12-1996

A preview of EI and EzI: Programs for ecological inference

Kenneth BENOIT
Singapore Management University, kbenoit@smu.edu.sg

Gary KING

Citation

BENOIT, Kenneth, & KING, Gary.(1996). A preview of EI and EzI: Programs for ecological inference. *Social Science Computer Review, 14(4)*, 433-438.
Available at: https://ink.library.smu.edu.sg/soss_research/4010

SOFTWARE

askSam. askSam Systems, P.O. Box 1428, Perry, FL 32347; telephone: 800-800-1997; 904-584-6590 (support); fax: 904-584-7481.

HyperREASEARCH. ResearchWare, Inc., P.O. Box 1258, Randolph, MA 02368-1258; telephone: 617-961-3909; e-mail: paul@bcvms.bc.edu.

NUD.IST. Qualitative Solutions and Research Party Ltd. [Developer], Box 171, La Trobe University Post Office, Melbourne, Victoria 3083, Australia; telephone: 61-3-459-1699; fax: 61-3-479-1441; e-mail: NUDIST@qsr.com.au. U.S. and Canada distributor: Scolari, Sage Publications Software, 2455 Teller Road, Thousand Oaks, CA 91310. Phone: 805-499-1325; fax: 805-499-0871; e-mail: NUDIST@SAGEPUB.COM.

Storyspace–s. Eastgate Systems, Watertown, MA.

The Ethnograph. Qualis Research Associates, P.O. Box 2070, Amherst, MA 01004; telephone: 413-256-8835; fax: 413-256-8472; e-mail: qualis@mcimail.com.

WordPerfect. Corel Corporation, 1600 Carling Ave., Ottawa, Ontario, Canada K1Z8R7; telephone: 800-772-6735; fax: 613-728-9176; Internet site: http://www.corel.com/.

REFERENCES

Fielding, N. G., & Lee, R. M. (Eds.). (1991). *Using computers in qualitative research.* Newbury Park, CA: Sage.

Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory.* Chicago: Aldine.

Hesse-Biber, S. (1995). Unleashing Frankenstein's monster? The use of computers in qualitative research. In R. G. Burgess (Ed.), *Studies in qualitative Methodology* (Vol. 5, pp. 25-42). London: JAI Press.

Hesse-Biber, S., DuPuis, P., & Kinder, T. S. (1991). *HyperRESEARCH:* A computer program for the analysis of qualitative data with an emphasis on hypothesis testing and multimedia analysis. *Qualitative Sociology, 14,* 289-306.

Lee, R. M. (Ed.). (1995). *Information technology for the social scientist.* London: UCL Press.

Lee, R. M., & Fielding, N. G. (1995). Users' experience of qualitative data analysis software. In U. Kelle (Ed.), *Computer-aided qualitative data analysis: Theory, methods, practice* (pp. 29-40). London: Sage.

Mangabeira, W. (1995). Computer assistance, qualitative analysis and model building. In R. M. Lee (Ed.), *Information technology for the social scientist* (pp. 129-146). London: UCL Press.

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook* (2nd ed.). Thousand Oaks, CA: Sage.

Miles, M. B., & Weitzman, E. (1991). *Computer-aided qualitative data analysis: A review of selected software.* New York: Center for Policy Research.

Richards, L. (1995). Transition work! Reflections on a three-year *Nud.ist* project. In R. G. Burgess (Ed.), *Studies in qualitative methodology: Computing and qualitative research* (Vol. 5, pp. 105-140). London: JAI Press.

Richards, T. J., & Richards, L. (1994). Using computers in qualitative research. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (pp. 445-462). Thousand Oaks, CA: Sage.

Strauss, A., & Corbin, J. (1990). *Basics of qualitative research: Grounded theory procedures and techniques.* Newbury Park, CA: Sage.

Beverly A. Smith is a doctoral candidate in the Department of Sociology at Boston College. She received her MBA degree in 1993 from Boston College.

