Revisiting two charts of the Statistical Atlas 1874

Heike Hofmann*,[a], Ryan Goluch[b]

[a]*Department of Statistics and Statistical Laboratory, Iowa State University, Ames, IA 50011-1121*

[b]*Department, Iowa State University, Ames, IA 50011-1121*

Abstract

This is the abstract.

It consists of two paragraphs.

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.

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 (Friendly 2002), e.g. Minard's (Minard 1844) plate #3 (General Research Division accessed in Feb 2017), Mosaic plots in their modern form of use were not published until 1877 (Mayr 1877). However, the descriptions given on both plates #31 and #32 makes it clear 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."

XXX What we are planning on doing here

1. *Discuss the charts:*
2. *Recreate the charts:* that involves to get the data, check that it is the right data and re-create the charts (to the degree that modern charts will allow us to do that). In order to be able to check the data for correctness, we have to digitize the information provided in the chart (by measuring the relevant geometric objects).
3. *Re-display the data:* some of the visualization choices made in the original charts are cognitively questionable. Re-displaying the data also allows us to introduce a spatial component.

Both plates #31 and #32 are based on population totals for the 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, we are making use of the 1% microsample of the ninth census provided by the Integrated Public Use Microdata Series (IPUMS-USA) provided through the Minnesota Population Center (Ruggles et al. 2015). Using 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 applying the proportions gained from the sample.

*Corresponding Author

Email addresses: hofmann@mail.iastate.edu (Heike Hofmann), rgoluch@mail.iastate.edu (Ryan Goluch)

Chart Discussion

The charts in the Statisitcal Atlas were created using extremely high-precision methods for the time it was published. Color images were produced by Julius Bien's publishing house (Associates, n.d.,(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 miniscule fraction of the time it would take to draw the same plot by hand.

XXX - I'm sure there are better reproducible research references. I don't know about citing the knitr book alone (don't want to be tool specific) but I've added a couple of placeholder references.

Reproducible research is a frequent topic of discussion in data visualization and data science(Donoho 2010,Baggerly and Berry (2011),Xie (2015)). This paper sets out to reproduce with as much fidelity as possible 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 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 addiitonal depth to the 1870 statistical atlas graphics without investing hundreds of hours of artistic work for each additional map and chart.

Plate #31: Church Accommodations

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

Figure 2 shows a miniature of the chart published as plate #32 in the Statistical Atlas of 1874 (Walker 1874) produced from data collected in the 9th US Census. The chart is set-up in form of small multiples (Tufte 1991), also known as lattice or trellis plots (Becker, Cleveland, and Shyu 1996), one for each state and an enlarged plot as with an overview of the nation-wide aggregates. States are represented by squares of the same size, representing "the total population over 10 years of age", as detailed in the zoom-in in Figure 3, which shows the description at the top of the plate.

With the help of the description and the legend of Figures 3 and 4, we can interpret the details of each of the squares at the example of Figure 5. This figure shows an overview of type of occupation by gender across the US in 1870. It is essentially a mosaic (Hartigan and Kleiner 1981, Friendly (2002), Wickham and Hofmann (2011)) or Marimekko plot [citation?] of type of occupation (horizontal) and gender (vertical), but with a twist: 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. The choice to show this part of the population by a band around is somewhat unfortunate, as it breaks the overall metaphor of the mosaicplot and thereby prevents any direct comparisons across charts except for area comparisons, which are cognitively harder and more error prone than comparisons of lengths (Cleveland and McGill 1984). It also masks the size of the population that is thus *unaccounted* for by visually cutting it into a quarter of the size it actually is. The percentage of unaccounted individuals is at about 30% nation-wide higher than any of the other groups. It is also made up of about 97% women and girls.

The Data

The data used to recreate the visualization of plate #32 is retrieved from Minnesota Population Center (2016), in particular, from table NT13 on "Employed Population by Occupation by Age by Sex". These numbers are state-level aggregates of population numbers by occupation, gender, and sex. We aggregate across ages to get the values included here. The size of the population not "gainfully employed" or attending school is based on estimates based on the 1% IPUMS sample. Figure 11 shows that the numbers combined from NT13 and the microsample closely match the information on plate #32. For all occupation levels and school attendance the numbers are *very* close. For population not accounted for, the numbers are estimated from the 1% IPUMS microsample. this inflates the variability in these numbers, but the relationship to the visual measurements is still very strong.

Big picture goals

- **reproducibility**: can we re-produce results, i.e. images, of the Statistical Atlas based on the resources we have access to today? The resources are: ipums, NHGIS, the high resolution images of the digitized version of the Statistical Atlas provided by the National Library of Congress.

Maps

The R package `USAboundaries` (Mullen 2016) provides historical US state borders. This is used to map the US in 1870 in Figure 9.

Acknowledgment

Software used:

- for creating mosaic plots (Jeppson, Hofmann, and Cook 2017),
- for creating charts in general (Wickham 2016),
- for combining the text, data and the code into a single document to ensure a higher chance of future reproducibility (Xie 2015)

References

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

Baggerly, Keith A, and Donald A Berry. 2011. "Reproducible Research." *Amstat News*. http://magazine.amstat.org/blog/2011/01/01/scipolicyjan11/.

Becker, Richard A., William S. Cleveland, and Ming-Jen Shyu. 1996. "The Visual Design and Control of Trellis Display." *Journal of Computational and Graphical Statistics* 5 (2): 123–55.

Cleveland, William S., and Robert McGill. 1984. "Graphical Perception: Theory, Experimentation and Application to the Development of Graphical Methods." *Journal of the American Statistical Association* 79 (387): 531–54.

Donoho, David L. 2010. "An Invitation to Reproducible Computational Research." *Biostatistics* 11 (3): 385. doi:10.1093/biostatistics/kxq028.

