MODELING CENSUS MAILBACK QUESTIONNAIRES, ADMINISTRATIVE RECORDS, AND SAMPLED NONRESPONSE FOLLOWUP, TO IMPUTE CENSUS NONRESPONDENTS

Elaine Zanutto, Harvard University and Alan M. Zaslavsky, Harvard University
Elaine Zanutto, Department of Statistics, 1 Oxford St., Cambridge MA 02138

KEY WORDS: Loglinear Models, Iterative Proportional Fitting, 1995 Census Test

1. Introduction

The use of sampling for nonresponse followup (NRFU) in Census 2000 will create an unprecedented amount of missing data. Therefore, it is important to synthesize all available information to estimate the complete roster with acceptable accuracy. In particular, administrative records are a relatively inexpensive source of detailed information. However, they differ systematically in coverage, content, and reference period from the census, so simply replacing nonresponding households with administrative records may introduce biases into the completed roster. To complete the roster, we propose fitting a hierarchical loglinear model to model characteristics of nonsample nonresponding households using low-dimensional covariates at the block level and more detailed covariates at more aggregated levels. Model estimates are then used to impute the characteristics of households at nonsample nonresponding addresses. We incorporate administrative records in this estimation and imputation method using data from sampled NRFU to correct for systematic differences between the information sources. We evaluate our methods through simulations using data from the 1995 Census Test.

2. Background

Several methods have been proposed for completing the census roster when NRFU is conducted in only a sample of blocks (Fuller, Isaki, and Tsay 1994, Schafer 1995, Zanutto and Zaslavsky 1995a,b). Recently, Zanutto and Zaslavsky (1996a,b) extended this list of papers by considering estimation when one of the data sources is a file of administrative records and when a housing unit sample design is used for NRFU. Zanutto and Zaslavsky (1996b, henceforth "ZZ") evaluate this potential use of administrative records using data from the 1995 Census Test and a preliminary version of the corresponding administrative records database, the "Phase I" database. Through simulation, they found that the RMSE of block level estimates of the number of households with various characteristics was smaller using their method of incorporating administrative records than using a comparable method that ignores all administrative records. They also found that using administrative records in a statistical model is the key to estimates with reduced RMSE. Directly substituting all available administrative records for the corresponding nonrespondents can result in estimates with very large biases. Our current work confirms these findings.

In this paper, we apply the methods of ZZ to data from the 1995 Census Test and an updated version of the corresponding administrative records database, the "Phase II" database. A limitation of the Phase I database was that the records were not grouped into households. To obtain preliminary results, ZZ artificially grouped the administrative records into households using a match to census records based on name, sex, and date of birth. Each administrative record that could be matched to a census record was assigned the same housing unit identification (huid) number as the census record to which it was matched. Any administrative records that could not be matched to census records were discarded. As a result of this matching process, the administrative records used in ZZ more accurately match the characteristics of the census households than can be expected in practice.

Both the Phase I and Phase II administrative records databases were built by combining records from federal, state, and local files. Files from commercial vendors were also used for the Phase I database. To form the final database, records from all sources were combined into one master file that was then unduplicated with the goal of having no more than one administrative record per person. The Phase II processing used an improved unduplication algorithm. Finally, during Phase II, administrative records were assigned huid numbers using the same algorithm applied to census records (Wurdeman and Pistiner 1997). The resulting Phase II database contains information about address (District Office (DO), tract, block, huid), sex, race, Hispanic origin, date of birth, and marital status. More details about the administrative records sources and the unduplication process appear in Wurdeman and Pistiner (1997), and Neugebauer, Perkins, and Whitford (1996).

The results of ZZ show that administrative records have the potential to contribute to the accuracy of estimates at very detailed levels of geography. The question addressed by this research is whether or not the administrative records that are available in practice (e.g. the Phase II database) are useful in improving the estimates.

3. General Estimation and Imputation Procedure

Data collection under sampling for NRFU occurs in two stages. At the first stage, census data are collected by mailout-mailback questionnaires. At the second stage, followup (field or telephone) is carried out for a sample of the nonresponse cases from the first stage. The followup determines whether a housing unit physically exists at the address, and if so, collects data about the unit and any residents. Because the characteristics of nonrespondent households that are not in the followup sample remain unknown after the two stages of data collection, the census roster is completed by imputing the characteristics of these nonsample nonrespondents.

