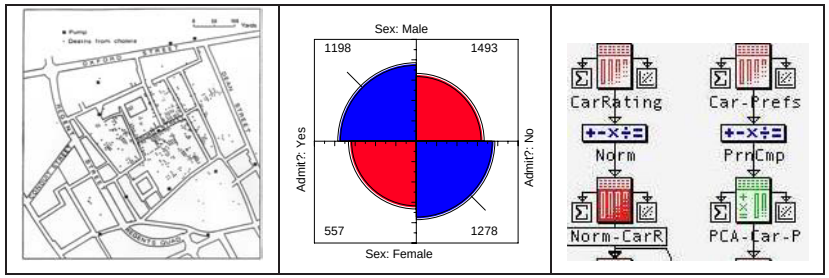
The Past, Present and Future of Statistical Graphics

(An Ideo-Graphic and Idiosyncratic View)



Michael Friendly

York University

http://www.math.yorku.ca/SCS/friendly.html

VIEWS, London, Nov, 2004

Outline

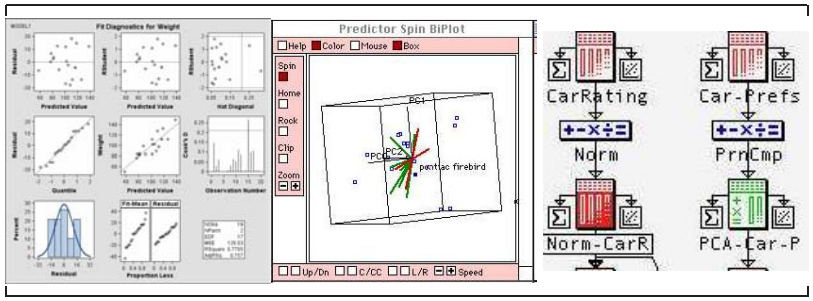
- SAS graphics: The power to grow?
  - Different strokes: graphics *user* vs. *developer*
  - Current models: SAS “Solutions,” graphic PROCs, SAS/IML
  - ODS Graphics
- Statistical graphics: Models for growth?
  - Minard’s lessons for statistical graphics
  - JMP— Model summary = Graphs + Numbers + · · ·
  - ViSta— Dynamic, interactive graphics (spreadplots, workmaps)
  - Innovation and Graphical excellence
- Wider visions
  - Visions from the Forrest
  - Visions for graphic users and developers
- Conclusions

Part 4: Visions of the Future

Prediction is very difficult, especially about the future — Niels Bohr

The best way to predict the future is to invent it — Alan Kay

- SAS graphics: The power to grow?
- Statistical graphics: Models for growth?
- Wider visions
- Conclusions



Different strokes: Business user, Analyst, Developer

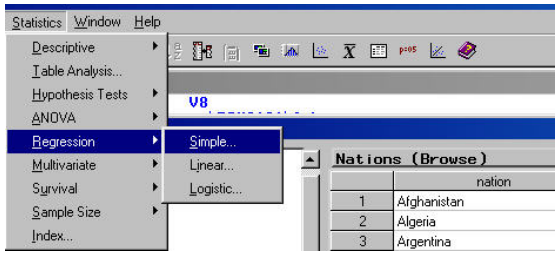
- **Different graphs for different tasks and goals**
  - Presentation graphs
    - Summarize, simplify, record information
    - Persuade, highlight a message
  - Analysis graphs
    - Reconnaisance: overview of large, complex datasets
    - Expose: detect patterns, trends, anomalies
    - Model diagnosis: departures from assumptions, corrective actions
- **Business user**
  - Small number of standard graph types
  - Ease of understanding & communication
  - Ease of producing them
- **Analyst, Statistician**
  - Wider range of graphs, tailored to analysis
  - Some need/want control of graphic styles, rendering details
- **Graphics developer**
  - Freedom to invent new methods of visualization with ease
  - Elegant connections between statistical analysis (summarization) and visualization (exposure)

SAS Graphics: The power to grow?

