EDITOR: Alexander B. Magoun, a.b.magoun@ieee.org

DEPARTMENT: ANECDOTES

The Effects of Increases in Computing Power on Demographic Analysis Over the Last 50 Years

Barbara A. Anderson, *University of Michigan, Ann Arbor, MI, 48103, USA*

This anecdote discusses the relationship of changes in computer power to demographic analysis over the past 50 years, based on my work as a demographer in that time. Increases in processing power, and the growing complexity of software that this increased power enabled, opened new opportunities for researchers, but the greater ease of computing sometimes led them to be sloppy. Increase in computing power and development in statistics also led to tradeoffs between data usefulness and confidentiality.

Computing for demographers in the late 1960s and 1970s relied on mainframe computers. Analysis was run by inputting programs, and often punched cards, at a computer center. Later iterations involved running a program from floppy discs and calling up data from a terminal connected to a mainframe, which evolved to desktop standalone computers, and most recently to laptop computers, which can now leverage shared computing power and storage space through the cloud. More computing power and increased storage allow the more complex analysis of larger datasets. As computing became much less cumbersome and time consuming, more of the data analysis could be accomplished by researchers working independently in their homes or offices.

Mainframe computers imposed substantial constraints, due to the limitations of the technology. In the 1970s, using university-based mainframe computers, it typically took eight hours to receive output during the day. At night it usually took one hour, leading to a great deal of computing in the middle of the night.

Programs were run from 80-column punched cards. If you dropped the cards, you needed a way to put them back in the correct order. Many of us put codes in the first few columns of every punched card that made it possible to put them in order using a counter sorter, which sorted the cards based on the values in specified columns.

When I was completing the statistical analysis for my dissertation at Princeton University in 1973, I went to the computer center at 11:00 P.M. every night to take the advantage of the shorter time to receive output. I went home at 7:00 A.M., when the mainframe ran systems tests. I slept until noon and then devoted half of my brain to constructing tables from my output and the other half to watching the Watergate hearings.

In the early 1970s, there were only rudimentary statistical programming packages. Anything more complicated required either substantial knowledge on the part of the researcher or the employment of a programmer, which also imposed limitations on and obstacles to demographic analysis.

In the mainframe era, computer processing time was charged to researchers in dollar equivalents that, for those who did not have large external grants, was called "funny money." Funny money was allocated to researchers from a departmental pool, supported by the university's general budget. The funny money charges were based on processing time, which increased with the number of cases and the complexity of what was being estimated. There was usually an allocation for each faculty member and graduate student, and there were controls on not exceeding one's funny money allocation [11].

The lag time for receiving output along with the financial costs encouraged care in preparing for a computer run. It was common to create a small test file, such as 50 cases. The costs in time and money for errors also encouraged the examination of distributions on individual variables for the entire dataset in order to check for errant values.

Increases in computing power on mainframe computers led to the development of programs for more complex analyses. While it had been a routine to run

1058-6180 © 2022 IEEE
Digital Object Identifier 10.1109/MAHC.2022.3214737
Date of current version 29 November 2022.

ordinary least squares regression, programs were developed for logistic regression, which was more appropriate for the analysis of dichotomous dependent variables, especially when the proportion of "yes" values in the dependent variable was small. Logistic regression was quite expensive in computer processing time, and thus in funny money, and the cost increased rapidly with the number of independent variables.

Yet in the mainframe era, demographers depended on statisticians and programmers to implement the analyses we wanted to run. In the mid-1970s Jim McCabe and I performed a logistic regression using data from a household survey in Kinshasa, Zaire, to examine factors related to the waiting time between the first and second birth [1]. We studied women aged 20–24 who had had at least one live birth. Starting nine months after the first birth, we examined six-month periods of experience. This approach is now common in event history analysis [3]. But in 1976, when we discussed our analysis with Yale University statisticians, we encountered a great deal of skepticism about whether we really had the right number of degrees of freedom, since different women contributed a different number of units of analysis.

