FIDDLING AROUND WITH NONMATCHES AND MISMATCHES

Fritz Scheuren and H. Lock Oh
Social Security Administration

The necessity of linking records from two or more sources arises in many contexts. One good example would be merging files in order to extend the amount or improve the quality of information available for population units represented in both files. In developing procedures for linking records from two or more sources, tradeoffs exist between two types of mistakes: (1) the bringing together of records which are for different entities (mismatches), and (2) the failure to link records which are for the same entity (erroneous nonmatches). Whether or not one is able to utilize one's resources in an "optimal" way, it is almost certainly going to be true that in most situations of practical interest some mismatching and erroneous nonmatching will be unavoidable. How to deal with these problems depends, of course, to a great extent on the purposes for which the data linkage is being carried out. Because these reasons can be so diverse, no general strategy for handling mismatches and nonmatches will be offered here. Instead, we will examine the impact of these difficulties on the analysis of a specific study. The study chosen is a large-scale matching effort, now nearing completion, which had as its starting point the March 1973 Current Population Survey (CPS).

THE 1973 CENSUS - SOCIAL SECURITY EXACT MATCH STUDY

The primary identifying information in the 1973 Census-Social Security study was the social security number (SSN). The problems which arise when using the SSN to link Current Population Survey interview schedules to Social Security records differ in degree, but not in kind, from the problems faced by other "matchmakers."

In the 1973 study, as in prior CPS-SSA linkages, the major difficulty encountered was incompleteness in the identifying information [1]. Manual searches had to be carried out at SSA for over 22,000 individuals for whom no SSN had been reported by the survey respondent [2]. Another major problem was reporting errors in the social security number or other identifiers (name and date of birth, etc.). SSN's were manually searched for at SSA in cases where severe discrepancies between the CPS and SSA information were found after matching the two sources using the account number initially provided [3]. Because of scheduling and other operational constraints, an upper limit of 4,000 manual searches had to be set for this part of the project. Therefore, it was possible to look for account numbers only in the most "likely" instances of CPS misreporting of the SSN. The cases sent through this search procedure were those for which both name and date of birth were in substantial disagreement. For social security beneficiaries, computerized (machine) searches at SSA were also conducted for both missing and misreported SSN's. This was made possible through an administrative cross-reference system which links together persons who receive benefits on the same claim number. About 1,000 potentially usable SSN's were obtained in this way.

Operational Restrictions on the Matching.-- One of the concerns the 1973 work has in common with earlier Census-SSA linkage efforts is the great care that is being taken to ensure the confidentiality of the shared information. The laws and regulations under which the agencies operate impose very definite restrictions on such exchanges, and special procedures have been followed throughout, so as to adhere to these provisions--in particular, to ensure that the shared information is used only for statistical purposes and not for administrative ones.1/ Another major restriction on the study was, of course, that it had to be conducted using data systems which were developed and are used principally for other purposes. The CPS, for instance, lacks a number of pieces of information that would, if available, have materially increased the chances of finding the surveyed individual in SSA's files. Finally, the manual searching for over 26,000 account numbers at Social Security imposed a sizable addition to the normal administrative workload in certain parts of the agency. Therefore, in order to obtain a reasonable priority for the project, numerous operational compromises were made which precluded the employment of "optimal" matching techniques [e.g., 4, 5, 6, 7, 8]. One of the most serious of these was the decision basically not to "re-search" for the missing and misreported SSN's of individuals for whom no potentially usable number was found after just one search.

Basic Match Results.--There were 101,287 interviewed persons age 14 or older who were included in the 1973 Census-Social Security Exact Match Study. Of the total, about 2 percent had not yet been issued an SSN at the time of the interview and, hence, were not eligible for matching. In another 8 percent of the cases, no potentially usable social security numbers could be found even though one was believed to exist. For the remaining 90,815 sampled individuals, an SSN was available, and CPS and SSA data could be linked. Of these account numbers, 77,465 were supplied by CPS respondents initially. There were also 3,347 cases where the SSN provided originally was replaced with an account number obtained from the manual and machine searches of SSA's files which were described above. In a few of these cases--about 200--the SSN's used as replacements were taken from a supplementary Census source. Finally, there were 10,003 sampled individuals for whom no account number had been provided initially, but one was obtained subsequently by a search of SSA's files.

