The ultimate quantitative extreme in textual data analysis uses scaling procedures borrowed from item response theory methods developed originally in psychometrics. Both Jon Slapin and Sven-Oliver Proksch's Poisson scaling model and Burt Monroe and Ko Maeda's similar scaling method assume that word frequencies are generated by a probabilistic function driven by the author's position on some latent scale of interest and can be used to estimate those latent positions relative to the positions of other texts. Such methods may be applied to word frequency matrixes constructed from texts with no human decision making of any kind. The disadvantage is that while the scaled estimates resulting from the procedure represent relative differences between texts, they must be interpreted if a researcher is to understand what politically significant differences the scaled results represent. This interpretation is not always self-evident.

Recent textual data analysis methods used in political science have also focused on classification: determining which category a given text belongs to. Recent examples include methods to categorize the topics debated in the U.S. Congress as a means of measuring political agendas. Variants on classification include recently developed methods designed to estimate accurately the proportions of categories of opinions about the U.S. presidency from blog postings, even though the classifier on which it is based performs poorly for individual texts. New methodologies for drawing more information from political texts continue to be developed, using clustering methodologies, more advanced item response theory models, support vector machines, and semisupervised and unsupervised machine learning techniques.

Kenneth Benoit
Trinity College
Dublin, Ireland

See also Data, Archival; Discourse Analysis; Interviews, Expert; Party Manifesto

Further Readings

Baumgartner, F. R., DeBoef, S. L., & Boydstun, A. E. (2008). *The decline of the death penalty and the discovery of innocence*. Cambridge, UK: Cambridge University Press.

Klingemann, H.-D., Volkens, A., Bara, J., Budge, I., & McDonald, M. (2006). *Mapping policy preferences II: Estimates for parties, electors, and governments in Eastern Europe, European Union and OECD 1990–2003*. Oxford, UK: Oxford University Press.

Laver, M., Benoit, K., & Garry, J. (2003). Extracting policy positions from political texts using words as data. *American Political Science Review, 97*(2), 311–331.

Leites, N., Bernaut, E., & Garthoff, R. L. (1951). Politburo images of Stalin. *World Politics, 3,* 317–339.

Monroe, B., & Maeda, K. (2004). *Talk's cheap: Text-based estimation of rhetorical ideal-points* (Working Paper). Lansing: Michigan State University.

Slapin, J. B., & Proksch, S.-O. (2008). A scaling model for estimating time-series party positions from texts. *American Journal of Political Science, 52*(3), 705–722.

Data Analysis, Exploratory

John W. Tukey, the definer of the phrase *exploratory data analysis* (EDA), made remarkable contributions to the physical and social sciences. In the matter of data analysis, his groundbreaking contributions included the fast Fourier transform algorithm and EDA. He reenergized descriptive statistics through EDA and changed the language and paradigm of statistics in doing so. Interestingly, it is hard, if not impossible, to find a precise definition of EDA in Tukey's writings. This is no great surprise, because he liked to work with vague concepts, things that could be made precise in several ways. It seems that he introduced EDA by describing its characteristics and creating novel tools. His descriptions include the following:

1. "Three of the main strategies of data analysis are: 1. graphical presentation. 2. provision of flexibility in viewpoint and in facilities, 3. intensive search for parsimony and simplicity." (Jones, 1986, Vol. IV, p. 558)
2. "In exploratory data analysis there can be no substitute for flexibility; for adapting what is calculated—and what we hope plotted—both to the needs of the situation and the clues that the data have already provided." (p. 736)
3. "I would like to convince you that the histogram is old-fashioned. . . ." (p. 741)

4. "Exploratory data analysis . . . does not need probability, significance or confidence." (p. 794)
5. "I hope that I have shown that exploratory data analysis is actively incisive rather than passively descriptive, with real emphasis on the discovery of the unexpected." (p. lxii)
6. "'Exploratory data analysis' is an attitude, a state of flexibility, a willingness to look for those things that we believe are not there, as well as those we believe to be there." (p. 806)
7. "Exploratory data analysis isolates patterns and features of the data and reveals these forcefully to the analyst." (Hoaglin, Mosteller, & Tukey, 1983, p. 1)
8. "If we need a short suggestion of what exploratory data analysis is, I would suggest that: 1. it is an attitude, AND 2. a flexibility, AND 3. some graph paper (or transparencies, or both)." (Jones, 1986, Vol. IV, p. 815)

This entry presents a selection of EDA techniques including tables, five-number summaries, stem-and-leaf displays, scatterplot matrices, box plots, residual plots, outliers, bag plots, smoothers, reexpressions, and median polishing. Graphics are a common theme. These are tools for looking in the data for structure, or for the lack of it.

