Models for combining longitudinal data from administrative sources and panel surveys

Gad Nathan
Central Bureau of Statistics,
66 Kanfey Nesharim St., Corner Bachi St.
91342 Jerusalem, Israel
gad@huji.ac.il

1. Introduction

While during the first half of the last century official statistics were based predominantly on censuses and administrative data, the last fifty years have been characterized by the predominance of sample surveys as the major source of data published by official statistical agencies. The development of advanced sample survey methodology for design and analysis of sample surveys and the technological advances in automated collection of data, via telephone and laptops, have enhanced the viability of the sample survey as a high quality and efficient data collection system. On the other hand, the expanded availability of large administrative data systems, in suitable electronic form, the exponential demand for detailed statistical data on national and local levels, and the increasing difficulties encountered in direct collection of data from respondents have all contributed to the reconsideration of the use of administrative data as a prime source for official statistics. A parallel development has been the increased interest in longitudinal data in order to investigate temporal relationships at the individual level (Binder, 1998). While some administrative sources can supply longitudinal information, data on the major topics of interest in longitudinal studies, such as gross flows in educational, social and economic characteristics, are not available in administrative sources and the longitudinal sample survey is still the major source for such data. The need for sample surveys still exists also for some important cross-sectional data, such as family expenditures, leisure activities and health status. Thus the role of sample surveys will continue to be important in conjunction with the use of data from administrative sources.

Similarly, the very high costs of the conventional population and housing census and its increasing undercoverage rates have led to the examination of alternatives for this major source of important detailed geographical data. In particular the increased use of administrative sources as a major component of the census is being considered in many countries (see Scheuren, 1999), together with an increased role for sample surveys, not only for evaluation and supplementation of content, but also to adjust for undercoverage (e.g. Brown et al., 1999). In this paper we consider the combination of a census, primarily based on administrative data sources, and sample surveys to supplement and complement the administrative information. This is the basis for the new strategy now being developed for the 2006 Census in Israel. The census is envisaged as a basis for continuing post-censal data collection, via the ongoing use of the administrative data, combined with a large-scale continuing longitudinal panel survey. Model-based analysis is proposed to combine these data sources. In section two we describe briefly the background and methodological issues involved in the design for this new approach. Section three proposes model-based imputation methods designed to combine the data from the longitudinal surveys and the administrative data and presents some initial results of simulations carried out to assess the proposed models.

2. The new strategy for the Israel 2006 Census

Israel's first population census was carried out in 1948, shortly after the foundation of the state and formed the basis for the national Population Register incorporating a unique permanent ID for each person. Since then four traditional censuses were carried out in 1961, 1972, 1983 and 1995, all linked to the continuing population register, to improve collection, for imputation and for evaluation. Since all births and new immigrants are immediately included in the register, it is virtually complete

but suffers from over-coverage (about 5-7%) and deficient data on current addresses (ca. 20% errors). Lack of cooperation by large parts of the population (2.5% overall non-response and 13% partial non response - in the last census), continuously increasing costs and a forecasted continuous decline in cooperation and resulting further deterioration of data quality dictated the need to consider a new strategy. The basic underlying idea is to take advantage of the population register and many other administrative sources (National Insurance, Driving licenses, Municipal tax systems etc.) as a combined census database and to evaluate, correct and complement it with an extensive set of sample surveys (possibly of size up to 300,000 households - 15% of the population). The expected benefits from the new approach are a considerable cost reduction (in the long term up to 10% of the traditional census costs). This would allow the investment of the savings in the improvement of census-like information, in increased timeliness of census data, in higher frequency of censuses, and additional surveys, including longitudinal surveys, utilizing the continuing system of administrative data.

A system is being developed for the evaluation of various sources of information and to combine the information to obtain reliable census-like estimates. The main components of the system include the evaluation of the administrative sources and their matching to obtain a joint administrative database. This is to be complemented by the design of samples for field enumeration, matching and estimation procedures to combine information and evaluation procedures. Estimation procedures to be developed, based on matching area sample results with administrative data base, will include synthetic estimation procedures to estimate small area characteristics and weighting for census and sample data for individuals and households to conform to small area estimates. Finally evaluation procedures will be designed for coverage evaluation, by area samples, and for content evaluation, via a re-interview for a sample of complete and incomplete records. The studies will provide conclusions regarding the continuous use of administrative data for current demographic estimates and other post-censal longitudinal data.

3. Combining longitudinal data from administrative sources and surveys

The expected situation following the census is that complete unit-level data will be available for all census variables linked with the administrative data, via the unique ID, by means of updating procedures for the major administrative systems. In addition, one or more longitudinal panel surveys are planned to be operative from the census onwards (possibly starting before the census). These will be linked with the combined census file and updated administrative systems. The longitudinal survey will, most probably, be designed as a rotating annual panel survey with a continuous inclusion period of 5-6 years (similar to the Canadian SLID). Thus at each point in time data for sampled units would be available for the current year, for a continuous period of previous years and for the census year. These would be supplemented by administrative data (for selected variables) for all non-sampled units and periods. The basic idea for continuing analysis of the longitudinal data is to combine sample data for the current period and for previous periods with administrative data for sampled and non-sampled units to obtain small-area estimates for the whole population. The method would basically be that of model-based "mass imputation" (data fusion), proposed by Colledge et al. (1978). The proposed method assumes a stable population over time, (with separate treatment for new entrants), the availability of updated administrative data for all units at each time and of the panel survey data for continuous series of time points (possibly with some gaps).