Current models

- SAS/AF "Solutions" (e.g., Analyst, Market Research, . . .)
  - + Menu-driven interface to a wide variety of SAS procedures
  - + Plots (somewhat) integrated with analysis steps
  - + Some provide the SAS code → save, edit, re-submit
  - − Separate applications, often inconsistent, no coherent structure
  - − AF interface unappealing, awkward, often not well-designed
  - − Options, controls limited by what the developer thought would suffice
  - ⇒ need a top-down, not bottom-up design

e.g., Analyst

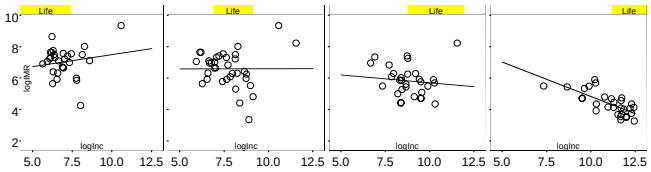


SAS Graphics: The power to grow?

- SAS PROCs (e.g., GPLOT, GCHART, . . .)
  - + Statistical procs provide analytics, ODS → stats
  - + Annotate provides all graphics primitives to customize displays
  - − Statistical context divorced from graphical context
  - − Multi-panel displays are difficult— no provision for *overall* scaling, axes, etc.

  e.g., Coplot (Trellis display) of $\log(IMR) \sim \log(\text{Income}) \mid \text{Life Exp.}$

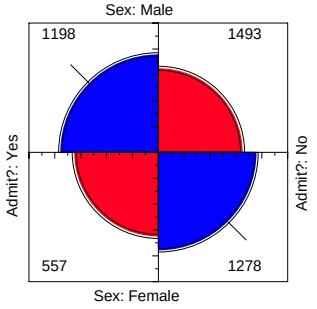
```
%coplot(data=nations,
    x=logInc, y=logIMR, given=Life,
    interp=rl, slices=4, rows=1);
```

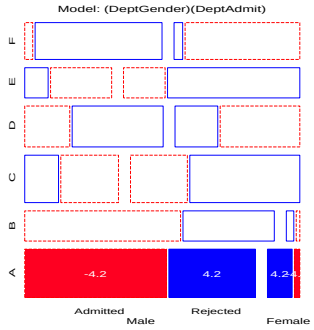


  - Macros ease the pain— new graphical methods, enhance available graphics
    - Utilities: labels, errbars, lines, ...
    - Apps: mosaic, lowess, inflglim, ...

SAS Graphics: The power to grow?

- PROC IML graphics— My favorite environment for new graphics



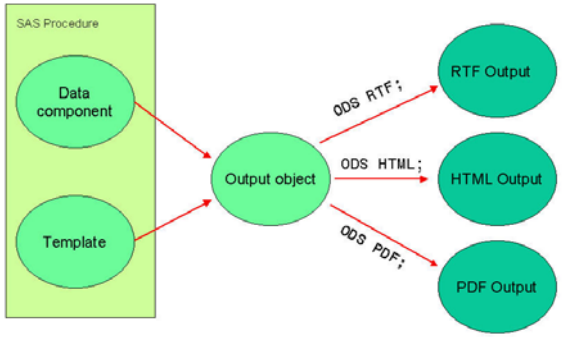


  - + All graphics primitives: `gpoint()`, `gdraw()`, `gpoly()`, ...
  - + Powerful matrix and statistical functions
  - + User-defined modules act like primitives (mostly)
  - − Embarassing gaps in communication with SAS data sets
    - no missing data,
    - no formatted values (1 ='Low' 2='Med' 3='High'),

ODS Graphics

- **SAS 8.2**: Introduces the Output Delivery System (ODS)
  - All SAS procedures produce (nested) output objects
  - Output objects can be rendered in a variety of formats (listing, HTML, RTF, LaTeX)
  - Output can be customized via templates

