DEPARTMENT OF INTERNATIONAL ECONOMIC AND SOCIAL AFFAIRS

STATISTICAL OFFICE

WORKING PAPER

THE USE OF MICROCOMPUTERS FOR CENSUS DATA PROCESSING

UNITED NATIONS
New York, 1989

PREFACE

The present publication contains a description of the use of a new computer technology, microcomputers, for processing population census and housing data.[1]

This document has been prepared in response to the continuing concern of the United Nations Fund for Population Activities and the United Nations Statistical Office to provide appropriate and cost effective advice to developing countries in the preparation for and execution of their censuses of population and housing. The rapid advances of computing technology in the last decade, especially those emerging in the area of microcomputer technology, provide significant new opportunities for developing countries to improve substantially the effectiveness of their statistical data processing operations. Such improvements are capable of being realized in all phases of operation, from planning and management through data processing to data and analytical services that can now be delivered to end users of the data.

The publication consists of seven chapters and an annotated bibliography of further references in the subject and related areas. Chapter 1 is an introduction to the paper and various aspects of the subject material. Chapters 2 and 3 focus upon the technological factors that are leading to the growth of microcomputer based markets, the substantial decentralization of functions that the technology encourages, and the redistribution of responsibilities and requirements that these shifts produce. Chapter 4 discusses the application of microcomputers to specific areas in census data processing, with emphasis upon the functionality of the software that should be used in specific areas to obtain maximum assistance from a computer based processing strategy. Chapter 5 raises a number of new issues that arise in introducing microcomputers into statistical and other data processing activities. Chapter 6 provides a glimpse into a number of areas in which microcomputers may be useful in the future, and which are likely to be of some importance soon for national statistical offices. Chapter 7 provides a concluding summary of the document.

The work for this and a complementary companion paper[2] were sponsored by the United Nations Fund for Population Activities.

[1] This document has been prepared by George Sadowsky, who is currently Director of Academic Computing and Network Services at Northwestern University, Evanston, Illinois, U.S.A., and was formerly a staff member of the United Nations Statistical Office as Technical Adviser in Computer Methods. The opinions expressed in this paper are those of the author and are not necessarily those either of The United Nations or of Northwestern University.

[2] International Statistical Programs Center, *Population Census Data Processing on Microcomputers*. Prepared for and available from the United Nations under project INT/88/P09. U.S. Bureau of the Census, February 1989.

CONTENTS

Page

The Use of Microcomputers for Census Data Processing

1. INTRODUCTION

Background

The year 1990 marks the one hundredth anniversary of the first application of automatic computing equipment to population censuses in the world. In 1880, frustrated by the long delays inherent in processing manually census data collected from a population of then about 45,000,000 persons, the U.S. Bureau of the Census commissioned Herman Hollerith to design and construct data processing machinery capable of tabulating census results rapidly and automatically. Hollerith's equipment was available in time for the U.S. 1890 census, and was successful in reducing the overall processing time by approximately a factor of four.

Hollerith's electromechanical accounting equipment of 1890 evolved in complexity, capacity, and speed, and found increasing uses in other areas of government and business as well. A major step forward occurred in the 1940's with the invention of the electronic digital computer. Operations that took seconds on electromechanical processing equipment now took thousandths of seconds on the newly invented all-electronic equipment. The first commercial computing company, the Eckert-Mauchly Computer Company, was founded in 1947 and was absorbed by the Remington Rand Corporation in 1949. So important was the population census that the first unit of its first computer model, a Univac I, was delivered to the U.S. Bureau of the Census in 1951 and was used to process the 1950 U.S. Decennial Census of Population.

Today we take for granted the use of electronic computing equipment not only for processing census data, but also for processing data from many large scale statistical and business operations. Much of the world's administration, research, and commerce depends critically for success upon current large scale computing equipment that is directly descended from the pioneering developments of the 1940's. The resulting computing techniques can provide substantial assistance to developing countries in improving the quality of governmental administration, the rate of economic development, the effectiveness with which scarce national resources are managed, and the adequacy of the statistical information base for policy planning and implementation.

Major changes have taken place in electronic computing technology since the 1940's. Computer *hardware* has become smaller, more reliable, and less expensive each year; over the entire period, the performance-price ratio has increased by about 25-30% each year. Likewise, computer *software*, i.e. the programs and user environments that instruct the hardware to perform useful work, has evolved substantially in the direction of ease of learning and use. Such evolution has made possible the widespread introduction to and use of computers by individuals who are not computer professionals and who have little if any technical training.

Prior to the 1980's, statistical data processing was performed on either *mainframe* computers or *minicomputers*. These computers were large, complex, and expensive. In addition to their large capital cost, they required substantial investments in site preparation and operation in order to function properly. They also required staff with professional skills in computer operations, and system and application programming in order to exploit their capabilities usefully.

In the late 1970's and early 1980's, a different computing environment began to emerge, one that was qualitatively and quantitatively very different from its predecessors. *Microcomputers* are characterized by small physical size, relatively low cost, tolerance toward a broad range of environmental conditions, and a lessened dependence upon specialized technical personnel. Microcomputing today is often characterized as *desktop computing*, reflecting the fact that microcomputers are sufficiently small to occupy only a fraction of the work surface of a typical office desk. The method of use of microcomputers is different; it provides *personal computing*, in which the new computer systems are oriented to serve one person at a time, employing a user interface often more powerful than what was previously available on microcomputers or minicomputers.

The rapid evolution of microcomputing technology and its increasing importance within the overall spectrum of information processing is much more than just a continuation of the overall technical progress in computing technology. Size, cost, reliability and convenience have reached the point where there has been a very significant qualitative change in the nature of computing services available. The emergence of a mass market in hardware and software and the resulting explosion of entrepreneurial effort and products, as well as the emergence of radically different and more productive computer interfaces and environments, can be regarded more as a fundamental revolution in information technology than as a continuation of earlier trends.

Objective

This document addresses the nature of this revolution and its implications in particular for the use of microcomputers for census data processing in developing countries. Much of this material is also applicable to other forms of statistical data processing, and a part is applicable to the use of microcomputers by anyone, either in

INTRODUCTION

developing or developed countries, for any of a wide range of applications.

The primary objective of this document is to provide a guide to experience in the use of microcomputer technology for the processing of censuses of population and housing in developing countries. The primary audience for the document consists of administrators and managers of statistical activities who may wish to exploit this technology, and the text is written with the intent of providing a view for them of the managerial and operational implications of such use based upon technical and product alternatives that currently exist.

There is a substantial secondary audience toward whom this document is also addressed, consisting of technical and professional staff members who may be responsible for detailed implementation of microcomputer-based statistical activities, including censuses of population and housing. In practice, these two groups must work together in order to implement successfully any large scale statistical operation such as a census, and their spheres of knowledge must overlap to some extent so that each group will have an appreciation of considerations affecting the overall processing strategy. In order to address the concerns of the second type of reader as well as enrich the technical knowledge of the first, specific technical material and technical elaboration of text appears for the most part in footnotes throughout the document.

This document is a companion to the document, *Population Census Processing on Microcomputers*, produced by the International Statistical Programs Center of the U.S. Bureau of the Census. That document is an extension of the POPSTAN series, which provides a thorough and comprehensive operational treatment of census activities in a hypothetical country. In contrast to the ISPC document, this document does not provide a methodical or a cookbook approach to census data processing activity; rather, it attempts to reflect actual experience gained in developing countries that have used microcomputers for census data processing as well as for other statistical activities.

It is important to note that this version of this publication has been written in early 1989. Its contents therefore reflect the state of computing technology, both hardware and software, at that time, as well as the perceived state of census data processing knowledge and practice then in developing countries. The material contained in this publication is therefore, like much information in computing and data processing, subject to relatively rapid decay. In a few years, the products and other options available are likely to combine to make this publication increasingly obsolete.

Developing countries are a relatively heterogeneous group, each having strengths and weaknesses that affect their rate and direction of development. The present document attempts to span a wide variety of skills and needs present in developing countries, with the knowledge that in its entirety, it is not applicable directly to any one such country. Nevertheless, it is hoped that much of the content will be useful for obtaining an effective and operational understanding of the role that microcomputer technology can play in census data processing, and the various ways in which developing countries can harness this technology to promote their interests. In this regard, this document provides a technical and organizational overview of the new technology and its implications, and a context within which countries can understand the alternatives available to them and make informed choices among them.

Finally, this document tells only a small part of a much larger story involving statistics, census data processing, and computing products and techniques. An annotated bibliography is provided at the end of this document, with annotations providing some guidance regarding the content and relevance of material in each of the references.

Background

The microcomputer environment available today is the result of technical progress that started in the mid-1940's with the development of the first electronic digital computer. In the intervening four decades, electronic data processing (computing) equipment has become affordable and usable by most if not all national statistical offices in the world. This progress has been achieved largely due to a rapid and sustained rate of technical progress in the underlying electronics technology coupled with continuous intellectual and commercial efforts to make computing equipment easier to use.[1] The cumulative effect of technical progress over 40 years has driven the cost of raw computational power and capacity to very low levels, while advances in software and the now emerging mass market for computer based tools has caused software prices to fall almost correspondingly.

Two major components contribute to the power and utility of computing equipment, hardware and software. The *hardware* comprising a computing system consists of physical equipment including electronic circuitry, magnetic storage media, input and output devices including devices for user interaction such as a keyboard and screen, and devices for communication using external transmission facilities to exchange information with other computing systems. The *software* that is executed on a computer system consists of the specific instructions that tailor the computer to be useful for specific tasks, such as data editing or data tabulation.

Hardware is often characterized by its generation. First generation hardware was prevalent in the 1950's and consisted of circuits based upon vacuum tube technology. As a result, first generation computers were very large and expensive, required very large amounts of electrical power, and failed frequently. Second generation hardware in the early and mid-1960's was characterized by the use of discrete transistor technology, while third generation technology was characterized by some initial small scale integration of components in the fabrication process.

Given the substantial progress in electronics technology, it is difficult and perhaps meaningless to categorize current hardware in terms of precise generation numbers. Each successive significant identifiable improvement has led to equipment that is smaller and less expensive, consumes less power, and is more reliable than its predecessors. However, a very major step forward occurred in the early 1970's with the invention of solid state integrated circuits that could be reproduced automatically and in large quantity by precise photographic techniques, replacing the predominantly manual assembly methods that had characterized prior computer manufacture. In particular, such solid state circuits could be designed that would include all circuitry necessary for a complete computer processor; such circuits are called microprocessors or microprocessor chips. Similarly, memory circuits and circuits for other computer hardware tasks were designed and have provided recent and current computer designers with a wide range of miniature circuits on chips from which to construct computing equipment of all sizes.

While advances in electronics technology have benefitted all sizes of computing systems, the emergence of microchip technology have made possible economical and useful small scale computing. The *granularity* or minimum size in which computing power can be obtained has changed radically; while in the 1960's it was not possible to purchase complete useful computing systems for less than US$250,000, complete systems having far more power can now be purchased for US$3,000. These inexpensive computing systems, based upon microelectronic chips including microprocessors, are generally referred to as *microcomputers* or *microcomputer systems*.

Microcomputers are best characterized by their physical size rather than by the functionality or power of computing that they provide. Microcomputers are generally placed on a part of a desktop, although sometimes the main unit may be placed on the floor or on a shelf nearby. Microcomputers are often modular, i.e., they may consist of separate processor, display, and keyboard parts, but they may also be integrated. Most microcomputers are meant to be used by one person at a time, but there are now multi-user microcomputers that are starting to have an impact on statistical data processing operations.[2]

It is more difficult to characterize what a microcomputer is by its power or functionality, since current microcomputers now often have as much computing power as minicomputers and mainframes that were considered quite powerful machines ten or more years ago. Technical progress in the underlying computing and electronics industries affects all scales of computing machinery, so that the power and functionality of any scale of computing is destined to migrate to smaller scales of computing over time.

[1] While there are many measures of rates of change in the underlying technology, a good rule of thumb in the past has been to assume that the cost/performance ratio of computing equipment will be declining by approximately one-half every two years.

[2] In particular, variations of the Unix operating system have been implemented that run on larger microcomputers and that provide a time-sharing environment on an inexpensive hardware platform. Although such systems are not in common use at present for statistical data processing, the benefits of shared computing environments that make shared minicomputers so useful are equally applicable to the microcomputer environment. Multi-user microcomputers are therefore likely to play a larger role in the future of statistical and census data processing than they do in the present in both developed and developing countries.

MICROCOMPUTERS AND COMPUTING IN DEVELOPING COUNTRIES

Low cost is a characteristic of microcomputer technology that has existed since its beginnings and that is likely to continue. Microcomputers are inexpensive relative to larger computers; their relatively low cost makes it economically practical for them to be used more in a role of assistant to individuals rather than being regarded as a scarce resource that must be kept constantly busy to justify its economic existence, as was generally the case in the past.

Microcomputers are physically compact and relatively light, and can therefore be transported easily. In addition, microcomputer technology is generally not at the frontier of technical progress or extremely fast or capacious, making it relatively easy to obtain export permission and licenses for microcomputers and related equipment.

Microcomputers may be connected in such a way that they can exchange programs and data electronically. The mechanism by which they perform the exchange is called a *local area network*, or LAN. The adjective *local area* refers to the geographic compactness of the network; generally such networks are installed within the same room or set of adjacent rooms in a building, and they are almost always contained within the same building. They are meant to serve a specific work group, consisting of persons working on the same or interrelated tasks. The use of a local area network allows the sharing of resources and data by all microcomputers attached to the network and make statistical data processing operations more effective, although it is not a necessary component of using microcomputers.

Local area networks generally connect a number of smaller microcomputers to a larger microcomputer which functions as a *server*. This system provides services to the other computers. It is likely to have a large fixed disk, and so can provide file access and storage services for the other systems on the network; if so, it is referred to as a *file server*. It may also have a printer attached to it that can be shared via the network connection by the other systems; it can then be referred to as a *print server*. The servers are the primary mechanism by which resources on the network are shared.

Because of the importance of the services provided by such servers to the rest of the network, sometimes two such servers are included in a network configuration so that if one fails, the other can take over and provide the services required to keep the network operational. Such a configuration is somewhat similar to having a spare central computer that can be pressed into service almost immediately should the primary central computer fail.

For the purposes of statistical data processing activities, networks characterized by moderate transmission speed will generally be sufficient. Such networks are moderately easy to set up and operate, and are increasingly easy to obtain export licenses for, for shipment to most countries.

The small size and cost of microcomputers allow data processing tasks to be distributed in a way that was not possible with larger computer systems. Computing power can be decentralized geographically, since each microcomputer can be made physically independent of others. Computing power can also be specialized; specific microcomputers can be dedicated to specific functions rather than requiring a large central computer to perform several different functions. Perhaps most important, demands for additional computing power can be satisfied by acquiring more microcomputer systems rather than having to replace a large computer system with a larger one.

Parallel changes in software technology are equally important. Software consists of instructions, written by programmers and users, that computing hardware executes in order to perform tasks. These instructions are organized into modules commonly called computer programs. These programs are executed quickly (at electronic speeds) and unforgivingly; an error in the set of instructions comprising a program will be executed as specified at the same electronic speeds as the correct sequence of instructions would have been executed. One of the major challenges of software technology today is to develop methodology that will allow the construction of correct programs efficiently.

Early software technology was rudimentary and expensive; software specialists called programmers were required to instruct computing systems in a language closely corresponding to the circuitry of the hardware on which the program would be executed. Over time, intermediate layers of software have been created which have allowed users to specify programs and procedures to computers in a form more closely resembling the task to be performed. The most important of these intermediate layers are: (1) the operating system, which controls the hardware resources of the computer directly; (2) language processors, which allow programs to be written using language constructs reflecting the logic of the program rather than the structure of the hardware; and (3) the user interface (including documentation), which allows less technical users to control the computer directly.[1] For most people familiar with computers today, the characteristics of the user interface and the applications for which programs exist are the most important aspects of a computer system.

Because of their small size and relatively low cost and their use in playing games, microcomputers are sometimes regarded as frivolous and not serious alternatives to earlier

[1] The operating system and the user interface often overlap. An example is provided by Apple's Macintosh operating system, which hides the basic operating system as much as possible under a graphic, icon-oriented user interface that is easy to use and subsumes functionally many operating system characteristics. A somewhat different approach is being taken by IBM/Microsoft in their development of OS/2, in which the basic operating system is readily available, but the Presentation Manager layer is available for users who prefer to be a level removed from operating system complexity.

forms of computing. This was an attitude adopted in the early days of microcomputing by many computer professionals who were then oriented strongly to using large computers, and who in addition perceived the breakdown of central control over computing resources and computer based operations that microcomputer technology would eventually lead to.

It is extremely important to recognize that microcomputers are not frivolous and are not toys; nothing could be further from the truth. Rather, they are the leading edge of a fundamental set of changes in information processing technology that will result in a widespread distribution of computing power in much the same way that the development of electrical power led to a widespread distribution of mechanical power and energy.

Microcomputer technology is in many respects as powerful as earlier mainframe computer technology. Microcomputers can execute instructions at very rapid speeds, and the newer models can accommodate substantial immediate access or primary memory. Microcomputers are less adept currently about providing substantial mass storage. They are also more limited in the area of shared use, which occurs when a group of users requires concurrent access to a shared resource. Although development is occurring in these and other areas, for many organizations residual deficiencies in current microcomputer environments will dictate an effective data processing strategy including the use of both microcomputers and larger computer technology.

Both computer hardware and software are evolving and developing at a rapid rate. Even though today's microcomputers are by many standards the equivalent of much larger computers of the past, they will almost certainly continue to develop at the same rate in the foreseeable future as they have in the past. This expectation has a number of implications for developing countries in the areas of technology exploitation and training.

More than any other form of computer technology, the commercialization of microcomputer systems has created a mass market and a consumer market in computing hardware and software. The rapid growth of such a large market has led to a large and diverse low end computing industry, characterized by a large set of products, historically low prices, and rapid market response to consumer demand. The consumer surplus generated by this growth has been enormous, and has benefitted microcomputer users in all sectors and in all countries.

Microcomputer software on the whole is significantly better in its orientation toward users compared with mainframe software. Because it is meant to be used by people who are generally not computer professionals, much more effort has been invested by software producers to ensure that their software is much easier to learn and easier to use. The term *user friendly* originated with microcomputer software. Prior to microcomputers, most software was used by data processing professionals, and power and functionality were generally more valuable than ease of use. With the growth of microcomputer technology and the mass, non-professional user market, ease of use often became the deciding factor between success or failure of a software product. Software producers quickly understood this characteristic of their marketplace, and much of the software available today is easy to learn and easy to use by people who are not computer professionals. A very important result is that the time necessary to train people to use microcomputers is lessened because microcomputer based software is easier to use.

Computing in Developing Countries

Computing in developing countries has historically lagged the application of computing in developed countries, often significantly. Much of the lag can be explained by the fact that, until fairly recently, computers have represented advanced technology requiring an infrastructure and a body of knowledge that was being created with some difficulty even in the developed countries. The emergence of microcomputer technology signals a fundamental shift in the nature of the technology and the infrastructure required to support it, and has the promise of shortening the technology transfer lag substantially.

It is useful to examine this issue within the context of the factors affecting the rate of successful absorption of computer technology within developing countries. Important and relevant factors include: (1) physical infrastructure; (2) human resources infrastructure; (3) size of market; and (4) access to information resources. Such an examination is useful in spite of the fact that developing countries as a group are quite heterogeneous, differing widely in the extent to which they have introduced computer technology, and the extent to which the necessary infrastructure exists for exploiting it. As a consequence, for almost any statement one can make on this subject, there is highly likely to be a group of countries for which it may be either not meaningful or not applicable.

Physical infrastructure. The degree to which physical infrastructure is important in exploiting computing technology is now decreasing somewhat in large part due to the introduction of microcomputer technology. Mainframe and minicomputer systems require a reliable supply of clean electrical power and reasonably stable environmental conditions such as temperature and humidity levels in order to operate reliably. The developing countries' dependence in the past upon these larger systems created major difficulties in installing computers and keeping them in operation. While large computers still require controlled environmental conditions, with some exceptions microcomputers can operate under a wider range of environmental conditions, making the elaborate and expensive site preparations of previous years largely unnecessary today.[1]

Almost all developing countries now have the necessary physical infrastructure to support operations of one or more types of computer systems, at least in one or more principal cities. Where electrical power is unreliable, power regulators and/or standalone generating capacity may be required for larger computer systems, making use of such systems more expensive than it otherwise would be. Power requirements for microcomputers can sometimes be met under such conditions by systems with built-in battery power, or with ordinary motor vehicle storage batteries providing an electrical buffer between an unreliable public power or intermittent private power source and the computing installation. While such arrangements involve some inconvenience, use of the technology on at least a limited scale is still viable.

Transportation and communications links to suppliers of hardware and software resources are also required for effective use of computers. Where communications and transport links with foreign countries are weak, some compensation in the form of additional redundancy of equipment, a larger spare parts and supplies inventory, special communications equipment and additional travel for education and obtaining information may be important. Such operations can add to the expense of using computers, but they make possible their use.

The use of microcomputers has alleviated this situation to a considerable extent, since they are generally used in groups rather than singly. Providing additional equipment redundancy can be easily accomplished by providing additional systems; these systems can be used productively when their constituent parts are not needed as spares. The burden of providing such redundancy is highest for small countries which are relatively isolated; their need for capacity is small to begin with, and they must have sufficient backup because of the cost and time delays in servicing their needs from outside the country. Small and remote countries, such as island nations in the Pacific, are probably most affected by such considerations. However, countries for which transportation, importation or exportation of goods, or communication are difficult for other reasons may be affected equally or more severely.

Lack of adequate physical infrastructure is rarely a major problem today, although remedying deficiencies in it will generally increase the costs of providing computing capability to a country. Providing a stable and reliable supply of electrical power is perhaps the most common problem, which often requires substantial additional effort and cost to solve. As overall economic development occurs, the ability of developing countries to provide an adequate physical infrastructure for computing activities should improve at a moderate rate.

Human resources infrastructure. The skills required to exploit computing technology effectively depend upon minimal levels of competence within the country in fields such as engineering and mathematics. The level of literacy among the most educated people may be an important factor. In some countries, the ability to work in a foreign language used by suppliers of computing equipment and software may be an important determinant of success. Aspects of the cultural milieu within the country may also be important. Responsiveness to educational opportunity, the strength of the work ethic, and attitudes and policies toward achievement, employment, production and productivity are all important in the successful transfer of this technology.

In general, a sufficient human resources infrastructure exists in most if not all developing countries to be able to make use of computer based systems. However, it is often thin, and there is substantial competition for the scarce more talented individuals within both the public and the private sector as well as between them. Emigration to better labor markets, a phenomenon sometimes referred to as the *brain drain*, causes a depletion of the resources necessary to exploit this technology; in countries initially having a limited set of human resources with which to work, such a drain may cause data processing activities to be in a state of suspense for periods of time. The lack of priority given to computing in many developing countries, along with difficulty in developing personnel policies adequate to attract and hold qualified specialists in the public sector hurts the development of the human resources infrastructure.

While the allocation of human resources between sectors and objectives is an internal matter reflecting a government's priorities, external assistance can be used to increase the supply of trained manpower in computer-related skills. If the infrastructure is robust, there will exist individuals who can benefit from specialized training abroad. More often, however, the primary need is for more basic intensive training of larger numbers of people within the country. Such training is often necessary to obtain any degree of effective use of a computer system.

In many countries, the need for such training is not adequately met at the present time, in part because it is a labor intensive activity requiring skilled manpower. Given the limited budgets and multiple objectives of technical assistance agencies, both the costs and the opportunity costs of such training are perceived as large. Nevertheless, it is a critical investment for long run effective development of computer based activities that affect not only statistical of-

[1] Microcomputers are surely hardier and more tolerant of environmental extremes than larger systems; however, some conditions jeopardize their successful operation. Extreme humidity, for example, may cause condensation on circuit boards. A coastal salt water atmosphere may result in salt deposits on computer components and media, rendering data irretrievable. Static electricity is a hazard to integrated circuit components in environments with inadequate humidity.

fices but also many other activities important for national development.

Fortunately advances in both hardware and software technology have made it easier for developing countries to leverage their existing human resources. Microcomputers are easier to install, operate and maintain than mainframes or minicomputers. The newer microcomputer application packages are much easier to use than their previous mainframe counterparts and cost less to acquire as well. Thus, persons having limited technical skills can be used to staff microcomputer based activities whereas it would have been difficult if not impossible to have used staff with similar levels of technical capability in the past.[1]

Size of market. The current and potential size of the market for computer systems has been an important determinant of the potential profitability of entry into the country by suppliers. In the past, countries that had limited markets generally received attention from one supplier; some countries had no suppliers of computing equipment and had to depend upon sales, support, and service to come from another country. Within countries having limited markets for computing equipment, the price of conventional equipment supplied in the past has been quite high.[2] Small markets have discouraged suppliers from making documentation available in a special local language, which has in turn limited the number of people who can locally learn how to use the supplier's products. Countries with relatively high rates of growth attracted more than one supplier and generally enjoyed the benefits of a more active domestic computer industry, including product variety, more competitive markets, and greater responsiveness of the market to demands for specific products and services.

The introduction of microcomputers into developing countries has modified the situation quite substantially. Microcomputer systems are in general relatively inexpensive, easily shipped, and easy to maintain. The cost of entry into a microcomputer business in most countries is substantially lower than for a mainframe or a minicomputer company. Even in the absence of a local supplier, such units can be imported easily and, with some special training, maintained by lower level staff than would be necessary for larger systems.[3] Because the cost of microcomputer based systems is so low, and because many computing tasks are better done using microcomputers, developing countries with limited mainframe computing markets now enjoy more sizeable microcomputer markets. Coupled with the low barriers to entry faced by suppliers, microcomputers have proliferated in such environments and have taken over significant segments of the computing marketplace.

The most serious constraint upon computer hardware acquisition have been and continues to be the necessity for and availability of computer hardware maintenance services. At the mainframe and minicomputer level, some developing countries have sufficient engineering talent to be able to dedicate one or more national staff to be trained as maintenance engineers, but most do not have such resources. Furthermore, training is generally expensive and can result in locking a country into a particular computer line for some time. Further specialized training will then be necessary with every significant shift in computer hardware. At the microcomputer level, hardware maintenance is considerably easier and cheaper, and can be handled in large part by means of redundant hardware capacity.

In the less developed countries, the constraint imposed by the necessity for maintenance used to favor large, established manufacturers and the more conventional of their product offerings, initially standard batch processing systems of moderate size and later gaining some limited interactive access. The introduction of the microcomputer and the economics of that part of the computing industry should allow them to leapfrog over some phases of the historical development of computers and utilize the more modern and productive techniques that are becoming common practice in developed countries.

