Journal of Official Statistics, Vol. 18, No. 4, 2002, pp. 559–576

Using Administrative Records to Improve Small Area Estimation: An Example from the U.S. Decennial Census

Elaine Zanutto[1] *and Alan Zaslavsky*[2]

We present a small area estimation strategy that combines two related information sources: census data and administrative records. Our methodology takes advantage of administrative records to help impute small area detail while constraining aggregate-level estimates to agree with unbiased survey estimates, without requiring the administrative records to be a perfect substitute for the missing survey information. We illustrate our method with data from the 1995 U.S. Decennial Test Census, in which nonresponse follow-up was conducted in only a sample of blocks, making small area estimation necessary. To produce a microdata file that may be used for a variety of analyses, we propose to treat the unsampled portion of the population as missing data and impute to complete the database. To do so, we estimate the number of nonrespondent households of each ''type'' (represented by a cross-classification of categorical variables) to be imputed in each small area. Donor households for these imputations can be chosen from the sampled nonresponse follow-up sample, the respondent households, or the administrative records households (if they are of sufficient quality). We show, through simulation, that our imputation method reduces the mean squared error for some small area (block-level) estimates compared to alternative methods.

Key words: Imputation; missing data; nonresponse follow-up; iterative proportional fitting; loglinear models; mass imputation.

1. Introduction

Small area estimation, that is estimation for small geographic areas or small subpopulations, is challenging because usually few or no units are sampled in some of the areas. The usual direct estimators, based only on data from units in the corresponding area, are likely to yield unacceptably large standard errors, if they are defined at all. Hence indirect estimators are needed that ''borrow strength'' by using data from other areas or auxiliary data (Ghosh and Rao 1994).

In the 1995 U.S. Decennial Test Census, small area estimation was necessary because nonresponse follow-up was conducted in only a sample of blocks, leaving the data incomplete in the remaining blocks. We present a small area estimation strategy applicable to this sample design that combines two information sources: census data and administrative records. Our goal is to produce a microdata file that may be used for a variety of analyses. To accomplish this, we treat the unsampled portion of the population as missing data and

[1] Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA, 19104, U.S.A. Email: zanutto@wharton.upenn.edu
[2] Department of Health Care Policy, Harvard Medical School, 180 Longwood Avenue, Boston, MA, 02115, U.S.A. Email: zaslavsky@hcp.med.harvard.edu

complete the database by imputation. (This resembles the use of ''mass imputation'' by Statistics Canada in surveys or censuses where some items are not collected for a large portion of the population in an effort to reduce the overall response burden (Whitridge, Bureau, and Kovar 1990; Royce, Hardy, and Beelen 1997; Rancourt and Hidiroglou 1998)). Our methodology uses administrative records as a covariate to help impute small-area detail while constraining aggregate-level estimates to agree with unbiased survey estimates.

Administrative records have a variety of forms and a broad range of potential statistical uses, with some distinct advantages and disadvantages (Zanutto and Zaslavsky 2002; Brackstone 1987). They are typically inexpensive to process in large-scale applications, relative to field data collection, and they can have excellent coverage for the part of the population to which they apply. Technology has improved our ability to link large administrative data sets to surveys and censuses. On the other hand, their content, coverage, accuracy, and reference periods, as well as the definitions of included variables, are determined by the needs of the program for which they are collected. Consequently, these data characteristics can differ from those that are desired for statistical purposes.

In this application, administrative records are available at the unit (here household) level for some, but not necessarily all, units and can be linked to the census data. We model the relationship between survey and administrative data for respondents; the estimated relationship can then be used to impute data for the nonrespondents based on their administrative records. The administrative records must be correlated with the survey responses of interest in order to improve estimation, but might contain systematic errors. The model corrects for such systematic differences between the two data systems.

More specifically, our methodology estimates the number of nonrespondent households of each ''type'' (represented by a cross-classification of categorical variables) to be imputed in each small area. Donor households for these imputations can be chosen from the sampled nonresponse follow-up sample, the respondent households, or the administrative records households (if they are of sufficient quality).

In Section 2, we describe our small area estimation strategy to improve block-level estimates when only some blocks are sampled. In Section 3 we show, through simulation, that our imputation method reduces the mean squared error for some small area estimates compared to alternative methods.

2. Using Administrative Records to Improve Block-Level Estimates in a Census with Sampled Nonresponse Follow-up

The ''traditional'' U.S. Decennial Census process (as practiced from 1970 to 2000) consists of a mailed questionnaire which is mailed back by respondents, followed by field nonresponse follow-up (NRFU) to collect information on households at nonresponding addresses. The feasibility of this approach is challenged by increasing nonresponse rates to the mailed questionnaire and increasing costs per household for NRFU. These trends, in the context of sharpened budgetary constraints, have driven the U.S. Census Bureau to consider alternatives that would permit creation of the necessary data products using less complete, and therefore less expensive, data collection methodologies. In particular, sampling for NRFU was a proposed innovation for census methodology for the year 2000.

In place of the traditional method of sending field enumerators to follow up on all households that did not respond to the census mailout questionnaire, this proposal required contact with only a random sample of the nonresponding households. Although sampling could reduce costs substantially, it would also create an unprecedented amount of missing data. Consequently, the characteristics of the nonrespondent households omitted from the follow-up sample would need to be estimated.

Although sampling for NRFU was not used in the 2000 Census, for legal reasons only marginally related to the statistical merits of the plan, our proposed methodology illustrates the use of administrative records to improve small area (block-level) estimation. Because census data have a variety of uses, accuracy at very detailed levels of geography is necessary so that estimates formed by aggregating block estimates are also accurate. A similar strategy could be used whenever auxiliary information is available for nonsample units and some respondents. Since auxiliary data like administrative records are unlikely to cover all households, we first demonstrate how estimation can be conducted for the households with such records. Then we describe, in Section 2.5, a strategy which uses a second model for the remaining households.

