

Framed! Reproducing and Revisiting 150 year old charts

Susan VanderPlas

CSAFE, Iowa State University

and

Ryan C Goluch

Department of Electrical & Computer Engineering, Iowa State University
and

Heike Hofmann

Department of Statistics, Iowa State University

October 16, 2018

Abstract

The Statistical Atlases published by the Census Bureau in the late 1800s utilized a number of novel methods for displaying data. In this paper, we examine the use of framed spine and mosaic plots used in two plates of the Statistical Atlas of 1870. We use forensic statistics to recreate the data using available census information, and then use that data to create framed charts using modern plotting methods. We then examine the effectiveness of the framed charts compared to other alternatives with a user study. The data and code for this study are available online.

Keywords: Census Bureau, mosaic plots, pie charts, spine plots, Statistical Atlas, statistical graphics

1 Introduction

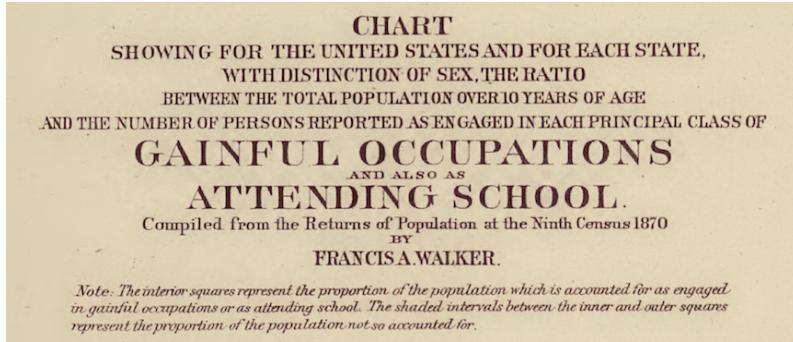
Between 1870 and 1920, the US Census Bureau published a Statistical Atlas to map the state of the Union based on data collected in each census¹. Each of these atlases represents a masterpiece

¹A Census Atlas of the United States was published by the Census Bureau in 2007 [Suchan et al., 2007].

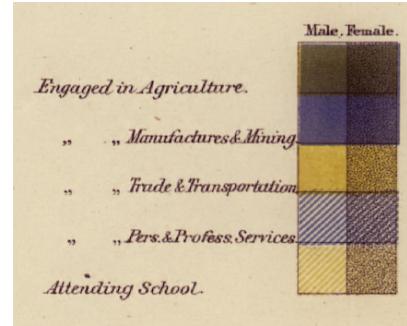
in science and technology. The ninth census, supervised by Francis A. Walker, assessed a country of about 38.5 million people. The 1870 Statistical Atlas is a graphical compendium of census information laid out over more than 100 pages, and corresponds to the ninth census. Most of the illustrations are overlaid maps (38 of 107 pages), but some consist of more abstract and, at that time, novel visualizations [Kostelnick, 2004], including age pyramids (bilateral frequency charts) and ring charts to show population growth. This paper addresses two types of charts from the 1870 Statistical Atlas which were presented as small multiples; that is, a separate chart was shown for each state and territory.

The charts in the 1870 Statistical Atlas were printed by Julius Bien's publishing house [Cartography Associates, 2009, Walker, 1875] using lithography. This process involved creating separate plates for each color utilized in the chart by hand, and then lining up each color precisely when the images were printed. Modern methods are much quicker and easier on the visualization designer; we only have to write computer code to describe the plot, and the computer renders the plot in a minuscule fraction of the time compared to what it would take to draw the same plot by hand.

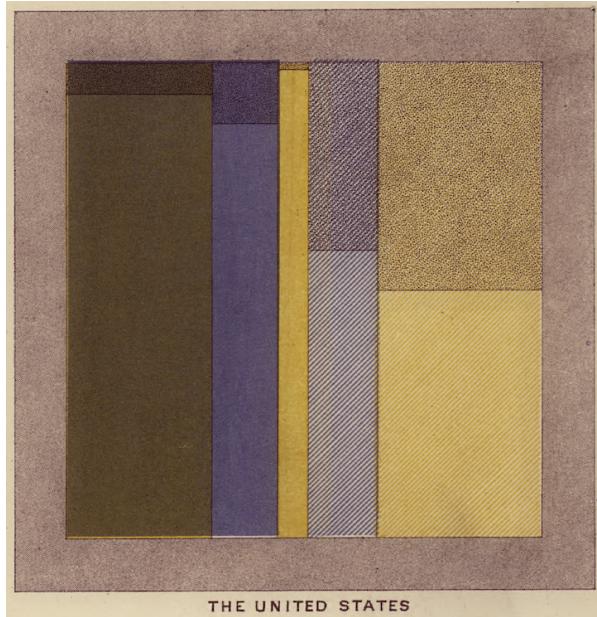
In this paper, we describe the methodology used to re-create plots from the 1870 Statistical Atlas, first by recreating the plots as they were presented, and then re-imagining the plots to provide a more visually optimal display that is similar to the original. We assess the accuracy of judgments made from these plots with a user study, and discuss the implications of the study on modern plot design.



(a) Description section of plate #32.



(b) Legend to plate #32.



(c) US wide distribution of genders across occupations.



(d) Miniature of plate #32 of the Statistical Atlas.

Figure 1: Overview of plate #32, with zoom-ins to different components.

2 Statistical Archaeology

Plates 31 (Figure 2) and 32 (Figure 1d) of the Statistical Atlas are very similar in their setup and structure: at the top of the chart we find an in-depth description (see the legend for plate #32 in Figure 1a) and a legend detailing the color choices (see Figure 1b). In the top left corner, there is an US-wide aggregate of the situation (shown in Figure 1c), and a series of small multiples [Tufte,

1991] or trellis plots [Becker et al., 1996], one for each state. In Plate 32, the plots are what are now known as mosaic [Hartigan and Kleiner, 1981, Friendly, 2002, Wickham and Hofmann, 2011] or Marimekko plots [Ribecca, 2018]. In 1874 these were a novel type of chart; Georg von Mayr, who is sometimes credited with publishing the first mosaic plot [see Friendly, 2002], didn't publish his form of mosaic plots until 1877.

Plate 31 (see Figure 2) uses spine plots²[Hummel, 1996] to give an overview of the percentage of religious sittings³ by denomination for each state (colored stripes in the square). Plate 32 (see 1d) shows the gender ratio of the population over the age of ten in different types of occupations using a mosaic plot.

In both plates, a grey band is drawn around each one of the states' squares representing essentially missing information. In plate 31, the grey band represents the size of religiously unaccommodated population over the age of ten. In plate 32, the area of the grey band is proportional to the population “unaccounted” for, i.e. the difference between the total population over the age of ten and the population gainfully employed in one of five categories or attending school.

In order to reproduce these charts using modern plotting methods, some statistical archaeology is necessary – we must reconstruct the data underlying the charts, using sources available nearly 150 years later. The National Historical Geographic Information System (NHGIS) [Minnesota Population Center, 2016] provides a wealth of tabulated data, and most of the data needed to reproduce plate #32 is available directly from table NT13 on “Employed Population by Occupation by Age by Sex” from the 1870 Census: Religious Bodies, Occupation & Government Data

²Spine plots are horizontal stacked bar charts; the width of the bar is proportional to the percent of the population represented by the category. They can be seen as one-dimensional mosaic plots with the initial split on the x-axis.

³‘Sittings’ is the term used in the Atlas to denote the number of individuals accommodated by a church. Presumably, this corresponds to the number of physical seats available.



Figure 2: Plate #31 from the Statistical Atlas of 1870: church accommodation by state and denomination for the population of over ten years of age.

(1870_sROG). Unfortunately, the size of the unemployed population is not published with this information. To get an estimate for the size of the unemployed population in each state, we make

use of the 1% microsample of the 1870 census provided by the Integrated Public Use Microdata Series (IPUMS-USA) through the Minnesota Population Center [Ruggles et al., 2015].

Using the microsample, we obtain counts of the male and female population above ten as well as state population numbers, which allows us to get estimates of the state-level population totals for individuals above the age of ten for each state by extrapolating from the sample proportions. This information is then used to calculate the size of the unaccommodated populations in each set of charts – either those who are not gainfully employed (plate #32) or those who are not accommodated religiously (plate #31):

Let $n_U^{(X)}$ be the –unfortunately unknown– size of the unaccounted population of state X . Then we have the relationship

$$n_U^{(X)} = n_{\text{age} \geq 10}^{(X)} - n_A^{(X)}, \quad (1)$$

where $n_{\text{age} \geq 10}^{(X)}$ is the number of over ten year olds in the state, and $n_A^{(X)}$ is the size of the population accounted for (i.e. gainfully employed or going to school, or those who are religiously accommodated). $n_A^{(X)}$ is known for each state from table NT13. We get an estimate for the unknown number of over ten year olds in each state $n_{\text{age} \geq 10}^{(X)}$ from:

$$n_{\text{age} \geq 10}^{(X)} = n_{\text{total}}^{(X)} \cdot p_{\text{age} \geq 10}^{(X)}.$$

where $p_{\text{age} \geq 10}^{(X)} \in [0, 1]$ is the proportion of the population in state X who is over ten years of age. While this proportion is also not known, we can estimate it from the 1% microsample as the ratio

of individuals over ten and the total number of individuals in the state.

$$\hat{p}_{\text{age} \geq 10}^{(X)} = \frac{\# \text{ individuals age ten and over in state } X}{\# \text{ individuals in state } X}.$$

Piecing this result back into (1), we get both estimates for the size of the unaccounted population as well as a standard error for it:

$$\hat{n}_U^{(X)} = n_{\text{total}}^{(X)} \cdot \hat{p}_{\text{age} \geq 10}^{(X)} - n_A^{(X)}, \quad (2)$$

$$SE(\hat{n}_U^{(X)}) = \sqrt{\hat{p}_{\text{age} \geq 10}^{(X)} \cdot (1 - \hat{p}_{\text{age} \geq 10}^{(X)}) n_{\text{total}}^{(X)}}. \quad (3)$$

Estimates for the number of unaccounted women and men over ten years of age in each state can be achieved similarly. We also use these results to get estimates and standard errors for the size of the religiously unaccommodated population over ten years of age, used to recreate plate 31.

