Component of Statistics Canada
Catalogue no. 12-001-X Business Survey Methods Division

Article

Administrative records and census taking

by Fritz Scheuren

December 1999

Statistics Canada Statistique Canada

Canada

Survey Methodology, December 1999
Vol. 25, No. 2, pp. 151-160
Statistics Canada, Catalogue No. 12-001

Administrative records and census taking

Fritz Scheuren [1]

Abstract

The shift in the use of administrative records from an incidental role in census applications to an essential one is now well along in many European countries. The challenges are greater in Canada and the U.S., as this paper discusses. Progress in the U.S. in developing a modified administrative census paradigm is dealt with in some detail and contrasts made to what has already been done elsewhere, notably in Canada. A research agenda is set out and some connections made across a whole gambit of U.S. census-connected statistical programs – including current surveys, intercensal estimates, and the measurement of the census undercount. Privacy concerns are prominent among the issues that are addressed. The role of low cost computing and advanced record linkage software also are given their due. The changing status of central statistical agencies as the information age advances is also touched on.

Key Words: Record linkage; Privacy; Surveys and intercensal estimates.

1. Introduction

The use of tax and other governmental administrative records in census-taking turns out to be quite old, even though most of the real advances have occurred in the last 25 years or so – that is, roughly during the same time span as the publication of *Survey Methodology*.

The introduction of modern sampling – or the representative method, as Kiaer called it (Bellhouse 1988) – was tied, it seems, to the matching of samples of Norwegian tax records to the census of 1890 in Norway (Johnson and Kotz 1997). Of course, the mathematics of Kiaer's approach, rather than the particular application, was the focus of much of the later work (and remains the usual focus in *Survey Methodology*). This was appropriate, given that administrative records were often inaccessible and hard to use.

Over time, though, there has been legislation (like the Canadian Statistics Act) that has made access to administrative records routinely possible by Central Statistical Offices in many countries. Advances in computing and recordkeeping in government and elsewhere, while bringing new problems, have certainly also made administrative records easier to use for statistical purposes and this trend seems likely to continue or even accelerate (*e.g.*, Kenessey 1994).

Traditional censuses have been replaced in whole or in part in some of the Nordic countries by administrative records (*e.g.*, Myrskla 1991; Thomsen and Holmoy 1998). This has not occurred in Canada or United States – partly because of the nature of the administrative records available and partly because of the sheer size and complexity of the undertaking. In fact, even in one of the most ambitious North American administrative record census (ARC) proposals (*e.g.*, Alvey and Scheuren 1982), the complete elimination of conventional census-taking was not advocated. Rather a mixed mode approach was suggested.

The current paper attempts to recount briefly the "state of the play" on administrative record census proposals (See Steffey and Bradburn 1994 for an additional perspective). The paper focuses mainly on the U.S. but with Canadian parallels (Leyes and Elsl-Culkin 1994). In some respects this is a follow-on or update to a piece in *Survey Methodology* ten years ago (Scheuren 1990).

Before going into details, it may be worth providing a context to the changes needed to achieve some form of an administrative record census. First, it may make sense to discuss the nature of scientific revolutions generally. Bellhouse (1988) cites Kuhn (1970) in this regard relative to sampling itself. Scheuren (1990) also drew independently on Kuhn concerning ARC ideas.

For example, it was not really until the paper by Neyman (1934) – aided by Sukhatme (1935) and, through Deming's advocacy, plus the all important paper by Hansen and Hurwitz (1943) – that Kiaer's randomization-based approach to the representative method might be argued to have been accepted.

At least in North America, ARC ideas may not yet have found their Neyman. Still, there have been many hard-won successes to celebrate and with the fuller emergence of enabling technologies (like record linkage and low-cost computing), the shape of the future can be characterized as encouraging.

Organizationally, the present paper is divided up into 8 sections, beginning with this introduction (section 1). Section 2 sets out some ARC background and section 3 develops a few assumptions about issues that go beyond the operational feasibility of an ARC. The rest of the paper consists of suggestions in four areas: the 2000 U. S. Census (section 4), the intercensal population estimates program for the coming decade (section 5), the current surveys program (section 6), and the planning for the 2010 U.S. Census (section 7). There is also a concluding section (section 8)

1. Fritz Scheuren, The Urban Institute. Mail: 1402 Ruffner Road, Alexandria, VA, 22302, U.S.A. E-mail: scheuren@aol.com.

that discusses priorities. Finally, some references are provided.

2. Background on original ARC proposal

In Europe, several countries are quite far along in developing administrative record censuses, having begun back in the 1970's (*e.g.*, Jensen 1983; Jabine and Scheuren 1987; Redfern 1989; Blum 1999). Those countries are much smaller and differ in many other ways from the U.S. or Canada – especially in the social contract that underlies census-taking. What they have done, therefore, is hard to apply directly. Still, their pathbreaking efforts have much to teach.

The original idea for a partial ARC in the United States was first made publicly at an American Statistical Association meeting in 1982 (Alvey and Scheuren 1982). The work of John Leyes and Doug Norris at Statistics Canada was one of the inspirations for that proposal. Basically, the paper advocated research on how –

> To link U.S. Internal Revenue Service (IRS) tax return data to wage and retirement earnings, unemployment compensation records, and U.S. Health and Human Services (HHS) administrative files to obtain a "bare bones" population census.

