Wang, Wong & Woodruff (1999), Symanzik, Carr, Axelrad, Wang, Wong & Woodruff (1999).

Micromaps at the USDA–NASS Web Site

The USDA–NASS Research and Development Division released a Web site (<http://www.nass.usda.gov/research/sumpant.htm>) in September 1999 that uses interactive micromaps to display data from the 1997 Census of Agriculture. The USDA–NASS Web site displays acreage, production, and yield of harvested cropland for corn, soybeans, wheat, hay, and cotton. [...]

Micromaps at the NCI Web Site

The NCI released the State Cancer Profiles Web site in April 2003 that provides interactive access to its cancer data via micromaps. This Web site is Java-based and creates micromaps “on the fly”. Wang et al. (2002) and Carr

[§]Archived versions of these Web pages can be found at http://web.archive.org/web/*/http://www.epa.gov/CumulativeExposure/.

et al. (2002) provide more details on the design of the NCI Web site that is accessible at <http://www.statecancerprofiles.cancer.gov/micromaps>. [...]

Micromaps at Utah State University

Micromaps and other graphical displays were found to be very useful for the display and analysis of the geographic spread of the West Nile Virus (WNV) and other diseases (Symanzik et al. 2003) across the U.S. For this reason, researchers at Utah State University (USU) obtained the NCI Java micromap code and adapted it for the display of WNV data (Chapala 2005). Similar to the NCI micromap application, a user can now select among WNV infection rates and counts, and death rates and counts, starting with the WNV data for the U.S. for the 2002 season. A drill-down into U.S. counties is possible given that data at the county-level is available. New features of the USU Web site include the plotting of the data for two years side-by-side in one panel and additional sorting criteria such as sorting from the highest increase over no change to highest decrease in two selected years. The USU WNV micromap Web site can be accessed at <http://webcat.gis.usu.edu:8080/index.html>. [...]"

4.2.5 Linked Micromaps in Java

The National Cancer Institute (NCI) states at <http://gis.cancer.gov/tools/micromaps/>:

“Linked Micromaps is a graphing program written in Java. It allows users to view multiple variables interactively and compare statistics across regions (states, counties, registries, hospitals) as well as across time. It supports six types of graph:

- bar graphs;
- box plots;
- raw data tables;
- point graphs;
- point graphs with arrow; and
- point graphs with confidence intervals.

In order to use Linked Micromaps, you must have Java installed on your PC. Your input files must be in a delimited (such as Comma-Separated Values [CSV]) or fixed-width text format.”

4.2.6 Linked Micromaps in R

Linked micromaps were initially implemented in S-Plus (and not in R). While higher-level functionality is similar, if not identical between both software packages, the internal implementations are rather different. Due to such internal differences, it took many years before the appearance of first examples of micromaps created in R. Three Web pages accompanying Carr & Pickle (2010) provide access to numerous data sets, shape files, and R code used in the book: <http://www.crcpress.com/product/isbn/9781420075731>, <http://www.statnetconsulting.com/micromaps.html>, and <http://mason.gmu.edu/~dcarr/Micromaps/>.

Example 1:

Execute this code to create a simplified micromap with one statistical panel, written by Mike Minnotte:

```
library(RColorBrewer)
library(maps)

display.brewer.pal(6, "Set1")
pal <- brewer.pal(6, "Set1")
pal[6] <- "#DDDDDD"

data(state)
murder <- state.x77[,5]
murder.name <- state.name[order(murder, decreasing = T)]
murder.name
murder <- sort(murder, decreasing = T)

pdf("Ch9_micromap_Ex1.pdf", width = 7.5, height = 10)

layout(matrix(1:36, nrow = 12, byrow = T), widths = c(1, 1, 2),
        heights = c(rep(4, 5), 1, rep(4, 5), 3))
#layout.show(36)

for (i in 1:10)
{
```

```

# compute colors, plot map (column 1)

par(mex = 0.5, mar = rep(.01,4))
if (i <= 5) m.col <- c(rep(pal[6],25),rep(0,25))
  else m.col <- c(rep(0,25),rep(pal[6],25))
m.col[(i-1)*5+1:5]<-pal[1:5]
map.m.col<-m.col[match.map("state",murder.name)]
map("state", fill = T, col = map.m.col, border = 0,
    xlim = c(-125, -65), ylim = c(25, 50))

# plot labels (column 2)

par(mar=rep(.1,4))
plot(0,0,xlim=c(0,1),ylim=c(0,1),type="n",bty="n",
    xaxt="n",yaxt="n",xlab="",ylab="")
points(rep(.1,5),seq(.9,.1,by=-.2),pch=21,bg=pal[1:5],cex=2)
text(rep(.18,5),seq(.9,.1,by=-.2),murder.name[(i-1)*5+1:5],pos=4,cex=1.5)

# plot dotplot of values (column 3)

par(mar=rep(.1,4))
if (i==10) plot(0,0,xlim=range(murder),ylim=c(0,1),type="n",
    yaxt="n",xlab="",ylab="")
else plot(0,0,xlim=range(murder),ylim=c(0,1),type="n",xaxt="n",
    yaxt="n",xlab="",ylab="")
abline(h=seq(.9,.1,by=-.2),lty=3,col="grey")
points(murder[(i-1)*5+1:5],seq(.9,.1,by=-.2),pch=21,bg=pal[1:5],cex=2)

# separate states above and below median

if (i==5) {for (j in 1:3){
plot(0,0,xlim=c(0,1),ylim=c(0,1),type="n",bty="n",
    xaxt="n",yaxt="n",xlab="",ylab="")
abline(h=.5,lwd=3,col=pal[6])
}}}

```

```
# Plot through remaining (empty) cells

if (i==10) for (j in 1:3)
plot(0,0,xlim=c(0,1),ylim=c(0,1),type="n",bty="n",
     xaxt="n",yaxt="n",xlab="",ylab="")
}

# Label for dot plots
text(0.5, 0.25, "Murders per 100K Population", cex = 1.5)

dev.off()
```

Example 2:

More sophisticated R code for micromaps, created by Dan Carr. The original version of this R code can be found at <http://classweb.gmu.edu/dcarr/eda/schedule.html> (Week 6). The R code posted here has been modified to run from our course Web page.

```
# Load R Functions

load(url("http://www.math.usu.edu/~symanzik/teaching/2011_stat6560/RDataAndScripts/Ch9_panelLayout.Rdata"))

# Load Data

stateUnemploy95 =
  read.csv(url("http://www.math.usu.edu/~symanzik/teaching/2011_stat6560/RDataAndScripts/Ch9_stateUnemployment95.csv"),
    row.names = 1, header = T)

stateNamesFips =
  read.csv(url("http://www.math.usu.edu/~symanzik/teaching/2011_stat6560/RDataAndScripts/Ch9_stateNamesFips.csv"),
    row.names = 1, header = T)

stateVBorders =
  read.csv(url("http://www.math.usu.edu/~symanzik/teaching/2011_stat6560/RDataAndScripts/Ch9_stateVisibilityBorders.csv"),
    row.names = NULL, header = T)

nationVBorders =
  read.csv(url("http://www.math.usu.edu/~symanzik/teaching/2011_stat6560/RDataAndScripts/Ch9_nationVisibilityBorders.csv"),
    blank.lines.skip = F, row.names = NULL, header = T)

# Create pdf (or jpg) Output

pdf("Ch9_micromap_Ex2.pdf", width = 7.5, height = 10)
#jpeg("Ch9_micromap_Ex2.jpg", width = 7.5, height = 10, units = "in", res = 72)

source(url("http://www.math.usu.edu/~symanzik/teaching/2011_stat6560/RDataAndScripts/Ch9_linked_micromaps.R"))

dev.off()
```


4.3 Conditioned Micromaps

Carr & Pickle (2010), p. 7, state:

“The primary purpose of the conditioned micromap is data exploration, unlike the linked micromap plot, which is used most often for presentation. A major difference between these designs is that linked micromaps use a single ranked variable to partition the mapped regions into a linear sequence of panels. In contrast, conditioned micromaps use two ranked variables to partition the regions into a two-way grid of panels, with the rank order of one variable determining the row membership and the rank order of the second variable determining the column membership. Two slider bars set category cutpoints for these variables, allowing the analyst to dynamically explore the geographic patterns on the map, based on categorized values of the two auxiliary variables.”

In the literature, conditioned micromaps, also have been called Conditioned Choropleth Maps (CCMaps). Additional examples of CCMaps can be found in Carr, Wallin & Carr (2000), Carr et al. (2002), Carr et al. (2003), Carr et al. (2005), and Symanzik & Carr (2008).

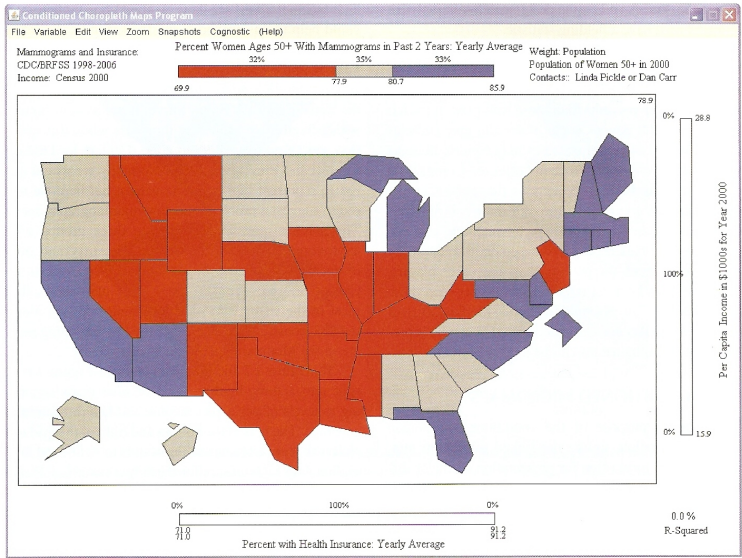


FIGURE 1.6 Full map of the percent of women ages 50 and over who have had a mammogram in the past two years, by state, which we partition into conditioned micromaps in Figure 1.7. Note that red is used as the low (bad) value, contrary to traditional use.

Figure 103: Carr & Pickle (2010), p. 8, Figure 1.6.

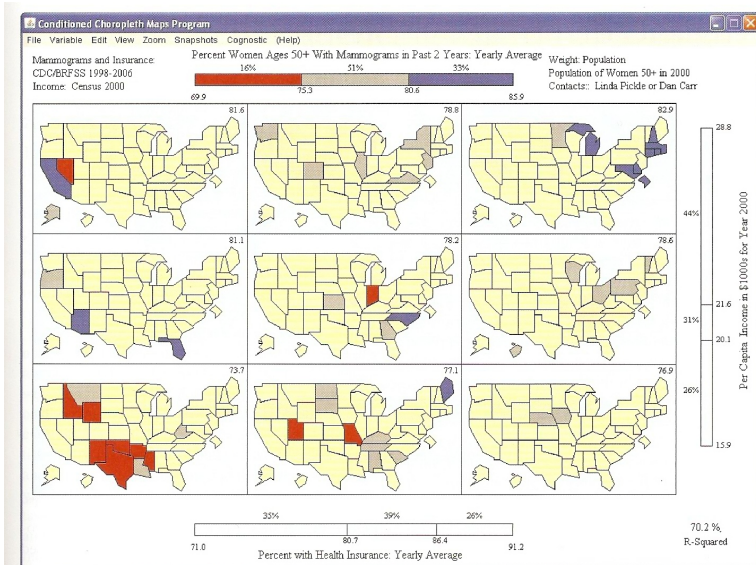


FIGURE 1.7 Conditioned micromap of the mammography data shown in Figure 1.6, conditioned on the per capita income (rows) and health insurance coverage (columns) as defined by the slider bars.

Figure 104: Carr & Pickle (2010), p. 9, Figure 1.7.

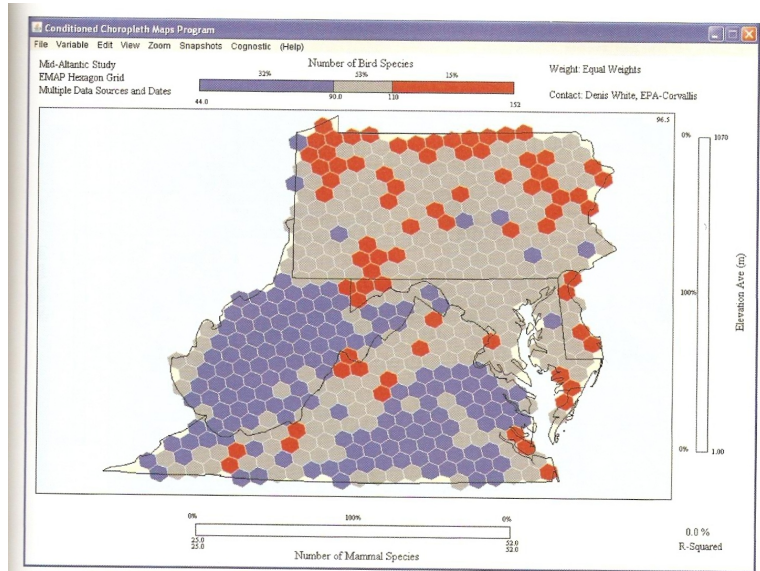


FIGURE 5.10 Use of a hexagon grid in conditioned micromaps to display the number of bird species in local regions. The blue hexagons have 90 or fewer bird species, the gray regions have 91 to 110 species, and the red hexagons have over 110 species (up to 152).

Figure 105: Carr & Pickle (2010), p. 91, Figure 5.10.

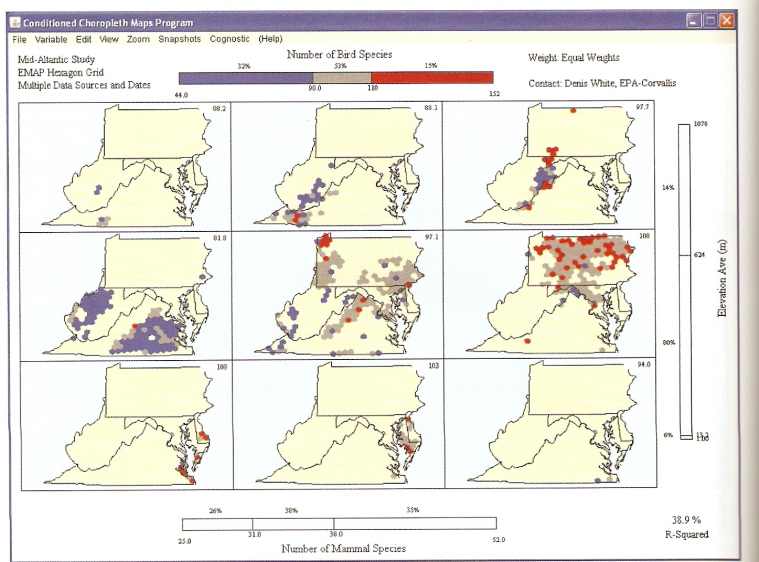


FIGURE 5.11 Hexagon bin conditioned choropleth micromap of bird species data (Figure 5.10) conditioned on the number of mammal species and average elevation.

Figure 106: Carr & Pickle (2010), p. 92, Figure 5.11.

4.4 Comparative Micromaps

Carr & Pickle (2010), p. 8, state:

“Comparative micromaps are one- and two-way sequences of maps indexed by time or other attributes. The emphasis is on comparisons across maps rather than comparisons within one map. The most common variant of comparative micromaps is a sequence of time-specific maps of values accompanied by a corresponding series of maps of the class or value differences of consecutive maps, allowing the reader to see explicitly the amount and location of changes. It is these difference maps that distinguish comparative micromaps from a standard time series of maps.”

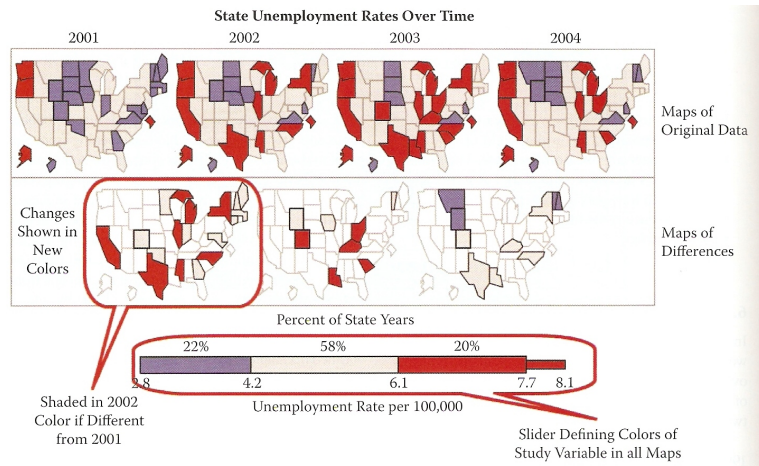


FIGURE 6.1 Comparative micromaps of annual state unemployment rates, 2001–2004, illustrating the basic design, annotated to show basic components.

Figure 107: Carr & Pickle (2010), p. 110, Figure 6.1.



FIGURE 1.8 A typical comparative micromap plot design, with time series maps supplemented by explicit difference maps.

Figure 108: Carr & Pickle (2010), p. 10, Figure 1.8.

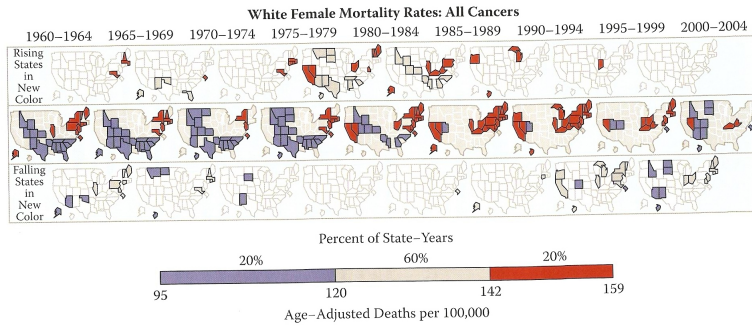


FIGURE 6.6 Rates of mortality among white females due to any malignancy, 1960–2004, by state, with changes shown separately for rising and falling rates.

Figure 109: Carr & Pickle (2010), p. 116, Figure 6.6.

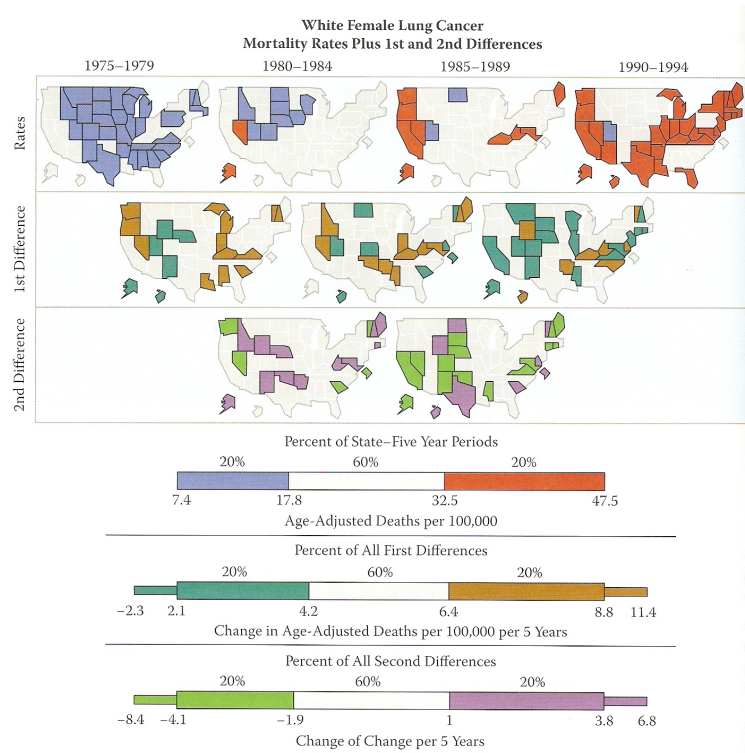


FIGURE 6.10 First differences (middle row) of the times series of white female lung cancer mortality rates by state (top row). Bottom row displays the second differences (differences of the second row).

Figure 110: Carr & Pickle (2010), p. 122, Figure 6.10.

4.5 Further Reading

Additional sources for choropleth maps and micromaps are:

- Work related to Pickle et al. (1996): Pickle & Herrmann (1999), Wainer (2008), and Pickle (2008)



Figure 111: http://www.cartoonstock.com/cartoonview.asp?start=4&search=main&catref=bven112&MA_Artist=&MA_Category=&ANDkeyword=statistics&ORkeyword=&TITLEkeyword=&NEGATIVEkeyword=, Cartoon.

5 Categorical Plots

5.1 Which Plot Type to Choose?

Often, there exist many valid options how to display (categorical) data.

Zelazny (2001), p. 12, suggests the following project:

“Sketch as many charts as you can think of using these data: the more the better.”

Percentage of January Sales by Region

	<u>Co. A</u>	<u>Co. B</u>
North	13%	39%
South	35%	6%
East	27%	27%
West	25%	28%

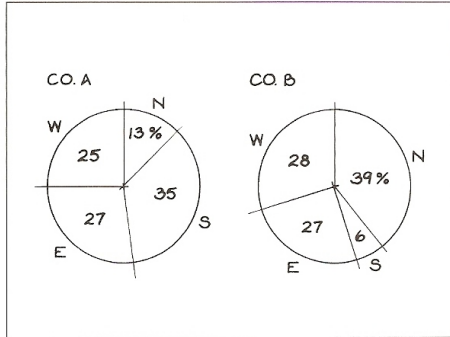
Worksheet

Your Name: _____

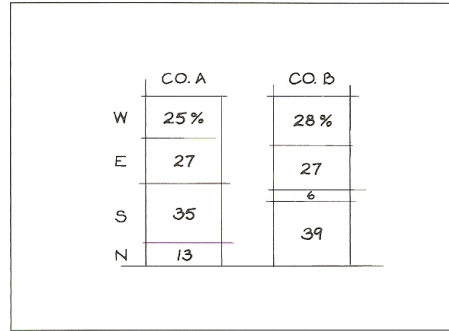
Worksheet Answers

WHICH CHART WOULD YOU CHOOSE?

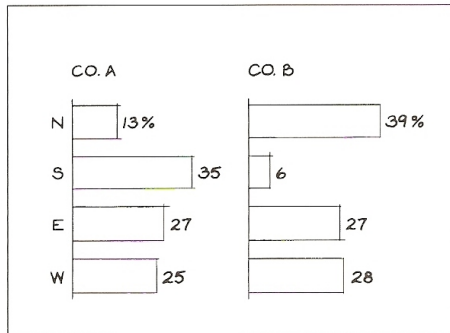
► 1



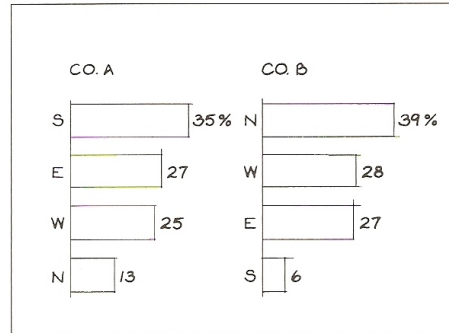
► 2



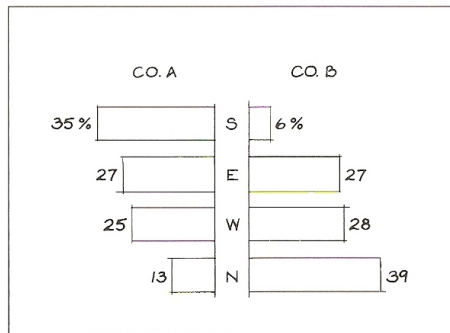
► 3



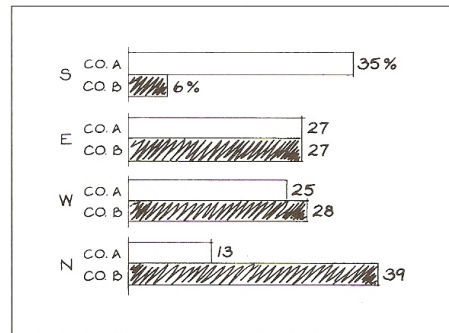
► 4



► 5



► 6



14

Figure 112: Zelazny (2001), p. 14, Figure.

The charts shown on the facing page may be among those you sketched. All the better if you thought of others. But a question remains.

WHICH CHART WOULD YOU CHOOSE?

It all depends! It all depends on the specific point *you* want to make—*your* message. Each chart shown, simply as a function of the way it's organized, is best equipped to emphasize a particular message.

For instance, showing the data as a couple of pie charts or 100 percent columns, you would be emphasizing that:

▶ **1 ▶ 2** The mix of sales is different for Companies A and B.

Or you may have shown the data as two sets of bar charts, sequencing the bars in the order the data were presented in the table. Now the chart is stressing the message that:

▶ **3** The percentage of sales for both Companies A and B varies by region.

On the other hand, you could have ranked the percentage of sales for each company in descending (or ascending) order, now stressing the point that:

▶ **4** Company A is highest in the South; Company B is highest in the North. Or, Company A is lowest in the North; Company B is lowest in the South.

By structuring the bars in a mirror image around the regions, we now demonstrate that:

▶ **5** Company A's share of sales is highest in the South where Company B's is the weakest.

By grouping the bars against a common base, we now compare the gaps by region, showing that:

▶ **6** In the South, Company A leads B by a wide margin; in the East and West, the two are competitive; in the North, A lags B.

Now, it's possible—even probable—that in the early stages of deciding what your message should be, you may need to sketch a number of charts that look at the data from various points of view. A more efficient approach is to highlight the aspect of the data that seems most important and settle on the message that brings out that aspect.

15

Figure 113: Zelazny (2001), p. 15, Text.

5.2 Categorical Plots in R

Recall Section 2.4, “*Sex Bias in Graduate Admissions*”, from Freedman et al. (2007), pp. 17–20, many of us are using in our introductory Stat 1040 class.

These data represent aggregate data on applicants to graduate school at Berkeley for the six largest departments in 1973, classified by admission and sex. These data are often used to discuss the issue whether the data show evidence of sex bias in admission practices. There were 2691 male applicants, of whom 1198 (44.5%) were admitted, compared with 1835 female applicants of whom 557 (30.4%) were admitted. Ultimately, this data set is frequently used for illustrating Simpson’s paradox and does not show any sex bias when properly analyzed. An effective graphical way to explain Simpson’s Paradox is the BK–Plot, summarized in Wainer (2002).

In R, the data are stored in a 3–dimensional array resulting from cross–tabulating 4526 observations on 3 variables. The variables and their levels are as follows:

No	Name	Levels
1	Admit	Admitted, Rejected
2	Gender	Male, Female
3	Dept	A, B, C, D, E, F

In R, this data set is accessible via:

```
UCBAdmissions
```

A better tabular representation can be obtained via:

```
ftable(UCBAdmissions)
```

To obtain the totals as represented in Freedman et al. (2007), p. 18, we have to sum over dimensions 2 and 3 in this 3–dimensional array:

```
apply(UCBAdmissions, c(2, 3), sum)
#
# also, margin.table produces the same result
#
margin.table(UCBAdmissions,2:3)
```

To better understand over which dimensions we sum, replace the `c(2, 3)` option with possible other indices, e.g., `1` or `c(1, 2)`. Try a few more.

Question:

How can we calculate in R the percent admitted, as shown in Freedman et al. (2007), p. 18, Table 2? This can be done via a single command line and does not require any loop! And, which single digit do we have to change in our previous R command to obtain the percent rejected?

Table 2. Admissions data for the graduate programs in the six largest majors at University of California, Berkeley.

<i>Major</i>	<i>Men</i>		<i>Women</i>	
	<i>Number of applicants</i>	<i>Percent admitted</i>	<i>Number of applicants</i>	<i>Percent admitted</i>
A	825	62	108	82
B	560	63	25	68
C	325	37	593	34
D	417	33	375	35
E	191	28	393	24
F	373	6	341	7

Note: University policy does not allow these majors to be identified by name.
Source: The Graduate Division, University of California, Berkeley.

Figure 114: Freedman et al. (2007), p. 18, Table 2.

Answer:

```
# Percent admitted
UCBAdmissions[1, ] / apply(UCBAdmissions, c(2, 3), sum) * 100
# Percent rejected
UCBAdmissions[2, ] / apply(UCBAdmissions, c(2, 3), sum) * 100
```

5.2.1 Pie Charts

Let us concentrate on the popularity of the six majors first, i.e., the total number of admissions for each of these majors.

In R, these application numbers can be calculated via:

```
apply(UCBAdmissions, 3, sum)
```

Many people would immediately think of a pie chart as a possible graphical representation:

```
pie(apply(UCBAdmissions, 3, sum))
```

Note that there is no sorting here. Can you easily order the slices by visual inspection, i.e., which major has the largest number/percentage of admissions, which is second, third, etc.?

A better representation is to sort the data from largest to smallest and then plot the slices in clockwise direction, starting with the largest slice at 90°.

```
pie(sort(apply(UCBAdmissions, 3, sum), decreasing = TRUE),  
     clockwise = TRUE,  
     main = "UC Berkley Admissions by Major")
```

This is somewhat better, but still not perfect. The R help page for pie charts indicates:

“Pie charts are a very bad way of displaying information. The eye is good at judging linear measures and bad at judging relative areas. A bar chart or dot chart is a preferable way of displaying this type of data.”

Moreover, Cleveland (1985), p. 264, states:

“Data that can be shown by pie charts always can be shown by a dot chart. This means that judgements of position along a common scale can be made instead of the less accurate angle judgements.”

And what about the extremely popular 3D-pie charts that often can be found in business reports and the media? The answer is a clear **Don't**.

Wallgren et al. (1996), p. 70, provide a striking example why not to use 3D-pie charts. Guess the percentages associated with the four different areas:

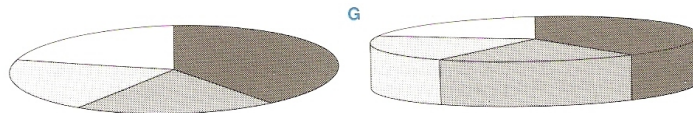


Figure 115: Wallgren et al. (1996), p. 70, Figure G.

And here is the answer:

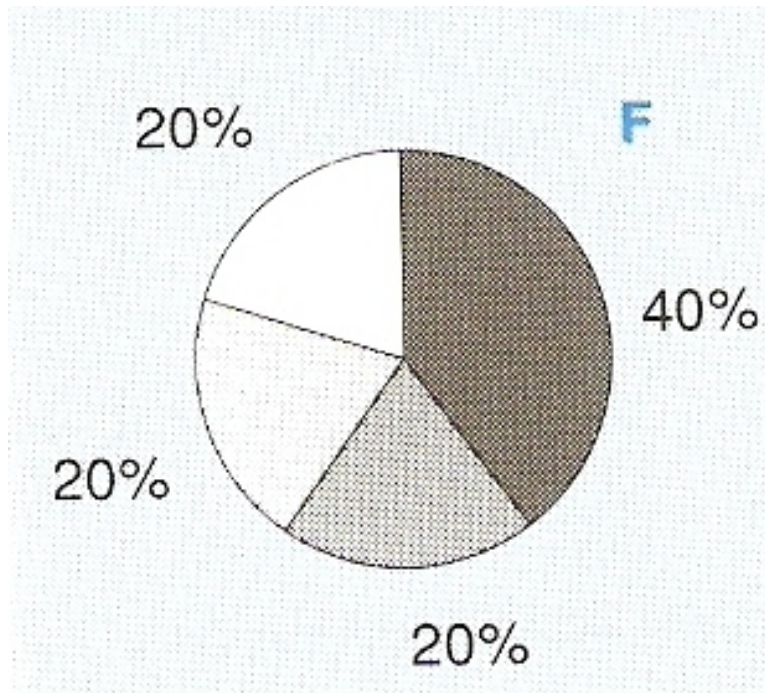


Figure 116: Wallgren et al. (1996), p. 70, Figure F.

5.2.2 Bar Charts

The R help page for `barplot` indicates:

“Creates a bar plot with vertical or horizontal bars.”

```
UCBAd = margin.table(UCBAdmissions, 1:2)
```

```
UCBAd
```

```
barplot(UCBAd, legend.text = T)
```

```
barplot(UCBAd, legend.text = T, beside = T)
```

The following commands create (divided) bar charts that show the percentage admitted/rejected for each gender.

```
barplot(UCBAd / rbind(margin.table(UCBAd, 2), margin.table(UCBAd, 2)),  
        legend.text = T)
```

```
barplot(UCBAd / rbind(margin.table(UCBAd, 2), margin.table(UCBAd, 2)),  
        legend.text = T, beside = T)
```

Warning:

Cleveland (1994), Section 4.10, “Pop Charts”, p. 262, strongly advises against the use of pie charts, divided bar charts, and area charts:

Three graphical methods — pie charts, divided bar charts, and area charts — are widely used in mass media and business publications but are used far less in science and technology. Because of their use, we will call these graphical methods *pop charts*.

Any data that can be encoded by one of these pop charts can also be encoded by either a dot plot or a multiway dot plot that typically provides far more efficient pattern perception and table look-up than the pop-chart encoding. Interestingly, the better pattern perception results from a detection operation, a phenomenon that has been missed in previous studies of pop charts.

5.2.3 Dot Charts

The R help page for `dotchart` indicates:

“Draw a Cleveland dot plot. [...]

This function is invoked for its side effect, which is to produce two variants of dotplots as described in Cleveland (1985). Dot plots are a reasonable substitute for bar plots.”

```
dotchart(UCBAd)
```

```
UCBMajor = margin.table(UCBAdmissions, 2:3)  
dotchart(UCBMajor)
```

```
UCBMajorsort = UCBMajor[, order(UCBMajor[1,], decreasing = TRUE)]  
dotchart(UCBMajorsort, color = c("red", "blue"))
```

5.2.4 Mosaic Plots

The R help page for `mosaicplot` indicates:

“Plots a mosaic on the current graphics device. [...]

shade a logical indicating whether to produce extended mosaic plots, or a numeric vector of at most 5 distinct positive numbers giving the absolute values of the cut points for the residuals. By default, *shade* is FALSE, and simple mosaics are created. Using *shade* = TRUE cuts absolute values at 2 and 4.”

```
mosaicplot(UCBAd)
```

```
mosaicplot(UCBAd, shade = T)
```

```
mosaicplot(UCBAdmissions, shade = T)
```

```
mosaicplot(aperm(UCBAdmissions, 3:1), shade = T)
```

5.2.5 Spine Plots and Spinograms

The R help page for spineplot indicates:

“Spine plots are a special case of mosaic plots, and can be seen as a generalization of stacked (or highlighted) bar plots. Analogously, spinograms are an extension of histograms.”

```
#
# compare the use of this command without () ...
#
spineplot(UCBAd)

#
# ... and with ()
#
(spineplot(UCBAd))

(spineplot(t(UCBAd)))

(spineplot(margin.table(UCBAdmissions, c(3, 2)), main = "Applications at UCB"))

(spineplot(margin.table(UCBAdmissions, c(3, 1)), main = "Admissions at UCB"))
```

5.2.6 Four Fold Plots

The R help page for fourfoldplot indicates:

“Creates a fourfold display of a 2 by 2 by k contingency table on the current graphics device, allowing for the visual inspection of the association between two dichotomous variables in one or several populations (strata).
[...]

std a character string specifying how to standardize the table. Must be one of “margins”, “ind.max”, or “all.max”, and can be abbreviated by the initial letter. If set to “margins”, each 2 by 2 table is standardized to equate the margins specified by margin while preserving the odds ratio. If “ind.max” or “all.max”, the tables are either individually or simultaneously standardized to a maximal cell frequency of 1.”

```
fourfoldplot(UCBAd, std = "a")
```

```
fourfoldplot(UCBAd)
```

```
fourfoldplot(UCBAdmissions, std = "m")
```

```
fourfoldplot(UCBAdmissions, std = "a")
```

5.3 Categorical Plots in Mondrian

According to <http://rosuda.org/mondrian/>:

“Mondrian is a general purpose statistical data–visualization system. It features outstanding visualization techniques for data of almost any kind, and has its particular strength compared to other tools when working with **Categorical Data, Geographical Data and LARGE Data**.