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

General Research Division. accessed in Feb 2017. *Tableau Figuratif Du Mouvement Commercial Du Canal Du*

Centre En 1844 Plate 3. The New York Public Library. doi:http://digitalcollections.nypl.org/items/d8979ef0-ee85-0131-9ccb-58d385a7bbd0.

Hartigan, J. A., and B. Kleiner. 1981. "Mosaics for Contingency Tables." In *Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface*, 268–73. Fairfax Station, VA: Interface Foundation of North America, Inc.

Jeppson, Haley, Heike Hofmann, and Di Cook. 2017. *ggmosaic: Mosaic Plots in the 'ggplot2' Framework*. http://github.com/haleyjeppson/ggmosaic.

Mayr, Georg von. 1877. *Die Gesetzmässigkeit Im Gesellschaftsleben, Statistische Studien*. München: Oldenbourg Verlag.

Minard, Charles Joseph. 1844. *Tableaux Figuratifs de La Circulation de Quelques Chemins de Fer, Lith. (N.s.)*. doi:ENPC: 5860/C351, 5299/C307.

Minnesota Population Center. 2016. *National Historical Geographic Information System (Nhgis), Version 11.0 [Database]*. Minneapolis: University of Minnesota: https://www.nhgis.org/. doi:10.18128/D050.V11.0.

Mullen, Lincoln. 2016. *USAboundaries: Historical and Contemporary Boundaries of the United States of America*. https://CRAN.R-project.org/package=USAboundaries.

Ruggles, Steven, Katie Genadek, Ronald Goeken, Josiah Grover, and Matthew Sobek. 2015. *Integrated Public Use Microdata Series: Version 6.0*. Minneapolis: University of Minnesota. doi:10.18128/D010.V6.0.

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

Walker, Francis Amasa. 1874. "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/.

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

Wickham, Hadley, and Heike Hofmann. 2011. "Product Plots." *IEEE Transactions on Visualization and Computer Graphics (Proc. InfoVis '11)* 17 (12): 2223–30.

Xie, Yihui. 2015. *Dynamic Documents with R and knitr*. 2nd ed. Chapman; Hall/CRC.

1875. *The North American Review* 121 (249). University of Northern Iowa: 437–42. http://www.jstor.org/stable/25109950.

