A Brief History of the Mosaic Display

Michael Friendly*

September 25, 2001

Abstract

This paper provides an illustrated history of the visual and conceptual ideas leading to the development of mosaic displays. We trace the origins of the use of rectangles and area to depict data quantities and their relations, of early forms of mosaic displays including sub-divided bar-like charts and various cartograms, to the modern forms used in log-linear analysis and in space-filling tree maps.

Key words: data visualization, space-filling displays, cartogram, thematic cartography, log-linear models, mosaic matrix, tree map

1 Introduction

Mosaics are space-filling designs composed of contiguous shapes ("tiles"). From the earliest Greek and Roman pictorial mosaics, to the intricate, non-representational Islamic mosaics of the Alhambra, to the playfully mathematical graphic works of M. C. Escher, and the beautifully chaotic architectural renderings of Antonio Gaudi, mosaic tilings have long been objects of beauty, wonder, and instruction.

In statistical graphics, as in any other field, lessons and achievements from the past should inform workers in the present— if only we knew and appreciated them. In like token, present developments may serve to stimulate future work, particularly if the bird's-eye view of history can put them in perspective— where they came from, where they might go. What follows is an attempt at a small piece of historiography in data visualization: the use of space-filling mosaic-like designs to portray quantitative and categorical data.

This account is thematic, rather than (restrictively) chronological: we wish to emphasize the development of visual and statistical ideas for data display. To motivate this paper, the remainder of this section briefly outlines the present use of mosaic plots for *n*-way contingency tables, listing the salient graphical features and principles employed. Section 2 gives a conceptual history of the visual and statistical ideas leading to the present uses. Section 3 describes the graphical and statistical innovations introduced in modern data visualization.

1.1 Background

In statistical graphics, the mosaic display, attributed to Hartigan and Kleiner (1981), is a graphical method to show the values (cell frequencies) in a contingency table cross-classified by one or more "factors". Figure 1 shows the basic form of the display for a two-way table of individuals classified by hair color and eye color (data from Snee (1974))—perhaps a trite, but by now canonical example (Friendly, 1991, 1994, 1995, 2000b, Spence, 2001). As explained below, the area of each tile is proportional to the cell frequency, n_{ij}, and if hair color and eye color were independent, n_{ij} would be proportional to the product of the row and column marginal totals, $n_{i+}n_{+j}$, so the tiles in each column would align horizontally. The fact that they *do not* align reveals an association between these two variables.

As extended to show both the data, and residuals from a log-linear model (Friendly, 1994, Theus and Lauer, 1999), mosaic displays have become a primary graphical tool for visualization and analysis of categorical data in the form of *n*-way contingency tables. Theus and Lauer (1999) and Friendly (2000b, §4.5) show how mosaic displays can be used to understand the structure of log-linear models themselves.

*Michael Friendly is Professor, Psychology Department, York University, Toronto, ON, M3J 1P3 Canada, E-mail: friendly@yorku.ca.

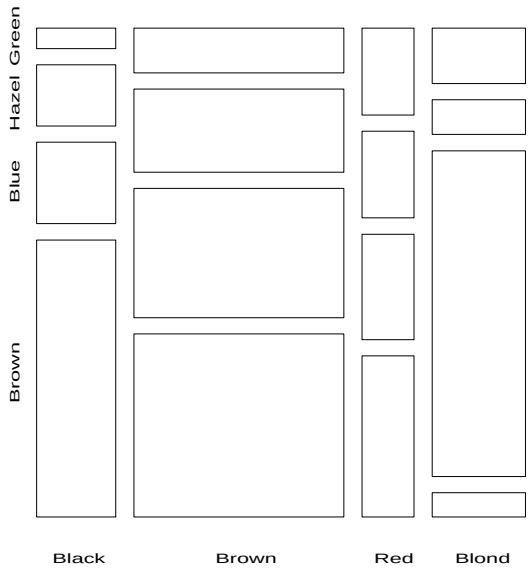

Figure 1: Basic mosaic display for hair color and eye color data. The area of each rectangular "tile" is proportional to the frequency in that cell.

Some examples are shown in Figure 2, for two-way and three-way tables, showing the relations among the categories of hair color, eye color and sex. In these figures, the area of each tile is proportional to the observed cell frequency, as in Figure 1, but the tiles are shaded according to the residuals from a particular log-linear model, thus showing the pattern of associations which remain. The interpretation of these figures, presented here simply to illustrate this graphic form, is described in Section 3.2.

The graphical features of the mosaic display are:

- A unit area (square or rectangle) is divided into bars, whose widths represent the quantities (marginal frequencies, n_{i+} or probabilities $p_{i+} = n_{i+}/n_{++}$) of one variable.
- Those bars are sub-divided into "tiles", whose heights represent the quantities (conditional probabilities $p_{j\,|\,i} = n_{ij}/n_{i+}$) of a second variable.
- This process of sub-division can be extended to any number of variables.
- At any stage, the *area* of each tile represents the total quantity (cell frequency, n_{ij}) in the cross-classification of the variables included.
- For cross-classified frequency data, the tiles in the mosaic will align when the variables are statistically independent.

The principal graphical ideas are:

- using area = height $\times$ width, to represent a quantity which depends on a product of two other variables, each of interest,
- using recursive sub-division to show any number of variables,
- using shading to display some other attribute of the data,
- purely multiplicative relations (e.g., $p_{ij} = p_{i+}p_{+j}$) produce equal sub-divisions,
- for two or more variables, the levels of sub-division are spaced with larger gaps at the earlier levels, to allow easier perception of the groupings at various levels, and to provide for empty cells.

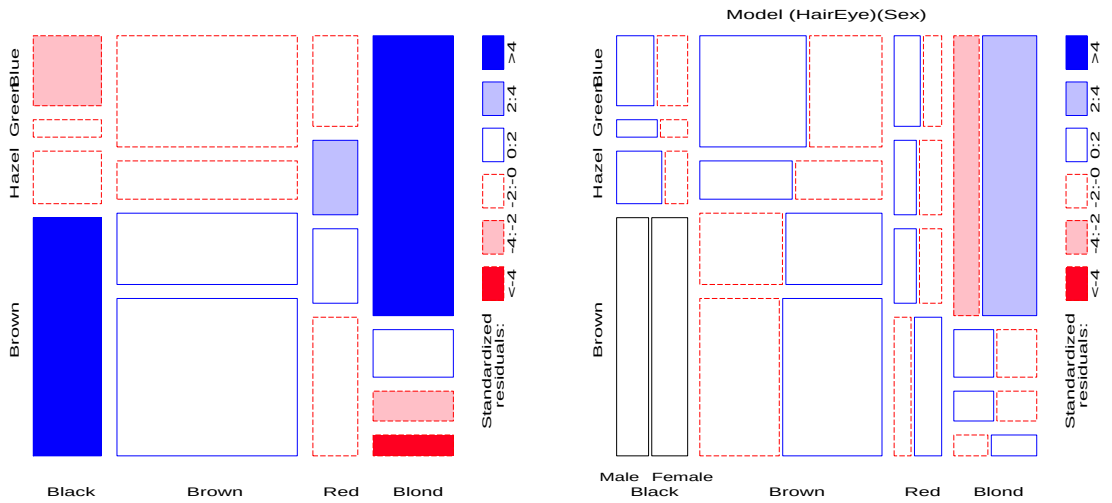

Figure 2: Mosaic displays for frequencies of hair color, eye color and sex. Left: two-way table, hair color by eye color; right: three-way table, divided by sex.