Information resources. People living in developed countries live in an information rich environment. A rich variety of information is available daily from newspapers, magazines, books, government publications, technical and scholarly journals, radio, and television. Information is further available from friends and colleagues, no matter whether they are local or distant. Reliable telephone systems provide relatively economical voice communication between most or all parts of the country. Emerging electronic mail systems connect professionals of many kinds together in an effective manner for passing written messages;

[1] While it is true that maintenance of microcomputer equipment is easier than maintenance of previous generation equipment, such maintenance does require a minimum basic understanding of technology sufficient to repair basic mechanical and electronic components. Microcomputer technology is not quite yet at the "plug and play" stage achieved by some consumer goods and electronics industries.

[2] Such high prices for conventional equipment have in the past been cited as examples of exploitative behavior, typically by multinational companies in their dealings with developing country economies. While such exploitative behavior cannot be denied, a sufficient reason for higher than average prices at that time was the high level of support costs incurred by the supplier in servicing a limited market. There are economies of scale in supporting mainframe computers in compact geographic areas, and markets in many countries were too limited to realize such economies. Limited markets also were conducive to the perpetuation of monopoly power, and therefore monopoly pricing, by sole suppliers in a country.

[3] The ability to import microcomputers even in the absence of local sources leads to mixed results, since doing so is likely to retard growth of a local commercial infrastructure that can supply and service microcomputers. Such an infrastructure is a desirable outcome of the introduction of microcomputer technology, especially since the probability is larger than with mainframe systems that local firms can compete successfully in this market. The most desirable outcome of the introduction of microcomputer technology -- both for statistical offices and within a context of national development -- would be the healthy growth of an active local industry of microcomputer sales and support services, including systems integrators, software specialists, and training facilities and programs.

such systems are growing in number and in size to cover larger subscriber bases and provide wider geographic coverage.

The rapid rate of technical progress affecting the computing industry, coupled with the growth of personal computing based on inexpensive microcomputers, has caused the computing industry to be one of the most information intensive in the world. In developed countries, computing magazines, electronic bulletin boards, computer clubs and user groups, and computing courses and training opportunities have all proliferated widely. The market for these goods and services is thriving in response to a large demand for information which makes it possible to apply the the technology to a wide variety of disciplines and applications.

Developing countries are limited in varying degrees in their ability to access this rich flow of information about technical issues, products, and reports of experiences. Delivery of printed material is delayed because of distance; international air mail rates are high. Foreign exchange may be limited and not available for such material. Local postal delivery systems may be unreliable, so that printed matter may not reach its intended recipient. Information in the national language may be limited. Further, the ability to comprehend such material and fit it into a specific environment may be impaired by a lack of expertise in the country; reasonable questions that could be answered with a little effort in developed countries may go unanswered in developing environments, blocking further exploitation of a technique, product, or idea.

Even with the problems within developing countries of obtaining access to rich sources of information about computer technology, the current situation is still considerably better than it was when countries had one or just a few computer installations and probably only one supplier of equipment. In the past, knowledge about the computing industry and computing techniques most often came directly from the local supplier of computers; it was in the supplier's interest to promote knowledge associated with that supplier's products and approach to computing. At that time only a limited number of people were qualified to work with computers, and foreign training, which was another source of knowledge, was expensive and limited to a few individuals. Access to knowledge is surely better now, with multiple and relatively inexpensive sources of information. This benefit is derived directly from the emergence of the mass market for inexpensive microcomputer systems.

Information poverty is still one of the more important and insidious obstacles in developing countries to effective exploitation of information processing and other types of technology. The lack of adequate information regarding developments in other countries and other environments is often not noticed; in the absence of new information, old techniques and procedures are continued without conscious knowledge of alternatives.[1] And, while developing countries may not be hurt in an absolute sense by lack of information, they are certainly negatively affected by any relative measure.

The issue underlying information poverty is how to overcome it to provide a sufficient flow of information into developing country environments so that individuals and governments can make good choices from among realistic alternatives. There are a number of approaches that could assist countries in obtaining a richer and more up-to-date flow of relevant information for statistical data processing. They include: (1) stronger partnerships with national universities and similar organizations; (2) greater exploitation of both short and long term visits by experts from other countries; (3) effective sharing and exploitation of locally available technical skills and knowledge; (4) establishment of an informal statistical data processing newsletter; and (5) recognition of the importance of release time for continued on-the-job learning.

Microcomputers and developing countries

From its inception, microcomputer technology has been appropriate technology for developing countries. Their size, cost, modularity, ease of transport and environmental requirements, as well as the variety and power of user-oriented software, make them appropriate for the smaller and simpler environments found in many developing countries.

Support requirements for the mainframe and minicomputer systems that were previously installed in developing countries were significant, and computer suppliers would generally offer products that were somewhat older and more well understood than the leading edge products being introduced in developed countries at the same time. Such a strategy worked well in part also for the developing country; the equipment and software supplied was well understood and more stable than when it was first introduced. Suppliers could afford to offer and support such systems far from their geographic areas of technical expertise because the systems were well understood and trouble free relative to newer models.

With the rapid development of the microcomputer industry, the minimum practical lag between the introduction of current technology in developed and developing countries has been compressed to a matter of months instead of

[1] Readers who wear eyeglasses and whose vision is changing may feel some sympathy with this condition. Deterioration of corrected vision is generally sufficiently slow that it is not noticed *until* new eyeglasses with improved correction are first worn. Often the effect of first using such new eyeglasses is startling; objects and details snap into sharp focus and the wearer is amazed that such improvement is possible. With respect to information, the effect is similar; it is often difficult to understand the losses due to not having relevant information until you have it, and only then is it possible to assess the opportunity cost of not having had it.

years.[1] This lag no longer depends very much upon the ability of the original manufacturer to support its products in a specific geographic region, since support can now often be delivered either within the country or with the assistance of a third party, perhaps an international aid agency, either locally or remotely. Instead, the lag depends more upon the ability of trained personnel in the country who are able to exploit the microcomputer systems and train others to do so also. This function can initially be provided by foreign expertise and later, as the results of training accumulate, by local expertise. For the first time, emerging computing technology can be placed at the disposal of developing countries almost as soon as it is readily available and understood in the developed countries.[2]

[1] This argument is of course valid only if the funds exist to acquire the equipment in the first place. Assuming however that funds *are* available, the range of appropriate information technology products that can now be supported within developing countries is relatively larger than it has been in the past because the newer technology requires less dependence upon the original supplier of the equipment.

[2] Progress in the computer base for statistical data processing depends upon advances in hardware and software, the rate of technical innovation, and the rapidity with which these advances are translated into marketable products at affordable prices. The rate at which the technology is exploited within any given environment is another matter, and reflects the status of the local infrastructure, especially the human resources component, and other conditions between countries.

Overview

Microcomputer technology is appropriate for many areas of statistical data processing, including census data processing. As the capacities and speeds of microcomputer systems increase, an increasing proportion of statistical data processing activities become tractable by applying this new technology. In addition, the personal interactive nature of the microcomputer and available software for it make it in many respects a qualitatively better computer platform on which to base such processing activities. Microcomputers can be used effectively for data collection, editing, tabulation, analysis, publication, control, and management. This technology provides the potential for a fundamental breakthrough in the effectiveness of many statistical data processing operations.

Nevertheless, microcomputers are not a panacea for data processing problems, just as the older mainframes did not solve all problems. Microcomputer technology brings with it a new set of responsibilities which are not immediately obvious to an initial user. It is necessary to understand these responsibilities — as well as the new opportunities offered by exploiting the technology — and develop an appropriate strategy for dealing with them. This section discusses these issues within the context of statistical data processing and operations in general, and is directly applicable to census data processing activities.

Responsibilities of Microcomputer Operations

The introduction of microcomputer technology in an organization implies a significant transfer of responsibility from a central computer support staff to individual users of the microcomputer systems. The transfer of responsibility has both benefits and costs for the user and for the organization.

In organizations that use mainframe and minicomputer systems, both the suppliers of the equipment and the central computer support staffs provide a variety of functions that may be partially or totally invisible to the users of the system. These functions include:

- Performing a *requirements analysis* for either the specific dedicated activities or for the organization as a whole and specifying the characteristics of the system to be acquired;[1]
- *System specification and negotiation with suppliers*; evaluation of responses to the request for proposals to supply hardware and software components;
- *Site preparation*, including designing and constructing or adapting the physical site for the computer system, supplying electrical power and possibly backup power as required, and providing air conditioning and other environmental requirements;
- *Installation and testing* of the system, including co-ordinating and assisting in delivery, installation, and acceptance testing;
- Providing *system programming* skills and services, generating the operating system and updating it periodically, and installing application program packages;
- Providing *system resource management*, including management of peak loads, priority scheduling, and allocations for scarce resources such as disk storage space;
- Providing *education and training*, including user consultation and education of various kinds, and possibly a more turnkey operation including system analysis, design, programming, testing, implementation and documentation;
- Providing *operating services*, including machine operation, tape and disk volume storage and control, file backup, and an inventory of expendable supplies;[2]
- Providing *hardware maintenance services*, including fault diagnosis (either internally or through a external supplier);
- Providing *applications implementation and management*, including choosing and implementing specific choices of programs in the computing system, as well as establishing operational controls over their use and managing staff use of those programs; and
- Assisting in long term *capacity planning and configuration management*; monitoring the market for additional or replacement hardware and software to maintain an appropriate and cost effective data processing operation.

Users of microcomputers are faced with the same set of responsibilities, although at a much lower level, and often with substantially less risk to themselves or to their orga-

[1] This task is considerably simpler for the microcomputer user. Not only are the tasks simpler and more homogeneous, but given the relatively low level of expenditure, often several different small computers can be purchased. Such diversity of hardware or software may be more effective in meeting institutional needs than earlier purchases of single systems.

[2] The importance of device and file backup, routinely performed centrally at most mainframe installations, is routinely ignored by most microcomputer operators, and at their own peril. Since the data processed on such machines is institutional, distributed data base management must ensure that microcomputer users are not only conscious of the possibility of data loss through failure to backup, but also perform backups on schedule.

nization if non-optimal decisions are made. Nevertheless, these responsibilities have been by and large transferred to users albeit on a considerably smaller scale, and users must be aware of them if decentralized microcomputer use is to be viable within an organization. This issue becomes immediately visible when malfunctions occur, and it is up to the user to diagnose the problem and solve it. Organizations with non-programmer users may want to retain a central data processing services staff for dealing with more difficult problems that may arise, while leaving the operational responsibility for substantively oriented work with the user.

The following sections deal with each of the issues raised above.

Requirements Analysis

Most classical computing centers were created and operated to serve a spectrum of needs. Such centers were created either for the use of one organization or for several organizations. In either case, the problem of understanding the combined requirements of all users and uses of the center was among the first to be dealt with.

The first step in analyzing the requirements that will be placed upon a computer system is to examine existing activities that are likely to benefit from being implemented on the new system. Such applications may already be running on another computer system, or they may be activities that are either performed manually or not at all.

Analyzing requirements that will be placed upon a computer system has a hidden dimension that is difficult for the non-computer professional to appreciate fully. That dimension is the analysis of latent demand, i.e., demand not foreseen at the time that the requirements for the computer are assembled. Experience indicates that over the life cycle of a computer system, which is in the range of 5-10 years, the requirements not foreseen at the beginning of the cycle become increasingly numerous and can come to dominate the total activity for which the computer is used. Such growth in unforeseen applications can eventually force a change or upgrade of the computer system because of the importance of the originally unseen capabilities of the system.

Such a process represents a healthy learning response to using a new and more powerful tool. It is important, however, not to pay too high a cost for the experience. During the process of analyzing requirements which the computer will be expected to fulfill, it is therefore advisable to obtain professional assistance in deciding which activities, existing or not, might benefit from the presence of the computer. To the extent that latent applications can be identified as early as possible, the computer system will be more responsive to the users' requirements over a longer period of time, and will provide a more stable and responsive computing environment.

The analysis of requirements for procurement of microcomputer systems is somewhat easier, since such systems are often selected for specific tasks. It is often convenient to provide a number of microcomputers to perform a number of different tasks. In this case, the hardware and software configurations of the microcomputers may be specialized to adapt them suitably for the different tasks. As latent demand is realized, additional systems can be acquired in the short run without incurring a significant cost penalty. Nevertheless, from the point of view of setting a direction for the acquisition of microcomputers and the way they are used as a whole to attack the problems of an organization, a relatively complete and professional requirements analysis is likely to lead to an acquisitions strategy, possibly phased over time, that satisfies the needs of the organization well.

For processing a census of population and housing, such a requirements analysis generally begins with the software that is to be used for the various phases of data processing, which in turn is determined by the amount of processing to be performed and the size and variety of the outputs required. Considerations of data volume generally dictate the number and size of computer systems that will be required to perform the work in a targeted period of time. Excess hardware capacity will be determined in part by the country's ability to repair computers locally and in part by the amount of trouble anticipated in effecting replacement parts and maintenance services from sources outside the country.

System Specification and Negotiation with Suppliers

Regardless of the scale of computing equipment used, requirements identified during the above analysis must be turned into a specification of what type and amount of computing equipment is necessary to accomplish the job. The specification must include hardware which has the speed, capacity, and devices to support the required activities, and software which has the functionality and capacity to perform the applications required. The choices of specific hardware and software are interdependent, making the selection process sometimes difficult.

When there are two or more possible suppliers and the procurement is of non-negligible value, public sector organizations are generally required to go through a process of qualifying suppliers and obtaining competitive bids for the system. Often such a procedure is based upon an official document called a *request for proposals* (RFP) which reflects the results of the requirements analysis as well as service, training, and related considerations deemed necessary as a part of the procurement. The content of the RFP is critical to assuring that the proposal ultimately accepted really meets the requirements well. Depending upon the legal requirements that pertain to specific procurements of equipment, an incomplete or inaccurate RFP documents can lead to having to accept a system that is sub-optimal or inappropriate for the tasks at hand.

Microcomputer procurements are generally simpler, although the same cautions are in order. The requirements of the activity to be performed must be translated into hardware capacity and software functionality. Once it is determined which set of software packages would best be used to process the census data, hardware requirements can be specified. With computer hardware, there is generally a tradeoff between buying from a recognized industry leader and an off-brand imitation at lower price. If the latter alternative appears attractive, then either the bid document or the contract with the supplier must be written in a way to protect the user from incompatibilities that interfere with use of the hardware and software together. Off-brand systems *may* require off-brand spare parts; if so, this must be foreseen at the time of procurement. Later expansion of the system, including adaptation to new hardware products and operating systems,[1] may be useful and the bid or contract document should include the purchaser's expectations in this regard.

Most developing countries using microcomputers to process their population censuses will use hardware that runs the MS-DOS operating system. With the currently available MS-DOS based software, there are several strategies that can be used. Although hardware requirements will tend to be somewhat similar, the approaches used by different editing and tabulation programs may require more or less hardware to do a specific job.[2] The total hardware specified should reflect the requirements of the particular software strategy chosen for the data processing.

Site Preparation

Site preparation for mainframe and minicomputers is often a complex and technical task. Mainframes require a physical environment in which temperature and humidity are controlled within moderately narrow limits. Any substantial amount of dust in the air may interfere with physical storage media such as disk or magnetic tape. The stability and quality of electrical power supplied to the computer must meet the requirements of the computer system. There may be a sufficient number of cables interconnecting the various cabinets comprising the computer system that a raised floor may be useful or necessary for safety. While minicomputers are generally less demanding of the environment in which they operate, preparation of a site for a minicomputer is non-trivial.

Large computer suppliers have been aware of the difficulties faced by individual users in preparing a site for their computer systems, and provide significant education and site planning and preparation assistance to their customers. They perform this function both for their customers and in their own self-interest; their systems will not perform effectively if the site is inadequately prepared.

Microcomputers are more insensitive to variations in their environment than are larger computer systems. They can tolerate wider swings in temperature and humidity; in general, they will work well in any environment in which their human operator is moderately comfortable. On the other hand, microcomputer diskettes are more fragile, and can be harmed by dust, sand, salt water condensation, and other particulate matter carried in the air.

Microcomputers require a stable electrical power supply, but the amount of power required by each individual unit is relatively low. A typical microcomputer consumes as much power as a bright incandescent electric bulb. Power outages and serious power fluctuations can cause loss of data and damage to the computer equipment, so that users must ensure an electrically clean and steady power source for their equipment. Different types and brands of microcomputers vary in their ability to tolerate electrical disruptions, and it is the purchaser's and user's responsibility to ensure that the quality of power supplied matches what the microcomputer system requires so that neither data nor equipment will be compromised or damaged.

Power analysis equipment is available for rent or for sale that will monitor existing electrical power precisely and will provide the data for an analysis of current power quality in any given location. The results of this analysis can be used to determine what types of power conditioning equipment or standby generating capacity, if any, will be required to provide acceptable power for the computer systems.[3] This ap-

[1] For example, currently most systems in developing countries are IBM PCs or PC clones running the PC-DOS or MS-DOS operating system. As programs become larger and users' needs evolve, there will be migration from MS-DOS to the OS/2 operating system. This migration will be slowly forced upon users who want to utilize new programs and new versions of programs, which to an increasing extent will not fit into the program address space provided by MS-DOS. The upgrade path for PC-compatible equipment may therefore be important, depending upon the tasks to be performed and the expected lifetime of the applications and the equipment.

[2] The fact that a specific set of programs requires less hardware than an alternative strategy does not necessarily mean that it should be chosen. The additional hardware required may reflect more efficient operations, or it may allow additional flexibility and options in the production process and in its outputs.

[3] The general subject of electric power conditioning is too broad to be discussed in detail here. There is a range of possible power problems that could either disrupt computer operations or damage data or equipment. *Voltage stabilizers* are used to make constant the voltage of a reliable source of power that varies in voltage. *Surge protectors* are used to clip short surges or drops in power often caused by other electrical devices on the power line or by external events such as stormy weather. *Uninterruptible power supplies* are often used to bridge short periods of power outage, using a bank of storage batteries that constantly buffer spare electrical capacity. Independent *electric generators*, often attached to diesel motors, provide continuous power subject to the availability of petrol to run the diesel or other engine that powers the generator. Depending upon the electric power environment that the computers needs to operate,

proach is generally effective, but has one major weakness; it is based upon the assumption that future power quality will mirror past quality. If in the future there is a significant degradation of power quality, then any power conditioning methods used may not be sufficient to cope with the degraded input quality.[1]

Fortunately it is easier and probably less expensive for developing countries to prepare a site for installing microcomputer systems for census processing than it was for their mainframe predecessor. While physical hazards to the data media (sand, salt, dirt) are still present, the temperature constraints and to some extent humidity constraints will be easier to meet. Local area networks may provide a way to minimize the detrimental effects of dirty environments by minimizing the transport of diskette media. Unstable power, unfortunately, will continue to be a problem which statistical agencies need to address.

Installation and Testing

Mainframe systems are almost always installed by the system supplier. Such an installation is technically relatively demanding, and the cost of making a mistake is potentially high. Minicomputers are less difficult to install, but the task is best done by technical specialists. Acceptance testing for such systems is generally done jointly by the customer and the supplier, and usually includes verification that all installed hardware and software operate functionally as promised, as well as a more lengthy test to meet a minimum downtime requirement.

Microcomputer installation is almost always offered as an additional service by the supplier, and the price of such service is not included in the price of the computer. Many microcomputers are relatively easy to set up and test, and require only a moderate amount of knowledge about the system. However, some installations may include installing special electronic circuit cards or other special devices which may pose some difficulties.[2] If the microcomputer system is made up of parts from different manufacturers, then the possibility of some level of incompatibility may complicate the installation and testing. Installation and testing has a software dimension, which is discussed below.

If a microcomputer system is procured at a remote or foreign location, it may be wise to specify testing at the point of purchase as a part of the procurement procedure. Such testing serves to identify initial problems before the equipment has been shipped and when it can easily be returned to the supplier for a replacement at little cost and with minimum delay.[3]

Statistical agencies should assure themselves that the procurement process for microcomputer systems and system software includes an acceptable strategy for installation and testing of the systems. Such a period is also useful for educating staff who will supervise the computer operation when it is commissioned. These staff members should ensure that they understand exactly how the systems are working and why they are working correctly before they formally accept the systems from the organization responsible for their procurement.

System Programming

System programming refers to the process of understanding and modifying the instructions comprising the operating system under which the computer operates. The operating system is a software program, generally supplied by or through the manufacturer of the computer hardware, that gives the user effective control over system resources. Classical operating systems generally include commands of the type "print", "copy", "create file", etc. Operating systems support execution of programming languages, applications programs, and all other software. They are an essential part of a computer configuration.

Mainframe computer centers and, to a somewhat lesser extent, minicomputer systems have one or more technical staff members who understand the operating system structure and modify and update it to meet the special needs of the software and users of that computer. They also install new software and ensure that it works well with the operating system of the computer. This is a highly specialized skill; operating systems for large computers are very complex and contain millions of instructions.

---

one or more of these techniques, as well as others, may be required to provide power that is sufficiently clean to operate the computer systems.

[1] Unfortunately, the cost of bad electrical power compounds itself in the following manner. In addition to obtaining whatever power conditioning equipment is required to condition the electrical power appropriately, the organization must also be prepared to support the power conditioning equipment itself should it malfunction. Such support can be provided in the same manner as supporting microcomputers, through maintaining a small inventory of spare units on site. For this reason, it may be be better to condition power to microcomputers using a number of transformers, regulators, and micro-UPS units rather than relying upon a power conditioning system based upon non-redundant large units. Power conditioning equipment is notoriously heavy, so that significant transport flexibility can be had by using multiple units. While the cost of multiple smaller units may be somewhat greater, the higher degree of certainty of continuous operation provided by having several identical units is almost always worth the additional cost if the application is at all time critical.

[2] The manner in which a microcomputer is packaged can affect ease of installation. Some types of microcomputers are relatively easy to open physically so that additional components can be added and other changes made. Some microcomputers, on the other hand, are designed to remain closed and are consequently more difficult to modify. Indeed, such modifications made directly by the user may void the manufacturer's warranty. However such voiding of the warranty may be unimportant when the maintenance strategy for the equipment relies more upon local resources than upon return to a supplier or a supplier's representative at a remote location.

[3] A strategy of assembling and testing each microcomputer supplied through its technical co-operation backstopping activities was adopted by the United Nations Statistical Office during 1979-1984. Such a pretesting policy remains valuable in cases where the system configuration is sufficiently new or different and may not be understood or work properly at its final destination.

STRATEGIC ISSUES IN THE USE OF MICROCOMPUTERS

Some of the tasks necessary for installing and maintaining a microcomputer environment can be considered in the realm of system programming. It is necessary to: (1) install the operating system on a fixed disk or diskettes; (2) make choices regarding how the system is to be configured; (3) configure a startup sequence for the computer; (4) install software links to special pieces of equipment that may be attached to the microcomputer;[1] (5) load and manage memory resident software, ensuring no memory conflicts with application programs; (6) configure application software to match specific characteristics of the microcomputer,[2] especially the video and printer interfaces; and (7) update the operating system periodically as new releases are issued, ensuring that the new release works consistently within the context of the changes made and the application programs being run.

In a large computer operation, the system programmer serves as the local expert on software issues. When a computer procedure consistently does not work, very often the cause is software based. Typically the system programmer is best situated technically to analyze the problem, determine why it does not work, and either change the procedure or create a work-around solution. Many such errors can be remedied by training users, but some require substantial analysis. In a microcomputer environment, the same kind of problems will occur, and it will be important to have at least one technically trained staff person who can understand the interaction of the system and application software as well as the structure of the hardware well enough to be able to assist users. Failure to have such skills available reduces users to guesswork, inefficient and limited use of the computer, and potential loss of information.

Most system software and applications software products are being revised or updated on a continuous basis by their manufacturers. Periodically, a specific *version* or *release* of such a product is made available to individuals and organizations that have previously purchased the product. New releases of software products are useful; they contain new features and functions that users have been asking for, as well as eliminating problems, sometimes referred to as *bugs*, that were detected in previous versions.

However, installing new versions of software may cause new problems to appear, possibly because of incompatibilities with other software products on the same system. Such modules should therefore not be installed while a large application (such as a population census) is being processed without extensive testing to ensure that substitution of the new module for the old has no adverse side effects. If the new functionality offered by the new release is not required and there have been no previous problems with the software, it may be better to freeze the hardware-software configuration for a time and defer the installation of the new release until after the entire large activity has been completed.

Census agencies should assign this function to one or more staff members at the beginning of the process of procuring the equipment. These staff members should be responsible for ensuring that they learn as much as possible about the aspects of system software and its interaction with both the underlying hardware and the applications programming environment during the procurement and installation process. These staff members should then be ready to serve as a local cadre of systems experts during the entire census data processing process.

System Resource Management

Mainframe computers are complex systems. From the point of vie

Mainframe resources are often managed through the operating system. For example the processing unit is generally rationed according to priorities which are associated with specific projects or users of the computer. Management of peak loads may be handled jointly between the operating system and operations staff. Disk storage resources may be managed automatically, or they may be managed by a staff person who functions part or full time as a storage administrator, parceling out storage to projects and users according to certain rules. In large computer installations, the management of system resources is generally performed by the computer operations staff working with the system programming staff.

Microcomputer resources are inherently easier to manage since, unlike larger computers, they are almost always established as single user machines. However, microcomputers with large fixed disks are often used in serial fashion by more than one person, necessitating management of fixed disk resources. Likewise, printers, communications devices and other specialized equipment will be shared over time; policies are needed for apportioning and assigning those resources and taking action when a resource is depleted. If a number of microcomputers are configured in a connected electronic network, the possibility of sharing more resources increases, and so do the responsibilities for sharing those resources.[3] Peak load situations must be managed manually

[1] A concrete example is provided by the addition of device drivers within an MS-DOS environment and placing appropriate commands within theCONFIG.SYS system file.

[2] Even the installation of some advanced applications in simple microcomputer environments may cause some difficulty. Often the most effective way of obtaining prompt and expert assistance is to contact the supplier's technical staff, which is likely to be in another continent. Having reliable communications links such as telephone, telex, and other electronic mail linkages with other countries is very useful on such occasions.

[3] In one sense it is *more critical* to manage shared microcomputer resources than large computer resources. Mainframe and minicomputer operating systems are built with the assumption that there will be multiple concurrent users of the computing resources; these operating systems therefore generally provide a comprehensive set of interlocks and protection mechanisms that ensure that one user's activities cannot interfere with another's. For example, the system will not provide one user with

by assigning available microcomputer systems to tasks to be performed in order of decreasing urgency.

Census agencies should be aware of the need for managing a sizeable number of microcomputers. One or more staff members should be given the responsibility for ensuring that the the microcomputer resources are used in a balanced manner for meeting the overall needs of the census process. These staff members should be one close to the overall management of the process, and the manner in which he or she allocates resources should reflect the overall plan for finishing the census processing. They will also be concerned with identifying machines needing repair, and adjusting the work schedules of the staff to compensate for any delays in obtaining access to machines. Depending upon the size of the processing operation, these functions could be performed on a part time basis in a small country or they could require several persons, some perhaps assigned on a per shift basis, in a larger country.