ALTERNATIVE COMPUTERIZED MATCH RULES

In general, aside from certain obvious errors (which have already been eliminated), it is not

possible to determine whether the SSN we have for a particular individual is his own or has been erroneously ascribed to him. One can, however, estimate the likelihood that a potentially usable account number is incorrect. To do this, five confirmatory variables common to both data sets were used: surname (first six characters), age attained in 1972 (in years), race, sex, and month of birth. The pattern of agreements and disagreements that might be expected between the CPS and SSA reporting on these variables depends, of course, on whether the records brought together are "mismatches" or "truematches." (See figure 1 below for definitions.)

Table 1.--ESTIMATED NUMBER OF MISMATCHES AND ERRONEOUS NONMATCHES BY MATCH RULE FOR MARCH 1973 CPS INTERVIEWED PERSONS 14 YEARS OF AGE OR OLDER
(Based on an unweighted CPS sample of all individuals with potentially usable SSN's including a small number of Armed Forces members excluded from the weighted figures in the remaining tables)

Item	Perfect agreement rule	Surname agreement rule	CPS-SER agreement rule	Potentially usable rule
Total.....	90,815	90,815	90,815	90,815
Matched, Total	76,294	85,293	86,910	90,815
Truematches.....	76,276	84,784	86,537	88,962
Mismatches.....	18	509	373	1,853
Mismatches as a percent of total matches.....	0.02	0.60	0.43	2.04
Nonmatches, Total	14,521	5,522	3,905	-
True Nonmatches.....	1,835	1,344	1,480	-
Erroneous Nonmatches.....	12,686	4,178	2,425	-

Figure 1 -- Match Definitions

TRUEMATCH -- A match between a Social Security Administration (SSA) record and a Current Population Survey (CPS) interview schedule where the two sets of documents were for the same individual.

MISMATCH -- The erroneous matching of data from the two sources when the information brought together was not for the same individual.

TRUE NONMATCHES -- Individuals in the Current Population Survey who have not yet been issued a social security number (SSN) and therefore do not have a Social Security Administrative record.

ERRONEOUS NONMATCH -- A case where either no SSN could be found even though it had been issued (making it impossible to match the sources together) or the two sources were brought together but because of the rule used to decide what would be called a "match" they were treated erroneously as nonmatches.

Mismatches.--If mismatches arise on a purely chance basis, then the probability of agreement on any one variable would depend just on the marginal distribution of that variable in the two data sets being linked. This is the assumption we have made here. The conditional probability given a mismatch of a particular combination of agreements (disagreements) on the confirmatory information, denoted by $\{P^{MM}\}$, was thus estimated as the product of the observed marginal proportions of agreement and disagreement for each variable separately.

Two separate mismatch models were fit: one for SSN's obtained in manual searching and one for all other SSN's. This was necessary because of the nature of SSA's manual searching procedures where, for a number to be returned from the search, there usually must be at least rough agreement on surname and age. (Hence, these two variables could not be used for evaluating mismatches among persons with SSN's obtained from manual searching.)

Truematches.-- Differences between the CPS and SSA variables can arise quite frequently even when the data is for the same person. The information in the two systems is collected at very different times; perhaps as long as 30 or more years separate the two observations. Furthermore, the respondent on the two occasions may very well be different. For the most part, the Social Security variables were obtained from the individual himself, while in the CPS, over half the information was obtained by proxy.

The extent of agreement for "truematches" has also been modelled by assuming independence among the confirmatory variables. However, the conditional probabilities of agreement, given a truematch, denoted by $\{P^{TM}\}$, cannot be estimated separately from the overall mismatch rate, "α," that exists among the 90,815 individuals with potentially usable SSN's. To obtain estimates an Information Theoretic approach was taken; the $\{P^{TM}\}$ and α were obtained by (iteratively) fitting the observed proportions $\{\pi\}$ for each of the combinations of agreement or disagreement on the confirmatory variables that were found in the sample. The estimating equation was of the form

$$(1) \qquad \pi = (1 - \alpha)\, P^{TM} + \alpha\, \hat{P}^{MM}$$

where the $\{\hat{P}^{MM}\}$ were calculated as described above, with α and the $\{P^{TM}\}$ being chosed such that

$$(2) \qquad I(\hat{\pi};\pi) = \Sigma\, \hat{\pi} \ln \frac{\hat{\pi}}{\pi}$$

was a minimum. The $\{\hat{\pi}\}$ are given by the expression

$$(3) \qquad \hat{\pi} = (1 - \hat{\alpha})\, \hat{P}^{TM} + \hat{\alpha}\, \hat{P}^{MM}$$

and were used in obtaining table 1 below.

