CHAPTER NINETEEN

Beautiful Political Data

Andrew Gelman, Jonathan P. Kastellec, and Yair Ghitza

Some of the earliest historical examples of data analysis involve politics and government; even the word "statistics" reveals the connection of data collection for and about the state. Statistical pioneers, including Playfair, Laplace, and Galton, devoted much of their effort to designing and analyzing public data, and, in the 20th century, statistics was associated with Gallup polls, economic and military organization (Five Year Plans and all that), and even Svengali-like political consultants (as in *The 480*, a novel from 1964 by the coauthor of *The Ugly American*, *Fail-Safe*, and other Cold War–era bestsellers). More recently, TV viewers have become accustomed to colored maps and charts of the latest polls and election results broken down by locality and demographic slices. And at the next level of sophistication are *USA Today*, the *New York Times*, and blogs such as FiveThirtyEight.com.

This chapter gives some examples where data visualization has increased our understanding of politics, along with a discussion of the factors involved in making each choice. Here we are focusing on the uses of graphics for research as well as presentation.

We try to apply the following template:

- "Figure X shows…"
- "Each point (or line) in the graph represents…"
- "The separate graphs indicate…"

- “Before making this graph, we did…which didn’t work, because…”
- “A natural extension would be…”

We do not have a full theory of statistical graphics—our closest attempt is to link exploratory graphical displays to checking the fit of statistical models (Gelman 2003)—but we hope that this small bit of structure can help readers in their own efforts. We think of our graphs not as beautiful standalone artifacts but rather as tools to help us understand beautiful reality.

We illustrate using examples from our own work, not because our graphs are particularly beautiful, but because in these cases we know the story behind each plot.

Example 1: Redistricting and Partisan Bias

Figure 19-1 shows the estimated effect on partisan bias from redistricting (redrawing of the lines dividing the districts from which legislators get elected). Each point in the graph represents a state legislative election year (such as Missouri in 1972), with the vertical and horizontal axes displaying an estimate of partisan bias in that election and in the previous election, two years earlier.

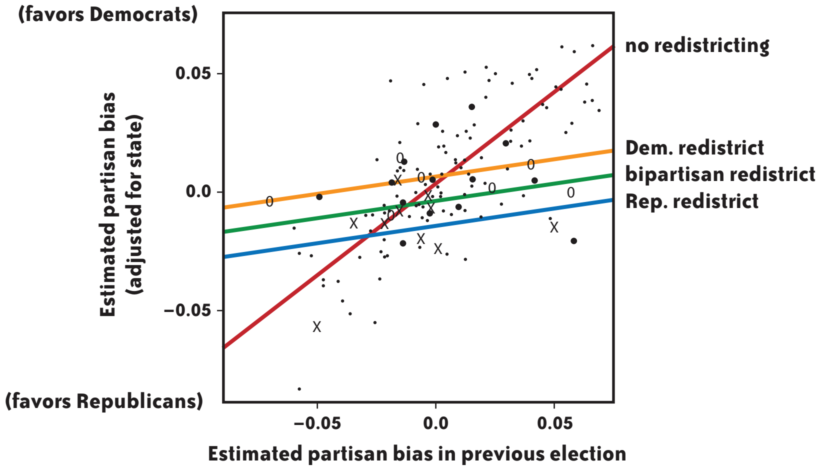

FIGURE 19-1. Effect of redistricting on partisan bias. Each symbol represents a state election year, with dots indicating controls (years with no redistricting) and the other symbols corresponding to different types of redistricting. As indicated by the fitted lines, the “before” value is much more predictive of the “after” value for the control cases than for the treated (redistricting) cases. The dominant effect of the treatment is to bring the expected value of partisan bias toward 0, and this effect would not be discovered with the usual approach, which is to fit a model assuming parallel regression lines for treated and control cases. This graph is just beautiful enough to reveal the key pattern in the data.

“Partisan bias,” as defined here, is a measure of how much the electoral system favors the Democrats or Republicans, after accounting for their vote share. Roughly speaking, the partisan bias is the expected Democratic share of the seats won in the legislature, if they were to average 50% of the vote. Biases are typically between –5% and 5%, implying that a party that wins half the vote for a state legislature will win between 45% and 55% of the seats.