2 History

Harry S. Truman is quoted as saying, "The only new thing in the world is the history you don't know." Indeed, the visual and conceptual ideas leading to mosaic displays have been re-invented many times, but their history is not well-known. We trace a few of the strands which led to this visual representation below. There are two subthemes to this historiography: the use of area to display numerical quantities, and rectangular displays where height and width depict two primary quantities, whose product (area) is also to be represented.

2.1 Use of area to represent statistical quantities

The earliest known pictorial representation using rectangular area to display a quantity derived from height $\times$ width was Edmund Halley's (Halley, 1693) diagram (see Figure 3) to show the chances for survival and death for two independent lives. Halley analyzed data from the best-known, and most reputable life-tables of survival at the time (Graunt, 1662), with a view to informing the English government on the value to be set for the purchase of an annuity.

Having done so, he remarks that the value of an annuity based on two independent lives, may be found from the same table, since "the number of Chances of each single Life, found in the Table, being multiplied together, become the Chances of Two Lives." Figure 3 shows the partition of a unit area according to living or dying in a given interval of time, for two individuals, assumed to do so independently. With an understanding of the graph, the annuities for any combination of life and death could be readily calculated from the table for a single life. Halley generalizes this result to three independent lives, foreseeing the notion of mutual independence in a three-way contingency table by over one hundred years.

The visual representation of actual data by areas was introduced in the early 1780s, in France by Charles de Fourcroy (Palsky, 1996, Fig. 15) and in Germany by August Friedrich Wilhelm Crome (1785). Fourcroy used proportional superimposed squares to compare demographic quantities (see Figure 4), while Crome used areas— with a literal interpretation— in a graph showing the areas of the European states by superimposed squares, so that these could be compared more easily than when shown separately on a map. However, these are simple area diagrams, rather than proto-mosaics.

A later development was the use of rectangles as thematic cartographic symbols, where height and width represented two variables, whose product (area) should also be visually prominent. The first use of this type is not known, but a fine early example is a map of Paris by Jacques Bertillon (Bertillon, 1896) of the foreign-born population in 1891, shown in Figure 5.

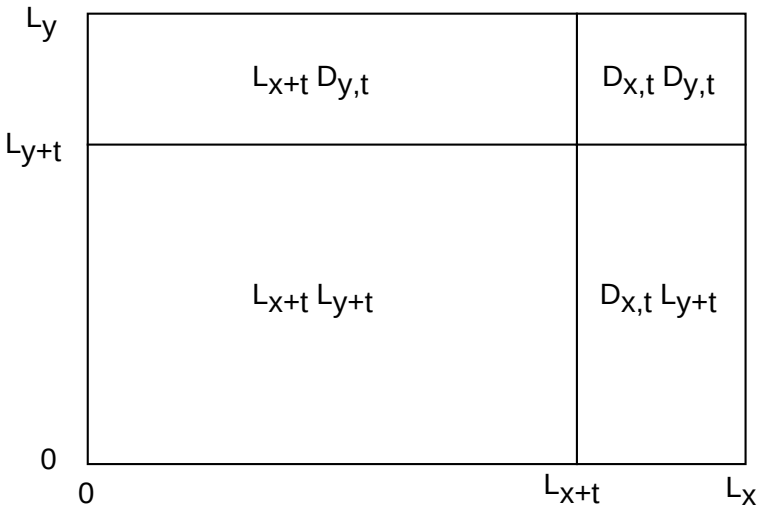

Figure 3: Halley's diagram, on the chances for two independent lives, x and y. L_x (L_y) are the numbers living at a given time, L_{x+t} (L_{y+t}) are the numbers living t years later, and $D_{x,t} = L_x - L_{x+t}$ ($D_{y,t}$) are the numbers dying in that interval of time, t. *Source*: redrawn from Hald (1990, Fig. 9.3.3)

For each sector (some arrondisements are sub-divided) the height of the rectangle is proportional to overall population, and width to percent foreigners; so area is proportional ($\sim$) to the absolute number of foreigners. Many interesting details may be read easily from the map, e.g., small total numbers of foreigners on the left bank, very small percentage of foreigners in the 1^e and 2^e arrondisements (right bank, left of Ile de la Cité), large percentage of foreigners in the northern region, particularly the 18^e (top), large variation within some arrondisements (12^e, bottom right), and so forth.

Of course, area was also used to represent frequencies, but in a circular form in Florence Nightingale's coxcomb diagrams showing number of deaths in the Crimean War. However, as these early examples illustrate, rectangular displays have the advantage of showing two primary quantities by lineal extent and their product as an area.

Most recently, psychophysical experiments comparing judgments of graphical attributes (length, position along an axis, area, color, texture, etc.) by Cleveland and McGill (Cleveland and McGill, 1984, 1985) have shown that, for tasks of magnitude estimation ("how much is x?), judgments of area are less accurate than attributes of length or position along an axis. Others, e.g., Simkin and Hastie (1987), Lewandowsky and Spence (1989) demonstrate that the ordering of visual attributes depends strongly on the viewer's task ("what fraction of the total is x?", "which is greater, x or y?"). However, *all* of these experimental results focus on what Bertin (Bertin, 1977, 1981) calls "elementary" or "intermediate" tasks. The true virtue of mosaic displays, as used today, is for more complex tasks: assessing patterns, trends, and anomalies. Moreover, Friendly (1995) shows that area is a natural visual representation for frequency data, with strong links to statistical theory (maximum likelihood estimation) and phenomena (power), and an underlying physical model which likens counted observations to gas molecules in a pressure chamber.

2.2 Early mosaic displays

By the early 1800s, all the modern forms of statistical graphics had been invented (pies, bars, line graphs, scatterplots) (Beniger and Robyn, 1978, Friendly and Denis, 2001). On the surface, mosaic plots descend from bar charts, but it was not until 1844 that Charles Joseph Minard introduced two simultaneous innovations: the use of divided and proportional bars in his "Tableau-graphique," showing the transportation of commercial traffic along canal routes in France by variable-width, divided bars Minard (1844). In these displays (Figure 6) the width of each vertical segment shows distance; the divided bars have height $\sim$ amount of goods, so area $\sim$ cost of transport. Minard, a true visual engineer (Friendly, 2000a), developed such diagrams to argue visually for setting differential price rates for partial vs. complete runs (Minard, 1842). Playfair had made data "speak to the eyes," but Minard wished to make them "calculer par l'oeil."

