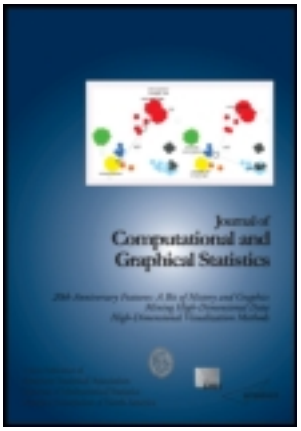

Journal of Computational and Graphical Statistics

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/ucgs20

The Generalized Pairs Plot

John W. Emerson [a], Walton A. Green [b], Barret Schloerke [c], Jason Crowley [c], Dianne Cook [c], Heike Hofmann [c] & Hadley Wickham [d]

[a] Department of Statistics, Yale University, New Haven, CT, 06520

[b] Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA, 02138

[c] Department of Statistics, Iowa State University, Ames, IA, 50011

[d] Department of Statistics, Rice University, Houston, TX, 77251

Accepted author version posted online: 22 May 2012.Version of record first published: 27 Mar 2013.

To cite this article: John W. Emerson , Walton A. Green , Barret Schloerke , Jason Crowley , Dianne Cook , Heike Hofmann & Hadley Wickham (2013): The Generalized Pairs Plot, Journal of Computational and Graphical Statistics, 22:1, 79-91

To link to this article: http://dx.doi.org/10.1080/10618600.2012.694762

Supplementary materials for this article are available online.
Please go to www.tandfonline.com/r/JCGS

The Generalized Pairs Plot

John W. EMERSON, Walton A. GREEN, Barret SCHLOERKE,
Jason CROWLEY, Dianne COOK, Heike HOFMANN, and Hadley WICKHAM

This article develops a generalization of the *scatterplot matrix* based on the recognition that most datasets include both categorical and quantitative information. Traditional grids of scatterplots often obscure important features of the data when one or more variables are categorical but coded as numerical. The *generalized pairs plot* offers a range of displays of paired combinations of categorical and quantitative variables. A mosaic plot, fluctuation diagram, or faceted bar chart may be used to display two categorical variables. A side-by-side boxplot, stripplot, faceted histogram, or density plot helps visualize a categorical and a quantitative variable. A traditional scatterplot is suitable for displaying a pair of numerical variables, but options also support density contours or annotating summary statistics such as the correlation and number of missing values, for example. By combining these, the generalized pairs plot may help to reveal structure in multivariate data that otherwise might go unnoticed in the process of exploratory data analysis. Two different R packages provide implementations of the generalized pairs plot, `gpairs` and `GGally`. Supplementary materials for this article are available online on the journal web site.

Key Words: Exploratory data analysis; Grammar of graphics; Graphics; Multivariate data; Scatterplot matrix; Visualization.

1. INTRODUCTION

This article contributes to the development of the *pairs plot*, which first appeared in the article by Hartigan (1975). It is also referred to as the *generalized draftsman's display* by Tukey and Tukey (1981) and Chambers et al. (1983), and as the *scatterplot matrix* (SPLOM) by Cleveland (1993) and Basford and Tukey (1999). The pairs plot is a grid of scatterplots showing the bivariate relationships between all pairs of variables in a multivariate dataset. Although the authors of this article (and many other academics and data analysts) regularly use this graphical display, it is not clear how widely it is used in practice. Our informal

John W. Emerson is at the Department of Statistics, Yale University, New Haven, CT 06520 (E-mail: *john.emerson@yale.edu*). Walton A. Green is at the Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138. Barret Schloerke is at the Department of Statistics, Iowa State University, Ames, IA 50011. Jason Crowley is at the Department of Statistics, Iowa State University, Ames, IA 50011. Dianne Cook is at the Department of Statistics, Iowa State University, Ames, IA 50011. Heike Hofmann is at the Department of Statistics, Iowa State University, Ames, IA 50011. Hadley Wickham is at the Department of Statistics, Rice University, Houston, TX 77251.

Journal of Computational and Graphical Statistics, Volume 22, Number 1, Pages 79–91
DOI: 10.1080/10618600.2012.694762

survey of several statistics texts that include multiple regression revealed inconsistent use of pairs plots.

Most datasets consist of both quantitative and categorical variables. When all variables of interest are quantitative, the scatterplot matrix is a natural tool for graphical exploration. Friendly (1994) proposed an alternative based on the mosaic plot (Hartigan and Kleiner 1984) for displaying pairwise relationships among a set of categorical variables. Emerson, Green, and Hartigan (2006) presented the first *generalized pairs plot*, addressing the need for a more flexible display of a mixture of quantitative and categorical variables. Though our use of "generalized" is in contrast with its usage by Chambers et al. (1983), the name seems most appropriate and we recommend it be adopted for this display.

Section 2 presents the basic design of the generalized pairs plot. Sections 3 and 4 then discuss two implementations available in extension packages for the R language and environment for statistical computing (R Development Core Team 2012): `gpairs` (Emerson and Green 2012b) and `GGally` (Schloerke et al. 2012). The former approach was a methodological development for exploratory data analysis (EDA). The latter presents an implementation for the same graphical exploratory purposes, but develops these plots as a contribution to the framework of Wilkinson's grammar of graphics (Wilkinson 1999b) as implemented by Wickham (2009). Both packages are built using R's grid graphics system (Murrell 2005), but each will likely appeal to different segments of the community. Section 5 concludes with a discussion. Supplementary materials available online include datasets presented in this article, the commands used to produce each of the displays, additional examples, and performance benchmarks.

