THE DEVELOPMENT OF SOCIAL DATA FROM ADMINISTRATIVE RECORDS: THE CANADIAN EXPERIENCE

Douglas Norris, John Leyes and Nelson Kopustas, Statistics Canada

1. INTRODUCTION

The development of administrative records on individuals for statistical purposes has become a topic of increasing interest in recent years. The question is: "Why have administrative records become a focus for statisticians?" There are at least three reasons:

a) the need for social data at more frequent intervals than the data available from the Census of Population;
b) the need for social data for smaller geographical areas than can be obtained from existing household sample surveys; and
c) the need to reduce respondent burden.

Furthermore, the cost of direct data collection has been mounting rapidly in recent years. On the one hand, Statistics Canada is alleged to be collecting some information that is already available in administrative record files. On the other hand, there was the promise that the use of administrative records could lead to reductions in the direct expenditures associated with data collection.

Of course, the use of administrative records for statistical purposes is not new to Statistics Canada. Administrative records have been used for a long time for statistics on institutions and selected economic statistics (e.g., education, public finances, hospitals, international trade, and so on). But for statistics on households, families and individuals, almost complete reliance had been placed on household sample surveys and censuses of population. However, with the development of large machine-readable files, the exploitation of these records for social statistics loomed ever larger as a potentially rich and inexpensive source of annual small area social statistics.(1)

A further stimulus to the developmental work with administrative records can be attributed to the growing demand for more and more social data. Prior to World War II, the number of social programs was relatively limited. Since that time, however, social programs have proliferated and grown. Now, as a result, there is a much larger demand for geographically detailed data in the social statistics domain as social scientists are developing, analysing, and evaluating social policies with respect to national goals and priorities.

In summary, then, Statistics Canada is studying administrative records as a means to produce more frequent small area data with less response burden. Indeed, it must also be recognized that administrative records hold the potential to reduce direct data collection costs, and administrative records may also prove instrumental in helping to meet the virtually insatiable appetite for more and more social data.

In pursuit of this potential, a small research and development unit, the Administrative Data Development Staff was established by Statistics Canada in 1979 to explore the feasibility of developing small area social data from administrative records that would be comparable to those derived from censuses of population and household surveys. The areas identified for initial developmental work included population, migration, income, family composition, and the labour force.

The remainder of this paper has been divided into four sections. Following this introduction there is an overview of the principal administrative records that are currently being used to develop small area data on households, families and individuals. In the next section, three main areas of developmental activity are described and for each there is a discussion of progress to date and an indication of where the research effort is headed. This is followed by an outline of some of the limitations and technical problems that have been encountered in developing the administrative records for statistical purposes. Finally, there is a discussion of the potential roles administrative data may play in meeting the Canadian social data needs in the 1980s and beyond.

2. MAJOR ADMINISTRATIVE RECORDS SYSTEMS

In its broadest sense, administrative data development for statistical purposes could include a vast array of record sets and an almost unlimited array of production possibilities. Attention, however, has initially focussed on those record sets currently held and maintained by federal government departments. In particular, four sources were identified for development: personal income taxation records, unemployment insurance beneficiary records, family allowance records and old age security records. Each file contains the mailing address and/or postal code, the key to geocoding the data at a small area level.(2)

The Revenue Canada personal income tax records hold the greatest social statistics potential of those files currently available in Canada. There is a high coverage of the population (in excess of 90% if dependents are included), especially of the labour force; an acceptable range of demographic information; and detailed income information by type.

There are several important shortcomings of the tax file: individuals are only required to file under prescribed circumstances; many individuals with low income do not file; and because of tax law and regulation changes over time, the intertemporal comparison of social data derived from the personal income tax files must be qualified accordingly.