Some of these tools of EDA will be illustrated here employing U.S. presidential elections data from 1952 through 2008. Specifically, Table 1 displays the percentage of the vote that the Democrats received in the states of California, Oregon, and Washington in those years. The percentages for the Republican and third-party candidates are not a present concern. In EDA, one seeks displays and quantities that provide insights, understanding, and surprises.

Table

A table is the simplest EDA object. It simply arranges the data in a convenient form. Table 1 is a two-way table.

Five-Number Summary

Given a batch of numbers, the five-number summary consists of the largest, smallest, median, and upper and lower quartiles. These numbers are useful for auditing a data set and for getting a feel for the data. More complex EDA tools may be based on them. For the California data, the five-number summary in percents is shown in Figure 1.

These data are centered at 47.6% and have a spread measured by the interquartile range of 8.75%. Tukey actually employed related quantities in a hope to avoid confusion.

Stem-and-Leaf Display

The numbers of Table 1 provide all the information, yet condensations can prove better. Figure 2

Table 1 Percentages of the Votes Cast for the Democratic Candidate in the Presidential Years 1952–2008

	Year														
State	*1952*	*1956*	*1960*	*1964*	*1968*	*1972*	*1976*	*1980*	*1984*	*1988*	*1992*	*1996*	*2000*	*2004*	*2008*
California	42.7	44.3	49.6	59.1	44.7	41.5	47.6	35.9	41.3	47.6	46.0	51.1	53.4	54.3	61.0
Oregon	38.9	44.8	44.7	63.7	43.8	42.3	47.6	38.7	43.7	51.3	42.5	47.2	47.0	51.3	56.7
Washington	44.7	45.4	45.4	62.0	47.2	38.6	46.1	37.3	42.8	50.1	45.1	49.8	50.2	52.8	57.7

Source: Statistical Abstracts of the U.S. Census Bureau.

Minimum	*Lower Quartile*	*Median*	*Upper Quartile*	*Maximum*
35.9	43.50	47.6	52.25	61.0

Figure 1 A Five-Number Summary for the California Democrat Percentages

Note: The minimum, 35.9%, occurred in 1980 and the maximum, 61%, in 2008.

3 | 6	3 | 6
4 | 12345688	4 | 1234
5 | 01349	4 | 5688
6 | 1	5 | 0134
	5 | 9
	6 | 1

Figure 2 Stem-and-Leaf Displays, With Scales of 1 and 2, for the California Democratic Data

provides a stem-and-leaf display for the data of the table. There are stems and leafs. The stem is a line with a value. See the numbers to the left of the " \| ". The leaves are numbers on a stem, the right-hand parts of the values displayed.

Using this exhibit, one can read, off various quartiles, the five-number summary approximately; see indications of skewness; and infer multiple modes.

