ALTERNATIVE SOCIO-ECONOMIC DATA COLLECTION METHODS IN THE 1980's: THE CONCEPT OF A SYNTHETIC DATA BASE

Mukund Nargundkar and Anis Ashraf
Statistics Canada*

This paper makes a proposal to create a new type of information bank, the "Synthetic Data Bank". This type of bank would involve linking information from two data banks to create a third. The result would be that much greater use could be made of existing data banks in conjunction with new data collection activities. This would mean a significant reduction in the amount of data to be collected which, in effect, could potentially reduce both data collection costs and response burden. The paper suggests a number of considerations in developing statistical techniques to facilitate the creation of such an information linkage concept. Some of these techniques are to be found in modern literature' others may well have to be developed.

INTRODUCTION

Ever since the dawn of history data collection or information gathering has played an important role in human activities and decision making processes. In ancient times, feudal lords used such information to determine the manpower needs for the protection and preservation of their domains and sources of revenue, as well as to serve their system of distribution at times of feast or famine.

In modern times, the fundamental principal of data collection remains unchanged. Today, data is collected by all human beings in their capacity as public officials, private entrepreneurs, academicians and at international levels by United Nations. Modern data collection activities, however, could be categorized as follows:

*The views expressed in this paper are those of the authors only, and do not represent departmental policy.

a) administrative data, i.e., collection of revenue, law enforcement, etc.
b) socio-economic data, i.e., census of population, housing, agriculture, manufacturing, balance of payments, etc.
c) data on physical sciences, i.e., soil, minerals, fisheries, etc.

Although it is an established fact that all components of the data mentioned above are essential for the betterment of mankind in one way or another, for this paper we would like to concentrate mainly on administrative and socio-economic data.

First of all, the question arises: why is it imperative to collect socio-economic data? The answer lies in the needs of the different sectors of our economy. In the public sector, this type of data serves the purpose of policy formulation at the executive, judiciary and legislative level. In the absence of reliable data on pertinent subject matter areas, no meaningful laws can be made or policies formulated for the good of society at large. Production, resource allocation, marketing and advertising strategies in the private sector would be severely hampered if reliable information on socio-economic variables was not readily available. Basic research by academia is also reliant on similar data.

One may ask what are the source of socio-economic information? The sources could be innumerable but ultimately, the one basic source is still man, as were his forefathers in ancient times. It is he who eventually bears the brunt of all data collection agencies, as a householder, as an entrepreneur, as a consumer and even as an alimony-paying husband. However, it is imperative that such information be collected from individuals since no modern society can continue to progress in socio-economic and administrative areas without reliable information. Contrary to this, at the same time, it is the total commitment of society not to infringe unduly upon issues that are perceived to be private or sensitive by an individual. This then is the dilemma confronting data collection agencies and the freedom of individual rights and obligations.

In our times, statisticians have developed scientific methods of data collection that, to a large extent, satisfy the demands of the data users in terms of a) appropriateness, b) timeliness, c) measure of error (quality)

and d) the efficiencies of cost. Up until recently, this was all dependent on a hundred percent count of the population under study. Consequently, a very large commitment of resources was required for the collection, compilation, tabulation, analysis, dissemination and storage of data. In other words, a large staff was continuously buried under statistical data and computer printouts. As you can well imagine it was a managerial nightmare to produce reliable data for governments and industries to help them formulate their policies and develop their strategies within a given time frame. It again created tremendous pressure on the individual who under law was obliged to continually supply data on practically every aspect of his life. Legislators, on one hand, required the information, while on the other hand, they felt obliged to look after the interest and welfare of their electorate (constituents). The collectors of data were directed by law to implement the programs of information gathering by the policy planners with the policy planners being so directed by the electorate. Thus, it became a dilemma for all and a vicious circle with no solution in sight.

In the face of "mutual harrassment" among major segments of the society, statisticians in certain countries were forced to come up with a solution to achieve their objectives without going through the entire process and which would reduce response burden to its minimum, and yet optimize (or at worst maintain) the efficiency, reliability and timeliness of the statistical output for the users.

The pioneers in the field, such as Professor Mahalanolis and Sir Ronald Fisher propounded the theory that a probability sample survey could provide satisfactory and reliable results using considerably limited resources and in a shorter time period. Thus, this was the dawn of a new era in the field of reliable data collection.

Now, we have to examine the development of sampling theory and its application in the context of modern industrial society, and the answers it provides to the questions raised earlier. The statistics estimated from a sample survey are subject to two kinds of error: a) sampling error, and b) nonsampling error, generally expressed as Mean Square Error (MSE) of the estimated statistics, more commonly known as the measure of the quality of the data.

