ESTIMATING A POPULATION ROSTER FROM AN INCOMPLETE CENSUS USING MAILBACK QUESTIONNAIRES, ADMINISTRATIVE RECORDS, AND SAMPLED NONRESPONSE FOLLOWUP

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

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

1 Introduction

The feasibility of the traditional census process is challenged by two trends: increasing nonresponse rates to the mailed questionnaire, and increasing costs per household for field nonresponse followup (NRFU). These trends make it important to synthesize information that is incomplete or imperfect due to sampling, nonresponse, and/or the unreliability of some data sources, 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.

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 (1996) extended this list of papers by considering estimation when one of the data sources is a file of administrative records. This paper applies these methods to census data and administrative records from the 1995 Census Test, and extends this methodology to incorporate a housing unit sample design for NRFU sampling. Zanutto (1996) provides a more detailed description of this research.

2 General Estimation and Imputation Procedure

Data collection under NRFU sampling occurs in two stages. At the first stage, census data are collected by mailout-mailback census 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. The problem that this research addresses is that it is necessary to impute the characteristics of nonrespondent households that are not in the followup sample. Once the census roster is completed by imputation, all tabulations prepared from the completed roster are guaranteed to be consistent with each other.

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

1. Housing units not responding to the census mailout questionnaire are sampled according to a predetermined scheme.
2. A model is fit and predicted counts are calculated for each block.
3. Counts are rounded.
4. Households are imputed for the nonsample nonrespondents.
5. The completed rosters are used to prepare tabulations and microdata samples.

This research focuses on Step 2 of the above process, in order to explore, through simulations, the gains in accuracy that are possible by incorporating information from administrative records into the model.

Step 2 of the above framework can be further broken down into the following steps:

2.1 Classify households into a small number of "types".

2.2 Estimate a vacancy model (using logistic regression) to estimate the number of nonsample nonrespondent housing units that are vacant in each block.

2.3 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.

Because it is difficult to model, simultaneously, all of the household characteristics of interest, Step 2.1 above classifies households into a small number of "types". We use 18 types based on a cross-classification by race (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). Also, because the primary goal of this research is to evaluate the performance of the household type model, all vacant households in the census data sets used for simulations were deleted, thus eliminating the need for Step 2.2. The remainder of this discussion is concerned with comparison of alternative models for Step 2.3.

3 The Model

The loglinear model we use to estimate the number of nonsample nonrespondent households of each type in each block, using the standard generalized linear models notation of Wilkinson and Rogers (1973), is of the form

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

The left hand side is the logarithm of the expected count for block i, 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, ARA (Address Register Area) or 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 which are based on the categories for household type j. More generally, x_1, x_2, and x_3 can be model expressions in the variables that define household type.

As in Zanutto and Zaslavsky (1996), this model can be used to estimate the household types of nonsample nonrespondents using respondents as predictors, ignoring administrative records, or using administrative records for nonrespondent households as predictors. Therefore, r can represent response status (respondent, nonrespondent) or it can represent data source (census, administrative record).

The x_1, x_2, and x_3 terms in this model, allow us to model detailed household types at large levels of geography, such as the tract or District Office (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 incorporates the fact that respondents and nonrespondents in the same block are similar in the characteristic represented by x_2. This feature is the essential difference from the Fuller, Isaki, Tsay (1994) method.

This 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×response 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 which fits the model.

The key difference between a housing unit and block sampling design for NRFU is that with a block sample we have complete information (i.e. information about respondents and nonrespondents) for all blocks in the NRFU sample. With a housing unit sample, on the other hand, there are no blocks for which we have complete information, except for those blocks which, by chance, have all nonrespondents in the NRFU sample. With a block sample, the unobserved cells in the block×type×response table are not a problem because nonsample nonrespondents contribute to the likelihood only through the the total number of nonrespondents in each block. Therefore, to maximize this part of the likelihood we need only ensure that the fitted number of respondents in each block equals the observed number, which is automatic because the model includes a block by response $(i * r)$ interaction term. With a housing unit sample, however, because all of the cells involving nonrespondents are partially observed, it is necessary to form estimates of the margins involving nonrespondents by weighting up the NRFU sample.

Also, when fitting the model, we incorporate a small amount of empirical Bayes smoothing to ensure that the model can be fit in every case. This smoothing adds one respondent household (or administrative record household) to each block. This household is divided among the 18 household types according to the overall DO proportions of the respondents (or administrative records).

