2017

Methods for exploring and presenting contingency tables: A case study visualizing the 1949 Great Britain occupational mobility table

Millicent Alexa Grant
Iowa State University

Methods for exploring and presenting contingency tables:

A case study visualizing the 1949 Great Britain occupational mobility table

by

Millicent Alexa Grant

A thesis submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Major: Sociology

Program of Study Committee:
Shawn Dorius, Major Professor
Cassandra Dorius
J. Arbuckle

The student author, whose presentation of the scholarship herein was approved by the program of study committee, is solely responsible for the content of this dissertation/thesis. The Graduate College will ensure this dissertation/thesis is globally accessible and will not permit alterations after a degree is conferred.

Iowa State University

Ames, Iowa

2017

DEDICATION

I would like to dedicate this thesis to my sister Madison whose faith in me helped me to complete this work. I would also like to thank all of my friends and family for their loving guidance and support during the writing of this work.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

ACKNOWLEDGEMENTS

I would like to take this opportunity to express my thanks to those who helped me with various aspects of conducting research and the writing of this thesis. First and foremost, Dr. Shawn Dorius for his patience, guidance and support throughout this research and the writing of this thesis. I would also like to thank my committee members for their efforts and contributions to this work: Dr. Cassandra Dorius and Dr. J. Arbuckle. I would additionally like to thank Deborah Burns for her support as a writing consultant and friend throughout this process.

ABSTRACT

In recent years, researchers in the social sciences have re-embraced data visualization as a tool for exploring data and communicating results. This development has focused primarily on quantitative data with continuous variables. In general, there has been little work toward establishing a standard for categorical data. This paper tackles a small, but important, component of data visualizations for categorical data: data visualization of contingency tables. The purpose of this paper is to review visualization techniques used for categorical data to determine which techniques are appropriate for contingency tables by ensuring the graphics follow the standards that have been established for continuous data.

CHAPTER 1. OVERVIEW

1.1 Introduction

A common strategy employed by social scientists in the study of the relationship between two categorical variables is the analysis of a contingency table or cross-classification table. A contingency table is a table of frequencies of observations cross-classified by two or more variables. Researchers often analyze these tables using log-linear models and almost exclusively report their results in tabular format. However, I propose that this customary practice be changed. In addition to tables, social scientists should present data visualizations. Furthermore, these graphical displays should be used not only to present information for publication purposes but used by the researchers themselves to develop a better understanding of the data.

There are many definitions of visualization from the simple definition "the presentation of data in a pictorial or graphical format" (SAS 2016) to the more complex definition "the use of computer-supported, interactive visual representations of data to amplify cognition" (Card, Mackinlay, & Shneiderman 1998). A popular definition comes from Edward Tufte who defines data visualization as "[the visual display of] measured quantities by means of the combined use of points, lines, a coordinate system, numbers, symbols, words, shading, and color" (2001). In this paper, I use a combination of Tufte's definition and Card, Mackinlay, & Shneiderman's definition. Specifically, I define data visualization as the visual display of measured quantities by means of the combined use of points, lines, a coordinate system, numbers, symbols, words, shading, and color to amplify cognition. This definition captures two key points. First, it captures the idea that data visualization is a representation of numeric information, and second, it encompasses the purpose of visualization to facilitate understanding of the numeric information.

In an effort to be viewed as a discipline that observes scientific rigor, sociologists have become dependent on using numerical tables to represent quantitative data and view visualizations as

merely pictures rather than as a powerful tool of insight and communication. The information provided in tables has not been further highlighted with visual displays. In essence, sociologists have forgotten the purpose of visualization. In a review of data visualization in sociology, Healy and Moody (2014) provide historical evidence that shows that during the early years of the field, data visualization was embraced as a way to obtain a better understanding of the data. It is time that sociologists follow in the footsteps of those researchers and take advantage of what data visualization has to offer.

Much of the data visualization techniques that exist are designed for continuous variables, and hence, a well-defined standard of practice has already been established for these type of variables. Unfortunately, the same is not true for categorical variables. Although, graphical methods have been developed for categorical variables, a standard has yet to be fully established outlining best practices for producing insightful visual displays for this type of data. To address this issue, I will discuss data visualizations designed for categorical data represented in contingency tables. As an example of the methods, I will present a case study using the visualization techniques to analyze a mobility table which is a special type of contingency table where the table has an equal number of categories for each variable. Note the methods that will be discussed can be applied to any contingency table; I focus on mobility tables to provide a substantive background to the discussion. This subject matter also provides a great example of the dominance of the tabular format.

Many of the prominent social mobility methodologists present results in tabular format. Over the years, they have scarcely included data visualizations. This point is evident with a review of publications that discuss this type of analysis. For instance, in Hauser's (1980) article, he discusses models that can be used to analyze mobility table and proceeds with an example using a mobility table of American sons and fathers. The author does not include any visualizations, but his article does contain nineteen tables. The tables do contain pertinent information, however, the inclusion of visualizations may have furthered his arguments without giving his readers table fatigue. Hout's (1988) article contains seven rather lengthy tables and two visualizations, which is an improvement. However, it demonstrates the dominance of tables as the means of delivering analysis results. In

1998, Goodman and Hout (1998) published an article that introduced the value of graphical displays to assist in the analysis and understanding of model fit results. Unfortunately, they use many of the same type of graphics, and the visualizations may serve better for technical users and not for the presentation of results to non-technical users. These examples were published when producing graphically excellent visualizations was not easy to do given the technology available at the time. Reviewing more recent articles reveals that although graphical displays were introduced, tables are still the dominate method to present results. This point will be discussed further at the end of this chapter.

In this paper, I determine which visualizations are most helpful for analyzing data that can be represented in a contingency table by using theories of data visualizations to obtain guidelines for creating excellent graphs. The first chapter contains a discussion of contingency tables, their use in understanding mobility, and how analysis of contingency tables are often reported in sociological publications. The second chapter concerns data visualization. Specifically, the chapter provides a brief history of data visualization with special attention given to those innovations important to the social sciences, theories of data visualizations, and data visualization methods. The third chapter discusses the data analyzed for the case study on the 1949 Great Britain Occupational Mobility Table. The last chapter contains the results of the case study and discusses the strengths and weaknesses of the various visualizations.

1.2 Contingency Tables

Contingency tables are a popular statistical tool used by social scientists because they do not require strict distributional assumptions, they are relatively simple to implement, interpretation is fairly straightforward, and the results can provide powerful quantitative insights into complex interactions. The analysis of categorical data examines relationships or associations among a set of categorical variables with the purpose of determining if the distribution of one variable changes the distribution of one or more other variables (Rosenthal 2011). In this section, I will discuss the terminology, notation, and distribution properties of contingency tables followed by a brief

discussion of mobility tables to demonstrate the unique characteristics that differentiate mobility tables from other contingency tables.

Let X and Y denote two categorical variables, X with R categories and Y with C categories. Classifications of subjects on both variables have RC possible combinations. The cells of the table represent these possible outcomes. Let $\{\pi_{ij}\}$ denote the probability that (X,Y) occurs in the cell in row i and column j. The probability distribution $\{\pi_{ij}\}$ is the joint distribution of X and Y. The marginal distributions are the row $\{\pi_{i.}\}$ and column $\{\pi_{.j}\}$ totals that result from summing the joint probabilities where the subscript "." denotes the sum over that index; that is,

$$\pi_{i.} = \sum_j \pi_{ij} \quad \text{and} \quad \pi_{j.} = \sum_i \pi_{ij} \tag{1.1}$$

where $\sum_i \pi_{i.} = \sum_j \pi_{.j} = \sum_i \sum_j \pi_{ij} = 1$. Table 1.1 provides an example of such a table.

Table 1.1: R-by-C Contingency Table

	Y (Columns)				
X (Rows)	1	2	...	C	Total
1	n_{11}	n_{12}	...	n_{1C}	$n_{1.}$
2	n_{21}	n_{22}	...	n_{2C}	$n_{2.}$
⋮	⋮	⋮	...	⋮	⋮
R	n_{R1}	n_{R2}	...	n_{RC}	$n_{R.}$
Total	$n_{.1}$	$n_{.2}$	...	$n_{.C}$	N

If X and Y are both response variables, the focus of the analysis is their joint distribution. However, if one of the variables, let's say, Y, is a response variable and the other variable, X, is an explanatory variable, the focus of the analysis is the conditional distribution of Y and how it changes as the categories of X change (Agresti 2013). The row totals (i.e. $n_{R.}$ in Table 1.2) and column totals (i.e. $n_{.C}$ in Table 1.1) describe the distribution of each variable disregarding the other. Given that a subject is classified in row i of X, let $\{\pi_{j|i}\}$ denote the probability of classification in column j of Y for all $j = 1, \ldots, C$. Then,

$$\pi_{j|i} = \pi_{ij}/\pi_{i.} \tag{1.2}$$

denotes the conditional distribution of Y given the i^{th} level of X. This setup is often how variables in mobility tables are treated.

1.3 A Special Case of the Contingency Table

Square tables, where the row and column variables have the same number of categories, comprise a special case of contingency tables, and describe the unique characteristic of social mobility tables or intergenerational occupation mobility tables. These type of tables cross-classify people "according to their occupations at two points in time. The earlier point is usually referred to as the origin; the later point is known as the destination" (Hout 1988) and are used to study social stratification.

Social stratification describes the hierarchical ranking of individuals that defines the social structure within a society (Parsons 1940). There are three key concepts used to understand social stratification position, status, and strata. Position refers to a person's place in the social system as a whole or can be thought of in terms of occupation as the person's role in an organization. For the remainder of this paper, I use the later definition of position. Status is defined as "one's generalized position (the sum total of one's major positions) in the structure" (Davis 1942). Stratum refers to a group of people in a society that have roughly the same status. For example, Dr. Joe's position is doctor in hospital X whose status is classified as professional and who is part of the upper stratum. Social mobility is an avenue for looking at how individuals move between strata and can be examined with mobility tables.

Mobility tables allow researchers to gauge the amount of openness present in a society with the idea being that an open society's occupational success is independent of an individual's socioeconomic background (Hout 1988). The examination of a society's openness is done by relating the occupational position of parents to their child's occupational position in adulthood. Some of the common origin variables are father's occupation and first occupation, and some of the common destination variables are current occupation and first occupation. In studies of intergenerational occupational mobility, father's occupation serves as the origin and the son's first or current oc-

cupation serves as the destination. In the study of intragenerational occupational mobility, the first occupation of the participant represents the origin while the destination is represented by the participant's current occupation. The case study in this paper uses data from a intergenerational occupational mobility study. The relationship between father's and son's occupational status is specified in the table by treating the father's occupational status as the origin, or independent variable, and the son's occupational status as the destination, or dependent variable. The cells of the table, then, provide the frequency of fathers and sons that share each combination of occupational classification. The objective is to measure the amount of mobility present in the table.

Early studies of social mobility relied on data collected by national census organizations such as the UK Office of Population Censuses and Surveys and the US Census Bureau. The classic Great Britain 1949 mobility table first examined by D.V. Glass came from the United Kingdom's census organization's Labor Mobility Study. In 1962 and 1973, the United States' census organization collected mobility data as supplements to the March Current Population Survey. The National Opinion Research Center's General Social Survey is a well-known source of mobility data. Recently, mobility data in the Unites States has come from the Panel Study of Income Dynamics (PSID) 1968-1997, "which contains annual descriptions of occupation and industry affiliation for a panel of individuals representative of the population of the United States in each year" (Kambourov & Manovskii 2008).

A dominant part of the methodological research aims to evaluate measures of mobility and/or discover new ways of measuring mobility. There are many different measures of social mobility including absolute mobility rates, relative mobility rates, inflow and outflow rates, and mobility ratios. Absolute mobility rates (also known as total mobility rates or just mobility rates) are defined as the observed total number of people that move between classes or the number of individuals in the cells off the main diagonal over the total number. Relative mobility rates are a measure of the association between father's occupational status and son's occupational status. It is taken as a measure of social fluidity. Inflow rates or inflow percentages are the row percentages taken from the mobility table, and outflow rates or outflow percentages are the column percentages taken

from the mobility table. Inflow and outflow percentages reveal information about the flow of labor. Specifically, inflow percentages represent the labor flowing into the given destination occupation, and outflow percentages represent the labor flowing out of the given origin occupation. Mobility ratios are of the observed frequencies and the expected frequencies under the model of statistical independence also known as the model of perfect mobility. The reason the ratio is appealing is that "as the ratio of an observed quantity to that expected when there is no association in the table, it suggests itself as an index of the extent of association" (Hout 1983).

Unfortunately, there are disadvantages to most of these measures of social mobility. A major disadvantage of using absolute mobility to measure social mobility is that the rate is heavily affected by the marginal distribution. It is also dependent on how the occupational categories are formed. Relative mobility rates are calculated using odds ratio which are invariant to proportional changes in marginal distributions. Mobility ratios are also flawed because mobility data often does not exhibit patterns of perfect mobility making it an ill choice as an index of association. Inflow and outflow percentages are only informative at a low level of analysis making them a weak measure of social mobility. Unsurprisingly, the problems with absolute mobility rates, inflow and outflow percentages, and mobility ratios has led to the increased use of odds ratio. Most of the statistical models developed for the analysis of social mobility lead to simple interpretation of the association parameters in terms of local odds ratios. Such models will be discussed with greater detail in the methods section of this paper. For now, it is important to note that most researchers use multiplicative log models to examine the data (Goodman, Hauser, Erikson & Goldenthrope, Xie, and Hout). However, there have been a few who forgo the use of mobility tables and opt for a regression analysis such as Mazumder & Acosta or path analysis such as Blau and Duncan. In addition to different statistical techniques used to examine mobility tables, researchers have also varied in their focus of study.

There have been many studies where researchers have focused on fathers and sons with no comparisons made over time or space such as D.V. Glass' (1954) study of social mobility in the Great Britain. However, as time as progressed, comparative studies have become much more common.

Social mobility across different countries has been studied such as Berent's (Glass 1954) study of social mobility and marriage in England and Wales, Ziegel and Hall (Glass 1954) comparison of social mobility in England, Wales, Italy, France and the US, and Grusky and Hauser (1984) comparison of social mobility in 16 different countries. The key finding across the literature is that relative social mobility remained constant in industrialized nations. A common feature in the literature, recently, is to try to explain patterns in mobility by analyzing additional variables that could contribute to changes in mobility such as education, race, gender, and age. Erikson & Goldenthrope (2002) found that educational attainment had a major impact on mobility when considering education in terms of the level of qualification, i.e. vocational vs. academic; Featherman & Hauser (1974) determined that the differences between whites and nonwhites cannot be attributed to their low-income origins but rather to their unfavorable patterns of occupational mobility; Hout (1988) reports that when using unbiased methods those researchers who study the differences in social mobility between men and women find that there are differences among men and women with white-collar occupations and between farm and non-farm classes; and Mazumder & Acosta (2015) find that there exist a life cycle bias where the age of the son or father can mislead the amount of mobility present in the data. These studies and others have furthered the field to obtain a more complete understanding of occupational social mobility.

