DISCUSSION

Rudolph C. Mendelssohn, Bureau of Labor Statistics

In the October 1970 issue of the American Statistician, there is a paper which I wrote entitiled The Principles of Processing Statistical cal data.1/ In it I said that there are certain functional operations, some, or all of which, typically appear in computer systems that process statistical data. Let me review that paper to give background and scale to my comments on the papers presented here.

BLS collects about 250,000 reports each month. This substantial collection of micro data is run through our computers to give figures on employment, pay, hours worked, and other data that describe the economic conditions of the rank-and-file worker. The computers put the incoming reports and subsequent operations through five or six diferent but clearly identifiable steps--or functional program modules--designed to yield a predetermined set of statistical tables at the end of the line.

Screening: Error Detection and Correction
The first of these program module steps is called a screening or edit run. The computer is programmed to examine the micro data for each report and mark those which appear to be of doubtful validity, or clearly erroneous, due to respondent error or incorrect transcription to the report form or to the punched cards, and so forth.

Tabulating and Estimating
The second functional task is a summary or cross tabulation of all the reports, separately, of course, for each survey subject; for example, let's take the five or six micro items reported by establishments on their total employment, number of production workers, and the total hours they worked, their total pay and overtime hours. When tabulated, they tell us something about the total U.S. employment in about 400 nationwide industry categories, ranging from drugstores to steel mills. Average weekly pay and hours of work are also compiled at the same time for these industries.

Of course, the summation of reported data does not necessarily give numbers which are an estimate of the universe from which the sample cases came. A third program module may be used to expand the reported figures to reflect the universe estimate, usually by one of two methods, "blow-up" or link-relative.

Storage, Retrieval, and Analysis
Our last three modules are programs for storage, retrieval, and analysis. In the BLS, programs are available which permit our economists and statistcians to do analytical research using the computer. For example, they can retrieve the data for an industry or set of industries and process them through seasonal adjustment programs, regression programs, growth rate programs, and so forth.

In all essential aspects, systems for processing occupational wage data, wholesale and retail price data for our Producer Price and Consumer Price Indices, and so forth, are like the sequence I have outlined. In some way or other, micro data are screened, tabulated to get macro figures, estimates are computed and the final figures are stored, retrieved, and analyzed by common functions.

Computer Power
In 1970, each subject survey system required a separate tailor-made program for each functional module. My 1970 article asks: "Have we used the electronic computer in a sensible way? For example, why not just one general program to screen all micro data? After all, computer works equally well with numbers from any source, regardless of thier substantive meaning. And the kinds of tests which can be made are limited, when all is said and done, to the simplest arithmetic operations."

"Why not general programs to compute macro data from micro files, and to store, retrieve, and anlayze results? A statistical table is a simple tool: It is a matrix of rows and columns with a stub and heading. Yet, beyond certain narrow limits, we need a new set of computer programs for each new table. It seems to me the computer should be able to make any table for us if we only tell it how the one we want differs in particulars from the general idea of a table."

During the eight years since I wrote that paper, much work has been done to design and implement general tools and these are now in use, some more widely than others. For example, the Swedish Central Bureau of Statistics has a powerful generalized cross-tabulation system called TAB-68, which is in use in the national statistical agencies of several European Countries. BLS has a similar system called Table Producing Language (TPL), widely used in both North America and Europe. Statistics Canada, the Canadian central statistical agency, itself has several very useful generalized tabulation packages. Data Base Management Systems to control, store, and retrieve data are available commercially. TOTAL, the one we use, is widely available and there are others suited to varying needs. There are not many generalized editing programs but one outstanding example is CANEDIT, again from Statistics Canada. It is gaining kudos in North America and Europe.

Although the review of general statistical data processing functions that I have just taken you through implies a yardstick against which I would have liked to measure the three systems reported today, the structures on which these systems are based are not clearly laid out in two papers. So, I shall confine most of my comments to the one paper that allows me to look inside, with a few brief comments on each of the other two papers.

Monitoring Survey Field Operations

First, a brief comment on Monitoring Survey Field Operations. The report by the Research Triangle staff, by intent of the authors, seems to cover, mainly, a limited portion of the full spectrum of statistical data processing; namely, that of survey control. There is brief mention of Data Edit and in this connection, I wonder if they thought of looking into Statistics Canada's CANEDIT program. This highly regarded program has attracted international attention and seems, on the surface, ideally suited to the Research Triangle problem.

I know of no general solution to the problem of automated control of survey respondents. If there is one I should like to know about it because we need one at BLS. Following my own advice, it is clear that we should take a close look at the work done by the Research Triangle Institute. Perhaps we can use or build on the work they have done. I find the fact that the system has expectations about the response from each participant especially novel and useful. Most systems fly blind, not knowing what to expect.