The general framework of the estimation and imputation procedure assumed in this paper is as follows:

1. Sample from housing units that did not respond to the census mailout questionnaire.
2. Classify households into a small number of "types".
3. Estimate a vacancy model (using logistic regression) to estimate the number of nonsample nonrespondent housing units that are vacant in each block.
4. Estimate a household type model (using a loglinear model) to estimate the number of nonsample nonrespondent nonvacant households that are of each type in each block.
5. Round the estimated counts of the number of nonrespondent households of each type to obtain integer counts
6. Impute households for the nonsample nonrespondent households according to the estimated (rounded) counts.

Because it is difficult to model, simultaneously, all of the household characteristics of interest, Step 2 above classifies households into a small number of "types". We use 18 types based on a cross-classification by race of the household (Black, non-Black Hispanic, Other), number of adults in the household (0-1 adults, 2 adults, 3 or more adults), and number of children in the household (0, 1 or more). The imputations in Step 6 fill in values for nonrespondent household characteristics that are not explicitly modeled in Step 4. This results in a completed roster that is suitable for preparing tabulations or microdata samples.

The remainder of this discussion compares alternative models for Step 4. We explore, through simulations, the gains in accuracy that are possible by incorporating information from administrative records into the model. Because the primary goal of this research is to evaluate the performance of the household type model, all vacant households are deleted from the simulation data sets, thus eliminating the need for Step 3. The rounding and imputation phases in Steps 5 and 6 are also omitted.

4. Estimation Model

As in ZZ, we fit a hierarchical loglinear model to model the characteristics of nonsample nonresponding households using low-dimensional covariates at the block level and more detailed covariates at more aggregated levels of geography. Specifically, to estimate the number of nonsample nonrespondent households of each type in each block, we use a model of the following form, expressed in the standard generalized linear models notation of Wilkinson and Rogers (1973):

$$\log \mathbf{E}n(i,j,r) \sim i+r+i*r+r*x_3+i*x_2+a*r*x_1. \quad (1)$$

The left hand side is the logarithm of the expected number of households in block i, of household type j and response status (or data source) r. The right hand side represents a linear predictor determined by the block index i, response status or data source indicator r, tract indicator $a=a(i)$, and $x_1=x_1(j)$, $x_2=x_2(j)$, and $x_3=x_3(j)$ which are categorical variables for classifications of household types that are based on the categories for household type j (e.g. x_2=race). More generally, x_1, x_2, and x_3 can be model expressions in the variables that define household type. For example, x_2=race×adults+children results in separate block×race×adults and block×children interactions in the model through the $i * x_2$ term.

This model can be used to estimate the number of nonsample nonrespondent households of each type in each block using respondents as predictors (ignoring administrative records), or using administrative records for nonrespondent households as predictors (ignoring respondents). Therefore, we can use this model with r representing response status (respondent, nonrespondent) or representing data source (census, administrative record).

The x_1, x_2, and x_3 terms allow us to model detailed household types at large levels of geography, such as the tract or DO levels, and more aggregated household types at smaller levels of geography, such as the block level. In particular, including the $i * x_2$ term represents the fact that respondents and nonrespondents in the same block are similar in the characteristics represented by x_2. This feature of the model is the essential difference from the Fuller, Isaki, Tsay (1994) method.

Our model is motivated by the following principle of maximum likelihood estimation in loglinear models: In a hierarchical loglinear model (i.e. one in which for every interaction effect, all main effects or interactions marginal to it are also included in the model), the expected values for every margin corresponding to an effect in the model are equal to the corresponding observed margins. Therefore, since each of the terms in this model can be interpreted as a margin of the block×type×r table, if we fit the model by maximum likelihood, the estimated values for these margins will match those observed in the data.

Under sampling for NRFU, however, not all margins of the block×type×response table are fully observed. Specifically, we have information for all responding households but for nonresponding households only in the NRFU sample. Therefore, when the NRFU sample is a housing unit sample, we weight up the sample households to obtain unbiased estimates of the margins involving nonrespondents. These margins are then treated as observed and used in an iterative proportional fitting (IPF) algorithm to fit the model. Further details about fitting this model, including the case where the NRFU sample consists of all nonresponding units in a sample of blocks (i.e. a cluster sample of housing units rather than an unclustered housing unit sample) can be found in ZZ and Zanutto (1997).

