Uplink

April 1996

A newsletter for the National Institute for Computer-Assisted Reporting

Statistics

Uplink update

Congratulations to the *Orange County Register* staff, who won the Pulitzer Prize for investigative reporting for "Fertility Fraud," and to *Raleigh News & Observer* reporters Pat Stith, Joby Warrick and Melanie Sill, who won the Pulitzer Prize for public service for "The Price of Pork," a look at the health risks of waste disposal in the hog industry. The next issue of Uplink will feature these and other contest winners and finalists who used computer-assisted reporting.

This issue of Uplink looks at statistics. Phil Meyer, author of "The New Precision Journalism," begins by giving an overview of statistical techniques and their use for journalists. Neill Borowski of the Philadelphia Inquirer provides instructions for doing economic analysis on the fly, and NICAR's Natalya Shulyakovskaya demonstrates how to parse using Excel.

Also check out our new data offerings from the NICAR library. We are providing the most complete and up-to-date FEC data available, including data you can't get online, in monthly updates.

Inside

CAR Trips
Page 6
On the Internet
Page 10
Bits, Bytes and Barks
Page 12

Stats add muscle to computer-assisted reporting

Feel the power

By Stan Dorsey
NICAR Staff

In the fast-moving world of computer-assisted reporting, journalists each day defy the widely held notion that reporters and numbers don't mix. They break down databases and computer record layouts with virtual ease, using programs such as FoxPro, Paradox, Access and Excel.

Just when it seems like they've done it all, something new comes along — something like statistics.

The use of statistics for research and analysis in the social sciences and in most medical fields has been going on for years. But it is a tool that is relatively unfamiliar to journalists, even those who are savvy in the general applications of CAR.

That does not have to be the case, says Philip Meyer, author of "The New Precision Journalism" and William Rand Kenan, Jr., professor of journalism at the University of North Carolina at Chapel Hill. The power of statistical analysis can be tapped by any journalist who is willing to learn, Meyer says.

Both Meyer and John Bare, a media consultant who works closely with Meyer and specializes in the use of statistics, agree that the first step in interpreting statistical analyses is to understand the basic model on which it based — the linear model.

It's a linear world

"Understanding the linear model," Bare says, "is the key to understanding other tools like linear regression or logistic regression. You should always know it first."

The linear model, Meyer says, "is based on the idea that many relationships in life fit a straight line, even when they don't seem to at first."

One way to conceptualize the linear model is to think back to the old days of high school or college algebra. On a

Continued on page two

SAS vs. SPSS

Playing with the big guns

By Steve Doig
The Miami Herald

If you can touch-type your way around a spreadsheet, and you speak fluent SQL to your database program, it's time to add a new power tool to your CARpentry workbench: One of the major statistical analysis software packages, either SAS or SPSS.

Here's where I'd like to tell you that, despite what you've heard, SAS and/or SPSS are cheap and easy to master. I'd like to, but I'd be telling a lie.

Actually, they're damned expensive, and they can do things you'll have trouble pronouncing (say "partial canonical correlation analysis"

Continued on page four

Uplink
April 1996
Volume 8, Number 4
A newsletter for the National Institute for Computer-Assisted Reporting

Editor
Brant Houston
brant@nicar.org
Managing Editor
Jo Craven
jo@nicar.org
Assistant Managing Editor
Andrew Lehren
andy@nicar.org
Senior Contributing Editors
Richard Mullins
richard@nicar.org
Rosemary Armao
jourrca@muccmail.missouri.edu
Staff
Wendy Charron
Stan Dorsey
Glaston Ford
Neil Reisner
Natalya Shulyakovskaya
Bryan Venable
Wallace Winfrey
Copy Editor
Marilyn Lehren

Uplink is published every month by the National Institute for Computer-Assisted Reporting, P.O. Box 838 Columbia, MO 65205. (573) 882-0684. Subscription price is $35 (U.S.) annually, $50 outside the United States. Postmaster: Please send address changes to NICAR. Send e-mail to nicar@muccmail.missouri.edu.

NICAR is a joint effort of Investigative Reporters and Editors and the University of Missouri School of Journalism. NICAR services include hands-on newsroom training on computer-assisted reporting, special academic and advanced training in data analysis. NICAR is supported by a grant from The Freedom Forum and other foundations intended to help the institute deliver its services nationwide to news organizations and associations.

From page one:

Feel the power of stats

basic graph with a y axis and an x axis, the linear model posits that as the x, or independent variable, changes, and so changes the y ,or dependent variable. In other words, the amount that y increases or decreases is directly dependent on changes in the x variable.

Consider the real life application offered by Meyer, which deals with state SAT score averages.

"The scores from each state would be directly dependent upon how many students took the test. The fewer students who take it, generally that means a higher score, and vice versa," Meyer says.

The degree to which a comparison fits a straight line is referred to as a linear correlation. So, for example, there is a strong linear correlation between the temperature outside and how much you sweat.

Linear correlation leads to the discussion of the first and simplest type of statistical analysis journalists use. It is called linear regression and follows the exact concept explained by the linear model. Regression refers to the relationship between two or more correlated variables.

Meyer says that a good use of linear regression is for dealing with economics. It allows you to make a prediction about something based on information you are sure about. For example, it is known that when the interest rate increases, the demand for housing decreases. Based on this linear correlation, we should be able to predict a rise or fall in the demand for housing by simply looking for a trend in interest rates.

Logistic regression

A more complicated application of statistics is called logistic regression. The main difference between linear and logistic regression is in the nature of the x and y (independent and dependent) variables.