2.1 Plate 32: Gender Ratio in Agriculture, Trade, Service, Manufacturing, and Schools

Figure 1d shows a miniature of the chart published as plate #32 in the Statistical Atlas of 1870 [Walker, 1874] produced from data collected in the 9th US Census. The chart shows the gender ratio of population over the age of ten in different types of occupations.

With the help of the description (Figure 1a) and the legend (Figure 1b), we can interpret the details of each of the squares in Figure 1c. This figure shows an overview of type of occupation by gender across the US in 1870. The percentage of population in a particular type of occupation is represented by the width of the rectangles; the height is proportional to the percentage of gender,



Figure 3: Nation-wide overview of gender ratio by occupation based on gainfully employed population over ten.

with men shown in the bottom rectangle and women in the top rectangle. The area of the grey band around each one of the states' squares is proportional to the size of the “unaccounted” population, i.e. the difference between the total population over the age of ten and the population gainfully employed in one of the five categories or attending school.

Showing the unaccounted population as a frame masks the size of the corresponding population by visually cutting it into quarters. The percentage of unaccounted individuals nation-wide is about 30%. This number is higher than any of the other groups. Assessing the size of the unaccounted population in form of the area of the grey frame is mentally strenuous, as discussed in section 3. The task becomes much easier if we incorporate the information directly into the mosaic plot, as shown in Figure 3b. This figure reveals another previously hidden finding: about 97% of the unaccounted population are female!

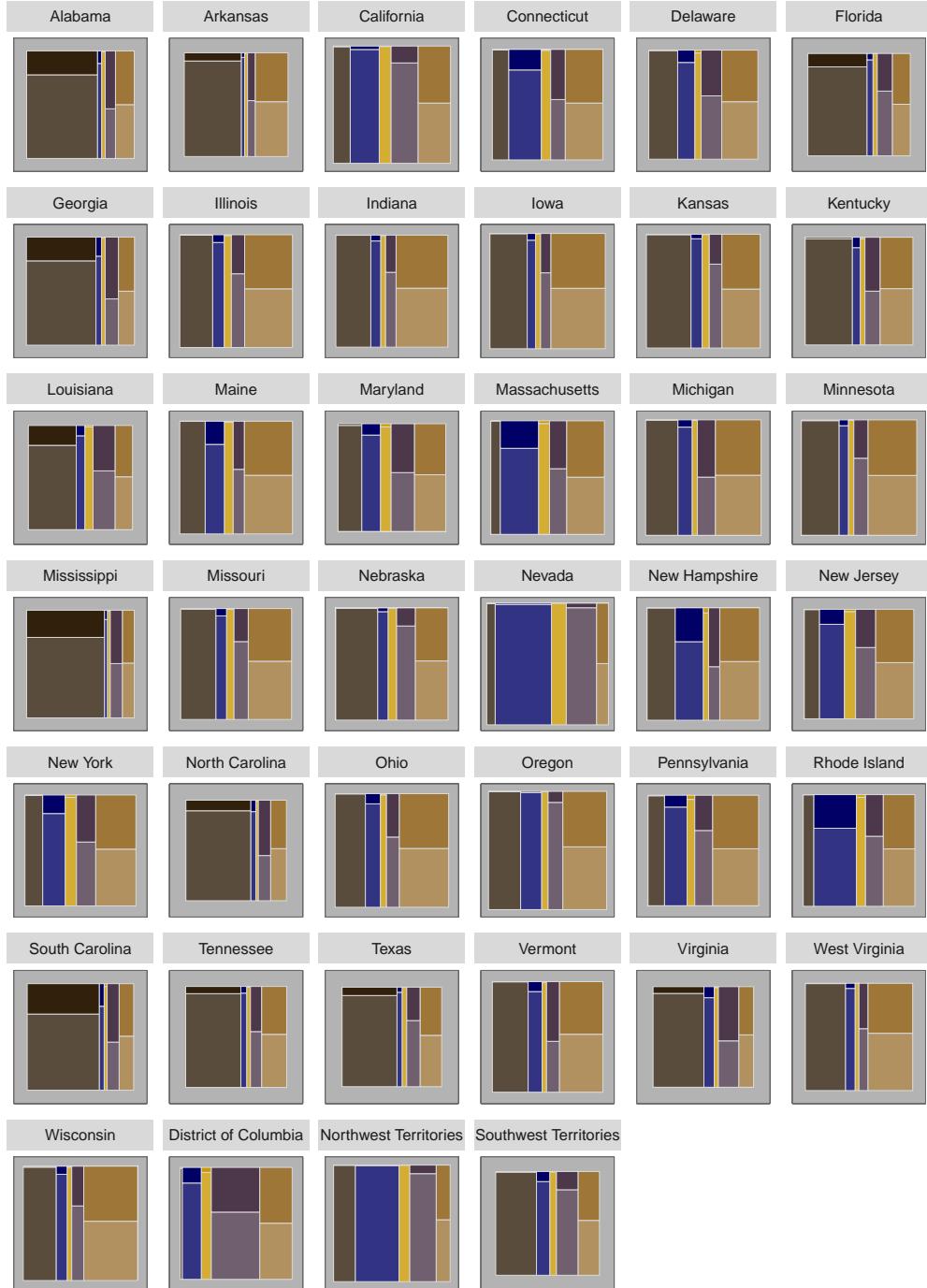


Figure 4: Reproduction of plate #32.

Note that using our estimation method outlined in Equation 2 and Equation 3, we are able to get estimates for the size of the unaccounted population by gender. This information is not shown on plate #32, but must have been available to the creators of the chart at the time.

Clearly the gender breakdown here is interesting: when this information is included as a bar in the reconstructed plot shown in Figure 3, it becomes clear that almost all of this unaccounted population is female. This design decision may have been made because of the perceived aesthetic appeal of the unaccounted border region (which makes showing the gender of unaccounted persons difficult), but it may also have been made because the unpaid work by women in the home was not visible (those persons who kept house for other individuals would have been included in the personal and professional services category). This logic is still present today: women who are homemakers are not included in the labor force participation rate, and thus their contributions to the overall economy are not typically counted [Bureau of Labor Statistics, 2015].

Visual inspection tells us that the data used to render the nation-wide overview in Figure 3 and the state-level aggregates in Figure 4 is “the right one”, i.e. the figures are similar to the charts shown in plate #32. A more precise evaluation of the similarity between the old and the new figures can be achieved by using pixel measurements of the high-resolution digitized image of plate #32 and directly compare the percentages for each of the occupations.

Figure 5 shows this comparison: for each of the occupations we draw a scatterplot of the percentages calculated from the pixel measurements (x -axis) and compare them to the percentages obtained from NT13 and the 1% microsample (y -axis). We can see in Figure 5 that the numbers closely match; the regions in blue are based on point-wise Agresti-Coull 95% confidence intervals [Agresti and Coull, 1998]. None of the points are outside these confidence bands. For all occupation levels and school attendance the numbers are *very* close. For population not accounted for, the numbers are estimated from the 1% microsample. This increases the variability of the estimates, but the relationship to the pixel measurements is still very strong.

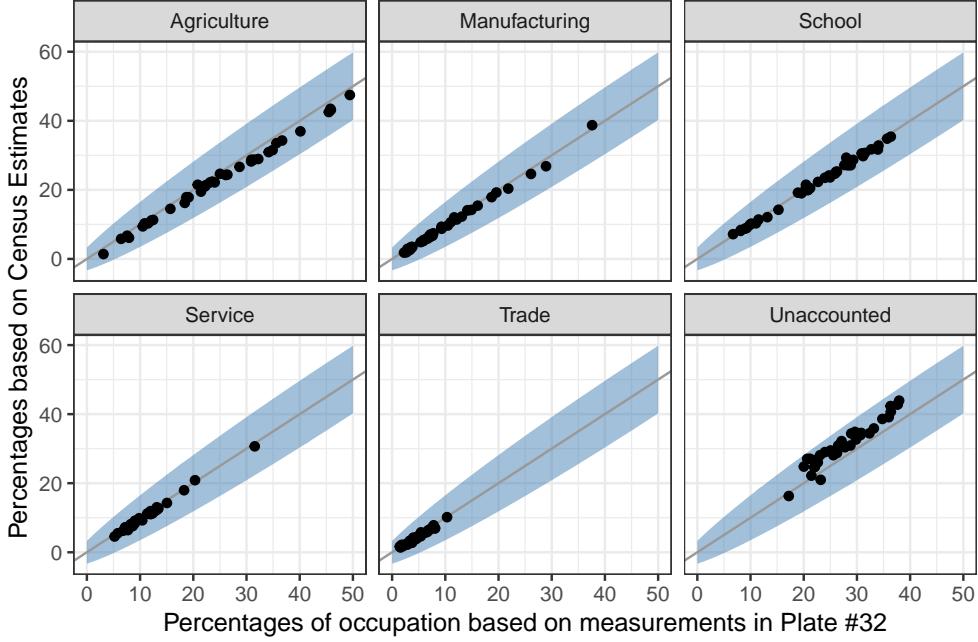


Figure 5: Differences/similarities between numbers from the Statistical Atlas and the modern reproduction.