Sharlene Hesse-Biber, Ph.D., is associate professor, graduate director, and assistant chair, Department of Sociology, Boston College, Chestnut Hill, Massachusetts. She is a codeveloper of HyperREASEARCH, *a computer-assisted software program for qualitative data analysis. She has coauthored several articles which describe the features of this program: "HyperResearch: A computer program analysis of qualitative data using the Macintosh: Hypermedia features" (Hesse-Biber, DuPuis, & Kinder, 1991) in* Qualitative Sociology, 14; *"Hypothesis testing" (Hesse-Biber & DuPuis, 1995) in* Computer-aided Qualitative Data Analysis: Theory, Methods and Practice, *edited by Udo Kelle and published by Sage (London).* HyperRESEARCH *is available for the Macintosh and IBM/Windows environment. Hesse-Biber is currently conducting research on the impact of computer-assisted instruction in the classroom.*

REPORTS AND COMMUNICATION

A Preview of *EI* and *EzI*

Programs for Ecological Inference

KENNETH BENOIT
GARY KING
Harvard University

Ecological inference, as traditionally defined, is the process of using aggregate (i.e., "ecological") data to infer discrete individual-level relationships of interest when individual-level data are not available. Existing methods of ecological inference generate very inaccurate conclusions about the empirical world—which thus gives rise to the ecological inference *problem*. Most scholars who analyze aggregate data routinely encounter some form of this problem.

EI (by Gary King) and *EzI* (by Kenneth Benoit and Gary King) are freely available software that implement the statistical and graphical methods detailed in Gary King's forthcoming book *A Solution to the Ecological Inference Problem* (King, in press). These methods make it possible to infer the attributes of individual behavior from aggregate data. *EI* works within the statistics program *Gauss* and will run on any computer hardware and operating system that runs *Gauss* (the *Gauss* module, *CML*, or constrained maximum likelihood—by Ronald J. Schoenberg—is also required). *EzI* is a menu-oriented stand-alone version of the program that runs under MS-DOS (and soon *Windows 95*, *OS/2*, and *HP-UNIX*). *EI* allows users to make ecological inferences as part of the powerful and open *Gauss* statistical environment. In contrast, *EzI* requires no additional software, and provides an attractive menu-based user interface for non-*Gauss* users, although it lacks the flexibility afforded by the *Gauss* version. Both programs presume that the user has read or is familiar with *A Solution to the Ecological Inference Problem*.

THE PROBLEM OF ECOLOGICAL INFERENCE

The best prior methods of ecological usually lead to very inaccurate conclusions about the empirical world; indeed, frequently, they give impossible answers, such as 20% of Israeli Labor Party voters remaining loyal between the last two elections. Ecological inferences are required in political science research when individual-level surveys are unavailable (e.g., local or comparative electoral politics), unreliable (racial politics), insufficient (political geography), or infeasible (political history). They are also required in numerous areas of public policy (e.g., for applying the Voting Rights Act) and other academic disciplines ranging from epidemiology and marketing to sociology and quantitative history. The ecological inference problem has been among the longest standing, most actively pursued,

Social Science Computer Review, Vol. 14 No. 4, Winter 1996 433-438

and consequential in all of quantitative social science. It was originally raised more than 75 years ago as the first statistical problem in the nascent discipline of political science (Ogburn & Goltra, 1919) and has held back research agendas in most empirical subfields of the discipline ever since.

Nearly half a century ago, a single article by William Robinson (1950) diverted the stream of academic research. By popularizing "the ecological fallacy," greatly clarifying the ecological inference problem, and (appropriately) convincing generations of scholars that aggregate data should never be used (with methods then available) to infer information about individuals, he fundamentally changed the nature of social science research. Vibrant fields of scholarship that relied on aggregate data withered. Once important traditions of political geography in France, Germany, and the United States largely collapsed. The scholars who continue to work in some of these fields—such as those attempting to explain who voted for the Nazi Party, or which social groups support each political party in the newly emerging democracies—do so because of the lack of an alternative to ecological data, but they toil under a cloud of great suspicion.

Research based on aggregate data was succeeded at mid-century with the then emerging methodology of survey research. Surveys have not only taught us a great deal; they may well represent the single most important methodological contribution of the social sciences of this era. Because better data beat more sophisticated statistics every time, the ecological inference problem does not arise if accurate survey data are available. Yet an exclusive focus on available surveys has had some unavoidable costs. Surveys necessarily force analysts to study recent periods, and thus to miss long-term trends and some large-scale patterns. Moreover, when we choose our subject to study based on available survey data, we study random collections of isolated individuals, and only rarely gather knowledge of local communities, contextual effects, or geographic patterns. Creative combinations of quantitative and qualitative research are also much more difficult when the identity of, and rich qualitative information about, individual respondents cannot be revealed to readers. Indeed, in most cases, respondents' identities are not even known to the data analyst. If "all politics is local," political science is missing much of politics.