In linear regression the two variables are continuous, meaning that there is a great range in their potential value. In the example of the SAT scores, the dependent variable can range from 400 to 1600, and the independent variable can be any number of students.

In logistic regression, however, at least one of the variables involved is referred to as a categorical or binary variable. In stark contrast to continuous variables, categorical variables have dichotomous values such as male/female, yes/no or dead/alive.

Logistic regression still involves finding a relationship between two or more factors, but it allows you to do it in a far more powerful, meaningful way.

"The beauty of logistic regression," Meyer says, "is that you are able to take several independent variables and use them to estimate the odds that your dependent variable will be one or zero."

Using the different independent variables, Meyer says, enables you to have control over certain variables that you ordinarily may not.

A case where this is most evident is in medical reporting. Very often reporters are interested in investigating mysterious death rates (death is a categorical variable) at a particular medical institution. Using logistic regression allows them to control for certain independent factors such as the patient's age, level of income, health history or insurance coverage.

Controlling for these factors in a comparison of death rates has the effect of leveling the playing field for all the institutions being examined. This is precisely what New York Times reporter Josh Barbanel attempted to do in his investigation of the unusually high rate of newborn deaths at certain New York City hospitals.

Multiple regression

Multiple regression is very similar to logistic regression with the exception that you are able to include more than one dependent variable.

Many education beat reporters have found this technique particularly valuable, Meyer says, when attempting to measure the efficiency of particular schools in a district by using graduate rates and test scores of students.

In these studies, variables such as the amount of money per pupil and the level of income per family can be controlled for. Dave Davis of The Plain Dealer+ found this out when he embarked upon a study of the test scores for 600 Ohio public schools.

Davis and three other education reporters controlled for variables such as poverty, parents' education level, and the percentage of renters in a district as a measure of how often families move. Their study found that these factors these account

Continued on page three

Looking at police stats

It ought to be a crime

By Bill Bell
Missouri School of Journalism

Last Christmas, Jeff Green of *The Oakland Press* in Pontiac, Mich., learned an important lesson: Do it yourself.

In the Christmas to New Year's lull, Green wrote a series of stories on growth and crime with fellow reporter Kathy Gray. Green got the jump on his competitors by circumventing the FBI and going straight to his local police departments.

Over the course of several weeks, Green asked about 60 law enforcement agencies in Oakland County for their tallies in eight crime categories: murder, rape, robbery, assault, burglary, larceny, auto theft and arson. Green pooled his numbers with reporters from a local television station.

Using his IBM laptop, Green put together his own database on Excel 5.0 and Fox Pro 2.6. He estimated it took seven hours to clean up and track down errors in the data. Green found that many data discrepancies stemmed from the structure of the police department.

Overall, Green said, his data was cleaner and more accurate than the crime statistics from the FBI. "It's better than the FBI database because the FBI has blanks and stuff," Green said. "If you are going to start comparing Community A to Community B, you better go to Community A and Community B."

Green also researched the correlation between commercial growth and crime. He used data on property tax values from the Oakland County Planning Department.

Green, a science, demographics and environment reporter with a background in statistics, said he is planning to follow up the series by seeing if the number of police officers is keeping up with community growth. Research last year showed the fire department increasingly understaffed.

Green advised reporters to look at the statistical significance of the data. Reporters are notorious for writing stories about startling numbers without looking at the whole statistical picture, he said. Small communities may experience a large percentage increase in crime but still have relatively low crime rates.

"A real pitfall in these type of stories is paying too much attention to the change and not to the actual numbers," Green said. "A lot of places with the fastest growing crime rates were also still the safest communities."

Jeff Green can be reached at (810) 745-4647, or send e-mail to user@cris.com.

The following week-long bootcamps offer hands-on training in computer-assisted reporting skills:

- **Advanced Computer-Assisted Reporting Seminar, May 5-10, featuring statistics and mapping, offered jointly with the University of North Carolina at Chapel Hill, N.C.,**
- **NICAR Bootcamps, featuring training in the use of spreadsheets and database managers, accessing data in various media, such as nine-track tapes, and negotiating for data; May 19-24; July 14-19, Columbia, Mo.**

These dates are open to all journalists. For more information, call NICAR (573) 882-0684, or send e-mail to nicar@muccmail.misssouri.edu

From page two:

Statistics add muscle to CAR

for nearly 60 percent of the difference in test scores between districts.

But more importantly, they were able to target schools that were not performing up to the standards which Davis' demographic analysis predicted they should. Some schools, on the other hand were performing well above predicted levels, despite difficult social circumstances.

Reliability analysis

A final tool, which Meyer says he finds very useful in making his stories relevant to their readers, is reliability analysis. It is based on the principle of intercorrelation, which measures the degree to which several different items measure the same thing.

This is particularly helpful in doing surveys.

"Reliability analysis allows you to take several different survey questions and find out how related they are to each other," he says. "Once you do that, you can compute a single number, which expresses your outcome."

By expressing complete survey results with only one figure, reporters can easily compare results from different segments of the population without drowning the reader in numbers.

Regardless of which method reporters decide is best for their story, the use of statistics is not as scary as it may sound. But Davis is quick to note that it is only part of the process.

"It's great for getting numbers crunched," he says, "but you still have to put everything in a relevant context for your readers."

Stan Dorsey can be reached at (573) 882-0684, or send e-mail to stan@nicar.org.

From page one:

Playing with the big guns

three times), much less readily understand how to use.