List of Figures

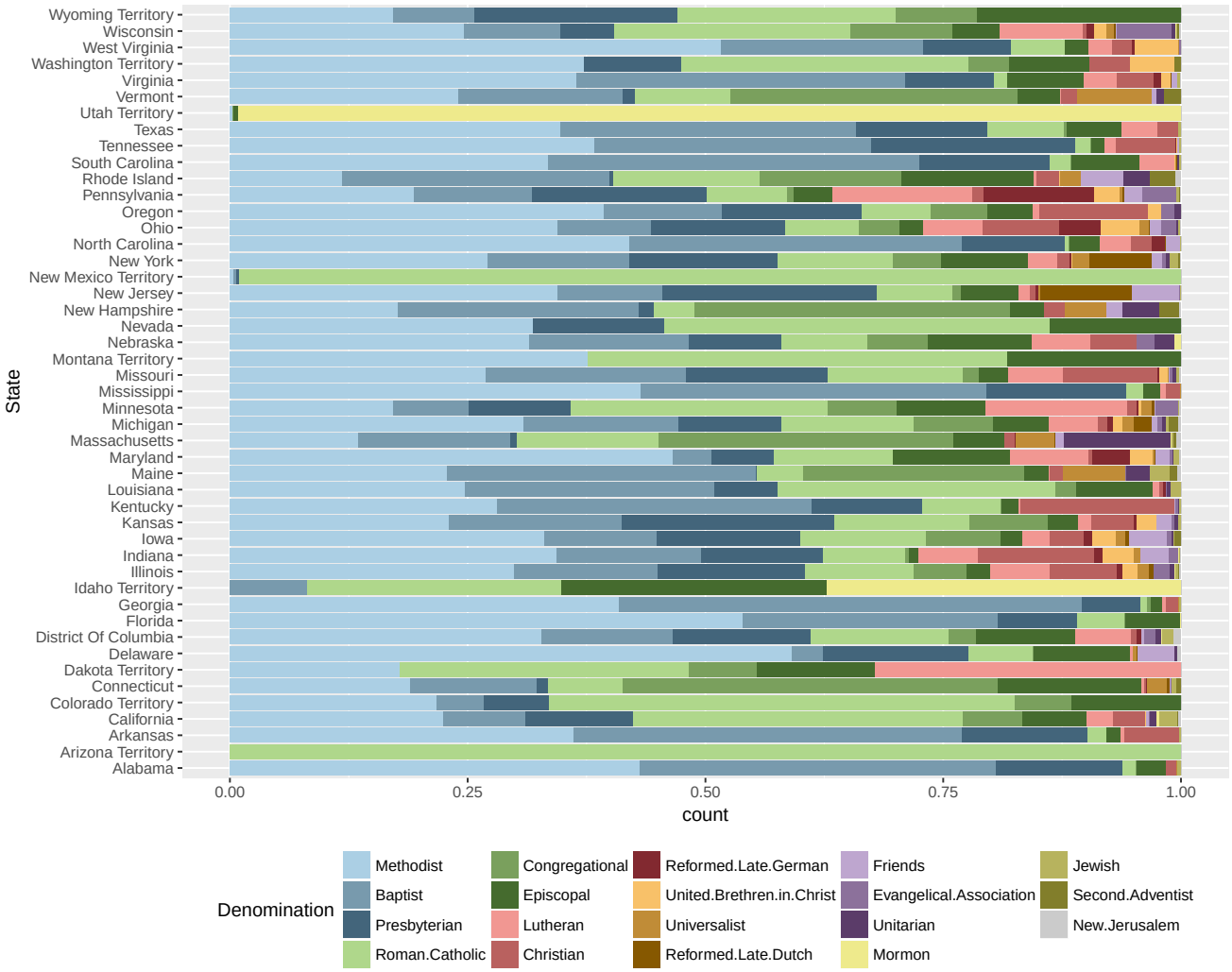


Figure 1: Proportion of religious sittings by Denomination

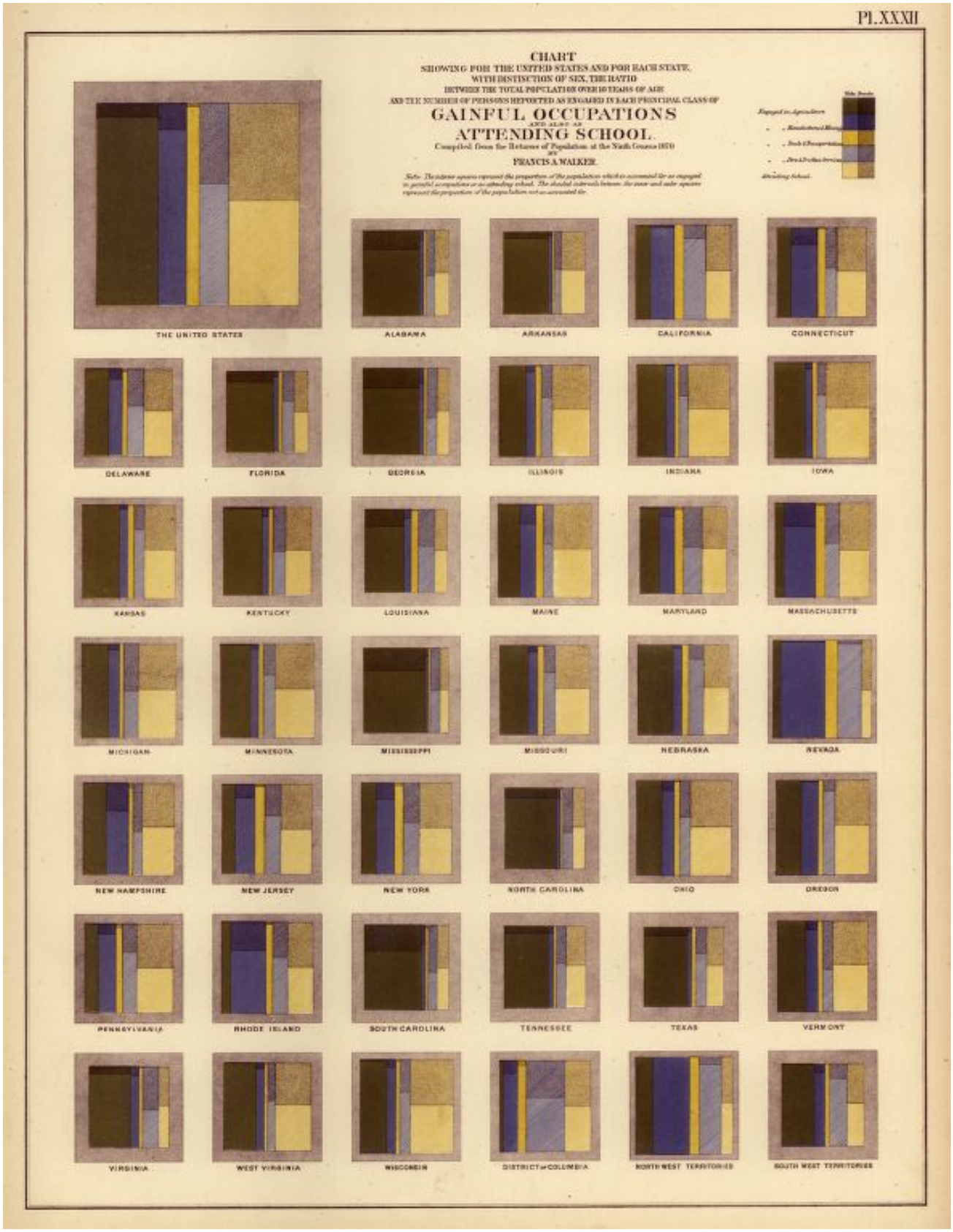


Figure 2: Plate #32 from the Statistical Atlas of 1874: Gender ratio of population over the age of 10 in different types of occupation.

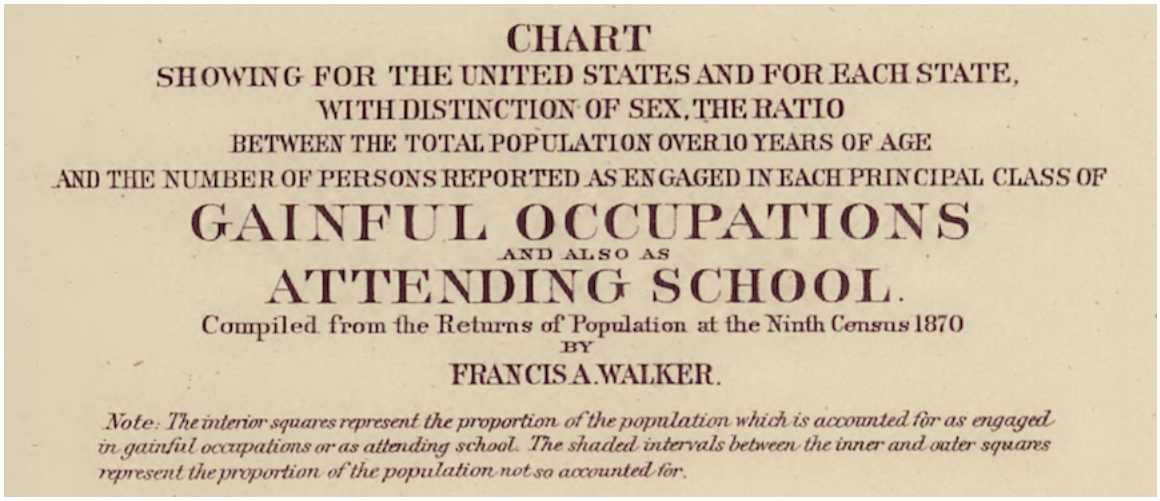


Figure 3: Zoom-in to the description section of plate #32.