The unemployment insurance beneficiary file is a second potential source of information at the small area level. This administrative file includes all beneficiaries including the seasonally unemployed (e.g., fishermen), voluntary quits, and women on maternity leave. Clearly, the unemployment insurance program covers persons not generally considered to be unemployed in surveys of the labour force. At the same time it may exclude some persons usually counted as unemployed (e.g., new entrants to the labour force). The long-term potential of this file will be determined in large measure by the success achieved in reconciling such coverage differences and using the file in conjunction with monthly Labour Force Survey

statistics to model the labour force in small areas.

The third major administrative record file is the family allowance file, a file of virtually all children resident in Canada, under the age of 18 years. The basic strength of this file is the almost universal coverage and the principal shortcoming is the limited information on the file.

The fourth administrative record file is the old age security file. It includes selected information on Canadian residents 65 years of age or older who meet the residency requirements of the Old Age Security Act. This virtually universal program for the population of Canadian residents 65 years of age and older also includes income information on a low income sub-group qualifying for supplemental benefits. Perhaps the most important shortcoming of this file arises because the mailing address can differ considerably from the residential address of old age security beneficiaries, especially when the benefits are paid to trustees rather than to the elderly themselves.

3. PRINCIPAL DEVELOPMENTAL ACTIVITIES

The principal developmental activities can be summarized as follows:

a) development of infrastructure,
b) file development, and
c) production of experimental data series.

a) Infrastructure

The development of administrative records for statistical purposes required an investment in infrastructure -- e.g., geocoding and investigation of content and coverage of files.

Geocoding is achieved by using the six-character alpha-numeric postal code of the mailing address. In urban areas the postal code identifies a block face. In all the other areas, the postal code identifies a post office and its hinterland service area. A substantial amount of work has been done on:

i. investigating the accuracy and completeness of the postal code on various files;
ii. developing files to convert postal codes to a variety of area systems ranging from census tracts to counties and provinces; and
iii. assessing the accuracy of these conversion processes.

A second investment in infrastructure has required an assessment of the content and coverage of the administrative records -- the investigation of concepts, collection and processing procedures, editing and imputation; and an assessment of the extent to which the data elements are internally consistent and are consistent with the data from alternative sources such as censuses and household surveys. To be successful, this activity requires the close cooperation of program departments that have collected and processed the data.

b) File Development

i) Cross-Sectional Data Base

The Revenue Canada-Taxation personal income tax file, because of its large taxfiler and dependent-to-population coverage, has been designated the nucleus for file development. The file contains demographic information on the age, sex and marital status of each taxfiler, and a detailed array of income information by source. In particular, it is possible to identify persons reporting income from employment activities and from unemployment insurance. These can be used as proxy variables to indicate gross labour force activity.

The work on geographic coding has permitted the tax file to be coded at the county or census division level and at the census tract level for a few metropolitan areas. Other ongoing work is dedicated to extending the range and detail of the geographic codes on the file.

Work is currently underway to augment and enrich the tax file by identifying households and families and by coding the industry of employment from establishment records. Meanwhile, the pilot work at IRS to code occupation is being monitored as a possible avenue for future Canadian development. (See Sailer, et al, 1980.)

Plans are also being made to undertake a pilot study to investigate the feasibility of integrating information from other administrative records.(3) For example, linkage between the tax file and the family allowance file would provide the ages of children. The linkage of the tax file with the old age security file would enable the addition of information (including income information in some cases) on the one-third of old age persion benefit recipients who do not file tax returns. Finally, the direct linkage of the tax file with the unemployment insurance beneficiary file would allow the coding of the unemployment insurance benefits by type (e.g., fishing, maternity, sickness and disability). The latter linkage may enable a measure of unemployment that is more consistent with the conventional definition, and enable inferences to be made about the number of periods of unemployment and their duration.

ii) Longitudinal Data Bases

Administrative records with a unique record key offer the potential for creating longitudinal files by linking the same individuals over time. Since 1967, a 10% longitudinal file of taxfilers has existed in Canada. This file consists of demographic and detailed income data by source, including income from employment activity and unemployment insurance benefits. The file currently covers 12 years and is updated annually. For purposes of comparison, it is similar to the Continuous Work History Sample (CWHS) that has been widely used in the United States. (See U.S. Bureau of Commerce, 1976.)