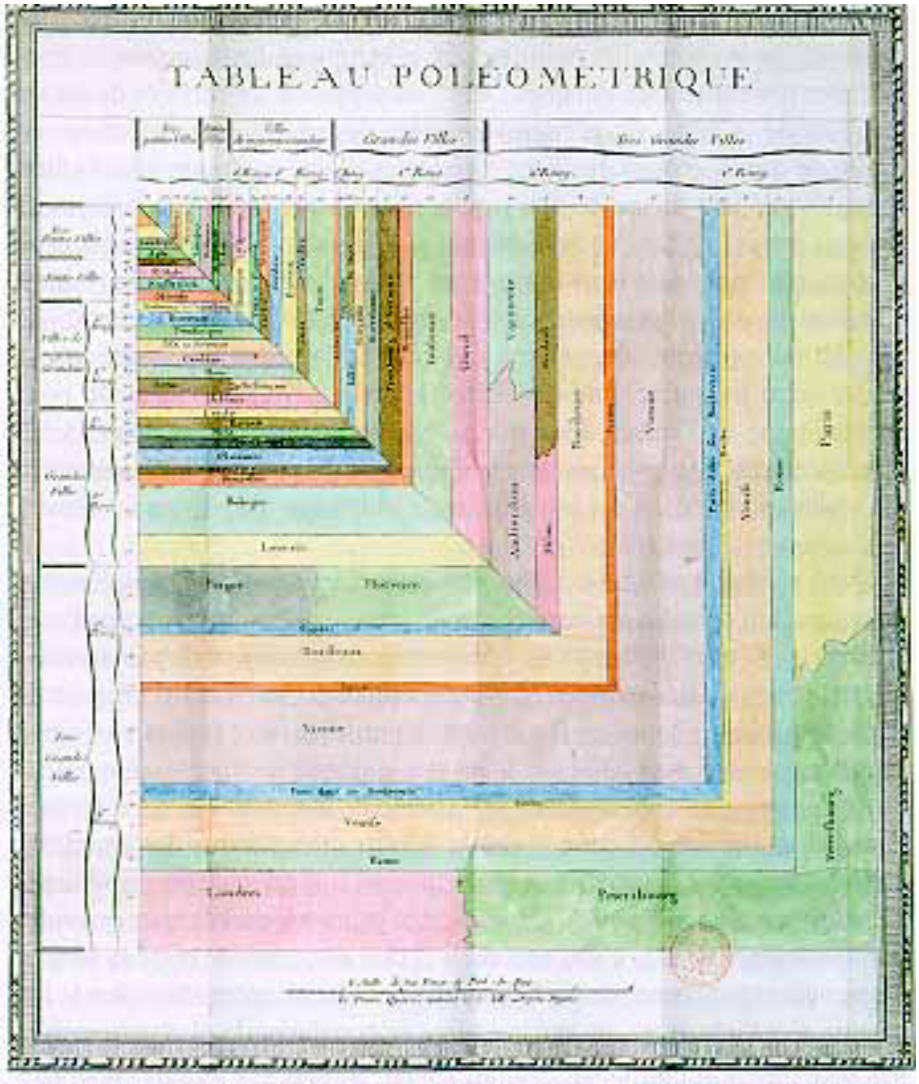

Figure 4: Charles de Fourcroy's Tableau Poléométrique. *Source*: Palsky (1996, Fig. 15).

From Ostermann (1999), it appears that Georg von Mayr (Mayr, 1877) was the first to use the format of the mosaic display in its modern form, at least for two variables. In Mayr (1877, S. 80) (see Figure 7) he shows the representation of a 3×3 table, where a total count of 1000 is subdivided first into categories A, B, C of size 600, 300, and 100. Each of these bars is then subdivided by categories a, b, c, and von Mayr uses different cross-hatch patterns to distinguish the levels of the second variable. It is not entirely clear whether von Mayr contemplated the display of additional variables, but the graph title, "Area diagram with two splittings" does suggest this extension.

These proportional, divided squares soon appeared again in the French *Albums de Statistique Graphique* and Swiss *Graphische-statistischer Atlas*, though in quite different forms.

In the French *Album* of 1885 (Ministère des Travaux Publics, 1879–1899), the mosaic form was used in a remarkable graphic tour-de-force to show the distribution of passengers and goods throughout France, starting in Paris, with one mosaic showing the breakdown of this variable according to the principal railway routes (Gare du Nord, Gare d'Est, etc.). Each destination was then represented by a proportionally smaller divided square showing the distribution there (color coded by origin), with lines connecting that square to its' origin and destinations (see Figure 8). This is certainly the finest early example of the use of mosaics to portray quantitative data, and it is the image which led to the present historical account.

In the Swiss *Atlas* (Statistischen Bureau, 1897), Tableau III, titled "Superficie territoriale de la Suisse et

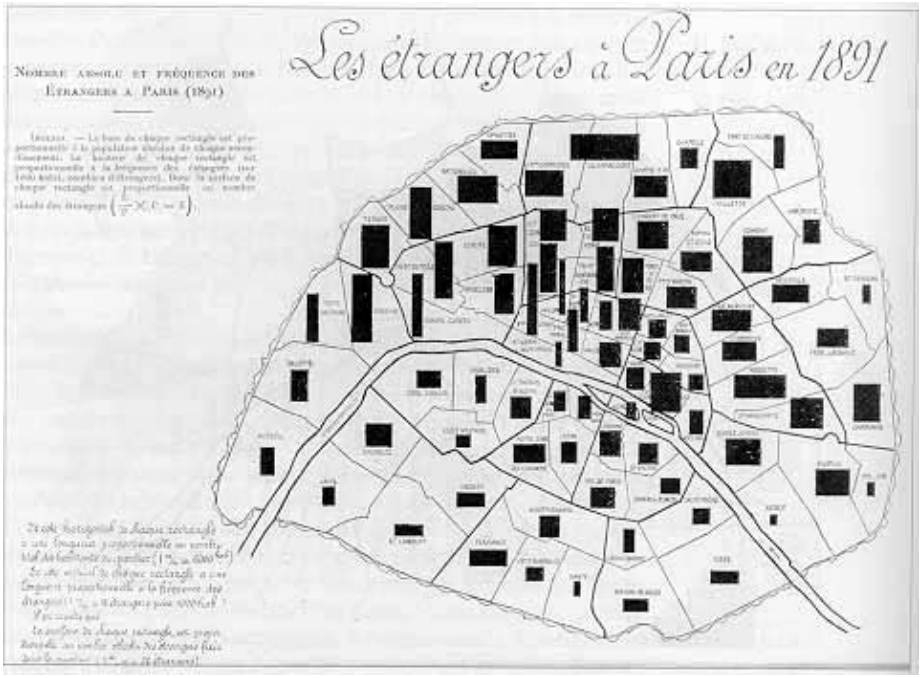

Figure 5: Bertillon's 1896 map of the population of foreigners in Paris. *Source*: Palsky (1996, Fig. 85).

densité de la population par cantons" shows one-way proportional squares of the area of each Swiss canton, broken down as productive vs. unproductive land, together with bar graphs of population density. (These were not displayed on a map, but rather on a page, arranged by area. It is of some graphical interest that the smallest cantons had the greatest population density, so, in order to keep the scales consistent, the population density bars for small cantons were bent and folded to make them fit the available space.)

2.3 Hundred-per-cent squares

By the early 1900s, the use of charts and graphs had become commonplace, and a variety of texts on their construction and use were written. In 1925, we find Karl Karsten's "Charts and Graphs" (Karsten, 1925) devoting one chapter to what he called "hundred-per-cent bars" and another to "hundred-per-cent squares" (true mosaics). These may have been described in earlier texts, but we know of no other which states the principles (and their limitations) so clearly. He says,

> We have so consistently inveighed against the use of areas to illustrate quantities that the reader will indeed be surprised at some coming retractions... But the fact is that we now propose to turn to advantage the very feature of areas which has previously been their greatest fault. ...
>
> We now come to data in which we wish to show simultaneously three ratios or sets of ratios, one of which is always the product of the other two. In other words, we wish to show two factors or sets of factors and their product.

Karsten then illustrates (in painstaking detail) the breakdown of a table of occupational categories (managers, clerks, skilled-workers, semi-skilled, laborers) by sector (agriculture, mining, manufacturing, etc.), and shows several other examples, including one of the estimated unmined world coal supplies, shown both as a mosaic and in nested circular form.

