Discussion

by Fritz Scheuren, Urban Institute

Technological changes have historically led to organizational change. The two arguably are inseparably connected (Scheuren 1996). Not so long ago now, central statistical agencies were at the hunter-gather stage in their use of administrative data. It is from this earlier era that the phrase exploiting administrative data comes (e.g., FCSM Report No 6). Around the world, statistical agencies are now moving rather quickly through the agricultural and industrial stages into the electronic (or dot.com) stage in their use of such records (Kennessey 1994). This pattern is occurring in the private sector too. Profitable information uses are being found for operational data, often integrated across what used to be separate silo or stovepipe structures. Some of this effort is captured by the term data mining (e.g., Barry and Linoff 1997).

Administrative agencies have moved in directions similar to large private corporations, since they have the same technology. There are natural limits, however, to a program agencys use of these new opportunities and nearly everyone agrees that if administrative data are to be integrated across agencies, then this should be the role of the central statistical office. The three papers today -- from Norway, Canada and the United States -- are based on that premise. The level of integration described is different but all are on the same path.

Bissett in the Canadian paper captures this trend nicely when he repeats what he calls a mantra. Specifically he says and I fully agree we are moving from an era where we depend on surveys producing aggregate statistics to census data from administrative sources producing detailed estimates. The Prevost- Leggieri paper from the United States says something similar. The Thomsen-Zhang paper from Norway, where we see the process at its most advanced, takes the (to them obvious) points made above for granted, simply citing earlier work (e.g., Thomsen and Holmoy 1998).

The paper by Thomsen and Zhang is discussed first, since, as I have just argued, Norway is the furthest along. Bissett captures nicely the transition now in full swing in Canada. Prevost and Leggieri offer us more of a plan than an accomplished fact. Even so, a lot has been done, as they point out, and as other US Census Bureau papers at this Conference attest.

Thomsen and Zhang Paper

Thomsen and Zhang throw us into the actual mathematics of the statistical use of an administrative register. Their application is to improve the estimation in the Norwegian Labour Force Survey (LFS). Post-stratification to the register is shown to lead to a marked (50%) reduction in variances for level estimates, but virtually no improvement for measures of change.

The first result on the LFS variance for level estimates was bigger but still in line with what I would have predicted in my work with Roger Herriot, now long ago, on the US counterpart of the LFS the Current Population Survey (Scheuren 1980). The lack of improvement in the variance of change was not what I expected. Despite my surprise, however, the result became quite intuitive -- once Thomsen and Zhang explained it.

The authors nicely followup their estimation success with some practical insights for design, notably the suggestion for a direct use of unemployment records. Charlie Jones, when at the US Census Bureau, long advocated such a linkage to the US Current Population Survey. The authors are to be especially commended for their treatment of nonresponse bias issues and the way they have employed register-based estimation to reduce that bias. (For more on this issue, the 1999 Portland Conference on Nonresponse would be another useful source.)

I would have liked the authors to have discussed the use of calibration estimation (e.g., Brewer 1999), mass imputation (Kovar and Whitridge 1995), or better yet a combination. The register data look exceedingly rich and just using post-stratification may not capture this benefit fully, despite the impressive gains found. Mass imputation, invented in Canada over 20 years ago (Colledge et al 1978), may be useful for small subdomains, especially if nonresponse bias is serious. Using the two together is an area that may be worth study too.

To follow, in the US, the path set out by Thomsen and Zhang we would have to embed record linkage directly into the Current Population Survey estimation by using the social security numbers (SSNs) already being collected. That survey then could be matched to unemployment wage (and payment) records collected monthly by the states (which are or could be made available to the Bureau of Labor Statistics). Even the linkage of annual wage records supplied by the US Social Security Administration (from W-2s) could lead to important variance reductions.

The Current Population Survey suggestions are not small steps and may remain, as they were when I first proposed variations on them, a bridge too far. Even if they were feasible right now there will be a lag in data availability that would reduce the gains over those Thomsen and Zhang found. Because of its different focus, the place to try these research ideas might well be the US American Community Survey and not the US Current Population Survey, at least not now.