Education and Training

Education and training activities deserve very significant attention in national statistical offices and in census agencies for a number of reasons. First, they represent the investment in human capital that is absolutely necessary to reach a self-sustaining mass of skills in statistical data processing that is essential for supporting census and other data processing activities. Furthermore, such activities often suffer relative neglect, often because of the necessity to proceed with operational activities at the expense of acquiring new knowledge. In addition, leakage of data processing skills from the statistical office and the government as a whole to other sectors and other countries is significant and demands that the level of investment in these sectors exceed the rate of outflow. Finally, the emergence of the microcomputer and its companion software industry has made it practical and desirable for subject matter specialists to be trained to use computers directly in their work.

An effective training strategy in data processing skills is an essential condition for both survival and development for a national statistical office. It must be accompanied by a strategy that assures retention of the skills within the office to as great an extent as possible, else the training will not directly benefit the purpose for which it was intended. Among the important issues are: (1) who is to be trained, in what subjects and to what depth; (2) who can best provide the training required; (3) what should be the relationship between conceptual training and detailed operational training; (4) how should the training be distributed between local and foreign locations; (5) how can indigenous training capacity be built up and/or utilized; and (6) how can changes in the organizational structure assist in ensuring continuity of the required data processing skills being available to the statistical office?

Selection of students and curriculum. Ever since the introduction of automatic data processing machinery into statistics in the 1880's, statisticians have had to rely upon a set of skills that have not fit properly in the professional repertoire of a statistician. To be sure, the gains made possible by employing such equipment have been enormous, but the cost to the statistician has been either to learn the additional skills required or to share responsibility for producing the results with a data processing specialist. Because of the inherent complexity of mainframe data processing equipment, statisticians have generally chosen to work with data processing specialists to produce statistical outputs.

The division of responsibility between statisticians and data processing experts has been generally a useful and productive one in the past, representing a division of labor based upon specialization of function. It has not been without its problems, which include problems of communication between disciplines, some incompatibility in career goals, and difficulties in determining the demarcation points between areas of responsibility. In the past, this division has generally been necessary because of the substantial technical difficulties of using earlier generation data processing equipment in an efficient manner.

The introduction and proliferation of microcomputer systems and mass produced microcomputer based software is causing this situation to change in a major way. Although professional data processing assistance will be needed by statistical offices for at least some activities into the indefinite future, the relative and increasing ease with which these systems can be manipulated argues strongly for direct intervention and manipulation by professional statisticians. In order to capture a mass market for products, software designers have had to create programs having an easy to use and forgiving interface. This is often referred to as "user friendly" software.[1, 2]

access to another user's data files without the express permission of the owner of the files. The owner may provide this permission selectively and may limit access to reading the data, updating it, or erasing it. In contrast, almost no microcomputer environments offer this protection; unless explicit and cumbersome precautions are taken manually, each user's information is accessible to all other users of the computer and can be inadvertently modified or erased quite easily.

[1] In putting more emphasis upon the quality of the user interface in microcomputer software, software designers have essentially substituted additional hardware and software capital for the labor of the user in manipulating the program. This substitution not only recognizes the changing characteristics of users as the computing market broadens, but it also reflects the continuing shift in the factor prices of capital and labor towards the higher opportunity costs of time as perceived by computer users.

[2] The sharp distinction between current user friendly software and previous presumed user unfriendly software is to some extent not justified. Since the invention of electronic computers, both the languages used for programming and the user interfaces available have undergone a relatively continuous evolution. This evolution has been driven by two major factors, evolution of experience and the relative factor prices of capital and labor. Experience with early programs led to successive stages of refinements in user interfaces, culminating at the mainframe and minicomputer levels in 4th generation languages, text based menu driven systems, rich interactive operating systems, and comprehensive statistical packages with user

STRATEGIC ISSUES IN THE USE OF MICROCOMPUTERS

A variety of commercially produced microcomputer programs exist which address directly processes familiar to the statistician: data entry and validation, creation of frequency distributions and tabulations, creation of simply structured data bases, derivation of more complex statistics and hypothesis testing. In addition, productivity tools in the form of spreadsheets and elementary data base systems are available as well as more complex numeric and visual data analysis tools. These commercial offerings are complemented by offerings from governmental bodies and not-for-profit organizations that are designed for direct application to specific problems of national statistical offices.[1] All these programs can be learned and used effectively by statisticians having only elementary computer knowledge.[2]

There are substantial implications for data processing training strategy based upon a shift in the relative responsibilities for processing the data. If statisticians are to take more direct responsibility for the data processing aspects of their work, then they must be prepared to use computer based tools to advantage in getting their work done. They do not necessarily need to learn to program computers in the classical sense, although interest in this direction should not be discouraged.[3] They do need to be trained in use of a variety of personal productivity tools that have direct applicability to statistical data processing, as well as in programs that are oriented specially to the major functions of statistical processing, such as microdata capture, statistical editing, and tabulation. Each practicing statistician with access to microcomputers should be familiar with at least one program from each of these classes, and should have the ability to apply it to any statistical data collection which can be effectively processed by it.

The implications for data processing specialists differ considerably. Their role will change considerably, from a role which included active responsibility for a significant part of the processing to a role which is more consultative and advisory in nature. The data processing specialist will have to have a significantly broader view of the technical alternatives made available by the computer industry and ways in which they might be applied to assist the statistician. Knowledge of processing techniques must be supplemented by a more in depth knowledge of techniques and of a broader market of products.

Conceptual and operational training. Training for both data processing staff and for statisticians must include both conceptual and specific operational training. Conceptual training is important in that it provides a model into which a professional can place new experiences and understand them as part of a coherent whole. For example, it allows a data processing professional to comprehend the important aspects of a new programming language within the context of what programming languages do, as opposed to learning the language as an isolated tool of his or her work. Data processing staff clearly need considerable training of both types to be good at what they do and to keep up with a technology that is rapidly evolving and requires continuous updating of skills and experience.

Statisticians on the other hand have generally been satisfied in the past with mostly conceptual training when the operational aspects of data processing were delegated to others. The conceptual aspects of most statistical data processing operations are relatively easy to understand, and it is not necessary to understand some of the more esoteric technical aspects of packages to understand what they do and how they can be applied to a statistical process currently at hand. With the shift in emphasis to microcomputer based processing, statisticians will benefit greatly from adding a significant amount of operational knowledge to their repertoire of skills. In particular, every statistician should at any point in time be familiar operationally with at least one representative program from each of a number of classes, such as data entry, editing, cross-tabulation, spreadsheet, word processing, and elementary data base manipulation programs.[4]

Operational training is important for statisticians for another important reason. A person who does not have operational experience in implementing data processing proce-

---

command languages of various flavors. As the factor prices of capital (hardware and software) and labor (opportunity cost of users' time) shifted substantially, it became more important to provide additional software capital to improve labor productivity. In the market for microcomputer systems, in which the most user labor is currently employed, this has led to major investments in the user interface, culminating in application oriented productivity tools that can be used effectively by a large number of individuals with little or no previous computing experience. Statisticians in general are among the more sophisticated in this substantially expanded category of new users of information technology.

[1] Examples are the family of CENTS programs, CONCOR, and CONTROL and CENTRY produced and distributed by the International Statistical Programs Center of the U.S. Bureau of the Census in Washington, COXTALLY and PC-EDIT distributed by the U.N. Department of Technical Co-operation for Development. These programs have been designed for various aspects of census and survey data processing.

[2] Even stronger statements can be made in some cases. It is now not unusual for a subject matter specialist to spend more time specifying a relatively simple task to a computer professional then it does for the specialist to perform the task directly.

[3] Some understanding of programming concepts and programming ability is useful. A number of microcomputer programs depend for their flexibility upon the user being able to embed high level computer language statements within a larger framework oriented toward the solution of a specific class of problems. Examples are the use of Pascal-like statements within EntryPoint, dBASE, and BLAISE (from the Netherlands Central Bureau of Statistics) for customizing these programs for specific applications. Such programming knowledge on the part of the statistician is not absolutely necessary, since such applications can often be customized by data processing consulting staff to the point where further customization by statisticians is a routine task. However, statisticians having some basic understanding of the underlying programming process will be able to exploit such tools in a more powerful and effective manner.

[4] With some exceptions, it matters less which of the programs from each class is selected than the fact that at least one program from each class is absorbed. Adequate data processing training at the conceptual level should provide statisticians with sufficient understanding to make the conceptual leaps from product to product in each class, and to understand the relevance and effectiveness of new products when they come to market.

dures really does not have an adequate understanding of data processing, regardless of the amount of conceptual knowledge acquired. Gaining experience (as opposed to appreciation) in data processing requires learning by doing; often it is best done using an apprenticeship model in which the student apprentices himself or herself to an experienced person and works under that person for a period of time.

Local and foreign training. Appropriate locations for training depend upon the number of persons participating in a course, availability of and requirements for special facilities, and the costs of providing training at competing locations.

In general, the larger the class, the greater the efficiencies inherent in local training. For groups of trainees, often the cost of foreign travel and subsistence is considerably larger than the cost of supporting a trainer in the country itself. Using microcomputer technology specialized facilities, especially for basic training, can generally be supplied locally. Further, basic training either at the conceptual or operational level is the type of training that should be given in a widespread manner throughout the statistical office. Such training is therefore generally best done locally both for reasons of coverage and of cost.

A significant disadvantage to local training, even at the elementary level, is that trainees remain within their own technical and cultural context during the training period and are not exposed to the wide variety of influences that can manifest themselves in a technically different and more advanced environment. One of the non-quantifiable, nevertheless important benefits of education is provided by exposing students to a wealth of new ideas, causing them to question assumptions and attitudes that may have in the past restricted the scope of their learning and their ability to apply their efforts to problem solving. Education and training in a foreign environment will provide new models of technical approaches and solutions, and is more likely to broaden the intellectual and technical horizons of students, with consequent benefits but at a higher cost. Developing countries should strive to take full advantage of the expertise found in other countries by absorbing both the specialized knowledge and the flavor of more advanced technical environments, as well as building durable and long lasting links with institutions in those countries having specialized and useful expertise.

Specialized and advanced training is generally better performed at remote or foreign locations. Often such specialized training requires special equipment or knowledge that is not easily made available locally; there may also be only a limited number of candidates for each specialized field. Furthermore, at more advanced levels, there are significant intangible benefits that come from students being exposed to different and generally more advanced environments as well as to students from still other countries.

Learning is, of course, a life long process. Like many other fields of professional endeavor, computing science and technology has been in the process of rapid change ever since its inception, and persons who have not reinvested substantially in maintaining their professional skills have seen their professional careers atrophy to some extent. Given the limited opportunities to invest in professional training in developing countries, it is very important to inculcate this idea in this emerging professional class. Professional training is *not* a series of discrete opportunities to participate in an in-country course or to travel elsewhere for a course; it is a life long activity which professionals must take active responsibility for themselves. To the extent this attitude can be fostered and adopted, training will become an active, self-perpetuating component of individual behavior.

*Indigenous training capacity***.** Training does not necessarily have to be performed by foreign individuals or institutions. There are often good sources of training and trainers within the same country. In countries with some commercial infrastructure in computing, there are often a number of sources of expertise in the private sector that can be used to execute a large part of the training program required, especially for training programs at the more elementary levels.

Co-operative relationships between statistical offices and local universities appear to be underutilized. Statistical offices and university departments have different objectives, but their interests in knowledge and learning are not competitive; on the contrary, they complement each other. Statistical offices have everything to gain from strong university departments in statistics and computer science and from an effective computer service center. Such relationships should be explored and exploited; they represent a two-pronged approach toward achieving a critical mass of skills in statistics and data processing which either institution would have a much harder time achieving without the other.

As an example, local universities are in a good position to introduce the use of computers, especially microcomputers, at the university level either as part of a curriculum in computer science or in support of other curricula in which the computer is used as an important tool, e.g. statistics. If this is possible, it would allow many students to become acquainted with computer based techniques prior to obtaining employment. It could provide the critical mass of trained computer users that would have a significant effect upon slowing leakage from the public sector. It would also allow earlier identification of persons with a good aptitude for computer use and help them to determine how they would best like to exploit these skills in their professional life. Government offices should work with universities in assuring that the curriculum, where appropriate, is consistent with post-university employment opportunities within the country.

Staff continuity. One of the most difficult problems faced by organizations in all countries, but especially in developing countries, is the problem of leakage of trained personnel. The classical scenario is as follows. A national statistical office invests resources in training statisticians and data processing staff. Shortly after the training has been completed and the trainees return to their full time positions, they find other more interesting positions either elsewhere in the government, in the private sector if one exists, or in another country if they are able to emigrate.[1]

To the extent that this such behavior occurs a great deal, it represents both a drain upon the resources of the statistical office and a disruption of its work program. It further acts as a depressant on the more optimistic expectations that were formed when the training was planned and commenced. Work goals of the statistical office may not be realized. It may indicate that the strategic plan for the training was overly optimistic or incorrect.

The emergence of microcomputer technology has helped to some extent to alleviate this problem, although the problem still exists. First, it is generally easier to identify staff who can be trained to use microcomputers, and to train them, than it was to identify and train staff on earlier generation large computers. It is therefore easier and less costly to adopt a strategy of training more staff in the technology than may be required at the time, and relying upon the additional trained personnel to fill in for the leakage that occurs. Second, with involvement of subject matter specialists in computing activities increasing, a proportionately larger amount of an organization's computing knowledge will reside in staff other than computer professionals, resulting in better integration of subject matter and computing knowledge as well as lessening the severity of the effects of leakage of pure data processing personnel.

Sources of training. Training in data processing has historically come from a variety of sources. Equipment suppliers have been active in promoting the use of their products. Bilateral and multilateral foreign aid have contributed both by providing on-site experts and in funding training in foreign countries. Local universities and firms in the private sector have occasionally been important contributors of expertise.

Historically, most training in basic data processing skills has been provided by suppliers of mainframe equipment. Such an arrangement has been in both the customer's and the supplier's interest, since the computing equipment will not perform satisfactorily unless the customer had a technical staff that knew how to use it. With the trend to smaller and more inexpensive computers, suppliers of computing equipment have diminished their training offerings and, significantly, have unbundled them from the hardware. Training can generally still be purchased, but at a higher price relative to the price of the hardware.

The shift in technology toward microcomputer use has altered the balance between these sources of training in a fundamental way. During earlier years when countries and organizations were acquiring mainframe computers, substantial training was provided by the equipment supplier as a part of the acquisition. Such training was required because, *inter alia*, successful exploitation of the equipment depended upon a trained staff, and there were often few sources of training other than the supplier. Customer satisfaction was an important marketing consideration, and therefore the supplier considered adequate training an essential part of the goods and services that had to be delivered. The cost of this training was real, and was bundled into the purchase or rental price of the equipment.[2] However, it generally did not appear explicitly as a charge to the customer.

The evolution of computing technology up to the present pervasiveness of microcomputers has substantially changed this balance between sources of training. First, the breadth and depth of computing expertise by users of the technology has increased over time, causing training activity to be less necessary for the success of a computer installation. As a result, suppliers have increasingly unbundled their training offerings from equipment sales, allowing customers with either a trained staff or an internal training capability to be able to provide such training internally. Second, as customers were increasing in sophistication, suppliers' reputations no longer were so dependent upon their providing sufficient training directly; it was increasingly perceived that users were responsible for obtaining whatever training was necessary to use their computing equipment appropriately.[3] With the introduction of the microcomputer, the high cost of

[1] The argument is made that, with the exception of leakage outside of the country, such leakage is not as detrimental as it may seem because the country nevertheless benefits from the increased training in some manner. While this may be true, continued substantial leakage from the statistical service specifically may result in a perpetual shortfall of output relative to expectations. If such leakage is systematic, the continuing shortfall in output can be dealt with either by explicitly scaling back expectations from the statistical service or by shifting the responsibility of such training to a unit more central in government, since such training is useful in many sectors within government and not just within statistics.

[2] The relative prices of hardware and services encouraged such bundling. The cost of such training in earlier years was small relative to the substantial financial commitments made by customers to rent or acquire such equipment. The real costs attributable to training could therefore easily be absorbed with little or no notice within an overall price that was dominated by hardware production costs.

[3] Since software manufacturers depend more upon satisfied users recommending their products to others, they have exhibited more concern about training users. Such training has taken the form of good documentation as well as on-line tutorial programs which lead users through the steps of using their programs. In developed countries there is now a small industry that supplies tutorial training material for a wide variety of commonly used microcomputer software packages; such products are sometimes inexpensive and useful.

labor relative to the cost of equipment alone guaranteed an end to bundled training; the cost of training was a substantial fraction of the cost of the hardware and often exceeded it. System acquisition and training were therefore almost totally separated.

The shift in sources of training also has advantages. Training provided by suppliers of computing equipment was often narrow and technical, suitable for computer specialists but less appropriate for statisticians wanting to learn something about data processing. The existence of only such narrow training may have assisted and perpetuated the role of data processing staff in operational activities and the partial abdication of statisticians from them. Now that developing countries and providers of external aid have more pressure put upon them to provide such training, the training can be shaped more appropriately for the situation and for the relative roles that statisticians and data processing staff will play in the future in national statistical offices.

Operating Services

Users of a large computer system use it either in batch mode by placing prepared jobs in a batch queue or by interacting with it through a computer terminal. Using either method, the computer is given a series of instructions to execute that are specific to the task at hand. In general, the user does not see and has no need to see the actual operation of the computer, which is attended to by an operations staff. That staff operates the physical equipment that comprises the computer system, including changing disk packs and tapes, stocking the printer with paper and separating and delivering output, responding to exception conditions, executing diagnostic programs if machine malfunctions are suspected, monitoring environmental conditions, scheduling the order of tasks, and powering the system up and down as required. The operations staff also mediates conflicts between demands for access to computing resources during periods of peak loads, and balances how resources are allocated among activities having different priorities.

Operators provide a very important service generally referred to as *backup*. Periodically, each file stored within the computer is written, or backed up, onto magnetic tape. In busy computer installations, this procedure may be executed daily. The tapes produced are labeled by date and time of creation, and are stored in a safe place. Some computer installations make two copies and store one in a different building, thereby protecting against the possibility of destruction of a sole copy. The backup tapes are used to restore files in the event of an error by either the user or the computer system which contaminates or erases one or more of the user's files. While such errors are not common, they do occur, and the cost of losing irrevocably programs and data in which much has been invested is too great a cost to bear.

Backup services generally are provided automatically to users of large computer installations. One of the advantages of having microcomputers connected to a local area network is that system and network resources for all computers on the network can be concentrated in the server system. Backup of these computers then reduces to backup of the server which is a network operation and can be performed by the network manager, rather than being an individual user operation which requires the co-operation of many individuals. Reliability and dependability of the process is enhanced.

A microcomputer user is faced with simpler operating chores. The machine must be initialized for work properly. Diskettes and disks (and possibly other forms of input) must be arranged in the right order and given to the executing program as required. The printer must have paper installed in the correct path, and must be kept clear of paper jams. Occasionally the ribbon will need to be replaced. Diskette and other computer media need to be stored in safe places. Some additional similar chores may be required, depending upon the complexity and the configuration of the system.

Microcomputers are vulnerable to the same kinds of problems as larger computers, although the frequency and severity of the problems may not be the same. It is therefore important to ensure that all information stored in a microcomputer on a fixed disk be backed up regularly. While such a function is routine, most users perform it only occasionally, and as a result valuable data or programs are not infrequently lost. Backup is an operational service that either must be performed by the user, or must be provided by a central operations service on a decentralized base of microcomputer systems.

Census agencies should ensure that the above operating functions are assigned explicitly and the staff responsible for carrying out these functions are trained to perform them properly. In particular, file backup procedures should be firmly institutionalized, and the backup file copies should be stored in a secure place not immediately adjacent to the site of the central data processing activities.

Hardware Maintenance Services

Hardware maintenance services are of two kinds: (1) preventive; and (2) problem diagnosis and repair. Preventive maintenance takes several forms, including physical cleaning of parts of the computer, checking and oiling mechanical parts as required, replacing parts with limited lives on a scheduled basis, executing diagnostic computer programs that check the correct functioning of the various units comprising the computer system, and installing engineering changes issued by the supplier's manufacturing and design center to correct error prone circuitry and provide for greater reliability. Problem diagnosis and repair consists of examining situations in which the computer is apparently not performing correctly, diagnosing the problem, and effecting the necessary repair. To perform the latter function properly may require having some specialized tools, manuals containing detailed hardware and software descriptions of the computer system, and a supply of spare parts and a communications and transport link to a regional or national diagnostic and parts service.

It is important to realize that when a complex computer system behaves in an erratic or incorrect manner, it may not be obvious initially even to specialists whether the problem is due to faulty hardware, faulty software, incorrect operations procedures, or even a problem with an application program. Sometimes the problem may arise from an interaction between hardware and software modules that has never occurred in the past. The problem may be due to incompatible versions of several software modules being executed together; programs from different suppliers may embody assumptions that conflict when they are joined. The major challenge facing technical staff who perform such diagnosis and repair is to identify at which level the problem exists and then to isolate and understand what is causing it.

Historically, mainframe computer installations have contracted directly with the supplier of the mainframe for hardware maintenance. The supplier is in the best position to understand the intricacies of the equipment, provide spare parts, and train a force of service technicians who can be stationed at installations of that particular computer model. In general, minicomputer users have followed the same strategy, although with simpler minicomputers users have assumed somewhat more of the responsibility for maintaining their own systems.

Given the relatively high cost of contracting directly with a computer supplier to provide hardware maintenance, some developing countries have trained national technicians to provide these services, but in most cases this strategy has not worked well. Technicians who have been trained so narrowly on only the current model of computer are unable to move to new systems without substantial retraining. Also, once trained, there are strong financial incentives for the technician to move to the private sector or even to another country.

The situation is the reverse in the microcomputer industry. Microcomputers are relatively simple and inexpensive, and because of both the simplicity and advanced technology, the rate of failure of any single piece of equipment is relatively low. Maintenance contracts may be offered by the manufacturer of the equipment, but are most often fulfilled by local dealers who are likely to be independent of the manufacturer. There is a retail level and often a wholesale level in the distribution channels for microcomputers that separates the original manufacturer and the customer.

Where competent and dependable local dealers exist, microcomputer hardware maintenance can be obtained from them, although the price may be high. Self-maintenance strategies are more feasible with microcomputers, and may be necessary if no local or reliable source of external maintenance is available. In any case, customers should have some level of awareness of common hardware problems and how to deal with them, simply to maintain an efficient working environment.

Organizations choosing to perform self-maintenance for their microcomputers can take advantage of a well known technique of isolating hardware faults which depends upon redundancy of equipment. Microcomputers are generally employed in clusters, very often with groups of machines having identical configurations. When a particular computer gives erratic or suspicious results, the problem can be investigated by moving components of the suspect machine to another machine that is functioning properly. For example, first the software copy that is being executed can be moved to another machine, followed one by one by the data file, the printer, the video monitor, the special printed circuit cards if any, the connecting cables, and any other components of the system. At some point, the problem will move, and the component needing more investigation or repair will be identified as the one that was moved when the problem moved.

When dealing with erratic or incorrect behavior in a microcomputer system, it also may not be obvious whether the problem is due to faulty hardware, faulty system software, incompatibility between application program and operating system, incorrect operations procedures, or a problem with an application program. Since hardware and software are likely to come from several different suppliers in a microcomputer environment, the possibility of two modules not working together properly is slightly higher than in a mainframe environment. It is necessary to be able to perform most diagnoses like this within the organization, since reliance upon external assistance for every such problem will be quite expensive. The best and most practical way to address situations such as these is to have a minimum critical mass of knowledge and experience in both microcomputer hardware and software within the organization.

The overriding issue in supporting a microcomputer installation is, when the system malfunctions or appears to malfunction, whether the situation can be diagnosed and fixed relatively quickly most of the time. If an organization has such a capacity, then reliance upon microcomputing technology is likely to be useful and effective. If this condition cannot be met, then such reliance is more likely to lead to frustration, periods in which activity is blocked, and higher costs of operation.

Having adequate hardware maintenance support is an essential part of any census data processing strategy, whether the processing is performed on a large scale computer or on microcomputers. Census agencies should ensure that such a strategy is formulated at the planning stage, and that any necessary training, establishment of relationships, and procurement of tools and spare parts occurs before the processing begins. Fortunately, there are a number of strategic directions that can be taken, the simplest of which is to obtain a significant number of redundant systems to be placed into service as others malfunction. If countries are able either to train local staff in board level fault isolation or better, or to take advantage of other local technical resources such as might be found in the national telecommunications agency or similar technically oriented group, then reasonably effective hardware maintenance can generally be assured at reasonable costs.

Applications Implementation and Management

The responsibility for implementing the actual set of application programs and managing their use is the part of the computing task that has historically been closest and most familiar to the user or organization using the computing system. Statistical offices have become intimately involved with these tasks of necessity in the past. Nevertheless, there are considerations that must be taken into account in a microcomputer based environment that are somewhat different from those in a mainframe environment.

The integration of system and applications software, and its continued functioning over multiple system releases and possible hardware configuration changes has been discussed above. The need to maintain a stable and accurate computing environment in such dynamic conditions may have a bearing upon the choice of applications software and especially the amount of external support that is available to assist in this task. In particular, it is likely to shift preferences in favor of purchasing popular and well known commercial software as well as software from technical co-operation agencies rather than relying upon internal software development. Such a shift leads to having to make more choices regarding the relative advantages of certain programs over others; one dimension to be evaluated in making such choices is ease of implementation and stability within the microcomputer environment chosen. For census data processing activities in the short run, these choices are relatively limited, but the principle applies to statistical applications beyond the next population census.

User organizations need to maintain the critical mass of computing skills that will allow them to succeed with the applications task. In previous years, the computing activities in many developing countries were centered around one or a few computers. Computing skills that existed in the country were located near one of those computers. Each computer center had not only access to hardware and software, but also access to a small but critical mass of knowledge that allowed the computing organization to function. This critical mass was important for learning, obtaining others' opinions, and for reinforcing positively the progress made by the organization.

With the distribution of hardware through independent microcomputer systems, the human skills and knowledge have also had to become distributed. This intensifies the danger that too widespread a distribution will dilute the critical mass too much, leading to a sub-optimal use of the computing resource and the overall activity completed not as well as it could have been. Statistical organizations should assess this risk in their own context and take whatever measures are possible to minimize it. In particular, it is dangerous to assume that distributing computing hardware in the form of individual microcomputers will by itself lead to solutions to computing problems, irrespective of the way in which staff support can be distributed with it.

Subject matter activities that are moved to microcomputers will require at least as much organization imposed upon them as they did when implemented upon central computers. With multiple machines in operation, it will be tempting for individuals to make changes to their local microcomputing environment for real or imagined local efficiencies, without understanding that they may be affecting the entire statistical production process detrimentally. While individual initiative in improving the use of the computing resource should be encouraged, improvements must be introduced in an organized and controlled manner to protect both the integrity and quality of the data that are being processed. Data processing, subject matter, and operations staff must all be made to understand and subscribe to this method of operation.

