USING MATCHED SUBSTITUTES TO IMPROVE IMPUTATIONS FOR GEOGRAPHICALLY LINKED DATABASES[1]

Calvin W.F. Chiu, Harvard University
Recai M. Yucel, Harvard Medical School
Elaine Zanutto, The Wharton School, University of Pennsylvania
Alan M. Zaslavsky, Harvard Medical School

Calvin W.F. Chiu, Department of Statistics, One Oxford Street, Cambridge, MA 02138

Key Words: Nonresponse, Multiple Imputation, Administrative Records, Contextual Variables.

Abstract

Local-area characteristics from the census are often a useful supplement to variables in databases created from administrative records, when geographical links to census block groups can be established. In large databases, some records might not have adequate address information to permit geocoding beyond zip code; hence, no links could be made to census block groups. Treating these ungeocodable cases as unit nonrespondents, we propose a new method that uses matched substitutes and regression modeling to create multiple imputations for such missing values.

1. Motivation

In a study of treatment patterns for colorectal cancer patients, characteristics such as age, gender and race/ethnicity were available from hospital files and insurance records. In this study, investigators also believed that variables such as income and education level could be useful in model building and prediction. Unfortunately, no individual measurements for these covariates were available from the administrative records. Instead, mean values of these variables were obtained from U.S. Census Bureau records for small geographical areas (census block groups or tracts) including the subject's residence were used as regressors to estimate income and education effects. Use of such "contextual variables" is a common procedure in epidemiological and health services research (Krieger et al, 1997). Such analyses often produce broadly similar results 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 context; in a purely contextual analysis, these effects are confounded. Nonetheless, associations between contextual characteristics and quality of care would suggest an equity problem, regardless of whether they primarily reflect individual or community-level relationships.

In the colorectal cancer database, a small but substantial percentage of records (about 4.1%, or 2084 cases) were not geocodable due to insufficient address information, and hence no values could be obtained for these cases through linkage to their corresponding census block groups. As suggested by Zanutto (1998), the availability of information from both administrative records and the census for geocodable cases (respondents) makes it possible to fit a model to estimate the relationship between the information in the two sources. This model can then be used to impute data for the ungeocodable cases (nonrespondents) based on their administrative records. We propose a similar strategy for using matched substitutes to impute data that are missing for ungeocodable cases in geographically linked databases. The matched substitutes allow us to incorporate small area effects into the imputations without having to explicitly model separate effects for each small area. This work is the first real-data application of the methodology proposed by Rubin and Zanutto (2001).

2. Imputation Methodology

Rubin and Zanutto (2001) proposed a method called "matching, modeling, and multiply imputing" (MMM) that uses matched substitutes to help impute for missing data due to nonresponse in sample surveys. In this approach, substitutes are selected for nonrespondents using background covariates, which are available prior to the survey and are convenient for matching, to obtain responses from survey units that appear to be similar to the nonrespondents. Hence, they referred to their substitutes as "matched" substitutes. Rather than the usual approach of using the substitutes directly to replace nonrespondent data, the method uses the matched substitutes along with respondent information

[1] This research was supported 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).

and the background covariates of the nonrespondents to build a model to multiply impute the missing data. To help fit this model, substitutes are also chosen for some respondents. Once the missing responses have been multiply imputed, the substitutes (for both respondents and nonrespondents) are discarded.

The methodology is designed to work well in realistically complex situations. In particular, it accommodates systematic differences between respondents and nonrespondents as well as between nonrespondents and their substitutes. In addition to the fact that substitutes are, by definition, respondents and therefore may be systematically different from their matching non-respondents, it is impractical to match substitutes to nonrespondents on all relevant covariates. For example, suppose that age and address are available for all units in the population prior to sampling. It may be feasible to choose substitutes for nonrespondents by matching on address (e.g., choosing a neighbor to be a substitute), but it may not be easy to include address information in a statistical model. Covariates like address are referred to as matching covariates, whereas covariates that can be included in statistical models are called modeling covariates (Rubin and Zanutto, 2001). Though age can be both a matching and modeling covariate, it may be difficult to find substitutes that are similar to the nonrespondents on both address and age. Therefore, one may choose not to match on age, and match only on address. If both the probability of response and the value of the survey outcome are related to age, then the outcomes for nonrespondents and their substitutes will be systematically different due to differences in age. In that case, age is treated as a modeling covariate and will be included in the multiple imputation model to adjust for the observed differences.

