DISCUSSION

John M. Leyes, Statistics Canada

Is it possible to conduct an administrative record census? Well, if the words are interpreted literally, the answer is obviously "Yes." But, if the notion of this proposal is to replace, in some way, a census of population, then, the answer must be an unequivocal "No."

In particular, and with regard to the second answer, three particular problems stand out amongst all others: coverage of the total population; identification of the characteristics of the identified population; and the geographical location of all identified members of the population.

COVERAGE OF THE POPULATION

In the absence of a census of population, then, the coverage of an administative records census (ARC) would be unknown. For this reason, it would appear that at least one test (or dry-run) would be required at the outset to assess the coverage. For this to occur, and for one to entertain the possibility of small area data, probably the earliest that an ARC could be undertaken would be 1986 in Canada and in the U.S., 1990. Moreover, and in both cases, an ARC would be only a trial run for comparison at the most.

Additionally, for an ARC to be undertaken with the most favourable opportunities for success in the coverage domain, it would be equally necessary for administrative records to be modified to enable improved estimation in the domains of (a) population characteristics and (b) geography.

CHARACTERISTICS OF THE ARC POPULATION

As is well-known, administrative records are notably deficient in the quality of data for items generally unrelated to the program being administered. For example, the purpose of a tax collection program is the collection of tax dollars, and to conduct various kinds of procedures to minimize tax evasion and tax avoidance (using the words of Revenue Canada). As a result, and quite reasonably, the efforts of tax system administrators are dedicated to revenue generation and not towards high quality control standards for, for example, demographic data, or to the collection and capture of additional data on variables such as household composition and the relationship of household members. Furthermore, even if detailed data would be obtained through the tax collection program, serious questions could be raised about the justification of including these costs in a tax administration program.

The latter point, quite obviously, is irrelevant if the costs of collection, capture, editing and so forth, are borne by a statistical program. Nevertheless, in the Alvey-Scheuren paper it was unclear as to whether additional data would be collected and who would pay the additional costs.

In the final analysis, an ARC would seem to have less value for market research, for demographic analysis, for economic analysis, and for an array of other typical uses of census data in the absence of much improved demographic data. In the absence of much-improved demographic data, it would seem clear that an ARC would represent a poor attempt to replace a census of population.

GEOGRAPHICAL LOCATION OF THE POPULATION

At the present time, the Canadian six-character postal code is a powerful geographical identifier in urban areas with door-to-door mail delivery. In rural areas, the postal code only enables a crude indication of the location of the population. By contrast, the five-digit zip code in the United States is similar to Canada's rural postal code problem. If, however, a nine-digit zip code is adopted by the U.S., then the U.S. can expect to have a series of successes and challenges that should be similar to the Canadian experience with the six-character postal code. For this reason, it seems interesting to describe some of the current experiences with the postal code in Canada.

The use of the postal code has not been universally rewarding as a means of allocating individual administrative records to geographical areas. First, and as Alvey-Scheuren have correctly noted in their paper, the geographical information on administrative records is based on mailing addresses and not on residential addresses. There is, therefore, a need to obtain both the mailing address and the residential address if the two are not the same. In Canada, a preliminary step has been taken to investigate the possibility of obtaining the residential postal code when it differs from the postal code of the mailing address.

Second, some mailing addresses have incorrect postal codes (e.g., the reported postal code may be K1A OT6 when it should have been K1A OT8). If the incorrect postal code has not been assigned by the post office, then the incorrect postal code can be identified and the code deleted. But if the incorrect postal code does exist (i.e., the post office has assigned the code), then there is no known method to (a) determine that the postal code is incorrect or (b) to assign the correct postal code.

