DISCUSSION

Joseph Steinberg, Survey Design, Inc.

INTRODUCTION

The three papers presented illustrate three of a number of varying objectives of exact matching:

(1) addition of data from second file to host file for the same IRS business tax unit;
(2) construction of a more comprehensive frame by combining files; and
(3) addition of variables on establishment economic data to data for individuals in the Survey of Income and Program Participation (SIPP).

This discussion primarily comments on earlier drafts of these papers.

These papers describe the files used and how the matching was done in fine detail. I leave it to those more expert to comment on these matters; I will not try to comment on that.

PERSPECTIVE OF COMMENTS

The point of view taken in preparing these comments was:

(1) How does the quality (or likely results) of the exact matching conform to statistical standards used to judge a statistical study or to judge completeness of a frame?
(2) After reading or listening to the paper, what is known about factors (and their magnitudes) affecting the nonsampling error component of the results?
(3) What additional information should be made available to judge the nonsampling error?
(4) What more (should) might possibly be tried to reduce the nonsampling errors?
(5) Further, if a sample reinterview program is considered useful in measuring coverage and content (net and gross) differences in a sample survey or census, why not use a sample reinterview program for evaluation and calibration in matching studies?
(6) Is the matching approach optimal or is it better to collect data through a survey process?

In view of the review approach, you will see that this discussion provides some comments and a series of questions for the presenters.

GREENIA

Nick Greenia has an interesting problem, even though both files come from IRS forms. The supplementary forms for individuals (C, F, and 4835), which are of interest, may not show the EIN or, if EIN is shown, it may be incorrect. What is known (if anything) about false nonmatches or false matches as a result (since only the 1979/1980 files of the Forms 941/943 were used, and not 1978/1979)? What is known about the false nonmatch rate which resulted?

It is interesting to observe that many identifier systems have similar problems -- here it is the "sole propietorship/corporation connection" re the EIN. There used to be (and may still be) the problem in the SSN: multiple people gave an identical SSN as a result of the purchase of a wallet that had a valid SSN on a specimen identification card.

I noted that matched cases were dropped when the 941/943 payroll was greater than the sole proprietorship's business deductions. Was any effort made to contact any sole proprietorship when this was found? Wouldn't it be of interest to know, for a small sample, at least, under what circumstances this situation arose? May not treating such cases as unmatched eliminate an important class of novel situations? Why do you think, Nick, that reweighting overcomes the problem?

Given the assertion in the paper "... that a significant portion of true matches remained to be found ..." (Section V), would the analytic objectives be served if the tabulations of "matched" data are based on not much more than the original set of matches? Would the nonsampling error of the results be too large?

I have often wondered whether information on the Forms W-3 was available on any accessible file. Since the Form 941 employment is only for employees for the pay period ending March 12, would a more useful source of employment and payroll be:

(1) the number of statements--counts of Forms W-2 and
(2) total payroll for the year from the summary W-3 process?

Incidentally, if any of these questions suggest a need for contact with a business (as re 941/943 payroll greater than business deductions), a statistical study (perhaps conducted by a third party) should be considered the vehicle, with results available to IRS only in tabulations (screened for disclosure problems). Consider, a statistical reinterview program may be a useful means for evaluating overall quality and not just for special issues.

HIRSCHBERG

Now I turn to Dave Hirschberg's paper. In the paper, I found the interesting points:

(1) that the Master Establishment List (MEL) is unique in its representativeness of small businesses of all size categories, and
(2) that the total number of businesses included in the MEL exceed more than half of the population or universe of all (small and large) businesses reporting to the Internal Revenue Service.

My question is: How complete is MEL? The tables show the relation of the Duns Market Identifiers (or DMI) to County Business Patterns. How do the distributions of MEL compare with some standard? And, by Standard Industrial Classification (SIC code) and employment size?