His use of the phrase "or sets of factors" anticipates the extension to three or more dimensions, but his hand-drawn two-way diagrams were, evidently, difficult enough that a multi-way mosaic would not be constructed until computer hardware and software eased the burden.

Looking forward, Karsten's hundred-per-cent squares were recently re-discovered by Hummel (1996) as "spine-plots," an alternative to stacked bar charts where the bar widths are proportional to the magnitudes of one variable, and some present workers derive the mosaic by generalization of the spine-plot (Hofmann, 2000). However, the present review shows that this derivation is conceptual rather than historically based.

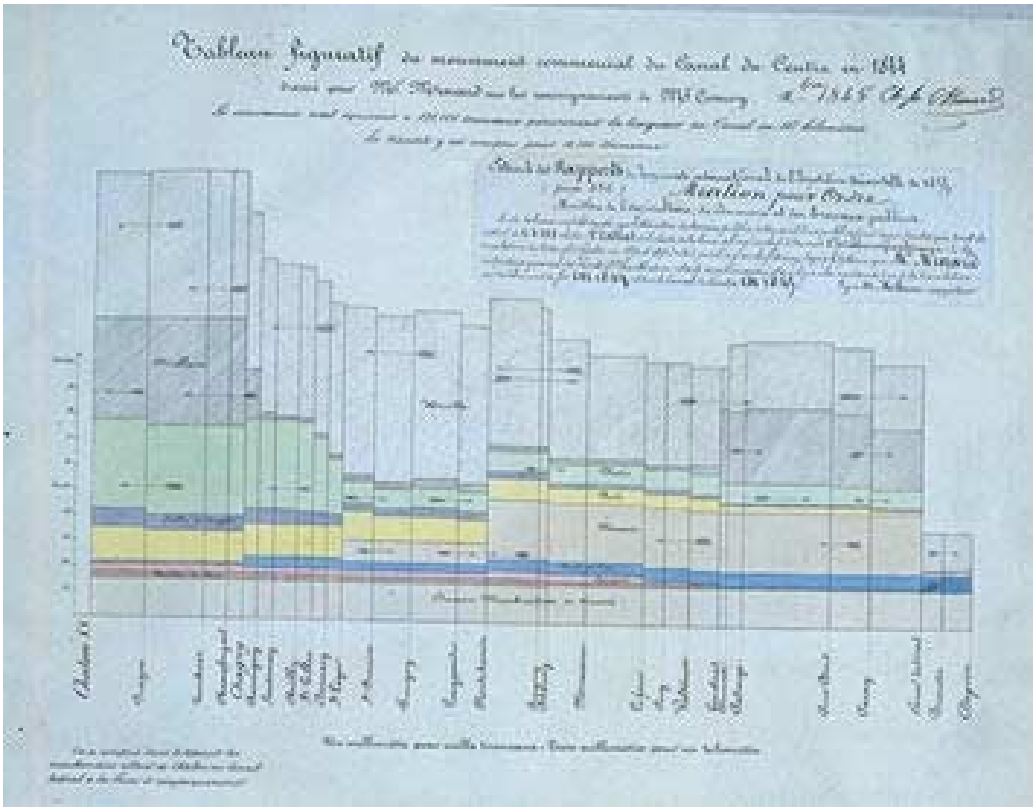

Figure 6: Minard's Tableau Graphique, showing the transportation of commercial goods along the Canal du Centre (Chalon–Dijon). Intermediate stops are spaced by distance, and each bar is divided by type of goods, so the area of each tile represents the cost of transport. Arrows show the direction of transport. *Source:* ENPC:5860/C351 (Col. et cliché ENPC; used by permission)

2.4 The rectangular statistical cartogram

In geography and cartography, these ideas seem to have been re-discovered by Raisz (1934) as the "rectangular statistical cartogram." He says,

> The idea ... occurred to the author when he had occasion to prepare maps of the United States showing the distribution of various economic units, such as steel factories, textile mills, power plants, banks, etc. These maps were far too crowded in the northeast to be useful, while elsewhere, for the most part, they were relatively empty. ... The system used here starts always with the larger divisions and by "proportional halving" arrives at the smaller ones.

Raisz goes on to show cartograms representing land area, population, national wealth, value from manufacture, agriculture, etc. These cartograms represent the United States by a rectangle, 1×1.5, to which he added smaller rectangles for New England (upper right) and Florida (lower right). Vertical divisions separated the Pacific states, Mountain States, West Central, East Central, and Atlantic states, while other divisions were made within each region as he deemed appropriate for a particular display. It is of some interest that each cartogram is recognizable both as a schematic map, and as a mosaic sub-division of a given quantitative total.

The rectangular cartogram apparently became sufficiently well-known in cartography for Birch (1964) to include in his treatment of "diagrams and diagram maps." As an example, he shows (Figure 9) a cartogram of land use in eight states of Australia, stratified by type of use (held under lease, unoccupied, etc.).

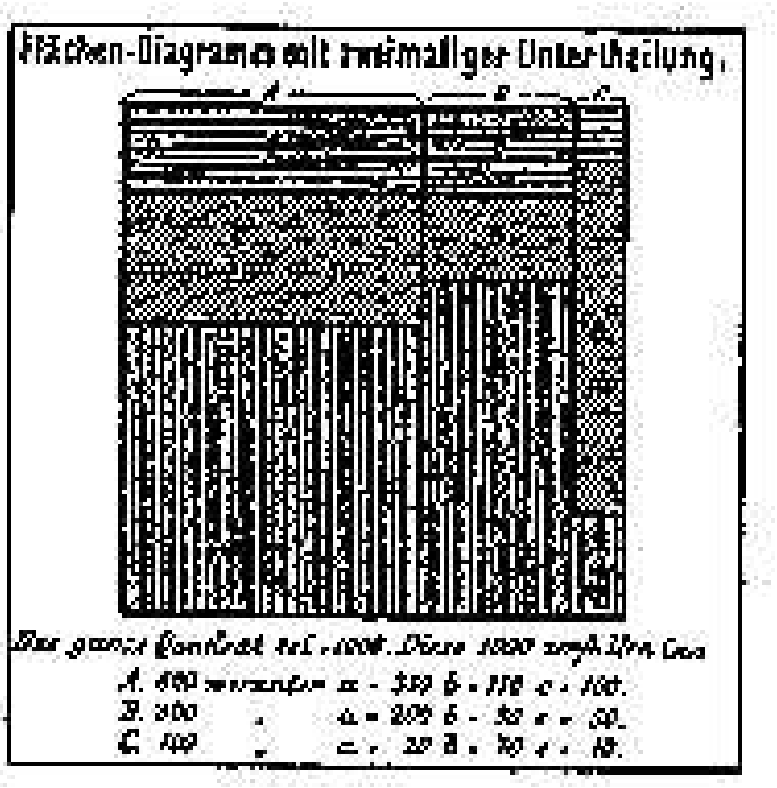

Figure 7: von Mayr's Area Diagram, the earliest-known modern mosaic. *Source*: Ostermann (1999, Fig. 3.3).

2.5 Sieve diagrams

The original form of the mosaic display (Figure 1) shows the observed cell frequencies, but departure from independence is shown only by differences in height among corresponding bars. Riedwyl and Schüpbach (1983, 1994) proposed the "sieve diagram" (later called a "parquet diagram") to provide a direct, visual comparison of observed and expected frequencies under independence. However, unlike the mosaic, this form does not generalize readily to more than two variables.