This review demonstrates that the discipline has been diverse in terms of the data and methods used to study mobility. What has lacked in variety is how the results are presented. The tabular format continues to be the most popular method for presenting results. I used Web of Science to identify the thirty most influential articles in the occupational mobility literature to see how researchers reported their results. I define articles as influential based on the number of times the article has been cited, and I narrowed the search by eliminating any articles that excluded both tables and articles since those articles were of a review rather than analysis in nature. I found that sixteen out of the thirty articles contained only tables in the presentation of results, twelve out of the thirty contained both tables and graphics, and two of the thirty contained only graphs. For the researchers who employed both tables and graphs, only two of the articles had the same

number of graphs as tables and the other ten had more tables than graphs. This point is not to say that tables should not be used, but to show their dominance and to suggest that more researchers should include both tables and data visualizations. As Tufte (2006) says, "a deeper understanding of human behavior may well result from integrating a diversity of evidence". It should also be noted that this trend does not exist solely in studies of occupational social mobility. It is prevalent across the entire field of sociology as described by Healy and Moody (2014) in their extensive analysis of visualization methods in sociology. They found the lack of visualization is evident in comparison of natural science journals and sociology journals. Specifically, the authors compare the *American Sociological Review* and the *American Journal of Sociology*, both of which were included in the meta-analysis, to the *Proceedings of the National Academy of Science*, *Science*, and *Nature*. The sociological journals are typically filled with tables and few visuals while the natural science journals centralize articles around a figure. They found that in the early quantitative days of sociology that visualization methods were relatively common. With the advent of surveys and a trend toward statistical quantification, qualitative methods, including visualization techniques, fell in disuse. Healy and Moody also make the point that the relationship between statistics and graphics is not consistently taught or included in the analysis process which results in a new generation of sociologists that cannot fully embrace data visualizations and a continuation of articles that do not include any visual displays. The authors urge sociologists to "think about how visualization could be more effectively integrated into all stages of our work" (2014).

CHAPTER 2. REVIEW OF LITERATURE

2.1 Historic Overview of Data Visualization

To obtain a better understanding of the usefulness of data visualization, it helps to learn its history within the social sciences. Data visualization has progressed as innovations and growth in technology and statistics have grown. At the beginning, the stars were of most interest to scientists who created graphs that display the position of the sun, moon, and planets (Friendly 2005). The next development in the field came in the 1600s when demography and statistics advanced greatly. In 1637, Descartes developed a coordinate system that became critical to the advancement of statistical graphs, and in 1669, Christiaan Huygens created a graphic of a function of life expectancy using data from John Graunt's book *National and Political Observations on the Bills of Mortality* (Friendly & Denis 2001). Huygens' graph marks the first time a graph was used to enhance the understanding of social scientific data (Friendly 2005). The 1700s and 1800s can be described as the rise of empirical problem solving initiated by William Playfair.

William "Playfair is credited with producing the first chartbook of social statistics" (Wainer 2001). Around 1785, he invented the statistical bar chart and creates line and bar charts of economic data. The reason his contributions were so integral to the development of data visualization is that he popularized the use of statistical graphics when the majority of researchers viewed graphics as less than valuable. He favored graphics over tables because they provided a more comparative way to view the data (Tufte, 2001). Playfair continued to advance the field by inventing the pie chart and circle graph. Essentially, most of the statistical graphic forms used today are due to Playfair. He was also known for his time series graphs of finance data. Baron Charles Dupin also contributed to advances in data visualization in the 1800s. He is credited with the use of shading from white to black to show the distribution of literacy in France. It was "the first unclassed choropleth map, and perhaps the first modern-style thematic statistical map" (Friendly & Meyer 2016).

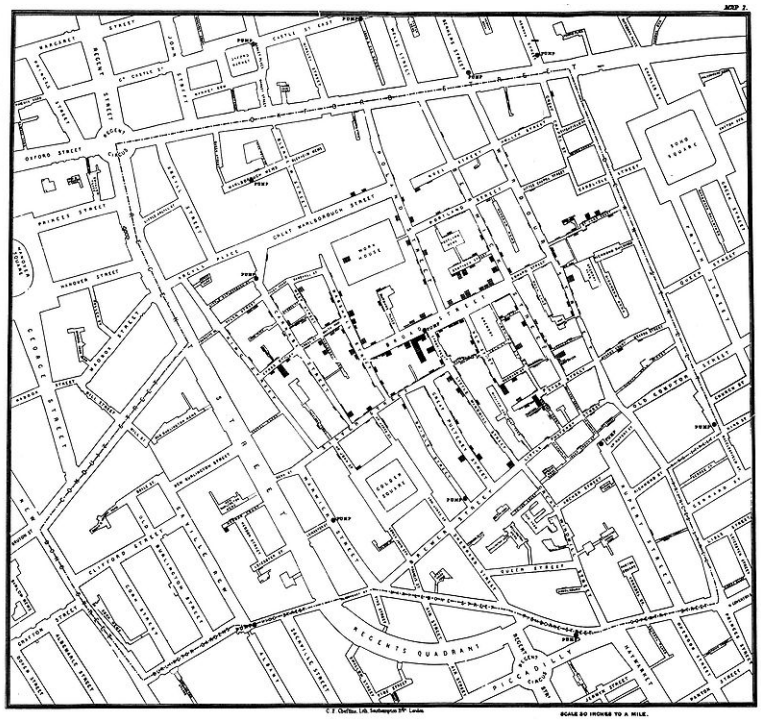


Figure 2.1: John Snow's Map of Cholera Death in Central London, 1854

All of these innovations in graphics along with advances in statistics contributed to the development of official state statistical offices throughout Europe, which translated to a plethora of data relating to social, industrial, commercial, and transportation planning. "With the usefulness of graphical displays for understanding complex data and phenomena established, many new graphical forms were invented and extended to new areas of inquiry, particularly in the social realm" (Friendly & Meyer 2016). One of the first examples of the these innovations came from Dr. John Snow in his investigation of the spread of cholera in London in 1854. Snow doubted the theory that the disease was caused by pollution or bad air. To gain a better understanding on how cholera spread, he plotted the locations of those who died from cholera. His map is shown in Figure 2.1. He concluded that most of the deceased lived near and/or drank from a water pump on Broad Street. After removing the handle from the pump, there was a major decline in people afflicted with cholera (Tufte 2001). His study marked the beginning of epidemiology, and it was a major event in the history of public health and geography.

Charles Minard provides more examples of the graphical innovations during this time period. He established the use of circles and flow lines to visualize different quantitative variables. Often regarded as the best graphic ever, he produces a tableau graphique illustrating the campaign of Napoleon against Russia in 1812 shown below in Figure 2.2. The width of the lines are proportional to the number of surviving soldiers with a beige line showing the path towards Russia and a black line showing the soldiers' return (Tufte 2001 & Tufte 2006). It is of such great importance because Minard was able to communicate the details of what happen during the battle almost entirely through his visualization. He provides a detailed title that provides his credentials, a summary of the image, and the type of diagram illustrated. Minard also includes a paragraph explaining how to interpret the graph and its sources (Tufte 2001 & Tufte 2006). His graphic is exemplary of graphical excellence and graphical integrity. His work shows how beneficial data visualization can be to everyone.

The late 1800s is also of historical significance because sociology makes its debut. The American Journal of Sociology was first published in 1895, and one of the first articles included in the journal was "Immigration and Crime" by Hastings H. Hart. He used bar charts to show the distribution of the population in the U.S. and as a comparative tool to illustrate the difference between two methods of analysis. Antonio Marro incorporates a line graph in his article presented in the fifth volume of the journal. Additionally,"Du Bois's (1898 [1967]) *The Philadelphia Negro* is filled with innovative visualizations, including choropleth maps, table-and-histogram combinations, time series, and others" (Healy & Moody 2014). His project on the Georgian Negro is also exemplary. Figure 2.3 shows a hand drawn horizontal stacked bar chart of the income and expenditure of African-American families in Atlanta, Georgia. Through the choice of color and detailed labelling, he shows the differences among families of different classes. It is a superb example of how social stratification can be displayed visually. The visualization was part of a compilation of materials W. E. B, DuBois displayed at the 1900 Paris Exposition (Smith 1999). He used a series of photographs and visual displays to show the economic, social, and cultural differences among African Americans to combat the presumed superiority of whites based on claims of biological race scientists. By using

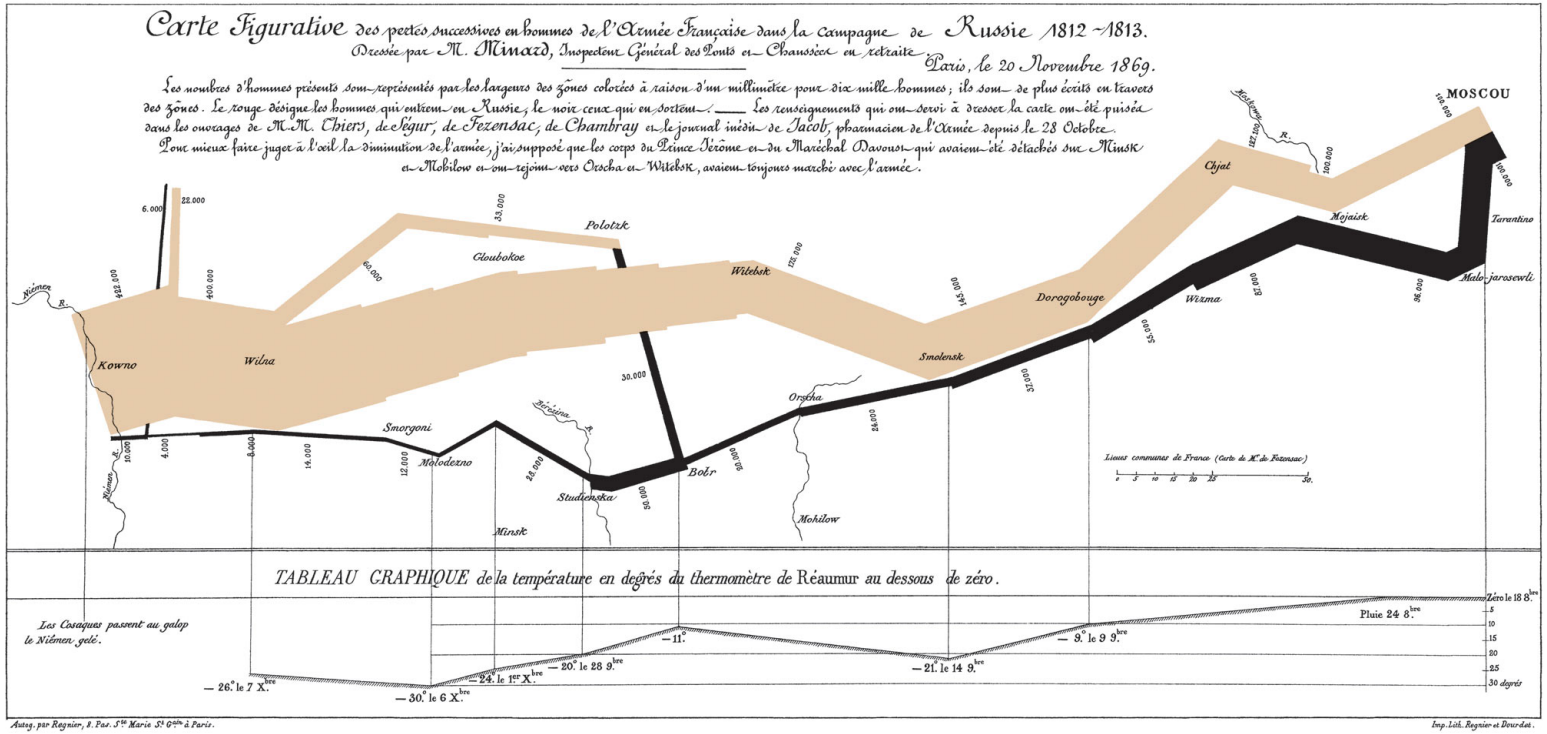


Figure 2.2: Charles Minard's figurative map of the losses in men of the French Army in Russia. Source: Tufte, 2006

a visual technique, he gave viewers direct evidence that the stereotypical views of blacks at the time were wrong. His exposition showed the power of presenting data visually rather than only using a table.

The methods of these sociologists should be of no surprise when also considering the history of sociological research methods during this time. The early years of the discipline consisted of social scientists who were often members of statistical societies or social reform institutions that developed statistically-based social studies. Hence, if visualizations were common to statisticians, they would also be common to sociologists. The substantive topics addressed by these researchers were not physically tangible like those of the physical sciences, so graphics helped explained those things that people cannot physically interact with.

Unfortunately, in the early 1900s, there was a lack of interest in graphical innovations possibly due to a "growth in quantification and formal models" (Friendly & Denis 2001), which was indeed reflected in sociology. Although, there were a few exceptions (Chapin 1924 and Sletto 1936), data visualizations pretty much disappeared from publications in the field. Sociologists embraced tables and suppressed visualizations perhaps to fulfill a sense that the discipline needed to be more legitimate. Whatever the reason, the discipline lagged behind and although statisticians began to re-embrace data visualization in the 1960s, sociology did not, at least not to the same extent.

What I consider the two most important reasons for data visualizations popularity are John W. Tukey's work on exploratory data analysis and the advent of computers. Tukey recognized that data visualizations could be used for model diagnostics and to gain a better understanding of the data before fitting it to a model. He is accredited with the creation of box-and-whisker plots, stem-and-leaf plots, and rootograms, and he helped establish ways to improve the quality of graphical displays (Tufte, 2001 & Wainer 2001). Tukey said, "the best single device for suggesting, and at times answering, questions beyond those originally posed is the graphical display" (Cook, 2015).