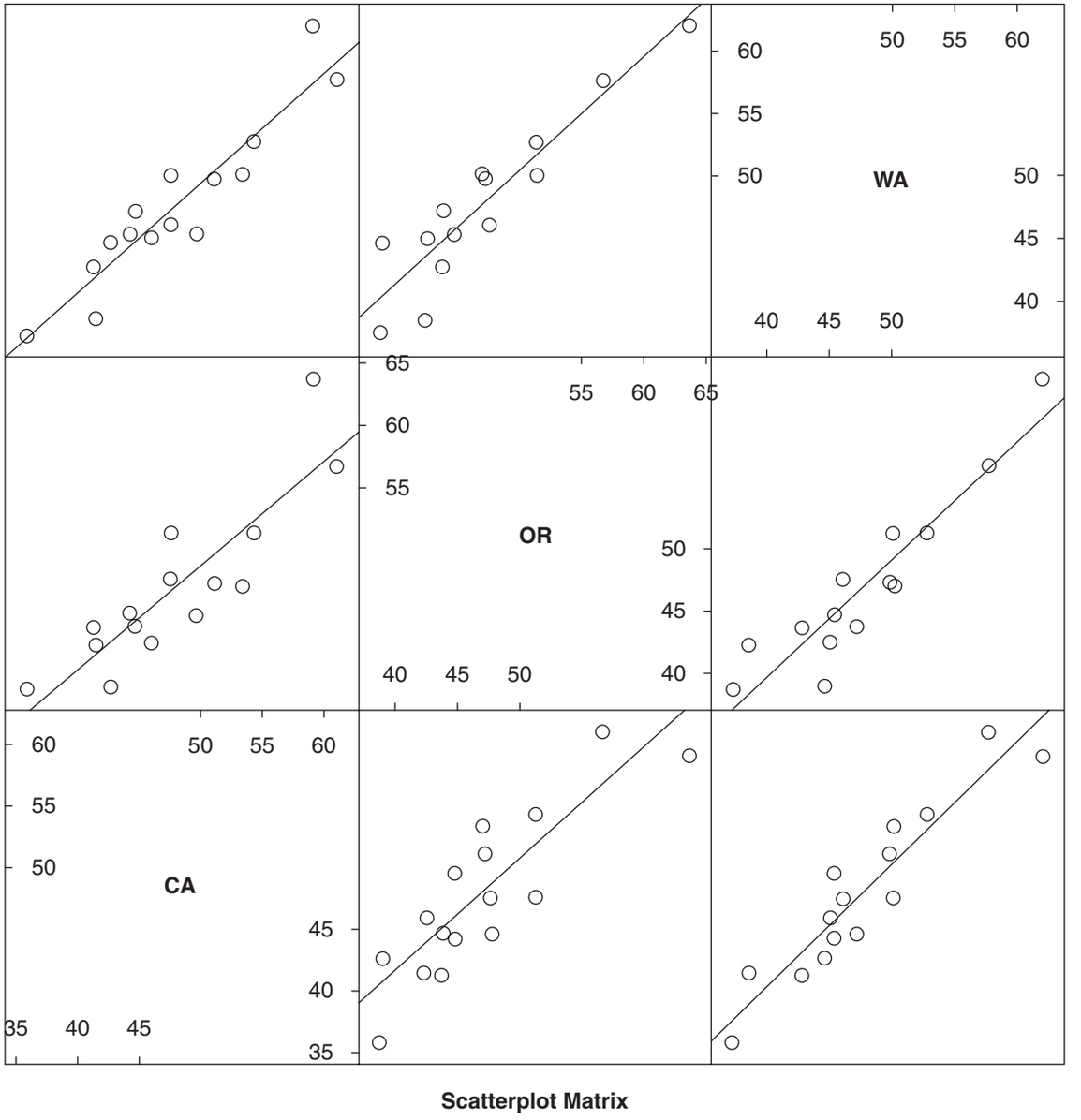

Figure 3 Scatterplots of Percentages for the States Versus Percentages for the States in Pairs

Notes: A least squares line has been added as a reference. CA = California; OR = Oregon; WA = Washington.

Scatterplot Matrix

Figure 3 displays individual scatterplots for the state pairs (CA, OR), (CA, WA), and (OR, WA). A least squares line has been added in each display to provide a reference. One sees the x and y values staying together. An advantage of the figure over three individual scatterplots is that one sees the plots simultaneously.

Outliers

An outlier is an observation strikingly far from some central value. It is an unusual value relative to the bulk of the data. Commonly computed quantities such as averages and least squares lines can be drastically affected by such values. Methods to detect outliers and to moderate their effects are needed. So far, the tools discussed in this entry have not found any clear outliers.