The small dots in the graph represent “control” cases in which there was no redistricting, and the larger symbols correspond to different kinds of redistrictings, which here we lump together as “treated” cases. Elections come every two years, and redistricting typically happens every 10 years, so most of the data points are controls. The correlation between before and after measurements is much larger for controls than treated cases. The difference in slopes for the two groups should be no surprise at all. In the control cases with no redistricting, the state legislature changes very little, and so the partisan bias will probably change very little from the previous election. In contrast, when the legislative districts are redrawn, larger and more unpredictable changes occur. It was crucial to model the variation in the treatment to see this effect.

The simplest way to get partisan bias from redistricting is for Democrats, say, to draw the district lines so that they are winning with 60% of the vote in each of their districts, with Republicans packed together so that they are winning their seats with close to 100% of the vote. However, such manipulation (“gerrymandering”) may not be possible in practice, given constraints including equal population and contiguity of districts, as well as the potential for egregious gerrymanders to be overturned in court challenges.

The graph in Figure 19-1 is beautiful because, until we made it (in Gelman and King 1994), the discussion of partisan redistricting had focused on whether or not parties could make large gains and whether districting reduced the competitiveness of the electoral system (because legislators who are drawing the district lines can try to preserve “safe seats” for themselves and their colleagues).

In our first attempt to use this data to model the consequences of redistricting, we fit a linear regression model with no interaction—thus completely missing the most important part of the story. It was only after plotting the data and the fitted regression line that we noticed the elephant in the room and fit a more appropriate model.

Our graph showed that the main consequence of redistricting was to reduce the magnitude of partisan bias (and also to make the electoral system more responsive to voters, but that is the subject of a different graph, not shown here).

Example 2: Time Series of Estimates

Figure 19-2 illustrates a problem with classical logistic regression (a standard statistical tool for predicting yes/no outcomes) and how it can be resolved using a so-called weakly informative Bayesian approach. Using polling data in each presidential election from 1952 through 2000, we fit a separate logistic regression model to each year's data, predicting Republican vote choice given race, income, and several other variables.

Within each of the little graphs, each dot displays a logistic regression coefficient with a vertical line indicating the uncertainty in the estimate. The series of dots shows separate estimates for each election, and the two rows of graphs show the time series of estimated coefficients for race and income. (For simplicity, we do not display the other coefficients here.) The left column of the display shows classical estimates, and the two right columns show different Bayesian estimates (which in this case give essentially identical answers).

The estimates in Figure 19-2 look fine except in 1964, where there is complete separation, with all black respondents supporting the Democrats. As a result, the coefficient for race is estimated at negative infinity—that is, an inference that being black results in a 0% chance of voting Republican that year. 1964 was indeed a year in which Republicans did not do well among black voters (the Republican candidate that year was Barry Goldwater, who had opposed the Civil Rights Act), but they certainly received more than 0% of the black vote. The purpose of this regression, as in nearly all survey analysis, is to draw conclusions about the general population, not merely the small sample surveyed, and, as such, we cannot be satisfied with the classical estimate of negative infinity. (The estimate displayed in the left column of Figure 19-2 is not actually infinite, but that is because the software used to fit the model is iterative and stopped at some point before diverging.)

The Bayesian approach, as shown in the rightmost two columns of Figure 19-2, stabilizes the coefficient for black voters in 1964 at a reasonable value—lower than in any other year from 1952–2000 and with a larger uncertainty bound but not infinite. While fixing this problem, the Bayesian procedures did not mess up the coefficient estimates for other years or for other variables in the model (as illustrated by the coefficients for income in the second row of plots).

This graph is hardly beautiful, but it illustrates an important and general principle, which is that graphing isn't just for raw data. The usual practice in the statistical literature is to display this sort of result in a table, but a well-made graph can show more information in less space (Gelman et al. 2002).

From our own perspective, the graph of parameter estimates was useful both for conveying to others the effectiveness of our method and to reassure ourselves that our series of estimates was reasonable, in a way that a table of coefficient estimates (or, more typically, a long series of computer output) would not.