Thus far, the 10% longitudinal file has not been used extensively. An important barrier has been the massiveness and complexity of the data file. Currently, software is being developed to allow a more effective and efficient manipulation of this file in the future.

c) Production of Experimental Data Series

An important aspect of the developmental program has been the identification of promising data series and the production of a limitted number of experimental series. These data are being made available to expert users and producers of data to allow them to assess the quality and relevance of the data with respect to their own data needs.

The production of annual small area migration data was deemed a most promising area for development. The tax data have been used to model migration flows between counties or census divisions for the total population by age group and sex. To date, experimental annual data have been produced for the period 1971-1979.

A second promising area was the production of income and gross labour force data for small areas. In this case the tax records have been used to produce experimental tabulations by age, sex and marital status. Initial developmental work has been completed for 1976 and data for 1977, 1978 and 1979 are currently being produced.

The tax records have been used to generate money income estimates for counties or census divisions.(4) Plans have also been completed to undertake a pilot study to investigate the feasibility of using the tax records and the unemployment insurance records to estimate small area unemployment rates.

4. SOME LIMITATIONS AND POSSIBLE SOLUTIONS

In developing the statistical potential of administrative records, a variety of well-known problems must be resolved.(5) A few specific limitations and technical problems relative to the Canadian experience are outlined below.

a) Geographic Coding

The successful exploitation of administrative records for producing small area data depends on the mailing address as an indicator of residence address. In general, this approach produces acceptable results, although it is clearly more difficult to identify rural areas precisely. Problems also are more frequent in urban fringe areas and overall, reliability is inversely related to the size of the community. Improvements in the use of mailing address are possible through inter-file linkages, or through the addition of a question on residence address to the administrative form. The latter approach is used by the Bureau of the Census to obtain the mailing and residence addresses from individual tax returns, although this approach is a costly one. (see U.S. Bureau of the Census (1980).)

b) Coverage

The tax file, when taxfilers and dependents are combined, covers about 90% of the Canadian population. It is expected that an additional 5% might be added through inter-file linkages. Nevertheless, selected sub-populations may not be covered by any administrative records system. For example, it may not be possible to identify low income persons who do not file tax returns, or the unemployed who do not qualify for unemployment insurance benefits. Furthermore, data may be incomplete since different record sets cover different populations and contain different information. These problems are not unique to administrative data. Census and survey data also contain inconsistencies that are at least partly resolved through edit and imputation procedures. Similar procedures may be just as effective for administrative records as well.

c) Accuracy

Are administrative data accurate? In general, administrative data are quite accurate. In fact, the perceived accuracy of administrative data is reflected in the fact that some survey and census questionnaires instruct respondents to check administrative records in supplying requested information. This accuracy stems, in part, from the recognition that information collected for program purposes can be, and often is, validated from independent sources.

There are, however, differences between the data required for program purposes and those that are not. Moreover, although the original data files used for program purposes may contain highly accurate data, there may be some slippage in accuracy on derivative files created solely for statistical purposes. Errors may begin to creep in during subsequent processing steps, especially when statistical codes are added. Clearly, the derived data elements are not subjected to the same level of scrutiny as the data required for program purposes. The assessment of accuracy and the resolution of any problems require the close cooperation of both the statistical agency and the primary collection department that provide the data files.

d) Comparability

An important characteristic of any data series is its comparability to other data. Since administrative data are generally based on specific program concepts and definitions, these will often differ from traditional statistical measures. The analytical importance of these differences must be carefully assessed by both the producers and users of data. Although new and different data series may seem inconsistent with the policy of standardizing concepts and definitions, analysts are provided with increased flexibility in choosing the series most useful for the work at hand. Additionally, the existence of complementary statistics provides the opportunity to use statistics to validate and extend the studies and analyses of social problems and policies. From this perspective, as long as data analysts and users have a thorough description of the sources, methodologies, limitations and so on for the alternative choices, their ability to select the most relevant is enhanced.