Software with strength

So why would I recommend that CAR specialists who already are happy with programs like Excel and FoxPro begin learning such powerful software? For that very reason — the power.

As you might expect, both SAS and SPSS are loaded with a huge array of functions for doing sophisticated statistical analyses. With them, you can do just about anything from simple crosstabs and frequency tables up through multiple linear regression and on to such exotica as cluster, factor and discriminant analyses.

For the past six years, most of my serious CAR work has been done with SAS. (I've also experimented with SPSS.) In that time, I've used perhaps 10 percent of SAS's capabilities. And I'm well aware there is considerable danger in all that power. Without a solid grounding in statistics, which I'm still struggling to acquire, you can unknowingly commit all sorts of career-threatening errors. As someone once noted, statistics can be made to say anything if you torture them enough.

Built for speed

But I recommend SAS or SPSS for much more than their statistical capabilities. Most important to CAR folks, both programs are filled with features designed to easily handle the myriad messy ways we receive our data. Signed packed decimals? Japanese date formats? Octals? Hierarchical data layouts? Thousands of variables per record? SAS or SPSS eat these like candy.

And they're built for speed. Recently, I was working with 24 fixed-length ascii files totaling about 19 megabytes. Microsoft Access 2.0 needed a total of 124 seconds to import one 27,000-record file and produce a simple crosstab. It took SAS just 14 seconds to do the same thing on the same 486-66 computer. In fact, SAS batch-imported all 24 files in the time it took Access to import one.

There isn't room here to do a full and fair comparison of all the features of SAS and SPSS. But based on my experience, here's the short version: SPSS 7.0 for Win95 costs less than SAS and is easier to begin using, thanks to a very user-friendly, point-and-click interface. The SPSS basic module lists at $695, which includes most of the statistical procedures and graphics that you're likely to need. For hundreds of dollars more, you can buy advanced modules that handle such problems as analysis of time series data.

SAS, on the other hand, can't even be purchased. Instead, you pay an annual license fee (which entitles you, though, to the frequent upgrades.) I use Base SAS and the SAS/Stat module, which together cost $1,935 the first year and $870 each year thereafter. Other modules, such as Graphics and GIS, add to the cost.

And while SAS has a nice point-and-click working environment, you have to write programs to do the actual work. In SPSS, the interface does the programming for you, at least for the easier stuff. (However, don't be intimidated by programming. It's like learning a new language, but it looks a lot like English and has a vocabulary of only a few dozen words.)

And the winner is ...

So why not just get cheaper-and-easier SPSS and forget about SAS? For one thing, SAS offers a somewhat wider range of the data import and manipulation functions you might need someday. That's why SAS is a widely used standard in thousands of big commercial and government data operations, such as the Bureau of the Census. In fact, chances are SAS is used in your own business-side computer shop; that's where I discovered it.

An even better reason for choosing SAS is SAS-L, a very active e-mail list of SAS users around the world. Countless times, SAS-L experts have given me fast solutions — including fully written programs — to data problems I've tossed to the list.

For my money (well, actually the *Herald's* money), the considerable expense and steeper learning curve of SAS is more than outweighed by its vast functionality and large corps of glad-to-help-you users.

In sum, using SAS or SPSS for routine CAR problems may seem a little like using a construction crane to plant flowerpots. But when the day comes that you need to do some really heavy data-lifting, you'll be glad you have either one.

Steve Doig can be reached at (305) 376-3476, or send e-mail to 0005038929@mcimail.com

Here are several books suggested by Philip Meyer: "SPSS Guide to Data Analysis" by Marija J. Norusis, whom Meyer says is currently the best writer on statistics; "Data Analysis for Politics and Policy" by Edward Tufte; "Primer of Statistics for Political Scientists" by V.O. Key Jr.

For a lively discussion on the use of SAS software, subscribe to SAS-L by sending e-mail to listserv@uga.cc.uga.edu Your message should read: subscribe SAS-L yourname.

To tune into an SPSS discussion, subscribe to SPSSX-L by sending e-mail to listserv@uga.cc.uga.edu Your message should read: subscribe SPSSX-L yourname.

A look at widgetmaking

Economic analysis on the fly

By Neill A. Borowski
The Philadelphia Inquirer

When it comes to local economies, perceptions may not be realities.

Community leaders may THINK their region is a leader in computer manufacturing, or financial services or other industries. But we don't really know this simply by looking at employment totals. Enter the "location quotient" û a well-worn tool in regional economic analysis. We recently used this statistic to see just how important computer industries are to the Philadelphia area.

Location quotients are fairly simple to compute. They show the concentration of employment in a region compared to the United States.

The quotients, which are measures of concentration, also can be used with total payroll dollars (from County Business Patterns) rather than employment. Or they can be used to calculate the concentration of one population (race, etc.) compared with the nation (or other areas).

Let's look at widgetmaking. The quotient calculation first requires calculating what percentage widget employment is in this metro and the percentage widgetmaking is nationally.

If there are 10,000 widget workers and 100,000 workers in a metro, the percent is 10 percent. If there are 5 million widget workers nationally and the total workforce is 100 million, 5 percent of the national workforce is in widgetmaking. Just by looking at the data (10 percent vs. 5 percent) you can see widget employment is twice as concentrated in the metro than in the nation as a whole.

You convey this with another quick calculation: (10 percent/5 percent)*100 = 200. This is the location quotient, or index, of the metro, with the U.S. average = 100.

With a spreadsheet or database manager you can quickly compute the quotients for all sectors in your local economy. If your quotients are around 100 (cut some slack for data reporting error), the concentration is equal to the national concentration. If your quotient is, say, 50, then your concentration is half that of the nation.

