IMPUTATIONS, RESPONSE ERRORS, AND MATCHING IN DUAL SYSTEM ESTIMATION

Howard Hogan and Charles D. Cowan, U.S. Bureau of the Census

1.1 Introduction

A good way to begin this discussion is by quoting from a paper delivered to the ASA some 15 years ago. Talking then of the 1960 census, Walter Perkins and Charles Jones began: "Perhaps it is true that 'a rose is a rose is a rose' this is a little outside our field. We do, however, disagree with anyone who makes the same claim for a match." They then add:

"There is a rather mysterious category - 'the impute;' the census interviewer has been told that a given address is occupied, but has been unable to find anyone home on repeated visits. The computer has imputed at this address a young couple plus a five year old daughter. On a revisit, however, we find the occupant at the time of the census was an elderly man. Was he counted - or wasn't he? It would seem clear that he wasn't; yet, from another point of view, it appears that he has been counted - not wisely, but too well!!"

The problem of the impute epitomizes a general problem in dual-system estimation: the problem of insufficient or erroneous information. Nonresponse is, of course, a standard problem in survey samples. Indeed, imputations were created as a way to deal with this problem. Further, every survey will contain erroneous responses.

1.2 Review of Dual-System Estimation

Although many people here are familiar with dual-system estimation, a quick overview may be helpful to provide perspective. As the name implies, we have two record systems, each of which records some but not all of the population. Thus, in our case, we have a census (System 1) and a followup survey (System 2). By careful case-by-case matching we can ascertain that a given number of cases are in both systems, but certain cases are only in one or the other of the two systems. Let:

N_1 be the number of cases recorded in System 1 (here the census count);

N_2 be the number of cases recorded in System 2, and

M be the number of cases recorded in both systems.

If system 2 is a sample survey, one normally inflates the results by the sampling fraction, although in this simple case it is not necessary. Now, let the probability of being recorded in one system (P_1) be independent of the probability of being recorded in the other (P_2); then a maximum likelihood estimate of the total population is:

$$N_1 / \frac{M}{N_2} = \frac{N_1 N_2}{M}$$

That is, we inflate the census results by the proportion matched in the other system. This assumes a perfect system where each case can be identified and matched exactly. One can refine this system somewhat by making separate estimates for separate subgroups, as pointed out by Chandrasekar and Deming in their pioneering 1949 paper.

1.3 Overview of Problem

But what does one do if there are records in either system which cannot be uniquely identified? Consider again the case of the imputed family. They are clearly included in the census count, but can we match to them? Even if the elderly man was recorded in the survey, can we match that entry against the imputed people in the census?

A similar problem is that some of the cases counted in either system may be out-of-scope, or should otherwise not have been counted. A case which commands our attention is that of erroneous enumerations. That is, households which were counted more than once, or in the wrong place. In addition, there are those cases fabricated by clever or lazy enumerators. These cases must be subtracted from the total counts for each system since they inflate the totals artificially with persons who don't really exist.

These two classificatory problems lead to total counts from each system which contain persons who are not matchable because of imputes or erroneous enumerations. A census case which was imputed in the total census count via the Hot-Deck procedure because of nonresponse cannot be matched to a record for the same household in a followup survey. Insufficient information in the census for an exact count also means that there is insufficient information for matching. This is true even if the followup survey has sufficient information for matching. Each system has to have enough information to determine that the identified person is in both systems.

Similarly, there will be unmatchable followup survey cases, chiefly refusals where it was not possible to at least record the name, but also including persons who have moved for whom we did not get a satisfactory address. In addition, there will be uniquely identifiable (and thus seemingly "matchable") cases which were counted but should not have been. Most prominent among them are the ficticious persons, but persons born after the census are also included here; as well as multiple enumerations.

Our proposed solution is rather simple - one should determine the number of nonmatchable cases and subtract them from the counts of both systems. A normal dual-system estimate would be made on the residual. While simple, and seemingly obvious, this solution is not without its problems. A good statistician hates to throw away data, as this procedure does. One strives to devise matching rules which use all the data. But procedures which try to match on little or no information, perhaps by balancing erroneous matches with erroneous nonmatches, are bound to increase the variance of the dual system estimate, and if they fail, to lead to an even bigger bias.