At another point, the author indicates that businesses not represented in the MEL are mostly smaller businesses or individuals that might be located in their homes or who, due to limited activities, would not appear in the credit markets nor advertise in the yellow pages.

In view of this, what problems are there in the Small Business Administration (SBA) use of MEL? Also, what is known about the rate of inclusions in MEL files of firms no longer in existence (given the slowness of purge of the DMI and Market Data Retrieval, Inc's "yellow-page" listings)? What is the duplication rate still in the file? (One source paper says "... hopefully relatively few.") Further, what is known of the proportion of false matches -- discards from one file or the other that really didn't match? This is not to suggest that "Findit" as a match program has any discernible problems -- at least to my knowledge.

Now, I turn to another matter. This project, creation of MEL, was initiated since there was essentially no single file available to SBA which satisfied its needs--and it is understandable why various agencies have statutes (Census) or regulations which require confidentiality of frames, privacy being deemed more important than government-wide efficiency.

What is the confidentiality status of MEL? Does SBA have a regulation which prohibits disclosure? What are any other possible public uses - could another government agency, say, Department of Energy, or could a research firm doing a study for a government agency have access? At what price? How does this compare to your costs?

On another matter -- what improvements in file completeness would there be from access to the UI files in the 25 states willing to share their files? Has anyone explored the possibility that uniform files for these 25 states may exist in a Federal agency's hands -- the Bureau of Labor Statistics (BLS)? And what cooperation can be worked out between SBA and BLS, given written agreement by these 25 states to permit SBA access?

The paper recognizes that data collection is "non-rigorous" and, therefore, employment, and possibly SIC codes, too, may be inaccurate. What, if anything, can be said about the effects of possible inaccuracies on the use of subsets of MEL as survey frames? Consider the value of a sample reinterview program to check on quality.

Finally, the paper mentions that some checks were planned, e.g., MEL vs. University of Michigan, Survey Research Center's sample of their nonhousehold establishment list. Are there any results of such checks available? What do they show about the completeness of MEL?

SATER

Now concerning Doug Sater's paper; first, I turn to the SIPP information collection to be used for the match. Has Census considered the desirability of expanding the questions being asked (name of employer, address, employer identification number)? Perhaps, in addition to address (or, if not available), they could consider getting nearest street intersection; asking for telephone number at place of employment -- for possible use, when no EIN is given, for calling the employer; or, if no address, calling to establish an address?

Also, re SIPP-collected data -- what steps are taken to assure that SIPP-collected EIN is consistent with SIPP-collected information on employer name and address?

The paper discusses a prospective matching project, and it is interesting to read about the decision process that leads to the decision concerning the source file and matching method. It will be interesting to hear, in the future, what actually took place: the degree of manual effort and the various costs. Incidentally, what is the relative budget planned for this matching activity compared to the SIPP data collection phase? It would be interesting to know, both here and in other matching projects, about relative budgets for matching vs. data collection of source surveys.

In view of the author's contention that they expect to obtain (in the SIPP) valid EINs about 40 percent of the time and that there is a need to use a variety of methods to try to determine the EIN in the remainder, how will the match validity be tested? (The paper says error measurement will be the subject of future development. And evaluation strategies will be the subject of future development.) What about considering a sample reinterview program as part of the evaluation strategy?

The paper describes a small scale familiarization test. Admittedly, it was not a true test, since address and EIN had not been collected in the nonprobability set of units used for the test.

How secure are you, Doug, in the rates of exact matching cited in the paper? Do you have plans for another, truer, test, using a subsample of the SIPP that you plan to use, before mounting the full-scale matching project? Suppose the results are not as good as in the small-scale familiarization test; what if the results suggest a 60-70 percent match rate. Would you recommend the project move forward?

The paper notes that adjustments are planned for matching problems. What order of magnitude of matching problems do you believe are likely to occur, for which allocation or reweighting is the preferred solution? What do you anticipate will be the net effect on the level of nonsampling error in some principal result?