2.1 Matching

Matches for ungeocodable cases can be obtained by making random selections from a pool of all geocodable cases in the same zip code; when the desired number of matches could not be achieved within the same zip area, the selection process is expanded to the nearest zip areas until all matches have been found. In our analysis, we used two substitutes per nonrespondents, but theoretically one could use any number of substitutes. As suggested by Rubin and Zanutto (2001), substitutes are also chosen, in similar fashion, for some randomly selected geocodable cases. In this study, all matches were obtained from the same colorectal cancer database. In general, substitutes need not be necessarily drawn from the same population where the nonrespondents and respondents originated. For example, one can select substitutes for colorectal cases from a general population of cancer patients, and then fit a model to correct for differences.

2.2 Modeling and Multiply Imputing

Suppose our model is

$$y_{ij} = \beta_0 + \mathbf{x}_{ij}\boldsymbol{\beta} + d_i + \varepsilon_{ij},$$

where i indexes small area (e.g., zip code), j indexes unit within area, and $\mathbf{x}_{ij}$ and $\boldsymbol{\beta}$ are covariate and coefficient vectors. This model includes a regression prediction $\beta_0 + \mathbf{x}_{ij}\boldsymbol{\beta}$, a small-area effect d_i, and a unit-specific residual ε_{ij}. We assume that d_i follows some distribution F_d with $E(d_i) = 0$, and ε_{ij} follows some distribution F_ε with mean zero and variance σ^2 (assuming for the moment that y is univariate, an assumption which we will relax shortly).

If we are given pairs of units y_{i1} and y_{i2} within the same small area, i.e.,

$$y_{i1} = \beta_0 + \mathbf{x}_{i1}\boldsymbol{\beta} + d_i + \varepsilon_{i1},$$
$$y_{i2} = \beta_0 + \mathbf{x}_{i2}\boldsymbol{\beta} + d_i + \varepsilon_{i2},$$

then we estimate $\boldsymbol{\beta}$ from the within-area regression

$$(y_{i1} - y_{i2}) = (\mathbf{x}_{i1} - \mathbf{x}_{i2})\boldsymbol{\beta} + (\varepsilon_{i1} - \varepsilon_{i2}),$$

where the constant term and the small area effect drop out. The residuals from this regression have a symmetrical distribution with variance $2\sigma^2$. Assume for the moment that we have a way of drawing from the posterior distribution of $\boldsymbol{\beta}$ and β_0, and we carry out all the rest of this analysis conditional on that draw.

Now suppose that we are interested in imputing for a third unit in the same small area. Assuming a flat prior for d_i, the posterior distribution for $d_i \mid y_{i1}, y_{i2}, \mathbf{x}_{i1}, \mathbf{x}_{i2}, \beta_0, \boldsymbol{\beta}$ has mean

$$\frac{y_{i1} + y_{i2}}{2} - \beta_0 - \left(\frac{\mathbf{x}_{i1} + \mathbf{x}_{i2}}{2}\right)\boldsymbol{\beta},$$

and variance $\sigma^2/2$. Hence the predictive distribution for $y_{i3} \mid y_{i1}, y_{i2}, \mathbf{x}_{i1}, \mathbf{x}_{i2}, \mathbf{x}_{i3}, \boldsymbol{\beta}$ has mean

$$\frac{y_{i1} + y_{i2}}{2} + \left(\mathbf{x}_{i3} - \frac{\mathbf{x}_{i1} + \mathbf{x}_{i2}}{2}\right)\boldsymbol{\beta},$$

and variance $3\sigma^2/2$ which is the sum of σ^2, the predictive variance under the model conditional on all parameters, and $\sigma^2/2$, the posterior variance of d_i. One also could use predictors that only utilize part of the data (i.e. only y_{i1} or only y_{i2}). Since the estimator presented above uses data from both y_{i1} and y_{i2}, it is therefore more efficient than an estimator that uses only one of y_{i1} and y_{i2}. 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 the multivariate-outcome case).