The invention of the computer meant that graphics no longer had to be hand drawn, and the development of software and computer systems allowed for dynamic and interactive visualiza-

tions that people across all disciplines have come to appreciate. These advancements explain why graphical displays are starting to again become popular in sociology, but the lack of many articles that include data visualizations shows that these techniques are not being embraced fully. This statement does not mean that sociologists do not use data visualizations at all

2.2 Data Visualization Theories and Methods for Contingency Tables

Traditionally, sociologists have communicated the contents of data through the presentation of tables. However, the amount of information that has become available to sociologists has increased to the point that the presentation of results only in tabular format is no longer an effective way of communicating, especially if using a large dataset. The brief history of data visualization shows that communication through graphs is not new. Today, they are commonplace in a variety of disciplines, but many graphs are produced poorly. "Modern data graphics can do much more than simply substitute for small statistical tables. At their best, graphics are instruments for reasoning about quantitative information. Often the most effective ways to describe, explore, and summarize a set of numbers - even a very large set - is to look at pictures of those numbers" (Tufte 2001). Therefore, it is important to understand theories and methods of data visualizations in order to create graphics that facilitate the analysis of large-scale data and to provide insights that may be missed with a purely quantitative assessment of the data.

2.2.1 Theories of Data Visualizations

Much like its history, theories of data visualization draw from several disciplines. A theory that reflects the cross-disciplinary nature of data visualization is the Gestalt theory. Wertheimer introduced the theory as a way to understand nature, and he developed the theory using his observations not only from psychology, but from the physical sciences and art. He explains, "There are contexts in which what is happening in the whole cannot be deduced from the characteristics of the separate pieces, but conversely; what happened to a part of the whole is, in clear-cut cases, determined by the laws of the inner structure of its whole" (Wertheimer & Riezler 1944). Essentially,

the total understanding of a particular item may not be obtained simply from its parts and vice versa. That is, understanding the whole may or may not give one enough information to understand its parts. The theory has been elaborated to design principles as they relate to visual perception.

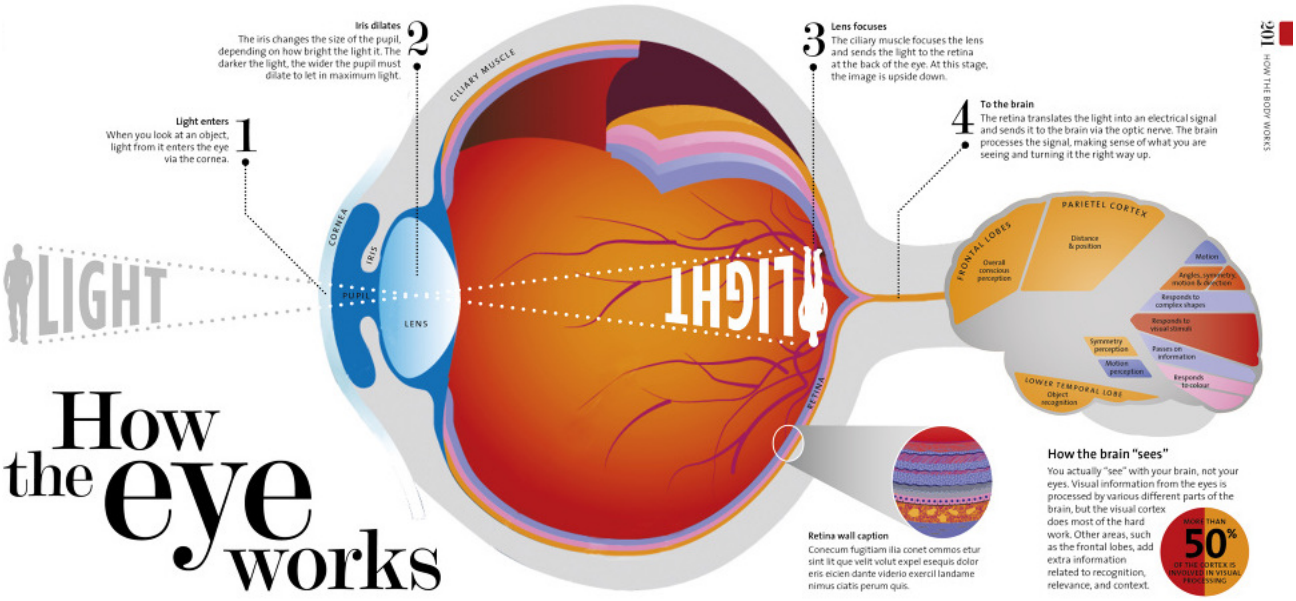


Figure 2.4: Illustration of the Visual Process. Source: Parker 2007

Visual perception has an impact on how human beings interpret information. Therefore, it is important to have a basic understanding of how people see. The reality of an object and what an individual sees when looking at that object can differ based on how the image appears to the eyes. The diagram shown in Figure 2.4 illustrates the visual process. The light rays reflected from an object passes through the cornea of the eye and after the pupil adjusts to the amount of light, the now inverted image reaches the lens and enters the retina where it is sent to the brain for processing (Few 2012). Due to the complexity of this process and the brain, an object may not appear as it does in the real world. For example, the retina contains three types of cones that detect different amounts of red, green, and blue light, but if a person is missing a type of cone, the individual will not be able to make distinctions between all colors meaning the person will be color blind.

Limitations such as these should be accounted for when creating data visualizations to ensure that every viewer can interpret the picture in the same way, and to ensure that happens, researchers can follow various principles including the Gestalt Design Principles. Stephen Few (2006) summarizes these principles succinctly,

- proximity: objects that are close together are perceived as a group
- similarity: objects that share similar attributes are perceived as a group
- enclosure: objects that appear to have a boundary around them are perceived as a group
- closure: open structures are perceived as closed, complete, and regular, whenever there is a way that they can be reasonably interpreted as such
- continuity: objects that are aligned together or appear to be a continuation of one another are perceived as a group
- connection: objects that are connected are perceived as a group
- figure/ground: objects are differentiated from the surrounding area

These concepts are illustrated in Figure 2.5. Each principle explains how human beings perceive objects in relation to each other. The Gestalt Design Principles do not guarantee graphically excellent data visualizations. Additional theories are needed that incorporate statistics and design to achieve that goal.

Graphical excellence is defined as "the efficient communication of complex quantitative ideas" (Tufte 2001), and it means that data visualizations should be clear, informative, and honest. Graphically excellent data visualizations allow a large quantity of data to be interpreted with ease by displaying layers of detail that not only provide a broad overview of the data, but also an intricate view of the data. These types of graphics make viewers think primarily about the substance behind the image, and they follow the fundamental principles of analytical design:

- display comparisons and differences
- show more than one or two variables
- integrate evidence

- thoroughly describe the evidence

The last point implies that all data visualizations should include a detailed title that indicates the author(s) and sponsor(s), cite data sources, include measurement scales, and point out anything unique or important. Obtaining a graphically excellent visualization is not instantaneous but requires an iterative process of revising and editing.

2.2.2 Data Visualization Methods

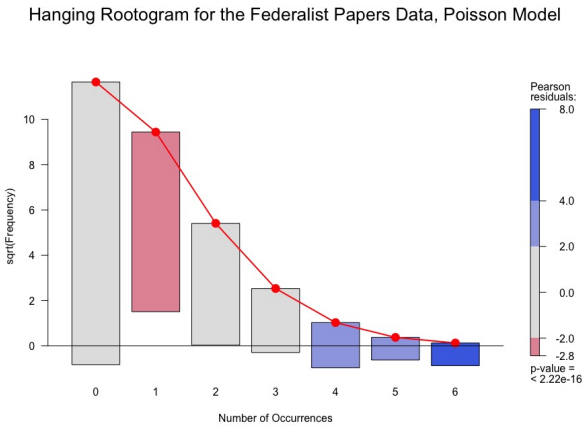
Methodology used for data visualization is dependent on the purpose of the visualization: exploration or presentation. Exploration refers to data visualizations that guide the statistical analysis by providing summaries of responses, facilitate the creation of new hypotheses, provide support of previously formed hypotheses, and check the validity of a chosen model (i.e. model diagnostics). Presentation refers to data visualizations used to communicate characteristics of the data and results to others, especially to non-technical viewers. In both categories, the value of the graphic can be determined by ensuring that the visual demonstrates graphical excellence and graphical integrity. In the case of graphical methods specifically designed for contingency tables, many of the visualizations can be used for both exploration and presentation needs. In this section, I describe a wide range of data visualizations and discuss their advantages and disadvantages.

Data visualizations for contingency tables should be able to reveal any trends, patterns or unexpected properties in the data, help the researcher determine if the data follow a particular probability distribution, and ascertain possible models that could be used to describe the data. The list of data visualizations available to accomplish these goals is extensive, so my discussion highlights just a few of the many options. The graphical displays discussed below are bar plots, rootograms, tile plots, spineplots, sieve diagrams, association plots, fourfold displays, corrplots, and mosaic diagrams.

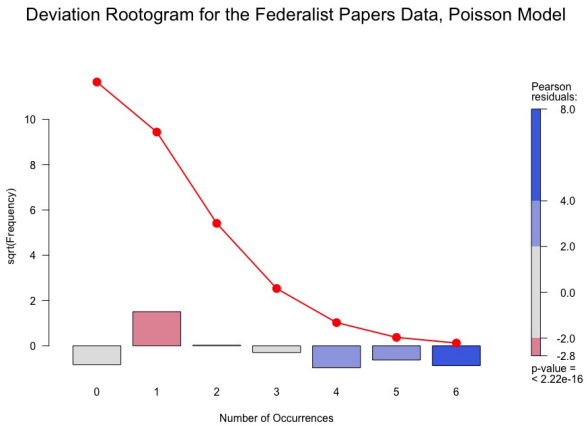
Bar plots visualize the frequency distribution of a dataset, provide an easy way to summarize a large data set in visual form, and expose trends better than tables. Often, group bar charts and stacked bar charts are used when there are multiple categories and more than one variable. In a

grouped bar chart, the bars representing each category are displayed side by side, and in a stacked bar chart the categories are shown in a single bar with different colors used for each category. In both types of bar plots, the height of the bars corresponds to the frequency of each category. Although, bar plots provide a simple way to visualize the data, they do not extend well to more than one dimension, which makes comparisons across and within groups difficult. Also, they do not usually match the tabular structure of the data, so they tend to require additional explanation to make the comparison of the graph to the raw data meaningful. Another disadvantage of bar charts are that they can be easily manipulated to yield misleading interpretations which violates the theory of graphical integrity. An example of these type of plots will be displayed in the results chapter of this paper.

Tukey (1977) defines a rootogram as "a stack of columns whose heights are proportional to the square root of counts", so it is classified as a frequency graph in which one axis is scaled by the square root of the frequencies to emphasize the smaller values. In a rootogram, the observed frequencies are displayed as bars and the fitted frequencies as a curved line. An extension of a rootogram is a hanging rootogram where the top of the bars begin at the expected frequency. The advantage of the hanging rootogram is that it makes it easier for the viewer to judge departures from independence since the reference line becomes a horizontal line at zero rather than the curve. Another option is a deviation rootogram where bars are drawn to show the gap between (observed - expected) and the reference line. In general, a rootogram is best for plotting observed and fitted frequencies as a way to measure the fit of a particular distribution. A drawback to using rootograms is that the square-root scale can limit what values can be seen in the graph, and at times it makes more sense to have graphs on a log-scale. An example of this type of graph is shown in Figure 2.6.



(a) Example of Hanging Rootogram



(b) Example of a Deviation Rootogram

Figure 2.6: Source: Friendly & Meyer 2016

A tile plot is a matrix of tiles. For each tile, either the width, height, area, or squared area is proportional to the corresponding entry. The design of the plot makes comparisons easy column-wise, row-wise, and overall. It allows for a direct comparison of the raw data to the plot since the rectangles representing the table frequencies in the tile plot are arranged in the same tabular form as the raw data unlike in a bar plot (Friendly & Meyer 2016). Although tile plots offer a better solution to visualizing contingency tables, they do not reveal much about how variables are associated, making them better as a first step in the analysis rather than a tool to help identify explanatory models. An example of these type of plots will be displayed in the results chapter of this paper.

Spineplots are a special case of mosaic plots and can be considered as a generalization of a stacked bar plot where the widths, rather than the heights of the bars correspond to the relative frequency. The heights of the bars then correspond to the conditional relative frequency. These types of graphs provide a way to visualize row percentages and to gain insight into possible associations since "departures from independence are shown by failure of alignment" (Cox 2008). They also offer an excellent way to visualize the difference between observed and expected frequencies. The problem with spineplots is that they lack color, being displayed exclusively in grayscale. If

independence is unlikely, as the case with mobility tables, spineplots will not be effective because of the coloring. Differences in color are also hard to discern if cells have small frequencies, or if cells have zero frequencies then they are not represented in the graph at all. An example of these type of plots will be displayed in the results chapter of this paper.

A sieve diagram shows the frequencies in a two-way contingency table in relation to expected frequencies under independence, and highlights the pattern of association between the row and column variables. In a sieve diagram, the area of each rectangle is proportional to the expected frequency, while the observed frequency is shown by the number of squares in each rectangle (Friendly & Meyer 2016). An advantage of representing the frequencies in these ways is that interpretation can be determined by the intensity of the shading. Cells whose expected frequency is greater than the observed frequency appear less intense, than those cells where the observed frequency is greater than the expected frequency. Deviations from independence can also be easily determined by color using one for positive deviations and the other for negative deviations. There are two limitations to a sieve diagram. First, it does not extend well beyond two variables. Second, the order of the categories has an impact on the pattern of association such that an illogical ordering of the categories can lead to a different interpretation. An example of these type of plots will be displayed in the results chapter of this paper.

Association plots visualize the table of Pearson residuals: each cell is represented by a rectangle that has height proportional to the corresponding Pearson residual r_{ij} and width proportional to the square root of the expected counts. Thus, the area is proportional to the raw residuals. The rectangles representing each cell in the table are positioned relative to a line representing independence. Cells with observed frequency greater than expected frequency are shown above the line and cells with observed frequency less than expected frequency are shown below the line (Friendly & Meyer 2016). Color is also used to shade the boxes according to the value of the Pearson residual. Although association plots are great for determining patterns of deviation from independence, they do not reveal much about possible models and like sieve diagrams, order matters. An example of these type of plots will be displayed in the results chapter of this paper.

A fourfold display is a special case of a polar area chart designed for the display of 2 by 2 and more generally 2 by 2 by k tables. In this graph there are four quadrants representing each cell of a fourfold table. The radius of the quadrants are proportional to the square root of the cell frequency so that the area is proportional to the cell count. These types of graphs are used to visualize the sample odds ratio. Association is present between the variables of interest if diagonally opposite, same color quadrants differ in size. Confidence rings can be added to the display to visually test the null hypothesis of independence. If the rings of adjacent quadrants overlap, then the observed counts are consistent with the null hypothesis (Friendly & Meyer 2016). These graphs are useful when doing a stratified analysis because a fourfold display can be created for each stratum. This visualization will allow the viewer to easily detect if the association between two variables is homogeneous across strata. However, if the goal of the analysis is to determine how the odds ratio varies within a quantitative strata, the fourfold display is not as useful as just plotting the odds ratio themselves.

