Appl. Statist. (1993)
42, *No.* 3, *pp.* 545–556

Analyses of Public Use Decennial Census Data with Multiply Imputed Industry and Occupation Codes

By NATHANIEL SCHENKER† and DONALD J. TREIMAN

University of California, Los Angeles, USA

and LYNN WEIDMAN†

United States Bureau of the Census, Washington DC, USA

[Received October 1990. Final revision October 1992]

SUMMARY

This paper gives a brief introduction to multiple imputation for handling non-response in surveys. We then describe a recently completed project in which multiple imputation was used to recalibrate industry and occupation codes in 1970 US census public use samples to the 1980 standard. Using analyses of data from the project, we examine the utility of analysing a large data set having imputed values compared with analysing a small data set having true values, and we provide examples of the amount by which variability is underestimated by using just one imputation rather than multiple imputations.

Keywords: Missing data; Multiple imputation; Non-response; Sample surveys

1. Introduction

A standard technique for handling non-response in surveys is to impute (i.e. to fill in) a value for each missing datum. This procedure is especially well suited to databases distributed to the public by an organization such as the United States Bureau of the Census because

(a) the non-response problem is fixed by the data producer, who typically knows more about the reasons for non-response and has more information available than the subsequent data analysts and
(b) the filled-in data set can be analysed by using standard methods designed for complete data.

A drawback of the procedure, however, is that the application of standard complete data methods to a data set completed by imputation ignores the uncertainty due to using imputed rather than true values for the missing data. Even assuming that an appropriate model for the missing values has been formulated, a single imputed value cannot represent the random variability conditional on the model. This results

†*Address for correspondence*: Department of Biostatistics, School of Public Health, University of California, Los Angeles, CA 90024-1772, USA.

0035–9254/93/42545 $2.00

in inferences that are too sharp; for example, interval estimates typically have lower than nominal coverage.

Multiple imputation (Rubin, 1978, 1987a) alleviates this drawback by replacing each missing datum with two or more values drawn from the predictive distribution of values under the posited model for non-response. The result is two or more completed data sets, each of which is analysed by using the same standard method. The analyses are then combined in a way that reflects the extra variability due to imputation. Multiple imputation can be carried out under several different models for non-response to display sensitivity to the choice of model as well.

In this paper, we describe the use of multiple imputation to create public use data from the US census with industry and occupation classifications that are comparable over time. Section 2 discusses how multiple imputations are created and analysed in general. In Section 3, we describe the industry and occupation project. Some simple analyses of data from the project are presented in Section 4 to demonstrate the gains due to multiple imputation. We conclude with a discussion in Section 5.

2. Multiple Imputation

This section discusses briefly the ideas underlying multiple imputation and the methods used to analyse a multiply imputed data set. For further discussion, see Rubin and Schenker (1986) and Rubin (1987a).

2.1. *General Ideas*

Multiple imputation was developed and can be justified most easily from the Bayesian perspective. Let Y_{obs} and Y_{mis} denote the observed and missing values in a survey respectively, and let Q denote the population quantity of interest. Suppose that, if the missing values were known, inferences would be in the form of a distribution for Q with density $g(Q \mid Y_{\text{obs}}, Y_{\text{mis}})$ (the 'complete data posterior density'). Since the missing values are not known, however, a model must be fitted which predicts the missing values given the observed values; suppose that this results in a posterior predictive density $f(Y_{\text{mis}} \mid Y_{\text{obs}})$. The ultimate goal is to obtain the posterior density of Q (given the observed data), which can be expressed as

$$h(Q \mid Y_{\text{obs}}) = \int g(Q \mid Y_{\text{obs}}, Y_{\text{mis}}) f(Y_{\text{mis}} \mid Y_{\text{obs}}) \, \mathrm{d}Y_{\text{mis}}. \tag{2.1}$$