Capacity Planning and Configuration Management

Mainframe computer installations almost always have a professional staff containing system programmers, operations staff, and user assistance staff. They monitor the activity of their computing equipment and often measure the utilization rate of various system resources. Such measurements provide guidelines regarding present and future demand for specific resources, whether it be computing capacity, communications bandwidth, disk mass storage, or printing capability.

The monitoring process allows the management of the computing facility to modify its configuration over time to provide balanced and effective service for its users. It also allows them to identify when and how quickly user demand is approaching the fundamental constraints built into the model of computer system they have installed. As such barriers are approached, management generally begins concrete planning either for replacement of the system or for a more significant augmentation of the facilities in order to supply its users with a more continuous service than would otherwise be possible.

This function is largely alleviated for microcomputer users because of the modularity of the technology. Often, additional capacity required means purchasing one or more additional computers. However, some constraints are somewhat more difficult to deal with, such as limitations on disk file sizes and disk space, network congestion, and others. While such problems are infrequent, the responsibility is upon the user to anticipate such circumstances and devise a plan for either alleviating capacity constraints or modifying the procedures to be performed in such a way that the constraints can be bypassed or worked around.

The staff members within the census data processing operation who have responsibility for systems work and operations management need to be sensitive to these issues so that they can alert census management when there appear to be resource constraints that are affecting completion of census data processing activities according to the planned schedule. Such a responsibility is consistent with their other roles in managing the overall data processing operation and resolving operational problems that arise during the course of the processing activities.

Summary

Microcomputer technology is appropriate technology for many areas of computing in developing countries. Statistical agencies in both developed and developing countries are beginning to adopt this technology, and experience to date has generally been very positive. Some countries have pioneered in employing it to process their censuses of population and housing, with positive results.

Microcomputers are, however, not a panacea for all data processing problems. Establishing a microcomputer based data processing operation entails fulfilling the same responsibilities that existed in previous mainframe or minicomputer environments. While the underlying problems and processes have not changed, the responsibilities have often diminished to the point of not being directly visible. This is due largely to the increased ease of use and simplicity which characterizes many microcomputer environments.

Successful exploitation of microcomputer environments, whether for population census data processing or for other applications, depends upon recognizing that these responsibilities associated with computing continue to exist in all computing environments, although with microcomputers in diminished or different form, and that that they must be met. This is often a matter of their proper recognition in the analysis and planning of a data processing operation. If responsibilities are defined during planning and are allocated to ensure that they will be met, then microcomputers can be used with good success in addressing the various aspects of the census data processing operation. Census agencies in developing countries are therefore strongly advised to consider how the functions discussed above in this chapter can be best met in their own census data processing activities, regardless of the processing strategy chosen or the scale of the processing operation.

Background and History

The movement of the classical activities of statistical data processing – data collection, editing, tabulation and analysis – to microcomputers has taken place in several ways.

First, established programs that previously executed in mainframe and minicomputer environments have been moved without significant changes to microcomputers. Examples of such programs are CENTS, CONCOR, TPL, SPSS, BMDP, PSTAT, SAS, and the population projection program sets from both the U.S. Bureau of the Census and the National Academy of Sciences.[1] These migrations are possible because the microcomputer system configurations available to developers today often are at least the equal of the mainframe environments in which the programs were developed originally. Limitations in the microcomputer versions are generally related to mass storage limitations in microcomputers generally[2] and possibly the speed of scientific calculations in microcomputers not having specialized hardware to execute them efficiently.[3]

In addition, new programs not specifically developed for use in statistical data processing have been adapted to perform such work. The most prevalent example appears to be the use of Lotus 1-2-3 to process aggregate data and surveys of modest size and to provide a framework for simply structured statistical data banks. Lotus 1-2-3 presents a spreadsheet interface to its user which is more suited for accounting than statistical data processing operations, yet it has high value for work in statistical offices.[4] Similarly, some simple data base management programs such as dBASE have been used to implement both numeric storage and processing of small data collections as well as administrative operations that were either performed with text editors or special programs in previous computer environments or were performed manually. Such new tools are not only useful for some statistical activities *per se*, but also for the *management* of large statistical activities which were previously not easily automated in an efficient manner.

Finally, for data entry procedures that were already being executed on primitive microcomputers using a combination of physically oriented controls and arcane languages,[5] the move to more general and powerful microcomputer environments must have been very welcome. A new generation of data entry programs, typified by Entrypoint, Rode-PC and CENTRY, provide considerable power, integration of data entry and editing, and use of multi-purpose hardware that can also be used for most other functions in the office. Portable computers allow the concrete possibility of substantial decentralization of the data collection activity up to the actual point of collection.[6]

History of Census Data Processing on Microcomputers. Initial use of microcomputers for census data processing began in the year 1979. The first countries to adopt this technology were small and geographically remote. In general, they had no installed computer base, and the market for computer systems was insufficient to attract any computer supplier to enter the country. Further, even if a computer

[1] It is interesting to note different paths that were taken by these programs en route to a microcomputer environment. CENTS and CONCOR were originally written to fit into small mainframe computers, and were rather easily adapted to microcomputer environments. SPSS and PSTAT had their origins in the 1960's and were first implemented on the largest mainframes available at that time; the adaptation to microcomputers was more difficult because of the substantial amount of code that had to be converted. One of the most complex programs, TPL, written by the U.S. Bureau of Labor Statistics was not converted; the system was entirely rewritten for a microcomputer environment by QQQ Software Inc. of Arlington, Va. by one of TPL's primary developers. There are now few mainframe statistical systems of general use that have not either begun or completed a migration in some form to a microcomputer based implementation.

[2] Within the developed countries, current product development work and actual products brought to market indicate that the mass storage restrictions are being alleviated. The marketing problem in doing so has been to provide individual users with adequate mass storage on a personal basis. The comparison for statistical offices is between microcomputers with large mass storage and an earlier mainframe with very substantial mass storage, but on a *shared* basis. While this issue is not yet resolved, it may be that statistical offices will need or want to rely upon large disk and file servers connected to individual microcomputers through local area networks.

[3] Cost considerations are interesting in this area. Within the last 10 years, additional such hardware, generally called *floating point* hardware, could be added to larger computers for about US$5,000 or more. The equivalent hardware in a microcomputer concept is a floating point chip; these currently cost about US$75 to $300.

[4] An interesting related development has been the recent emergence of programs which make use of the spreadsheet metaphor within the context of official statistics represented in tabular form. A good example is provided by PRTAB from Prospect Research in New Haven, Conn. Such a model allows data to be represented in spreadsheet or tabular form, with the advantages of being able to manipulate the table in standard ways; it *also* provides the additional attributes and superstructure that characterizes official statistical tabulations: footnotes attached to rows, columns, and entries, grouped headings; an expanded notion of units of measure; non-cardinal data classifications for categories such as missing, insignificant, and not applicable, with the spreadsheet algebra extended to cover such categories. While the extent to which such an approach is useful remains to be determined, it is a promising development for the tabular processing and storage of certain common types of statistical entities.

[5] Classical data entry devices consisted of card punch machines, later replaced by diskette recording devices such as the IBM 3742 and the IBM 5280. The 3742 did little more than modify the recording medium adding limited programmability. The IBM 5280 was programmable in a non-standard dialect of RPG which was not an easy language to use, requiring professional programmers.

[6] There is some controversy regarding the degree to which the computer should itself be interjected directly into the interview or primary data collection process; this issue is not discussed here. However, there are clearly a significant number of instances at present for which it would be beneficial to couple the data collection and recording processes. The benefit has three possible dimensions: (1) increased speed of recording and therefore earlier completion of the activity and dissemination of results; (2) cost savings due to elimination of intermediate data recording steps; and (3) increased quality of data by using the same person to collect and record data. The last benefit is the most intangible and least measurable quantitatively, but may be the most important in some activities.

were to be installed in the country, service and maintenance calls would be prohibitively expensive because of the remoteness of the computer system from service and maintenance centers. Prior to microcomputer technology, it was easier and more efficient to process such a census in another country than to process it locally.

The following countries are some of the earliest that have used microcomputer technology to process their population census data:

Cape Verde Islands (pop. c. 300,000). Cape Verde was the first country to use microcomputer technology. A Billings Microcomputer system was delivered to Cape Verde in 1979 to begin preparations for the census[1]. Four additional systems were delivered in 1980, and were used for data entry and editing operations. Two multi-user Onyx systems completed the installation in 1982 to prepare the tabulations and print them for publication.

Cook Islands (pop. c. 23,000). A single user Altos microcomputer system was installed in 1981 to process the population census data.

Comoros Islands (pop. c. 420,000). Two multi-user Altos microcomputer systems were installed at the beginning of 1981 to process all phases of the census data processing.

Sao Tome and Principe (pop. c. 90,000). Two multi-user Altos computer systems were installed early in 1982 to process all phases of the census data processing.

Laos (pop. c. 4,200,000). PC-compatible Honeywell-Bull microcomputers were used in 1984-85 for all phases of the census data processing.

Burkina Faso (pop. c. 8,000,000). The 1985 population census was processed by 28 IBM personal computers of different sizes and capacities, using Bernoulli cartridges and magnetic tapes for large file storage.[2]

Other countries have applied microcomputer technology in the past to portions of their census data processing activities. Furthermore, a number of countries such as Cameroon, Ethiopia, Malawi, Niger, Senegal, and Swaziland, *inter alia*, are now employing microcomputers to assist them in this area.

As microcomputer technology has improved, it has been possible to use the technology in larger countries but also for more functions. The following sections describe the census data processing applications that are now feasible using microcomputers and that should be considered by countries wanting to exploit the technology for processing population census data. Chapter 6 discusses additional activities that are just becoming feasible or will become feasible in the near future.

Census Management

Conducting a population census is an enormous task, regardless of the size or degree of development of a country. A successful population census demands involvement on the part of all of its inhabitants. Substantial organization and resources are required to bring about such involvement in a structured manner, on a given timetable, and to produce useful results. Microcomputer technology can help substantially to increase the efficiency with which the managerial and organizational aspects of the census process are performed.

Project planning. Complex activities such as a population census consist of a large number of smaller tasks which are interdependent in many ways. Each smaller task depends upon predecessor tasks being completed, consumes resources, and produces one or more outputs. Tasks must be scheduled in such a manner that their preconditions are met and resources are available. For all tasks together, sufficient resources of all types – human, physical, and financial – must be available to permit the overall census activity to proceed on schedule.

A variety of microcomputer based project planning and scheduling programs now exist that can automate a large fraction of the planning activity. Most such programs provide for different views of the process, providing schedules of required resources, time line charts, task dependency tables, cash flow analyses, and critical path charts. More useful programs generally provide good graphical and pictorial output that is easily understood and digested by planners and operational administrators.[3]

The use of such a program has several advantages. First, it provides a relatively comprehensive framework into which concrete assumptions can be placed, and within which the implications of the assumptions can be measured. If there are different assumptions regarding the productivity or delivery of a resource, then the planning program can be used easily to provide the implications of the two different

[1] Microcomputer systems made by Billings, Altos, and Onyx based upon Z-80 microprocessor chips and 64KB to 256KB memory sizes were used in the early 1980s in large part because of their support of the OASIS operating system. These systems were specified before IBM entered the microcomputer market and before the MS-DOS operating system existed. Since then, MS-DOS dominance of a large part of the microcomputer market has encouraged convergence to it as a standard operating system environment in which to work. Of the above three computer firms, only Altos still exists.

[2] A comprehensive retrospective description of the use of microcomputer technology in the Burkina Faso census is contained in "Traitement des Données d'un Recensement par Micro-ordinateur," United Nations Statistical Office, Project INT/88/P09, June 1988.

[3] Two examples of such packages are Harvard Project Manager on the IBM PC and Claris' MacProject on the Macintosh. There are many other such programs available on the market today.

assumptions, including whether timing has been modified for the critical path through the entire process.

Furthermore, the planning model serves as a basis for keeping the plan for the activity up-to-date. Complex activities rarely proceed exactly as planned; deliveries are late, unexpected difficulties arise, and some work may be completed ahead of schedule. By continuing to modify the plan, updated plans, schedules, and resource requirements can be generated easily from the revised plan and can be used to inform participants in a uniform manner about the current state of the project.

Finally, a microcomputer based planning tool can be used to provide a more accurate retrospective assessment of the process. Each time the census plan is changed, a new version of the schedule is created, with different resource requirements, possibly new activities that were not initially thought of, and new relationships between the simpler constituent tasks. Each revised plan and schedule contain more historical information about the course of the census to that time, and the final version of the plan will reflect the entire course of the project. By comparing the initial plan with the final plan, a relatively accurate retrospective assessment can be obtained about the adequacy of the original plan, both structurally and in terms of resources required. Such an assessment can be a valuable tool in assessing planning processes in general in a statistical office, and can lead to better, more realistic plans for future statistical activities.

Geographic data base. For the purposes of census enumeration and organization of reporting census results, the geographic area of a country is subdivided in a hierarchical manner into smaller and smaller units. The smallest of such units is generally called an *enumeration area (EA)*, corresponding to the area which is assigned to a single enumerator for data collection. While each country has a slightly different method of geographic disaggregation and naming of the units at each level, the practice of hierarchical disaggregation is required for both substantive and managerial reasons.

The enumeration area is the basic unit of aggregation for census results. Data from enumeration areas are further grouped and the results aggregated to provide information about larger areas of a country. The enumeration area is also administratively important. Enumerators are hired and assigned to enumeration areas. Enumerators require certain kinds of training prior to enumeration. Questionnaires, and sometimes additional resources such as fuel or transportation, must be provided to them prior to enumeration day. After the enumeration has been completed, the questionnaires must be returned to a processing center at a higher geographic level.

Information such as this regarding the geographic organization of the country is a basic data collection that can benefit from being created and maintained in computer form. Such a collection is often called a register; in this case, it is a register of enumeration areas. Such a register should include a record for each enumeration area in the country; this record should include information regarding the geographic hierarchy so that information about areas can be aggregated to produce information at higher levels, including the entire country. For example, the record for each EA could contain a field for the initial count of persons and families enumerated; the initial and provisional population count could then be obtained by adding all such values as soon as all EA initial count data were available.

Either the EA register or a duplicate of it can be used for administrative aspects of the census that are applicable to the EA level, such as the hiring and training of the enumerator, payment for services, and delivery and return of questionnaires. The enumeration area register may be kept on a microcomputer using any one of a number of data management tools and techniques. Relatively straightforward data base management systems are one appropriate tool for this function.

It is very important to build such a register of enumeration areas with which the processing of enumeration area files can be directed and controlled. In processing a population census of any significant size, it is relatively easy for operational mistakes to be made which result in the inclusion of a particular area either more than once or not at all in the final population microdata file. The importance of having an effective control mechanism for tracking enumeration area data accurately through the various stages of its processing cannot be overestimated.[1]

A computer based EA register provides significant benefits. First, it provides an official list of all data collection areas in the country, with officially designated names that can be used to cross reference other data, such as administrative responsibilities and maps.[2] Further, such a list will be useful and often necessary to provide labelling information for tabulation program output. In addition, such a register can be used to maintain the current status and location of data collected from the enumeration area for

[1] In the early 1980s, the International Statistical Programs Center of the U.S. Bureau of the Census distributed a program, CONTROL, that provided many of the functions for establishing and managing such an enumeration area data base. CONTROL was written to run on an IBM 370 type mainframe, and served as a prototype program in this application area. ISPC is developing a successor to CONTROL that will run on IBM-compatible microcomputers; such a program would be an important part of their IMPS (Integrated Microcomputer Processing System) package currently being produced.

[2] Assembling such a list of enumeration areas is not a trivial task, and the time to do it must be included in the overall plan and time schedule for census activities. The composition of the list will change, perhaps substantially, between decennial censuses. Naming of enumeration areas may be subject to certain national and local procedures. If the national alphabet cannot be represented directly in the computer system that will be used to process the census data and print the results, substantial translation or transliteration of these names into an alternate character set will have to be performed.

retrieving both physical questionnaires and data in machine readable form.

Finally, such an enumeration area control file can be used as a convenient way of providing initial aggregate statistics for the country as a whole. For example, adding fields for initial counts by sex and by gross age ranges would provide a quick capability to produce initial population figures in those groups. Such information might be entered as the initial batches of questionnaires and batch summary statistics were received from enumerators and field supervisors, providing a record of receipt of the forms.

Cartographic data base. Almost all developing countries maintain a collection of maps for census enumerations. Often these maps must be updated significantly prior to a census enumeration.

Microcomputers can be used to create and maintain an index of the collection of maps. Data can be added to the enumeration area register to point to the most current map of the area. The registry of maps can include the year in which the map was last updated, as well as other characteristics. Some maps may have been created for multiple purposes, or may display information beyond what is required for census enumeration; it may be useful to include such information in the map registry so that it can be retrieved electronically.

Logistics. Although part of the overall planning process, some logistical operations deserve individual attention. In particular, the delivery of essential material to support the census can be a significant problem in some countries. Such material includes questionnaires, vehicles, and petrol, for example. Microcomputers can be used to create and maintain requirements lists, delivery schedules, and inventories. If the situation is complex, they can be used to construct submodels of the delivery process that with some programs can then be integrated into the overall computer based model of the plan described above.

Personnel control and payment. A substantial number of people are involved in the census taking and processing activity, many on a temporary basis for the duration of one or more large tasks. Computers can be used to register the names and personal characteristics of the workers and to compute their net payments. They can also be used to project the rate of payments for some activities, such as data entry, with the percentage of the work completed (from use of the EA register to track data entry progress, for example) and with the amount of time and money allocated to the task (obtained from the computer based plan). If all of this information recorded in microcomputer files, it can be brought together fairly easily to provide feedback for judging how the overall process is being executed.

Cycles of Data Collection and Processing Activity

There are three cycles of data collection and processing activity a

Pilot census. The purpose of a pilot census is to test all aspects of a planned census, including data processing operations, for one or a few very small areas in the country. It is a prototype, or forerunner of the full census. The results of the pilot census provide important feedback to planners, and allow them to modify procedures to improve the administration and operation of the full census, including taking measures to improve the quality of the final data collected.

Pilot census operations provide an important test for data processing procedures. Even though the bulk of data processing activities are performed after the enumeration as been completed, census offices should regard data processing as an integral part of the pilot census test. Because the volume of data collected is intentionally small, it can be processed relatively quickly. Using a processing strategy based upon microcomputers helps; no longer must a large computer be installed prior to a pilot census, but rather one or a very few microcomputers can now be used to process the limited data.

Other benefits accrue from processing the pilot data promptly. The process of obtaining and installing the initial microcomputers can be tested prior to the enumeration. If the computers are IBM PC-compatibles, their compatibility with the processing software selected can be tested at a time when there are not substantial pressures to proceed with a heavy workload. If substantial difficulties are encountered in processing pilot census data, both the financial and the cost in time of having discovered them will still be moderate. Likewise, if any of the initial choices of processing software do not meet prior expectations, there is still time to obtain and evaluate an alternative strategy before the full enumeration begins.

Long and short forms. Population and housing censuses generally use two collection instruments to gather data, called the long and short enumeration forms. The short form contains the basic set of questions that are asked of everyone, and is used to enumerate a very substantial majority of the people and housing units in the country. If use of a long form is a part of a country's census strategy, it is used to collect additional information describing a relatively small proportion of the country's inhabitants and housing units. The long form should contain all of the questions included in the short form, with no changes, so that this core or common information may be used to compute information covering 100% of the population and housing units.

The combination of long and short form satisfies a country's needs for 100% coverage of basic information, while providing greater detail in some areas yet limiting both excessive respondent and processing burden. The presence of two different forms, however, must be taken into account in the data processing strategy. In particular, each of the basic stages of processing – data entry, data editing, and tabulation, must have both basic and extended stages so that the additional long form information receives appropriate attention.

Two basic strategies present themselves. The first is to process both short and long form information combined through all stages of processing. If this strategy is chosen, the various software modules chosen must be able to accommodate this difference and treat both types of records based upon their type. Software for processing complex surveys, such as ISSA (Integrated System for Survey Analysis) should have little trouble dealing with such a record and file structure, since it is simpler than many survey file structures. More elementary software will probably not be able to process such files.

The second strategy is to create two files, consisting only of short form records and long form records respectively. Then from the long form records, short form records can be derived since short form information is a proper subset of long form information. The complete short form file, consisting of the original short forms and the extracts from the long form file, will provide basic census information for 100% of the population, while the separate long form file can be used to obtain more detailed information on a subset of the population. Projections of this more detailed information for the entire population can be obtained by computing and applying weighting factors to each record based upon the sampling framework used to assign the distribution of long forms.

The second strategy permits less complex software to be used in processing the short form data, but requires that the operation of extracting the long form records and ensuring that the short form information remain in the short form file be carefully managed and controlled. Given the large number of basic enumeration areas being processed, inadequate controls could result in an improper mixture of record types on some files, the possibility of some long form records being removed from the short form file without leaving in their short form equivalent, or the possibility of some long form records being reintroduced in the short form file more than once. A formal control structure is necessary to ensure that each processing step required is performed on each processing unit in exactly the correct sequence. The enumeration area register, with extensions to provide information on processing steps taken, is perhaps the best approach to providing such a control structure. Such a control system can run on a microcomputer, and should be able to be queried interactively to determine the production status of the data for any enumeration area or work unit at any time.

Short form data results in relatively short records with a simple, two-level hierarchical structure, and can be processed by a number of census and survey processing programs as well as programs written primarily for statistical analysis. Long form data may have a more complex structure, and may benefit from processing by programs more adept at handling more complex record structures and data representations.

The choice of processing strategies may depend upon whether long forms are distributed more or less randomly throughout the country, or whether they are strongly clustered in specific enumeration areas or larger geographic aggregates. In the latter case, the task of separating the forms for different types of processing is simplified, and the two record types are somewhat easier to manage, since computer files up to some level of aggregation only contain one type of record.

Post-enumeration survey. Finally, a short time after the actual enumeration a post-enumeration survey is generally taken to assess the accuracy of the enumeration. The post-enumeration survey is generally a survey of moderate size. Its limited size often permits it to be processed by a more comprehensive survey processing package if its structure is complex and if there are other advantages in doing so.

Data Collection and Recording

There are a number of different technologies that can be used for collecting census data and recording it in machine readable form. They include manual recording of responses in the field with pencil or pen, mark sense documents that provide for automatic scanning and entry of responses,[1] and computer assisted data collection. The first two have been used in developing countries; the third method is still experimental and has long run promise.

Questionnaire design. Microcomputers can be used to assist in preparation of census questionnaires, using the technology commonly known as *desktop publishing*. This technology exploits the graphics screen capability that many microcomputers have, combined with medium resolution inexpensive laser printers and software that has extensive page layout capabilities.[2] Using such software, a forms designer can lay out a page electronically, using lines, boxes, graphic images, and text of various sizes and fonts, more easily than it can be done manually. Output printed using

[1] OMR (Optical Mark Reader) technology has been used effectively for population census data collection in the past even in some large developing countries. OMR products now exist for smaller scale data collection activities, and they can be connected to and controlled by microcomputers. Exploiting OMR technology on a large scale requires close attention to the quality of all phases of questionnaire preparation, as well as a field staff that is trained to handle the OMR forms properly prior to entry into a computer system.

[2] Examples of such desktop publishing software are Quark Xpress, Ready Set Go, and Pagemaker for the Macintosh, and Ventura and Pagemaker for the IBM PC and compatibles.

the laser printer is generally of sufficiently good quality that it can be used as copy for bulk printing of the questionnaires.

Page layout software allows questionnaire design to proceed iteratively. An initial version of the questionnaire can be created and copies circulated for review. Changes can be made to the page images that are stored electronically and new copy can be produced using a laser printer without any typesetting or other mechanical printing operations. The ease of revision is likely to lead to correct and more readable questionnaires, which can influence the quality of the data collected.

Desktop publishing software can be used to design and print questionnaires designed for manual or mark sense data entry. In the case of questionnaires that will be read by a mark sense scanner, some specialized assistance may be needed to lay out the form in accordance with the restrictions of the mark sense reading equipment. In addition, the type and quality of paper used for the eventual bulk printing will have to meet the tolerances of the equipment. However, the design and modification of the layout may be done relatively easily using desktop publishing techniques.

Page layout software also provides an effective mechanism to produce the internal operational control forms that are needed when dealing with the quantity of paper and machine readable information that characterizes the processing of census data. Word processing software will be sufficient for this task if page layout software is not available. Using laser printers, which are becoming increasingly inexpensive, attractive and readable forms are relatively easy to produce. This advantage goes beyond aesthetics; well designed forms can assist in minimizing operational errors and therefore improve data quality.

Data entry. For the 1990 round of censuses, the majority of countries are likely to use questionnaires that will be completed by an enumerator and which will not rely upon mark sense technology. The data from these questionnaires will need to be entered manually into computer form. Microcomputers are ideal machines with which to execute this process.[1]

There are a number of strategies that can be employed in using microcomputers for data entry. Strictly speaking, it is not necessary to use a specialized program for this purpose; a line editor can be used to create records from individual and household data by recording the information, character by character, in the appropriate positions of the line. In practice, however, this procedure is inefficient and error prone, and should not be used.

Another approach is to write specialized programs that will perform the data entry procedure for the specific task of recording data from a specific census. Such programs can be written in higher level languages such as Basic or Pascal or even within an applications program such as dBASE and may be efficient in entering complex information types. This techniques was used for entering census data into microcomputers prior to the development of specialized data entry programs, e.g. for the initial censuses processed by microcomputers listed above in this chapter.

A more recent possibility is provided by specialized data entry programs that are specifically designed to assist in large scale data entry. Two such initial commercial programs are EntryPoint and Rode-PC. These are comprehensive data entry programs that were used prior to the emergence of equally effective and more inexpensive alternatives.[2] The Blaise system at the Netherlands Central Bureau of Statistics offers a similar data entry module; two other programs have just been released by the U.S. Bureau of the Census and by the United Nations Department of Technical Co-operation for Development. All of these programs operate on an IBM PC or compatible system. These programs have the advantage that they are tailored to enter large batches of microdata directly, using operators having relatively high keying rates. Some of these programs require some basic program statements to be written, describing either questionnaire flow, error conditions, or both.[3] Some of the programs in this class also have a verification mode. Such a mode is important as a quality control mechanism on all data entry operators at the beginning of the process and selectively thereafter to ensure consistent quality throughout.[4]

Prior to the use of microcomputers, coding of some responses such as industry and occupation was performed manually. Coders would use a table of industry and occupa-

[1] It should be noted that previous data entry equipment bears some similarity to microcomputer systems. The IBM 3742 key to disk system was essentially a microcomputer with very limited built-in programmed functions. The IBM 5280 data entry system was a general purpose microcomputer with its own programming language, a subset of RPG-II, with which the data entry process could be programmed. Multiple key to disk systems, which were quite prevalent 10 to 15 years ago, used unintelligent key stations and placed the intelligence in a small shared minicomputer that maintained the disk files of completed work. These latter systems are somewhat similar to networks of personal computers that may be used to enter data.