ODS Output Objects



Example:

open RTF output →

```
ods rtf file='odsex1.rtf';
ods select factorANOVA;

data odor;
   input Odor Temperature GasLiquidRatio PackingHeight @@;
datalines;
66 -1 -1  0   39 1 -1  0   43 -1  1 0   49 1 1 0
58 -1  0 -1   17 1  0 -1   -5 -1  0 1  -40 1 0 1
65  0 -1 -1    7 0  1 -1   43  0 -1 1  -22 0 1 1
-31 0  0  0  -35 0  0  0  -26  0  0 0
run;
proc rsreg;
   model odor=Temperature GasLiquidRatio PackingHeight;
run;

ods rtf close;
```

close RTF output →

Produces:

Factor	DF	Sum of Squares	Mean Square	F Value	Pr > F
Temperature	4	5258.016026	1314.504006	2.60	0.1613
GasLiquidRatio	4	11045	2761.150641	5.46	0.0454
PackingHeight	4	3813.016026	953.254006	1.89	0.2510

ODS Graphics

- **SAS 9.1**: Introduces ODS Graphics
  - SAS/STAT procedures modified to produce *some* graphs internally (à la PROC REG)
  - Output graphs can be rendered in a variety of *formats* (HTML, RTF, LaTeX), and in a variety of *styles* (Analysis, Journal, Statistical)
  - Graphs can be customized via templates

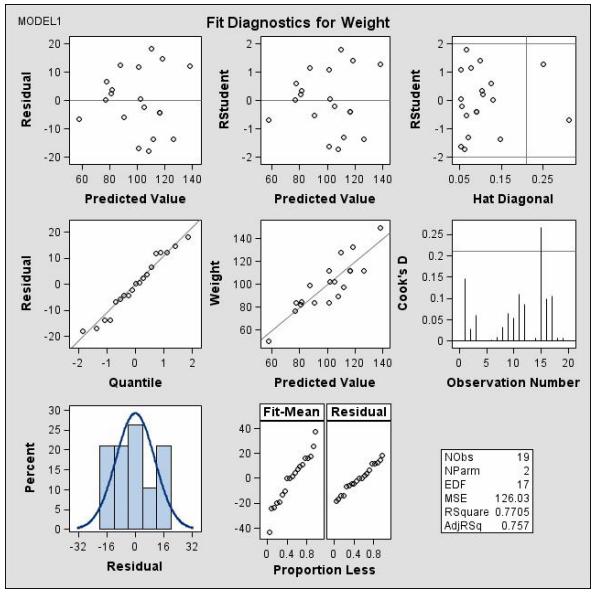
Example:

```
ods html style=Default;      /* or, style=Journal */
ods graphics on;

proc reg data = sashelp.class;
   model Weight = Height;
run;
quit;

ods graphics off;
ods html close;
```


Output:



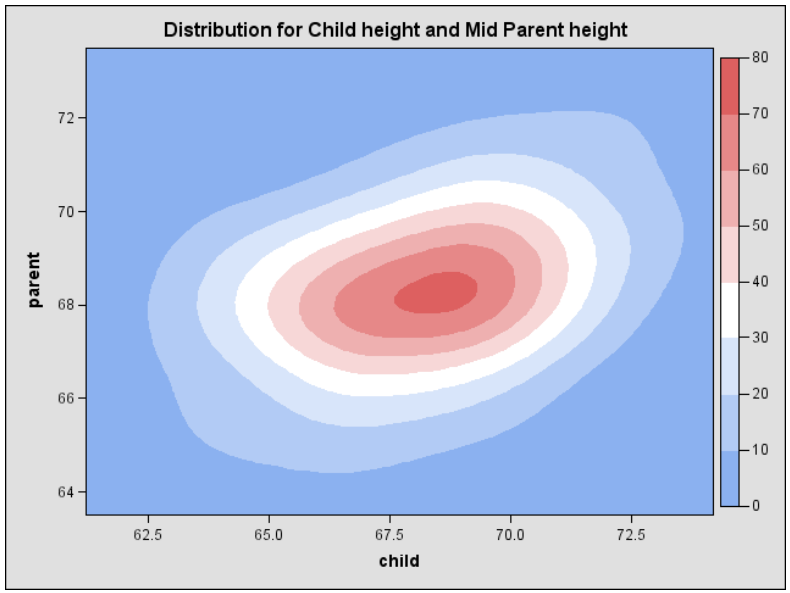
ODS Graphics: Galton's data

The contour plots of Galton's data can now be produced more easily using ODS Graphics:

```
ods html file = "galton.html";
ods graphics on;

proc kde data=galton;
  bivar child (bwm=1.5) parent (bwm=1.5) /
    ngrid  = 80
    levels = 2.5 5 10 20 25 33 40 50  60 68 75 80 90 95 97.5
    plots  = contour contourscatter;
  freq frequency;
run;

ods graphics off;
ods html close;
```