All plots in Mondrian are fully linked, and offer various interactions and queries. Any case selected in a plot in Mondrian is highlighted in all other plots.

Currently implemented plots comprise **Mosaic Plot, Scatterplots and SPLOM, Maps, Barcharts, Histograms, Missing Value Plot, Parallel Coordinates/Boxplots and Boxplots y by x.**”

Main references for Mondrian are Theus (2002), Theus (2003), and Theus & Urbanek (2009).

Theus & Urbanek (2009) has an associated Web page at <http://www.interactivegraphics.org>:

“This site is the web resource for the book “Interactive Graphics for Data Analysis — Principles and Examples”.

There are links to the most important software tools, all datasets used in the book for easy download, and a set of slides which may be used together with the book for a lecture.

The R–code used in the book can be found here as well.”

5.3.1 Installation

Go to <http://rosuda.org/mondrian/>, then follow the link to the download section on this page. Look over the license condition. If you agree, then download the most recent version of Mondrian (currently 1.0 as of 12/18/2008) by right mouse–clicking on the operating system you use. Save *Mondrian.exe* into a directory of your choice. You can start Mondrian directly (without any additional installation) by mouse–clicking on *Mondrian.exe*.

As a test data set, work with the *Titanic* data available under the *Mondrian Titanic* link or directly from <http://stats.math.uni-augsburg.de/Mondrian/Data/Titanic.txt>. Save these data locally as *Titanic.txt*. Then load them into *Mondrian*.

5.3.2 The Titanic Data in Mondrian

The *Mondrian* description at <http://rosuda.org/mondrian/> indicates:

“Titanic

Data set on the 2201 passengers of the Titanic. Pure categorical with data on class, gender, age and survival.”

The interactive exploration of the *Titanic* data via *Mondrian* has been further discussed in Theus & Urbanek (2009), Examples D: The Titanic Disaster Revisited, pp. 183–191.

Task:

Interactively recreate the nine plots from Figure 117 using *Mondrian*.



Figure 117: Theus & Urbanek (2009), p. 186, Figure.

Answer:

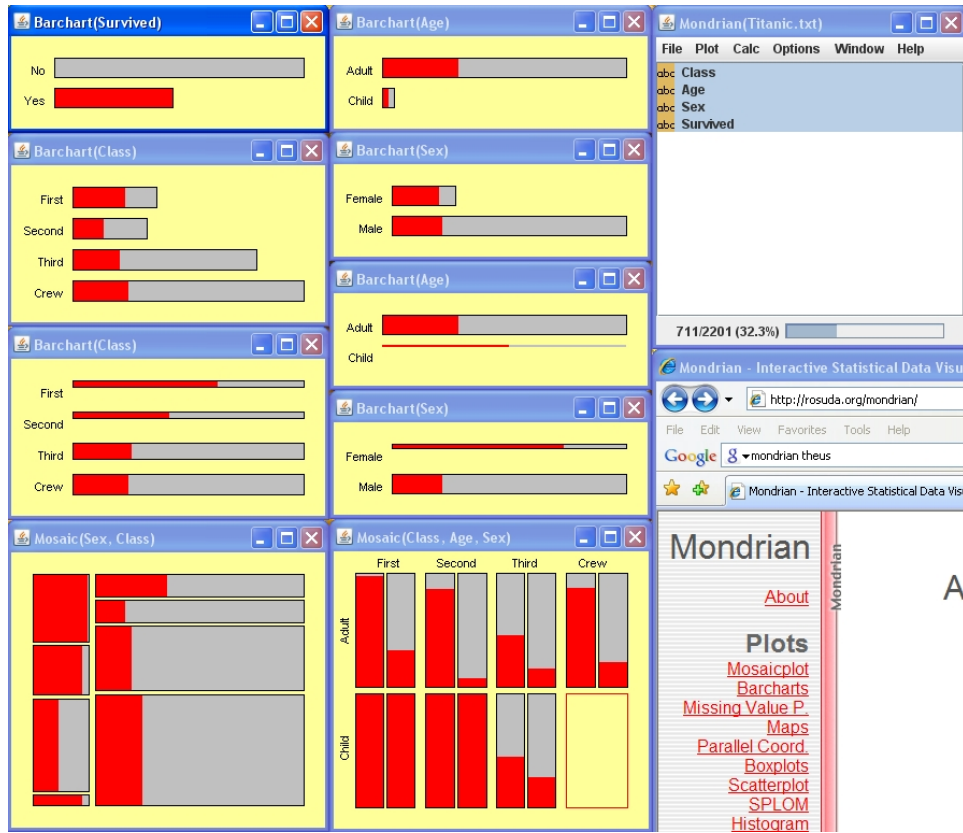


Figure 118: *Mondrian* output related to Theus & Urbanek (2009), p. 186, Figure.

5.4 Further Reading

Additional sources for the visualization of categorical data are:

- Blasius & Greenacre (1998)
- Friendly (2000*b*)
- Hofmann (2007)
- Theus & Urbanek (2009)

5.5 R Code and Output

The automatic output with R code, numerical results, and graphical output in this section and sections in the next chapters are created via the R tool *Sweave*. According to <http://www.statistik.lmu.de/~leisch/Sweave/>:

“What is Sweave?”

Sweave is a tool that allows to embed the R code for complete data analyses in latex documents. The purpose is to create dynamic reports, which can be updated automatically if data or analysis change. Instead of inserting a prefabricated graph or table into the report, the master document contains the R code necessary to obtain it. When run through R, all data analysis output (tables, graphs, etc.) is created on the fly and inserted into a final latex document. The report can be automatically updated if data or analysis change, which allows for truly reproducible research.”

In the recent past, the idea of Reproducible Research (RR) has been introduced at several major research centers, such as the MD Anderson Cancer Center (Baggerly & Berry 2011).

Depending on the hardware platform where you are running *Sweave*, you may have to adjust paths — or you can simply copy the *Sweave.sty* file from the appropriate R directory (e.g., `C:\Program Files\R\R-2.8.1\share\texmf`) into your working directory where the \LaTeX file is located.

Then, within R, you can invoke *Sweave* (without having to install anything else) via:

```
Sweave("lect_chapter4_sweave1.snw")
```

When successful, you should obtain a response like this:

```
You can now run LaTeX on 'lect_chapter4_sweave1.tex'
```

A brief overview on *Sweave* can be found at Adele Cutler’s Web page at <http://www.math.usu.edu/~adele/s6100/sweave.ppt>. For full details including frequently asked questions, visit the official *Sweave* homepage at <http://www.statistik.lmu.de/~leisch/Sweave/>.

The first version of *Sweave* is described in Leisch (2002). A preprint is available at <http://www.statistik.lmu.de/~leisch/Sweave/Sweave-compstat2002.pdf>.

5.5.1 Example 1: UCBAmissions

The R description indicates:

Student Admissions at UC Berkeley

Aggregate data on applicants to graduate school at Berkeley for the six largest departments in 1973 classified by admission and sex.

```
> UCBAmissions
```

```
, , Dept = A
```

	Gender	
Admit	Male	Female
Admitted	512	89
Rejected	313	19

```
, , Dept = B
```

	Gender	
Admit	Male	Female
Admitted	353	17
Rejected	207	8

```
, , Dept = C
```

	Gender	
Admit	Male	Female
Admitted	120	202
Rejected	205	391

```
, , Dept = D
```

	Gender	
Admit	Male	Female
Admitted	138	131
Rejected	279	244

```
, , Dept = E
```

```
      Gender
Admit  Male Female
Admitted  53   94
Rejected 138  299
```

```
, , Dept = F
```

```
      Gender
Admit  Male Female
Admitted  22   24
Rejected 351  317
```

```
> ftable(UCBAdmissions)
```

```
      Dept  A  B  C  D  E  F
Admit  Gender
Admitted Male    512 353 120 138  53  22
        Female    89  17 202 131  94  24
Rejected Male    313 207 205 279 138 351
        Female    19  8 391 244 299 317
```

```
> apply(UCBAdmissions, c(2, 3), sum)
```

```
      Dept
Gender  A  B  C  D  E  F
Male   825 560 325 417 191 373
Female 108  25 593 375 393 341
```

```
> #
```

```
> # also, margin.table produces the same result
```

```
> #
```

```
> margin.table(UCBAdmissions, 2:3)
```

```
      Dept
Gender  A  B  C  D  E  F
Male   825 560 325 417 191 373
Female 108  25 593 375 393 341
```

```
> # Percent admitted
> UCBAmissions[1,,] / apply(UCBAmissions, c(2, 3), sum) * 100
```

	Dept					
Gender	A	B	C	D	E	F
Male	62.0606	63.03571	36.92308	33.09353	27.74869	5.898123
Female	82.4074	68.00000	34.06408	34.93333	23.91858	7.038123

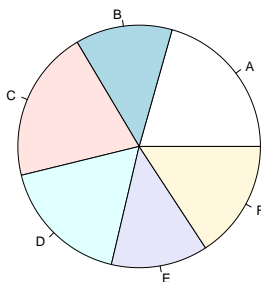
```
> # Percent rejected
> UCBAmissions[2,,] / apply(UCBAmissions, c(2, 3), sum) * 100
```

	Dept					
Gender	A	B	C	D	E	F
Male	37.93939	36.96429	63.07692	66.90647	72.25131	94.10188
Female	17.59259	32.00000	65.93592	65.06667	76.08142	92.96188

```
> apply(UCBAmissions, 3, sum)
```

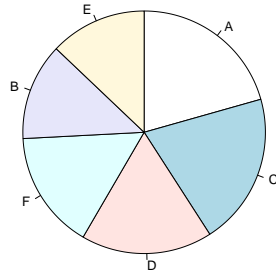
	A	B	C	D	E	F
	933	585	918	792	584	714

```
> pie(apply(UCBAmissions, 3, sum))
```



```
> pie(sort(apply(UCBAmissions, 3, sum), decreasing = TRUE),
+ clockwise = TRUE,
+ main = "UC Berkley Admissions by Major")
```

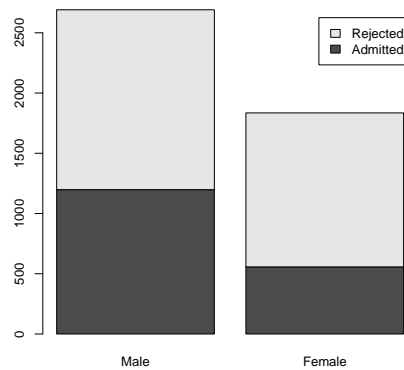
UC Berkley Admissions by Major



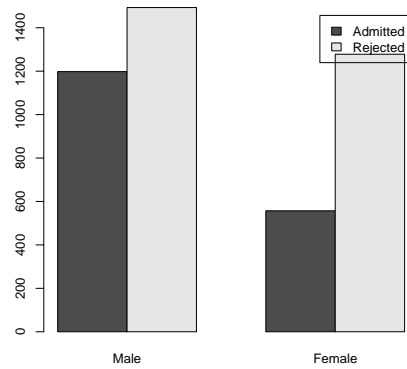
```
> UCBAAd = margin.table(UCBAAdmissions, 1:2)  
> UCBAAd
```

	Gender	
Admit	Male	Female
Admitted	1198	557
Rejected	1493	1278

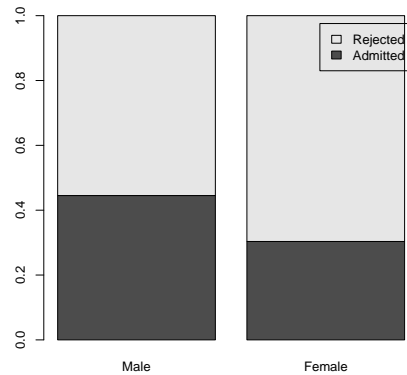
```
> barplot(UCBAAd, legend.text = T)
```



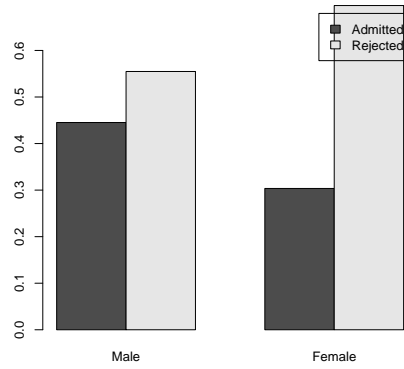
```
> barplot(UCBAAd, legend.text = T, beside = T)
```



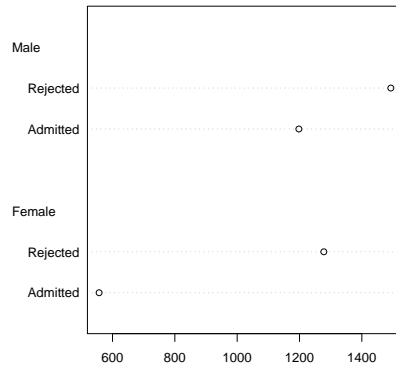
```
> barplot(UCBAd / rbind(margin.table(UCBAd, 2), margin.table(UCBAd, 2)),
+ legend.text = T)
```



```
> barplot(UCBAd / rbind(margin.table(UCBAd, 2), margin.table(UCBAd, 2)),
+ legend.text = T, beside = T)
```

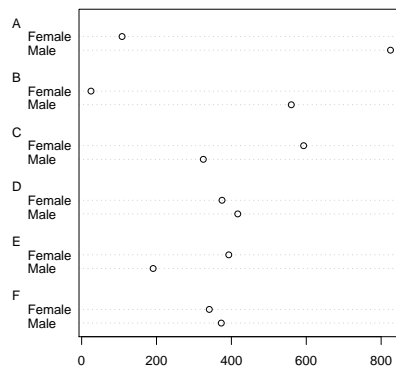


> dotchart(UCBAd)



> UCBMajor = margin.table(UCBAdmissions, 2:3)

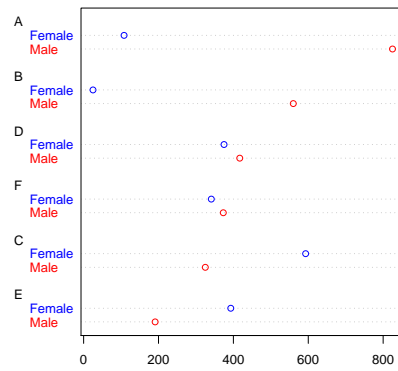
> dotchart(UCBMajor)



```

> UCBMajorsort = UCBMajor[, order(UCBMajor[1,], decreasing = TRUE)]
> dotchart(UCBMajorsort, color = c("red", "blue"))

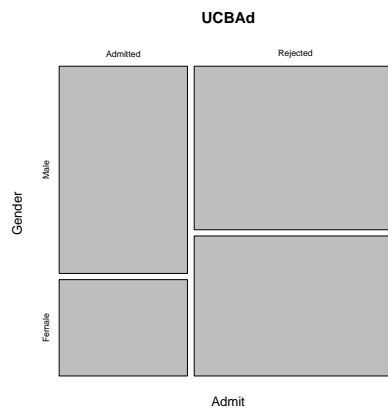
```



```

> mosaicplot(UCBAd)

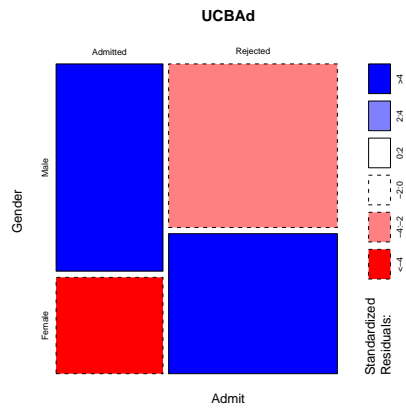
```



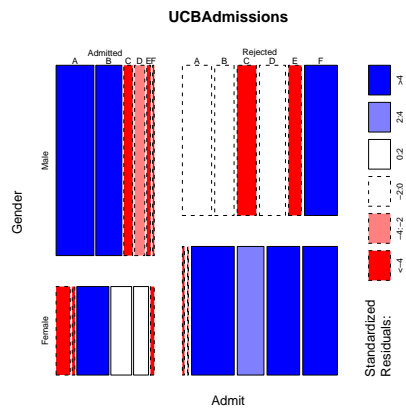
```

> mosaicplot(UCBAd, shade = T)

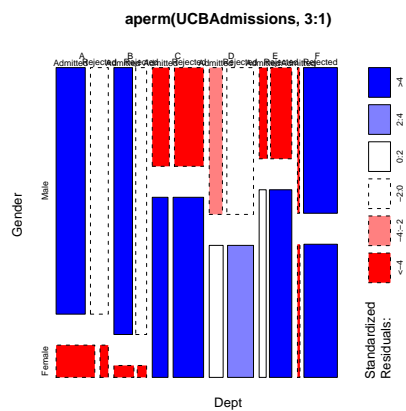
```



```
> mosaicplot(UCBAadmissions, shade = T)
```



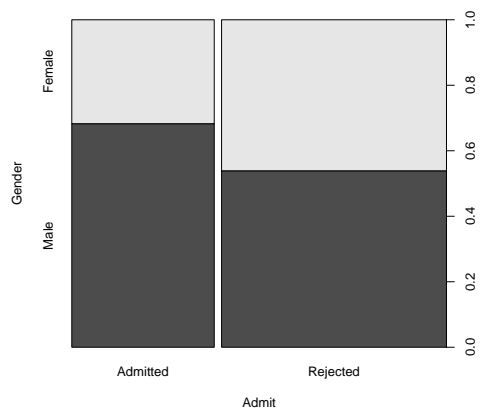
```
> mosaicplot(aperm(UCBAadmissions, 3:1), shade = T)
```




```

> #
> # compare the use of this command without () ...
> #
> spineplot(UCBAd)

```



```

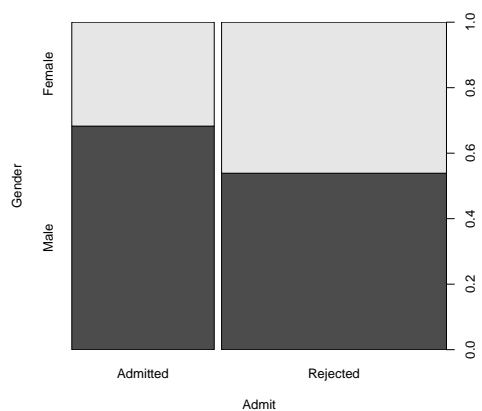
> #
> # ... and with ()
> #
> (spineplot(UCBAd))

```

```

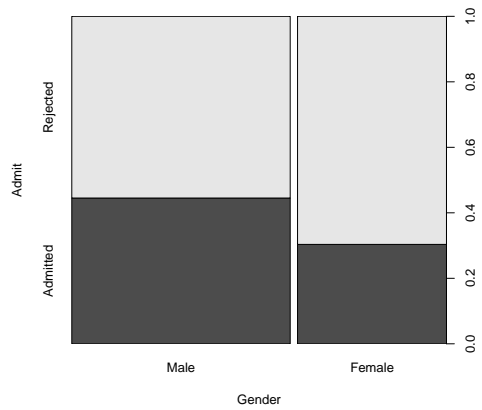
      Gender
Admit  Male Female
Admitted 1198  557
Rejected 1493 1278

```



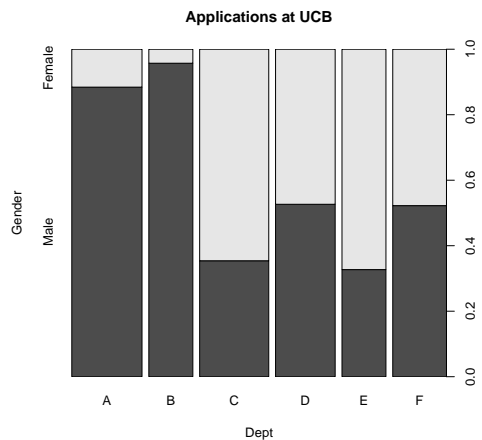
```
> (spineplot(t(UCBAd)))
```

Admit		
Gender	Admitted	Rejected
Male	1198	1493
Female	557	1278



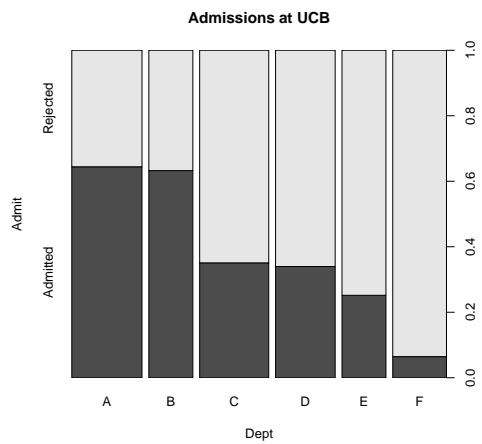
```
> (spineplot(margin.table(UCBAdmissions, c(3, 2)), main = "Applications at UCB"))
```

Gender		
Dept	Male	Female
A	825	108
B	560	25
C	325	593
D	417	375
E	191	393
F	373	341

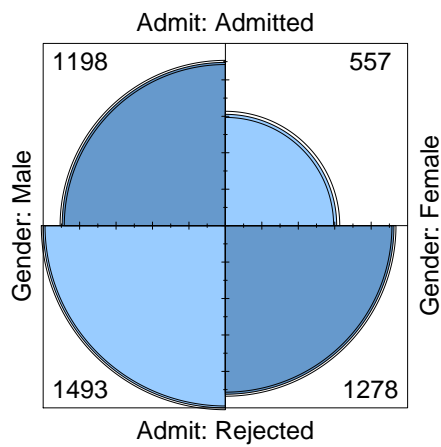


```
> (spineplot(margin.table(UCBAdmissions, c(3, 1)), main = "Admissions at UCB"))
```

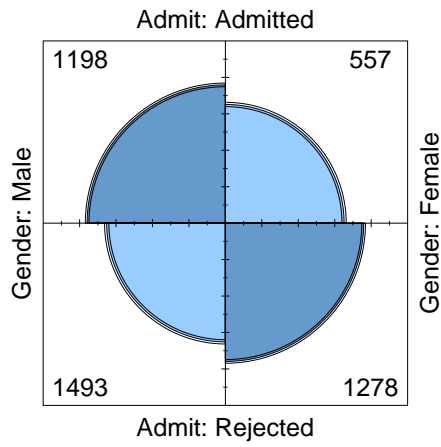
Admit		
Dept	Admitted	Rejected
A	601	332
B	370	215
C	322	596
D	269	523
E	147	437
F	46	668



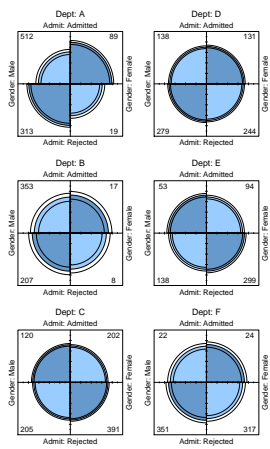
```
> fourfoldplot(UCBAd, std = "a")
```



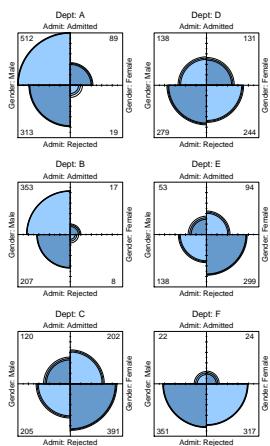
```
> fourfoldplot(UCBAd)
```



```
> fourfoldplot(UCBAAdmissions, std = "m")
```



```
> fourfoldplot(UCBAAdmissions, std = "a")
```



5.5.2 Example 2: Titanic

The R description indicates:

Survival of passengers on the Titanic

This data set provides information on the fate of passengers on the fatal maiden voyage of the ocean liner “Titanic”, summarized according to economic status (class), sex, age and survival.

```
> Titanic
```

```
, , Age = Child, Survived = No
```

```
      Sex
Class Male Female
1st     0     0
2nd     0     0
3rd    35    17
Crew    0     0
```

```
, , Age = Adult, Survived = No
```

```
      Sex
Class Male Female
1st   118     4
2nd   154    13
3rd   387    89
Crew  670     3
```

```
, , Age = Child, Survived = Yes
```

```
      Sex
Class Male Female
1st     5     1
2nd    11    13
3rd    13    14
Crew    0     0
```

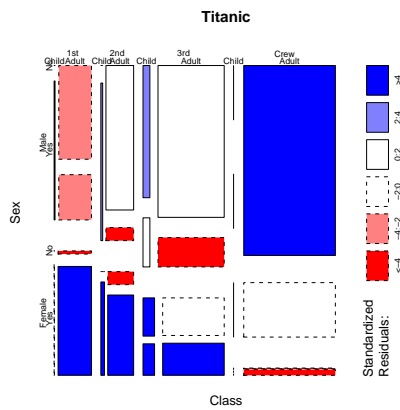
```
, , Age = Adult, Survived = Yes
```

```
Sex
Class Male Female
1st    57    140
2nd    14     80
3rd    75     76
Crew  192     20
```

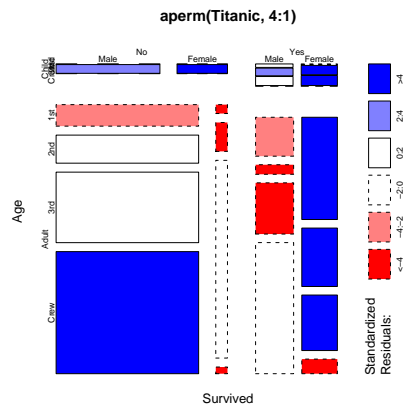
```
> margin.table(Titanic,1)
```

```
Class
1st  2nd  3rd Crew
325 285 706 885
```

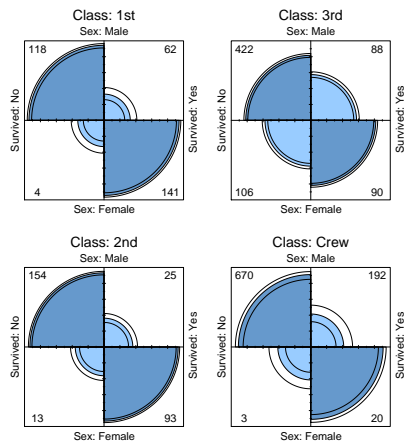
```
> mosaicplot(Titanic, shade = T)
```



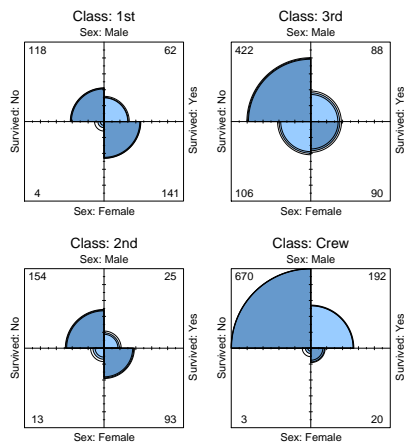
```
> mosaicplot(aperm(Titanic, 4:1), shade = T)
```



```
> Titanic2 <- margin.table(aperm(Titanic, c(2,4,1,3)), 1:3)
> fourfoldplot(Titanic2)
```



```
> fourfoldplot(Titanic2, std = "a")
```



5.5.3 Example 3: HairEyeColor

The R description indicates:

Hair and Eye Color of Statistics Students

Distribution of hair and eye color and sex in 592 statistics students.

```
> HairEyeColor
```

```
, , Sex = Male
```

```
      Eye
Hair   Brown Blue Hazel Green
Black   32   11   10    3
Brown   53   50   25   15
Red     10   10    7    7
Blond    3   30    5    8
```

```
, , Sex = Female
```

```
      Eye
Hair   Brown Blue Hazel Green
Black   36    9    5    2
Brown   66   34   29   14
Red     16    7    7    7
Blond    4   64    5    8
```

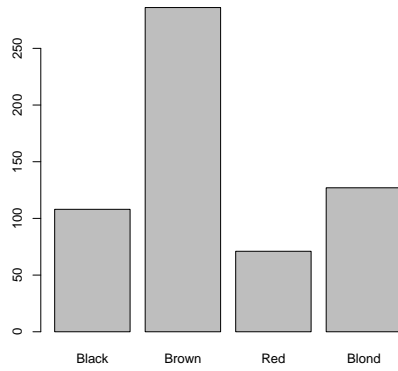
```
> HairCol <- margin.table(HairEyeColor, 1)
```

```
> sort(HairCol)
```

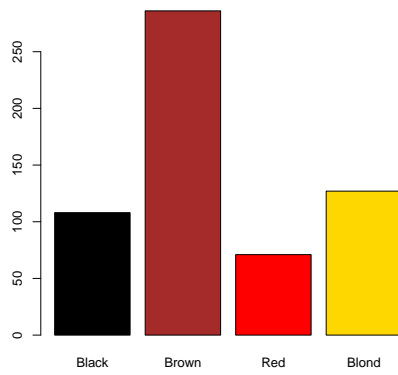
```
Hair
```

```
Red Black Blond Brown
  71  108  127  286
```

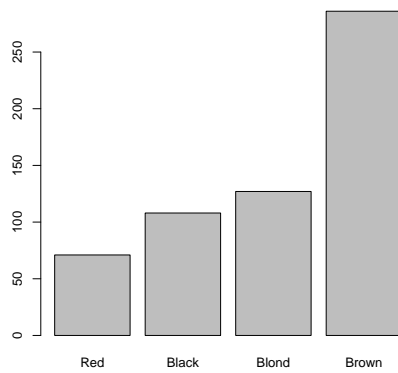
```
> barplot(HairCol)
```

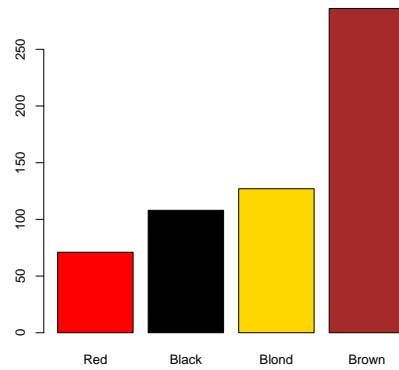
```
> barplot(HairCol, col = c("black", "brown", "red", "gold"))
```



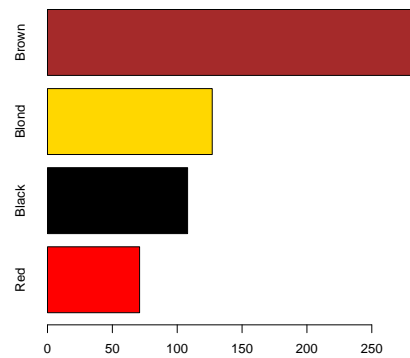
```
> barplot(sort(HairCol))
```



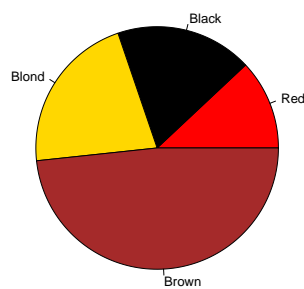
```
> barplot(sort(HairCol), col = c("red", "black", "gold", "brown"))
```



```
> barplot(sort(HairCol), col = c("red", "black", "gold", "brown"), horiz = T)
```



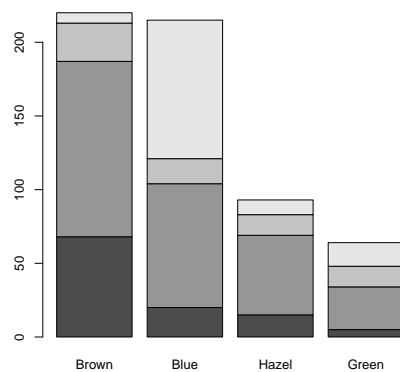
```
> pie(sort(HairCol), col = c("red", "black", "gold", "brown"))
```



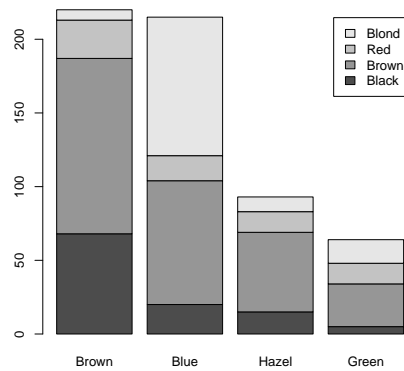
```
> HairEye <- margin.table(HairEyeColor, 1:2)
> HairEye
```

	Eye			
Hair	Brown	Blue	Hazel	Green
Black	68	20	15	5
Brown	119	84	54	29
Red	26	17	14	14
Blond	7	94	10	16

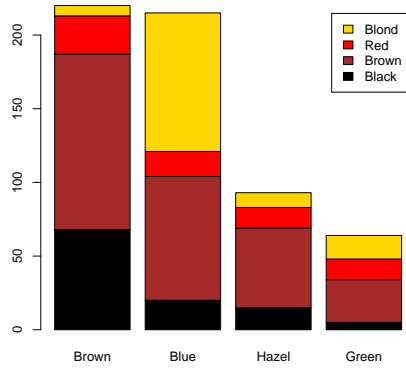
```
> barplot(HairEye)
```



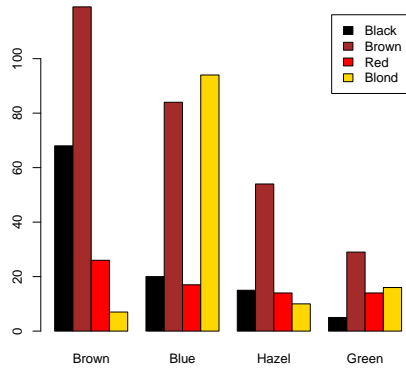
```
> barplot(HairEye, legend.text = T)
```



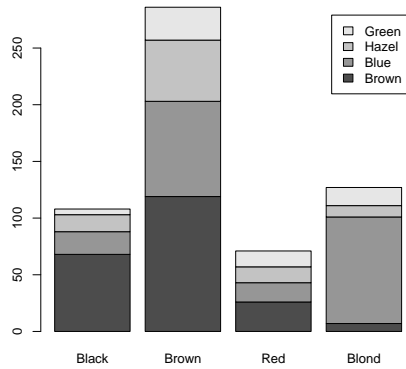
```
> barplot(HairEye, legend.text = T, col = c("black", "brown", "red", "gold"))
```



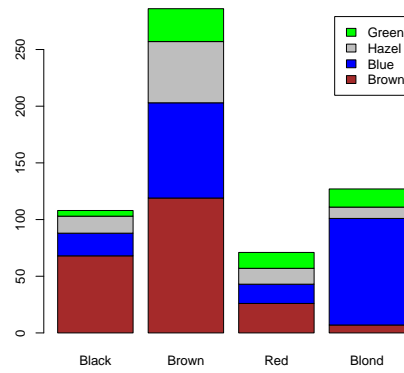
```
> barplot(HairEye, legend.text = T, col = c("black", "brown", "red", "gold"), beside = T)
```



```
> barplot(t(HairEye), legend.text = T)
```



```
> barplot(t(HairEye), legend.text = T, col = c("brown", "blue", "grey", "green"))
```

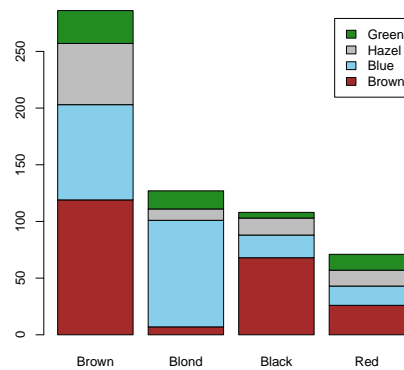


```
> sort(HairCol)
```

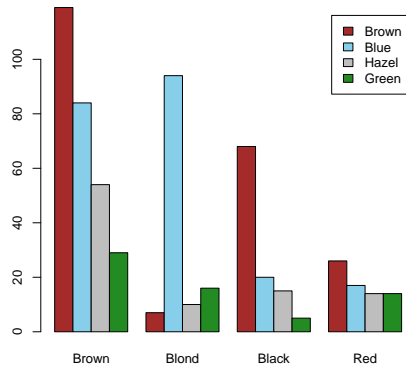
Hair

```
Red Black Blond Brown
  71  108  127  286
```

```
> barplot(t(HairEye[c(2,4,1,3),]), legend.text = T,
+ col = c("brown", "skyblue", "grey", "forestgreen"))
```



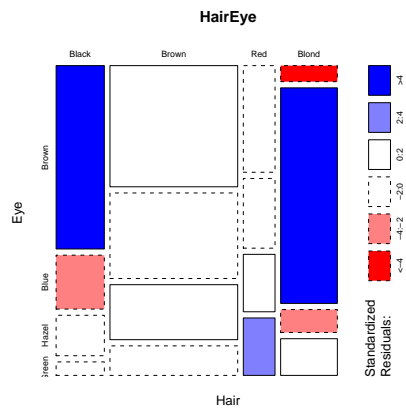
```
> barplot(t(HairEye[c(2,4,1,3),]), legend.text = T,
+ col=c("brown", "skyblue", "grey", "forestgreen"), beside=T)
```



```
> mosaicplot(HairEye)
```



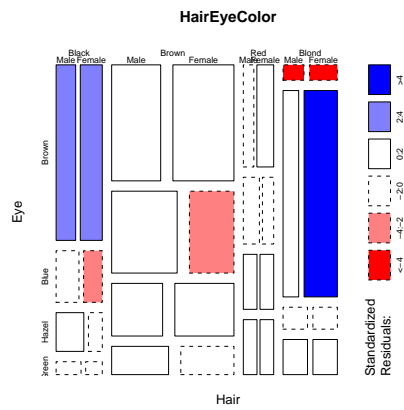
```
> mosaicplot(HairEye, shade = T)
```



```
> mosaicplot(HairEyeColor)
```



```
> mosaicplot(HairEyeColor, shade = T)
```



6 Univariate Plots

6.1 Histograms

Example 1:

Four histograms of the same data set, showing the weights in pounds of 132 professional male athletes.