Equation (2.1) shows that the posterior density of Q can be obtained by averaging the complete data posterior density over the predictive distribution of the missing values. Thus, one way to compute $h(Q \mid Y_{\text{obs}})$ (say, at a particular value Q_0) would be to draw values of Y_{mis} repeatedly from $f(Y_{\text{mis}} \mid Y_{\text{obs}})$, to calculate $g(Q_0 \mid Y_{\text{obs}}, Y_{\text{mis}})$ separately for each draw and then to average the values. In principle, multiple imputations are repeated draws from the predictive distribution of Y_{mis} under the posited model for the missing data. Hence, multiple imputation allows the data analyst to approximate equation (2.1) by analysing the completed data set under each draw and then combining the analyses.

2.2. *Approximation to Posterior Distribution*

The exact computation of equation (2.1) by simulation would require

(a) an infinite number of draws of Y_{mis} from its predictive distribution and
(b) the calculation of $g(Q_0 \mid Y_{\mathrm{obs}}, Y_{\mathrm{mis}})$ for every value Q_0. This section describes an approximation to equation (2.1) given by Rubin and Schenker (1986) for use with only a small number of imputations of Y_{mis}.

The complete data inference for Q (corresponding to g in equation (2.1)) is typically based on a point estimate $\hat{Q}$, a variance U and a normal reference distribution. In the presence of non-response, however, with several (say M) imputations of the missing values Y_{mis} under the posited model for the missing data, there are M completed data sets and hence M sets of complete data statistics, say $\hat{Q}_m$ and U_m, $m = 1, \ldots, M$.

The M sets of complete data statistics are combined to create one multiple-imputation inference as follows. Let

$$\bar{Q} = M^{-1} \sum_{m=1}^{M} \hat{Q}_m$$

be the average of the M complete data estimates of Q,

$$\bar{U} = M^{-1} \sum_{m=1}^{M} U_m$$

be the average of the M complete data variances and

$$B = (M-1)^{-1} \sum_{m=1}^{M} (\hat{Q}_m - \bar{Q})^2$$

be the between-imputation variance of the complete data estimates of Q. Then the total variance of $Q - \bar{Q}$ is given by the sum of a within-imputation component and a between-imputation component:

$$T = \bar{U} + (1 + M^{-1})B.$$

Interval estimates and significance levels are obtained by using a t-distribution with centre $\bar{Q}$, scale $T^{1/2}$ and degrees of freedom

$$\nu = (M-1)(1 + r^{-1})^2,$$

where

$$r = (1 + M^{-1})B/\bar{U}$$

is the ratio of the between-imputation component of variance to the within-imputation component. Thus, for example, a $100(1-\alpha)\%$ interval estimate for Q is

$$\bar{Q} \pm t_\nu(1-\alpha/2)T^{1/2}, \tag{2.2}$$

where $t_\nu(1-\alpha/2)$ is the $(1-\alpha/2)$-quantile of the t-distribution with ν degrees of freedom. When $\bar{U}$ is large compared with B, indicating that there is little extra variability due to missing data, ν is large and the multiple-imputation analysis is based on a nearly normal reference distribution, as used in the standard complete data analysis. When B is large relative to $\bar{U}$, however, ν is close to the degrees of freedom $(M-1)$ associated with B.

2.3. *Choosing Number of Imputations*

Although multiple imputation was developed within the Bayesian framework (Rubin, 1978), it has been shown to have good frequentist properties. Theoretical and empirical work reported in Herzog and Rubin (1983), Rubin and Schenker (1986, 1987), Raghunathan (1987), Rubin (1987a), chapter 4, and Heitjan and Little (1991) has shown that multiple imputation is far superior to single imputation (i.e. imputing only one value per missing item) with regard to validity of interval estimates and significance levels.

An issue that arises in using multiple imputation concerns the number of draws (M) of Y_{mis} that should be created. The larger M is, the more precisely equation (2.1) can be simulated. A large number of imputations also necessitates a massive amount of computation in performing the multiple-imputation analysis, however.