2.1. Background

The primary task of the U.S. Decennial Census is to create a full roster of the population of the United States, grouped into households (or allocated to non-household units) and with characteristics (notably age, sex, and race) attached. This roster is the basis for tabulations at the block level and successively more aggregated levels of geography (tracts, political divisions, and ultimately states) of totals and counts by person and household characteristics. (A block is a unit of census geography roughly corresponding to a city block or a compact rural area, averaging 15 households. A census tract is a neighborhood averaging about 140 contiguous blocks.) The goal of statistical methodologies in the census is to estimate this roster. The relevant criteria of the accuracy of these estimates, however, concern the accuracy of the various aggregates (including both tabulations and geographically aggregated microdata releases) that are prepared from them.

Fuller, Isaki, and Tsay (1994), Schafer (1995), Zanutto and Zaslavsky (1995a, 1995b; henceforth ''ZZ'') and Zanutto (1998) have proposed methods for completing the roster when NRFU is conducted in only a sample of blocks. These methods use information from census respondent households and from census nonrespondent households in the NRFU sample to impute the characteristics of census nonrespondent households that are not in the NRFU sample. We extend these methods by considering estimation when one of the data sources is a file of administrative records.

Research on estimation of the roster when NRFU is sampled has followed one of two basic strategies. Fuller, Isaki, and Tsay (1994) and ZZ pursue what might be called a ''top-down'' strategy, which starts with aggregates of households and subdivides them in a manner that maintains consistency with estimates calculated at the aggregate level. Simple ratio models or more complex loglinear (raking) models are used to estimate counts for small areas and detailed demographic groups, for which direct estimates are not possible. These ad hoc models do not describe the full complexity of the units, but

they are designed to maintain consistency of the aggregates which are considered most important.

Schafer (1995) develops a ''bottom-up'' strategy in which households are built up from individual persons and their characteristics and relationships, each of which must be described by its own model. This strategy gives a more complete and detailed description of the population, and if carried out successfully it can support full probability (e.g., Bayesian) inferences about its unobserved characteristics. However, this approach, unlike the other, requires that a fairly complex set of models be built before any imputations can be made. Furthermore, in this framework it is more problematic to maintain consistency between microdata and aggregate controls. Zaslavsky (1989, Part II) and Rubin and Zaslavsky (1989) also develop a model-based strategy for imputation of individual households, using a semiparametric approach to description of household types that is simpler than Schafer's.

Our approach extends the ''top-down'' approach to estimating the census roster. To describe our approach, we first briefly summarize all available data sources in Section 2.2. In Section 2.3 we outline a general strategy for combining data sources to estimate the characteristics of census nonrespondents. In Section 2.4 we describe our proposal to complete the census roster by fitting a hierarchical loglinear model to model characteristics of nonrespondent households that are not in the NRFU sample using low-dimensional covariates at the block level and more detailed covariates at more aggregated levels. We incorporate administrative records in this estimation process as covariates for predicting the characteristics of the corresponding nonrespondent households. Data from households in the NRFU sample for which we have both census and administrative records information are used to estimate the systematic differences between the two information sources. Model estimates can then be used to impute the characteristics of nonsample nonrespondent households. Section 2.5 describes two other estimation strategies that are evaluated, for comparison, along with our method in simulations summarized in Section 3.

2.2. *Data sources*

We assume the following data sources are available:

1. Responses are available for all respondents to the *mailout census*. (Mailback response rates are likely to be in the range 50–80%.) Responses to this form are the ''truth,'' in the sense that the definitions implicit in its completion are the standard for what is ultimately reported. In other words, our objective is to obtain data consistent with what would have been obtained with 100% mailback response.
2. Data obtained through *nonresponse follow-up* (NRFU) are available for selected households that did not respond to the mailout census. This sample will either be an unclustered unit sample, consisting of a sample of nonresponding households, or a block sample (cluster sample of units) consisting of all nonresponding households in a sample of blocks. Responses to NRFU, like mailback responses, are regarded as ''truth'' for the covered households.
3. *Administrative records* are available for all NRFU and nonsample nonrespondent households. These records have address information that makes possible a fairly close match to the census address list.

Census data have many uses, and both accuracy of demographic counts aggregated across broad areas and accuracy of geographically detailed counts are important. Our strategy builds on the strengths of each of the data sources. Geographically aggregated estimates are constrained to agree with unbiased estimates based on the relatively sparse sample, while local detail is completed using more detailed data sources, even if we have less confidence in their validity.

2.3. Outline of the estimation and imputation procedure

Because the characteristics of nonrespondent households that are not in the NRFU sample remain unknown after the two stages of data collection (the initial mailout questionnaire and NRFU), the census roster is completed by imputing the characteristics of these households. The estimation and imputation procedure assumed in this article moves from the coarsest description of the nonresponding housing units (whether they are occupied or not) to the finest (detailed composition).

1. *Vacancy model*: We fit a logistic regression model for the fraction of nonresponding addresses that are vacant, using data from the NRFU sample. Potential covariates include mailback rate, characteristics of households that responded by mail, and characteristics of the block as a whole and of nonresponding households in particular as set forth in administrative records. The vacancy model is kept separate from the model for household types in the next step because administrative records do not indicate whether a housing unit is actually vacant, only that there are no data for it in our record system. Since many nonvacant households are not reflected in our administrative record system, lack of a record does not predict vacancy in the same way that the characteristics of an included record predict the characteristics of a resident household. Furthermore, vacant housing units do not fit the structure of our model for household types, which is based on a cross-classification of characteristics.
2. *Household type model*: We classify households into ''types,'' because it is difficult to model, simultaneously, all of the household characteristics of interest. In our simulations, we define 18 household types by cross-classifying race of the household (Black, non-Black, Hispanic, Other), number of adults (0–1, 2, or 3 or more adults), and whether or not children are present in the household. Calculating imputed counts by household types involves three substeps:
   (a) *Model fitting*: Fitting a loglinear model for the prevalence of the various types of households, using data from the NRFU sample and administrative records.
   (b) *Model predictions*: Calculating predictions under the model for nonsample blocks. These predictions give numbers of households by type for each block.
   (c) *Rounding*: Rounding the (noninteger) counts predicted by models, using an unbiased controlled rounding algorithm. ''Unbiasedness'' here means that the rounding algorithm is stochastic, and expected rounded counts by block and household type are equal to model predictions. The rounding algorithm is ''controlled' in the sense that certain aggregates in the rounded table agree (within one unit) with the corresponding aggregates before rounding. It would be desirable if all the control totals in the models (Steps 1 and 2) were also

controlled in rounding, but this may be beyond the capabilities of present algorithms (Cox 1987; Fischetti and Salazar-González 1998).