In this display the tiles have height and width proportional to the marginal frequencies, n_{i+} and n_{+j} respectively, so the area of each rectangle is proportional to expected frequency, $m_{ij} = n_{i+}n_{+j}/n_{++}$. The observed frequency is shown by the number of cross-ruled squares in each rectangle. Hence, the difference between observed and expected frequency appears as variations in the density of shading. Cells whose observed frequency n_{ij} exceeds the expected m_{ij} appear denser than average. The pattern of positive and negative deviations from independence can be more easily seen by using color, say, red for negative deviations, and blue for positive, as shown in Figure 10 for the same data as in Figure 1.

3 Modern mosaic displays

3.1 Mosaic displays for log-linear analysis

As mentioned at the outset, the modern mosaic display in statistical graphics is usually attributed to Hartigan and Kleiner (1981). However, it is fair to say that Bertin's (1977, 1981) "la matrice pondérée" is very similar to, and contains most of the graphic features of present-day mosaics. It may also be said that much of the present interest in mosaic plots stems from their use as statistical visualization tool for frequency data and log-linear modeling (Friendly, 1992a, 1994, 2000b) along with recent extensions of these methods (Friendly, 1999, Theus and Lauer, 1999) and a wide variety of computational realizations, listed next.

Computationally, the implementation was first described in FORTRAN (Wang, 1985). This method

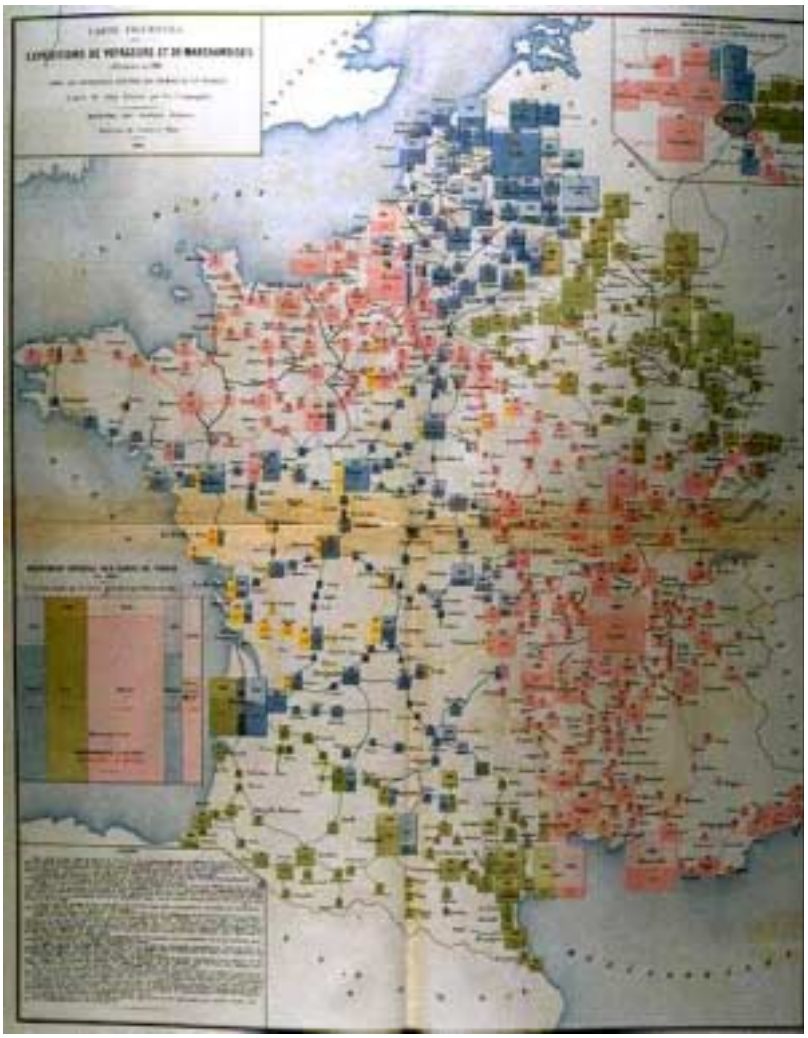

Figure 8: Map-based multi-mosaic of the distribution of passengers and goods in France. *Source*: Author's collection

has now been implemented in many statistical systems and programming languages, including SAS/IML (Friendly, 1992b), SAS/INSIGHT, JMP (SAS Institute, Inc., 2000), S-Plus (Emerson, 1998), Java (MONDRIAN, Theus (1997)), VISTA (Young and Bann, 1996, Young et al., 2000), and MANET (Unwin et al., 1996, Hofmann, 2000). The last three are notable for providing dynamic, interactive visualization of contingency table data (as do SAS/INSIGHT and JMP, but only for two-way tables). MANET and MONDRIAN are also notable for providing an explicit representation for empty cells.

3.2 Graphical innovations

For two-way tables, one variable is sub-divided on the horizontal axis and the other on the vertical, so whether the tiles are contiguous or separated by small gaps is immaterial. For larger, multi-way tables (3 or more variables), two or more variables must be assigned in a cross-classification to one or both axes, making it more difficult to discern (and label) the various subcategories. Hartigan and Kleiner (1984) illustrated the mosaic display of a six-way table of television ratings, where the tiles were spaced with larger gaps between levels of variables split earlier. This turns out to be a perceptually important innovation for showing the subdivisions of several variables simultaneously, because the viewer can readily distinguish larger from smaller groupings of the mosaic tiles. The un-spaced versions are sometimes called "Mondrian diagrams" (from Theus (1997)) to distinguish them.

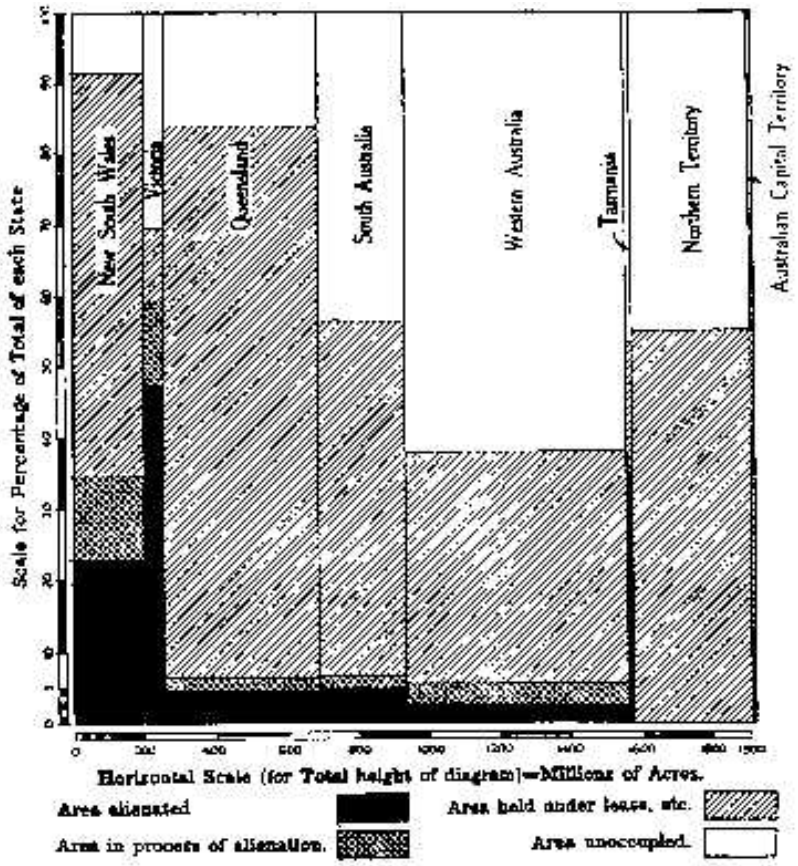

Figure 9: Birch's cartogram of land use in Australia. *Source*: Birch (1964, Fig. 94).