In the late 1980s, I went to Tallinn, Estonia, to collaborate with a sociologist/demographer. At the time Estonia was still in the mainframe era, and we could access Estonian survey data only through a mainframe computer, at Tartu University, which was over two hours away by bus. We had access only through a programmer at the university. To run an analysis we called the programmer and explained what we wanted done. If she did not agree, she would not do the analysis. Sometimes it took quite a while to convince her to do what we wanted. After a day or three, she would run the analysis and send the output to us on the Tartu-to-Tallinn bus. It might take five days or more between requesting an analysis and receiving the output. If she had misunderstood what was to be done, another round of analysis was required.

We compared the results from the Estonian survey data with data from the American National Longitudinal Survey of Youth (NLSY), which I brought with me on a PC. We easily analyzed the NLSY data, using the results of one analysis to decide what to do next. The difference in the ease of the two modes of computing was striking.

Over the next 15 years, the benefits of increased computing power and the attendant developments in hardware and software were notable. In 1990, I returned to Estonia with Brian Silver, bringing the most powerful personal computer that was legal to export to the Soviet Union at the time (restrictions on taking computers there were based mainly on processing speed). It was a one-piece computer, including the monitor, and was the size of a small suitcase. It came in a Hartmann carrying case with a handle. We also took an external disk drive with a large capacity by the standards of the time. We became somewhat expert on luggage that could fit various kinds of computer hardware.

Statistics Estonia had received permission from the Soviet government to download the data from the 1989 Soviet Census for Estonia. Estonia then had a population of about 1.5 million people. We ran the program on a test file with 50 cases and then downloaded all of the data for Estonia, which took about 26 hours. The only other computers with the processing power, workspace, and storage necessary to do this task in Estonia were mainframes. The downloaded data provided the sampling frame for representative surveys, including the Estonian Family and Fertility Survey (1994 female, 1997 male), the Estonian Labor Force Survey (1995), the Estonian Health Interview Survey (1996), and the Survey of Ethnic Minorities (1997), which greatly expanded the analytic possibilities for Estonian demographers.

> *YET IN THE MAINFRAME ERA, DEMOGRAPHERS DEPENDED ON STATISTICIANS AND PROGRAMMERS TO IMPLEMENT THE ANALYSES WE WANTED TO RUN.*

Reduction in the costs and time constraints involved in computing sometimes led to egregious errors, even in analyses that were published in prestigious journals. A well-known case of an analysis led astray because of failure to eliminate miscodes and errant values was published in the mid-1980s by Jasso [7]. The study analyzed coital frequency. Married women were asked how many times in the previous four weeks they had had sexual intercourse. The study yielded the surprising finding that coital frequency increased with the logarithm of woman's age. There were about 2000 cases. The missing value code was 99, and those cases were properly treated by Jasso as missing values. There were also four cases with a recorded value of 88. It seems likely these were miscodes—data entry errors—but they were included in the analysis. Four other cases had recorded values above 40, indicating sex 10 times a week on average or more, and were also included by Jasso as valid values. A reanalysis of the data [9] showed that the elimination of those eight cases removed the statistical significance of the logarithm of woman's age. If the values for individual variables had been examined before the analysis was done, the cases coded 88 and perhaps the cases coded greater than 40 would have been eliminated from the analysis. Probably the ease of moving directly to a multivariate analysis led the researcher to skip preliminary inspection of individual variables.

The increased ease of performing more complicated statistical analyses with more cases also increased the temptation for mindless data mining for statistically significant results, a bad practice often referred to as p-hacking [6]. Recently, in response to p-hacking, some journals have begun to require researchers to file their hypotheses and planned statistical analyses in advance. Work in data science and data mining has addressed the problem as a philosophical issue [12]. Researchers would do well to remember Thomas Kuhn's [9] advice about the roles of exploratory versus hypothesis testing research, and the different approaches to statistical analysis in the two kinds of research.