Rubin and Schenker (1986, 1987) and Rubin (1987a), chapter 4, have shown that, if the imputation procedure follows the Bayesian paradigm of drawing Y_{mis} from its posterior predictive distribution as suggested by equation (2.1), then the multiple-imputation interval (2.2) generally has close to the nominal coverage level, even when only a few imputations (say $M=3$) are used. In fact, results in Rubin and Schenker (1986) suggest that improvements in the coverage of multiple-imputation intervals due to increasing M are linear in $1/(M-1)$. Thus the gains from using a large number of imputations are small.

3. Industry and Occupation Imputation Project

In this section, we describe what is to date the largest application of multiple imputation. This recently completed project imputed industry and occupation codes based on the 1980 coding scheme into individual records in two public use files from the 1970 US census to achieve comparability between 1970 and 1980 public use files.

3.1. *Lack of Comparability of Industry and Occupation Codes over Time*

In each decennial census of the USA, information is collected on occupations and industries by asking respondents to the census questionnaire to provide a narrative report of the type of work that they do and the type of enterprise within which they perform it. These narrative reports are then coded into several hundred occupation categories and several hundred industry categories, following a classification scheme devised by personnel of the Bureau of the Census. Because the primary concern of the Bureau is to provide up-to-date information on the current characteristics of the population, the occupation and industry classifications are revised for each census.

Major changes were made for the 1980 census, especially for occupations. For example, the 1980 occupation classification contains 503 categories, compared with 441 categories in the 1970 classification, and fewer than one-third of the categories in the 1970 occupation classification (covering about 15% of the labour force) map into single categories in the 1980 classification. Although changes in the industrial classification were not as extensive (there were 228 categories in 1970 and 232 categories in 1980), only about three-quarters of the 1970 categories (covering about 64% of the 1970 labour force) map into single categories in the 1980 classification.

The result of these classification changes has been to make analyses of change in the industry and occupation structures, and change in the relationship of industry and occupation to other phenomena, very difficult. Researchers interested in social change commonly analyse public use data from two or more censuses. However, given the complexity of changes in the classification systems, there has been no adequate way to compare results based on different classifications.

3.2. *Multiple Imputation to Achieve Comparability*

The Bureau of the Census and the Social Science Research Council in the early 1980s convened the Joint Subcommittee on Comparability of Occupation Measurement to consider ways to achieve comparability of industry and occupation codes in public use samples from various censuses. In a 1983 report, the Subcommittee suggested two possible methods. The first was to assign codes from the 1980 coding scheme to public use samples from previous censuses by directly coding the narrative responses for individuals. This would be a very accurate method, but it would be prohibitively expensive for the 1960 and 1970 censuses, because the files released for public use do not include the narrative responses and these responses would be very costly to retrieve.

The second method was to assign codes from the 1980 coding scheme directly to smaller samples from previous censuses, to use these samples to estimate models predicting the 1980 codes from the old codes and other covariates and then to use the models to impute 1980 codes for larger public use samples from the previous censuses. To allow the assessment of uncertainty associated with the imputation of 1980 industry and occupation codes to public use samples from censuses before 1980, the Subcommittee suggested creating multiple imputations.

A project for multiple imputation of industry and occupation codes based on the 1980 coding scheme to public use samples from the 1970 census has been completed recently with funding from the National Science Foundation and support from the Census Bureau. To build imputation models, the project used a sample of size 127 125 from the 1970 census for which the narrative responses were directly assigned industry and occupation codes by using both the 1970 and the 1980 classifications. This 'double-coded' sample had already been created by the Bureau for other purposes. The imputation models were used to create multiple imputations of 1980 codes for two public use samples from 1970 with a combined total of 1.6 million cases.

