

Visualizing Data

Department of Government
The University of Texas at Austin

Fall 2010

Mondays 12:30-3:30 PM Batts Hall 5.102 GOV 385L (38820)

Professor

Dr. Samuel Workman

Office: Mezes 3.128

Email: sworkman@austin.utexas.edu

Phone: 512-232-1445

Office Hours: Tuesdays 11:00 AM to 2:00 PM, and by appointment.

Course Description

Visual displays of data inform each part of the process of research from exploration and description through causal inference and communication of results. Visual displays of data are an inherent part of communicating research findings to the broader academic audience oftentimes making, or breaking, an argument in a presentation or research paper. Despite the centrality of visuals to the conduct of good research, social scientists seldom put the same care into their figures and tables as they do in crafting their arguments. This course takes seriously the role of good graphics in both the process of analyzing data and communicating the results of research to a broader academic audience. The course focuses on the role of visual displays of data both in exploratory data analysis as well as in summarizing statistical results. Emphasis is placed on the principles of effective visualization with examples from the social sciences. The course addresses innovative visual displays and ways of thinking about, or approaching visualization. We will implement these techniques in graphical and statistical packages.

Grades

Grades for the course will be based on four components: class participation, three homework assignments, a final paper, and an in-class presentation. Class participation includes presenting material that we cover in class, asking questions, participating in discussion, and participating in our weekly discussion of graphics brought into class by the students. The three homework assignments will cover exploratory and descriptive graphical techniques, programming of new graphics, and the evaluation of statistical models with graphics. The final paper will explore a research question of interest to the student. The paper will begin with a very brief description of the central question guiding the research, relevant theories and hypotheses, and then move to the empirical portion of a larger research paper. The course will culminate in the students applying

the visual techniques learned in the course to present their research to the rest of the class. Grades will be comprised of 45 percent homework assignments (15 percent each), 35 percent for the final paper, 10 percent for the final presentation, and 10 percent for participation in the seminar. I will be using plus/minus grading in the course.

Assignments

Students will have two weekly assignments. Each student will submit to me by 12AM Sunday night two questions that pertain to the readings for each week. These questions should promote discussion or raise interesting questions about the material and its application to the conduct of research in the social sciences broadly or the student's particular subfield. Note that only ONE of the two questions may be a clarifying question (e.g. Can you please explain the concept of "small multiples"?). The students will also write a half page (single-spaced) *synthesis* of the readings that seeks to identify the fundamental principles of visualization and relate them to broader issues in the social sciences including inference and research design. In addition, over the course of the semester each student will twice submit a figure or graphic (it could contain data analysis or be a causal diagram) for discussion at the beginning of each course. The students may draw the first graph or figure from any source they choose (e.g. newspapers, textbooks, research, etc.). The second submission of a figure or graph for discussion should be drawn from the student's own research, or alternatively, from the student's subfield or particular area of interest. These two weekly assignments, along with class discussion will constitute 10 percent of the student's final grade.

During the course of the semester, students will also complete three homework assignments geared toward making the students think analytically about the role of visualization in the research process and apply techniques developed in the course using statistical and graphical computer programs. These three homework assignments will be collectively worth 45 percent of the student's final grade.

Students will also develop a research paper for the course. The research paper will lay out, in about two or three pages, the major questions, theories, and hypotheses brought to bare in the paper. The papers do not require the standard literature review and so are somewhat shorter than the typical submission to an academic journal (however, if students wish to write the full paper, they may do so). The central focus of the papers will be on what we typically associate with the back-end or data and methodology sections of research papers. In short, the paper is a journal submission or a paper that is moving in that direction, without the literature review and tedious citation of all the greats and gods. This final paper will comprise a hefty 35 percent of the final grade. The paper should integrate knowledge and techniques developed over the length of the course.

The culmination of the course is the presentation of the students' individual research projects to the class. It is of little value to create a provocative and powerful visual display and stumble in an attempt to communicate the information conveyed in the analysis. The final presentation will be worth 10 percent of the final grade.