2.2 Plate 31: Church Accommodations by State

Figure 2 shows a miniature of plate #31 from the 1870 Statistical Atlas. This plate shows the percentage of religious sittings by denomination for each state (colored stripes in the square) as well as the percentage of unaccommodated population over the age of ten (area of the grey outer frame). For each state a square of the same size is drawn. The four most common denominations within each state are shown as colored stripes, the width of each is proportional to the number of people accommodated. Each denomination is shown by one of eleven different colors, all other denominations are represented jointly by a twelfth color. The color scheme chosen in the Statistical Atlas is essentially that of paired colors, i.e. each hue is represented with a lighter shade (using hatching) and a darker shade (see zoom-in to the legend of plate #31 in Figure 6a).

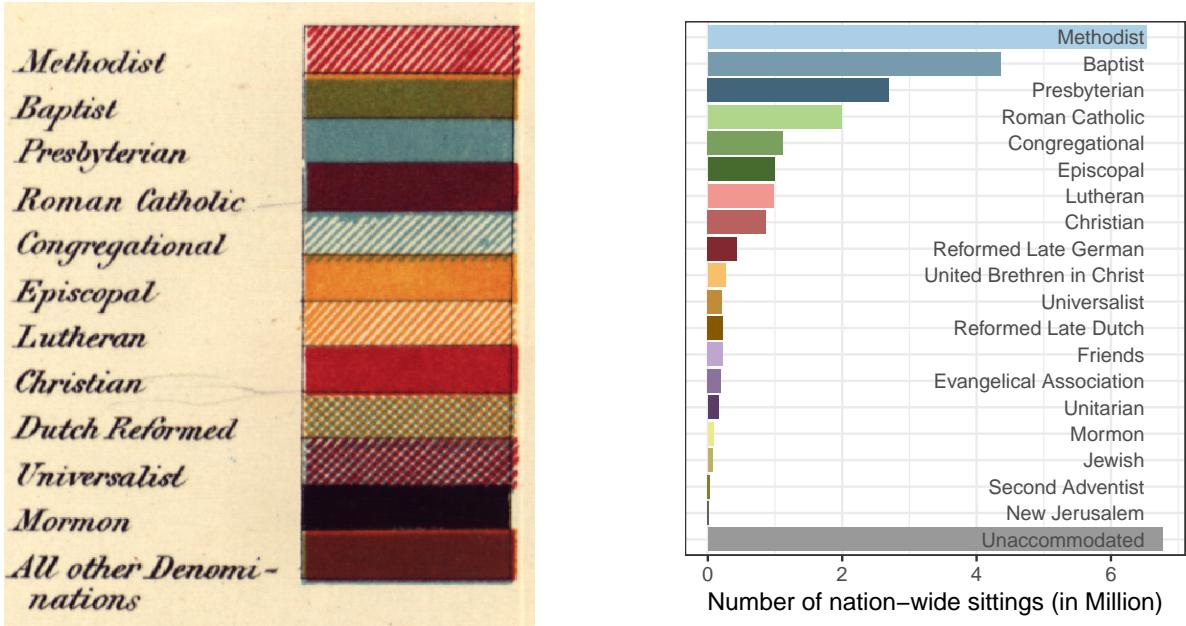
The inside squares of plate #31 are simpler in structure than those of plate #32 – they are one-dimensional mosaic plots, also known as rotated stacked bar charts or spine plots [Hummel,

1996]. As in the previous example, a grey frame is drawn with an area proportional to the size of the unaccommodated population over the age of ten. Additionally, other design choices were made which impact the readability of the charts:

Top four only: A decision was made to only show the top four denominations in each state and the top eight denominations in the United States. As a result, some denominations are shown which have lower overall accommodation, while other denominations are excluded even though they accommodate more people. For example, in 1870 the Reformed Late Germans accommodated 431,700 people nation-wide; the 10th highest number, as shown in Figure 6b. However, they are not mentioned in plate #32. The Mormon Church is locally (Utah Territory) strong enough to be represented, even though it accommodates only 87,838 people (17th place nationally) and is far behind the German Reformed Church. This decision was likely due in part to technical constraints: it would be very hard to align very thin bars with even high-precision lithography, as each color layer of each chart must be drawn by hand and carefully aligned. Limiting the choices to five total internal bars ensured that it was generally possible to render the image and match the colors to the denominations listed in the legend.

Re-ordering: Within each state denominations are ordered from largest to smallest. This makes comparisons of the state-wide religious makeup between states more difficult, as it re-orders the colors in the bar code-like color strips representing each state, thus small differences in numbers might result in rather large visual differences.

We initially computed the unaccommodated population for each state as described in equation (2), but found that there was a clear difference in the percentage reported by each state in the



(a) Legend section of plate #31.

(b) Church sittings by denomination.

Figure 6: Old and new side-by-side: denominations are ordered from top to bottom according to the nation-wide number of sittings. On the right, all denominations are shown that reported accommodations in 1870. We also see that the number of unaccommodated people over the age of ten is higher than the number accommodated by any single denomination.

Statistical Atlas plots and the proportion estimated by using the 1% microsample; this difference was present for every state and territory. This difference was resolved by considering only the population above age 11 when computing estimates for plate #31. There is some ambiguity in the phrase “population over 10 years of age”, as an individual attains 10 years of age on their 10th birthday (that is, $x \geq 10$), but in any other context, “over 10” would indicate $x > 10$. In plate #32, the interpretation of $x \geq 10$ is used, but in plate #31, it appears that the interpretation $x > 10$ or $x \geq 11$ is used instead. This modification to the calculations resolves the difference in estimates (shown in Figure 7) for all states except Minnesota, which we discuss below.

A close examination of plate #31 will also reveal that the different color layers of the lithographic prints are not as well aligned as in plate #32. In addition, the state plots are much more

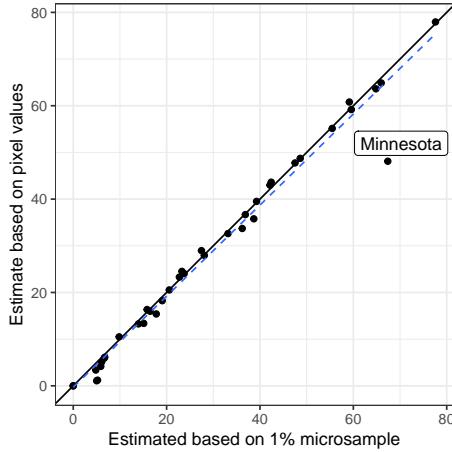


Figure 7: Percentage of state-wide unaccommodated population based on estimates from the 1% Microsample (x -axis) and pixel values measured from the digitized version of plate #31 (y -axis).

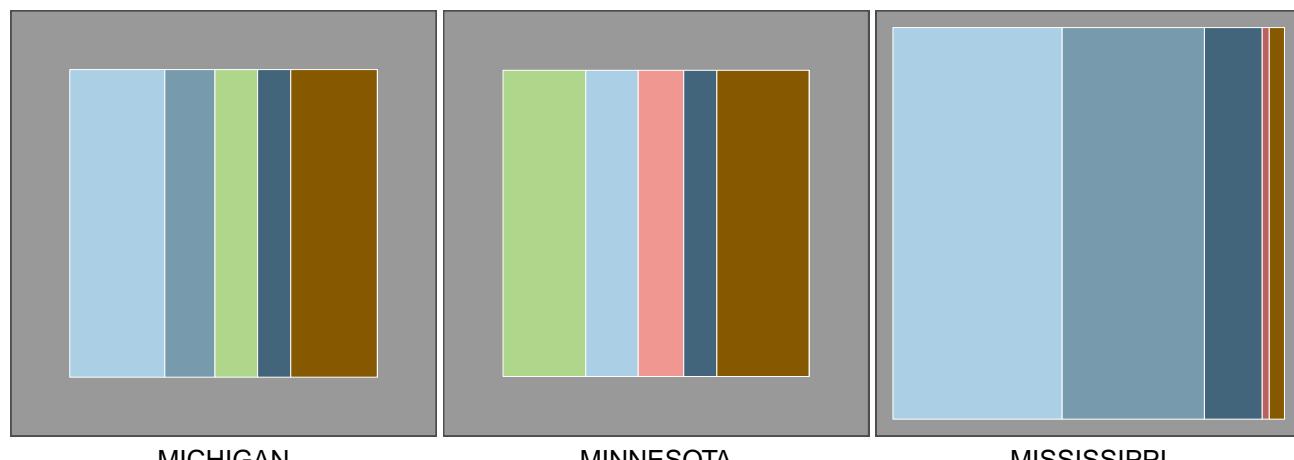
variable in size in plate #31 than in plate #32. These differences suggest that plate #31 might have been created by a less experienced lithographer than plate #32. Both plate #31 and plate #32 are based on data compiled from the 1870 census by Francis A. Walker, but it appears the artistry and methodology may not have been consistent between the two plates.