Let us examine the sampling error first. Ever since the inception of sampling theory, survey statisticians have, almost exclusively devoted their energies towards minimizing sampling error in order to increase the accuracy, efficiency and reliability of the statistics. Furthermore, numerous methods for the selection of probability samples have been developed, together with "estimators" for estimating the characteristics and their sampling error (variance). On the other hand, the nonsampling error has remained an elusive and nonquantifiable entity, inasmuch as direct measurement is concerned. The theoretical development for estimating the nonsampling error, is justifiably empirical in nature and still in its infancy. Techniques such as interpenetrating samples, experimental designs or other statistical methods have had limited success. Nevertheless, survey statisticians in Canada and elsewhere are still engaged in perfecting this area. Their problems and frustrations centre around factors contributing to the nonsampling error. These are: a) respondents' inaccessibility, refusal to provide any information or providing incomplete information, and b) errors introduced in recording, transcribing and processing the information.

Considering these factors, it is our opinion that the basic premises of data collection methods have to be re-evaluated in terms of their efficiency and reliability in estimating the total error of a statistics, i.e., sampling as well as the nonsampling errors. It is conceivable, in a number of situations, that nonsampling error in fact may exceed the sampling error. In all such cases, inferences or decisions based only on sampling error could well be misleading. While this phenomenon is under study, we do not foresee a radical change in the attitude of a respondent towards expending his energies in satisfying the insatiable data requirements of the data collection agencies in the foreseeable future, namely the 1980's. Techniques such as self-enumeration, use of incentives, telephone and mail interviews have, until now, produced only marginal efficiencies. If the present practices of data collection are continued with only patchwork modifications to overcome present difficulties as well as meet anticipated changes in the attitudes of respondents towards collection of data, it will not only skyrocket the cost of data collection but will also deteriorate the quality of data to the point where it will be of little use in the decision-making process.

For the statisticians of the 1980's it is, therefore, essential to minimize response burden through a) alternative methods of data collection, and by b) maximizing the analytical value of the existing data bases without compromising the principles of timeliness, quality, appropriateness and cost efficiency of the necessary statistics.

THE SUGGESTED ALTERNATIVES

In our opinion, the first alternative towards producing statistics of high quality without reverting to the traditional methods (personal interview, etc.) would be to obtain relevant statistics on socio-economic variables from existing administrative data bases. As was pointed out earlier, a tremendous amount of information exists on a variety of subject in one form or another in various centralized government and industrial data banks. Moreover, the quality of administrative data, because its provision is mandatory rather than voluntary, is understandably superior.

In order to derive socio-economic statistics from administrative data banks, it would be necessary first of all to develop certain procedures. These could generally fall into four categories:

- develop methods to insure that the derived statistics are, appropriate (relevant), up-to-date (proper reference period), and complete in terms of their coverage and content

- develop a wide variety of coordination techniques in order to link two independent data bases

- synthesize two or more data bases, administrative and/or socio-economic, for the purposes of creating a new data base for a given set of socio-economic variables

- develop "chartered" statistical data banks

Generally speaking, all socio-economic surveys require, in one form or another, variables such as age, marital status, income and similar time-oriented characteristics. These facts can no doubt be obtained from administrative data banks without necessarily making them a part of a survey questionnaire. Besides these, time-invariant characteristics such as mother tongue, place of birth, blood type, etc., can also be obtained from similar sources without any additional effort.

To further elaborate the basic points raised earlier, it appears logical to give scientific examples for each category here.

Let us, therefore, now examine how we could make an administrative data base appropriate for a given socio-economic study.

For example, let us assume that we want to convert an administrative data base generated from tax returns to create income distribution of census families. Here it could be possible to reorganize the data set to create "census families" from the data files by grouping people (including dependents) with the same last names and same addresses and hence create a data set appropriate for generating family income distribution. In a situation where the data set cannot be reorganized to create "families", a statistical model with available auxiliary and/or external information, could synthetically match people to create "families". A model used to create the data base appropriate for statistical purposes would depend upon the content of the data base and the statistical requirements for which it has to be used. Thus, one would be converting the data base to be appropriate for family income statistics. This created data base is a synthetic data base in a sense that it is generated from a basic administrative data base using a statistical simulation procedure and not from an "actual" data collection procedure. Also it is important to note that the synthetic data base represents information about the population at some aggregate level and not at a micro or individual level.

Another possibility for making administrative data appropriate for a statistical purpose is to modify the objectives of statistical analysis so that the available data set can become appropriate. For example, if the objectives are modified such that a distribution of individual income could be used instead of family incomes, the data set generated by the tax returns would become appropriate.

The use of administrative data depends also upon the compatibility of the reference period to which the data refer and the reference period for which statistics have to be derived by using the data set. For example, the data set generated from tax returns relate to income during the previous calendar year while the statistics to be derived from the data set refer to the current fiscal year. However, in "actual" data collection, information sought about income often relates to or is provided relative to the last full taxation year.

