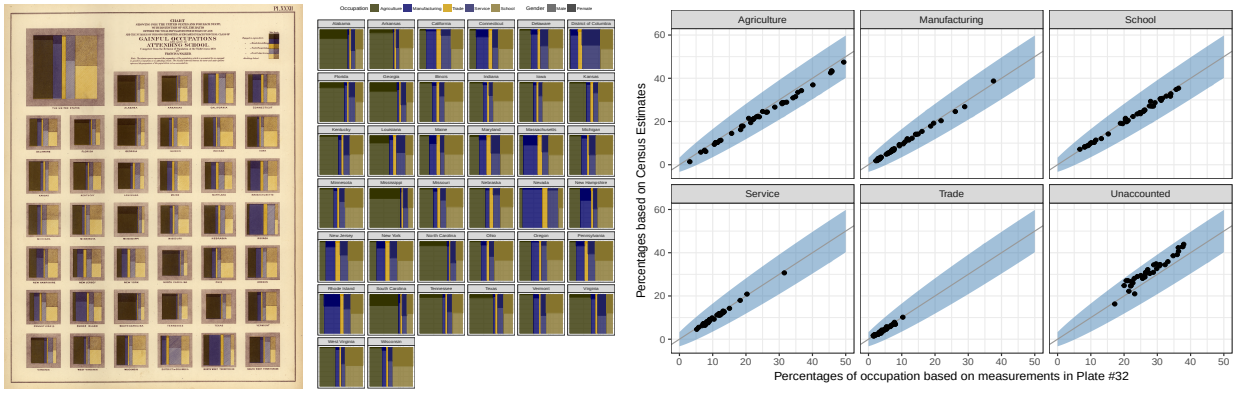
A call for computational reproducibility in InfoVis

Heike Hofmann *Member, IEEE*, Susan R. VanderPlas, and Ryan C. Goluch



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

(b) Reproduction of plate #32.

(c) Differences/similarities between numbers from the Statistical Atlas and the modern reproduction.

Fig. 1: Example of translational reproducibility: from left to right we have the original chart in Fig. 1a, its reproduction in Fig. 1b, and the number differences between the two in Fig. 1c.

Abstract—Computational Reproducibility is a fundamental aspect of the scientific method. One question that we need to therefore ask ourselves is "how reproducible are our charts?". Recent developments have made it much easier to ensure computational reproducibility of results and visualizations. In this paper, we investigate reproducibility of charts created almost 150 years ago based on data collected from the US census in 1870. Three times in the past, the US Census Bureau published a Statistical Atlas to map the state of the Union based on data collected in the 9th, 10th, and 11th US census. Each of these atlases represents a masterpiece in science and technology. The atlases also introduced novel ways of visualizing data. In this paper, we *discuss* two plates of the Statistical Atlas of 1874, show a way to *re-create* the charts using modern tools and freely accessible data, and *re-display* the data to emphasize missing values.

Index Terms—Mosaic plots, translational reproducibility, statistical graphics, Census Bureau.

1 Introduction

Three times in the past, the US Census Bureau published a Statistical Atlas to map the state of the Union based on data collected in the 9th, 10th, and 11th US census (in 1870, 1880, and 1890). Each of these atlases represents a masterpiece in science and technology. Here, we want to focus on the ninth Census, supervised by Francis A. Walker. At this time, the United States had a population of about 38.5 million people. The Atlas represents a graphical compendium of the census information prepared in more than 100 lithographic plates. Most of these plates are overlaid maps, but some consist of more abstract and, at that time, novel visualizations. Of particular interest are plates #31 and #32. Both of these plates have a very similar structure: they show small multiples, one for each state, of what are now known as mosaic plots or Marimekko charts.

The charts in the Statistical Atlas were created using extremely high-precision methods for the time it was published. Color images were produced by Julius Bien's publishing house [1, 9] 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.

- *Heike Hofmann is with the Department of Statistics and faculty member of the Human Computer Interaction Program, Iowa State University. E-mail: hofmann@mail.iastate.edu.*
- *Susan R VanderPlas is with the Department of Statistics, Iowa State University. E-mail: skoons@iastate.edu.*
- *Ryan C Goluch is with the Department of Software Engineering, Iowa State University. E-mail: rgoluch@iastate.edu*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

2 Reproducibility

Reproducible research is a frequent topic of discussion in data visualization and data science [3, 7, 11, 18, 32]. This paper sets out to reproduce with as much fidelity as possible two of the hand-drawn charts of the Statistical Atlas, using modern methods. In some cases reproducibility focuses on whether the results of a study can be replicated from the exact same data set using the same methods and computer code; this is not the approach we are taking in this paper. Rather, here we are exploring whether it is possible to access the data from the 1870 census (or a sample thereof) and, using that data, re-create some of the charts in the Statistical Atlas using modern methods. In addition to sampled data, we also extrapolate data from digitized versions of the original charts by measuring the geometric objects. This type of reproducibility might be termed "translational reproducibility", as it examines whether methodology and data can be translated across more than a century to produce the same results. Faced with a similar problem in Bioinformatics research and results, Baggerly and Coombes [4] used the term "forensic bioinformatics" to define situations in which researchers

attempt to replicate study results using other methods.