This model is designed to be used when the NRFU sample is a random sample of housing units within large geographical areas (e.g. tracts). If the NRFU sampling plan is modified to guarantee that the sample contains some nonrespondents from each block, then further research is required to determine what, if any, advantages estimates from this model have over direct estimates of blocks formed by simply weighting up the sampled units in each block.

4 Modeling Strategies

In this section, we describe our proposed modeling strategy for Step 2.3 of Section 2. We also describe two alternative strategies, for comparison. These estimation methods take into account the fact that, in practice, many households are not be represented in the administrative records database. In each method, tract and DO level estimates are formed by aggregating block level estimates.

We propose the following modeling strategy, which we call the "two model" method:

1. Divide 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.

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 household types of the remaining nonsample nonrespondent households are 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 completely ignore all administrative records. In this method we fit loglinear model (1) using respondents to predict the household types of all nonrespondents (e.g. r=respondent, nonrespondent, for the whole data set). We call this the "one model method".

5 Simulation Design

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

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.

In these simulations, models using administrative records to predict the characteristics of nonsample nonrespondents use x_2=(race+adults+children). All models using respondents to predict the characteristics of nonsample nonrespondents use x_2=race. All models use x_1=household type, which leads to the $r * x_3$ being absorbed into the $a * r * x_1$ term. Here, a represents pseudo-tracts formed by aggregating ARAs into groups of the same approximate size as tracts. For processing reasons, actual tract information was not used.

To evaluate the estimates for the nonsample nonrespondents, we calculate measures of overall error, error due to bias, and error due to variance. Specifically, we calculate Root Mean Weighted Root Mean Square Error, Root Mean Weighted Squared Bias, and Root Mean Weighted Variance using the formulas from Zanutto and Zaslavsky (1995b, 1996). These measures have several desirable properties as described in Zanutto and Zaslavsky (1995b, 1996).

6 The Data

Data from the 1995 Census Test and the corresponding administrative records database are used in these simulations. The 1995 Census Test was conducted in three locations: Oakland, California; Paterson, New Jersey; and six parishes in northwest Louisiana. The data from the Oakland and Paterson sites are used in these simulations.

Because sampling for nonresponse followup was conducted for the Census Test, only those nonrespondents in the followup sample are used in these simulations. This is because, since we only know the true characteristics of the nonrespondents in the followup sample, these are the only nonrespondents we can use to evaluate our estimation procedures. Specifically, the simulation populations consist of all blocks in the block followup sample, which was conducted in Paterson and in half of Oakland, and all blocks with housing units in the housing unit followup sample, which was conducted in the other half of Oakland. For processing reasons, the blocks included in Integrated Coverage Measurement (ICM) operations were excluded from the simulation populations. Descriptions of the simulation populations from the two test sites broken down by size, nonresponse rate, and demographic characteristics are given are Table 1.

The administrative records database for the 1995 Census Test consists of records from federal government files, state government files, local files, and files from commercial vendors. Records from all of these sources were combined into one master file which then underwent an unduplication process with the goal of having no more than one administrative record per person. The resulting database contains information about address (District Office, ARA, block), sex race, Hispanic origin, date of birth, age, marital status, and relationship to first person listed on the census questionnaire. More details about the administrative records sources and the unduplication process can be found in Neugebauer, Perkins, and Whitford (1996) and Leslie (1995).

A limitation of the processing is that the person level administrative records were not grouped into households, although this is required to carry out the modeling described in this paper. Administrative records can be grouped into households based on addresses. In fact, 1995 Census Test administrative records are being reprocessed to assign housing unit identification numbers to them. The reprocessed administrative records database, called the "Phase 2 database", will be used for future research.

For these simulations, however, administrative records were grouped into households using a match to census records on name, sex, and date of birth (Neugebauer, Perkins, and Whitford 1996), which was carried out for reasons independent of this research. Each administrative record that could be matched to a census record was assigned the same housing unit identification number as the census record to which it was matched. Any administrative records that could not be matched to census records could not be assigned housing unit identifiers and therefore could not be used in these simulations.

The results of the matching process are used only to group administrative records into households. The actual address information contained in each administrative record is used to place people in a blocks, tracts, and DOs. All modeling and all comparisons made between census and administrative records (such as those described in the next paragraph) are based on matching census and administrative records by address only.

Figure 1 compares the distributions of the various household characteristics in the administrative records and the census NRFU, where both are available. In Oakland, 30% of the nonrespondents have administrative records and in Paterson, 24% do. (Only administrative records that contain complete address and race information are counted in these percentages.) Figure 1 shows that the distribution of households in each of the three race categories is about the same in the census and administrative records for the Oakland simulation data set, but in the Paterson data, the administrative records understate the number of Black and Hispanic households. Also, in both data sets, the administrative records understate the number of households with children and overstate the number of households with 0-1 adults.