In such situations, the data set can be updated by the use of projection and/or forecasting techniques to create a synthetic data base having a reference period compatible with the objectives. Again there are numerous projection nd forecasting procedures available from which an appropriate one can be selected. The appropriateness of a particular procedure would be determined in terms of the ease in calculating and controlling the quality of the statistics and the availability of "external" information needed. It should be noted that progressive as well as regressive projection in time can be achieved according to the needs.

The use of adminstrative data likewise depends on the compatibility of the target population of a statistical study with the population covered by an administrative data set. For example, an administrative data set generated from family allowance information would not cover those families in Canada who do not have dependents eligible for family allowance. Hence, this administrative data set may not be useful for statistical purposes where the target population is all census families in Canada. In such situations, administrative data set has to be supplemented and/or missing units have to be generated using auxiliary information.

Also, in situations where an administrative data set has most of the variables required for a study but is lacking some variables, other data bases, auxiliary information or statistical techniques have to be introduced. The information on missing variables can be generated by regression models or other information methods. The supplementary data to overcome the shortcomings of the coverage and/or content of a given data base by generating synthetic data base could be at micro or macro levels. On a larger scale, this leads us to the second phase of our discussion.

The problem of linkage of data sets has in the past related to linkage at micro levels. This is more commonly known as the problem of second linkage. However, the linkage considered in generating a synthetic data base need not be restricted only at micro levels but can be extended to "statistical linkage". The extension of statistical linkage enhances the use of data sets to a greater extent.

Extensive work has already been done in the area of micro record (data) linkage and is not further discussed here. Some work has also been carried out with regard to data linkages which can be termed as the "statistical linkage". Developments in the areas of data simulation, model analysis, multi-variate regression and such statistical techniques can be employed in carrying out the Statistical Linkage.

The Statistical Linkage is a statistical technique that can be effectively used to generate a synthetic data set by "combining" two or more data sets while providing quality measures for the statistics derived from the synthetic data set.

With this general definition of a statistical linkage system, one could say that the sky is the limit on the ways in which statistical linkage can be carried out. It is not possible in this paper to go through in detail the ways of adopting statistical techniques that can be used for generating a synthetic data set. We will, however, provide an overview with a few examples.

It should be noted that the statistical techniques presently available are not all developed for statistical linkage per se but have to be adapted. Further new statistical techniques specifically for use with data linkage require to be developed. This is the new area in the development of information for decision-making that survey statisticians have to recognize and develop. If this is not done, those sociologists, psychologists, market researchers and other decision makers, etc., who are not statisticians but are heavy data users, will have to develop their own methods of generating the information as they have done in developing and using statistical techniques in such areas as psychometrics, conjoint analysis, etc.

We must make it again explicitly clear that the synthetic data base generated would represent the target population at aggregate level and would be useful only for the derivation of statistical inferences. We also emphasize again at this point that a unit in a synthetic data set does not exist individually but represents the population at aggregate level. Also the statistical linkages we are discussing are not restricted to administrative data sets but to statistical and other types of data sets. Therefore, when we refer to a data set it could be either administrative, statistical or any other data set. The only conditions that a data set has to satisfy as a potential member for statistical linkage are:

i) each unit (record) is uniquely identified with a set of information;
ii) the "target population" of the data set is clearly defined;
iii) the form and format in which the data set is available can easily be made computer processible.

The third phase in this exercise is to synthesize more than two data sets.

Now we will consider some examples to classify the generation of a synthetic data base.

Suppose that there are two data sets on a population. One data set, D1 has basic demographic characteristics of the units (individuals) along with a set of characterisrics C1, and the other data set D2 has also some basic demographic characteristics of units (individuals) and another set of characteristics C2. Suppose also that the objective of a statistical analysis requires that information of characteristics set C1 and C2 be both available on a data set, along with the basic geographic characteristics.

If all units on D1 are on D2 and there is a common identifier to link two sets at micro level, and if the characteristic sets C1 and C2 refer to the same time period, the required data set with both characteristics sets C1 and C2 can be obtained without any statistical linkage. One of the other possibilities may also exist as a combination of one or more of the following situations:

i) the reference periods for C1 and C2 are not the same;

ii) not all units of D1 and D2 are common and hence neither D1 nor D2 completely cover the target population, P. However P = D1UD2;

iii) there is no common record identifier to link the two sets of data using a micro linkage procedure.

