ESTIMATION EQUATIONS FOR THE NUMBER OF DUPLICATE HU LISTINGS IN THE 1980 CENSUS, BASED ON A SAMPLE OF ED CLUSTERS

David W. Chapman
U.S. Bureau of the Census

1. Introduction

Based on a comparison of housing unit (HU) counts from the 1980 Census and independent estimates based on updates of 1970 Census HU Counts, it appeared that there might be a national overcount of housing units in the 1980 Census of between 1.5 and 2.0 million. Consequently, some immediate investigation of the sources and levels of HU duplication was initiated.

Part of this investigation included a search for duplicate listings in two Washington, D.C. area local district offices. This search revealed some duplication of HUs between bordering enumeration districts (EDs) that apparently occurred because the quality of the maps was poor enough so that the boundaries between bordering EDs were not adequately defined.

Consequently, a survey was designed to estimate the number of duplicate HU listings in the 1980 Census, using a sampling plan that provided for the selection of clusters of "bordering" EDs. The focus of the data collection activities was to check for duplicates both within EDs and across ED boundaries.

Specifically, a sample of 80 1980 Census EDs--20 per census region--was selected in November 1980 in order to estimate the number and percent of duplicate HU listings.* In addition to the initial sample of 80 EDs, each ED that bordered a selected ED was also included in the sample. As a result, a total sample of about 550 EDs was selected.

The primary objective of this project is to obtain rough estimates, in a short period of time, of the amount of duplication of HU listings in the 1980 Census. Consequently, a relatively small sample size was chosen and the data collection procedures were designed to exclude any field work. Checks for duplicates consisted of matching the individual entries in the master address registers of the EDs selected in the sample. The address registers of the selected EDs were keypunched and then a computer match of ED address listings was carried out to identify potential duplicate listings. The questionnaires for the potential duplicate listings were then checked in order to determine whether or not these listings were actually duplicates. Since no field work was included in this project, the final determination of whether two or more HU address listings were duplicates was based on the questionnaire comparisons.

It was recognized at the beginning of the project that certain types of duplicate listings would often be missed in this address-match procedure. For example, a duplicate listing that occurs because a housing unit has two distinct addresses, such as a street address and a rural route address, would not generally be identified. In some rural areas names were available on the address registers and were included in the matching process. In these areas some matches involving different addresses were identified. This partial weakness in the survey's capability to identify these types of duplicates was assumed to be a reasonable price to pay to obtain rough estimates of the duplication rates relatively quickly and inexpensively.

The estimates of the level of duplication obtained from this address-match procedure will tend to underestimate the total number of duplicates in the 1980 Census. Even so, the estimates obtained from this survey should give an indication of whether or not the number of duplicate HU listings in the 1980 Census is excessive.

In the next section the definitions of some terms are given. In Section 3 a description of the sample selection procedure is presented. The estimation formulas are given in Section 4, followed in Section 5 by the application of these formulas to three hypothetical populations. The final section includes some conclusions and recommendations.

2. Terminology

In order to discuss the sampling plan and present the estimation formulas, some definitions are needed. The target EDs are the 80 EDs initially selected for the sample. Two EDs are adjacent or bordering if their boundaries have at least one point in common. The bounding EDs are those that are adjacent to a target ED. A target ED and its bounding EDs will be referred to collectively as an ED cluster.

A within-ED duplicate is a duplicate listing that appears in the same ED as the initial listing. If two or more duplicates of a listing appear within an ED, each extra listing counts as a separate duplicate. A between-ED duplicate is a duplicate listing for which the initial and duplicate listings appear in different EDs. If more than one duplicate listing appears in the second ED, only one between-ED duplicate is counted. The other duplicate in the second ED would be counted as a within-ED duplicate. However, if a listing in one ED is duplicated in each of two separate EDs, two between-ED duplicates would be counted.

Most of the 1980 Census was carried out by mail. Mailing lists for EDs were generated in two basic ways. In many areas commercial lists were purchased by the Census Bureau from private companies. Areas serviced in this way are referred to as tape address register (TAR) areas. In other areas the address lists for the EDs were obtained from a census prelist operation. The prelist activity involved a canvassing and careful listing by Census Bureau personnel of the housing units in the ED.

3. Sample Selection

The sample was selected in November 1980 from those EDs in the most recent Field Count Capture 2 File, except those in conventional census areas and those with zero population. The conventional areas include only about 5% of the HUs in the country.

Prior to selection, the EDs in each region were sorted by the following characteristics:

(1) TAR vs. Census Prelist
(2) District Office Code
(3) ED Code

After the EDs in a region were sorted, a straightforward systematic 1 in k sample of 20 EDs was selected, using a random start. The four selection (or skip) intervals used in this selection were 2606 (Northeast), 3626 (North Central), 4663 (South), and 2073 (West). The use of systematic sampling applied to a sorted list provides the sample with a "stratification effect" from the sort variables. In addition to the 80 "target" EDs , all EDs adjacent to the target EDs were included in the sample. The total sample consisted of 568 EDs.