When we compare the proportions of state populations shown in the Statistical Atlas and the proportions of different denominations in the 1% microsample data another interesting anomaly emerges. The proportion of unaccommodated individuals in Minnesota is much higher in the Statistical Atlas plot than it is in the 1% microsample, while the proportions are comparable for every other state. Figure 7 shows the relationship between the 1% microsample population and the pixel proportions in plate #31. The most likely explanation for this large discrepancy is that the lithographer made an error when creating the state sub-plot for Minnesota.

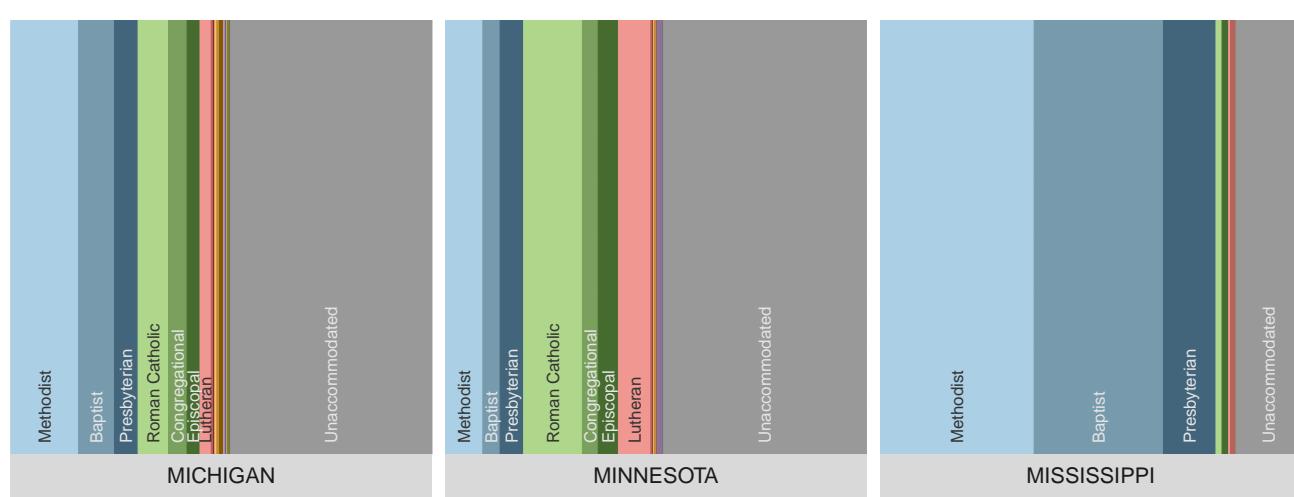
Figure 8a shows three sub plots from plate #31 of the 1870 Statistical Atlas; Figure 8b shows a modern replication of the original charts, and Figure 8c shows an improved version of the same plots. We have improved the charts by adding an additional bar for the proportion of



(a) Statistical Atlas Panels



(b) State religious accommodation plots generated with the `ggplot2` package [Wickham, 2016b].



(c) Re-imagined religious accommodation plots.

Figure 8: Panels from Michigan, Minnesota, and Mississippi, in original, modern, and reimaged form.

the population who are unaccommodated, and we have included all denominations which were part of the 1870 census. These changes produce much more complicated plots, but allow us to notice that Mississippi is much more religiously homogeneous than Michigan or Minnesota, and also that Mississippi has significantly fewer unaccommodated persons. We can also see that Minnesota has a higher proportion of Lutheran church accommodations than Michigan, but the same denominations seem to be present in both states. These details were much less obvious on the original plots due to the choice to show only 4 denominations surrounded by a rectangle representing unaccommodated persons.

Both sets of re-imagined church accommodation plots use a triple color scheme. The lightest colors were drawn from the Colorbrewer [Harrower and Brewer, 2003] “Paired” palette’s light colors (a grey hue was also added); each color was then darkened twice using the `munsell` package [Wickham, 2016a] to produce a color scheme composed of 21 colors and 7 hues. This scheme is similar to the color scheme in the Statistical Atlas, which is composed of 7 hues and uses crosshatching to create lighter and darker bars. A one-to-one mapping of the hues is not feasible, however, as the re-imagined version of the plots shows all denominations reporting accommodations, rather than the top 8 denominations nationwide.

3 Perception of Framed Plots

While the use of framed plots in plates 31 and 32 of the Statistical Atlas is aesthetically pleasing, it causes several problems:

1. The frame cuts the unaccommodated population into quarters, making it difficult to estimate the size of the unaccommodated population. Mosaic and spine plots require the viewer

to make judgments along the x -axis, but this is difficult to do for the unaccommodated population, as the portion of the x -axis devoted to an unaccommodated population of p is $1 - \sqrt{p}$.

2. The frame makes it difficult to compare between states. Judgments comparing two states using a framed chart require the viewer to judge area, rather than length, and also break up any common scale between two states, as the inside portion of the plots are not aligned between states. Even if the viewer (incorrectly) uses length as a cue, they are comparing length on an un-aligned scale, which increases the difficulty level [Cleveland and McGill, 1984].
3. Numerical estimation in framed plots requires a two-stage estimation procedure. In order to estimate the proportion of the chart attributed to one occupation or denomination, one must first estimate the proportion of the overall square dedicated to the internal chart, the proportion of the internal chart dedicated to the category of interest, and finally multiply these two quantities together to obtain the overall estimate. In comparison, estimating the proportion of an unframed spine plot dedicated to one category requires only that the viewer estimate the width of the rectangle.
4. The frame decreases the perceptual salience of the dimension of interest. Information is primarily conveyed on the x -axis in both spline and mosaic plots, though secondary information is conveyed through the y axis in mosaic plots. In unframed plots, there is variation along the x -axis in color and in width, with additional unaligned variation along the y -axis in mosaic plots. The frame disrupts this by disconnecting the x axis and the variation in

bar width, increasing the perceptual effort required to map numerical information to the visual depiction. Carpenter and Shah [1998] discuss the effects of increased cognitive load resulting from additional complexity among the dimension of interest.

5. The frame cannot easily be divided into male and female portions, which means that information is lost when considering gender differences, as shown in plate 32.

Bias in area estimation and comparison has been well studied. Krider et al. [2001] showed that viewers estimate area based on the most salient dimension of an object, rather than on its' total area. Piaget [1968] showed this bias is stronger in young children, but that the bias decreases with age and cognitive development. Teghtsoonian [1965] examined the perceived size of squares, parallelograms, and other shapes, and found that area estimation was consistent with the psychophysics equation $A_{est} \propto A_{act}^\beta$, with a scaling coefficient of around $\beta = 0.8$. That is, perceived area increases more slowly than the actual area of the estimated region. Lewandowsky and Spence [1989] found that area judgments are approximately accurate in certain conditions: namely, when the variation in size occurs along a single dimension, and other dimensions are aligned. Taken together, these studies suggest that the use of framed charts can be expected to produce biased estimates, and that these estimates should be considerably worse for the mosaic plots than for spine plots, as information in mosaic plots is represented in two dimensions. In addition, Lewandowsky and Spence [1989] suggests that we can reduce these biases by integrating the information about unaccommodated individuals into the main part of the plot, treating it as just another category to be displayed.

In order to quantify the perceptual biases induced by framed charts, we designed a study which compares the framed and unframed versions of mosaic plots, spine plots, and pie charts, using

data reconstructed from the Statistical Atlas charts.