The key element here was researching a partial replacement for a conventional census – not to completely replace it. The administrative records were, moreover, limited to those already legislatively available in whole or in part to the U.S. Census Bureau. Speculations were offered that some administrative system changes might be possible to accommodate an ARC use; even so, this proposal never contemplated "content-wise" that the resulting ARC would be much more than a bare bones population count.

The anticipated coverage of an ARC was believed, however, to be good but not treated as perfect. In fact, the ARC proposal always assumed some form of sampling to adjust the population for completeness. The prediction was made that the proposed ARC would cover well over 95% of the population covered in a conventional census. The 1993 and 1998 papers by Sailer and his colleagues confirmed this conjecture (Sailer, Weber, and Yau 1993; Czajka, Moreno, and Schirm 1997; Sailer and Weber 1998).

The bare bones aspect might be best illustrated by the fact that no provision was made for the housing census that is conducted along with the current U.S. population census. Housing would have to be dealt with in some other way. Among the weaknesses of the proposal, acknowledged at the time, was the quality of the race data in administrative records and the problem of having mailing rather than actual residential addresses.

Considering these limitations, why proceed? Well, the ARC originally proposed not only would reduce the cost and burden of a decennial census, but has the potential for producing a total population count more frequently than every 10 years. It also might provide improved coverage for some of the populations traditionally undercounted in a decennial census. Moreover, Bye's 1997 work (Bye 1997), plus his recent detailed look at Social Security Administration data on race and ethnicity (Bye 1998a and 1998b), put these weaknesses into perspective and go a long way to suggesting how they could be overcome or at least lived with (See also Bye 1999.)

The most important point about the proposal was that it advocated research towards a potential ARC 10 or even 20 years down the road. Implementing an ARC was not proposed, although some of the reaction raised this concern. Privacy and confidentiality aspects were prominently mentioned in the proposal as also requiring research.

There is no need here to carry the story forward in detail from the original Alvey and Scheuren paper until now. That has been done elsewhere (Scheuren 1995a). What is important to mention is the shift in the tone of the research over the years, from "proving" an ARC could not work to trying to find ways that it might. Bye, for example, in his excellent report to the Census Bureau (Bye 1997), fully spelled out a way to implement such a census. While it has many researchable elements, Bye's approach demonstrates that the idea is operationally feasible.

3. Assumptions

Certainly the technology of record linkage and the widespread availability of massive fully-computerized record systems make the creation of alternatives to conventional censuses possible outside the U.S. Federal sector. State governments have incentives to be sure that every resident is counted (Biskupick 1998) and certainly could construct partial ARCs using their own record systems. The motivation to challenge the Census Bureau monopoly is definitely present with the devolution of Federal activities to the states and the financial incentives involved in Federal grant programs. Nearly $200 billion in Federal aid is distributed annually based on population.

3.1 Massive data sets

The mass marketers and telephone survey organizations also have extensive data systems that might be tapped into. Private data sources unheard of a few years ago (*e.g.*, even from grocery chains!) are expanding rapidly and extensive statistical use of these private sources is already occurring (National Academy of Sciences 1996). With the worldwide revolution in electronic recordkeeping practices, there will be many new entrants in the emerging information industries. The "hurdle" price has been lowered and the value of information has been growing.

Some recent work done for the State of Connecticut might be worth illustrating the general points just made. In White, Mulrow and Scheuren (1999), the authors describe an effort commissioned by the State of Connecticut to use state administrative records to improve Connecticut's jury selection system. It is important to note at the outset that the goal of that work was not to do an ARC. Still the exercise has a lesson in it about the ease with which a partial ARC could be developed for a state.

Formerly, Connecticut employed voter registration and motor vehicle files with a labor-intensive process to unduplicate the two systems, so as to form a list from which to draw potential jurors for duty. The new effort, described in White *et al.* (1999), involved employing probability-based linkage technologies (Jaro 1989) with four state-level files: the two mentioned already, plus the State's income tax file and the State's unemployment file. The files were all created in early 1998.

To evaluate the Connecticut linked data, a comparison was made to 1996 Census Bureau population projections by township, brought forward to 1997. The administrative record population coverage obtained by the combined file was surprisingly good, given that an ARC was not the goal. In fact, the linked administrative record counts by township were highly correlated with population projections. The simple correlation was $\rho = 0.946$. When four of the 169 townships are removed as outliers, the correlation went up to $\rho = 0.977$.

3.2 Privacy considerations

Of course, privacy assumptions bear on direct use of administrative lists and on linkages across them. Obviously, ARC considerations about personal privacy will impact linkages of data from different sources for census purposes. In 1985, early results of the privacy research on linkage issues were presented (Scheuren 1985), followed by a great deal of other work, notably by the Census Bureau – reported on, for example, by Gates and Bolton (1998), Gates (1999) and Singer (1999).

It looks reasonable, despite concerns, such as those in Scheuren (1997), that a careful introduction of greater and greater linkage will succeed in gaining wide acceptance as a policy. In fact, Statistics Canada is already experimenting with this now through their Survey of Financial Security, where respondents are given an opportunity to authorize access to tax and pension records instead of responding to selected survey questions (Statistics Canada 1999). In that survey, they are finding very high acceptance of the idea. This reference is just an example of the success that has already been achieved in direct uses of administrative records in Canadian surveys. For example, the option of accessing tax records has been standard in the Canadian Survey of Labour and Income (SLID) since May 1995. (See Statistics Canada 1993-1996.)

