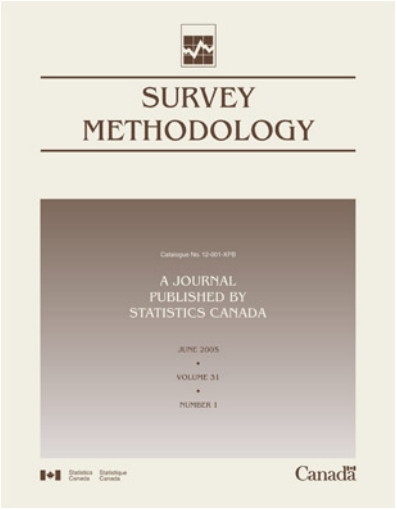
Catalogue no. 12-001-XIE

Survey Methodology

June 2005



Statistics Canada Statistique Canada

Canada

How to obtain more information

Specific inquiries about this product and related statistics or services should be directed to: Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6 (telephone: 1 800 263-1136).

For information on the wide range of data available from Statistics Canada, you can contact us by calling one of our toll-free numbers. You can also contact us by e-mail or by visiting our website.

National inquiries line	1 800 263-1136
National telecommunications device for the hearing impaired	1 800 363-7629
Depository Services Program inquiries	1 800 700-1033
Fax line for Depository Services Program	1 800 889-9734
E-mail inquiries	infostats@statcan.ca
Website	www.statcan.ca

Information to access the product

This product, catalogue no. 12-001-XIE, is available for free. To obtain a single issue, visit our website at www.statcan.ca and select Our Products and Services.

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner and in the official language of their choice. To this end, the Agency has developed standards of service that its employees observe in serving its clients. To obtain a copy of these service standards, please contact Statistics Canada toll free at 1 800 263-1136. The service standards are also published on www.statcan.ca under About Statistics Canada > Providing services to Canadians.

Statistics Canada
Business Survey Methods Division

Survey Methodology

June 2005

Published by authority of the Minister responsible for Statistics Canada

May 2006

Catalogue no. 12-001-XIE
ISSN 1492-0921

Frequency: semi-annual

Ottawa

Cette publication est disponible en français sur demande (nº 12-001-XIF au catalogue).

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued cooperation and goodwill.

Vol. 31, No. 1, pp. 65-72
Statistics Canada, Catalogue No. 12-001

Using Matched Substitutes to Improve Imputations for Geographically Linked Databases

Wai Fung Chiu, Recai M. Yucel, Elaine Zanutto and Alan M. Zaslavsky [1]

Abstract

When administrative records are geographically linked to census block groups, local-area characteristics from the census can be used as contextual variables, which may be useful supplements to variables that are not directly observable from the administrative records. Often databases contain records that have insufficient address information to permit geographical links with census block groups; the contextual variables for these records are therefore unobserved. We propose a new method that uses information from "matched cases" and multivariate regression models to create multiple imputations for the unobserved variables. Our method outperformed alternative methods in simulation evaluations using census data, and was applied to the dataset for a study on treatment patterns for colorectal cancer patients.

Key Words: Unit nonresponse; Multiple imputation; Contextual variables; Matched substitutes; Administrative records.

1. Introduction

In a study on treatment patterns for colorectal cancer patients, income and education are desired variables for constructing statistical models of relevant scientific interest. Unfortunately, individual measurements for these variables are not directly observable from the cancer registry databases that are compiled from hospital records, which like many administrative databases contain primarily information required for administrative purposes. Instead, mean values of these variables for small geographical areas (census block groups or tracts) including the subject's area of residence are used as regressors to estimate income and education effects. Analyses using such "contextual variables" are common in epidemiological and health services research (Krieger, Williams and Andmoss 1997), and often produce results broadly similar to those based on individual variables. If both individual and contextual variables were available, it might be possible to separate the effects of individual characteristics and contexts; in a purely contextual analysis, these effects are confounded. Nonetheless, associations between contextual socioeconomic characteristics and quality of care would suggest an equity problem, regardless of whether such associations primarily reflect individual or community-level relationships.

In the colorectal cancer treatment study, each contextual variable for a given patient record is assumed to be the variable's census group (or tract) mean value obtained by geographically linking the record's address to a census block group (or tract). A small but substantial percentage of patient records (about 3.3% or 1,696 records) have insufficient address information to permit links with census block groups, hence making the corresponding contextual variables unobservable. Such records will be called *ungeocodable* records, while records that can be linked to census block groups will be referred to as *geocodable*. To generate multiple imputations for the unobserved contextual variables, we propose a strategy that uses information from more than one "matched case" to help build parametric/nonparametric imputation models. In particular, information from the matched cases accounts for small area effects in our imputation models, so that there is no need to explicitly model such effects.

Rubin and Zanutto (2001) use the term "matched substitute" instead of "matched case", and propose a parametric imputation model using only one matched substitute per record. The analyses resulted from their model were compared to those given by other analytic methods in an extensive simulation study, but was not applied to real data. We extend Rubin and Zanutto's method by (1) allowing use of information from more than one matched case per record and (2) using an empirical rather than a parametric distribution of residuals.