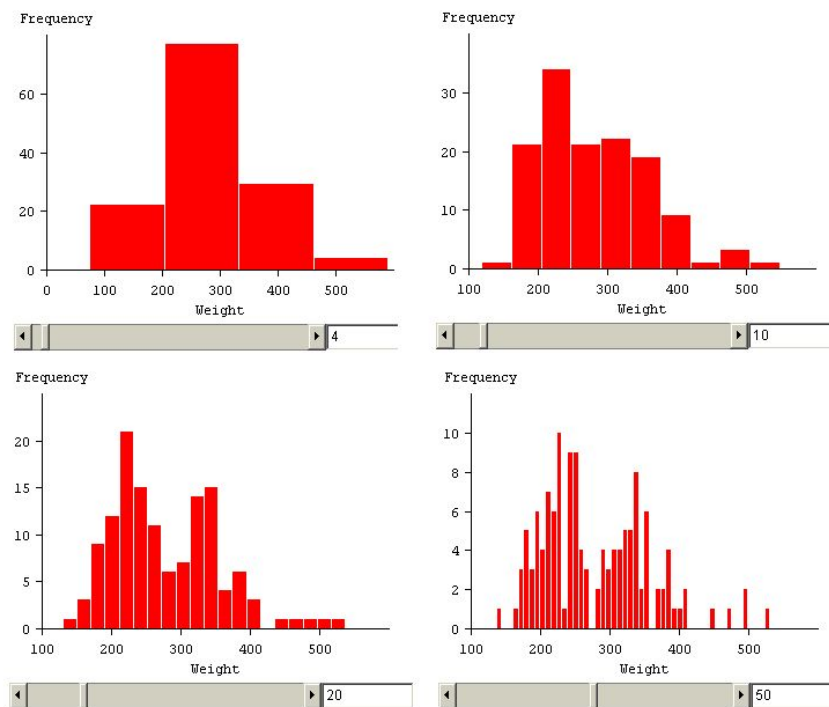


Figure 119: Symanzik, Stat 1040 Lecture Notes, Chapter 3: Four Histograms of the same data set.

Question:

What can we conclude about the underlying data? And which of these four histograms best reveals this fact?

Example 2:

An interactive applet that allows to change the number of classes in a histogram via a slider can be found at <http://www.stat.sc.edu/~west/javahtml/Histogram.html>:

“Histogram Applet:

This applet is designed to teach students how bin widths (or the number of bins) affect a histogram. The histogram below is for the Old Faithful data set. The observations are the duration (in minutes) for eruptions of the Old Faithful geyser in Yellowstone National Park. Students should interactively change the bin width by dragging the arrow underneath the bin width scale. For large bin widths, the bimodal nature of the dataset is hidden, and for small bin widths the plot reduces to a spike at each data point. What bin width do you think provides the best picture of the underlying data?”

Example 3:

Example data sets from Weber (2008):

```
# Weber (2008), Set 1:
```

```
data1 = c(.968, .982, .991, .993, .998, .999, 1.004, 1.004,  
  1.007, 1.010, 1.012, 1.015, 1.017, 1.019, 1.021,  
  1.035, 1.037, 1.037, 1.039, 1.039, 1.042, 1.042,  
  1.047, 1.053, 1.055, 1.059, 1.081, 1.107, 1.1305)
```

```
par(mfrow = c(1, 3))
```

```
hist(data1) #Default
```

```
hist(data1, breaks = 0.9356 + (0:7) * 0.0325) #A
```

```
hist(data1, breaks = 0.9600 + (0:7) * 0.0300) #B
```

```
summary(data1)
```

```
# Weber (2008), Set 2:
```

```
data2 = c(2.05, 2.27, 2.50, 2.95, 3.18, 3.41, 3.64, 3.86, 4.09, 4.32,  
  5.68, 5.91, 6.14, 6.36, 6.59, 6.82, 7.05, 7.50, 7.73, 7.95)
```

```
par(mfrow = c(2, 6))
```

```
hist(data2)
```

```
hist(data2, breaks = 1.425 + (0:3) * 3.2075) #C
```

```
hist(data2, breaks = -1.048 + (0:3) * 3.2075) #D
```

```
hist(data2, breaks = 1.9767 + (0:4) * 1.9789) #E
hist(data2, breaks = 0.1078 + (0:4) * 1.9789) #F

hist(data2, breaks = 1.9829 + (0:5) * 1.4750) #G
hist(data2, breaks = 0.6421 + (0:5) * 1.4750) #H

hist(data2, breaks = 1.9619 + (0:4) * 1.9060) #I
hist(data2, breaks = 0.8944 + (0:4) * 1.9060) #J

hist(data2, breaks = -0.6800 + (0:4) * 2.8400) #K
hist(data2, breaks = 1.9542 + (0:4) * 1.5229) #L

hist(data2, breaks = 1.9619 + (0:7) * .9060) #New 1

summary(data2)
```

Conclusion: “Never believe any statistics you haven’t falsified yourself.”

(<http://sanmateorealestateblog.com/real-estate/statistics-real-estate/never-believe-any-statistics-you-havent-falsified-yourself/>)

The R help page for `hist` indicates:

“The generic function `hist` computes a histogram of the given data values.”

The R help page for the Iris data set indicates:

“This famous (Fisher’s or Anderson’s) iris data set gives the measurements in centimeters of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of 3 species of iris. The species are *Iris setosa*, *versicolor*, and *virginica*.

`iris` is a data frame with 150 cases (rows) and 5 variables (columns) named `Sepal.Length`, `Sepal.Width`, `Petal.Length`, `Petal.Width`, and `Species`.”

Choosing the number of classes for a histogram:

As seen in the previous three examples, a bad choice for the number of classes (`nclass` or `breaks` in the R command) in a histogram or the starting point of an interval and its width can almost entirely hide the most interesting information of the underlying data.

Several suggestions for the number of classes exist and are summarized in Venables & Ripley (2002), p. 112. We define $\text{range} = x_{(n)} - x_{(1)}$, where n represents the number of observations.

- Sturges’ formula (default in R):

$$\text{nclass} = \lceil \log_2 n + 1 \rceil, \quad \text{bin width} = \frac{\text{range}}{\text{nclass}},$$

where $\lceil \dots \rceil$ indicates the ceiling function.

- Scott’s 1979 formula (“`scott`” in R):

$$\text{bin width} = 3.5 \hat{\sigma} n^{-1/3}, \quad \text{nclass} = \frac{\text{range}}{\text{bin width}},$$

where $\hat{\sigma}$ is the estimated standard deviation.

- Freedman and Diaconis 1981 formula (“`fd`” in R):

$$\text{bin width} = 2 \text{IQR} n^{-1/3}, \quad \text{nclass} = \frac{\text{range}}{\text{bin width}},$$

where IQR is the inter-quartile range.

- Sometimes, the use of $n_{class} \approx \sqrt{n}$ is suggested:

<http://www.qimacros.com/qiwizard/how-to-determine-histogram-bin-interval.htm> suggests: *“Take the square root of the number of data points and round up to determine the number of bins required.”*

<http://www.moresteam.com/toolbox/t417.cfm> suggests: *“Calculate the square root of the number of data points and round to the nearest whole number. In the case of our height example, the square root of 50 is 7.07, or 7 when rounded.”*

http://www.micquality.com/introductory_statistics/int08.htm states: *“There are various ways of calculating the number of bins. I find that using the square root of the number of data values gives as good a result as the more complicated methods. The value is usually on the low side, but you can adjust it upwards to get convenient bin boundaries. Treat the calculated number of bins as a starting point, and adjust it as necessary to give the result you prefer.”*

Example:

```
data(iris)
head(iris)
plength <- iris[,3]
n <- length(plength)

par(mfrow = c(3, 2))

hist(plength, freq = F,
     main = "Default (Sturges) Breaks")
hist(plength, breaks = as.integer(sqrt(n)), freq = F,
     main = "sqrt(n) Breaks")
hist(plength, breaks = "scott", freq = F,
     main = "Scott Breaks")
hist(plength, breaks = "fd", freq = F,
     main = "FD Breaks")

nclass.Sturges(plength)
sqrt(n)
nclass.scott(plength)
nclass.FD(plength)
```

```

h <- 3.5 * sd(plength) * n^(-1/3)
h # bin width
range(plength)
tk <- seq(.9, 8, by = h)
tk
length(tk) - 1 # nclass
hist(plength, breaks = tk, freq = F,
     main = "Exact Scott Breaks")

tk2 <- seq(.9, 8, by = h/2)
tk2
length(tk2) - 1 # nclass
hist(plength, breaks = tk2, freq = F,
     main = "Adjusted Exact Scott Breaks")

```

Finally, how do the histograms for the three species look like?

```

par(mfrow = c(2, 2))

hist(plength, main = "all")

hist(plength[1:50], main = "setosa")

hist(plength[51:100], main = "versicolor")

hist(plength[101:150], main = "virginica")

```

Note:

- These various methods (Sturges, Scott, FD, sqrt) provide suggestions for the number of classes only. To enforce particular breaks, we have to provide a vector giving the exact break points between the histogram cells. However, good software will use the suggestions and then make further adjustments to obtain meaningful class breaks for a human reader, e.g., use integers (and multiples of 5 or 10, etc.) as the boundaries.

- Carefully check whether class intervals are left–open or right–open. R class intervals by default are left–open whereas most readers prefer right–open intervals. Also, check in which class interval the minimum and maximum of a data set are included. For continuous data, there will be little differences in the appearance of a histogram, but for discrete data, different settings may result in a dramatically different visual appearance of a histogram. R provides arguments (`include.lowest` and `right`) to adjust these options.
- Wallgren et al. (1996), p. 30, state: **“Since it is relatively complicated both to draw and to read histograms with classes of different size we recommend that, as far as possible, both tables and charts should be made with classes of equal length.”**

6.2 Averaged Shifted Histograms

Symanzik (2004), p. 307, states:

“Scott (1992) provides a general overview on techniques for density estimation, including averaged shifted histograms (ASH) and kernel density estimators, including possible visualization techniques via contour surfaces, (transparent) α -level contours, and contour shells.”

ASH plots were originally introduced in Scott (1985). They are created by averaging several shifted histograms and further smoothing the result. Details are provided in Chapter 5 of Scott (1992).

ASH plots may not be easy to explain to non-statisticians, but they may help to determine which histograms may be closest to the underlying data.

Weber (2008), Set 1:

```
data1 = c(.968, .982, .991, .993, .998, .999, 1.004, 1.004,  
  1.007, 1.010, 1.012, 1.015, 1.017, 1.019, 1.021,  
  1.035, 1.037, 1.037, 1.039, 1.039, 1.042, 1.042,  
  1.047, 1.053, 1.055, 1.059, 1.081, 1.107, 1.1305)
```

```
library(ash)
```

```
f1 = ash1(bin1(data1, nbin = 50), 5) # compute ash estimate
```

```
par(mfrow = c(1, 3))
```

```
hist(data1, freq = F, ylim = c(0,14)) #Default  
lines(f1 , type="l" ) # line plot of estimate
```

```
hist(data1, breaks = 0.9356 + (0:7) * 0.0325, freq = F, ylim = c(0,14)) #A  
lines(f1 , type="l" ) # line plot of estimate
```

```
hist(data1, breaks = 0.9600 + (0:7) * 0.0300, freq = F, ylim = c(0,14)) #B  
lines(f1 , type="l" ) # line plot of estimate
```



```

# Weber (2008), Set 2:

data2 = c(2.05, 2.27, 2.50, 2.95, 3.18, 3.41, 3.64, 3.86, 4.09, 4.32,
  5.68, 5.91, 6.14, 6.36, 6.59, 6.82, 7.05, 7.50, 7.73, 7.95)

f2 = ash1(bin1(data2, nbin = 50), 5) # compute ash estimate

par(mfrow = c(2, 6))

hist(data2, freq = F, ylim = c(0, 0.25))
lines(f2 , type="l" ) # line plot of estimate

hist(data2, breaks = 1.425 + (0:3) * 3.2075, freq = F, ylim = c(0, 0.25)) #C
lines(f2 , type="l" ) # line plot of estimate

hist(data2, breaks = -1.048 + (0:3) * 3.2075, freq = F, ylim = c(0, 0.25)) #D
lines(f2 , type="l" ) # line plot of estimate

hist(data2, breaks = 1.9767 + (0:4) * 1.9789, freq = F, ylim = c(0, 0.25)) #E
lines(f2 , type="l" ) # line plot of estimate

hist(data2, breaks = 0.1078 + (0:4) * 1.9789, freq = F, ylim = c(0, 0.25)) #F
lines(f2 , type="l" ) # line plot of estimate

hist(data2, breaks = 1.9829 + (0:5) * 1.4750, freq = F, ylim = c(0, 0.25)) #G
lines(f2 , type="l" ) # line plot of estimate

hist(data2, breaks = 0.6421 + (0:5) * 1.4750, freq = F, ylim = c(0, 0.25)) #H
lines(f2 , type="l" ) # line plot of estimate

hist(data2, breaks = 1.9619 + (0:4) * 1.9060, freq = F, ylim = c(0, 0.25)) #I
lines(f2 , type="l" ) # line plot of estimate

```

```
hist(data2, breaks = 0.8944 + (0:4) * 1.9060, freq = F, ylim = c(0, 0.25)) #J
lines(f2 , type="l" ) # line plot of estimate
```

```
hist(data2, breaks = -0.6800 + (0:4) * 2.8400, freq = F, ylim = c(0, 0.25)) #K
lines(f2 , type="l" ) # line plot of estimate
```

```
hist(data2, breaks = 1.9542 + (0:4) * 1.5229, freq = F, ylim = c(0, 0.25)) #L
lines(f2 , type="l" ) # line plot of estimate
```

```
hist(data2, breaks = 1.9619 + (0:7) * .9060, freq = F, ylim = c(0, 0.25)) #New 1
lines(f2 , type="l" ) # line plot of estimate
```

6.3 Stem-and-Leaf Plots

The R help page for stem indicates:

“stem produces a stem-and-leaf plot of the values in x.”

Venables & Ripley (2002), p. 113, further specify: “A *stem-and-leaf plot* is an enhanced histogram. The data are divided into bins, but the ‘height’ is replaced by the next digits in order.”

```
sort(plength)
```

```
stem(plength)
```

6.4 Boxplots (or Box-and-Whisker Plots)

The R help page for `boxplot` indicates:

“Produce box-and-whisker plot(s) of the given (grouped) values.

`range`: this determines how far the plot whiskers extend out from the box. If `range` is positive, the whiskers extend to the most extreme data point which is no more than `range` times the interquartile range from the box. A value of zero causes the whiskers to extend to the data extremes.”

The default for `range` is 1.5.

Venables & Ripley (2002), p. 115, further specify: *“A boxplot is a way to look at the overall shape of a set of data. The central box shows the data between the ‘hinges’ (roughly quartiles), with the median represented by a line. ‘Whiskers’ go out to the extremes of the data, and very extreme points are shown by themselves.”*

```
par(mfrow = c(2, 3))
```

```
boxplot(plength)
```

```
boxplot(plength, range = 0)
```

```
boxplot(plength, range = 0.1)
```

```
boxplot(plength ~ iris$Species)
```

```
boxplot(plength ~ iris$Species, range = 0)
```

```
boxplot(plength ~ iris$Species, range = 0.5)
```

6.5 Dot Charts for Univariate Data

The R help page for `dotchart` indicates:

“Draw a Cleveland dot plot.”

The R help page for `UScereal(MASS)` indicates:

“Nutritional and Marketing Information on US Cereals:

The `UScereal` data frame has 65 rows and 11 columns. The data come from the 1993 ASA Statistical Graphics Exposition, and are taken from the mandatory F&DA food label. The data have been normalized here to a portion of one American cup. ”

```
library(MASS) # for cereal data
data(UScereal)
head(UScereal)

Kel.carbs <- UScereal[UScereal$mfr == "K", 7]
Kel.carbs
names(Kel.carbs) <- row.names(UScereal[UScereal$mfr == "K",])
Kel.carbs

dotchart(Kel.carbs)
dotchart(sort(Kel.carbs))
dotchart(sort(Kel.carbs), xlim = c(10, 35),
  xlab = "g carbohydrates per 1 cup serving")
```

Now, activate the `lattice` package and produce similar graphics:

```
library(lattice)

dotplot(Kel.carbs) # from lattice library
dotplot(sort(Kel.carbs), xlim = c(10, 35),
  xlab = "g carbohydrates per 1 cup serving")
```

The R help page for `barley(lattice)` indicates:

“Yield data from a Minnesota barley trial:

Total yield in bushels per acre for 10 varieties at 6 sites in each of two years.”

```
library(lattice) # for barley data
data(barley)
head(barley)

dotplot(variety ~ yield | site, data = barley, groups = year,
  key = simpleKey(levels(barley$year), space = "right"),
  xlab = "Barley Yield (bushels/acre)",
  aspect = 0.5, layout = c(1, 6), ylab = NULL)

levels(barley$site)

# alphabetical sorting of sites (top to bottom)
dotplot(variety ~ yield | site, data = barley, groups = year,
  key = simpleKey(levels(barley$year), space = "right"),
  xlab = "Barley Yield (bushels/acre)",
  aspect = 0.5, layout = c(1, 6), ylab = NULL,
  index.cond = list(c(6,3,4,1,2,5)))
```

Question:

What is the most striking (unusual) feature in these plots? Look carefully!

6.6 Kernel Density Plots for Univariate Data (with Rug Plot)

The R help page for density indicates:

“Kernel Density Estimation:

The (S3) generic function density computes kernel density estimates. Its default method does so with the given kernel and bandwidth for univariate observations. [...]

bw: the smoothing bandwidth to be used. The kernels are scaled such that this is the standard deviation of the smoothing kernel. (Note this differs from the reference books cited below, and from S-PLUS.)

bw can also be a character string giving a rule to choose the bandwidth. See bw.nrd.

The specified (or computed) value of bw is multiplied by adjust.”

The R help page for rug indicates:

“Adds a rug representation (1-d plot) of the data to the plot.”

Chambers & Hastie (1993), p. 548, further specify: *“rug: [...] a univariate histogram or rugplot is displayed along the base of each plot, showing the occurrence of each x-value; ties are broken by jittering.”*

```
par(mfrow = c(3, 2))

plot(density(plength), xlim = c(-1, 9))           # default nrd0 bw
rug(plength, ticksize = 0.05)
plot(density(plength, bw = "nrd"), xlim = c(-1, 9)) # normal reference rule bw

plot(density(plength, bw = "ucv"), xlim = c(-1, 9))
                                     # unbiased cross-validation rule bw

plot(density(plength, bw = "bcv"), xlim = c(-1, 9))
                                     # biased cross-validation rule bw

plot(density(plength, bw = "SJ-ste"), xlim = c(-1, 9))
                                     # Sheather-Jones ("solve-the-equation") bw
```

```

plot(density(plength, bw = "SJ-dpi"), xlim = c(-1, 9))
      # Sheather-Jones ("direct plug-in") bw

par(mfrow = c(3, 2))

plot(density(plength), xlim = c(-1, 9))
plot(density(plength, adjust = 1/2), xlim = c(-1, 9)) #adjust default bandwidth
plot(density(plength, adjust = 1/4), xlim = c(-1, 9))
plot(density(plength, adjust = 1/8), xlim = c(-1, 9))
plot(density(plength, adjust = 2), xlim = c(-1, 9))
plot(density(plength, adjust = 4), xlim = c(-1, 9))

```

Now, use the lattice package and produce similar graphics:

```

densityplot(plength) # lattice, takes all parameters from density (above)
densityplot(plength, n = 512)
densityplot(plength, n = 512, bw = "SJ")

```

n is the “number of points at which density is to be evaluated” and the default is 50.

6.7 Quantile–Quantile Plots (Q–Q Plots)

One of the best ways to compare the distribution of a sample \underline{x} of size n with an assumed theoretical distribution F is to use a Quantile–Quantile Plot (Q–Q Plot). In such a plot, we plot the pairs of points

$$\left(F^{-1} \left(\frac{i - 0.5}{n} \right), x_{(i)} \right), \quad i = 1, \dots, n.$$

Example 1:

Convergence of a t distribution with df degrees of freedom towards a normal distribution.

```
# set seed of random number generator to be able to reproduce results
```

```
set.seed(1234)
```

```
par(mfrow = c(3, 2))
```

```
tdf1 = rt(100, df = 1)
```

```
qqnorm(tdf1)
```

```
qqline(tdf1)
```

```
tdf2 = rt(100, df = 2)
```

```
qqnorm(tdf2)
```

```
qqline(tdf2)
```

```
tdf5 = rt(100, df = 5)
```

```
qqnorm(tdf5)
```

```
qqline(tdf5)
```

```
tdf10 = rt(100, df = 10)
```

```
qqnorm(tdf10)
```

```
qqline(tdf10)
```

```
tdf20 = rt(100, df = 20)
```

```
qqnorm(tdf20)
```

```
qqline(tdf20)
```

```
tdf30 = rt(100, df = 30)
```

```
qqnorm(tdf30)
qqline(tdf30)
```

Note:

The closer the points from the sample fall to a straight line, the closer the sample distribution and the theoretical distribution are related. However, here, the greater spread of the extreme quantiles for the sample (for $df = 1, 2, 5, 10, 20$) is an indicator of a long-tailed distribution.

Example 2:

Recall: What is the relationship between exponential distributions and Gamma distributions? Verify this graphically!

```
par(mfrow = c(3, 2))
```

```
set.seed(1234)
```

```
exp1 = rexp(100, rate = 1)
plot(qgamma(ppoints(exp1), 1, 1), sort(exp1))
abline(0, 1)
```

```
exp2 = rexp(100, rate = 2)
plot(qgamma(ppoints(exp2), 1, 2), sort(exp2))
abline(0, 1)
```

```
exp5 = rexp(100, rate = 5)
plot(qgamma(ppoints(exp5), 1, 5), sort(exp5))
abline(0, 1)
```

```
exp10 = rexp(100, rate = 10)
plot(qgamma(ppoints(exp10), 1, 10), sort(exp10))
abline(0, 1)
```

```
# Some major misspecifications
```

```
# a) Swapping shape and rate parameters
plot(qgamma(ppoints(exp2), 2, 1), sort(exp2))
```

```
abline(0, 1)
```

```
# b) Using 1/rate instead
```

```
plot(qgamma(ppoints(exp2), 1, 1/2), sort(exp2))
```

```
abline(0, 1)
```

Example 3:

Compare iris plength sample data with an underlying assumed normal distribution. For which of the species is the assumption of normality justified?

```
par(mfrow = c(2, 2))
```

```
qqnorm(plength)
```

```
qqnorm(plength[1:50])
```

```
qqnorm(plength[51:100])
```

```
qqnorm(plength[101:150])
```

6.8 Empirical Cumulative Distribution Functions (ECDFs)

Recall from Stat 6720:

Definition 7.1.3:

Let X_1, \dots, X_n be a sample of size n from a population with distribution F . The function

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^n I_{(-\infty, x]}(X_i)$$

is called **empirical cumulative distribution function (empirical cdf, ECDF)**. ■

Theorem 7.1.7: Glivenko–Cantelli Theorem

$\hat{F}_n(x)$ converges uniformly to $F(x)$, i.e., it holds for all $\epsilon > 0$ that

$$\lim_{n \rightarrow \infty} P\left(\sup_{-\infty < x < \infty} |\hat{F}_n(x) - F(x)| > \epsilon\right) = 0.$$

■

Verify this theorem for samples from a normal distribution:

```
par(mfrow = c(2, 2))

set.seed(1234)

xvals = seq(-4, 4, 0.01)

norm10 = rnorm(10)
plot(ecdf(norm10), xlim = c(-4, 4))
lines(xvals, pnorm(xvals))

norm25 = rnorm(25)
plot(ecdf(norm25), xlim = c(-4, 4))
lines(xvals, pnorm(xvals))

norm100 = rnorm(100)
plot(ecdf(norm100), xlim = c(-4, 4))
lines(xvals, pnorm(xvals))
```

```

norm1000 = rnorm(1000)
plot(ecdf(norm1000), xlim = c(-4, 4))
lines(xvals, pnorm(xvals))

# Animation

par(mfrow = c(1, 1))

normgrow = NULL
for (i in 1:20)
{
  normgrow = c(normgrow, rnorm(10))
  plot(ecdf(normgrow), xlim = c(-4, 4), main = length(normgrow))
  lines(xvals, pnorm(xvals))
  Sys.sleep(1)
}

```

Example:

And here the ECDF for iris plength.

```
plot(ecdf(plength))
```

6.9 Graphics and Small Sample Sizes

Worksheet

Your Name: _____

Question:

The data shown in these four histograms originate from which distribution?

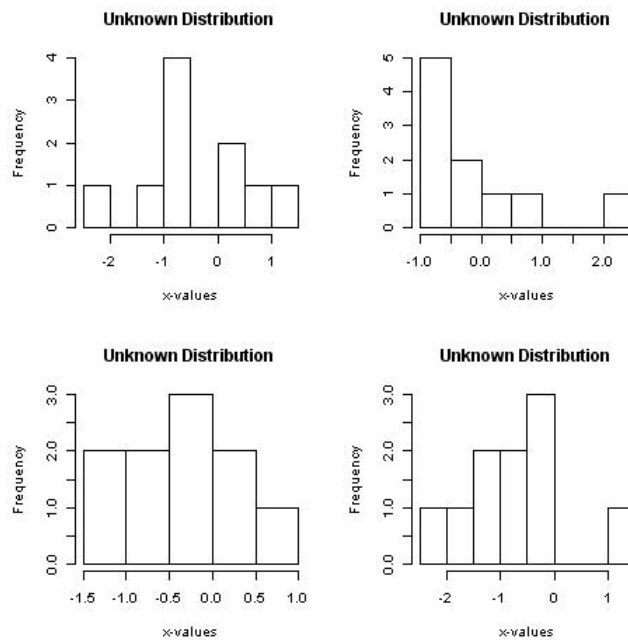


Figure 120: Histograms for data from four unknown distributions.

The corresponding distributions are:

Upper left: _____

Upper right: _____

Lower left: _____

Lower right: _____

Worksheet

Your Name: _____

Question:

Do the data shown in these four qqplots follow a normal distribution?

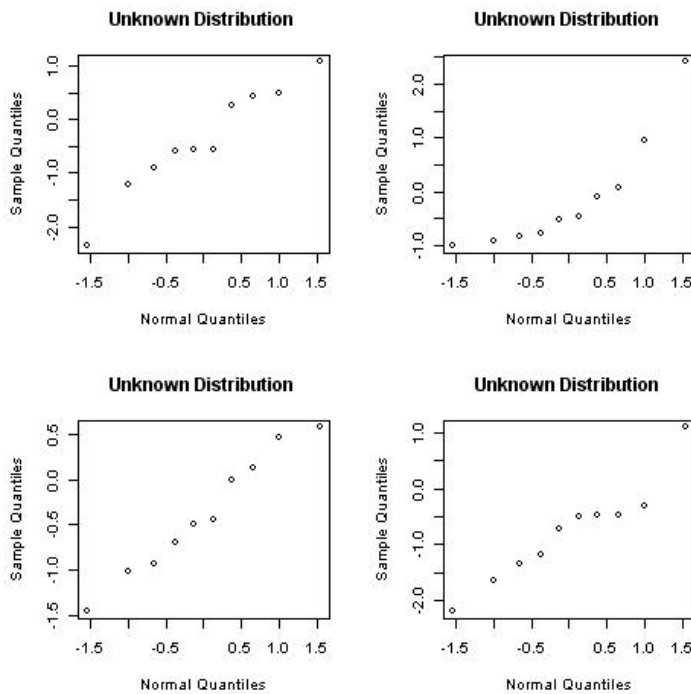


Figure 121: Normal QQplots for data from four unknown distributions.

So, does a particular qqplot suggest that the data originate from a normal distribution?

Circle your answer:

Upper left: **yes** / **no**

Upper right: **yes** / **no**

Lower left: **yes** / **no**

Lower right: **yes** / **no**

Answers:

The following R code was used to create the figures shown on the previous two pages:

```
jpeg("Chapter5_unknown_hist.jpg")

par(mfrow = c(2, 2))
set.seed(1234)
xvect1 = NULL

for(i in 1:4)
{
  x = rnorm(10)
  xvect1 = c(xvect1, x)
  hist(x, main = "Unknown Distribution",
       xlab = "x-values")
}

dev.off()

jpeg("Chapter5_unknown_qqplot.jpg")

par(mfrow = c(2, 2))
set.seed(1234)
xvect2 = NULL

for(i in 1:4)
{
  x = rnorm(10)
  xvect2 = c(xvect2, x)
  qqnorm(x, main = "Unknown Distribution",
        xlab = "Normal Quantiles")
}

dev.off()
```


When we jointly plot all 40 observations, we start to see that the underlying distribution indeed is a normal distribution. In fact, all eight plots on the previous two pages show 10 samples each drawn from the standard normal distribution!

```
# Plot all data combined
```

```
par(mfrow = c(1, 1))  
hist(xvect1)
```

```
par(mfrow = c(1, 1))  
qqnorm(xvect2)
```

6.10 Further Reading

Additional sources for Trellis Graphics are:

- <http://cm.bell-labs.com/cm/ms/departments/sia/project/trellis/index.html>
- Murrell (2006), Chapter 4
- William G. Jacoby's Web page on Dot Plots: <http://polisci.msu.edu/jacoby/research/dotplots/dotlist.htm>

7 Bivariate Plots

7.1 Scatterplots

Wallgren et al. (1996), p. 46, state:

“Scatterplots are used to show relationship (causal relationship or covariance) between two *quantitative* variables. The data consists of a number of pairs of coordinates (x, y) . Each pair of coordinates is indicated by a dot or circle in a system of coordinates.”

In the next example, we resume work with the iris data introduced in the previous chapter.

Example 1:

```
data(iris)
head(iris)
slength <- iris[,1]
swidth <- iris[,2]
species <- iris[,5]

par(mfrow = c(2, 2))

plot(slength, swidth,
     xlim = c(4.0, 8.0), ylim = c(2.0, 5.0))

plot(slength, swidth, pch = 21, bg = unclass(species),
     xlim = c(4.0, 8.0), ylim = c(2.0, 5.0))
legend("topright", levels(species), pch = 21, pt.bg = 1:3)

plot(jitter(slength), jitter(swidth), pch = 22, bg = unclass(species),
     xlim = c(4.0, 8.0), ylim = c(2.0, 5.0))
legend("topright", levels(species), pch = 22, pt.bg = 1:3)

plot(jitter(slength, 5), jitter(swidth, 5), pch = unclass(species),
     xlim = c(4.0, 8.0), ylim = c(2.0, 5.0))
legend("topright", levels(species), pch = 1:3)
```

In the next example, we look at samples of sizes 10, 100, 1,000, 10,000, 100,000, and 1,000,000. from a bivariate normal distribution with correlation $\rho = 0$. Up to 1,000 observations, we can still identify regions of higher and lower density, but for 10,000 (and more) observations, we just have a black center in the plot. We will need some different graphical representation when we have too many observations (and thus too much overplotting) for an ordinary scatterplot.

Example 2:

```
set.seed(1234)

par(mfrow = c(2, 3), pty = "s")

x10 = rnorm(10)
y10 = rnorm(10)
plot(x10, y10,
     xlim = c(-4.0, 4.0), ylim = c(-4.0, 4.0))

x100 = rnorm(100)
y100 = rnorm(100)
plot(x100, y100,
     xlim = c(-4.0, 4.0), ylim = c(-4.0, 4.0))

x1000 = rnorm(1000)
y1000 = rnorm(1000)
plot(x1000, y1000,
     xlim = c(-4.0, 4.0), ylim = c(-4.0, 4.0))

x10000 = rnorm(10000)
y10000 = rnorm(10000)
plot(x10000, y10000,
     xlim = c(-4.0, 4.0), ylim = c(-4.0, 4.0))

x100000 = rnorm(100000)
y100000 = rnorm(100000)
plot(x100000, y100000,
     xlim = c(-4.0, 4.0), ylim = c(-4.0, 4.0))
```

```
x1000000 = rnorm(1000000)
y1000000 = rnorm(1000000)
plot(x1000000, y1000000,
      xlim = c(-4.0, 4.0), ylim = c(-4.0, 4.0))
```

7.2 Hexagon Binning

(Based on Carr et al. (1987))

The R help page for hexbin indicates:

“Creates a “hexbin” object. Basic components are a cell id and a count of points falling in each occupied cell.”

Let us apply hexbin to data from our bivariate normal distribution for 10,000 and 1,000,000 observations.

Example 3:

```
library(hexbin)
library(colorspace)

plot(hexbin(x10000, y10000))

plot(hexbin(x10000, y10000), colramp = heat.colors)

plot(hexbin(x10000, y10000), style = "lattice")

plot(hexbin(x10000, y10000), style = "centroid")

plot(hexbin(x1000000, y1000000), colramp = terrain.colors)

plot(hexbin(x1000000, y1000000, xbins = 16), colramp = topo.colors)
```

7.3 Bivariate Histograms

Execute this code once to create the `bivhist3d` function, written by Mike Minnotte:

```
bivhist3d <- function(x, y, nbins = nclass.Sturges(x),
  col = c(heat.colors(19)), xlab = "x", ylab = "y")

# Bivariate histogram plot. Uses the gplots and rgl libraries.
# Written by Mike Minnotte
{
h2d <- hist2d(x, y, nbins, show = F)
xb <- h2d$x
yb <- h2d$y
count <- h2d$counts
xs <- (xb[2]-xb[1])/2
ys <- (yb[2]-yb[1])/2
xb <- (xb+xs)
yb <- (yb+ys)

if (length(col)>1)
  zcol <- matrix(cut(count, length(col), labels=F), length(xb))
else
  zcol <- matrix(1, length(xb), length(yb))

cube <- cube3d()
clear3d()
bg3d(col="grey")

xr <- (max(xb) - min(xb))
yr <- (max(yb) - min(yb))
zr <- (max(count))
mr <- max(xr, yr, zr)*.8

aspect3d(mr/xr, mr/yr, mr/zr)

for (i in 1:length(xb))
  for (j in 1:length(yb))
  {
    scube <- scale3d(cube, xs, ys, count[i,j]/2)
    tcube <- translate3d(scube, xb[i], yb[j], count[i,j]/2)
    if (count[i,j] == 0)
      shade3d(tcube, color = "black")
    else
      shade3d(tcube, color = col[zcol[i,j]])
  }

axes3d()
title3d(xlab = xlab, ylab = ylab, zlab = "count")
return()
}
```

Let us apply `bivhist3d` to data from our bivariate normal distribution for 10,000 observations.

```
library(gplots)
library(rgl)
```

```
bivhist3d(x10000, y10000, col = "red")
```

```
bivhist3d(x10000, y10000)
```

```
bivhist3d(x10000, y10000, col = terrain.colors(15))
```

If you want to create a snapshot of an interesting 3D view, use the command:

```
rgl.snapshot("bivhist1.png")
```

Finally, let us do some automatic animation:

```
start <- proc.time()[3]
while ((i <- 36*(proc.time()[3] - start)) < 360)
{
  rgl.viewpoint(i, 75, fov = 10)
}
```


8 Trivariate Plots

8.1 Scatterplot Matrix

The R help page for `pairs` indicates:

“A matrix of scatterplots is produced. [...] The ij th scatterplot contains $x[i]$ plotted against $x[j]$.”