Attendance and Assignments

Obviously, one cannot garner the points for participation if one is not in class. Do not miss class and do not be late. . . this is graduate school. As for late assignments, I do not accept them.

Texts

Required Books:

Colin Ware. 2004. *Information Visualization: Perception for Design*. 2nd Edition. Morgan Kaufmann.

William S. Cleveland. 1993. *Visualizing Data*. Hobart Press.
Ben Frye's MIT Dissertation: Available on Blackboard.
Christopher Adolph's Harvard Dissertation: Available on Blackboard.

Required Articles:

King, Gary, Michael Tomz, and Jason Wittenberg. 2000. "Making the Most of Statistical Analyses: Interpretation and Presentation." *American Journal of Political Science*. 44(2) 341-355.
Andrew Gelman, Cristian Pasarica, Rahul Dodhia. 2002. "Let's practice what we preach: Turning tables into graphs." *The American Statistician*. 56:2(May): 121-130.
Jake Bowers and Katherine W. Drake. 2005. "EDA for HLM: Visualization when Probabilistic Inference Fails." *Political Analysis*. 13(4): 301-326.
John O'Loughlin. 2002. "The Electoral Geography of Weimar Germany: Exploratory Spatial Data Analysis (ESDA) of Protestant Support for the Nazi Party." *Political Analysis*. 10(3): 217-243.
Kastellec, Jonathan P. and Leoni, Eduardo L. 2007. "Using Graphs Instead of Tables in Political Science." *Perspectives on Politics*. 5(4): 755-771
Tremmel, Lothar. 1995. "The Visual Separability of Plotting Symbols in Scatterplots." *Journal of Computational and Graphical Statistics*. 4(2): 101-112.
Gelman, Andrew. 2003. "A Bayesian Formulation of Exploratory Data Analysis and Goodness-of-Fit Testing" *International Statistical Review*. 71(2): 369-382.
Gelman, Andrew. Discussion Paper. "Exploratory Data Analysis for Complex Models." Posted on Blackboard.
Gelman, Andrew. Discussion Paper. "Rejoinder." To Buja on the role of Bayesian methods in model checking. Posted on Blackboard.
Buja, Andreas. Discussion Paper. Response to Gelman. Posted on Blackboard.

Optional:

Paul Murrell. 2006. *R Graphics*. Chapman & Hall.

Recommended:

Edward R. Tufte. 2001. *The Visual Display of Quantitative Information*. Graphics Press. 2nd ed.
Edward R. Tufte. 1997. *Visual And Statistical Thinking: Displays Of Evidence For Making Decisions*. Graphics Press.
Edward Tufte. 1997. *Visual Explanations*. Graphics Press.
Edward Tufte. 1990. *Envisioning Information*. Graphics Press.
Leland Wilkinson. 1999. *The Grammar of Graphics*. Springer-Verlag.
Michael Friendly. 2000. *Visualizing Categorical Data*. SAS Publishing.
R. Dennis Cook. 1998. *Regression Graphics*. Wiley Interscience.

Statistical and Word-Processing Resources for Social Scientists

Most statistical and word-processing software is not good when it comes to generating and displaying statistical graphics. Most packages (e.g. MS Excel, SPSS) combine the deadly sins of inflexibility and poor default settings. Students are free to use any statistical or graphical package they like, but should choose one which allows flexibility in generating nearly any graphic, a command-line or code interface (which may or may not have a graphical interface), and generates widely useable output such as Postscript or PDF. I use R and most of the examples in class and setup for homework will be based on R. I am proficient in some other statistical program languages, but students should be aware that choosing to work with a program other than R will limit my ability to aid them as problems arise (and problems always arise when working with data).