First the characteristics out of sets C1 and C2 can be identified that are time invariant (e.g., mother tongue, place of birth, etc.). Those which are time-variant can be adjusted to the required reference period by a variety of estimation or projection methods. To carry out the estimation/projection, auxiliary information may be used. For example, suppose one of the elements of C1 is annual income data which is one year old while current income is required. The simplest (but crude) procedure to up-date the income would be to adjust the old income by a factor derived from such auxiliary information as the increase in GNP, CPI, etc. For that matter, a set of regression curves based on previous data can be used to project the income to the current period. The auxiliary information used and the procedure adopted to carry out the projection would certainly depend upon the objectives of the statistical study. Similarly, the basic demographic information can be brought up-to-date for both sets of data.

The data set D1 or D2 does not cover the target population P, but as we assume that P = D1UD2, i.e., collectively they cover the target population. Therefore, before conducting a statistical data linkage of characteristics sets C1 and C2, we have to adjust data sets so that each set would cover the target population.

This can be done by generating f_1 and f_2, two sets of dummy units to add to D1 and D2 so that they will cover total population where

$f_1 = D2 \rightarrow D1 \,\&\, D2$

$f_2 = D1 - D1 \,\&\, D2$

To generate f_1 and f_2, the common units between D1 and D2 have to be identified. This can be done, for example, by comparing age-sex distribution of units from data sets D1 and D2. It should be pointed out that the identification of common units may not be based on the exact match or the age-sex value but would be based on some statistical rule such as the rules used in identifying the outliars or inliars in a data set.

After adjusting the sets to cover total target population, a "hot deck" imputation technique can be used to imput the characteristics for dummy units within each set. To carry out statistical data linkage, the "hot deck" imputation technique can further be used to impute "missing" characteristics on one set using the other set as "hot deck" and using basic demographic characteristics as correlated variables and maintaining the distributions of imputed characteristics same as on the original data sets. With this procedure we would have two synthetic data sets. These two synthetic data sets would represent the same population and hence may be used to evaluate the quality of the statistics derived from these sets.

In our example, we assume that a common set of demographic characteristics exists on both sets, however this does not have to be the case. However, without some common set of characteristics, use of external auxiliary information has to be made, for example correlation parameters between C1 and C2 if correlation exists.

Also, it was assumed that D1UD2 = P. However, set D1 and/or D2 could be a sample data base. In this situation the derived synthetic data set could also be a sample data base and therefore one would need to generate "weights" for estimation purposes.

Similar statistical linkage can be extended for more than two data sets to generate a synthetic data base.

The development of "chartered" statistical data banks is the fourth and final process in the discussion of alternative sources of information.

The data banks referred to here are not just the data banks as presently understood but data banks in the real sense as the chartered banks, where various data collection agencies (i.e., government, industry, academia) would deposit the data. Then the banks which are created, would develop the synthetic data bases for data users for statistical purposes.

This is not a far-fetched proposition, as presently such data banks exist within the Federal Government, e.g., CANSIM in Statistics Canada. In the private sector, the Better Business Bureau is a data bank of this type. The information banks of the future could be cooperative data banks where members would deposit data to generate wider bases of synthetic data using the data sets of various members which could be withdrawn by the members for use in their decision making. In the advent of the technology such as TELIDON, such a reality is potentially imminent.

CONCLUDING REMARKS

It can be seen from the above observations that:

a) socio-economic data can be obtained from the administrative data already available from the government agencies;

b) linkage of data bases can be developed without too much effort, although it would be necessary to develop highly refined and sophisticated procedures, and of course it would be necessary to train and develop additional expertise in this new field;

c) the establishment of "chartered" statistical data banks requires further research for implementation procedures. For these to function efficiently, the fullest cooperation of all concerned is a must, whether they be custodians of administrative data or socio-economic data.

It is our opinion, based on the presentation of our analysis of this area, that with these methods having been adopted and made functional we will not only save millions of dollars, hundreds of man-hours, but above all will save and conserve the source of information, in this case, the respondent himself, from the continuous pressure of data collection agencies.

Now it is up to us, the statisticians, to relieve the respondents of response terror and to relieve us, the data collection agencies, of mathematical error. Perhaps the application of the above-suggested alternatives could help solve the problem.

ACKNOWLEDGEMENT

The authors wish to acknowledge the valuable contribution of Mr. S.M. Ashraf in the development of this paper.

RESUME

Ce document propose la création d'une bangue de renseignements d'un nouveau genre, la "bangue de données synthétiques". Il s'agirait de coupler les renseignements de deux banques distinctes pour en créer une troisième. Il en résulterait une utilisation beaucoup plus grande des banques de données existantes dans le cadre des activités de collecte de données nouvelles. On pourrait ainsi réduire considérablement la quantité de données recueillies et, par conséquent, les coûts de collecte et le fardeau de réponse. Ce document recommande diverses considérations pour l'élaboration de techniques statistiques susceptibles de faciliter la création d'un tel concept de couplage de l'information. Certaines techniques pourraient se retrouver dans les ouvrages modernes, alors que d'autres devraient être élaborées.