The R help page for `states` indicates:

“Data sets related to the 50 states of the United States of America. [...] `state.x77`: matrix with 50 rows and 8 columns giving the following statistics in the respective columns.

Population: population estimate as of July 1, 1975
Income: per capita income (1974)
Illiteracy: illiteracy (1970, percent of population)
Life Exp: life expectancy in years (1969–71)
Murder: murder and non-negligent manslaughter rate per 100,000 population (1976)
HS Grad: percent high-school graduates (1970)
Frost: mean number of days with minimum temperature below freezing (1931–1960) in capital or large city
Area: land area in square miles”

Example 1:

```
data(state)
head(state.x77)

tristate <- state.x77[, c(3,6,5)] # trivariate data - illiteracy, hs grad, murder

illiteracy <- tristate[,1]
grad <- tristate[,2]
murder <- tristate[,3]

pairs(tristate) # pairwise scatterplot
pairs(tristate, col = unclass(state.region))
```

```
pairs(tristate, pch = 21, bg = unclass(state.region))
pairs(tristate, pch = 21, bg = unclass(state.region), cex = 2)

pairs(tristate, pch = 21, bg = unclass(state.region), panel = panel.smooth)
```

8.2 3D Scatterplots

Example 2:

```
## lattice

library(lattice)

cloud(murder~illiteracy*grad)

## scatterplot3d

library(scatterplot3d)

scatterplot3d(tristate)

scatterplot3d(tristate, highlight.3d = T)

## rgl

library(rgl)

plot3d(tristate)

plot3d(tristate, type = "s")

plot3d(tristate, type = "s", radius = 0.5)

plot3d(tristate, type = "h")
```

```
plot3d(tristate, type = "s", radius = 0.5, col = "red")

plot3d(tristate, type = "s", radius = 0.5,
       col = c("red", "yellow", "blue", "green")[unclass(state.region)])

text3d(illiteracy, grad, murder + 1.0, text = state.name)
```

Example 3:

The R help page for `randu` indicates:

“Random Numbers from Congruential Generator RANDU: 400 triples of successive random numbers were taken from the VAX FORTRAN function RANDU running under VMS 1.5.”

```
data(randu)
head(randu)

pairs(randu)

plot3d(randu, type = "s", radius = 0.025, col = "red")
```

8.3 Co-Plots

Cleveland (1993), p. 182, states:

“The concept of *conditioning* [...] is a fundamental one that forms the basis of a number of graphical methods developed in the past [...]. And it forms the basis for the *conditioning plot*, or *coplot*, a particularly powerful visualization tool for studying how a response depends on two or more factors. ”

Example 4:

```
# create dataframe
tristatenew = matrix(c(murder, illiteracy, grad), 50, 3)
colnames(tristatenew) = c("murder", "illiteracy", "grad")
```

```

tristatnewdf = as.data.frame(tristatnew)

coplot(murder~illiteracy|grad, data = tristatnewdf)

coplot(grad~illiteracy|murder, data = tristatnewdf)

m.interval = co.intervals(tristatnewdf$murder, number = 6, overlap = 0)
coplot(grad~illiteracy|murder, data = tristatnewdf, given.values = m.interval)

m.interval = co.intervals(tristatnewdf$murder, number = 6, overlap = 0.25)
coplot(grad~illiteracy|murder, data = tristatnewdf, given.values = m.interval)

m.interval = co.intervals(tristatnewdf$murder, number = 4, overlap = 0)
coplot(grad~illiteracy|murder, data = tristatnewdf, given.values = m.interval)

m.interval = co.intervals(tristatnewdf$murder, number = 4, overlap = 0)
coplot(grad~illiteracy|murder, data = tristatnewdf, given.values = m.interval,
       rows = 1)

m.interval = co.intervals(tristatnewdf$murder, number = 9, overlap = 0)
coplot(grad~illiteracy|murder, data = tristatnewdf, given.values = m.interval)

m.interval = co.intervals(tristatnewdf$murder, number = 4, overlap = 0)
coplot(grad~illiteracy|murder, data = tristatnewdf, given.values = m.interval,
       panel = function(x, y, ...) panel.smooth(x, y, span = 0.3, ...))

m.interval = co.intervals(tristatnewdf$murder, number = 4, overlap = 0)
coplot(grad~illiteracy|murder, data = tristatnewdf, given.values = m.interval,
       panel = function(x, y, ...) panel.smooth(x, y, span = 0.7, ...))

m.interval = co.intervals(tristatnewdf$murder, number = 4, overlap = 0)
coplot(grad~illiteracy|murder, data = tristatnewdf, given.values = m.interval,
       panel = function(x, y, ...) panel.smooth(x, y, span = 1.5, ...))

```

8.4 Trivariate Density Estimation

Execute this code once to create the tkde (three-dimensional kernel density estimation) function, written by Mike Minnotte:

```
tkde <- function(x, y=NULL, z=NULL, hx=NULL, hy=NULL, hz=NULL,
xlist=NULL, ylist=NULL, zlist=NULL, extend=.1,
nx=20, ny=20, nz=20, adjust=c(1,1,1),
xlab=NULL, ylab=NULL, zlab=NULL,...)

{data <- xyz.coords(x,y,z)
x <- data$x
y <- data$y
z <- data$z

n <- length(x)

if (is.null(hx)) hx <- bw.SJ(x,method="dpi")
if (is.null(hy)) hy <- bw.SJ(y,method="dpi")
if (is.null(hz)) hz <- bw.SJ(z,method="dpi")

if (is.null(xlist) )
{dx <- (max(x)-min(x))*extend
xlist <-seq(min(x)-dx,max(x)+dx,length=nx)}
else nx <- length(xlist)

if (is.null(ylist) )
{dy <- (max(y)-min(y))*extend
ylist <-seq(min(y)-dy,max(y)+dy,length=ny)}
else ny <- length(ylist)

if (is.null(zlist) )
{dz <- (max(z)-min(z))*extend
zlist <-seq(min(z)-dz,max(z)+dz,length=nz)}
else nz <- length(zlist)

f <- array(0, c(nx,ny,nz))

kx <- matrix(0, n, nx)
for (i in 1:nx) kx[,i]<-dnorm(x,xlist[i],hx*adjust[1])

ky <- matrix(0, n, ny)
for (j in 1:ny) ky[,j]<-dnorm(y,ylist[j],hy*adjust[2])

kz <- matrix(0, n, nz)
for (k in 1:nz) kz[,k]<-dnorm(z,zlist[k],hz*adjust[3])

for (i in 1:nx) for (j in 1:ny) for (k in 1:nz)
f[i,j,k] <- mean(kx[,i]*ky[,j]*kz[,k])

maxf <- max(f)
levs <- maxf*c(.99,.9,.75,.5,.25,.1)

plot3d(x,y,z,type='n')
```

```
contour3d(f, levs, xlist, ylist, zlist,  
color=c("black", "red", "orange", "yellow", "cyan", "blue"),  
alpha=c(1, 1, .4, .4, .25, .15), add=T, ...)  
#decorate3d()  
  
invisible()  
}
```

Example 5:

Let us apply tkde to a subset of the states data.

```
library(rgl)
library(misc3d)

tkde(illiteracy, grad, murder)
```

Example 6:

Let us apply tkde to data from a multivariate normal distribution.

```
# 3 independent standard normals

mnorm1 <- matrix(rnorm(3000), ncol = 3)

tkde(mnorm1)

tkde(mnorm1, adjust = c(1.5, 1.5, 1.5))

tkde(mnorm1, adjust = c(2, 2, 2))

## dependent normals

library(MASS)
Sigma <- matrix(c(1,.8,.8,.8,1,.8,.8,.8,1),3,3) # Covariance matrix
Sigma

mnorm2 <- mvrnorm(1000, mu=rep(0,3), Sigma = Sigma) # MASS library

tkde(mnorm2)

tkde(mnorm2, adjust = c(1.5, 1.5, 1.5))
```

Example 7:

Let us work with the ethanol data from the lattice library.

The R help page for ethanol indicates:

“Engine exhaust fumes from burning ethanol. Ethanol fuel was burned in a single-cylinder engine. For various settings of the engine compression and equivalence ratio, the emissions of nitrogen oxides were recorded. A data frame with 88 observations on the following 3 variables.

NOx: Concentration of nitrogen oxides (NO and NO2) in micrograms/J.

C: Compression ratio of the engine.

E: Equivalence ratio — a measure of the richness of the air and ethanol fuel mixture.

```
library(lattice)
```

```
data(ethanol)
```

```
head(ethanol)
```

```
compress<-ethanol$C
```

```
equiv<-ethanol$E
```

```
NOx<-ethanol$NOx
```

```
plot3d(compress, equiv, NOx, type = "s", radius = 0.15, col="red")
```


Execute this code once to create the `trilpr` and `trillr` functions for local polynomial regression estimates, written by Mike Minnotte:

```
#####
#
# Functions for calculating and plotting trivariate local
# polynomial regression estimates.
#
# Save as text, then source into R
#
# trilpr - local polynomial regression of arbitrary order
# trillr - local linear regression, much faster than trilpr
#
#####

trilpr<-function(x,y,z,xh,yh,p=1,xlist=NULL,ylist=NULL,extendx=.1,nx=51,
  extendy=.1,ny=51,doplot=T,xlab='x',ylab='y',zlab='m(x,y)',phi=30,...)

{# trivariate local polynomial regression (loess) estimate
# x, y - data (explanatory variables)
# z - response variable
# xh, yh - bandwidth (smoothing parameters; standard deviation of normal kernel)
# p - order of locally fitted polynomials (default - linear)
# xlist, ylist - points of evaluation
# extendx, extendy - if xlist (ylist) NULL, how far beyond the range of
# the data to calculate estimate
# nx, ny - number of points of evaluation
# doplot - if T, plot perspective plot of results, else return (x,y) list for
# later plotting
# xlab, ylab, zlab - axis labels
if (length(x)!=length(y) | length(x)!=length(z))
  stop("Lengths of x, y, and z must be equal")
if (is.null(xlist))
  {xr<-range(x)
  xd<-(xr[2]-xr[1])*extendx
  xlist<-seq(xr[1]-xd,xr[2]+xd,length=nx)}
else
  nx<-length(xlist)
if (is.null(ylist))
  {yr<-range(y)
  yd<-(yr[2]-yr[1])*extendy
  ylist<-seq(yr[1]-yd,yr[2]+yd,length=ny)}
else
  ny<-length(ylist)

par(terr=-1)
mhat<-matrix(0,nx,ny)
n<-length(x)
Y<-matrix(z,n,1)

for (i in 1:nx) for (j in 1:ny)
  {W<-dnorm(x,xlist[i],xh)*dnorm(y,ylist[j],yh)
  X<-matrix(rep(1,n),ncol=1)
  if (p > 0) for (k in 1:p) for (m in 0:k)
    X<-cbind(X,(x-xlist[i])^(k-m)*(y-ylist[j])^m)
```

```

betas<-lsfit(X,Y,wt=W,intercept=F)$coef
mhat[i,j]<-betas[1]}

if (doplot) {
persp(xlist,ylist,mhat,phi=phi,...)
return()}
else {
fitted.values<-rep(0,n)
residuals<-rep(0,n)
for (i in 1:n)
{W<-dnorm(x,x[i],xh)*dnorm(y,y[i],yh)
X<-matrix(rep(1,n),ncol=1)
if (p > 0) for (k in 1:p) for (m in 0:k)
X<-cbind(X,(x-x[i])^(k-m)*(y-y[i])^m)
betas<-lsfit(X,Y,wt=W,intercept=F)$coef
fitted.values[i]<-betas[1]
residuals[i]<-z[i]-fitted.values[i]}
return(list(x=xlist,y=ylist,z=mhat,fitted.values=fitted.values,
residuals=residuals))}
}

#####

trillr<-function(x,y,z,xh,yh,xlist=NULL,ylist=NULL,extendx=.1,nx=51,
extendy=.1,ny=51,doplot=T,xlab='x',ylab='y',zlab='m(x,y)',phi=30,...)

{# trivariate local linear regression (loess) estimate
# x, y - data (explanatory variables)
# z - response variable
# xh, yh - bandwidth (smoothing parameters; standard deviation of normal kernel)
# xlist, ylist - points of evaluation
# extendx, extendy - if xlist (ylist) NULL, how far beyond the range of
# the data to calculate estimate
# nx, ny - number of points of evaluation
# doplot - if T, plot perspective plot of results, else return (x,y) list for
# later plotting
# xlab, ylab, zlab - axis labels
if (length(x)!=length(y) | length(x)!=length(z))
stop("Lengths of x, y, and z must be equal")
if (is.null(xlist))
{xr<-range(x)
xd<-(xr[2]-xr[1])*extendx
xlist<-seq(xr[1]-xd,xr[2]+xd,length=nx)}
else
nx<-length(xlist)
if (is.null(ylist))
{yr<-range(y)
yd<-(yr[2]-yr[1])*extendy
ylist<-seq(yr[1]-yd,yr[2]+yd,length=ny)}
else
ny<-length(ylist)

par(err=-1)
mhat<-matrix(0,nx,ny)

```

```

n<-length(x)
Y<-matrix(z,n,1)

for (i in 1:nx)
{Kx<-dnorm(x,xlist[i],xh)
dx<-x-xlist[i]
dx2<-dx^2
for (j in 1:ny)
{K<-Kx*dnorm(y,ylist[j],yh)
dy<-y-ylist[j]
T0<-sum(K)
Tx<-sum(K*dx)
Txx<-sum(K*dx2)
Ty<-sum(K*dy)
Tyy<-sum(K*dy^2)
Txy<-sum(K*dx*dy)
Tz<-sum(K*z)
Txz<-sum(K*dx*z)
Tyz<-sum(K*dy*z)
Sxxyy<-Txx*Tyy-Txy^2
Sxyy<-Ty*Txy-Tx*Tyy
Sxxy<-Tx*Txy-Ty*Txx
mhat[i,j]<-(Tz*Sxxyy+Txz*Sxyy+Tyz*Sxxy)/
(T0*Sxxyy+Tx*Sxyy+Ty*Sxxy)}
if (doplot) {
persp(xlist,ylist,mhat,phi=phi,...)
return()}
else {
fitted.values<-rep(0,n)
residuals<-rep(0,n)
for (i in 1:n)
{K<-dnorm(x,x[i],xh)*dnorm(y,y[i],yh)
dx<-x-x[i]
dy<-y-y[i]
T0<-sum(K)
Tx<-sum(K*dx)
Txx<-sum(K*dx^2)
Ty<-sum(K*dy)
Tyy<-sum(K*dy^2)
Txy<-sum(K*dx*dy)
Tz<-sum(K*z)
Txz<-sum(K*dx*z)
Tyz<-sum(K*dy*z)
Sxxyy<-Txx*Tyy-Txy^2
Sxyy<-Ty*Txy-Tx*Tyy
Sxxy<-Tx*Txy-Ty*Txx
fitted.values[i]<-(Tz*Sxxyy+Txz*Sxyy+Tyz*Sxxy)/
(T0*Sxxyy+Tx*Sxyy+Ty*Sxxy)
residuals[i]<-z[i]-fitted.values[i]}
return(list(x=xlist,y=ylist,z=mhat,fitted.values=fitted.values,
residuals=residuals))}
}

```

```
NOx.fit <- trillr(compress, equiv, NOx, 1, 0.07, doplot = F)
persp(NOx.fit, phi = 30, theta = -70)

persp3d(NOx.fit, col = "grey", xlab = "C", ylab = "E", zlab = "NOx")
spheres3d(compress, equiv, NOx, type = "s", radius = 0.15, col = "red")

persp3d(NOx.fit, col = "grey", xlab = "C", ylab = "E", zlab = "NOx", alpha = 0.8)
spheres3d(compress, equiv, NOx, type = "s", radius = 0.15, col = "red")
```

9 “Hypervariate” (High–Dimensional) Plots

9.1 Scatterplot Matrix (for $n \geq 4$)

Example 1:

```
data(state)
head(state.x77)
```

```
pairs(state.x77)
```

```
pairs(state.x77[,2:7])
```

Example 2:

```
data(iris)
head(iris)
```

```
species <- iris[,5]
iris <- iris[,1:4]
```

```
pairs(iris)
```

```
pairs(iris, pch = 21, bg = unclass(species))
```

9.2 Parallel Coordinate Plots

Symanzik (2004), p. 303, states:

“Parallel coordinate plots (Inselberg 1985, Wegman 1990) [...] are a geometric device for displaying points in high-dimensional spaces, in particular, for dimensions greater than three. The idea is to sacrifice orthogonal axes by drawing the axes parallel to each other resulting in a planar diagram where each d -dimensional point (x_1, \dots, x_d) is uniquely represented by a continuous line. The parallel coordinate representation enjoys some elegant duality properties with the usual Cartesian coordinates and allows interpretations of statistical data in a manner quite analogous to two-dimensional Cartesian scatterplots. This duality of lines in Cartesian plots and points in parallel coordinates extends to conic sections. This means that an ellipse in Cartesian coordinates maps into a hyperbola in parallel coordinates. Similarly, rotations in Cartesian coordinates become translations in parallel coordinates.

The individual parallel coordinate axes represent one-dimensional projections of the data. We can isolate clusters by looking for separation between data points on any axis or between any pair of axes. Because of the connectedness of the multidimensional parallel coordinate diagram, it is usually easy to see whether or not this clustering propagates through other dimensions.”

Example 3:

```
library(MASS)

state<-state.x77[,2:7]
state2<-state
head(state2)

state2[,c(2,4,6)] <- -state2[,c(2,4,6)] #large = good for all vars
head(state2)

parcoord(state)
```

```
parcoord(state2)
```

```
parcoord(state2, col = unclass(state.region))
```

Example 4:

```
data(iris)
```

```
iris <- iris[,1:4]
```

```
parcoord(iris)
```

```
parcoord(iris, col = unclass(species))
```

```
parcoord(iris, col = unclass(species), var.label = TRUE)
```

Example 5:

```
library (colorspace)
```

```
x = seq(0, 2 * pi, length.out = 30)
```

```
sinx = sin(x)
```

```
cosx = cos(x)
```

```
par (mfrow = c(1, 2))
```

```
plot(sinx, cosx, pch = 22, col = topo.colors(30))
```

```
parcoord(cbind(sinx, cosx), col = topo.colors(30),  
var.label = TRUE, pch = rep(1, 2))
```

9.3 Faces, Star Plots, and other Glyph Representations

Venables & Ripley (2002), p. 314, state:

“There is a wide range of ways to trigger multiple perceptions of a figure, and we can use these to represent each of a moderately large number of rows of a data matrix by an individual figure. Perhaps the best known of these are Chernoff’s faces (Chernoff 1973) [...] and the star plots as implemented in the function stars [...].

These glyph plots do depend on the ordering of variables and perhaps also their scaling, and they do rely on properties of human visual perception. So they have rightly been criticised as subject to manipulation, and one should be aware of the possibility that the effect may differ by viewer. Nevertheless they can be very effective as tools for private exploration.”

Other glyph representations discussed in the literature are based on “trees” and “castles” (Kleiner & Hartigan 1981).

Example 6:

```
data(state)

state <- state.x77[,2:7]
state2 <- state
state2[,c(2,4,6)] <- -state2[,c(2,4,6)] #large = good for all vars

stars(state, key.loc = c(15, 1.5)) #star plot of raw state data

stars(state2, key.loc = c(15, 1.5))
```

Example 7:

```
data(iris)

species <- iris[,5]
iris <- iris[,1:4]

stars(iris, key.loc = c(25, 1.3))

stars(iris, key.loc = c(25, 1.3), col.stars = unclass(species))
```


The R help page for faces indicates:

“Chernoff Faces: [...] Explanation of parameters: 1-height of face, 2-width of face, 3-shape of face, 4-height of mouth, 5-width of mouth, 6-curve of smile, 7-height of eyes, 8-width of eyes, 9-height of hair, 10-width of hair, 11-styling of hair, 12-height of nose, 13-width of nose, 14-width of ears, 15-height of ears. For details look at the literate program of faces.”

Example 8:

```
library(aplpack)

faces(state2)

# look at different permutations of the variable order
faces(state2[,c(2:6, 1)])

faces(state2[,c(3:6, 1:2)])
```

Example 9:

```
data(iris)

species <- iris[,5]
iris <- iris[,1:4]

faces(iris)
```

You should decide yourself how useful these are!

9.4 Andrews Plots

Symanzik (2004), p. 306, states:

“The Andrews (multidimensional data) plot, as introduced in Andrews (1972) is based on a series of Fourier interpolations of the coordinates of multi-dimensional data points. Points that are close in some metric will tend to have similar Fourier interpolations and therefore will tend to cluster in the Andrews plot. Thus, the Andrews plot is an informative graphical tool most useful to detect clustering.

Ideas underlying the Andrews plot and the grand tour are quite similar. However, in contrast to the grand tour, the Andrews plot is a static plot while the grand tour is dynamic. Although dynamic renditions of the Andrews plot exist, and these sometimes also are (incorrectly) referred to as one-dimensional grand tour (Crawford & Fall 1990), the Andrews plot is not a grand tour since it cannot sweep out all possible directions as pointed out in Wegman & Shen (1993). Three-dimensional generalizations of the Andrews plot and other pseudo grand tours have been introduced in Wegman & Shen (1993) as well.”

Execute this code once to create the Andrews function, written by Mike Minnotte:

```
#####
#
# Function for calculating and plotting Andrews curves for      #
# visualization of multivariate data                          #
#                                                              #
# Save as text, then source into R                            #
#                                                              #
#####

Andrews<-function(x,ord=1:ncol(x),nt=101,col=1,lty=1,...)

{# calculates and plots Andrews curve representations of multivariate data
# x - matrix of data, rows are observations, columns are variables
# ord - order of variables for Fourier representation. Earlier variables
# are lower frequency terms.
# nt - number of horizontal points to plot at (default 101).
x<-x[,ord]
minx<-apply(x,2,min)
maxx<-apply(x,2,max)
rangex<-maxx-minx
for (i in 1:ncol(x))
x[,i]<-(x[,i]-minx[i])/rangex[i]
x<-2*x-1
t<-seq(-pi,pi,length=nt)
nx<-nrow(x)
Aout<-matrix(0,nx,nt)
for (i in 1:nx)
Aout[i,]<-rep(x[i,1]/sqrt(2),nt)

np<-floor((ncol(x)-1)/2)

if (np > 0)
for (j in 1:np)
for (i in 1:nx)
Aout[i,]<-Aout[i,]+x[i,2*j]*sin(t*j)+x[i,2*j+1]*cos(t*j)

if (ncol(x) > 2*np+1)
for (i in 1:nx)
Aout[i,]<-Aout[i,]+x[i,ncol(x)]*sin(t*(np+1))

#plot(t,Aout[1,],type='l',ylim=range(Aout))
#for (i in 2:nx)
# lines(t,Aout[i,])

matplot(t,t(Aout),type='l',xlab='t',ylab='A(t)',main='Andrews Curves',
col=col,lty=lty,...)

return()
}
```

Example 10:

```
data(state)
```

```
state <- state.x77[,2:7]
```

```
Andrews(state)
```

```
Andrews(state, ord = 6:1)
```

```
Andrews(state, ord = 6:1, col = unclass(state.region))
```

```
legend("bottomleft", levels(state.region), lty = 1, col = 1:4)
```

Example 11:

```
data(iris)
```

```
species <- iris[,5]
```

```
iris <- iris[,1:4]
```

```
Andrews(iris, col = unclass(species))
```

```
legend("bottomleft", levels(species), lty = 1, col = 1:3)
```

```
Andrews(iris, ord = 4:1, col = unclass(species))
```

```
legend("bottomleft", levels(species), lty = 1, col = 1:3)
```

9.5 Data Images

Minnotte & West (1998), p. 25, state:

“The color histogram was first introduced as a tool for visualizing higher dimensional data by Wegman (1990). A new version of this concept, called a data image, is discussed. Each variable is transformed into a greyscale or color range so that a high-dimensional data set may be viewed as an image, with observations on one axis and variables on the other. The rows and columns of the image may be rearranged to highlight relationships between observations and variables. New ways of displaying the image based on various linear orderings of a data set are discussed.”

A Java version of the data image software, developed by Mike Minnotte and Webster West, could be accessed at <http://www.math.usu.edu/~minnotte/research/java/dimage.html>, but no longer is available from this location.

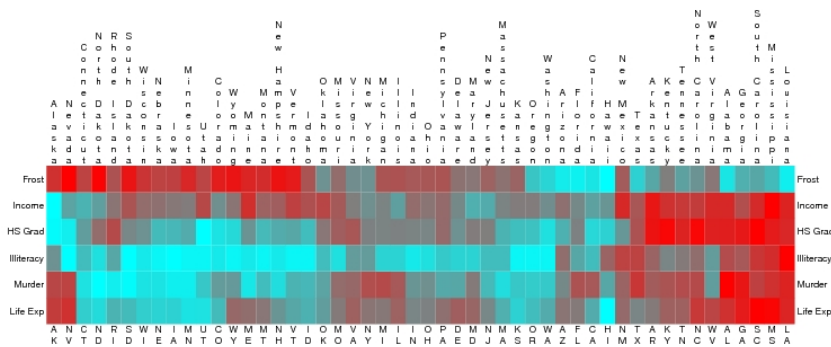


Figure 6: Data image of 1970’s states data. Cyan is high for income, high school graduation, and life expectancy; low for frost days, illiteracy rate, and murder rate. Both observations and variables are sorted by complete linkage clustering algorithm. Identification of observations as well as variables allows recognition of expected and surprising clusters.

Figure 122: Minnotte & West (1998), p. 30, Figure 6: Screenshot of color version taken from <http://www.math.usu.edu/~minnotte/research/preprints/DataImage-col.ps> on 3/26/2009.

The basic idea of color histograms and data images has been extended for multivariate time series (Peng 2008) and for circular-spatial data (Morphet & Symanzik 2010).

Example 12:

Source the R code once to create the data image function, written by Mike Minnotte, then execute the remaining code:

```
source(url("http://www.math.usu.edu/~symanzik/teaching/2011_stat6560/RDataAndScripts/Ch8_Dataimage.R"))
```

```
library(stats)
```

```
data(state)
```

```
state <- state.x77[,2:7]
```

```
state2 <- state
```

```
state2[,c(2,4,6)] <- -state2[,c(2,4,6)] #large = good for all vars
```

```
dataimage(state2)
```

Example 13:

```
data(iris)
```

```
iris <- as.matrix(iris[,1:4])
```

```
dataimage(iris)
```

9.6 From Tables to Plots

Converting complex statistical tables, often with hypervariate data, to effective plots may be a challenge. General guidelines were given by Carr (1994) and Carr & Nusser (1995). However, as Wainer (2009*b*) pointed out, a good table can beat a bad graph, in particular when following guidelines for good tabular presentations, such as those given in Wainer (1993). Fortunately, a good graphical presentation almost certainly will beat a good tabular presentation!

The following two examples are based on two Powerpoint slides from a presentation titled “*Test Methods and Metrology for Evaluating Human Detection and Tracking Systems*,” given at the Interface conference in Seattle, Washington, on June 18, 2010, authored by Tsai Hong, National Institute of Standards and Technology (NIST), Manufacturing Engineering Laboratory, Intelligent Systems Division, and Barry Bodt, Army Research Laboratory (ARL). Thanks are due to Paul Murrell for his help with creating and improving the R code for the two resulting figures.

Example 14:

Object Type	RCTAAlgorithm					
	Alg1	Alg2	Alg3	Alg4	Alg5	Alg6
Human	97.3%	90.8%	98.4%	98.0%	89.5%	85.7%
Mannequin	10.2%		97.7%	98.4%	91.4%	62.5%
Cones	0.0%		4.7%	0.0%	65.6%	0.0%
Barrels	14.1%		54.7%	70.3%	89.1%	0.0%
Crates	46.9%		100.0%	90.6%	100.0%	50.0%
Trucks	25.0%		100.0%	25.0%	100.0%	75.0%
Tripods	1.3%	46.7%	53.6%	60.7%	58.9%	29.8%
False Positives	29.8	77.9	155.0	37.3	29.8	1.3

NIST • Manufacturing Engineering Laboratory • Intelligent Systems Division

Figure 123: Table 1 from Hong and Bodt's 2010 Interface presentation.

Execute the R code below to obtain a graphical representation of the data presented in Figure 123:

```
# Figure 1

library(lattice)

dat1 = read.csv(url("http://www.math.usu.edu/~symanzik/teaching/2011_stat6560/RDataAndScripts/Ch8_Hong_Table1.csv"))

dat2 = data.frame(
  Object = rep(dat1$Object.Type, 6),
  Alg = c(rep("Alg1", 8), rep("Alg2", 8), rep("Alg3", 8), rep("Alg4", 8),
    rep("Alg5", 8), rep("Alg6", 8)),
  Val = c(dat1$Alg1, dat1$Alg2, dat1$Alg3, dat1$Alg4,
    dat1$Alg5, dat1$Alg6)
)

median(dat2$Val[dat2$Object == "Human"]) # 94.05 -> Pos 5
median(dat2$Val[dat2$Object == "Cones"], na.rm = T) # 0 -> Pos 6
median(dat2$Val[dat2$Object == "Tripods"], na.rm = T) # 50.15 -> Pos 7
median(dat2$Val[dat2$Object == "Barrels"], na.rm = T) # 54.7 -> Pos 8
median(dat2$Val[dat2$Object == "Trucks"], na.rm = T) # 75 -> Pos 1
median(dat2$Val[dat2$Object == "Crates"], na.rm = T) # 90.6 -> Pos 2
median(dat2$Val[dat2$Object == "Mannequin"], na.rm = T) # 91.4 -> Pos 3
median(dat2$Val[dat2$Object == "FALSE Positives"], na.rm = T) # 33.55 -> Pos4

dat2$Alg = factor(dat2$Alg,
  levels = dat2$Alg[dat2$Object == "Human"][order(dat2$Val[dat2$Object == "Human"])]))

#pdf("Table1_asPlot.pdf")
```



```

dotplot(Alg ~ Val | Object , data = dat2,
  layout = c(4, 2),
  main = "Percent Detection & Misclassification",
  xlab = "Percentage (Blue) / Mean (Orange); Green Line = Median for each Object",
  index.cond = list(c(8, 3, 6, 4, 5, 2, 7, 1)),
  scales = list(x = "free"),
  xlim = list(c(-10, 110), c(-10, 110), c(-10, 110), c(-10, 170), c(-10, 110), c(-10, 110), c(-10, 110), c(-10, 110)),
  panel = function(x,y, ...) {
    if (panel.number() == 1)
      panel.abline(v = median(dat2$Val[dat2$Object == "Trucks"], na.rm = T), col = "light green")
    else if (panel.number() == 2)
      panel.abline(v = median(dat2$Val[dat2$Object == "Crates"], na.rm = T), col = "light green")
    else if (panel.number() == 3)
      panel.abline(v = median(dat2$Val[dat2$Object == "Mannequin"], na.rm = T), col = "light green")
    else if (panel.number() == 4)
      panel.abline(v = median(dat2$Val[dat2$Object == "FALSE Positives"], na.rm = T), col = "light green")
    else if (panel.number() == 5)
      panel.abline(v = median(dat2$Val[dat2$Object == "Human"], na.rm = T), col = "light green")
    else if (panel.number() == 6)
      panel.abline(v = median(dat2$Val[dat2$Object == "Cones"], na.rm = T), col = "light green")
    else if (panel.number() == 7)
      panel.abline(v = median(dat2$Val[dat2$Object == "Tripods"], na.rm = T), col = "light green")
    else if (panel.number() == 8)
      panel.abline(v = median(dat2$Val[dat2$Object == "Barrels"], na.rm = T), col = "light green")

    if (panel.number() == 4)
      panel.dotplot(x, y, xlim = c(-10, 170), col = c("red", rep("orange",5)),
        cex = 1.0, pch = c(17, rep(16, 5)), ...)
    else if (panel.number() == 5)
      panel.dotplot(x, y, xlim = c(-10, 170), col = c("purple", rep("blue", 5)),
        cex = 1.2, pch = c(17, rep(16, 5)), ...)
    else
      panel.dotplot(x, y, xlim = c(-10, 110), col = c("purple", rep("blue", 5)),
        cex = 1.0, pch = c(17, rep(16, 5)), ...)
  }
)
#dev.off()

```

A possible description of this figure could be as follows:

This figure shows the detection and misclassification rates (in %) of seven objects (as well as the average number of false positives) for six different algorithms.

The upper left panel shows the human detection rate. Rows, representing the six different algorithms, in each panel are sorted decreasingly (from top to bottom) with respect to this human detection rate. The vertical green line in all panels shows the median percentage (or median average), obtained for the different algorithms for a particular object.

The next six panels show misclassification rates for different objects for each algorithm. These panels have been arranged from lowest median misclassification rate (0% for Cones) to highest median misclassification rate (about 91% for Mannequin).

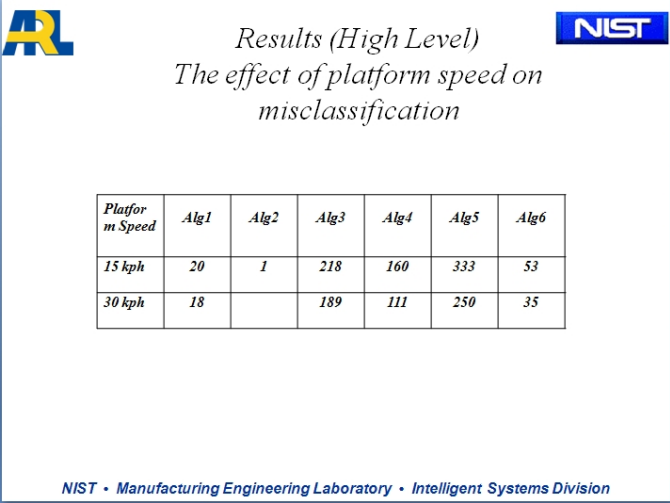
The last panel shows the average number of false positives per run for each algorithm. To contrast this from the percentages shown in the other panels, different colors have been selected.

One of the algorithms (Alg1) has been highlighted via different colors and symbols. This algorithm is among the best three with respect to human detection rate and it only lagged by about 1% to the best algorithm — confidence intervals for these point estimates have not been reported. Moreover, this algorithm has the lowest misclassification rates (possibly shared with some other algorithms) for five of the six objects (Cones, Tripods, Trucks, Crates & Mannequin) and the second lowest misclassification rate for Barrels. Its number of false positives per run also is the second lowest (split with Alg5).

It should be noted that misclassification rates for Alg2 have not been recorded for five of the six objects.

Note that this is my personal interpretation of the original table. If a different sorting is needed, or a different algorithm needs to be highlighted for various reasons, this obviously can be done.