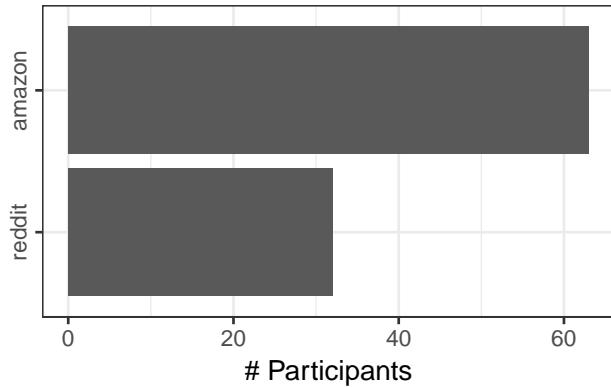
4 Frames in Practice

4.1 Study Design

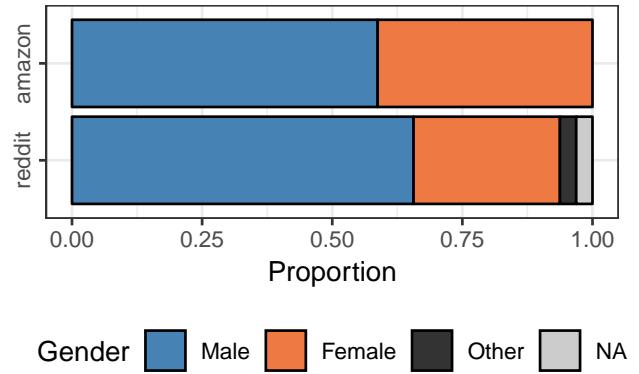
Participants A total of $n = 98$ participants were recruited using reddit ($n = 32$) and Amazon Mechanical Turk ($n = 63$). Participants were required to be over 18. We collected information about their age range, gender, and education levels. Compared to demographics of the general US population [as published in United States Census Bureau, 2016] we see in Figure 13 that survey participants are generally younger and more highly educated. This effect is even more pronounced among participants recruited from reddit than amazon turkers. In addition, redditors are slightly more likely to identify as male or non-binary gender. Reddit participants were recruited from two subreddits which are data and visualization related, and therefore might be expected to be more educated and proficient at reading charts and graphs than the general population; they were also volunteers and did not receive any remuneration, whereas Amazon Mechanical Turk users received some minimal compensation for their participation.

Survey Composition Each participant was asked to estimate the area of a marked region of 18 plots (details below). For each estimation, participants had to answer within 25 seconds before the survey automatically advanced to the next question. This ensured that individuals did not measure sections of the plot or otherwise calculate their answers.

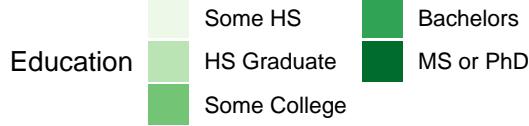
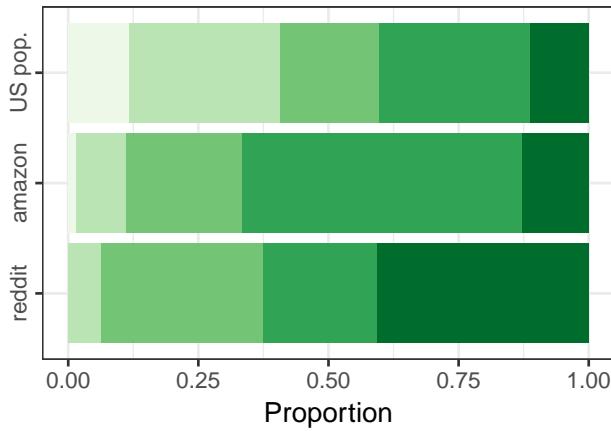
Three plot types were used in this study – mosaic plots (as in Figure 1d), spine plots (as in Figure 2), and pie charts (using data from Figure 2). Each plot type was rendered in a



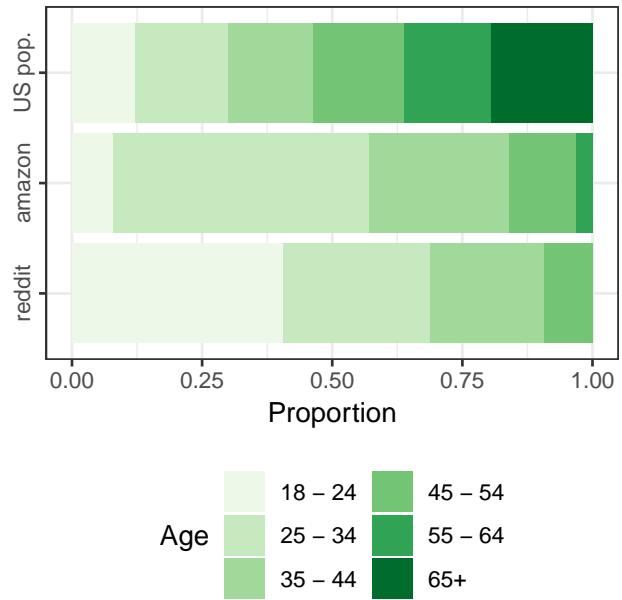
(a) Number of participants by source.



(b) Gender distribution.



(c) Educational background.



(d) Age distribution.

Figure 9: Demographics of participants by source (amazon or reddit) compared to the general US population.

framed version and in an unframed version. For each plot type, three states were chosen as “representative” states (representative in the sense that the three selected states provided an example of small, medium, and large missing/unaccounted populations), to provide replication⁴. Finally, three sections of each plot were selected for estimation – a small, medium, and large

⁴Pie charts: Alabama, Illinois, Indiana; Mosaic plots: California, Maine, Ohio; Spine plots: Connecticut, Michigan, Oregon

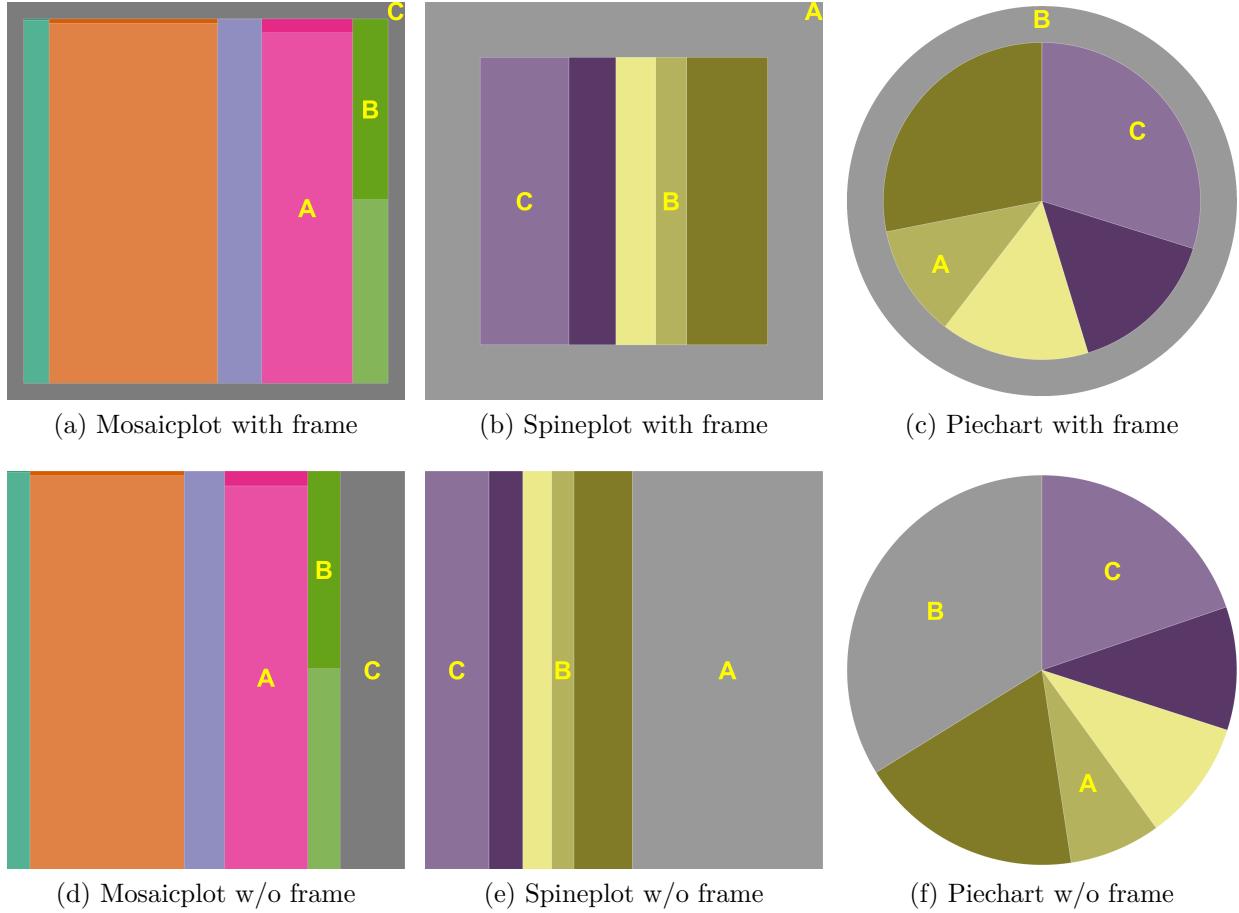


Figure 10: Overview of the six plots shown to one third of the participants: all six combinations of plot types and designs are covered. The yellow ‘A’, ‘B’, and ‘C’s mark the areas that participants are asked to estimate. Three areas in each of the plots were selected for estimation.

section. Thus, a total of 54 images were used in the study. These images were assembled into three sets of 18 plots, and each participant evaluated a single set of plots.

Each set of 18 plots was chosen such that all combinations of framed and unframed plot types were included. For each plot type \times frame combination, the participant evaluated a small, medium, and large region in the same plot. Thus, if a participant evaluated the framed pie chart for Illinois, they might evaluate the unframed pie chart for Alabama. At the end of the evaluation, a participant saw plots from six states, corresponding to each combination of plot type and frame condition.