Another dimension of comparability pertains to the inter-temporal stability of a data series. This can be a serious problem with administrative data since the coverage and contents of administrative records can change suddenly as a result of legislative and/or administrative changes. It is essential, therefore, that data series be selected and defined to minimize the impact of sudden changes in legislation or regulations.

5. ADMINISTRATIVE SOCIAL DATA: SOME POSSIBLE ROLES AND CAVEATS

What will be the role of administrative records in meeting the social data needs in the 1980s and beyond?

Historically, administrative data have played a limited role in meeting data needs in the domain of households, families and individuals. But with the evolving computer technology, it is clear that more small area administrative social data will become available in the form of direct tabulations and modelled estimates. In assessing the further potential of administrative social data, several issues must be considered.

The first issue pertains to the actual data required by users. The informational requirements of society are increasing and changing rapidly in response to the proliferation of government programs. Data are needed to plan, implement, evaluate, and modify these programs. This

demand for data has been reflected in requests for more frequent and timely data at the small area level. And clearly, census data produced once every five or ten years have a diminishing value in a rapidly changing society. Administrative data, on the other hand, can be produced annually and with somewhat better timeliness.

However, the statistical limitations of administrative data are clear: concepts and definitions are based on program requirements; coverage of some segments of the population may be weak; and geographic delineation is based on mailing addresses. Nevertheless, one might ask, "How important are these limitations?" If the objective is to have data that represent the correct level, then administrative data can fall short. But if the data need only reflect consistent spatial and temporal relationships for policy purposes, then administrative data may meet a much larger array of user needs than seems obvious at first glance.

A second issue pertains to the possibility of enhancing the usefulness of administrative records for statistical purposes. One approach to increasing the quality and content of record files is to engage in micro-record linkages.(6) Although the results might be a richer and integrated cross-sectional data base, the idea is somewhat awesome as it could be interpreted to represent a massive invasion of privacy. One way of avoiding this threat of Big Brother is to engage in micro-linkages on a much reduced scale. For example, micro-linkages could be made between the enriched cross-sectional file discussed above and the monthly Labour Force Survey, say rolled-up over two or three years. While this micro-linkage would not be very powerful for small areas, it would not be a very massive micro-linkage project either. At the same time, it would offer an array of tabulations that would exceed the possibilities of either administrative data or survey data alone.

In the longer term, another approach to increasing the statistical potential of administrative records is to ensure that the content, collection and processing of administrative records include both statistical and administrative objectives. Unless statistical objectives are formally incorporated, administrative data will suffer an array of deficiencies that will be beyond the scope of traditional statistics.

A third issue relates to the suggestion that the use of administrative records represents a way to reduce the cost of census-taking. In a strict definitional sense, administrative records cannot be considered a replacement for the census. Some populations are not covered, and the concepts and definitions depart from those of the census. Consequently, data derived from censuses and administrative records are different.

However, a combination of imputations for non-covered populations and conceptual and definitional reconciliations can reduce the differences. If the apppropriate imputations and reconciliations are successful, some administrative data may become acceptable surrogates for census information at a small area level of detail. At this point, judicious assessments would seem necessary for determining whether census replacement should be undertaken, and if so, how to best redeploy any resources freed by the cost savings. Should the resourcces be used to collect entirely new census data or should the resources be used for some other statistical objectives?

On the other hand, fiscal considerations may transcend the issue of data trade-offs. Funds for census-taking may be reduced and force a reduction in the size and scope of the census. If such a scenario should occur, then administrative records may represent the only feasible alternative for the production of some small area social data.