Example 15:



The slide displays the following table:

<i>Platform Speed</i>	<i>Alg1</i>	<i>Alg2</i>	<i>Alg3</i>	<i>Alg4</i>	<i>Alg5</i>	<i>Alg6</i>
<i>15 kph</i>	20	1	218	160	333	53
<i>30 kph</i>	18		189	111	250	35

NIST • Manufacturing Engineering Laboratory • Intelligent Systems Division

Figure 124: Table 3 from Hong and Bodt's 2010 Interface presentation.

Execute the R code below to obtain a graphical representation of the data presented in Figure 124:

```
# Figure 3

library(lattice)

dat31 = read.csv(url("http://www.math.usu.edu/~symanzik/teaching/2011_stat6560/RDataAndScripts/Ch8_Hong_Table3.csv"))

dat32 = data.frame(
  Alg = c(rep("Alg1", 2), rep("Alg2", 2), rep("Alg3", 2), rep("Alg4", 2),
    rep("Alg5", 2), rep("Alg6", 2)),
  Val = c(dat31$Alg1, dat31$Alg2, dat31$Alg3, dat31$Alg4,
    dat31$Alg5, dat31$Alg6),
  Speed = rep(c("15 kph", "30 kph"), 6)
)

#dotplot(Alg ~ Val | Speed, data = dat32,

dat32$Alg = factor(dat32$Alg,
  levels = dat32$Alg[dat32$Speed == "15 kph"][order(- dat32$Val[dat32$Speed == "15 kph"])])

#pdf("Table3_asPlot.pdf")

dotplot(Alg ~ Val, data = dat32, groups = Speed,
  main = "The Effect of Platform Speed on Misclassification",
  xlab = "Results (High Level)",
  layout = c(1, 1),
  pch = 16,
  cex = 1.2,
```

```

# tell key to match symbols to those used in the plot
par.settings = list(
  superpose.symbol = list(cex = 1.2, pch = c(16, 16) ) ),
  auto.key = list(border = TRUE, between = 4, levels(dat32$Speed), space = "right", title = "Platform Speed",
    cex.title = 0.9),
)

#dev.off()

```

A possible description of this figure could be as follows:

This figure shows the effect of platform speed on misclassification for six different algorithms.

Rows, representing the six different algorithms, are sorted increasingly (from top to bottom) with respect to the number of misclassifications at a speed of 15 kph. The number of misclassifications at 30 kph are consistently below the corresponding number at 15 kph.

No data are available for Alg2 at 30 kph. Recall that misclassification rates were missing for five of the six objects in Figure 123 — so the low number for Alg2 at 15 kph should be taken with care.

Note that if you want to be consistent with Figure 123 and you want to point out that Alg1 overall is the recommended algorithm to use, I would highlight the two symbols for Alg1 with the same colors/symbols as in Figure 123. Also, if my interpretation of the result for Alg2 is correct, one might use some fainter blue for this algorithm to indicate some high uncertainty.

10 Interactive and Dynamic Graphics

Symanzik (2004), p. 295, states:

“Interactive and dynamic statistical graphics enable data analysts in all fields to carry out visual investigations leading to insights into relationships in complex data. Interactive and dynamic statistical graphics involve methods for viewing data in the form of point clouds or modeled surfaces. Higher-dimensional data can be projected into one-, two- or three-dimensional planes in a set of multiple views or as a continuous sequence of views which constitutes motion through the higher-dimensional space containing the data.

Strictly, there is some difference between interactive graphics and dynamic graphics. When speaking of interactive graphics only, we usually mean that a user actively interacts with, i.e., manipulates, the visible graphics by input devices such as keyboard, mouse, or others and makes changes based on the visible result. When speaking of dynamic graphics only, we usually mean that the visible graphics change on the computer screen without further user interaction. An example for interactive graphics might be the selection of interval lengths and starting points when trying to construct an optimal histogram while looking at previous histograms. An example for dynamic graphics might be an indefinitely long grand tour with no user interaction. Typically, interactive graphics and dynamic graphics are closely related and we will not make any further distinction among the two here and just speak of interactive and dynamic statistical graphics.

Two terms closely related to interactive and dynamic statistical graphics are exploratory data analysis (EDA) and visual data mining (VDM).

EDA, as defined by Tukey (1977), “*is detective work — numerical detective work — or counting detective work — or graphical detective work.*” Modern techniques and software in EDA, based on interactive and dynamic statistical graphics, are a continuation of Tukey’s idea to use graphics to find structure, general concepts, unexpected behavior, etc. in data sets by looking at the data. To cite Tukey (1977) again, “*today, exploratory and confirmatory can — and should — proceed side by side.*” Interactive and dynamic statistical graphics should not replace common analytic and inferential statistical methods — they should rather extend these classical methods of data analysis.”

10.1 Software for Interactive and Dynamic Graphics

Symanzik (2004), pp. 308–309, states:

“In this section, we concentrate on three main streams of software for interactive and dynamic statistical graphics: Software developed by researchers affiliated with the University of Augsburg, in particular REGARD, MANET, and Mondrian; software developed by researchers affiliated with George Mason University (GMU), in particular ExplorN and CrystalVision; and software developed by researchers affiliated with Bell Labs, AT&T, and Iowa State University (ISU), in particular XGobi and GGobi. Wilhelm et al. (1996) contains an in depth review of software for interactive statistical graphics. Wilhelm et al. (1999) is one of the few publications where the different interactive graphical concepts provided by these three main streams (represented by MANET, ExplorN, and XGobi, respectively) are applied to the same data set and thus allow a direct comparison of their features and capabilities in visual clustering and classification.”

10.1.1 REGARD, MANET, and Mondrian

Symanzik (2004), pp. 309–310, states:

“In this section we present a series of software developments that was initiated in the late 1980’s by John Haslett and Antony Unwin at Trinity College, Dublin, and later was continued by Antony Unwin and his collaborators at the Institut für Mathematik, University of Augsburg. Other main collaborators that contributed to the development of these software tools that should be mentioned here are Heike Hofmann, Martin Theus, Adalbert Wilhelm, and Graham Wills.

Some of the early developments are Diamond Fast (Unwin & Wills 1988) and Spider (Craig et al. 1989). Diamond Fast is a software package for the exploration of multiple time series with interactive graphics. Spider is a software package for the exploration of spatially referenced data. One of its main features are moving statistics, an extension of brushing for spatial data (Craig et al. 1989). Spider also supports histograms, density estimates, scatterplot matrices, and linked brushing. It runs on Macintosh computers only.

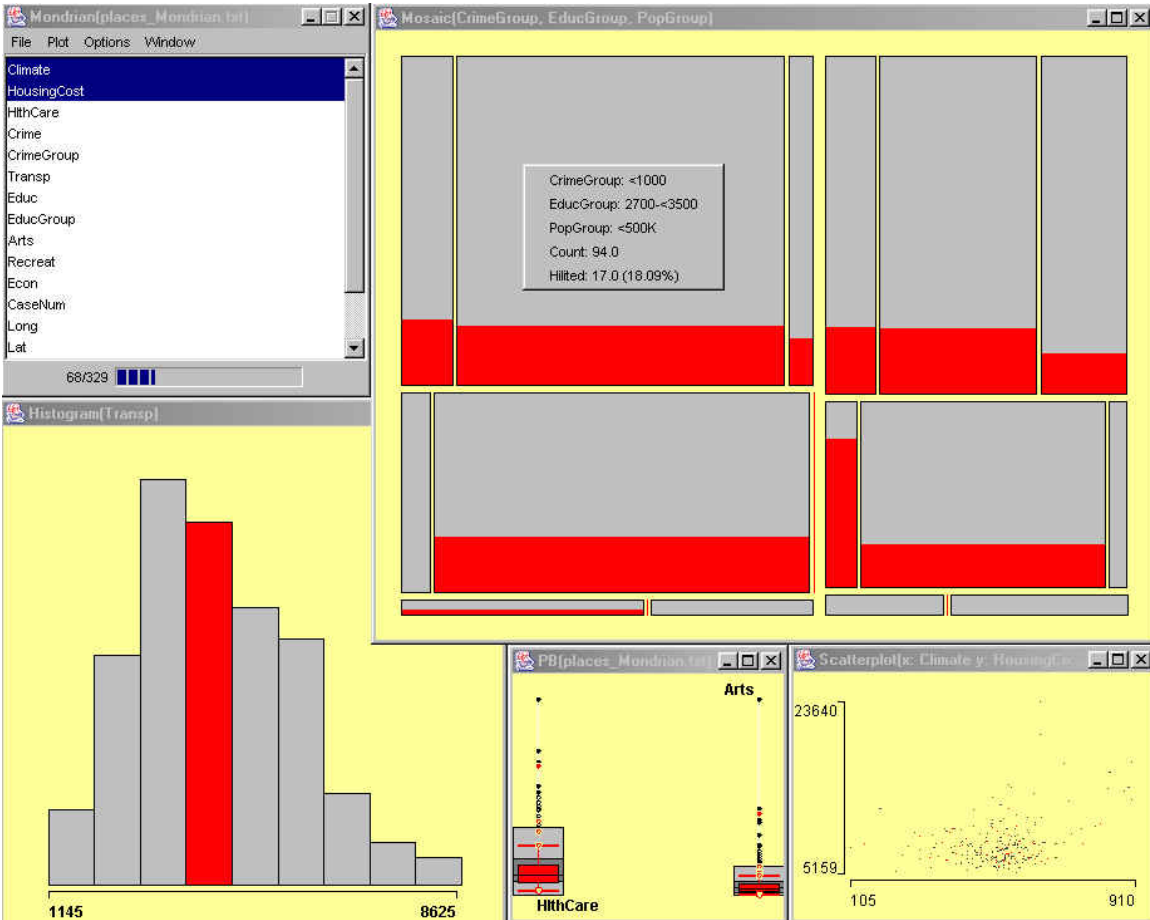


Figure 125: Symanzik (2004), p. 308, Figure 10.4: Screenshot of the “Places” data in Mon-drian. The variables Crime, Education, and Population have been discretized for this figure. A mosaic plot of Crime (first vertical division, grouped as below 1000 [left] and above 1000 [right]), Education (first horizontal division, grouped as 2700 to 3500 [top], below 2700 [middle], and above 3500 [bottom]), and Population (second vertical division, grouped as 500,000 to 1,000,000 [left], below 500,000 [middle], and above 1,000,000 [right]) is displayed at the top right. A histogram of Transportation is shown at the bottom left, boxplots of HealthCare and Arts are shown at the bottom middle, and a scatterplot of Climate (horizontal) vs. Housing-Cost (vertical) is shown at the bottom right. The mosaic plot shows that Crime, Education, and Population are not independent. The different displays show how average Transportation (that has been brushed in the histogram) is related to the other variables. All displays have been linked.

REGARD (Unwin et al. 1990, Unwin 1994) is a software package that also provides high interaction graphics tools for spatial data. REGARD stands for “Radical Effective Graphical Analysis of Regional Data” and runs

on Macintosh computers only. REGARD supports four types of layers of spatial data, i.e., points, regions, lines, and pictures. The central display in REGARD is the map window that is linked to statistical displays such as boxplots, scatterplots, and rotating plots. A map may be loaded as one picture in a picture layer or as several pictures in several layers, thus allowing to turn on or off different aspects of a map (such as state boundaries or a road network). Additional interactive features are interrogation, highlighting, resizing, and rescaling. Advanced features include zooming into submaps, animation across ordered variables, cross-layer linking, network analysis tools, and interactive query tools across all graphical displays.

MANET (Unwin et al. 1996) is a statistical graphics research program for EDA and written in C++. MANET stands for “Missings Are Now Equally Treated” and runs on Macintosh computers only. It is freely available from the following Web site: <http://www1.math.uni-augsburg.de/Manet/>.

MANET offers all standard one- and two-dimensional graphics for continuous data as well as for discrete data: dotplots, scatterplots, histograms, boxplots, bar charts. Some special graphics for discrete and spatial data are integrated: spine plots, mosaic plots and polygon plots. MANET grew out of a project to keep track of missing values in statistical graphics. In MANET all displays are fully linked and instantaneously updated. Displays are kept as simple as possible to not distract the user.

The standard use of linked views in MANET is to highlight clusters that are apparent in one dimension and to see these one-dimensional clusters in the light of other variables. By systematically subsetting the sample points, we can also detect two- and higher-dimensional clusters. Once a cluster has been detected, a classification rule can be set up by taking the boundary values of the cluster. In MANET those values can easily be obtained by interrogating the plot symbols.

One-dimensional views show the one-dimensional clusters directly. Two-dimensional clusters become visible by highlighting a subset in one variable and conditioning another plot on this subset. For three- and higher-dimensional clusters, we have to combine various subsets in different plots into one conditioning set and then we have to look at the remaining plots to check for clusters. The generation of such combined selections is not only possible in MANET but it is also very efficiently implemented through selection sequences.

In MANET, both dotplots and boxplots are drawn in a non-standard way. In dotplots the brightness of a point shows the frequency of its occurrence. This method, called tonal highlighting, is used to visualize overplotting of points. A bright color represents many points while a dark color represents just a few points. There is no tonal highlighting for selected points in MANET. The layout of boxplots is changed so that a standard boxplot can be superimposed for selected points. The inner fifty percent box is drawn as a dark grey box. The outer regions, usually represented as whiskers, are drawn as light grey boxes.

A recent new development, Mondrian (Theus 2002, 2003), is a data visualization system written in JAVA and therefore runs on any hardware platform. Mondrian is freely available from the following Web site:

<http://www.rosuda.org/Mondrian/>.

The main emphasis of Mondrian is on visualization techniques for categorical and geographical data. All plots in Mondrian (see Figure 125) are fully linked and offer various interrogations. Any case selected in one plot in Mondrian is highlighted in all other linked plots. Currently, implemented plots comprise mosaic plots, scatterplots, maps, bar charts, boxplots, histograms, and parallel coordinate plots. Mosaic plots in Mondrian are fully interactive. This includes not only linking, highlighting and interrogations, but also an interactive graphical modeling technique for loglinear models.”

10.1.2 HyperVision, ExplorN, and CrystalVision

Symanzik (2004), pp. 310–311, states:

“In this section we present a series of software developments that was initiated in the late 1980’s by Daniel B. Carr (initially while at Battelle Pacific Northwest Laboratories) and Edward J. Wegman at GMU. Other main collaborators that contributed to the development of these software tools that should be mentioned here are Qiang Luo and Wesley L. Nicholson.

EXPLOR4 (Carr & Nicholson 1988) is a research tool, originally implemented on a VAX 11/780 and written in FORTRAN. Its main features are rotation, masking, scatterplots and scatterplot matrix, ray glyph plots, and stereo views.

HyperVision, presented in Bolorforoush & Wegman (1988), is a software product that has been implemented in PASCAL on an IBM RT under the

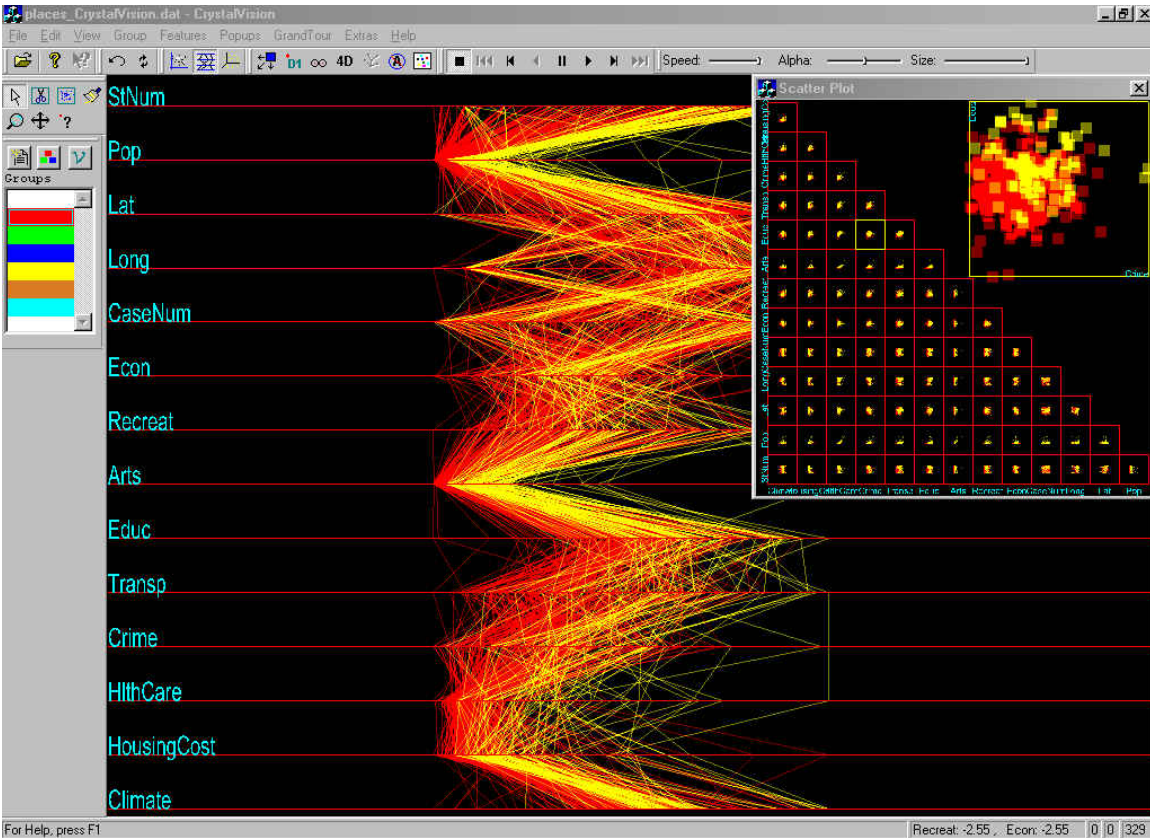


Figure 126: Symanzik (2004), p. 304, Figure 10.3: Screenshot of the “Places” data in CrystalVision. A parallel coordinate plot of all variables is shown as the main plot. A scatterplot matrix of all variables with a scatterplot of Crime (horizontal) vs. Education (vertical) is shown as a popup in the top right. The data has been brushed according to high and low Population. According to the parallel coordinate plot, higher Population is associated with higher Arts and HousingCost. The scatterplot of Crime and Education seems to reveal that higher Population is also associated with higher Crime and higher Education. All displays have been linked.

AIX operating system as well as for MS-DOS machines. The latter implementation has a mouse-driven painting capability and can do real-time rotations of 3D scatterplots. Other displays are parallel coordinate plots, parallel coordinate density plots, relative slope plots, and color histograms. The main interactive features in HyperVision in addition to linked brushing are highlighting, zooming, and nonlinear rescaling of each axis.

ExplorN (Carr et al. 1997) is a more advanced software package than HyperVision and EXPLOR4, but with similar basic features. It runs on SGI workstations only, using either the GL or the OpenGL tools. ExplorN

is freely available from the following ftp site: <ftp://www.galaxy.gmu.edu/pub/software/>.

ExplorN supports scatterplot matrices, parallel coordinate plots, icon-enhanced three-dimensional stereoscopic plots, d -dimensional grand tours and partial grand tours (i.e., tours based on a subset of the variables with the remaining variables being held fixed), and saturation brushing all in a high interaction graphics package.

The ExplorN software is intended to demonstrate principles rather than to be an operational tool so that some refinements normally found in operational software are not there. These include history tracking, easy point identification, identification of mixture weights in the grand tour, relabeling of axes during and after a grand tour as well as simultaneous multiple window views.

Although ExplorN also supports conventional scatterplots and scatterplot matrices, one of its outstanding features are parallel coordinate displays and partial grand tours. Since it is easy to see pairwise relationships for adjacent variables in parallel coordinate plots, but less easy for nonadjacent variables, a complete parallel coordinate investigation would require running through all possible permutations. Instead of this, we recommend using the d -dimensional parallel coordinate grand tour that is implemented in ExplorN. An important interactive procedure for finding clusters using parallel coordinate plots is via the brush-tour.

CrystalVision is a recently developed successor of ExplorN, freely accessible at <ftp://www.galaxy.gmu.edu/pub/software/>.

Its main advantage over the older package is that it is available for PCs. Similar to ExplorN, CrystalVision's (see Figure 126) main focus is on parallel coordinate plots, scatterplots, and grand tour animations. Examples of its use, e.g., its EDA techniques applied to scanner data provided by the U.S. Bureau of Labor Statistics (BLS), can be found in Wegman & Dorfman (2003)."

10.1.3 Data Viewer, XGobi, and GGobi

Symanzik (2004), pp. 311–313, states:

“In this section we present a series of software developments that was initiated in the mid 1980's by Andreas Buja, Deborah F. Swayne, and Di-

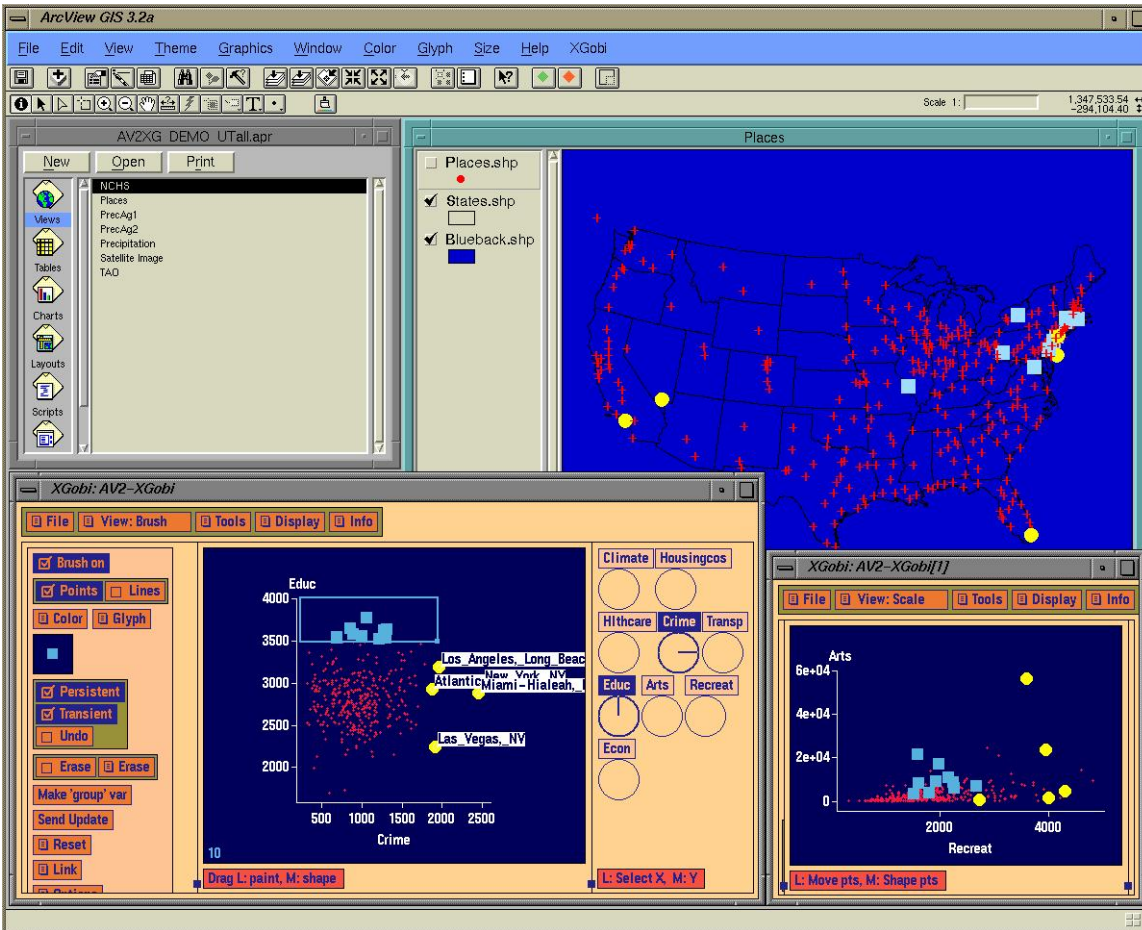


Figure 127: Symanzik (2004), p. 299, Figure 10.1: Screenshot of the “Places” data in ArcView/XGobi. A map view of the 329 spatial locations is displayed in ArcView at the top. The two XGobi windows at the bottom are showing scatterplots of Crime (horizontal) vs. Education (vertical) [left] and Recreation (horizontal) vs. Arts (vertical) [right]. Locations of high Crime have been brushed and identified, representing some of the big cities in the U.S. Also, locations of high Education (above 3500) have been brushed, mostly representing locations in the northeastern U.S. All displays have been linked.

anne Cook at the University of Washington, Bellcore, AT&T Bell Labs, and ISU. Other main collaborators that contributed to the development of these software tools that should be mentioned here are Catherine Hurley, John A. McDonald, and Duncan Temple Lang.

The Data Viewer (Buja et al. 1986, 1988, Hurley 1988, 1989, Hurley & Buja 1990) is a software package originally developed on a Symbolics Lisp Machine that supports object-oriented programming. The Data Viewer is a

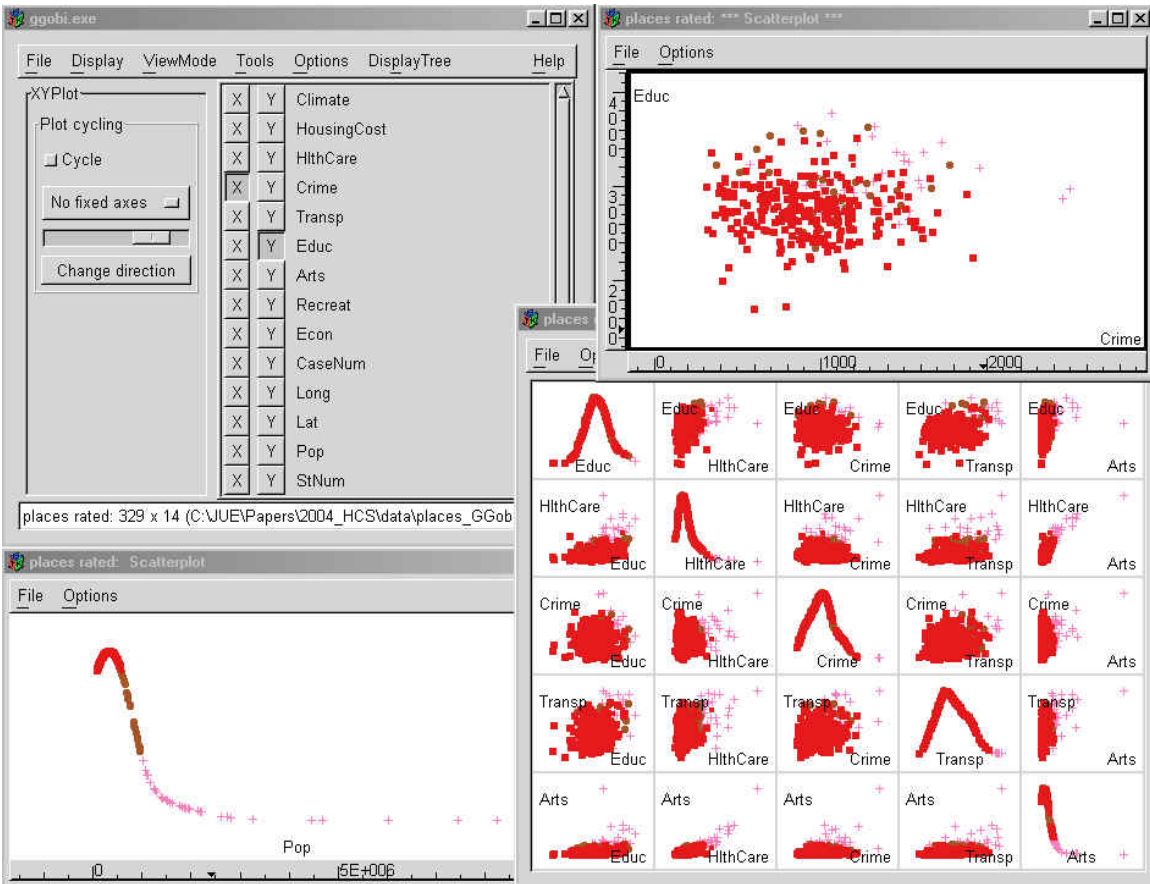


Figure 128: Symanzik (2004), p. 300, Figure 10.2: Screenshot of the “Places” data in GGobi. A scatterplot of Crime (horizontal) vs. Education (vertical) is displayed at the top right, a scatterplot matrix of five of the variables is displayed at the bottom right, and a density (1D) plot of Population is displayed at the bottom left. The data has been brushed with respect to Population: one group for a Population less than 500,000, one group for a Population between 500,000 and 1,000,000, and one group for a Population above 1,000,000. The scatterplot of Crime and Education seems to reveal that higher Population is associated with higher Crime and higher Education. The scatterplot matrix seems to reveal that higher Population is also associated with higher Arts and higher HealthCare. All displays have been linked.

system for the exploratory analysis of high-dimensional data sets that allows interactive labeling, identification, brushing, and linked windows. Additional features are viewport transformations such as expanding or shrinking of the data and shifting of the data. The Data Viewer supports several types of projections, including simple 3D rotations, correlation tour (Buja et al. 1988), and grand tour.

Many of the design and layout concepts of the Data Viewer as well as parts of its functionality provided the basic ideas for the follow-up XGobi (see Figure 127), first described in Swayne et al. (1991) and Swayne & Cook (1992). Development on XGobi took place for about a decade; its almost final version is documented in Swayne et al. (1998). XGobi is implemented in the X Windows System, so it runs on any UNIX system, and it runs under Microsoft Windows or the Macintosh operating system if an X emulator is used. XGobi can be freely downloaded from

<http://www.research.att.com/areas/stat/xgobi/>.

XGobi is a data visualization system with interactive and dynamic methods for the manipulation of views of data. It offers 2D displays of projections of points and lines in high-dimensional spaces, as well as parallel coordinate plots. Projection tools include dotplots and ASH of single variables, scatterplots of pairs of variables, 3D data rotations, and grand tours. Views of the data can be panned and zoomed. Points can be labeled and brushed with glyphs and colors. Lines can be edited and colored. Several XGobi processes can be run simultaneously and linked for labeling, brushing, and sharing of projections. Missing data are accommodated and their patterns can be examined; multiple imputations can be given to XGobi for rapid visual diagnostics (Swayne & Buja 1998). XGobi can be cloned, i.e., an identical new XGobi process with exactly the same data and all brushing information can be invoked.

Rotating plots are nowadays implemented in most statistical packages, but the implementation in XGobi goes beyond most of the others. In addition to the standard grand tour, XGobi supports the projection pursuit guided tour. More details on projection pursuit indices available in XGobi can be found in Cook et al. (1993) and Cook et al. (1995). Additional index functions that result in speed improvements of the calculations have been presented in Klinke & Cook (1997).

GGobi (Swayne et al. 2003) is a direct descendant of XGobi, but it has been thoroughly redesigned. GGobi (see Figures 128) can be freely downloaded from <http://www.ggobi.org/>.

At first glance, GGobi looks quite unlike XGobi because GGobi uses a newer graphical toolkit called GTK+ (<http://www.gtk.org>), with a more contemporary look and feel and a larger set of user interface components. Through the use of GTK+, GGobi can be used directly on Microsoft Win-

dows, without any emulator. In addition, GGobi can be used on any UNIX and Linux system.

In contrast to XGobi, the plot window in GGobi has been separated from the control panel. In XGobi, there is in general a single plot per process; to look at multiple views of the same data, we have to launch multiple XGobi processes. In contrast, a single GGobi session can support multiple plots of various types: scatterplots, parallel coordinate plots, scatterplot matrices, and time series plots have been implemented thus far. Other changes in GGobi's appearance and repertoire of tools (when compared to XGobi) include an interactive color lookup table manager, the ability to add variables "on the fly", and a new interface for view scaling (panning and zooming). At this point, some of the advanced grand tour and projection pursuit guided tour features from XGobi have not been fully reimplemented in GGobi (but hopefully will be available in the near future)."

10.2 Concepts of Interactive and Dynamic Graphics

Symanzik (2004), p. 298, states:

“This section will provide some deeper insights into concepts of interactive and dynamic graphics mentioned in the previous sections. Buja et al. (1996) contains a taxonomy of interactive data visualization based on the notions of focusing, linking, and arranging views of data. Unwin (1999) discusses some of the main concepts in the context of interactive graphics software.”

10.2.1 Scatterplots and Scatterplot Matrices

Symanzik (2004), pp. 298–299, states:

“Perhaps the most basic concepts for statistical graphics are scatterplots (see Figures 127, 128, 126, and 125). In a simple scatterplot, we place different symbols (sometimes also called glyphs) at x - and y -positions in a two-dimensional plot area. These positions are determined by two of the variables. The type, size, and color of the symbols may depend on additional variables. Usually, explanatory information such as axes, labels, legends, and titles are added to a scatterplot. Additional information such as a regression line or a smoothed curve can be added as well.

If the data consist of more than two variables (e.g., somewhere between three to ten), the data can be displayed by a scatterplot matrix (see Figures 128 and 126) that shows all pairwise scatterplots of the variables. The essential property of a scatterplot matrix is that any adjacent pair of plots have one of their axes in common. When plotting the full array of all $n \times (n - 1)$ pairwise scatterplots, each plot in the upper triangle of plots has a matching plot in the lower triangle of plots, with the exception that the axes in these pairs of plots have been flipped. Therefore, sometimes only the upper or lower triangle of scatterplots is displayed; thus gaining plotting speed and allowing each individual plot to be somewhat larger. Early examples of scatterplot matrices can be found in Chambers et al. (1983) and Cleveland (1985) for example. In fact, Chambers et al. (1983) initially called an array of pairwise scatterplots for three variables a draftsman’s display and for four (or more) variables a generalized draftsman’s display. In

their (generalized) draftsman’s display, each point is plotted with the same symbol. When encoding additional information through the use of different plotting symbols, Chambers et al. (1983) speak of symbolic (generalized) draftsman’s displays. Today, we hardly make any distinction of these different types of displays and just speak of scatterplot matrices.

Murdoch (2002) and Unwin (2002) discuss features good scatterplots and related interactive software should provide, e.g., meaningful axes and scales, features for rescaling and reformatting, good handling of overlapping points and missing data, panning and zooming, and querying of points. Carr et al. (1987) describes techniques for scatterplot matrices particularly useful for large numbers of observations.”

10.2.2 Brushing and Linked Brushing/Linked Views

Symanzik (2004), pp. 300–301, states:

“Brushing, as introduced in Becker & Cleveland (1988) and Becker et al. (1988), initially was considered as a collection of several dynamic graphical methods for analyzing data displayed in a scatterplot matrix. The central idea behind brushing is a brush, usually a rectangular area on the computer screen, that is moved by the data analyst to different positions on the scatterplot or any other graphical display. Four brushing operations were introduced in Becker & Cleveland (1988): highlight, shadow highlight, delete, and label. The most commonly used brushing technique is highlighting — often in the context of linked brushing, i.e., for linked views. All points that are inside the brush in the currently selected display are highlighted, i.e., marked with a different symbol or color. Simultaneously, points that correspond to those points are automatically highlighted with the same symbol/color in all linked views.

A very useful brushing technique is the transient paint mode. As the brush is moved, the new points that come inside the brush are highlighted while points that move outside the brush are no longer highlighted.

While brushing initially was only developed for scatterplot matrices, it quickly has been adapted to other types of linked graphical displays. Linked brushing among different displays is one of the most useful techniques used within dynamic and statistical graphics. Linked brushing can be applied

to graphical representation of continuous data, summary data such as histograms (Stuetzle 1988), or even displays of categorical data such as mosaic plots (Hofmann 2000, 2003). All dynamic statistical graphics software packages support linked brushing among different types of graphical displays these days.

When dealing with massive data sets, it is often beneficial to focus on particular subgroups of the data and also be able to quickly return to a previous stage of the analysis. Selection sequences (Theus et al. 1998, Hofmann & Theus 1998) are an extension of the conventional linked–highlighting paradigm as they store the whole hierarchical path of a selection and allow an easy editing, redefinition, and interrogation of each selection in the path of the analysis. In a selection sequence, we can easily jump from one branch of the hierarchic selection tree to another.”