2. THE GENERALIZED PAIRS PLOT

The generalized pairs plot should not be confused with the generalized draftsman's display by Chambers et al. (1983); we regard the latter as a traditional pairs plot or scatterplot matrix of quantitative information. Figure 1 shows an example of a scatterplot matrix of Fisher's iris data (Fisher 1936), originally collected by Anderson (1935). Here, the species is treated numerically (1 for *Iris setosa*, 2 for *I. versicolor*, and 3 for *I. virginica*). This plot could be improved by using color to identify the species instead of explicitly including the numerical representation of species as a quantitative variable. Doing so uncovers striking clusterings of petal and sepal measurements by species, shown in Figure S1 in the supplementary materials available online.

When a dataset includes one or more categorical variables, the traditional display offers limited flexibility. Friendly (1994) proposed a grid of mosaic tiles for displaying sets of entirely categorical variables. Our generalization takes this a step further, recognizing the need for different types of panels that together display a wider range of features in a collection of continuous and categorical variables. There are three general types of displays. A display (or tile, or panel) containing a graphic or other summary information corresponding to two quantitative variables is called *quantitative–quantitative* display. A panel for two categorical variables is called *categorical–categorical*. The last type corresponds to one categorical and one quantitative variable, called a *quantitative–categorical* panel.

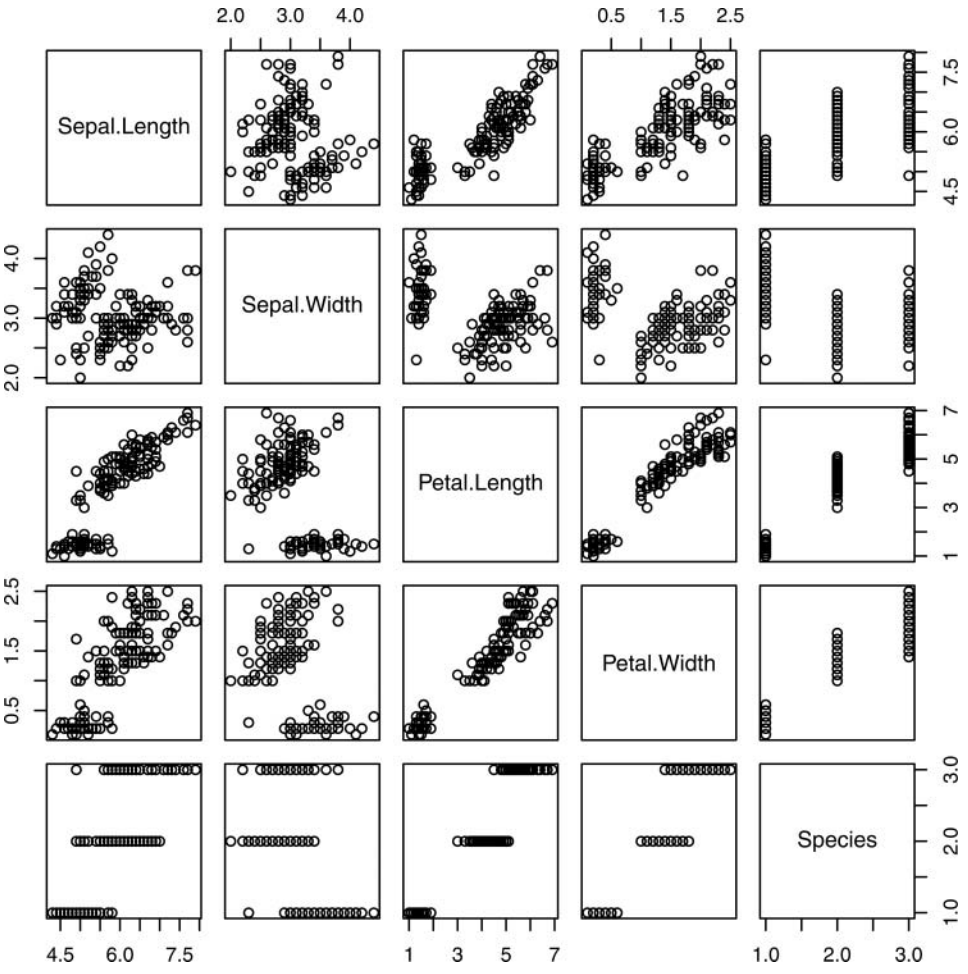

Figure 1. A traditional pairs plot of Fisher's iris data. All variables except `Species` are quantitative. All pairs of variables are plotted as scatterplots, both above and below the diagonal. Clustering can be seen in several plots, and a strong positive association can be seen between petal length and width.

Scatterplots are naturally used in quantitative–quantitative panels, but various options or alternatives include displaying density contours, information on correlation, missing values, or linear or nonlinear fits. Mosaic plots (Hartigan and Kleiner 1984) provide a graphical display of counts in a contingency table for two categorical variables where areas are proportional to counts. A categorical–categorical display may be used to emphasize either the joint distribution or one of the conditional distributions. Finally, the association between a categorical and a quantitative variable may be depicted using a box-and-whisker plot (Tukey 1977) or some variation thereof showing the conditional distribution.

