Proceedings of Statistics Canada Symposium 2002
Modelling Survey Data for Social and Economic Research

INFERENCES FOR FINITE POPULATIONS USING MULTIPLE DATA SOURCES WITH DIFFERENT REFERENCE TIMES

Jana Asher[1], Stephen E. Fienberg[2], Elizabeth Stuart[3] and Alan M. Zaslavsky[4]

ABSTRACT

While censuses and surveys are often said to measure populations as they are, in fact most reflect information about individuals as they were at the time of measurement, or even at some prior time point. Inferences from such data therefore should take into account change over time at both the population and individual level. Specifically, at the population level, the estimands of interest, such as the size or mean characteristics of a population, might be changing. At the same time, individual subjects might be moving in and out of the frame of the study or changing their characteristics. Such changes over time can affect statistical studies of government data that combine information from multiple data sources, including censuses, surveys and administrative records, an increasingly common practice. Inferences from the resulting merged data bases often depend heavily on specific choices made in combining, editing and analyzing the data that reflect assumptions about how populations of interest change or remain stable over time. In this paper we provide a first effort at a unifying framework for such inference problems, illustrating it through a diverse series of examples: (1) Combining administrative records for estimating the size of the U.S. population, (2) Estimating residency status on census day using multiple administrative records, (3) Estimating the prevalence of human rights abuses, (4) Using rolling averages from the American Community Survey.

KEY WORDS: Administrative records; American Community Survey; Hierarchical models; Multiple systems estimation methods; Stochastic models.

1. INTRODUCTION

In the statistical literature on sample surveys and their design, we usually claim that censuses and surveys measure populations as they are, either in terms of facts or opinions (e.g., see Cochran, 1977; Lyberg and Cassel, 2001). Coupled to this perspective, the standard finite sampling theory treats measurements as fixed values (c.f., Särndal, Swensson, and Wretman, 1992). But the reality is typically different from these pronouncements and the standard theory. More often than not, we gather data about individuals as they were at prior times, using measurements at yet other times. In such contexts inferences are affected by temporal changes. At the population level, estimands of interest, e.g., size or mean characteristics of a population, might be changing. At the individual level, subjects might move into and out of the frame of study or change their characteristics.

Such changes over time can affect statistical studies of government data that combine information from multiple data sources, including censuses, surveys and administrative records, an increasingly common practice. Inferences from the resulting merged databases often depend heavily on specific choices made in combining, editing and analyzing the data that reflect assumptions, usually implicit, about how populations of interest change or remain stable over time. To put these inferences on firm footing, we need explicit stochastic models that capture changes over time. The need for such stochastic models has long been recognized in the context of panel surveys where over-time changes are the focus of interest (e.g., see Fienberg, 1980; Stasny, 1990; Pfeffermann et al., 1998; Ybarra and Lohr,

[1]Department of Statistics, Carnegie Mellon University, Pittsburgh PA, U.S.A., 15213-3890 (asher@stat.cmu.edu)
[2]Department of Statistics and Center for Automated Learning and Discovery, Carnegie Mellon University, Pittsburgh PA, U.S.A., 15213-3890 (fienberg@stat.cmu.edu)
[3]Department of Statistics, Harvard University, Cambridge MA, U.S.A., 02138 (stuart@stat.harvard.edu)
[4]Harvard Medical School, Department of Health Care Policy, Boston MA, U.S.A., 02115 (zaslavsk@hcp.med.harvard.edu)

2002), but it has been by and large ignored for situations where the focus is the estimation of a single population quantity at a specific point in time or a sequence of such quantities.

In this paper we provide a first effort at a unifying framework for such inference problems involving multiple measurements and systems, illustrating it through a diverse series of examples including:

1. Combining administrative records for estimating the size of U.S. population.
2. Estimating residency status on census day using multiple administrative records.
3. Estimating killings in Kosovo.
4. Using rolling averages from the American Community Survey for formula allocations.

In the following sections we describe each of these examples and then we return, in a final section, to their commonalities and broader modelling issues.