5. Modeling Strategies

In this section, we describe our proposed modeling strategy for Step 4 of Section 3. For comparison, we also describe two alternative strategies. All three estimation methods allow for the fact that, in practice, many households are not represented in the administrative records database. In each method, tract and DO level estimates are formed by aggregating block level estimates.

The following modeling strategy uses administrative records through statistical modeling:

1. Group nonrespondent households into those with and without administrative records.
2. To estimate the household types of the nonsample nonrespondent households that have administrative records, fit loglinear model (1) using the available administrative records for nonrespondents and any corresponding census records from the followup sample (e.g. r=census, administrative records).
3. To estimate the household types of nonsample nonrespondent households without administrative records, fit loglinear model (1) using all census records from respondent households, and census records from the NRFU sample for households that do not have administrative records (e.g. r=respondent, nonrespondent).

Combining the estimates from Steps 2 and 3 gives estimates for all nonsample nonrespondents. We call this the "two model" method.

An alternative strategy is to naively substitute administrative records, whenever possible, for census nonrespondents not in the NRFU sample. In this method, if a nonsample nonrespondent household has an administrative record, it is substituted for the missing census record. The number of households of each type among the remaining nonsample nonrespondent households that do not have administrative records is estimated using loglinear model (1) with respondents as predictors (e.g. r=respondent, nonrespondent). We call this the "substitution" method.

Another alternative strategy is to ignore all administrative records and fit loglinear model (1) using respondents to predict the number of nonsample nonrespondent households of each type in each block (e.g. r=respondent, nonrespondent, for the whole data set). We call this the "one model method".

6. Simulation Design

The goal of this study is to evaluate the bias, variance, and RMSE of the estimates of demographic aggregates (such as number of households by race, number of adults, and number of children) at the block, tract, and DO levels, using estimated household compositions for nonsample nonresponding addresses. Because it is not feasible to answer these questions analytically, we approach these evaluations through simulation.

Using data for which we know the characteristics of all respondents and nonrespondents (described in Section 7), the steps of the simulation are as follows:

1. Simulate NRFU sampling by selecting a 1 in 3 sample of nonrespondent households in each tract using simple random sampling.
2. Fit the model(s).
3. Estimate the number of nonsample nonrespondent households of each type in each block.
4. Compare estimates to the truth.

These steps are repeated 30 times for each estimation method. This yields sufficiently accurate estimates of Root Mean Weighted Root Mean Square Error (RMSE), Root Mean Weighted Squared Bias, and Root Mean Weighted Variance, which we calculate using the formulas from Zanutto and Zaslavsky (1995b, 1996). These measures have several desirable properties as described in Zanutto and Zaslavsky (1995b, 1996).

In these simulations, all loglinear models use x_2=race. Experimentation with several other specifications of x_2 did not result in estimates with smaller RMSE. All models also use x_1=household type, which leads to the $r * x_3$ being absorbed into the $a * r * x_1$ term.

7. The Data

Data from the 1995 Census Test and the corresponding Phase II administrative records database are used in these simulations. The Census Test occurred in three sites: Oakland, California; Paterson, New Jersey; and six parishes in northwest Louisiana. We focus on results from the Oakland site. Similar results were obtained for the Paterson site.

Our simulations use only a subset of the data from each site. Because sampling for NRFU was conducted in the 1995 Census Test, we know the actual characteristics only of those nonrespondents in the NRFU sample. Therefore, these are the only nonrespondents we can use to evaluate our estimation procedures. As a result, the subset of the data we use consists of all blocks containing nonrespondent households in the followup sample, i.e. all respondents in these blocks and all nonrespondents in the followup sample. This followup sample followed a block sampling design in Paterson and in half of Oakland, and a housing unit sampling design in the other half of Oakland. Overall, one-sixth of the nonresponding housing units in Paterson and two-sevenths of the nonresponding housing units in Oakland were selected for followup (Vacca, Mulry, and Killion 1996). Descriptions of the simulation populations from the two test sites broken down by size, nonresponse rate, and demographic characteristics are given are Table 1.