These models were judged to be adequate except for cases where there was perfect or near perfect agreement on the confirmatory variables. For such individuals, research from other SSA studies indicated that the estimated number of mismatches was probably too small, and some upward adjustments were made to the fitted results.2/

Alternate Match Rules.--The match rules considered in the remainder of this paper all use the extent of agreement on age, race, sex, month of birth, and surname to determine whether CPS and SSA records linked by common SSN's should be treated as "matches" or "nonmatches." Four ad hoc rules were examined:

1. "Perfect" Agreement Rule.--For this rule all five confirmatory variables had to agree within tolerance. For surname, which depends on a character-by-character agreement of the first six letters of the last name, a tolerance of two letters was allowed. Similarly, a difference of four years was permitted in defining agreement on age. For sex, race, and month of birth, no tolerance was allowed.

2. Surname Agreement Rule.--This rule requires at least four of the first six letters of the surname to be the same. (The other confirming variables were not considered.) The surname rule is based on a modified version of the administrative procedures now in use at IRS and SSA to verify the correctness of the social security number supplied.

3. CPS-SER Agreement Rule.--This rule basically requires that four out of the five confirmatory variables agree (within the tolerances mentioned in the first rule above). In selected cases (361 altogether), agreement on just three variables was enough to consider the individual a match. It was this rule, discussed in report no. 4 of SSA's Series on Studies from Interagency Data Linkages, which has been employed for the first public-use match file prepared from the project and described in reports nos. 5 and 6 of that Series.

4. Potentially Usable Rule.--This is the least stringent of the rules in that no restrictions are placed on what is to be called a "match."

IMPACT OF ALTERNATE MATCH RULES ON EARNINGS

In assessing the four match rules being considered, it is not enough simply to look at them in terms of their respective mismatch and erroneous nonmatch rates. What we need to do is to take account of the bias and variance implications of the matching error on some of the chief variables to be provided by the linkage. Among the most important of these data items are the 1972 earnings information reported to the Census Bureau and to Social Security. In this section, therefore, we will compare these earnings data under each match rule. First, we will examine the extent to which one's overall "level" estimators of the CPS or SSA earnings distribution are affected by the different match rules. The level estimates are of interest principally because a standard exists for these against which a comparison can be made. What is crucial to our evaluation, however, is the sensitivity of the relationships between CPS and SSA earnings amounts to the match rule chosen. Here, of course, no outside standard exists, since it was to examine these relationships that the study was mounted.

Level Comparisons.--Tables 2 and 3 below compare the percentage distributions of CPS and SSA earnings for each procedure with preliminary overall survey or administrative control figures. No correction has been made for erroneous nonmatches or mismatches, but the sample has been reweighted to make a rough adjustment for differences which arise because of survey undercoverage [9].

Sizable discrepancies among the various estimates can be observed in the tables. For example, from table 2, it can be seen that the difficulty of obtaining an SSN may have been relatively greater for individuals who were not identified in the CPS as having worked in 1972. Large differences (statistically significant at $\alpha = 0.01$) exist, in fact, between each of the match results and the control for the "no earnings" category of the CPS classifier. On the other hand, both tables 2 and 3 show that persons with CPS or SSA earnings of $9,000 or more are always proportionately overrepresented in the sample. For the SSA classifier the observed differences for the $9,000 or more class are all significant at the $\alpha = 0.01$ level.

Table 2--UNADJUSTED CPS EARNINGS PERCENTAGE DISTRIBUTIONS FOR CIVILIANS 14 OR OLDER WITH SSN'S UNDER ALTERNATE MATCH RULES AS COMPARED TO THE OVERALL SURVEY ESTIMATE
(Based on weighted sample counts for civilians adjusted as explained in the text. Note detail may not add to totals because of rounding.)