The PSID System

I come now to the PSID system. Most of my comments will be directed to it because the authors of that paper helpfully laid it all out so that the warts as well as the virtues are clearly seen. First, a few comments on the control file, the direct data entry approach, and then a more detailed review of the data base approach.

The control file, giving immediate information on the status of the interview process, is well thought out and is something we at BLS would like to see in more detail and perhaps copy. And fitting this in with the compilation of interview cost figures for accouting purposes is unique in my experience and, no doubt, useful in appeasing those who hold the purse strings.

The Direct Data Entry approach is highly innovative. I know of only one similar effort, an experiment by the U.S. Census Bureau over a year ago. Some of you may have heard their report in their session here. In the Census case, on-line CRT's were used to prompt the interviewer in a household telephone survey. Since the PSID data entry system is still under development, I would encourage the ISR folks to compare their experiences with those of the Census Bureau. As far as BLS is concerned, I would like to have us keep in touch with this project as it has productive implications for our work.

I am afraid I cannot be as complimentary about the PSID data base approach. Here is why: In my experience, roughly speaking, there are two kinds of statistical data processing requirements; first, batch processing where the records are processed only once to compile results. Thereafter, the files are archived as they are likely not to be used again. For this need, tape sequential processing is perfectly appropriate--you pass the file once, you get your results and give the tape to your librarian. The second kind of processing is called transactional. Here, a beehive of activity continuously impinges on the basic record. There is a constant flow of information into records on a selective basis--added data and corrections come in and selected cases are the target of queries and other sets are used to compile results at frequent varying intervals.

Despite the author's claim to the contrary, I see the PSID as a transactional system which has been placed in the straight jacket of a tape sequential approach. Naturally, such a distortion required an ingenious escape mechanism. A much simpler, less tortuous, and more direct solution would have been to use any one of a number of data base management systems for the direct access, transactional approach. Then, it would have been possible to get into records directly and readily as information about them becomes available or is needed on an individual or set basis. There are probably 2,000 of these data base systems from a dozen or more vendors in use around the world.

But let me take the view that the system is indeed naturally tape sequential. Then, I ask, why not use a generalized tabulation tool, such as TPL which the BLS developed precisely for dealing with tape sequential hierarchical files five years ago? The system is used in over 200 installations, mainly in North America, but also in Europe and the Far East. The Michigan folks might argue that their file structure and survey responses antedate TPL by several years and so they are hampered by past commitments. Well, they have a flexible "Extract" program that could retrieve and reformat the file rather easily, I gather.

In summary, the PSID system is innovative and imaginative in those areas where solutions to functional requirements have not been met elsewhere but perhaps shortsighted in failing to take advantage of what has been done elsewhere.

The Canadian Labor Force Survey

My comments on the Canadian system are very much like those I have just made about PSID. It's an interesting mixture of novel, forward-looking development of new tools and a disappointing reinvention of old ones. The use of minicomputers in regional offices to generate the questionnaires and for subsequent entry of returns via visual display terminals is up-to-the-minute use of the state of the art. But why did the architects of this system plan, design, and implement a 26,000 line editing program called HOPS, rather than use the excellent CANEDIT program, in use elsewhere in Statistics Canada and in some national statistical agencies in Europe? Why did they not use one of several generalized tabulation systems? The answer, I think, is that many of us in our business are infected by an illness that is epidemic in the data processing field, especially in statistical data processing. It's called the NIH disease--Not Invented Here!

Conclusion

I can agree that some might question the notion of building systems from blocks of functional modules. Well, we are doing it at BLS. In one graphic instance, an entire system was constructed from off-the-shelf items. Not a single line of newly written code is used. The system tests data for validity with a routine from the collection of statistical analysis routines which we got from North Carolina State University; updates and extracts data with a prototype of the file manipulation system; manages its data base with TOTAL which we bought; uses TPL and its Codebook for cross tabulation; seasonally adjusts data with the X-11 program which we got from the Census Bureau; and TPL for table display by line printer or through an electronic photocomposing device where high-quality publication standards must be met. In addition, if the system managers wish to do analysis they can use a packaged charting device called DISSPLA, which we bought, and an interactive, on-line macro data retrieval and manipulation language called MDL (for Macro Data Language), which we got from the Federal Reserve Board.

In summary, in many cases it really is not necessary to reinvent the wheel or repeat mistakes of the past.

----

[1] Rudolph C. Mendelssohn, "The Principles of Processing Statistical Data," The American Statistician, Vol. 24, October 1970.