While, in the United States, perhaps a sixth of the population will object, their views may not be listened to. Despite this, it appears likely that there will be no outcry and the "taking" of these privacy rights will proceed with little incident. Fellegi (1997), in his opening address at the 1997 International Record Linkage Conference, gave a sound analysis of this possibility.

3.3 Access considerations

For the Census Bureau to do an ARC would require new legislation to mandate cooperation by various government agencies with the Bureau. Currently, except for the IRS, the Bureau may receive administrative data if other agencies choose to provide them; but, unlike the Statistics Act in Canada, there are no laws that require agencies to cooperate. Continuing this arrangement, of course, would be untenable if the Census Bureau were to try an ARC.

The development and enactment of such legislation would provide the opportunity for a public debate on ARC ideas, something that must occur before an ARC could be done. In any case, legislation is required that would mandate cooperation with ARC research; otherwise, the Bureau may never get to do the required preparatory work. This suggests legislation "now," if the Census Bureau is to prepare for an ARC in 2010.

3.4 Technological advances

The assumption is that there will also be continuing advances in record linkage techniques, led by Bill Winkler at the U.S. Census Bureau and Martha Fair, among others, at Statistics Canada. The data mining "craze" can be anticipated to lead to a very wide dissemination of these techniques. Large privately-held data sets will be increasingly combined and in an increasingly statistically satisfactory way. Tied to this growth will be a realization, as in Scheuren and Winkler (*e.g.*, 1993 and 1997), that the goal of linkage is not mainly the matched data, but a way to combine disparate sources to produce information otherwise unattainable because of cost.

There will continue to be an expansion of access to and uses of improving Geographic Information System (GIS) software, especially in small area estimation applications, both within and outside of Central Statistical Offices. We are entering a new "data-dense" world, where the amount of information available geographically is exploding. Much of this will be estimated, but the overall quality will be superb. Increasingly, isolated estimates (as in Schaible 1996) will be replaced by sets of interlocked covariates that are coherent together. In all likelihood, market forces will drive this. The impact of cheaper and more powerful computing will mean that the handling of very large files and burdensome computations will not be seen as barriers, even in government – albeit there will be a lag in the public sector.

If these scenarios happen, the world of high cost data gathering (like a conventional census) will increasingly be replaced by a world of frugal reuse of data – often automatically obtained (*e.g.*, as predicted in International

Statistical Institut (1994). Widespread reuse applications will spur even better techniques and, combined with competition and cheap computing, will reduce greatly the power of data producers, including Central Statistical Offices.

4. Census 2000 suggestions

To develop an administrative record census, much research is clearly needed. This section sets out suggestions for administrative record research to be done as part of the 2000 Census in the United States. These are grouped into process (section 4.1) and content (section 4.2) suggestions.

4.1 Process observations

The 2000 Census "kicks off" a decade of potential activity in getting ready for Census 2010. The observations made here on these possible activities fall under four headings: acquiring more administrative data, strengthening the safeguards on use, building cooperative arrangements for staff exchanges, and establishing a precedent of modifying existing administrative systems to enhance their information uses.

4.1.1 Data acquisition

There certainly is a history of greater cooperation at census time by other government agencies. While there were many complications, it is no coincidence that the first IRS Individual Master File extract that the U. S. Census Bureau received was obtained for income year 1969 of returns filed in the decennial census year 1970. The occasion of the 2000 Census should be used (and has), therefore, to advance the 2010 agenda, by acquiring data and exploring how to use them to develop an ARC.

The Census Bureau's recent precedent in obtaining the full Social Security Number (SSN) application or Numident file from the Social Security Administration (SSA) is a particularly important example of the kind of acquisition needed, since the file contains age and other demographic data items on all persons who have SSNs. As Prevost and Leggieri (1999) discuss, there are many efforts underway which have led to the Census Bureau obtaining still more Federal record systems.

Obtaining pilot access to state program records for the medically indigent (Medicaid) should be a priority; this is so despite the quality issues that such systems have. Make no mistake, however; a wholesale acquisition policy could be perceptually dangerous (*i.e.*, violating the privacy assumption mentioned in section 3). Only systems for which there are clear, sustainable research objectives (and financial support) should be sought.

It is important to point out that a census requires not only a full count of the population but must include correct geographic location at a point in time. For apportionment, state-level geography is required; for redistricting, geographic location well below the state-level is required. The implication of this for data acquisition is twofold. First, the administrative record files must attempt collectively to cover the "entire" population. Second, the files must provide good information on low-level geographic location at chosen points in time. Sometimes, even, files should be obtained just because they provide better geographic location for some part of the population (see Bye 1997 for more details.).

IRS acquisitions might be of two types: small incremental additions, as well as acquisitions of full-scale tax files already being received by the Census Bureau. The small additions are technical and procedural, involving working level staffs; the larger acquisitions have policy elements and need a different approach – with involvement at the highest level.

Regularly since 1969, the Census Bureau has obtained an extract from the IRS Individual Master File system. Late returns, not filed in time for that extract, are becoming increasingly important and might be added to the data from IRS. Second, the prior year returns should also be obtained and introduced into the longitudinal samples recommended in sections 5 and 6 below. Marginally increasing item content to include more types of income is also suggested. Obtaining all or a large sample of information master file documents electronically is recommended. Getting all wage and social security information records, as has been done, is an exceedingly good start and certainly seems a plausible compromise for 2000, but interest and dividend records are important too.

In any case, a major effort should be made to provide budget support in non-census years for sustaining this system – a problem that the administrative records program at the Census Bureau has had historically. It can be argued, until recently in fact, that the Bureau already has had more administrative record data than it had resources and people to use fully.