The public use samples for which codes have been imputed are the two 'standard metropolitan statistical area' (SMSA)/county group' samples from the 1970 census (United States Bureau of the Census, 1972). Each data set includes, for a 1% sample of US households, all information (with the exception of certain identifying geographical detail) obtained from one of two 'long form' versions of the census questionnaire: one administered to 5% of all households and the other administered to 15% of all households. The computer files for these samples, known as the '1970 census SMSA/county group one percent public use samples augmented with 1980-basis industry and occupation codes', may be obtained from the Census Bureau's Division of User Services or from the Inter-University Consortium for Political and Social Research.

Preliminary studies using just the double-coded sample suggest that analyses of the 1970 public use files with multiply imputed 1980 codes should be approximately valid (Rubin and Schenker, 1987; Weld, 1987; Treiman *et al.*, 1988).

3.3. *Outline of Imputation Procedures*

In this section, we describe briefly how the imputations of industry and occupation codes were created. For more details, see Weidman (1989) and Clogg *et al.* (1991). Industry and occupation codes were imputed in two stages. First, industry codes were imputed; then occupation codes were imputed conditional on the imputed industry codes, to ensure consistency between the two. Our emphasis here will be on describing the imputation of industry codes, as the procedures used for occupation codes are analogous.

The double-coded sample was split into subsamples, each corresponding to a particular 1970 industry code. For each subsample and thus each 1970 industry code, a model predicting 1980 industry code was estimated with covariates describing age, race, sex, class of worker (private industry, government or self-employed), amount of work and geography. Then, for each unit in the 1970 public use samples which had the same 1970 industry code, the results of the estimation and the unit's covariates were used to create multiple imputations of 1980 industry codes. Thus, each 1970 industry was considered separately; no pooling of information across 1970 industries was attempted. As mentioned earlier, imputations of 1980 occupation codes were created analogously, except that the covariates for occupations also included earnings variables and variables indicating the 1980 industry code that had been imputed.

The covariates used in the imputation models were chosen by substantive researchers including Census Bureau staff and members of the Subcommittee on Comparability of Occupation Measurement. They were chosen with two goals in mind:

(a) to include variables that would be good predictors of 1980 codes;
(b) to include variables that were likely to be used often by researchers analysing the public use samples, so that uncertainty in prediction across levels of these variables would be represented in the multiply imputed codes.

Whenever more than two 1980 industry codes were possible for a given 1970 code, imputations were performed by using a nested sequence of dichotomous imputations; thus the process of imputing 1980 codes for each unit in the 1970 public use samples can be described in terms of imputing a dichotomy. Modelling for this task was performed by using logistic regression. Once a model predicting a specific dichotomy, say IND1 *versus* IND2, was estimated from the appropriate subsample of the double-coded sample, multiple imputations for the corresponding subsample of the public use sample were created as follows. To impute one set of codes for the public use sample:

(a) a value β^* was drawn from the approximate posterior distribution of the logistic regression coefficients by using a multivariate normal approximation (along with the sampling–importance resampling algorithm of Rubin (1987b) in some cases);
(b) for each unit in the subsample of the public use sample, the values of the unit's covariates and β^* were used to obtain a probability π^* of the unit having 1980 code IND1, and then the 1980 code was imputed as IND1 with probability π^* or IND2 with probability $1-\pi^*$.

Multiple imputations were created by repeating steps (a) and (b) independently $M=5$

times. For this project, the use of five imputations was chosen because it was felt that most of the gains of multiple imputation would be achieved without too much burden for data analysts (see Section 2.3).

A two-step procedure was used in imputing values of Y_{mis}. First, a value of the parameter of the imputation model was drawn from its posterior distribution; then, a value of Y_{mis} was drawn conditional on the drawn value of the parameter. Using a two-step process to draw Y_{mis} from its posterior predictive distribution corresponds to writing the predictive density as the average of the density given β over the posterior distribution of β:

$$f(Y_{\text{mis}} \mid Y_{\text{obs}}) = \int f(Y_{\text{mis}} \mid Y_{\text{obs}}, \beta) f(\beta \mid Y_{\text{obs}}) \, \mathrm{d}\beta.$$

3.4. *Special Features of Missing Data Problem*

The industry and occupation code problem can be viewed as a missing data problem. Consider just the 1970 and 1980 censuses. The units in the 1970 public use samples can be thought of as 'non-respondents', since they are missing 1980 industry and occupation codes, whereas the units in the 1970 double-coded sample can be thought of as 'respondents'.

