SOME IMPLICATIONS OF COMPUTER PROCESSING OF ECONOMIC CENSUSES AND SURVEYS

Jack A. Scharff, Nathaniel Swersky, Eugene L. Wendt, Bureau of the Census

In 1951 a small group of individuals including the Chairman of this session, Dr. Daly, pioneered at the Bureau of the Census in the planning and programing of the Annual Survey of Manufactures to be run on a new piece of equipment called UNIVAC. They had no idea that from this small beginning there would be proposed such a wide range of computer applications as have been developed in the last 10 years. The consequences of the work of this group of people have had far reaching effects in many areas, not the least of which is the economic statistics surveys of the Census Bureau. The work that is now being done on the computer far exceeds the dreams of the people who were working on this one project ten years ago. It might be said that we are now celebrating an anniversary of the processing of economic statistics on the computer.

Why can it be so candidly stated that the original pioneers in programing had so little concept of the wide range of contemporary applications? The answer is found in that the implications and applications of computer processing were not fully understood at that time, and, in fact, currently, computer applications stretch as far as the imagination and desires of programers and management.

In Appendix A there is a partial list of the economic statistics projects that have been processed in whole or in part on the computer at the Bureau of the Census in the last few years. The size of this list is only a fair indication of the extent of computerization in the field of economic surveys. The estimates of work to be performed on the computer in the next few years is much greater as is indicated on this list, and would tax the capacity of existing Census computers without considering any other surveys or censuses which the Bureau will conduct. In order to overcome this shortage of computer hours there is under way at this time a feasibility study to determine how much additional capacity is required and which computer will best fulfill the requirements.

In order to more fully appreciate the work involved in the list of projects mentioned, one must consider some of the implications of this processing effort and of computer processing in general. Of paramount importance are the following:

1. Types of personnel involved in the staffing of a programing unit and their functions.
2. Programing methods utilized.
3. Systems analysis and data analysis as related to computer programs.
4. Interest of management in furthering objectives.
5. Scheduling of operations, programing and production by the computer.

Staffing a Programing Unit

There are three basic types of skills required within a successful programing unit. First is the executive or unit administrator who has the primary responsibility to receive all incoming requests for programing, to set up the necessary meetings with project sponsors in order to clarify and solidify specifications, to allocate the project to a responsible systems analyst or project planner, to establish a schedule for the completion of the project, to allocate test periods and computer priorities to insure the culmination of each job within the prescribed period, and to set programing policies, principles and techniques for the unit.

Next is the systems analyst or project planner who has the responsibility of evaluating the particular project, in order to determine the best method or computer system to be used in deriving final results and to prepare an operational flow-chart which will guide all programing efforts.

Last is the coding programer who works with the analyst to prepare and test the programs dictated by the operational flow-chart.

Programing Methods for Computer Results

One should never forget that implicit in the use of computers is the ultimate goal of obtaining results. A successful program is one which essentially fills the real needs of management within the desired time limits and as budgeted.

In order to attain this end some programing policies are implied. In the Economic Operations Division of the Bureau of the Census four basic programs are produced. The first three listed below are capable of producing almost immediate results, that is to say within a day and less than a week from inception to completion of project.

The first type of program and usually the easiest to write, is that which merely requires insertion of parameters into a generator which, after computer processing, provides a running program. Typical of this application are sorts and merges which are more or less standardized.

The next type utilizes a basic input/output framework program including subroutines for input/output, housekeeping and conversion. The actual processing steps required are coded manually and then entered into a provided storage area of the program.

The third type, different from the one above, utilizes as a basic source a previously completed program which accomplishes results similar to the current needs. Input data is tailored to fit the conditions and revisions are made in the program to serve the input.

Finally there is the full-fledged complex program required for high volume processing or for periodic surveys. Within this category of programs it has been found necessary to fully document every stage of operations so that modifications may be made in the future by persons other than the one who originated the program.

Within these four categories, during the course of the last six months, supposedly a slack period between economic censuses, there have been 500 programs written by fewer than 25 people.

Systems Analysis and Data Analysis as Related to Application to Computers

By far the most critical implication to successful processing is the preliminary work done prior to the program coding effort. This is the "leg work" in which the executive of the programing effort gathers together the requirements from management usually in the form of "specifications" and evaluates these specifications in concert with statisticians, methods determination experts, program analysts and the like, in order to determine how the particular program may best be resolved.

The duration of this initial effort is a variable depending upon the conciseness of specifications and the intricacy of the problem. The following summary of some actual applications is illustrative of some but not all of the implied considerations in computer processing.

General Effects of Computer Processing of Economic Statistics

There are a series of questions which can be asked when exploring whether a survey should be processed on the computer, but the most important one is, "What can I obtain from computer processing that I cannot get from another technique, whether in timing, or cost, or quality of work?". The first possible answer to this is consistency.