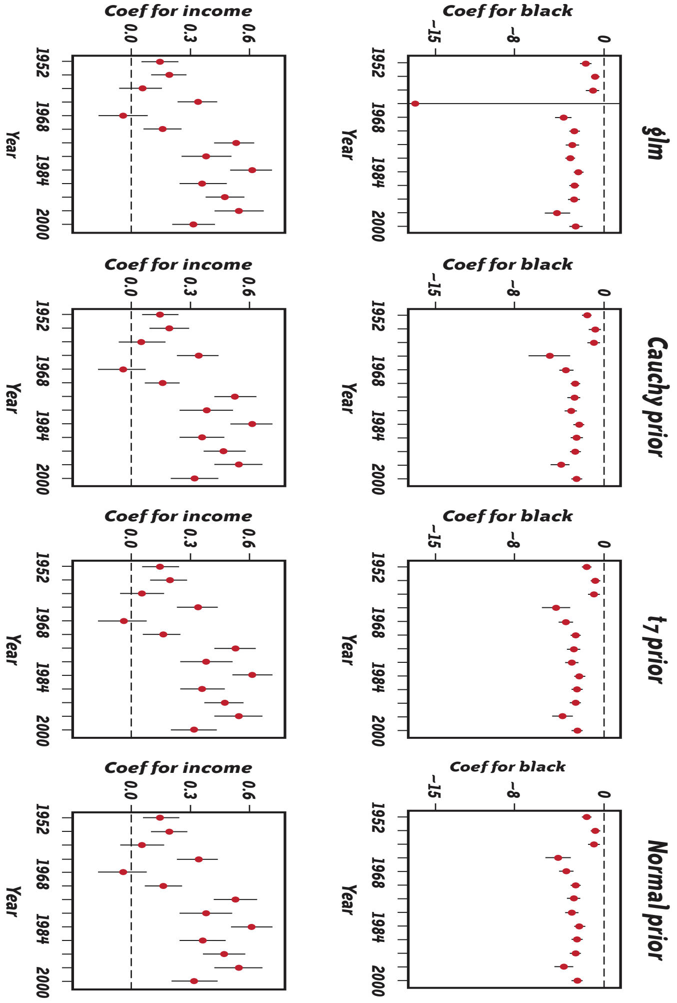

FIGURE 19-2. The left column shows the estimated coefficients (±1 standard error) for two predictors in a logistic regression predicting probability of Republican vote for President given demographics, as fit separately to data from the National Election Study for each election 1952 through 2000. The numerical variable income (originally on a 1–5 scale) has been centered and then rescaled by dividing by two standard deviations. There is complete separation in 1964 (with none of the African-American respondents supporting the Republican candidate, Barry Goldwater), leading to a coefficient estimate of −1 that year. (The particular finite values of the estimate and standard error are determined by the number of iterations used by glm function in R before stopping.) The other two columns show Bayesian estimates for the same model using different "weakly informative" prior distributions. The Bayesian inferences fix the problem with 1964 without doing much to the estimates in the other years.

The beauty of this graph, and others like it, is that its strict parallelism (the "small multiples" idea discussed by Tufte, 1990, and Bertin, 1967) allows the reader—and also the creator of the graph—to make many comparisons at once.

Example 3: Age and Voting

Immediately after Barack Obama's historic election, there was speculation about the role of young voters in the winning coalition. Exit poll data showed that Obama did particularly well among the young, but was this really newsworthy? For example, political consultant Mark Penn wrote on the *New York Times* website, "Sure, young people voted heavily for Mr. Obama, but they voted heavily for John Kerry." Was Penn right?

As always, the clearest way to make a comparison is using a graph. Figure 19-3 shows the results, with four versions: first the basic graph that we made on election night (pulling exit poll data off the CNN website), then an improved version posted by a student who had noticed our graph on the Web, then to more time series plots of our own. In each of these graphs, points are connected with lines, with points representing the Republican candidate's share of the two-party vote among each of four different age groups in several recent elections. 2008 clearly was different, and so Mark Penn was wrong—another case of a pundit looking at numbers and not seeing the big picture. This is what graphics is all about: showing the details and the patterns all at once.

To get to the even larger picture, there is a huge amount of research in this area, and we do not mean to imply that these graphs, which reveal some simple patterns, are in any sense a replacement for more serious study of patterns of age cohorts and voting over time.

Example 4: Public Opinion and Senate Voting on Supreme Court Nominees

Few decisions made by U.S. senators are as visible to the public as votes to confirm or reject a Supreme Court nominee. Whereas the outcomes of many Senate votes, such as spending bills or the modification of a statute, are ambiguous or obscured in procedural detail, the result of a vote on a Supreme Court nomination is stark: either the nominee is confirmed, allowing her to serve on the nation's highest court, or she is rejected, forcing the president to name another candidate (Kastellec et al. 2008). Do senators follow state-level public opinion when casting such votes?