2.0 The Model

2.1 Notation

We can denote our census and survey counts as:

N_{1_A} the census count for group A

$\hat{N}_{2_A}$ estimate of unmatchable census cases

$\hat{EE}_{1A}$ estimate of erroneous and out-of-scope census cases

$N_{1A}-\hat{II}_{1A}-\hat{EE}_{1A}$ estimated number of persons correctly enumerated and matchable in the census for group A

N_{1A} estimate of population of followup survey persons (inflated to population total)

$\hat{II}_{2A}$ estimate of unmatchable followup survey cases

$\hat{EE}_{2A}$ estimate of erroneous and out-of-scope followup survey cases

$N_{2A}-II_{A}-EE_{A}$ estimated number of persons correctly enumerated and matchable in the followup survey

In addition we will have

M_A = estimate of number of followup survey persons enumerated in (i.e., matched to) the census for group A.

2.2 Estimation without Errors

Table 1 shows the usual dual-system estimation problem. The top panel gives the counts. N_A (the true population total) is, of course, not observable and must be estimated. The lower panel gives the probabilities. The quantities a, and b, are the probabilities of being enumerated in each source.

Now, it is only necessary to make the usual dual-record-system assumption of independence between the event of being enumerated in System 1 (the census) and the event of being enumerated in System 2 (the survey) to derive population estimates. Table 1 shows the derivation of the dual-system estimate in a perfect system. In terms of the expected values, we have

1) $N_{1A} = a_1 N_A$

2) $N_{2A} = b_1 N_A$

3) $M_A = a_1 b_1 N_A$

4) $N_A = (a_1 N_A)(b_1 N_A)/M_A$
$= N_{1A}\, N_{2A}/M_A$

which we estimate by

5) $\hat{N}_A = \dfrac{\hat{N}_{1A}\,\hat{N}_{2A}}{\hat{M}_A}$

which is the classic dual system estimator given above.

2.3 Estimation with Imperfect Information

Table 2 shows the counts actually obtained in the census and the followup survey as the sum of three marginal totals, only two of which are directly observable. To obtain the number of persons in the census or survey who are also matchable, one has to subtract the values not to be included in the estimation. Thus:

Number of Truly Matchable Persons in Source i = $N_{iA} - II_{iA} - EE_{iA}$

The process of dual system estimation works exactly the same way in the expanded table as with the two by two tables shown in Table 1. The difference is that the probabilities on Table 2 are more detailed as to the categories they represent. A case in either source can be correctly enumerated, not enumerated, or erroneously enumerated, and if correct the case may or may not have sufficient information for matching. This gives four possible categories in which a record or case may be put, and so cases in each system are multinominally distributed.

These probabilities are represented in the marginals on the right of Table 2 as a_1' to a_4' and b_1' to b_4'. The values EE_{1A} and EE_{2A}, however, will be subtracted from the total counts for the census and the followup survey and are not to be represented in the matching or the total population count for the U.S. because they are erroneous, and so the probabilities to be used in the derivation of the dual system estimate of the total population are just those involved in being correctly enumerated, matched or not, and not enumerated. (Probabilities a_1 to a_3 and b_1 to b_3 on the right side of Table 2.) Again assuming independence between the two distributions one can derive an estimator of the total population:

1') $N_{1A} - II_{1A} - EE_{1A} = a_1 N_A$

2') $N_{2A} - II_{2A} - EE_{2A} = b_1 N_A$

3') M = prob (In/Matchable & In/Matchable)$\cdot N_A$

4') $M = a_1 b_1 N_A$
$= (a_1 N_A)\ (b_1 N_A) / N_A$
$= (N_{1A} - II_{1A} - EE_{1A})(N_{2A} - II_{2A} - EE_{2A}) / \hat{N}_A$

so,

5') $N_A = (N_{1A} - II_{1A} - EE_{1A})\,(N_{2A} - II_{2A} - EE_{2A})/M_A$

But this assumption of independence is less believable that the assumption made with only two categories for each source. Besides correlation bias due to being included in both sources or neither source, correlation bias can now also be due to being included, but not collecting enough information for matching purposes.

2.4 An Example

To make this presentation a bit more concrete three examples of estimation are presented in Tables 3a, 3b, and 3c.

Table 3a presents the case of complete independence, both between counted and not counted; and, between correctly enumerated and insufficient information. (Erroneous enumerations are assumed to have been subtracted from the table.) In this case, the estimator yields the correct total population.

In Table 3b correlation bias only exists as between counted and not counted. However, the conditional probabilities of being correctly enumerated given the person was counted are independent. Here the estimator is biased downward. But, this is a problem with which we must deal regardless of how we handle imputations, closeouts, etc.

Table 3c is the more realistic. Correlation bias exists both between counted and not counted and between correctly enumerated and insufficient information. The latter bias is assumed not to be so severe. Our estimator is biased further downward.