3. *Imputation*: Impute household characteristics for nonrespondent households according to the rounded counts in Step 2c. Donor households for these imputations can be chosen from the sampled NRFU households, the respondent households, the administrative records households, or a combination of these sources. The imputations fill in values of nonrespondent household characteristics that are not explicitly modeled in Steps 1 and 2.

The product of this process is a roster in which every address either is listed as vacant or contains a household that mailed back a form, was interviewed in NRFU, or was imputed. This completed roster is suitable for preparing tabulations or microdata samples.

This modeling strategy relies on special properties of the logistic and loglinear models under maximum likelihood estimation. The logistic regression model of Step 1 has the property that the predicted rate of vacants under the model is equal to that observed, averaged across the NRFU sample blocks (as a whole, or in any area for which there is an indicator variable in the model). The loglinear model of Step 2 has the property that, provided it is a hierarchical loglinear model (i.e., one in which for every interaction effect, all main effects or interactions marginal to it are also included), the expected values for every margin corresponding to an effect in the model are equal to the corresponding observed margins (Birch 1963). Because model predictions for the included effects are constrained to agree with observed rates based on a probability sample, the corresponding estimates have very little bias. (Exact unbiasedness is not obtained because of the nonlinearity of the prediction model and because there may be a correlation between the number of housing units for which predictions are made in a block, i.e., nonsample nonrespondents, and some characteristics of the nonresponding households in the block.) The remaining steps are designed to maintain these properties as much as possible while completing the required detail in the roster.

The model used at Step 2 differs from those in ZZ in that information from administrative records, instead of information from mail respondents, is used to predict the characteristics of the households in those units. The administrative record information has a qualitatively different relationship to the ''truth'' for the nonresponding households than does the respondent information used in ZZ, because the former at least purports to tell us about the actual households for which we are making imputations, while the latter only tells us about the households by describing the general characteristics of the block. We must still use models to correct the administrative records, because of the records' known biases, but the variability of the differences between administrative records and the truth should be much smaller than with respondent data.

To use administrative records in estimation, we specify a joint loglinear model for nonrespondent households and the corresponding administrative records, fitted to the table whose dimensions are geographical area (down to the block level), household type (itself a cross-classification of several variables), and record source (census or administrative). The model is designed so that all block-level parameters (interactions between characteristics or mailback response and block) can be estimated even in the case of a block sampling NRFU design which results in a lack of NRFU information for some

blocks. This ensures that predictions can be made for blocks not in the NRFU sample. Heuristically, the distribution of household types observed in NRFU households in the surrounding area is shifted, using administrative record information, to predict the distribution of types among nonresponding households in the target block. Another way of looking at the same process is that the administrative characteristics of nonsampled nonrespondents in the target block are shifted, using information about differences between census and administrative data in the NRFU sample, to predict characteristics of the corresponding group as they would have been measured in the census. Such a model is appropriate if administrative records are relatively complete and accurate, with fairly consistent biases of coverage and content across the estimation area.

2.4. Estimation model

We use a model of the following form for estimation, fitted to data from one large area:

$$\log \mathbf{E}(n_{ijd}) \sim x_1 + i * x_2 + i * d + d * x_3 + d * a * x_4 \tag{1}$$

In the standard generalized linear models notation of Wilkinson and Rogers (1973), the ''*'' operator indicates that the main effects and all interactions that are marginal to the given interaction are included in the model. The left-hand side of (1) is the logarithm of the expected number of nonrespondent households in block i of household type j, according to data source d (NRFU census response or administrative record). The linear predictor on the right-hand side is determined by the block index i, data source indicator d, tract index $a = a(i)$, and categorical variables $x_1 = x_1(j)$, $x_2 = x_2(j)$, $x_3 = x_3(j)$, and $x_4 = x_4(j)$ that group the household types (e.g., $x_2 =$ race). More generally, x_1, x_2, x_3, and x_4 can be model expressions in the variables that define household type, in our case $j =$ race $\times$ children $\times$ adults. 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 administrative records for nonrespondent households as predictors of the nonrespondents (ignoring respondents). Estimates of the number of nonsample nonrespondent households of each type in each block depend on the characteristics of those households, as described by their administrative records, and the characteristics of nonrespondents in the NRFU sample in the same tract, as measured by the NRFU.

The x_1, x_2, x_3, and x_4 terms allow us to model detailed household types at large levels of geography, such as the tract or ''site'' (the overall area in which the estimation is carried out) 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 sampled nonrespondents and their administrative records, within the same block, are similar in the characteristics represented by x_2. This term is the essential difference from the Fuller, Isaki, and Tsay (1994) method, which can be regarded as a special case of our approach. ZZ show that our loglinear model approach results in estimates with smaller MSE compared to the stratified ratio approach of Fuller, Isaki, and Tsay (1994).

The rationale for this specification of the model is that all interactions are potentially included except any interaction of the form $d * i * x$, where x represents a model expression in the variables that define household type. Interactions of this form depend on the

margins determined only by nonrespondent households in a single block, and these are unavailable in nonsample blocks under the block sample design, and based on a very small sample under the unit sampling design. This model generalizes two simple theories that are contained as submodels. If there are no differences between administrative records for nonrespondents and their census (NRFU) records, then interactions with d are zero and nonrespondents are imputed in the same proportions as in the administrative records. Also, if there are no block effects then interactions with i and a are zero and nonrespondent households are imputed in the same proportions in each block based on the proportion of nonrespondent households in the NRFU sample in each of the x_3 categories.

Because not all margins of the block $\times$ type $\times$ source ($i \times j \times d$) table are fully observed under sampling for NRFU, to fit the model, assuming the NRFU sample is a housing unit sample, we weight the sampled nonrespondent households by their inverse probabilities of selection to obtain unbiased estimates of the census $\times$ tract $\times$ type margin for nonsample nonrespondents; these estimated margins are then used in a modified iterative proportional fitting (IPF) algorithm to fit the model.