Size of CPS earnings	Overall Survey Estimate	Match rule			
		Perfect agreement rule	Surname agreement rule	CPS-SER rule	Potentially usable rule
TOTAL	100.0	100.0	100.0	100.0	100.0
None	35.0	32.8	33.6	34.0	34.2
$1 to $999 or loss	10.9	10.5	10.6	10.7	10.6
$1,000 to $1,999	5.8	5.9	5.9	6.0	6.0
$2,000 to $2,999	4.4	4.5	4.5	4.5	4.5
$3,000 to $3,999	4.4	4.5	4.6	4.6	4.6
$4,000 to $4,999	4.4	4.5	4.5	4.5	4.5
$5,000 to $5,999	4.5	4.7	4.7	4.7	4.7
$6,000 to $6,999	4.1	4.3	4.3	4.2	4.2
$7,000 to $7,999	4.2	4.3	4.3	4.2	4.2
$8,000 to $8,999	3.5	3.6	3.5	3.5	3.5
$9,000 or more	18.9	20.4	19.5	19.2	19.0

Table 3--UNADJUSTED SSA EARNINGS PERCENTAGE DISTRIBUTIONS FOR CIVILIANS 14 OR OLDER WITH SSN'S UNDER ALTERNATE MATCH RULES AS COMPARED TO THE ADMINISTRATIVE CONTROLS
(Based on weighted sample counts for civilians adjusted as explained in the text. Note detail may not add to totals because of rounding.)

Size of SSA earnings	Administrative Control	Match rule			
		Perfect agreement rule	Surname agreement rule	CPS-SER rule	Potentially usable rule
TOTAL	100.0	100.0	100.0	100.0	100.0
None	40.9	39.2	40.0	40.6	41.0
$1 to $999	10.2	9.7	9.8	9.9	9.8
$1,000 to $1,999	6.5	6.3	6.3	6.2	6.2
$2,000 to $2,999	4.7	4.6	4.7	4.7	4.6
$3,000 to $3,999	4.4	4.4	4.4	4.4	4.4
$4,000 to $4,999	4.3	4.5	4.4	4.4	4.4
$5,000 to $5,999	4.1	4.2	4.1	4.1	4.0
$6,000 to $6,999	3.7	3.9	3.9	3.8	3.8
$7,000 to $7,999	3.3	3.6	3.5	3.5	3.5
$8,000 to $8,999	3.1	3.0	3.0	2.9	2.9
$9,000 or more	14.8	16.5	15.8	15.5	15.3

Relationship Comparisons.--The relationships between CPS and SSA reported earnings can be investigated in a number of ways. One of the standard methods is to cross-classify the two amounts by the same dollar size-classes and count the fraction of cases which fall into the same interval or into a higher or lower interval [11]. Table 4 provides a summary of such cross-tabulations for each match rule where the dollar size-classes used are the same as those shown in tables 2 and 3.

Table 4.--PERCENTAGE DISTRIBUTION OF EARNINGS CLASS AGREEMENT BETWEEN CPS AND SSA REPORTED AMOUNTS FOR CIVILIANS 14 OR OLDER WITH SSN'S UNDER ALTERNATE MATCH RULES BEFORE ADJUSTMENT
(Based on weighted sample counts for civilians adjusted as explained in the text. Note detail may not add to totals because of rounding.)

Extent earnings class agreement	Perfect agreement rule	Surname agreement rule	CPS-SER agreement rule	Potentially usable rule
Total	100.00	100.00	100.00	100.00
SSA earnings in higher interval than CPS	10.84	11.35	11.05	11.70
CPS and SSA earnings class agree	68.08	67.13	67.42	66.05
CPS earnings in higher interval than SSA	21.08	21.52	21.53	22.25

As can be seen from table 4, marked differences exist among the procedures in the proportion of individuals whose CPS and SSA earnings class agree. The percentages vary from a high of 68 percent for the perfect agreement rule to a low of 66 percent for the potentially usable one, with the surname and CPS-SER rules having class agreements of around 67 percent. The standard errors for the four estimators of the extent of earnings class agreement average about 0.25 percentage points. The range of the agreement figures (at 2.0 percentage points) is thus eight times the standard error.

Since our focus is on the matching process itself, we will leave to others [12, 13] a detailed study of the relationships between the earnings distributions shown in table 4. Instead, we will proceed (in the next section) to examine the bias and variance impact of adjustments designed to lessen the effect of errors in the matching.