This missing data problem differs from standard problems in two major ways. One difference, which *decreases* the difficulty of this problem, is that the respondents (i.e. the double-coded sample) were selected by probability sampling, so that the reasons for non-response are known. In most missing data problems, the exact reasons for non-response are not known, and it is therefore important to explore sensitivity to different models of non-response to assess the possible biases under any particular model. The second difference, which *increases* the difficulty of this problem, is that the double-coded sample is about one-seventh the size of each of the public use samples for which imputations were created, so that the 'non-response rate' is much higher than in standard surveys. In fact, since only the public use files (and not the double-coded sample) are ultimately used for analysis, the non-response rate could be viewed as 100% for 1980 industry and occupation codes. Because of this high level of non-response, the uncertainty due to missing data can be unusually high, as will be seen in Section 4.

4. Analyses of Changes in Sex Composition of Occupations between 1970 and 1980

Suppose that it is desired to estimate the change between 1970 and 1980 in the sex composition of an occupation defined according to the 1980 coding scheme. Estimates of this kind are of considerable interest to sociologists and economists concerned with explaining patterns of participation of females in the labour force, to students of particular occupations or professions and to policy-oriented researchers wishing to assess changes in the extent of gender segregation of the labour force into different occupational categories.

4.1. *Analyses Performed*

Let p_{70} and p_{80} denote the proportions of people in the occupation that are female in 1970 and 1980 respectively; then the population quantity of interest is $Q = p_{80} - p_{70}$.

If samples from the 1970 and 1980 censuses are available, both having 1980 occupation codes, then Q may be estimated by

$$\hat{Q} = \hat{p}_{80} - \hat{p}_{70}, \tag{4.1}$$

where $\hat{p}_{70}$ and $\hat{p}_{80}$ are the sample proportions. The estimated variance of $\hat{Q}$ is

$$U = \hat{p}_{70}(1 - \hat{p}_{70})/n_{70} + \hat{p}_{80}(1 - \hat{p}_{80})/n_{80}, \tag{4.2}$$

where n_{70} and n_{80} are the numbers of sample cases in the occupation. If n_{70} and n_{80} are not small, the standard complete data inference for Q is given by interval (2.2), using equations (4.1) and (4.2).

For data from the 1980 census, we used a 2% public use sample with 1980 occupation codes assigned directly to each case. With regard to data from the 1970 census, two sources of data were available. The first source was the 1970 double-coded sample. Since this sample has 1980 codes assigned directly, the standard complete data analysis outlined above was used. The second source was the multiply imputed 1970 public use sample. With five sets of imputed 1980 codes on this sample, there were five values of $\hat{Q}$ and U for each occupation, corresponding to the five values of $\hat{p}_{70}$ and n_{70}. ($\hat{p}_{80}$ and n_{80} do not change across imputations because the true codes are known for the 1980 public use sample.) Using the five sets of statistics, $\hat{Q}_m$ and U_m, $m = 1, \ldots, 5$, the multiple-imputation analysis discussed in Section 2 was performed. The occupations examined and the sample sizes involved are displayed in Table 1.