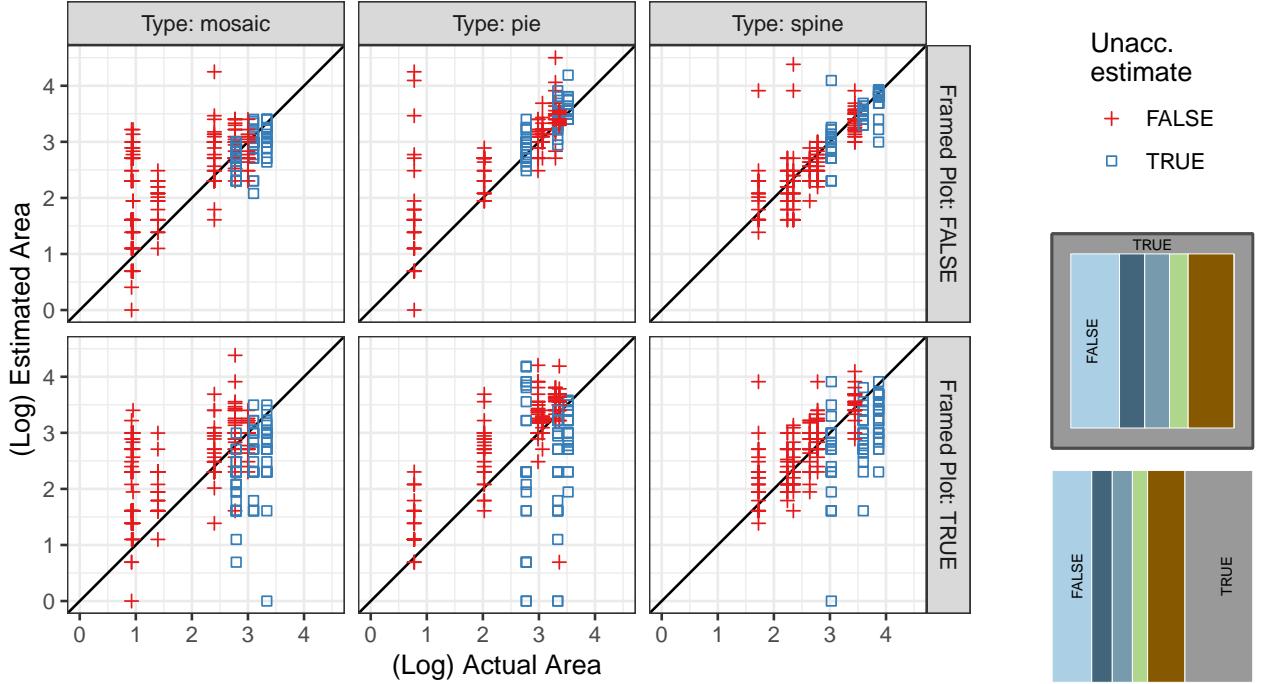


Figure 11: Scatterplots of actual versus estimated areas. The color and shape of points distinguish between areas that represent unaccommodated or unaccounted populations (abbreviated Unacc. in the chart). In framed plots, these areas are part of the frame; in unframed plots, these areas are part of the main portion of the plot.

The survey was administered in form of a branched Qualtrics survey, where each set constituted one of the branches. The survey software randomly assigned one of the branches to a participant (while balancing out participants across branches).

4.2 Results

Figure 11 shows scatterplots of estimated area size versus actual size of the area for each of the plot types. Plots in the top row are designs in which the frame piece is incorporated in the design, the bottom row are plots with a frame. Color and shape distinguish between pieces on the inside of a plot and its frame. What we see is that there does not seem to be a difference between frame and inside pieces in designs without a frame. We also see that in framed designs, inside pieces

have a tendency to be overestimated, while the size of the frame is underestimated.

There are also quite a few very extreme differences between estimated and actual size of areas apparent, particularly for small areas in mosaic plots and un-framed pie charts. These extreme differences may have resulted from typos, interrupted answers (where the question advances before the participant has finished typing their answer), or participant confusion. In the most extreme example, a participant estimated the size of a pie slice to be 70%, when it was in fact less than 3% of the chart's area. In addition, several participants entered an estimate of zero for a plot area, which would also indicate a lack of understanding of some portion of the task. Estimates which are more than five times above or below the actual magnitude have been excluded from further analysis in this study (66 out of a total of 1652 evaluations). The specific value of 5x as a cutoff is somewhat arbitrary - we identified clear errors (e.g. 0 as an estimate, a response of 60% for an actual area of about 10%, etc.) and used a cutoff value which excluded those obvious errors but which preserves much of the rest of the data. We cannot determine why these estimates were so far off of the actual, but it is likely that there was some fundamental issue affecting the participant responses which is not related to the psychophysical effects of area estimation.

Because estimates of pieces in unframed designs corresponding to the frame of framed designs do not appear different from pieces inside, we only distinguish between three types of pieces: pieces of an unframed plot, pieces of the frame in a framed plot and pieces within the plot of a framed plot. As an example, in Figure 10, region C of subplot (a) would be a framed plot frame piece, region B of subplot (a) would be a frame plot inside piece, and regions B and C of subplot (d) would be the corresponding unframed pieces.

As a measure of the error we compare the log estimated size of a portion of the chart to

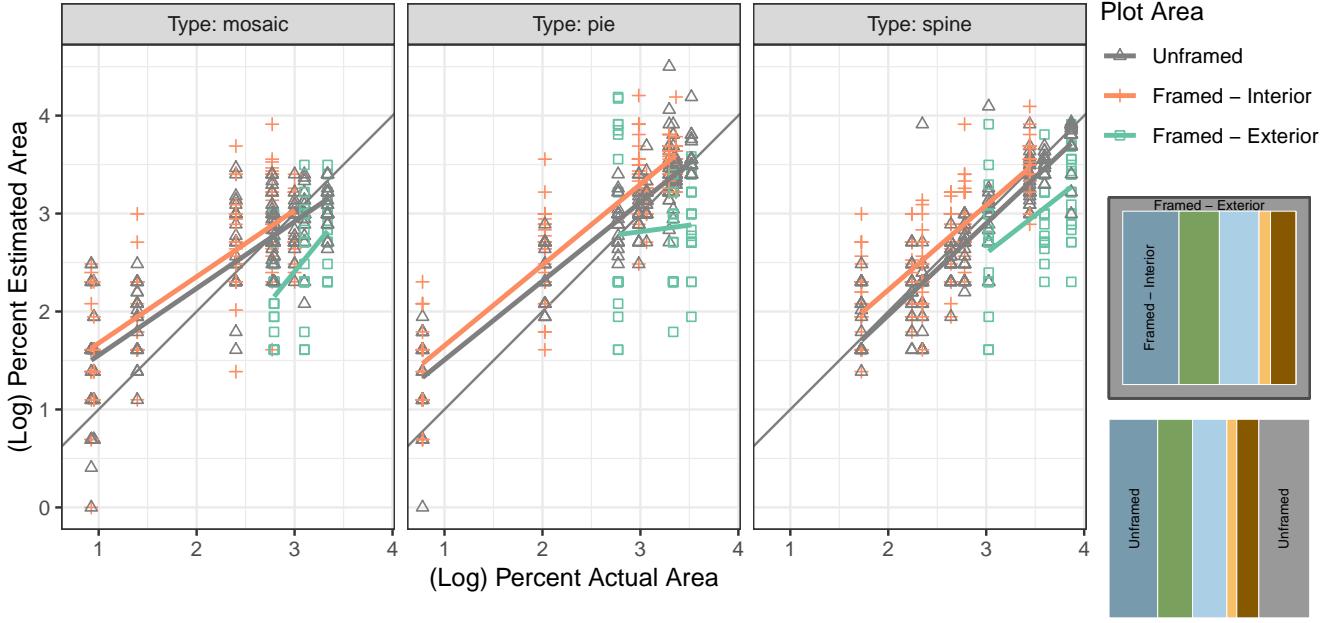


Figure 12: Scatterplot of difference between actual and observed percentages plotted against actual percentages. Colors and shapes show different types of areas, as labeled in the sample figures below the legend. The lines correspond to simple linear regressions of the form y versus x for each type of area within plot types.

the log actual size of the portion of the chart. This approach has the advantage of relating to the psychophysics equation used in Teghtsoonian [1965], $A_{est} \propto A_{act}^\beta$, where β has historically been estimated to be between 0.6 and 0.9, depending on the task. We add a constant to this proportional relationship to create an equation which can be fit using standard statistical methods; $A_{est} = \tilde{\alpha}A_{act}^\beta$, which can be written in terms of logs as $\log A_{est} = \alpha + \beta \log A_{act}$. Figure 12 shows scatterplots of log estimated versus log actual sizes of areas, with simple linear regressions summarising the relationship between estimated and actual size for each of the plot types and each of the types of pieces.