In order for research to be reproducible, it is important to have not only a description of what was done, but also access to as much of the environment in which the analysis was conducted as possible. Buckheit & Donoho [7] note that:

> An article about computational science in a scientific publication is not the scholarship itself, it is merely advertising of the scholarship. The actual scholarship is the complete software development environment and the complete set of instructions which generated the figures.

Extending this, scholarship in information visualization is then the combination of data (and data collection framework, environment, and methodology), the software and code used to analyze the data and generate results, and the final collection of interpretations and "advertisement" that summarizes the results in publication form.

Reproducing scholarship from almost 150 years ago relies heavily on the preservation of critical infrastructure used to generate the original images in the 1870 Statistical Atlas. The Library of Congress preserved the original "advertisement": the pages of the Statistical Atlas, available online through very high-resolution scans. The Census Bureau has preserved the data from the 1870 census, but equally importantly, they have also preserved the basic framework for data collection, which is still used today to compile the modern census (the questions have changed a bit, and the country has expanded considerably, but the basic goal of the census has not changed). While researchers do not have access to the full 1870 census data, the Integrated Public Use Microdata Series (IPUMS-USA) provided through the Minnesota Population Center [24] releases a 1% microsample for research purposes. This sample allows us to reproduce the graphics in the Statistical Atlas using data read from the original prints as well as from estimates extrapolated from the 1% sample of the data. By duplicating the graphics using both methods, we are able to compare the data to the final product, examining the fidelity of the printed result compared to the recorded data.

Reproducibility depends on openness: open data and open publications. Without both repositories of information, it would be difficult or impossible to go back and compare past data visualizations with today's methodology.

[25] highlights the importance of the entire framework for reproducibility with ten rules for reproducible computational research:

1. For every result, keep track of how it was produced.
2. Avoid manual data manipulation steps.
3. Archive the exact versions of all external programs used.
4. Version control all custom scripts.
5. Record all intermediate results, when possible in standardized formats.
6. For analyses that include randomness, note underlying random seeds.
7. Always store raw data behind plots.
8. Generate hierarchical analysis output, allowing layers of increasing detail to be inspected.
9. Connect textual statements to underlying results.
10. Provide public access to scripts, runs, and results.

With R [22] and its add-on packages (knitr, packrat), it is incredibly easy to implement these ten suggestions. `knitr` [32] makes it easy to store code and results while tying computational steps to textual statements, `packrat` [27] allows researchers to easily archive software, and the use of git with R projects makes version control and intermediate results available to even novice users. The developments in R packages to facilitate reproducible research makes reproducibility much easier for the modern researcher than it would have been for those compiling the 1870 census. These rules are focused on the computational era, but many of the underlying ideas can be translated to the creation of the Statistical Atlas. Intermediate results are recorded in tabular summaries of the census data, raw data was stored (though only a 1% sample is made available to the public), and hierarchical summaries are shown: plates generally include an overview of the country as well as individual states and territories. While the specific method for producing the plots has not (to our knowledge) survived to the modern era, the plots used in the two plates of the Statistical Atlas examined in this paper are fairly easy to reverse-engineer, which allows us to compare the data shown in the plots to the data provided in the 1% microsample of the 1870 census, resulting in several surprising discoveries.

This study is intended to examine the persistence of data and methodology across nearly 150 years and several technological revolutions. As the study of statistical visualization has developed considerably over the past 147 years, we also examine the visualization decisions made for the 1870 statistical atlas and create improved graphics which more clearly display the same data. The technological advances since the 1870 census also allow us to more easily add a spatial component, as it is now much easier to display the census data in map form. These improvements allow us to add additional depth to the 1870 statistical atlas graphics without investing hundreds of hours of artistic work for each additional map and chart.

3 Statistical Archaeology

Most of the plates of the Statistical Atlas are overlaid maps, but some consist of more abstract and, at that time, novel visualizations. Of particular interest for us are plates #31 and #32 . Plate #31 (see Fig. 6) gives an overview of 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), plate #32 (see Fig. 1a) shows the gender ratio of the population over the age of ten in different types of occupations. Both of these plates have a very similar setup and structure: at the top of the chart we find an in-depth description (see the legend for plate #32 in Fig. 2) and a legend detailing the color choices (see Fig. 3). The structure of the visualization is that we have a big square in the top left with an US-wide aggregate of the situation, and a series of small multiples [26] or trellis plots [5], one for each state, of what are now known as mosaic [13, 16, 31] or Marimekko plots [23]. There is a twist to both plates, though: a grey band is drawn around each one of the states' squares representing essentially missing information. In case study 1 the grey band is proportional to the number of population "unaccounted" for, 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. In the second case study, the grey band represents the size of religiously unaccommodated population over the age of ten. The choice to show the unaccounted/unaccommodated part of the population by a grey band around the square is somewhat unfortunate, as it breaks the overall metaphor of the mosaic plot and thereby prevents any direct comparisons across charts except for area comparisons, which are cognitively harder and more error prone than comparisons of lengths [10]. It also masks the size of the population that is thus unaccounted/unaccommodated for by visually cutting it into quarters. This is one of the design choices that we re-visit in our re-creation of the chart.