Third, some mailing addresses do not have a postal code at all -- it is missing. Although considerable efforts have been made to assign postal codes automatically (about 10 percent of the tax records in Canada do not have mailing address postal codes), the efforts have resulted in a mixed series of successes. As is well-known, mailing addresses need not and do not conform to a standard format. Apartment numbers, for example, can be included in a number of different locations. It is difficult to write computer programs that can

decipher apartment numbers every time. Moreover, and by way of example again, street address directions (i.e., North, South, Southwest, etc.) cause considerable complications in the automatic assignment of postal codes when they are missing. Of course, it is possible to have the postal code assigned through clerical resources. But, with hundreds of thousands of addresses in which automatic assignment fails, and given that it takes a clerk about eight hours to code 200 missing postal codes, the consumption of resources is clearly prohibitive. So, there is a real conundrum -- postal codes are missing at a rate of about 10 percent, automatic coding is proving difficult, and manual coding is prohibitively expensive.

Fourth, the postal code is less than powerful in new surburban areas that do not have door-to-door mail delivery. Typically, mail service occurs in suburban areas through the use of rows of green boxes (with individual locks). The postal code identifies where the mail is to be sorted for direction to a particular post office or postal station. The postal code does not indicate where the mail is to be delivered. For example, in metropolitan areas, it is possible that the mail will be sorted and delivered to a post office in one municipality while the mail is destined for delivery in an adjacent municipality. When this occurs, it is not even possible to use the postal code to assign the administrative record to the right municipality. Thus, the postal code for rapidly growing areas is useless. And to the extent that lags of several years can occur between when a postal code is changed from suburban service to door-to-door mail delivery, to this extent considerable imprecision is built into the use of postal codes in developing suburban areas.

Fifth, some people do not obtain their mail through door-to-door mail delivery even when it is available -- some people use post office boxes. As is the case with suburban service, the postal code for a post office box number only identifies a post office or postal station.

Sixth, there is no known way to resolve the allocation problem in rural areas. For example, the same postal code is used in small communities by everyone, those residing in the community and those residing on rural routes. Ideally, one would wish, at the very crudest level of geographical detail, to be able to distinguish farmers from non-farmers in rural areas. The current single six-character postal code does not allow one to distinguish those mailing addresses associated with the population indigenous to the community and those who are not indigenous to the community. There is no simple solution to the problem. And, given the small population of rural communities and their hinterlands, a simple and cost-effective solution does not seem feasible from a central location.

The only apparent solution to the imprecision of the geographical allocation problem in rural areas is to conduct some kind of survey of these areas. Perhaps one might call this a "Census of Rural Canada (America?)." That is, if any precision is to be built into an ARC in non-urban areas, a substantial field collection or field identification component would be required. These costs would necessarily be substantial, and might exceed those incurred for similar areas in a more usual census of population.

Nevertheless, the total population in small communities and their rural hinterlands is not that great. Perhaps, rough head counts is all that may be required for the use of the data in these areas. If this is true, then, the issue of geographical imprecision may not be important at all. In fact, in the Canadian work, the rural problem has been treated from this perspective -- although important, given that there is no simple and cost-effective solution, the problem is being accepted as one of the data deficiencies when data are derived from administrative records. In the U.S., for example, it can be expected that similar difficulties will arise in rural areas, even with a nine-digit zip code. An ARC would still represent a low cost alternative to a full-blown census of population even with the known and predictable geographical problems. Furthermore, an ARC with geographical problems may be the only way to achieve mid-decade (but non-comparable) small area data for a relatively modest cost.

Based on the overall Canadian experience in converting postal codes into standard geographical areas, it is clear that millions of Canadian taxfilers can make the statistician's geographical challenges difficult. They can fail to use their postal code; they can use the wrong postal code; they can use a code other than that of their residence (e.g., that of a third party such as a tax accountant). In addition, the code may be captured incorrectly. They may also use the postal code (and address) of group residences. In fact, and in summary, in the absence of strict control over the geographical information included on administrative forms, the statistician must face a series of conundrums -- conundrums about which there are in some cases, little means of correction, and for still others, no means of detection.

While expensive fixes may exist for addressing and resolving geographical assignment challenges, the inherent attractiveness of an ARC -- its potentially low cost -- is diminished, and perhaps considerably.