Figure 1 compares the distributions of the basic household characteristics in the administrative records and the census NRFU sample, where both are available. In Oakland, 50.9% of the nonrespondents have administrative records and in Paterson, 21.5% do. (Only administrative records that contain complete address and race information are counted in these percentages.) Figure 1 shows that in the Oakland simulation data set the distribution of households in each of the three race categories in the administrative records agrees with the distribution in the census data, but in the Paterson data, the administrative records slightly understate the number of Black and Hispanic households. In both data sets, the administrative records severely understate the number of households with children. Also, in Oakland, the number of households with 3 or more adults is severely overstated in the administrative records, and in Paterson the number of households with 0-1 adults is overstated in the administrative records. Closer examination of the data for Oakland reveals that the administrative records contain many out-of-date records. This results, in many cases, in the current residents being listed at an address in the administrative records as well as the previous occupants thereby overstating the number of households with 3 or more adults. The Paterson situation is also not unusual. Because administrative records often do not contain information for all members of the households, many people are omitted from the administrative records database and hence the number of people in a household can be understated.

Agreement rates between the administrative record and census household type classification for nonrespondents, where both records are available, were also tabulated. The agreement rates for Oakland and Paterson are, respectively 29.9% and 24.7% agreement on household type, 84.0% and 74.8% agreement on race, 42.7% and 48.9% on adult category, and 77.0% and 56.0% agreement on children category.

8. Simulation Results

Simulation results for the Oakland site are shown in Figure 2. The three bar charts in this figure show the RMSE for the estimates of the total number of households in each of the race, adult, and children categories at each of the block, tract, and DO levels of geography. (Only the zero children category is shown since the results for the 1+ children category are identical.) The height of the bar represents the percent RMSE, and this percent is also printed at the top of each bar. All three charts are on the same scale. The three shaded bars represent the three estimation methods, as indicated by the legend.

The results for the block level estimates show that the substitution method performs well for estimates for the race categories, but results in block level estimates with large RMSE for the children and adult categories. These adult and children categories are critical because they determine total population. The results are even more dramatic at the tract and DO levels, where it is clear that substitution produces estimates with much larger RMSE that the other two estimation methods. These large RMSEs are due to a large bias component that results from the biases in the administrative records seen in Figure 1.

Figure 2 also shows that the one and two model methods both produce estimates with smaller RMSE than the substitution method for all household characteristics, except Black, at all levels of geography. For this reason, the remainder of our comparison will focus on the one and two model methods. A comparison of these two methods must be limited to block level estimates only, because the fitting algorithm for the loglinear models constrains the tract level estimates to equal their unbiased estimates from the NRFU sample. Therefore, the one and two model methods produce the same estimates at the tract and DO levels. They can, however, differ at the block level. At the block level the two model method produces estimates with smaller RMSE than the one model method for the race categories. (These differences are significant with $p < .0001$.) This smaller RMSE is due to a smaller bias component. Both methods produce block level estimates with comparable RMSE for the children and adult categories.

In Figure 2, the differences in RMSE for block level race estimates between the one and two model methods may appear small. However, our measures of RMSE, bias, and standard deviation are based on the difference between the estimated total number of households of a given type and the truth, relative to the total number of households in the area (block, tract, or DO). The estimated total number of households of a given type in an area is the sum of the number of respondent households of that type, the nonrespondent households in the NRFU sample of that type, and the estimated number of nonsample nonrespondent households of that type. Therefore, in the Oakland data, since the nonresponse rate is only 19.3% and of these nonrespondents only 50.9% have administrative records, the administrative records affect only 9.8% of the households used in these calculations. If more nonrespondents had administrative records, the difference between the two methods would be larger. Also, if these measures were calculated based only on nonrespondents, the difference between the two methods would be easier to see.

9. Conclusions

This work confirms the previous findings of ZZ that administrative records contribute to accuracy at very detailed levels of geography, such as the block level, and that the key to obtaining estimates with reduced RMSE is using administrative records in a statistical model. Direct substitution of administrative records for nonrespondents can lead to estimates with large biases. Despite the small benefit from using administrative records with these data, this work confirms the potential usefulness of administrative records even though they contain imperfect information. More research is needed into possible uses of administrative in other census and general survey situations.

