ORGANIZING CENSUS RESULTS FOR USE BY SOCIAL SCIENTISTS*

Everett S. Lee, University of Georgia
Richard C. Taeuber, Oak Ridge National Laboratory

We stand at the beginning of an era in social research, unfortunately one which we are not prepared to enter. The Census Bureau has adopted a policy of releasing elaborate tabulations for small areas on magnetic tapes and has prepared a series of one-percent samples of individual records for 1970 (in which no person can be identified, of course), that permit the investigator to make whatever tabulations or correlations he wishes. Furthermore, the Census Bureau is providing comparable data for 1960 and has arranged for the release of Current Population Survey Data, after combining surveys in such manner that the statistical variation is not distressingly great. There is perhaps some hope that samples will be prepared from the records of earlier censuses so that much needed historical research can be performed with the same ease as that based on the 1970 Census.

The Census Bureau is not the only agency to have taken an advanced view of data dissemination. The Social Security Administration has released tapes which include one-percent samples of social security records, going back more than a decade, and there is experimentation with the cost and value of increasing the sample size, perhaps to ten percent. Data from the Census of Manufactures and the Census of Agriculture are also available on tape, as are the materials published in County Business Patterns. It may well be that the National Center for Health Statistics will follow suit. Already the amount of material available on magnetic tape is such that major efforts are necessary to catalog the available materials.

While we have long been handicapped more by lack of imagination than by lack of data, the situation of the social scientist and planner in this country is greatly improved over one that was already excellent. As Mary Jean Bowman has noted:

"The leadership of American economists in the development and empirical testing of models for analysis of investment in education rests firmly upon the imaginative and yet unwitting contributions of sociologists, demographers, and very practical market analysts in private business. It is these people who persuaded the U.S. Census to put earnings and educational attainments into the 1940 Census, along with breaks by sex, age, race, and region. These data were improved in the 1950 and especially the 1960 Census, and will be even better in 1970. They are the envy of economists in other lands concerned with human resource problems."1

This is a true statement but it deserves an addendum. Some of the contributions were not unwitting, and many of them were made by social scientists within the Census Bureau itself, who were led to them by their own analytical work. One of the great strengths of the Census Bureau has been that its leaders have distinguished themselves in the analysis of census and related materials. Consequently, they have been made aware of the pitfalls and limitations of the data, and they have had the easy access to other leaders in social science research that comes from intellectual communion and cooperation. While there were many people outside the Bureau who urged the development of public use samples and other magnetic tape releases, it was through the efforts of Census officials, themselves analytically oriented, that we have the current wealth of data.

We have moved so rapidly into this new era of data manipulation that none of us are prepared. The Census Bureau has understandably encountered difficulties and delays in making the tapes available in usable form. The spate of private research and data processing organizations that sought to use the census tapes in providing services to their clients has been decimated by inflation and recession. Only a few remain and probably all of these have channelled their efforts much more narrowly than originally intended. Universities and nonprofit research organizations have also had to curtail their plans, as the costs of assembling the census tapes, providing the necessary programming, and developing working tapes have become clear. There is not only a shortage of money, there is an even more disturbing lack of properly trained personnel. Furthermore, all of us have to be made aware of the multiplicity of opportunities we are offered, though as E. B. Wilson once remarked of mathematics, statistical technique under the spur of the computer is like a horse with so many and so varied gaits that the entranced rider is likely to ride off in all directions. Until we become better accustomed to both tools and data, we may fully expect to see regressions and multiple regressions without end or reason, and we will find certain variances explained by several hundred percent.

There are lessons to be learned from the Surveys of Economic Opportunity, imaginative undertakings sponsored by the Office of Economic Opportunity and conducted by the Census Bureau. After tapes from these surveys were released by the Census Bureau, it took the joint efforts of the Brookings Institute and the ASSIST Corporation to make the tapes usable to non-census people. Even so, important questions including sample weighting and variances, remain to plague investigators. Programming turned out to be no easy matter, and it is only now, through the combined efforts of social scientists from the University of Georgia, the University of Wisconsin, the Department of Agriculture and elsewhere, that data from these surveys can be brought to bear on problems of immediate importance.