In earlier mosaics, the order of the categories of each variable is arbitrary, as is the scheme used in coloring or shading the tiles. In contrast, the mosaics in Friendly (1994) introduced two graphical features designed to facilitate the perception of patterns of relations among the variables: (a) each tile is shaded according the residual from a particular model, using a bipolar color mapping to show both the sign and magnitude of the residual. (b) the order of the categories of the variables is permuted to place cells with similar residuals contiguously. These mosaics show both the frequencies in the cells of a contingency table (area of the tiles), and the pattern of association between the variables (color and shading intensity).

Figure 2 illustrates these innovations. In the left, two-way display, the categories of hair color and eye color were reordered according to best one-dimensional representations of the association between these variables, which orders both from dark to light (suggesting an explanation for the association). The model of independence has been fit to the table, and opposite-corner pattern of the shading shows how the frequencies deviate from independence, i.e., the *pattern* of association between hair color and eye color.

In the right panel, the log-linear model $[HairEye][Sex]$ has been fit to the three-way table, which asserts that the combinations of hair color and eye color are independent of sex. Only one pair of cells have residuals large enough to be shaded, the blue-eyed blondes, where there are more women and less men than would be the case under this model of independence. Observe that we can still see the relative sizes of the hair-eye combinations because of the spacing between the tiles.

3.2.1 Interactive, visual fitting

The interactive use of shading and coloring of the tiles to convey some additional information deserves further comment. In some implementations (e.g., JMP, S-Plus), the tiles are colored nominally according to the levels of one variable, merely to keep the categories of that variable visually distinct.

Others (e.g., Friendly (1994), VISTA), as just mentioned, use color and shading to display the residual

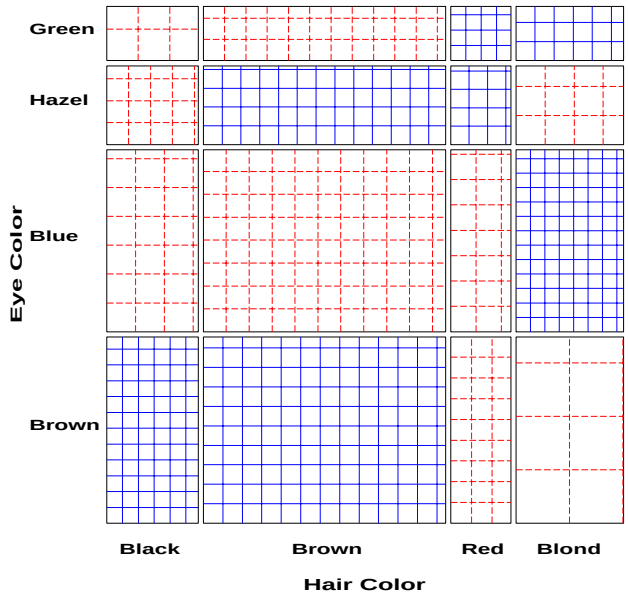

Figure 10: Sieve diagram for hair-color, eye-color data. Observed frequencies are equal to the number squares in each cell, so departure from independence appears as variations in shading density.

from a particular log-linear model. With a bipolar color mapping, ranging from deep blue (say) for large positive residuals, through white, to deep red for large negative residuals, a good-fitting model will have most tiles unfilled.

Hence, model search may be characterized as "cleaning the mosaic," and interactive systems (VISTA, MONDRIAN) provide a perceptual basis for "visual fitting." (Valero et al., 2001) extend this idea considerably in VISTA with a spreadplot (a collection of windows linked algebraically and through interaction) containing the list of possible model terms, an influence plot and mosaic plots of observed and fitted frequencies for the current model, and finally a deviance plot of χ^2/df for all models.

The interactive methods introduced in MANET are designed more for visual exploration than for model fitting. Using linked mosaics, a tile selected in any one plot is highlighted (shaded) there and in all others. When one variable in the contingency table is a binary response, selecting one level of that response then shows the (conditional) proportion of that outcome in all other plots, and variations in the heights of the highlighted portions reveals dependence on the categories of the other variables. For example, in the well-known *Titanic* data (Friendly, 1999, 2000b, Hofmann, 1998), selecting the survivors can reveal how survival on the *Titanic* depended on the factors Age, Gender, and Class. However, this strategy is less successful when there is no natural response variable, or for a polytomous response. On the other hand, MANET (Hofmann, 2000) provides a rich variety of interactive methods for selection, queries, reordering variables and levels, and grouping, as well as variations of the mosaic construction for special purposes.

3.3 Statistical innovations

Mosaic displays have also been extended both for teaching statistical concepts and for visualizing the relations in large, complex frequency tables.

3.3.1 Teaching

In "Seeing Statistics", McClelland (1999) includes an interactive mosaic designed for teaching about categorical data analysis. The interactive design (using Java) allows students to change the value of any cell or marginal frequency in the table, or the size of any tile in the mosaic, and see the change in the visual display of frequencies and residuals, and associated numerical statistics.

For example, increasing or decreasing the total frequency in the table changes all cells proportionally (keeping the association the same), but changes the statistical strength of the association. This is immediately reflected visually in the strength of shading of the tiles, and numerically in the χ^2 value and its' significance level, and provides a tangible appreciation of the concept of statistical power. On the other hand, changing the marginal total in one category does not affect the χ^2 or shading of the tiles.

3.3.2 Links between quantitative and categorical data

As well, the links in statistical theory for quantitative and categorical data have been extended by recent work on mosaic displays (Friendly, 1999). Mosaic matrices provide discrete analogs of scatterplot matrices, and partial mosaic arrays stratified by the levels of one or more variables give analogs of Trellis displays. Both of these are mosaics composed of tiles whose elements are themselves mosaics, taking the recursive nature one step further.

For example, Figure 11 shows a mosaic matrix for the data on hair color, eye color and sex— a mosaic of all two-way marginal tables of the three-way table. The panels in row, column (1,2) and (2,1) show the same data (but with different order of splitting)— the relation between hair and eye color, displayed in the left panel of Figure 2. The bottom row and the last column show the association between each of hair and eye color with sex. Because all the tiles are unshaded, and are nearly aligned when split by sex, we see at a glance that there are no overall associations with sex in this data.

Figure 11: Mosaic matrix for hair color, eye color and sex. Each off-diagonal panel shows the two-way mosaic for the corresponding row and column variables.

Finally, we note that mosaic displays for *n*-way contingency tables provide a hierarchical decomposition of association in a way analogous to sequential fitting in regression. Mosaic plots assume that the table variables have been hierarchically ordered, $v_1, v_2, \ldots v_n$. By construction, the joint cell probabilities in the full *n*-way table are recursively decomposed as

$$p_{ijk\ell\cdots} = \underbrace{\overbrace{p_i \times p_{j|i}}^{\{v_1 v_2\}} \times p_{k|ij}}_{\{v_1 v_2 v_3\}} \times p_{\ell|ijk} \times \cdots \times p_{n|ijk\cdots} \tag{1}$$

The braces in (1) indicate that the first two terms provide a mosaic for the marginal frequencies of variables v_1 and v_2; the first three terms yield a mosaic for the $\{v_1 v_2 v_3\}$ marginal table, and so forth up to the display of the full *n*-way table.