Figure 4: Zoom-in to the legend section of plate #32

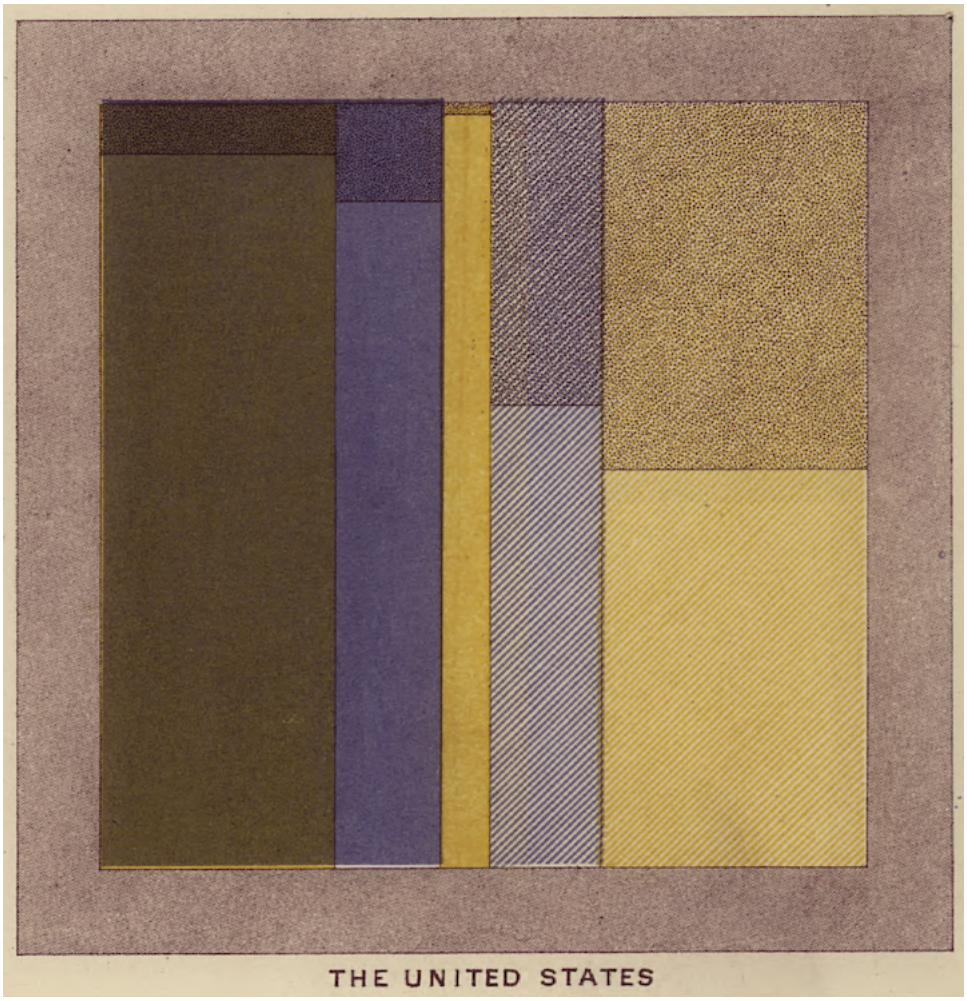


Figure 5: Zoom-in to the overview of the US wide distribution of genders across occupations.