The standard IPF algorithm (Darroch and Ratcliff 1972; Csiszár 1989; Fienberg and Meyer 1983) successively adjusts fitted cell counts so that they match each observed marginal table in the set of minimal sufficient statistics for the model. This iterative procedure continues until the fitted values of sufficient statistics and their observed values are sufficiently close, converging to maximum likelihood estimates. Our modified IPF algorithm iteratively fits three margins. The block $\times$ source margin is fully observed and unbiased estimates for the census $\times$ tract $\times x_4$ margin are obtained for nonsample nonrespondents by applying sampling weights to the sampled NRFU cases. The block $\times x_2$ margin is observed for administrative records, but unobserved for nonsample nonrespondent households. For this incompletely observed margin, predictions for nonsample nonrespondent households in each block are obtained by applying, during the fitting algorithm, the same fitting proportions to those missing households as to the administrative households in each block. This modified IPF algorithm produces maximum likelihood estimates because the nonsample nonrespondent households contribute to the likelihood only through the total number of nonrespondent households in each block.

Further details about the properties of this model and an alternative method of fitting this model, including the case of a block sampling design for NRFU (where all households in selected blocks, and none in other blocks, are followed up), appear in Zanutto (1998) and ZZ.

2.5. *Modeling strategies*

We compare the use of loglinear Model (1), ''modeling with administrative records,'' to two other strategies: one which uses administrative records without modeling, and one which uses modeling without administrative records. In the ''substitution method,'' we substitute household types from administrative records for the nonsample nonrespondents. ''Modeling without administrative records'' ignores administrative records and fits a loglinear model, similar to Model (1), using census respondents to predict the number of nonsample nonrespondent households of each type in each block. More specifically we use Model (1) with d representing response status (respondent or nonrespondent) so

that estimates of the number of nonsample nonrespondent households of each type in each block depend on the characteristics of respondents in the same block and nonrespondents in the NRFU sample in the same tract.

Our ''modeling with administrative records'' strategy is unavoidably complicated by the incomplete coverage of the administrative record database, which requires use of a two-part model. We first divide nonrespondent households into those that can (Group A) and cannot (Group B) be linked to administrative records. To estimate household types in Group A, we fit a loglinear model in which variable d indicates administrative records versus nonresponse follow-up responses. From the fitted model we predict the types of nonrespondent households that have administrative records but are not in the NRFU sample. We fit a loglinear model to Group B identical to the ''modeling without administrative records'' described above. Combining the estimates for Groups A and B gives estimates for all nonsample nonrespondents. This strategy uses administrative records, whenever they are available, as predictors of the characteristics of the nonrespondents, and census respondents otherwise. The ''substitution method'' is modified in a similar fashion so that we substitute household types from administrative records for all households in Group A, with model-based estimation for Group B as in the two-part model.

3. Simulation Study with 1995 Census Test Data

Analytical evaluation of the estimation and imputation strategy we propose is unlikely to be feasible due to the complexity of both the proposed models and the relationship between census data and administrative records. Instead we explore through simulations the gains in accuracy that are possible by incorporating information from administrative records into the modeling process. Because the primary goal of this research is to evaluate the performance of the loglinear (household type) model, all vacant households are deleted from the simulation data sets, thus eliminating the need for Step 1 of Section 2.3. Similarly, because we are interested in evaluating the use of administrative records to predict the characteristics of nonsample nonrespondents we restrict the simulations to include only households with administrative records. Steps 2c and 3 are also omitted, since they are unaffected by the choice of model. Since the most recent proposals for NRFU sampling in the U.S. Decennial Census specified a unit sampling design for NRFU (U.S. Bureau of the Census 1997; Farber 1996) our simulations also use that design. However, these models can also be used under a block sampling design.

3.1. Data

Our simulations use census data and administrative records from the Oakland, California and Paterson, New Jersey sites of the 1995 U.S. Decennial Census Test. The administrative records databases combine records from federal government files (Housing and Urban Development files, 1993 Individual Tax Return Master File, Social Security Administration files, Medicare files, food stamp files), state government files (drivers' licenses), and local files (public school enrolment, voter registration, parolee lists, probationer lists) (Wurdeman and Pistiner 1997). Neugebauer, Perkins, and Whitford (1996) describe the difficulties of acquiring the various administrative files. To form the final database, person-level records from all sources were standardized and combined into one master file

that was then unduplicated with the goal of having no more than one administrative record per person. Finally, administrative records were assigned housing unit identification (HUID) numbers using the same algorithm that was applied to census records (Wurdeman and Pistiner 1997). The resulting database contains person-level information about address (HUID, and census area divisions such as tract and block), sex, race, Hispanic origin, date of birth, and marital status. The consolidated administrative record for a person may contain information from several different sources. Because the final database did not record the source for each item, nor did it record the number of sources that corroborated this information, incorporating these additional pieces of information has been recommended for future administrative record databases (Neugebauer, Perkins, and Whitford 1996; White and Rust 1997), and such data have been analyzed with later versions of the database (Larsen 1999). White and Rust (1997) summarize the development of the 1995 Census Test administrative records databases and evaluate the administrative data.

Since sampling for NRFU was conducted in the 1995 Census Test, we know the actual characteristics for all mail-back respondents and for all nonrespondent households in the NRFU sample. Hence, our simulation population is limited to blocks containing nonrespondent households in the NRFU sample, including all respondents in these blocks and sampled nonrespondents. A block sampling design was used for NRFU 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 follow-up (Vacca, Mulry, and Killion 1996). Table 1 describes the simulation populations from the two test sites. In this article, to focus on the potential improvement that can be made by using administrative records, simulation results are presented only for estimates of the characteristics of nonsample nonrespondents with administrative records. Results for the entire simulation population are presented by Zanutto and Zaslavsky (2002).