An example of this type of graph is shown in Figure 2.7. Friendly and Meyer use a dataset of graduate school admissions to the University of California, Berkley to examine if men are actually admitted more than women. At the aggregate level, men appear to be admitted at a higher rate than women. The fourfold display shows, however, that by examining admissions rates for each department, men and women are equally likely to be admitted except for in the case of Department A. In this department, women are actually more likely to be admitted. At the aggregate level, men and women are assumed to apply equally to each department, but in fact there is a difference in what departments men and women apply to and this difference is immediately evident in the visual display.

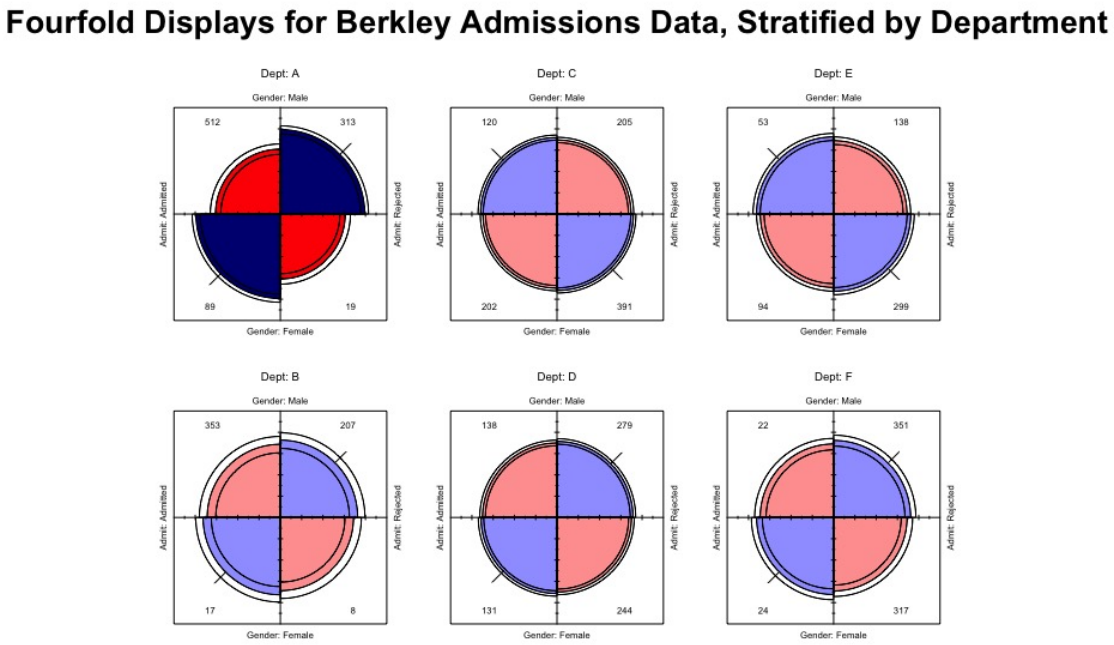


Figure 2.7: Example of fourfold display. Source: Friendly & Meyer, 2016

A corrplot is a graphical display of a correlation matrix that can be generalized to visualize any matrix. The purpose of the plot is to detect highly correlated variables or to detect patterns of associations if examining a matrix of log odds ratios. There are different visualization techniques that may be used in the corrplot. The value in the cells of the matrix can be represented by shape, text or color. If using the shape method, the shape grows with the size of the correlation. That is, if using circles, for instance, the larger the circle the larger the correlation. It should be noted that the shape method allows for emphasis not only by size of the shape but by the color of the shape, so depending on the color gradient, darker bigger shapes signify a larger correlation than a small light colored shape. If using the text method, the actual cell value is plotted in the corrplot

with shading to emphasize different values. The last method differentiates between values in the matrix through color intensity. If examining correlations, the matrix can be reordered according to the correlation coefficient, which can help to reveal hidden patterns in the matrix. It can also be used as a visual test of significance by combining the correlation matrix with the corresponding p-values to obtain a corrplot where insignificant values are left blank or marked with an X. The flexibility of the corrplot is what makes it so useful, however, what makes it useful can also limit its understanding. That is, if a non-sensical color palette is used then differences in values will be hard to discern.

Mosaic diagrams visualize the frequencies of a table such that the size of each tile is proportional to the cell frequency. The way mosaic diagrams are crafted is that a unit area is divided into bars such that the widths of the bars represent the marginal frequencies of one variable. Then, those bars are sub-divided into tiles so that the height of the tiles represent the conditional probability of the second variable. Note that this processes can be continued to extend the diagram beyond only two variables. For cross-classified data, the tiles will align when there is no association among the variables. "For two or more variables, the levels of sub-division are spaced with larger gaps at the earlier levels, to allow easier perception of the groupings at various levels, and to provide for empty cells" (Friendly 2002). If using the mosaic diagram to visualize the structure of a given loglinear model, the tiles can then be shaded in various ways to reflect the residuals (lack of fit) of the particular model. The pattern of residuals can then be used to suggest a better model or understand where a given model fits or does not fit (Friendly 2016). Thus, the reason mosaic diagrams are extremely useful in data visualizations of contingency tables is that they are multi-functional. These graphs can be used simply as a way to visualize frequencies or to visualize the structure of a model. Despite the many strengths of mosaic plots, they are not without their weaknesses. The mosaic diagram must be arranged side by side along a common baseline in order to clearly compare the heights of the tiles. Additionally, these plots are sensitive to both the order of variables used in the sub-division process as well as the order of categories within a variable. Lastly, the close nature of tiles in a mosaic diagram can make labeling an issue, so axes labels may

have to be abbreviated to avoid overlapping. Example of these type of plots will be displayed in the results chapter of this paper.

Data visualizations that are used for the presentation of results require careful consideration of the possible audiences. Will the data visualizations be viewed in a presentation using slides? Will the data visualizations be viewed in a journal publication? Will the data visualizations be printed or copied? The answer to these questions can alter choices made through the analysis process. A major point to consider is color. Fortunately, there are tools available online and through the R statistical software that can be used to make informed decisions. Cynthia Brewer and Mark Harrower (2009) created the website Color Brewer 2.0 that provides advice for coloring maps, and their scheme has been extended to an R package called RColorBrewer. The setup of Color Brewer 2.0 makes choosing an appropriate color palette easy. It has an option for the number of data classes or groups to be considered, the nature of the data (i.e. sequential, diverging, and qualitative), and choice of color scheme (i.e. multi-hue or single hue). In addition, the user can choose to only show colorblind safe, print friendly safe, and/or photocopy safe colors. RColorBrewer essentially has the same options, but the options are chosen through R code rather than the point and click method available to online users.

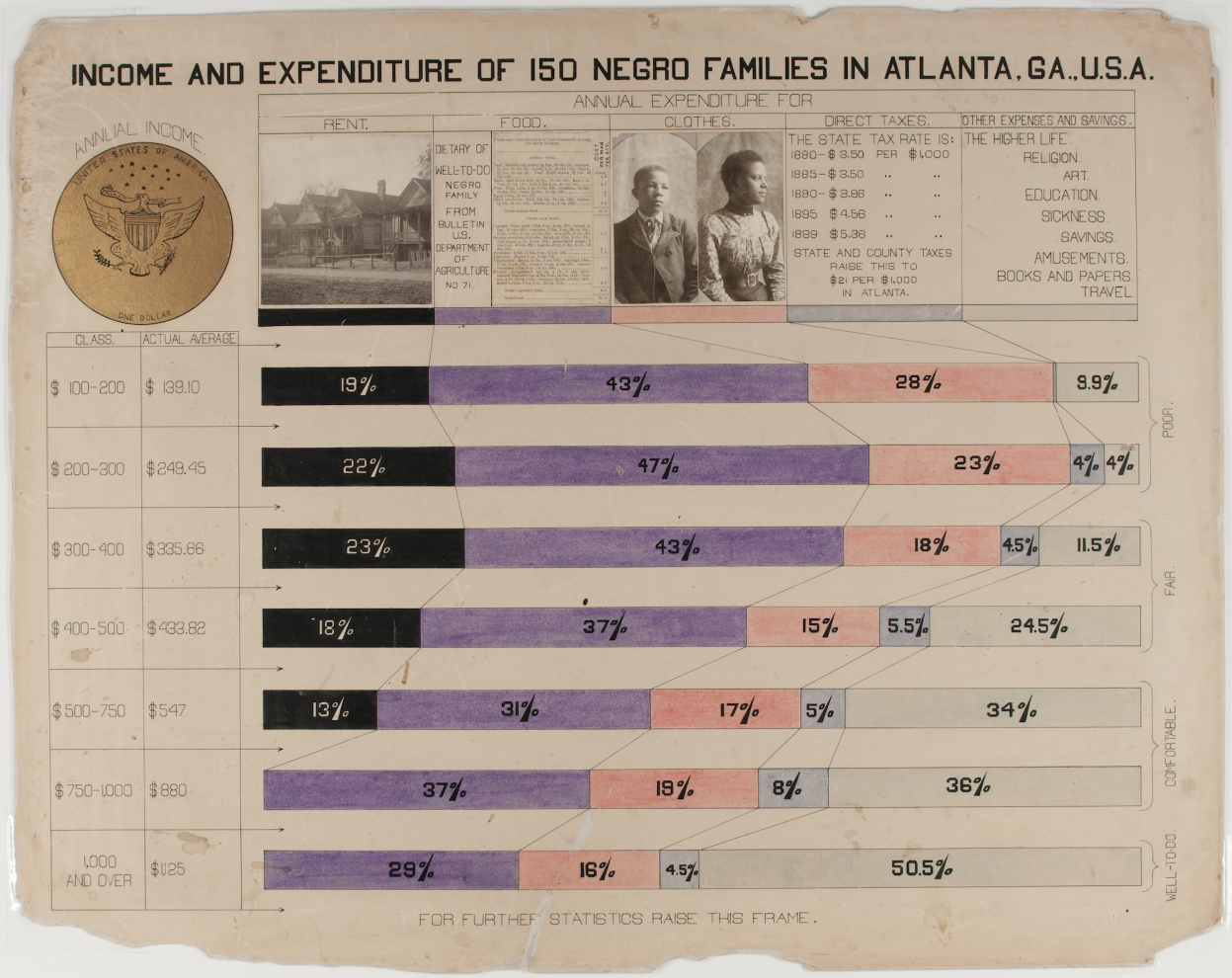


Figure 2.3: W.E.B. DuBois's Bar Chart of the Expenditures of the Georgian Negro. Source: DuBois, 1900

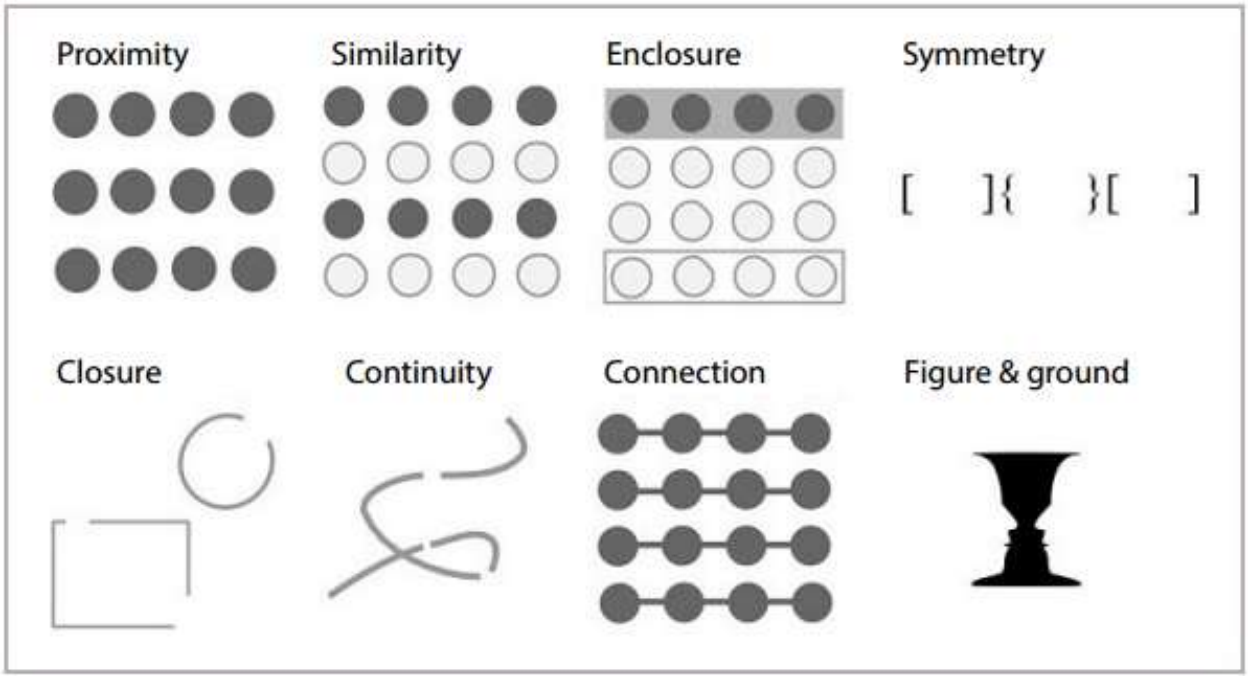


Figure 2.5: Illustration of the Gestalt Design Principles. Source: Few, 2006

CHAPTER 3. CASE STUDY

The focus of this research is to consider how visualization methods can enhance the exploration and presentation of sociological data. I use a classic sociological dataset, the 1949 Great Britain Occupational Mobility Table, as a case study to illustrate the utility of these methods. The basics of contingency tables reveal that they are useful when analyzing the relationship or association of a variable for each level of another variable. This concept is of interest to researchers who study social mobility since persons are usually classified by their social class and by the social class of one of their parents (Hauser, 1980). In these studies, social scientists analyze tables of counts where subjects have been cross-classified using the same set of categories (e.g. social class). This cross-classification is another reason why contingency tables are of use to sociologists. To interpret the counts, one has to view the counts as products of prevalence and interaction effects. Prevalence and interaction can be analyzed using models whose parameters correspond to those concepts (Hauser, 1980). These models are typically log-linear or extensions of such models.