We ran a bar chart with the Philadelphia metro vs. San Jose in several computing and other industries. A broken line through the charts marked U.S.=100.

Why use index numbers when you could run a table showing 10 percent and 5 percent? For widgetmaking, you could. But what about ceramic-plated widgetmaking? The numbers are 0.0043 percent vs. 0.0021 percent. Still twice as concentrated, but a lot more difficult to understand in percentages (the index would be 205). In general, the percentages will be relatively small. By the way, round the index numbers; don't use decimals, which serve to confuse.

Of course, there are assumptions and problems with this. I used only non-agricultural businesses. Government employment wasn't a factor. If the metro is small enough, it could have a high concentration (index of 200 or more) but actually have only a few thousand employees in that industry. Or, the employment might be all at one company. (However, the County Business Patterns will suppress employment and payroll data if there are only one or two employers in a given industry.) Some purists might also say you have to subtract your metro from the nation before making the national computations.

You don't have to use County Business Patterns. Your regional Bureau of Labor Statistics office can supply establishment employment (series 790) statistics. Or you could use data from the Bureau of Economic Analysis (U.S. Commerce Department). I chose the Patterns because they are seen as a pretty accurate measure. Another measure is the series of economic censuses from the U.S. Census Bureau (they're done every five years in years ending in "2" and "7"; the data for 1992 came out last fall).

Some possibilities:

- If agriculture is big in your area, use the Census Bureau's Census of Agriculture to compute location quotients (dairy revenues in your metro vs. other leading dairy metros).
- If the federal government is talking of more cuts, determine how concentrated federal employment is in your metro. Is it much more concentrated than the nation as a whole?
- Is upscale retail hot in your area? Use the Census of Retail Trade to compute the relevant location quotients.
- For those fortunate enough to have motor vehicle databases, is there a concentration of luxury cars in your county vs. other counties?

Neill A. Borowski can be reached at (609) 779-3884, or send e-mail to borowski@voicenet.com

Practice your skills on the sport that loves statistics: baseball. You can look at baseball player's salaries and performance figures. Or, if you prefer, see an example of a simple regression on the taste of cheddar cheese, and look at how to use a scatterplot to probe the history of Olympic winners. That's just one of the offerings from Carnegie Mellon University's Statistics Department. A great resource for reporters just starting is the Data and Story Library, a collection of real-life ways that experts crunched data. There are serious subjects, too — from breast cancer to crime to air pollution — and examples about how to investigate them with stats. Start at http://lib.stat.cmu.edu/

CAR trips: First adventures

Exposing 'Little White Lies'

By Aaron Elstein
Illinois Times

Aaron Elstein attended NICAR Bootcamp in January. In "Little White Lies," one of his first CAR stories, Elstein built his own database of state "work-force reports" and looked at whether Illinois lives up to its affirmative action standards.

Every state agency in Illinois is supposed to file a "work-force report" with the secretary of state's office by Jan. 1. The reports detail the past year's hirings and promotions and categorizes employees by race, gender and salary level. I had written about the issue a year earlier, but this year, I had the benefit of FoxPro.

There are almost 100 state agencies, commissions and boards, and 65 of them had filed their reports. The records were not available electronically, so I typed in the information I wanted into about 20 different fields. I wanted to look at percentages of minorities in every agency, and at how salaries for minorities compared from agency to agency and track hiring and promotion trends. It took about three days to enter this using my own Mac laptop for the project.

Affirmative action is a hot issue, and state employment is always a lively topic around Springfield, where one-third of the city's population works for the government. There's a law on the books saying each state agency's work force should reflect the state's population as a whole (which is 25 percent nonwhite), but the groups who represent minority employees believe the state has been lax in enforcing this.

My data analysis revealed that, in some ways, the state has done pretty well hiring minorities and women: 23 percent of the work force is minority, and Illinois Gov. Jim Edgar has appointed some women and minorities as cabinet directors.

Despite that, the number of minorities working for the government has actually declined since his administration started five years ago. And right away, the analysis showed just how few minorities can get a job outside the urban social

Continued on page nine

Riding the A-train

By Jim Morris
CNN

Jim Morris attended the NICAR Bootcamp in January.

Like a bolt of lightning, it struck. It was my first day at NICAR Bootcamp, and I had just had a revelation of Biblical proportions: I really could use this stuff.

I was working on a railroad safety story, and I was anxious to use my new skills on my return to Atlanta. I got on the phone with the systems operator at the Federal Railroad Administration. I knew the FRA maintained several databases on train safety, and I wanted to know how I could access the information.

I was expecting a bureaucratic nightmare, but what I found was computer nirvana. The FRA maintains a page on the Department of Transportation's website. The page includes ftp versions of all of its safety databases for 1990-95.

First, I downloaded the file layout and code sheets. Then I downloaded the databases. When I opened the first file, voila, nothing. Something was wrong with the layout.

I began experimenting with the layout values and quickly discovered that things didn't add up. My record lengths were all wrong. The layout sheet listed fields that just didn't exist! After much fussing, I finally got it all worked out. Until I checked the codesheet.

Ahh, the codesheet. Not even close to reality. I got back on the horn with the FRA and had them fax me a copy of the incident/accident report sheet from which the data is entered. Having the report form saved me from near insanity.

By now, my eyes were burning, my fingers cramped. But I finally was ready to ask some questions. I wanted to know where the most deadly railroad crossing in the United States was. So I asked the computer to count the number of accidents at each highway crossing. That was a mistake. As I've since realized (I feel stupid admitting I didn't catch this right away), highways often cross railroad tracks more than once.