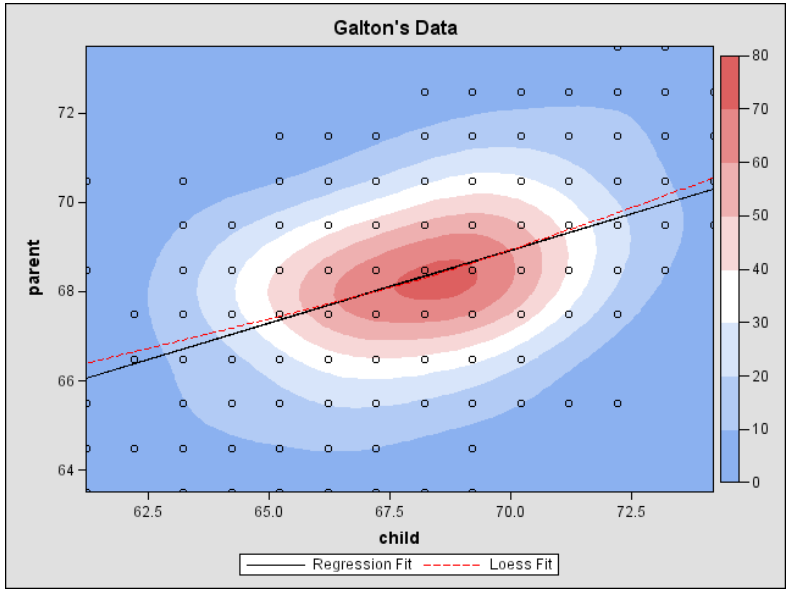


ODS Graphics

Procedures supporting ODS Graphics (SAS 9.2):

- Base SAS
  - CORR
- SAS/ETS
  - ARIMA
  - AUTOREG
  - ENTROPY
  - EXPAND
  - MODEL
  - SYSLIN
  - TIMESERIES
  - UCM
  - VARMAX
  - X12
- High-Performance Forecasting
  - HPF
- SAS/STAT
  - ANOVA
  - CORRESP
  - GAM
  - GENMOD
  - GLM
  - KDE
  - LIFETEST
  - LOESS
  - LOGISTIC
  - MI
  - MIXED
  - PHREG
  - PRINCOMP
  - PRINQUAL
  - REG
  - ROBUSTREG

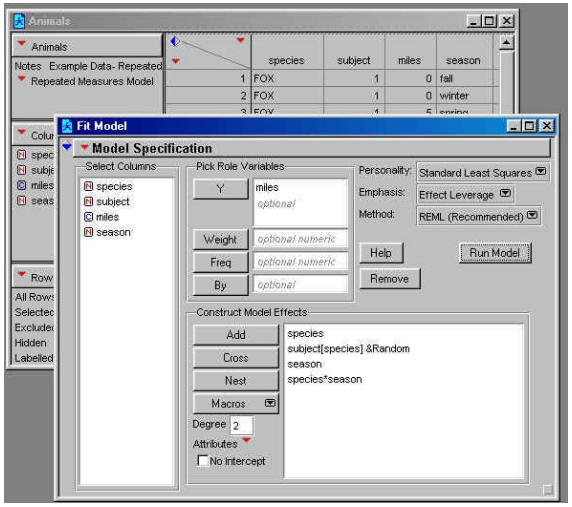
Customizing via templates:

