Journal of the American Statistical Association, 85, 398–409.

George, E.I., and McCulloch, R.E. (1993), "Variable selection via Gibbs sampling," *Journal of the American Statistical Association*, 88, 881–889.

Hart, J.D. (1994), "Automated kernel smoothing of dependent data using time series cross-validation," *Journal of the Royal Statistical Society, Series B*, 56, 529–542.

Loader, C, *LocFit*, available at `http://cm.bell-labs.com/stat/project/locfit/adap.html`

Mitchell, T.J., and Beauchamp, J.J. (1988), "Bayesian variable selection in linear regression," *Journal of the American Statistical Association*, 83, 1023–1036.

Ruppert, D., Sheather S., and Wand M. (1995), "An effective bandwidth selector for local least squares regression," *Journal of the American Statistical Association*, 90, 1257–1270.

Smith, M., and Kohn, R. (in press), "Nonparametric regression using Bayesian variable selection," *Journal of Econometrics*.

Smith, M., and Kohn, R. (1996b), "Nonparametric bivariate regression," Manuscript.

Smith, M., Sheather S., and Kohn R. (in press), "Finite sample performance of robust Bayesian regression," *Computational Statistics*.

Smith, M., Wong C., and Kohn R. (1996), "Additive nonparametric regression with autocorrelated errors," under revision for *Journal of the Royal Statistical Society, Series B*.

Mike Smith
Department of Econometrics
Monash University
`mikes@smith.ecom.monash.edu.au`

Robert Kohn
Australian Graduate School of Management
`R.Kohn@unsw.edu.au`

TOPICS IN INFORMATION VISUALIZATION

Plot Interpretation and Information Webs: A Time-Series Example From the Bureau of Labor Statistics

By Daniel B. Carr, Richard Valliant, and Daniel Rope

1. Plot Interpretation, Metadata, and Software

Plot interpretation is an interesting topic because the ability to interpret plots is a measure of scientific literacy and because plots themselves can be informative. Tufte (1983) observes that the popular media in the U.S. considers the scatterplot beyond the common reader. Noting that some elementary school curricula include Cartesian plots, and even box and whisker plots, there is hope that the statistical literacy movement will eventually force a well-justified reassessment.

There is one form of scatterplot that routinely appears in the newspapers, the times series plot. This implies that when one variable is time something magical happens and people, especially those with economic interests, can understand. Section 2 in this paper suggests that when the other variable is measured in floating units, such as dollars, understanding the times series at more than a superficial level requires background information.

Interpretation of times series is problematic when articles suppress required metadata. Popular articles typically advocate a position and strive to be entertaining. Guidance from Entertainment 101 reads: "How to entertain with statistics: Omit 95%." Too often the role of statistics in an article is to suggest credibility and little more. Actual content may not be required. For the omissions one can chose from the seven basic building blocks of background information (metadata): who, what, where, when, why, how, and how well (W^5H^2). Who and why are important to omit whenever the answers raise issues of bias. How and how well are extremely important to omit in an advocacy context because they so often raise the undesirable issues of representativeness, complexity and uncertainty. In fairness, the motivation behind omissions may be brevity more than it is advocacy, but the negative effect can still be the same.

Publications at the federal statistical agencies often provide a striking contrast to popular media articles. At the federal agencies appropriate interpretation of plots is important and publications can devote significant space to metadata descriptions. However, even in lengthy government publications, the metadata description is often provided in text that is separated from summary plots and tables. This puts agencies at risk of having their graphics pulled out of context. It is easy to copy a plot and reproduce it without all the surrounding documentation. The article including the graphic will likely cite the source agency as part of the credibility gambit. Few are likely to pursue the information trail to the metadata to see if the graphic really supports the article.

Attaching metadata to every plot is problematic. In the static document context, Carr (1994a) suggests attaching icons that provide immediate warnings or serve as a cue for finding additional description. The emerging world of interactive documents provides additional alternatives that allow brevity and information access. To provide access to metadata and better graphics at the Bureau of Labor Statistics (BLS), Dan Rope and Dan Carr are developing a Java™-based Graphics Production Library (GPL). The resulting software provides metadata access through mousable metadata icons.

Historically BLS developed a table producing language (TPL). If there is an analogous graph producing language for the federal agencies, it is spreadsheet graphics. Spreadsheet graphics have remained to a large extent 1970's business graphics. Since the 1970's the statistical graphics community has provided much guidance about graphics and developed replacements for inferior graphics. Unfortunately this has had limited effect. For example the elegant dot plots described by Cleveland (1985, 1993a) remain hard to find in either the public media or government publications.