Experience with these surveys and with a few 1960 census tapes, including the 1/1000 sample, has saved us many difficulties, but there are still many steps and costs involved in proceeding from the census tapes to research findings. In the first place, tapes from the Census Bureau are high in cost and sometines slow in delivery, because the copying and documentation of tapes has to be fitted into the Bureau's heavy operational work load. As received from the Census the tapes are not in the best form for processing. For the sake of economy in storage and handling they should be packed and converted for particular tape drives. Programming suitable for several types of computers and for varying materials (the Fourth Count as against the Public Use Samples, for example) is necessary.

Fortunately, these needs were foreseen by the Ford Foundation, the National Science Foundation, and the Center for Research Libraries,who took steps to facilitate the widest possible access to the Census data by providing grants to universities to obtain tapes and programs and by arranging with a non-profit organization, DUALabs, to provide compressed tapes and adequate programming at minimal costs. Arrangements were also made to provide extract or working tapes at the lowest possible costs to researchers. Perhaps just as important was the provision of an organization which could make copies of tapes without the long delays that are associated with governmental operations. To the researcher quick cost estimates and quick delivery of materials after they have been purchased are crucial concerns.

But even with these developments it was clear that there were alarming as well as desirable features in the new methods of data dissemination. It remains true that some of the most important papers will be written by researchers who, thumbing through a census volume, are struck with the relevance of a single table or combination of tables to an important concern, and thereafter with pencil and paper, and perhaps with desk calculator, perform the necessary recombinations and calculations. Nevertheless, the lone researchers or groups of researchers at smaller universities have been placed at a considerable disadvantage vis-a-vis those in places with better computers, more programming assistance, and money to amass tapes. What is being ushered into the social sciences is the development of Big Science, performed in large establishments with large research staffs and experienced support personnel, as is already the case in the physical sciences. Again, we emphasize that much can be done by the imaginative researcher who has nothing more than pencil, paper, and a census volume, but he may have his productivity crippled by the awareness that others have better access to data and can obtain their results more easily. We can only foresee an increase in the psychological blocks to productivity for researchers in small and poorly supported institutions.

Although the physical sciences have moved in the direction of Big Science, partly because of costs and partly because of the many advantages to assembling a group of scientists to work on a specific problem, ways have been worked out to facilitate cooperative efforts and increase the possibilities of faculty members and students at smaller or faraway institutions. The reactors at AEC National Laboratories have been used for experiments by scientists from many universities, and fellowships have permitted students from all over the country to use the complex and expensive equipment found in national laboratories. In such ways Big Science, as represented by large research organizations, has aided in research and training in many other institutions.

In the South an opportunity for cooperative arrangements among social scientists is being provided by the development of a data center at the Oak Ridge National Laboratory (ORNL) in connection with work being done for the Department of Housing and Urban Development. At this center all of the tapes from the 1970 Census are being amassed, along with available tape files from the 1960 Census, the Social Security One-Percent Sample, County Business Patterns tapes, and a number of valuable files made available by individual researchers, such as the Bowles-Tarver County Migration tapes for 1950-60 and the SEA by SEA migration flows for 1955-1960. Cooperating in this endeavor have been the Department of Agriculture, the Rand Corporation, the Tennessee Valley Authority, and a number of universities throughout the country.

The Oak Ridge Associated Universities, a corporation sponsored by 43 Southern colleges and universities with graduate schools, was established to further cooperation between the National Laboratory and educational institutions. Recognizing the possibilities afforded by the data bank and by the computer facilities and technology at Oak Ridge, a group of demographers asked to be allowed to utilize the Oak Ridge Associated Universities in somewhat the same way as physical and biological scientists. The Southern Regional Demographic Group resulted, a group which now has over 150 individual members from many colleges and universities. Its purpose is to ease access to ORNL data and capabilities for members of the group, to organize and participate in studies which were beyond the facilities of a single university, and to organize conferences and symposia addressed to research problems and needs.

Initial funding was received from the Ford Foundation, and a series of conferences have been held. In March, 1971, a conference on Research and the 1970 Census was sponsored by the National Center for Population Research of NICHD, the papers from which have been published. A meeting scheduled for September 28-29, 1972, deals with "Research Needs for a Southern Population and Urbanization Policy." A central office with a full-time executive secretary supplies needed materials to member researchers, arranges for the dissemination of information on

research carried out by its members, and facilitates cooperative endeavors among universities.