Figure 2 shows a generalized pairs plot of a dataset containing measurements taken on dining parties in a restaurant by a single waiter (Bryant and Smith 1995). Variables include total bill ($), tip ($), gender of the bill payer, day of the week, and the tip as a percentage of the total bill. As with scatterplot matrices, candidate "dependent" variables (when applicable)

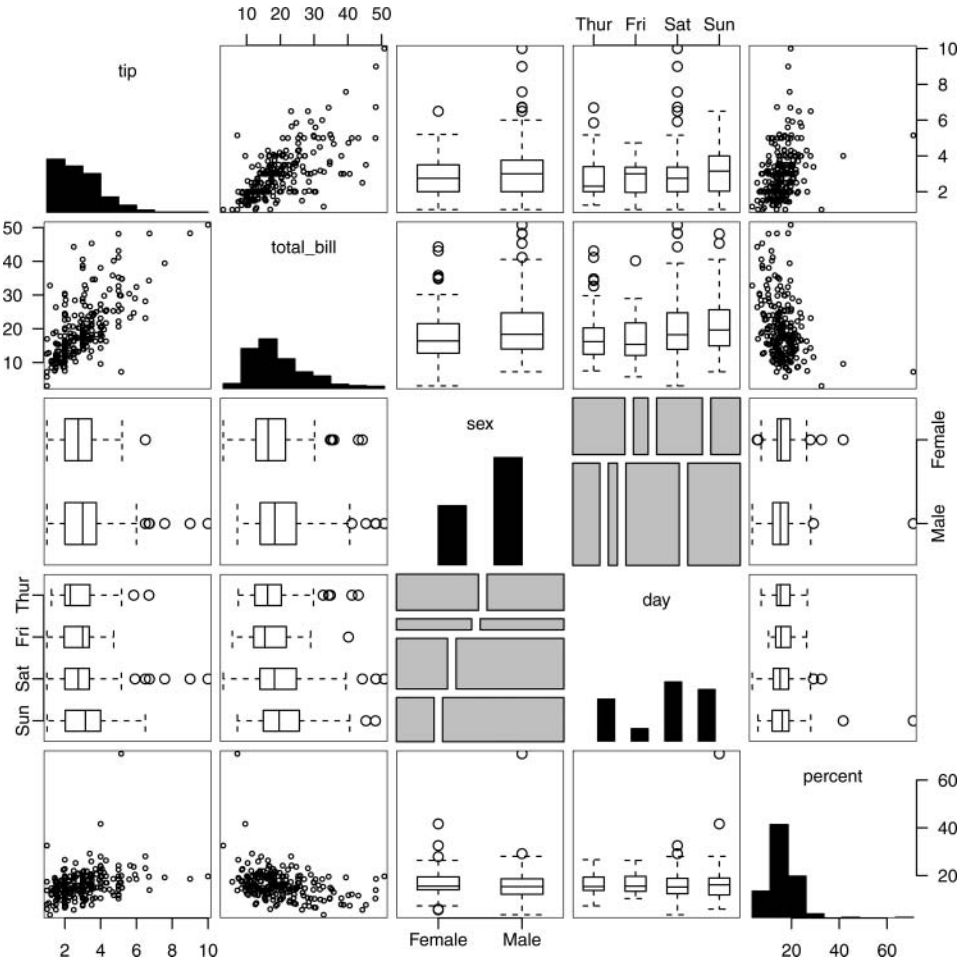

Figure 2. A first example of the generalized pairs plot. The dataset contains a mixture of quantitative and categorical variables that are reflected in the types of plots displayed: scatterplots for quantitative–quantitative; side-by-side boxplots for quantitative–categorical; and mosaic plots for categorical–categorical.

are usually placed in the upper-left positions. For quantitative–quantitative and quantitative–categorical panels, the information in the upper and lower diagonals of this particular plot is redundant. However, the mosaic tiles between `sex` and `day` show both of the conditional distributions; the tile in row 3, column 4 gives the distribution of `day` conditional on `sex`, for example. Histograms and bar charts on the diagonal reflect the marginal distributions of the variables. Total bill size and tip are positively associated (as shown by the scatterplots), but not as strongly as one might expect because there is increasing variability in tip as bill increases. Both `tip` and `total_bill` have skewed distributions (evident in the histograms), which might lead the analyst to consider log-transforming these variables. Males spend more on average than females and bills are higher on the weekend (shown in the side-by-side boxplots). The 70% tip on a very small bill by a male on a Sunday may be an outlier. Much can be learned about tipping behavior by studying this first example of a generalized pairs plot.

Table 1. A summary of a subset of the 2010 Environmental Performance Index data using the `whatis` function of R extension package `YaleToolkit` (Emerson and Green 2012c)

Variable	Type	Missing	Unique	Precision	Min	Max
`Country`	Character	0	231	NA	AFG	ZWE
`EPI`	Numeric	68	163	1e-08	32.12	93.48
`Landlock`	Pure factor	0	2	NA	No	Yes
`HighPopDens`	Pure factor	0	2	NA	No	Yes
`ENVHEALTH`	Numeric	49	173	1e-08	0.06	95.09
`ECOSYSTEM`	Numeric	68	163	1e-08	0.06	95.09

3. EXPLORATORY DATA ANALYSIS

Our development of the generalized pairs plot follows in the EDA tradition by John Tukey. At the most basic level, every exploration should begin by asking what is (in) a dataset. In most datasets, the answer includes a description of the contents of "rows" (cases, observations, subjects, . . .) and "columns" (variables, characteristics, measurements, . . .) as typically arranged in a table or spreadsheet. Are there missing values or obvious data entry errors? Where do they occur? Are there both quantitative and categorical variables? Simple descriptions often reveal important features and surprises that may demand attention prior to further analyses.

A summary such as that shown in Table 1 is a good starting point; these data are from the 2010 Environmental Performance Index (Emerson et al. 2010). Each of 231 countries from around the globe is classified as being landlocked (`LandLock`, having no direct access to an ocean) or not, and as having a high population density (`HighPopDens`) or not. Indices reflect overall environmental performance (`EPI`) as well as performance on two subcategories, environmental health (`ENVHEALTH`) and ecosystem vitality (`ECOSYSTEM`). The indices can range from 0 to 100, but no country achieves these extremes. The subcategory indices of environmental health and ecosystem vitality were scaled to share the same range. Missing values impede construction of the indices for many of the countries.