Of course, we know census closeouts and imputations do differ from the correctly enumerated population, just as completely missed people differ. There will be correlation bias, and the bias will depend, in part, upon which cases are excluded as unmatchable.

3.0 A Model of Bias

In this section, we will present a model to show the precise effects of correlation and error on the estimation procedure. In the case with only two categories for each source, there is only one parameter to represent net correlation and error when the marginals of the table are fixed. In the table below, the probabilities in the table represent the probability of cell membership under the assumption of independence plus or minus a factor (α) representing the correlation between the two sources, error in matching or categorization, or other sources of bias.

		Followup Survey		
		In	Out	Total
C	In	$a_1b_1+\alpha$	$a_1b_2-\alpha$	a_1
E	Out	$a_2b_1-\alpha$	$a_2b_2+\alpha$	a_2
N	Total	b_1	b_2	1.0
S				
U	$-a_2b_2 \leq \alpha \leq (1-a_1b_1)$			
S				

In this case as before, the number of matches observed is a proportion of the total population size:

$$= (a_1b_1+\alpha)\, N_A$$

$$M_A = \frac{N_A a_1 b_1 N_A + \alpha N_A^2}{N_A}$$

$$= \frac{N_{1A}\, N_{2A}}{N_A} + \alpha N_A$$

so

$$7) \quad N_A = \frac{M_A - \sqrt{M_A^2 - 4\,\alpha\, N_{1A}\, N_{2A}}}{2\,\alpha}$$

Estimator (7) indicates how one could evaluate the effects of errors in matching or assuming independence between sources. Departures from independence lead to significant biases, as can be seen by taking the difference between (7) above and the estimate (5) given above. Note that (7) is equal to (5) in the limit as α approaches zero, using L'Hôpital's rule. The value α used above is prespecified (really a function of our procedures in the field and matching) and so yields a family of estimators.

The most unfortunate aspect of this formulation is the extremely rapid change in the estimate of true value of the total population in the region where we would like the assumption of independence to hold (where $\alpha = 0$). In fact, at $\alpha = 0$, the slope of the function is infinite, implying that the estimator is extremely volatile, relying heavily on independence. Nonrandom errors in matching, or correlation bias between the two data sources lead to large changes in the estimate.

Now consider the estimator in the presence of problems in identifying where a case is matched. If insufficient information is a problem, such that some cases can be included for marginal totals from either source, but are not considered matchable, then several sources of bias can affect the estimator. In the table below, departures from independence for the probabilities of cell membership are represented as α, β, η, γ, which are biases of identification and completeness in the followup survey.

The value η represents a bias of completeness in both surveys (the census and the followup survey). The value α represents a bias of identification in both surveys. The values β and γ are interaction terms, indicating biases of both completeness and identification. The value β is a bias of completeness in the census, but identification in the followup survey; the value γ is the obverse.

A bias of completeness is the bias due to cases being missed in the census having a higher than expected probability of being missed in the followup survey. A bias of identification is missing more information than expected if missing information is truly a random event for each of the two sources.

		Followup Survey			
		Correctly Enumerated			
		In	II	Out	Total
C	In	$a_1b_1+(\alpha+\beta+\eta+\gamma)$	$a_1b_2-\eta-\gamma$	$a_1b_3-\alpha-\beta$	a_1
E	II	$a_2b_1-\eta-\beta$	$a_2b_2+\eta$	$a_2b_3+\beta$	a_2
	Out	$a_3b_1-\alpha-\gamma$	$a_3b_2+\gamma$	$a_3b_3+\alpha$	a_3
	Total	b_1	b_2	b_3	1.0

Following the same procedure as for the two by two table, the estimator of total population if the correlations were known would be:

$$8) \quad N_A = \tfrac{1}{2}(\alpha+\beta+\eta+\gamma)^{-1}\left[M - \left[M_A^2 - 4(\alpha+\beta+\eta+\gamma)(N_{1A}-II_{1A}-EE_{1A})(N_{2A}-II_{2A}-EE_{2A})\right]^{\frac{1}{2}}\right]$$

As in (7), this estimator is very volatile in the presence of departures from independence. Furthermore, any of the sources of bias (correlation) will lead to drastic differences between the true population value and the estimate (5') obtained by assuming independence.