This research was motivated by our need for multiple imputations for the partially observed variables in the study of treatment patterns for colorectal cancer patients. Ayanian, Zaslavsky, Fuchs, Guadagnoli, Creech, Cress, O'Connor, West, Allen, Wolf and Wright (2003) analyzed a dataset that included imputations generated by our method,

1. Wai Fung Chiu, Department of Statistics, Harvard University, One Oxford Street, Cambridge MA 02138. E-mail: wfchiu@post.harvard.edu; Recai M. Yucel, Department of Biostatistics and Epidemiology, 408 Arnold House, School of Public Health and Health Sciences, University of Massachusetts, 715 North Pleasant Street, Amherst, MA 01003-9304. E-mail: yucel@schoolph.umass.edu; Elaine Zanutto, The Wharton School, University of Pennsylvania, 466 Jon M. Huntsman Hall, 3730 Walnut Street, Philadelphia PA 19104. E-mail: zanutto@wharton.upenn.edu; Alan M. Zaslavsky, Department of Health Care Policy, Harvard Medical School, 180 Longwood Avenue, Boston MA 02115. E-mail: zaslavsky@hcp.med.harvard.edu.

referring to Rubin and Zanutto (2001) and a preliminary version of this paper that appeared in a proceedings publication (Chiu, Yucel, Zanutto and Zaslavsky 2001). This paper is the first comprehensive publication of our methodology and the first published report that describes an application of Rubin and Zanutto's method to real data.

The organization for the rest of this paper is as follows. Section 2 summarizes Rubin and Zanutto's method and gives a general description of our method. Section 3 outlines the application of our method to the colorectal cancer study. Section 4 illustrates in a simulation study the performance of our method relative to three other commonly-used nonresponse adjustment methods.

2. Imputation Methodology

This section will begin with a summary of Rubin and Zanutto's method, followed by a general description of our method that includes a discussion on out-of-sample versus within-sample matching, the details of the modeling and multiply-imputing tasks, and an analysis of efficiency as a function of the number of matched cases used.

2.1 Matching, Modeling and Multiply Imputing

Rubin and Zanutto (2001) proposed a method called "matching, modeling, and multiply imputing" (MMM) that uses matched substitutes to help generate multiple imputations for nonrespondents in sample surveys, without requiring that substitutes be perfect replacements for the nonrespondents. Matched substitutes are responding survey units chosen to match the nonrespondents on one or more "matching covariates" – variables that are available prior to the survey and are convenient for matching but not necessarily for modeling. As a result of matching, nonrespondents and their substitutes may share similar values in their "field covariates" – variables that are only implicitly observed and are therefore not available for data analysis. "Modeling covariates" are variables that can be included in statistical models to adjust for observed differences between nonrespondents and their substitutes, but that may not be available or used for matching. The essence of MMM is that both matching and modeling covariates are used, in the context of proper multiple imputation (Little and Rubin 1987, pages 258 – 259 and references therein).

Consider a simple example where age and address covariates are available for all units in a population prior to sampling. Finding substitutes matching nonrespondents with respect to both age and address may be difficult. An alternative is to match only on address (*e.g.*, choosing a neighbor to be a substitute) and adjust for systematic age differences between nonrespondents and matched substitutes through statistical modeling. If neighboring households were chosen as matched substitutes for nonresponding households, the substitutes and nonrespondents might have similar socioeconomic contexts (*e.g.*, levels of crime, access to public transportation, *etc*.) even though these characteristics might have not been recorded. In this example, address is a matching covariate, age is a modeling covariate, and the contextual socioeconomic characteristics are field covariates.

In summary, MMM (i) chooses matched substitutes for nonrespondents and some respondents based on matching covariates, (ii) uses modeling covariates to fit a model estimating the systematic differences in responses between pairs of respondents and substitutes, (iii) multiply-imputes the unobserved values using the model in (ii) under the assumption that the same relationship holds between pairs of nonrespondents and substitutes, and (iv) discards all matched substitutes after imputation.

2.2 Out-of-Sample Versus Within-Sample Matching

Matched cases may be obtained from out-of-sample data or within-sample data. In the Rubin and Zanutto approach, matched substitutes are obtained from out-of-sample data *after* the missingness is detected. Their description emphasizes that the matched substitutes must be discarded after imputation since including such additional cases in inferences would modify the sample design by adding extra cases in the "blocks" that contain unobserved data. Matched cases are considered within-sample data if they are obtained from the database that is available *before* imputing or even finding out which records in the database have unobserved variables. As far as the overall inferential goals are concerned, these matched cases are not additional cases, but are part of the original data collection, and therefore will be included in scientific analyses.