In contrast, aggregate data analysis can almost always be based on extensive, detailed, and local, geographic patterns; it can extend over very long time frames; and it can be supplemented with qualitative information at any level of richness or detail. In fact, with a working method of ecological inference, there is little reason for any division between quantitative and qualitative approaches to a problem. Systematic quantitative analyses can be easily joined with richer qualitative data at the local geographic level and any study will be improved as a result. Going into the field and visiting villages and communities from which data are available becomes again a central task. In fact, if some limited survey data are available, it too can be used with the method proposed to improve ecological inferences.

Moreover, the quantity and quality of unanalyzed aggregate data sets waiting around for enterprising political scientists to take notice is staggering. In the United States alone, there are political data on dozens of electoral offices and 3000+ census variables available from each of 190,000 electoral precincts. Most of these data have been untouched by social scientific hands. The situation is not much different in most parts of the world. When the Berlin Wall came down, scholars of Eastern Europe and the former Soviet Union found that behind it all these years was a flood of data, and those standing nearby were inundated. With these and other massive unanalyzed aggregate data sets pouring in from all over the world, and a solution to the ecological inference problem, much opportunity remains.

The methods implemented in *EI* and *EzI* give accurate estimates as well as reliable assessments of the uncertainty of all inferences. It is also robust to numerous data problems

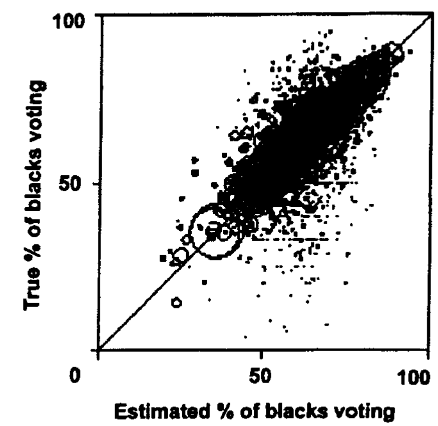

Figure 1: Model Verification: Voter Turnout Among African Americans in Louisiana Precincts
NOTE: This figure represents 3,262 precincts in 1990, with each circle size proportional to the number of voting-age African Americans in the precinct. That the vast majority of circles fall near the diagonal line, indicating that the estimated and true fractions of blacks voting are nearly identical, is strong confirmation of the model.

such as high levels of aggregation bias. Because the ecological inference problem is caused by the lack of individual-level information, all methods of ecological inference, including those introduced in *EI* and *EzI*, will always entail some risk. These risks are minimized in the approach taken by new models that incorporate far more available information, intuitive graphics and diagnostics to evaluate when assumptions need to be modified, easy methods of modifying the assumptions, and realistic uncertainty estimates for all quantities of interest. *Solving the Ecological Inference Problem* verifies the method by using a variety of new data sets for which the true, individual-level answer is known. This makes possible more than 16,000 comparisons between estimates of individual-level relationships from aggregate data and the known individual-level answer. (This compares to 49 such comparisons in the history of research on this subject.) The method works in practice.

AN EXAMPLE

Given these marginal percentages from Louisiana in 1990, what are the cell entries?

	Vote	No Vote	
Black	?	?	26.6%
White	?	?	73.4%
	68.5%	31.5%	

The goal of ecological inference in this example is to fill in the unobserved cell entries given only the observed aggregate percentages from the column and row totals in this table (from Louisiana in 1990) and the corresponding tables from each of Louisiana's 3,262 precincts (i.e., without survey data). For example, the upper-left cell entry is the (normally unobserved) percent of Blacks who voted in 1990. Because the cell entries in this case are known from public records, we know that the true value of this cell is 64%. The estimate from the method reported in the book, based only on aggregate data, is only a fraction of a percentage point under this mark. In this example, like numerous others reported in the book, the method gives accurate statewide estimates, which has been the goal of past research.

But the solution to the ecological inference problem provides much more interesting information than accurate statewide estimates. It also provides, in these data for example, accurate estimates of the fraction of Blacks who vote (and fraction of Whites who vote) for each of Louisiana's 3,262 electoral precincts. For example, Figure 1 compares *estimates* from aggregate data of the percent of Blacks who vote to the true percent of Blacks who vote in all Louisiana precincts (with each precinct represented by a circle sized proportional to its black population). The center of almost all of the circles falls on or near the diagonal line, indicating that the estimated percent of Blacks voting is very close to the true individual-level percentage. This figure is not merely a plot of the observed values of a variable by the fitted values of the same variable used during the estimation procedure; it is a much harder test because the true fractions of Blacks voting (the vertical dimension in the figure) were not used during the estimation procedure.