Table 3.1: Observed Frequencies of the 1949 Great Britain Mobility Table

	Son's Occupation				
Father's Occupation	Upper NonManual	Lower NonManual	Upper Manual	Lower Manual	Farm
Upper NonManual	50	45	8	18	8
Lower NonManual	28	174	84	154	55
Upper Manual	11	78	110	223	96
Lower Manual	14	150	185	714	447
Farm	0	42	72	320	411

Table 3.1 from Hauser (1978) provides the classic 1949 Great Britain five-by-five son's by father's occupational mobility table. The cells of the table give the counts of persons that share each combination of occupational category, and the schemes represented by each category are shown in the list below (Hout 1983). The schemes are constructed by considering class positions to be designated by employment relations and further dictated by the mode of employment (Erikson

& Goldenthorpe 2002). In general, there are not inherit hierarchical differences between Lower NonManual and Upper Manual classes. That is, those employees in the Lower NonManual classes may have lower average incomes than do those in the Upper Manual classes but employees in the Lower NonManual classes may have more stable levels of income. However, the employees considered to be apart of the Upper NonManual category that represents the "salariat", can be considered to be advantaged over employees in other classes.

> "Members of the salariat are advantaged over members of the working class in that they experience; i) greater long-term security of income through being less likely to lose their jobs and to become unemployed; ii) less short-term (week-to-week or month-to-month) fluctuation of income through being less dependent on piece rates, shift premiums, overtime payments and less exposed to loss of pay on account of absence or illness; and iii) better prospects of steadily increasing income over the life course- into their 50s rather than their 30s- through having employment contracts that are conducive to an upward-sloping age-earnings profile (Lazear, 1995) with in turn better prospects for the accumulation of wealth" (Erikson & Goldenthorpe 2002).

The data shown in Table 3.1 originated from the 1949 Labour Mobility Study conducted as part of a census of Great Britain. The objective of the survey was to determine "the rate of occupational, industrial, and geographical change in the employed population in Great Britain in 1949, to compare it with the frequency and change in the past and to ascertain the factors associated with change" (Office of Population Censuses and Surveys 1949). The original researchers obtained a sample of 4207 men aged eighteen years or older through a systematic random sample. Details of the sampling procedure are discussed in Glass (1954).

Let i index the rows and j the columns, then m_{ij} denotes the number of persons with father's occupational category i and son's occupational category j. The cells in the main diagonal of the table refer to fathers and sons with the same occupational category, and this group is important because it measures the total amount of mobility exhibited by the son. The observed frequencies are used to estimate the expected frequency of each cell of the table under the null model (denoted

Table 3.2: Five-Category Classification of Occupation

Upper NonManual (UpperNoM)	Professionals, self-employed
	Professionals, salaried
	Managers
	Salespersons, nonretail
Lower NonManual (LowerNoM)	Proprietors
	Clerical workers
	Salespersons, retail
Upper Manual (UpperM)	Craftsmen, manufacturing
	Craftsmen, other
	Craftsmen, construction
Lower Manual (LowerM)	Service workers
	Operatives, other
	Operatives, manufacturing
	Laborers, manufacturing
	Laborers, other
Farm	Farmers and farm managers
	Farm laborers

as E_{ij}) of perfect mobility.

$$E_{ij} = \frac{(\text{row } i \text{ total})(\text{col } j \text{ total})}{(\text{table total})} = \frac{n_i n_j}{N} \tag{3.1}$$

Let π_{ij} represent the proportion of the population classified into row i column j. Then, the null hypothesis of the perfect mobility model is defined as $H_0 : \pi_{i.}\pi_{.j}$, i.e., H_0: row and column classifications are independent.

3.1 Statistical Tests and Residuals

The Chi-square test statistic compares observed and expected frequencies for all cells in the table.

$$X^2 = \sum_{i=1}^{R}\sum_{j=1}^{C} \frac{(\text{observed - expected})^2}{\text{expected}} \tag{3.2}$$

$$= \sum_{i=1}^{R}\sum_{j=1}^{C} \frac{(n_{ij} - E_{ij})^2}{E_{ij}} \tag{3.3}$$

Given a large sample size, under the null hypothesis of no association, this test statistic can be compared to a central chi-square distribution with (R-1)(C-1) degrees of freedom.

The chi-squared test may be associated with large degrees of freedom and a small p-value that leads to the rejection of the null hypothesis, but this information does not show how the test fails. To obtain a better understanding of why the test fails, one can look at the pattern of residuals. For cell i, j, the raw residual is the difference between the observed and fitted frequencies, i.e.

$$\text{observed - expected} = n_i j - E_{ij}. \tag{3.4}$$

The signed contribution to the Pearson χ^2 for cell i, j is

$$r_{ij} = \frac{n_{ij} - m_{ij}}{\sqrt{m_{ij}}} \tag{3.5}$$

The deviance residual is defined as

$$g_{ij} = sign(n_{ij} - E_{ij})\sqrt{(2n_{ij}log(\frac{n_{ij}}{E_{ij}}) - 2(n_{ij} - E_{ij}))}. \tag{3.6}$$

Since the residuals for cell with small expected frequencies have a larger sampling variance, adjusted residuals provide a standardized version of the Pearson and deviance residuals. Dividing the Pearson and deviance residuals by its estimated standard error gives the adjusted residuals.

Another test statistic that is often computed to evaluate mobility models is the likelihood ratio statistic.

$$L^2 = 2\sum_{i=1}^{R}\sum_{j=1}^{C} n_{ij}\log(\frac{n_{ij}}{m_{ij}}) \tag{3.7}$$

Given a large sample size, under the null hypothesis, this test statistic can also be compared to a central chi-square distribution with (R-1)(C-1) degrees of freedom. The likelihood ratio statistic is often preferred over the chi-square test statistic because it can be decomposed into substantively and statistically interpretable parts (Hout, 1983).

Index of dissimilarity does not test a particular hypothesis but is used as a way to measure the proportion of cases misclassified by the model.

$$\Delta = \frac{1}{2N_{ij}}\sum_{i=1}^{R}\sum_{j=1}^{C} |\pi_{ij} - m_{ij}| \tag{3.8}$$

Association for a two by two table can be measured by the odds ratio.

$$\phi = \frac{\omega_2}{\omega_2} = \frac{E_{11}E_{22}}{E_{12}E_{21}} \tag{3.9}$$

An odds ratio higher than one means that the second categories of the row and column variables are positively associated. An odds ratio of one indicates a null relationship between the two variables, corresponding to statistical independence. For a general two-way table of dimension R by C, there are (R-1)(J-1) non-redundant odds ratios, from which other odds ratios can be derived.

3.2 Mobility Models

The model selection process for analyzing mobility tables begins with the model of perfect mobility. The null hypothesis associated with the model of perfect mobility for the data discussed in the previous chapter is that the son's occupational status is independent from the father's occupational status (i.e. destination is independent from origin). When perfect mobility exists, each row of outflow percentages is the same (Hout 1983). All other models are tested against the model of perfect mobility.

3.2.1 Quasi-Independence Model

Often, the model of perfect mobility is rejected because of the prevalence of the diagonal cells. That is, there are many fathers and sons with the same occupational status. Therefore, the next question is whether or not independence is achieved by ignoring the diagonal cells. The quasi-independence model "specifies independence only in the off-diagonal cells" (Friendly & Meyer 2016). It is expressed as

$$\log(m_{ij}) = \mu + \lambda_i^A + \lambda_j^B + \delta_i I(i = j) \tag{3.10}$$

The added parameter, δ_i, measures the deviation from independence in the diagonal cells. The model works by assigning a certain number of sons with the same occupational status as their fathers, and the other sons are given occupational status without regard to their father's status.

3.2.2 Quasi-Symmetry, Symmetry, and Marginal Homogeneity Models

Another important model to test is the model of quasi-symmetry, and it is expressed as

$$\log(m_{ij}) = \mu + \lambda_i^A + \lambda_j^B + \lambda_{ij} \tag{3.11}$$

where $\lambda_{ij} = \lambda_{ji}$ for all $i \neq j$, i.e., all interaction effects are symmetric. The parameter λ_{ij} represents the idea that sons are just as likely to move from one of their father's occupations to another. Two related models are the model of symmetry and the the model of marginal homogeneity. The model of symmetry requires that the interaction effects and the marginal effects are symmetric. If the interaction effects are not symmetric, but the marginal effects are equal, then the model of marginal homogeneity is obtained. It can be shown that the model of symmetry can be decomposed into the model of quasi-symmetry and the model of marginal homogeniety, i.e., symmetry = quasi-symmetry + marginal homogeneity (Friendly & Meyer 2016).

3.2.3 Uniform Association Model

The uniform association model takes into account the ordering of the categories in the table. Ordered scores are assigned to the categories so that the ordinal nature of the variables is included in the model. Suppose the row variable is assigned scores, $\mathbf{a} = (a_i), a_1 \leq a_2 \leq ... \leq a_I$, and the column variable is assigned scores, $\mathbf{b} = (b_j), b_1 \leq b_2 \leq ... \leq b_I$, then the uniform association model is defined as

$$\log(m_{ij}) = \mu + \lambda_i^A + \lambda_j^B + \gamma a_i b_j \tag{3.12}$$

where $\gamma a_i b_j$ describes the pattern of associations such that deviations from independence increase linearly with a_i and b_j in opposite directions towards the corners of the table. The model can be extended by adding a parameter to fit the main diagonal cells so that these cells are ignored in the model.

3.2.4 Row, Column, Row + Column, and Row-and-Column Effects Models

Row and column effect models stem from the uniform association model when only one variable is assigned an ordered score. In the row effects model, the column variable, B, is assigned ordered scores and the row variable, A, is treated as nominal. This model is denoted as

$$\log(m_{ij}) = \mu + \lambda_i^A + \lambda_j^B + \alpha_i b_j, \tag{3.13}$$

where α_i represent the row effects and are constrained so that $\sum_i \alpha_i = 0$. Similarly, in the column effects model, the row variable, A, is assigned ordered scores and the column variable, B, is treated as nominal. Thus, the model is expressed as

$$\log(m_{ij}) = \mu + \lambda_i^A + \lambda_j^B + a_i \alpha_j, \tag{3.14}$$

where α_j denote the column effects and $\sum_j \alpha_j = 0$. A row plus column effects model is a related model where the scores for the row and column variables are specified.

A generalization of these models is given when the assigned scores are treated as parameters and is known as the row-and-column effects model. It is denoted as

$$\log(m_{ij}) = \mu + \lambda_i^A + \lambda_j^B + \gamma \alpha_i \beta_j, \tag{3.15}$$

and the ordering restriction is no longer assumed since the scores are estimated using the data. The additional parameters are subjected to the following constraints for identifiability and interpretation purposes:

$$\sum_i \alpha_i = \sum_j \beta_j = 0, \tag{3.16}$$

and

$$\sum_i \alpha_i^2 = \sum_j \beta_j^2 = 1 \tag{3.17}$$

3.2.5 Crossings Model

The crossings model "hypothesizes that there are different difficulty parameters for crossing from one category to the next, and that the associations between categories decreases with their separation. In the crossings model for an I x I table, there are I - 1 crossings parameters, $\nu_1, \nu_2, ..., \nu_{I-1}$.

The association parameters, λ_{ij}^{AB} have the form of the product of the intervening ν parameters" (Friendly & Meyer 2016). The model is expressed

$$\log(m_{ij}) = \mu + \lambda_i^A + \lambda_j^B + \lambda_{ij}^{AB} \tag{3.18}$$

where

$$\lambda_{ij}^{AB} = \begin{cases} \prod_{k=i}^{k=j-1} \nu_k & \text{if} i < j \\ \prod_{k=j}^{k=i-1} \nu_k & \text{if} i > j \end{cases}$$

A quasi form of this model can be produced by adding a diagonal term to fit the main diagonal cells.

3.3 Model Selection

No model will perfectly fit the data. However, model selection techniques can be used to identify the model with relative little bias, describes the truth well, and provided more accurate estimates of the quantities of interest (Agresti 2002). These techniques extend beyond significance tests and judge a model by how close the fitted values are to the true values. Two common criteria are Akaike Information Criterion (AIC) and Schwarz Bayesian Information Criterion (BIC). AIC is a criterion that selects the model that best minimizes

$$AIC = -2L(\hat{\theta}) + 2p \tag{3.19}$$

where L($\hat{\theta}$) represents the maximum likelihood and p is the total number of model parameters. The +2k part of AIC is viewed as a penalty for model complexity. BIC is similar to AIC except the penalty for model complexity is greater and grows with n. It selects the model that best minimizes

$$BIC = -2L(\hat{\theta}) + plog(n) \tag{3.20}$$

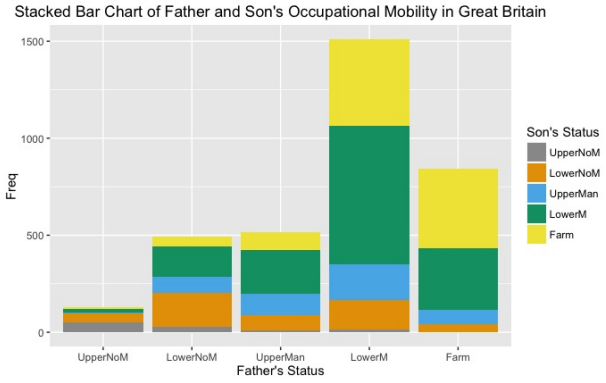
where L($\hat{\theta}$) represents the maximum likelihood, p is the total number of model parameters, and n is the sample size. Small values of AIC and BIC are preferred, but a nearly best model is acceptable if preferred by previous research in the discipline.

CHAPTER 4. RESULTS AND CONCLUSIONS

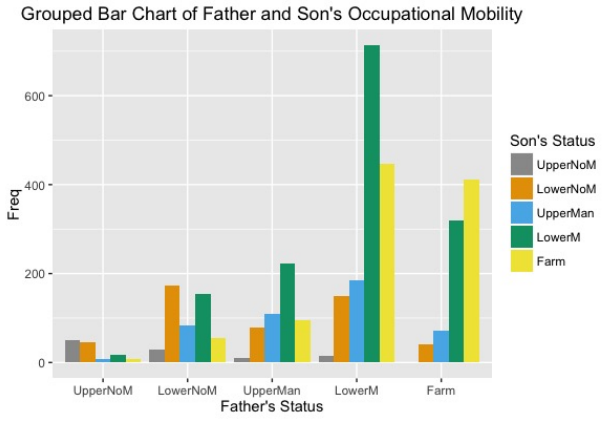
Figure 4.1 shows four different visualizations of the observed frequencies of father and son's occupational mobility in Great Britain in 1949. Figure 4.1(a) is a stacked bar chart depicting several bars each subdivided by different colored blocks. The bars represent the occupational mobility status of the fathers, and the blocks represent the occupational mobility status of the sons. There are five bars to differentiate between occupational categories of the fathers, and five different colors for the blocks to differentiate between each occupational category of the sons within each bar. The height of the bars and blocks are proportional to the observed frequency. This type of graph provides a way to display the data visually but beyond that the graph is not that useful. It is hard to read specific values, and it requires effort to compare groups. For instance, in the case of fathers classified in the Upper NonManual category, it is difficult to determine the number of son's classified as farmers.