- R. Most, if not all, in-class examples of statistical visualizations will use the R statistical language. R is superior to all comers in the respects mentioned above. R may be obtained from <http://www.r-project.org>. Students should be aware that, although they may use a statistical or graphical package of their choosing, I can only promise detailed help in the R package.
- STATA. This one is very popular for political scientists. Where visualization is concerned, it has gotten much better since I first encountered STATA 6.0 back in the day. Still, STATA is no match for R in offering complete control of all aspects of visualization. Most, but not all, visualizations are possible in STATA. STATA model estimators and algorithms are as good as most any statistical package. As such, one potential strategy is to model in STATA and bring the output into R for visualization.
- SAS, Eviews, RATS, etc. See comments for STATA.
- Igor. An exception to the rule, Igor provides all the power and functionality of R with a graphical user interface. Igor may be obtained at <http://www.wavemetrics.com> for a student license fee of \$85.00.
- Adobe Illustrator. Great for important graphics generated by R and other programs for touch up, but pricey at \approx \$200.00 for a student license.
- LaTeX. It does no good to generate effective visualizations and be limited by the ability to include them seamlessly into research papers and presentations. A package that allows the smooth formatting of statistical graphics for inclusion in research papers and presentations is the powerful typesetting package LaTeX. LaTeX runs from a text editor such as WinEdt or Emacs. WinEdt may be obtained from <http://www.winedt.com> for an educational licensing fee of \$40.00. The Emacs text editor is free. Finally, LaTeX itself is also free and easy to install from this web site <http://www.latex-project.org/ftp.html>. If you choose to install LaTeX and use it, check out the powerful Beamer package for LaTeX presentations—my seminar slides are done in Beamer.

Academic Dishonesty

Students who violate University rules on scholastic dishonesty are subject to disciplinary penalties, including the possibility of failure in the course and/or dismissal from the University. Since such dishonesty harms the individual, all students, and the integrity of the University, policies on scholastic dishonesty will be strictly enforced.

University of Texas Honor Code

The core values of The University of Texas at Austin are learning, discovery, freedom, leadership, individual opportunity, and responsibility. Each member of the university is expected to uphold these values through integrity, honesty, trust, fairness, and respect toward peers and community.

Students with Disabilities

Students with disabilities may request appropriate academic accommodations from the Division of Diversity and Community Engagement, Services for Students with Disabilities at 471-6259 (voice) or 232-2937 (video phone) or <http://www.utexas.edu/diversity/ddce/ssd>

Organization of the Course

In general, there are two approaches to data visualization. This course takes a decidedly deductive, rather than inductive, approach to visual methodologies. Visualization is taught as a tool for investigating our theories and research questions. We will cover the approaches and techniques of visualization as it pertains to and comes to bear on the major phases of a research project both before and after attempts at causal modeling. The major areas covered are exploratory data analysis, descriptive inference, visualization in model selection and validation, and finally presenting quantitative results in articles and presentations. Each seminar will split somewhat evenly into the theory and conceptualization of visualization and the implementation of techniques for visualizing descriptive and causal inference. At all junctures, we will maintain the tight deductive connections between theory, conceptualization, measurement, and hypothesis testing.

0.1 Monday, August 30: What is this course and why am I here?

0.2 Monday, September 6: Labor Day Holiday (i.e. Don't Show Up)

1 General Theory of Data Visualization

In this section of the course, we will build a deductive approach to the use of visualization and contrast this with inductive approaches, especially data mining and demonstrate the superiority of thinking deductively with visuals. We will situate this approach within the broader process of conducting research. We will further examine how visuals are useful tools at each stage of the process from conceptualization and theoretical development through delivering this information in the form of papers and presentations.

1.1 Monday, September 13: Foundations of a Deductive Approach to Visualization

Topics: general approaches to visualization; visual perception and human cognition; types of data and associated plots.

Required Reading: Ware, Preface & Chapter 1; Cleveland, Preface & Introduction; Frye, Chapters 1 & 2.

Optional: Murrell, Preface, Chapter 1, & Appendix A.

1.2 Monday, September 20: Visualization in at the Front-End of the Research Process

Topics: visualization for conceptual and theoretical development; visualization as measurement (part 1); the fit between theory, measurement, and visuals.