EDA typically begins with tabulation of categorical variables and univariate summaries such as histograms for quantitative variables. Bivariate associations are often explored with scatterplots and side-by-side boxplots, as appropriate, with two-way tables and mosaic plots used for pairs of categorical variables. For example, the boxplot shown in Figure 3 provides a standard graphical exploration of the bivariate association between a categorical variable (landlocked status, in this case) and a continuous variable (the environmental health index). A pair of stacked histograms would also show that the environmental health index is lower on average for the landlocked countries. However, both methods of display are based on data reduction that can obscure information in the conditional distributions.

An alternative quantitative–categorical display that maintains the full data resolution is the *barcode plot* (Emerson, Green, and Hartigan 2006). The barcode plot was originally developed by Hartigan in the spirit of the rug and stripplot (see Chambers and Hastie 1992, for example) and named because of its similarity to the Universal Product Code (UPC) on commercial packaging. Figure 4, produced using the `barcode` function of R extension package `barcode` (Emerson and Green 2012a), shows the barcode plot for the same data

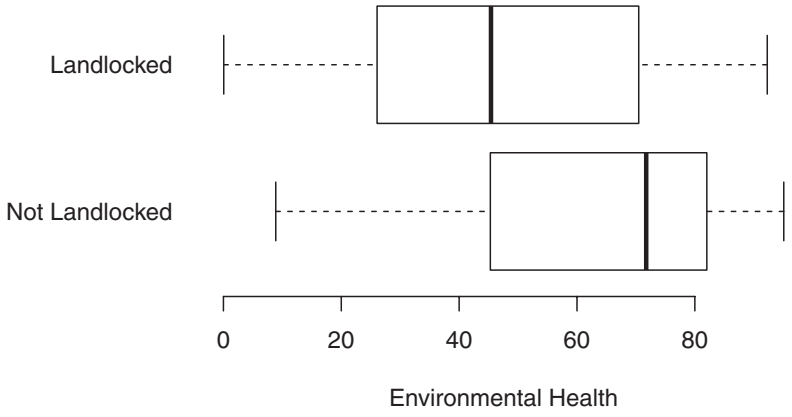

Figure 3. The association between environmental health and landlocked status in the 2010 Environmental Performance data is explored using a side-by-side boxplot. The Environmental Health Index is lower on average for the landlocked countries.

displayed in Figure 3. A single stroke represents each data value, like dots in a dotplot (Tukey and Tukey 1990). The slim stroke helps alleviate overplotting in dense regions. If ties are present, histogram-like stacked segments depict the cardinality of the ties, with one represented by the primary stroke and the remainder building the stacked segments. The ties represented by the tall spike in the bottom right of Figure 4 reveal an interesting aspect of the data not evident in the boxplot and obscured by a regular histogram. Germany, Finland, France, Luxembourg, Norway, and New Zealand have identical values of the environmental health index (of these, only Luxembourg is landlocked), so the tall spike is of height $5 - 1 = 4$ above the initial stroke. In addition, several other pairs of countries were tied with similarly high values of environmental health, indicated by the single strokes surrounding the tall spike for nonlandlocked countries.

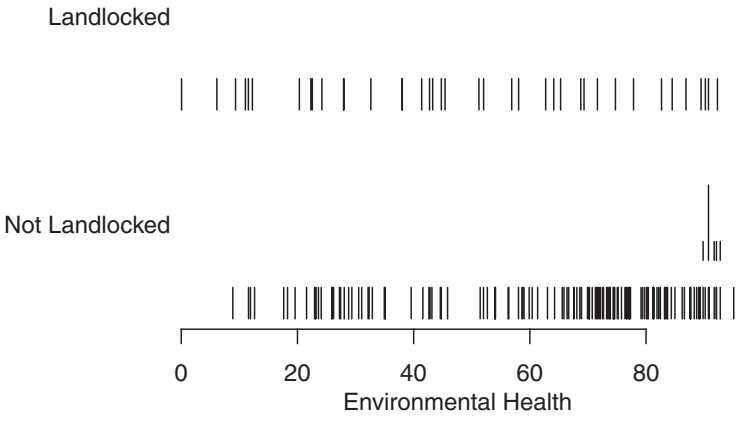

Figure 4. A barcode as an alternative to the side-by-side boxplot shown in Figure 3. If k cases are identical, $k - 1$ smaller strokes above the primary stroke denote the ties. Here, the tallest spike consists of four strokes above the primary stroke, where five countries share the same level of environmental health.