This formulation differs from that of Jabine and Bershad, who use the phi coefficient as a measure of correlation between the two sets of observations. In their formulation, similar problems arise with their estimator, though they are not explored in the 1968 paper. Their correlation coefficient (ρ) can only range between certain values ($-(1-a_1) < \rho < a_1$) which are similar to the restrictions placed on the value of γ above, and their estimator (which also reflects the correlation between sources) becomes infinitely large (has a singularity) within the viable range of alternatives for the correlation coefficient. In either the Jabine/Bershad formulation, or the one used in this paper, the conclusion is that the dual system estimator is badly behaved if the assumptions underlying its use are even moderately violated.

4.0 Operational Considerations

The important constraint in designing a survey to measure erroneous enumerations, or defining sufficient information for matching is that the independence of the census and the followup survey must not be compromised. We must be careful not to turn the dual record system into a double entry accounting system. Not being able to locate a record in the other system cannot be allowed to be grounds for defining it II or EE.

This may seem obvious, but there is a strong tendency to do just that. It is extremely easy to set up a matching procedure which first searches the other system in an attempt to find a match. If a match is found, the record is coded "Matched." If no record is found, the personal and address information is carefully examined, and it is often determined that there was not really enough information for matching, and the case is coded II instead of being considered a true nonmatch (out of one system). The errors introduced by this biased procedure can be of the same magnitude as the number of true misses.

4.1 Measuring Insufficient Information Cases

Checking for sufficient information is a simple clerical problem. Clear rules can be established for minimal information. Establishing optimal rules for sufficient information however, is one of the most important decisions to be made. Overly stringent rules lead to increased correlation bias, as well as increased variance by reducing the size of the usable sample. Overly loose rules of sufficient information are bound to lead to an increase in erroneous nonmatches and erroneous matches, and these in all probability will not balance out. This is not a trivial problem--even in the case of imputations.

First, it is clear that we need to know the number of cases with insufficient information for matching for each subpopulation for which we will be controlling. This requires an extensive cross tabulation of imputations--as well as of counts.

Further, there is not an exact correspondence between imputations and insufficient information for matching. An imputation may be matchable. There may be nothing wrong with the questionnaire except that the entries were made too lightly to be machine readable. On the other hand, some census questionnaires may pass edit, but will lack name or other information vital to the matching.

4.2 Measuring Erroneous Enumerations

Measuring erroneous enumerations is a more difficult matter. It involves, in essence, second guessing the census. One must return to the field, usually months later, and find out whether these people exist and whether they were correctly enumerated.

4.2.1 Whether someone exists or not is conceptually a simple problem. Operationally, it is not so easy. The difficulty is to ensure that we do not throw out, as erroneous enumerations, members of groups with low social visibility. We do not want to throw out those cases most likely to be missed by either system. The fact that someone was missed by the followup survey is not sufficient evidence of erroneous enumeration. We need additional information, for example, that a building does not exist, or another family was living in that house. If no one in a neighborhood knows a household, we can throw them into the category EE--albeit at some risk.

4.2.2 Determining whether someone was correctly enumerated is conceptually and operationally more difficult. One must decide not only that someone exists and should have been enumerated, but also where they should have been enumerated. The Bureau has identified two approaches:

Definition I - A person is "correctly enumerated" if he should have been enumerated and was enumerated once and only once, even though it might have been in an incorrect location. A person is "missed" if he should have been enumerated in the census but was not enumerated in any location. An enumeration is considered to be an "erroneous enumeration" if the person should not have been enumerated but was (e.g., he did not exist, lived outside the U.S., was born after the census or died before the census), or the person should have been enumerated but was enumerated more than once.

Definition II - A person is "correctly enumerated" if he was enumerated in the census at the address reported by the followup survey as the census date residence. A person is "missed" if he was not enumerated at the census date residence that was reported in the followup survey. An enumeration is considered to be "erroneous" if the followup survey reports that the person was not living at the location where the census recorded him. For example, the followup survey could report that no such person exists, or that the person was born after the census, died before the census or was living elsewhere on census date.

The Census Bureau has found that it is impossible to search all locations where a person might have been enumerated. So we are forced into Definition II. But, while seemingly clear for the purpose of defining misses, the definition must be carefully used in dealing with erroneous enumerations. In theory, where one reports one should have been enumerated should be the same regardless of how one is sampled for a followup survey (System 2). But for the people who move between the census and the followup survey, serious problems can arise. This brings us to our next issue: misstatement of address.