Figure 4.1(b) is a grouped bar chart. In this visualization, for each category of father's occupational mobility, a group of bars correspond to each category of the son's occupational mobility. It is easier to compare groups within each category in the grouped bar chart versus the stacked bar chart. Specific values are also easier to read in the grouped bar chart. However, the grouping suggests a causal relationship which can be misleading. Both figures 4.1(a) and 4.1(b) have another disadvantage. Neither graph offers a graphical representation that matches the tabular data structure which can complicate comparisons with the raw data. Figure 4.1(c), however, does match the tabular data structure. The table frequencies are represented by the area of rectangles arranged in the same tabular form as the raw data, facilitating comparisons between tiles across both variables. Figure 4.1(d) of the spineplot offers another way to display the data in a visual way. It shows the row percentages of the son's occupational mobility for each category of the

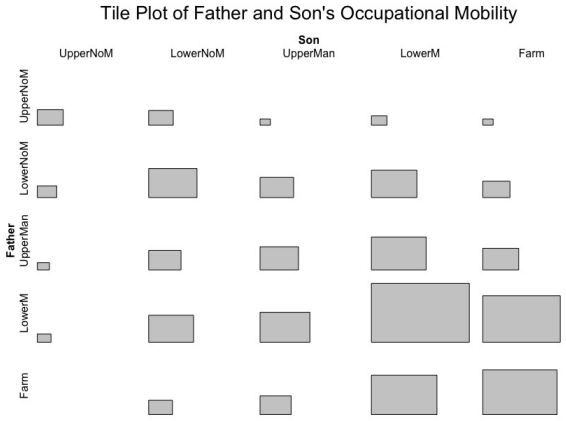
father's occupational mobility, and the widths of each bar is proportional to the overall percentage of father's occupational mobility.



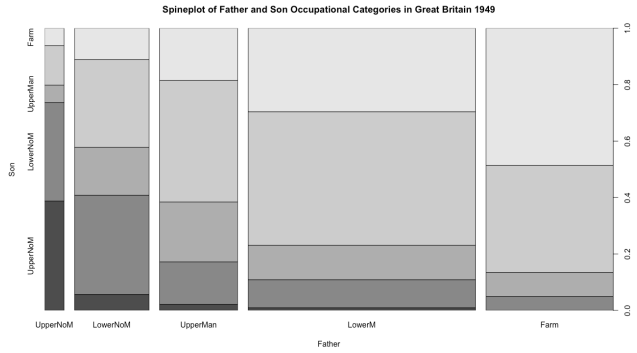
(a) Stacked Bar Chart



(b) Grouped Bar Chart



(c) Tile Plot

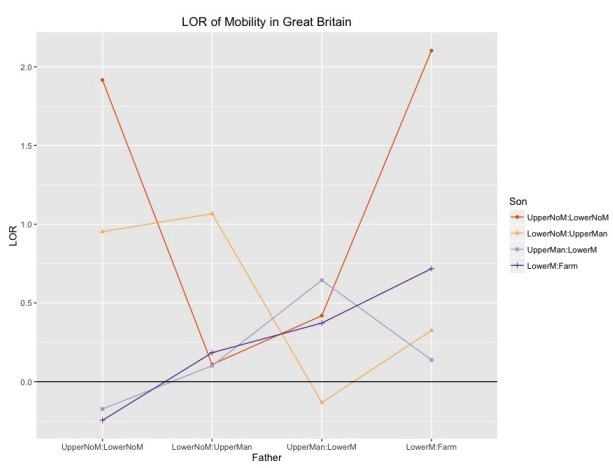


(d) Spineplot

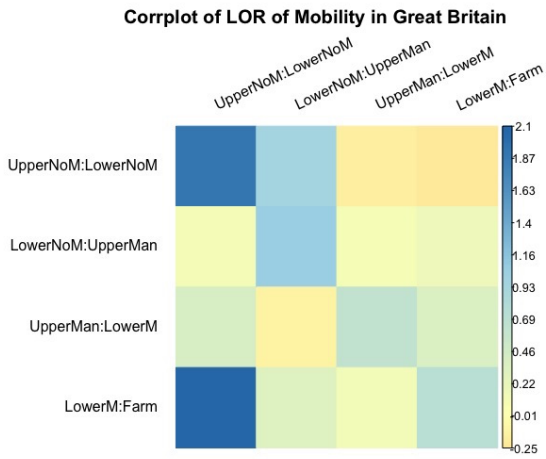
Figure 4.2: Options for simple ways to display the frequencies in a mobility table visually

Before fitting any models, it is useful to calculate and plot the observed local log odds ratios to see the patterns in the data that need to be accounted for. These values are graphed and shown in

Figure 4.2. Recall, if the log odds ratio is equal to zero, then there is no association between father and son's occupational statuses, if the log odds ratio is greater than zero then there is a positive association between the two variables, and if the log odds ratio is less than zero there is a negative association. Using this information, we can interpret the patterns displayed in the graphs. The visualization on the left (Figure 4.2 (a)) is of the observed local log odds ratios in the Great Britain 1949 data. The various occupational categories are depicted with different color lines and different shaped points to help get a clear distinction of each category. Also, note the solid horizontal line at zero that signifies local independence. Overall, there appears to be a positive association between the father and son's occupational statuses. There are two locations where the log odds ratios dip below zero. The first location is at the comparison of sons in non-manual categories with fathers in the manual categories, and the second at the comparison of sons in the manual categories with fathers in the non-manual categories. These cases may indicate that sons are not likely to be downwardly mobile. The high log odds ratios of sons in the non-manual and the lowest manual categories with fathers in the highest non-manual categories may need attention. The graph on the right (Figure 4.2(b)) is a corrplot of the log odds ratio. This type of plot is typically used to visualize correlation matrices, but the R function can be generalized to visualize any matrix. The pattern displayed in the corrplot is essentially the same as for Figure 4.2(a). However, there are certain aspects that are easier to discern in the corrplot. The points of interest stand out more in the corrplot because of the distinctive colors than in the line graph, but the colors in the gradient tend to mix when the log odds ratios are close to zero. Therefore, it is easier to see indications of negative association in the line graph.



(a) LOR of Mobility in Great Britain



(b) Corrplot of Log Odds Ratios

Figure 4.4: Visual Displays of the Log Odds Ratios

Sieve diagrams are quite useful as a starting point because these type of plots display the observed frequencies in relation to the expected frequencies. The area of each rectangle in the diagram is proportional to the expected frequency under independence because it is constructed such that the widths are proportional to the total frequency in each column and the heights are proportional to the total frequency in each row. Figure 4.3(a) is a sieve diagram of data from Hauser (1978). Observed frequencies are shown by the number of squares in each rectangle, and the difference between observed and expected frequencies is shown through the density of the shading. The colors represent positive and negative deviations from independence. To preserve the usefulness of the plot to all viewers, colorblind friendly and black and white printer friendly colors were chosen. Orange signifies positive deviations from independence, and purple represents negative deviations from independence. Given this information, an interpretation of the diagram is simple. There is a high frequency of sons and fathers with the same occupational status, and it is highly unlikely that a son will have a status higher than that of his father's.

Association plots follow a similar scheme to that of sieve diagrams. Association plots display boxes where area are proportional to the difference between the observed and expected frequency. Figure 4.3(b) is an association plot for the 1949 Great Britain Occupational Mobility Table. A dotted line for each row of the table is drawn to symbolize independence, and the boxes are positioned relative to this line. If a cell within the table has an observed frequency greater than the expected frequency, the box is shown above the line, and if the opposite is true, the box is shown below the line. The color is determined by the value of the cell's Pearson residual. Cells with residual values greater than four are shaded with the colorblind/printer friendly orange color, cells with residual values less than negative four are shaded with the colorblind/printer friendly purple color, and cells with residual values between negative four and positive four are shaded with a colorblind/printer friendly gray color. The association plot confirms the observations made using the sieve plot. That is, sons are likely to have the same occupational status as their father, and they rarely will have an occupational status greater than that of their fathers.

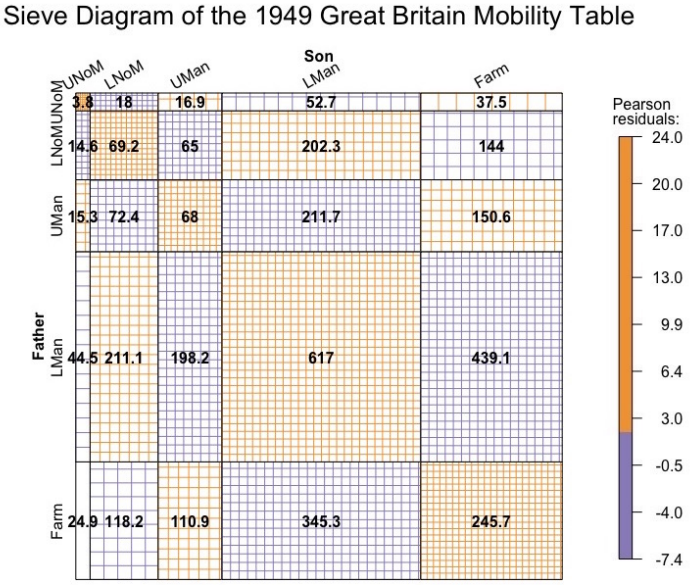


Figure 4.5: Sieve Diagram of Mobility in Great Britain

Table 4.1: Observed Frequencies of the 1949 Great Britain Mobility Table

	Son's Occupation				
Father's Occupation	UpperNoM	LowerNoM	UpperM	LowerM	Farm
UpperNoM	50	45	8	18	8
LowerNoM	28	174	84	154	55
UpperM	11	78	110	223	96
LowerM	14	150	185	714	447
Farm	0	42	72	320	411

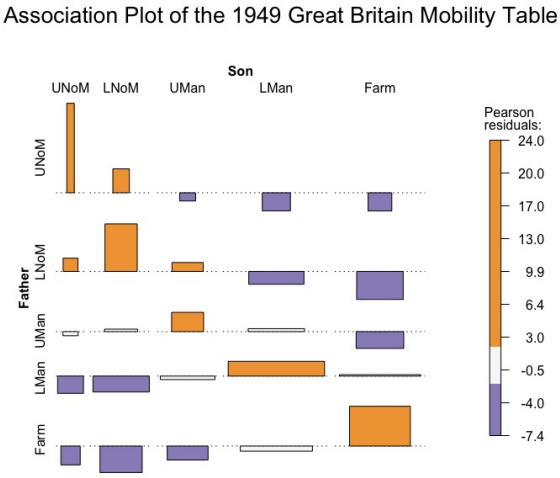


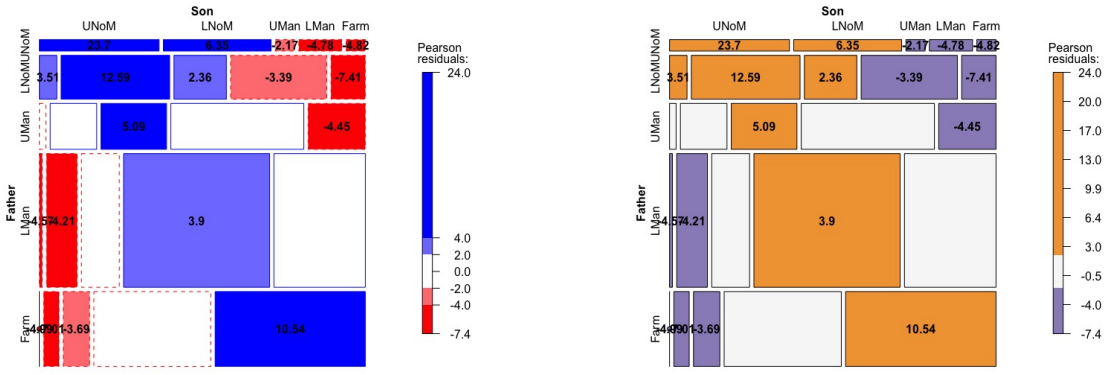
Figure 4.6: Association Plot of Mobility in Great Britain

Table 4.2: Observed Frequencies of the 1949 Great Britain Mobility Table

	Son's Occupation				
Father's Occupation	UpperNoM	LowerNoM	UpperM	LowerM	Farm
UpperNoM	50	45	8	18	8
LowerNoM	28	174	84	154	55
UpperM	11	78	110	223	96
LowerM	14	150	185	714	447
Farm	0	42	72	320	411

The mosaic diagrams for father and son's occupational mobility is shown in Figure 4.4. The two plots differ in the method used to shade the tiles. Figure 4.4(a) was created using the *shad-*

ingFriendly option for the generating function, and Figure 4.4(b) was created using a generating function I created with colors that are colorblind friendly, printer friendly, and photocopy safe. Previously, I emphasized the importance of using colors that are interpretable by anyone which is why I customized the diagram. However, my generating function is limited in how the shading is implemented. The *shadingFriendly* function alters line type as well as



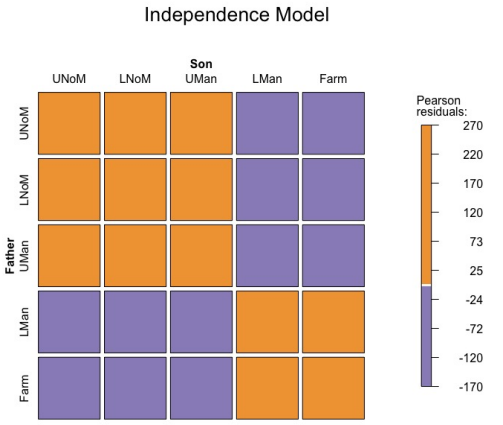
(a) Mosaic Diagram Using Default Settings

(b) Mosaic Diagram Using Customized Settings

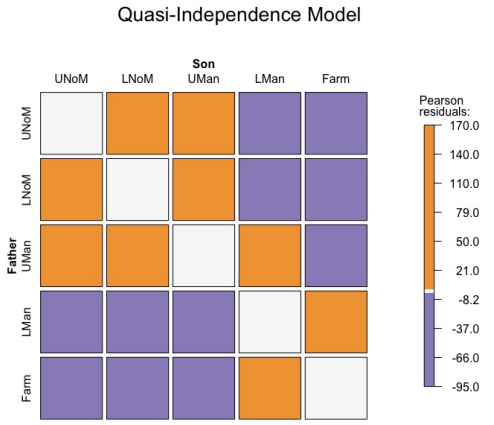
Figure 4.8: Mosaic Diagrams of Mobility in Great Britain in 1949

shading based on the value of the residual, so it adds another layer to the graph to make it easier to interpret. The user must decide if the difference when using a created function versus a built-in function is big enough to make a true impact on how the results are interpreted. The graphs can be interpreted as follows: the area of each tile is proportional to the cell frequency so if father and son's occupational statuses are independent, the tiles in each column would align horizontally. It is clear from both Figure 4.4(a) and Figure 4.4(b) that the tiles do not align horizontally which means there is an association between the two variables. Additionally, the pattern down the main

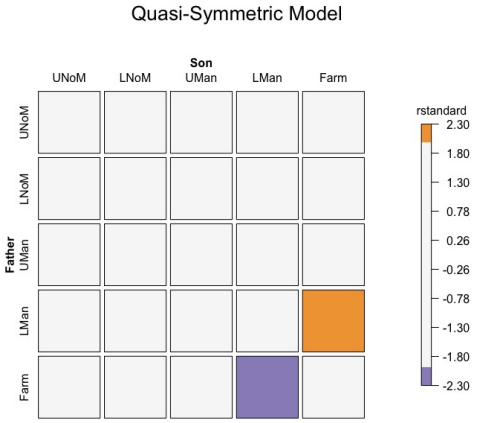
diagonal suggests that models that ignore the effects of the main diagonal values may be of interest when selecting the appropriate model.