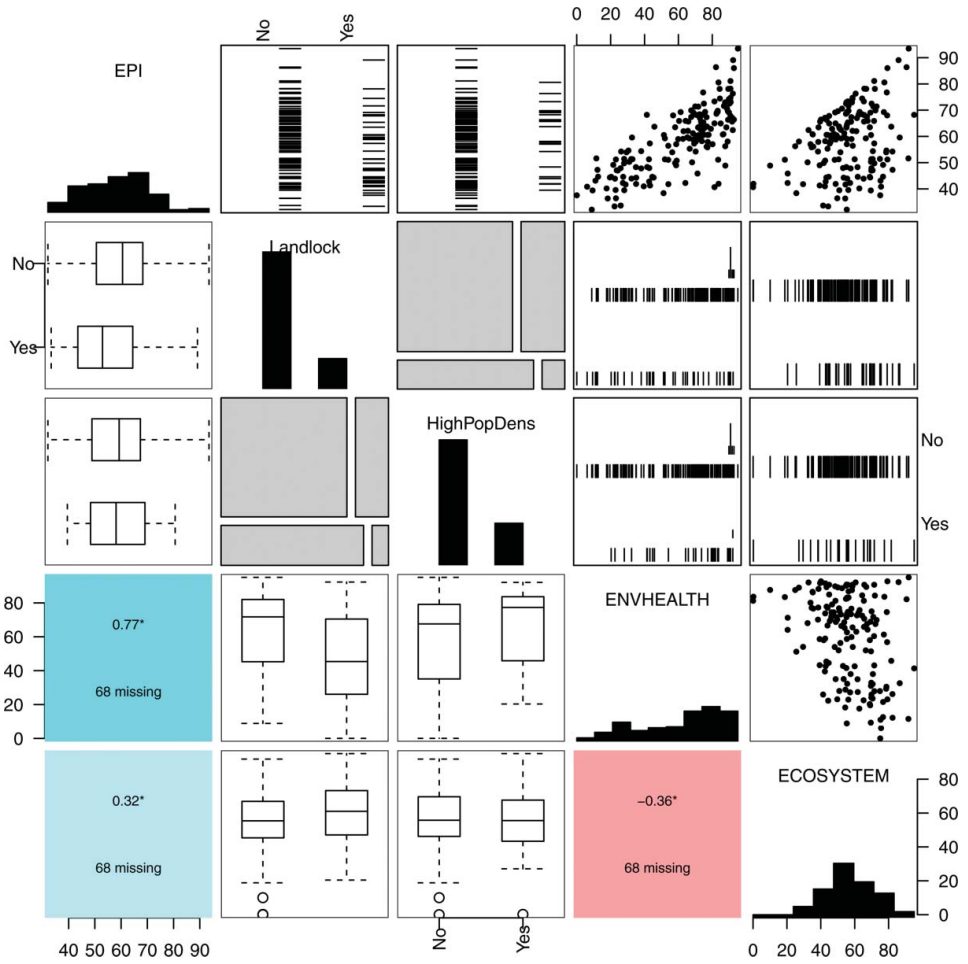

Figure 5. Generalized pairs plot of five variables in the 2010 Environmental Performance Index data. Choices of arguments ensure that different plots are used in the upper and lower triangle. Quantitative–quantitative pairs are shown as scatterplots and summarized by the correlation. Quantitative–categorical pairs are displayed as side-by-side boxplots and barcode plots, and the one categorical–categorical pair of plots uses mosaic tiles with a different conditioning variable above and below the diagonal.

The generalized pairs plot can combine scatterplots, mosaic plots, and the detailed barcode plots with the higher-level summary of traditional boxplots. Figure 5 displays selected variables from the 2010 Environmental Performance Index, showing some options provided by the `gpairs` function of R extension package `gpairs` (Emerson and Green 2012b). Scatterplots are displayed above the diagonal for pairs of quantitative variables. Below the diagonal, text in the cells shows the correlations and numbers of pairwise missing values. Statistical significance of the correlation at the 5% level is indicated by an asterisk, though caution must be exercised if such a test is not justified (as is the case here). Color shading and saturation (red for negative, blue for positive, as shown in Figure S5 in the supplementary materials available online) visually reinforces the nature of the linear associations between these variables. Both `ENVHEALTH` and `ECOSYSTEM` are positively

associated with the overall environmental performance index (EPI) by construction. The difference in their correlations with the EPI, 0.77 and 0.32, respectively, is indicative of a known weakness of the 2010 EPI: it suffered from an imbalance in the influence of the two policy objectives on the EPI. This problem was addressed in the subsequent 2012 EPI. Finally, we note that the negative association between ENVHEALTH and ECOSYSTEM reveals an interesting facet of environmental performance: wealthier countries enjoy better access to health care and score better on environmental health, whereas their protection of the ecosystem is far less predictable and often worse than for poorer, less-developed countries.

Mosaic tiles in Figure 5 display the two different conditional distributions for the categorical variables; Landlock conditional on HighPopDens is shown below the diagonal, with HighPopDens conditional on Landlock appearing above the diagonal. These illustrate that countries with higher population densities are somewhat less likely to be landlocked. Finally, the boxplots and barcode panels show the quantitative–categorical variable associations. These illustrate that countries that are not landlocked have generally higher EPI and health values and lower ecosystem values, for example.

Other plotting options are supported by the gpairs function. Stripplots may be used in place of boxplots or barcode plots. Points may be customized in scatterplot panels using alternative symbols, sizes, and colors for the exploration of high-dimensional patterns. A companion function, corrgram, is also provided by package gpairs (see Friendly (2002) for a nice discussion of these plots).

4. AN EXTENSION OF THE GRAMMAR OF GRAPHICS

The generalized pairs plot is also well suited for the the grammar of graphics ideas first described by (Wilkinson 1999b) and recently realized in the package ggplot2 (Wickham 2009). The grammar of graphics defines a language for describing graphical displays. The language is designed to reveal common elements among disparate plot types and provides an efficient way to describe a new plot.

Wickham's interpretation of the grammar of graphics treats the scatterplot matrix as a faceted plot. Faceting involves partitioning data and displaying the resulting subsets in separate plots. Originally, this technique was designed for studying conditional distributions such as the scatterplots of X versus Y conditional on a categorical variable W. Faceting is provided by trellis plots (Becker, Cleveland, and Shyu 1996) and lattice plots (Sarkar 2008). Making a scatterplot matrix using faceting requires a little sleight of hand, because a scatterplot matrix is a plot of the joint rather than conditional distributions. The data need to be expanded into a long form with four columns, the first two containing the variable names and the other two with the data values for the horizontally and vertically displayed variables. Faceting is then applied to the first two columns of variables names, yielding each pair of scatterplots. This approach, taken by the function plotmatrix in ggplot2, is too limited for the generalized pairs plot because it does not adapt to a mixture of variable types.