There are several approaches to draw residuals. For single dimension, one could estimate the residual variance and make independent draws under univariate normality. To generalize to multiple dimensions, the corresponding approach would estimate the residual covariance matrix and then draw under multivariate normality. To save investigators from having to model a covariance matrix and to relax the normality assumption, we propose sampling the residuals *jointly* from the within-area regressions and multiply them by $\sqrt{3/4}$ to adjust residuals with variance $2\sigma^2$ so that they have variance $3\sigma^2/2$ as desired. This may be a little overdispersed if the residuals are long-tailed since $\varepsilon_{i3} - (\varepsilon_{i1} + \varepsilon_{i2})/2$ may be closer to normality than $\varepsilon_{i1} - \varepsilon_{i2}$. On the other hand, the former is asymmetrical and the latter is symmetrical. Simulation results suggest that the above simple rescaling gives a reasonably good approximation for our data. If we believe the model might be heteroskedastic (in the general sense that the residual distribution is related to x, not necessarily just a change of scale as in univariate normal models), we could draw residuals within classes of observations believed to be "similar" with respect to residual variation.

To create multiple imputations for missing values of a unit in the same small area, we first fit a within-area regression for each dimension and save the residuals. Then we repeat the following two steps several times:

1. Draw β_0 and $\boldsymbol{\beta}$ under the model. For example, $(\beta_0, \boldsymbol{\beta})^T \sim N((\hat{\beta}_0, \hat{\boldsymbol{\beta}})^T, (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2)$ if ε_{ij} are iid $N(0, \sigma^2)$.
2. For each missing case, calculate the predictive mean under the model and then add a randomly-sampled residual times $\sqrt{3/4}$.

3. Application: Colorectal Cancer Study

The main colorectal cancer database has a total of 50,740 patient records. Approximately 96% are geocodable and 4% are ungeocodable. Among the ungeocodable, about 50% have P.O. box addresses (often in a rural area); the rest have mistyped addresses, or addresses that lie in a new housing development and therefore is not in the address databases.

Researchers were interested in obtaining local-area characteristics from the census as contextual predictors of treatment processes. For geocodable cases where links to census block groups could be established, the following census measurements were available:

$Y_1 =$ Median Household Income,

$Y_2 =$ Percent with no High School Diploma,

$Y_3 =$ Percent in Poverty;

each of which had values for each of six race groups (Asian/Pacific Islanders, Blacks, Hispanics, Native American/Eskimo/Aleutian Islanders, Whites, and Others). No contextual values could be obtained for ungeocodable cases because their address information were not sufficient to identify the corresponding census block groups. We applied the methodology described in Section 2 to create multiple imputations for these unobserved values.

3.1 Transformations

To better fit the regression model, a scaled logit transformation (see "transforming variables" in help topics for the imputation software NORM: Schafer, 1999) was applied to each of the two percentage variables y_2 and y_3. The transformed values were obtained by

$$\log\left(\frac{(y_i - a)/(b - a)}{1 - (y_i - a)/(b - a)}\right) \text{ for } i = 2, 3,$$

with $a = -0.5$ and $b = 100.5$. Upon completion of imputations, applying the inverse transformation and rounding to the nearest integer ensured that all imputed values were integers between 0 and 100 inclusively.

For the race-specific median income variables, we observed from the main database that they were truncated (bottom- and top-coded) at \$2,500 to \$100,000. Though in some blocks certain race groups were shown as having zero representation and hence were coded as having \$0 median household income, it did not necessarily mean that there were actually no members observed for these groups. Rather it is very likely that the observed counts were so small that they were rounded down to zero for confidentiality reasons; and as a result these groups were labeled as having \$0 median household income. In fact, in the main database, many of the blocks showing zero representation in certain race groups had the bottom-coded median income for other races at the same block. Because of this and the fact that less than 1% of our data have zero median incomes, for simplicity, these zeros were replaced with the bottom-coded value \$2,500; then a log-transformation was applied to these observed median household income y_1, i.e. $\log y_1$. To avoid clumsy notation, hereafter y_1, y_2 and y_3 represent their transformed versions.