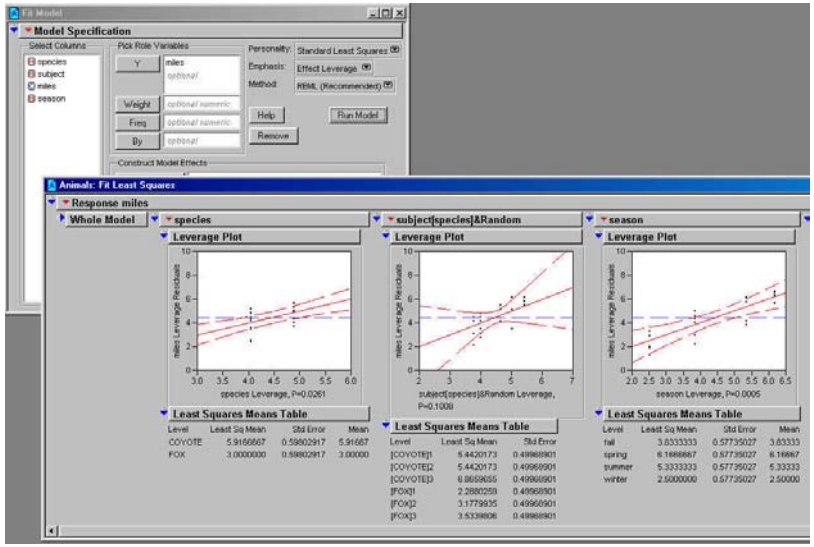


JMP— Model summary = Graphs + Numbers + · · ·

- **No need to beg for graphs**
  - Every analysis ⟶ graphs + tables
  - Graphs for different “personalitites,” and “emphasis”
  - All graphs have associated menus to control (some) details and options
- **I like menus, but I need to do this again, and again ...**
  - All menu/dialog actions can be saved to a script
  - Scripts can be generalized to be used with any similar dataset
- **Interactive graphics: linking and brushing**
  - All views of a data table are linked— selecting observations in one view ⟶ selected in all other views
  - Selected observations can be hidden, excluded, colored, labeled, ...

JMP— Model summary = graphs + numbers

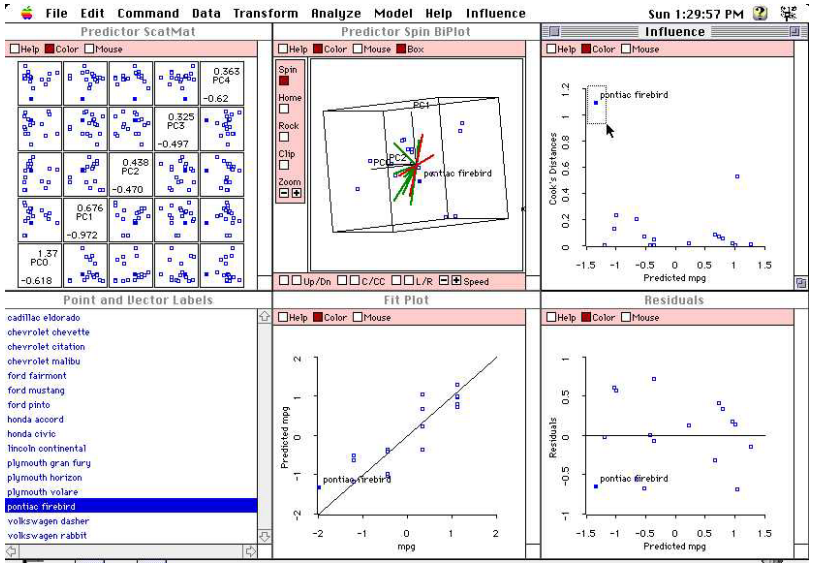




ViSta— spreadplots, work maps

- Spreadplots
  - Graphic equivalent of a spreadsheet
  - Dynamically linked views of *data* and *model* objects
  - Highly interactive: every action → data, model, plots
  - (Message passing architecture)
- e.g., Spreadplot for multiple regression
  - Scatterplot matrix— overview
  - 3D spin predictor biplot— leverage, collinearity
  - Influence plot, fit plot, residual plot— influential cases
  - Observation, variable labels, interactive brushing, etc.