A LONGER TERM VIEW OF AN ARC

At the present time, an ARC offers promise but also incorporates a number of fundamental difficulties, particularly with respect to geographical identification. From this short-term, narrow and myopic perspective, an ARC falls well short of the traditional understanding of a census. At the same time, the information and data processing industries have been making incredible strides in recent years. More and more, individual information is being stored in data banks, and there exists an increasing capacity to link data in these data banks at rapidly declining costs.

With the evolution of the new technology

and the data banks, it seems relevant to ask whether any one single person can avoid inclusion in one or more data banks? Is one of the prices of "welfare statism" and a modern and high technology-based society that no single person can escape one or more data banks? Is a further price that individuals in modern society will increasingly relinquish their claims to privacy as these claims have pertained to the linkage of government records? Have citizens already and largely foregone their claim to privacy?

With the evolution of modern data banks, and with the evolution of entitlement and tax collection programs, will the day come, perhaps in one or several generations, when statistical collection will be largely replaced, as it is currently known and understood, by the widespread use of satellite communications systems, data banks, universal and unique record keys, and so on? While this may sound like science fiction today, it is probable that statistical collection will undergo significant changes in the next one or two generations and that an ARC, however ill-advised it may seem today, will be the way to undertake censuses in the future along with virtually all other kinds of statistical collection.

REJOINDER

We would like to thank John Leyes for providing such excellent comments on some of the limitations which are apparent in an Administrative Record Census (ARC).

Administrative record environments are not alike as evidenced by the Canadian and American systems. Therefore, some of the concerns John raises differ in importance, depending on the country in which one tries to implement the idea. For this reason, we thought it might be appropriate to clarify some of the issues he raises.

Coverage of the Population

John rightly calls attention to the need for a protest of the reliability of administrative records as (even) a partial substitute for a conventional census. Because of this need for testing, we agree that it is highly unlikely that an ARC population approach could be conducted in the United States in 1990 on anything other than a trial basis. Furthermore, changes in the conventional 1990 census would probably be needed in order to make the testing workable--for example, asking for Social Security Numbers (at least on a sample basis).

Characteristics of the ARC Population

The administrative record systems that would be employed in an ARC effort are not of uniformly high quality on demographic and economic variables. Some research has been conducted, however, on the quality of the more important variables in such systems--like, for instance, the CPS-IRS-SSA Exact Match Project [1]. It appears that, while the concerns John raises are partly justified, the content errors in administrative records may be comparable to--and, possibly in some cases, even smaller than--those that would be encountered in a conventional population enumeration. The big problem with the ARC is that the concepts used for administrative purposes do not relate closely to those employed in past censuses.

Geographic Location of the Population

Perhaps the most important concern that many of us have about the population census (aside from the basic coverage question, itself) is the quality and codability of geographic data.

In our work at the Internal Revenue Service and the Social Security Administration, we have experienced many of the same kinds of problems with mailing addresses that John points out as existing in Canada. We agree, also, that, even if a nine-digit zip code becomes part of each address, there will still be many problems to overcome. Perfecting addresses in the United States for tax and program administration purposes (e.g. mailing social security checks) is something that goes on routinely, at present; however, the quality of these administrative programs is probably not, in our opinion, sufficient to be relied on in order to conduct an administrative record census. Unquestionably, therefore, it will be necessary to spend additional resources to improve this address information.

In our paper, we have mentioned two procedures, among the many that will be needed, that will undoubtedly be quite expensive. One of these is to repeat, as was done for 1980, questions on the individual income tax return about the taxfiler's residential address (as distinct from his or her mailing address). A followup with Social Security and Medicare beneficiaries on residential addresses also seems needed.

Longer View of an ARC

We agree that a longer view of our proposal may be needed to put it in the proper context. We think, however, that the technology already exists to carry out the proposal we have made (at least for the U.S. administrative record systems mentioned in the paper). Spectres of 1984 are clearly a big concern and, indeed, a compelling criticism of the approach we describe; however, administrative programs