While education is important, the biggest barrier to routine use of preferable graphics is the limited availability of convenient software. We hope that the Java™-based GPL will provide easy access to sound graphics and that use of inferior graphics, such as 3-D bar plots, will diminish. A prototype applet derived from the GPL is scheduled for completion in August. What appears here is suggestive but preliminary. The carefully designed object-oriented library facilitates revision and the software will evolve as BLS statisticians, economists, cognitive scientists and icon designers provide input.

In what follows, Section 2 discusses the web of information behind a BLS time series and provides some interesting templates for time series graphics. Section 3 describes current web access to BLS series. Section 4 returns to the topic of metadata and provides an example of an applet with metadata access through icons and user interactivity.

2. Toward Understanding a BLS Time Series

Time series at BLS can be quite complicated. Computational techniques include adjusting series at benchmark periods, weighted averaging, seasonal adjustments, intervention analysis, scaling, and lagged ratioing (e.g. percent change over 12 months). Behind all this lies sampling methodology that determines the benchmarks, the weights, and other adjustments. Much of this is embedded in the time series shown in Figure 1. This figure, adapted from Employment Cost Indexes and Levels, 1975-1995, shows two series. The series are quarterly values depicted as occurring at mid-March, mid-June, mid-September and mid-December. The omission of values prior to 1980 provides resolution in subsequent monthly plots. The first series in the plot is marked "Current" and gives the 12 month change in wages and salaries for people in private industry as expressed in current dollars. The units are percent change and this calculation involves lagged ratioing.

Tufte (1983) warns that a series expressed in dollars can deliver that wrong message unless it is reexpressed in terms of constant dollars. The second series in Figure 1 is reexpressed in constant dollars. The message is quite different. Although the percentage changes in current dollar salaries are always positive, ranging from about 3% to 9%, the changes in constant dollars are often negative because of high inflation. The current dollar change of 9% in the first quarter of 1980 is actually more that a 4%

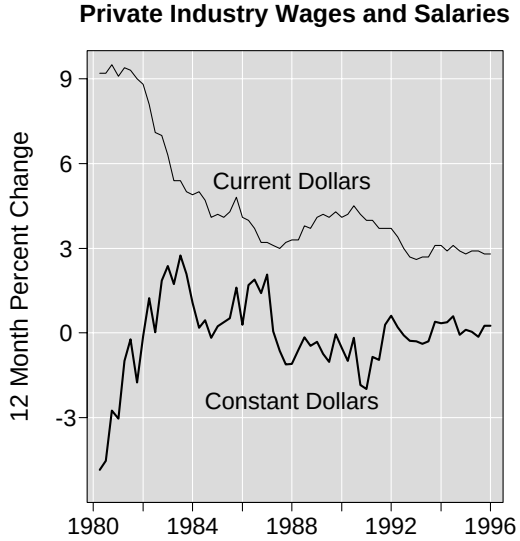


Figure 1: Simple Time Series?

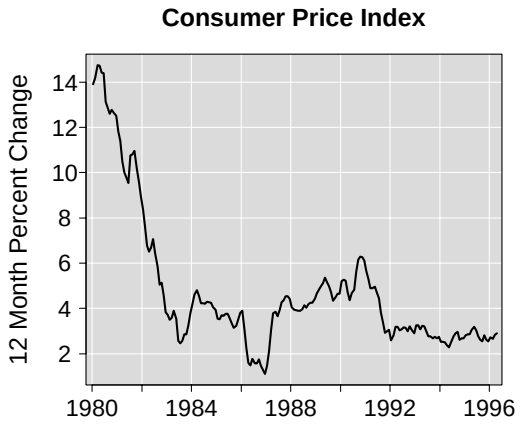


Figure 2: Key Series For Conversion To Constant Dollars

decrease in real dollars. The 1980 to mid-1983 period of deceleration in current dollar pay raises translates into a period of accelerating constant dollar pay changes.

Figure 1 suggests that the constant dollar salaries are pretty flat in the recent years. Of course this may not hold for all subsets of people, such as chief executive officers (CEOs). The decomposition of the wages and salaries series is of interest but the focus here is on the adjustment to constant dollars.