It is well recognized that one of the many advantages of the computer is the absolute consistency between the handling of one item and the next. Once the computer has been correctly instructed to do a job it performs its mission tirelessly and efficiently.

Consistency, however, is only a virtue if the results are consistently correct. The necessity to provide a system which represents a carefully developed plan with no logical loopholes creates the need for a different emphasis by statistical personnel than that required in a non-computer philosophy. The pre-test of plans is of greater importance; the need for investigation of all possible consequences of specifications for a survey is magnified; and the requirement for liaison between the statistician and the computer technician is mandatory.

Thus under a computer oriented system a person not only has to be proficient and understand the subject matter, but also should be able to translate this understanding into a format adaptable to the computer. Sometimes even the best plans cannot be translated into computer language because of what is called the lack of judgment of the computer. A person who is not only familiar with the subject matter but also with the computer can usually devise means of building a pseudo-judgment factor which can then be adapted for the computer. As an example, there are many surveys which use a consistency edit as one of the techniques for determining the reliability of input data. The computer makes decisions by applying a procedure to data and using as a base its recent experience with similar data, much the same way as a clerk would use his judgment.

Another of the results of computerization has been to transfer the work load from the clerk to the systems analyst. This phenomenon has had one advantageous effect in the Census Bureau where traditionally there have been periods of peak employment for comparatively short periods of time. It is now possible to conceive of a type of operation which will minimize the effects of this peak employment and will create a more stable working force. This requires a greater lead time to properly prepare for a major census or survey. A wholesale change can never be made without sufficient lead time to thoroughly check the results usually by processing on the computer a small portion of the data and reviewing the output to determine whether the work has been performed satisfactorily.

The computer, which is equally effective in accounting and statistical processing, can decide whether a particular set of conditions are acceptable and, if not, can substitute for the unacceptable data a closely related series. In the Foreign Trade Exports processing, in order to eliminate inconsistencies in the statistics every detail record is checked for unit price. Most of the items for each commodity are within the established range of acceptable prices. Previously, those items

that fell outside the price range were reviewed clerically and were re-inserted in the statistics in the following month. These specifications produced statistics which were considered to be misleading because of the carry-over from one month to the next. In order to overcome this problem, the computer now derives a new quantity based on the characteristics of historic data. The value is left unchanged because in numerous studies it has been found to be one of the most reliable of data items on the record. After this imputation the detail is reasonably consistent, both internally and between months, however, this has had the effect of minimizing the legitimate, large differences which might occur in the universe.

Relationship of Computerization to Quantity of Output and Operation

There is a sign on the door of a programmer's office which may be attributed to an anonymous source: "Never begin a vast project with half vast ideas." This statement is particularly appropriate when applied to computer work. The amount of information and quality of operation that can be achieved with computers is only limited by the ability of the humans to conceive ideas. Certainly, it is true that projects as large as the surveys conducted by the Census Bureau deserve a considerable amount of preliminary thinking in order to make the use of the computer worthwhile. In addition to the miles of paper that can be generated by the computer, there is also the ability to utilize analytical tools which have previously been too costly to consider.

The Annual Survey of Manufactures, prior to 1959, was a probability sample designed to publish general statistics with considerable detail but with only limited coverage for product statistics or for statistics on small areas. The sample selected for the current panel of the Annual Survey of Manufactures increased the coverage sufficient to permit publication of data on product classes and local areas in detail not previously considered publishable because of the high variances.

Subsequent to the selection of the sample, a dual control was established on the Annual Survey of Manufactures panel. First, a mailing register was initiated to provide for printing the 1958 data on the report forms. Separately, using the same source material, a data register was established to be used for the computer editing and tabulating routines. This dual register has provided the means of controlling each file with an independent verification by matching the two files.

The mailing register of the Annual Survey of Manufactures is used for controlling and recording the receipt of correspondence and reports. Furthermore, it is a source of information to identify those companies to be included in a mail followup for delinquent reporting. This file contains sufficient information to facilitate selection of sub-samples from the Annual Survey of Manufactures universe; to study response patterns by selected categories; and to maintain a reference file of delinquent reporting characteristics.

The data register containing the prior year statistics for each company is used in the tabulation of the Annual Survey of Manufactures to control status changes and to provide a base for the editing of current data. After identifying and correcting duplicate and mispunched reports the prior year's register is used in the balancing and editing operations as a stabilizing influence to prevent the insertion of erroneous data. Likewise, for those plants for which no report has been received, it provides the base for imputation. The prior year's data register is also used in the tabulation program as a base year for estimates.