UTILITY OF POST-HOC ADJUSTMENT PROCEDURES

In this section a combination of procedures is examined which is designed to adjust for mismatching and erroneous nonmatches. Successive adjustments will be made to the data: first, by reweighting to account for the nonmatches; then, by "raking" the results to the overall survey and administrative controls shown in tables 2 and 3; and, finally, by "subtracting out" estimates of the effect of the mismatching. The utility of each step taken will be evaluated in terms of its bias and variance impact.

Reweighting for Nonmatches.--No matter which of the four match rules is used, important differences exist between those who are treated as "matches" and those believed to have SSN's but for whom no usable account number could be determined. This is evident not only from tables 2 and 3, but also from previous papers which have discussed the reporting of social security numbers in the March 1973 Current Population Survey [i.e., 1, 2, 3]. For example, large differences exist between the two groups by earnings, age, race, sex, and respondent status.3/

One way to "correct" for these differentials (the method adopted in this paper) is to consider the cases where SSN's were obtained through manual searching as a sample from the entire group of individuals who "should" have usable numbers but do not. The exact procedure followed was to subtract from the estimated total with SSN's, the weighted number of adults who had an acceptable SSN but who had not obtained it from the manual search. The weighted manual search cases were then ratioed up to this difference and added to the estimates obtained from the rest of the sample. These steps were carried out for each of the eight CPS rotation groups separately in order to be able to come up with an approximation to the variance.4/ The overall adjustment factors applied are shown below for each match rule along with the (weighted) fraction of sample cases with SSN's but for which no usable SSN could be found.

Match Rule	Percent with No Usable SSN Found	Weighting Factor for Manual Search Cases
Perfect agreement rule	26.9	3.4
Surname agreement rule	13.2	2.2
CPS-SER rule	10.9	2.0
Potentially usable rule	5.9	1.5

The reweighting procedure just described, while crude in many respects, does have a certain logic to it since the great bulk of the cases for whom no SSN is available were searched for manually in SSA's files. It might also be noted in passing that such an approach is quite analogous to the classical method for utilizing follow-up samples of those persons who, in the survey's initial wave, were nonrespondents [14].

To help evaluate the impact of the reweighting scheme, table 5 is provided below. As can be seen, for all match rules, the reweighting reduces

Table 5.- PERCENTAGE DISTRIBUTION OF EARNINGS CLASS AGREEMENT BETWEEN CPS AND SSA REPORTED AMOUNTS FOR CIVILIANS 14 OR OLDER WITH SSN'S UNDER ALTERNATE MATCH RULES AFTER REWEIGHTING
(Based on weighted sample counts for civilians adjusted as explained in the text. Note detail may not add to totals because of rounding.)

Extent earnings class agreement	Perfect agreement rule	Surname agreement rule	CPS-SER agreement rule	Potentially usable rule
Total	100.00	100.00	100.00	100.00
SSA earnings in higher interval than CPS	11.99	12.01	11.59	12.01
CPS and SSA earnings class agree	66.74	66.34	66.81	65.70
CPS earnings in higher interval than SSA	21.26	21.65	21.60	22.29

the amount of CPS-SSA earnings-class agreement. In fact, the average declined by about 0.8 percent, from 67.17 percent to 66.40 percent. From internal evidence in the CPS, there seems to be a definite tendency for persons who provide

usable SSN's to be better respondents than those who do not. Thus, this reduction in earnings-class agreement (with accompanying increases elsewhere) probably reduces the overall nonmatch bias which exists for all of the estimators. There is, of course, no way of knowing whether the magnitude of the changes is appropriate, but it is encouraging to note that the net effect of the reweighting is to bring the estimates for the four rules closer together. (The range of the percentages for earnings-class agreement dropped from 2.0 percent to 1.1 percent.

For the probable reduction in the nonmatch bias, a price has been paid in increasing the standard error of nearly all the estimators shown in the table. These increases range from small to moderate for the potentially usable, surname, and CPS-SER rules. However, for the perfect agreement rule, the increase is sizable; if such a rule were seriously being contemplated, some other method of adjustment would, in all likelihood, be desirable.

Raking Adjustment for Nonmatches.--The reweighting scheme just described tends to bring the matched CPS and SSA earnings distributions closer to the control totals shown in tables 2 and 3. However, the remaining discrepancies are still large. Unlike biases in the CPS-SSA interrelationships, which can only be adjusted indirectly and incompletely, it is possible to alter the sample earnings marginals so they conform simultaneously to both sets of controls more or less exactly. There are a number of well-known procedures for doing this. The approach employed here is due to Deming and Stephan [15], and we have referred to it, following the practice at the Census Bureau, as "raking." (Perhaps it is better known elsewhere as "the method of iterative proportions" [16].)