Box Plots A box plot consists of a rectangle with top and bottom sides at the levels of the quartiles, a horizontal line added at the level of the median, and whiskers, of length 1.5 times the interquartile range, added at the top and bottom. It is based on numerical values. Points outside these limits are plotted and are possible outliers. Figure 4 presents three box plots. When more than one box plot are present in a figure, they are referred to as *parallel box plots*.

Figure 4 presents a parallel box plot display for the presidential data. The California values tend to be higher than those of Oregon and Washington. Those show a single outlier each and a skewing toward higher values. Both the outliers are for the 1964 election.

Residual Plots

A residual plot is another tool for detecting outliers and noticing unusual patterns. Suppose there is a fit to the data, say, a least squares line. The residuals are then the differences between the data and their corresponding fitted values.

Consider the percentages in the table depending on the year of the election—that is, consider the data as a time series (Figure 5).

The time series of these three states track each other very well, and there is a suggestion of an outlier in each plot.

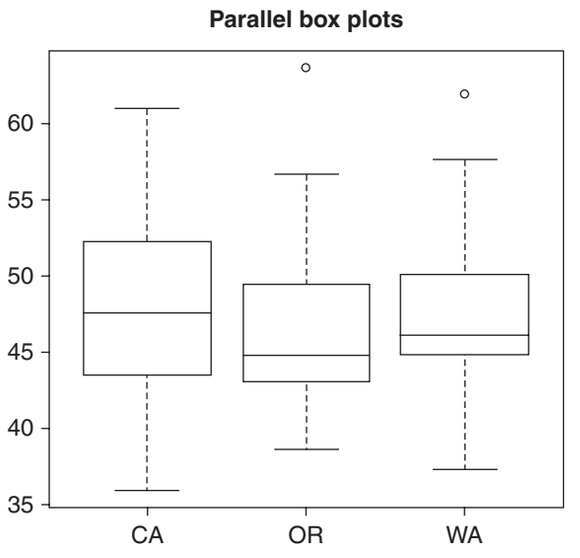

Figure 4 Parallel Box Plots for the Percentages, One for Each State

Note: CA = California; OR = Oregon; WA = Washington.

Figure 6 shows the residuals for the three states.

Each display in Figure 6 shows an outlier near the top. They all correspond to year 1964. This was the first year after John Kennedy was assassinated, and Lyndon Johnson received a substantial sympathy vote. There also is a suggestion of temporal dependence.

With today's large data sets, one wishes for automatic ways to identify and handle outliers and other unusual values. One speaks of resistant/robust methods, resistant methods being those not overly sensitive to the presence of outliers and robust ones being those not affected strongly by long tails in the distribution. In the case of bivariate data, one can consider the bag plot.

Bag Plots

The bag plot is a generalization of the box plots of Figure 4. It is often a convenient way to study the scatter of bivariate data. In the construction of a bag plot, one needs a bivariate median, analogs of the quartiles, and whiskers. Tukey and his collaborators developed these. The center of the bag plot is the Tukey median. The "bag" surrounds the center and contains the 50% of the observations with the greatest depth. The "fence" separates inliers from