Assuming within-sample matching, we treat the ungeocodable records as nonrespondents and the geocodable records as respondents. For each ungeocodable record, a given number of matched cases are randomly chosen from a pool of geocodable records within the same small geographical area (*e.g.*, zip code, which is a postal delivery code usually representing an area served by a single main US post office). Similarly, the same number of matched cases are also chosen for each of the randomly sampled geocodable records (see Rubin and Zanutto (2001) for recommendations on the size of such a sample relative to the total number of ungeocodable records in a given dataset). If more matched cases were needed than those are available in the same small area, the selection pool would be extended to the "nearest" geographical areas until the required number of matched cases was achieved.

All matched cases in the colorectal cancer study came from the same cancer database. In general, matched cases need not be drawn from the same population in which the nonrespondents and respondents originated. For example, matched cases for colorectal cancer records can be obtained from a general population of cancer patients, and a model can then be fitted to correct for systematic differences. Note that, with matched cases from a more similar population, stronger models can be built with more covariates. In our example, since we used other patients with the same cancer type, relationships to treatment process and outcome variables are likely to be consistent.

2.3 Modeling and Multiply-imputing

A simple example of our method is given here to convey the basic idea; in practice, more complex models may often be required. Suppose the following relationship holds in the population,

$$y_{ik} = \boldsymbol{x}_{ik}^T \boldsymbol{\beta} + \delta_i + \varepsilon_{ik}, \tag{1}$$

where i indexes small geographical area, k indexes unit within area, and y_{ik} and $\boldsymbol{x}_{ik}$ are respectively the response and the characteristics of the k^{th} unit in geographical area i. This model includes a regression prediction $\boldsymbol{x}_{ik}^T \boldsymbol{\beta}$, a small-area effect δ_i, and a unit-specific residual ε_{ik}. We assume that ε_{ik} follows some distribution F_ε with mean zero and variance σ^2. Note that this development generalizes directly to multivariate $\boldsymbol{y}_{ik}$.

We extend Rubin and Zanutto's method to allow more than one match in the same small area, because having several matches in small areas is possible (often convenient and inexpensive) in census data or in large administrative datasets. Rubin and Zanutto's assumption of a single match is appropriate to survey data collection that requires additional field work for each match.

The regression coefficients in equation (1) are estimated using any collection of observations with two or more records per small area to fit the regression model in which the δ_i are treated as fixed effects. With only two cases per area, $\boldsymbol{\beta}$ can instead be estimated from the within-area regression

$$(y_{i1} - y_{i2}) = (\boldsymbol{x}_{i1}^T - \boldsymbol{x}_{i2}^T)\boldsymbol{\beta} + (\varepsilon_{i1} - \varepsilon_{i2}), \tag{2}$$

where the small area effect drops out. The residuals from this regression have a symmetrical distribution with variance $2\sigma^2$.

Assuming for the moment that we have a draw from the posterior distribution of $\boldsymbol{\beta}$, we carry out the rest of this analysis conditional on that draw. Now suppose that we are interested in imputing for a new unit (indexed as $k = 0$) in area i, and that we have obtained $K_i \geq 1$ matched cases for this unit. Denote the outcomes of these matched cases by the vector $\boldsymbol{y}_i = (y_{i1}, ..., y_{iK_i})^T$ and the corresponding characteristics by the matrix $\boldsymbol{X}_i = (\boldsymbol{x}_{i1}, ..., \boldsymbol{x}_{iK_i})^T$. With a flat prior for δ_i, the posterior distribution for $\delta_i \mid \boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{\beta}$ has mean

$$\bar{y}_i - \bar{\boldsymbol{x}}_i^T \boldsymbol{\beta} \tag{3}$$

and variance σ^2 / K_i, where $\bar{y}_i = \sum_{k=1}^{K_i} y_{ik} / K_i$ and $\bar{\boldsymbol{x}}_i = \sum_{k=1}^{K_i} \boldsymbol{x}_i / K_i$. Hence, the predictive distribution for $y_{i0} \mid \boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{x}_{i0}, \boldsymbol{\beta}$ has mean

$$\bar{y}_i + (\boldsymbol{x}_{i0}^T - \bar{\boldsymbol{x}}_i^T)\boldsymbol{\beta} \tag{4}$$

and variance $(1 + 1/K_i)\sigma^2$ which is the sum of the predictive variance under the model conditional on all parameters and the posterior variance of δ_i. These statements assume that the mean of the residuals is a sufficient statistic for δ_i. This assumption is true for the normal distribution (or natural observations of any exponential family distribution); we assume it is at least approximately true for F_ε, so that we can base inferences on that mean. Note that use of a flat prior leads to overdispersed draws relative to what would be obtained with a proper prior from a hierarchical model, but is much simpler (especially in analyses with the multivariate outcomes).