4.1.2 Physical and perceptual security

Clearly enhanced physical security of administrative data goes hand in hand with more data acquisitions. The Census Bureau recently established a secure restricted access environment for its demographic administrative records (Clark and Gates 1999). However, the Census Bureau must not stop there. An outside auditing firm should be hired to test the new physical security. In fact, such efforts should be an ongoing part of the Census Bureau's new data steward role for administrative records. Assuring protection of the data is critical to the success of an ARC. It is important to recognize that linked administrative record databases are inherently more valuable than individual agency files; employees are subject to more temptation or at least the suspicion of being vulnerable. In fact, violations by IRS employees which came to light several years ago (see Scheuren 1995b), led to anti-browsing legislation specific to tax data. The Census Bureau must take every precaution to enforce such rules for all of its administrative

records. There is also a need to keep up public opinion survey research and conduct more focus groups with the various stakeholders and the general public, as well as with the Bureau's own employees. The cost of maintaining massive administrative record systems involves both physical and perceptual maintenance of data security. And neither of these comes with a small price tag.

4.1.3 Building cooperative arrangements for staff exchanges

Human capital improvements are also key to any administrative record initiative. Professional statisticians outside the administrative agency too often think of just the data products they obtain rather than the system as a whole. Some would argue that the unfortunate phrase, "exploiting administrative data," grows out of this narrow (and denigrating) view. Whether the phrase is unfortunate or not, it reflects the hunter-gather phase in the use of administrative records for statistical purposes. That age is ending.

The real (or new) goal should be to turn "administrative systems into information systems" (Scheuren and Petska 1993). This means we need to move, continuing the analogy, to the next or agricultural stage in the use of administrative records.

One way this new phase might be speeded up would be through something like an American Statistical Association fellows program. A sabbatical might be paid for by the Census Bureau and offered to operating administrative agency staff – perhaps from around the world. This could involve having IRS, SSA, and other administrative record stewards in residence at the Census Bureau for short periods. Among the goals would be to give them an understanding of the importance of the information services that their administrative systems made possible. A by-product would be the invaluable insights the administrative agency staff could provide regarding assumptions about and use of their data for statistical purposes.

More important still could be the reverse exchange – Census Bureau staffers going to work at the operating agency for an extended period of time. Unlike in Canada, which has a great deal of professional migration into and out of Statistics Canada to administrative agencies, the U.S. has very little. Anyway, more is needed. Think of the stimulus that this could give the statistical imaginations of the individuals sent. Deming talked about the need for systems thinking (Deming 1986). How better for people to obtain such thinking in connection with administrative systems than by such an experience, repeated periodically every few years.

4.1.4 Establishing a precedent for modifying existing administrative systems to enhance information uses

Improving the statistical data products derived from administrative systems can be achieved in many ways. One is to add an item (and the associated burden) to an existing administrative system. Naturally, this is a two-edged sword. Obtaining residential addresses on tax returns, for example, as was done in 1981, would be an obvious example; however, see Bye (1997), where another – and perhaps better – approach is advocated that would involve a direct followup for addresses that are clearly not residential.

Another potential addition to the tax return might be a conventional (or landline) residential telephone number. While the growing use of cellular phones may make such numbers of only temporary value, they still might be worth obtaining. In the U.S. at least the shift to cellular has not been accompanied by the abandonment, yet, of earlier technologies. In any case, it can be predicted that the administrative uses of these numbers could more than pay for their value as a statistical tool in record linkage during the census and later on in an ongoing survey program. Moreover, for listed numbers, there would be a valuable check on the address.

While probably very hard to accomplish, changing third party wage reports (IRS Forms W-2s; T-4s are the Canadian counterpart) so that they have the date of the last pay period on them, would be an enormously valuable addition from an ARC perspective. The addition of the date of the last pay period covered could remove much of the ambiguity associated with multiple addresses on such documents. Of course, accessing the quarterly unemployment system wage records, through the U.S. Bureau of Labor Statistics, might be even better and would not increase existing burden.

These suggestions, while perhaps feasible, will require a great deal of work to implement, since there are many other stakeholders and costs to consider. One observation, which Bye included in his 1997 report on a possible ARC, is to obtain from the U.S. Social Security Administration the mailing address files that are used by them to send SSNs back to the parents of newborns. Here the burden is slight and the value sizable, since it would give access to current addresses for families with new borne children.

Clearly, some tradeoffs are easier to make than others. It is essential, whenever possible, is to find ways that better join an information purpose to an existing administrative one – thereby obtaining something of value for everyone.

4.2 Research suggestions

There are many worthy research ideas that could be recommended. Two important ones are (1) obtaining SSNs on the post-censal quality check samples to be drawn, so that a triple-systems estimate can be obtained of the undercount; and (2) producing a limited ARC estimate during 2000 for cross-checking with the official "counts." For these to be fully effective the results from both are needed on the same schedule as the official Census Bureau counts and undercount adjusted estimates – due in December 2000 and March 2001 respectively.

4.2.1 Triple systems estimation

It has long been advocated (*e.g.*, Scheuren 1995a) that a triple-systems estimate be attempted (Zaslavsky and Wolfgang 1993). The three systems would be the quality check sample, the census itself, and an amalgam of unduplicated administrative records. For triple systems estimation to succeed, all the matching needs to be of high quality. Without SSNs obtained in the quality check post-enumeration survey, the matching to administrative records will be a lot harder and, for doubtful cases, perhaps fatally ambiguous. People with multiple addresses and common names would be particularly challenging in the absence of SSNs.