At the time the Statistical Atlas for the ninth census was created, Mosaic plots were a novel way of visualizing data. Even though area plots as a means of visualizing data had been in use before [13], e.g. Minard's [20] plate #3 [14], Mosaic plots in their modern form of use were not published until 1877 [19]. However, the descriptions given on both plates #31 and #32 clearly state the intention of their creators that areas are designed to be proportional to the population they represent:

> The interior squares represent the proportion of the population which is accounted for as engaged in gainful occupations or as attending school. The shaded intervals between the inner and and outer squares represent the proportion of the population not so accounted for.

Both plates #31 and #32 are based on population totals for (as stated on the legend, see Fig. 2) "the total population over 10 years of age". Because state-level aggregates of the number of total population above the age of ten are not available directly; instead we are extrapolating these numbers based on the 1% microsample of the ninth census provided by the Integrated Public Use Microdata Series (IPUMS-USA) provided through the Minnesota Population Center [24]. Using

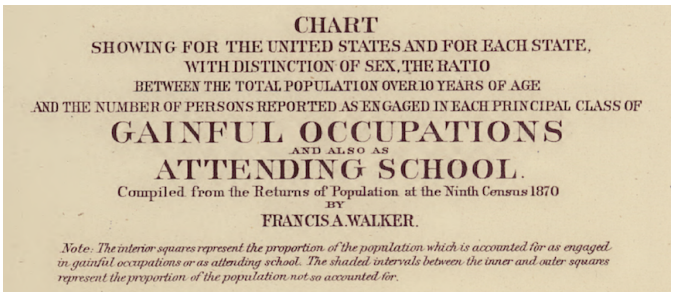


Fig. 2: Zoom-in to the description section of plate #32.

Fig. 3: Zoom-in to the legend of plate #32.

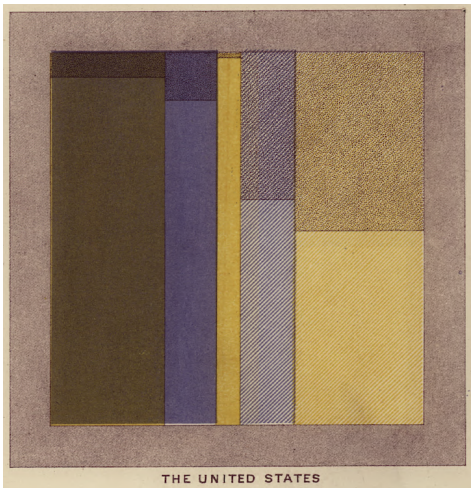


Fig. 4: Zoom-in to the overview of the US wide distribution of genders across occupations.

the small sample, we get counts of the male and female population above ten as well as state totals. This allows us to get estimates for the size of the male and female population above ten for each state by extrapolating from the sample proportions.

Let $n_{UA}^{(\mathrm{XY})}$ be the –unfortunately unknown– size of the unaccounted population of state *XY*. Then we have the relationship

$$n_{UA}^{(\mathrm{XY})} = n_{\mathrm{AGE}\geq 10}^{(\mathrm{XY})} - n_{\mathrm{acc}}^{(XY)}, \tag{1}$$

where $n_{\mathrm{AGE}\geq 10}^{(\mathrm{XY})}$ is the number of over ten year olds in the state, and $n_{\mathrm{acc}}^{(XY)}$ is the size of the population accounted for (i.e. gainfully employed or going to school). $n_{\mathrm{acc}}^{(XY)}$ is known for each state from table NT13 on "Employed Population by Occupation by Age by Sex" from the 1870 Census: Religious Bodies, Occupation & Government Data (1870_sROG) provided by the National Historical Geographic Information System (NHGIS) [21]. Unfortunately, the total number of over ten year olds in each state $n_{\mathrm{AGE}\geq 10}^{(\mathrm{XY})}$ is not known. We do have the relationship to $n_{\mathrm{total}}^{(\mathrm{XY})}$, the total population size of state *XY*:

$$n_{\mathrm{AGE}\geq 10}^{(\mathrm{XY})} = n_{\mathrm{total}}^{(\mathrm{XY})} \cdot p_{\mathrm{AGE}\geq 10}^{(\mathrm{XY})}.$$

where $p_{\mathrm{AGE}\geq 10}^{(\mathrm{XY})} \in [0,1]$ is the proportion of the population in state *XY* 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.

$$\widehat{p_{\mathrm{AGE}\geq 10}}^{(\mathrm{XY})} = \frac{\#\text{ individuals age ten and over in state } XY}{\#\text{ individuals in state } XY}.$$

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

$$\widehat{n_{UA}}^{(\mathrm{XY})} = n_{\mathrm{total}}^{(\mathrm{XY})} \cdot \widehat{p_{\mathrm{AGE}\geq 10}}^{(\mathrm{XY})} - n_{\mathrm{acc}}^{(XY)}, \tag{2}$$

$$s.e.(\widehat{n_{UA}}^{(\mathrm{XY})})^2 = \widehat{p_{\mathrm{AGE}\geq 10}} \cdot (1 - \widehat{p_{\mathrm{AGE}\geq 10}}) n_{\mathrm{total}}^{(\mathrm{XY})}. \tag{3}$$

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 get estimates and standard errors for the size of the religiously unaccommodated population over ten years of age in case study 2.