3.2 Matching, Modeling & Multiply Imputing

Preliminary analyses reported about 91% (1,888 out of 2,084) of the ungeocodable cases have zip code information. For simplicity, we used zip code as a convenient definition for neighborhoods, our matching covariate. In some situations, the numerical sequence of zip codes does not correspond to the implied neighborhood relationships. For example, locally we have a 02138 post office that also uses the 02238 zip code for mailboxes; there is also a 02215 zip code that was carved out by splitting the 02115 area. To capture more realistically the distances between neighborhoods, we used the latitude and longitude of the post office that goes with each zip code in our neighborhood definition. Fortunately, all the 1,888 cases have such latitude and longitude information. To help fit the model, some geocodable cases (1,882 in total) were randomly selected from the main database.

According to the procedure described in Section 2.1, two matches (first match, second match) were selected for each of the 1,888 ungeocodable cases and each of the 1,882 geocodable cases. Figure 1 displays the structure of the data after matching.

Figure 1: Data Structure

Data	*Administrative Information*			*Census Contextual variables*		
	AGE	…	ACOS99	y_1	y_2	y_3
Geo. 1,882 cases	✓	…	✓	✓	✓	✓
	⋮	⋱	⋮	⋮	⋮	⋮
	✓	…	✓	✓	✓	✓
First Matches for **Geo.**	✓	…	✓	✓	✓	✓
	⋮	⋱	⋮	⋮	⋮	⋮
	✓	…	✓	✓	✓	✓
Second Matches for **Geo.**	✓	…	✓	✓	✓	✓
	⋮	⋱	⋮	⋮	⋮	⋮
	✓	…	✓	✓	✓	✓
UnGeo. 1,888 cases	✓	…	✓	**?**	**?**	**?**
	⋮	⋱	⋮	⋮	⋮	⋮
	✓	…	✓	**?**	**?**	**?**
First Matches for **UnGeo.**	✓	…	✓	✓	✓	✓
	⋮	⋱	⋮	⋮	⋮	⋮
	✓	…	✓	✓	✓	✓
Second Matches for **UnGeo.**	✓	…	✓	✓	✓	✓
	⋮	⋱	⋮	⋮	⋮	⋮
	✓	…	✓	✓	✓	✓
✓ = Observed, ? = Missing						

Census contextual variables refer to the race-specific census measurements Y_1, Y_2 and Y_3 defined earlier. The values for these three variables are observed for geocodable cases, but are missing for the ungeocodable records. Administrative information is a collection of patient characteristics in the main colorectal cancer database. To avoid complexity, subsequence analysis will be carried out using only the seven modeling covariates presented in Table 1. These covariates are believed to be somewhat correlated with the three dependent variables Y_1, Y_2 and Y_3.

Each pair of matches corresponds to y_{i1} and y_{i2} (defined in Section 2.2) with observed modeling covariates $\mathbf{x}_{i1}$ and $\mathbf{x}_{i2}$ respectively. All matches are race-specific to allow for projection to the race of an ungeocoded person at the imputation stage. Following the steps described in Section 2.2, separate regression models were fitted for each race. For a particular race, we regressed $(y_{i1} - y_{i2})$ on $(\mathbf{x}_{i1} - \mathbf{x}_{i2})$ separately for each of the three dimensions: median household income Y_1, percent with no high school diploma Y_2, and percent in poverty Y_3. The residuals from these within-area regressions were then saved. Each imputed value was the sum of the predictive mean for the 3×1 vector $(Y_1, Y_2, Y_3)^T$ and $\sqrt{3/4}$ of a randomly-sampled triple of residuals. Five sets of imputations were created. In short, each block group had six sets of the three variables, one for each race; we fitted models separately for each of these and then imputed whichever one was needed.