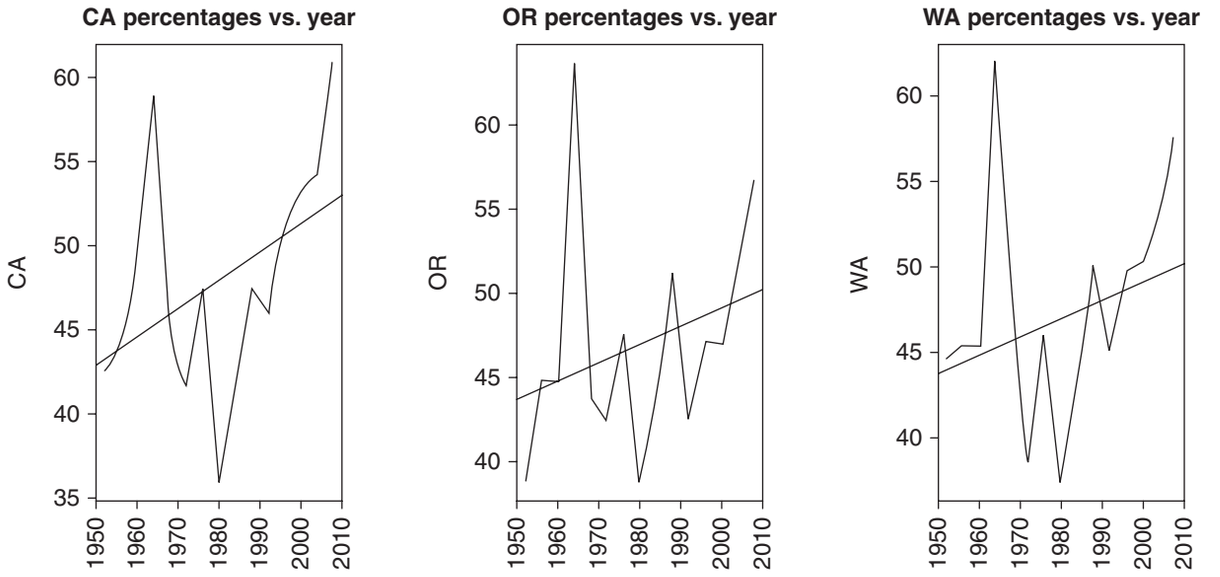

Figure 5 Graphs of the Individual State Democrat Percentages Versus the Election Years

Notes: A least squares line has been added as a reference. CA = California; OR = Oregon; WA = Washington.

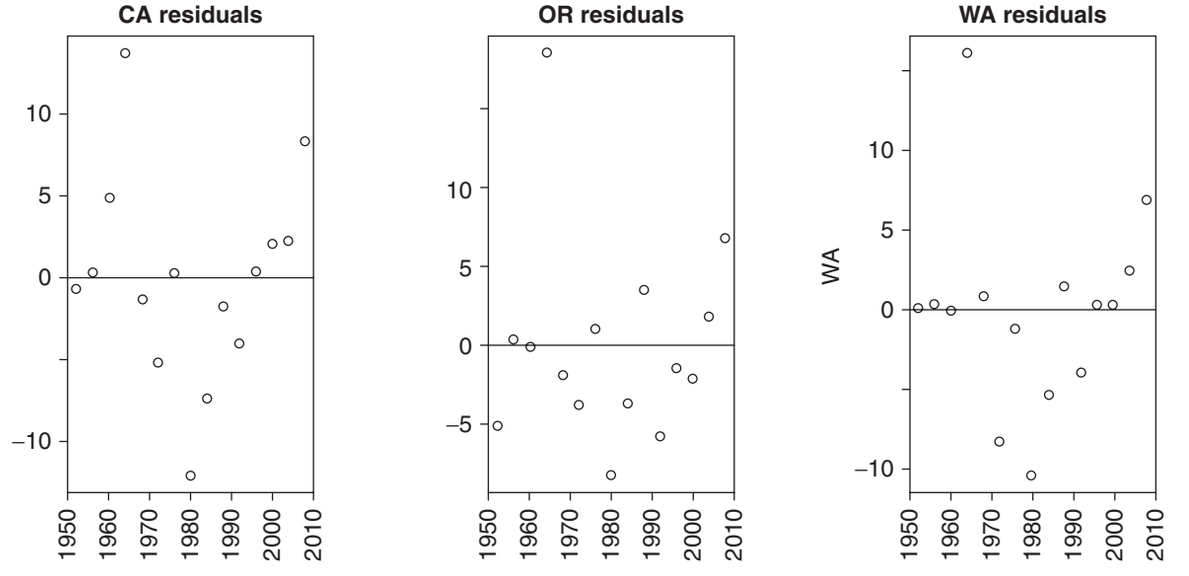

Figure 6 Residuals From Least Squares Line Versus Year With 0-Line Added

Note: CA = California; OR = Oregon; WA = Washington.

outliers. Lines called whiskers mark observations between the bag and the fence. The fence is obtained by inflating the bag, from the center, by a factor of 3.

Figure 7 provides bag plots for each of the pairs (CA, OR), (CA, WA), and (OR, WA).