4.2.2 Concurrent partial state level ARC

Even without attempting a triple systems approach, a concurrent limited ARC has potential in any post-census review. The need to have an immediate check on the statewide counts could be accomplished using the methods employed twice now by Sailer, Webber and Yau 1993 and Sailer and Weber 1998 and could be done quickly, if given a high enough priority. The needed IRS administrative records are expected to be essentially in place by the early fall of 2000 and could be processed by the Census Bureau on a flow basis.

Specifically, it is recommended that the Census Bureau receive its normal IRS Individual Master File extracts monthly, so matching can begin early. Information documents on wages earners and social security recipients could also be received on a flow basis from the Social Security Administration (even before being compiled at IRS). Many of the decennial census misses that are in the IRS data bases could well be earned income tax credit (EITC) recipients who may move. Continuous matching and sample checking will be key for finding such individuals. Certainly those EITC filers who use refund anticipation loans will need extra attention, if followup is going to be successful.

There are, of course, many other worthy 2000 Census research ideas that might lay the groundwork for an eventual U.S. Administrative Record Census (see Prevost and Leggieri 1999). The two mentioned above seem, however, far and away the most important. For a recent paper on the use of administrative records in the Canadian Census, see Carter and McClean 1996.

5. Intercensal implications

Paradoxical as it may sound, to make revolutionary advances in the use of administrative records an evolutionary approach is needed – especially in the intercensal estimates program.

5.1 Annual administrative record portion of ARC

First of all, the Census Bureau should continue annually, on at least a sample basis, the ARC estimation of state totals mentioned in section 4.2. Eventually, depending on data acquisitions and funding, these could be enlarged and deeper geography obtained.

5.2 Large longitudinal administrative sample

Second, large longitudinal administrative record samples should be mounted. Following the Canadian example, the Census Bureau could begin with a straightforward longitudinal sample of tax return records, matched to the U.S. Social Security Administration's Numident file, containing demographic information for all those with SSNs. Statistics Canada has long had a 10% longitudinal sample of T1 returns (Leyes and Elsl-Culkin 1994), which in the U.S. would translate into a 1% sample, given relative country sizes. In fact, tying this longitudinal sample to the U.S. Social Security Administration's 1% Continuous Work History Sample (CWHS) could give it a very long (time) footprint, indeed.

Eventually, this longitudinal sample might be extended across other Federal administrative systems (at IRS and SSA, but perhaps elsewhere too). A caution, though. The chore of matching changing administrative units over time may require more resources and patience than might be anticipated and so should proceed incrementally with smaller efforts, say the 0.1% CWHS for example – as has already been partially implemented (Czajka and Walker 1989).

5.3 Integrated administrative statistical sample

Scheuren (1979) has a much more ambitious 20-year old proposal for a set of interlocking administrative samples. Perhaps this should be re-examined and updated. His ideas involved both standalone efforts and efforts potentially supportive directly of traditional intercensal and current survey programs. Unlike the basic ARC concept, they would have expanded item content as their main objective, rather than complete or near-complete population coverage. They could also be used as starting points for various current survey efforts, as is the case now in the dual frame Survey of Consumer Finances (SCF) mounted by the Federal Reserve Board (Kennickell and Woodburn 1997).

5.4 Transaction-based system

Fourth, for the long term, the current intercensal administrative records program should move, to the extent it can, from annual data systems with year by year matches towards direct transaction-based adjustments of the administrative counts. Consider, for example, an effort to follow a sample of SSNs over the decade. This clearly would be a move towards a partial population register. Despite possible public concerns about massive databases, having a statistical population register as a goal might be a good way to rationalize and prioritize intercensal activities. The goal of a household address register, updated transactionally, seems evident already in the work that the Census Bureau is undertaking with its improving Master Address File (MAF) system.

These ideas for decennial uses of administrative records can be intertwined with suggestions regarding the Census Bureau's current survey program, as we will discuss below. In any case, much greater coordination (and positive synergy) between these two separate efforts is needed than has been true traditionally. See Alexander and Chand (1999) for the kind of effort that could really pay off.

6. Current survey implications

The introduction of administrative records into the design and estimation of the Census Bureau's continuing surveys seems a natural step towards an ARC. Some examples of how this can be done include:

6.1 Sampling frame uses

The American Community Survey (ACS) might be a natural starting point for an effort to use administrative records in a multiple frame context. ACS' use of the Master Address File could be supplemented, for example, with tax return addresses and, potentially, Social Security recipient addresses – and for more than just updating addresses.

6.2 Matching poststratified samples

Some time ago Scheuren (1980) advocated that the CPS might be routinely matched to administrative records and that administrative controls be used as poststratifiers. The pilot for this was the 1973 Exact Match Study. The approach would be much more workable today. Work, like that of Thomsen and Zhang (1999), might form an up-to-date prototype. A related approach is found in Kennickell and Woodburn (1997).

6.3 Linking current survey program to intercensal goals

Whether you start from an administrative frame or match back to an administrative list, each effort will provide information on coverage weaknesses in the administrative records that will make it possible for them to be better used in the intercensal period. Also, such joint operations will point out where to concentrate coverage research for 2010.