Table 1: Seven modeling covariates extracted from the main colorectal cancer database

Variable	**Type**	**Range/Possible values**
Patient's Age at diagnosis	C	11 – 104
Patient's Gender	N	1, 2
Patient's Marital Status at diagnosis	N	1, 2, 3, 4, 5, 9
Type of Cancer and Radiotherapy Treatment	N	C = Colon, RR = Rectum with Radiotherapy, R = Rectum without Radiotherapy
Cancer Stage	O	00, 10, 20, 25, 30, 40
Chemotherapy Treatment	N	0 = No, 1 = Yes
ACOS category of approval, 1999	N	1, 2, 3, 4, 5, 9
C = Continuous, N = Nominal, O = Ordinal		

3.3 Multiple-Imputation Inference

To illustrate and evaluate the multiple-imputation inference, we treated the data used in these analyses as if they were the entire dataset and performed inferences for means of each of the three contextual variables Y_1, Y_2 and Y_3. Based on the rules for combining complete-data inferences, we present in Table 2 a summary of the multiple-imputation $(m = 5)$ inferences, where

Q = the complete-data point estimate,

$\overline{Q}$ = the average of the complete-data point estimates over the five imputed datasets,

$\overline{U}$ = the within-imputation variance estimate,

B = the between-imputation variance estimate,

$\hat{\lambda}$ = an estimate of the fraction of missing information about Q.

Detailed expressions for the above quantities can be found in Rubin (1987, Chap. 3) or Schafer (1997, p. 109-110).

We can see from Table 2 that the estimates of the fraction of missing information about the Q's are significantly less than the fraction of missing data, which is $1{,}888/(3\times1{,}888+3\times1{,}882)\approx 0.17$. This implies that, compared to complete data estimates for each Y, we have achieved more efficient estimates for the Q's using the sets of imputed data generated from our methodology described in Section 2.

Table 2: Multiple-Imputation Inference

Q	$\overline{Q}$	$\overline{U}$	B	$\hat{\lambda}$
$\overline{Y}_1$	0.1375	0.1120	0.0186	0.0581
$\overline{Y}_2$	0.1120	0.1300	0.0246	0.0742
$\overline{Y}_3$	38,060	110,400	48,383	0.0938

All $\hat{\lambda} < 0.17 =$ *Fraction of Missing Data*

4. Summary

Motivated by Rubin and Zanutto (2001), we propose a similar strategy that uses matched substitutes to improve imputations. The methodology is practical, flexible, and easy to be implemented under multiple dimensions. It has been successfully implemented for the colorectal cancer database. Five sets of imputed data were created from the procedures described in Section 2. Based on these imputed data, we estimated the fraction of missing information about the mean for each of the three contextual predictors. Each of which was substantially less than the fraction of missing data. This suggests that the imputed datasets generated by our model produce more efficient estimates than the corresponding complete-data version. The imputed data have been used in our analyses of distribution of services for colorectal cancer patients.

In future work, we might attempt to fit more sophisticated models using more covariates from the main colorectal cancer database, and quantify the gains relative to the main effects model fitted in Section 3. Methods and quantitative measures should be developed to assess the properties and to evaluate the "goodness" of the imputed values generated by the presented methodology. In conclusion, we have demonstrated that imputation methodology can be useful for researchers working with geographically linked databases when some cases cannot be fully geocoded to the level at which the linkage is being made.

References

Krieger N., Williams D., Moss N. (1997). *Measuring Social Class in US Public Health Research: concepts, Methodologies, and Guidelines*. Annual Review of Public Health, **18**: 341-378.

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

Rubin, D.B., and Zanutto, E. (2001). Using Matched Substitutes to Adjust for Nonignorable Nonresponse through Multiple Imputations. To appear in *Survey Nonresponse*, edited by R. Groves, R.J.A. Little, and J. Eltinge. New York: John Wiley.

Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data*. 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 from: http://www.stat.psu.edu/~jls/misoftwa.html.

Zanutto, E. (1998). Modeling matched substitutes to create multiple imputations for unit nonrespondents. *ASA Proceedings of the Section on Government Statistics and Section on Social Statistics*, 197-202.