A comparison of the distributions of the basic household characteristics in the administrative records and the census NRFU sample, for households where both sources

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

Test Site	Oakland	Paterson
Number of Households	58,387	11,096
Number of Blocks	1,803	292
Number of Tracts*	91	31
Nonresponse Rate	19.3%	49.8%
Hispanic Households	10.9%	35.8%
Black Households	36.3%	36.2%
Other (Race) Households	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 Oakland site and 33 in the Paterson site but several small tracts were combined to form larger tracts for the simulations.

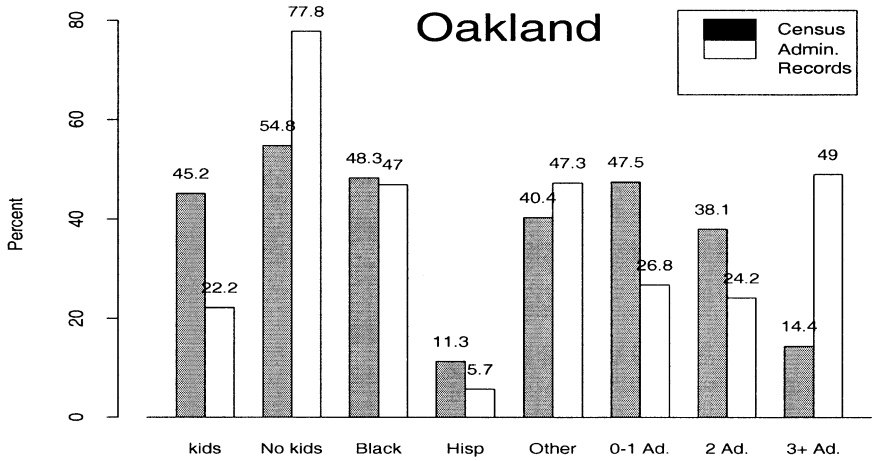

Fig. 1. Prevalence of household characteristics in administrative records for nonrespondent households and in the corresponding census records for the Oakland, California simulation data

of information are available, illustrates some of the common problems with administrative records (Figures 1 and 2). Only 50.9% of the nonrespondents in Oakland and 21.5% in Paterson had usable administrative records. The administrative records severely understate the number of households with children in both data sets, a consequence of relying on sources that contain few or no children. In Oakland, the proportion of households with three or more adults is overstated in the administrative records, because many of the records were outdated, so both current and previous occupants were listed at the same address. In Paterson the proportion of households with 0–1 adults is overstated in the administrative records, due to undercoverage of adults when records contained information for only a single household member, rather than all household members. On the other hand, the administrative records agree fairly well with the census on the distribution of households by race.

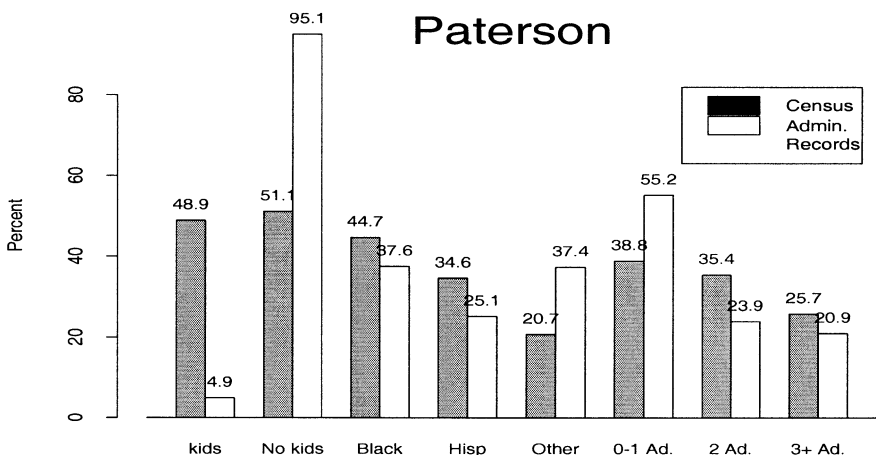

Fig. 2. Prevalence of household characteristics in administrative records for nonrespondent households and in the corresponding census records for the Paterson, New Jersey simulation data

Similar patterns were found for agreement between census and administrative data for individual nonrespondent households. The agreement rates for Oakland and Paterson are, respectively, 29.9% and 24.7% on household type, 84.0% and 74.8% on race, 42.7% and 48.9% on adult category, and 77.0% and 56.0% on child category.

3.2. *Simulation design*

We evaluated the bias, variance, and mean squared error of the alternative estimates of demographic aggregates (such as number of households by race, number of adults, and number of children) at the block, tract, and site levels, using estimated household compositions for nonsample nonrespondent households. Estimates at tract and site levels were formed by aggregating block-level estimates.

Using data for which we know the characteristics of all respondents and nonrespondents, we simulated NRFU sampling by drawing a one in three simple random sample of nonrespondent households in each tract. We fitted the models, estimated the number of nonsample nonrespondent households of each type in each block, and compared aggregates at the block, tract, and site levels to the truth. We repeated these steps 30 times for each estimation method to obtain sufficiently accurate estimates of Root Mean Weighted Mean Squared Error (RMWMSE). This loss function is based on the relative error in estimates for nonrespondents in household category j (a type or combination of types) in geographical unit i (a block or collection of blocks):

$$d_{ijs} = \frac{\hat{Y}_{ijs} - Y_{ij}}{Y_{i+}} \tag{2}$$

where Y_{ij} is the true number of nonrespondent households of category j in geographical unit i, $\hat{Y}_{ijs}$ is the estimated number of nonrespondent households of category j in geographical unit i using the model fit from sample s, and Y_{i+} is the total number of nonrespondent households in geographical unit i. For example, $\hat{Y}_{ijs}$ could be the estimated number of nonrespondent households of Type 3 in block i or it could be the estimated number of nonrespondent childless households in tract i.

The RMWMSE for the estimate of the number of nonrespondent households of category j in a geographical unit (e.g., block, tract, site) is estimated by

$$\widehat{\text{RMWMSE}}_j = \sqrt{\frac{\sum_i Y_{i+}(1/S \sum_s d_{ijs}^2)}{\sum_i Y_{i+}}} \tag{3}$$

where Y_{ij}, $\hat{Y}_{ijs}$, Y_{i+}, i, and $S = 30$ are defined as above. (The two ''means'' are over geographical units (i) and over samples (s).) This quantity, like the corresponding measures of bias and variance, may be interpreted as the average error for percentages in a category over geographical units. This type of measure has several desirable properties as described in Zanutto (1998) and ZZ.