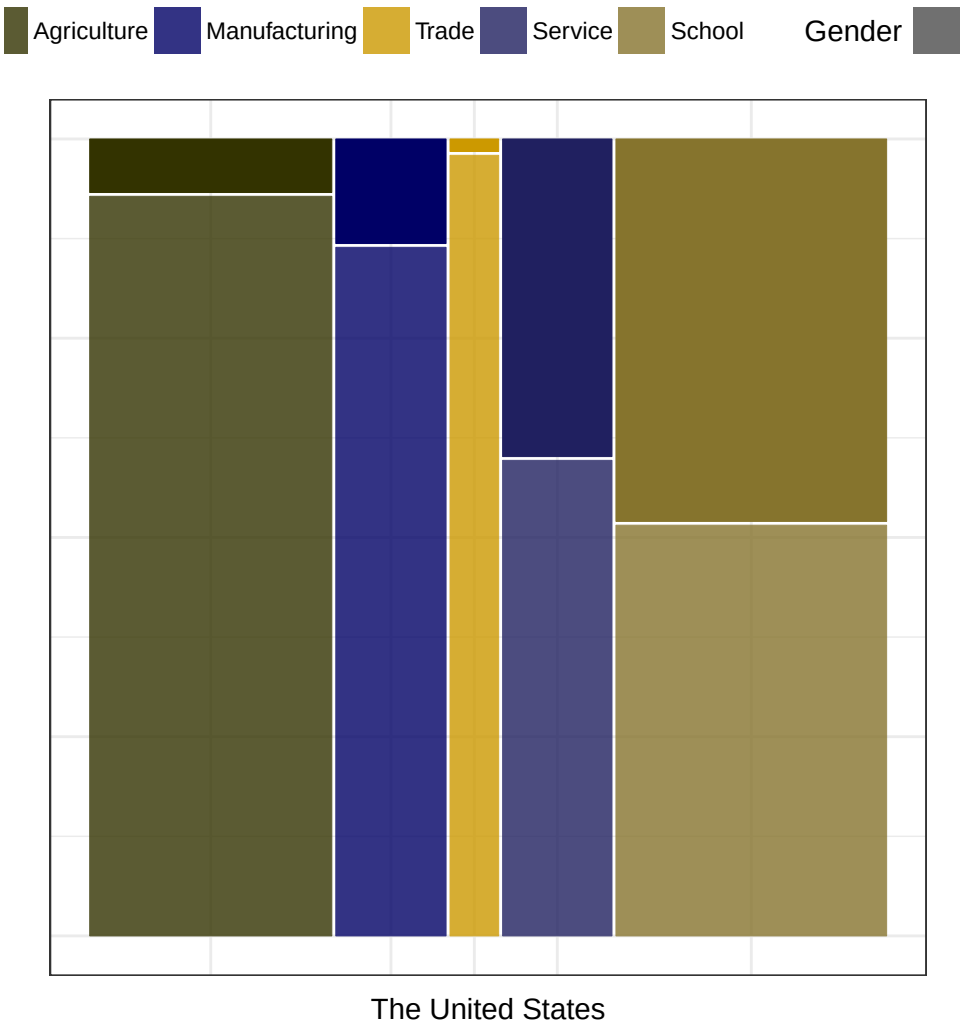


Figure 6: Recreation of the mosaicplot based on gainfully employed population over ten.

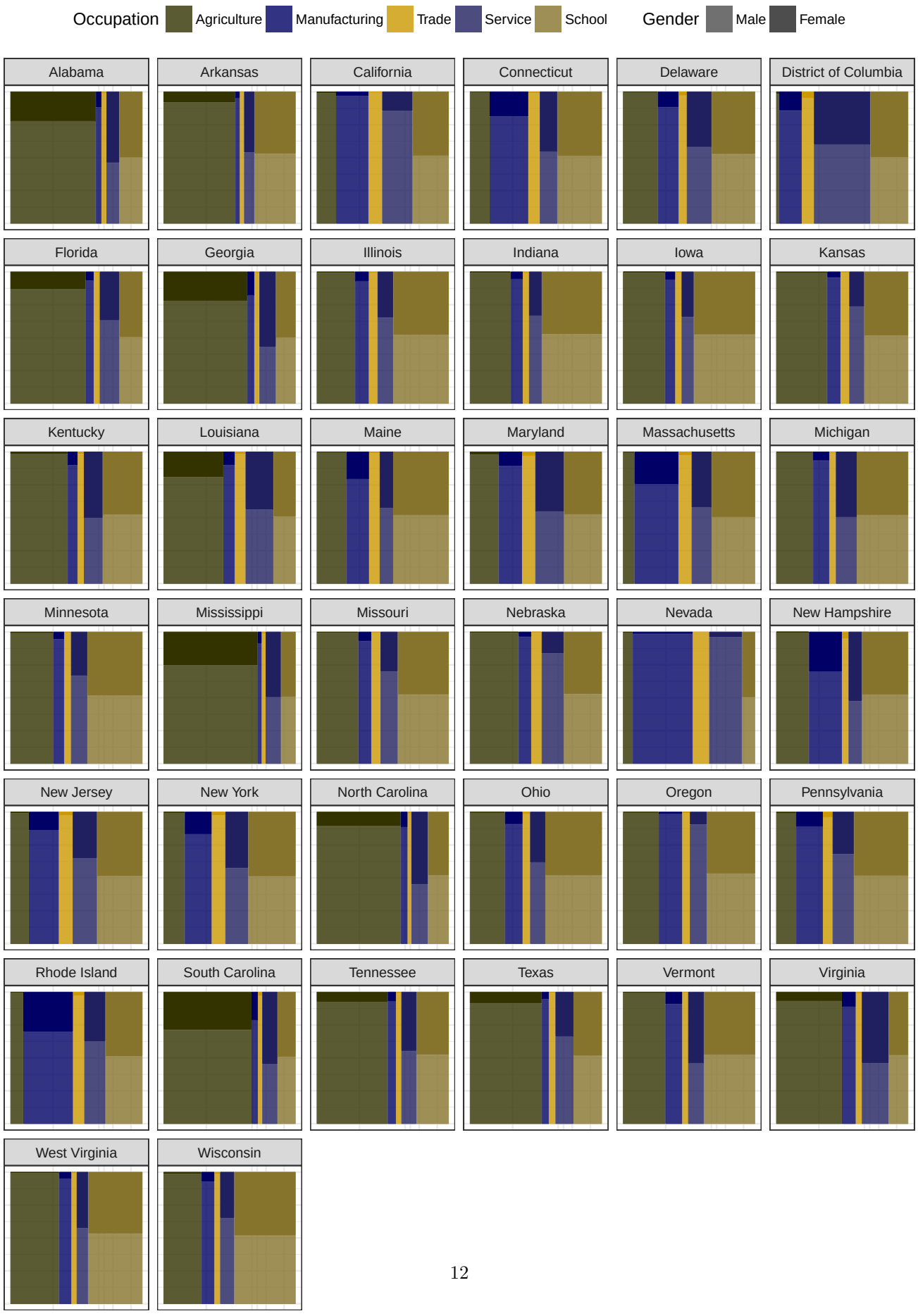


Figure 7: Recreation of mosaics of gainful occupation by states.

Figure 8: Mosaics of gainful occupation by territories.