3.1 Case Study 1: Gender Ratio in Agriculture, Trade, Service, Manufacturing, and Schools

Fig. 1a shows a miniature of the chart published as plate #32 in the Statistical Atlas of 1874 [28] 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 and the legend of Fig. 2 and Fig. 3, we can interpret the details of each of the squares at the example of Fig. 4. 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 shown as the widths of the rectangles, the heights are proportional to the percentage of gender, with men shown in the bottom rectangle and women in the top rectangle. The grey band around each one of the states' squares is proportional to the number of population "unaccounted" for, 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 cognitively a hard task. The task becomes much easier, if we incorporate the information directly into the mosaic plot, as shown in Fig. 5. This figure reveals another previously hidden finding: about 97% of the unaccounted population are women and girls! 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 Fig. 5, 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 [8] participation rate, and thus their contributions to the

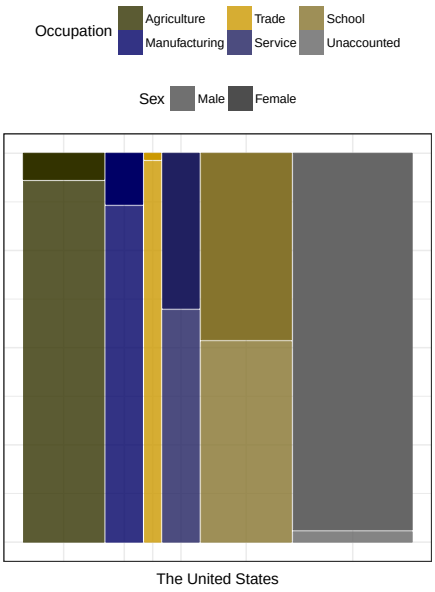


Fig. 5: Recreation of the mosaic plot nation-wide overview of gender ratio by occupation based on gainfully employed population over ten.

overall economy are not typically counted.

Visual inspection tells us that the data used to render the nation-wide overview in Fig. 5 and the state-level aggregates in Fig. 1b 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. Fig. 1c 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 Fig. 1c that the numbers closely match; the regions in blue are based on point-wise Agresti-Coull 95% confidence intervals [2]. None of the points are outside these confidence bands. For all occupation levels and school attendance the numbers are *v*ery 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.

On another note: the Census information about territories is available at a higher resolution than shown on plate #32. Charts with mosaic plots of all territories are available in the online supplement.

3.2 Case Study 2: Church Accommodations by State

Fig. 6 shows a miniature of plate #31 from the 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 are shown as colored stripes, the width of each is proportional to the number of their sittings. 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 Fig. 7a).

The inside squares of plate #31 are simpler structured than those of plate #32 – they are one-dimensional mosaic plots, also known as rotated stacked bar charts or spine plots [17]. 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. This is one design choice, that we are going to address in our reproduction of it. However,

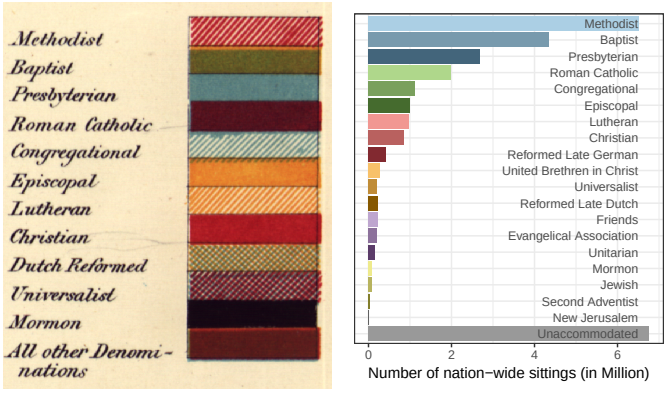
Fig. 6: Plate #31 from the Statistical Atlas of 1874: church accommodation by state and denomination for the population of over ten years of age.

there are other design choices that have questionable(?) consequences:

Top four only: A decision was made to only show the top four denominations in each state and the top eight denominations overall. This has the effect, that some denominations are shown but others are not, even though the ones not shown have an overall higher number of sittings. For example, the Reformed Late Germans have in 1870 nation-wide 431,700 sittings, which puts them into tenth place However, they are not mentioned in plate #32, whereas the Mormon Church is locally (Utah Territory) strong enough to get featured, even though it is with 87,838 sittings far behind the German Reformed Church in a nation-wide 17th place. This decision was likely at least partially due 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 5 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 cognitively more difficult, as it re-orders the colors in the bar code-like color strips representing each state and 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 Statistical Atlas plots and the proportion estimated by using the 1% microsample. This difference was resolved by considering only the population above age 11 when computing estimates for plate #31. There is some ambiguity



(a) Legend section of plate #31. (b) Church sittings by denomination.