The adjustment to constant dollars involves scaling the salary series (before lagged ratioing) by the Consumer Price Index (CPI). Figure 2 shows the 12 month change in the CPI with values depicted at the middle of each month. Note the variability of the CPI as compared to wages and salaries. Since the CPI is variable and provides the basis for scaling to constant dollars, interested readers will want to know how the CPI is derived.

2.1 The CPI Subseries and Graphical Design

The CPI is the weighted average of seven major subseries. Figure 3 shows the CPI, the weights and the seven subseries. (These series are not seasonally adjusted–a procedure described in Section 2.2.) The bar plots to the left of the series show the weights. For example the housing cost has the largest weight of around 42%. In terms of graphic design, Figure 3 is a derivative of the row-labeled plots discussed in Carr (1994a). The design places the summary (the CPI series) first and places the subseries with the largest weights closest to this series. The figure communicates this graphically via the bars in the bar plots.

Figure 3 shows the time series as filled polygons. The fill is relative to the value 100. This calls attention to that value and the time period when the series were close to 100. As suggested by the plot, the benchmarking for the series occurred during the time period 1982 to 1984 where indexes were set to 100. Increases in the index reflect the relative increase in price for benchmark goods. The grid lines in the plot help in judging values and making comparisons (see Cleveland 1993b and Carr 1994b). Adding the polygon fill risks hiding the useful grid lines below the series. Figure 3 gets around this by using a translucent fill. That is, the grid lines show through the polygon fill.

Perhaps the most controversial design feature is the inclusion of text in the bar plots. Common media practice is to include both graphics and text in the attempt to communicate to both right brain and left brain, respectively. Unfortunately text is a distraction to right-brained pattern perception. The design here works towards a compromise by placing the text in the top right of the graph. Sorting of bars in decreasing order typically leaves this space free. An earlier version of this figure drew strong attention to the text by plotting white letters on a black background. Figure 3 attempts to reduce the text contrast with the rest of the graph. The box surrounding the text still draws undesirable attention, but omitting the box risks having the text misinterpreted as a special grid line label. The purest approach, of course, is to remove the text from the plot and put it with the series label.

The graph design provides grid labels for the CPI bar plot, but avoids repeating the labels for the identically scaled bar plots in the seven subseries. The reader may be reluctant to assume the scales are identical at first glance, but the similarity of plots and written values for bar heights provide strong evidence of identical scales. The inference that the highlighted bar represents the weight for the series should be straightforward. The grid-line labels are on the right for the time series. This facilitates reading the most recent values that are likely to be of greatest interest.

The series in Figure 3 show several interesting patterns. The increase in medical costs is dramatic and the subject of considerable media attention. However, the weight is small compared to that of housing. Controlling the cost of housing would have more effect on the CPI.

2.2 Seasonality

The wiggle in the Apparel and Upkeep Series draws attention. There are four cycles between the two year grid lines indicating two cycles per year. The first dip is in January when after-holiday sales are held. Prices then rise to a peak, typically in April, as winter fashions are replaced by ones for spring and summer. The second valley is in July when summer fashions are being phased phased out. The second peak then occurs in October or November as stores stock up on winter clothing.