USING THE SOFTWARE

Installation of the software is straightforward. Installing *EI*, which requires that the user already have *Gauss* installed, involves copying the library and source code files for *EI* into the appropriate subdirectories of the *Gauss* program directory tree. The *EI* library is then accessed using the *Gauss* command `library ei`. The whole process involves only four copy commands. Getting the *EzI* standalone program working is also trivial. Simply extract the *ezi.zip* file to the directory where you want it installed and it runs immediately.

A postscript file containing a detailed reference manual is available with both programs. For users who lack access to a postscript printer, a copy of the manual is also available on-line at Gary King's Web site (`http://GKing.Harvard.Edu`). A short supplement describing the differences between *EI* and *EzI* is also included with the latter. Anyone with access to *Gauss* and little experience in this statistical programming environment will prefer to use *EI*, because the *EI* offers all of the advantages of remaining within the open *Gauss* environment while still being very easy to use. For example, the output quantities from an *EI* model can easily be fed to other models available in *Gauss*, something which the stand-alone version cannot offer. If you prefer not to use *Gauss*, however, you can still use *EI*. *EzI* is easier to use than its direct-*Gauss* counterpart, although it offers less flexibility. It has an attractive color menu system and provides context-based help messages that summarize the material included in the *EI* manual. *EzI* is written in *Gauss* and uses the *EI* library directly, so there are no differences in their statistical procedures. All of the procedures available in *EI* are implemented in *EzI*.

One of the best features of the *EI* library is the ease with which it makes it possible to distribute the information necessary to replicate your empirical results. All data are stored in a single "data buffer"—*Gauss* data construct that permits data of different types to be stored along with identifying labels in a single file. *EI* data buffers contain not only all of the input data, but the values of all of the globals set by the user before estimating the ecological inference model, as well as the quantities resulting from the estimation. The data buffer acts as a large container for all of the inputs and outputs of the model run. Subsequent user requests for resulting quantities or for graphs are simply read quickly from the data buffer, eliminating the need for any significant recalculation once the model has already been run. The *EI* command `eirepl` actually takes a data buffer containing an already-run model, reads the settings of all of the globals from that file, and reestimates the model. This means that someone using *EI* or *EzI* in a study who wants to make his or her data open to replication need simply make the data buffer available. Other users with *EI* will be able to replicate the original analysis, or perform variations of it, directly from the distributed data buffer file. This structure is designed to make it as convenient as possible to follow ***the replication standard*** by making it easy to make publicly available all of the data and information necessary to replicate your published analyses. The output data buffer thus serves as an automatically created "replication data set" (see King, 1995).

It is a relatively simple matter to read in data for use in *EI* and *EzI*. Both programs easily import an ASCII data set that is delimited by tabs, spaces, or commas. *EzI* also lets users include a separate variable definition (**.var*) file that includes variable names and even variable specifications—such as which variables will correspond to which standard elements for the model—that is loaded automatically upon import of the ASCII file. Data may also be loaded from a data buffer (created from a previous model estimation) for use in a new model.

Estimating the ecological inference model requires only a single command. Once data have been loaded from either an ASCII or a data buffer file, the model is run simply by specifying which variables will constitute the five input items, which the *EzI* model requires. The length of time that the model requires to estimate varies in proportion to computer speed, model complexity, and data set size. In our development of the software, this varied in practice from several minutes to nearly an hour for very large data sets. For a typical problem of about 300 observations, the estimation procedure takes about 5 minutes on an HP 715/80 workstation. All subsequently computed graphics and numerical results are then instantaneously available.

EI and *EzI* include a large number of predefined graphs and summary statistics. Built into both programs are more than 50 detailed graphs for visualizing the results, and nearly 100 numerical items that may be read from the data buffer after estimating the model. In many cases, the user will want to perform subsequent statistical estimations on these quantities. In *EI*, this is done directly using standard *Gauss* techniques; in *EzI*, the user can collect quantities from the model into an export data set that is written in a form that can easily be reloaded as an ASCII data set. *EzI* even created the corresponding variable name definition file for this purpose.

We present a short example here showing how easy *EI* is to use. From *Gauss*, only a few lines of commands are required to initialize the library, load in a data set, run the model, and view the results.

```
library ei,cml,pgraph;          @ initialize EI @
loadvars sample.asc t x n;      @ load variables t, x, and n into memory @
dbuf = ei(t,x,n,1,1);           @ run main procedure, save results in dbuf @
eigraph(dbuf,"tomog");          @ draw tomography graph, with data in dbuf @
v = eiread(dbuf,"betab");       @ extract precinct-level estimates from dbuf @
```

In *EI*, these steps involve selecting analogous menu items. The user first selects a menu item to load in an ASCII data set. If a `sample.var` is present specifying the variable names and their assignment as the required elements of t, x, and n, then the next step can be choosing to run the model. Otherwise the user can specify or respecify the model and variable names from a model editing screen. Once the model is finished, *EzI* graphs and quantities are available from their corresponding pull-down menus.