An ongoing program embedded in the current survey effort to enhance already excellent demographic methods is essential (and seems to be under consideration by Prevost and Leggieri 1999). Resistance to adjustment can be worn down by repeated and open experiments over the decade, accompanied by continuous coverage and content improvements. A goal should be set to develop an annual fully projected ARC beginning no later than 2005. Funding for a large enough sample to supplement administrative records ought to be sought, perhaps through the American Community Survey. Frankly, though, this budget strategy may require that the Census Bureau promise to make major savings in 2010 – a risky proposition but necessary psychologically and fiscally.

7. Additional 2010 research implications

Specific suggestions for additional 2010 research are hard to make, since much will depend on how successful the Census Bureau is in making their other administrative record uses serve multiple purposes. Nonetheless, two observations may be worth highlighting in any case, since they are not mentioned above.

7.1 Tracing sample from 2000 census

A large tracing sample should be followed over the decade. The starting point might be the post-censal quality check sample, after matching it to administrative records and augmenting it with cases, to the extent feasible, found only in the census or only in an administrative record. The kind of administrative steps outlined above would be followed, plus the actual use of tracing methodologies in at least a subsample. Fieldwork would be necessary to sort out all the problems in "cross-footing" satisfactorily from one census to another. Most of the work would be done by matching in successive waves of administrative records (in a manner similar to that touched on in section 5.2). Again, a big issue would be privacy concerns (as set out already in subsection 3.2).

7.2 Special censuses

To prepare for 2010, there will be a need to conduct special censuses that begin with administrative records and attempt to complete them using sampling. In structure, these would not be very different from the pretests done before every census. However, because the ARC paradigm is new, there would need to be more tests and, especially, more testing time. The first 2010 tests should be built into the 2000 Census and should continue uninterrupted through the decade. Early on, general feasibility issues need to be addressed. For example –

7.2.1 Developing a way to efficiently use an administrative amalgam of addresses and individual names as a frame, so that addresses not on the administrative list are oversampled.

7.2.2 Developing a way to efficiently handle multi-unit dwellings, since the administrative addresses usually do not have apartment numbers. (It may be that some of the sampling will have to be independent of the lists, as in the census quality check sampling, then matched-in after the fact.)

7.2.3 Developing an approach for dealing with problem populations (*e.g.*, low-income minority children) will need special attention (Medicaid data, mentioned earlier, might be of help here but this is unclear).

7.2.4 Developing a means to deal with problem locations in the 2000 Census (*e.g.*, inner city neighborhoods) may need to be looked at individually.

The notion of designing these special censuses as a rolling sample (*e.g.*, Kish 1990) would allow – say, by the end of 2005 – a way to obtain a "gold standard" for evaluating the ARC approach that evolves. Note, there is no reason that two or more methods cannot be tried simultaneously to speed up the process of testing. Indeed, it may turn out that the 2010 Census should employ multiple approaches simultaneously – including sampling. Given the diversity of circumstances that exist, multiple approaches may prove inherently better than any single approach.

8. A summary and some possible priorities

The U.S. Census Bureau's major efforts (*e.g.*, Prevost and Leggieri 1999) to research an administrative record census are deserving of applause. Even though the Census Bureau is now well underway in its ARC research, it still might be of value to reiterate key points and priorities.

8.1 Constancy of purpose

Deming, in setting out his famous 14 points, lists "Constancy of Purpose" or, in the words of the old Negro spiritual, "Keep your eyes on the prize." With all the extra challenges of running a census in 2000, keeping focused on the future may be the hardest task facing the excellent staff assembled. Temptations to cut budgets or reassign key people must be avoided. After the decennial census, separate budgeting should be sought and the sums involved will need to be large. Thinking that the big efforts are connected with the 2000 Census and could then slack off for a while is just flat wrong. The research effort will need to grow and grow.

8.2 Environmental scan

While the responsibility for the official census count will not change any time soon (if ever), census-taking will no longer be the monopoly it has been. Ways to integrate independent information sources will be essential to how the Census Bureau's success is measured. Levels of accountability can be predicted to increase. The Connecticut case study, discussed in section 3, is just one example.

What is crucial to see is that, ironically, central statistical agencies – including places like the Census Bureau – could well be left behind in the information age. Census Bureau market share in the information sector has been falling for decades and, *ceteris paribus*, a steeper drop is quite likely during the next ten years, given the slow pace of change inherent in a government agency. The Census Bureau's administrative record research and its continuing emphasis on being the leader in key information technologies could mitigate this trend, but probably not reverse it.

8.3 Assumptions

The privacy assumptions are the ones to worry about the most, as leaders at the Bureau, like Gates, have long been saying. Careful watching and listening are needed. The use of an advisory Institutional Review Board, not mentioned earlier, might be considered – to provide independent oversight with the public's interest in mind, especially on record linkage. Alternatively, now that there is a Census Monitoring Board, the Board might be the natural place to focus advice on handling privacy concerns and in doing priority setting.

The real concern is not that the Census Bureau will proceed rashly, but that it might be too timid. To quote Emerson, "Be bold, be bold, be not too bold." Certainly the Census Bureau should seek legislation like the Statistics Act in Canada, in order to assure the cooperation of administrative agencies in providing data for an ARC. As already noted, the development and enactment of such legislation would provide the opportunity for a public debate on ARC ideas, something that must occur before an ARC could be done.

8.4 Suggestions for 2000 Census

The inclusion of the SSN question in the quality check sample for 2000 is crucial to building the bridge from old to new. The suggestion to produce a simultaneous partial ARC estimate will help not only in testing an approach that will be needed in 2010, but also in validating both the ARC and the census itself. Regarding other administrative record research, carpe diem – seize the moment – especially the opportunity to acquire key files (an effort already well underway according to Prevost and Leggieri 1999) and to strengthen long-term partnerships that build human capital.