Figure 19-4 presents a preliminary answer to this question by graphing the relationship between state-level public opinion on nine recent Supreme Court nominees and senators' votes on whether to confirm those nominees. On each graph, the curve shows the probability that a senator votes to confirm the nominee as a function of public opinion in the senator's state. The solid black line displays the estimated curve from a fitted logistic regression, and the clusters of light-gray lines depict uncertainty in this estimation. The hash marks (or "rugs") indicate votes of approval ("1") and rejection ("0") of nominees, while the numbers in the lower-right corner of each plot denote the overall vote tally by the Senate. The bottom plot pools all nominees together. We order the plots across and down by increasing mean support for each nominee.

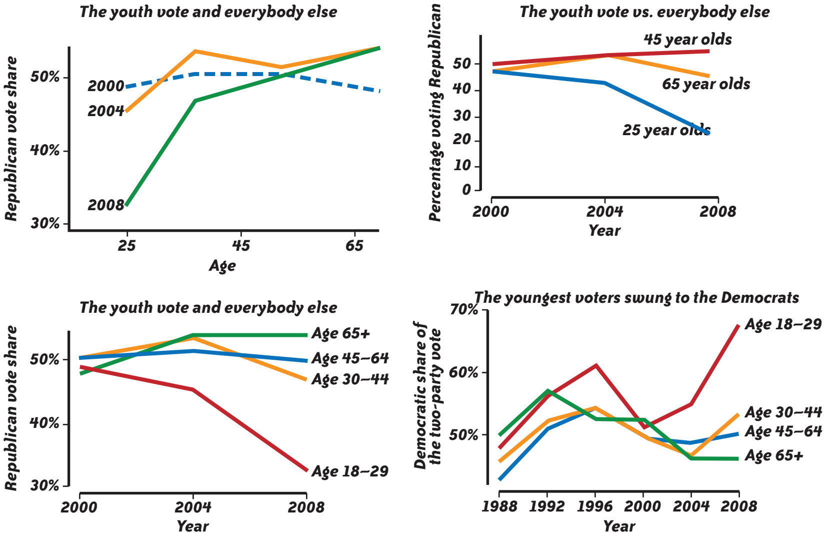

FIGURE 19-3. Some graphs showing recent patterns of voting by age. The top-left graph shows my first attempt, created on election night based on immediate exit poll data. The top-right graph was created by Hober Short, a student who saw my graph on the Web and made his own, displaying time on the x-axis. The lower-left graph is my cleaned-up version of Short's graph, labeling all four age categories directly on the lines of the graph. All these graphs show the dramatic difference between 2008 and the two previous elections. Finally, the lower-right graph extends the data back to 1988, showing that Bill Clinton in 1996 also did well among young voters—like Barack Obama, he was a young Democrat facing older Republican opponents—but not so well as Obama in 2008.

These graphs show the choices involved in making even the simplest possible graphs. As in many political settings, the largest gains come from incorporating additional data—in this case, the comparison of 2008 with earlier years, the comparison of young voters with those of other ages, and the comparison of the three other age groups with one another (with the lack of variation in this last comparison being a motivation to focus on trends among young voters in particular).

In addition, we improved our final graph by focusing on Democratic rather than Republican vote (more appropriate given the focus on Obama's strength among young voters) and by giving the graph a more descriptive title.

The graph shows that the relationship between public opinion and confirmation is generally positive, though it varies across nominees. Not surprisingly, there is greater uncertainty for nominees with lopsided confirmation votes. At the same time, the plot for "All Nominees" shows that, in general, as state public support for a nominee increases, a senator is more likely to vote yes. (This relationship holds even if one controls for other predictors of roll call voting, such as nominee quality and ideological distance between the senator and the nominee.)

The beauty of this graph is that it combines raw data with a simple inferential model in a single plot. Typically, bivariate relationships are presented in tabular form; in this example, doing so would require either nine correlation coefficients or regression coefficients and standard errors from nine regression models, which would be ungainly, make it difficult to visualize the relationship between opinion and voting for each nominee, and create difficulties in making comparisons across nominees. The only actual numbers we include