Table 6.--PERCENTAGE DISTRIBUTION OF EARNINGS CLASS AGREEMENT BETWEEN CPS AND SSA REPORTED AMOUNTS FOR CIVILIANS 14 OR OLDER WITH SSN'S UNDER ALTERNATE MATCH RULES AFTER REWEIGHTING AND RAKING
(Based on weighted sample counts for civilians adjusted as explained in the text. Note detail may not add to totals because of rounding.)

Extent earnings class agreement	Perfect agreement rule	Surname agreement rule	CPS-SER agreement rule	Potentially usable rule
Total	100.00	100.00	100.00	100.00
SSA earnings in higher interval than CPS	11.78	11.82	11.47	11.98
CPS and SSA earnings class agree	66.01	65.89	66.36	65.45
CPS earnings in higher interval than SSA	22.21	22.30	22.17	22.57

Table 6 provides a summary of the impact of the raking on the extent of agreement between CPS and SSA earnings. As will be seen, our estimators of the amount of agreement have declined still more as a result of this additional adjustment (from an average of 66.4 percent after reweighting to 66.2 percent after raking). The range in the extent of agreement has also narrowed further, from 1.1 percent to 0.9 percent, respectively, with the largest proportion on the main diagonal being 66.4 percent (CPS-SER) and the smallest, 65.5 percent (potentially usable rule). Again, we believe that this change represents a further reduction in the nonmatch bias. Not unexpectedly, the raking has also produced reductions in the standard errors, although not uniformly so. (For 8 of the 12 estimators in the table, there was some reduction. In the four instances where increases occurred, they were slight.)

Mismatch Adjustment.--If two linked records have been brought together just by chance, then it is highly unlikely for them to agree on earnings class. Thus, a "natural" consequence of the mismatching which exists under each rule is that the estimates of the extent of agreement, as shown in table 6, understate the true underlying amount of agreement. Some further adjustment, therefore, is necessary. There are a number of ways of taking account of the mismatches, depending on the assumptions one is willing to make about their affect on the relationship between the CPS and SSA classifiers. The model chosen here is a fairly simple one which may not be too unrealistic. Basically, it assumes that the mismatch rates do not depend on earnings levels and that, when a mismatch occurs, the matched CPS and SSA amounts are independently distributed. Put another way, the mismatches can be thought of as having the same row $\{P_{i.}\}$ and column $\{P_{.j}\}$ marginal proportions for CPS and SSA earnings, respectively, as the truematches; but such that the proportion of mismatches for any particular combination ij of CPS and SSA earnings classes, denoted $\{P^{MM}_{ij}\}$, is given by

$$(4) \qquad P^{MM}_{ij} = P_{i.} P_{.j} .$$

The expected value of the observed relationship between the two classifiers is assumed to consist of two components. First, there is an estimate of the truematch proportion in the $(ij)^{th}$ cell of the earnings cross-tabulation, denoted P^{TM}_{ij}, times the fraction of the total sample that were truematches, denoted by $(1 - \alpha)$. The second term consists of the mismatch proportion P^{MM}_{ij} times the fraction of the total sample that were mismatches (i.e., "α"). Thus, we have that the observed cell proportions $\{\pi_{ij}\}$ can be expressed as

$$(5) \qquad E\pi_{ij} = (1 - \alpha) P^{TM}_{ij} + \alpha P^{MM}_{ij} .$$

From (4) this becomes

$$(6) \qquad E\pi_{ij} = (1 - \alpha) P^{TM}_{ij} + \alpha P_{i.} P_{.j} .$$

Since estimates of the mismatch rate α, the CPS marginal $\{P_{i.}\}$, and SSA marginal $\{P_{.j}\}$ were all readily available (tables 1 to 3), it was a simple matter to obtain estimates of the $\{P^{TM}_{ij}\}$ by substituting $\hat{\alpha}$, $\hat{P}_{i.}$, and $\hat{P}_{.j}$ in (6). The $\{P^{TM}_{ij}\}$ so obtained were then used to produce the results in table 7. 5/

For the perfect agreement rule, the mismatching had only a small effect, but, for the other rules, changes in the percent with CPS and SSA earnings

Table 7.--PERCENTAGE DISTRIBUTION OF EARNINGS CLASS AGREEMENT BETWEEN CPS AND SSA REPORTED AMOUNTS FOR CIVILIANS 14 OR OLDER WITH SSN'S UNDER ALTERNATE MATCH RULES AFTER ALL ADJUSTMENTS INCLUDING THE ADJUSTMENT FOR MISMATCHING (Based on weighted sample counts for civilians adjusted as explained in the text. Note detail may not add to totals because of rounding.)