One sees an apparent outlier in both the California versus Oregon and the California versus Washington cases. Interestingly there is not one for the Oregon versus Washington case. On inspection, it is seen that the outliers correspond to the 1964 election. One also sees that the points

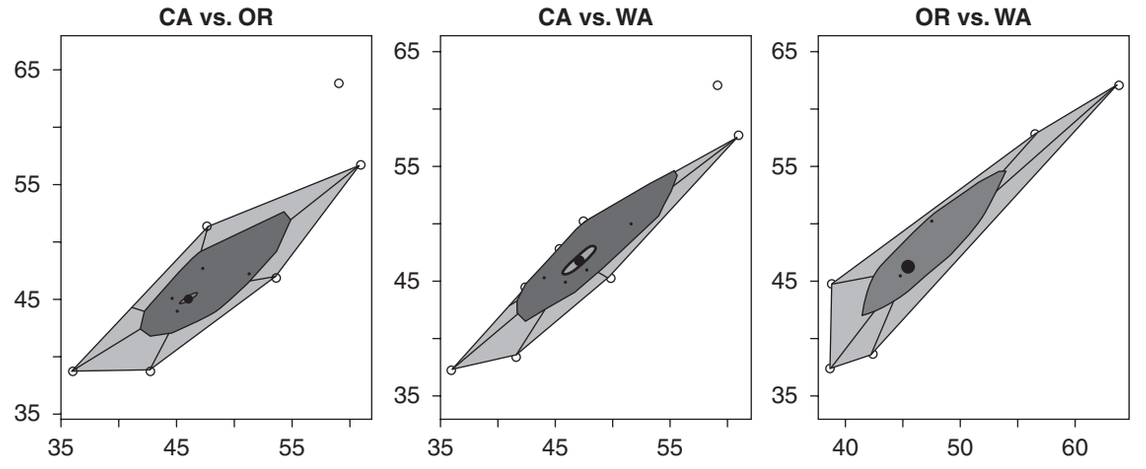

Figure 7 Bag Plots for the State Pairs, Percentages Versus Percentages

Note: The bivariate median is the black spot within the bag (darker shading), and the fence is the outer boundary.

vaguely surround a line. Because of the bag plot's resistance to outliers, the unusual point does not affect its location and shape.

Smoother

Smoothers have as a goal the replacement of a scatter of points by a smooth curve. Sometimes the effect of smoothing is dramatic and a signal appears. The curve resulting from smoothing might be a straight line. More usefully, a local least squares fit might be employed with the local curves, $y = f(x)$, a quadratics. The local character is often introduced by employing a kernel. A second kernel might be introduced to make the operation robust/resistant. It will have the effect of reducing the impact of points with large residuals.

Figure 8 shows the result of local smoothing of the Democrat percentages as a function of election year. The loess procedure was employed.