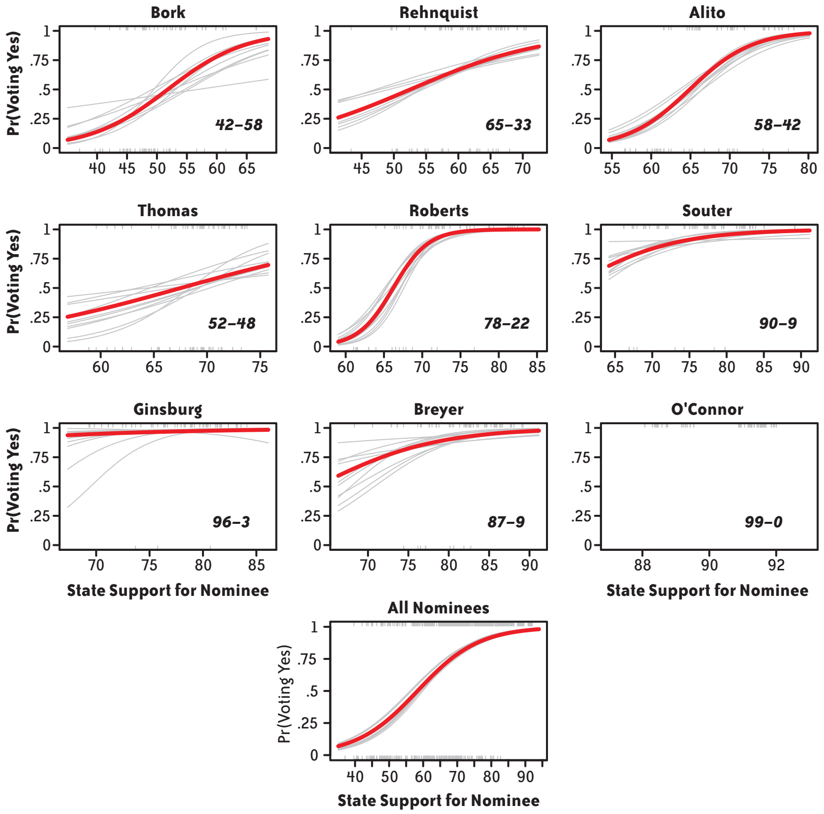

FIGURE 19-4. Correlation between state opinion and Senate roll call voting on Supreme Court nominees. For each nominee, the black line depicts the estimated logit curve from regressing senators' votes on state public opinion. Light-gray lines depict uncertainty in the estimates. Hash marks indicate votes of approval ("1") and rejection ("0") of nominees, while the numbers in the lower-right corner of each plot denote the overall vote tally by the Senate. The bottom plot pools all nominees together. The beauty of this graph is that it combines raw data with a simple inferential model in a single graph.

in the plot (which we do in an unobtrusive manner that does not distract from the plots themselves) are the roll call margins, which are both easily interpretable and give the reader a sense of how contentious each nomination was. Finally, as with Figure 19-2, the use of small multiples in the display allows the reader to make several comparisons at once, and prevents the information overload that can occur with a single plot.

Example 5: Localized Partisanship in Pennsylvania

In 1986, political strategist James Carville, who later ran Bill Clinton's first presidential campaign, described Pennsylvania as Paoli and Penn Hills with Alabama in between. Paoli is a suburb of Philadelphia, and Penn Hills is a suburb of Pittsburgh, and so Carville was referring to the two urban centers of this long-standing "swing state" as Democratic strongholds, with the remaining rural areas of the state as Republican territory.

Carville's words are indicative of the broader desire of both the public and the highest level of political punditry to divide the country into red and blue areas. For most Americans with an even passing familiarity with elections in the 21st century, one of the defining images of recent American politics has been the ubiquitous electoral map from 2000 and 2004, featuring slivers of blue states along the north and west coast, and a sea of red states in the south and the heartland. Despite President-elect Barack Obama's insistence that we are not a collection of red and blue states, this salient imagery is difficult to overcome.

Figure 19-5 presents a clarification of sorts for Carville's description of Pennsylvania and a different way of looking at geographic partisanship, based on a new and exciting type of data and a rich visualization technique. The bottom layer of the map shows Pennsylvania counties shaded by their 2004 presidential election returns, with blue indicating higher support for the Democratic candidate John Kerry, red indicating higher support for the Republican candidate George W. Bush, and shades of purple in between. By using the continuous red-purple-blue scale instead of the more common solid red or solid blue indicating each county's winner, we can better visualize the varying degrees of partisanship across the state.