Continued on page nine

One news group where mostly mathematicians bicker, cajole and help each other on running statistical tests is sci.stat.math

There are a variety of listservs offering information about statistics. To drop in on the pros, you may want to join allstat, a UK-based discussion on general statistics. Send a message reading "join allstat your name" to mailbase@mailbase.ac.uk

For a discussion on demography and population studies, send a message reading "subscribe demographic-list your name" to majordomo@ coombs.anu.edu.au

We've come a long way, Baby

Two years at Pentium speed

By Brant Houston
NICAR Managing Director

We're taking a moment here at the National Institute for Computer-Assisted Reporting to hit the break key and review what we've accomplished in the past two years with the help of our staff and volunteers, the funding from foundations and the support from the farsighted and adventurous throughout the news profession.

Begun in 1989 by the Missouri School of Journalism, the institute was revitalized in 1994 after the school and Investigative Reporters & Editors joined forces to promote computer-assisted reporting (CAR) through expanded training and services.

When I arrived in February 1994, there were two graduate assistants at NICAR. Since then, our full-time and part-time staff has increased to a dozen. We have a training director, several graduate assistants, a World Wide Web site administrator and web designer, an assistant systems administrator, and an office coordinator. We also share support staff with IRE, of which NICAR is an administrative division, including a bookkeeper, office manager and seminar coordinator.

With this staff, we have conducted more than 125 seminars throughout the United States for more than 5,000 print and broadcast journalists during a two-year period. With funding from the Freedom Forum and Knight Foundation, we have initiated a minority training program.

Our database library has increased from a handful of offerings to more than 20 different databases that we have distributed to more than 90 news organizations. And we have done data work for organizations ranging from small newspapers to national TV magazines.

On the paper publishing side, this newsletter's circulation has more than quadrupled from 80 subscriptions to more than 400. We also have published a book on CAR stories, "100 Computer-Assisted Stories," that shows how such stories are done, and IRE and NICAR supported my work on a basic primer on CAR, "Computer-Assisted Reporting: A Practical Guide," which is published by St. Martin's Press.

While all this activity was going on, we have developed online offerings for the profession with the financial help of IRE supporter Greg Hillman and the Freedom Forum.

You can go to our web site at http://www.reporter.org and find your way not only to NICAR's and IRE's web pages (where you can obtain sample data and do research on investigative reporting), but also to web pages for minority journalists' organizations. You can also join

Continued on page nine

Miss something? Check out the IRE-L and NICAR-L mailing list archives on our websites at http://www.ire.org and http://www.nicar.org. You can see posts to both lists organized by thread, author and date. The list archives are available in html or in plain text format, which is useful for downloading a particular month.

New books for online journalism

Just in time for NICAR's newest training push — computer-assisted reporting for broadcasters — come two books for using online resources for news stories.

The first book, "Wired Journalist" by longtime broadcast and print reporter Mike Wendland, is aimed at TV and radio journalists. Wendland is a full-time reporter at WDIV-TV in Detroit where he practices CAR on a daily basis. He also teaches at NICAR seminars and knows how to make computer technology easy to understand.

His book, published by the Radio and Television News Directors Foundation, is practical and to the point. He avoids "geek speak," using a relaxed and friendly tone to explain the Internet and how to use it for news stories. It's must read for broadcast journalists just starting to Net surf, and it's a handy reference book for those who have already begun.

Also out is Randy Reddick and Elliot King's revision of "The Online Journalist," which is now called "The Online Student." Reddick and King, two university professors, had the misfortune to publish "The Online Journalist" right before the World Wide Web exploded. Now they are back with a valuable excursion into online resources. More detailed than Wendland's book, this would be a good follow-up companion. For professional journalists, the chapters targeting student concerns might seem unnecessary, but this still is a good introductory text.

Meanwhile, don't forget Nora Paul's excellent web book, which you can find at poynter.org, or order the hard copy from the Poynter Institute in St. Petersburg, Fla.

— By Brant Houston

"The Online Student: Making the Grade on the Internet," Randy Reddick and Elliot King, Harcourt Brace College Publishers, order at (800) 782-4479 or (800) 433-0001 in Florida.

"Wired Journalist: Newsroom Guide to the Internet," Mike Wendland, Radio and Television News Directors Foundation, order at (202) 659-6510.

Tech tips: It may look like a table — but it's not

Parsing is the answer

By Natalya Shulyakovskaya
NICAR Staff

Parsing comes into play when you need to turn a text file into a table with rows and columns. There are plenty of those text files scattered all over the Web with numbers and other information organized into columns, or, once in a while, you get a disk from a bureaucrat with data in report format: The files look like tables because the information is organized into separate columns, but they are flat text files. They may have a .txt extension, the extension .dat or some other creative three-letter extension — or no extension at all.

You can open these files with a text editor, but no database manager or a spread sheet will recognize them as tables. At this moment, you know it's time to parse.

One word about downloading files using Netscape: If you are saving a table-looking file from the Web, make sure to save it as a text file and not as an html file. You don't want to waste time fishing out html coding from the document; let the browser do it for you. If you are not using Netscape, you may have to strip out the html coding using an html browswer or text editor before proceding.

Straight to the business

Lets download and parse one of the tables from the Census Bureau Web site (http://www.census.gov/)

The easiest tool to use for parsing is Excel, but remember: Excel can only handle up to 16,000 records in one table.