2. CENSUS 2000 MULTIPLE SYSTEMS ESTIMATION

For at least the last six decennial censuses, the U.S. Census Bureau has used a variety of methods for assessing the accuracy and coverage of the decennial census counts. In 2000, this assessment took the form of the Accuracy and Coverage Evaluation Survey (ACE), a survey of approximately 314,000 housing units. In a separate effort, the Bureau has investigated the use of administrative records either as a substitute for or as a supplement to the traditional enumeration and other surveys. Specifically, the U.S. Census Bureau's Planning, Research and Evaluation Division (PRED) has developed a program to produce, on an annual basis, an administrative records "superlist" called the Statistical Administrative Records System (StARS), which is the end result of the careful merging of six or more administrative records sources. As part of the Census Bureau's evaluation program for Census 2000, PRED developed an administrative records experiment (AREX 2000) in which the feasibility of using StARS for an administrative records census was explored in five counties of Colorado and Maryland.

As part of AREX 2000, an effort was undertaken to develop multiple systems estimation methods that would combine decennial census 2000 data, ACE survey data, and the AREX 2000 Bottom-Up administrative records file in order to create block-level population counts. Several issues arose during this research. One issue was the coverage afforded by ACE; matching across the three files creates 3-way cross-classifications for ACE sample blocks and 2-way cross-classifications for non-sample blocks. A modeling umbrella for both these data types had to be developed. A more serious issue related to the reference time for each list; the AREX data were collected well prior to 2000, the Census 2000 data were collected around April 2000, and the ACE data were collected later in 2000. Combining these data into a cross-classification table therefore proved problematic.

2.1 Record Linkage and estimation

The first step in the multiple systems estimation project was to create a cross-classification table of population counts for each block using the three data sources described above. Because each data source references a different time period, it was possible for an individual to have different addresses in different lists, and therefore be in different blocks in different lists. To address this issue, we assumed multiple addresses are always the result of different reference times for lists (not multiple residency). Since the desired reference point was that of the decennial census (April 1, 2000), we placed all triple matches in their Census 2000 addresses. Double matches mirrored the same strategy; Census 2000-ACE and Census 2000-AREX 2000 matches were put in their Census 2000 addresses, but ACE-AREX 2000 matches were put in their ACE addresses.

Several multiple systems estimation frameworks were subsequently developed for analyzing the resulting cross-classification tables. These frameworks incorporate assumptions about missingness of ACE sample blocks by including either stratification or covariates based on the ACE sampling frame. Modeling within each ACE stratum allows the missingness to be ignored; including covariates based on the characteristics by which blocks were stratified for the ACE sample allows the missingness to be accounted for within the model. The models include a common set of parameters (within stratum for the stratified model) across all blocks for the ACE effect and interaction effects between ACE and the other lists, and individual parameters for each block for the decennial

census and AREX 2000 effects and interaction. The results of this modeling effort will be published in a future paper; further information about the models is available in Asher and Fienberg (2002).

2.2 Benefits and issues

One benefit of this modeling framework is its flexibility; these models can be adapted to reflect not only omissions, but also erroneous inclusions or errors in each data source. These errors are inherently stochastic in nature; over time, population counts shift for particular blocks as births, deaths, and migrations occur. Because our goal is to estimate the block population counts for census day, we chose to project forwards for the AREX individuals and backwards for the ACE individuals during the merging of the data. In other words, the stochastic nature of these data was accounted for in the creation of the cross-classification table, not in the model itself. In future work, we will attempt to incorporate the stochastic nature of these data directly into the model.

To that end, an issue with this modeling framework is that the models are really not for "estimating" data on individuals or even households, but for aggregates. As a result, the stochastic information available at the individual level (e.g., address changes) must be incorporated into a model that relies on block-level cross-classifications. One potential solution is a further cross-classification by migration status.