Extent earnings class agreement	Perfect agreement rule	Surname agreement rule	CPS-SER agreement rule	Potentially usable rule
Total	100.00	100.00	100.00	100.00
SSA earnings in higher interval than CPS	11.77	11.63	11.34	11.46
CPS and SSA earnings class agree	66.03	66.25	66.62	66.45
CPS earnings in higher interval than SSA	22.20	22.12	22.05	22.10

in the same interval were substantial. For the potentially usable rule, where the amount of mismatching was estimated to be greatest, that proportion increased by 1 percent, from 65.45 percent to 66.45 percent. Increases for the CPS-SER and surname rules were smaller but still sizable (0.3 and 0.4 percentage points, respectively). The range of the four estimates of the extent of agreement narrowed again as a result of this final adjustment (from 0.91 percent after raking to 0.59 percent). The "cost" of the mismatch adjustment was a very slight increase in the variance over that of the raked estimator.

Summary of Impact of Adjustments.--Overall, when we look at the combined affect of all three adjustments, we see that the range of earnings class agreement under the four rules has been reduced to less than one-third of what it was to begin with (i.e., from 2.0 percent to 0.6 percent). This narrowing of the range of agreement suggests that the techniques employed may have been "moderately" successful in reducing the various biases which affect each rule (and may even have some merit in general). However, since the range in earnings-class agreement after adjustment is still about twice the standard deviation, it seems likely that residual uncorrected biases remain an important part of the total mean square error.

Except for the perfect agreement rule, the price that was paid for this bias reduction appears to be "small." The median increase in the standard errors was about 10 percent of the original standard errors. (However, since the sample sizes involved are so large, this amounted to only 0.025 percentage points.)

In the light of our computations, it might be of interest to comment on which match rule is "best." Because the final results are so close, this question has lost some of its force but is still worth pursuing. By and large, the results suggest that in this case, and for the statistics considered, the best choice of the four match rules examined is the potentially usable rule. 6/ It tends to have the smallest standard error after all adjustments; its initial and final estimates change the least; and, its initial and final estimates are the closest of any rule to the overall average for all rules after adjustment. Partly as a consequence of this finding, all subsequent public-use data tapes to be prepared from the 1973 Census-Social Security Study will be made available with all the potentially usable "matches" included. 7/ Also, since information on the extent of agreement on the confirmatory variables is available on these data tapes, another consequence of this decision is that users will have the option of choosing the match rule best suited for their purposes.

Conclusion.--Matched statistical samples have much in common with other surveys and, as we have seen, adjustment techniques normally encountered in standard practice (e.g., raking), can be applied successfully to linked data sets as well. The problems of choosing a suitable match rule and of dealing with mismatches are, however, unique to record linkage studies. Usually, in the literature on data linkage, match rules (and mismatching) have been dealt with in the context of the research design and how to choose "optimal" strategies for allocating resources. With few exceptions [17], there has been insufficient attention given to the analysis aspects of imperfectly matched samples. In the 1973 Census-Social Security Study, the administrative (and, to some extent, confidentiality) constraints imposed on the design and execution of the data linkage make these analysis issues particularly pointed. Our approach to them has, of course, been quite applied. Obviously, theoretical examinations are warranted as an adjunct to the empirical work on matching commented on here. We invite participation in this endeavor.

FOOTNOTES

*The authors would like to thank Wendy Alvey and Gina Savinelli for their assistance, especially for helping to prepare the basic tabulations. Thanks also must be extended to Ben Bridges and Dean Leimer for their careful reading of an earlier draft.

1/ For details on the confidentiality precautions taken, see the invited paper session on the Reconciliation of Survey and Administrative Sources through Data Linkage shown elsewhere in these Proceedings.