Fig. 7: 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 were accommodated for in 1874. We also see that the number of unaccommodated people over the age of ten is higher than the number of sittings in any of the denominations.

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 Fig. 8) 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. In addition, the state plots are much more 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 were based off of 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 Fig. 8 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 Fig. 9 shows 3 sub plots from the 1874 Statistical Atlas; figure **??** shows an improved version of the same plots. We have added an additional bar for the proportion of the population which is unaccomodated, 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 unaccomodated persons.

If we re-display the data from plate # 31, unconstrained by the original format, we might choose to show all of the states on the y-axis, with denominations on the x axis, using stacked bars, as shown in Fig. 11. This allows for comparison between states much more readily than the small multiples approach, because states are aligned on a common axis, and denominations are shown in the same order in every stacked bar.

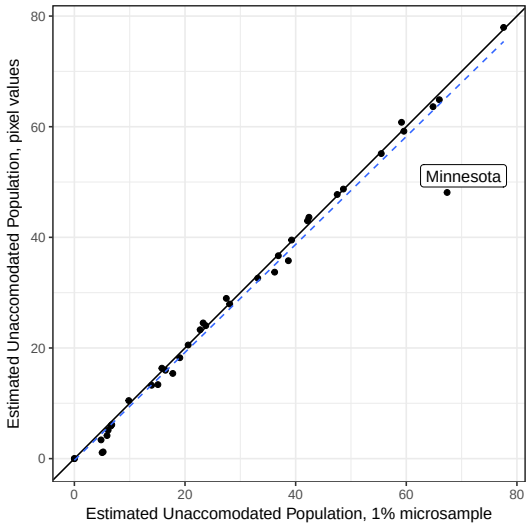


Fig. 8: Percentage of state-wide unaccommodated population based on estimates from the 1% Microsample (x-axis) and pixel values measured from the digitized version of plate #31 (y-axis).

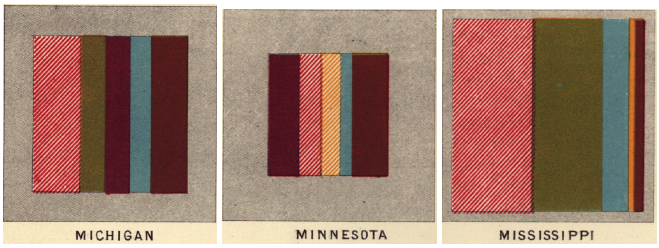


Fig. 9: Panels from three states: Michigan, Minnesota, and Mississippi.

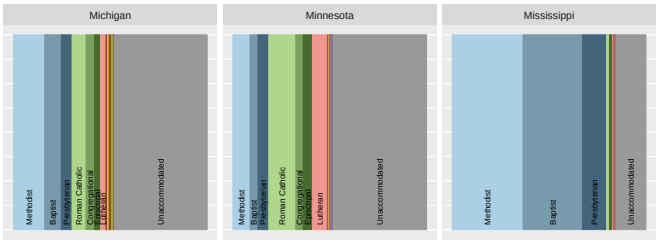


Fig. 10: State religious accommodation plots generated with the extttg-gplot2 package [30].

Fig. 11 does not show any information about unaccommodated persons: this information is shown in Fig. 12, with states ordered by the proportion of the accommodated population. The combination of these two plots allows us to compare states by denominations and accomodated population.

Both sets of re-imagined church accomodation plots use a triple color scheme. The lightest colors were drawn from the Colorbrewer [15] "Paired" palette's light colors (a grey hue was also added); each color was then darkened twice to produce a color scheme composed of 21 colors and 7 hues using the `munsell` package [**?**]. 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.

The re-imagined church accomodation plots provide the ability to make comparisons between states, but there is one obvious ingredient which is still missing: geographic context. Fig. 13 shows the geographic distribution of church accomodations; this spatial information allows us to see that the mormon church is highly localized geograph-