3. MODELING RESIDENCY AND MIGRATION

This section describes an effort to model the individual-level address changes mentioned above. As mentioned in Section 2, in recent years there has been discussion regarding the use of administrative records to supplement U.S. Census Bureau operations, particularly the decennial census enumeration. One of the drawbacks of administrative records is that their coverage period does not coincide with census day, and may extend considerably earlier. Due to migration, individuals might be included in the records who are no longer residents on census day. For example, somebody might renew a driver's license at a particular address but have no further input to a driver's license record system until the next renewal is due some years later. Estimated census counts that include all such individuals will be biased. We develop a model of migration and observation in administrative records that utilizes the full information in a set of records, particularly the dates of the records, to predict whether someone is still a resident of the address on census day. In this way, census day population counts can be estimated for the areas of interest, accounting for the time discrepancy.

3.1 Hierarchical stochastic modeling with migration

We present here an overview of the model; further details, algorithms for fitting the models, and simulation results appear in Stuart and Zaslavsky (2001, 2002). The model is an extension of multiple-system estimation, which typically treats several data systems as effectively contemporaneous, and of capture-recapture methods for estimation of animal populations, which rely on the "capture" of individuals using similar methods at multiple time points and may include a component for migration of individuals in and out of the capture zone. The data is assumed to consist of several systems, each of which includes a set of records that reflect the presence of an individual at an address on a particular date. The record systems might be administrative record systems, the census, or a coverage evaluation survey. We consider each administrative record system or census to be a "capture." The process that generates the records is represented by a 3-level hierarchical model, which can be described generically as follows. Within this general framework, a variety of specific models can be accommodated.

Level I (data): Observation in a particular record system in area of interest and within time frame;
Level II: Presence of the individual in a particular system at any place or time, and the time the individual would be observed if in the system;
Level III: The individual's migration history.

Level III describes the distribution of in- and out- migration dates from the area of interest. A simple and reasonably flexible model that we have used in simulations describes the population as a mixture of two types of individuals: stayers and movers. Stayers never move to or from the area of interest, whereas movers have a constant hazard of

moving out and are replaced at an equal rate by in-movers. The assumption of constant hazard is likely to be a reasonable approximation over short intervals, but other migration models can also be used.

Levels I and II represent specific observation models for each type of record. Level II variables for each person by each record system include an indicator variable for inclusion of the person in the record system and a variable for the date on which the person would be recorded, if included. The parameters thus include the probability that an individual has that type of record (which may depend on covariate values through a regression model) and parameters for the distribution of dates associated with that record type. For example, if driver's licenses are renewed on birthdays, the renewal record dates would be uniformly distributed through the year; income tax filing dates, on the other hand, have a non-uniform distribution peaked at the annual filing deadline. The Level I model describes whether a given person will appear in a given system as a function of the Level II variables and the person's migration history (Level III variables); this part of the model might be deterministic. The specific models will depend on the details of the record systems. Examples can be found in Stuart and Zaslavsky (2001, 2002). Multiple record systems can be easily modeled concurrently by modeling each separately in Levels I and II and then combined assuming independence of the systems conditional on the covariates and migration dates.

Due to the hierarchical structure, several levels of inference are possible using this model. Global parameters such as the coverage probabilities and migration rates can be estimated given adequate data. For the purposes of census enumeration, the posterior probability that each individual was a census day resident can be obtained. Census day population can then be estimated at a census block or tract level as the expected number of residents on census day.

This model has been implemented with promising results using both simulated data and Census Bureau administrative records data. See Stuart and Zaslavsky (2001, 2002) for a summary of simulation results. In unpublished work, a set of record systems from the Statistical Administrative Records System (StARS) at the U.S. Bureau of the Census gave population estimates close to the decennial census counts. At an individual level, results are similar to what would be expected. For example, an individual observed in the area only in a Medicare enrollment file 2 years prior to census day is relatively unlikely to be a census day resident of that area. However, an individual observed in a Medicare enrollment file 2 years prior to census day and in an Internal Revenue Service tax return file at the same address 6 weeks before census day is very likely to be a census day resident of that area.