Instead, it is advantageous to consider the generalized pairs plot as a type of layout of multiple different plots—call the complete layout a *composite* plot. The scatterplot matrix is then a special case, where all of the plots are uniformly scatterplots. This is the approach

adopted by `ggpairs` in the package `GGally` (Schloerke et al. 2012). Other types of multiple layout plots are in common use. For example, JMP's (SAS Institute 2010) default display of univariate distributions shows a boxplot stacked above a histogram, and for bivariate distributions JMP makes it easy to display histograms along the margins of the scatterplot. Multiple time series are often displayed in a vertical layout, with different variables plotted against time in separate plots. Side-by-side boxplots, parallel coordinate plots, and the slug plot (Grosjean, Spirlet and Jangoux 2003)—used for displaying quantiles overlaid on side-by-side histograms—might also be considered to be composite plots.

Composite plots allow the user to be creative in each panel of the matrix. Categorical–categorical panels can display mosaic plots, faceted bar charts, or fluctuation diagrams. When one variable is categorical and the other quantitative, side-by-side boxplots, faceted histograms, or density plots can be used. The grammar of graphics can be used to define the plot for each cell. In this way, `ggpairs` is effectively a wrapper to `ggplot2`'s primary plotting methods, building upon its language for defining plots and allowing the user to develop a complex display of selected pairs of variables in the data.

Figure 6 shows an example of a generalized pairs plot created with `ggpairs`. The data comes from the latest National Research Council report on 61 statistics graduate research programs in the United States (National Research Council 2010). Table 2 summarizes the variables selected for the plot. Two different types of rankings are shown: the 5th percentiles of so-called "R" and "S" rankings. `Time.to.Grad` measures the average number of years students take to graduate from the program. `Workspace` is a binary variable indicating whether all students get some private space in which to work in the department. Finally, `Prizes.Awards` is categorical with four levels reflecting the opportunities for the graduate students to receive awards.

In the example, scatterplots are used for quantitative–quantitative panels below the diagonal, and correlations are displayed in corresponding panels above the diagonal. Quantitative–categorical panels use side-by-side boxplots and faceted density plots. Categorical–categorical panels use faceted bar charts. In the spirit of EDA described in Section 3, we can observe several things about the program rankings. Although the correlation between the two ranking systems is moderately positive, the ranking methods frequently disagree. For one program, the S method provides a rank of 10 while the R method ranks the program 45th. Time to graduate has no apparent relationship to either program rank. The boxplots show that highly rated programs (i.e., programs with lower ranks) often provide all students with workspace and have more award opportunities. However, it is also evident that very few programs have limited workspace or fail to offer award opportunities. The density plots corroborate the observations made using the boxplots. For

Table 2. A summary of a subset of the 2010 National Research Council rankings of statistics graduate programs

Variable name	Type	Num unique	Precision	Min	Max
R.5th	Numeric	39	1.00	1	56
S.5th	Numeric	34	1.00	1	61
Time.to.Grad	Numeric	29	0.01	3.5	7
Workspace	Factor	2	NA	<100%	100%
Prizes.Awards	Factor	4	NA	Both	Prog

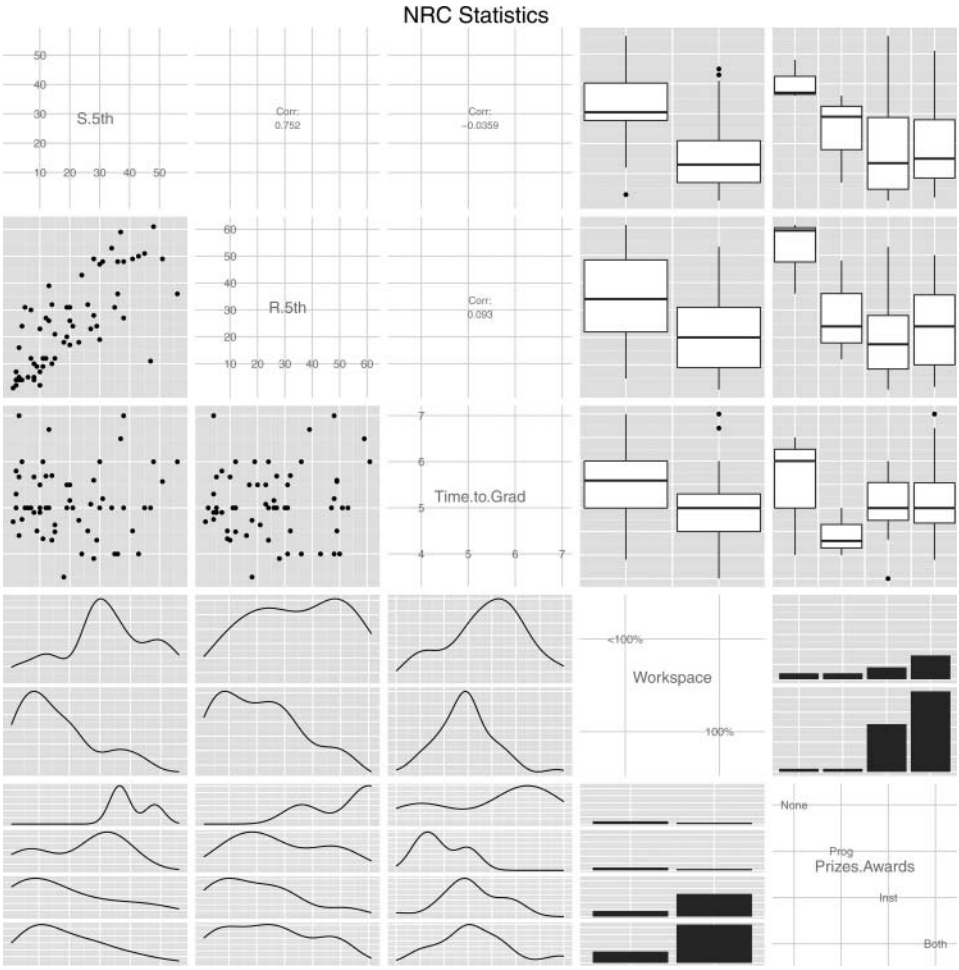

Figure 6. National Research Council rankings of statistics graduate programs. Five variables are plotted: S and R 5th percentile rankings, time to graduate, workspace provided to students, and types of prizes and awards available.

example, it can be seen that students tend to finish sooner in programs that give all students workspace.