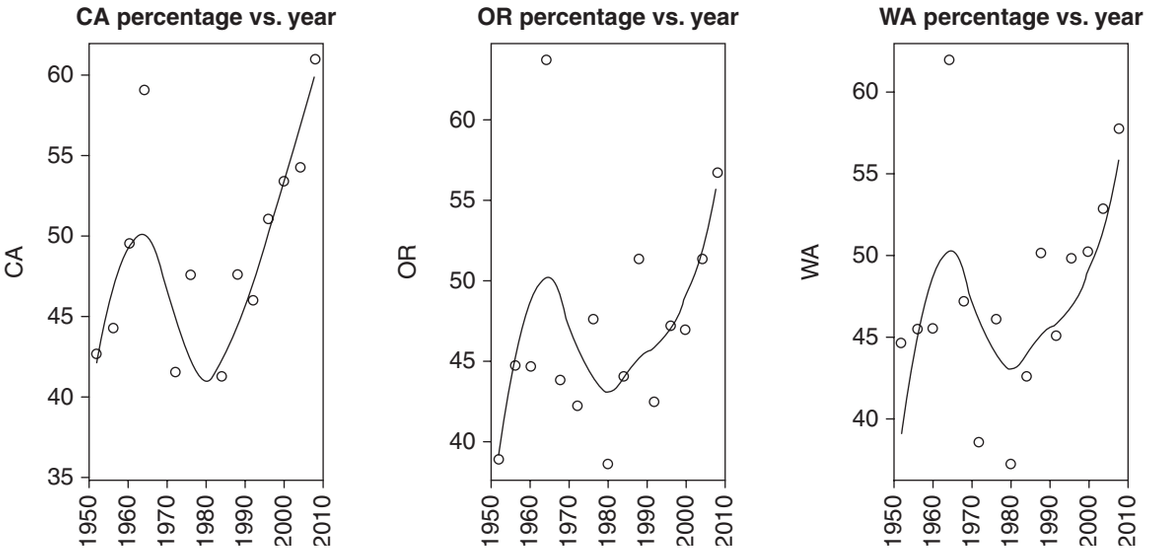

Figure 8 Percents Versus Year With a Loess Curve Superposed

Note: CA = California; OR = Oregon; WA = Washington.

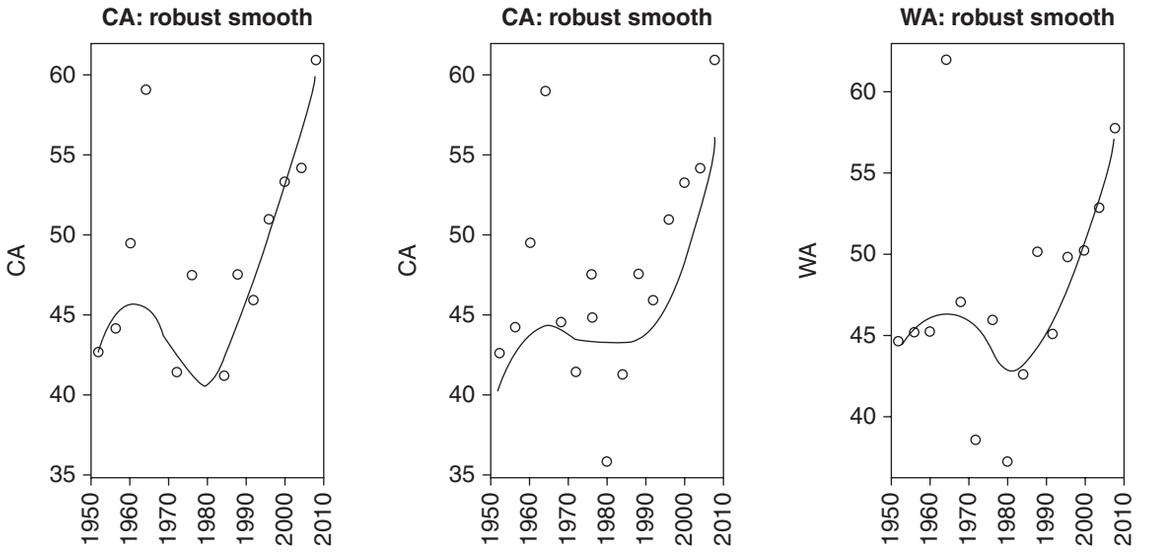

Figure 9 The Curves Plotted Are Now Resistant to Outliers

Note: CA = California; OR = Oregon; WA = Washington.

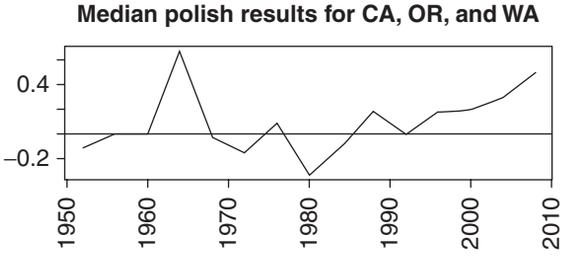

Figure 10 The Year Effect Obtained for the Election Data via Median Polishing

Note: CA = California; OR = Oregon; WA = Washington.

The curves have a similar general appearance. They are pulled up by the outlier at the 1964 point.

Robust Variant

The behavior in 1964 being understood to a degree, one would like an automatic way to obtain a curve not so strongly affected by this outlier. The loess procedure has a robust/resistant variant. The results follow in Figure 9.