[2] For example, CENTRY and ISSA offer data entry capability that is appropriate for population census data entry. In addition, each of these programs has the added advantage that it is integrated with other processing functions or programs that meet the needs of other stages of census and survey data processing.

[3] Some of these programs require the syntax of the statements to be supplied are Pascal-like and use only a few algebraic forms and control structures. The overall program environment and data structures are embedded in the program and are not a major concern to the user. For these programs, some familiarity with modern programming languages is sufficient to use the programs effectively. Other programs such as CENTRY and PC-EDIT require no programming skill, accepting the data entry and logic specifications in a more tabular form.

[4] The proper rate of verification depends upon the level of quality desired and the accuracy and skill of the data entry staff. In some countries, an appropriate rate of verification might be 100%.

tion categories to determine numeric values for the enumerator's written response. Often, such lookups would be done at the same time that other data from the original census questionnaire would be copied onto coding sheets in preparation for data entry. In one Latin American country, coding was performed conventionally, but as the industry and occupation codes were entered the computer retrieved and displayed to the data entry operator the text corresponding to the numeric code. While such mechanisms are not commonplace now and involve an additional expense in computing resources, it is possible that in the future such coding assistance could become routine and lead to an improvement in data quality.[1]

Error correction philosophy. A significant operational issue in statistical data entry is when to detect and correct perceived errors in data. Some recorded values are almost certain to be errors, such as a sex code not corresponding to male or female, while others are open to substantial doubt, such as a 14 year old child enrolled in a university.

The conventional wisdom with respect to this issue is that for population censuses, the volume of data, the time pressure for producing complete national results, and the generally temporary nature of the data entry work force argue strongly for delaying error detection and correction until after data entry and accomplishing it by automatic methods.[2] The cost that is paid for adopting such a policy is that errors cannot be corrected by an examination of the source questionnaire, with a resulting (invisible) loss of data quality. The reason that it has been possible to pay such a price is that the error rate is generally low, the volume of data is large, and it is believed that current methods of error correction have acceptably limited bias.[3]

Microcomputer based data entry programs have the possibility of improving upon this situation. The improvement requires that, even though the erroneous value has been recorded in the data entry process, the resulting record is flagged either automatically or manually as having a possible error. An exception verification stage can be added for each batch; the batch is verified but only those records that are flagged by the data entry process as having a possible error are verified. If this exception verification procedure is performed by an operator knowledgeable in the substance of the census data, with the original source questionnaire available for reference, it should be possible to supply correct values for the majority of erroneous values.

Data control and the unit of processing. Data entered into a microcomputer are most easily and inexpensively recorded onto diskettes, which can contain between 360,000 and 1,4000,000 characters of information. A diskette can hold information contained in a number of files, which often correspond to data from a batch of questionnaires or all of the questionnaires from an enumeration area. A rapid data entry operator in a developing country will take approximately two days to enter enough information to fill a diskette.

Diskettes are a convenient form of initial storage for data; they are inexpensive, removable, and can generally be transferred from computer to computer.[4] In the absence of other considerations regarding where the computers will be used after the census processing is completed, systems with diskettes only are an economical choice for data entry. After the data have been entered, they will (eventually) be transferred to other storage media, such as tape cassettes, computer compatible 1/2 inch tape drives, large fixed disks, or higher capacity removable disk cartridges such as those used by Bernoulli boxes.

A reasonable policy is to store exactly one enumeration area's worth of data on one diskette until the entry, verification, and perhaps some of the editing steps have been performed and are complete. During this time, the diskette can then move physically with the questionnaires for the enumeration area until there is no further need for or gain in referring to the source documents. After that, the data may be merged with data from other enumeration areas and aggregated into new files on diskette or some more capacious medium.

The aggregation of enumeration area data into *work units* or *data processing units* is an important part of census data processing strategy. The need for aggregation arises because of the physical number of diskettes that would be required if the data for each enumeration area were to be stored on a unique diskette. For example, the data for a census of a country of 10,000,000 people might require 10,000 diskettes, assuming that there were about 1,000 people in an average enumeration area. While the cost of the diskette media itself is not prohibitive, maintaining the

[1] It might be possible in the future to enter part or all of the enumerator's written response and have the computer display the full occupation or industry description and enter its numeric equivalent automatically.

[2] The strategy embedded in the ISPC CENTRY package reflects this orientation to doing little significant editing at the time the data are entered.

[3] For example, a common methodology for dealing with imputation of missing or erroneous values, the "hot deck-cold deck" methodology developed at the U.S. Bureau of the Census, is capable of preserving the variance of all distributions used to create the imputation functions.

[4] In reality, the situation is somewhat more complicated. There are three different sizes, or form factors, for diskettes: 8 inch, 5 1/4 inch and 3 1/2 inch. The 8 inch form factor has largely disappeared, while the other two sizes are both heavily used. Of the two, the larger size is the oldest, and has lower storage capacity and a less rugged casing. The larger 5 1/4 inch diskettes and drives may be preferred if there is reason to maintain compatibility with existing equipment of that size. New investments should be channeled toward 3 1/2 inch diskette media and drives. In addition, for each of these sizes, there are a number of different ways to format the corresponding diskette, not all of which are compatible with each other. The 5 1/4 inch diskettes can hold either 360,000 or 1,200,000 characters of information, whereas the 3 1/2 inch diskettes, formatted for use on MS-DOS computers, can hold either about 720,000 or 1,440,000 characters and, formatted for use in the Apple Macintosh, can hold either 400,000, 800,000, or 1,600,000 characters. While the higher capacity formats allow more information to be stored, they are not essential for successful operation.

data on so many distinct units of physical media creates a large operational control problem. The work unit is an aggregate of data from several enumeration areas that is created to make the data processing operation more manageable and more convenient.

The optimal size of the work unit chosen depends upon a number of factors. For all processing steps that require physical access to the source questionnaires themselves, the enumeration area is generally the most convenient work unit due to the bulk and weight of the questionnaires and the convenience of being able to associate one diskette or one file with specific sets of paper questionnaires.

The stages of data processing beyond data entry and possibly initial editing are more automatic. For these stages, the work unit is likely to be defined by the storage capacity of the medium used to contain it. Typical choices available to developing countries include:

- *diskettes*, containing from 360,000 to 1,440,000 characters per diskette;

- *cartridge diskettes*, containing from 5,000,000 to 20,000,000 characters per cartridge;

- cartridge tape drives; generally containing up to 40,000,000 characters per tape; however smaller files of information generally must be moved onto a microcomputer fixed disk for processing; and

- standard 1/2"magnetic tapes, containing 45,000,000 characters (or up to about 160,000,000 characters at the highest current recording density).

Enumeration units may be aggregated in such a way that work units comprise computer files that occupy most of the space that will be available for processing, leaving some slack space for additions and changes to each file. If the largest external storage medium available is the diskette, then the work unit should occupy most of the space of a diskette. If larger external storage media are available, then a reasonable work unit might consist of a file of about 10,000,000 characters.

The precise strategy governing the choice of work units for a specific processing load and environment is more complex than can reasonably be addressed here. Considerations involve cost, physical volume of media involved, fixed disk capacity and speed of microcomputers that will process the work units, and editing and tabulation strategy. Many developing countries may wish to seek external advice on this matter, given the significant operational payoff in making a good choice.

Just as it is important to control precisely the status and location of enumeration areas being processed, it is equally important to maintain control over the content of each work unit and the processing stages that it undergoes. For this purpose, the enumeration area register should associate with each enumeration area the identification of the work unit which contains it. Another register of work units should be created to track the progress and location of each of the work units. The work unit register will be smaller in size, since there will be less work units than enumeration areas, possibly many less.

Data being entered, verified, and changed through editing, should be backed up with regularity. The amount of work invested in enlarging and correcting a machine readable file of data should not exceed what the data processing operation can tolerate to lose in the event of a computer or operational problem. At the enumeration area level, such backup is best accomplished by regularly copying the current contents of the data file either onto another diskette or onto a fixed disk contained in the same machine. At the larger work unit level, backup may require the use of another medium such as cartridge diskette or magnetic tape. Backup procedures such as this need to be made an integral part of the work of data entry and editing.

If the data entry systems are organized into a computer network, then movement of the data can take place through the network links. A typical computer network contains a dedicated microcomputer system, usually called a file server. All of the microcomputer systems in the network are connected together by cabling that allows data and programs to be transmitted between any two systems at relatively rapid speeds.[1] In practice, much of the transmission is between the file server and the individual workstations. In such a configuration, workstation operators can transmit files between the file server and their workstation, obviating the need to physically carry diskettes between machines and locations. Depending upon how such a network is configured and managed, it is possible for any workstation to access any file of data on the server, thus allowing for a more flexible operation.

While it is not necessary to construct a computer network for processing census data, there are advantages in doing so if the resources are available and the network hardware and software technology can be supported locally. In particular, with appropriate data base software, it is possible to keep the registers of enumeration areas and work units together with the processing status of data from each, on commonly accessible files.[2] It is also possible for a data

[1] Typical transmission speeds that can be achieved inexpensively using ordinary twisted pair wiring range from 30,000 to 500,000 characters per second.

[2] Appropriate software in a networked, or multi-user environment, has the property that any set of authorized workstations may have access to a specific data file in order to read it, but only one workstation, if any, has the power to change the contents of the file at any one time. This prevents the possibility of inconsistent updates being performed concurrently. More sophisticated such systems provide locks against concurrent updating at the record or household level, rather than locking the entire file.

entry operator and a supervisor to examine the same file at the same time from different workstations and discuss problems that may occur.

Data Editing

Microcomputer systems provide a good computing environment for performing statistical data editing. They can be used effectively for exploration and construction of appropriate edit rule sets and for performing the large scale editing of all census records. They owe their suitability to several factors: (1) they provide a natural interactive environment which can be made relatively friendly to the user; (2) they are as powerful individually as many of the minicomputers and small mainframe computers that were used for these tasks 10 years ago; and (3) large scale editing of this type can be decomposed into smaller tasks that can be executed in parallel.

Determining appropriate edit rule sets. Formulating editing rules for computer application is generally a complex process that requires considerable substantive work. Such a process depends upon experience in subject matter areas such as demography, education and labor markets as well as country and region specific knowledge.

While the formation of the set of census data edit rules may begiven substantial time and resources, the testing and convergence of the set of rules to its final form may be more neglected.[1] This may be due to the difference in the two processes; creating the initial rule set relies much more upon substantive knowledge and very little upon computer processes. Refining the rule set, on the other hand, relies upon having useful computer based tools to apply the rule set to the data and analyze the outcome of such applications.

Interactive environments such as microcomputers offer the possibility of improving the process of refining the set of rules, leading to an improvement in final data quality.[2] Such a procedure would be based upon a machine coded library of potential data edit rules, together with a program that would retrieve a set of rules, apply them in some order to a test sample of census records, and display enumeration units that failed one or more rules and identify the rules that were failed. The subject matter specialist could then determine if the failure condition correctly identified a data condition that should be corrected, or whether the rule incorrectly predicted an error. Such a process should lead to a slow evolution of the initial set of rules into a set with more statistical power in separating units with correct data relationships from those with relationships to be corrected.

Existing edit programs can be used for such iterative edit rule refinement, although they are generally not well suited for it. Another approach would be to write a skeleton program to read the data from a family or household into memory and display it, and code the trial edit rules in the same programming language as the skeleton program. Selection of the edit rules would then consist of collecting all of the program source code in the correct order, and translating and executing it. This approach has the advantages of minimizing the overhead of the process and being adaptable to change; it does, however, require that the final set of edit rules be completely recoded to match the specifications of the program that will be used for the full scale error detection and correction process.

Performing large scale editing. Once the edit rules have been made final, they need to be applied to the entire population census data file. In the past programs such as CONCOR that have performed such editing were written and implemented on mainframes and minicomputers. With the growth in capability of microcomputers, such programs have migrated from larger computing environments to microcomputers.

Larger scale computing environments have had an advantage for performing census data editing in that they can easily be configured with large storage devices. Magnetic tape is used to store data sequentially; by using multiple reels of such tape, data files of arbitrary size can be accommodated. Magnetic disks allow large amounts of data to be stored in a randomly accessible fashion and are also quite useful for processing very large collections of data. In the past, error correction and detection rules have been applied either to entire population files or to very large subsets of population files.

Microcomputer systems can overcome this relative disadvantage in two ways. First, it should be recognized that storage devices being made available with microcomputers are increasing in capacity at least as fast as the rate of increase of such devices for larger computers. The relative disadvantage is decreasing. For example, it is now possible to obtain an industry standard tape transport[3] that can be attached to a microcomputer such as an IBM PC-compatible machine or an IBM PS/2 system, with software which will allow the data on the magnetic tape to be read and written under control of an application program. Fixed disk drives with capacities of 50-300 million characters of information are available and prices are dropping substantially. Other

[1] Sometimes the formation of data edit rules does not receive the attention that it deserves in view of its importance in improving the overall data quality and consistency of the ultimate census outputs. Microcomputer based software may make it easier to test the implications of a set of data edit rules, making it easier and more attractive to devote more time and subject matter input to this area.

[2] The advantage offered in this respect by microcomputers is shared by minicomputer and time sharing technology, both present and past. One reason that this approach has not emerged in minicomputer environments is because census data processing methodology has been based historically upon massive automatic batch processes.

[3] The term industry standard refers to the mainframe industry and its long term standardization upon 1/2" wide magnetic tape, about 2400 feet long, with 9 data recording tracks, at densities of 1600 and 6250 bits per inch.

devices are available that provide different solutions to this problem.[1] Thus, it is possible to process large data files, perhaps segmented, using currently available microcomputer hardware, and the situation will continue to improve.

Second, the process of statistical error detection and correction is capable of being divided into parallel processes and executed concurrently on a number of independent microcomputer systems. However, some caution is in order due to the nature of the editing process, which generally consists of an intermixed set of deterministic correction procedures and stochastic imputations. Deterministic rules are exact; they specify that if a certain combination of precisely defined circumstances occurs, a specific action is to be performed. Stochastic imputations are triggered by specific conditions, but the resulting action is not deterministic.[2] Instead, it is chosen from among a number of possibilities based, *inter alia*, upon the order in which the enumeration unit data appear in the data file.

If the editing process consists of entirely deterministic correction procedures, then there is no inter-unit dependency in what corrections will be made, and the correction procedures can be applied to the population units in any order and however they may be aggregated into work units. Under these conditions, the file may be split (or not combined), and independent identical editing processes can be executed in parallel on multiple computers to edit the data. If stochastic imputation is involved, then the results of independent editing of subfiles will yield numerically different results than editing the entire population file sequentially. Care must be taken to ensure that the consequences of a specific method of imputation are understood no matter how it is applied, but the possibility of using independent editing procedures raises additional possibilities which need to be understood before committing to such a strategy.

The choice and size of the processing work units chosen may have an effect upon the results of stochastic imputations performed. In particular, if a "hot deck, cold deck" approach to imputation is used, then valid values used for imputation will be taken from individual and household data contained in the same work unit This process has implications for the validity of the resulting corrected microdata. It is therefore important that subject matter specialists who understand geographically related differences in the country examine both the strategy underlying the formation of processing work units and the application of imputation procedures.

Data Tabulation

Historically, the results of a population census have been displayed and disseminated by means of publishing a wide variety of tabulations of the collected and edited population microdata. This method of dissemination has reflected the usefulness of such a collection of cross-section tables describing a population at a specific point in time, as well as reflecting the relative advantage of batch data processing methods in supporting this approach.

The creation of census tabulations can be approached in several ways using microcomputer systems. First, many of the tabulation programs that were used in large computer environments for producing such tables in earlier years are now available for use with microcomputers. This migration of programs downward is in reality more of a measure of the upward evolution of microcomputer environments to increasing levels of power and capacity.[3] It also reflects to some extent the history of such programs. COCENTS, a precursor of CENTS, was originally written for mainframes of only modest size because it was targeted for use in developing countries. XTALLY was originally written for the IBM System/3 and System/32 which had a capacity equivalent to early microcomputers. However TPL, a large and comprehensive tabulation program targeted to a very rich and developed mainframe computing environment, has also been ported successfully to microcomputers.

Second, microcomputing environments consisting of a number of similar computer configurations can be effectively used to generate tabulations without any regard to existing software. The process of tabulation is aggregative,

[1] Removable cartridge diskettes, sometimes referred to by example as Bernoulli box diskettes, hold 10 to 20 million bytes of information and have been used in a number of developing countries for census and large survey processing. Cartridge tape drives, such as QIC 1/4" standard drives, have capacities of 20 to 100 million characters, can be used for backup of fixed disks. Many QIC units can also be used to provide file by file backup and restore of data files, thus providing an indefinitely expandable reservoir of off-line storage for census work units. The emergence of digital casette tape storage, the computer equivalent of digital audio tape, will occur when the current controversy in the consumer electronics market regarding copy protection for digital audio tape is resolved. WORM (Write Once, Read Many times) optical disk technology is stabilizing; optical platters for the smaller versions of these drives can store 200 to 400 million characters. Erasable optical drives are just now becoming available with capacities of 200 to 600 million characters; this technology may be sufficiently reliable for use in developing countries within 2 to 3 years.

[2] Stochastic imputations depend upon sampling from a distribution of one or more variables and assigning the value(s) selected to the variable value(s) to be corrected. If the distribution is sampled using a random number generator, then a different set of values will be selected for different orderings of the population data file; the characteristics of the distribution sampled, however, will be preserved. Likewise, if a "hot deck-cold deck" approach is used, then the values in the imputation matrix at any time will depend upon the ordering of the previously processed cases that were judged correct and which therefore contributed their valid values to the matrix.

[3] Specifically, microcomputer systems now support essentially equivalent language translators for Cobol, Fortran and other common production languages as do larger computers. The recent growth of mass storage capacity for microcomputers allows sorting of relatively large files, which is often needed by tabulation strategies and is used extensively by CENTS and TPL. Still missing is the ability to support sorting of files of arbitrary length using sequential magnetic tape technology. The need for such a capability, however, is less than it was in earlier batch environments because of new capabilities in the microcomputer environment which were either not present or not exploited earlier.

and proceeds according to the hierarchical geographic organization of the country that is used for census data collection and reporting purposes. The generation of a specific table for an enumeration area is easily performed on a microcomputer. The microdata involved occupy only a part of a diskette, the amount of computation required is modest, and the amount of output is small and is capable of being printed on one or at most a few pages.[1] Such tables can be generated in parallel on sets of enumeration areas, and the results can either be written onto additional diskettes or accumulated on a fixed disk.[2] Larger work units may be used in such a process if each work unit consists of a collection of complete enumeration areas.

The second part of such a tabulation strategy could be to aggregate the tables generated for each enumeration area according to the definition of the hierarchical subdivision of the country. This definition should be contained in the machine readable register of enumeration areas in such a form that an aggregation program could use it automatically. In principle, once a particular table has been generated for each enumeration area, it can then be generated and printed for all other levels of aggregation defined, including the entire country.[3] This step is both conceptually and operationally simpler than the initial step of preparing the tabulations at the enumeration area level.

At present, no general purpose computer software packages exist that implement the above conceptual approach to generating census tabulations, although specific programs have been written in some countries that follow similar strategies. In general, the strategy of using existing tabulation programs has been more widely adopted. Such an approach has provided continuity of data processing methods and use of existing knowledge within census and statistical offices. Such approaches, having evolved from earlier computing environments, also tend to be operationally simpler and generally more efficient. Nevertheless, the approach of tabulation at the individual enumeration area level followed by aggregation of tables may be useful for certain countries having special circumstances or having an interest in developing it.

Microcomputers may be as or more useful than larger computing environments for producing *ad hoc* tabulations on demand after the completion of the original program of table generation and publication. For moderate size countries, the original microdata are likely to be stored on diskettes which can be accessed manually and processed further.[4] Given a tabulation program that recognizes the file format of the census microdata, it is as easy to generate special tables as it was to generate the original published tables. Having manually accessible data through diskettes has other advantages, such as being able to distribute appropriate microdata to government offices at lower geographic levels for local processing in response to specific local needs. This last step depends, of course, upon maintaining at the local level the confidentiality and privacy requirements imposed upon the census process and required by law.

Preparation of Census Publications

The traditional form of dissemination of statistical results has been through publications. Traditionally these have been in printed form, although microform substitutes have been used in developed countries and for large publications. Although newer and considerably more powerful forms of dissemination are emerging, traditional statistical publication will continue, albeit in modified and possibly diminished form, for many years.

Statistical publications initially were produced using master copies in typewritten or typeset form. Both forms of production require transcription of the data from either manual worksheets or computer output, increasing the possibility of transcription errors. More recent methods include preparing camera ready copy directly on the computer[5] and preparing output from a computer in machine readable form that is then read as input by photocomposition equipment. Both methods eliminate the data transcription step; however, both methods are relatively inflexible and have other disadvantages. Direct photography of computer output sheets places a formatting burden upon statistical data processing programs that they are generally ill equipped to handle, while a computer supported linkage to photocomposition equipment must take into account the

[1] While the specific programming steps that generate the numeric results are relatively simple, such a program must also take into account formatting and labeling considerations, as well as how to handle exception conditions such as missing data or inapplicable observations. These additional considerations often dominate the amount of work performed in preparing such programs.

[2] Such a strategy would generally not be used in practice to compute a large number of tables for publication one by one for each enumeration area. It is worth noting, however, that such tables can be produced for each enumeration area using relatively simple microcomputer systems.

[3] Such an approach depends upon having sufficient fixed disk or other mass storage available to the tabulation system so that all of the input and output data can be accessed without manual intervention. If the space is not sufficient, then external storage will be required, and tabulation scheduling – probably through sorting tables – will be required. If this is required, the operation resembles more the approach taken by CENTS, which assumes such a situation and contains automated measures to deal with it.

[4] The extent to which the strategy of keeping and using data on diskettes for further use depends upon the size of the country and the amount of data collected for each respondent. In any case, a copy of the data should be stored in a different location. In all but the smallest countries, it will probably be stored on a different medium to take advantage of more compact methods of storage and lower storage costs.

[5] The preparation of camera ready copy directly on the computer had mixed results. On the one hand, it eliminated the data transcription process, which improved accuracy of the publications. On the other hand, it substituted a rough and relatively unattractive single font and single point size monospace format for previously typeset pages which were easier to read and were more economical in their use of page space. Many users may have perceived the change as negative because of the degradation of form, even though the substance was improved. Form does have importance; in the past, one may have had to make a tradeoff between them, but with the advent of newer techniques they may no longer be substitutes.

idiosyncratic nature of specific photocomposition equipment.

Microcomputers have made possible a new alternative, which is often referred to as *desktop publishing.* The technology depends upon two elements: (1) inexpensive laser printing technology which has recently been developed that provides reasonably crisp images of facsimiles of typeset pages; and (2) the emergence of page description languages, which allow pages to be composed in a fairly flexible and interactive manner.[1] Current standard laser printers produce page images at a resolution of 300 dots per inch, but printers with finer resolution exist and are decreasing in price as the market for such devices grows.[2] Laser printers can generally produce output with a standard page orientation (*portrait* format) or rotated 90° (*landscape* format) with equal ease. Combined with other image manipulation tools on microcomputers, the user has the flexibility of using page layout software that is already moderately powerful along with additional tools for publication layout.[3] Further, the content of the pages can be directly transferred from spreadsheets and data bases, as well as being generated by special purpose programs.

Microcomputer based publishing also allows both the creation and the publication of graphical material in an integrated form. Graphs are readily produced on microcomputers, either using procedures coupled to spreadsheet programs or independent programs. Using computers which have a common bit mapped base for both text and graphics such as the Apple Macintosh, integration of text, tables, and graphics is straightforward, as is the production of high quality pages containing all three types using desktop publishing techniques. The incorporation of graphics into statistical publications often adds to their attractiveness, readability, and ability to convey to the reader the meaning of the statistics reported.

A particularly interesting and powerful aspect of using computers which are based upon bit mapped graphics output is the capability to represent non-Latin national character sets. For example, it is relatively easy to define character set fonts for many national character sets and then intersperse them easily and flexibly within tables and mix them with Latin characters and variations thereof.[4] This capability is an integral part of the computer's architecture, in particular: (1) the integration of text and graphics at the most fundamental level through use of bit mapping for both screen and printed output; and (2) the ability to separate and redefine the size and shape of the graphic symbol associated with each key and combination of keys on the keyboard.[5] Publications prepared with such systems can therefore respond well to local needs when two or more different languages and/or character sets would assist readers.

Desktop publishing and image manipulation tools implemented on microcomputers are generally easy to use, and require no specialized computer training. Responsibility for the production of publications using such tools can often be assigned to statisticians and their colleagues, with appropriate levels of recommendations and consulting assistance from a central data processing staff. Desktop publishing technology is powerful and relevant for the publication of census results as well as for other statistical publications and deserves thorough consideration and adoption.

The Post-enumeration Survey

The post-enumeration survey is taken to assess the accuracy of the enumeration process. The enumeration area register can provide a frame from which the sample for the survey can be drawn, assuming that the basic population count information has been added to the register as it was computed, either by the enumeration field staff or by computer programs.

For some analytical purposes, it may be desirable to link the information collected in the post-enumeration survey on a record by record basis with the information collected for the same persons and households during the census enumeration. This may result in a file and record matching process that is uncomfortably large to perform on microcomputers at present. If the post-enumeration survey is clustered, then it will be easier to extract the original census data corresponding to the sample design of the post-enumeration survey and attempt to obtain the desired record matches.

[1] There are several heavily used page description languages available in the market today. The most widely used and a likely long term standard is Postscript, produced by Adobe Systems.

[2] As an example of what is possible with desktop publishing techniques, the contents of this publication were prepared and formatted using Microsoft WORD on an Apple Macintosh computer, using an Apple Laserwriter which produces printed output at a density of 300 dots per inch.

[3] The low cost leader in desktop publishing is Apple Computer using its Macintosh, although IBM has recently begun to offer such products with its PS/2 line of systems. Apple's lead in both desktop publishing and image manipulation tools is sufficiently great that it can be considered the vendor of choice wherever Apple support is available. Microcomputer data files can be transferred between MS-DOS and Macintosh computers with relative ease, so that the relative advantages of the Macintosh software environment can be exploited even if the majority of processing is performed in an MS-DOS environment.

[4] The development and display of non-Latin character sets is advanced and well integrated with other programs on the Apple Macintosh, but can also be performed on an IBM PC or compatible system.