Bissett Paper

The Bissett paper makes quite clear how actionable the administrative mantra has been made in Canada While my Canadian relatives might not fully appreciate it, the still recent goods and services tax (GST) -- and now in some provinces the Harmonized Sales Tax (HST) -- was a major opportunity to improve economic information. The level of integration achieved across survey and administrative records has been astonishing in scope and speed.

Despite some risks, mentioned in the paper, Statistics Canada went ahead in cooperation with Revenue Canada Taxation to put these new sources to immediate use. The estimation implications of the new approach are enormous, especially for small areas/small industries. There may be a modest shakeout period, though, as the author implies. Nonetheless, the result is clearly a win-win with both more information and lower respondent burden.

The Bissett paper covers a wide range of definitional and policy topics. At least four may deserve comment because they have elements that connect to United States practices. These are:

1. Enterprise Reporting. The paper discusses the fact that in the future enterprise rather than establishment reporting will be dominant in the ongoing Canadian statistical system. This is primarily, of course, because of the administrative records being used. The shift may have been inevitable anyway for some enterprises, given the changing demography of business -- notably the emergence of dot.com operations which have virtual rather than physical locations. Allocations or imputations will be needed, as the author indicates, when an enterprise has multiple lines of business. While this may be only a small weakness in a Canadian context, it is unclear how well such allocations would work for US businesses.

2. Business Number. Originally the Canadians did not have a common business number, as we do in the US, for reporting to all taxing authorities. This has now changed and its introduction rectifies an historical accident that the pension system and tax system were not fully integrated at the business level. In the US, the national tax authority, the US Internal Revenue Service (IRS), simply took over the US Social Security Administration (SSA) Employer Identification Numbers (EINs) in 1958 and made them general business identifiers.

3. Industry Coding. Statistics Canada codes all business enterprises as to their industrial activity, with the results shared in full with Revenue Canada Taxation. This is enormously better than the multiple administrative and statistical codes now put on businesses in the United States (statistically by the US Census Bureau and the US Bureau of Labor Statistics -- each separately; administratively by state unemployment offices, SSA and IRS -- again mainly separately). Incidentally, the proposed US data sharing legislation, now working its way through Congress, would not completely solve this problem, calling as it does only for sharing of codes in one direction (from administrative agencies to statistical ones).

4. Confidentiality and Data Access. While the Canadian Statistics Act makes for a much better structure for sensible partnering relationships, there are still gaps between the Canadian federal and provincial levels of government, for example. When and how these get resolved it can be safely predicted that, even in Canada, stewardship roles have to continue changing. As time goes on, all of us (in both countries and elsewhere) will, hopefully not from hard experience, learn a great deal more about privacy, physical security, and our ability to keep the confidentiality promises we make. In this connection an active research program seems needed. Included could

be efforts (1) to simulate breaking into what are thought to be secure files and facilities and (2) to keep measuring the views of all stakeholders, especially data respondents (Mulrow and Scheuren 1999).

To summarize, the Bissett paper introduces us to a world that in many ways is similar to our own here in the United States. The differences are fascinating, though, and could be helpful to us, since clearly for business surveys Canada is well along on its transition to a new approach that relies mainly on administrative data, using survey information to help interpret what the administrative data says i.e., as a Rosetta Stone.

Prevost and Leggieri Paper

Let me start off my discussion of the Prevost-Leggieri paper on an administrative record census by confessing a strong self-interest that makes me hard to be objective. After all, the authors are attempting to carry out ideas that Wendy Alvey and I sketched quite some time ago (Alvey and Scheuren 1982). In fact, I have just updated my views in a paper to appear in the December issue of *Survey Methodology* (Scheuren 1999). For both these reasons I will be very brief here.

The Prevost-Leggieri paper expresses the administrative record mantra just as was done in the Bissett paper, albeit not quite yet with the same level of experience and consequentially self-confidence. Much of their paper consists of listing the extensive administrative sources that could inform a decennial administrative census. Certainly the research plans that exist to develop these records for a census application are quite impressive. The work by Bye (1997-99) gives me many reasons to believe that the considerable challenges to be faced in a partial US administrative record census can be overcome. However, for many of the record systems the Census Bureau is still at the data acquisition phase, learning what is going to be possible.