Change from Detail Analysis to Summary Analysis

Computers have provided the means by which it is possible to change from a review of data at the detail level to an intensive analysis of a small number of summaries. It is no longer necessary to construct a model within which the individual report is retained until the final stages of the summarization, but now it is feasible to assume that the raw data can influence the compilation only at the early stages and later can be subjugated to the effect of the cell changes. An excellent example of this method of non-reliance on detail may be found in some of the techniques utilized in the production of the 1959 County Business Patterns. For this publication an edit was made against historical data from the 1956 County Business Patterns and from the 1958 Censuses. If cell totals were found to be inconsistent with anticipated relationships to the historical data, the detail items for that cell were manually reviewed.

In the County Business Patterns tabulations many innovations were introduced or expanded. In the first place, the County Business Patterns were run on three types of computers -- IBM 705; Univac-I; and Univac 1105. The communications between the computers presented a number of problems. The resolution of these provided the experience needed to eliminate this problem for the future. This flexibility in computer hardware is the outgrowth of recent improvements but the ability to accomplish this is becoming more common.

For the 1959 County Business Patterns the computer was instructed to analyze and select a single set of codes and data from five files obtained from Census and Social Security. The processing of these files on the computer through a series of matches, edits, corrections and sorts required several hundred programs.

In order to facilitate the program writing there was developed for Univac-I an input/output framework generator which freed the programer from the necessity of debugging at least the housekeeping half of his program. After corrections were assembled, the more powerful 1105 computer was utilized to provide final tables. A single summary program provided, in three successive passes, county industry tables, state industry tables, and U. S. industry tables.

Capacity to Explore

One of the features which has been developed on the computer in the last few years has been that of experimenting with available data to determine what would be the effect of certain program changes. Foreign Trade statistics have been compiled on the computer for more than four years and these historical data have been frequently utilized in various studies to test theories. Many of the innovations which have been introduced into the Foreign Trade program have been the result of some of these studies.

The replacement of the logical approach utilizing quasi-accepted procedures, by an empirical approach limited only by the imagination and by the availability of data, has released surveys from tradition and allows them to operate within a freer framework. Objectives can then be tested based upon a new standard and an old data instead of the previous reliance on the opposite, namely, old standards and new data.

The limitation of cost can quite often be overcome in the computer by designing a program specifically for a given objective. Many illustrations of this could be easily cited. In the course of the 1958 Census of Manufactures and Minerals when the industry volume tables were being planned, instead of using a full record to process all of the tables it was thought to be more economical to produce a number of individual records each designed to fulfill the requirements of a specialized series of tables. For example, a twelve-digit record was devised which was intended for use in producing two tables of area by industry by employee size statistics. The entire Census of Manufactures universe was contained on two partial tapes, whereas it would ordinarily require 40 tapes. This same record was later used to produce the publication "Location of Manufacturing Plants." Similar economies made it possible to utilize the computer in areas which previously were not deemed efficient.

One of the many problems which exist in producing statistics from reports which are extensive in the relationship of one item to another is the difficulty of coordinating all of the interrelationships. In the 1958 Census of Manufactures it was decided that the entire record should be kept together until such time as the interrelationships had been completely edited. This necessitated making provisions for a maximum of 50 data fields of General Statistics, 250 Product items with 8 data fields for each item, 60 Material items with a maximum of 4 data fields for each item, 50 Special Inquiry items with a maximum of 24 data fields for each item, and a Fuels and Electric Energy item with 15 data fields required. The handling of these records during the course of industry coding and editing illustrates the potential to which the computer may be directed.

Minimization of Pre-Computer Processing of Input

In addition to quantity and quality of data there is a possibility now to extend the computer work both forwards in the process in order to take advantage of the raw data and backwards to deliver from the computer a finished product. The system which merely reflected a transfer of operations from punch card equipment to the computer is now being converted to one which examines the entire process from the beginning to end and allows the computer to be used freely as a tool within the system.

The use of check digits, alphabetic coding, analysis of complex interrelations, editing, nonsense balancing and the like has extended further the utilization of the computer. Also, the output of the computer can now be presented in such a form that with only minor manual editing the computer output may be published as produced. The first successful experiments of presenting data directly from the copy produced on the high speed printer indicated not only the feasibility of this approach but also the desirability and practicability.

The Interest of Management in Extending Objectives. The Dynamic Qualities of Computer Utilization.

It is particularly significant to observe the growth in interest by management in the use and functions of computer operation. Although enthusiasm may be present, there is also some trepidation in the transition from conventional punch card and clerical procedures to computer operations. Apparently a great number of procedures done clerically defy definition. Comments like "this is based on experience," or "we know intuitively that this is wrong," and "actual documents must be examined to resolve problems" are by constant questioning formalized into a computer system. Even so, the initial objectives are not too far extended from the conventional system.