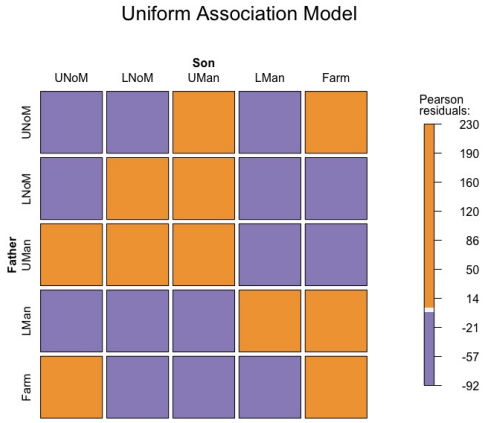
(a) Mosaic Diagram of the Independent Model



(b) Mosaic Diagram of Quasi-Independent Model

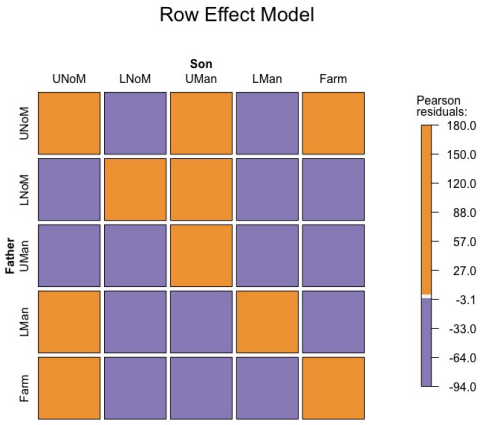


(c) Mosaic Diagram of Quasi-Symmetric Model

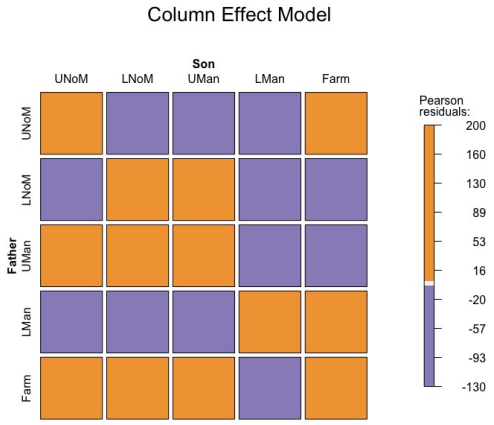


(d) Mosaic Diagram of the Uniform Association Model

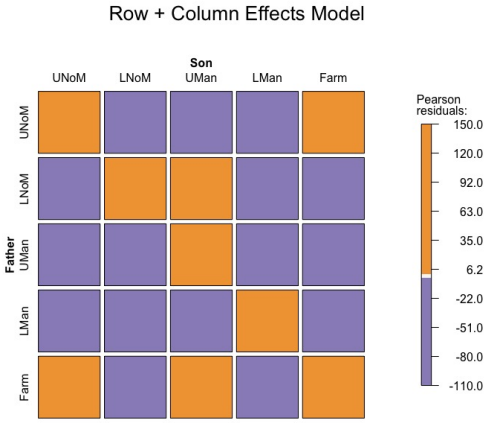
Figure 4.10: Part 1 Mosaic Diagrams Showing Model Structure



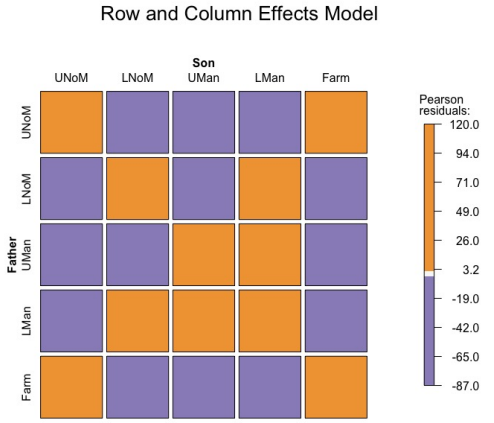
(a) Mosaic Diagram of the Row Effect Model



(b) Mosaic Diagram of the Column Effect Model

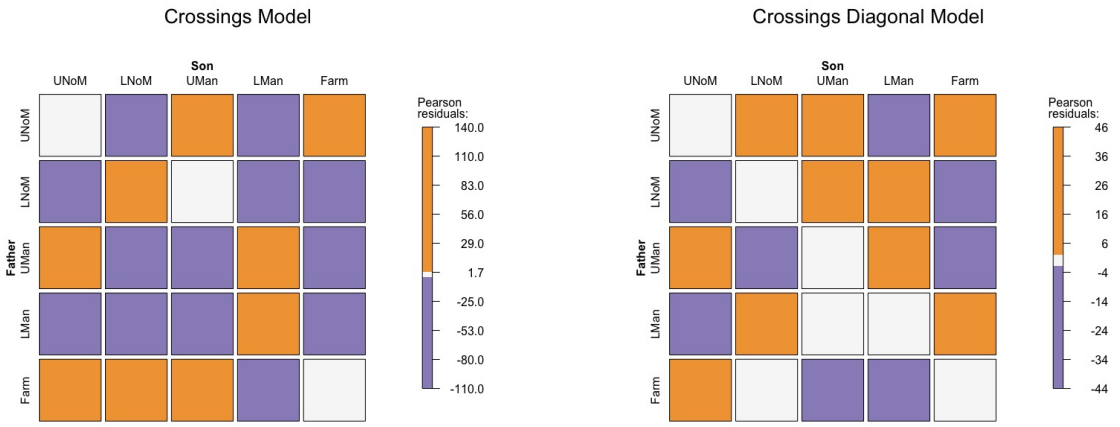


(c) Mosaic Diagram of the Row + Column Effect Model



(d) Mosaic Diagram of the Row and Column Effects Model

Figure 4.12: Part 2 Mosaic Diagrams Showing Model Structure



(a) Mosaic Diagram of the Crossings Model (b) Mosaic Diagram of the Crossings Diagonal Model

Figure 4.14: Part 3 Mosaic Diagrams Showing Model Structure

The first model fit to the data is the model of perfect mobility i.e., the independence model, and the mosaic diagram of Pearson residuals is shown in Figure 4.5 (a). The figure shows the opposite-corner pattern of signs and magnitudes of the residuals indicating a poor model fit, which is not surprising given the large frequencies in the main diagonal.. The poor fit is further supported by the large test statistic, $G^2 = 268,868$ with 16 degrees of freedom (d.f.). To determine if the table exhibits independence disregarding the cells in the main diagonal, I fitted the model of quasi-independence ($G^2 = 83,758$ with 11 d.f.), and the mosaic display corresponding to this model is shown in Figure 4.5(b). The residuals on the main sub-diagonal are mostly positive meaning that when the occupational status of sons and fathers differ, they are most likely to differ by one category. Additionally, the pattern of residuals is somewhat symmetric. The large frequencies of the main diagonal cells coupled with symmetry displayed in the mosaic diagram of the quasi-independence model suggest a test of the quasi-symmetric model. This model asserts that when fathers and sons do not have the same occupational status, sons are equally likely to fall in the

adjacent category without assuming the marginal distribution of fathers and sons are the same. The test statistic of the quasi-symmetric model is $G^2 = 5340$ with 6 degrees of freedom which is a vast improvement in fit when compared to the other models. The mosaic diagram, shown in Figure 4.5(c), shows a fairly consistent pattern of residuals on the off-diagonals as demonstrated by the gray tiles indicating an acceptable fit. Although, the quasi-symmetric model is acceptable, I will fit the other models discussed in the Data and Methods chapter before making a final decision about the best fitting model. The results of the likelihood ratio statistics for each model is shown in Table 4.1. It provides the AIC, BIC, test statistic, d.f., and p-values. The mosaic diagrams corresponding to the uniform association model, row effect model, column effect model, row plus column effects model, row and column effects model, the crossings model, and the crossings diagonal model are displayed in Figure 4.6 and 4.7, respectively. It is easy to see from the mosaic graphs that none of these models provide a better fit than the model of quasi-symmetry. Since there are so many models, a more useful visualization to compare all the models is a model comparison plot that measures AIC and BIC against degrees of freedom. This plot is shown in Figure 4.8. It is easy to discern from this graphic that the model of quasi-symmetry provides the best fit when compared to all other models discussed here in terms of both AIC and BIC values (lower left quadrant of Figure 4.8). Re-observing the mosaic diagram for the quasi-symmetric model, one can see that the only cells that show a lack of symmetry are those for the Lower Manual (i.e., service workers, operatives, and laborers) and the Farm (i.e., farmers, farm managers, and farm laborers) categories. At these cells, the son of a Lower Manual employee is less likely to move to the Farm class than the reverse. I conclude that there does exist a lack of mobility among sons in Great Britain in 1949, which means there is not much movement between stratum in that society.

Table 4.3: Likelihood Summary Table

Model	AIC	BIC	LR χ^2	Df	P-value
Independence	323.02	335.21	268,868	16	0.000
Quasi-Independence	303.87	322.15	83,758	11	0.000
Quasi- Symmetric	245.05	269.43	5,340	6	0.000
Uniform Association	305.49	318.90	123,097	15	0.000
Column Effects	307.03	324.10	102,981	12	0.000
Row Effects	302.34	319.40	85,344	12	0.000
Row+Column Effects	301.62	322.34	65,228	9	0.000
Row and Column Effects	290.15	310.87	41,230	9	0.000
Crossings	288.41	305.48	48,902	12	0.000
Crossings Diagonal	260.00	280.72	12,344	9	0.000

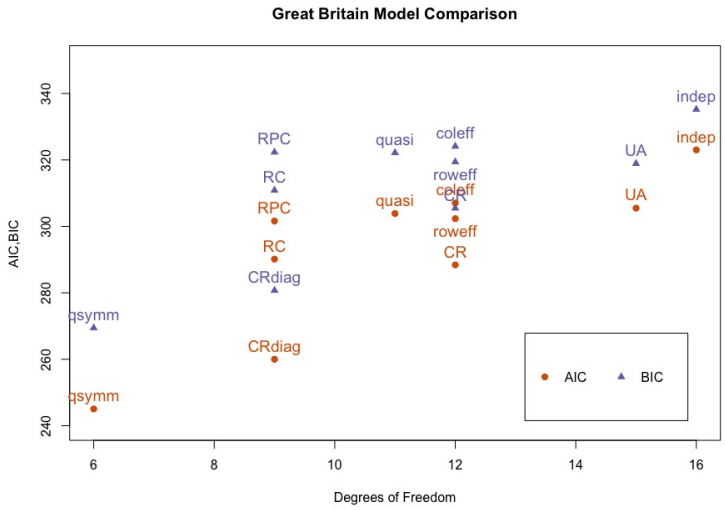


Figure 4.15: Model Comparison Plot for the models fir to the 1949 Great Britain Mobility Table

In respect to data visualizations, the results show that the first display should be one that visualizes the table such that it mirrors the structure of the raw data so that comparisons between

groups can be easily discerned such as in the tile plot. To obtain a view of the local log odds ratio, the corrplot provides a more graphically excellent visualization because the shading highlights differences among the categories better than that of the line plot. Essentially, the patterns important for model development are seen faster and more clearly in the corrplot. Both the sieve diagram and association plots hold value for learning about the relationship between the observed and expected frequencies. The sieve diagram holds more information in one space while the association plot's simplicity communicates possible associations in a way that is more straightforward. Since both have strong advantages, the decision between which to use should be based on what needs to be communicated to the audience. Mosaic diagrams should definitely be used to visualize the data and the structure of the model. By doing both, the researcher can acquire knowledge about what model should be used to describe the data, and it can assist the researcher in selecting the best model and communicating that choice to the audience. Lastly, when there are more than five models under consideration, a model comparison plot should be used to rapidly and easily determine the best model.

4.1 Conclusion

I have advocated for the increase use of visual displays to enhance the analysis of categorical data in the form of contingency tables through the application of data visualizations of mobility tables. The notion that visual displays reduce the scientific rigor of the discipline is erroneous as evident by the popularity of graphics in physical science publications and by the sociologists who incorporated data visualizations into their research during the early life of the field. Integration of graphs into the research process has been well established in terms of continuous variables, and it has been shown that the same is possible for categorical data. There are multiple ways to show data visually that can inform the analysis such as a bar plot, tile plot, or spineplot. These types of graphs can give the researcher a general idea of what the data looks like. Sieve diagrams, rootograms, fourfold displays, association plots, and corrplots are useful research tools for gaining a greater understanding of possible trends or patterns in the data. Such graphics can be informative

at multiple stages of the research process, as in the case of mosaic diagrams. These types of displays can be customized not only to visualize the data, but also to gain a deeper understanding of the structure of the statistical models. Using a classic dataset from the occupational mobility literature, I demonstrated how graphs could be presented in place of a massive amount of tables. The model of quasi-symmetry best fits the data, and sons have not exhibited much more mobility than their fathers. Although, this conclusion does not differ from other researchers who have examined this dataset, it does provide a new approach to how the information is presented to the reader. Graphics enhance the texts by giving readers a tool that summarizes the data rather than overwhelming the reader with lengthy tables. Unfortunately, this application only tackles a fraction of the data visualizations available to researchers that study contingency tables. A possible next step would be to expand data visualization techniques to display comparative studies of social mobility. For instance, visualizations of social mobility differences across countries using maps would be a great avenue for displaying information visually. There are many other possibilities. Therefore, sociologists should consider this paper as a reminder of what visualization can do and incorporate it into each aspect of their research process.