The basic model we propose for imputation is based on the combination of a hierarchical mixed model at each point in time and a time series model over points of time. At each point of time, *t*, the hierarchical model is:

$$y_{hjt} = \mathbf{x}'_{hjt}\mathbf{b}_t + \mathbf{z}'_{ht}\mathbf{v}_t + \mathbf{z}'_{ht}\mathbf{u}_{ht} + e_{hjt},$$

where y_{hjt} is the value of the response variable at time *t*, for individual *j*, belonging to household *h*, $\mathbf{x}_{hjt}$ and $\mathbf{z}_{ht}$ are vectors of individual and household level explanatory variables, respectively; $\mathbf{b}_t$ and $\mathbf{v}_t$ are fixed

vector coefficients; and $\mathbf{u}_{ht}$ and e_{hjt} are household and individual level random effects , respectively. The individual and household level random errors are assumed to follow independent first order autoregressive models:

$$\mathbf{u}_{ht} = \mathbf{A}\mathbf{u}_{ht-1} + \mathbf{d}_{ht}; \quad \mathbf{d}_{ht} \sim \mathrm{N}(\mathbf{0}_q, \mathbf{D}); \quad e_{hjt} = \rho e_{hjt-1} + \varepsilon_{hjt}; \quad \varepsilon_{hjt} \sim \mathrm{N}(0, \sigma_e^2).$$

The model can be defined in terms of a state-space model and the random components can be predicted by the application of a Kalman filter to provide imputed values for missing data, i.e for units for which only administrative data is available. Full details on the model and estimation of its parameters are given in Pfeffermann and Nathan (2001), as well as other imputation methods, which could be considered for the imputation of non-sampled units - mean imputation, nearest neighbour, and simple and augmented regression.

An initial small simulation study was carried out to test the feasibility of the use of these methods. The study was based on 100 samples of size 1000 households each generated from the model with household sizes randomly allocated as two or three, data assumed available for four time points, random effects selected from uniform distributions and arbitrary values of fixed parameters. Bernoulli sampling at the rate of 20% was applied and the remaining values imputed by each of the methods proposed. These were compared with the true nonsampled values to estimate average bias and root mean square error (RMSE) for each of ten arbitrary equal-sized "small areas" (of about 100 households each). The following table gives the averages (over simulations) of the mean, the minimum and the maximum over the ten "small areas".

		BIAS			RMSE		
No.	Imputation method	Mean	Min.	Max.	Mean	Min.	Max.
1	Mean imputation	-0.07	-2.36	2.20	18.76	17.50	20.19
2	Nearest neighbor	0.19	-2.10	2.37	21.94	20.57	23.35
3	Simple regression	-0.07	-2.14	1.95	15.15	13.82	16.54
4	Augmented regression	0.09	-2.03	2.16	15.29	14.00	16.64
5	State space	0.32	-2.79	3.27	15.43	13.34	17.62

The results from such limited simulation runs are not conclusive. Although the biases are similar, the model-based imputation methods (regression and state-space) have smaller mean square errors. It is obvious that considerable efforts will have to be invested in searching for suitable models and methods of imputation.

The imputation method selected will be used to create 'complete' population files for each time period to include the data from the current administrative files, supplementary data for sampled units and imputed supplementary data for non-sampled units. Derived statistics will include small area cross-sectional statistics, longitudinal gross-flow data and case history data. A variety of problems still remain to be dealt with. These include the treatment of changes in the population (primarily via the administrative data), the fitting of suitable models and their evaluation and the study of the robustness to departures from the models. Finally attention will have to be given to variance estimation and to the evaluation of non-sampling errors – due to non-response and attrition, rotation group biases etc.

REFERENCES

Binder, D. A. (1998). Longitudinal surveys: why are these surveys different from all other surveys *Survey Methodology* **24**, 101-108.

Brown, J. J., Diamond, I. D., Chambers, R. L., Buckner, L. J. and Teague, A. D. (1999). A methodological strategy for a one-number census in the UK. *Journal of the Royal Statistical Society*, **A 162**, 247-267.

Colledge, M. J., Johnson, J. H., Pare, R. and Sande, I. G. (1978). Large scale imputation of survey data. *American Statistical Association, Proceedings of the Section on Survey Research Methods*, 431-436

Pfeffermann, D. and Nathan, G. (2001). Imputation for wave nonresponse - existing methods and a time series approach. **In:** *Survey Nonresponse*. (eds. R. Groves, D. Dillman, J. Eltinge, and R. Little), Chap. 28. New-York: Wiley.(in press).

Scheuren, F. (1999). Administrative records and census taking. *Survey Methodology*, **25**, 151-160.

SUMMARY

The Israel 2006 population census will be based primarily on administrative systems and completed by a set of supplementary sample surveys and evaluation studies, followed up by longitudinal surveys, all linked with a high degree of accuracy via a unique personal identity number. We consider the possible use of a combination of time series methods with hierarchical modelling, in order to utilize efficiently all available information. The models will be used to impute data for a current point in time for units for which administrative data are available but which are not included in the panel surveys.

RÉSUMÉ

Le recensement israélien de la population de 2006 se basera surtout sur les données venant des fichiers administratifs. Une série d'enquêtes par sondage pour compléter et évaluer les données administratives sera suivie par des enquêtes longitudinales. Toutes ces données seront liées par le numéro personnel d'identité. Nous considérons l'emploi des modèles hiérarchiques multi-niveaux en combinaison avec les méthodes d'analyse des séries temporelles. Les modèles seront employés pour l'imputation des données des individus pour lesquels on a des données administratives, mais qui ne sont pas inclus dans les enquêtes longitudinales.