Even so, the authors describe several places where the Census Bureau has already begun to directly confront key issues, some of which are unique to a US context. Social contract issues, for example, are not the same as in Canada or in most European countries. We are quite a bit more distrustful of government in the United States; hence the Census Bureau has had to proceed more slowly in doing what would be straightforward elsewhere or, indeed, has in some Scandinavian countries already has been done (e.g., Redfern 1989; Blum 1999).

Some technologies being developed for conventional census and survey purposes, like GIS software, will be portable to this new application. The worldwide methodological developments in small area estimation will help too. Record linkage techniques, pioneered partly at the Census Bureau by Winkler and Jaro, should also be directly applicable. In any case, there are many reasons to be optimistic (The 1997 International Conference on Record Linkage, co-sponsored by the FCSM has more on this).

Concluding Comments

These three papers have been fun to read and comment on. I hope you enjoyed them too. I expect them to offer value for quite some time, even though this area of statistical practice is entering an era of explosive change. When we look back in five or ten years, naturally there will be places where we learned we were too optimistic. Similarly there will be opportunities, obvious after the fact, that we do not see now, even though they are right in front of us. My thanks again to the authors today for their efforts at showing us so much of what is coming.

References

Alvey, W. and Scheuren, F. (1982). Background for an Administrative Records Census.

Statistics of Income and Related Administrative Record Research. Washington, DC: U.S. Department of the Treasury, Internal Revenue Service. See also the discussion of this paper by John Leyes of Statistics Canada, appearing in the same volume.

Barry, M. and Linoff, G. (1997) ***Data Mining Technique***, John Wiley and sons, inc. New York.

Blum, O. (1999) Combining Register-based and Traditional census Processes as a Pre-defined Strategy in Census

Planning, ***FCSM Conference Proceedings***.

Brewer, K. R. W.(1999) Cosmetic Calibration With Unequal Probability Sampling, ***Survey Methodology***, 25, 2, 205-212.

Bye, B. (1997). ***Administrative Record Census for 2010: Design Proposal***. Prepared for the U.S. Bureau of the Census, Rockville Md., Westat Inc.

Bye, B. (1998a). ***Race and Ethnicity Modeling with SSA Numident Data: File Development and Tabulations***, Prepared for the U.S. Bureau of the Census, Rockville MD., Westat Inc.

Bye, B. (1998b). ***Race and Ethnicity Modeling with SSA Numident Data: Individual-level Regression Model Version 2,*** Prepared for the U.S. Bureau of the Census, Rockville MD., Westat Inc.

Bye, B. (1999).***Race and Ethnicity Modeling with SSA Numident Data: Two-level Regression Model,*** Prepared for the U.S. Bureau of the Census, Rockville MD., Westat Inc.

Colledge, M., Johnson, R., Pare, R., and Sande, I.. (1978) Large-Scale Imputation of Survey Data, ***Survey Methodology***.

Gates G. and Bolton, D. (1998) Privacy Research involving Expanded Statistical Uses of administrative Records, ***Proceedings Government and Social Statistics Section, American Statistical Association.*** See also, Gates, G.. (1999). Data Mining, Panel on Privacy and Statistics in the New Millennium, panel presentation at the Joint Statistical Meetings, Baltimore, MD.

Kenessey, Z. (1994) -- ed. ***The Future of Statistics: An International Perspective***. Voorburg: International Statistical Institute.

Kovar, J. G. and Whitridge, P. J.(1995) Imputation of Business Survey Data, ***Business Survey Methods,***403-424, John Wiley & Sons, Inc. New York.

Mulrow, J. and Scheuren, F. (1999) The Confidentiality Beasties, ***Proceedings of the section on Survey Research Methods, American Statistical Association.***

Redfern, P. (1989). Population registers: Some Administrative and Statistical Pros and Cons, ***Journal of the Royal Statistical Society***, Series A, 153, 1-41..

Scheuren, F. ed. (1980) ***Studies from Interagency Data Linkages***, Social Security Administration. See especially Report No. 10.

Scheuren, F. (1996) Review of Organizing to Count, ***American Statistician***.

Scheuren, F. (1999) Administrative Records and Census Taking, ***Survey Methodology***.

Thomsen, I. And Holmoy, A. (1998) Combining Data from Surveys and Administrative Record systems. The Norwegian Experience, ***International Statistical Review***, 66, 2, 201-221.