2/ A paper is in preparation which provides more details on the procedures employed in estimating the number of mismatches with particular attention to other estimation methods.

3/ In the public-use file (with the CPS-SER match rule), the reweighting adjustment being made attempts to take account of most of these factors. See report nos. 5 and 6 in Studies from Interagency Data Linkages for details.

4/ The raking and mismatch adjustments were also carried out separately by CPS rotation group to make it possible to approximate their variance impact as well.

5/ The mismatch rates used were not those shown in table 1 but were calculated (by rotation group) in terms of the weighted data after having taken account of the adjustments for nonmatches.

6/ Readers should carefully note the qualifications on this "endorsement" of the potentially usable rule. While for the example chosen here the nonmatch and mismatch errors of this rule tended to cancel each other out, this would not always be the case. In fact, the potentially usable rule, if not adjusted for mismatches, in many situations might even be the worst rule one could choose.

7/ For reasons of confidentiality, social security information for CPS respondents who refused to provide their SSN's to the Census Bureau are not includable on the public-use files from this project, even though it was possible to find on account number for them. With the CPS-SER rule, 619 such cases were eliminated. With the potentially usable rule, 641 cases would have to be treated as nonmatches for this reason.

REFERENCES

[1] Vogel, L., and Coble, T., "Current Population Survey Reporting of Social Security Numbers," 1974 Amer. Stat. Assn. Proc. Soc. Stat. Sec., 1975, pp. 130-136.

[2] Kilss, B., and Tyler, B., "Searching for Missing Social Security Numbers," 1974 Amer. Stat. Assn. Proc. Soc. Stat. Sec., 1975, pp. 145-150.

[3] Cobleigh, C., and Alvey, W., "Validating Reported Social Security Numbers," 1974 Amer. Stat. Assn. Proc. Soc. Stat. Sec., 1975, pp. 137-144.

[4] Tepping, B. J., "A Model for Optimum Linkage of Records," J. Amer. Stat. Assn., vol. 63, 1968, pp. 1321-1332.

[5] Felleghi, I. P., and Sunter, A. B.,"A Theory for Record Linkage," J. Amer. Stat. Assn., vol. 64, 1969, pp. 1183-1210.

[6] DuBois, Jr., N. S. D., "A Solution to the Problem of Linking Multivariate Documents," J. Amer. Stat. Assn, vol. 64, 1969, pp. 163-174.

[7] Wells, B., Optimum Matching Rules, University of North Carolina, 1974.

[8] Nathan, G., "Outcome Probabilities for a Record Matching Process with Complete Invariant Information," J. Amer. Stat. Assn., vol. 62, 1967, pp. 454-469.

[9] Vaughan, D. R., and Ireland, C. T., "Adjusting for Coverage Errors in the March 1973 Current Population Survey," 1975 Amer. Stat. Assn. Proc. Soc. Stat. Sec.

[10] Mosteller, F., "Association and Estimation in Contingency Tables," J. Amer. Stat. Assn., vol. 63, 1968, pp. 1-28.

[11] Scheuren, F. J., and Oh, H. L., "A Data Analysis Approach to Square Tables," Comm. in Stat., July 1975.

[12] Alvey, W., and Cobleigh, C., "Exploration of Differences Between Linked Social Security and Current Population Survey Earnings Data for 1972," 1975 Amer. Stat. Assn. Proc. Soc. Stat. Sec.

[13] Johnston, M. P., "Evaluation of Current Population Survey Simulations of Payroll Tax Changes," 1975 Amer. Stat. Assn. Proc. Soc. Stat. Sec.

[14] Hansen, M., and Hurwitz, W., "The Problems of Non-Response in Sample Surveys," J. Amer. Stat. Assn., vol. 41, 1946, pp. 517-528.

[15] Deming, W. E., and Stephan, F. F., "On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Tables are Known," Annals Math. Stat., vol. 11, pp.427-444, 1940.

[16] Feinberg, S. E., "An Iterative Procedure for Estimation in Contingency Tables," Annals Math. Stat., vol. 41, 1970, pp. 907-1017.

[17] Neter, J., Maynes, E. S., and Ramanathan, R., "The Effect of Mismatching on the Measurement of Response Errors," J. Amer. Stat. Assn., vol. 60, 1975, pp. 1005-1027.