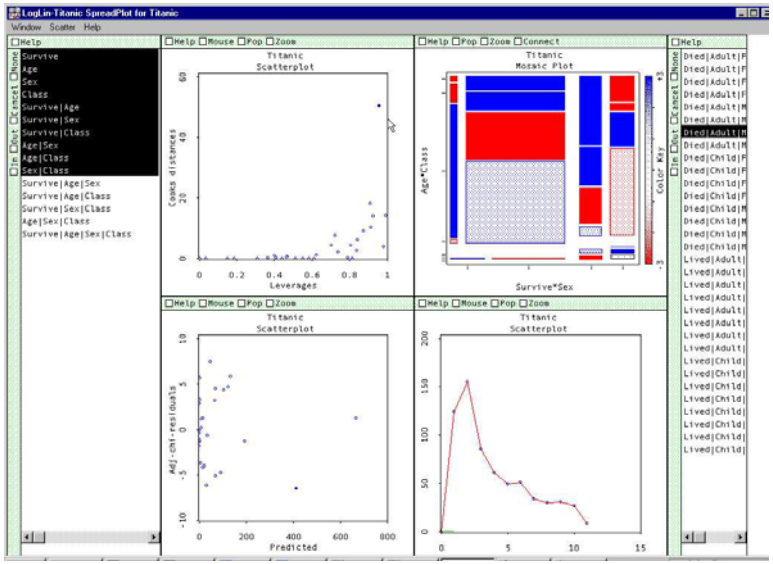
See: `http://forrest.psych.unc.edu/research/`



ViSta— Categorical data

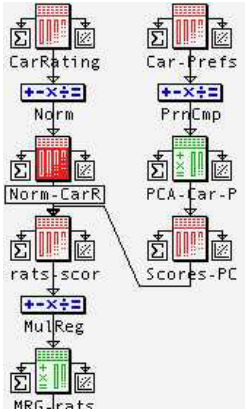
- Visual model fitting— select terms
- Mosaic display for current model
- Influence plot: Cook's D vs. Leverage (Hat values)
- Model summary graph: Deviance vs. df
- All dynamically linked, manipulable!

See: Valero et al. (2003),
`http://www.math.yorku.ca/SCS/Papers/viscat.pdf`



ViSta— Workmaps

- Workmap— visual GUI for path(s) of analysis
- Each item: dynamic links to table-view, numerical summary, spreadplot visualization



ViSta— Expandability

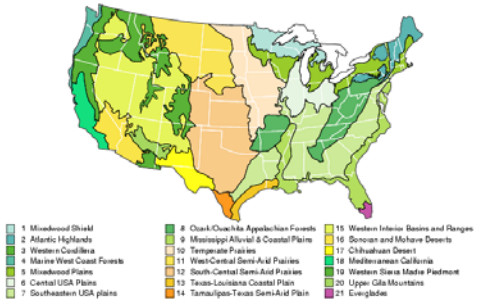
- Other features:
  - Plugins — add new analysis and visualizations
  - Web Applets, Scripts
  - Data analysis language

See: `http://forrest.psych.unc.edu/research/`

Innovation and Graphical Excellence

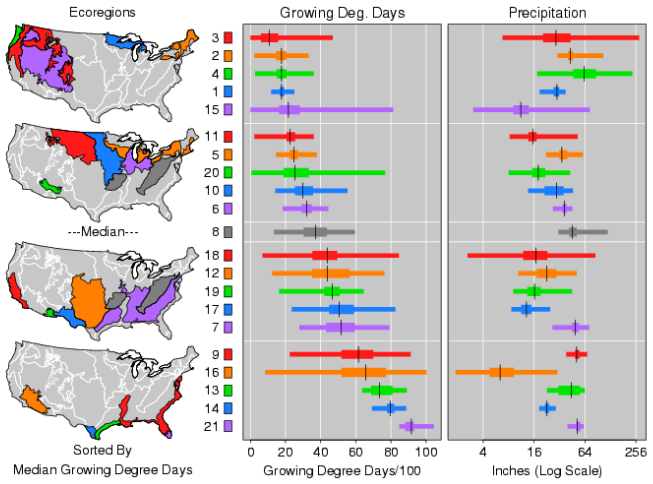
e.g., Dan Carr (Carr et al., 1998)

- Omernick ecoregions - 21 ecological distinctive areas



- Problems:
  - Linking regions with labels is difficult
  - Hard to use distinct colors
  - How to show spatial variation of analysis variables?