The question posed at the beginning of this section remains unanswered. "What will be the role of administrative records in meeting the social data needs in the 1980s and beyond?" There is no simple answer to this question. Data trade-offs and the fiscal trade-offs must be made. At the same time, there are pressures to produce more small area data, to reduce the burden on respondents, to ensure that the principle of privacy is not violated, and to reduce the costs of collecting and producing data. These objectives and needs are frequently in direct conflict. Therefore, the role of administrative records in the Canadian statistical system is likely to evolve as a product of political, fiscal and statistical trade-offs. Nevertheless, it seems clear that whatever the trade-offs are, administrative records will be of growing importance in meeting the social data needs of the future.

ACKNOWLEDGEMENTS

The authors would like to thank many colleagues at Statistics Canada who reviewed earlier drafts of this paper and made many helpful suggestions for improvements. Many thanks also to Louise Saucier who patiently typed several drafts of the paper.

FOOTNOTES

1. For a discussion of the role of administrative data in social statistics in Canada, see Rowebottom (1980) and Fellegi (1980).

2. For more detailed descriptions of the contents of the administrative records see Leyes(1980).

3. An excellent example of the use of record linkage is provided in the series, Studies from Interagency Data Linkage. A summary of this work is provided in Kilss and Scheuren (1978).

4. See Statistics Canada (1974). Similar estimates for more recent years (1976 and 1977) are currently being produced and will be available in the near future.

5. For a discussion of some of these problems see U.S. Department of Commerce (1980), especially Chapter VII.

6. The problems of creating large national data banks based on matched records have been extensively considered by Dunn (1967) and Hansen (1971).

REFERENCES

Dunn, Edgar S. (1967) "Statistical evaluation report No.6 - review of proposals for a national data center." In The Computer and Invasion of Privacy, pp.254-276, Hearings before a Subcommittee on Government Operations, House of Representatives. U.S. Government Printing Office, Washington, D.C.

Fellegi, I.P. (1980) "Data, statistics, information - some issues of the Canadian social statistics scene." Statistical Reporter, 80-7, pp.168-181.

Hansen, Morris (1971) "The role and feasibility of a national data bank, based on matched records and alternatives." In Federal Statistics, Report of the President's Commission, Volume II, pp.5-61.

Kilss, Beth and Scheuren, Fritz (1978) "The 1973 CPS-IRS-SSA exact match study: past, present and future." Policy Analysis with Social Security Research Files, Proceedings of a Workshop held March 1978 at Williamsburg, Virginia, United States Department of Health, Education and Welfare, Office of Research and Statistics, Research Report No.52, pp.163-194. Washington, D.C.

Leyes, J.M. (1980) "Developing administrative data: some research and development prototypes underway in Statistics Canada." Proceedings, Statistics Canada Geocartographics Workshop, Ottawa, Canada, August 12, 1980 pp.50-64. Carleton University, Ottawa, Canada.

Rowebottom, L.E. (1980) "The utilization of administrative records for statistical purposes." Survey Methodology. Vol.4, No.1, pp.1-15.

Sailer, Peter J., Orcutt, Harriet, and Clark, Phil (1980) "Coming soon: taxpayer data classified by occupation." Economic and Demographic Statistics. Selected papers given at the 1980 Annual Meeting of the American Statistical Association in Houston, Texas. United States Department of Health Services, Social Security Administration, pp.187-192. Washington, D.C.

Statistics Canada (1974) Income Estimates for Counties and Census Divisions (Catalogue 13-204), Information Canada.

United States Bureau of the Census (1980) Current Population Reports, Series p-25, No.699. Population and Per Capita Money Income Estimates for Local Areas: Detailed Methodology and Evaluation. U.S. Government Printing Office, Washington, D.C.

United States Department of Commerce (1976) Regional Work Force Characteristics and Migration Data: A Handbook on the Social Security Continuous Work History Sample and Its Application. U.S. Government Printing Office, Washington, D.C.

United States Department of Commerce (1980) "Technical problems in the statistical use of administrative records." In Statistical Policy Working Paper No. 6: Report on Statistical Uses of Administrative Records, Chapter VII, pp.81-90. U.S. Government Printing Office, Washington, D. C.