At ORNL preliminary attention has been focused upon methodological and data processing questions. The public use samples are being converted into binary form as an aid in compression and in speed of processing. Available computer programming packages such as SPSS are being modified and extended for use with the public use samples, as is Howard Brunsman's CENTS program. In these endeavors we are cooperating closely with DUALabs and research organizations in NIH and the Department of Commerce. Tapes have been received, catalogued, combined, extracted, compressed, and otherwise processed for immediate research or for wide spread usage.

A number of pilot studies have been made which were designed to explore the possibilities of varied materials and to discover the difficulties in their use. One such study, carried out in cooperation of the Tennessee Valley Authority, dealt with migration into and out of the Tennessee Valley, and gave us valuable insight into the use of Social Security materials in assessing the flows of labor. We then secured the cooperation of David Hirschberg and the Bureau of Economic Analysis in making a study of changes in the labor force in metropolitan areas.

We have now moved to the construction of a special tape for the Atlanta Metropolitan Area which permits us to follow over a ten-year period anyone who was ever in covered employment in that city. We can determine where he was before he came to Atlanta, what his annual earnings were, what he earned while he was in Atlanta, and where he went when he left Atlanta, all that by age, sex, color, and industry. Once such a study is made we hope to encourage comparable studies by other researchers in other cities. Our aim is not to direct the course of research, but to permit replication and improvement by studies in other areas. Our hope is that the results of these initial efforts will be multiplied and improved upon so that large initial processing costs need not be borne by researchers with little equipment or limited funds.

Similar exploration has been done with census materials and a study of the usefulness of city directories as a supplement to census information is under way. Examinations of black suburbanization have been made in several metropolitan areas, while other studies deal with the shift of industry from central cities to the rings. 'Growth by size of place has been studied as it relates to past growth, to location in regard to metropolitan areas, or in regard to interstate highways. As an aid in understanding areal patterns we have experimented with computer graphing. For example, we have worked out ways of graphing migration rates for counties by age, sex, and race so that we can get graphs for all 3000 counties in a few minutes of computer time. When we have the basic data on hand we will make similar graphs for 1970 and offer sets for 1960 and 1970 for publication in the various states. Another program converts density by enumeration district or block group to density per kilometer, giving us density maps for 1960 and 1970 that can be compared to see which areas lost and which gained population. Once done for a single city, processing for a large number of cities is straightforward. Furthermore, the same technique can be extended to black population, the aged, broken families, or to any other quantity we choose to relate to area.

Although our efforts have been going on for more than two years, we feel we are just beginning. We have experienced enough irritation from members of our own group at our slowness and at our fumbling with unfamiliar materials that we have come to sympathize greatly with the Census Bureau and other organizations that are delayed in the processing of eagerly awaited data. Frustration, failure, procrastination, and sheer stupidity among our own ranks has taught us humility. We shall not soon master the necessary techniques for processing and analysis of such a flood of data but there is an encouraging acceleration in our efforts.

Furthermore, we are even more convinced that cooperation among institutions and the establishment of non-profit and semi-public data centers are essential to the proper exploitation of census and other materials. One of the great failures of social science research is the lack of replication of studies in other places and for different times. We believe that more research per dollar and per scientist-year can be obtained through the kind of cooperation we have initiated and that it can be done better and more quickly. We are moving to institute fellowships and summer research opportunities. In cooperation with two different universities we have now produced two Ph.D.'s and we have brought students from a number of instiutions into our ongoing research. We expect to cooperate with the graduate schools of our member universities in providing data and facilities for students who will work with our help but under the direction of their own faculty committees.

Finally, we are encouraged enough by results to date to recommend the establishment of regional centers for the receipt and processing of data tapes. The kinds of data we have described will multiply far faster than the persons who can process and analyze them, or even find them. All too often the establishment of a computer center and the turn toward magnetic tape has slowed down research and increased expenses, rather than the reverse. We have never been so much in need of guides to available materials, of methodological studies of particular sets of data, and of the development of quick and rapid processing. As social scientists, we should follow the lead of the physical scientists to combine the efforts of Big Science with the furtherance of individual research.

Footnotes

*Paper presented at the meeting of the American Statistical Association in Montreal, August 1972.

[1]M. J. Bowman in Mark Blaug, Economics of Education, Volume I. Cambridge, Cambirdge University Press, 1968, pp. 114-115.