An imputation for y_{i0} can be generated by first drawing σ^2 from its posterior distribution, second drawing $\boldsymbol{\beta}$ conditional on the draw of σ^2, third computing the predictive mean in equation (4) from the draw of $\boldsymbol{\beta}$, and finally adding a residual of variance $(1 + 1/K_i)\sigma^2$ to the predictive mean. In simple surveys with $\boldsymbol{\beta}$ estimated by equation (2), the posterior distribution of $\boldsymbol{\beta}$ (conditional on σ^2 and the data) under a flat prior is approximately $\mathrm{N}(\hat{\boldsymbol{\beta}}, (\boldsymbol{X}^T \boldsymbol{X})^{-1}\sigma^2)$ where the i^{th} row of $\boldsymbol{X}$ is $(\boldsymbol{x}_{i1}^T - \boldsymbol{x}_{i2}^T)$. In more complex designs, the posterior distribution of $\boldsymbol{\beta}$ can be approximated using the point estimate and sampling variance calculated under the associated design.

The residual can be obtained through modeling *or* sampling. Modeling involves estimating σ^2 using the residual variance of equation (1) and drawing the residual under univariate normality (see Rubin and Zanutto (2001) for the special case where only one matched case was obtained for each record) or some other parametric distribution. We refer to such an approach as **parametric MMM** (PMMM). An alternative is to randomly sample a regression residual from *any* area j whose residuals might be regarded as exchangeable with those from area i (Rubin 1987 pages 166 – 168). See also Lessler and Kalsbeek (1992, section 8.2.2.4), Kalton and Kasprzyk (1986), and Kalton (1983). Since the variance of such a residual is $[(K_j - 1)/K_j]\sigma^2$, we multiply the randomly-sampled residual by $\sqrt{[(K_i + 1)/K_i][K_j/(K_j - 1)]}$ to obtain the

correct predictive variance. We call this approach **nonparametric MMM** (NpMMM).

In summary, our method consists of three basic steps:

1. Draw matched cases for the ungeocodable records and for some randomly sampled geocodable records;
2. Use the sampled geocodable records and their matched cases to fit equation (1) where the δ_i are treated as fixed effects, and save the residuals;
3. Repeat the following for m (usually 5 to 10) times:
   (a) Draw σ^2 from its posterior distribution, then $\boldsymbol{\beta}$ conditional on the draw of σ^2;
   (b) For each ungeocodable record, treat the sum of the vector of predictive means obtained from equation (4) and a vector of residuals drawn using either PMMM or NpMMM as a realization of the unobserved vector of contextual variables.

2.4 Efficiency

The efficiency of an imputation is related to the number of matched cases used. Let V_K be the predictive variance of an imputation model where K matched cases per record are used. For the model in section 2.3, $V_K = (1 + 1/K)\,\sigma^2$. Define efficiency as

$$E_K = \frac{V_\infty}{V_k} = \frac{\sigma^2}{(1 + 1/K)\,\sigma^2} = \frac{K}{K + 1}, \qquad (5)$$

for any positive integer K. Efficiency increases as the number of matched cases per record increases; for example, $E_2 \approx 0.67$, $E_4 = 0.8$, $E_{10} \approx 0.91$, and $E_{20} \approx 0.95$.

Theoretically each record can have as many matched cases as permitted by available resources. In practice, the number of matched cases used often depends on the cost of matched cases and the cost of computation involved in model fitting. In our method, the cost of computation for each added matched case per record is negligible. In the colorectal cancer study, while the matched cases were free, the ability to do the imputation based on a limited number of matched cases was crucial because confidentiality restrictions prevented investigators from using the entire dataset in modeling with zip codes (even in a coded form) attached. For illustrative purposes, we will use two matched cases per record in subsequent analyses.

3. Application: The Colorectal Cancer Study

The colorectal cancer database has a total of 50,740 patient records, of which approximately 3.3% are ungeocodable. Among these, about half have P.O. box addresses (often in a rural area), and the rest are mistyped addresses or addresses from newly developed areas that are not in address databases. In a study of factors predicting provision of chemotherapy for colorectal cancer patients, investigators believed that the following three census block-group means would be useful contextual variables:

Y_1 = median household income,

Y_2 = percent with no high school diploma, and

Y_3 = percent below poverty level.

These variables were observed in geocodable records but unobserved in ungeocodable records. The task was to generate multiple imputations for the unobserved census variables using the methods in section 2.

Each of the block-group means was reported in the census data for six race/ethnic groups, and the scientific analyses used only the set of block-group means corresponding to the race/ethnicity of each patient. For imputations used in Ayanian *et al.* (2003), we therefore fitted six separate models to impute all $18 (6 \times 3)$ values for each ungeocodable patient and then selected the three variables pertinent to each patient; joint distributions for different race/ethnic groups were not important because each imputation only used values for a single group. An alternative would have been to use race as a matching variable, but this would have forced us to seek some matches at a much greater distance geographically, diluting the predictive value of the geographical match.

For expository purposes, we assume henceforth that only the block-group mean corresponding to the race of each respondent is available, but not the means corresponding to the other five races that are available simultaneously in the census data. This is more typical of data that would be collected directly from the respondent, where the race variable itself (as a modeling variable) is quite predictive because income data for people of different races reflect differences in income associated with race.