- → Linked micromaps
  - Boxplots of growing degree days & precipitation
  - Effect ordering: sorted by median growing degree days
  - Color linking is clear; attention to detail exemplary

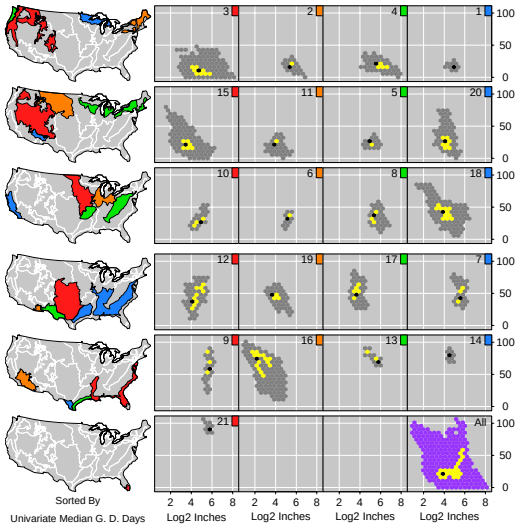


Innovation and Graphical Excellence

- Relationship of growing days and precipitation hard to see in univariate views.
- Bivariate density estimation (481K grid cells)
- Bivariate boxplots (50% high-density region, bivariate median)
- Sorted by median growing degree days

Figure 2: LM Bivariate Boxplots

1961-1990 Precipitation (x) versus Growing Degree Days/100 (y)



Visions from the Forrest

The Statistician's 3D Virtual-Reality Workroom

- A 3D, VR statistical analysis environment:
  - Data sources, data streams, data views
  - Tools (and a glove?) for manipulating data
  - Analysis and visualization devices
  - An amenuensis— virtual assistant
- Data sources, data streams, data views
  - Visual, manipulable building blocks (Stat/Graph "Lego")
  - Snap together to form statistical objects (tables, datasets)
  - Spigots for incoming streams, trapdoors to the data mine, hoses, valves, connectors...
  - Lassos and windows for data views
- Tools for manipulating data *directly*:
  - transformations (sliders on power transforms)
  - subset, merge, join, ... (slice, drag/drop)
  - $\longrightarrow$ new data objects, views, ...

Visions from the Forrest

The Statistician's 3D Virtual-Reality Workroom

- Analysis and visualization devices
  - Data toasters: data $\longrightarrow$ toast (model summary) $+$ crumbs (residuals)— all plug 'n play
  - Data/Model/Residual VCR's, with controls: pop in the data, out comes a visualization.
  - Recepticles for making new connections, plugging in new appliances
  - Hand-held devices— controls to interact with transformations, models, summaries, residuals, ...
  - Workmaps to show you where you've been, Guidemaps to show you where you might want to go
- An amenuensis— virtual assistant
  - take notes,
  - offer guidance,
  - suggest visualizations,
  - summarize results,
  - write results section,
  - serve virtual coffee, ...

The Future for Graphics Users

- Statistical procedures extensively developed— will continue
  - regression $\longrightarrow$ GLM $\longrightarrow$ GENMOD $\longrightarrow \{$MIXED, GAM$\}$
  - PCA $\longrightarrow$ FA $\longrightarrow \{$Lisrel, SEM (`PROC CALIS`)$\}$
- Simplify the environment— for most users, but allow for growth
  - 80–20 rule: 80% of a graph takes 20% of effort. The last 20% is hard work.
  - $\Rightarrow$ provide the 80% by default, no need to beg
  - $\Rightarrow$ provide the tools to customize, extend, combine, annotate, ...
- Statistical graphics is on the right track when ...
  - it allows you to picture what your data have to say
  - the picture is faithful to some (possibly complex) model
  - the picture leverages the perceptual and cognitve capabilities of the viewer.