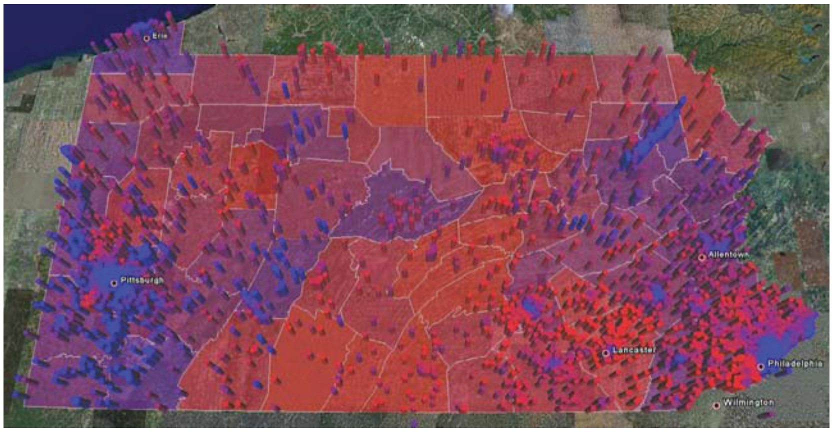

FIGURE 19-5. Geographic partisanship in Pennsylvania. The base layer shows Pennsylvania counties shaded by their 2004 presidential election returns, with blue indicating higher support for the Democratic candidate John Kerry, red indicating higher support for the Republican candidate George W. Bush, and shades of purple in between. The scattered cylinders represent localized partisanship for 4,000 random registered voters in the state, defined as the percentage of people living within a 1-mile radius who are registered Democrats. Each cylinder is located on the voter's household and has a radius of 1 mile, thus replicating the region for the partisanship measure. Again, blue cylinders indicate highly Democratic regions—this time with regard to individual-level registration—red cylinders indicate highly Republican regions, and shades of purple indicate regions in the middle. The beauty of this graph is that it reveals complexity in the idea of red and blue regions of the country, of individual states, and even of individual counties. (See Color Plate 69.)

The top layer of the map—the scattered cylinders—displays *localized partisanship* for a random sample of 4,000 registered voters in the state. Localized partisanship is a measure of the concentration of Democrats or Republicans in each neighborhood. Specifically, it is defined as the percentage of people living within a 1-mile radius who are registered Democrats. Each cylinder is located on the voter's household and has a radius of 1 mile, thus replicating the region for the partisanship measure. Again, blue cylinders indicate a highly Democratic region—this time with regards to individual-level registration—red cylinders indicate a highly Republican region, and shades of purple indicate regions in the middle.

The beauty of this graph is that it reveals complexity in the idea of red and blue regions of the country, of individual states, and even of individual counties. Although it is sometimes convenient to think of red and blue states, this graph reveals that there are shades of purple going down to the neighborhood (and even the individual) level. Just outside Philadelphia, the biggest city in the state, you can easily find pockets of red neighborhoods. Conversely, even in the reddest counties in the middle of the state, there are areas of purple and blue.

The graph is also beautiful because it demonstrates how our commonly held beliefs can be challenged and our understanding can be deepened through the careful analysis and visualization of data. This particular graph uses data provided by Catalist, a company that maintains a national database of all voting-age individuals in the United States. As detailed and large-scale data sources become increasingly accessible, multilayered visualization techniques will be instrumental in our abilities to use data to understand the world around us.

Conclusion

Political data is increasingly accessible and is increasingly being plotted and shared in the media and on the Web. At the research level, articles in political science journals are starting to make use of graphical techniques for discovery and presentation of results. And online tools ranging from NationMaster.com to the Name Voyager (*http://www.babynamewizard.com/voyager*) are becoming increasingly accessible, with data dumps such as Hans Rosling's TED talk (*http://www.ted.com/index.php/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen.html*) becoming cult favorites. We expect statistical visualization to become more important and more widespread in political analysis.

References

Bertin, J. (1967). *Semiology of Graphics*. Translated by W. J. Berg (1983). Madison: University of Wisconsin Press.

Gelman, A. (2003). "A Bayesian formulation of exploratory data analysis and goodness-of-fit testing." *International Statistical Review* 71, 369–382.

Gelman, A., A. Jakulin, M. G. Pittau, and Y. S. Su (2008). "A weakly informative default prior distribution for logistic and other regression models." *Annals of Applied Statistics*, to appear.

Gelman, A. and G. King (1994). "Enhancing democracy through legislative redistricting." *American Political Science Review* 88, 541–559.

Gelman, A., C. Pasarica, and R. Dodhia (2002). "Let's practice what we preach: turning tables into graphs." *American Statistician* 56, 121–130.

Kastellec, J., J. Lax, and J. Phillips (2008). "Public opinion and Senate confirmation of Supreme Court nominees." Technical report, Department of Political Science, Columbia University.

Tufte, E. R. (1990). *Envisioning Information*. Cheshire, CT: Graphics Press.