3.2 Benefits and issues

A key benefit of the approach described here is the modeling at an individual level. The full information in the records can thus be used, including covariate and date information, and all address information for each individual. Previous use of administrative records at the U.S. Census Bureau relied on the selection of a "best" address for each individual, based on a series of somewhat ad hoc selection rules. This model uses all address information available to determine the probability of census day residency for each individual.

A second major benefit of this approach is the flexibility of models for various record systems, which can fully reflect each system's characteristics. The general framework accommodates a variety of record types, such as tax records, Medicare enrollment files, and Selective Service registration lists. For example, to model Internal Revenue Service 1040 tax returns, a model can be developed that describes the probability of an individual filing a tax return, which can depend on available covariates such as age, region of the country, sex, and family size. The distribution of filing dates could be modeled parametrically or estimated non-parametrically using the empirical distribution of filing dates. The migration model is also very general, and can allow for seasonal migration or regional variations in migration patterns. All of these models can be made more realistic and complex as desired.

By using individual records and the timing of these records to obtain estimates of the population size on census day, the framework described here affords us with an approach to mitigate the effects of the discrepancy in record system coverage time frames.

4. KILLINGS IN KOSOVO, 1999

As part of the prosecution's case against Slobodon Milosevic in his trial before the International Criminal Tribunal for the Former Yugoslavia, an analysis of available data on ethnic Albanian migrations and deaths in Kosovo during the period of March–June of 1999 was undertaken by a team of American scientists (see Ball, et al. 2002). The purpose of this analysis was to support or refute each of three theories as to the cause of these migrations and deaths: 1) actions of the Kosovo Liberation Army (KLA) resulted in ethnic Albanian deaths and migrations, 2) air attacks by the North Atlantic Treaty Organization (NATO) resulted in ethnic Albanian deaths and migrations, and 3) a systematic campaign by Yugoslav forces resulted in ethnic Albanian deaths and migrations. Because of the nature of the questions raised by the prosecution, dates were an essential aspect of the analysis undertaken. This date information, however, was not exact in many cases. Migration information was obtained from two sources; records kept by the border guards on the Kosovo-Albania border, and UN reports. Four sources of data were available on ethnic Albanian deaths, each of which was compiled from several interviews of survivors and/or exhumations. Further, matching across these different data sources resulted in additional date conflicts.

As a result, each of the dates for an individual death was assigned a descriptor to indicate the potential accuracy of that date: exact, approximate, imprecise, and unknown. This allowed a date to be chosen for each individual according to a logical system. Roughly, the date of death assigned to an individual was either the one with the most precision or the earlier date if more than two dates had the same level of exactness. The assumptions underlying this system are that 1) bodies were not buried right away, therefore interviewees didn't necessarily see bodies the day of death and 2) more exact dates were most probably collected closer to the true time of the death. There were, however, 204 individual records that had no date of death assigned to them. We chose to estimate the date of death for these individuals using a hot-deck procedure that randomly assigned a date to the individual from donor individuals whose deaths were geographically closest to the individual with the missing date. As part of this procedure, three dates were selected for each record, and a weight of 1/3 was assigned to each date.

Once a date had been chosen for each individual death, multiple systems estimation was used to estimate the number of deaths for two-day periods using a cross-classification of the four data sources available. The main result of this analysis was a series of two-day estimates of the numbers of migrations and deaths in Kosovo in March–June of 1999. As can be seen in the following figure, the dates assigned to each death and each migration are essential in showing the association between migrations and deaths. Further, since lulls in deaths may correspond to exact historical events (e.g., April 6-7 represents a unilateral cease-fire on the part of the Yugoslav army in observance of orthodox Easter), an understanding of the accuracy of the date information is essential.