Moreover (Friendly, 1994, §3.5), when sequential models of joint independence, $[v_1][v_2]$, $[v_1 v_2][v_3]$, $[v_1 v_2 v_3][v_4]$, ... are fit by maximum likelihood, the likelihood ratio G^2s for these models (and the corresponding mosaics) provide an additive decomposition of the total association, $G^2_{[v_1][v_2]\ldots[v_p]}$ (mutual independence), in any ordered subset of the first p variables:

$$G^2_{[v_1][v_2]\ldots[v_p]} = G^2_{[v_1][v_2]} + G^2_{[v_1 v_2][v_3]} + G^2_{[v_1 v_2 v_3][v_4]} + \cdots G^2_{[v_1 \ldots v_{p-1}][v_p]} \tag{2}$$

Thus, mosaic displays rely on two fundamental operations:

marginalization: The mosaic for variables $v_1 \ldots v_p$ presents the joint distribution of those variables, but ignores (collapses over) variables $v_{p+1} \ldots v_n$.

conditionalization: The mosaic for variables $v_1 \ldots v_p$ provides a visualization of the conditional distribution of $v_p \,|\, v_1, \ldots, v_{p-1}$.

3.4 TreeMaps

As it often turns out, a solution to one problem provides a solution to other, related problems. The mosaic display relates most naturally to a cross-classification, but the same visual ideas apply to a nested classification, or tree structure.

In 1991, Shneiderman (1991) developed the idea of representing a tree by recursive sub-division, rather than by the traditional approach using a rooted, directed graph with the root node at the top or left of the diagram. We describe this here simply to suggest the wider applicability of mosaic-like displays for future work in statistical graphics. For example, clustering problems, and classification and regression trees may benefit from this perspective.

The space-filling property of the mosaic allows much larger, and more complex trees to be usefully displayed (and provides for greater interactive use, such as zooming in on a portion of the display to see more detail) than the traditional rooted-tree. Figure 12 shows the treemap of file storage on the HCIL server classified by year and subject.

Shneiderman describes the origin of the idea (`http://www.cs.umd.edu/hcil/treemaps/`):

> During 1990, in response to the common problem of a filled hard disk, I became obsessed with the idea of producing a compact visualization of directory tree structures. Since the 80 Megabyte hard disk in the HCIL was shared by 14 users it was difficult to determine how and where space was used. Finding large files that could be deleted, or even determining which users consumed the largest shares of disk space were difficult tasks.
>
> Tree structured node-link diagrams grew too large to be useful, so I explored ways to show a tree in a space-constrained layout. I rejected strategies that left blank spaces or those that dealt with only fixed levels or fixed branching factors. Showing file size by area coding seemed appealing, but various rectangular, triangular, and circular strategies all had problems. Then while puzzling about this in the faculty lounge, I had the Aha! experience of splitting the screen into rectangles in alternating horizontal and vertical directions as you traverse down the levels. This recursive algorithm seemed attractive, but it took me a few days to convince myself that it would always work and to write a six line algorithm.

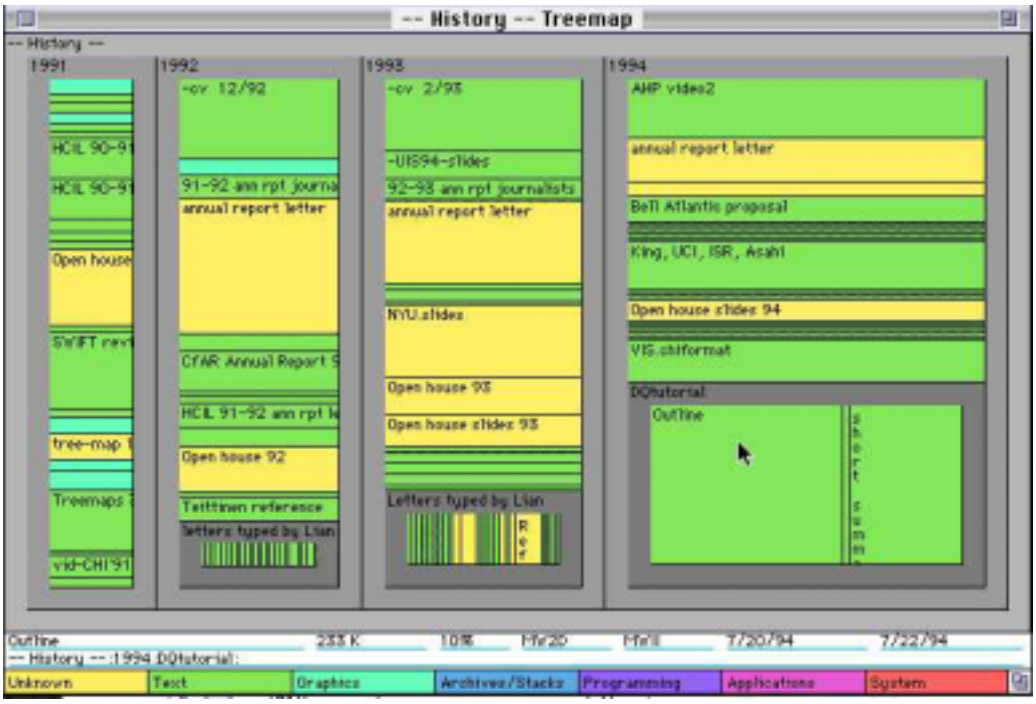

Figure 12: TreeViz display of a file storage system. Files are classified first by year, then by subject, with area proportional to file size.

In the "TreeMap" software (Johnson and Shneiderman, 1991), this idea was implemented so that each node is a rectangle whose area is proportional to some attribute such as node size (number of documents, disk space, etc.). The cartograms by Raisz (1934) are actually treemaps, rather than true mosaics.

4 Conclusions

The history of statistical graphics has deep roots (Friendly and Denis, 2001), of which this paper charts a slender tendril. This account of mosaic displays shows that this graphic form arose in Halley's diagram to depict the joint probabilities of mortality under independence, and was later used in other contexts to show the product of two quantities represented by height and width of rectangles. The general ideas of showing conditional proportions, and of recursive sub-division to show three or more factors were also introduced, at least implicitly, before the 20^{th} century.

The modern history of mosaic displays has progressed from simple plots of observed frequencies in *n*-way tables, to mosaic plots showing the lack-of-fit of a given log-linear model, to interactive systems providing visual fitting and exploration.

These graphic developments for categorical data, traditionally the poor cousin of quantitative graphics, raise interesting questions for future research, e.g., scatterplot matrices for mixtures of categorical and quantitative variables, and (bivariate) marginal vs. conditional views (Friendly, 1999). As well, we may look forward to the continued development of interactive and dynamic methods for exploration, model specification, fitting, and diagnosis with categorical data.

The development of tree maps for nested (vs. cross-classified) data structures suggests that space-filling graphic designs have wider applicability in data visualization, including clustering problems, classification and regression trees, network visualization, and so forth.

5 Acknowledgments

I am grateful to *les Chevaliers des Album de Statistique Graphique*, particularly Antoine de Falguerolles, Gilles Palsky, Ian Spence, Ruediger Ostermann, and Antony Unwin for historical background, references, and access to images used here. The École Nationale des Ponts et Chausées (ENPC) generously allowed access to their archives, and permission to reproduce Minard's "Tableau Graphique." Two anonymous reviewers helped me to strengthen the overall framework and attention to details. This work is supported by Grant OGP0138748 from the National Sciences and Engineering Research Council of Canada.