10.2.3 Focusing, Zooming, Panning, Slicing, Rescaling, and Reformatting

Symanzik (2004), pp. 301–302, states:

“Focusing techniques, as introduced in Buja et al. (1991), are based on the idea that it often might be easier for a human analyst to understand several individual displays, each focused on a particular aspect of the underlying data, rather than looking at the full data set. Focusing techniques include subset selection techniques, e.g., panning and zooming or slicing, and dimensionality reduction techniques, e.g., projection. Methods for focusing can be automatic, interactive, or a combination of both. While focusing shows only part of the data at a time, it is important to display multiple linked views of the data, perhaps each focusing on a different aspect of the data, to maintain the full picture of the data.

Zooming is a technique that can be used for inspecting details of the data when overplotting arises. Zooming can be done via some kind of a magnifying glass or by manually selecting subsections of the visible axes, e.g., via sliders. The main idea behind zooming is that when several points overplot in the full display, it may indeed turn out that these points are exactly the same when zooming into the neighborhood of these points — or, what most frequently happens, that these points have a particular structure and are not exactly the same.

Panning is closely related to zooming. An analyst should know which subset of the data is currently visible. Therefore, an information plot should reveal the current location on which subregion we have zoomed.

Slicing, as described in Furnas (1988) and Furnas & Buja (1994), is a technique that takes sections (or slices) of a high-dimensional data set. While slicing (and projections) are useful means for an exploratory data analysis, these techniques also have their limitations. However, these limitations may be overcome by combining slicing and projections in so-called projections (Furnas & Buja 1994). An extension of individual projection views is the projection matrix (Tweedie & Spence 1998), some kind of a density plot summarizing multi-dimensional volumetric information. The projection matrix is a useful representation for engineering design, allowing an analyst to interactively find a design that leads to a maximal manufacturing yield.

Rescaling is a technique to quickly change the scale of the displayed variables, e.g., by taking the log, square root, standardize, or by mapping to a 0–1 scale. When looking at multiple variables, it might also be beneficial to have a common scale (from the minimum across all variables to the maximum across all variables). By interactively rescaling variables, an analyst may identify useful transformations for a follow-up modeling step of the data.

Reformatting includes features as simple as swapping x and y axes in a scatterplot or changing the order of coordinate axes in a parallel coordinate plot.

Unwin (2002) provides more details on several of the techniques described above.”

10.2.4 Rotations and Projections

Symanzik (2004), p. 302, states:

“Rotation, as introduced in Fisher et al. (1974) and later refined in Becker et al. (1988), is a very powerful tool for understanding relationships among three or more variables. The familiar planar scatterplot is enhanced by rotation to give the illusion of a third dimension. We typically rotate plots in search of some interesting views that do not align with the plot axes and therefore cannot be seen in a scatterplot matrix. Usually, a three-dimensional point cloud representing three of the variables is shown rotating on a computer screen. The rotation shows us different views of the points and it produces a 3D effect while moving, allowing us to see depth. Basic rotation controls with a mouse have been introduced in Becker et al. (1988).

Mathematically speaking, each rotation within a 3D space onto a 2D computer screen is based on a projection. Obviously, it is mathematically possible to project high-dimensional data onto low-dimensional subspaces and gain insights into the underlying data through dynamic visualizations of such projections. One particular example of a continuous sequence of projections, the grand tour, will be discussed in the next section. Cook & Buja (1997) discuss methods how to manually control high-dimensional data projections. Cook (1997) provides a variety of training data sets that help new users get a visual feeling of the underlying high-dimensional data set when seen as a projection into low-dimensional space.”

10.2.5 Grand Tour

Symanzik (2004), p. 303, states:

“Often, simple plot rotation, as discussed in the previous section, does not suffice to see all interesting views of the data. To produce a plethora of possible interesting views, the grand tour has been introduced in Asimov (1985) and Buja & Asimov (1986). In Asimov (1985), the grand tour has been described as “*a method for viewing multivariate statistical data via orthogonal projections onto a sequence of two-dimensional subspaces. The sequence of subspaces is chosen so that it is dense in the set of all two-dimensional subspaces.*” Some of the features the grand tour can be used

for are examining the overall structure and finding clusters or outliers in high-dimensional data sets.

In the context of the grand tour, an alternating sequence of brushing, looking at additional projections from the grand tour, brushing, and so on, is referred to as the brush-tour strategy in the remainder of this chapter. We can only be sure that a cluster visible in one projection of the grand tour really is a cluster if its points remain close to each other in a series of projections and these points move similarly when the grand tour is activated. If points move apart, we probably found several subclusters instead of one larger cluster.

Wegman (1992) discusses a form of the grand tour for general d -dimensional space. The algorithms for computing a grand tour are relatively computationally intensive. Wegman & Shen (1993) discuss an approximate one- and two-dimensional grand tour algorithm that was much more computationally efficient than the Asimov winding algorithm. That algorithm was motivated in part by a discussion of the Andrews (multidimensional data) plot, discussed in Section 9.4, which can also be regarded as a highly restricted pseudo tour.”

10.3 Spatial Data Analysis in the Linked ArcView 2.1 and XGobi Environment

The material on the next few pages consists of material from a workshop held in November 1996 at the University of Michigan. XGobi (see Section 10.1.3 for more details) is the software that was used at that time. Our task is to translate the historic instructions for use in GGobi.

Linking cartographic displays with interactive statistical graphics has been widely discussed in the literature. Symanzik, Cook, Lewin-Koh, Majure & Megretskaia (2000), pp. 471–472, state:

“Linking statistical plots with geography for analyzing spatially referenced data has been discussed widely in recent years. Monmonier (1988) describes a conceptual framework for geographical representations in statistical graphics and introduces the term *geographic brushing* in reference to interacting with the map view. But geographic brushing does not only mean pure interaction with the map. In addition, this term has a much broader meaning, e. g., finding neighboring points and spatial structure in a geographic setting.

A good overview of dynamic graphics for multivariate data is given by Buja et al. (1996). In addition, many software solutions have been developed for exploring multivariate spatially referenced data. We describe a sampling of this work here.

In McDonald & Willis (1987), a grand tour (Asimov 1985, Buja & Asimov 1986) is linked to an image to assess the clustering of landscape types in the band space of a LandSat image taken over Manaus, Brazil. In Carr et al. (1987) and Monmonier (1989), a scatterplot matrix is linked to a map view. In REGARD (Unwin et al. 1990), map views are linked with histograms and scatterplots and, moreover, diagnostic plots for assessing spatial dependence are also available. Another exploratory system that links histograms and scatterplots with latitude and longitude (and depth) coordinates is discussed in MacDougall (1992). In Carr et al. (1992), (bivariate) ray-glyph maps have been linked with scatterplots. Klein & Moreira (1994) report on an interface between the image program MTID and XGobi, used for the exploratory analysis of agricultural images. DiBiase et al. (1994) provide an overview on existing multivariate (statistical) displays for geo-

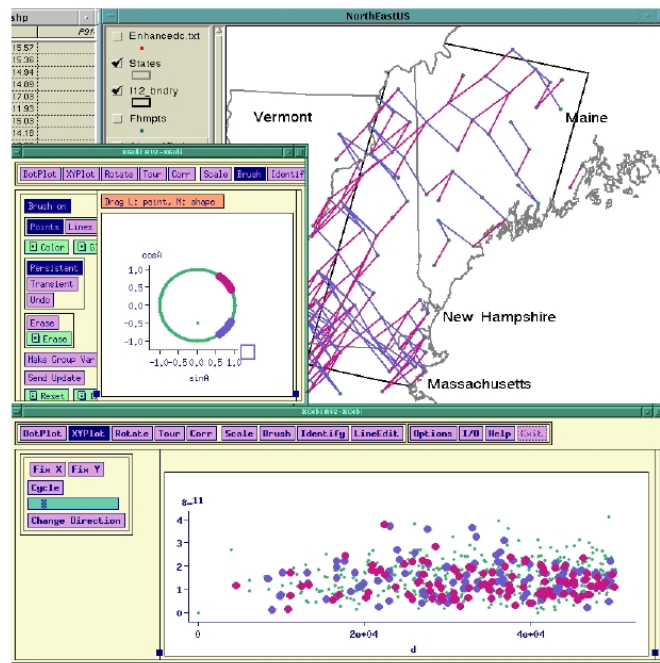
graphic data. Some recent developments are the cartographic data visualizer, *cdv* (Dykes 1996), where a variety of plots are linked with geography, the Space–Time–Attribute Creature/Movie, STAC/M (Openshaw & Perrée 1996), that searches for patterns in GIS data bases under the control of a Genetic Algorithm, and an exploratory spatial analysis system in XLisp–Stat (Brunsdon & Charlton 1996). The reader of these references should note that in most geographically influenced publications authors distinguish between *geographic (or cartographic) space* and *attribute (or data) space* but rarely use the statistical expression *variable* to relate to the latter one.

In addition to the ArcView/XGobi link, there are several other examples where GISs and (graphical) statistical packages have been linked. Williams et al. (1990) demonstrate how S and the GRASS GIS can be jointly used for archaeological site classification and analysis. Scott (1994) links STATA with ArcView and the spatial data analysis software SpaceStat has been linked with ARC/INFO^{TM¶} (Anselin et al. 1993) and with ArcView (Anselin & Bao 1996, 1997). In Haining et al. (1996), the designing of a software system for interactive exploration of spatial data by linking to ARC/INFO has been discussed, and in Zhang & Griffith (1997), a spatial statistical analysis module implemented in ArcView using Avenue has been discussed. MathSoft (1996) describes the S+GISLink, a bidirectional link between ARC/INFO and S–PLUS[®],^{||} and Bao (1997) describes the S+Grassland link between S–PLUS and the Grassland GIS. Finally, a comparison of the operational issues of the SpaceStat/ArcView link and the S+Grassland link has been given in Bao & Anselin (1997).”

¶ *ARC/INFO* is a trademark of Environmental Systems Research Institute, Inc.

|| *S–PLUS* is a registered trademark of StatSci, a division of MathSoft, Inc.

Spatial Data Analysis in the Linked ArcView 2.1 and XGobi Environment



Material provided by
Dianne Cook
Iowa State University
Department of Statistics
dicook@iastate.edu

Presented by
Jürgen Symanzik
Iowa State University
Department of Statistics
symanzik@iastate.edu

<http://www.public.iastate.edu/~dicook> <http://www.public.iastate.edu/~symanzik>

GIS SEMINAR SERIES @ THE UNIVERSITY OF MICHIGAN
Workshop, November 1996

10.3.1 Basic XGobi Layout

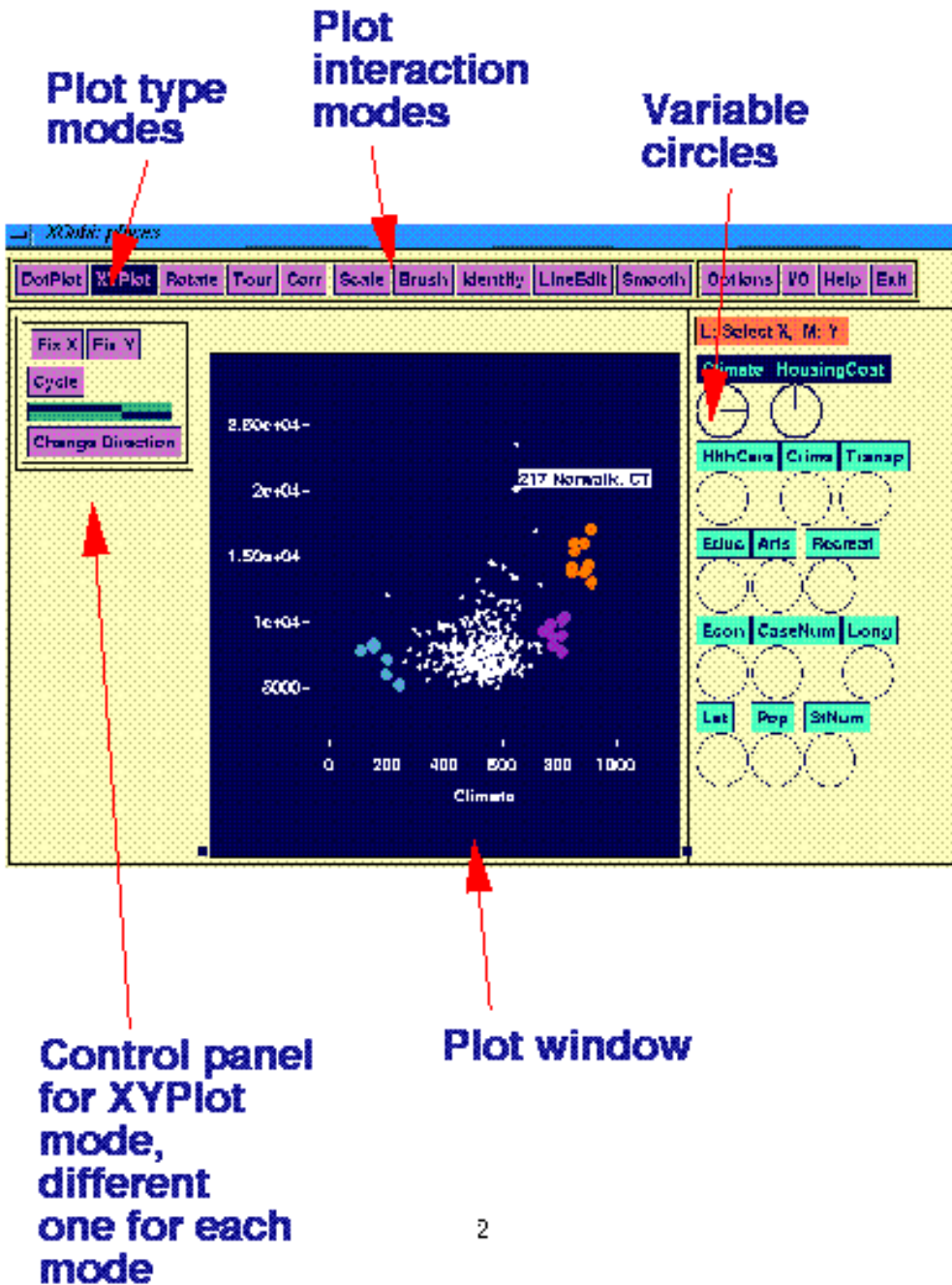


Figure 129: XGobi main window with main components labeled.

10.3.2 Graphical Methods

A most useful way to tackle the visualization of multivariate data is by breaking the task down into smaller pieces, use simple graphics (*focus*) but *link* several of them to understand how each view relates to each other.

Focusing: is placing attention on a particular aspect of the data by selecting subsets (panning and zooming or slicing) or dimension reduction (projection or variable selection).

Linking: is connecting multiple focused views, in parallel (simultaneous) by brushing or as a sequence over time by using animation and motion.

10.3.2.1 Linked Brushing (and Identification)

Linked brushing is the dynamic changing of symbols (glyphs) or colors in one plot which simultaneously changes corresponding points in other plots. Most classic example is brushing in a matrix of pairwise scatterplots. Linked identification is where brushed points are identified, for example, by labels rather than colored.

Example 1 (*Figure 130*)

About the data: *The “places data” were distributed to interested American Statistical Association (ASA) members a few years ago so that they could apply contemporary data analytic methods to describe these data and then present results in a poster session at the ASA annual conference. Latitude and longitude have been added by Paul Tukey.*

The first dataset is taken from the Places Rated Almanac, by Richard Boyer and David Savageau, copyrighted and published by Rand McNally. This book order ISBN number is 0-528-88008-X, and it retails for \$14.95 (Boyer & Savageau 1981). The data are reproduced on disk by kind permission of the publisher, and with the request that the copyright notice of Rand McNally, and the names of the authors appear in any paper or presentation using these data.

The nine rating criteria used by Places Rated Almanac are: Climate and Terrain, Housing, Health Care and Environment, Crime, Transportation, Education, The Arts, Recreation, Economics.

For all but two of the above criteria, the higher the score, the better. For Housing and Crime, the lower the score the better.

The scores are computed using the following component statistics for each criterion (see the Places Rated Almanac for details):

Climate and Terrain: very hot and very cold months, seasonal temperature variation, heating- and cooling-degree days, freezing days, zero-degree days, ninety-degree days.

Housing: utility bills, property taxes, mortgage payments.

Health Care and Environment: per capita physicians, teaching hospitals, medical schools, cardiac rehabilitation centers, comprehensive cancer treatment centers, hospices, insurance/hospitalization costs index, fluoridation of drinking water, air pollution.

Crime: violent crime rate, property crime rate.

Transportation: daily commute, public transportation, Interstate highways, air service, passenger rail service.

Education: pupil/teacher ratio in the public K-12 system, effort index in K-12, academic options in higher education.

The Arts: museums, fine arts and public radio stations, public television stations, universities offering a degree or degrees in the arts, symphony orchestras, theatres, opera companies, dance companies, public libraries.

Recreation: good restaurants, public golf courses, certified lanes for tenpin bowling, movie theatres, zoos, aquariums, family theme parks, sanctioned automobile race tracks, pari-mutuel betting attractions, major- and minor- league professional sports teams, NCAA Division I football and basketball teams, miles of ocean or Great Lakes coastline, inland water, national forests, national parks, or national wildlife refuges, Consolidated Metropolitan Statistical Area access.

Economics: average household income adjusted for taxes and living costs, income growth, job growth.

Purpose: Which places in the USA have good climate and moderate housing cost?

Action: Start up two XGobi's on the places data.

```
prompt% xgobi places &  
prompt% xgobi places &
```

In one window show an XYPlot of Latitude vs Longitude (click on the variable circle of Latitude with the middle mouse button, and the variable circle of Longitude with the left mouse button). Scale the plot so that the shape of the USA is more apparent (click on scale, then use the arrows in the control panel to change the plot shape).

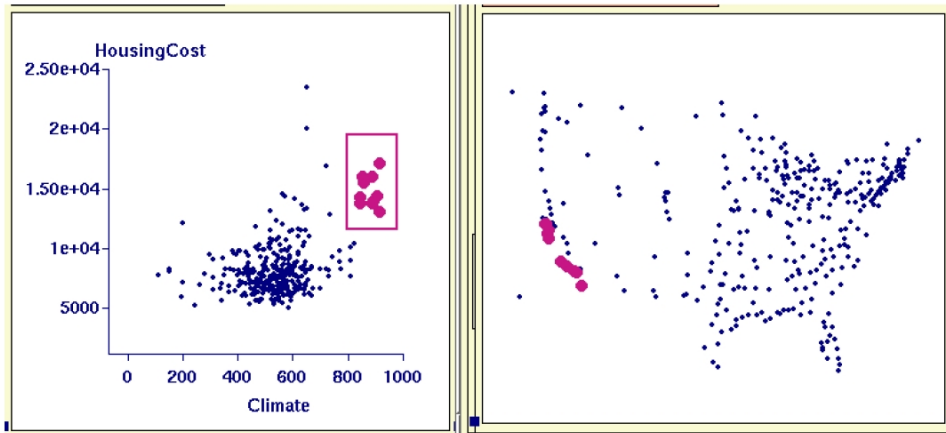


Figure 130: Places rated data: nice climate, expensive housing places are all in California.

In the second window show an XYPlot of Climate vs Housing Cost.

In the plot of Climate vs Housing Cost a small cluster can be seen to the far right about the middle vertically. Brush this cluster (click on brush, select a color by dragging the mouse down the color menu, select a glyph similarly, then move the rectangular brush over the cluster with the left mouse button, reshaping the size of the brush with the middle mouse button).

All of these nice climate, expensive places are in California!

Identify the place names (click on identify, move the cursor over the brushed points and the names will popup).

Exercise 1

- (i) Are there some nice climate, cheap places to live? Where are they?*
- (ii) Now think for a second.... From your current affairs knowledge, given what you know about the growth of Microsoft and other high-tech industries in the area under study is this information current? What else is wrong with this picture, do you know anything about the weather in the region (look back at the way Climate is evaluated)?*
- (iii) Find the most expensive place to live. What are the other attributes of this place?*
- (iv) What place has the highest rating on the Arts?*

Example 2

About the data:

This data set consists of the percentage composition of 8 fatty acids (palmitic, palmitoleic, stearic, oleic, linoleic, eicosanoic, linolenic, eicosenoic) found in the lipid fraction of 572 Italian olive oils. An analysis of this data is given in Forina et al. (1983). There are 9 collection areas, 4 from southern Italy (North and South Apulia, Calabria, Sicily), two from Sardinia (Inland and Coastal) and 3 from northern Italy (Umbria, East and West Liguria).

Purpose: *Did you know that the olive oils from different regions of Italy can be distinguished by their fatty acid composition? The aim of the study on this data is to find combinations of the fatty acids which distinguish the oils from different regions.*

Action:

Start up XGobi on the olive oil data.

```
prompt% xgobi olive &
```

- (i) Look at how the regions are distinguished by color and areas are distinguished by glyph (from initial plot and using identify).*
- (ii) The region of southern Italy's oils can be recognized as different from all other regions by the presence or absence of one fatty acid. Which fatty acid is it? (Using dotplot change between variables until one variable shows a separation. A dotplot is a univariate plot, like a histogram to be read sideways. The horizontal axis is used for randomly jittering points that have the same numerical value.)*
- (iii) Erase the points corresponding to the southern Italy (go into brush and select erase, and move the brush over the red points - easier if you have the dotplot of region showing - then pull down the erase menu and select "delete erased points").*

10.3.2.2 Dynamic Linking using Tours

10.3.2.2.1 Grand Tour

Definition 1 *A grand tour is a continuous 1-parameter family of d -dimensional projections of p -dimensional data which is dense in the set of all d -dimensional projections in \mathbb{R}^p . The parameter is usually thought of as time.*

This means that each projection shown can be indexed by a time parameter. As time is allowed to wander off to ∞ the grand tour will show all possible d -dimensional projections of the data, which is the meaning of “dense in the set of all projections”.

A grand tour offers a multitude of aspects simultaneously in relationship to one another. If the data is intrinsically 0-, 1-, or 2-dimensional (that is, clusters, curves or surfaces) the human eye can pick up the “gestalt” almost instantly. (We are adept at detecting and recognizing moving objects.)

Three-dimensional rotation can be considered a special case of the tour, where the dimension of the data is $p = 3$.

Example 3 *Watch the olive oil data in a grand tour (click on the variable circles of all fatty acids except eicosenoic, and also those of region and area so these are removed from the plot).*

- (i) Keep the axes on the plot and watch how the variables fade in and out of view.*
- (ii) Change the speed of the tour by dragging on the top scrollbar.*
- (iii) Turn off the axes and watch the shapes that the data forms. See how the points separate into clusters in some projections.*
- (iv) Stop the tour at a view where the green group (Sardinia) is separated from the purple group (North Italy). Turn on the axes again, and look for the variable(s) which contribute most in the direction of the separation. Compare the view given in a pairwise plot of these variables and see if the separation can be seen similarly here.*
- (v) If you have difficulty stopping at a projection which separates Sardinian oils from those of Northern Italy use projection pursuit (click on projection pursuit and then*

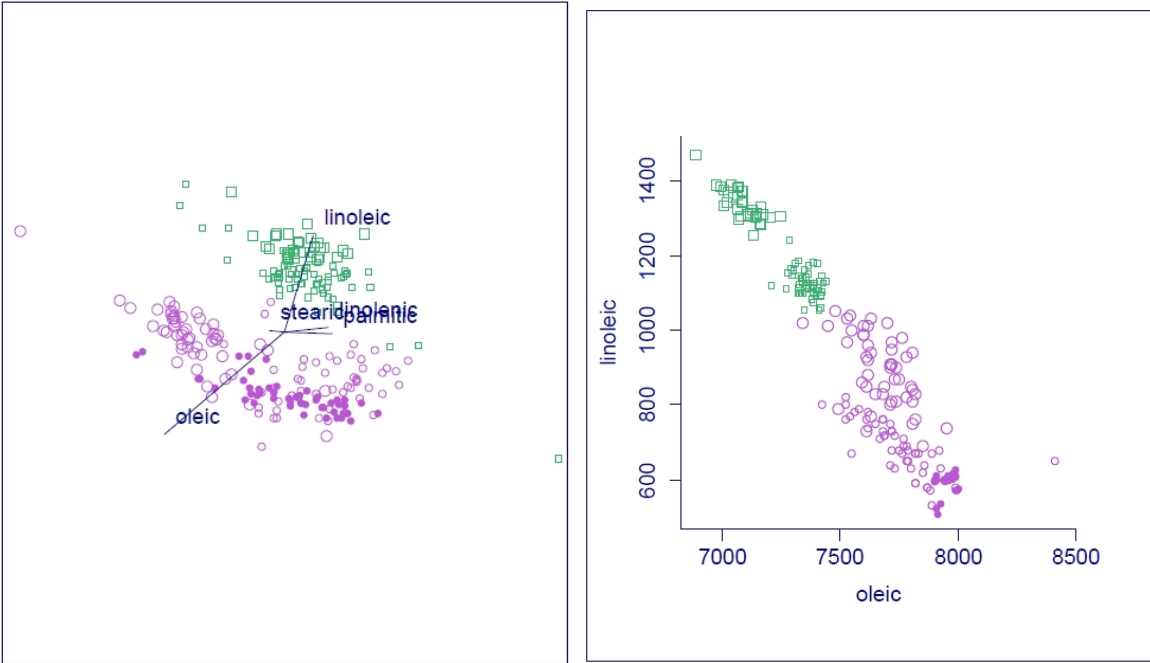


Figure 131: Olive oil data: projection showing separation of Sardinian oils from those of Northern Italy shows that the difference is in two fatty acids, oleic and linoleic.

the optimize button - you may have to turn optimize on and off several times to get to the same maximum as shown in Figure 131). You should soon see a projection where the two regions are fairly well separated. Turning the axes (I/O menu on top line of controls) on you will find that two acids, oleic and linoleic are the main ones contributing to the difference (Figure 131). Look at the XYPlot of these two acids.

Interpretation: *The northern Italian oils contain small amounts of linoleic acid and large amounts of oleic acid, the reverse is true for Sardinian oils. You could set up a discrimination rule for oils from the two regions based on these two variables.*

Exercise 2 See if you can find a composition of acids that would allow you to distinguish Inland Sardinian oils from coastal Sardinian oils!

10.3.2.2.2 Some Additional Optional Exercises

Exercise 3

(theoretical) Standard 5-dimensional normal sample.

```
prompt% xgobi norm.500.5 &
```

Note the circular nature of the data in most projections, but also some “apparent” structure like two outliers.

Exercise 4

(theoretical) 5-dimensional normal with large variance differences between variables.

```
prompt% xgobi norm.diffvar.500.5 &
```

Note that the data looks linear in some views, especially in variables 4 and 5 which have very small variance in comparison to variables 1,2,3.

Exercise 5

Look at the flea beetle data in a tour.

```
prompt% xgobi flea &
```

These are 6 measurements on 3 species of flea beetles (3 leg measurements, 2 antennae measurements and 1 head measurement). The aim is to find a projection of the data that can be used to discriminate between the three species, so newly found beetle can be classified. (Turn the axes drawing off, Options menu.). (Paint all the points to be the same color.) See if you can see three groups of points moving separately. Stop the tour when you see one group standing out, and use Identify to check whether the points do correspond to one species.

Click on Projection Pursuit, and then Reinit and Optimiz. The tour should stop when the view shows three separate groups. Use Identify to see that these groups correspond to the three species. Which variables contribute most to the horizontal projection, and which to the vertical?

Exercise 6

Look at the laser data in a tour.

```
prompt% xgobi laser &
```

This is data collected at Bellcore, New Jersey on the behavior of laser beams, in terms of output current and frequency, based on the input power to the front and back of the laser. Looking only at the first 3 variables there are (at least) three important features of this data. See if you can find them.

10.3.3 Linking Between More Complicated Objects for Data Analysis

10.3.3.1 Panel Data: Shopping Frequency — 52 Weeks of Customer Shopping

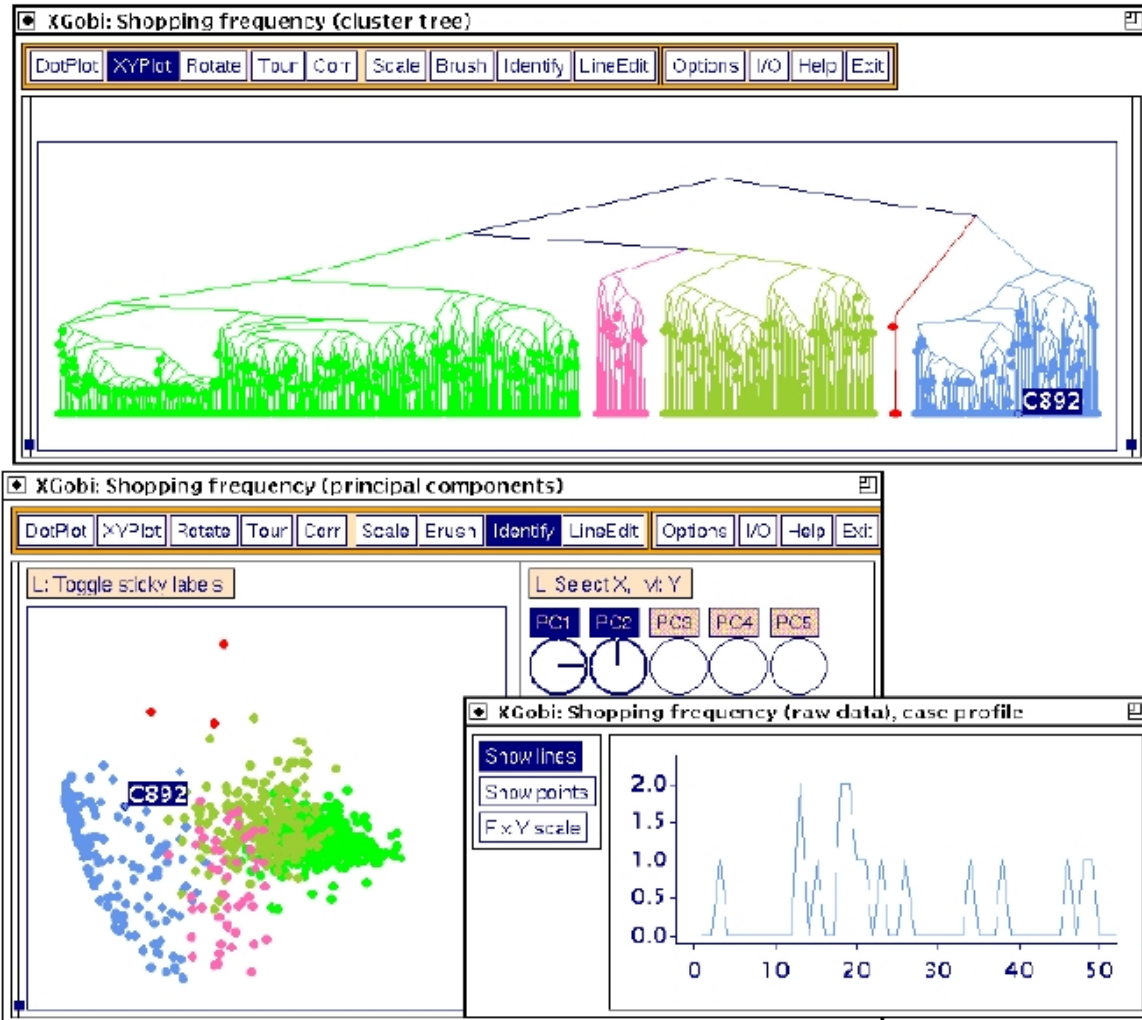


Figure 132: Panel data of customer shopping for ketchup: clusters of customers brushed in the top plot of cluster tree, principal component 2 vs principal component 1 in bottom left shows division of clusters, time profile of one blue customer in bottom right plot shows sporadic shopping for ketchup.

Linked views in XGobi (Koschat & Swayne 1996):

Cluster Tree - grouping

Principal Components - clusters “separated” in first two axes, some interpretability on which variables create separation.

Time Series - Red customer differs from Blue customer by more shopping in the first ~ 30 weeks.

10.3.3.2 Links with a GIS for Exploring Spatial Data

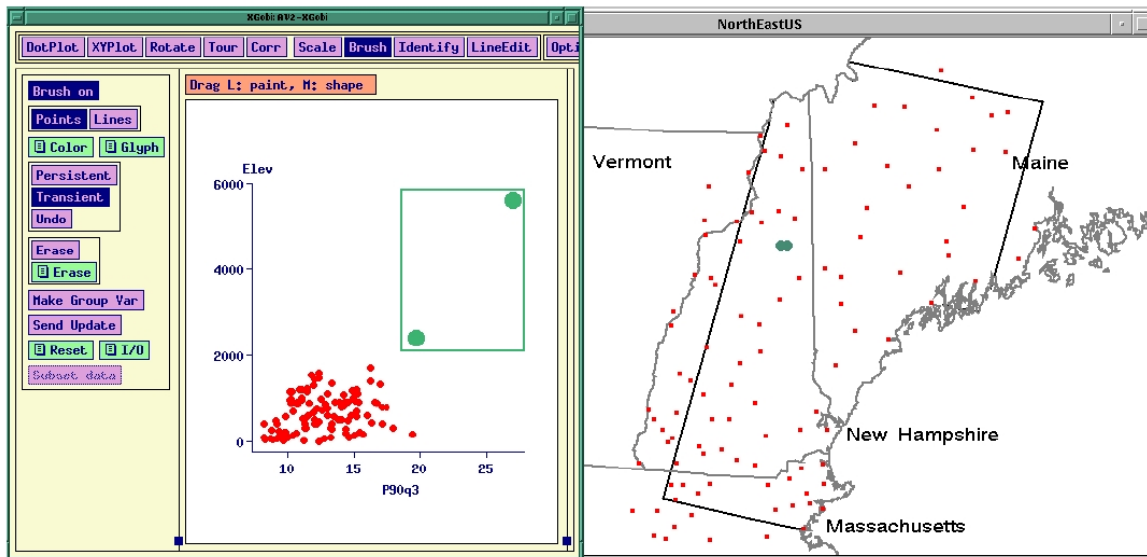


Figure 133: Link between ArcView and XGobi where multiple attributes are passed to XGobi for multivariate exploration. Here we see a large precipitation value which also has a high elevation value. It is a recording station on top of a mountain which has unusually high precipitation.