Figure 2: Estimated total refugee migration and killings over time

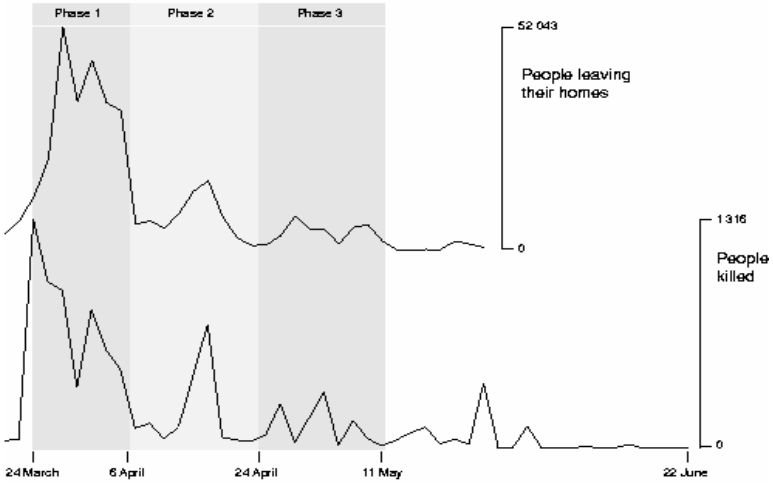

Figure from page 6 of Ball, Betts, Scheuren, Dudukovich, and Asher (2002).

4.1 Kosovo date assignments

In order to test the sensitivity of the shape of the curve of deaths over time to the date assignment mechanism, the dates were shifted and the analysis re-run to determine the resulting curve. Two different sets of dates were determined, an "early" set and a "late" set. The dates used were the 25th percentile and 75th percentile dates for each individual with three or more dates of death and the two dates available for individuals with two dates of death. Individuals with one or no dates of death available were assigned a range of dates via hot deck imputation; the early and late dates for these records were taken as plus or minus half the hot decked range. The following figure shows the original curve over time compared to the "early" date and "late" date curves. The substantive results of the analysis are not changed by the perturbation.

Figure 22: Estimated killings over time with alternative

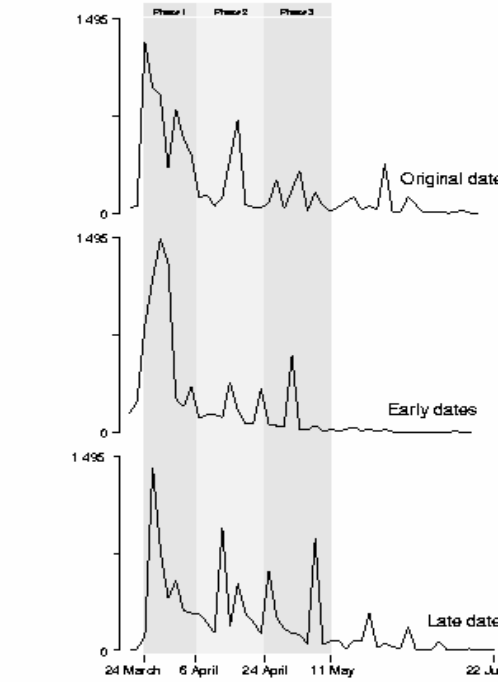

Figure from page 61 of Ball, Betts, Scheuren, Dudukovich, and Asher (2002).

4.3 Benefits and issues

One serious issue with these data is that they are formed from a retrospective reconstruction in spite of their inherently stochastic nature. It is possible that interviewees could forget details of the deaths they witnessed as time passed, altering the results of the analysis significantly. As a result, the analysis team had to be careful to give every benefit of the doubt to the defendant; data that was unreliable was necessarily not included in the analysis. As an example, all reports of deaths for which a unique identifier (i.e., full name) could not be determined were not used.

When deaths could be uniquely identified, differences in the reported date of death were problematic. As described, a logical scheme was used for determining a "best" date of death, and a sensitivity analysis suggests that taking more extreme values for death dates does not significantly change either the pattern of deaths or the conclusions of the analysis. A potentially better solution would have been to use formal models that incorporate the uncertainty surrounding the dates of death.

This example differs from the previous two examples in that estimates of population counts (in this case, of death counts) are required for several time points instead of just one time point. The next example provides the most potent case of data that supports a stochastic model.