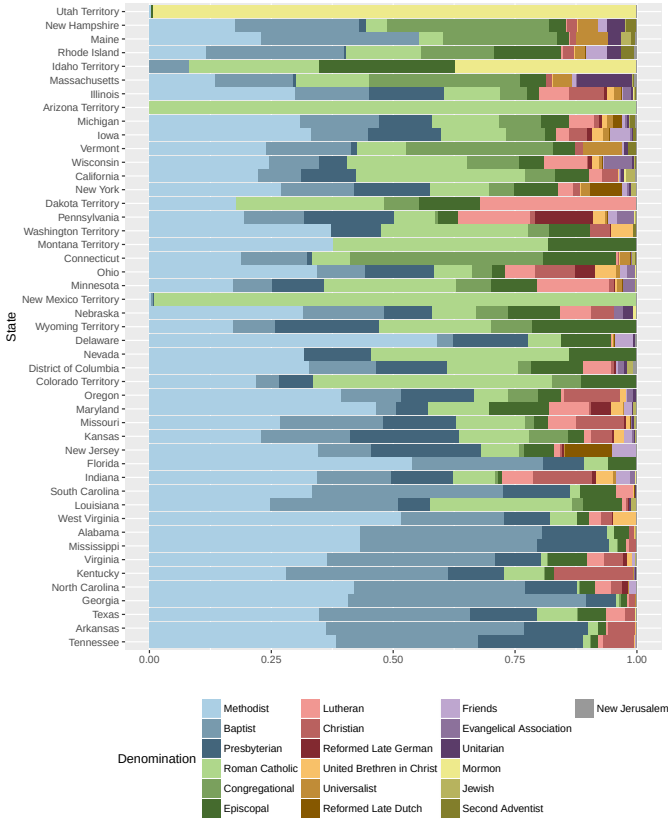


Fig. 11: Proportion of religious sittings for each state by denomination.

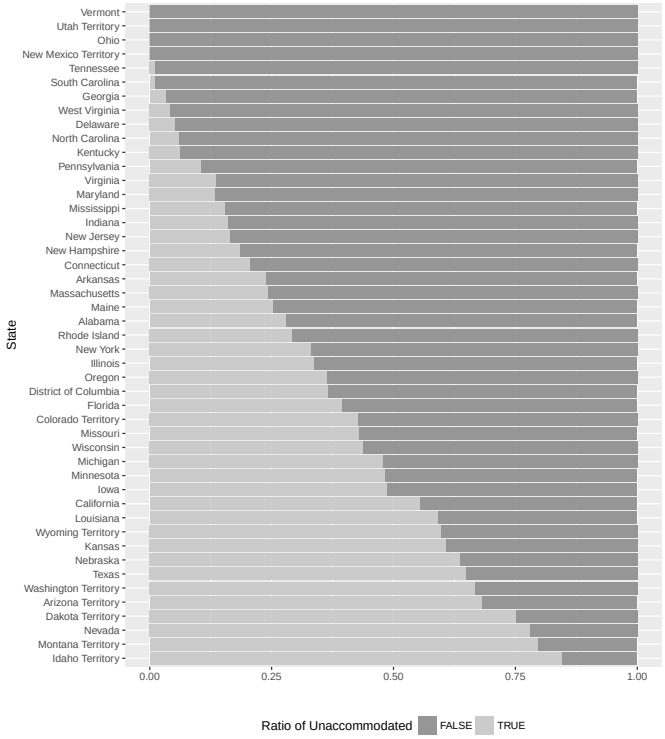


Fig. 12: Proportion of unaccommodated population for each state.

ically, while the Methodist church is relatively popular across most of the country. Numbers of unaccommodation population are – not surprisingly – particularly high in the new territories, with the exception of New Mexico. Fig. 13 is an example for a density dot map. The location of the dots within each state is random, but the number of the dots within each state is proportional to the population. Density maps are not new - some of the first examples appear in the Statistical Atlas accompanying the tenth US census of 1880 [**?**].

Fig. 13 is similar to maps included in the 1880 and 1890 Statistical Atlases that show the geographic strength of the most common denominations. These maps are accompanied by bar charts showing the proportion of individuals in each state who identify with each denomination, providing a very nice picture of the religious affiliations of the residents of the United States.

4 Discussion

Plates # 31 and # 32 of the statistical atlas are reproducible using modern methods and a bit of exploratory analysis. Reproducing these charts provides important lessons on the importance of archival information to reproducible analyses. Additionally, the reproduction process highlights some interesting features of historical chart creation, such as the use of grey bands around the plots to represent unaccounted or unaccommodated population, and whether 10 year olds are counted in the population or not.

The statistical atlas based on the 1870 census is the only version which contains mosaic plots. The atlases based on the 1880 and 1890 censuses contained pie charts, maps, and bar charts instead of mosaic plots. The purpose of the charts may have changed or developed over time as well: in 1870, the statistical atlas shows church sittings, which might be interpreted as the physical building space; in 1880, the atlas shows the geographical strength of each denomination, and in 1890, denominational statistics are displayed as pie charts. This difference may be due to a changing focus: first, it's necessary to make sure the infrastructure is in place as westward expansion occurs, then religious affiliation becomes more of a demographic concern. The mosaic plots for occupational statistics in 1870 give way to state bar charts in 1880 and stacked bar charts in 1890; segregation by gender is no longer a concern, and educational statistics are separated out into different charts entirely. It is likely that the chart design changed due to several factors: different goals, new technological capabilities, and individual preferences. Each atlas was published by a different publishing house, and was likely assembled by different individuals, each with their own preferred methods of data display. The use of different methods for displaying similar data led to discussions of which methods were best [6, 12, 29] and the formation of a community devoted to statistical visualization. True reproducibility of these figures would also include the "why": the goals of the visualization, known only to the individuals who made the decision to display the data using mosaic and spine plots. We return to Buckheit & Donoho's [7] insight:

> The actual scholarship is the complete software development environment and the complete set of instructions which generated the figures.