However, once the particular program has been computerized and proves successful, there ensues a feed-back of ideas from management. Initial "fears" vanish and of paramount interest is the question of what else can the computer accomplish.

Illustrative of this feature of feed-back is the monthly Foreign Trade Program conducted by the Census Bureau. Approximately four years ago

each of 400,000 commodity classification codes appearing on incoming documents was verified clerically and each unit of quantity was reviewed for consistency. In addition, approximately 8,000 items which were rejected by the computer for nonconformance to pricing criteria during an edit were manually reviewed and corrected for inclusion in the survey of the following month.

Today, because of management's questioning of clerical processes, approximately 10,000 items are manually coded per month and rejects for pricing are imputed by the computer eliminating clerical intervention and what is more important included in the current month's survey. Problems of this nature are resolved as they are recognized in concert with "statistical specialists" who analyze the situation and reduce the problem to a mathematical formula for computer resolution. In this instance the computer itself contributed in the resolution by evaluating data for an extended period in the past and introduced new levels for pricing criteria.

In the same program, historically, alphabetic information appearing on incoming documents were clerically encoded in accordance with two coding manuals. Effective in January 1962 this coding operation will be performed by the computer by use of 4 alphabetic characters from the document to express country and 5 characters to express foreign port. In essence these are the first characters of the country and the foreign port name. Today this function is successfully accomplished on 10% or 40,000 records. Once again it was the result of application of "specialists" cognizant of the problem, and the computer utilizing past data to analyze and assist in the preparation of a translation matrix. The implications for the future indicate the end is not in sight. Document "scanners" open a new vista for the computer.

Utilizing current and past data the computer has been developing statistical aides to be used as a guide in our current tariff dilemma and in the management of import quotas in the area of cotton and cotton textiles. These special surveys are a few of the illustrations of "feedback" of data produced by the computer to be utilized in a manner not initially contemplated.

Scheduling of Operations, Programing and Production on the Computer

Conceding that an efficient working programing organization exists, that the preparation of computer programs poses no problem; one is faced with the final hurdle of getting results.

In a unit there is more than one project entailing many computer programs being processed at the same time. Some order must be maintained; this implies the establishment of a unit whose sole responsibility is to maintain order. It is the responsibility of the head of this unit to evaluate the number of hours required, usually derived from experience or in consultation with the executive of the programing unit and to schedule computer time based on the priority needs of management.

Having scheduled time, work must be set up in advance and personnel made available at the computer to perform the necessary tasks. This is accomplished by following project operational flow charts and operating instructions previously prepared by the systems analyst and coding programer. Attendant also is the maintenance of records of computer time, controls to insure data has been processed properly and the like.

In conclusion, the solution of problems in existing surveys increases the potential of applying the successful techniques to newer surveys or to those not yet programed for the computer. The limitation of the computer a few years ago has diminished. Each year a different set of criteria is used to provide the basis for determining the range of computer applications. Automatic programing such as COBOL is a new feature which has been introduced and eventually should overcome some of the existing shortages of trained personnel. The acceptance of the computer as a powerful tool by educational institutions and businesses, as well as government has created the desire to further exploit this tool. Tomorrow we should be in a position to look back on today and sympathize with our present limitations.

APPENDIX A

ECONOMIC CENSUSES AND SURVEYS CONDUCTED ON THE COMPUTER AT THE BUREAU OF THE CENSUS

	COMPUTER HOURS	
	Used During Period July 1958-Nov. 1961 (Univac I & 1105 hours)	Estimated for Period Jan. 1962-Dec. 1966 (1105 hours)
Censuses of Business, Manufactures and Minerals		
Basic Censuses	18,000	22,500
Business Supplemental Program	800	
Industry Supplemental Program	300	
Concentration Statistics	170	
Central Business Districts	240	
Company Statistics	450	
Census of Governments	20	1,000
Current Business Program		
Monthly Accounts Receivable Survey	900	700
Monthly Retail Trade Report	1,800	1,300
Monthly Wholesale Trade Survey	800	600
Other Business Surveys	75	300
Business Trust Funds	650	-
Current Industry Program		
Annual Survey of Manufactures	2,700	2,200
Monthly Industry Survey	90	1,000
Annual Lumber Survey	120	1,000
Other Industry Surveys	60	5,500
Industry Trust Funds	650	-
Foreign Trade Program		22,700
Exports	4,600	
Imports	4,400	
Shipping	5,300	
Exports of Manufactured Products	100	
Other Foreign Trade	300	
Foreign Trade Trust Funds	1,600	
County Business Patterns	4,000	2,500
Construction Program	50	1,100
Miscellaneous	-	7,800