5. AMERICAN COMMUNITY SURVEY

5.1 Small-area data, timeliness, and the American Community Survey

For the past 6 decades, the statistical system of the United States has relied on the "long form" of the decennial census for small-area information on income, characteristics of housing, and other topics besides the basic demographic information included in the basic census form. The long form is distributed to a sample of about 1/6 of households, with higher sampling rates in areas of relatively sparse population. A major deficiency of the long form as a source of small-area data is that the data collection only takes place once per decade. Thus, for example, in 2001 the most recent available data had been collected as part of the 1990 decennial census for which the reference year for income was 1989, and therefore the data were 12 years out of date, a poor basis for planning and allocation of resources.

Two strategies have been considered for obtaining more timely estimates. One is to make use of statistical small-area estimation techniques to update decennial long-form estimates using annual data from administrative systems such as income tax forms and enrollment in the food stamp program together with current survey data, and combining the sources through regression models. This approach was used to develop estimates of numbers of school-aged children in poverty, part of the basis for allocation of federal education dollars to local education authorities (Citro and Kalton, 2000a, 2000b). Estimates derived from this program were substantially more accurate for most years than those carried forward from the previous decennial census. The predictive error of these models, however, is larger than the error of the census for the year in which it is conducted.

The second strategy is to initiate a new survey that collects data in a more timely way. The American Community Survey (ACS), inspired by the proposal of Kish (1990, 1998) for a rolling census, would replace the decennial census long form with a survey conducted continuously over the decade, leaving only the basic population count and "short form" demographic information to the decennial data collection (Alexander 2001). (The ACS would also contribute to intercensal population counts because it would count residents of existing units and its sampling frame would include addresses of newly constructed units.) In addition to yielding more timely data, such a program might have significant operational and quality advantages, by removing the long-form data collection from the overstressed decennial operation and placing it in the hands of a permanent professional staff.

When fully implemented, each month's sample for the ACS would be a national probability sample. Over the course of 5 years, the cumulative ACS sample would be equal in size to the decennial long-form sample. For large areas, a single year's data would provide adequately precise estimates, while for smaller areas, a 5-year cumulation (pooling of data over time) might be required. Although statistics based on such cumulation would not be representative of any single point in time, even a 5-year cumulation is on the average about 3 years old, compared to decennial data that are on the average more than 5 years old. (Instead of simple cumulation, equivalent to an unweighted moving average estimate, a weighted moving average can be used that gives less weight to more distant years, making the data more current at the cost of a slight increase in variance.) Thus, the transition to the ACS would replace a sharp snapshot with a slightly fuzzy (in small details) movie, with one frame per year (or in principle, one per month, although it is unlikely that the data would be made available on such a schedule).

5.2 Using continuously collected data for planning and allocation

It might seem obvious that more timely data would be a better basis for planning – for example, one would not build roads on the basis of the distribution of population 10 years ago if more recent data were available. Even for such purposes, the transition to a new schedule of data release requires substantial changes in procedures, and perhaps more critically a change in mindset to effectively use data that represent a temporally moving target.

The switch from decennial long-form estimates to annual estimates based on the ACS has some particularly interesting implications for allocation of funds from the federal government to local areas using formulae based on area statistics such as poverty rates or numbers of children in poverty. In the remainder of this section, we focus on

these implications, drawing on conclusions from Zaslavsky and Schirm (2002), where more detailed analyses and simulation results are presented.

One important difference between decennial and continuously sampled data is that the former represents a one-year temporal sample from the decade while the latter provides data for every year. This distinction can affect the equity of allocations. A single year is not necessarily representative of the distribution of the populations of interest over the course of the decade because cycles of employment and the rise and fall of locally dominant industries are not synchronized across the nation. With decennial data, areas that happen to demonstrate lesser need in the census year are disadvantaged for the entire decade. Continuous measurement protects against such affects, because each year's data receives about the same cumulative weight in the long run, regardless of whether the estimate in a single year is based on that year's data or a cumulation.

The most surprising unintended consequences arise through interactions among the data source, the estimation procedure, and the formula for distribution of funds. These components of the allocation process are intimately connected and cannot always be sharply distinguished. For example, if procedures specify that a 3-year cumulation is to be used as the basis for allocation, is this part of the estimation procedure or of the formula? However, there are some specific types of provisions often used in allocation formulae that make them sensitive to changes in the frequency of reallocations.