The `ggpairs` software leverages a modular design. Each cell contains a plot that is described by a single character string. The dataset is stored separately from the plot definition, and the plot is produced only when the string is evaluated with the corresponding dataset. By maintaining separation between the data and the plot description until production time, memory management is cleaner and may reduce the number of spurious copies. The design enables customization—any cell in the matrix can be substituted with any plot created by `ggplot2`, using the `getPlot` and `putPlot` functions. This additional flexibility does come with a time penalty compared to the `ggplot2` `plotmatrix` approach.

As with many R functions, arguments recognized by `ggplot2` can be provided to `ggpairs` and passed through to the lower-level plotting functions. When a plot is rendered,

the title, legend, and axis labels are removed from the display for more efficient use of space. All of this information is kept internally though, so that a user can easily inspect or modify each individual plot. Indeed, any plot can be retrieved from the structure, modified, and placed back into the matrix. Using `ggplot2` as the base plays a large role in making this possible because it defines the plot as an abstract quantity, with values populated by data when data are provided. The color choices from `ggplot2` are available, although traditional legends are not displayed. Legends can be inferred when the correlation is displayed in one of the cells, because the correlation is calculated and displayed separately for each color group.

The coordination of axis scales and labels is an important and often challenging aspect of most complex graphical displays; `ggpairs` uses global limits to ensure that all panels of the generalized pairs plot are aligned appropriately on each axis. In addition, variable names and axis labels (whether scales or categories) are inserted on the diagonal by default, providing an alternative to the marginal distributions displayed in diagonal panels by `gpairs`.

The composite display is the conceptual basis for the `GGally` package, which will eventually provide many other types of plots. In addition to the `ggpairs` plot, it includes the `ggparcoord` plot, implementing the parallel coordinate plot (Inselberg 1985; Wegman 1990) using a composite plot construction. This display supports different choices of univariate plots for each axis, scaling of each variable, and reordering of variables by several different algorithms.

5. DISCUSSION

This article introduces the generalized pairs plot as a tool for graphical EDA and offers two implementations that evolved separately. Each implementation could be expanded with further options. For example, time series might be displayed using lines rather than points, a capability currently supported in the basic `pairs` plot of R for panels corresponding to time-quantitative pairs of variables when the time variable is represented as an object of class `ts`. When the nontime variable is categorical, however, new types of displays will need to be developed. Similarly, additional features could offer specialized behavior for ordered factors or spatially distributed data. Other future extensions include dot plots (Wilkinson 1999a) and more flexible options for identifying specific points across the various types of panels.

EDA is also enhanced by interactive graphics. The generalized pairs plot introduced here is a static plot, but each point or category is naturally associated with other points or categories in the display. An interactive generalized pairs plot would require brushing of objects for selection and linking across different panels of the display. The original pairs plot was one of the first to be adapted for interactivity (Becker and Cleveland 1988), but the generalized plot offers a unique set of challenges. Would a highlighted subset be displayed as a separate boxplot? Overlaid on the boxplot of the full data? Considerable work has already been done with interactive graphics (e.g., see Unwin (1999), Theus (2003), Theus and Urbanek (2008), and Swayne et al. (2003)). None of these work addresses linking of plots as required in a generalized pairs plot.

Data exploration should not be automated or optimized in a solely algorithmic fashion. Effective EDA requires human intervention and adaptation to inevitable surprises and diversity of features in the data. For example, the automated selection of an "ideal bandwidth" for a density estimate conflicts with the spirit of EDA. Multiple bandwidths should be investigated in the context of real-world questions about the data, and different reasonable choices can each serve useful purposes. Although no single version of a pairs plot is likely to be best for all applications, the generalized pairs plot is a promising addition to the field of multivariate analysis and can help guide and inform subsequent modeling and statistical inference.

SUPPLEMENTARY MATERIALS

Data and scripts: Datasets along with the commands used to produce the displays in this article are available online in a .zip archive file.

R packages: Each of the R packages used in this article (`barcode`, `gpairs`, `YaleToolkit`, `GGally`, and `ggplot2`) are available online (URLs are provided in the bibliography).

ACKNOWLEDGMENTS

The authors thank John Hartigan, Antony Unwin, and many students for advice and testing of these graphical displays. This work was partially supported by an unrestricted fellowship from Novartis, and by National Science Research grant DMS0706949.

[Received July 2011. Revised May 2012.]