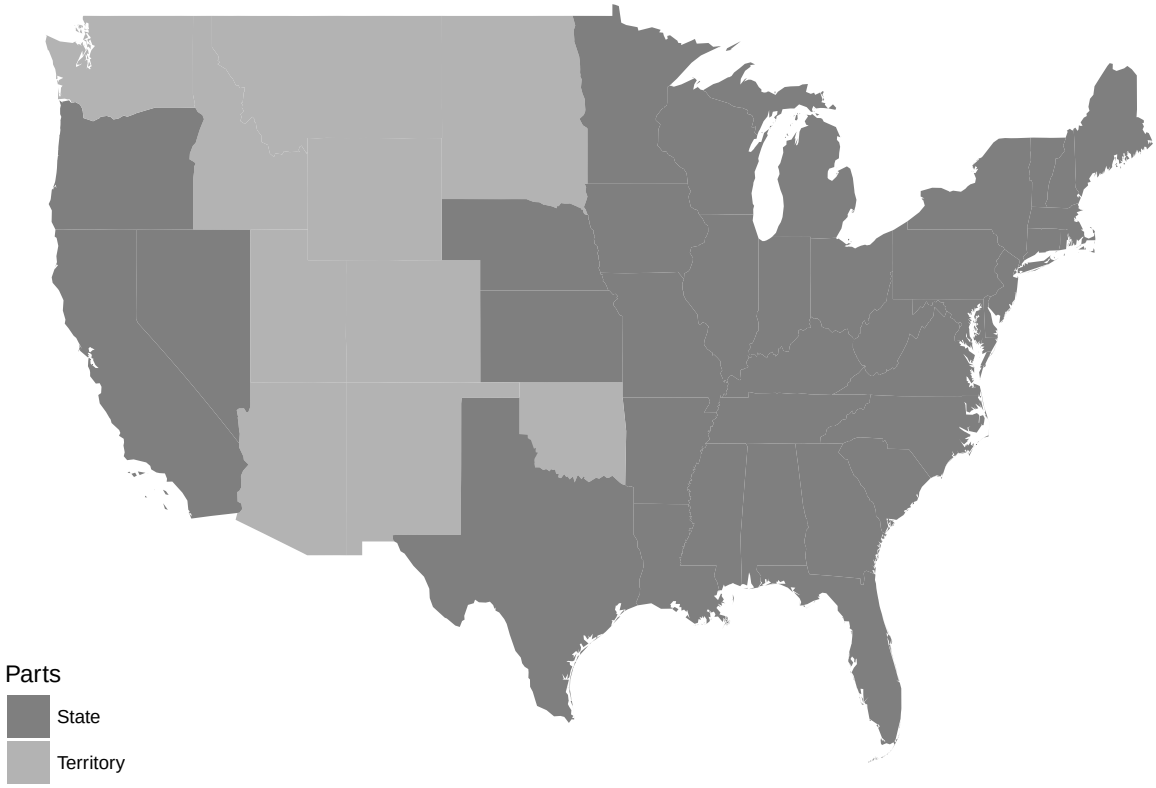


Figure 9: Map of States in the US in 1870

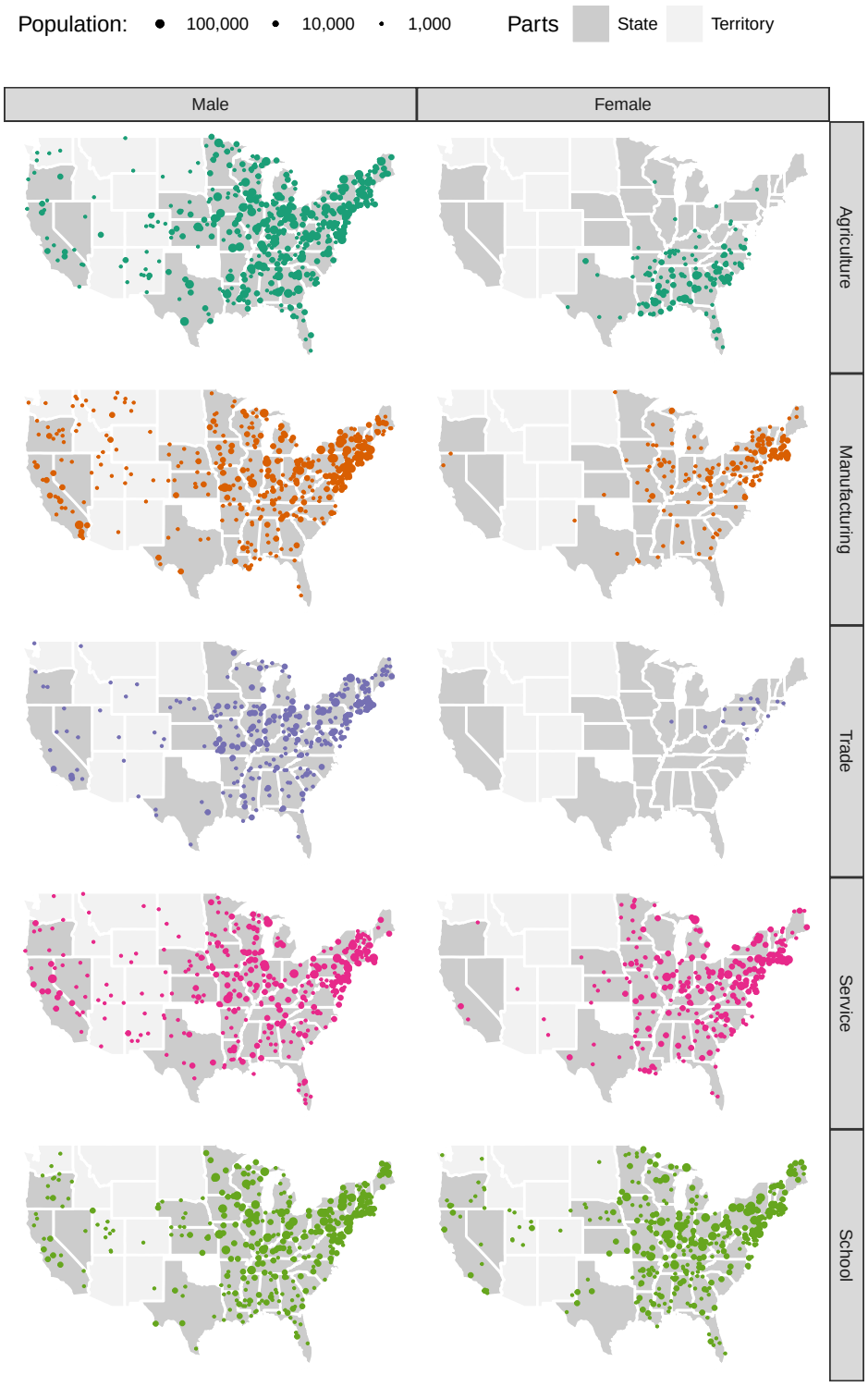


Figure 10: Density plots: each dot represents (according to its size) the number of people employed in each occupation or going to school. For women in particular, strong geographic patterns emerge.