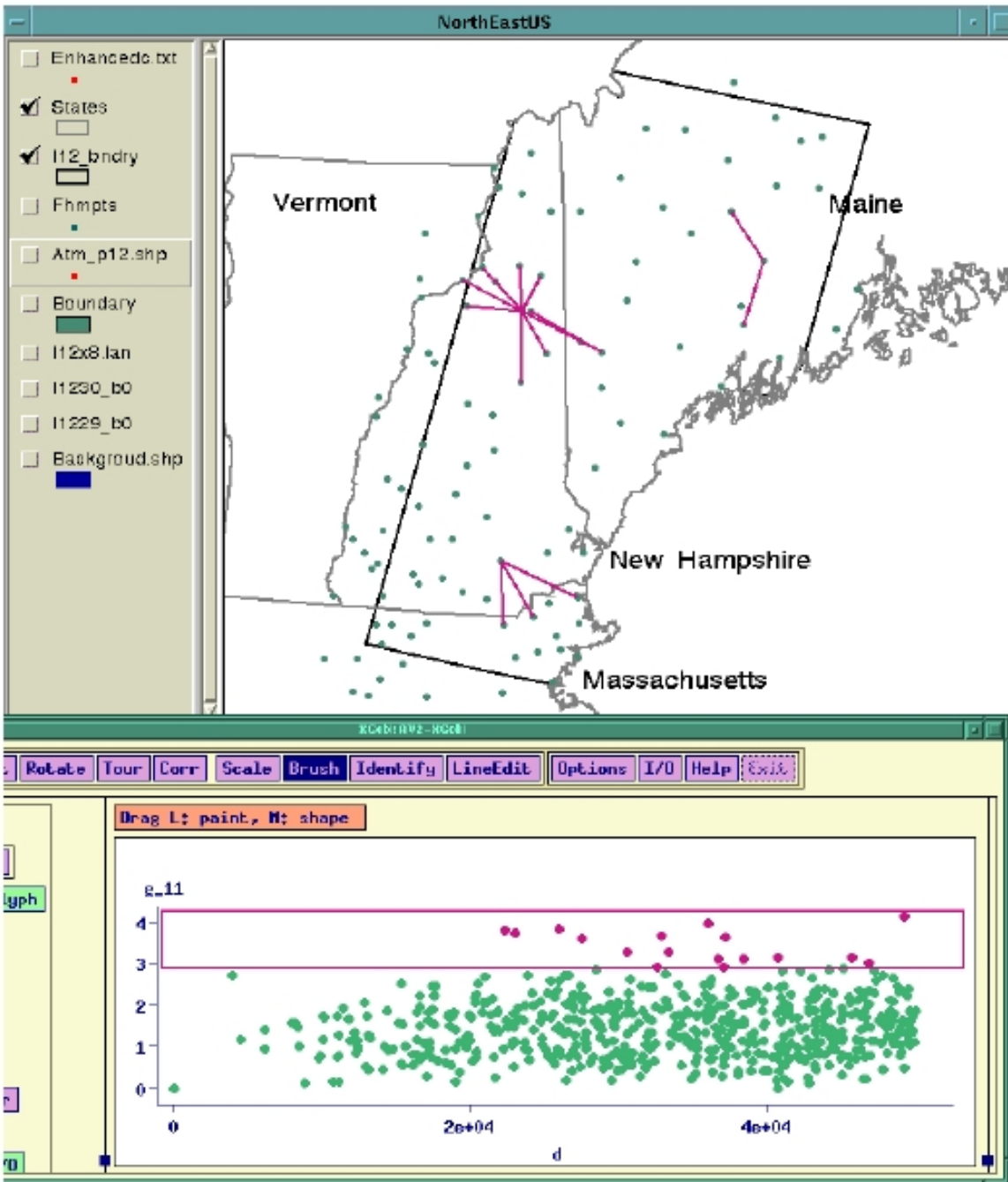


Figure 134: Variogram cloud of precipitation shown in XGobi. Large precipitation differences are brushed in red, and the pairs of points are connected by lines in the ArcView window. Several potential spatial outliers are revealed - these are the locations which have two or more lines running from them.

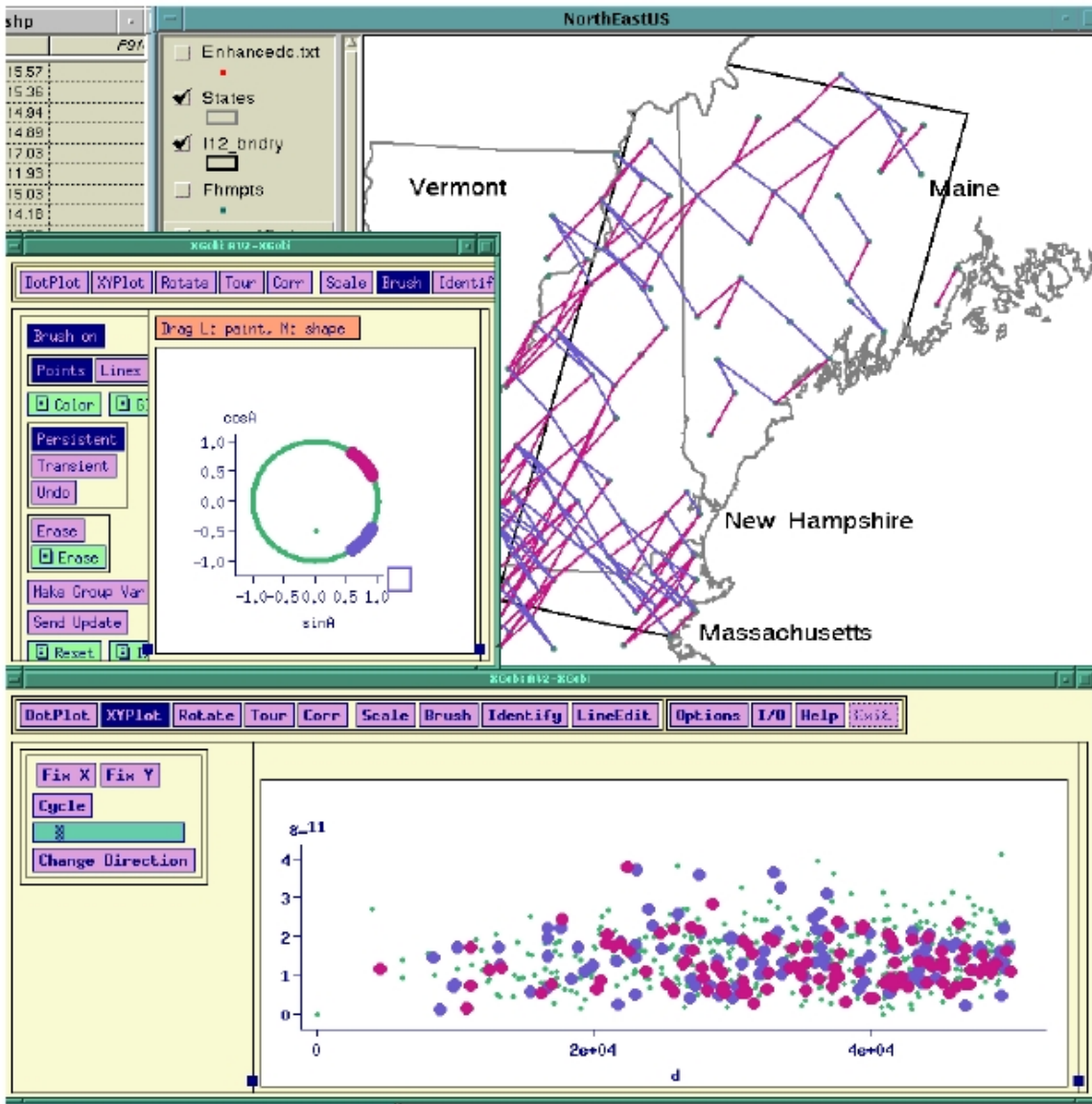


Figure 135: Variogram cloud of precipitation shown in XGobi. Two different angle classes are brushed in the cloned XGobi window: north-east (red), and south-west (blue) directions. In the variogram cloud it appears that there is a small difference between the two angle classes: there seems to be slightly more spatial variability in the south-west direction because the blue points are a little higher.

10.4 Interactive and Dynamic Graphics in R

10.4.1 R Package *rggobi*

The R package *rggobi* is closely related to the Data Viewer/XGobi/GGobi family, introduced in Section 10.1.3. Wickham et al. (2008) introduce *rggobi*. Technical details are provided in Lawrence et al. (2009). Cook & Swayne (2007) deals with GGobi and *rggobi*. Additional details on *rggobi* can be found at <http://cran.r-project.org/web/packages/rggobi/index.html>.

Example 1:

```
library(rggobi)
```

Unfortunately, *rggobi* did not load on my laptop due to conflicts with other libraries I have installed. It should work with a “cleaner” R version than the one I have. Then experiment with the sample code from <http://cran.r-project.org/web/packages/rggobi/rggobi.pdf>.

10.4.2 R Package *tourr*

The R package *tourr* introduces the Grand Tour, discussed in Section 10.2.5, to R. Wickham et al. (2011) deals with *tourr*. Additional details on *tourr* can be found at <http://cran.r-project.org/web/packages/tourr/index.html>.

Example 2:

```
library(tourr)

# iris data
animate(iris[,1:4], grand_tour(), display_xy())

# olive data
animate(olive[,3:9], grand_tour(), display_xy())

# flea data
animate(flea[,1:6], grand_tour(), display_xy())
```

10.4.3 R Package *iplots*

The R package *iplots* is closely related to the REGARD/MANET/Mondrian family, introduced in Section 10.1.1. Theus & Urbanek (2004) introduce *iplots*. Theus & Urbanek (2009) deals with Mondrian and *iplots*. Additional details on *iplots* can be found at <http://cran.r-project.org/web/packages/iplots/index.html>.

Example 3: (Based on a Student Project by Rong Xia in Spring 2009)

```
#####  
##                               Input data                               ###  
#####  
  
data_url <- "http://www.interactivegraphics.org/Datasets_files/rent.txt"  
  
rents <- read.table(url(data_url), head = TRUE, sep = "\t")  
  
head(rents)  
  
#####  
##                               Activate iplots                               ###  
#####  
  
library(iplots)  
  
attach(rents)  
  
iplot(Size, Rent)  
  
iset.col(Num..Rooms)  
  
iabline(lm(Rent~Size))  
  
ibar(Num..Rooms)  
  
iset.col()  
  
ibar(Built, isSpine = TRUE)  
  
ihist(Rent)  
  
iset.select(Rent>855.6)  
  
iset.selectNone()  
  
ibox(Rent, Good.Neighborhood)  
  
ibox(Rent, Best.Neighborhood)  
  
ibox(Rent, Warm.Water)
```

```

ibox(Rent, Central.Heating)

ibox(Rent, Tiled.Bath)

ibox(Rent, Extra.Bath)

ibox(Rent, Premium.Kitchen)

iplot(Size, Rent)

d <- iplot.data()

iabline(lm(d$y ~ d$x), col = "black")
ilines(lowess(d$x,d$y), col = "#0000c0")
ilines(c(0,0), c(0,0), col = "marked", visible = FALSE)
cat("Select 'Break' from the menu of any plot to return back to R.\n")

while (!is.null(ievent.wait()))
{
  if (iset.sel.changed())
  {
    s = iset.selected()
    if (length(s) > 1)
      iobj.opt(x = lowess(d$x[s],d$y[s]), visible = TRUE)
    else iobj.opt(visible = FALSE)
  }
}

for(i in 1:3) iobj.rm()

iplot.off()

```

10.4.4 R Package *animation*

The R package *animation* introduces the concept of general animations to R. Xie & Cheng (2008) introduce animation. Additional details on animation can be found at <http://cran.r-project.org/web/packages/animation/index.html> and <http://animation.yihui.name/start>.

Example 4: (Based on a Student Project by Jin Ying in Spring 2009)

```

#####
###                                     ###
### the following code is modified based on the R documentation for ###
###                               the animation package                ###
###                                     ###
#####

#### install the package #####

```

```

library(animation)

##### example 1: buffon's needle #####

oopt = ani.options(nmax = 500, interval = 0)
opar = par(mar = c(3, 2.5, 0.5, 0.2), pch = 20, mgp = c(1.5, 0.5, 0))

par(mfrow=c(1,2))

buffon.needle(l = 0.8, d = 1, redraw = TRUE, mat = matrix(c(1, 3, 2, 3), 2),
heights = c(3, 2), col = c("lightgray", "red", "gray", "red", "blue",
"black", "red"), expand = 0.4)

##### swans data to apply for the k-means (given in the animation package) cluster analysis #####

data_url1 = "http://www.math.usu.edu/~symanzik/teaching/2009_stat6560/RDataAndScripts/jin_ying_project2_swans.txt"
swans = read.table(data_url1, header = T)

swan = swans[,c(1,3)]
oopt = ani.options(interval = 2, nmax = 1000)
op = par(mar = c(3, 3, 1, 1.5), mgp = c(1.5, 0.5, 0))
kmeans.ani(swan, centers = 2, pch = 1:2, col = 1:2)

##### iris data #####

irissub = iris[,c(1,3)]
oopt = ani.options(interval = 2, nmax = 1000)
op = par(mar = c(3, 3, 1, 1.5), mgp = c(1.5, 0.5, 0))
kmeans.ani(irissub, centers = 3, pch = 1:3, col = 1:3)

```


11 Graphics Galleries and Sources on the Web

Graphics Galleries:

- Michael Friendly' Statistics and Statistical Graphics Resources: <http://www.math.yorku.ca/SCS/StatResource.html>
- Romain François' R Graph Gallery: <http://addictedtor.free.fr/graphiques/>
- Paul Murrell's R Graphics Web page, accompanying Murrell (2006): <http://www.stat.auckland.ac.nz/~paul/RGraphics/rgraphics.html>

Social Data Analysis Web Sites:

- NameVoyager - Baby Name Wizard of Most Popular Baby Names (Martin and Laura Wattenberg): <http://www.babynamewizard.com/voyager>
- Many Eyes (IBM): <http://many-eyes.com>
- Swivel: <http://www.swivel.com/> — This Web site no longer exists, but there is an interview that describes the rise and fall of Swivel: <http://eagereyes.org/criticism/the-rise-and-fall-of-swivel>
- StatCrunch (Webster West): <http://statcrunch.com/>
- Talks: Hans Rosling: Debunking Third-World Myths with the Best Stats You've Ever Seen (TED.com — featured in Time, April 27, 2009, p. 44): http://www.ted.com/index.php/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen.html; see <http://www.gapminder.org/> for more details and examples

Statistical Graphics Section:

- Statistical Computing and Statistical Graphics Sections of the American Statistical Association (ASA) Home: <http://stat-computing.org/>
- Statistical Graphics Video Library: <http://stat-graphics.org/movies/>

Appendix

Homework Assignments

Homework Assignment 1 (1/12/2011)

15 Points — Due 1/21/2011 (pdf file via e-mail by 12noon)

(i) (15 Points — Individual Work) We will work with data from the Deepwater Horizon oil spill as one of our main data sets this semester. Think of at least ten other disasters since 1900 where technology was involved (i.e., no wars or diseases), such as the Titanic, the Challenger disaster, or the Chernobyl nuclear reactor disaster.

- Create a diagram in R that shows a timeline on the horizontal axis and marks your ten (or more) disasters. Use google Images search for *timeline* to get some idea about effective (and less effective) timeline plots.

R packages such as *diagram* (<http://cran.r-project.org/web/packages/diagram/>) or *igraph* (<http://cran.r-project.org/web/packages/igraph/index.html>) may be helpful to create your timeline. Or, you can simply draw your timeline “manually” by adding text, line segments, etc. to an empty plot area in R. All figures for my Stat 6710/20 lecture notes were created this way. For example, the figure for Theorem 1.2.1 (iv) on page 6 of my Stat 6710 lecture notes, accessible at http://www.math.usu.edu/~symanzik/teaching/2010_stat6710/lect_main_full.pdf, has been created as follows:

```
#R code to produce the graphic for Theorem 1.2.1 iv
#
pdf("lect_theorem1.2.1_iv.pdf", height = 8, width = 8)
plot(c(0,1), c(0,1), xlab = " ", ylab = " ", type = "n", axes = F)
lines(c(0, 1, 1, 0, 0), c(0, 0, 1, 1, 0))
deg = seq(0, 2 * pi, length = 1000)
x = sin(deg)
y = cos(deg)
lines(0.25*x + 0.4, 0.25*y + 0.6)
lines(0.20*x + 0.7, 0.20*y + 0.4)
text(0.4, 0.6, "A", cex = 3)
text(0.7, 0.4, "B", cex = 3)
title("Theorem 1.2.1 (iv)", cex.main = 3)
dev.off()
```

- Create two different plots that show the number of human deaths for your ten (or more) selected disasters. Sort your data in different ways and possibly transform the number of deaths (e.g., by taking the log). Provide meaningful titles and labels for your plots.
- Write a report in \LaTeX and translate via `pdflatex`. You should include an informative title page with your name, course information, and homework information.

Your main answer should contain a list of your ten (or more) disasters with clickable links (use `hyperref`) to an online reference where the disaster is described in more details.

Include your three figures (either in pdf or jpg format) into your \LaTeX document. Add a meaningful caption for each figure. Create some main text that briefly describes each of your figures and that links to these figures. Compare your two figures that show the number of human deaths and explain which of these is better (in your opinion).

Include your R code in the appendix (in a `verbatim` environment).

- Submit the single pdf file that is resulting from your work via e-mail to `symanzik@math.usu.edu` by 12noon on Fr 1/21/2011.
- Hint: If some of the expressions above mean nothing to you, google for them. Almost all documentation for R, \LaTeX , or any other software is nowadays freely accessible on the Web. It may take some time to search, but eventually you should find it! Also, if you have an idea how a graphic should look like, but have no idea how to get started, you may find it at the *R Graph Gallery* (<http://addictedtor.free.fr/graphiques/>). This Web site shows numerous graphics and it provides the R code how each graphic has been produced.

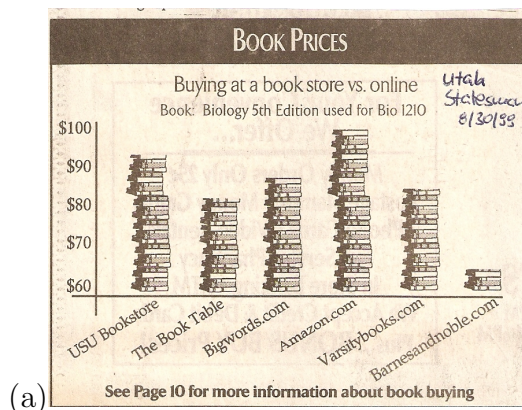
Homework Assignment 2 (2/1/2011)

30 Points — Due 2/11/2011 (pdf file via e-mail by 5pm)

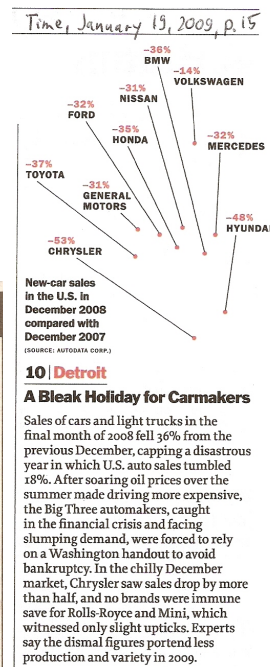
(i) (8 Points — Group Work) Read Chapters 1, 2, and 3 of Tufte (1983) “*The Visual Display of Quantitative Information*”. Then take a closer look at the figures on top of p. 55 (“Comparative Annual Cost ...”) and on top of p. 69 (“Accroissement ...”).

(*) For each of these figures, explain which rule(s) (how to construct a bad graphic) from our lecture notes the graphics designer has followed, i.e., list the rule(s) and explain why it has been followed. Demonstrate how these poor graphics might be improved. Using the data from the graphic (or your best approximation if necessary), construct a superior representation of the same information, using R. Include a short write-up (about half a page to a page) as to how you believe your version improves on the poor original.

(ii) (8 Points — Group Work) Repeat (*) for the two graphics included below.



(a)



(b)

- (iii) (8 Points — Group Work) Update your timeline plot from Homework 1 with events related to the Deepwater Horizon oil spill. Start with the actual explosion, the sinking, major attempts to stop the oil leakage, and the successful capping of the leak. The Data Expo 2011 Web page (<http://streaming.stat.iastate.edu/dataexpo/2011/>) may be a good starting point, but you should search for additional references beyond those listed on the Data Expo page. Also arrange (and summarize) the events from your timeline plot in a textual and/or tabular way.

When we take a closer look later at the actual data provided at the Data Expo 2011 Web page, we will see that the data are very heterogeneous. The events above may explain some of the sudden changes. There are likely other (environmental, climatic, etc.) variables that could lead to sudden changes. Hurricanes could be one such source. Find data related to hurricanes (since April 2010) that might have had an impact on the oil spill. These could be graphics (with dates) of hurricane paths, but even better, exact data with geographic locations of the hurricane centers. Document these data sources and download such data to your own computer. If there are other (environmental, climatic, etc.) variables that could have had an impact on the oil spill, do the same for these variables. Provide a brief textual summary of these additional data.

- (iv) (6 Points — Individual Work, due 2/18/2011, 5pm) Find a fresh example of a poor statistical graphic. Do not choose your example from one of the books for this class, or another book that specifically relates to graphics and charts, but from an original (preferably recent) source. Journal articles, newspapers, magazines, and scholarly books are all appropriate sources.

Each student must use a different graphic for this question. In fact, only the first person who notices a poor graphic can work on that particular graphic. If you notice a poor graphic, you have to send an e-mail to me, indicating something similar to the following: *“I found a poor graphic on page 1 of the Salt Lake Tribune on Wed 2/2/2011, titled ... that deals with ...”* Your bad graphic must meet at least three rules for bad graphics from Chapter 1. Good sources for bad graphics are CNN, Time magazine, the Utah Statesman, Wikipedia, and many other online sites, but also textbooks and journal papers.

Repeat (*) for your graphic. Include the original (electronic) figure or turn in a scan or high quality photo of the original together with a precise reference (source and page, URL, etc.) where you found that figure.

General Submission Rules: (for group and individual homework, reports, etc.)

All your submissions this semester must be typeset in \LaTeX . In fact, your submissions should translate via `pdflatex`. Figures (from scans or from graphical software) must accompany your \LaTeX document in electronic format. R code and data sets must be directly accessible from your document. You should assume that all documents reside in the same directory. For testing, one student should finalize all documents while another student checks the intended submission for completeness on a different computer. \LaTeX warnings are OK, but \LaTeX error messages will result in point deductions (depending on how much effort it takes on my side to fix a problem). Submit your files via e-mail to `symanzik@math.usu.edu`. In case your submission consists of four or more individual files, you have to collect these in a zip file and just submit this single zip file.

You will be allowed one original submission plus one revision where you should make changes, based on my feedback. For the first submission, just submit your resulting pdf file. For the second submission, do not submit a pdf file as I will retranslate all documents on my side. There will likely be opportunities for bonus points that will be awarded to the best group (or best individual) submission.

Your files should be named as follows (or in a closely related way):

```
groupI_hwJ_main.tex
groupI_hwJ_figK.pdf
groupI_hwJ_qL.R
groupI_hwJ_qM_data.xxx
```

```
lastname_firstname_hwJ_qL_main.tex
lastname_firstname_hwJ_qL_figK.pdf
```

```
lastname_firstname_projectN_main.tex
lastname_firstname_projectN_figK.pdf
```

where I, J, K, L, M, and N will be replaced by appropriate integers. xxx can be any acceptable extension for R data files. Include comments in your files where possible, e.g., dates, names, purpose of a file, etc. Also, include a `Readme.tex` that provides detailed instructions how I have to recreate a pdf file on my side.

Overall, please follow the general submission rules from Homework 1. In particular, please follow these instructions:

- Include a title page, including course name, student names, number of homework, and the submission date.
- Link your figures, tables, sections, R code, etc. within your document. Include meaningful figure and table captions. Provide clickable links to outside sources.
- Start answers to each question on a new page, clearly labeling which question is being answered. To combine multiple tex files created by different students, work with `\input`. Check that your overall document is consistent in appearance once you have combined individual tex files, e.g., consistent labeling, referencing, use of titles and headlines, etc.
- Include all R code in appendices so that I can run this code on my side. Test your code so that it runs correctly when copied from the pdf back into R. Incorrect R code will result in point deductions. I will sample some R code from the original submission, but will do some more careful testing only for the second submission.
- Include a References section. Make sure to include printed references, online references, and sources that inspired your R code. Do not forget to cite R and R packages you have used.

Homework Assignment 3 (4/15/2011)

30 Points — Due Sunday 5/1/2011 (zip of all source files via e-mail by 11:59pm)

- (i) (30 Points — Individual Work) This homework will be entirely related to data from the Deepwater Horizon oil spill.

Each student in class will be randomly assigned to one or two data sets. The following topics and data sets are available:

Number	Data Group	Data Set Name
1	A	glider
2	A	boats
3	A	floats
4	B	fish-baseline & fish-june2010
5	C	epa1 (sampling)
6	C	epa2 (monitoring)
7	D	birds
8	D	turtles & mammals

You need to provide a graphical summary of the data set(s) assigned to you. See the data expo Web site at <http://streaming.stat.iastate.edu/dataexpo/2011/> for questions (“*The Challenge*”) of general interest. You can use any additional information that you can find on the Web, but you cannot ask anyone else outside our class for help.

Your tasks are as follows:

- Understand your data well! These are raw data with possible outliers, invalid locations, missing (NA) observations, and possible other problems (think of the data from the Minnesota barley trial in Section 6.5). You may first want to create simple summaries (numerical and graphical) to check for data correctness before moving on to more sophisticated plots.
- All data sets contain latitude/longitude, so include at least one (but ideally more) maps, created via the *RgoogleMaps* package. Add other relevant information to your map(s).

- All data sets contain a date, so consider to create multiple maps and difference maps.
- We expect changes over time, but we cannot expect that these changes become immediately visible after the April 20, 2010, explosion of the Deepwater Horizon. Rather, changes may occur gradually over time or may suddenly occur on day x . You may have to aggregate data over time. The “week” information in some of the data sets may or may not be useful as sudden changes may have occurred in the middle of a week.
- How can we best visualize spatio-temporal data? Perhaps some of the ideas from the entries to the Data Expo 2006, posted at <http://stat-computing.org/dataexpo/2006/entries.html>, may be suitable for your data set as well. Of course, take a closer look at the three winning posters first.
- Data sets within the four data groups (A to D) are similar, so you should coordinate and discuss your approach with the other student(s) assigned to the same data group.
- Each student has to write a report related to his/her data set(s) that consists of exactly 10 pages of main text. Your sections should be titled from “Introduction” to “Conclusion”. Just a collection of figures is not an acceptable report. Your figures must be explained in the main text and there must be motivation behind your graphical analysis such that a reader not familiar with your data set(s) can follow your report.
- Don’t forget to verbally describe your data set(s) in your report, e.g., what are the variables, what are the units, whether there were any outliers (and what you did with those), etc.
- You are allowed extra pages (beyond the 10 pages of main text) for the title page, table of contents, list of figures, references, appendices, etc.
- Your R code must be placed in an appendix. Also, another appendix should contain all relevant figures you created that did not fit into the main part, e.g., diagnostic figures and numerical summary statistics you created to check the data correctness or any figures you created to check the overall data distributions and patterns. Even figures in an appendix need to be briefly mentioned in the text, often in a short paragraph at the start of an appendix that summarizes the content of this appendix.
- Follow all requirements and submission rules from Homeworks 1 & 2.

- You should download your assigned data set and work with a local copy. However, ultimately, your R code most work directly with the data set that is posted on the data expo Web site at <http://streaming.stat.iastate.edu/dataexpo/2011/>. This means that you cannot manually modify the data set, i.e., all modifications, sorting, deletions of outliers, etc. have to be done in R. Revisit Example 2 in Section 4.2.6 how to read data directly from the Web. If you do not recall from your Stat Computing class how to perform such tasks directly in R, you may want to take a closer look at Paul Murrell’s (2009) “*Data Technologies*” book (<http://www.stat.auckland.ac.nz/~paul/ItDT/>).
- You can use any software to initially explore the data, e.g., Mondrian or GGobi, but the final results of your figures must be created via R.
- When working with colors, choose meaningful color schemes.
- All figures in the main 10 pages must be of “professional” quality, i.e., fully labeled (axes & title), readable at 100% screen resolution, appropriate for the data type in your data set (some data sets are mostly quantitative while others are mostly categorical), and have good color choices. Figures in the appendix can be “working” figures with default R settings, but still should be meaningful. For example, a figure without a caption or any axis labels is not meaningful.
- Before you submit your assignment, make sure (on a different computer) that the zip file you created is complete, i.e., that you included all figures and all L^AT_EX source files. Make sure that I can run your R code out of the appendix (it is fine to reduce the R code to scriptsize or so). Provide instructions which steps are needed to retranslate your files on my side.
- The use of Sweave and bibtex is not required, but will result in bonus points if used correctly.
- Your submission will be graded based on the criteria listed above. There is no time for a resubmission, so there is just this is one final submission. However, feel free to step by during office hours or ask specific questions via e-mail. Please do not send your full submission before the deadline and ask whether this is OK — rather ask very specific questions.
- After grading, all your submissions will be passed on to Anvar so he can further work on the Deepwater Horizon data set over the summer as part of

our planned poster submission for the the 2011 Data Expo (and as his likely MS project). As a reminder, you will be a co–author of this poster and are strongly encouraged to provide additional suggestions, feedback, and new work yourself over the summer. If, for some reason, you do not want to be a co–author of this poster and are not willing to pass on your submission to Anvar after grading, please let me know as soon as possible.

- Our auditing students are invited to work on any of these data sets and provide their suggestions and results to Anvar after the end of the semester and they can also contribute over the summer. If any of you wants to join as a co–author of the poster, please let Anvar and me know.

General Submission Rules: (for group and individual homework, reports, etc.)

All your submissions this semester must be typeset in \LaTeX . In fact, your submissions should translate via `pdflatex`. Figures (from scans or from graphical software) must accompany your \LaTeX document in electronic format. R code and data sets must be directly accessible from your document. You should assume that all documents reside in the same directory. For testing, one student should finalize all documents while another student checks the intended submission for completeness on a different computer. \LaTeX warnings are OK, but \LaTeX error messages will result in point deductions (depending on how much effort it takes on my side to fix a problem). Submit your files via e–mail to `symanzik@math.usu.edu`. In case your submission consists of four or more individual files, you have to collect these in a zip file and just submit this single zip file.

Your files should be named as follows (or in a closely related way):

```
groupI_hwJ_main.tex
groupI_hwJ_figK.pdf
groupI_hwJ_qL.R
groupI_hwJ_qM_data.xxx
```

```
lastname_firstname_hwJ_qL_main.tex
lastname_firstname_hwJ_qL_figK.pdf
```

```
lastname_firstname_projectN_main.tex
lastname_firstname_projectN_figK.pdf
```

where I, J, K, L, M, and N will be replaced by appropriate integers. xxx can be any acceptable extension for R data files. Include comments in your files where possible, e.g., dates, names, purpose of a file, etc. Also, include a Readme.tex that provides detailed instructions how I have to recreate a pdf file on my side.

Overall, please follow the general submission rules from Homework 1. In particular, please follow these instructions:

- Include a title page, including course name, student names, number of homework, and the submission date.
- Link your figures, tables, sections, R code, etc. within your document. Include meaningful figure and table captions. Provide clickable links to outside sources.
- Start answers to each question on a new page, clearly labeling which question is being answered. To combine multiple tex files created by different students, work with `\input`. Check that your overall document is consistent in appearance once you have combined individual tex files, e.g., consistent labeling, referencing, use of titles and headlines, etc.
- Include all R code in appendices so that I can run this code on my side. Test your code so that it runs correctly when copied from the pdf back into R. Incorrect R code will result in point deductions. I will sample some R code from the original submission, but will do some more careful testing only for the second submission.
- Include a References section. Make sure to include printed references, online references, and sources that inspired your R code. Do not forget to cite R and R packages you have used.

References

- Andrews, D. F. (1972), ‘Plots of High–Dimensional Data’, *Biometrics* **28**, 125–136.
- Anscombe, F. J. (1973), ‘Graphs in Statistical Analysis’, *The American Statistician* **27**(1), 17–21.
- Anselin, L. & Bao, S. (1996), Exploratory Spatial Data Analysis Linking SpaceStat and ArcView, Technical Report 9618, West Virginia University, Morgantown, WV.
- Anselin, L. & Bao, S. (1997), Exploratory Spatial Data Analysis Linking SpaceStat and ArcView, *in* M. M. Fischer & A. Getis, eds, ‘Recent Developments in Spatial Analysis’, Springer, Berlin, pp. 35–59.
- Anselin, L., Dodson, R. F. & Hudak, S. (1993), ‘Linking GIS and Spatial Data Analysis in Practice’, *Geographical Systems* **1**(1), 3–23.
- Asimov, D. (1985), ‘The Grand Tour: A Tool for Viewing Multidimensional Data’, *SIAM Journal on Scientific and Statistical Computing* **6**(1), 128–143.
- Baggerly, K. A. & Berry, D. A. (2011), ‘Science Policy: Reproducible Research’, *Amstat News* **January 2011**(403), 16–17.
- Bao, S. (1997), *User’s Reference for the S+Grassland Link*, Mathsoft, Inc., Seattle, WA.
- Bao, S. & Anselin, L. (1997), Linking Spatial Statistics with GIS: Operational Issues in the SpaceStat–ArcView Link and the S+Grassland Link, *in* ‘1997 Proceedings of the Section on Statistical Graphics’, American Statistical Association, Alexandria, VA, pp. 61–66.
- Becker, R. A. & Cleveland, W. S. (1988), Brushing Scatterplots, *in* W. S. Cleveland & M. E. McGill, eds, ‘Dynamic Graphics for Statistics’, Wadsworth & Brooks/Cole, Belmont, CA, pp. 201–224.
- Becker, R. A., Cleveland, W. S. & Weil, G. (1988), The Use of Brushing and Rotation for Data Analysis, *in* W. S. Cleveland & M. E. McGill, eds, ‘Dynamic Graphics for Statistics’, Wadsworth & Brooks/Cole, Belmont, CA, pp. 247–275.
- Bertin, J. (1977), *La Graphique et le Traitement Graphique de l’Information*, Flammarion, Paris, France.

- Bertin, J. (2005), *Sémiologie Graphique: Les Diagrammes — Les Réseaux Les Cartes (4e édition)*, Les ré-impressions des Éditions de l'École des Hautes Études en Sciences Sociales, Paris, France.
- Blasius, J. & Greenacre, M., eds (1998), *Visualization of Categorical Data*, Academic Press, San Diego, CA.
- Bolorforoush, M. & Wegman, E. J. (1988), On Some Graphical Representations of Multivariate Data, in E. J. Wegman, D. T. Gantz & J. J. Miller, eds, 'Proceedings of the 20th Symposium on the Interface between Computing Science and Statistics', American Statistical Association, Alexandria, VA, pp. 121–126.
- Boyer, R. & Savageau, D. (1981), *Places Rated Almanac*, Rand McNally, Chicago, IL.
- Brewer, C. A. (1997), 'Spectral Schemes: Controversial Color Use on Maps', *Cartography and Geographic Information Systems* **24**(4), 203–220.
- Brewer, C. A. (1999), Color Use Guidelines for Data Representation, in '1999 Proceedings of the Section on Statistical Graphics', American Statistical Association, Alexandria, VA, pp. 55–60.
- Brewer, C. A. (2003), 'A Transition in Improving Maps: The ColorBrewer Example', *Cartography and Geographic Information Science* **30**(2), 159–162.
- Brewer, C. A., Hatchard, G. W. & Harrower, M. A. (2003), 'ColorBrewer in Print: A Catalog of Color Schemes for Maps', *Cartography and Geographic Information Science* **30**(1), 5–32.
- Brewer, C. A., MacEachren, A. M., Pickle, L. W. & Herrmann, D. J. (1997), 'Mapping Mortality: Evaluating Color Schemes for Choropleth Maps', *Annals of the Association of American Geographers* **87**(3), 411–438.
- Brewer, C. A. & Pickle, L. W. (2002), 'Comparison of Methods for Classifying Epidemiological Data on Choropleth Maps in Series', *Annals of the Association of American Geographers* **92**(4), 662–681.
- Brillinger, D. R. (2002), 'John W. Tukey: His Life and Professional Contributions', *The Annals of Statistics* **30**(6), 1535–1575.
- Brunsdon, C. & Charlton, M. (1996), Developing an Exploratory Spatial Analysis System in XLisp-Stat, in D. Parker, ed., 'Innovations in GIS 3', Taylor & Francis, London, U.K., pp. 135–145.