3.1 Matching and the Dataset

The addresses of over 90% of ungeocodable records have zip codes. Zip code was therefore chosen as a matching covariate. A simple diagnostic for its usefulness appears in section 3.2. The numerical sequence of zip codes does not always correspond to neighborhood distance relationships. For example, Cambridge, Massachusetts has a 02138 post office that also uses the 02238 zip code for mailboxes, and in nearby Boston there is a 02215 zip code that was carved out of the 02115 area. Instead of using the numerical sequence of zip codes, the distances between zip codes were computed based on latitudes and longitudes of their main post offices, under the assumption that two zip codes were closest to each other if their main post offices were closest to each other.

The colorectal cancer database has 1,696 ungeocodable records. The same number ($n^* = 1{,}696$) of geocodable records was randomly selected from the same database. For each of these 3,392 records, two matched geocodable cases were randomly chosen from its own zip code or (if necessary) neighboring zip codes. This created a dataset with $3{,}392 \times 3 = 10{,}176$ records. Note that n^* was a convenient choice, because the data were free. In general, the choice of n^* could affect both the total cost and the precision of the estimates. Both the randomly selected geocodable records and the matched cases were within-sample data and hence were retained in the analyses for Ayanian *et al.* (2003). We asked the cancer registry for these cases only because for confidentiality purposes we could not do the matching ourselves with the data (for the same cases) that we had in hand.

The modeling covariates used in the imputation model were the eight administrative-record variables: age, sex, race, marital status, cancer stage, chemotherapy treatment, cancer type and radiotherapy treatment, and category of treating hospital's American College of Surgeons accreditation as of 1999 (ACOS99). These variables are observed for all 10,176 records included in the imputation model. (Some of these variables are predictors and some are outcomes in the scientific models of the main analyses, but the distinction is irrelevant for imputation.) The census mean values Y_1, Y_2 and Y_3 are observed in geocodable records, but not in ungeocodable records. These variables were treated as outcome variables of the imputation model in section 2.3. The data structure is represented by Table 1.

Table 1
Structure of Data Used in Imputation for the Colorectal Cancer Study

Data*	Eight Modeling Covariates				Census Variables		
	Age	Sex	…	ACOS99	Y_1	Y_2	Y_3
Ungeocodable	√	√	…	√	?	?	?
First Match	√	√	…	√	√	√	√
Second Match	√	√	…	√	√	√	√
Geocodable	√	√	…	√	√	√	√
First Match	√	√	…	√	√	√	√
Second Match	√	√	…	√	√	√	√

* There were 1,696 records in each of the six types of data.
√ = observed ? = unobserved

Before we fitted the model, the percentage outcomes y_2 and y_3 were transformed using the scaled-logit function:

$$\log\left(\frac{(y-a)/(b-a)}{1-(y-a)/(b-a)}\right), \qquad (6)$$

with $a = -0.5$ and $b = 100.5$ so that after imputations the inverse transformation with rounding to the nearest integer would yield imputed values between 0 and 100 inclusive (Schafer 1999). Similarly, a log-transformation was applied to the income outcome y_1 so that the imputed incomes would be nonnegative. Note that the distributions of the transformed variables are closer to normality than they are on the original scale (Schafer 1997). To keep notation simple, we redefine y_1, y_2 and y_3 as their transformed versions.

3.2 Preliminary Diagnostics

A simple diagnostic test for the usefulness of the matching covariates is to compare the adjusted R^2 for the regression models predicting the three census variables with only the modeling covariates, the models with only the matching covariates, and the models with both. In this application, zip code was the only matching covariate. There were 1,133 distinct zip codes (hence 1,132 dummy variables) in the 8,480 fully observed records (the geocodable records and all first and second matches). Table 2 shows the adjusted R^2 for models with only the eight modeling covariates, models with only zip code, and models with both modeling covariates and zip code. The adjusted R^2 for models with both modeling covariates and zip code are higher than the corresponding ones for models with only one of the two covariate types. Our imputation procedure uses information from both matching and modeling covariates and thus can be expected to work better than procedures using only the matching or the modeling covariates (as shown by the simulation study in section 4). Although the contribution of the modeling covariates to R^2 is relatively modest, their inclusion is important for removing systematic biases and properly representing relationships that might be important in the scientific models.

Table 2
Adjusted R^2 for Alternative Regression Models

	Only Modeling Covariates	Only Matching Covariate (Zip Code)	Both Modeling and Matching Covariates
Median household income (INC)	0.091	0.453	0.496
Percent with no high school diploma (EDU)	0.115	0.452	0.503
Percent below poverty level (POV)	0.047	0.327	0.343
Model degrees of freedom[a]	26[b]	1,133	1,158
Sample sizes	8,480	8,480	8,480
Residual degrees of freedom	8,454	7,347	7,322

(a) With intercept.
(b) The modeling covariates are age, sex (2 levels), race (6 levels), marital status (6 levels), cancer stage (6 levels), chemotherapy treatment (2 levels), cancer type and radiotherapy treatment (3 levels), and category of treating hospital's American College of Surgeons accreditation as of 1999 (6 levels).