While we can no longer access the instructions from the designers of these figures, we can certainly generate comparable figures using what remains of the data and final product.

The reproducibility of the charts in the statistical atlas is only possible because of the infrastructure created by the US Census Bureau, the only government agency with constitutional mandate to gather data, as described in Article 1, Section 2:

> Representatives and direct Taxes shall be apportioned among the several States which may be included within this Union, according to their respective Numbers... The actual Enumeration shall be made within three Years after the first Meeting of the Congress of the United States, and within every subsequent Term of ten Years, in such Manner as they shall by Law direct.

The Census Bureau's dedication to preserving the historical data and images from the statistical atlas makes it possible to reproduce the

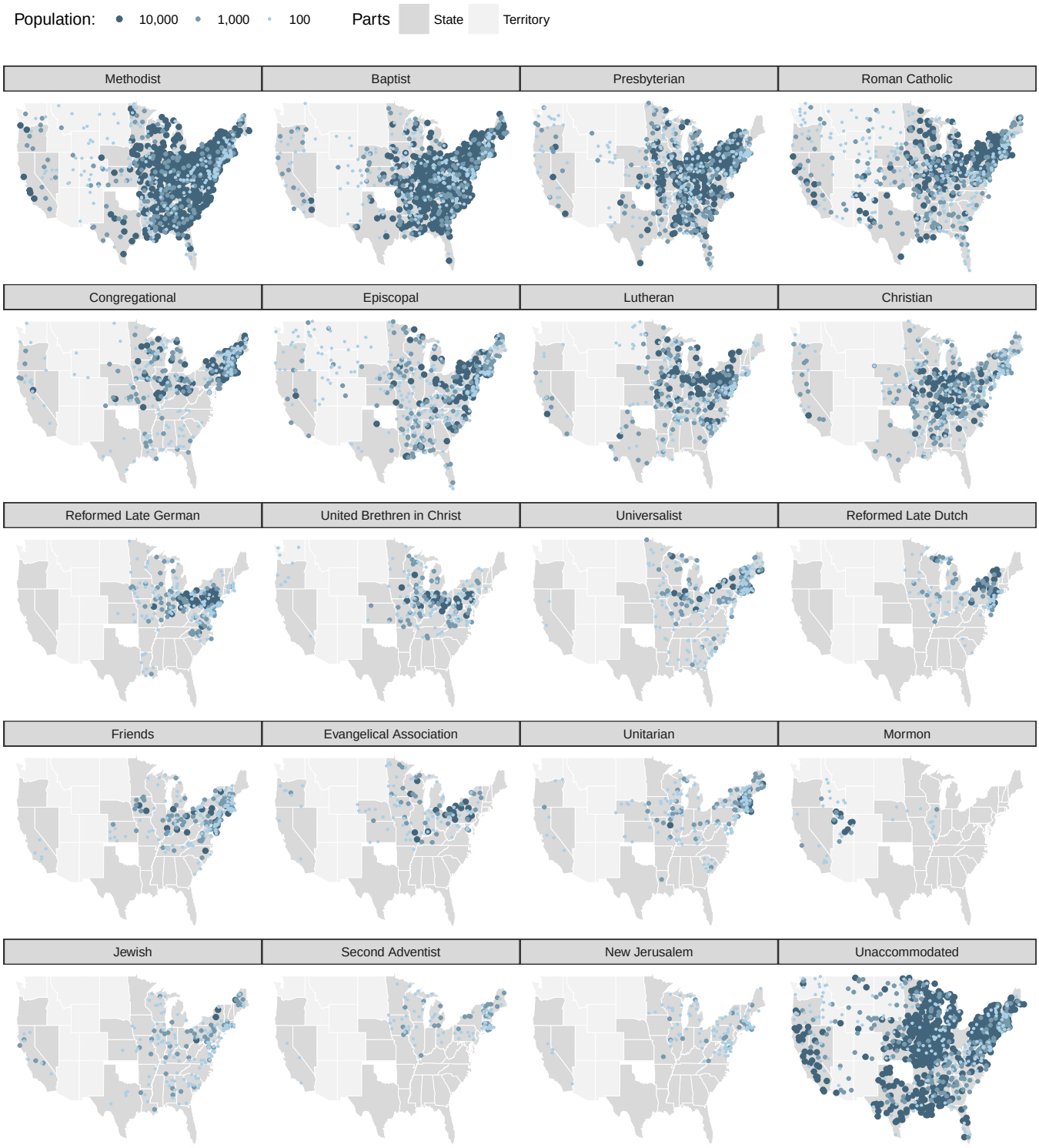


Fig. 13: Density dot maps of geographic expansion of religious sittings by denomination.