Having understood that 1964 was an unusual year, one can use a robust curve to understand the other values better. The plots have similar shapes. One sees a general growth in the Democrat percentages starting around 1980. In this two-step procedure, it is important to study both the outlier and the robust/resistant curve.

Reexpression

This term refers to expressing the same information by different numbers, for example, using logit = $\log(p/(1 - p))$ instead of the proportion p. The purpose may be additivity, obtaining straightness or symmetry, or making variability more nearly uniform.

The final method is a tool for working with two-way tables.

Median Polish

This is a process of alternately finding and subtracting medians from rows and then columns and perhaps continuing to do this until the results do not change much. One purpose is to seek an additive model for a two-way table, in the presence of outliers in the data.

The state percentages in Table 1 form a 3×15 table and a candidate for median polish. The resulting row (year) effects are shown in Figure 10.

These effects are not meant to be strongly affected by outliers. Figure 10 shows the same general curve as in Figure 5.

This entry ends with Tukey's 1973 rejoinder: "Undoubtedly, the swing to exploratory data analysis will go somewhat too far" (cited in Jones, Vol. III, p. lxii).

David R. Brillinger
University of California, Berkeley
Berkeley, California, United States

See also Cross-Tabular Analysis; Data Visualization; Graphics, Statistical; Statistics: Overview

Further Readings

Bashford, K. E., & Tukey, J. W. (1999). *Graphical analysis of multivariate data*. London: Chapman & Hall.

Brillinger, D. R. (2002). John W. Tukey: The life and professional contributions. *Annals of Statistics, 30,* 1535–1575.

Hoaglin, D. C., Mosteller, F., & Tukey, J. W. (1983). *Understanding robust and exploratory data analysis*. New York: Wiley.

Jones, L. V. (Ed.). (1986). *The collected works of John W. Tukey: Philosophy and principles of data analysis 1949–1964* (Vols. III & IV). London: Chapman & Hall.

McNeil, D. R. (1977). *Interactive data analysis*. New York: Wiley.

Tukey, J. W. (1962). The future of data analysis. *Annals of Mathematical Statistics, 33,* 1–67.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, PA: Addison-Wesley.

Velleman, P. F., & Hoaglin, D. C. (1981). *Applications, basics, and computing of exploratory data analysis*. Boston: Duxbury.

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). New York: Springer.

Data Visualization

The basic objective of data visualization is to provide an efficient graphical display for summarizing and reasoning about quantitative information. Data visualization should be distinguished from other types of visualization used in political science (more general information and knowledge visualization, concept visualization, strategy and work flow visualization, metaphor visualization, etc.) as it is more specific to the representation of quantitative data existing in the form of numerical tables. In the following sections, the different types and methods of data visualization and their application in political science are presented.

Chart Types and Methods

During the past decades, political science has accumulated a large corpus of various kinds of data such as comprehensive fact books and atlases, characterizing all or most of existing states by multiple and objectively assessed numerical indicators within certain time periods (e.g., *OECD Factbook* and *Political Atlas of the Modern World*). As a consequence, there exists a tendency for political science to gradually become a more quantitative scientific field and to use quantitative information in analysis and reasoning. Any analysis in political science must be multidimensional and combine various sources of information; however, human capabilities for perception of large amounts of numerical information are limited. Hence, methods and approaches for the visualization of quantitative and qualitative data (especially multivariate data) are an extremely important topic in political science. Data visualization approaches can be classified into several groups, starting from creating informative charts and diagrams (statistical graphics and infographics) and ending with advanced statistical methods for visualizing multidimensional tables containing both quantitative and qualitative information. Data visualization in political science takes advantage of recent developments in computer science and computer graphics, statistical methods, methods of information visualization, visual design, and psychology. Data visualization in political science has certain special features such as the frequent use of geographical maps, which creates a link with the well-developed field of geographic information systems (GIS). Furthermore, numerical tables in political science are often incomplete, which makes important the use of methods dealing with missing or uncertainly measured data entries.

There are two main types of numerical tables that can be the subject of data visualization. The first one is called an *object–feature table*, where each row represents an observation or an object and each column corresponds to a numerical feature or indicator commonly measured for the whole set of objects. An example of such an object–feature table