To determine whether a multivariate model was needed, we fitted a multivariate-outcome regression model with both

modeling covariates and zip code. The estimated correlations between the residuals were: $r_{12} \approx -0.194$, $r_{13} \approx -0.297$, and $r_{23} \approx 0.357$, where "variable 1" is median household income, "variable 2" is percent with no high school diploma, and "variable 3" is percent below poverty level. These estimates were significantly different from zero, which therefore indicated that multivariate versions of the methods in section 2.3 should be used to generate imputations.

3.3 Multiple Imputation Results and Comparisons

Imputations under NpMMM were used in the study of factors predicting provision of chemotherapy for colorectal cancer patients (Ayanian *et al.* 2003). Their model included three indicator variables for ranges of contextual income, together with 21 other variables representing patient and hospital characteristics. The multiple imputation analysis shows that the information loss due to missing information is always less than 0.1%, which is much smaller than the fraction of ungeocodable records (3.3%). As expected, the largest fractions of missing information appeared for the income variables. The scientific results in Ayanian *et al.* (2003) would not have changed dramatically if the incomplete cases had been dropped. In this type of research, however, every case is precious and expensive, and saving the 3.3% with missing data was a contribution to the study.

For comparison, variances of parameters under the complete-case analysis were on the average 4.0% larger than those under multiple imputation analysis. Such percentage differences are close to the fraction of incomplete cases deleted for this analysis. When the imputations generated by our method were included in the scientific analysis, the precision of the estimate of the "rural" effect was dramatically improved (using only the complete cases led to 41.6% increase in variance), due to the concentration of ungeocodable records in rural areas (21.6% of rural records are ungeocodable, but only 3.1% of nonrural records are ungeocodable).

4. A Simulation Study

This simulation study compares performance of our new method with three other commonly-used nonresponse adjustment methods. The population of this study was the 1,696 fully observed triples – the 1,696 geocodable records and the corresponding first and second matches (one row from each of the last three horizontal blocks in Table 1) – or 5,088 observations. For simplicity, we assumed that the triples were from distinct zip codes (clusters), hence $i = 1, 2, ..., I = 1{,}696$. Each cluster i contained three units $(u = 1, 2, 3)$, and the record of each unit consisted of $\boldsymbol{x}_{iu}$ (the covariates) and $\boldsymbol{y}_{iu}$ (the census variables).

4.1 Simulated Data and Response Mechanism

Assuming that the design was cluster sampling with sample size 800, we drew random samples of 800 clusters. For each random sample, about half of the 800 clusters were randomly selected to have an ungeocodable record in which the census variables were unobserved, with the probability of missingness depending on an individual's race and on the mean income of the cluster (zip code). We simulated missingness under a multinomial logit model where the outcomes are: nothing unobserved $(w_{i0} = 1)$, $\boldsymbol{y}_{i1}$ unobserved $(w_{i1} = 1)$, $\boldsymbol{y}_{i2}$ unobserved $(w_{i2} = 1)$, and $\boldsymbol{y}_{i3}$ unobserved $(w_{i3} = 1)$. Specifically, for each $i = 1, 2, ..., I$, let $z_{i0} = 0$ and

$$z_{iu} = a + b \times I(\text{unit } iu \text{ is White}) + c \times (\text{mean income in zip code } i) \quad (7)$$

where $u = 1, 2, 3$. Then

$$\Pr(w_{iu} = 1) = \exp(z_{iu}) \Big/ \sum_{u=0}^{3} \exp(z_{iu}) \quad \text{for } u = 0, 1, 2, 3. \quad (8)$$

The results of this simulation study were based on datasets generated by the mechanism with $a = -1$, $b = 11$ and $c = 0.0003$, which made about 17% of the units in a random sample ungeocodable, with probability of geocoding positively related to White race and higher block-level income. The task was to use the random sample to estimate $\bar{\boldsymbol{y}}$, the mean values of the population (1,696 clusters).

The simulation conditions described in the preceding paragraphs were designed to give a stringent test of the procedure and alternatives by exaggerating the impact of unobserved data and making the missingness strongly related to characteristics both of the individual and of the area. We were not attempting to simulate the exact conditions of the application in section 3 but rather to use an artificial population with similar distributions to those in the real population to illustrate the workings of our method and its competitors.