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, 53% and 35% agreement on household type, 96% and 72% agreement on race, 65% and 63% on adult category, and 83% and 78% agreement on children category.

7 Simulation Results

Some simulations results are shown in Figure 2. The three bar charts in this figure show the Root Mean Weighted Mean Squared Error (RMSE) for the estimates of the total number of households in each of the children (only the zero children category is shown since the results for the 1+ children category are identical), race, and adult categories at each of the block, tract, and DO levels of geography for the Oakland simulation data set. 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 results for each of the estimation methods are represented by the three shaded bars, as indicated by the legends.

The results for the block level estimates show that the substitution method performs well for estimates for the children and race categories, but results in block level estimates with large RMSE for the adult categories. The results are even more dramatic at the tract and DO levels, where it is clear that substitution produces much larger RMSE than the other two estimation methods. While not shown here, results from bias and standard deviation calculations show that the large RMSEs of these estimates are due to a large bias component. This bias 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 at the block, tract and DO levels for the adult categories, estimates with slightly larger RMSE for the children and race categories at the block level, and estimates with similar RMSE for the children and race categories at the tract and DO levels. A comparison of the one and two model 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. At the block level, however, the two model method produces estimates with smaller RMSE than the one model method, for all categories. This smaller RMSE is due to a smaller bias compared to the one model method.

In Figure 2, the differences between the results for the one and two model methods may appear to be small. It should be kept in mind, however, that only 30% of the nonrespondents in the Oakland simulation data set had administrative records. If more nonrespondents had administrative records, the difference between the two methods would be greater. Also, the measures of RMSE, bias, and standard deviation that we calculate 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). These calculations include both respondents and nonrespondents. If these measures were calculated based only on nonrespondents, the difference between the two methods would be easier to see.

Results for the Paterson data are not shown here, but are similar to the Oakland results. The substitution method performs well for block level estimates for the children categories, but has larger RMSE for the race and adult categories, compared to the two other estimation methods. Again, these results are more dramatic at the tract and DO levels where it is clear that the estimates from the substitution method have very large RMSE, due to a large bias in the administrative records, seen in Figure 1. Also, the two model method produces block level estimates with smaller RMSE and smaller bias in all categories compared to the one model method.

8 Conclusions

From these simulations, it can be seen that administrative records contribute to accuracy at very detailed levels of geography, such as the block level. These simulations also show that direct substitution of administrative records for nonresponses, as carried out in these simulations, introduces large biases into the estimates. While these results depends on such factors as the specific data sets, the bias and coverage of the administrative records, the sampling rates, the levels of aggregation used, and the choice of household type classification, we believe that these conclusions are generalizable.

9 Future Work

We plan to continue investigating these models in several ways. The most immediate plan is to reevaluate the performance of these estimation methods using the Phase 2 administrative records database for the 1995 Census Test.

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.

Leslie, T.F. (1995), "November 7, 1995 Debriefing Results, Automated Matching Evaluation," DMD 1995 Census Test Memorandum No. 22, February 1, 1995, United States Bureau of the Census.

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.

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

Zanutto, E. (1996), "Modeling Administrative Records for Estimation and Imputation of Census Nonrespondents under Sampling for Nonresponse Followup," unpublished report, United States Bureau of the Census.

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. (1996), "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.

Test Site	CA	NJ
Number of Households	57,760	10,672
Number of Blocks	1,826	280
Number of "Tracts"	36	6
Nonresponse Rate	20.3%	53.5%
Hispanic Households	8.7%	34.3%
Black Households	37.3%	37.3%
Households of Race Other	54.0%	28.4%
Households with Children	69.5%	46.0%
Households without Children	30.5%	54.0%
Households with 0 or 1 Adults	43.5%	33.2%
Households with 2 Adults	41.0%	38.5%
Households with 3+ Adults	15.5%	28.3%