Required Reading: Ware, Chapters 2-4 & Appendix C; Frye, Chapter 3; Adolph, Chapters 1 & 2.

Optional: Murrell, Chapters 2 & 3.

2 Visualization in Descriptive Inference

This section of the course examines visualization as a tool of descriptive inference and measurement. We will focus on the types of data scholars face and the array of plots that one might use in dealing with them. The theoretical focus of this section will be on keeping the tight connections between theorizing, conceptualization, and measurement. The section will cover visualization leading up to the choice, and execution, of a causal statistical model.

2.1 Monday, September 27: Exploratory Data Analysis and Descriptive Inference

Topics: what do you do with data?; visualizing the simple and most basic features of data; more on types of plots and their uses; visualization as measurement (part 2)

Required Reading: Ware, Chapters 5 & 6; Cleveland, Chapter 2; Frye, Chapter 4; Lothar (1995); Adolph, Chapter 3.

Optional Reading: Murrell, Chapter 4.

2.2 Monday, October 4: The Curse of Dimensionality

Topics: visualizing multivariate data; data reduction; how to visualize relationships—bivariate and trivariate data.

Required Readings: Ware, Chapters 7 & 8; Cleveland, Chapters 3 & 4; Frye, Chapter 5.

Optional Readings: Murrell, Chapter 5 & Appendix B.

2.3 Monday, October 11: The Curse of Time

Topics: special problems presented by time series data; techniques for visualizing time series data in R; examples from my own work.

Required Readings: Ware, Chapter 9; Frye, Chapters 6-8.

Optional Readings: Murrell, Chapter 5.

3 Visualization for Model Selection and Validation

So we have a model, how might visualization help us to understand the fragility and/or robustness of our statistical models? This section covers assessing our statistical models in terms of, well, statistics and in terms of their substantive importance—or lack thereof. In part, this section demonstrates how to convey that someone should “care” about your findings.

3.1 Monday, October 18: Non-parametric Visualization and Model Selection

Topics: visualizing non-parametric models, smoothing; EDA for model selection.

Required Reading: Ware, Chapters 10 & 11; Bowers & Drake (2005); O’Loughlin (2002).

Optional Reading: Murrell, Chapter Chapter 7.

3.2 Monday, October 25: Visualizing Inference-Statistical Implications

Topics: visualization for model checking, diagnostics, and uncertainty.

Required Reading: Cleveland, Chapters 5 & 6; Adolph, Chapter 4 & 5; Gelman (2003); Gelman & Buja Discussion.

3.3 Monday, November 1: Visualizing Inference-Substantive Implications

Topics: visualization for the substantive implications of models and the uncertainty surrounding these.

Required Reading: King, Tomz, & Wittenberg (2000); Adolph, Chapter 6 & 7.

3.4 Monday, November 8: Visualization for Model Validation

Topics: visualizing simulations; bootstrapping; forecasting; prediction uncertainty.

Required Reading: Adolph, Chapters 8 & 9.

4 Presenting and Communicating Quantitative Information

Assuming that we have a model, with findings that warrant dissemination, how might we craft papers and presentations in such a way that they convey the substantive and theoretical importance of our analyses? This section deals with communicating research findings in papers, books, and presentations. We will conclude this final section with each student presenting a mini job talk on the research they've conducted throughout the semester.

4.1 Monday, November 15: Visualization for Effective Presentation

Topics: preparing articles and presentations with effective visualizations; transforming tables into graphs; communicating research findings.

Required Reading: Gelman, Pasarica, and Dhodia (2002); Kastellec & Leoni (2007); selected readings from Gelman & Hill (provided on Blackboard).

4.2 Monday, November 22: Selected Topics in Visualization

Topics: review of the principles of visualization; selected topics chosen by the students or Professor; TBD.

4.3 Monday, November 29: Presentations of Student Research

Topics: students will present their research to the seminar.

4.4 Wednesday, December 8: Final Papers Due