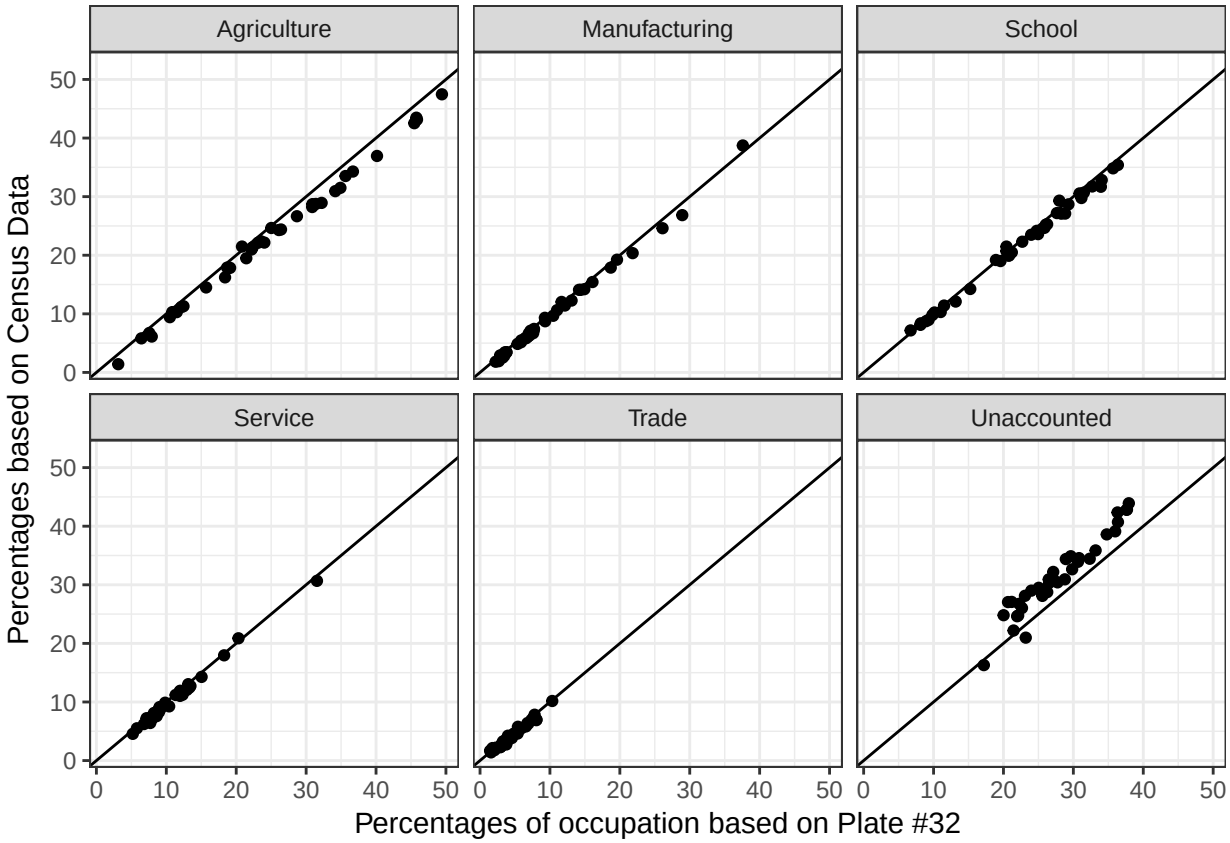


Figure 11: Set of scatterplots showing a comparison of estimates of occupation percentages based on the Census Data (y) and Chart measurements (x).