The Future for Graphics Developers

- Statistical graphics now well-developed, but many different systems— mostly incompatible, different capabilities
  - SAS $\longrightarrow$ macros, ODS Graphics, SAS/INSIGHT, ...
  - R/S-Plus $\longrightarrow$ general `plot()` methods, packages, connections to interactive graphics (`ggobi`)
- Need to provide paths of growth for new visualizations, methods of interaction, ...
- 80–20 rule: 80% of software development takes 20% of effort. The last 20% is hard work.
- Statistical graphics is on the right track when ...
  - it allows one to develop a new method of visualization or interaction with ease
  - it provides elegant connections between statistical analysis (summarization) and visualization (exposure)
  - it leverages the capabilities of different software systems

Conclusions

- The past history of statistical graphics teaches us that:
  - All modern methods have deep roots, and lessons for today:
  - Statistical graphics can have both *beauty* and *truth*
  - Graphics always had a purpose— tell a story, inform a decision, ...
  - We can often better understand these intellectual accomplishments by re-tracing their steps
- The present history of statistical graphics teaches us that:
  - We need graphical methods for categorical data on a par with those for quantitative data.
  - Users— Different strokes for different folks:
    - Most want *graphical toasters*: data in, picture out (but, what picture?)
    - Some want/need complete control of graphic styles, rendering details
    - Graphic developers want it all: freedom to invent!

... Conclusions

- The future of statistical graphics?
  - Statistical graphics is on the right track when ...
    - it allows one to construct a pretty picture of data,
    - the picture is faithful to some (possibly complex) model,
    - the picture leverages the perceptual and cognitve capabilities of the viewer.
  - Statistical graphics is on the right track when ...
    - it moves the 80–20 rule in favor of the user/developer,
    - it nurtures future growth of tools, techniques $\longrightarrow$ insight,
    - it allows for *beauty* as well as *truth*.

References

Bickel, P. J., Hammel, J. W., and O'Connell, J. W. Sex bias in graduate admissions: Data from Berkeley. *Science*, 187:398–403, 1975.

Carr, D., Olsen, A. R., Pierson, S. M., and Courbois, J.-Y. Boxplot variations in a spatial context: An Omernik ecoregion and weather example. *Statistical Computing & Statistical Graphics Newsletter*, 9(2):4–13, 1998.

Cleveland, W. S. *Visualizing Data*. Hobart Press, Summit, NJ, 1993.

Fox, J. Effect displays for generalized linear models. In Clogg, C. C., editor, *Sociological Methodology, 1987*, pp. 347–361. Jossey-Bass, San Francisco, 1987.

Friendly, M. *SAS System for Statistical Graphics*. SAS Institute, Cary, NC, 1st edition, 1991.

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

Friendly, M. Conceptual and visual models for categorical data. *The American Statistician*, 49: 153–160, 1995.

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

Friendly, M. Corrgrams: Exploratory displays for correlation matrices. *The American Statistician*, 56(4):316–324, 2002.

Friendly, M. Milestones in the history of data visualization: A case study in statistical historiography. In Gaul, W. and Weihs, C., editors, *Studies in Classification, Data Analysis, and Knowledge Organization*. Springer, New York, 2004. (In press).

Friendly, M. and Denis, D. The early origins and development of the scatterplot. *Journal of the History of the Behavioral Sciences*, 2004. (In press, accepted 7/09/04).

Friendly, M. and Kwan, E. Effect ordering for data displays. *Computational Statistics and Data Analysis*, 43(4):509–539, 2003.

Hartigan, J. A. and Kleiner, B. Mosaics for contingency tables. In Eddy, W. F., editor, *Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface*, pp. 268–273. Springer-Verlag, New York, NY, 1981.

Tufte, E. R. *Visual Explanations: Images and Quantities, Evidence and Narrative*. Graphics Press, Cheshire, CT, 1997.

Tukey, J. W. *Exploratory Data Analysis*. Addison Wesley, Reading, MA, 1977.

Valero, P., Young, F., and Friendly, M. Visual categorical analysis in ViSta. *Computational Statistics and Data Analysis*, 43(4):495–508, 2003.