One such provision is the "hold harmless", which limits downward changes in the allocation to any unit. For example, the formula might guarantee that no local authority would receive less than 80% of its previous allocation, in order to maintain the stability of program activities. Such a provision more tightly constrains allocations when they are recalculated every 10 years than when they are recalculated annually; thus if a new data source leads to more frequent reallocations, it will change the effective impact of the hold harmless provision. A more subtle point is that a high (close to 100%) hold harmless can tend to "ratchet up" allocations for units for which estimates have large sampling variability: when estimates of need are randomly higher than true values, the allocation goes up, and then the hold harmless keeps it up. Sampling variability can be reduced by cumulation (at the cost of less responsiveness to changing need); surprisingly, for a given amount of cross-sectional sampling variation the cumulated estimates give *less* "ratcheting up" because moving averages are more stable over time than independent estimates with the same variance.

Another common provision is the use of a threshold criterion for funding; for example, part of the budget for U.S. education aid is distributed only to local education authorities with at least 6500 poor children or a child poverty rate exceeding 15%. Again, such a provision can interact with sampling characteristics of the estimates. With more variable estimates, the threshold is "smoothed" in the sense that units whose actual rates are just below the threshold have a probability approaching one-half that their estimates will be above the threshold, and the reverse happens for units that are just above the threshold.

While this smoothing of a necessarily arbitrary threshold may not be entirely undesirable, the amount of smoothing, like the effects of "hold harmless", is dependent on the random variation in the estimates and therefore on the sample design and estimation procedure. We believe that it is undesirable that sample design should have implications for equity, other than the general consideration of accurate and as far as possible unbiased estimation for each area. Thus, estimation procedures and formulae should be designed to minimize such effects, which are particularly associated with nonlinear features such as hold harmless or threshold provisions. Use of linear estimation procedures and formulae tends to reduce such effects because each year's true levels of the variables affecting allocations then enters in a relatively predictable and symmetrical way into the allocations, aggregated over time.

6. SOME COMMON ISSUES AND OVERARCHING IDEAS

In this paper we have provided, through a diverse series of examples, a first effort at thinking about a unifying framework for inference problems involving multiple measurements and systems. These examples differed in a number of respects that illustrate the variety of inferential problems that must be addressed. For the ACS the date is an explicit component of the sample design, for administrative records dates are inferred from information in the

files, and in the Kosovo example, inferences about dates were a critical component of the analysis. The administrative records could be assembled to construct longitudinal data for individuals, while the ACS samples are longitudinal for areas but not for individuals. The migration model focused on longitudinal behavior, while the other examples aimed at inference for aggregates. In each case, however, the population size and/or characteristics were specific to the time of collection and the time of reference, and in each case we needed to understand the inference problem by formulating a stochastic framework for the underlying phenomenon.

More generally in such situations, inferences may focus on a specific time point or on estimating quantities or processes over time, but the data are specific to the time frame of collection. Dealing with such issues in the context of administrative data is clearly difficult since the "dates" of record may vary substantially relative to the dates associated with the data file as a whole. For example, the date of issuance of the renewal of a driver's license and not the date of the license list may be most relevant, unless there is some evidence about reporting of new addresses between dates of renewal. Thus, the data have an extra dimension of time that is part of the information contained in each observation.

The stochastic nature of survey, census, and administrative data requires: (a) treating all data as random variables, including data relating to editing and corrections, (b) simplifying assumptions, e.g., by allocating individuals or events to specific times or locations, and (c) model-based analyses rather than design-based ones. Regardless of the treatment of such specifics, when considering multiple data sources we need to tie the various data sources to the full stochastic framework, even if only to "project" the data to a date of interest, e.g., April 1, 2000 which is the referent for the 2000 U.S. decennial census. Further, multiple dates for events and multiple residences for individuals create additional problems that require use of more complex data structures and models.