References

Fuller, W.A., Isaki, C.T. and Tsay, J.T. (1994), "Design and Estimation for Samples of Census Nonresponse," *Proceedings, Bureau of the Census Annual Research Conference*, 10:289-305.

Neugebauer, S., Perkins R.C., and Whitford, D.C. (1996), "First Stage Evaluations of the 1995 Census Test Administrative Records Database," DMD 1995 Census Test Results Memorandum Series, No. 41, March 14, 1996, United States Bureau of the Census.

Schafer, J.L. (1995), "Model-Based Imputation of Census Short-Form Items," *Proceedings, Bureau of the Census Annual Research Conference*, 11:267-299.

Vacca, E.A., Mulry, M., and Killion R.A. (1996) "The 1995 Census Test: A Compilation of Results and Decisions," DMD 1995 Census Test Results Memorandum No. 46, April 1, 1996, United States Bureau of the Census.

Wilkinson, G.N. and Rogers, C.E. (1973), "Symbolic description of factorial models for analysis of variance," *Applied Statistics*, 22:392-9.

Wurdeman, K. and Pistiner A.L. (1997), "1995 Administrative Records Evaluation–Phase II," DMD 1995 Census Test Results Memorandum Series, No. 54, Revised, March 26, 1997, U.S. Bureau of the Census.

Zanutto, E. (1997), "Models for Imputing Nonsample Households with Sampled Nonresponse Followup," qualifying paper, Harvard Univ. Dept. of Statistics.

Zanutto, E. and Zaslavsky, A.M. (1995a), "Models for Imputing Nonsample Households with Sampled Nonresponse Followup," *Proceedings, Bureau of the Census Annual Research Conference*, 11:673-686.

Zanutto, E. and Zaslavsky, A.M. (1995b), "A Model for Imputing Nonsample Households with Sampled Nonresponse Followup," *Proceedings, Section on Survey Research Methods, American Statistical Association.*

Zanutto, E. and Zaslavsky, A.M. (1996a), "Estimating a Population Roster from an Incomplete Census using Mailback Questionnaires, Administrative Records, and Sampled Nonresponse Followup," *Proceedings, Bureau of the Census Annual Research Conference*, 12:741-760.

Zanutto, E. and Zaslavsky, A.M. (1996b), "Estimating a Population Roster from an Incomplete Census Using Mailback Questionnaires, Administrative Records, and Sampled Nonresponse Followup," *Proceedings, Section on Survey Research Methods, American Statistical Association.*

Test Site	CA	NJ
Number of Households	58387	11096
Number of Blocks	1803	292
Number of Tracts*	91	31
Nonresponse Rate	19.3%	49.8%
Hispanic Households	10.9%	35.8%
Black Households	36.3%	36.2%
Households of Race Other	52.7%	28.0%
Households with Children	30.9%	46.9%
Households without Children	69.1%	53.1%
Households with 0 or 1 Adults	44.3%	35.9%
Households with 2 Adults	41.4%	39.6%
Households with 3+ Adults	14.3%	24.5%
Households with Admin. Records	63.2%	28.0%

* There are actually 101 tracts in the CA site and 33 in the NJ site but several small tracts were combined to form larger tracts for the simulations.

Table 1: 1995 Census Test Site Summaries (for the subset of data used in simulations)