References

Beniger, J. R. and Robyn, D. L. Quantitative graphics in statistics: A brief history. *The American Statistician*, 32(1):1–11, 1978.

Bertillon, J. Fréquence des étrangers à Paris en 1891. In *Cours élementaire de statistique administrative*. Societé d'éditions scientifiques, Paris, 1896. (map).

Bertin, J. *La graphique et le traitement graphique de l'information*. Flammarion, Paris, 1977.

Bertin, J. *Graphics and Graphic Information-processing*. de Gruyter, New York, 1981. (trans. W. Berg and P. Scott).

Birch, T. W. *Maps: Topographical and Statistical*. Oxford University Press, Oxford, UK, 1964. (2nd. ed., 1976).

Cleveland, W. S. and McGill, R. Graphical perception: Theory, experimentation and application to the development of graphical methods. *Journal of the American Statistical Association*, 79:531–554, 1984.

Cleveland, W. S. and McGill, R. Graphical perception and graphical methods for analyzing scientific data. *Science*, 229:828–833, 1985.

Crome, A. F. W. Über die grösse and bevölkerung der sämtlichen europäschen staaten. Leipzig, 1785.

Emerson, J. W. Mosaic displays in S-PLUS: A general implementation and a case study. *Statistical Computing & Statistical Graphics Newsletter*, 9(1):17–23, 1998.

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

Friendly, M. Mosaic displays for loglinear models. In *ASA, Proceedings of the Statistical Graphics Section*, pp. 61–68, Alexandria, VA, 1992a.

Friendly, M. User's guide for MOSAICS. Technical Report 206, York University, Psychology Dept, 1992b. `http://www.math.yorku.ca/SCS/mosaics.html`.

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*, 8:373–395, 1999.

Friendly, M. Re-Visions of Minard. *Statistical Computing & Statistical Graphics Newsletter*, 11(1):1, 13–19, 2000a.

Friendly, M. *Visualizing Categorical Data*. SAS Institute, Cary, NC, 2000b.

Friendly, M. and Denis, D. J. Milestones in the history of thematic cartography, statistical graphics, and data visualization, 2001. `http://www.math.yorku.ca/SCS/Gallery/milestone/`.

Graunt, J. *Natural and Political Observations Mentioned in a Following Index and Made Upon the Bills of Mortality*. Martin, Allestry, and Dicas, London, 1662.

Hald, A. *A History of Probability and Statistics and their Application before 1750*. John Wiley and Sons, New York, 1990.

Halley, E. An estimate of the degrees of mortality of mankind, drawn from curious tables of the births and funerals at the city of Breslaw, with an attempt to ascertain the price of annuities on lives. *Philosophical Transactions*, 17:596–610, 1693.

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.

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

Hofmann, H. Simpson on board the Titanic? Interactive methods for dealing with multivariate categorical data. *Statistical Computing & Statistical Graphics Newsletter*, 9(2):16–19, 1998.

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

Hummel, J. Linked bar charts: Analyzing categorical data graphically. *Computational Statistics*, 11:23–33, 1996.

Johnson, B. and Shneiderman, B. Treemaps: A space-filling approach to the visualization of hierarchical information structures. In *Proc. of the 2nd International IEEE Visualization Conference*, pp. 284–291, San Diego, CA, 1991.

Karsten, K. G. *Charts and Graphs*. Prentice Hall, New York, 1925.

Lewandowsky, S. and Spence, I. The perception of statistical graphs. *Sociological Methods & Research*, 18: 200–242, 1989.

Mayr, G. v. *Die Gesetzmäßigkeit im Gesellschaftsleben*. Oldenbourg, 1877.

McClelland, G. H. *Seeing Statistics*. Duxbury, Pacific Grove, CA, 1999.

Minard, C. J. Importance du parcours partiel sur les chemins de fer, 1842. ENPC: 5773/C338.

Minard, C. J. Tableaux figuratifs de la circulation de quelques chemins de fer. lith. (n.s.), May 1844. ENPC: 5860/C351, 5299/C307.

Ministère des Travaux Publics. *Album de Statistique Graphique*. Imprimerie Nationale, Paris, 1879–1899.

Ostermann, R. Georg von Mayrs beiträge zur statistischen graphik. *Algemeines Statistisches Archiv*, 83(3): 350–362, 1999.

Palsky, G. *Des Chiffres et des Cartes: Naissance et développement de la Cartographie Quantitative Français au XIX^e siècle*. CTHS, Paris, 1996.

Raisz, E. The rectangular statistical cartogram. *Geographical Review*, 24:292–296, 1934.

Riedwyl, H. and Schüpbach, M. Siebdiagramme: Graphische darstellung von kontingenztafeln. Technical Report 12, Institute for Mathematical Statistics, University of Bern, Bern, Switzerland., 1983.

Riedwyl, H. and Schüpbach, M. Parquet diagram to plot contingency tables. In Faulbaum, F., editor, *Softstat '93: Advances In Statistical Software*, pp. 293–299. Gustav Fischer, New York, 1994.

SAS Institute, Inc. JMP: Statistical discovery software, 2000.

Shneiderman, B. Tree visualization with treemaps: A 2-D space-filling approach. Technical Report TR 91-03, University of Maryland, HCIL, 1991. (Published in *ACM Transactions on Graphics*, vol. 11(1): 92–99, 1992).

Simkin, D. and Hastie, R. An information-processing analysis of graph perception. *Journal of the American Statistical Association*, 82:454–465, 1987.

Snee, R. D. Graphical display of two-way contingency tables. *The American Statistician*, 28:9–12, 1974.

Spence, R. *Information Visualization*. Pearson Education Ltd, Essex, England, 2001.

Statistischen Bureau. *Graphisch-statistischer Atlas der Schweiz (Atlas Graphique et Statistique de la Suisse)*. Buchdruckeri Stämpfli & Cie, Departments des Innern, Bern, 1897.

Theus, M. Visualization of categorical data. In *Advances in Statistical Software*, volume 6, pp. 47–55. Lucius & Lucius, 1997.

Theus, M. and Lauer, S. R. W. Visualizing loglinear models. *Journal of Computational and Graphical Statistics*, 8(3):396–412, 1999.

Unwin, A., Hawkins, G., Hoffman, H., and Siegl, B. Interactive graphics for data sets with missing values—MANET. *Journal of Computational and Graphical Statistics*, 5(2):113–122, 1996.

Valero, P., Young, F., and Friendly, M. Visual log-linear analysis in ViSta. Technical report, University of Valencia, 2001. submitted, *Computational Statistics and Data Analysis*.

Wang, C. M. Applications and computing of mosaics. *Computational Statistics & Data Analysis*, 3:89–97, 1985.

Young, F. W. and Bann, C. M. ViSta: A visual statistics system. In Stine, R. A. and Fox, J., editors, *Statistical Computing Environments for Social Research*, pp. 207–236. Sage, 1996.

Young, F. W., Valero, P. M., and Ledesma, R. D. Visualizing categorical data in ViSta. In Oñate, E., García-Sicilia, and Ramallo, L., editors, *Métodos Numéricos en Ciencices Sociales*, pp. 196–207, Barcelona, Spain, 2000. Centro Internacional de Metodos Numericos En Ingenieria.