Our examples illustrate that new data-collection technologies, be they administrative or sample-survey based, create new statistical options and opportunities. Knowing that changes accumulate over time, we recognize that more timely data can be more useful. Thus, in the ACS, more recent data should be given more weight in the rolling cumulative samples. With multiple data sources for census estimation, various possibilities arise as we weigh both the accuracy of the different data sources for estimating parameters in the stochastic model and the timeliness of the data sources. Furthermore, developments in information technology make it possible to preserve temporally detailed information on transactions involving individuals, both in administrative records and for operations like the census that have been traditionally analyzed as if they took place at a single point in time. These new and richer sets will create new demands for statistical modeling.

ACKNOWLEDGMENTS

This research was supported in part by the U.S. Bureau of the Census through a contract with Research Triangle Institute supporting activities at Carnegie Mellon University and with National Opinion Research Center and Datametrics Research Inc. supporting activities at Harvard University.

REFERENCES

Alexander, C. H. (2001), "The American Community Survey for economic analysis", Washington, DC: U.S. Bureau of the Census.

Asher, J. and S. E. Fienberg (2002), "The Administrative Records Experiment in 2000: An Application to Population Count Estimation via Triple Systems Estimation", *Proceedings of the Government Statistics Section, American Statistical Association.*

Ball, P., Betts, W., Scheuren, F., Dudukovich, J., and J. Asher (2002), *Killings and Refugee Flow in Kosovo March – June 1999: A Report to the International Criminal Tribunal for the Former Yugoslavia*, Washington, DC: American Association for the Advancement of Science.

Citro, C. F. and G. Kalton, eds. (2000a), *Small-Area Income and Poverty Estimates: Evaluation of Current Methodology*, Washington, DC: National Academy Press, 2000.

Citro, C. F. and G. Kalton, eds. (2000b), *Small-Area Income and Poverty Estimates: Priorities for 2000 and Beyond*, Washington, DC: National Academy Press, 2000.

Cochran, W. G. (1977), *Sampling Techniques, Third Edition*, New York: Wiley.

Fienberg, S. E. (1980), "The measurement of crime victimization: Prospects for a panel analysis of a panel survey", *The Statistician*, 29, pp. 313-350.

Kish, L. (1990), "Rolling samples and censuses", *Survey Methodology*, 16, pp. 63-78.

Kish, L. (1998), "Space/time variations and rolling samples", *Journal of Official Statistics*, 14 , pp. 31-46.

Lyberg, L. and C. M. Cassel (2001), "Sample Surveys: The field", In N. J. Smelser and P. Baltes (eds.), *International Encyclopedia of the Social & Behavioral Sciences*, Vol. 20, Oxford: Elsevier, pp. 13475-13480.

Pfeffermann, D., C. Skinner, and K. Humphreys (1998), "The estimation of gross flows in the presence of measurement error using auxiliary variables", *Journal of the Royal Statistical Society, Series A*, 161, 13-32.

Särndal, C-E., B. Swensson and J. Wretman (1992), *Model Assisted Survey Sampling*, New York: Springer-Verlag.

Stasny, E. A. (1990), "Symmetry in flows among reported victimization classifications with non-random nonresponse", *Survey Methodology*, 16, pp. 305-330.

Stuart, E., and A. M. Zaslavsky (2001), "Using administrative records to predict census day residency", Proceedings of the Joint Statistical Meetings, American Statistical Assocation, Alexandria, VA: American Statistical Association.

Stuart, E., and A. M. Zaslavsky (2002), "Using administrative records to predict census day residency",in C. Gatsonis et al. (eds.) *Case Studies in Bayesian Statistics, Volume VI*, New York: Springer-Verlag, pp. 335-349.

Ybarra, L. M. R. and S. L. Lohr, (2002), "Estimates of repeat victimization using the National Crime Victimization Survey," *Journal of Quantitative Criminology*, 18, pp. 1-22.

Zaslavsky, A. M. and A. L. Schirm (2002), "Interactions between Survey Estimates and Federal Funding Formulas", *Journal of Official Statistics*, 18, pp. 371-391.