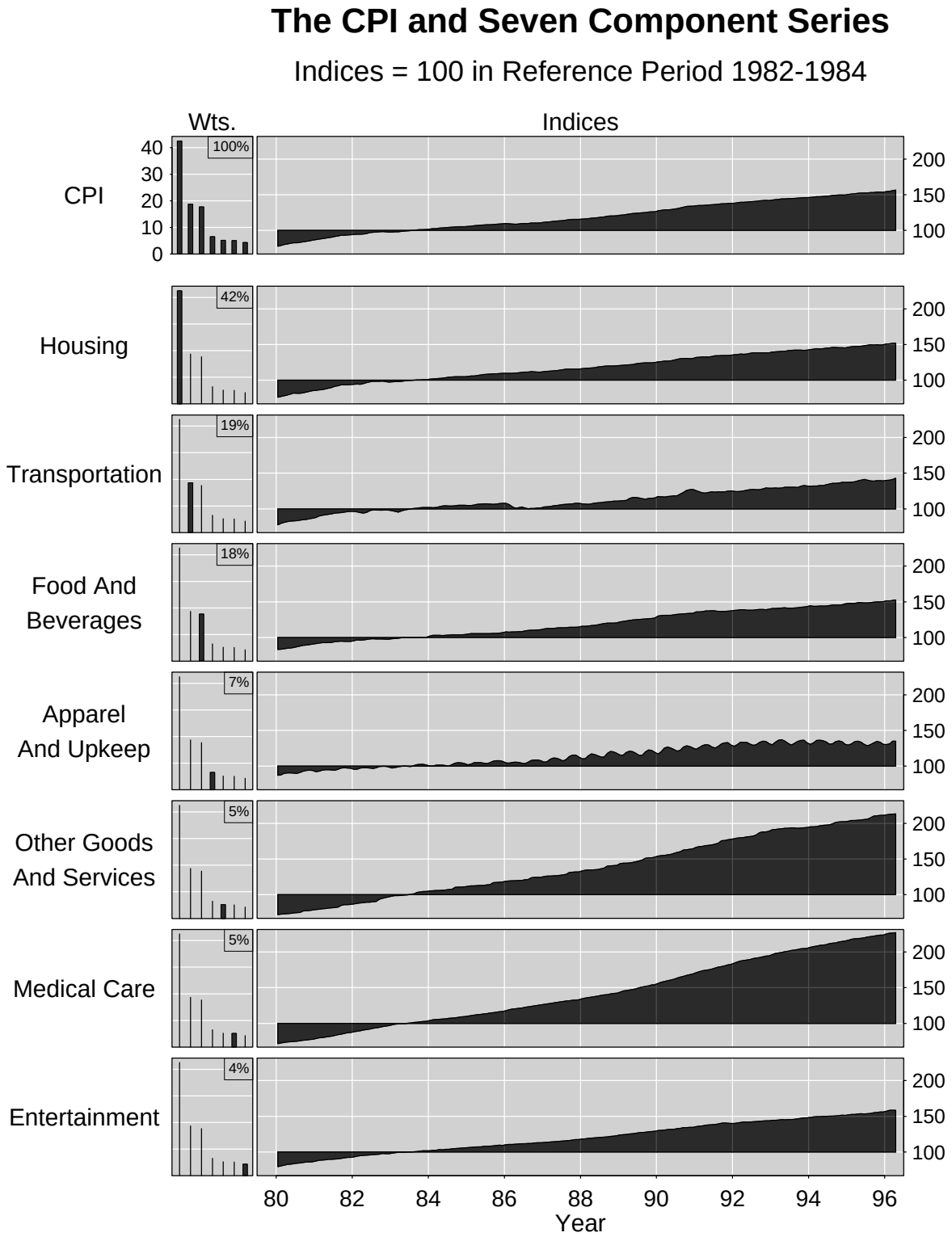


Figure 3: Multiple Time Series
With Relative Weights

Many BLS series have regular hops or skips that can be predicted with some accuracy. The unemployment rate, for example, rises every year in early summer as more students try (and fail) to find jobs. Prices drop in late December and early January for clothing as stores run after-holiday sales. Gasoline prices typically jump up in May or June when people begin to do more vacation driving. Price changes that normally occur about the same time and in about the same magnitude every year do not usually signal any general change in economic trends. For that reason, data series that have these seasonal fluctuations removed may be better for some analyses. Seasonal adjustments are additive or multiplicative factors applied to time series to reduce or eliminate the effects of regularly occurring short-term blips. Whether or not series have been seasonally adjusted is important metadata for the appropriate interpretation of BLS series.

2.3 Intervention Analysis and Seasonal Adjustment

The Transportation series has a distinct nonseasonal hump in the latter half of 1990. This occurred shortly after Iraq invaded Kuwait. The rise in prices was mainly associated with uncertainties in the supply of motor fuel and oil. From July through November, 1990, the Transportation index rose 7.2% but then decreased 3.9% from December, 1990, through April, 1991, as supply stabilized.

Other nonseasonal economic phenomena can also happen that cause a shift in level in a time series that is more long-lasting. Sales tax increases or oil embargoes are examples. Before doing seasonal adjustment, these level shifts are deleted from a series in a process called intervention analysis. Seasonal adjustment is then done on the series after which the level shift is injected back into the series.