Let E_{ijkl} be the estimated size of the area for a piece in plot type i (with 1 = mosaic, 2 = pie, and 3 = spine), location j (with 1 = piece in unframed plot, 2 = inside piece of a framed plot, and 3 = exterior(frame) piece of a framed plot), response k where $k = 1, \dots, n_{ij}$, where n_{ij}

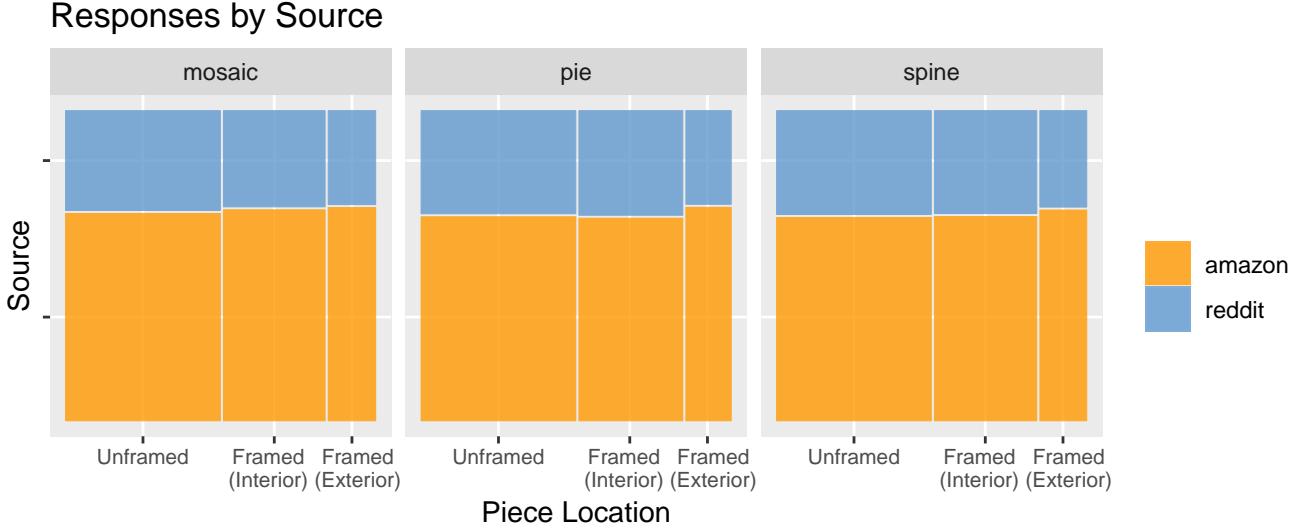


Figure 13: Breakdown of the number of evaluations by source (reddit or amazon) for all combinations of plot type and piece locations.

is the number of evaluations for plot-frame location combination ij by one of the participants ℓ , $\ell = 1, \dots, 95$. Let A_{ij} be the corresponding actual size of the piece.

The model is written in the form of regressions lines of the form $\log E = \alpha + \beta \log A$, with the unframed version of each of the types as the base line. The other parameters can then be written as contrasts to the base line, producing a set of easily interpretable results:

$$\log E_{ijkl} = \alpha_i + \beta_i \cdot \log A_{ij} + (\alpha_{ij} + \beta_{ij} \cdot \log A_{ij}) + e_{ijk} + u_\ell, \quad \text{Model (1)}$$

where

α_i can be interpreted as the estimated size of an area in an unframed design with an actual size of zero for each of the plot types i ($i = 1, 2, 3$),

β_i can be interpreted as the average change in the log ratio of estimated and actual size of a

piece in an unframed design of plot type i ($i = 1, 2, 3$) when the actual size increases by 1 percentage point,

α_{ij} , β_{ij} are the average multiplicative changes to α_i and β_i for a piece in location j ($j = 1, 2, 3$).

For estimability, $\alpha_{i1} = \beta_{i1} = 0$ for all plot types i ,

e_{ijk} is the error observed within each combination of plot type and piece location ij . We assume

$e_{ijk} \sim N(0, \sigma_e^2)$ i.i.d. for all $k = 1, \dots, n_{ij}$,

u_ℓ is a random effect introduced by participant ℓ , $\ell = 1, \dots, N$, where $N = 95$ participants taking part in the study. Each participant evaluated at least ten of the 18 pieces and gave, on average, 17 (valid) estimations out of 18 possible evaluations. We assume that the random effect is normally distributed, i.e. $u_\ell \sim N(0, \sigma_u^2)$ i.i.d. for all $\ell = 1, \dots, N$, and independent of the error, i.e. $\text{cov}(u, e) = 0$.

Estimates for all of the fixed effects parameters are shown in Table 1. Variance estimates show that performance varies between participants ($\widehat{\sigma}_u = 0.169$) but overall variability in evaluation of sizes is dominated by the residual error $\widehat{\sigma}_e = 0.444$.

Humans generally estimate using round numbers Baird et al. [1970]. Participants in our study showed the same preference, as can be seen in Figure 14b. We therefore fit a model of the same structure using the actual size rounded up to the nearest 5. Coefficient estimates and conclusions drawn from this model are similar to those using the (not rounded) actual percentage values, indicating that our results are not due to human rounding tendencies.

Assessing the fitted model, there are three main conclusions:

- $\alpha_1 + \beta_1 \log(A)$, $\alpha_2 + \beta_2 \log(A)$, $\alpha_3 + \beta_3 \log(A)$ represent estimates for a participant's assessment

Type	Location	Param	Estimate	LB	UB	Pr...t..	Signif.
mosaic	Unframed	α_1	1.274	1.127	1.421	0.000	***
		β_1	0.548	0.490	0.606	0.000	***
	Framed-inside	α_{12}	0.163	-0.051	0.378	0.136	
		β_{12}	-0.028	-0.124	0.067	0.561	
	Framed-frame	α_{13}	-1.959	-3.297	-0.620	0.004	**
		β_{13}	0.464	0.029	0.899	0.037	*
pie	Unframed	α_2	1.036	0.853	1.219	0.000	***
		β_2	0.700	0.638	0.761	0.000	***
	Framed-inside	α_{22}	-0.170	-0.437	0.097	0.211	
		β_{22}	0.106	0.011	0.201	0.029	*
	Framed-frame	α_{23}	-0.255	-1.279	0.769	0.626	
		β_{23}	-0.152	-0.470	0.165	0.348	
spine	Unframed	α_3	0.194	-0.039	0.426	0.103	
		β_3	0.905	0.827	0.984	0.000	***
	Framed-inside	α_{32}	0.310	-0.078	0.699	0.118	
		β_{32}	-0.043	-0.188	0.103	0.565	
	Framed-frame	α_{33}	-0.054	-1.045	0.937	0.915	
		β_{33}	-0.106	-0.391	0.179	0.467	

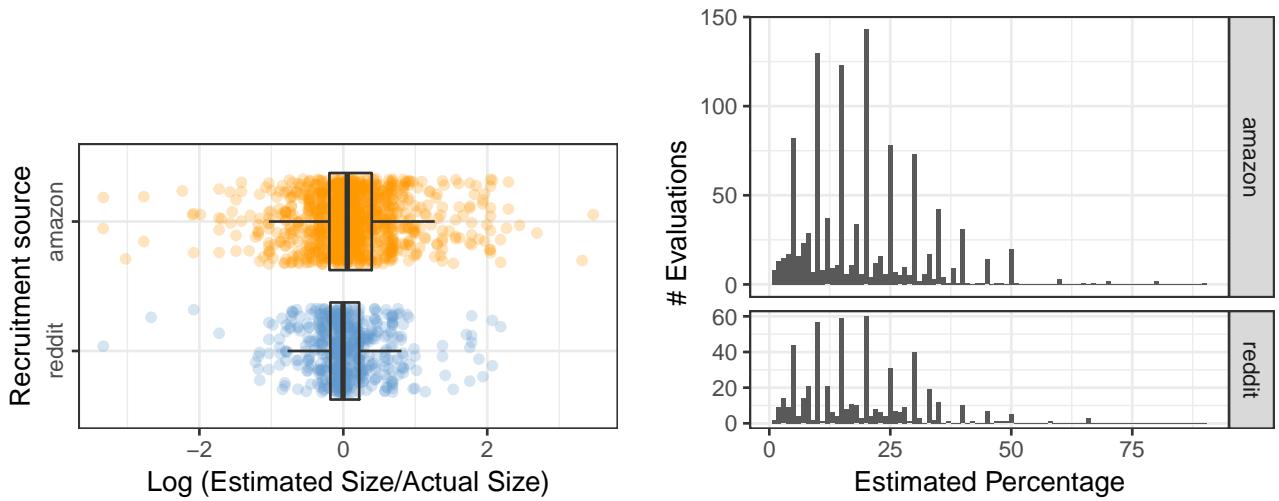
Table 1: Table of parameters and corresponding estimates for Model (1).

of an area of size $\log(A)$ in unframed (regular) mosaic, pie, and spine charts. Values of $\alpha_i = 0$ and $\beta_i = 1$ indicate no bias; estimates with $\alpha_i + \beta_i \log(A) < \log(A)$ indicate a visual under-estimation of the area, while estimates with $\alpha_i + \beta_i \log(A) > \log(A)$ indicate an over-estimate. All parameter estimates β_i , $i = 1, 2, 3$, are smaller than 1, indicating that participants tend to underestimate the size of large areas A . For small areas A this effect is reversed: all parameter estimates α_i , $i = 1, 2, 3$ are positive, indicating that participants tend to overestimated the size of small areas. Spine plots ($\hat{\alpha}_3 = 0.194$, $\hat{\beta}_3 = 0.905$) are less prone to these biases, in particular α_3 is not significantly different from zero; Mosaic plots ($\hat{\alpha}_1 = 1.274$, $\hat{\beta}_1 = 0.548$) on the other hand exhibit the worst biases. The two-dimensional nature of area assessment in mosaic plots results in poorer estimates, which is consistent with our original hypothesis.