One of our serious problems is that many people misstate their Census Day address. Many people report that they were living "here" during the followup survey even though they have moved. A less common problem is people who report their address as "there" during the followup survey even though they moved before Census Day. This phenomenon, known as telescoping, has been uncovered in other studies with with the same net result. Careful probing can reduce this problem, but it cannot eliminate it. Clearly, anyone who misstates their Census Day address will be counted as missed. This must be properly balanced in the E-Sample by treating people who misstate their address as erroneously enumerated. There are two ways of doing this: one potentially unbiased, but expensive, one potentially biased but cheap.

The potentially unbiased method is to followup out movers, and interview them at their new address. The interview would be a normal "System 2" survey interview. They would be asked "where they were living on Census Day." If they correctly reported their previous address, they would be counted as correctly enumerated there. If they incorrectly reported their old address, we would treat them as erroneously enumerated at the old address. Thus, the treatment of misreporting of address is the estimation of erroneous enumerations would be consistent with the estimation of omissions.

The other approach is to accept the word of the current occupant as to who was living there on Census Day. Thus, if the current occupant wrongly reports that he was living "here" on Census Day we accept this. If he also reports that the previous occupants moved out before he moved in, we accept that. Clearly, any other family enumerated in the housing unit at the time of the census was erroneously enumerated--if we accept the word of the current occupants! Again, the reports may be inaccurate but they are consistent and balancing.

5.0 Conclusion

The methods we have outlined are a way to handle a difficult problem. However, they do not solve the problem, any more than hot-decking has solved the problem of nonresponse.

As always, field work should be done so as to minimize nonresponse, and erroneous enumerations. Matching rules should be constructed to keep the insufficient information category as small as possible. But the problem will exist and all one can do is to attempt to handle it in an unbiased manner.

Table 1: Derivation of Dual System Estimates
A Perfect System
(complete information for each record and no erroneous enumerations)

Followup Survey Counts

CENSUS	In	Out	Total
In	M_A		N_{1A}
Out			
Total	N_{2A}		N_A

Followup Survey Probabilities

CENSUS	In	Out	Total
In	a_1b_1		a_1
Out			$1-a_1$
Total	b_1	$1-b_1$	1.0

Table 2: Derivation of Dual System Estimates
A System with Imperfect Information
(missing information for enumerated households and erroneous enumerations in the counts)

Followup Survey Counts

CENSUS	Correctly Enumerated: In	Correctly Enumerated: II	Correctly Enumerated: Out	Sub Total	EE	Total
Corr. Enum. In	M_A				∅	
Corr. Enum. II				II_{1A}	∅	II_{1A}
Out						
Sub Total		II_{2A}		N_A		N_{1A}
EE	∅	∅				EE_{1A}
Total		II_{2A}			EE_{2A}	

N_{2A}

Followup Survey Probabilities

	Correctly Enumerated: In	Correctly Enumerated: II	Correctly Enumerated: Out	Sub Total	EE	Total
Corr. Enum. In	a_1b_1	a_1b_2	a_1b_3	a_1		a_1'
Corr. Enum. II	a_2b_1	a_2b_2	a_2b_3	a_2		a_2'
Out	a_3b_1	a_3b_2	a_3b_3	a_3		a_3'
Sub Total	b_1	b_2	b_3	1.0		a_4'
EE						
Total	b_1'	b_2'	b_3'	b_4'		1.0

Table 3: Estimation of the Size of the Total Population

Table 3a: Estimation when the Assumption of Independence Holds

	PES				
	Counted				
Census	Total	Corr. Enum.	II	Not Counted	Total
Total Counted	720	675	45	180	900
Corr. Enum.	640	600	40	160	800
II	80	75	5	20	100
Not Counted	80	75	5	20	100
Total	800	750	50	200	1,000

$$\hat{N} = \frac{(900-100)(800-50)}{600} = 1,000$$

Table 3b: Estimation when Identification by Each Source is Correlated

	PES				
	Counted				
Census	Total	Corr. Enum.	II	Not Counted	Total
Total Counted	720	675	45	180	900
Corr. Enum.	640	600	40	160	800
II	80	75	5	20	100
Not Counted	60	50	10	40	100
Total	780	725	55	220	1,000

$$\hat{N} = \frac{(900-100)(780-55)}{600} = 967$$

Table 3c: Estimation when both Identification and Completeness of each Source is Correlated

	PES				
	Counted				
Census	Total	Corr. Enum.	II	Not Counted	Total
Total Counted	720	665	55	180	900
Corr. Enum.	640	600	40	160	800
II	80	65	15	20	100
Not Counted	60	50	10	40	100
Total	780	715	65	220	1,000

$$\hat{N} = \frac{(900-100)(780-65)}{600} = 953$$