4.2 Inferential Methods and Measures of Performance

Preliminary results indicated that the performance of PMMM and NpMMM is similar; NpMMM is, however, simpler (especially in analyses with multivariate outcomes), because the method does not require explicit parametric modeling of the residual variance. Our simulations compared performance of NpMMM (using two matched cases per record) with three other commonly-used nonresponse adjustment methods:

1. **Complete-case Method (CCM)**
   The population means are estimated from all geocodable units of a random sample.
2. **Substitute Single Imputation (SSI)**
   This is the traditional use of substitutes. The unobserved census variables of each ungeocodable unit are replaced by the values of the census variables of a randomly selected unit from the same cluster. The resulting sample is treated as if there had been no ungeocodable unit; all 800 clusters in such a sample are used for estimating the population means.
3. **Multivariate Normal Multiple Imputation (MNMI)**
   This method uses only one randomly selected unit from each of the fully observed clusters in a random sample to fit the multivariate normal linear regression

$$\boldsymbol{y}_i^T \sim \mathrm{N}(\boldsymbol{\beta}_0^T + \boldsymbol{x}_i^T\ \mathbf{B},\ \boldsymbol{\Sigma}),$$

   with a noninformative prior on the parameters. The model is then used to create m sets of multiple imputations for the unobserved census variables using a direct multivariate generalization of the algorithm given by Rubin (1987, page 167).

Note that CCM uses *neither* matching nor modeling covariates, SSI uses *only the matching covariate* (zip code), MNMI uses *only the modeling covariates*, and NpMMM uses *both* the matching covariate and the modeling covariates.

The CCM and SSI data are analyzed by the usual complete-data method which estimates the population mean from the data with the appropriate estimator for cluster sampling from a finite population, including the finite population correction (Cochran 1977, Chapters 9–10). Both MNMI and NpMMM produce m sets of complete data, each of which is analyzed by the same complete-data method used for the CCM and SSI data; the m sets of point and variance estimates are then combined using the multiple imputation combination rule (Rubin 1987; Schafer 1997, pages 108–110).

For each simulation $t \in \{1, 2, ..., T\}$, we denote the point estimates from the four methods by $\bar{y}_{\mathrm{CC}}(t)$, $\bar{y}_{\mathrm{SS}}(t)$, $\bar{y}_{\mathrm{MN}}(t)$, and $\bar{y}_{\mathrm{Np}}(t)$, and the means of these quantities across simulations are written as $\bar{y}_{\mathrm{CC}}$, $\bar{y}_{\mathrm{SS}}$, $\bar{y}_{\mathrm{MN}}$, and $\bar{y}_{\mathrm{Np}}$. Performance evaluation of the four nonresponse adjustment methods will be based on three measures:

1. **Percent reduction in the average bias of an estimator relative to the average bias of the CCM estimator**. Denote the average bias of an estimator by $\bar{\boldsymbol{b}}_{\mathrm{E}}$. Then

$$\bar{\boldsymbol{b}}_{\mathrm{E}} = \bar{\boldsymbol{y}}_E - \bar{\boldsymbol{y}},$$

   where $E \in \{\mathrm{CC}, \mathrm{SS}, \mathrm{MN}, \mathrm{Np}\}$. We define the percent reduction in the average bias of an estimator relative to the average bias of the CCM estimator as

$$R(\bar{b}_{\mathrm{E}},\ \bar{b}_{\mathrm{CC}}) = \frac{|\bar{b}_{\mathrm{CC}}| - |\bar{b}_{\mathrm{E}}|}{|\bar{b}_{\mathrm{CC}}|},$$

   where $\bar{b}_{\mathrm{E}}$ is an element of $\bar{\boldsymbol{b}}_{\mathrm{E}}$ and $\bar{b}_{\mathrm{CC}}$ is the corresponding element in $\bar{\boldsymbol{b}}_{\mathrm{CC}}$. By definition, $R(\bar{b}_{\mathrm{CC}},\ \bar{b}_{\mathrm{CC}})$ is zero.
2. **Estimated coverage of the nominal 95% confidence intervals for $\bar{\boldsymbol{y}}$**. Intervals produced by the CCM or SSI estimates were constructed under appropriate $t-$distributions. For intervals associated with the MNMI or NpMMM estimates, we followed the procedure outlined in Schafer (1997, pages 109–110) and replaced the degrees of freedom v with the updated version of Barnard and Rubin (1999).
3. **Estimated fraction of missing information about $\bar{\boldsymbol{y}}$**. For each of MNMI and NpMMM, we computed $\hat{\lambda}$, an estimate of the fraction of missing information about $\bar{\boldsymbol{y}}$ (see Barnard and Rubin (1999) for the most recent expression).

4.3 Results

The simulation procedure was implemented 2,000 times, and $m = 10$ was used for MNMI and NpMMM. The mean values of the census variables in the population were $\bar{\boldsymbol{y}} = (40{,}642,\ 21.65,\ 9.55)^T$. The average bias of the CCM estimator was $\bar{\boldsymbol{b}}_{\mathrm{CCM}} = (-5{,}405,\ -3.97,\ -1.79)^T$. Other results are summarized in Table 3. NpMMM achieved large percent reductions in relative average bias (95.0% to 99.5%). SSI reduced biases more than MNMI, because the matching covariate (zip code) was much more informative than the set of modeling covariates (section 3.2). Since the response mechanism was *nonignorable* (the response probabilities depended partly on income), the poor performance of MNMI, which did not use the geographical information to help predict income, was expected. Note that MNMI is biased, and the bias is large enough so that with the sample size considered in this paper the confidence intervals never covered the hypothetical population values.