In practice, the user will want to view some descriptive statistics on the data set before running the model, and possibly change some *EI* globals. In *EI*, these involve simple one-line commands, whereas in *EzI* these actions are available as menu choices.

HARDWARE AND SOFTWARE REQUIREMENTS

EI will run on any computer system that already runs *Gauss*. *EzI* is currently available for only MS-DOS, although 32-bit *Windows* and *HP-UNIX* versions are planned. The MS-DOS binaries run well in DOS sessions of *OS/2* and *Windows 95*, however. *EzI* requires 8 megabytes of free memory and about 2 megabytes of disk space.

OBTAINING THE SOFTWARE

Before attempting to use *EI* or *EzI*, we recommend that the user read *A Solution to the Ecological Inference Problem*. Until its publication, a postscript version of the manuscript is available for download from Gary King's home page (`http://GKing.Harvard.Edu`). The software is available from the same source. The *Gauss* statistical programming language and the *CML* module are available for *DOS/Windows*, *Windows 95*, *UNIX*, and other operating systems from Aptech Systems, Inc., 23804 S.E. Kent-Kangley Road, Maple Valley, WA 98038; phone: 206-432-7855; fax: 206-432-7832; e-mail: `sales@aptech.com`.

REFERENCES

King, G. (1995, September). Replication, replication. *PS: Political Science and Politics*, *XXVIII*(3), 443-499.

King, G. (in press). *Reconstructing individual behavior from aggregate data: A solution to the ecological inference problem*. Princeton: Princeton University Press. Preprint available at `http://GKing.Harvard.Edu`.

Ogburn, W. F., & Goltra, I. (1919). How women vote: A study of an election in Portland, Oregon. *Political Science Quarterly*, *3*(XXXIV), 413-433.

Robinson, W. (1950). Ecological correlation and the behavior of individuals. *American Sociological Review*, *15*, 351-357.

Assessing and Coping with Computer Anxiety in the Social Science Classroom

DAVID ALEXANDER BOWERS, JR.
VICKI MARTIN BOWERS

INTRODUCTION

One of the major challenges of using computers in the social science classroom is overcoming student fears about computers. This report will suggest ways to both measure computer anxiety and provide remedies to ameliorate this problem for students. Anxiety stems from the anticipation of danger (in the case of computers, social disgrace) which causes apprehension, tension, or uneasiness in cognition or behavior. In addition, the *Diagnostic and Statistical Manual of Mental Disorders* (American Psychiatric Association, 1987) says, "Anxiety may be focused on an object, situation, or activity" (p. 392). Computer anxiety can involve all of these three elements: fear of the computer itself, using the computer, or doing the computer homework assignments. According to psychologists, computer anxiety is a type of performance anxiety with its two components: discomfort anxiety—where the student fears physical or social discomfort—and ego anxiety—where the student feels that his worth as an individual will be damaged if he makes mistakes or cannot perform the computer assignment (Robin & DiGiuseppe, 1993).

It may be hard to believe, but a significant portion of the American population still suffers considerable anxiety when confronted with using computers. For instance, a recent poll taken by Dell computers "found that 27% of the adults surveyed had never used a computer; 55% had never bought one; 32% of adults 'feel intimidated using computers'; and 25% 'miss the days when we just had typewriters' " (Magid, 1993). Anxiety about using computers can be a major disability in both education and the workplace. Computer anxiety can be so paralyzing even students with some experience cannot perform well. In fact, Marcoulides (1988) has concluded that anxiety is a more important determinant of achievement than computer experience.

We will focus on how to measure computer anxiety— those attributes which tend to be associated with computer anxiety—and make suggestions for how to ameliorate computer anxiety. Finally, we will present some results from a study of two classes that the first author taught using computers.

MEASURING COMPUTER ANXIETY

The most obvious way to measure computer anxiety for the social science professor is to use a pencil-and-paper test. Educational researchers have devised a number of scales to measure computer anxiety and have provided fairly credible information for the reliability and validity of these scales. In fact, many of the questions on differing scales are similar and seem to measure the same attitudes.

Social Science Computer Review, Vol. 14 No. 4, Winter 1996 439-443