- Estimation of the inside portion of framed plots is not significantly different from estimation of portions of unframed plots in mosaics or pie charts, as both $\alpha_{.2}$ and $\beta_{.2}$ are not significantly different from the unframed plot estimates. It is possible that the known bias towards under-estimation of area is counteracted by the over-estimation of the area of the inside portion of the framed plot (and corresponding under-estimation of the framed region). For spine plots, estimation of the inside portion shows a larger bias towards over-estimation of small areas; that is, $\alpha_{22} = 0.31$. Even in framed spine plots, though, the small area over-estimation bias is smaller than in unframed mosaic or pie charts.
- Estimation of the framed portion of the plot is significantly worse than estimation of the same plot area in the unframed version of the plot for all three types of charts (though spine plots are only marginally significant). In the unframed version of mosaic and pie charts, the unaccommodated area tends to be over-estimated for small areas and approximately accurate for large areas; in the framed versions, the framed area tends to be under-estimated.

From these observations, we can conclude that framed plots are perceptually not an ideal choice, even though they are aesthetically pleasing. When spine plots are framed, area estimation does not suffer significantly, but with both pie and mosaic plots, it does; this inconsistency is likely due to the different perceptual demands of pie charts and mosaic plots compared to spine plots. Pie charts require estimation of the relative angle of the slice, which inherently requires consideration of two dimensions simultaneously. Spine plots require consideration of the relative length of a single dimension; Cleveland and McGill [1984] ranks this as one of the easiest psychophysical estimation tasks.

Our participants were drawn from two different populations – reddit and amazon mechanical



(a) Variability is somewhat lower in responses from participants recruited from reddit compared with participants recruited from Amazon Turk.

(b) Histograms show preference for rounding to 5.

Figure 14: Comparison of results from the two different sources of participant recruitment.

turk. On reddit, participants were recruited from /r/samplesize and /r/visualization; both communities are interested in data or data visualization, and thus are more likely to be data-literate than the general population. This may explain why responses from redditors have lower variability than responses from amazon mechanical turkers (see Figure 14a). Figure 14b shows responses from redditors and mechanical turkers. On the left side of the figure we see that both groups tend to round their area estimates to the nearest 5%, as seen in the modes of the histograms. We expected that redditors might be more accurate due to greater interest in the topic; however, other than the lower variability of responses, we did not detect any significant difference between the two populations. This may indicate that the biases responsible for estimation errors are perceptual in nature and may not decrease with education or exposure.

5 Conclusion

The United States Census Bureau did an incredible job at collecting, tabulating, visualizing, and maintaining the data presented in the Statistical Atlas. Reproducibility has become a highly discussed topic in science lately, but few researchers even try to preserve their data and methods in such a way that the results could be precisely duplicated 150 years later. The Census Bureau (in conjunction with the Library of Congress) has managed to make it fairly easy for researchers to access the data necessary to conduct this study online.

In addition to the technical feat of reproducibility, the Statistical Atlas, spearheaded by the vision of Francis Walker, represented a state-sponsored foray into the development and use of new types of charts. With 107 pages of charts, and several relatively new types of plots, it is not surprising that some of these experiments fell flat. While framed plots are aesthetically appealing, they are not effective at communicating numerical information accurately. In addition, there are several design choices made in plates 31 and 32 which can be attributed to the technological limitations of hand-drawn plots. These design choices are not ideal from a modern perspective, but are perfectly reasonable when viewed in light of the technological constraints of lithography. While we have established that using framed plots is a suboptimal design choice, we acknowledge that the plots have a certain visual appeal. It is possible that the creators of the Statistical Atlas recognized the failed experiment, as framed plots do not appear in the 1880 or 1890 atlases.

Of the successes which underlie the creation of the 1870 Statistical Atlas, it would be better to leave the framed charts in the annals of history and focus on the remarkably resilient infrastructure which allows us to test the effectiveness of these plots 150 years later.

SUPPLEMENTAL MATERIALS

Github Repository: https://github.com/srvanderplas/Statistical_Atlas contains code and data sets used in this analysis.

References

Alan Agresti and Brent A. Coull. Approximate is better than “exact” for interval estimation of binomial proportions. *The American Statistician*, 52(2):119–126, 1998. doi: 10.1080/00031305.1998.10480550. URL <http://dx.doi.org/10.1080/00031305.1998.10480550>.

John C Baird, Charles Lewis, and Daniel Romer. Relative frequencies of numerical responses in ratio estimation. *Perception & Psychophysics*, 8(5):358–362, 1970.

Richard A. Becker, William S. Cleveland, and Ming-Jen Shyu. The visual design and control of trellis display. *Journal of Computational and Graphical Statistics*, 5(2):123–155, 1996.

Bureau of Labor Statistics. How the government measures unemployment. https://www.bls.gov/cps/cps_htgm.htm#nilf, 2015.

Patricia A. Carpenter and Priti Shah. A model of the perceptual and conceptual processes in graph comprehension. *Journal of Experimental Psychology: Applied*, 4(2):75, 1998.

Cartography Associates. Julius Bien, Master Printer and Cartographer, 2009. URL <http://www.davidrumsey.com/blog/2009/9/13/julius-bien-master-engraver-and-cartographer>.

William S. Cleveland and Robert McGill. Graphical perception: Theory, experimentation and

application to the development of graphical methods. *Journal of the American Statistical Association*, 79(387):531–554, 1984.

Michael Friendly. A Brief History of the Mosaic Display. *Journal of Computational and Graphical Statistics*, 11(1):89–107, March 2002.

Mark Harrower and Cynthia A. Brewer. Colorbrewer.org: an online tool for selecting colour schemes for maps. *The Cartographic Journal*, 40(1):27–37, 2003.

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

Jürgen Hummel. Linked bar charts: Analysing categorical data graphically. *Journal of Computational Statistics*, 11:23–33, 1996.

Charles Kostelnick. Melting-pot ideology, modernist aesthetics, and the emergence of graphical conventions: The statistical atlases of the united states, 1874-1925. *Defining visual rhetorics*, pages 215–242, 2004.

Robert E. Krider, Priya Raghbir, and Aradhna Krishna. Pizzas: π or square? psychophysical biases in area comparisons. *Marketing Science*, 20(4):405–425, 2001.

Stephan Lewandowsky and Ian Spence. The perception of statistical graphs. *Sociological Methods & Research*, 18(2-3):200–242, 1989.

Minnesota Population Center. *National Historical Geographic Information System (NHGIS), Version 11.0 [Database]*. Minneapolis: University of Minnesota, 2016. data retrieved from

System: Version 11.0 [Database]. Minneapolis: University of Minnesota, <https://www.nhgis.org/>.

Jean Piaget. Quantification, conservation, and nativism. *Science*, 162(3857):976–979, 1968.

Severino (ed.) Ribecca. Marimekko chart. http://www.datavizcatalogue.com/methods/marimekko_chart.html, 2018.

Steven Ruggles, Katie Genadek, Ronald Goeken, Josiah Grover, and Matthew Sobek. *Integrated Public Use Microdata Series: Version 6.0*. Minneapolis: University of Minnesota, 2015.

Trudy A Suchan, Marc J Perry, James D Fitzsimmons, Anika E Juhn, Alexander M Tait, and Cynthia A Brewer. Census atlas of the unites states, series censr-29. *US Census Bureau*. Retrieved February, 1:2008, 2007.

Martha Teghtsoonian. The judgment of size. *The American Journal of Psychology*, 78(3):392–402, 1965. ISSN 00029556. URL <http://www.jstor.org/stable/1420573>.

Edward Tufte. *The Visual Display of Quantitative Information*. Graphics Press, USA, 2 edition, 1991.

United States Census Bureau. Table 1. Educational Attainment of the Population 18 Years and Over, by Age, Sex, Race, and Hispanic Origin: 2016. <https://www.census.gov/data/tables/2016/demo/education-attainment/cps-detailed-tables.html>, 2016.

Francis A. Walker. Statistical Atlas of the United States, Based on the Results of the Ninth Census, 1870, with Contributions from Many Eminent Men of Science, and Several Departments of the

Government by Francis A. Walker. *The North American Review*, 121(249):437–442, 1875. ISSN 00292397.

Francis Amasa Walker. Statistical Atlas of the United States, Based on the Results of the Ninth Census, 1870, with Contributions from Many eminent Men of Science, and Several Departments of the Government. digitized version provided through Library of Congress, <https://www.loc.gov/item/05019329/>, 1874.

Charlotte Wickham. *munsell: Utilities for Using Munsell Colours*, 2016a. URL <https://CRAN.R-project.org/package=munsell>. R package version 0.4.3.

Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2 edition, 2016b. ISBN 978-0-319-24275-0. doi: 10.1007/978-3-319-24277-4. URL <http://ggplot2.org>.

Hadley Wickham and Heike Hofmann. Product plots. *IEEE Transactions on Visualization and Computer Graphics (Proc. InfoVis '11)*, 17(12):2223–2230, 2011.