- Buja, A. & Asimov, D. (1986), ‘Grand Tour Methods: An Outline’, *Computer Science and Statistics* **17**, 63–67.
- Buja, A., Asimov, D., Hurley, C. & McDonald, J. A. (1988), Elements of a Viewing Pipeline for Data Analysis, *in* W. S. Cleveland & M. E. McGill, eds, ‘Dynamic Graphics for Statistics’, Wadsworth & Brooks/Cole, Belmont, CA, pp. 277–308.
- Buja, A., Cook, D. & Swayne, D. F. (1996), ‘Interactive High–Dimensional Data Visualization’, *Journal of Computational and Graphical Statistics* **5**(1), 78–99.
- Buja, A., Hurley, C. & McDonald, J. A. (1986), A Data Viewer for Multivariate Data, *in* T. J. Boardman & I. M. Stefanski, eds, ‘Proceedings of the 18th Symposium on the Interface between Computer Science and Statistics, Fort Collins, CO’, American Statistical Association, Washington, D.C., pp. 171–174.
- Buja, A., McDonald, J. A., Michalak, J. & Stuetzle, W. (1991), Interactive Data Visualization Using Focusing and Linking, *in* G. M. Nielson & L. J. Rosenblum, eds, ‘Proceedings of Visualization ’91, Los Alamitos, CA’, IEEE Computer Society Press, pp. 156–163.
- Carr, D. B. (1994), Converting Tables to Plots, Technical Report 101, Center for Computational Statistics, George Mason University, Fairfax, VA.
- Carr, D. B. (2001), ‘Designing Linked Micromap Plots for States with Many Counties’, *Statistics in Medicine* **20**(9–10), 1331–1339.
- Carr, D. B., Bell, B. S., Pickle, L. W., Zhang, Y. & Li, Y. (2003), The State Cancer Profiles Web Site and Extensions of Linked Micromap Plots and Conditioned Choropleth Map Plots, *in* ‘Proceedings of the Third National Conference on Digital Government Research’, Digital Government Research Center (DGRC), pp. 269–273. http://www.dgrc.org/conferences/2003_proceedings.jsp.
- Carr, D. B., Chen, J., Bell, B. S., Pickle, L. W. & Zhang, Y. (2002), Interactive Linked Micromap Plots and Dynamically Conditioned Choropleth Maps, *in* ‘Proceedings of the Second National Conference on Digital Government Research’, Digital Government Research Center (DGRC), pp. 61–67. http://www.dgrc.org/conferences/2002_proceedings.jsp.
- Carr, D. B., Littlefield, R. J., Nicholson, W. L. & Littlefield, J. S. (1987), ‘Scatterplot Matrix Techniques for Large N’, *Journal of the American Statistical Association* **82**(398), 424–436.

- Carr, D. B. & Nicholson, W. L. (1988), EXPLOR4: A Program for Exploring Four-Dimensional Data Using Stereo-Ray Glyphs, Dimensional Constraints, Rotation, and Masking, *in* W. S. Cleveland & M. E. McGill, eds, ‘Dynamic Graphics for Statistics’, Wadsworth & Brooks/Cole, Belmont, CA, pp. 309–329.
- Carr, D. B. & Nusser, S. M. (1995), ‘Converting Tables to Plots: A Challenge from Iowa State’, *Statistical Computing and Statistical Graphics Newsletter* **6**(3), 11–18.
- Carr, D. B. & Olsen, A. R. (1996), ‘Simplifying Visual Appearance by Sorting: An Example using 159 AVHRR Classes’, *Statistical Computing and Statistical Graphics Newsletter* **7**(1), 10–16.
- Carr, D. B., Olsen, A. R., Courbois, J. P., Pierson, S. M. & Carr, D. A. (1998), ‘Linked Micromap Plots: Named and Described’, *Statistical Computing and Statistical Graphics Newsletter* **9**(1), 24–32.
- Carr, D. B., Olsen, A. R., Pierson, S. M. & Courbois, J. P. (2000), ‘Using Linked Micromap Plots to Characterize Omernik Ecoregions’, *Data Mining and Knowledge Discovery* **4**(1), 43–67.
- Carr, D. B., Olsen, A. R. & White, D. (1992), ‘Hexagon Mosaic Maps for Displays of Univariate and Bivariate Geographical Data’, *Cartography and Geographic Information Systems* **19**(4), 228–236, 271.
- Carr, D. B. & Pickle, L. W. (2010), *Visualizing Data Patterns with Micromaps*, Chapman & Hall/CRC, Boca Raton, FL.
- Carr, D. B. & Pierson, S. M. (1996), ‘Emphasizing Statistical Summaries and Showing Spatial Context with Micromaps’, *Statistical Computing and Statistical Graphics Newsletter* **7**(3), 16–23.
- Carr, D. B., Wallin, J. F. & Carr, D. A. (2000), ‘Two New Templates for Epidemiology Applications: Linked Micromap Plots and Conditioned Choropleth Maps’, *Statistics in Medicine* **19**(17–18), 2521–2538.
- Carr, D. B., Wegman, E. J. & Luo, Q. (1997), ExplorN: Design Considerations Past and Present, Technical Report 137, Center for Computational Statistics, George Mason University, Fairfax, VA.

- Carr, D. B., White, D. & MacEachren, A. M. (2005), ‘Conditioned Choropleth Maps and Hypothesis Generation’, *Annals of the Association of American Geographers* **95**(1), 32–53.
- Carvalho, F. M., Lima, F. & Kriebel, D. (2004), ‘RE: On John Snow’s Unquestioned Long Division’, *American Journal of Epidemiology* **159**(4), 422.
- Centers for Disease Control and Prevention (2010), Cholera. Retrieved February 13, 2011 from <http://www.cdc.gov/cholera/index.html>.
- Chambers, J. M., Cleveland, W. S., Kleiner, B. & Tukey, P. A. (1983), *Graphical Methods for Data Analysis*, Wadsworth & Brooks/Cole, Pacific Grove, CA.
- Chambers, J. M. & Hastie, T. J., eds (1993), *Statistical Models in S*, Chapman & Hall, New York, NY.
- Chapala, G. K. (2005), ‘Development of Rich Features for Web-Based Interactive Micromaps’. Report, Department of Computer Science, Utah State University.
- Chernoff, H. (1973), ‘The Use of Faces to Represent Points in k -dimensional Space Graphically’, *Journal of American Statistical Association* **68**, 361–368.
- Cleveland, W. S. (1985), *The Elements of Graphing Data*, Wadsworth, Monterey, CA.
- Cleveland, W. S. (1993), *Visualizing Data*, Hobart Press, Summit, NJ.
- Cleveland, W. S. (1994), *The Elements of Graphing Data (Revised Edition)*, Hobart Press, Summit, NJ.
- Cook, D. (1997), ‘Calibrate Your Eyes to Recognize High-Dimensional Shapes from Their Low-Dimensional Projections’, *Journal of Statistical Software* **2**(6). <http://www.jstatsoft.org/v02/i06/>.
- Cook, D. & Buja, A. (1997), ‘Manual Controls for High-Dimensional Data Projections’, *Journal of Computational and Graphical Statistics* **6**(4), 464–480.
- Cook, D., Buja, A. & Cabrera, J. (1993), ‘Projection Pursuit Indexes Based on Orthonormal Function Expansions’, *Journal of Computational and Graphical Statistics* **2**(3), 225–250.
- Cook, D., Buja, A., Cabrera, J. & Hurley, C. (1995), ‘Grand Tour and Projection Pursuit’, *Journal of Computational and Graphical Statistics* **4**(3), 155–172.

- Cook, D. & Swayne, D. F. (2007), *Interactive and Dynamic Graphics for Data Analysis — With R and GGobi*, Springer, New York, NY.
- Craig, P., Haslett, J., Unwin, A. & Wills, G. (1989), Moving Statistics — An Extension of “Brushing” for Spatial Data, *in* K. Berk & L. Malone, eds, ‘Proceedings of the 21st Symposium on the Interface between Computing Science and Statistics’, American Statistical Association, Alexandria, VA, pp. 170–174.
- Crawford, S. L. & Fall, T. C. (1990), Projection Pursuit Techniques for Visualizing High-Dimensional Data Sets, *in* G. M. Nielson, B. Shrivvers & L. J. Rosenblum, eds, ‘Proceedings of Visualization in Scientific Computing, Los Alamitos, CA’, IEEE Computer Society Press, pp. 94–108.
- Dent, B. D. (1993), *Cartography: Thematic Map Design (Third Edition)*, William C. Brown, Dubuque, IA.
- DiBiase, D., Reeves, C., MacEachren, A. M., von Wyss, M., Krygier, J. B., Sloan, J. L. & Detweiler, M. C. (1994), Multivariate Display of Geographic Data: Applications in Earth System Science, *in* A. M. MacEachren & D. R. F. Taylor, eds, ‘Visualization in Modern Cartography’, Pergamon (Elsevier), Oxford, U.K., pp. 287–312.
- Dorling, D. (1995), *A New Social Atlas of Great Britain*, John Wiley and Sons, New York, NY.
- Dykes, J. A. (1996), Dynamic Maps for Spatial Science: A Unified Approach to Cartographic Visualization, *in* D. Parker, ed., ‘Innovations in GIS 3’, Taylor & Francis, London, U.K., pp. 177–187.
- Few, S. (2004), *Show Me the Numbers: Designing Tables and Graphs to Enlighten*, Analytics Press, Oakland, CA.
- FisherKeller, M. A., Friedman, J. H. & Tukey, J. W. (1974), PRIM-9: An Interactive Multidimensional Data Display and Analysis System, Technical Report SLAC-PUB-1408, Stanford Linear Accelerator Center, Stanford, CA.
- FitzPatrick, P. J. (1960), ‘Leading British Statisticians of the Nineteenth Century’, *Journal of the American Statistical Association* **55**(289), 38–70.
- Forina, M., Armanino, C., Lanteri, S. & Tiscornia, E. (1983), Classification of Olive Oils from their Fatty Acid Composition, *in* H. Martens & H. Russwurm Jr., eds, ‘Food Research and Data Analysis’, Applied Science Publishers, London, pp. 189–214.

- Freedman, D., Pisani, R. & Purves, R. (2007), *Statistics (Fourth Edition)*, W. W. Norton & Company, New York, NY.
- Friendly, M. (2000a), ‘Re-Visions of Minard’, *Statistical Computing and Statistical Graphics Newsletter* **11**(1), 1 & 13–19.
- Friendly, M. (2000b), *Visualizing Categorical Data*, SAS Publishing, Cary, NC.
- Friendly, M. (2005), Milestones in the History of Data Visualization: A Case Study in Statistical Historiography, in C. Weihs & W. Gaul, eds, ‘Classification: The Ubiquitous Challenge’, Springer, New York, NY, pp. 34–52.
- Friendly, M. (2008), A Brief History of Data Visualization, in C. Chen, W. Härdle & A. Unwin, eds, ‘Handbook of Data Visualization’, Springer, Berlin, Heidelberg, pp. 15–56 & 2 Color Plates.
- Funkhouser, H. G. (1937), ‘Historical Development of the Graphical Representation of Statistical Data’, *Osiris* **3**, 269–404.
- Funkhouser, H. G. & Walker, H. M. (1935), ‘Playfair and his Charts’, *Economic History* **3**, 103–109.
- Furnas, G. W. (1988), Dimensionality Constraints on Projection and Section Views of High Dimensional Loci, in E. J. Wegman, D. T. Gantz & J. J. Miller, eds, ‘Proceedings of the 20th Symposium on the Interface between Computing Science and Statistics’, American Statistical Association, Alexandria, VA, pp. 99–107.
- Furnas, G. W. & Buja, A. (1994), ‘Prosection Views: Dimensional Inference Through Sections and Projections (with Discussion)’, *Journal of Computational and Graphical Statistics* **3**(4), 323–385.
- Gebreab, S. Y., Gillies, R. R., Munger, R. G. & Symanzik, J. (2008), ‘Visualization and Interpretation of Birth Defects Data Using Linked Micromap Plots’, *Birth Defects Research (Part A): Clinical and Molecular Teratology* **82**, 110–119.
- Haining, R., Ma, J. & Wise, S. (1996), ‘Design of a Software System for Interactive Spatial Statistical Analysis Linked to a GIS’, *Computational Statistics: Special Issue on Computeraided Analysis of Spatial Data* **11**(4), 449–466.
- Hankins, T. L. (1999), ‘Blood, Dirt, and Nomograms: A Particular History of Graphs’, *Isis* **90**(1), 50–80.

- Harris, R. L. (1999), *Information Graphics — A Comprehensive Illustrated Reference*, Oxford University Press, New York, NY.
- Harrower, M. A. & Brewer, C. A. (2003), ‘ColorBrewer.org: An Online Tool for Selecting Color Schemes for Maps’, *The Cartographic Journal* **40**(1), 27–37.
- Henry, G. T. (1995), *Graphing Data: Techniques for Display and Analysis*, Sage Publications, Thousand Oaks, CA.
- Heyde, C. C. & Seneta, E., eds (2001), *Statisticians of the Centuries*, Springer, New York, NY.
- Hofmann, H. (2000), ‘Exploring Categorical Data: Interactive Mosaic Plots’, *Metrika* **51**(1), 11–26.
- Hofmann, H. (2003), ‘Constructing and Reading Mosaicplots’, *Computational Statistics & Data Analysis: Special Issue on Data Visualization* **43**(4), 565–580.
- Hofmann, H. (2007), ‘Interview with a Centennial Chart’, *Chance* **20**(2), 26–35.
- Hofmann, H. & Theus, M. (1998), ‘Selection Sequences in MANET’, *Computational Statistics: Special Issue on Strategies for Data Analysis* **13**(1), 77–87.
- Holmes, N. (1991), *Designer’s Guide to Creating Charts & Diagrams (Paperback Edition)*, Watson–Guptill Publications, New York, NY.
- Huff, D. & Geis, I. (1954), *How to Lie with Statistics*, W. W. Norton & Company, New York, NY.
- Hurley, C. (1988), A Demonstration of the Data Viewer, in E. J. Wegman, D. T. Gantz & J. J. Miller, eds, ‘Proceedings of the 20th Symposium on the Interface between Computing Science and Statistics’, American Statistical Association, Alexandria, VA, pp. 108–114.
- Hurley, C. (1989), The Data Viewer: A Program for Graphical Data Analysis, PhD thesis, Statistics Department, University of Washington, Seattle.
- Hurley, C. & Buja, A. (1990), ‘Analyzing High–Dimensional Data with Motion Graphics’, *SIAM Journal on Scientific and Statistical Computing* **11**(6), 1193–1211.
- Inselberg, A. (1985), ‘The Plane with Parallel Coordinates’, *The Visual Computer* **1**, 69–91.

- Jones, G. E. (2000), *How to Lie with Charts*, toExcel Press, Lincoln, NE.
- Klein, R. & Moreira, R. I. (1994), Exploratory Analysis of Agricultural Images via Dynamic Graphics, Technical Report 9/94, Laboratório Nacional de Computação Científica, Rio de Janeiro, Brazil.
- Kleiner, B. & Hartigan, J. A. (1981), ‘Representing Points in Many Dimensions by Trees and Castles (With Discussion)’, *Journal of the American Statistical Association* **76**, 260–276.
- Klinke, S. & Cook, D. (1997), ‘Binning of Kernel-based Projection Pursuit Indices in XGobi’, *Computational Statistics & Data Analysis* **27**(3), 363–369.
- Koschat, M. A. & Swayne, D. F. (1996), ‘Interactive Graphical Methods in the Analysis of Customer Panel Data (with Discussion)’, *Journal of Business and Economic Statistics* **14**(1), 113–132.
- Kosslyn, S. M. (1994), *Elements of Graph Design*, W. H. Freeman and Company, New York, NY.
- Kosslyn, S. M. (2006), *Graph Design for the Eye and Mind*, Oxford University Press, New York, NY.
- Krämer, W. (1991), *So lügt man mit Statistik (3. Auflage)*, Campus Verlag, Frankfurt/Main, Germany.
- Lawrence, M., Wickham, H., Cook, D., Hofmann, H. & Swayne, D. F. (2009), ‘Extending the GGobi Pipeline from R: Rapid Prototyping of Interactive Visualizations’, *Computational Statistics* **24**(2), 195–205.
- Leisch, F. (2002), Sweave: Dynamic Generation of Statistical Reports Using Literate Data Analysis, in W. Härdle & B. Rönz, eds, ‘COMPSTAT 2002: Proceedings in Computational Statistics’, Physica-Verlag, Heidelberg, pp. 575–580.
- Leslie, M. (2002), ‘Tools: A Site for Sore Eyes’, *Science* **296**(5567), 435.
- MacDougall, E. B. (1992), ‘Exploratory Analysis, Dynamic Statistical Visualization, and Geographic Information Systems’, *Cartography and Geographic Information Systems* **19**(4), 237–246.

- MacEachren, A. M., Brewer, C. A. & Pickle, L. W. (1995), Mapping Health Statistics: Representing Data Reliability, *in* 'Proceedings of the 17th International Cartographic Conference, Barcelona, Spain, September 3–9, 1995', Institut Cartographic de Catalunya, Barcelona, Spain, pp. 311–319.
- MacEachren, A. M., Brewer, C. A. & Pickle, L. W. (1998), 'Visualizing Georeferenced Data: Representing Reliability of Health Statistics', *Environment and Planning A* **30**(9), 1547–1561.
- MathSoft (1996), *S+GISLink*, MathSoft, Inc., Seattle, WA.
- McDonald, J. A. & Willis, S. (1987), 'Use of the Grand Tour in Remote Sensing', ASA Statistical Graphics Video Library (<http://stat-graphics.org/movies/>).
- Minnotte, M. C. & West, R. W. (1998), The Data Image: A Tool for Exploring High Dimensional Data Sets, *in* '1998 Proceedings of the Section on Statistical Graphics', American Statistical Association, Alexandria, VA, pp. 25–33.
- Monmonier, M. (1988), Geographical Representations in Statistical Graphics: A Conceptual Framework, *in* '1988 Proceedings of the Section on Statistical Graphics', American Statistical Association, Alexandria, VA, pp. 1–10.
- Monmonier, M. (1989), 'Geographic Brushing: Enhancing Exploratory Analysis of the Scatterplot Matrix', *Geographical Analysis* **21**(1), 81–84.
- Monmonier, M. (1993), *Mapping It Out: Expository Cartography for the Humanities and Social Sciences*, University of Chicago Press, Chicago, IL.
- Monmonier, M. (1996), *How to Lie with Maps (Second Edition)*, University of Chicago Press, Chicago, IL.
- Moore, D. S., McCabe, G. P. & Craig, B. A. (2012), *Introduction to the Practice of Statistics (Seventh Edition)*, W. H. Freeman and Company, New York, NY.
- Morphet, W. J. & Symanzik, J. (2010), 'The Circular Dataimage, a Graph for High-Resolution Circular-Spatial Data', *International Journal of Digital Earth* **3**(1), 47–71.
- Murdoch, D. J. (2002), 'Drawing a Scatterplot', *Chance* **13**(3), 53–55.
- Murrell, P. (2006), *R Graphics*, Chapman & Hall/CRC, Boca Raton, FL.

- Olsen, A. R., Carr, D. B., Courbois, J. P. & Pierson, S. M. (1996), Presentation of Data in Linked Attribute and Geographic Space, *in* ‘1996 Abstracts, Joint Statistical Meetings, Chicago, Illinois’, American Statistical Association, Alexandria, VA, p. 271.
- Openshaw, S. & Perrée, T. (1996), User-Centred Intelligent Spatial Analysis of Point Data, *in* D. Parker, ed., ‘Innovations in GIS 3’, Taylor & Francis, London, U.K., pp. 119–134.
- Palmer, S. E. (1999), *Vision Science, Photons to Phenomenology*, The MIT Press, Cambridge, MA.
- Peng, R. D. (2008), ‘A Method for Visualizing Multivariate Time Series Data’, *Journal of Statistical Software* **25**(Code Snippet 1). <http://www.jstatsoft.org/v25/c01/>.
- Pickle, L. W. (2008), ‘Commentary on “Improving Graphic Displays by Controlling Creativity”’, *Chance* **21**(2), 53.
- Pickle, L. W. & Herrmann, D. J. (1999), Cognitive Research for the Design of Statistical Rate Maps, *in* ‘1999 Proceedings of the Section on Survey Research Methods’, American Statistical Association, Alexandria, VA, pp. 186–191. http://www.amstat.org/sections/SRMS/proceedings/papers/1999_029.pdf.
- Pickle, L. W., Mungiole, M., Jones, G. K. & White, A. A. (1996), *Atlas of United States Mortality*, U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics, Hyattsville, MD.
- Playfair, W. (1805), *An Inquiry into the Permanent Causes of the Decline and Fall of Powerful and Wealthy Nations*, Greenland & Norris, London, U.K.
- Playfair, W. (2005), *The Commercial and Political Atlas and Statistical Breviary, Edited and Introduced by Howard Wainer and Ian Spence*, Cambridge University Press, New York, NY.
- Rensink, R. A. (2006), Attention, Consciousness, and Data Display, *in* ‘2006 JSM Proceedings’, American Statistical Association, Alexandria, VA, pp. 2412–2421. (CD).
- Robbins, N. B. (2005), *Creating More Effective Graphs*, Wiley, Hoboken, NJ.
- Robinson, A. H. (1967), ‘The Thematic Maps of Charles Joseph Minard’, *Imago Mundi* **21**, 95–108.

- Robinson, A., Sale, R. & Morrison, J. (1978), *Elements of Cartography (Fourth Edition)*, John Wiley and Sons, New York, NY.
- Rosenbaum, A. S., Axelrad, D. A., Woodruff, T. J., Wei, Y.-H., Ligoeki, M. P. & Cohen, J. P. (1999), 'National Estimates of Outdoor Air Toxics Concentrations', *Journal of the Air and Waste Management Association* **49**, 1138–1152.
- Rosling, H. & Johansson, C. (2009), 'Gapminder: Liberating the x-axis from the Burden of Time', *Statistical Computing and Statistical Graphics Newsletter* **20**(1), 4–7.
- Scott, D. W. (1985), 'Average Shifted Histograms: Effective Non-Parametric Density Estimation in Several Dimensions', *Annals of Statistics* **13**, 1024–1040.
- Scott, D. W. (1992), *Multivariate Density Estimation: Theory, Practice, and Visualization*, Wiley, New York, NY.
- Scott, L. M. (1994), 'Identification of a GIS Attribute Error Using Exploratory Data Analysis', *The Professional Geographer* **46**(3), 378–386.
- Simons, D. J. & Rensink, R. A. (2005), 'Change Blindness: Past, Present, and Future', *Trends in Cognitive Sciences* **9**(1), 16–20.
- Snow, J. (1936), *Snow on Cholera: Being a Reprint of Two Papers by John Snow, M.D. Together with a Biographical Memoir by B. W. Richardson, M.D. and an Introduction by Wade Hampton Frost, M.D.*, The Commonwealth Fund & Oxford University Press, New York, NY & London, U.K.
- Spence, I. (2004), Playfair, William (1759–1823), in H. C. G. Matthew & B. Harrison, eds, 'Oxford Dictionary of National Biography', Oxford University Press, Oxford, U.K. <http://www.oxforddnb.com/view/article/22370> (accessed 13 Aug 2009).
- Spence, I. (2006), William Playfair and the Psychology of Graphs, in '2006 JSM Proceedings', American Statistical Association, Alexandria, VA, pp. 2426–2436. (CD).
- Stuetzle, W. (1988), Plot Windows, in W. S. Cleveland & M. E. McGill, eds, 'Dynamic Graphics for Statistics', Wadsworth & Brooks/Cole, Belmont, CA, pp. 225–245.
- Swayne, D. F. & Buja, A. (1998), 'Missing Data in Interactive High-Dimensional Data Visualization', *Computational Statistics: Special Issue on Strategies for Data Analysis* **13**(1), 15–26.

- Swayne, D. F. & Cook, D. (1992), ‘Xgobi: A Dynamic Graphics Program Implemented in X With a Link to S’, *Computing Science and Statistics* **22**, 544–547.
- Swayne, D. F., Cook, D. & Buja, A. (1991), XGobi: Interactive Dynamic Graphics in the X Window System with a Link to S, *in* ‘1991 Proceedings of the Section on Statistical Graphics’, American Statistical Association, Alexandria, VA, pp. 1–8.
- Swayne, D. F., Cook, D. & Buja, A. (1998), ‘XGobi: Interactive Dynamic Graphics in the X Window System’, *Journal of Computational and Graphical Statistics* **7**(1), 113–130.
- Swayne, D. F., Temple Lang, D., Buja, A. & Cook, D. (2003), ‘GGobi: Evolving from XGobi into an Extensible Framework for Interactive Data Visualization’, *Computational Statistics & Data Analysis: Special Issue on Data Visualization* **43**(4), 423–444.
- Symanzik, J. (2004), Interactive and Dynamic Graphics, *in* J. E. Gentle, W. Härdle & Y. Mori, eds, ‘Handbook of Computational Statistics — Concepts and Methods’, Springer, Berlin, Heidelberg, pp. 293–336.
- Symanzik, J. (2008), ‘Interview with Andreas Buja’, *Computational Statistics* **23**(2), 177–184.
- Symanzik, J., Axelrad, D. A., Carr, D. B., Wang, J., Wong, D. & Woodruff, T. J. (1999), HAPs, Micromaps and GPL — Visualization of Geographically Referenced Statistical Summaries on the World Wide Web, *in* ‘Annual Proceedings (ACSM–WFPS–PLSO–LSAW 1999 Conference CD)’, American Congress on Surveying and Mapping.
- Symanzik, J. & Carr, D. B. (2008), Interactive Linked Micromap Plots for the Display of Geographically Referenced Statistical Data, *in* C. Chen, W. Härdle & A. Unwin, eds, ‘Handbook of Data Visualization’, Springer, Berlin, Heidelberg, pp. 267–294 & 2 Color Plates.
- Symanzik, J., Carr, D. B., Axelrad, D. A., Wang, J., Wong, D. & Woodruff, T. J. (1999), Interactive Tables and Maps — A Glance at EPA’s Cumulative Exposure Project Web Page, *in* ‘1999 Proceedings of the Section on Statistical Graphics’, American Statistical Association, Alexandria, VA, pp. 94–99.
- Symanzik, J., Cook, D., Lewin-Koh, N., Majure, J. J. & Megretskaia, I. (2000), ‘Linking ArcView and XGobi: Insight Behind the Front End’, *Journal of Computational and Graphical Statistics* **9**(3), 470–490.

- Symanzik, J., Fischetti, W. & Spence, I. (2009), ‘Editorial: Commemorating William Playfair’s 250th Birthday’, *Computational Statistics* **24**(4), 551–566.
- Symanzik, J., Gebreab, S., Gillies, R. & Wilson, J. (2003), Visualizing the Spread of West Nile Virus, *in* ‘2003 Proceedings’, American Statistical Association, Alexandria, VA. (CD).
- Symanzik, J., Wong, D., Wang, J., Carr, D. B., Woodruff, T. J. & Axelrad, D. A. (2000), Web-based Access and Visualization of Hazardous Air Pollutants, *in* ‘Geographic Information Systems in Public Health: Proceedings of the Third National Conference August 18–20, 1998, San Diego, California’, Agency for Toxic Substances and Disease Registry. <http://www.atsdr.cdc.gov/GIS/conference98/>.
- Tappin, L. (1994), ‘Analyzing Data Relating to the Challenger Disaster’, *The Mathematics Teacher* **87**(6), 423–426.
- Theus, M. (2002), ‘Interactive Data Visualization Using Mondrian’, *Journal of Statistical Software* **7**(11). <http://www.jstatsoft.org/v07/i11/>.
- Theus, M. (2003), ‘Abstract: Interactive Data Visualization Using Mondrian’, *Journal of Computational and Graphical Statistics* **12**(1), 243–244.
- Theus, M., Hofmann, H. & Wilhelm, A. F. X. (1998), ‘Selection Sequences — Interactive Analysis of Massive Data Sets’, *Computing Science and Statistics* **29**(1), 439–444.
- Theus, M. & Urbanek, S. (2004), ‘iPlots : Interactive Graphics for R’, *Statistical Computing and Statistical Graphics Newsletter* **15**(1), 11–14.
- Theus, M. & Urbanek, S. (2009), *Interactive Graphics for Data Analysis: Principles and Examples*, Chapman & Hall/CRC, Boca Raton, FL.
- Tufte, E. R. (1974), *Data Analysis for Politics and Policy*, Prentice–Hall, Englewood Cliffs, NJ.
- Tufte, E. R. (1983), *The Visual Display of Quantitative Information*, Graphics Press, Cheshire, CT.
- Tufte, E. R. (1990), *Envisioning Information*, Graphics Press, Cheshire, CT.
- Tufte, E. R. (1997), *Visual Explanations: Images and Quantities, Evidence and Narrative*, Graphics Press, Cheshire, CT.

- Tufte, E. R. (2006), *Beautiful Evidence*, Graphics Press, Cheshire, CT.
- Tukey, J. W. (1977), *Exploratory Data Analysis*, Addison Wesley, Reading, MA.
- Tweedie, L. & Spence, R. (1998), ‘The Prosection Matrix: A Tool to Support the Interactive Exploration of Statistical Models and Data’, *Computational Statistics: Special Issue on Strategies for Data Analysis* **13**(1), 65–76.
- Unwin, A. (1994), REGARDing Geographic Data, in P. Dirschedl & R. Ostermann, eds, ‘Computational Statistics’, Physica-Verlag, Heidelberg, pp. 315–326.
- Unwin, A. (1999), ‘Requirements for Interactive Graphics Software for Exploratory Data Analysis’, *Computational Statistics: Special Issue on Interactive Graphical Data Analysis* **14**(1), 7–22.
- Unwin, A. (2002), ‘Scatterplotting’, *Chance* **15**(2), 39–42.
- Unwin, A., Hawkins, G., Hofmann, H. & Siegl, B. (1996), ‘Interactive Graphics for Data Sets with Missing Values — MANET’, *Journal of Computational and Graphical Statistics* **5**(2), 113–122.
- Unwin, A. & Wills, G. (1988), Eyeballing Time Series, in ‘1988 Proceedings of the Section on Statistical Computing’, American Statistical Association, Alexandria, VA, pp. 263–268.
- Unwin, A., Wills, G. & Haslett, J. (1990), REGARD — Graphical Analysis of Regional Data, in ‘1990 Proceedings of the Section on Statistical Graphics’, American Statistical Association, Alexandria, VA, pp. 36–41.
- Vachon, D. (2005), ‘Doctor John Snow Blames Water Pollution for Cholera Epidemic’, *Old News* **16**(8), 8–10.
- Venables, W. N. & Ripley, B. D. (2002), *Modern Applied Statistics with S (Fourth Edition)*, Springer, New York, NY.
- Wainer, H. (1993), ‘Tabular Presentation’, *Chance* **6**(3), 52–56.
- Wainer, H. (1997), *Visual Revelations: Graphical Tales of Fate and Deception from Napoleon Bonaparte to Ross Perot*, Copernicus/Springer, New York, NY.
- Wainer, H. (2002), ‘The BK-Plot: Making Simpson’s Paradox Clear to the Masses’, *Chance* **15**(3), 60–62.

- Wainer, H. (2005), *Graphic Discovery: A Trout in the Milk and Other Visual Adventures*, Princeton University Press, Princeton, NJ.
- Wainer, H. (2007), ‘Improving Data Displays: Ours and the Media’s’, *Chance* **20**(3), 8–15.
- Wainer, H. (2008), ‘Improving Graphic Displays by Controlling Creativity’, *Chance* **21**(2), 46–52.
- Wainer, H. (2009a), ‘A Centenary Celebration for Will Burtin: A Pioneer of Scientific Visualization’, *Chance* **22**(1), 51–55.
- Wainer, H. (2009b), ‘A Good Table Can Beat a Bad Graph: It Matters Who Plays Mozart’, *Chance* **22**(4), 55–57.
- Wainer, H. & Francolini, C. M. (1980), ‘An Empirical Inquiry Concerning Human Understanding of Two–Variable Color Maps’, *The American Statistician* **34**(2), 81–93.
- Wallgren, A., Wallgren, B., Persson, R., Jorner, U. & Haaland, J.-A. (1996), *Graphing Statistics & Data: Creating Better Charts*, Sage Publications, Thousand Oaks, CA.
- Wang, X., Chen, J. X., Carr, D. B., Bell, B. S. & Pickle, L. W. (2002), ‘Geographic Statistics Visualization: Web–based Linked Micromap Plots’, *Computing in Science & Engineering* **4**(3), 90–94.
- Weber, J. S. (2008), Small Sample Histogram Possibilities and Paradoxes, in ‘2008 JSM Proceedings’, American Statistical Association, Alexandria, VA. (CD).
- Wegman, E. J. (1990), ‘Hyperdimensional Data Analysis Using Parallel Coordinates’, *Journal of the American Statistical Association* **85**, 664–675.
- Wegman, E. J. (1992), ‘The Grand Tour in k–Dimensions’, *Computing Science and Statistics* **22**, 127–136.
- Wegman, E. J. & Dorfman, A. (2003), ‘Visualizing Cereal World’, *Computational Statistics & Data Analysis: Special Issue on Data Visualization* **43**(4), 633–649.
- Wegman, E. J. & Shen, J. (1993), ‘Three–Dimensional Andrews Plots and the Grand Tour’, *Computing Science and Statistics* **25**, 284–288.

- Wickham, H., Cook, D., Hofmann, H. & Buja, A. (2011), ‘tourr: An R Package for Exploring Multivariate Data with Projections’, *Journal of Statistical Software* **40**(2). <http://www.jstatsoft.org/v40/i02/>.
- Wickham, H., Lawrence, M., Temple Lang, D. & Swayne, D. F. (2008), ‘An Introduction to rggobi’, *R News* **8**(2), 3–7. http://CRAN.R-project.org/doc/Rnews/Rnews_2008-2.pdf.
- Wilhelm, A. F. X., Unwin, A. & Theus, M. (1996), Software for Interactive Statistical Graphics — A Review, *in* F. Faulbaum & W. Bandilla, eds, ‘SoftStat ’95 — Advances in Statistical Software 5’, Lucius & Lucius, Stuttgart, pp. 3–12.
- Wilhelm, A. F. X., Wegman, E. J. & Symanzik, J. (1999), ‘Visual Clustering and Classification: The Oronsay Particle Size Data Set Revisited’, *Computational Statistics: Special Issue on Interactive Graphical Data Analysis* **14**(1), 109–146.
- Williams, I., Limp, W. F. & Briuer, F. L. (1990), Using Geographic Information Systems and Exploratory Data Analysis for Archaeological Site Classification and Analysis, *in* K. M. S. Allen, S. W. Green & E. B. W. Zubrow, eds, ‘Interpreting Space: GIS and Archaeology’, Taylor & Francis, London, U.K., pp. 239–273.
- Xie, Y. & Cheng, X. (2008), ‘animation: A Package for Statistical Animations’, *R News* **8**(2), 23–27. http://CRAN.R-project.org/doc/Rnews/Rnews_2008-2.pdf.
- Yoshioka, K. (2002), ‘KyPlot — A User-Oriented Tool for Statistical Data Analysis and Visualization’, *Computational Statistics* **17**(3), 425–437.
- Zeileis, A., Hornik, K. & Murrell, P. (2009), ‘Escaping RGBland: Selecting Colors for Statistical Graphics’, *Computational Statistics & Data Analysis* **53**(9), 3259–3270.
- Zelazny, G. (2001), *Say it with Charts: The Executive’s Guide to Visual Communication (Fourth Edition)*, McGraw-Hill, New York, NY.
- Zhang, Z. & Griffith, D. A. (1997), ‘Developing User-Friendly Spatial Statistical Analysis Modules for GIS: An Example using ArcView’, *Computers, Environment and Urban Systems* **21**(1), 5–29.

— THE END —

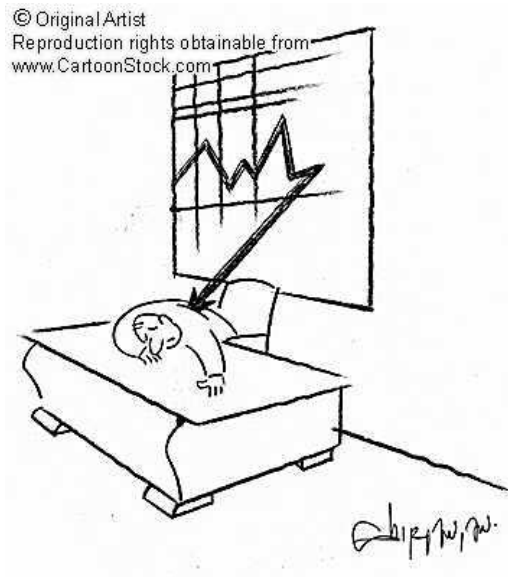


Figure 136: http://www.cartoonstock.com/blowup_stock.asp?imageref=vsh0184&artist=Shirvanian,+Vahan&topic=statistics+, Cartoon.