charts from plate # 31 and # 32, but most charts created today do not have the same commitment to historical preservation. It is not hard to imagine that most charts we create today will not be easily reproduced in 150 years - the code that we write will almost certainly not run on the operating systems which might be in use then; and our digitally stored data files may not be decodable either. We do not advocate for storing data on stone tablets, but we have made our code and aggregated data available on github to ensure at least near-term reproducibility. We have also opted to utilize packrat [27], which preserves the software package versions used to create this report. While these methods may not ensure reproducibility for 150 years, they should provide the ability for others to exactly replicate the steps we have taken for the next 5-10 years at least. Reproducibility longer-term requires a commitment to maintaining archival information in currently-readable

formats in perpetuity, and this commitment is (at this point) best left to governments, libraries, and other archives.

REFERENCES

[1] 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.

[2] A. Agresti and B. 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

[3] K. A. Baggerly and D. A. Berry. Reproducible research. *Amstat News*, 2011.

[4] K. A. Baggerly and K. R. Coombes. Deriving chemosensitivity from cell lines: Forensic bioinformatics and reproducible research in high-throughput biology. *Ann. Appl. Stat.*, 3(4):1309–1334, 12 2009. doi: 10.1214/09-AOAS291

[5] R. A. Becker, W. S. Cleveland, and M.-J. Shyu. The visual design and control of trellis display. *Journal of Computational and Graphical Statistics*, 5(2):123–155, 1996.

[6] W. Brinton. *Graphic Methods for Presenting Facts*. Industrial management Library. Engineering Magazine Company, 1919.

[7] J. B. Buckheit and D. L. Donoho. *WaveLab and Reproducible Research*, pp. 55–81. Springer New York, New York, NY, 1995. doi: 10.1007/978-1-4612-2544-7_5

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

[9] Cartography Associates. Julius Bien, Master Printer and Cartographer.

[10] W. S. Cleveland and R. McGill. Graphical perception: Theory, experimentation and application to the development of graphical methods. *Journal of the American Statistical Association*, 79(387):531–554, 1984.

[11] D. L. Donoho. An invitation to reproducible computational research. *Biostatistics*, 11(3):385, 2010. doi: 10.1093/biostatistics/kxq028

[12] W. C. Eells. The relative merits of circles and bars for representing component parts. *Journal of the American Statistical Association*, 21(154):119–132, 1926.

[13] M. Friendly. A Brief History of the Mosaic Display. *Journal of Computational and Graphical Statistics*, 11(1):89–107, Mar. 2002.

[14] General Research Division. *Tableau figuratif du mouvement commercial du Canal du Centre en 1844 Plate 3*. The New York Public Library, accessed in Feb 2017. doi: items/d8979ef0-ee85-0131-9ccb-58d385a7bbd0

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

[16] J. A. Hartigan and B. Kleiner. Mosaics for contingency tables. In *Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface*, pp. 268–273. Interface Foundation of North America, Inc., Fairfax Station, VA, 1981.

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

[18] F. Leisch. Sweave: Dynamic generation of statistical reports using literate data analysis. In *Compstat*, pp. 575–580. Springer, 2002.

[19] G. v. Mayr. *Die Gesetzmässigkeit im Gesellschaftsleben, statistische Studien*. Oldenbourg Verlag, München, 1877.

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

[21] 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/`. doi: 10.18128/D050.V11.0

[22] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017.

[23] S. e. Ribecca. Marimekko chart. `http://www.datavizcatalogue.com/methods/marimekko_chart.html`.

[24] S. Ruggles, K. Genadek, R. Goeken, J. Grover, and M. Sobek. *Integrated Public Use Microdata Series: Version 6.0*. Minneapolis: University of Minnesota, 2015. doi: 10.18128/D010.V6.0

[25] G. K. Sandve, A. Nekrutenko, J. Taylor, and E. Hovig. Ten simple rules for reproducible computational research. *PLOS Computational Biology*, 9(10):1–4, 10 2013. doi: 10.1371/journal.pcbi.1003285

[26] E. Tufte. *The Visual Display of Quantitative Information*. Graphics Press, USA, 2 ed., 1991.

[27] K. Ushey, J. McPherson, J. Cheng, A. Atkins, and J. Allaire. *packrat: A Dependency Management System for Projects and their R Package Dependencies*, 2016. R package version 0.4.8-1.

[28] F. 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. digitized version provided through Library of Congress, `https://www.loc.gov/item/05019329/`, 1874.

[29] H. Wickham. Graphical criticism: some historical notes. *Journal of Computational and Graphical Statistics*, 22(1):38–44, 2013.

[30] H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2 ed., 2016. doi: 10.1007/978-3-319-24277-4

[31] H. Wickham and H. Hofmann. Product plots. *IEEE Transactions on Visualization and Computer Graphics (Proc. InfoVis '11)*, 17(12):22232230, 2011.

[32] Y. Xie. *Dynamic Documents with R and knitr*. Chapman and Hall/CRC, 2 ed., 2015.