8.5 Intercensal steps

Creating a greater positive synergy between the current monthly and annual survey programs and the intercensal estimates program is key too. Transforming the intercensal estimates effort to one that is transaction-based, rather than essentially cross-sectional, would be the other major priority. Integrating special census results would also be crucial.

8.6 Current survey steps

The opportunities for administrative record applications in the American Community Survey are excellent, if they continue to be grasped quickly enough. However, the time from inception to results for new census surveys is often too long. Censuses have cycle times that extend arguably over more than ten years. The introduction of a new frame in the current survey programs has been growing. Whatever is done after the next census, it needs to be a lot quicker than after the last.

8.7 Additional possible steps

Of the two suggestions in the last section, the tracing sample has the most appeal as basic research. The Canadian experience can help here, but payoffs are uncertain. Tracing would help in addressing immigration flows, both legal and illegal. Obviously, as proof of concept and to make the ARC operationally feasible, special censuses will be critical. The budgeting will have to be a lot heavier in the early years of the decade than historically has been the case for the census pilots and dress rehearsals done in the 1980's and 1990's. Planning for the continuing research after 2010 should be a priority, as well, and might begin now and be revisited at least annually.

In the March 26, 1999, issue of *Science* (Cohen 1999), there is a news item entitled "The March of Paradigms." It tracks the growing number of scientific papers that use the phrase "new paradigm" in their titles or abstracts. It goes on to state that –

Many of these claims, however, may not be quite the kind of developments science philosopher Thomas Kuhn had in mind when he made the term new paradigm famous with his paradigm-shifting 1962 book, *The Structure of Scientific Revolutions*.

Despite this caution, the change to a partial ARC does qualify as a paradigm-shifting event and should be studied from that perspective. In this connection, compliments are due to all those who have already attempted and achieved it around the world. Best wishes to the U.S. Census Bureau in their research on it now.

Acknowledgements

For the ideas in this paper there are many who deserve thanks. Some of the key people are Wendy Alvey, Bonny Bye, John Leyes and Peter Sailer. Thanks are also owed to the Associate Editor and two very helpful referees.

References

Alvey, W., and Scheuren, F. (1982). Background for an administrative records census. (With discussion by John Leyes, Statistics Canada). *Statistics of Income and Related Administrative Record Research*. Washington, DC: U.S. Department of the Treasury, Internal Revenue Service.

Alexander, C., and Chand, N. (1999). Indirect estimation with administrative records and the American Community Survey. *1999 Federal Committee on Statistical Methodology Proceedings*. Washington DC: U.S. Bureau of the Census.

Bellhouse, D. (1988). *Hanbook of Statistics*: *Sampling*, *A brief history of random sampling methods*. New York: North-Holland.

Biskupick, J. (1998). Division of representation, funds at stake in census feud. *The Washington Post*. 27 November 1998.

Blum, O. (1999). Combining register-based and traditional census processes as a pre-defined strategy in census planning. *1999 Federal Committee on Statistical Methodology Proceedings*. Washington DC: U.S. Bureau of the Census.

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.

Carter, R., and McClean, K. (1996). Using administrative data in the Canadian census: Experiences and plans. *Statistical Journal of the United Nations*, 13, 4, 375-383.

Clark, C., and Gates, G. (1999). *Memorandum on Restricted Access Policy for Administrative Records*. U.S. Bureau of the Census, June 25, 1999.

Cohen, J. (1999). The march of paradigms. *Science*, 283, 1998-99.

Czajka, J.L., and Walker, B. (1989). Combining panel and cross-sectional selection in an annual sample of tax returns. *Proceedings of the Section on Survey Research Methods*, *American Statistical Association*, 463-468.

Czajka, J.L., Moreno, L. and Schirm, A. (1997). On the Feasibility of Using Internal Revenue Service Records to Count the U.S. Population. Washington, DC: Mathematica Policy Research.

Deming, W. (1986). *Out of the Crisis*. Cambridge MA: MIT Press.

Edmonston, B., and Schultze, C. (Eds.) (1995). Modernizing the U.S. Census, Panel on Census Requirements in the Year 2000 and Beyond. Committee on National Statistics, National Research Council, Washington, DC: National Academy Press.

Fellegi, I. (1997). Record linkage and public policy - a dynamic evolution. *Record Linkage Techniques*, 3 - 12. Arlington VA: Ernst and Young, LLP.

Gates, G. (1999). Data Mining, Panel on Privacy and Statistics in the New Millennium. Panel presentation at the Joint Statistical Meetings, Baltimore, MD.

Gates, G., and Bolton, D. (1998). Privacy research involving expanded statistical uses of administrative records. *Proceedings of the Government and Social Statistics Section*, American Statistical Association.

Hansen, M., and Hurwitz, W. (1943). On the theory of sampling from finite populations. *Annals of Mathematical Statistics*, 14, 333-62.

International Statistical Institute (1994). *The Future of Statistics*, (Ed., Z. Kenessey). ISBN: 90-73592-11-9.

Jabine, T., and Scheuren, F. (1987). Record linkages for statistical purpose: Methodological issues. *Journal of Official Statistics*, 2, 255-277.

Jaro, M. (1989). Advances in record-linkage methodology as applied to matching the 1985 Census of Tampa, Florida. *Journal of the American Statistical Association*.