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

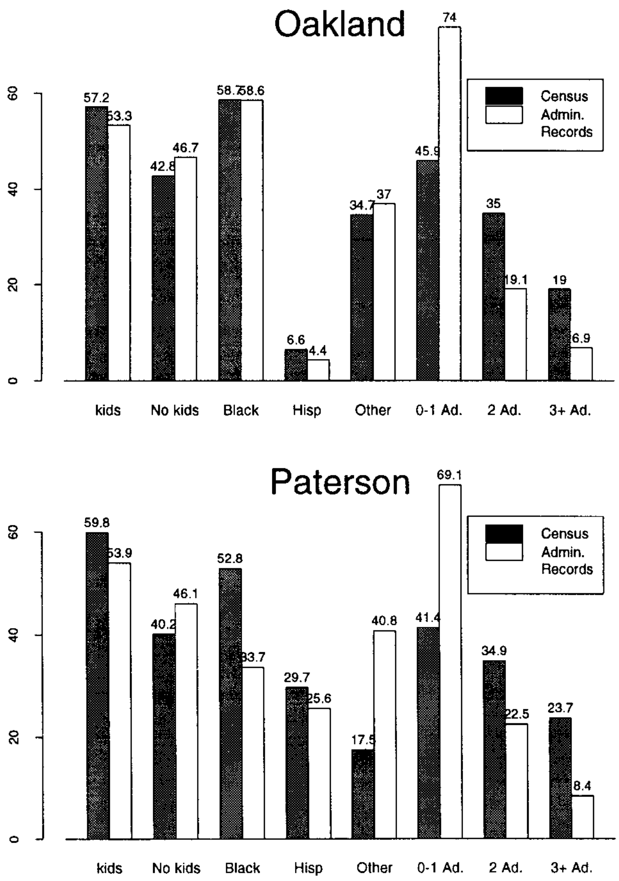

Figure 1: Prevalence of household characteristics in administrative records for nonrespondents households and in the corresponding census records

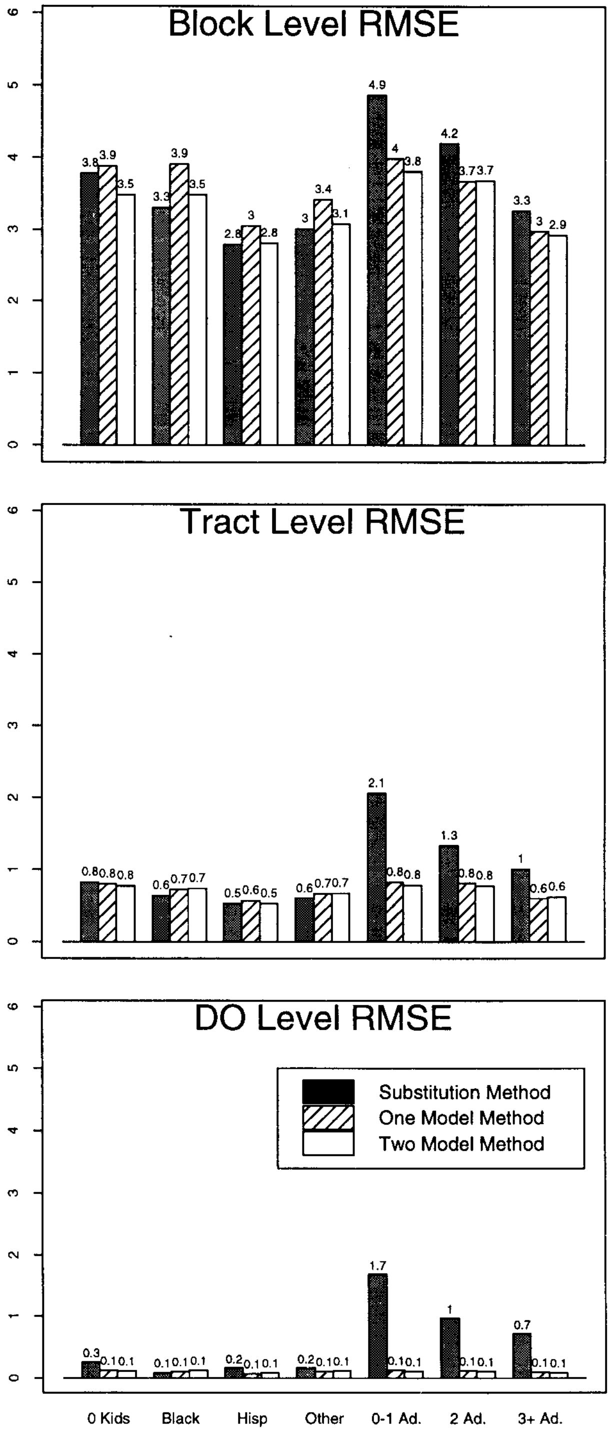

Figure 2: Root Mean Weighted Mean Squared Error (RMSE) at block, "tract", and DO levels, as a percent of total number of households in each area.