In these simulations, all loglinear models use $x_2 =$ race and $x_4 =$ household type, so the x_1 and $d * x_3$ terms are absorbed into the $d * a * x_4$ term. Experimentation with several other specifications of x_2 did not result in the reduction of RMWMSE overall.

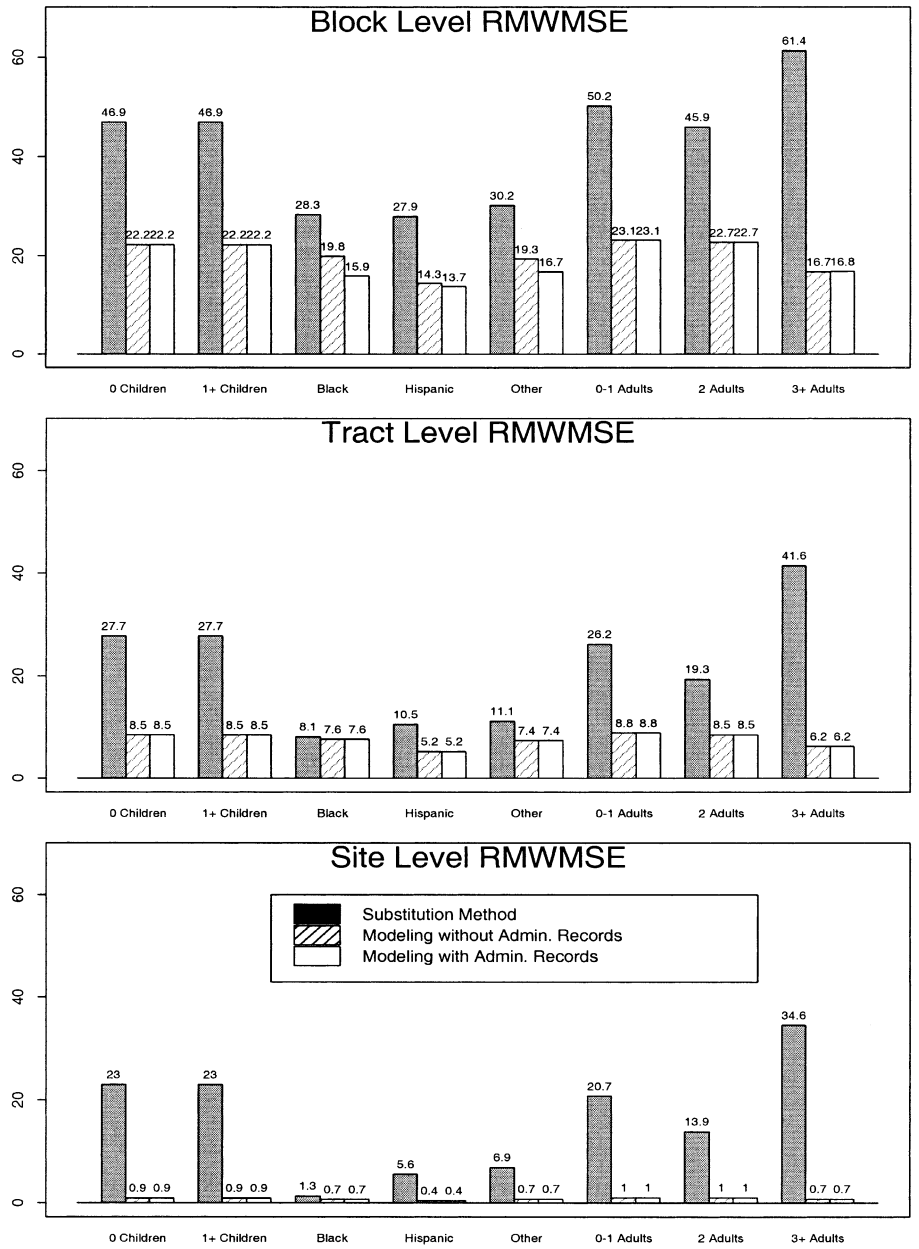

Fig. 3. RMWMSE, expressed as a percent, for estimates for each nonrespondent household characteristic at block, tract, and site levels, resulting from each of the three estimation methods, for the Oakland simulation data set

3.3. Simulation results

Simulation results are shown in Figures 3 and 4. The three bar charts for each site show the RMWMSE for the estimates of the total number of nonrespondent households in each of the race, adult, and children categories, at each of the block, tract, and site levels of geography.