Table 1. Projections of the Population of Voting Age, for States, by Sex, Race, and Selected Ages: November 1996: http://www.census.gov/ftp/pub/population/socdemo/voting/proj/votepg1.asc

Go to the File pull-down menu, pick Save As and for File Type, specify All files. Save the file, and go to Excel.

In Excel, under the File pull-down menu, choose Open.

Find your file, making sure to specify the correct file type — it should be text only(??). Excel will bring you to the Text Import Wizard.

On Step 1, you have an option of specifying Original Data Type. Here are your choices:

Delimited: All elements in the text are separated by special characters, such as quote marks, ampersands, tabs, spaces or commas.

Fixed Width: Fields are aligned in columns, with spaces between each column of text or numbers.

Choose Fixed Width (see Variation below).

If your file has a lot of text on the top that you don't need in your table, note a small Start Import at Row box. You can browse your file in a window below your text with the numbers of each line. For our Census Projection file, choose Row 8 to start importing. If your file has some junk at the bottom, we will kill it after the parsing by deleting rows. The Wizard adjusts the view in the window. Click on Next.

In Step 2, you will slice your file into the columns: Use the mouse and click once to insert a new line; double click to delete a line. In some cases, Excel will put black vertical lines subdividing the columns for you. Often, though, it doesn't guess right. Click and drag the lines within the window. If you see any repeating junk characters, such as vertical bars, you can separate them from the other data with the lines, or you can clean them out later using Excel's search and replace function under the Edit pull-down menu.

In Step 3, specify in what format to import the data. Excel usually sets all the data types to General and guesses the data types; however, it doesn't always guess right. We will import all the columns as text.

Now, save your table using Save As under the File pull-down menu. Under Save As Type, pick Microsoft Excel Workbook.

Variation

This file also is a good example of how you can use the Delimited option. Before, when you opened the file, you specified Fixed Width. To see how the Delimited option works, repeat the process with the following exceptions: During Step 1, examine your file — each column is separated by a series of vertical bars. Let's use them as delimiters. Check the Delimited button.

In Step 2 in the Delimiters part of the dialog window, check the Other box and type | (vertical bar) in the provided window.

In Step 3, Excel puts the separating lines in

Continued on page nine

To register online for IRE's 1996 National Conference in Providence, R.I., from June 13-16, go to the web page at https://www.ire.org/register.html

Note that you need a web browser, such as Netscape, that supports a secure connection; Lynx will not work with this location.

Your form data and credit card information will be sent through a restricted connection and encrypted by Netscape Commerce Server software. (In order to insure optimal protection for your credit card information, non-electronic registration is also available.)

Continued from page eight:

Tech tips: Parsing

the place of those pesky vertical bars. Let's let Excel do the data format. Never trust it completely, though, especially with dates.

Now, let's give the columns their names — just type them into the cells above the appropriate columns of information. If you plan to use your file later in FoxPro, limit the names to eight characters and don't use spaces.

Clean-up

Let's get rid of periods in our first column: Highlight the column by clicking on the gray bar above the column, and use the Replace option from the Edit pull-down menu. Put a period into the Find box and leave the Replace With box empty.Click Replace and see the periods dissapear into the ether. Let's save this file as Excel Workbook, too.

What we just did with Excel is possible with Visual Foxpro 3.0. In FoxPro 2.6, in order to import a text file, you must build a corresponding table structure and then append a text file to the table as an SDF (standard data-file format), herding the data into correct chutes.

Visual Fox lets you do the parsing without going to Excel for help. Use Import under the File pull-down menu. Choose Import Wizard from the pop-up menu and follow the steps. Visual FoxPro will ask you to name the columns and set up the data format within the Wizard.

Natalya Shulyakovskaya can be reached at (573) 882-0684, or send e-mail to natalya@nicar.org

For information on the June13-16 IRE/ NICAR National Conference in Providence, R.I., check out the Providence Journal's National Conference Homepage at: http:// www.projo.com/ire. You'll find dates, prices, hotel numbers, information on the shape of the program so far and fun things to do in New England.

Continued from page seven:

Reflections on NICAR

our listservs to discuss CAR and investigative reporting.

In the next few months, we hope to unveil new web pages that will provide even more help to journalists.

Most important, we have stuck to our goal of working with other groups to form alliances to do better work. In the past two year, we have offered seminars with Northwestern University, New York University, the University of North Carolina, the University of Maryland and Indiana University. We worked, too, with many other societies including the Society for Professional Journalists, the Society of Environmental Journalists, and the Unity Conference.

We hope to raise our work to new and higher levels in the coming months with your help and the continued financial support of the profession and its foundations.

If you have suggestions for the future or observations about the present, please send them to brant@nicar.org

And thanks again from NICAR for a terrific two years.

CAR trips: First adventures

Continued from page six:

Trains

Then I noticed the column marked "GXID," or Grade Crossing ID. That was the magic field. Within seconds, I had the correct answer.

If I've learned anything from this, it is DON'T GIVE UP. Whenever I got stumped, I'd take a break, get some air. If I was still stumped, I'd call NICAR or the FRA and ask for help.

I even believe this experience changed my life. As my wife will attest, I'm no longer afraid to ask for directions or admit I'm lost while driving my car. That's progress.

Jim Morris can be reached at (404) 827-4023, or send e-mail to jim.morris@turner.com

Continued from page six:

'White lies'

service agencies that have traditionally provided them with patronage work in welfare offices and mental hospitals.

Outside a few well-paid directors, few minorities could be found in management-level jobs in any state agency. Blacks actually represented a greater percentage of state employees than they did in the overall state population. But Hispanics, who now comprise at least 8 percent of Illinois' population, have yet to get state government jobs.