4.2. *Comparisons of Precision and Validity*

The main issues to be investigated here are the precision of inferences involving the double-coded sample relative to inferences involving the multiply imputed public

TABLE 1
Occupations analysed for changes in sex composition between 1970 and 1980

1980 code†	*Description of occupation*	*No. of cases in occupation for the following samples:* *1970 double-coded sample*	 *1970 public use sample*‡	 *1980 public use sample*
067	Statisticians	49	227	562
084	Physicians	482	2798	8645
095	Registered nurses	1162	7875	25695
263	Sales workers: motor vehicles and boats	345	2384	5632
375	Insurance adjusters, examiners and investigators	168	1110	3332
418	Police and detectives: public service	515	3194	8401
484	Nursery workers	33	204	693
583	Paper-hangers	25	138	315
686	Butchers and meat cutters	473	3059	5952
704	Lathe and turning machine operators	224	1280	2370
808	Bus drivers	408	2580	7844
889	Labourers, except construction	1323	9129	27992

†United States Bureau of the Census (1981).
‡For the 1970 public use sample, the number of cases varies across imputations. The minimum number is given here.

use sample and the underestimation of variability when a single-imputation analysis is performed rather than the valid multiple-imputation analysis. For completeness, however, we display point estimates of Q from the double-coded sample and the multiply imputed public use sample in Table 2.

4.2.1. *Double-coded sample versus public use sample*

Since the 1970 double-coded sample has true 1980 codes assigned to its cases, it is valid to perform the standard complete data analysis with this sample and the 1980 public use sample. A question of interest is whether inferences drawn in this way are more or less precise than inferences based on the multiply imputed 1970 public use sample. In other words, is the loss in precision due to the smaller size of the double-coded sample smaller or greater than the loss in precision due to having imputed rather than true codes for the public use sample?

Measures of variability for the analyses of changes in sex composition are given in Table 3. The standard errors obtained from the double-coded sample ($U^{1/2}$) and the public use sample ($T^{1/2}$) are displayed in columns (2) and (4) respectively; their ratio $U^{1/2}/T^{1/2}$ is given in column (7). Although both analyses are valid, the multiple-imputation analysis involving the public use sample tends to yield more precise inferences than the analysis involving the double-coded sample. In five of 12 cases the ratio is greater than 1.5. Only for occupation 263 ('sales workers: motor vehicles and boats') is the ratio less than 1, and only for occupations 263 and 484 ('nursery workers') are the 95% interval estimates based on the multiply imputed public use sample wider than those based on the double-coded sample. (For occupation 484, the interval based on the public use sample is wider because the t-distribution in the multiple-imputation analysis has only 7 degrees of freedom.) For each occupation, conditional on the proportion $\hat{p}_{70}$ of the double-coded sample that is female, we computed the factor by which the sample size for the double-coded sample must be

TABLE 2
Estimates of the sex composition (percentage female) of occupations in 1980 and changes in the sex composition between 1970 and 1980†

Occupation (1980 code)	$\hat{p}_{80}$	*1970 double-coded sample* ($\hat{Q}$)	*1970 public use sample* ($\bar{Q}$)
067	49.6	15.0	4.2
084	13.6	5.3	3.9
095	95.9	-2.6	-1.7
263	7.5	5.2	1.3
375	61.4	30.5	30.7
418	5.9	3.5	2.1
484	46.2	-2.3	-16.4
583	18.1	10.1	6.6
686	14.3	2.9	2.7
704	8.5	-4.0	-0.9
808	45.4	18.7	18.1
889	19.8	3.9	5.5

†Positive values of Q indicate an increase in percentage female from 1970 to 1980.

TABLE 3
Measures of variability for analyses of changes in the sex composition of occupations between 1970 and 1980

Occupation (1980 code) (1)	*1970 double-coded sample* $(U^{1/2})$ *(2)*	*1970 public use sample* $U_1^{1/2}$ *(3)*	$T^{1/2}$ *(4)*	r *(5)*	$\hat{\gamma}$ *(%) (6)*	*Ratio (2)/(4) (7)*	*Ratio (3)/(4) (8)*	*Sample size factor*† *(9)*
067	7.1	3.9	4.0	0.06	5	1.8	1.0	4.0
084	1.3	0.7	1.0	1.28	62	1.3	0.7	1.8
095	0.4	0.2	0.2	0.37	29	1.5	0.9	2.8
263	0.9	0.6	1.0	2.07	73	0.8	0.6	0.7
375	3.7	1.6	1.8	0.19	17	2.1	0.9	5.3
418	0.7	0.4	0.5	0.54	39	1.4	0.8	2.2
484	8.9	4.0	7.7	3.57	83	1.2	0.5	1.4
583	5.8	3.3	4.5	0.79	49	1.3	0.7	1.9
686	1.5	0.7	0.8	0.24	21	1.9	0.9	4.7
704	2.3	0.9	1.7	1.80	70	1.4	0.6	2.0
808	2.3	1.0	1.0	0.02	2	2.2	1.0	6.1
889	1.0	0.4	0.8	2.75	78	1.2	0.5	1.6