REFERENCES

Agresti, A. (2012). *Analysis of Ordinal Categorical Data: Second Edition*. New York: Wiley.

Andrews, D., Leigh, A. (2009). More Inequality, Less Social Mobility. *Applied Economic Letters*, *16*(15), 1489–1492.

Beller, E. (2009). Bringing Intergenerational Social Mobility Research into the Twenty-First Century: Why Mothers Matter. *American Sociological Review*, *74*(4), 507–528.

Beller, E., Hout, M. (2006). Intergenerational Social Mobility: The United States in Comparative Perspective. *The Future of Children*, *16*(2), 19–36.

Bertin, J. (1983). *Semiology of Graphics*. Madison: University of Wisconsin Press.

Biblarz, T. J., Raftery, A. E. (1993). The Effects of Family Disruption on Social Mobility. *American Sociological Review*, *58*(1), 97–109.

Biblarz, T. J., Raftery, A. E., Bucur, A. (1997). Family Structure and Social Mobility. *Social Forces*, *75*(4), 1319–1341.

Biblarz, T. J., Raftery, A. E., Bucur, A. (1996). Social Mobility Across Three Generations. *Journal of Marriage and the Family*, *58*(1), 188–200.

Brewer, C., Harrower, M. (2009). Color Brewer 2.0. URL: colorbrewer2.org/

Buja, A., Cook, D., Hoffman, H., Lawrence, M., Lee, E., Swayne, D.F., Wickham, H. (2009). Statistical Inference for Exploratory Data Analysis and Model Diagnostics. *Philosophical Transactions: Mathematical, Physical and Engineering Sciences*, *367*(1906), 4361–4383.

Card, S. K., Mackinlay, J. D., Shneirderman, B. (1999). *Readings in Information Visualization: Using Vision to Think*. San Francisco: Morgan Kaufmann.

Chan, T. W., Boliver, V. (2013). The Grandparents Effect in Social Mobility: Evidence from British Birth Cohort Studies. *American Sociological Review*, *78*(4), 662–679.

Chang, W. (2013). *R Graphics Cookbook*. Sebastopol, CA: O'Reilly.

Chapin, F. S. (1924). The Statistical Definition of a Societal Variable. *American Journal of Sociology*, *30*(2), 154–171.

Cook, D. (2015). Data Mining: Exploratory Data Analysis, Iowa State University, Lecture.

Cox, D. R., Jackson, M., Lu, S. (2009). On Square Ordinal Contingency Tables: A Comparison of Social Class and Income Mobility for the Same Individuals. *Journal of the Royal Statistical Society*, *172*(2), 483–493.

Cox, N. J. (2008). Speaking Stata: Spineplots and their Kin. *The Stata Journal*, *8*(1), 105–121.

Davis, K. (1942). A Conceptual Analysis of Stratification. *American Sociological Review*, *7*(3), 309–321.

Deary, I. J., Taylor, M. D., Hart, C. L., Wilson, V., Smith, G. D., Blane, D., Starr, J. M. (2005). Intergenerational Social Mobility and Mid-Life Status Attainment: Influences of Childhood Intelligence, Childhood Social Factors, and Education. *Intelligence*, *33*(5), 455–472.

Diprete, T. A. (2002). Life Course Risks, Mobility Regimes, and Mobility Consequences: A Comparison of Sweden, Germany, and the United States. *American Journal of Sociology*, *108*(2), 267–309.

Diprete, T. A. (1990). Adding Covariates to Loglinear Models for the Study of Social Mobility. *American Sociological Review*, *55*(5), 757–773.

DuBois, W. E. B. (1900). The American Negro at Paris. *The American Monthly Review of Reviews*, *22*(5), 575–577.

Duncan, O. D. (1979). How Destination Depends on Origin in the Occupational Mobility Table. *American Journal of Sociology*, *84*(4), 793–803.

Erikson, R., Goldenthorpe, J. H. (2002). Intergenerational Inequality: A Sociological Perspective. *Journal of Economic Perspectives*, *16*(3), 31–44.

Erola, J., Moisio, P. (2007). Social Mobility over Three Generations in Finland. *European Sociological Review*, *23*(2), 169–183.

Fekete, J., Wijk, J. J. Van, Stasko, J. T., North, C. (2008). The Value of Information Visualization. *Information Visualization*, *4950*(2), 1–18.

Few, S. (2012). *Show Me the Numbers: Designing Tables and Graphs to Enlighten*. Burlingame, CA: Analytics Press.

Few, S. (2006). *Information Dashboard Design, The Effective Visual Communication of Data*. Sebastopol, CA: O'Reily.

Friendly, M., Denis, D. J. (2001). Milestones in the History of Data Visualization: A Case Study in Statistical Historiography. URL: http://www.datavis.ca/milestones/.

Friendly, M. (2005). Milestones in the History of Data Visualization: A Case Study in Statistical Historiography. *Classification: The Ubiquitous Challenge*. New York: Springer.

Friendly, M. (2006). A Brief History of Data Visualization. *Handbook of Computational Statistics: Data Visualization*. Berlin: Springer-Verlag.

Friendly, M. (2003). Working with Categorical Data with R and the vcd and vcdExtra packages. *Insight*, *17*(2), 5987–5994.

Friendly, M. (2002). A Brief History of the Mosaic Display. *Journal of Computational and Graphical Statistics*, *2*(1), 89–107.

Friendly, M., Meyer, D. (2016). *Discrete Data Analysis with R: Visualization and Modeling Techniques for Categorical and Count Data*. Boca Raton: Taylor & Francis Group.

Garnier, M., Hazelrigg, L. E. (1974). Father to Son Occupational Mobility in France: Evidence from the 1960s. *American Journal of Sociology*, *80*(2), 478–502.

Gerber, T. P., Hout, M. (2004). Tightening up: Declining Class Mobility during Russia's Market Transition. *American Sociological Review*, *69*(5), 677–703.

Glass, D. V. (1954). *Social Mobility in Britain*. Abingdon, Oxon: Routledge.

Goodman, L. A. (1965). On the Statistical Analysis of Mobility. *American Journal of Sociology*, *70*(5), 564–585.

Goodman, L. A. (1969). How to Ransack Social Mobility Tables and Other Kinds of Cross-Classification Tables. *American Journal of Sociology*, *75*(1), 1–40.

Goodman, L. A. (1972). Some Multiplicative Models for the Analysis of Cross-Classified Data. In *Proceedings of the sixth Berkley symposium on mathematical statistics and probability* (1, 649-696). University of California Press Berkley.

Goodman, L. A., Hout, M. (1998). Statistical Methods and Graphical Displays for Analyzing How the Association between Two Qualitative Variables Differs among Countries, Among Groups, or Over Time: A Modified Regression-Type Approach. *Sociological Methodology*, *28*(1), 175–230.

Grusky, D. B., Hauser, R. M. (1984). Comparative Social Mobility Revisited: Models of Convergence and Divergence in 16 Countries. *American Sociological Review*, *49*(1), 19–38.

Hauser, R. M. (1978). A Structural Model of the Mobility Table. *Social Forces*, *56*(3), 919–953.

Hauser, R. M. (1980). Some Exploratory Methods for Mobility Tables and Other Cross-Classified Data. *Sociological Methodology*, *11*, 413–458.

Healy, K., Moody, J. (2014). Data Visualization in Sociology. *Annual Review of Sociology*, *40*(1), 105–128.

Higginbotham, E., Weber, L. (1992). Moving Up with Kin and Community: Upward Social Mobility for Black and White Women. *Gender and Society*, *6*(3), 416–441.

Hiroshi, I., Muller, W., Ridge, M. J. (1995). Class Origin, Class Destination, and Education: A Cross-National Study of Ten Industrial Nations. *American Journal of Sociology*, *101*(1), 145–193.

Hout, M. (1988). More Universalism, Less Structured Mobility: The American Occupational Structure in the 1980s. *American Journal of Sociology*, *93*(6), 1358–1400.

Isajiw, W.V. , Driedger, L. (1993). Ethnic Identity and Social Mobility: A Test of the "Drawback Model". *The Canadian Journal of Sociology*, *18*(2), 177–196.

Jonsson, J. O., Mills, C. (1993). Social Mobility in the 1970s and 1980s: A Study of Men and Women in England and Sweden. *European Sociological Review*, *9*(3), 229–248.

Kambourov, G., Manovskii, I. (2008). Rising Occupational and Industry Mobility in the United States: 1968-97. *International Economic Review*, *49*(1), 41–79.

Kleiber, C., Zeileis, A. (2016). Visualizing Count Data Regressions Using Rootograms. *The American Statistician*, *70*(3), 1–25.

Knigge, L., Cope, M. (2006). Grounded Visualization: Integrating the Analysis of Qualitative and Quantitative Data through Grounded Theory and Visualization. *Environment and Planning A*, *38*(11), 2021–2037.

Kostenlnick, C. (2007). The Visual Rhetoric of Data Displays: The Conundrum of Clarity. *IEEE Transactions on Professional Communication*, *51*(1), 116–130.

Long, J., Joseph, F. (2007). The Path to Convergence: Intergenerational Occupational Mobility in Britain and the US in Three Eras. *The Economic Journal*, *117*(519), 61–71.

Long, J., Joseph, F. (2013). Intergenerational Occupational Mobility in Great Britain and the United States since 1850. *American Economic Review*, *103*(4), 1109–1137.

Mazmunder, B., Acosta, M. (2015). Using Occupation to Measure Intergenerational Mobility. *The Annals of the American Academy of Political and Social Science*, *657*(1), 174–193.

Meyer, D., Zeileis, A., Hornik, K. (2006). The Strucplot Framework: Visualizing Multi-way Contingency Tables with vcd. *Journal of Statistical Software*, *17*(3), 1–48.

Nettle, D. (2003). Intelligence and Class Mobility in the British Population. *British Journal of Psychology*, *94*(4), 551–562.

Office of Population of Censuses and Surveys. Social Survey Division. *Labour Mobility Study, 1949* [data collection]. UK Data Service. SN: 196, http://doi.org/10.5255/UKDA-SN-196-1

Parker, S. (2007). *The Human Body Book*. New York: DK Publishing.

Parsons, T. (1940). An Analytical Approach to the Theory of Social Stratification. *The American Journal of Sociology*, *45*(6), 841–862.

Powers, D. A., Yu, X. (2000). *Statistical Methods for Categorical Data Analysis*. San Diego: Academic.

Rooth, D., Ekberg, J. (2006). Occupational Mobility for Immigrants in Sweden. *International Migration*, *44*(2), 57–78.

Rosenthal, J. (2011). *Statistics and Data Interpretation for Social Work*. New York: Springer.

SAS Software. (2016). Data Visualization. URL: https://www.sas.com/en_us/insights/big-data/data-visualzation.html

Saunders, P. (1997). Social Mobility in Britain: An Empirical Evaluation of Two Competing Explanations. *Sociology*, *31*(2), 261–288.

Selert, W. (1997). Occupational and Economic Mobility and Social Integration of Mediterranean Migrants in Germany. *European Journal of Population*, *13*(1), 1–16.

Sletto, R. F. (1936). A Critical Study of the Criterion of Internal Consistency in Personality Scale Construction. *American Sociological Review*, *1*(1), 61–68.

Smith, S.M. (1999). Library of Congress: African American Photographs Assembled for 1900 Paris Exposition. URL: http://www.loc.gov/pictures/collection/anedub/dubois.html

Solga, H. (2001). Longitudinal Surveys and the Study of Occupational Mobility: Panel and Retrospective Design in Comparison. *Quality and Quantity*, *35*(3), 291–309.

Torche, F. (2011). Is a College Degree Still the Great Equalizer? Intergenerational Mobility across Levels of Schooling in the United States. *American Journal of Sociology*, *117*(3), 763–807.

Tufte, E. R. (2001). *The Visual Display of Quantitative Information*. Cheshire, CT: Graphics Press.

Tufte, E. R. (2006). *Beautiful Evidence*. Cheshire, CT: Graphics Press.

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

Tukey, J. W. (1972). Some Graphic and Semigraphic Displays. In TA Bancroft (ed.) *Statisitcal Papers in Honor of George W. Snedecor*, Ames, IA: Iowa State University Press.

Von, S., Gale, C.R., Batty, G. D., Deary, I. J. (2009). Childhood Intelligence, Locus of Control and Behaviour Disturbance as Determinants of Intergenerational Social Mobility: British Cohort Study 1970. *Intelligence*, *37*(4), 329–340.

Wainer, H. (1984). How to Display Data Badly. *The American Statistician*, *38*(2), 137–147.

Warren, J. R., Hauser, R. M. (1997). Social Stratification across Three Generations: New Evidence from the Wisconsin Longitudinal Study. *American Sociological Review*, *62*(4), 561–573.

Wegener, B. (1991). Job Mobility and Social Ties: Social Resources, Prior Job, and Status Attainment. *Journal of Statistical Software*, *56*(1), 60–71.

Wertheimer, M., Riezler, K. (1984). Gestalt Theory. *Social Research*, *51*(1/2), 305–327.

Wickham, H. (2009). *ggplot: Elegant Graphics for Data Analysis*. New York: Springer.

Wickham, H. (2014). Tidy Data. *Journal of Statistical Software*, *59*(1), 1–23.

Wickham, H. (2011). ggplot2. *Computational Statistics*, *3*(2), 180–185.

Wu, X., Treiman, D. (2007). Inequality and Equality under Chinese Socialism: The Hukou System and Intergenerational Occupational Mobility. *American Journal of Sociology*, *113*(2), 415–445.

Xie, Y. (1992). The Log-Multiplicative Layer Effect Model for Comparing Mobility Tables. *American Sociological Review*, *57*(3), 380–395.

Xie, Y. (2010). Historical Trends in Social Mobility: Data, Methods, and Farming. University of Michigan Institute for Social Research, Discussion Paper.

Xie, Y., Goyette, K. (2003). Social Mobility and the Educational Choices of Asian Americans. *Social Science Research*, *32*(3), 467–498.