Jensen, P. (1983). Towards a register-based statistical system - some Danish experiences. *Statistical Journal of the United Nations*, 341-365.

Johnson, N., and Kotz, S. (Eds.) (1997). *Leading Personalities in Statistical Science*: *From the Seventeenth Century to the Present*. New York: John Wiley & Sons, Inc.

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

Kennickell, A., and Woodburn, L. (1997). Consistent Weight Design for the 1989, 1992, and 1995 SCF, and the Distribution of Wealth, available from the web site http://www.bog.frb.fed.us/pubs/oss/oss2/scfindex.html.

Kish, L. (1990). Rolling samples and censuses. *Survey Methodology*, 16, 1, 63-71.

Kuhn, T. (1970). *The Structure of Scientific Revolutions*. Chicago IL: University of Chicago Press.

Leyes, J., and Elsl-Culkin, J. (1994). Administrative social data in Canada: Some results and some implications. *Statistics of Income*: *Turning Administrative Systems into Information Systems*, U.S. Internal Revenue Service: Washington, DC.

Myrskla, P. (1991). Census by questionnaire – census by registers and administrative records: The experience of Finland. *Journal of Official Statistics*, 7, 457-74.

National Academy of Sciences (1996). Massive Data Sets: Proceedings of a Workshop Committee on Applied and Theoretical Statistics, Board on Mathematical Sciences, National Research Council: Washington, DC.

Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97, 558-625.

Prevost, R., and Leggieri, C. (1999). Expansion of administrative record uses at the census bureau; a long-range research plan. *1999 Federal Committee on Statistical Methodology Proceedings*, Washington DC: U.S. Bureau of the Census.

Redfern, P. (1989). Population registers: Some administrative and statistical pros and cons. *Journal of the Royal Statistical Society*, Series A, 153, 1-41.

Sailer, P., and Weber, M. (1998). The IRS population undercount: An update. *Proceedings of the Section on Survey Research Methods*, *American Statistical Association*.

Sailer, P., Weber, M. and Yau, E. (1993). How well can the IRS count the population? *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

Schaible, W. (Ed.) (1996). *Indirect Estimators in U.S. General Programs*. New York: Springer-Verlag.

Scheuren, F. (1979). Integrated linked administrative statistical sample. LASS Working Notes, U.S. Social Security Administration.

Scheuren, F., Oh, H.L., Vogel, L. and Yuskavage, R. (1981). Methods of Estimation for the 1973 Exact Match Study. Studies from Interagency Data Linkages. U.S. Department of Health and Human Services, Social Security Administration, Publication 13-11750.

Scheuren, F. (1985). Methodological issues in linkage of multiple databases. In *Record Linkage Techniques - 1985*: *Proceedings of the U. S. Internal Revenue Service*, 155-178.

Scheuren, F. (1990). Discussion of Kish (1990). *Survey Methodology*, 16, 1, 72-79.

Scheuren, F. (1995a). A U.S. Administrative records census. *Chance*, 8, 2, 43-45.

Scheuren, F. (1995b). Private lives and public policies: Confidentiality and accessibility of government services. In *Journal of the American Statistical Association*, 90, 386-387. Washington, DC: National Academy Press (1993).

Scheuren, F. (1997). Linking health records: Human rights concerns. *Record Linkage Techniques*. Washington DC: Ernst and Young, LLP.

Scheuren, F., and Petska, T. (1993). Turning administrative systems into information systems. *Journal of Official Statistics*, 9, 109-119.

Scheuren, F., and Winkler, W. (1993). Regression analysis of data files that are computer matched. *Survey Methodology*, 19, 39-58.

Scheuren, F., and Winkler, W. (1997). Regression analysis of data files that are computer matched - Part II. *Survey Methodology*, 23, 157-165.

Singer, E. (1999). Data Mining, Panel on Privacy and Statistics in the New Millennium. Panel presentation at the Joint Statistical Meetings, Baltimore, MD.

Statistics Canada (1993-1996). *Survey of Labour and Income Dynamics*: *Research Papers*. Catalogue No. 75F0002MIE, 93-01, 94-03, 94-11, 95-19 and 96-12.

Statistics Canada (1999). *Statistics Canada's Survey of Financial Security*: *Update - July 1999*. Catalogue 13F002MIE 99006.

Steffey, D., and Bradburn, N. (Eds.) (1994). *Counting People in the Information Age, Panel to Evaluate Alternative Census Methods*. Committee on National Statistics, National Research Council, Washington, DC: National Academy Press.

Sukhatme, P. (1935). Contributions to the theory of the representative method. *Journal of the Royal Statistical Society Supplement*, 2, 263-68.

Thomsen, I., and Holmoy, A.M.K. (1998). Combining data from surveys and administrative record systems, the Norwegian experience. *International Statistical Review*, 66, 2, 201-221.

Thomsen, I., and Zhang, L. (1999). The effects of using administrative registers in economic short term statistics: The Norwegian Labour Force Survey as a case study. 1999 Federal Committee on Statistical Methodology Proceedings. Washington DC: U.S. Bureau of the Census.

White, G., Mulrow, E. and Scheuren, F. (1999). *Connecticut Jury Record Linkage Research*. Washington DC: Ernst and Young, LLP.

Zaslavsky, A.M., and Wolfgang, G.S. (1993). Triple-system modeling of census, post-enumeration survey, and administrative list data. *Journal of Business and Economic Statistics*, 11, 279-288.