REFERENCES

Anderson, E. (1935), "The Irises of the Gaspe Peninsula," *Bulletin of the American Iris Society*, 59, 2–5. [80]

Basford, K. E., and Tukey, J. W. (1999), *Graphical Analysis of Multiresponse Data: Illustrated With a Plant Breeding Trial*, Boca Raton, FL: Chapman & Hall/CRC. [79]

Becker, R. A., and Cleveland, W. S. (1988), "Brushing Scatterplots," in *Dynamic Graphics for Statistics*, eds. W. S. Cleveland and M. E. McGill, Monterey, CA: Wadsworth, pp. 201–224. [89]

Becker, R. A., Cleveland, W. S., and Shyu, M. J. (1996), "The Visual Design and Control of Trellis Display," *Journal of Computational and Graphical Statistics*, 5, 123–155. [86]

Bryant, P. G., and Smith, M. A. (1995), *Practical Data Analysis: Case Studies in Business Statistics*, Homewood, IL: Richard D. Irwin. [81]

Chambers, J. M., Cleveland, W. S., Kleiner, B., and Tukey, P. A. (1983), *Graphical Methods for Data Analysis*, Belmont, CA: Wadsworth International Group. [79,80]

Chambers, J. M., and Hastie, T. J. (1992), *Statistical Models in S*, Pacific Grove, CA: Wadsworth & Brooks. [83]

Cleveland, W. S. (1993), *Visualizing Data*, Summit, NJ: Hobart Press. [79]

Emerson, J. W., Esty, D. C., Levy, M. A., Kim, C. H., Mara, V., de Sherbinin, A., and Srebotnjak, T. (2010), *2010 Environmental Performance Index*, New Haven, CT: Yale Center for Environmental Law and Policy. [83]

Emerson, J. W., and Green, W. (2012a), *barcode: The Barcode Plot*, R package version 1.1. Available at *http://CRAN.R-project.org/package=barcode*. [83]

——— (2012b), *gpairs: The Generalized Pairs Plot*, R package version 1.1. Available at *http://CRAN.R-project.org/package=gpairs*. [80,85]

——— (2012c), *YaleToolkit: Data Exploration Tools from Yale University*, R package version 4.1. Available at *http://CRAN.R-project.org/package=YaleToolkit*. [83]

Emerson, J. W., Green, W. A., and Hartigan, J. A. (2006), "Barcodes, Generalized Pairs Plots, and Sparkmats," UseR! 2006 conference presentation, Vienna. [80,83]

Fisher, R. A. (1936), "The Use of Multiple Measurements in Taxonomic Problems," *Annals of Eugenics*, 7, 179–188. [80]

Friendly, M. (1994), "Mosaic Displays for Multi-Way Contingency Tables," *Journal of the American Statistical Association*, 89, 190–200. [80]

——— (2002), "Corrgrams: Exploratory Displays for Correlation Matrices," *The American Statistician*, 56, 316–324. [86]

Grosjean, P. H., Spirlet, C., and Jangoux, M. (2003), "A Functional Growth Model With Intraspecific Competition Applied to a Sea Urchin, *Paracentrotus lividus*," *Canadian Journal of Fisheries and Aquatic Science*, 60, 237–246. [87]

Hartigan, J. A. (1975), "Printer Graphics for Clustering," *Journal of Statistical Computation and Simulation*, 4, 187–213. [79]

Hartigan, J., and Kleiner, B. (1984), "A Mosaic of Television Ratings," *The American Statistician*, 38, 32–35. [80]

Inselberg, A. (1985), "The Plane With Parallel Coordinates," *The Visual Computer*, 1, 69–91. [89]

Murrell, P. (2005), *R Graphics*, Boca Raton, FL: Chapman & Hall/CRC. [80]

National Research Council (2010), "Data-Based Assessment of Research-Doctorate Programs." Available at *http://www.nap.edu/rdp*. [87]

R Development Core Team (2012), *R: A Language and Environment for Statistical Computing*, Vienna: R Foundation for Statistical Computing. Available at *http://www.R-project.org/*. [80]

Sarkar, D. (2008), *Multivariate Data Visualization with R*, New York: Springer. [86]

SAS Institute (2010), *JMP*. Available at *http://www.jmp.com/*. [87]

Schloerke, B., Crowley, J., Cook, D., Hofmann, H., and Wickham, H. (2012), *GGally: Extension to ggplot2*, R package version 0.3.2. Available at *http://CRAN.R-project.org/package=GGally*. [80,87]

Swayne, D., Lang, D., Buja, A., and Cook, D. (2003), "GGobi: Evolving From XGobi Into an Extensible Framework for Interactive Data Visualization," *Computational Statistics & Data Analysis*, 43, 423–444. [89]

Theus, M. (2003), "Interactive Data Visualization Using Mondrian," *Journal of Statistical Software*, 7, 1–9. [89]

Theus, M., and Urbanek, S. (2008), *Interactive Graphics for Data Analysis: Principles and Examples*, London: Chapman & Hall/CRC. [89]

Tukey, J. W. (1977), *Exploratory Data Analysis*, Reading, MA: Addison Wesley. [81]

Tukey, J. W., and Tukey, P. (1990), *Strips Displaying Empirical Distributions: I. Textured Dot Strips*, Bellcore Technical Memorandum. [84]

Tukey, P. A., and Tukey, J. W. (1981), "Graphical Display of Data Sets in Three or More Dimensions," in *Interpreting Multivariate Data*, ed. V. Barnett, Chichester: Wiley, pp. 189–275. [79]

Unwin, A. (1999), "Requirements for Interactive Graphics Software for Exploratory Data Analysis," *Computational Statistics*, 14, 7–22. [89]

Wegman, E. (1990), "Hyperdimensional Data Analysis Using Parallel Coordinates," *Journal of American Statistics Association*, 85, 664–675. [89]

Wickham, H. (2009), *ggplot2: Elegant Graphics for Data Analysis*, New York: Springer. Available at *http://had.co.nz/ggplot2/book*. [80,86]

Wilkinson, L. (1999a), "Dot Plots," *The American Statistician*, 53, 276–281. [89]

——— (1999b), *The Grammar of Graphics*, New York: Springer-Verlag. [80,86]