[5] The 1-1 correspondence between keys and key combinations and graphic symbols places a limit on the number of characters that can be represented in this manner. Japanese and Chinese character sets and perhaps a few others cannot be handled in this manner. However, there exist systems more sophisticated than the Macintosh, also using bit mapped screens as user interfaces, which accommodate such character sets in a multi-font environment. Advanced work in this area is being performed by Joseph Becker at Xerox PARC in Palo Alto, California. The utility of the overall approach can be seen by noting that there now exist commercial fonts for many non-Latin character sets for the Macintosh available at relatively low prices.

APPLICATION OF MICROCOMPUTERS TO CENSUS DATA PROCESSING

Demographic Analysis

One of the first uses of the data collected through a population census is to analyze the demographic structure of the population. A variety of programs exist to perform such analyses.

Among the program sets that have been used for demographic analysis are two substantial packages of programs from the U.S. National Academy of Sciences and the International Statistical Programs Center of the U.S. Bureau of the Census. These programs were originally written for IBM mainframe computers in the early 1970s, using the Fortran programming language. Both sets have since been converted to run on IBM and IBM-compatible microcomputers. Since demographic analysis follows production of the published census results, some of the microcomputers used for processing the population census can be reassigned to a demographic analysis function or group and can continue to be used for analyzing the data.

Although the subject is beyond the scope of this document, there are other sources of programs for demographic analysis. The procedures underlying most if not all methods of analysis are within the scope of being implemented on current microcomputers. Newer programs are just now beginning to emerge that take advantage of visual interfaces to provide on-screen interactive two and three dimensional visualizations of demographic distributions.[1] As microcomputing environments evolve, programs to perform demographic analyses using microcomputers are likely to grow in number and become more powerful and easier to use.

Data Dissemination and Data Bases

The collection and processing of population data is, of course, not an end in itself, but represents the production of an intermediate product that is extremely valuable for national growth and development in many dimensions. By making the census data more accessible, its use can be encouraged. Creating data bases that are easy to access is an important step to making these data broadly usable. Computing technology can support the storage of population data in a form which is easily accessible and will encourage more widespread use.

Computer technology provides methods of census data dissemination and use in addition to a standard program of statistical publications. Although the collection, analysis and publication of statistical data on a piecemeal sectoral or cross-section basis can yield important information, the data are generally capable of much more useful exploitation. While the traditional and historical model of a statistical office as a collector and publisher of information still has some merit, the newer concept of an information service specializing in the collection, protection and dissemination of data capital has increasing validity and should be seriously considered by statistical offices and census bureaus.

Although the set of publications produced by a national statistical service is an early form of data base, albeit a printed form, the term *data base* as commonly understood now generally implies data stored in machine readable form and generally computer resident. The notion of data base is a general one in computing technology, and refers to an integrated set of data relationships, often complex, that are addressed indirectly through another set of data relationships, generally called a data dictionary.[2] The term data bank is an earlier term whose origin is more directly related to statistical data processing. The term is a vague one, referring to a significant or comprehensive collection of data in a specific subject area such as statistics, and has no implications regarding the structure of the data collection.[3] The notion of data bank has a somewhat more passive connotation than the notion of data base; yet another concept that is roughly synonymous, that of *information system*, connotes an active process orientation rather than an orientation toward storage and preservation. All of these terms are broad and cover a heterogeneous set of realizations of the concepts.

The terms *statistical* data base, bank, and information system encompass the notion of electronic manipulation and dissemination of subsets of the data collection. The existence of flexible computer based statistical data bases allows views of the data that are not related to the original purpose of the data collection. The addition of data collected from other sources, other ministries and other countries, can provide additional and comparative data for a statistical office. Manipulation of the data does not have to be confined to numeric form or to tabular presentation. Effective tools allow two and three dimensional graphs to be created and manipulated; data collections can be explored in the form of related collections of graphs. The geographical side of data can be exploited; much data collected by statistical offices has a geographic attribute and can therefore be located on a map and associated with other data measured at that and other locations. This last notion is now being developed and is discussed in a later section on future developments.

---

[1] In particular, the IPSS (Interactive Population Statistical System) from Bowling Green State University (Ohio, USA) provides an excellent example of how graphical user interfaces and point-and-click technology can be used to simplify greatly the task of the demographic analyst. Implemented on the Macintosh, the program provides for both graphical input and output, and an analyst can complete analyses considerably more rapidly and comfortably than with the more traditional batch and line oriented programs.

[2] Statistical data processing literature often refers to the contents of the data dictionary as metadata, that is, data describing data.

[3] In some countries having a significant concern with issues of individual privacy, the term has developed to have a somewhat pejorative connotation at times in the past because of an implication of loss of privacy through matching and analysis of individual data by government agencies.

Dissemination and use of machine readable results. In many countries, computer systems have been used to process census data for some time, and computers are now used for this purpose in almost all areas of the world. Much of this processing has occurred on large computer equipment. Dissemination of machine readable census data then required that the recipient of the data was also required to have similar large and costly equipment. In most developing countries, such a requirement often meant that there were no recipients who could use the data in this form; instead, requests were made to the census authority for special results which the authority was in a position to produce.

The availability of microcomputers has made machine readable data much more widely useful within all countries. Microcomputer systems that are capable of serious and useful analysis of many types of data are now available for less than U.S.$2,000, and their cost will decrease somewhat further.[1] Diskettes are an inexpensive and effective dissemination medium for statistical data.[2] To the extent that relatively stable electric power is available, microcomputers can be placed in other ministries and in regional and local offices for data dissemination and analysis, as well as data collection and editing. The same systems can be used for other work, including statistical and organizational administration. While microcomputers have not increased the amount of statistical information that can be disseminated, they allow the data to be disseminated widely and inexpensively in a form that is easily manipulated and used for analysis and for unanticipated purposes.

A wide variety of microcomputer based commercial software can be used to process population data on microcomputers. Among the useful programs in this regard are:

- *SPSS/PC*, a microcomputer version of the entire mainframe SPSS/X for performing statistical analysis;

- *PSTAT/PC*, a microcomputer version of the entire mainframe PSTAT for performing statistical analysis, and which has sufficient microcomputer processing functionality to be used for processing very large microdata samples;

- *Data Desk*, a Macintosh program embodying a powerful implementation of John Tukey's philosophy of interactive data exploration; and

- *SAS/PC*, a set of programs implementing parts of the large SAS mainframe based statistical analysis system.

These programs and similar ones are available for microcomputers at a moderate cost per copy. They can be useful in many ways, either with samples of census microdata or with aggregate data derived from the census microdata. Microcomputers used initially to process the census data can be diverted at the end of the processing phase to equip offices within the statistical agency having a need to perform data analysis. In some cases, the computers may need additional primary memory to accommodate the larger statistical packages. Such microcomputer based statistical analysis capability is now capable of substituting adequately for many if not most of the power that would be obtained from mainframe versions of the same software products (where they exist), generally with significant savings and often with greater ease of use.

Data bases consisting of aggregate tables. The amount of data collected by a population census is enormous. In almost all countries the original microdata are considered confidential and cannot be released in a form that would allow identification of any individual or family on the basis of inferences from the data. The challenge posed by this restriction is how a census authority can legally provide the most insight into the social and economic structure of its population. The problem is exacerbated by the fact that much of the data is of interest to local organizations, where the reduced population size makes potential disclosure more likely.

One method that has been used in the past has been the creation of a very large set of tables with different content at various levels of geographic disaggregation. Such a program would provide the largest number of tables with the greatest subject matter disaggregation at the national level, since it would be possible to do so and still maintain cell sizes sufficient to avoid breach of confidentiality. At the other extreme, tables for each enumeration area would be considerably limited both in number and in subject matter breakdown. Moving upward from the enumeration area to national totals, each unit at each level of geographic aggregation would be described by an increasing number of tables of increasing detail.

Such a strategy can lead to generation of a very large number of tables, more than can reasonably be published.[3]

[1] The cost of a microcomputer for reading and manipulating census and other statistical data is now less than 10 times the cost of a passive microform reader. Recording the data on diskette is easier, more flexible, and less expensive than preparing microfiche images of printed documents in developing countries.

[2] Statistical data released in machine readable form must be inspected to ensure that no national privacy or confidentiality policies are violated by its release. In particular, population microdata cannot be released without assurance that the records released cannot be identified with their source. In addition, detailed small area tabulations may indirectly disclose information about specific reporting units, and some technique such as random rounding may be necessary to prevent such identification. While the same precautions should be taken with published information, the ease with which large volumes of machine readable information can be produced and disseminated can sometimes lead to statistical organizations overlooking the non-disclosure issues implicit in its unmonitored release.

[3] For example, the original microdata collected during the 1970 population census in the U.S.A. could be stored on about 600 reels of magnetic tape. The set of tabulations generated for the aggregate data base,

APPLICATION OF MICROCOMPUTERS TO CENSUS DATA PROCESSING

Historically there have been such economies of scale in the tabulation process that it has paid to generate many more tables than are to be published and retain the entire set, appropriately indexed and therefore retrievable, in machine readable form. Such a collection may be considered a data base of population census tabulations. The data base contains a very large number of tables that may be requested, but it is not sufficiently extensive to supply information on all possible relationships. Only a strategy resting upon the use of the original microdata is capable of doing that.

Census authorities may want to consider generating considerably more tables than are planned for the census publication program, and then disseminating appropriate tables on diskette to the various regional and local governments and offices. If the tables were formatted in such a way that they could be read by either general purpose or special purpose spreadsheet and database programs, then the aggregated information could be further processed and analyzed at the local level. Such distribution of information in processable form is likely to make local data considerably more useful within local contexts.

Data bases consisting of microdata. The primary data base is, of course, the original population data collected by the census authority. This data base is rarely if ever released, but can be used by the census authority to provide special tabulations, extracts, and analyses as required by users with legitimate access to those results. Regardless of other methods of dissemination, this method should be made available by the census authority as a method of last resort. Often the computing tools used to satisfy such requests are similar or identical to those used to prepare published results or other byproducts.

The census authority may wish to restructure these microdata in ways that allow for more efficient and effective retrieval based upon the patterns of access to the data. Whereas in the past such files have been maintained in sequential mode, other file organizations that include transposition and inversion techniques, as well as use of relational data base systems, have proven to be very useful in large computing environments.[1]

One way of making census microdata available is to release controlled samples of the data. A number of countries now have programs of releasing 0.1% and 1% samples of microdata for research and instructional purposes within the country. As in the case of designing an aggregate data base plan, the larger the sample, the more concern there should be regarding breach of confidentiality, and the more likely that only a subset of the attributes of each respondent can be released.

Microdata samples can be provided relatively easily and efficiently for use with microcomputer systems. For moderate size countries, it is possible to provide useful microdata samples on diskettes. For example, if a country of population 10,000,000 supplies a 0.1% (1-in-1,000) random sample of 10,000 persons, basic short form information is quite likely to fit even on an older 5 1/4" 360KB diskette. A 0.1% sample of long forms of a country of 30,000,000 might occupy two to three high density 3 1/2" 1.44 MB diskettes, which is not an unwieldy amount of storage for microcomputer users. Even the long form data for a 0.1% random sample of a sizeable *developed* country of 200,000,000 can be stored on a fixed disk of between 20MB and 150MB capacity.[2] Disks of such sizes are now readily available for attachment either to individual microcomputer systems or local area network servers, and currently cost between US$500 and US$2000.[3]

In summary, microcomputers can be exploited by governments both to create data bases containing various forms of population census data and to disseminate it in a dynamic and usable form. Substantial improvements will occur in this area if software is developed for microcomputers that builds upon established statistical data processing methodology that has been implemented on larger computers for a number of years already.

Choosing a Processing Strategy

General considerations. Progress in technology and the emergence of a mass, low cost market for microcomputer hardware and software provide a wide and often bewildering variety of alternatives for statistical agencies. This trend will continue in the future. Choosing among

whoever, was stored on about 2,000 reels, an increase of a factor of more than 3.

[1] Statistics Canada's example in the 1970's is worth noting. In 1973, all of the short form Canadian population census data were stored in an innovative storage structure on-line on a large mainframe. A system, STATPAK, was written that was able to obtain *ad hoc* small area tabulations of any area in Canada within about one minute of elapsed time from when the request was entered into the computer until the results were displayed. This development resulted in the creation of the RAPID system, which uses relational data base software technology and can serve a central role in statistical data processing of aggregate, time series, and microdata. Until recently no software to perform this function in quite the same way existed on microcomputers. However, two new programs, ISSA and REDATAM, are now based upon these earlier data storage structures and can therefore in principle support such access methods and retrievals.

[2] In connection with any discussion of such public use microdata samples, it is important to note that dissemination of any such data must satisfy all privacy and confidentiality requirements of the national statistical agency. Countries that have considered distributing such samples have learned that such distribution is tolerable only if there is a strong public perception that individual data are kept confidential and are in no way compromised by the release of such statistical samples.

[3] There are limits to the amount of detail which can be obtained from processing a subsample of population data, and there are variances associated with any estimate obtained in this manner. The smaller the sample size, the larger the variance. Nevertheless, subsamples are quite useful in providing initial estimates and often final results because they can be processed quickly and locally on one or many microcomputers. Whereas providing ad hoc census tabulations was previously often a long process involving specialized programming and batching of results, the availability of subsamples on microcomputers with appropriate software makes this process easier and considerably more rapid.

these alternatives can be confusing. Furthermore, appropriate choices can shift quickly over time because the market for microcomputer hardware and software changes relatively rapidly. Finally, appropriate choices for developing countries depend further to some extent upon local support and the availability of bilateral and multilateral technical co-operation assistance.

The nature and extent to which a country's processing strategy depends upon microcomputers will depend upon the size of the country. It now seems clear that most small countries should plan to rely extensively if not wholly upon microcomputer technology for their census data processing. Larger countries may be better off committing some of their processing such as data entry, administration, control, planning, publishing, and demographic analysis to a microcomputer environment, while continuing to perform the more massive data processing operations on larger computers. The notion of what is a small and a large country is in this context dependent upon the the power of the underlying technology, and appropriate choices made in 1990 are likely to be very different from choices that will be appropriate to make in the year 2000.

The nature and extent to which a country's processing strategy shifts toward a new technology also depends upon the amount of risk that the country is willing to assume. While the use of microcomputer technology for processing population censuses is not inherently risky, there may be occasions when significant change of any kind may pose a risk to stability. Countries should assess the specific risks to them of effecting such a change, and compare those risks with the benefits that are likely to be realized by them specifically in deciding whether and to what extent to commit their census processing strategy to the use of microcomputer technology.

The following sections provide a view of what appears to be the situation which applies generally to the 1990 round of population censuses.

Target computing environment. The most prevalent microcomputer environment in the world is characterized by IBM Personal Computers or "clones" that are very compatible with them. The most prevalent and well understood and supported operating system that runs in this environment is MS-DOS, called PC-DOS for systems supplied by IBM. This processing environment should generally be selected as the target computing environment for the bulk of census data processing operations by countries committing in 1989 and possibly 1990 to the use of microcomputers for census data processing.[1] Countries should investigate with expert assistance the benefits of connecting their microcomputers using a local area network and how such a network could be supported, both technically and administratively, within the country.

For a small but significant set of tasks not directly related to routine processing of large data files, national statistical agencies and census offices should strongly consider investing in a few Apple Macintosh computers. There are at least two reasons for this. First, Macintosh computers enjoy a very substantial advantage in a number of areas which can be used to complement the use of IBM or IBM-compatible machines. Areas include graphical and cartographic outputs as well as text processing and desktop publishing. Second and more important, the user interface and software approach embodied in the Apple Macintosh operating system represent a very significant step forward in the way in which computing equipment can be used to increase productivity and functionality. National statistical agencies cannot afford to ignore an opportunity to learn about this technology on an initial but first hand basis, and in doing so, evaluate its potential for assisting them.

The hardware and software acquired by developing countries for processing a population census in or about 1990 will find widespread and productive use within the statistical agency for a number of years after the population census processing has been completed. However, national statistical agencies should not plan to reuse either the hardware or the software obtained for their current census for their next population census in about the year 2000. By that time, even if the hardware obtained now is still functioning, technical progress made in the intervening decade will have provided a significant new range of options that will be much more powerful and effective than what is available now and in the near future. Having committed to the use of microcomputers now, however, is an essential step to build the technical capability and understanding necessary within the statistical agency to exploit both the current generation of microcomputer technology as well as future generations whose shape is not yet known to any of us.

Software strategies. Countries need to choose a consistent set of software products that span the set of functions that they will need to perform in processing their population census data. There are three basic approaches to choosing and obtaining this set of software:

- development of special purpose software;
- selection of a variety of existing programs that, when used together, provide the data processing functions required; and
- use of an existing integrated package of modules that provides all of the functions required to support processing the census data.

[1] Countries having significant intellectual and technical infrastructure may want to investigate the possibility of using multi-user microcomputer configurations, often UNIX based, for some census data processing operations.

Strategies that mix elements from each of the above are also possible and perhaps the most practical, since there is at this time no one integrated system that can provide the functions that are needed for all aspects of census data processing.

Countries with substantial data processing expertise may have some interest in developing special purpose software, although existing programs may satisfy many of their needs. Developing special purpose software is generally a higher cost alternative than using existing software, and may represent a higher risk in terms of being dependent upon the continued employment of the individuals who develop the software. Nevertheless, there may be reasons why this strategy is sensible for specific aspects of census data processing

A strategy of selecting existing computer programs and providing the binding, or "glue," between them can be quite powerful and effective. The adoption of this strategy provides maximum choice, for there are hundreds if not thousands of microcomputer programs that could possibly be useful to statistical agencies in many fields of statistical data processing, including processing population census data.

In the process of implementing this strategy, agencies commit themselves to a substantial learning experience. In order to acquire a program, it and its competitors must be evaluated to make an appropriate choice. Once a program has been acquired, its actual strengths, weaknesses, and areas of applicability need to be determined. If it is to be used in conjunction with other programs, then the linkages between the programs including the way in which data are passed between programs must be studied and understood operationally.

Previous mainframe knowledge in the statistical agency may be helpful in making choices about the software pieces to be assembled. If, for example, the agency has used SPSS/X intensively on a mainframe in the past, then the cost of introducing and using SPSS/PC is low since the package and its control language are well understood by staff. However, such a strategy works against the selection of new programs which have no mainframe equivalent because they offer radically new and productive approaches to specific tasks. Data Desk is an example of such a software product.[1]

[1] Data Desk, written by Paul Velleman of Cornell University and distributed by Odesta Corp., is a recent highly innovative implementation of John Tukey's exploratory data analysis methodology. However, no mainframe program preceded it because such operations could not be implemented on mainframes at any reasonable cost.

Countries that have less data processing experience or which would prefer to process their census data with a minimum of time and experimentation would do well to select one of the integrated approaches to processing. At this time two programs, IMPS (Integrated Microcomputer Processing System) and ISSA (Integrated System for Survey Analysis) offer good functional coverage of the three major processing tasks (data entry, editing and tabulation) and are good candidates for evaluation and selection.

Integrated computer software packages offer some significant benefits. First, since the individual packages know about each others' characteristics, the seams between them are largely transparent. In IMPS and ISSA, this is implemented through the use of a common data dictionary. The data files to be created and processed need to be defined only once; each part of the system will use the common definition. Such an approach virtually guarantees an important dimension of consistency, since any change in the format and content of the data is introduced uniformly and consistently in all modules referencing the common dictionary.

The selection of an integrated package such as IMPS or ISSA to perform census microdata processing does not preclude the additional use of more special purpose programs after the microdata file has been created and the standard tabulations generated. Each package has its own set of relative advantages. For example, the microdata files produced by either of the above programs could be used by REDATAM to exploit its transposed data structures to provide relatively rapid generation of *ad hoc* tables after the regular publications program was completed. Such a strategy of exploiting the relative advantages of a collection of computer programs provides a rich functional computing environment, assuming that the technical capability exists to bridge the gap between the programs chosen.

The choice of a software strategy for population census data processing is not a simple one. Developing countries are strongly advised to regard this choice as a crucial one for the success of the data processing phase of the census and for the quality of the resulting data, and to exploit outside expertise if there are any doubts whatsoever about what an appropriate set of choices should be. External expertise can be used both to arrive at a set of good choices for data processing and to learn from the experience of other countries and other computing environments. The opportunity costs of not making good choices in this area are substantial, and with good external guidance, need not be borne by a developing country.

Microcomputer technology is new technology. Within the context of statistical and census data processing, on the other hand, the problems addressed by any level of data processing technology are generally familiar ones. Enumerations need to be taken for censuses and surveys regardless of the level of technology available. Statistical publications need to be produced to disseminate national, regional, and local information and to assist policy makers to make appropriate decisions. While the level of technology may improve staff productivity and make more activities feasible using the computer, the tasks to be performed are defined by national goals and activities.

On the other hand, each new set of technological improvements brings with it advantages that may improve the efficiency of statistical activities, lower their costs, or make entirely new options for dissemination and use available. In order to capture these advantages, there are likely to be costs in understanding, education of staff, new investment, and changes in organization of the activity. If the implications of using the technology are not well understood, the benefits from its use may not materialize; worse yet, uneducated use of the technology may lead to worse results than staying with a known and tested methodology for obtaining results.

The use of microcomputer technology requires changes in perception of how responsibilities are distributed among users, subject matter specialists, and computer experts. This has been discussed in a previous section. With special regard to statistical operations and population censuses, there are an additional set of considerations that are important to understand, generally centering upon the issue of data capital, its preservation, and its confidentiality. These are discussed below.

Issues of Data Control and Ownership

The data capital of a national statistical organization is one of its most important assets. The data are collected, stored, and used on behalf of and in support of national goals. Although the data are generally placed in the public domain to obtain full utilization from them, the statistical organization has the responsibility for their integrity and accuracy. Responsibility for the data collection is organizational, even though specific persons and groups in the office have the responsibility for collecting and processing them.

Centralized computing environments assist in meeting the above responsibilities. When the data are processed on one computer and stored there, they can be managed and kept secure with the assistance of the computer operations staff. Likewise, central processing procedures help to ensure both formally and informally that the data maintain coherence and that there are identifiable versions of data, including the version that is officially released. If central data base management systems are used, they further encourage consistency of the data capital through their data base management function and through system wide data dictionaries which are used by programs addressing the data base.

The content of organizational data bases within decentralized computing environments can be more difficult to control, especially when decentralization extends to collections of individual microcomputers. The essence of the typical microcomputer environment is that it is *not* shared, that it *is* individual, and that it supports primarily the individual user's productivity and only secondarily organizational objectives.[1] Within such an environment, it is easy for copies of data sets to proliferate; the copies may be of the same version or they may differ, with varying results for processing steps that involve them.[2]

The challenge of obtaining true organizational productivity from such an environment -- in addition to individual productivity -- is to reconcile the individual's personal interest in the data that he or she uses with a framework that supports the organization's interest in the integrity of its data. This is a complex issue, and it will not be solved easily. Recognition of the importance of the issue is the first step in addressing the problem, but further management controls are probably required. Balancing these interests for the overall benefit of the statistical service must be done individually in each country depending upon its circumstances and modalities of operation.

Three concepts relating to these issues are the concepts of data security, data integrity, and data confidentiality. These concepts and the implication of microcomputer technology upon them are discussed below.

Data security

Data are secure when they are retained with the confidence that they will not be changed or removed, either purposely or by accident. Data security differs from the notion of data confidentiality, which is concerned with

[1] This should not be taken to imply that the use of microcomputers does not support organizational objectives. Far from it; the underlying assumption is that by making individuals within an organization more productive, the effectiveness of the organization is enhanced. However, the organization is faced with the problems arising from how to combine the work effort of these individuals to realize the higher level of effectiveness on an organizational basis.

[2] As an example, consider what could happen in a decentralized environment in taking a survey of industrial production. Suppose the master register of producers were updated in one place, and copied several times to support separate groups drawing the sample, processing the results and following up non-responses. If the separate registers were not kept in synchronization, then bias might be introduced into the sample, errors might be introduced in the processing stage, and non-respondents for follow up might either not have been selected or they might no longer exist. The result could be either undetected errors, or administrative confusion, or both.

preventing the unwanted disclosure of data base content to an individual not authorized to have access to it.

Violations of data security can be either planned or inadvertent. Planned violations include unauthorized duplication and destruction of machine readable files. Inadvertent violations include loss of a data diskette, accidental deletion of a computer based file containing data, and incorrectly recording the storage location of a diskette, thus reducing the likelihood that it can be retrieved when it is needed.

Maintaining data security depends in part upon control of the physical and electronic forms of the information entrusted to the statistical office. Physical control implies not only ensuring that all transfers of material into and out of the office are authorized and proper, it also implies having a filing and indexing system for materials contained within the operation that is correct, secure, and up to date. Use of an enumeration area register and a tabulation register, described below, can be an important part in implementing such a system.

Maintaining data security also depends upon having adequate operational procedures to prevent inadvertent security violations or minimize the loss that occurs when they occur. Examples include:

- explicitly write protecting each diskette used, except for operations which require adding to or modifying the data.
- instituting frequent, mandatory, and supervised backup procedures for data on fixed disks, network file servers, and any data sets resident on larger computers that are part of the current work in process.
- training staff members to understand the importance of data capital, and therefore the importance of data security and related procedures such as file backup.
- regularly storing copies of important file backups in a secure place away from the main processing location so that the data will be secure from physical catastrophes.
- establishing specific controls to ensure that batches of data can be easily and quickly accounted for, whether on the premises or off.

When the final computer based census results are available, archival files should be made available and stored at a number of different locations in the country to ensure that the valuable end product is secure. If the data are to be retained for a long period of time, each set should be copied periodically to ensure that any long term degradation of the original magnetic recording does not jeopardize the country's ability to use the data in the long run.[1] In the future, other machine readable media such as optical recording may become more popular for archival purposes because of the considerably longer known retention rate of data recorded on such media.

Data integrity

Data integrity includes the notions of correctness, coherence, and synchronization. A set of data has integrity if it is internally consistent. Examples of lack of data integrity include mixed responses from two different time periods, responses missing from an enumeration area, duplicate responses, and responses coded according to the different sets of coding rules.

Achieving integrity of the data in the final census microdata tape depends upon having adequate control of operational procedures at every step in the creation of the tape, regardless of the technology utilized. Achieving integrity in the final tabulations depends upon the correctness of the tabulation program, the order in which microdata input are supplied to it, the preparation of table layouts, especially row and column descriptors and their placement with respect to data areas, and the method of aggregation of the lower level tables.