THE INCREASING USE OF LIFE COURSE ANALYSIS AND EVENT HISTORIES ALSO REQUIRES A DETAILED DATING OF EVENTS, SUCH AS MARRIAGE, DIVORCE, CHILDBIRTHS, AND PERIODS OF UNEMPLOYMENT.

The introduction of user-friendly statistical packages reduced the reliance of demographers on statisticians and programmers. However, limitations and errors in statistical packages sometimes created new problems. In 1985, two political scientists and I were looking at factors related to vote misreporting [13]. With a dichotomous dependent variable, it was clear to us that a logistic regression was appropriate. We were using SPSS for data analysis, and SPSS included a logistic regression procedure. The version of SPSS used at the University of Michigan was from the University of Alberta. The SPSS output was very clear and easily interpretable. The only problem was that it gave the wrong answers, reporting the degrees of freedom as the chi-squared value. At first the University of Michigan computer consultant refused to acknowledge that the SPSS results were wrong. Sometime later the SPSS program at the University of Michigan was fixed. It had been susceptible to such an error because logistic regression was a fairly new procedure. In the meantime, I did the logistic regression analysis using GLIM, a program produced by the Royal Statistical Society in Great Britain. The GLIM output was less easily interpretable and required more calculations by the researcher, but the answers were correct.

Changes in ease and autonomy in computing saved researchers an enormous amounts of time, and gave them the freedom to think more flexibly about alternative hypotheses in more creative ways without having to limit analysis plans artificially by time and cost constraints. This was especially helpful to graduate students and junior scholars who had less access to programmers and sometimes had difficulty getting likely costs approved.

With changes in computing power and statistics there were also changing concerns about protecting the identity of survey and census respondents. The U.S. Census Bureau has long masked aggregate data for categories with fewer than 10 residents. Surveys have often protected identities through reporting geographic location at coarse scales or reporting when events occurred by five-year intervals rather by single-year.

These approaches assume that researchers do not need data by small geographic unit or by year or even month of occurrence. However, the increasing use of GIS and of multilevel analysis, which combines individual responses with ecological areal data, calls for detailed location data for individuals or households [2], [10]. The increasing use of life course analysis and event histories also requires a detailed dating of events, such as marriage, divorce, childbirths, and periods of unemployment [4], [16].

Recently, the ubiquity of data in our lives has led to heightened concerns that individual survey or census respondents could be identified, which has led to recommendations of ever more restricted access to datasets, and ever more masked and perturbed data [5]. This alteration of the data results increasingly in a tradeoff between data usefulness and protection of respondent confidentiality.

To protect datasets with personally identifiable information, while simultaneously trying to maximize opportunities to conduct research, the Census Bureau established the Research Data Centers (RDC). The first was located at the Census Bureau's headquarters. The first RDC outside of the Census was established in Boston in 1994. In 2021, there were 30 RDCs located throughout the country [14]. These now offer individual-level data linked across many different agencies. This allows the use of confidential data but is also quite expensive and time consuming. The requirements and costs differ by the agency from which the data come and some RDCs do not actually collect the seat ticket fees, but these fees can be as high as $1800.

The researcher must write a proposal and get approval. For Census Bureau data, the proposal must:

"Provide benefits to Census Bureau programs
Demonstrate scientific merit
Require non-public data
Be feasible given the data
Pose no risk of disclosure" [15].

These are not unreasonable requirements, although providing "benefits to Census Bureau programs" is not a standard research requirement.

With an approved proposal, the researcher is given the results of the desired statistical analysis run on the original rather than the masked or perturbed data. The researcher never has an access to any of the data, and the analysis is run by an employee of the RDC.

With increasing concern about disclosure risk even for data not usually considered sensitive, this mode of data access could be extended to a wider range of data. The issue becomes more pressing with the collection of extremely sensitive data, such as biomarkers, and the increased desire and capability to merge data from a large variety of data sources. The future may see a starker choice between 1) limited data that a researcher can use on its own for free without anyone's permission, and 2) detailed data but with costs and time lags in return for the researcher, creating a situation similar to mainframe computing through a programmer.