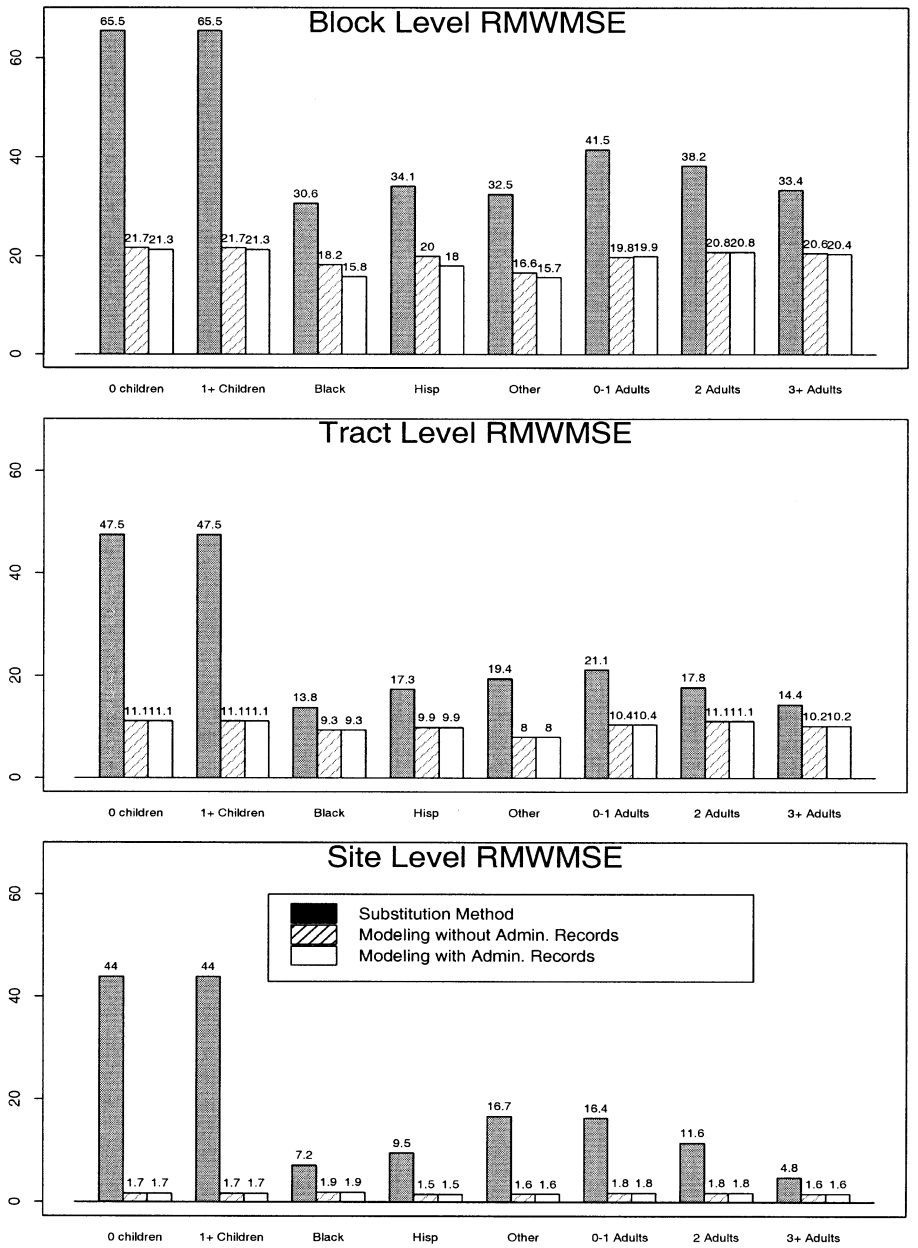

Fig. 4. RMWMSE, expressed as a percent, for estimates for each nonrespondent household characteristic at block, tract, and site levels, resulting from each of the three estimation methods, for the Paterson simulation data set

Substitution produces block-level estimates with substantially larger RMWMSE than the other methods for the children and adult categories, which are critical because they determine total population. These effects of bias are even more dramatic at the tract and site levels, where sampling error is a smaller component of error.

The model-based methods have smaller RMWMSE than substitution for almost all

household characteristics at all levels of geography. We compare these methods only at the block level, because the loglinear model constrains tract- and site-level estimates to equal the same unbiased estimates from the NRFU sample. Therefore, the modeling methods produce the same estimates at the tract and site levels, but may differ at the block level. Use of administrative records reduces RMWMSE for the race categories ($p < .0001$) compared to modeling without administrative records. Modeling without administrative records yields RMWMSEs of 19.85%, 14.34%, and 19.31% for the Black, Hispanic, and Other race categories in Oakland compared to 15.87%, 13.73%, and 16.73%, respectively, when administrative records are used. In Paterson, RMWMSEs were 18.24%, 20.00%, and 16.60% without administrative records compared to 15.82%, 18.00%, and 15.66% with. The differences are due to a smaller bias component. Using administrative records has little effect on RMWMSE for the children and adult categories.

4. Summary

This example illustrates that using administrative records through modeling can improve, albeit modestly in this application, the accuracy of small area (block-level) estimates. Direct substitution of administrative records for missing data can engender large biases. When an administrative records database with fairly complete and consistent records is developed, we can overcome concerns about bias relative to the gold standard of the survey estimates, because our methodology uses information from administrative records to construct estimates at detailed levels of geography while constraining these estimates to agree with unbiased survey estimates at higher levels.

Improvements in accuracy are modest in this application due to limitations of these administrative records databases, including limited coverage of households and selection of variables. Nevertheless, it is promising that even with these limitations, using administrative records through statistical modeling leads to gains in accuracy at the smallest level of detail (the most difficult to estimate). More dramatic benefits should be obtained as the quality of the administrative records databases improves. Furthermore, when administrative records are of sufficient quality, the administrative records households can be used as imputation donors, with our model estimating the number of households of each type to impute in each block. This allows actual observed households to be used for imputation, and avoids criticisms that imputed households are ''made up'' by the U.S. Census Bureau. Where administrative records are of such quality that they can be used as a primary data source, our model can be used to correct small biases in the administrative records.

Although sampling for nonresponse follow-up has been prohibited for the decennial census, administrative records are expected to play a more prominent role in census operations in the future. The U.S. Census Bureau is currently researching several potential applications of administrative records including nonresponse follow-up substitution and imputation, imputation for item nonresponse, reducing differential undercoverage, address list improvement, linkages to ongoing survey programs, and population estimation (Judson 2000; Panel on Future Census Methods, Committee on National Statistics 2001). To support this research, the U.S. Census Bureau conducted an Administrative Records Experiment in the 2000 Census (AREX 2000), is currently developing a ''Statistical Administrative Records System'' which is a database of personal and address

data using administrative records from various government agencies (Farber and Leggieri 2002), and is planning an Administrative Records Census Experiment in 2003 (Leggieri and Prevost 1999). These efforts promise to greatly improve the combined administrative record system by incorporating national files that in combination cover most of the population, such as Internal Revenue Service files of tax returns and information forms and files from the Social Security System covering recipients' retirement insurance. Better coverage will aid the performance of our model in two ways: first, more households will appear in the administrative database, and second, the classification of each household will be more accurate because the list of members will be more complete.

Looking beyond the decennial census, our methodology might be used to improve small-area estimates from the American Community Survey, a large survey that will be conducted continuously and is intended to replace the long form of the U.S. decennial census (Committee on National Statistics 2001). Use of administrative records in combination with this survey might improve the accuracy of small-area estimates, making it less necessary to roll up several years of data to obtain acceptable accuracy.