2.4 Further Series Decomposition

Each of the seven series are decomposed into more detailed series that are also available through the BLS Homepage as indicated in Section 3. Food and Beverages, for example, is broken down into these subseries:

Food
- Food at home
  - Cereals and bakery products
  - Meats, poultry, fish, and eggs
  - Dairy products
  - Fruits and vegetables
  - Other food at home
    - Sugar and sweets
    - Fats and oils
    - Nonalcoholic beverages
    - Other prepared food
- Food away from home

Alcoholic beverages

2.5 Prices for the CPI

The inputs to the decomposed subseries are individual items for which prices can be determined. BLS collects prices for the CPI through a set of sample surveys that are probably some of the more complex surveys conducted by the Federal government. BLS collects prices in 85 urban areas across the country for about 7,300 housing units per month and 22,500 retail establishments-department stores, supermarkets, hospitals, gasoline stations, and other outlets that sell goods and services. The items that are priced are selected in several stages including probability sampling of geographic areas, establishments, and specific items within establishments. Prices on most of the sampled items are then collected by personal visit or telephone either monthly or bimonthly for up to five years, at which time new samples are rotated-in for each geographic sample area.

The fact that the CPI is a sample survey means that the time series BLS publishes are sample estimates and, consequently, subject to variability. How to graphically portray the variance or standard errors of time series is a problem we have not yet considered, but one that definitely needs attention.

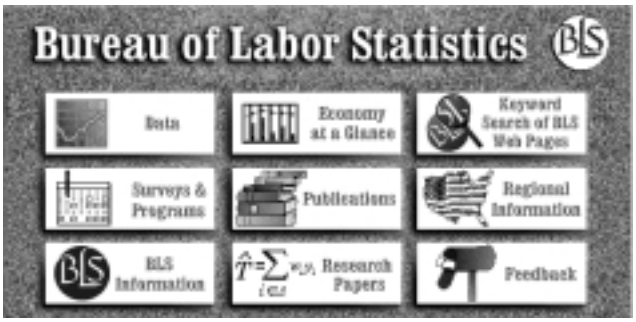
2.6 Relative Importance of Subseries

BLS publications often list the "relative importance" of each subseries. To compute the current relative importance, the procedure is to multiply each subseries benchmark weight by the current subseries index and normalize so the weights sum to one. One criticism of the CPI is that the benchmark weights in Figure 3 are not updated often enough to reflect changes in consumers' buying habits. As the relative price for gasoline falls over time, for example, consumers may increase their purchases and, as a result, gasoline's share in consumer budgets may exceed its benchmark weight in the CPI. The benchmark weights are updated every 10 years and will next be revised in 1998 as part of a major, regularly scheduled revision of the CPI.

3. Web Access to BLS Time Series

The Bureau of Labor Statistics (BLS) is the principal fact-finding agency for the Federal Government in the area of labor economics and statistics. It has a dual role as the statistical arm of the Department of Labor and

as an independent national statistical agency that collects, processes, analyzes, and disseminates economic and statistical data to the American public, Congress, other Federal agencies, State and local governments, business, and labor. In addition to the wage, salary, and consumer price data already mentioned, BLS data covers the areas of employment and unemployment, producer prices and prices of internationally traded good, compensation and benefits for productivity and technology, and employment projections.



BLS Homepage.

Much of the published data is available on the BLS Homepage (`http://stats.bls.gov`) along with links to general information on the Bureau, lists of paper publications, research papers, and other links. Users can retrieve data through their web browsers or gopher clients. For example, the unemployment rate for every month since January 1948, the monthly Consumer Price Index from 1913 to the present, and quarterly export and import price indexes beginning in 1983 are all retrievable time series published by BLS.

4. Future Graphics at BLS

BLS is not content with the current publications of tables and network distribution of data. BLS also seeks to provide user access to good graphical presentation of its time series and to provide more convenient access to metadata. The work in progress includes the development of a Graphics Production Library.

4.1 The Graphics Production Library

The GPL stems from a need for convenient tools for producing quality statistical graphics. Feedback from seminars shows that there is interest in applying new ideas for statistical graphics, but as suggested earlier, the reality is that people will use what is most convenient to create statistical graphs. The problem of producing quality graphics worsens when these graphics are included in World Wide Web pages. The rendered bitmaps are often barely legible on low resolution computer screens. To solve both of these problems at once we are developing platform-independent, extensible Java™ class libraries for producing graphics.

The library itself is composed of a set of reusable components (objects) that encapsulate drawing graphics. These objects can be "glued" together to create applets or applications, and the objects' underlying methods are used to change high level behavior and appearance. Applets are miniature software applications that run within Web browsers, and applications are standalone software products.