BIBLIOGRAPHY

[1] B. A. Anderson and J. L. McCabe, "Nutrition and the fertility of younger women in Kinshasa, Zaire," *J. Develop. Econ.*, vol. 4, pp. 343–363, Dec. 1977, doi: 0.1016/03014-3878(77)90014-1.

[2] M. T. Berg, C. H. Burt, M.-K. Lei, L. G. Simmons, E. A. Stewart, and R. Simons, "Neighborhood social processes and adolescent sexual partnering: A multilevel appraisal of Anderson's player hypothesis," *Social Forces*, vol. 94, pp. 1823–1846, Jun. 2016, doi: 10.1093/sf/sow032.

[3] M. Cancian, Y. Chung, and D. R. Meyer, "Fathers' imprisonment and mothers' multiple-partner fertility," *Demography*, vol. 53, pp. 2045–2074, Dec. 2016, doi: 10.1007/s13524-016-0511-9.

[4] B. Crepon, M. Ferracci, and D. Fougere, "Training the unemployed in France: How does it affect unemployment duration and recurrence?," *Ann. Econ. Statist.*, no. 107/108, pp. 175–199, Jul.–Dec. 2012, doi: 10.2307/23646576.

[5] S. Garfinkel, J. Abowd, and S. Powazek, "Issues encountered deploying differential privacy," in *Proc. Workshop Privacy Electron. Soc.*, Oct. 2018, pp. 133–137, doi: 110.1145/3267323.3268949.

[6] N. Hawkes, "Sixty seconds on . . . p-hacking," *Brit. Med. J*, vol. 362, p. K4039, Sep. 2018, doi: 10.1136/bmj.k4039.

[7] G. Jasso, "Marital coital frequency and the passage of time: Estimating the separate effects of spouses' ages and marital duration, birth and marriage cohorts, and period influences," *Amer. Sociol. Rev.*, vol. 50, pp. 224–241, Apr. 1985, doi: 10.2307/2095291.

[8] J. R. Kahn and R. Udry, "Marital coital frequency: Unnoticed outliers and unspecified interactions lead to erroneous conclusions," *Amer. Sociol. Rev.*, vol. 51, pp. 734–737, Oct. 1986, doi: 10.2307/2095496.

[9] T. Kuhn, *The Structure of Scientific Revolutions*. Chicago, IL, USA: Univ. of Chicago Press, 1962.

[10] F. I. Matheson, H. L. White, R. Moineddin, J. R. Dunn, and R. H. Glazier, "Neighborhood chronic stress and gender inequalities in hypertension among Canadian adults: A multilevel analysis," *J. Epidemiol. Community Health*, vol. 64, pp. 705–713, Aug. 2010, doi: 10.1136/jech.2008.083303.

[11] J. W. Ruden. *The History of Computing at Cornell University*. Ithaca, NY, USA: The Internet First Univ. Press, 2005, pp. 61–62.

[12] X. Shu, *Knowledge Discovery in the Social Sciences: A Data Mining Approach*. Oakland, CA, USA: Univ. California Press, 2020.

[13] B. D. Silver, B. A. Anderson, and P. R. Abramson, "Who overreports voting?," *Amer. Political Sci. Rev.*, vol. 80, pp. 613–624, Jun. 1986, doi: 10.2307/2095496.

[14] United States Census Bureau, *Research At the Center for Economic Studies and the Research Data Centers: 2000-2004*. Washington, DC, USA: U.S. Census Bureau, 2005.

[15] United States Census Bureau, 2020, Restricted-Use Microdata, Census Bureau website. Accessed: Mar. 21, 2021. [Online]. Available: https://www.census.gov/programs-surveys/ces/data/restricted-use-data/apply-for-access.html

[16] D. Vignoli, A. Matysiak, M. Styre, and V. Tocchioni, "The impact of women's employment on divorce: Content, selection, or anticipation?," *Demographic Res.*, vol. 38, no. 37, pp. 1059–1110, 2018, doi: 10.1080/13545701.2011.582822.