Aaron Elstein can be reached at (217) 753-2226, or send e-mail to 75574.247@CompuServe.COM

On the Internet

That's stats

By Andy Lehren
NICAR staff

There's a good chance you can learn a few things about statistics on the Internet.

It's just a matter of where you weigh in. A significant amount is at universities. Some of this has a tendency to drive beginners into fits of apoplectic boredom. Don't get mean. Odds are, if your patient, you'll get some good ideas for how statistics can help you're reporting.

Learning about statistics

A good starting point is Chance, designed to help teach statistics. For those who think number crunching has no place in the news, check out the archive of stories that use statistics. There's Stephen Jay Gould's "The Median Isn't The Message," plus a lot more. Visit http://www.geom.umn.edu/docs/snell/chance/welcome.html

For reality checks, visit How to Avoid Lies and Damned Lies at http://maddog.fammed.wisc.edu/pitfalls/

There are a lot of college statistic classes with some course material on the Internet. Some are actually worthwhile. One of the best is at Arizona State University, with everything from a simple program you can download for free to help learn about probability, to a clear course outline that you can actually learn from. Don't miss its concise links collection, too. It's at http://olam.ed.asu.edu/~glass/502/home.html

Another good course is from Concordia University at http://artsci-ccwin.concordia.ca/inte/inte298s/courses/index.htm

From the land down under, at the University of Newcastle, is http://frey.newcastle.edu.au/Stats/surfstat/noframes/surfstat.html

American Demographics does a lively job analyzing numbers. Visit http://www.marketingtools.com/ad_current/default.htm

A thorough dictionary of statistics is Prophet Statguide at http://www-prophet.bbn.com/statguide/sg_glos.html Scroll through the interactive dictionary, or do keyword searches. A short list written for journalism students is at http://excellent.com.utk.edu/~foley/stat.html

Download a program that will help you learn exactly how regressions work, and how to run them. Visit Click&Learn Regression at http://nsns.com/click/ From there, you can get a program that will walk you through regression — and even quiz you along the way. The site offers a free demo. The full program is $49.

For an online book about probability, check out the University of Maryland's Resampling Project at http://www.statistics.com/text.html It offers clear writing about some basics in sampling and statistics.

Tooling around

When it comes to getting the software to run statistical tests, the Internet harbors everything from freebie programs to how to contact the big software makers.

For SPSS and SYSTAT, visit http://www.spss.com

For SAS, go to http://www.sas.com/

For Statistica, go to http://www.statsoftinc.com/

Download and sample Resampling Stats software at http://www.statistics.com/software.html Right nearby, at http://www.statistics.com/resample/vendors.html, is a handy price guide for statistical software.

Get a review copy of KwikStat at http://www.dfw.net/homepage/

For Xlisp-Stat, download free copies starting at http://stat.umn.edu/~rcode/xlispstat-resources.html#win31 For StatView, try http://www.abacus.com/

The Journal of Statistics Education Information Service offers a variety of services, including software and the ability to check the archives of SAS and SPSS listservs. It's at http://www2.ncsu.edu/ncsu/pams/stat/info/infopage.html

The University of Minnesota's School of Statistics offers everything from free software to guides on running statistical tests at http://stat.umn.edu:80/ARCHIVES/archives.html

For those with time, check out extensive lists of universities and other organizations around the world that offer information on statistics at http://www.stat.cmu.edu/otherplaces/ Or try http://www.geom.umn.edu/docs/snell/chance/sources.html

Andy Lehren can be reached at (573) 882-0684, or send e-mail to andy@nicar.org

Want to use U.S. Bureau of Labor Statistics to calculate inflation rates? Check out this straightforward primer at http://www.science.gmu.edu/csi779/drope/csi804/theory.html

The BLS homepage on the consumer price index is at http://stats.bls.gov/cpihome.htm

Oregon State's Prof. Robert Sahr offers a variety of ways to look at inflation — from blockbuster movies to Congressional pay. There is even a handy Excel table you can download. http://www.orst.edu/Dept/pol_sci/sahr/sahr.html

You can point and click your way to compare housing prices at http://fly.HiWAAY.net/~rhuff/Postma/Calc/

Or, if you're moving, point and click to compare what your salary buys elsewhere in the U.S. at http://www.homefair.com/homefair/cmr/salcalc.html

Growing collection of federal databases

From the NICAR library

NICAR offers a number of federal government databases. Here is a list of our growing collection:

NEW • A monthly CD subscription for all 1995-96 Federal Election Commission campaign contributions by individuals and political action committees, plus all presidential matching fund requests.

NEW • The Health Care Financing Administion's 1995 database of all Medicare-funded inpatient work in U.S. hospitals.

• Federal Railroad Administration data for accidents, casualties, and highway crossings. 1991-1995.

• Coast Guard boating accidents, 1969-1994.

• Federal Aviation Administration data, including airplane maintenance work documented in the service difficulty report, pilot licenses and grades, and aircraft registration.

• Home Mortgage Disclosure Act records, for tracking who gets loans and who gets turned down, and finding redlining patterns.

• Federal procurement data, 1992-1994, includes breakdowns by agency.

• Alcohol, Tobacco and Firearms gun dealer records.

• National Bridge Inventory System data, includes inspection grades.

• FBI Uniform Crime Reports, a detailed compilation of crime data that includes statistical breakdowns of individual murders. This includes the new 1994 data.

• Social Security death records, by name and social security number, going back to 1937.