†Factor by which the sample size for the double-coded sample must be multiplied to achieve the same precision as for the public use sample.

multiplied to obtain the same standard error that was obtained from the multiply imputed public use sample; see column (9). One factor is smaller than 1 (occupation 263), whereas the rest of the factors range from 1.4 to 6.1.

The multiply imputed public use sample generally provides more precision because it contains many more cases, each having information about Q obtained from the covariates used in the imputation process, even though the 1980 industry and occupation codes are missing. Another advantage of the public use sample, however, is that the larger sample size makes more detailed analyses possible. For instance, although the analyses presented here are for the entire USA, the double-coded sample has only 25 observations for occupation 583 ('paper-hangers') whereas the public use sample has at least 138 observations. If, say, finer geographical detail were desired, many analyses would be impossible with the double-coded sample but still possible with the public use sample.

The ratio of the between-imputation component of variance to the within-imputation component of variance, r, is given in column (5) of Table 3. This ratio indicates the extra variability due to having imputed values rather than true values. A related measure, suggested by Rubin (1987), chapter 3, is the estimated fraction of information missing due to non-response,

$$\hat{\gamma} = \frac{r + 2/(\nu + 3)}{r + 1},$$

which is given in column (6). This gives the relative difference between the Fisher information about Q in the multiple-imputation t-distribution and the estimated Fisher information about Q that would exist with complete response. For a simple random sample with just one outcome variable and no covariates, $\hat{\gamma}$ is similar to the non-response rate for the outcome variable. When covariates are available, however, $\hat{\gamma}$ is typically smaller than the non-response rate because of the information contained in the covariates. In general, both r and $\hat{\gamma}$ depend on the non-response rate, the predictive power of the covariates and the specific inference problem being considered.

Although the simple non-response rate for the 1970 public use sample is 100% (all 1980 industry and occupation codes are missing), $\hat{\gamma}$ ranges from 2% to 83%. Note as well the tendency for the ratio $U^{1/2}/T^{1/2}$ to decrease as $\hat{\gamma}$ and r increase. The advantage of the public use sample over the double-coded sample decreases when the covariates used for imputation have little predictive power.

4.2.2. *Multiple imputation versus single imputation*

As mentioned in Section 2, previous work has shown that single-imputation analyses generally tend to yield inferences that are too sharp because the between-imputation variability is ignored. Column (3) of Table 3 displays the standard errors $U_1^{1/2}$ obtained by applying the standard complete data analysis using just the first set of imputations. Column (8) displays the ratios $U_1^{1/2}/T^{1/2}$. In six of 12 cases, the ratio is less than 0.75; this indicates, for example, that in these cases single-imputation interval estimates will be less than three-quarters as wide as they should be. The ratio tends to be smaller when r and $\hat{\gamma}$ are larger, and the ratio is nearly 1 when r and $\hat{\gamma}$ are very small (e.g. occupation 808, 'bus drivers'); this is because, when r and $\hat{\gamma}$ are large, a substantial proportion of the total variability is from the between-imputation component, which cannot be assessed by using single imputation. The range of ratios shows that, not only are inferences based on single imputation generally too sharp, but also the extent of excessive apparent precision varies substantially from occupation to occupation, in ways that cannot be determined by using single imputation.