5. References

Birch, M.W. (1963). Maximum Likelihood Estimation of a Linear Structural Relationship. Journal of the Royal Statistical Society, Series B, 25, 220–233.

Brackstone, G.J. (1987). Issues in the Use of Administrative Records for Statistical Purposes. Survey Methodology, 13, 29–43.

Committee on National Statistics (2001). The American Community Survey: Summary of a Workshop. Washington, DC: National Academy Press.

Cox, L.H. (1987). A Constructive Procedure for Unbiased Controlled Rounding. Journal of the American Statistical Association, 82, 520–524.

Csiszár, I. (1989). A Geometric Interpretation of Darroch and Ratcliff's Generalized Iterative Scaling. The Annals of Statistics, 17, 1409–1413.

Darroch, J.N. and Ratcliff, D. (1972). Generalized Iterative Scaling for Log-linear Models. The Annals of Mathematical Statistics, 43, 1470–1480.

Farber, J. (1996). A Comparison of Imputation Methods for Sampling for Nonresponse Follow-up. Proceedings of the American Statistical Association, Section on Survey Research Methods, 383–388.

Farber, J. and Leggieri, C. (2002). Building and Validating a National Administrative Records Database for the United States. Administrative Records Research Memorandum Series. Washington, DC, U.S. Census Bureau.

Fienberg, S.E. and Meyer, M.M. (1983). Iterative Proportional Fitting. Encyclopedia of Statistical Sciences, Vol. 4. New York: Wiley, 275–279.

Fischetti, M. and Salazar-González, J. (1998). Experiments with Controlled Rounding for Statistical Disclosure Control in Tabular Data with Linear Constraints. Journal of Official Statistics, 14, 553–565.

Fuller, W.A., Isaki, C.T., and Tsay, J.H. (1994). Design and Estimation for Samples of Census Nonresponse. Proceedings of the U.S. Bureau of the Census Annual Research Conference. Suitland, MD: U.S. Bureau of the Census, 289–305.

Ghosh, M. and Rao, J.N.K. (1994). Small Area Estimation: An Appraisal. Statistical Science, 9, 55–76.

Judson, D.H. (2000). The Statistical Administrative Records System: System Design, Successes, and Challenges. Presented at the NISS/Telcordia Data Quality Conference, November 30–December 1.

Larsen, M. (1999). Predicting the Presidency Status for Administrative Records that Do not Match Census Records. Technical Report. U.S. Census Bureau Administrative Records Research Memorandum Series #20, U.S. Bureau of the Census, Washington, DC.

Leggieri, C. and Prevost, R. (1999). Expansion of Administrative Records Uses at the U.S. Census Bureau: A Long-Range Research Plan. Presentation to the Census Advisory Committee of Professional Associations, October.

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

Panel on Future Census Methods, Committee on National Statistics (2001). Designing the 2010 Census: First Interim Report. Washington, DC: National Academy Press.

Rancourt, E. and Hidiroglou, M. (1998). The Use of Administrative Records in the Canadian Survey of Employment, Payrolls, and Hours. Statistical Society of Canada Proceedings of the Survey Methods Section, 39–47.

Royce, D., Hardy, F., and Beelen, G. (1997). Project to Improve Provincial Economic Statistics. Proceedings, International Symposium Series. Ottawa, Ontario, Canada: Statistics Canada, 21–24.

Rubin, D.B. and Zaslavsky, A.M. (1989). An Overview of Representing Within-household and Whole-household Misenumerations in the Census by Multiple Imputations. Proceedings of the U.S. Bureau of the Census Annual Research Conference, Vol. 5. U.S. Bureau of the Census, 109–117.

Schafer, J. (1995). Model-based Imputation of Census Short-form Items. Proceedings of the U.S. Bureau of the Census Annual Research Conference. U.S. Bureau of the Census, 267–299.

U.S. Bureau of the Census (1997). Census 2000 Operational Plan. Washington, DC.

Vacca, E.A., Mulry, M., and Killion, R.A. (1996). The 1995 Census Test: A Compilation of Results and Decisions. Technical Report DMD 1995 Census Test Results Memorandum #46, U.S. Department of Commerce, US Bureau of the Census.

White, A. and Rust, K. (1997). Preparing for the 2000 Census: Interim Report I of the Panel to Evaluate Alternative Census Methodologies. Washington, DC: National Academy Press.

Whitridge, P., Bureau, M., and Kovar, J. (1990). Use of Mass Imputation to Estimate for Subsample Variables. Proceedings of the American Statistical Association, Section on Business and Economics Statistics, 132–137.

Wilkinson, G.N. and Rogers, C.E. (1973). Symbolic Description of Factorial Models for Analysis of Variance. Applied Statistics, 22, 392–399.

Wurdeman, K. and Pistiner, A.L. (1997). 1995 Administrative Records Evaluation – Phase II. Technical Report DMD 1995 Census Test Results Memorandom Series #54, Revised.

Zanutto, E. (1998). Imputation for Unit Nonresponse: Modeling Sampled Nonresponse Follow-up, Administrative Records, and Matched Substitutes. Ph.D. thesis, Department of Statistics, Harvard University.

Zanutto, E. and Zaslavsky, A.M. (1995a). A Model for Imputing Nonsample Households With Sampled Nonresponse Follow-up. Proceedings of the American Statistical Association, Section on Survey Research Methods, pp. 608–613.

Zanutto, E. and Zaslavsky, A.M. (1995b). Models for Imputing Nonsample Households With Sampled Nonresponse Followup. Proceedings of the U.S. Bureau of the Census Annual Research Conference, 673–686.

Zanutto, E. and Zaslavsky, A.M. (2002). Using Administrative Records to Impute for Nonresponse. In R. Groves, D. Dillman, J. Eltinge, and R.J.A. Little (eds.), Survey Nonresponse. New York: Wiley, 403–415.

Zaslavsky, A.M. (1989). Multiple-system Methods for Census Coverage Evaluation. Proceedings of the American Statistical Association, Section on Survey Research Methods, pp. 681–686.

Received January 2001
Revised March 2002