While ensuring data integrity is a problem regardless of data processing technology, the greater potential distribution of semi-independent processing sites poses the danger of additional loss of control over operations. Data entry has always been a distributed operation, although it is often closely logically controlled by an on-line computer[2] or integrated into a local area network. With microcomputers, data entry is performed on independent or linked microcomputers, and the results of the operation are generally recorded on magnetic diskettes. After verification, in a large mainframe environment these data would be batched together on magnetic tape and would go through the remaining processing steps and cycles of editing and tabulation together. In a microcomputer environment, depending upon the geographic distribution of processing and the underlying processing strategy, these data might undergo some or all of these steps individually. In this case, the potential for introducing inconsistencies through

[1] It is difficult to obtain authoritative knowledge about the shelf life of recorded magnetic media that are not used. Early magnetic recording techniques were subject to a hazard known as "print through," which resulted when magnetized areas of magnetic tape wound in a reel affected the magnetization of areas of tape in the adjacent inner and outer wind of the tape. Beyond that, there exist opinions that the magnetic strength of the magnetically recorded spots on a tape or disk, especially at high densities, decays over time. Some commercial organizations have policies for periodically rewinding and re-recording information on valuable tapes to minimize the probability of loss of data. Authoritative information about this subject would be helpful to both national statistical offices and many other organizations having long term data security concerns.

[2] Examples are on-line entry under control of a general purpose multi-user interactive system, or use of a special purpose environment such as a key-to-disk system.

inadequate operational control of the batches of data is higher.

A useful way to think about the problem is as a production process in which a large number of objects – batches of data, each originating from the data collected in one enumeration area – are subjected to a series of steps. Some steps such as verification and editing are cyclic; data not meeting quality control standards are reprocessed until they do. Others, such as data entry, are linear, and need to be performed only once.

A key organizing device for such a production process is the register of enumeration areas described in a previous section. The register, which contains a unique entry for each enumeration area, is also constructed to record the current location of the data set for each area or work unit comprising a number of areas, a history of the steps that have been applied to that data set, including dates, times, and operators, and an indication of what step is to be performed next. Every time an operation is started or finished on one of these data sets, the enumeration area register must be updated promptly. There must be only one such register associated with a specific location, and it must be accessible only serially, so that two accesses to update the information in it cannot inadvertently create inconsistent or incorrect entries. The register is essentially an information model of the production process, and can be used to control the process and maintain the maximum amount of integrity in the microdata file being created.

If tabulation is also to be performed on a decentralized basis, then the register operation requires additional capabilities. First, it needs to be able to create entries in a computer based tabulation register that indicates which tabulations have been created on an enumeration area or work unit basis, when they were created, and where they reside. The time element is important, since it may become necessary to perform further editing on enumeration area microdata *after* a tabulation has been produced with it. In such a case, it is necessary to recreate all tables that were previously created, using the new data. The problem becomes larger if the old table at the enumeration area level has also been used by an aggregation program to create further tables for larger geographic areas; in such a case, the aggregate tables must also be recreated. Associating a definite date and time with the completion of every processing step applied to both enumeration area data and tabulation results allows a computer program associated with the two registers to determine exactly any pattern of dependencies that require some steps to be repeated.

Enumeration area and tabulation registers need to be computer based and unique. Their security is important, since in a sense their contents are used to drive the data processing operations. They should therefore not only be backed up frequently, but the transactions against them between backup times must be logged securely so that the registers can be recreated to within a second or so of the time when the failure occurred. Use of a local area network, with some spare computer capacity for register management, may be useful. However, the process of controlling the data processing operation does not require that the microcomputers are linked through local area networks.

Many of the problems of achieving integrity at the tabulation stage can be minimized by using existing software that has been tested and used extensively. Software that provides automatic labelling of table contents from a data dictionary that has guided earlier processing steps is perhaps the best way to maximize integrity at this stage. Such tabulation programs may be better at providing internal consistency than staff who must specify each table laboriously with no automatic consistency check provided by machine readable documentation, or metadata.

When census data processing is complete, then it is important to use the various registers to collect a complete copy of all of the census data in machine readable form, both for archival purposes and also for fashioning additional products, both printed and machine readable, for exploitation of the census data base.

Data confidentiality

The notion of data confidentiality is well known to statisticians. Most of the data collected by statistical offices is collected on a confidential basis, generally under the protection of law. Such protection is required in order to obtain data that otherwise would not be disclosed voluntarily. Statistical offices have an obligation to keep these data confidential. This obligation is generally interpreted to imply that not only should the original data remain confidential, or undisclosed outside of the statistical office, but that users of generally available statistical publications should not be able to infer or deduce original values reported by any one respondent from the aggregated data that appears in the publications.

The latter problem has been studied extensively by advanced statistical offices, and most offices have explicit tabulation and dissemination strategies that meet this requirement. Inadvertent disclosure of individual information from aggregate results is not related to the technology used to prepare the aggregate data.

The former problem is more closely related to changes in technology. Before microcomputers were available, data processing of statistical information was almost always performed centrally due to economies of scale of equipment as well as the importance of a critical mass of trained people under direct central supervision. Large countries may have decentralized their processing down to the state or provincial level, but only when there were economies of scale of both data processing equipment and trained staff. In such an environment, it was both desirable and possible to control the flow of both questionnaires and data in machine readable form closely. Physical containment of the data processing operation was, therefore, an important means to control access to data. In addition, there was insufficient other computing capability in the country to be able to read the data in machine readable form even if it could be obtained.

Microcomputers allow processing to be geographically decentralized, both within a central set of offices and to remote city and district offices. Local area networks, linking microcomputers together, can serve specialized functions remote from the central data processing organization. Each such operation must be as concerned with data confidentiality as the one central operation was in the past. This may mean replicating secure storage areas for original documents and establishing parallel control procedures for machine readable data, including standards for destroying media if it is no longer needed by the process. Further, transfer of documents or machine readable media between remote and central offices must take place in such a way that the working material is secure in transit. As long as the physical material on which the data are recorded is guaranteed to be secure, then data confidentiality is protected. If the machine readable material is not secure, then it may be useful to consider encryption of the data. Encryption programs exist for both microcomputer and large computer environments, and are used routinely to secure sensitive data in developed countries. In developing countries generally, maintaining control over access to the physical media on which the data are stored will be a sufficient measure to ensure confidentiality.

In developed countries, there is a *potential* threat posed to data confidentiality by having statistical data processing computer systems attached to data networks. Some types of data networks are more vulnerable to being tapped than others, and wide area networks operating between remote locations are also capable of being tapped. Such considerations are generally not important now for statistical data processing in developing countries, but use of microcomputer and other distributed technology requires an awareness of the nature of the problem. With respect to local area networks, which are generally the most effective way of linking microcomputers together, control of the physical path of the network connections[1] is sufficient to protect against penetration of the network. Although local data networks can be penetrated remotely, the cost is very high and is very unlikely to occur.

[1] The most common media for connecting microcomputers in a local area network are twisted pair telephone wire, thin wire cable capable of carrying Ethernet traffic, and RS-232 cables for point-to-point serial transmission.

Long distance transmission of confidential data can be intercepted, although at very significant risk and cost. In addition to the cost of interception, the interceptor would have to decode a complex transmission signal in order to obtain what might be useful data. Given the importance of the content of typical microdata, such penetration is highly unlikely to occur. In the unlikely event that important data are to be safeguarded during transmission, well-known encryption techniques can be used to increase the confidentiality of the transmission. This issue should not be a problem for most developing countries in the near future.

In general, added awareness of physical control of data processing materials over geographically dispersed locations is sufficient to ensure a level of data confidentiality consistent with that of the past. However, it is important to be aware of other concerns, including the availability of an increasing number of computers in many countries that are capable of reading and using microdata that could be obtained if a breach of security were to occur.

Revised Selection Criteria for Computing Environments

The degree to which microcomputer technology is applicable to national statistical office operations will change over time. In general, the technologies underlying computing hardware and the expanding market for mass distributed software will change faster than a statistical office's data collection, processing and dissemination plans. It is to be expected therefore that a greater fraction of a statistical office's processing will be tractable within a microcomputing environment as time progresses.

Advanced statistical offices as well as offices in large developing countries will still require mainframe or minicomputer resources for some time to come. Microcomputers still lack the very large file storage and sequential file processing capability that larger computer systems support well. In addition, there may be financial economies of scale in buying a specific software product for one large computer, shared by many users, rather than buying many copies of the same or a more limited software package to run on many microcomputers.

The *relative advantage* of larger computers will diminish over time for applications in statistical data processing.[2]

[2] One can already observe in developed countries the encroachment of microcomputer oriented large data storage devices into what has been mainframe territory for years. In particular, optical storage devices are now being used to store and process files of the order of 100 to 500 MB, a size that would have taken 3 to 12 reels of magnetic tape several years ago. The newest competition is between CD-ROM (Read Only Memory) disks, WORM (Write Once, Read Many times) disks, erasable optical disks, and digital audio tape media. One or more of these media will develop to provide effective large file storage for microcomputers, thus eroding another computer capability that was previously only associated with mainframes. The challenge for developing countries is to be able to understand when such technologies become sufficiently mature to

SPECIFIC ISSUES RAISED BY USE OF MICROCOMPUTER TECHNOLOGY

The difficult job of the statistical office is to match activities involving statistical data processing with the appropriate computing environment to perform the activity efficiently and effectively. Such decisions over time are not trivial ones; furthermore, they are interlinked with the office's investment strategy in computing equipment which is quite important given the limited funds most developing countries have to invest in such facilities.

Micro-Mainframe Connectivity and Co-operative Processing

Much of this document has treated microcomputer technology and mainframe computer technology as opposites. Such a comparison is useful because it provides a contrast between the various relative advantages and disadvantages of the technologies, as well as the distribution of responsibilities in exploiting them.

Having made these comparisons, it should be noted that the two technologies are more appropriately viewed as complementing and supplementing each other rather than being in competition for all aspects of any activity as complex as census data processing. Each technology has relative advantages and should be used to exploit those advantages whenever it is efficient to do so. Microcomputer technology is divisible, enjoys a very large low cost associated software market[1], and is generally easy to use. Mainframe technology offers large capacity and varied storage media, extensive processing power, and generally a broad range of support services provided by the affiliated central computing organization.

Microcomputers and mainframe computers can be used co-operatively in a variety of ways. First, it should be noted that there are generally moderately effective ways to transfer character data between mainframes and microcomputer systems. Files of modest size, such as all census data comprising an enumeration area, are often easy to transfer back and forth.[2] It is possible to consider moving data around a combined computer system to take advantages of these relative advantages if the benefit were worth the cost of transporting them back and forth.

A more common form of cooperation would be to perform data entry and editing on microcomputers and then decide whether or not to move the clean microdata file to the mainframe. Assuming that a mainframe already exists in the country and is available for census processing, larger countries may still find it to their advantage to move the microdata to the mainframe for final tabulation even though the entire work could be performed with microcomputers. This is a dramatic change from the last round of census processing, when a mainframe or powerful minicomputer was *necessary* for processing population census data.

There are other, more tightly coupled, co-operative uses of mainframes and microcomputers now beginning to be used in developed countries. Such uses rely upon more sophisticated programs which "know" where to obtain the resources required to perform a request which has been made in a system independent manner.[3] Such models will someday migrate to developing countries and emerge in powerful multi-microcomputer systems.

Software as Intellectual Property

Within the last several years, microcomputer systems have become important to many branches of government, both in developed and developing countries, as a source of efficient and productive computation in many areas. Rapid technical progress in the underlying technologies have made such systems affordable not only to government agencies and commercial firms, but also to individuals.

The rapid expansion of the market for microcomputers has made possible a mass market for high quality software which runs on those systems, and software producers have responded by offering a wide variety of products at very low prices relative to the value they provide their users. Without such software, microcomputers would be useless. The cost of this software is a legitimate and essential cost that must be borne in order to provide a productive environment for the use of microcomputer systems.

While software prices may be regarded as reasonable within the context of a developed economy, developing countries and/or international aid agencies may regard the same prices as exorbitant within the context of limited development aid or operating budgets. This conflict between wanting to use software and not being able to afford it may lead to the temptation to make copies of software already purchased rather than purchasing new copies.

incorporate them into their computing environment, and then what specific activities and problems to apply them to.

[1] The cost of software may be low, and at the same time amount to an appreciable fraction of the cost of an entire microcomputer system for performing a specific set of functions. In addition, while the *per copy* cost of microcomputer software is often low, supplying one copy for each of many machines (in the absence of a site license or volume purchase agreement) can lead to considerable financial expense.

[2] Files transferred as character data, usually in ASCII format, can often be transferred by connecting the computers using a serial (RS-232) connection and then using either a file transfer or terminal emulation program to transmit the data. Sometimes conversion between two competing standards for representing character data, ASCII and EBCDIC, may be required; this can often be done either on the microcomputer or the mainframe end of the connection. If the mainframe is an IBM mainframe without an ASCII 7171 or similar conversion box, an IRMA card can be installed in an IBM PC or IBM PC compatible system to provide SNA communication between the systems. Finally, for large data transfers it is possible to put an industry standard 9-track 1/2" compatible tape drive on a microcomputer and carry data on tape between the two systems. There are additional communications options, and more will emerge over time.

[3] An example is the implementation of distributed data bases and distributed data base processing over a network of different machines; the DB/2, Ingres, and Oracle data base systems all support some form of this model of storage and processing.

SPECIFIC ISSUES RAISED BY USE OF MICROCOMPUTER TECHNOLOGY

Computer software is an important form of intellectual property. It provides users considerable value at relatively low cost. Unfortunately, like a book, it can be copied quickly and easily. Doing so with the intent to use the duplicate copy on yet another computer at the same time the original copy is used will in almost all cases violate the license agreement entered into by the original user and will almost certainly be in violation of copyright laws.[1]

Through their activities, statistical offices produce significant intellectual property. The information summarized and prepared, often at substantial cost and with substantial effort, is of great use in preparing development plans, formulating policies, and in assessing progress toward meeting national goals. As creators of information, statisticians are acutely aware of the cost of producing information and of the value that it provides.

Statistical information produced by national statistical offices is generally considered a public good; its production and distribution are supported by the government, and it is generally placed in the public domain so that full and free use can be made of it. Private firms developing software have no such subsidy, and depend upon license fees to maintain their development and support activities. Developing countries depend heavily upon microcomputer software products to improve the productivity of, *inter alia*, their national statistical offices. In this regard, it will be very important for software producers to perceive that public institutions in developing countries respect their rights to intellectual property. Illegal copying of software is a significant problem in developed countries, and the consequences of users in developing countries being perceived as a part of this problem could restrict the supply of such products.

Software producers are aware of the difficulty that some customers have in paying for software, and have often made price concessions for either volume procurement or for procurement by not-for-profit organizations. Software producers are generally not unsympathetic to the problems faced by developing countries. When access to and use of multiple copies of software is important for developing countries, the companies can and should be approached either directly or through appropriate bilateral or multilateral aid channels to negotiate more favorable and affordable prices.

[1] Software products can sometimes be obtained under a site license or a right to copy license. Such licenses generally provide one copy of the software, but give the licensee the right to make as many copies of the software as desired within certain specific bounds, such as a university or a government agency. Such licenses are more expensive to purchase than single copy licenses, but are generally considerably cheaper for an organization or work group when there are a substantial number of copies being used concurrently.

Advances in technology help organizations to perform current activities better, where better performance may mean faster, or more inexpensively, or more efficiently, or some combination of these methods of extracting savings. In the long run, the more important benefits of technical progress are derived from new activities that such progress makes feasible; these are activities that either could not possibly have been performed before the breakthrough or could have been performed either at such exorbitant cost or under such unacceptable conditions that no attempt would have been made to do so. Inexpensive computing in the form of microcomputer technology allows additional activities to be performed that were not feasible in the past.

New Methods of Data Dissemination

The traditional form of dissemination of population census data has been to print volumes of tabulations. Countries preparing a large number of tables have often resorted to microform (microfilm or microfiche) to reduce production and distribution costs. Some more advanced countries have distributed both samples of microdata and a larger set of tabulations on magnetic tape; this has been effective in countries having a large number of computers capable of using the information in this form.

Countries using microcomputers to process their census data will soon be in a good position to distribute a number of census byproducts on diskette. Two formats suggest themselves, data base format and spreadsheet format. Both allow disaggregated data to be distributed. Using microcomputer based data base systems, information about individual enumeration areas and intermediate geographic aggregates can be distributed in a form suitable for further analysis or publication; the spreadsheet format allows final tables to be distributed in machine readable form for further manipulation. Both forms allow analytical manipulation, transformation, and analysis of the data rather than merely viewing it in static publication form.

Finally, national statistical services should begin now to consider how they structure themselves to deliver services using an internal data base and a set of appropriate computer based tools. Such a service may be able to produce responses to *ad hoc* queries rapidly and inexpensively. Depending upon the efficiency of such an operation, it might be possible to scale back the extent of the traditional publications program, preferring instead to rely upon recognizing explicitly expressed demands rather than an *a priori* notion of what tables and other outputs might be needed. Computer based tools such as REDATAM and the thematic mapping systems discussed below form an initial basis for the establishment of such a service bureau approach to meeting customer requirements.

Portable Data Entry in the Field

The most important objective of the data collection and entry process is to capture the data supplied by the respondent as fully and as accurately as possible. Errors of accuracy can intrude in the data collection process at any stage of the processing. Data collection and entry strategies should minimize the number of separate processes through which the data are put to minimize the introduction of errors into the data.

One method of minimizing the number of steps involved in data collection is to collect and record the data at the same time the interview occurs. In theory, such a procedure could be constructed to collect the data in machine readable form, provide interview *flow control*, i.e. allow the flow of questions put to the respondent to reflect his or her specific circumstances, such as only asking women above a specific age about their childbearing history.

The strategy of recording information in machine readable form at interview time has been implemented in developed countries for at least 10 years. In its most common form it is referred to as computer-assisted telephone interviewing or CATI. Because earlier computers have been strictly non-portable, the method implemented links the interviewer, who is stationed at a computer terminal, with the respondent by telephone. An interviewing program guides the interviewer through the series of questions to be directed to the respondent and records all answers directly in a computer based file. Consistency checks are generally a part of such systems, and can be used to correct either an inaccurate answer or an inaccurate path on the questionnaire. Such corrections can be made *during* the interview when the respondent can be asked to resolve any inconsistencies that arise. Such immediate feedback is believed to improve the quality of the data significantly and to place less of a burden on the later use of editing and imputation procedures.

During the last several years, advances in microcomputer technology have led to the manufacture of relatively powerful, battery operated portable microcomputers which are capable of supporting computer-assisted personal interviewing (CAPI) at the respondent's location.[1] Interviews can take place under control of an interviewing program stored in the portable computer; the responses entered are recorded on a diskette. At the end of the day, the data are either sent physically on diskette to a central collection point or they are transmitted there using telephone lines via modem. The batteries of the portable computer are then charged overnight, and the system is ready for another round of interviews.

[1] In early 1989, one of the most attractive of such computers was the Toshiba Model 1000, a 512KB one diskette 8088-based MS-DOS system that weighs less than 3 kg. and costs about US$750 in the United States.

FUTURE OPPORTUNITIES FOR THE USE OF MICROCOMPUTER TECHNOLOGY

CAPI is now in use in, *inter alia*, France and the Netherlands,[1] for a variety of survey data collection procedures. Although it was initially hypothesized that introducing a computer and a computerized technique into the interview situation would confuse or otherwise negatively affect the respondent, initial results appear to show that there are no significant adverse effects. The technique appears therefore to be a promising one that can only become less expensive and more effective based upon future technological progress.

It is very unlikely that CAPI techniques based upon portable microcomputers will be used for a complete census enumeration in the near future. The initial costs are quite high; the number of units required would be very substantial. Nevertheless, the techniques may have some applicability in some areas. For example, CAPI could be an effective technique to be used for the field work for the post-enumeration survey (PES), in which a limited number of households are revisited. To the extent that the PES sample is clustered geographically, then an investment in such machines could result in a more rapid and accurate survey.

At the moment, CAPI techniques based upon portable microcomputers appear to be suited for survey operations having a more limited number of respondents, where the interviewing process is complex, and in which there are a limited number of enumerators who receive thorough training[2]. While this technique is being used for some survey work in developed countries now, it will probably have an increasingly justifiable place in the statistical data processing toolkits of a significant number of developing countries within the next 5 to 10 years.

Geographic Information Systems

The potential for graphical representation of data has always existed, but the sets of tools that have historically been used to process statistical data have made it relatively difficult to obtain useful graphical output easily.[3] Some microcomputer based software products can now help to reverse that trend. Existing productivity tools such as Lotus 1-2-3 and Microsoft Excel have graphics capabilities that can be used to advantage in displaying statistical data. ASCII files can be used to transfer data between data base storage programs and a variety of standalone graphics display programs. In addition, programs such as Data Desk II are emerging which provide the capability to view and manipulate data powerfully in two, three, and four dimensions visually.[4] The use of statistical data capital as part of a geographic information system must not be ignored, and relates to the eventual use of the statistical data produced. While the classic form of statistical output consists of publications, tables are poor ways to present data which has geographic attributes and which can be visualized as data defined over regions and areas in the country.

Geographic information systems refer to computer based systems that bring together information with a spatial component and allow the information to be processed and displayed in map format. They allow information of different sorts to be merged, new attributes computed, and maps made on the basis of values of one or more spatial attributes.[5]

Within the context of a national statistical agency's activities, two types of mapping dominate: (1) *thematic mapping* for the display of geographically distributed data; and (2) *collection mapping* to support the data collection process. They are discussed below.

Thematic mapping. The emphasis upon explicit graphical display of statistical data in a map format makes possible a new and powerful dimension of information representation that can be considerably more productive for many users of

[1] See, for example, Denteneer, D., Bethlehem, J.G., Hundepool, A.J., and Schuerhoff, *Blaise, A New Approach to Computer Assisted Survey Processing*, Department for Automation, Netherlands Central Bureau of Statistics, Voorburg, BPA No. 13436-87-M3, November 6, 1987, 38 pp. The paper describes BLAISE, a microcomputer based system for survey data collection and processing, using a Pascal-like approach to codebook definition and use, which is now used for a significant volume of survey work in the Netherlands.

[2] The success of CAPI in the field depends upon having a corps of interviewers who are comfortable with computer-assisted interviewing, and who can be trained sufficiently in computer technology to feel confident using a computer without any support staff present. Well designed programs that do not fail will be necessary to support such interviewers. Countries should not attempt to use CAPI unless they can assure themselves of having some combination of a well trained interviewing staff and expert support that will ensure that the computer will not detract from the effectiveness of the interview process.

[3] Early mainframe technology was oriented strongly to the processing of numeric data, with concern for processing of text growing only relatively recently. Both forms of processing were essentially character-oriented, and both the hardware developments and the software tools associated with the technology reflect these concerns. The MS/DOS based microcomputer family essentially continues this tradition, with screens oriented to displays of characters and with the operating system user interface built upon complex character strings. Computing paradigms oriented around processing and display of visual images have existed for some time, but have in the past been quite expensive on a relative basis, and prices for corresponding software have been high because of greater complexity of production and more limited markets. This situation has changed radically in the last several years with the strong emergence of other microcomputer architecture paradigms that make extensive use of graphics concepts (examples include Apple, Atari, Commodore, and X-Windows under Unix). At present, there is little evidence that this alternative paradigm is either known to or understood by many statistical offices, to their disadvantage.

[4] Data Desk II is one of a class of programs implemented on the Macintosh rather than MS-DOS machines because of the more powerful user interface and graphics capabilities of the Macintosh. Just as the MS-DOS environment will probably become an accepted standard for the more classical forms of statistical data processing on microcomputers, the Macintosh is likely to and deserves to serve a very powerful adjunct presence in statistical offices to perform processing that is at present somewhat non-standard, but is so powerful and important that it should not be neglected.

[5] Such maps may be viewed on a computer screen, plotted, or printed using a laser printer. With some care, screen images may be photographed to make slide transparencies for presentations. If useful, the film processing operation can be bypassed by using a self-developing transparency film such as the slide film recently introduced by Polaroid.

the data.[1] Maps of this type are generally called thematic maps; they overlay a substantive theme or themes upon a map presentation format.

An example may be helpful. Consider a map of a country having the boundaries of all administrative subdivisions present down to the county level. Such a map might contain between 500 and 2,000 subdivisions.[2] From the population census data base, one can extract population summary tables for each of these areas. From other data bases developed by the country, one could determine the number and capacity of schools at various educational levels, by subdivision location. By making some assumptions about the demand for education if the supply were there, one could compute local demand functions for continuing education, and display them in map format as a function of their deviation from the norm. The result is a map whose thematic information is derived from separate statistical data collections, brought together and displayed *de novo* for a special analysis.

The example can be extended further. Fields of statistics dealing with the production and distribution of goods might benefit from an understanding of how the transportation facilities of a country are positioned to assist in distribution. Maps involving distribution could easily be overlaid with physical infrastructure data describing transport networks: highway and roads, railroad lines, canals and waterways, airports, and other means of transport. Fields of statistics dealing with balances, such as food and energy, can use maps to advantage by displaying thematically the geographic distribution of production and consumption. Adding the paths and capacities of distribution mechanisms graphically displays internal economic relationships of importance to the country.[3]

The generation of even relatively complex thematic maps using large configurations of existing microcomputers is now possible and becoming easier and less costly as software tools increase and as the underlying hardware costs decrease. Powerful programs for supporting aspects of geographic information systems are now available on large MS-DOS environments. The thematic map is for many purposes a superior data integration and display tool for users of statistics than are standard sectoral tabulations, and such maps can increasingly be generated relatively easily with existing software bases.

Geographic information systems can be used to prepare material such as thematic maps, graphs and charts for official publication. Such images, when designed appropriately, can enhance both the readability of a publication and assist the reader in understanding the statistical relationships presented. Many official government publications already use maps to advantage now.

Computer based geographic information systems can also be used to prepare *ad hoc* thematic maps to assist planners and policy makers. This capability may be more important in that, in conjunction with a statistical data base, it can be used when and where needed to study problems of specific interest at the time. It places the statistical service in a proactive rather than a static role, as a partner with other governmental agencies in satisfying the need for statistics in easily comprehensible form and in rapid response to current concerns.[4]

While thematic mapping capabilities are becoming easier to obtain using microcomputer based technology, their use is not yet as straightforward or as well understood as is the use of traditional census tabulation programs. Some developing countries may wish to delay introducing such techniques until their statistical data processing capabilities are more robust. Other countries may be interested in experimenting with the potential of this newer method of displaying census information, and are encouraged to do so. Such experimentation can occur with only modest investment of capital and labor, and can be extended incrementally based upon the ability of the statistical agency to absorb and use effectively the new hardware and software technology involved.

Computer-assisted Cartography

Collection mapping. Defining areas for data collection, or enumeration areas, is a major activity in preparing for a population census. During the period of enumeration, each

[1] Maps showing boundaries of major civil subdivisions of a country are now relatively easily prepared using inexpensive digitizing equipment attached to a microcomputer. Microcomputer based software is available to construct and display the digitized map base. Defining a map with several thousand geographic subdivisions for a computer by digitizing it should take a period of several months of one person's time.

[2] With appropriate hardware and software which is not too costly, the process of digitizing the borders of the principal civil subdivisions of such a country might take 2 to 4 months. The result would be a computer readable map data base consisting of the outlines of each civil subdivision. Maps of civil subdivisions for parts of the country or for the entire country could be retrieved, scaled, overlaid with thematic data, and displayed in picture or map format on a computer screen or on paper.