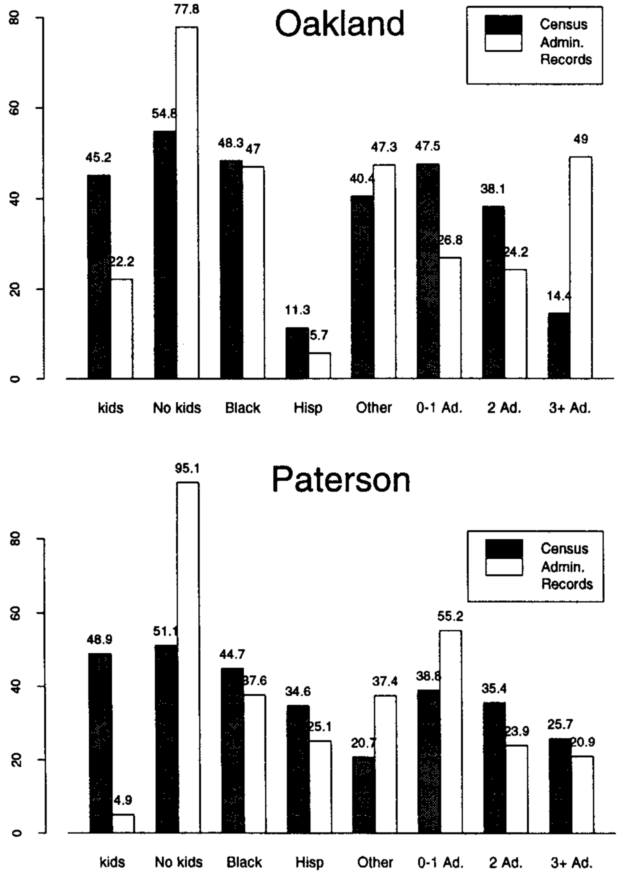

Figure 1: Prevalence of household characteristics in administrative records for nonrespondent households and in the corresponding census records

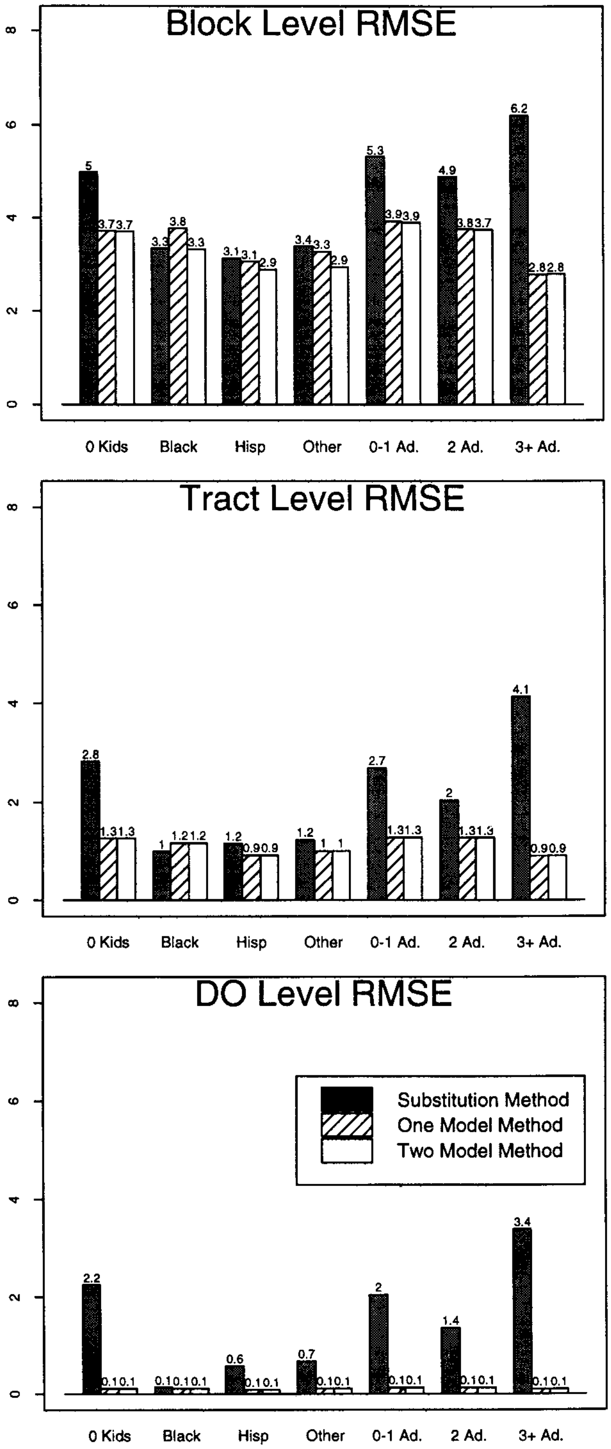

Figure 2: Root Mean Weighted Mean Squared Error (RMSE) at block, tract, and DO levels, as a percent of total number of households in each area.