5. Discussion

This paper has presented information on what is to date the largest application of multiple-imputation methods: the imputation of 1980-based industry and occupation codes into individual records in two 1% public-use samples from the 1970 census. The development of public use samples for 1970 containing imputed 1980-based codes makes possible for the first time a rigorous analysis of changes in the distribution of the labour force over industries and occupations and changes in the relation of industry and occupation to other attributes of individuals. As such, these files are of great potential value to sociologists, economists and other social scientists concerned with the analysis of the structure and character of the US labour force.

Our analysis of the change in the percentage female from 1970 to 1980 for various occupations has demonstrated that the use of the imputed data from even one of the 1% public use samples in general provides more precise estimates than the only available alternative, a smaller sample in which 1980-based industry and occupation codes were directly assigned by coders from the narrative accounts returned by census respondents. Further analyses of data from the multiply imputed public use file are available from the authors, including examples of testing for changes in the coefficients of the regression of earnings on occupational status and sex between 1970 and

1980. The results of these analyses are consistent with the finding of this paper that using the public use sample results in a gain in precision in most cases.

We have demonstrated in examples with real data what is known theoretically —that reliance on single imputation understates the true uncertainty in the imputation method. Our examples demonstrate that the degree of distortion is highly problem specific, varying substantially from one occupation to another in our simple analysis of the change in the percentage female from 1970 to 1980. This result should give pause to researchers tempted to correct for a quick and unrefined analysis based on a single imputation by arbitrarily inflating the size of standard errors by some fixed percentage.

References

Clogg, C. C., Rubin, D. B., Schenker, N., Schultz, B. and Weidman, L. (1991) Multiple imputation of industry and occupation codes in census public-use samples using Bayesian logistic regression. *J. Am. Statist. Ass.*, **86**, 68-78.

Heitjan, D. F. and Little, R. J. A. (1991) Multiple imputation for the fatal accident reporting system. *Appl. Statist.*, **40**, 13-29.

Herzog, T. N. and Rubin, D. B. (1983) Using multiple imputations to handle nonresponse in surveys. In *Incomplete Data in Sample Surveys*, vol. 2, *Theory and Bibliographies* (eds W. G. Madow, I. Olkin and D. B. Rubin), pp. 209-245. New York: Academic Press.

Raghunathan, T. E. (1987) Large sample significance levels from multiply-imputed data. *PhD Thesis*. Department of Statistics, Harvard University, Cambridge.

Rubin, D. B. (1978) Multiple imputations in sample surveys—a phenomenological Bayesian approach to nonresponse. *Proc. Survey Res. Meth. Sect. Am. Statist. Ass.*, 20-34.

——— (1987a) *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

——— (1987b) Discussion of Tanner and Wong. *J. Am. Statist. Ass.*, **82**, 543-546.

Rubin, D. B. and Schenker, N. (1986) Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *J. Am. Statist. Ass.*, **81**, 366-374.

——— (1987) Interval estimation from multiply-imputed data: a case study using census agriculture industry codes. *J. Off. Statist.*, **3**, 375-387.

Treiman, D. J., Bielby, W. T. and Cheng, M.-T. (1988) Evaluating a multiple-imputation method for recalibrating 1970 U.S. census detailed industry codes to the 1980 standard. *Sociol. Methodol.*, **18**, 309-345.

United States Bureau of the Census (1972) *Public Use Samples of Basic Records from the 1970 Census: Description and Technical Documentation*. Washington DC: United States Bureau of the Census.

——— (1981) *Census of Population Alphabetical Index of Industries and Occupations*, 2nd edn. Washington DC: US Government Printing Office.

Weidman, L. (1989) Final report: industry and occupation imputation. *Report Census/SRD/89/03*. Statistical Research Division, United States Bureau of the Census, Washington DC.

Weld, L. H. (1987) Significance levels from public-use data with multiply-imputed industry codes. *PhD Thesis*. Department of Statistics, Harvard University, Cambridge.