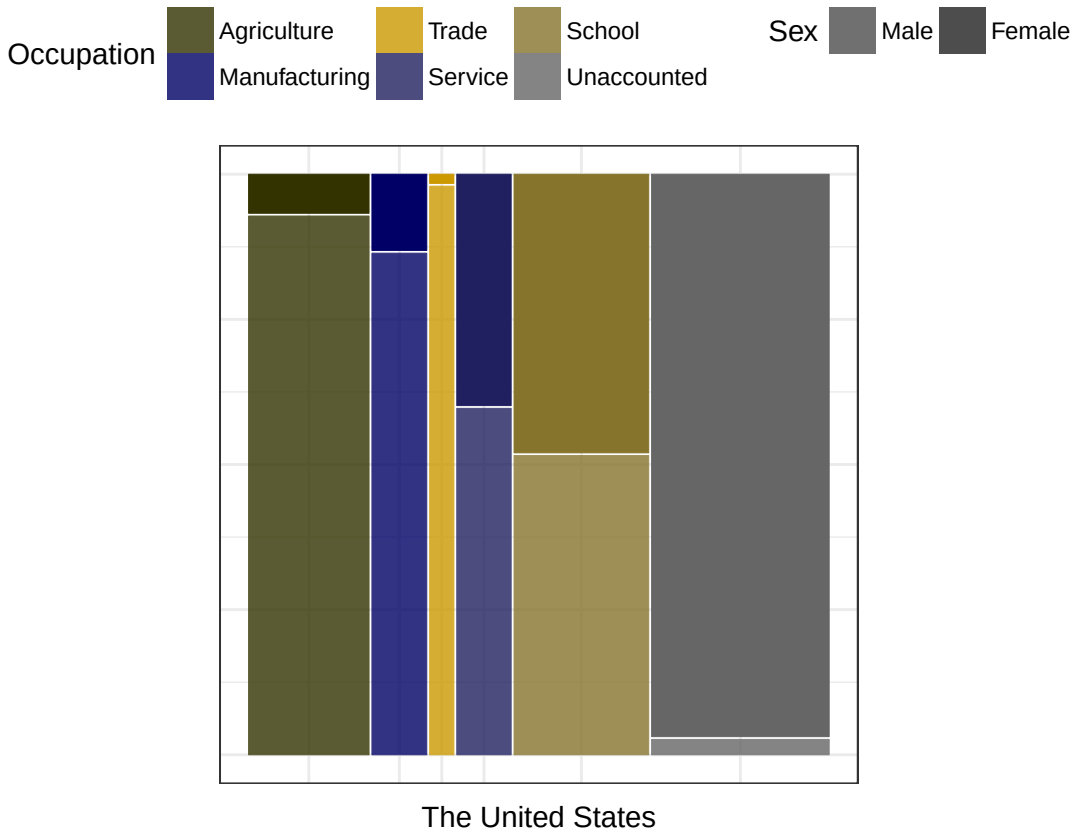
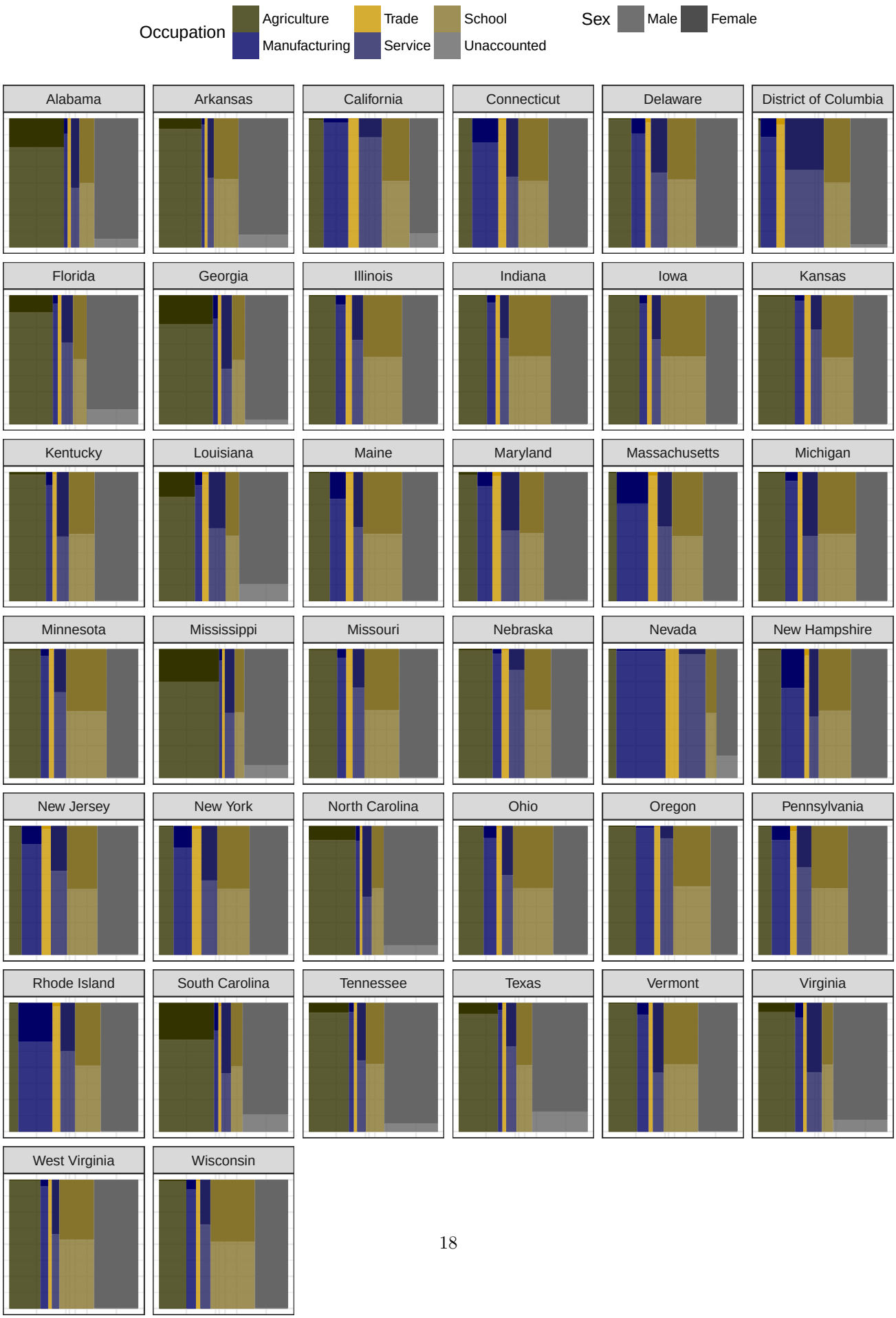


Figure 12: Mosaicplot of the gender ratio in different occupations of the population ten years of age and above.


Occupation
Agriculture
Manufacturing
Trade
Service
School
Unaccounted
Sex
Male
Female
Alabama
Arkansas
California
Connecticut
Delaware
District of Columbia
Florida
Georgia
Illinois
Indiana
Iowa
Kansas
Kentucky
Louisiana
Maine
Maryland
Massachusetts
Michigan
Minnesota
Mississippi
Missouri
Nebraska
Nevada
New Hampshire
New Jersey
New York
North Carolina
Ohio
Oregon
Pennsylvania
Rhode Island
South Carolina
Tennessee
Texas
Vermont
Virginia
West Virginia
Wisconsin