Under MNMI and NpMMM, the percent of missing information was much less than the average percent of unobserved data. The percent of missing information was smaller under NpMMM than under MNMI. Only NpMMM produced well calibrated intervals with correct coverage. In summary, NpMMM combines the best features of the other two methods – close-to-nominal coverage and less missing information.

Table 3
Simulation Results[a]: Bias Reduction, Coverage, and Fraction of Missing Information

Measure	Mean	Method NpMMM	MNMI	SSI
Percent bias	INC	99.5	44.6	95.2
Reduction	EDU	95.0	40.6	83.7
$100R(\bar{b}_E, \bar{b}_{\text{CCM}})$[b]	POV	96.8	32.6	80.3
Estimated	INC	95.1	0.00	89.8
Coverage of the	EDU	94.8	0.00	65.7
95% CIs[c]	POV	95.2	0.00	66.0
$100\times$Estimated	INC	1.00	9.92	
fraction of missing	EDU	0.05	0.07	
information $\hat{\lambda}$[d]	POV	0.07	0.08	

(a) Based on 2,000 replications and $m = 10$.
(b) By definition, $100R(\bar{b}_{\text{CCM}}, \bar{b}_{\text{CCM}}) = 0$.
(c) Results for the CCM estimates were all zeros.
(d) The average percent of unobserved data was approximately 17%.

5. Conclusion

This work extends Rubin and Zanutto (2001) in two respects. First, our method allows more than one matched case per record. We show theoretically that the efficiency of an imputation increases as the number of matched cases per record increases. When the cost of matched cases is relatively low, our method offers an option where information of more than one matched case per record is used to help fit imputation models at a negligible computational expense. Second, NpMMM does not require explicit parametric modeling of residual variance(s), hence simplifying the modeling task (especially for analyses with multivariate outcomes). This nonparametric approach makes it feasible to apply our method to datasets with complex model structures. In a simulation study, NpMMM estimates achieved substantial bias reductions, and NpMMM produced confidence intervals with correct coverage.

Although we have focused on geographically-based matching to complete unobserved geographically-linked variables, the procedures described in this paper can be generalized to other matching variables. For example, to impute clinical variables, it might be more appropriate to match to another patient in the same hospital, if clinical characteristics and therapies are likely to be more strongly associated with the hospital than with the geographic location of the patient's residence.

Acknowledgements

This research was supported in part by the Bureau of the Census through a contract with the National Opinion Research Center and Datametrics, Inc., and by a grant from the Agency for Healthcare Research and Quality (AHRQ) and the National Cancer Institute (HS09869). The authors thank John Z. Ayanian for leadership of the Quality of Cancer Care research project, Mark Allen and Robert Wolf for preparation of data, Bill Wright for his support to this research, and the associated editor and two anonymous referees for their helpful comments.

References

Ayanian, J.Z., Zaslavsky, A.M., Fuchs, C.S., Guadagnoli, E., Creech, C.M., Cress, R.D., O'connor, L.C., West, D.W., Allen, M.E., Wolf, R.E. and Wright, W.E. (2003). Use of adjuvant chemotherapy and radiation therapy for colorectal cancer in a population-based cohort. *Journal of Clinical Oncology*, 21, 1293-1300.

Barnard, J., and Rubin, D.B. (1999). Small-sample degrees of freedom with multiple imputation. *Biometrika*, 86, 948-955.

Chiu, W.F., Yucel, R.M., Zanutto, E. and Zaslavsky, A.M. (2001). Using matched substitutes to improve imputations for geographically linked databases. *Proceedings of the Survey Research Methods Section*, American Statistical Association.

Cochran, W.G. (1977). *Sampling Techniques*. New York: John Wiley & Sons, Inc.

Kalton, G. (1983). *Compensating for Missing Survey Data*. Research Report Series, Ann Arbor, MI: Institute for Social Research.

Kalton, G., and Kasprzyk, D. (1986). The treatment of missing survey data. *Survey Methodology*, 12, 1-16.

Krieger, N.,Williams, D. and Andmoss, N. (1997). Measuring social class in U.S. public health research: Concepts, methodologies, and guidelines. *Annual Review of Public Health*, 18, 341-378.

Lessler, J.T., and Kalsbeek, W.D. (1992). *Nonsampling Errors in Surveys*. New York: John Wiley & Sons, Inc.

Little, R.J.A., and Rubin, D.B. (1987). *Statistical Analysis with Missing Data*. New York: John Wiley & Sons, Inc.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons, Inc.

Rubin, D.B., and Zanutto, E. (2001). Using matched substitutes to adjust for nonignorable nonresponse through multiple imputations. In *Survey Nonresponse*, (Eds. R. Groves, R. Little and J. Eltinge), New York: John Wiley & Sons, Inc., 389-402.

Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall.

Schafer, J.L. (1999). NORM: Multiple imputation of incomplete multivariate data under a normal model, version 2. Software for Windows 95/98/NT available at http://www.stat.psu.edu/~jls/misoftwa.html.