[3] These examples describe maps that combine statistical data and administrative records. Other sources of data can also be integrated into the map model: remotely sensed data collected by satellites can be used to determine broad land use patterns, for example. Specific population subgroups could be used as a base population for generating thematic maps in order to study planning and policy issues relating to that subgroup. Examples are the provision of health facilities and services compared to incidence of diseases, and the geographic distribution of individuals in various occupations, such as teachers, doctors, and technical specialists.

[4] The ability of the statistical service to assume a proactive role by responding rapidly to requests is independent of the existence of an information system specifically incorporating a geographic component. However, many national concerns have a geographic component, and inattention to geography lessen the effectiveness of the statistical service to respond. Furthermore, the map framework is sufficiently general to be able to integrate information from other sources as well as provide pictures for assimilation instead of tables. If the pictures can be produced rapidly at low cost, then interactive dialogues can take place between planners and the information system, allowing a powerful mode of investigation and exploitation of the statistical data base.

enumerator must have a precise definition of the area which he or she is to enumerate. Enumeration is a labor intensive activity, and it is not unusual for a country to employ 0.1% of its population or more for a house-to-house enumeration process.

Maps are by far the preferred way of defining enumeration areas. As a result, each population census depends either upon the preparation of a cartographic data base for the country or upon an update of a previous set of maps, perhaps from a previous census or other statistical operation. Maps used for this purpose are generally referred to as collection maps, based upon their primary function of supporting the collection of statistical data. Such maps are also useful for sample frame definition and sample selection for surveys.

In some countries, a central cartographic agency has primary responsibility for mapping and making special purpose derivative maps available to the statistical authorities. However, in other countries, there may be no such authority and the statistical agency may have to accept the responsibility of manufacturing its own maps. If a set of up to date maps already exists, then the statistical authority's responsibility may be limited to defining the collection areas using those maps.

Some developed countries have established computer-based sets of maps that can be used for multiple purposes, including defining enumeration areas. Such a map is kept as a series of geographic co-ordinates, defining streets, street names, city or village blocks, address ranges within block faces, and natural boundaries, and may have additional information such as block face identifiers and building counts for each block face. These map data can be retrieved and used to draw an up-to-date map either on a computer screen or on printer or plotter paper.

Since the time required to digitize a map for a geographic information system may be less than the time needed to draw the map carefully, it is natural to consider using computers to create and store a base of maps which can be periodically updated for statistical activities. Some developing countries may wish to investigate this possibility further. Even on microcomputers, software exists that can produce street maps that would be useful for defining enumeration areas and guiding enumerators. However, the number of such maps for a country can be quite large, and the census budget may not allow for both drawing or updating maps manually and then digitizing the information collected. Once the map base has been digitized and moved onto the computer, then periodic updating becomes much easier, but the initial investment in preparing the map base is still labor intensive and time consuming. If attempted, the initial digitization of a map base should be given a sufficient period of time to be successful. Alternatively if resources permit, digitization could occur after the completion of either an enumeration or a complete population census cycle, at a time when there are slack resources available and no possibility of jeopardizing the timeliness of producing results from a major statistical field operation.

Statistical offices should be aware of the possibilities offered by computer based cartographic systems and plan to introduce them at some point in time of their development. Because of the substantial initial cost of entry into this activity, statistical offices should take care not to begin without a very good understanding of the costs and benefits, as well as expert guidance. Microcomputer technology provides a relatively inexpensive way to begin exploring computer mapping on a pilot basis, perhaps for one urban area, and to assess the costs and benefits on on the basis of a limited investment.[1]

Socioeconomic Microsimulation

The creation of census and survey tables provides population cross-section data which can be quite valuable in supporting public administration and planning, as well as providing a better knowledge of the characteristics of the country's population.

Some planning and research issues require access to the original census microdata, either for *ad hoc* tabulations not originally produced or for testing specific assumptions about behavior that require calculations to be made at the individual and family level and then aggregated for tabulation or other method of analysis.[2] The latter type of calculation is often called a static simulation, since it starts with a fixed population or sample and makes computations based upon how population units might react to a change in external circumstances which is defined as a specific set of calculations, such as the calculation of the effects of a change in a tax or transfer system.

Such models can be made dynamic by a process of starting with a given population sample, applying a synthetic stochastic[3] model of certain characteristics such as aging and death, birth, marriage and divorce, educational achievement, labor force participation, and income generation and asset formation. Such a model would be applied to an initial population. The result of the application would be what the model predicts the population would look like one time period, say one year, in the future. Assuming that the model

[1] The emphasis upon MS-DOS as the dominant microcomputer operating system has slowed a good understanding and appreciation of these possibilities. Computers having flexible, standard and adequate graphics capabilities, such as the Macintosh, have a decided advantage for this type of work.

[2] For example, one question might be to measure the financial effect of a tax or transfer system based upon assumptions of differential income changes across occupations or industries.

[3] A stochastic model is one that uses random draws from known probability distributions to determine whether to cause a given event to be simulated or not for each individual instance in which a decision must be made.

is valid from a subject matter point of view, repeated applications of it would age the population as far into the future as is required to answer questions of interest.

Dynamic microsimulation models of persons and families have been built in developed countries, but have in the past consumed large amounts of large computer resources.[1] The emergence of large and powerful microcomputers, combined with advances in programming knowledge, now allow dynamic microsimulation modelling to be performed using large microcomputers.[2] Such models are of increasing interest to statisticians, economists, and policy makers since they provide a sufficiently general and long term framework to be able to investigate many questions not easily answered in other frameworks. The possibility of implementing them in microcomputer environments brings this modelling technology considerably closer to those developing countries which have the interest and analytical and modelling abilities to exploit it.

Developing countries should have some awareness of the characteristics of this approach, and should understand the importance of the contribution of microdata files resulting from population census data processing to the original microsimulation data base. While such an approach to policy exploration may not now be feasible for many developing countries, for others it may become an increasingly useful analytic and modelling tool.

---

[1] See, for example, Sadowsky, George, *MASH: A Computer System for Microanalytic Simulation for Policy Exploration,* The Urban Institute, 1977 for a sense of the scale of such modelling activities in the past.

[2] See, for example, Statistics Canada, SPSD/M Product Overview. Ottawa, Canada, 1989. The SPSD/M model can be executed effectively on a population of 180,000 individuals using a medium speed microcomputer with Intel 80386 and 80387 processor chip set. Each iteration of a fairly complex tax and transfer model takes about 5 to 15 minutes of elapsed time.

7. CONCLUSION

National censuses of population and housing provide data which is a very important and essential part of the statistical infrastructure of a country. The planning and execution of a national census of population and housing is a massive undertaking, requiring both substantial resources and considerable organization and management. The size of the data processing portion of this task was historically sufficient to spur the creation and development of automatic data processing machinery. The use of both electromechanical and electronic computing technology has been essential to the processing of census data since their inception.

During the last decade a new type of computing environment, referred to as microcomputing and sometimes as personal computing, has evolved. The emergence of microcomputer technology is the result of technical progress in the computing and digital electronics industries, and this progress is expected to continue into the future. Microcomputing technology is in many ways appropriate technology for developing countries; it is small, inexpensive, easily maintainable, divisible, and the software and user interface make it generally much more attractive for persons who are not computer specialists to use.

Microcomputer based information technology now offers substantial hardware and software processing capabilities which are directly applicable to tasks in statistical and census data processing. It is the most important emerging computing technology today for intensive exploitation by statistical services in both developing and developed countries.

Developing countries are already exploiting microcomputer technology to process population censuses. Starting with the Cape Verde Islands and the Cook Islands in the early 1980's, microcomputer technology has increasingly been evaluated and used for processing census data. Based upon an assessment of the current state of this technology, it should now be considered for adoption by countries having populations of about 25,000,000 or less. Further technical progress is likely to raise this figure in the future.

Progress in statistical data processing is an important ingredient for effective planning and development in developing countries. It depends upon appropriate and effective exploitation of new computing technology such as microcomputers from which everyone benefits on a continuing basis. Such progress occurs and is modulated by a variety of factors specific to individual countries, including infrastructure considerations, incentives for learning and working, and the nature and extent of relationships among organizations having a stake in information technology in different sectors of the country.

While many developing countries are aware of the potential of microcomputing technology and have started to exploit it, insufficient information is available to countries describing in detail how this technology is being used in other developing and developed countries. Articles, books, case studies, documentation, programs and short courses would all be helpful if distributed both to and within developing countries. Training activities financed by technical co-operation programs oriented toward co-operation among developing countries could also be exploited to disseminate available knowledge more widely. Ameliorating information poverty in statistical services and their supporting computing services in developing countries is critical for improving statistical and census data processing activities.

The introduction and use of microcomputer technology represents a dramatic shift of control over technical resources and responsibility for it from the traditional central computing service to individuals and organizations who were previously end users of more centrally provided services. This shift of responsibility occurs across many dimensions; it includes responsibility for requirements analysis, system specification, negotiation with suppliers, site preparation, installation and testing, system programming, resource management, education and training, hardware maintenance, and capacity planning and management. Although the amount of responsibility shifted in each of these areas is small relative to its magnitude in a larger computing operation, the shift exists and is necessary, and failure to understand the nature and extent of such added responsibilities will make it very difficult to exploit the technology successfully.

Microcomputing technology is applicable to many aspects of census data processing. It can assist materially in the planning, logistics and management of the overall census process as well as in processing the statistical data collected. It can be applied to all phases of the data processing task that have traditionally relied upon earlier larger scale data processing equipment: data entry, data editing, correction and imputation, and generating tabulations. In addition, it can assist in producing high quality camera ready publication copy that will save printing costs and increase readability. A sufficient number of useful special purpose microcomputer based programs are available for aspects of census data processing. Developing countries can and generally should rely upon them for most aspects of their census data processing activities. Investment in such programs by international technical co-operation organizations is continuing, so that these products are quite likely to be extended and improved indefinitely.

New, non-traditional applications are also possible. Microcomputers can be used to extend the content of publications by producing graphs, charts and thematic maps. Data distributed in microcomputer readable form will be more widely usable by others in the government and in the country because of the more widespread availability of microcomputers to read and analyze the data. The traditional method of dissemination of statistics through paper-based publication can now be integrated more closely and more inexpensively with the statistical production process, with considerable potential savings of time and money and with no decrease in quality. Microcomputer based publishing tools can be integrated into the production of census outputs.

Issues of data ownership and control and data integrity, security and confidentiality, assume new dimensions in a distributed environment, and must be understood and satisfactorily dealt with by the national census office to ensure that both individual rights to privacy and institutional concerns about the quality and security of the census data are safeguarded. New forms of processing become available using the relative advantages of microcomputers and larger computers together. With the proliferation of microcomputers and the emergence of a mass market for microcomputer based software, the legal status of such software assumes substantially greater importance. Government offices need to ensure that the legal rights of software suppliers are not violated by unauthorized duplication of software.

Gains realized from using microcomputers to replace specific and traditional uses of earlier, large scale computers should not be allowed to obscure the more sizeable future benefits available from taking a considerably broader and longer view of the impact of microcomputing technology. Modern information technology provides inexpensive and effective means of data dissemination and presentation that transcend traditional methods of data dissemination through publication. In particular, models of processing dealing with image manipulation in a variety of contexts have enormous potential for making statistics usable in a dynamic manner to key users of official statistics. In this connection it should be noted that while most microcomputer environments now installed in developing countries are IBM PC compatible environments, there are a variety of other environments which are emerging and which appear to have significant current advantages over MS-DOS computing environments.. Developing countries need to begin studying and understanding such environments, like the Apple Macintosh and Unix, so that their longer term investment decisions in this technology will be informed ones and so that the relative advantages of these alternative environments can be effectively exploited in the country at the right time and for the right reasons.

Brodman, Janice, *Microcomputers in Developing Countries: Opportunities and Obstacles*, International Development Management Center, The University of Maryland, College Park, Md., April 1986, 52 pp. (An examination of the organizational issues in the adoption of microcomputer technology in Kenya and Indonesia; also an interesting bibliography.)

Broome, Frederick R. and Witiuk, Sidney W., "Census Mapping by Computer," in Taylor, D. R. F. (ed.), The Computer in Contemporary Cartography. John Wiley and Sons, Ltd., 1980, pp. 191-217. (Provides a good summary of issues in computer cartography, with interesting examples based upon the Canadian experience.)

Carlson, Beverley A. and Sadowsky, George, "Survey Data Processing in Developing Countries: Selected Problems and Issues," *Proceedings of the 100th Session of the International Statistical Institute*, Amsterdam, August 1985. (Contains a discussion of issues raised by the application of microcomputer technology to survey data processing.)

Cotton, P., Turner, M.J., and Hammond, R., "RAPID: A Data Base Management System for Statistical Applications," *Proceedings of the 42nd Biennial Congress of the International Statistical Institute*, Manila, December 1979. (A technically difficult but excellent article describing the evolution of transposed file systems and relational data bases for statistical data processing, culminating in RAPID. The development is important because such data organizations are very powerful and underlie future advances in effective data retrieval and processing of very large data bases such as census microdata files.)

Cotton, Paul, "A Comparison of Software for Editing Survey and Census Data," Presented at *Symposium 88: The Impact of High Technology on Survey Taking.* Ottawa, Canada, October 1988. (Presents criteria for evaluating microdata editing programs and compares four such programs that are implemented on the IBM PC.)

de Clercq, Robert, "Traitement des Données d'un Recensement par Micro-ordinateur," United Nations Statistical Office, Project INT/88/P09, June 1988, 45 pp. (Contains a description of the processing of the data from the 1985 Population Census of Burkina Faso by the United Nations Technical Adviser in Census Data Processing who assisted the government to execute the activity.)

Deecker, Gordon, Cunningham, Ron, and Kidd, Karole, "Computer Systems to Support Census Geography," *Journal of Official Statistics*, Vol. 2, No. 4, 1986, pp. 431-446. (A recent account of planning in Canada to extend computer based geographic mapping capacity in preparation for the 1991 Population Census. Contains a description of the evolution of geographic systems in North America and a framework for assessing enhancements.)

Denteneer, D., Bethlehem, J.G., Hundepool, A.J., and Schuerhoff, *Blaise, A New Approach to Computer Assisted Survey Processing*, Department for Automation, Netherlands Central Bureau of Statistics, Voorburg, BPA No. 13436-87-M3, November 6, 1987, 38 pp. (A description of Blaise, a microcomputer based system for survey data collection and processing, using a Pascal-like approach to codebook definition and use).

Diskin, Barbara, An Evaluation of Microcomputer Data Entry Software Packages for Use in Developing Countries for Census and Survey Processing. Computer Software Branch, International Statistical Programs Center, U.S. Bureau of the Census, Washington, 1986. (An interesting but by now dated evaluation of the first comprehensive data entry packages applicable to population census data entry).

Doucet, J. E., "The Use of New Technologies at Statistics Canada," Presented at *Symposium 88: The Impact of High Technology on Survey Taking*. Ottawa, Canada, October 1988. (A technical paper describing some of Statistics Canada's strategic thinking with regard to exploitation of new technology, including desktop workstations.)

Economic Commission for Africa, *Report of the Training Workshop on the Use of Software Packages in Statistical Computing for English-Speaking Government Statisticians and Planners: Addis Ababa, 21 September-2 October 1987*. United Nations Economic and Social Council, Document E/ECA/PSD.5/47. 28 December 1987, 20 pp.

Economic Commission for Europe, *Handbook on Statistical Data Processing Methodology*, Statistical Computing Project Report. Document ECE/UNDP/SCP/H.1. (A thorough overview of the classical model of statistical data processing in the public sector, containing both descriptions of various statistical data models and descriptions of programs developed by ECE member participants in the project.)

Engelbourg, Saul, *International Business Machines: A Business History.* Unpublished doctoral thesis, Columbia University, 1954. Ann Arbor, Michigan: University of Michigan Doctoral Dissertation Series No, 8806. 370 pp. (An excellent and thorough study of the work of Hollerith and the emergence of the pre-electronic electromechanical computing industry.)

Food and Agriculture Organization, *Microcomputer-Based Data Processing: 1990 World Census of Agriculture*. FAO Statistical Development Series no. 2a, Rome, 1987. 84 pp. (Contains an example of how microcomputer technology can be applied to processing agricultural census data; some of this material is relevant to a discussion of population censuses.)

Institute for Resource Development, *ISSA: The Integrated System for Survey Analysis*, Version 2.25 Documentation Manual, Demographic and Health Surveys Program, Westinghouse Corporation, Columbia, Maryland, 1989. (A thorough and well written manual describing ISSA, a comprehensive microcomputer based package that supports survey data processing and is in part applicable to census data processing in developing countries.)

International Cartographic Association, *Report of the Ad-Hoc Commission on Population Cartography*, delivered at the 13th International Conference of the International Cartographic Association, Morelia, Mexico, October 1987. (A document expressly prepared to provide guidance for national statistical offices in planning and executing the 1990 round of population censuses. Discusses both collection mapping and thematic mapping issues in developed and developing countries, including sources of technical and financial support. Contains a number of technical papers relevant for developing countries.)

International Statistical Programs Center, *CENTRY: CENsus data enTRY: User's Guide*. U.S. Bureau of the Census, Version 1.0, February 2, 1989. (Provides an overview of CENTRY, with major emphasis upon guidance for the user of the program)

International Statistical Programs Center, *CENPLAN: CENsus PLANning Spreadsheet for Data Processing: An Example*. U.S. Bureau of the Census, Version 1.0, 1989. (Provides an example of the use of CENPLAN in tutorial format.)

International Statistical Programs Center, *CENPLAN: CENsus PLANning Spreadsheet for Data Processing: Operator's Manual.* U.S. Bureau of the Census, Version 1.0, 1989. (Provides a detailed guide to the layout of the underlying spreadsheet and specific guidance for its use in reference manual format.)

International Statistical Programs Center, *CENTS: Installation Guide for PC Systems*. U.S. Bureau of the Census, Release 5.0, September 20, 1988. (In addition to installation instructions, the manual contains substantial examples of CENTS input and output, providing the user with a good understanding of what may be required to implement and use the system.)

International Statistical Programs Center, *CONCOR Edit and Imputation System: Installation Guide*. U.S. Bureau of the Census, Release 3.0, September 20, 1988. (In addition to installation instructions, the manual contains substantial examples of CENTS input and output, providing the user with a good understanding of what may be required to implement and use the system.)

International Statistical Programs Center, *Popstan: A Case Study for the 1980 Census of Population and Housing*. U. S. Bureau of the Census, Washington, D. C., 1979. (A comprehensive 10 volume set of statistical training documents providing extensive operational guidance for all aspects of population census activity, embodying a case study based upon the imaginary nation of Popstan.)

International Statistical Programs Center, *Population Census Data Processing on Microcomputers*. Prepared for and available from the United Nations under project INT/88/P09. U.S. Bureau of the Census, February 1989. (Contains a thorough description of population census data collection, processing, evaluation and distribution issues. Much of the document is an extension of the 1980 Popstan case study, with data processing methodology updated to reflect decisions and activities related to the availability of microcomputer based processing alternatives. Serves as a companion to this document, focusing upon the subject from an operational point of view.)

Johns, Robert, *Fundamentals of Computer Processing for Statisticians.* United Nations Economic and Social Commission for Asia and the Pacific, Publication ST/ESCAP 537. August 1987, 70 pp. (Training material used at ESCAP for training professional statisticians in the essentials of statistical data processing.)

Keller, Wouter J. and Bethlehem, Jelke G., "Computer Assisted Survey Processing on PC and PC-LAN," Presented at *Symposium 88: The Impact of High Technology on Survey Taking*. Ottawa, Canada, October 1988. (Describes the statistical production process at the Netherlands Central Bureau of Statistics and presents Blaise, a microcomputer based software system for processing survey data.)

Latin American Demographic Center (CELADE), *REDATAM: Retrieval of Census Data for Small Areas by Microcomputers User's Manual*. Economic Commission for Latin America and the Caribbean (ECLAC), Version 2.00, 31 March 1987 and *Supplemental Manual*, Version 3.1, 31 March 1989. (A description of the READATAM system, installation instructions, command language and user interface specifications, and tutorial.)

Latin American Demographic Center (CELADE), *REDATAM: Database Generation Manual.* Economic Commission for Latin America and the Caribbean (ECLAC), Version 2.00, March 31, 1987. (Used to instruct, through examples, in the generation of new data bases for analysis using REDATAM.)

Latin American Demographic Center (CELADE), *Small-Area Census Data Services by Microcomputer: Applications* *of the REDATAM System in Latin America and the Caribbean.* United Nations Document LC/DEM/R.57, Series A, No. 185, 22 August 1988 (A short report of the factors underlying the development of REDATAM and its initial uses.)

Meyers, Edmund D. Jr., *Time Sharing Computation in the Social Sciences*. Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1973. (Presents a tutorial on the concept of transposed files and why they are useful in statistical analysis. This concept underlies Statistics Canada's innovative exploitation of transposed files for rapid cross-tabulations and the development of the RAPID data base system.)

Orcutt, G. H., Greenberger, M., Korbel, J., and Rivlin, A. M., *Microanalysis of Socioeconomic Systems: A Simulation Study.* New York: Harper and Brothers, 1961. (The first complete description and implementation of a socioeconomic microsimulation model of the household sector.)

Orcutt, G. H., Caldwell, S., Wertheimer, R., Franklin, S., Hendricks, G., Peabody, G., Smith, J., and Zedlewski, S., *Policy Exploration Through Microanalytic Simulation.* Washington: The Urban Institute, 1976. (Detailed description of a complex, second generation microanalytic simulation model of the household sector, with emphasis upon the subject matter modeling in demography and economics.)

Ortuzar, Julio C. and Willard, Joseph L., ed., *Statistical Software Inventory Report*, prepared by Delta Systems Consultants, Coral Gables, Florida, under contract AID/DSPE 0027-SUB-1 to USAID, September 1983. (Provides descriptions of tabulation, editing and statistical analysis software available in 1983, most of which has by now been adapted to microcomputers. It was written to be used as a reference document for use by statistical institutions in developing countries. The booklet contains a good bibliography of further references.)

Reid-Green, Keith S., "The History of Census Tabulation." *Scientific American*, Vol. 260, No. 2, February 1989, pp. 98-103. (An interesting article describing the background and introduction of the first Hollerith machines into the U.S. Bureau of the Census in the late 1800's.)

Rex, Frederick J. Jr., "Herman Hollerith, The First Statistical Engineer," Computers and Automation, August 1961, pp. 10-13. (Provides an anecdotal description of Hollerith's initial work for the Bureau of the Census and an interesting bibliography on the subject.)

Sadowsky, George, *MASH: A Computer System for Microanalytic Simulation for Policy Exploration*, The Urban Institute, 1977, 300 pp. (A description of the use of computers for implementing dynamic microsimulation models of the household sector. This is an advanced application area, illustrating the usefulness of family and person microdata samples in computer readable form for addressing complex research and public policy issues.)

Statistical Commission of the United Nations, *Special Issues: Communications Between National and International Statistical Agencies and the Development of Data Bases*. United Nations Economic and Social Council, Document E/CN.3/1987/2. 2 December 1986, 22 pp.

Statistics Canada, SPSD/M Product Overview. Ottawa, Canada, 1989, English and French, 28 pp. (Provides an overview of the Social Policy Simulation Data base and Model (SPSD/M) developed by Statistics Canada, including a description of the synthetic microdata set created to combine tax, transfer, employment, income and expenditure data for 180,000 Canadian residents.)

Toro, Vivian and Melaney, Thomas, "The Use of Microcomputers for Census Processing in Developing Countries," paper presented at the 1987 annual meeting of the American Statistical Association, San Francisco, U.S.A. (Contains a statement of the strategic plan of the U.S. Bureau of the Census to provide microcomputer-based software to census data processing activities in developing countries, with discussion of field experience in several countries.)

Toro, Vivian and Chamberlain, Kathleen, "The Use of Microcomputers for Census Processing in Developing Countries: An Update." International Statistical Programs Center, U.S. Bureau of the Census. Presented at *Symposium 88: The Impact of High Technology on Survey Taking*. Ottawa, Canada, October 1988. (Contains information on the state of the U.S. Bureau of the Census microcomputer based international census data processing aid activities as of October 1988.)

United Nations, *The Application of Computer Technology for Development*. Department of Economic and Social Affairs, The United Nations, Document E/4800, ST/ECA/136, 1971. 122 pp. (Although dated in its references to technology, the document provides a discussion of policy, training, management and assistance issues that are still relevant today.)

United Nations, *The Application of Computer Technology for Development: Second Report of the Secretary-General*. Department of Economic and Social Affairs, The United Nations, Document ST/ECA/176, 1973. 89 pp. (An update of the first report of the Secretary General, containing a discussion of additional issues and also some comparative data on computing installations and activities in various countries which is now quite dated.)

United Nations Department of Technical Co-operation for Development, *PCEDIT*. Project for Software and Support for Population Data Processing, 14 pp. (Provides a brief overview to PCEDIT, a microcomputer program for data entry and editing, including operating instructions and a model layout file.)

United Nations Economic and Social Commission for Asia and the Pacific, "ESCAP/POP: A Computer Program for Projecting Populations by Age and Sex," *Population Research Leads*, No. 22, Bangkok, 1986. (Describes a population projection program that produces age-sex distributions of population in the future and runs interactively on an IBM-compatible personal computer. The document includes a copy of the program on a 5 1/4" diskette.)

United Nations Industrial Development Organization, *Microelectronics Monitor,* Development and Transfer of Technology Division, Department for Industrial Promotion, Consultation, and Technology, Vienna. Issued quarterly. (An informal compendium of many articles, abstracts, synopses, and news items in the general field of microelectronics including computers and software; a very useful periodic snapshot of developments in these areas, especially if alternative sources of information are thin.)

Van Est, Jan P., Maack, Udo, Witiuk, Sidney W., and Broome, Fred R., "Information Systems for Urban and Regional Planning." Prepared by SORSA (Spatially-Oriented Referencing Systems Association) in co-operation with UNESCO as the keynote paper presented at the *Seminar on Information Systems for Sub-National Planning*, Visakhapatnam, India, December 1985. (Addresses the development and use of information systems for national and subnational planning, with particular reference to the importance and use of geographic information and computer methods. A good background paper, embodying a historical and tutorial approach and an overview of current issues and techniques. Contains a useful bibliography including references to developing country experiences.)

Westlake, Andrew, "Data Dictionaries in Survey Analysis: The WFS Experience," *Statistical Software Newsletter,* Vol. 12, No. 2, July 1988, pp. 62-65. (An articulate and succinct description of specific microdata processing experience using data dictionaries as a unifying vehicle.)

Westlake, Andrew, *The Impact of 'New Technology' on Statistical Offices: Hardware, Software and Systems for Survey Data Processing.* International Statistical Institute Research Centre. August 1987, 79 pp. (An overview of some implications of new technology, much of it based upon microcomputers, for survey data processing in the future.)