Beyond the ability to draw directly on the user's computer screen, one of the most significant advantages of this approach is that we can introduce a whole new level of user interaction with statistical graphics. The user will be able to directly manipulate the graphic to more quickly answer his or her questions about the data. In a sense, the graphic will become the interface, since the interface is the graphic. For instance, as illustrated in Figure 4, moving the mouse over a data point yields its value in the extreme lower left corner of the display. Other ideas include:

- The ability to drag & drop actual time series from different panels onto one panel to facilitate comparison.
- The ability to drag panels around with the mouse to reorder them.
- Providing windows that pop-up with more information about a data value (such as metadata or hyperlinks to further information) when the user clicks on a particular data point.
- The ability to navigate through the data by dragging or stretching the dark gray box located in the "Pan & Zoom" control.
- The ability to highlight points of interest. For instance, clicking the "A" in the "Highlight Month" control in Figure 4 yields a small tic on every April in all the time series.

Other advantages of using Java™ as the development language include database connectivity and object embedding. As part of the Java™ Enterprise API, the JDBC™ (Java Database Connectivity) will allow for applets with seamless connectivity to existing proprietary database formats including Sybase, Oracle, and ODBC compliant databases. The proposed Java™ "Beans" API will allow applications to be embedded within other software products. So, for example, graphics applications built using the library could be embedded directly within popular word processing packages. Our plan is to have an initial prototype of an applet ready sometime during August for feedback. For more information, see `http://www.mnsinc.com/dan_rope/gpl`

4.2 Metadata Icons and Access

Figure 4 includes just two metadata icons. The first is a set of books that provide access to a technical guide to BLS procedures, the BLS Handbook of Methods (1992). The second is a yield sign that provides access to interpretation warnings. Another icon might call attention to features in the graphics. Experience suggests that people sometimes miss basic features in the graphics. Spotting the basic features is a function of graphical experience, background in the subject matter, motivation, and time.

BLS has many types of metadata to represent. Sarndal, Swensson and Wretman (1992) discuss quality declaration for survey data. They suggest providing quality descriptors for a number of areas including:

Coverage
Sampling Error
Response Rates
Comparability over Time

Benchmarking and Revisions
Comparability with Other Data Sources
Total Variance and/or its Components by source
Non-response Bias

Response Bias
Edit and Imputation Effects
Seasonal Adjustment

We do not elaborate the descriptors here. Suffice it to say, that providing access to such descriptors is important and the BLS research is addressing the topic.

5. Access and Comments

Data, Splus functions and script files producing the first three figures are available via anonymous ftp to galaxy.gmu.edu. Change directory to `submissions/newsletter/bls`. The page layout function, text functions, and panel access function called `talk()` may be of interest to Splus users. Examples from other newsletter articles are also stored under the newsletter directory. As always, I (Dan Carr) welcome constructive comments plus new graphics challenges.

Acknowledgments

The current research was supported by an ASA/NSF/BLS fellowship. It builds on methodology develop under EPA cooperative research agreement CR820820-01-0 with George Mason University. Thanks go to Sallie Keller-McNulty who suggested improvements to Figure 3.

References

Carr, D.B. (1994a), "Converting Plots To Tables," Technical Report No. 101, Center for Computational Statistics, George Mason University, Fairfax Va 22030.

Carr, D.B. (1994b), "Using Gray in Plots," Statistical Computing & Graphics Newsletter, Vol. 5, No. 2, pp. 11-14.

Cleveland, W.S. (1985), *The Elements of Graphing Data*, Monterey CA: Wadsworth Advanced Books and Software.

Cleveland, W.S. (1993a), *Visualizing Data*, Summit NJ: Hobart Press.

Cleveland, W.S. (1993b) "A Model for Studying Display Methods of Statistical Graphics," *Journal of Computational and Graphical Statistics*, Vol 2, No. 4 , pp. 323-343.

Sarndal, C.E., Swensson, B., and Wretman, J., (1992), *Model Assisted Survey Sampling*, New York, N.Y.: Springer-Verlag.

Tufte, E.R. (1983), *The Visual Display of Quantitative Information*, Cheshire CT.: Graphics Press.

U.S. Department of Labor, (1992), *BLS Handbook of Methods*, Washington, D.C.: Government Printing Office.

U.S. Department of Labor, (1995), *Employment Cost Indexes and Levels, 1975-1995*, Washington, D.C.: Government Printing Office.

Daniel B. Carr
George Mason University
dcarr@voxel.galaxy.gmu.edu

Richard Valliant
Bureau of Labor Statistics
vallianr@ore.psb.bls.gov

Daniel Rope
Bureau of Labor Statistics
rope_d@bls.gov