• Occupational Safety and Health Administration violation data includes worker accidents and exposures to hazardous chemicals by companies.

• U.S. Department of Transportation truck accident and census data. It includes accidents by company and road.

• U.S. Small Business Administration loan guarantees, 1989-1995. This includes the name of the business, address, amount covered by the SBA, and status, including whether the loan went bad.

• U.S. Small Business Administration disaster loan guarantees, 1989-1994. This includes individuals and businesses, the amount covered by the SBA, and the status, including whether the loan went bad.

• U.S. Small Business Administration's list of minority companies certified for SBA assistance in seeking federal contracts. It includes the name of the company, its address, the owner, type of business and phone number.

• U.S. Department of Transportation hazardous materials accidents database, a collection of roadway, rail, air and waterway accidents from 1971 to 1995.

• U.S. Department of Transportation fatal accident reporting system. It includes all roadway accidents from 1988 to 1994.

• U.S. Coast Guard directory of U.S. merchant vessels. It includes the name of the ship, the managing owner, home port and various descriptive information.

• National Endowment for the Arts, grants, 1989-1993.

For up-to-date prices
and more information, call (573) 882-0684,
or send e-mail to

Grants available for investigative journalism

IRE is donating $2,000 of proceeds from the IRE Awards Contest Entry fee to the Fund for Investigative Journalism.

IRE members Robert Baskin of CBS, author David Burnham, Brooks Jackson of CNN, Margaret Engle of the Alicia Patterson Foundation and Gene Roberts of the New York Times, sit on the board of the fund. For 25 years, IRE members have gone to the fund for support to do stories that otherwise would not be done.

For information about applying for a grant, call (202) 462-1844, or write to the fund at: 1755 Massachusetts Ave., NW, Room 324, Washington. D.C. 20036

The following two-day seminars, including one day of panel discussions and demonstrations and one day of hands-on training in basic spreadsheet and database skills, are open to all journalists. Cost ranges from $50-$75 for Day 1 and from $100-$135 for both days. Here is a list of dates, locations and phone numbers to call for more information:

**East Lansing, Mich.,
May 14-15,
(313) 259-0650.**

**Concord, N.H.,
May 14-15,
(603) 224-3327.**

**Aberdeen, S.D.,
May 17-18,
(605) 332-2111**

**Albany, N.Y.,
May 20-21,
(518) 458-7821**

**Rochester, N.Y., May 22-23,
(518) 458-7821**

**Vancouver, Wash.,
May 31-June 1,
(206) 682-1812**

**State College, Penn.,
June 7-8,
(215) 561-1133**

Bits, Bytes and Barks

Hate crimes

The Chicago Reporter, an investigative monthly, built its own database of hate crimes from police records obtained from the Chicago Police Department.

The paper chose 12 fields including type of crime, location, victim's age and race, offender's age and race.

The analysis found that hate crimes against whites had dropped, while those against blacks had risen, and there was a slight decrease in hate crimes overall.

Some information was missing, but even that turned into a story. The paper discovered the state of Illinois was violating state law by not collecting data on hate crimes.

— John Sullivan

JOBS: Systems editor in Des Moines

The Des Moines Register seeks a systems editor, part of the newsroom's management team, to supervise daily newsroom technical operations and planning.

The Register uses an SII/Coyote editing system, Quark pagination, networked Mac and IBM/Windows computers, a Digicol archive system, Photoshop and AP/Leaf imaging systems and Cumulus archiving.

The systems editor leads newsroom technology training, help-desk operations, installations and budgeting. Applicants must have a bachelor's degree in journalism, computer science or related studies plus three to five years experience at a newsroom or periodical publisher. Work with Apple's OS, Windows, traditional front-end systems and modem communications is a must. Experience with SII Styl, Unix, AppleScript or Internet applications is an asset.

The *Register* is a Gannett paper with 179,000 daily circulation, 302,000 Sundays. Send a resume and cover letter to: David Rhein, Deputy Managing Editor, *The Des Moines Register*, 715 Locust St., Des Moines, IA 50313.

JOBS: The Corpus Christi Caller-Times

The Corpus Christi Caller-Times has an opening for a computer-assisted reporter/coordinator. Applicants should have at least two years experience; knowledge of database, spreadsheet, Internet, modem communication and word processing software; basic hardware skills, including replacement of motherboards and memory chips; and experience negotiating for data and cleaning it up for CAR use. Expect to produce investigative and enterprise projects, help other reporters with projects and help other Harte-Hanks newspapers and the Harte-Hanks Austin Bureau with CAR projects.

The Caller-Times is a 65,000-circulation daily. An Internet version of the newspaper is at http://www.caller.com.

To apply, send a cover letter, resume and clips to: Scott Rothschild, Metro Editor, *Corpus Christi Caller-Times*, P.O. Box 9136, Corpus Christi, TX 78469. Send questions via e-mail to Rothschild at scottr@caller.com.

Keep up with NICAR online

Subscribe to our listserv and join in as reporters talk about how to do the job better. E-mail to LISTSERV@MIZZOU1.MISSOURI.EDU. In the message, on the first line, write: subscribe NICAR-L your name. To join IRE on the Internet, the instructions are the same except, on first line, write: subscribe ire-l your name.

IRE/NICAR is also accessible through CompuServe's Journalism Forum. Go to the JForum, Section 19. Also look into the IRE/NICAR files in Library 19.

Investigative Reporters and Editors, Inc.
100 Neff Hall
University of Missouri
School of Journalism
Columbia, MO 65211