Some consideration was given to the possibility of oversampling certain types of EDs, such as "large" EDs and EDs located near boundaries between TAR and prelist areas. However, due to the desire to carry out the survey in a relatively short period of time, and due to the uncertainty regarding the improvement in estimation precision that oversampling would have, it was decided to use the relatively simple procedure of systematic, equal probability selection.

4. Estimation Formulas

The following four estimators of the total number of duplicate HU listings in the 1980 Census, excluding conventional areas, are being considered:

Simple:

$$x_0' = \sum_{j=1}^{4} k_j \sum_{i=1}^{20} x_{jiw} + \sum_{j=1}^{4} k_j \sum_{i=1}^{20} x_{jib}/2 \qquad (1)$$

Multiplicity:

$$x_1' = \sum_{j=1}^{4} \sum_{i=1}^{20} \sum_{m=1}^{n_{ji}} \frac{k_j}{t_{jim}+1} x_{jimw} + \sum_{j=1}^{4} \sum_{i=1}^{20} \sum_{q=1}^{t_{ji}} \frac{k_j}{t_{jiq}+2} x_{jiqb} \qquad (2)$$

Ernst:

$$x_2' = \sum_{j=1}^{4} \frac{N_j}{\sum_{i=1}^{20} n_{ji}} \sum_{i=1}^{20} u_{jiw} + \sum_{j=1}^{4} \frac{k_j}{t_j + 2} \sum_{i=1}^{20} u_{jib} \qquad (3)$$

Chapman:

$$x_3' = \sum_{j=1}^{4} \sum_{i=1}^{20} \sum_{m=1}^{n_{ji}} \frac{k_j}{t_{jim}+1} x_{jimw} + \sum_{j=1}^{4} \sum_{i=1}^{20} \sum_{m=1}^{n_{ji}} \frac{k_j}{t_{jim}+1} \cdot \frac{t_{jim}}{r_{jim}} \cdot \frac{x_{jimb}}{2}, \qquad (4)$$

where

k_j = the selection interval for region j,

x_{jiw} = the number of duplicates within the i-th target ED selected from the j-th region,

x_{jib} = the number of duplicates between the i-th target ED selected from the j-th region and the ED's adjacent to it in region j,

u_{jiw} = the number of within-ED duplicates found in the i-th ED cluster selected from the j-th region,

u_{jib} = the number of between-ED duplicates found in the i-th ED cluster selected from the j-th region,

x_{jimw} = the number of duplicates within the m-th ED identified in the i-th ED cluster selected from the j-th region,

x_{jimb} = the number of duplicates between the m-th ED identified in the i-th ED cluster selected from the j-th region and the ED's adjacent to it in the sample,

x_{jiqb} = the number of duplicates between the q-th pair of adjacent ED's identified in the i-th ED cluster selected from the j-th region,

N_j = the total number of EDs in the population in region j, excluding those in conventional areas,

n_{ji} = the number of EDs in the i-th ED cluster selected from region j,

t_{ji} = the total number of pairs of adjacent EDs in the i-th ED cluster selected from region j,

t_{jim} = the total number of EDs in region j that are adjacent to the m-th ED identified in the i-th cluster selected from the j-th region,

r_{jim} = the number of EDs in the sample ED cluster that are adjacent to the m-th ED identified in the i-th cluster selected from the j-th region,

t_{jiq} = the total number of EDs in region j that are adjacent to the q-th pair of adjacent EDs identified in the i-th ED cluster selected from the j-th region and

t_j = the sample average number of EDs that are adjacent to both members of a pair of adjacent EDs in region j, one of which is a target ED. This average will be calculated over all adjacent pairs in the sample from region j that consist of a target ED and a bounding ED.

If there are no between-ED duplication cases in the population that involve more than two EDs, the simple estimator and multiplicity estimator are unbiased.** The other two estimators are biased.

The simple estimator is an unbiased estimator that uses only those duplicates that involve a target ED. Duplicates that occur within or between bounding EDs are not included.

The multiplicity estimator is an unbiased estimator that is based on all the duplicates identified in the sample clusters. Since it uses all the sample information, the multiplicity estimator has a considerably lower variance than the simple estimator. The disadvantage of the multiplicity estimator is that there is a considerable amount of additional map work that is required in order to be able to determine the adjacency counts (multiplicity factors) needed to apply the estimator. Specifically, it can be rather difficult and time consuming to determine from the maps the number of EDs adjacent to a bounding ED.

The Ernst estimator*** is a biased estimator that uses all the duplication data from the sample clusters. This estimator can be generated by replacing the two adjacency count (multiplicity) factors, t_{jim} and t_{jiq}, in the multiplicity estimator by regional averages of these factors, where the average is taken over all EDs and ED pairs involving the target EDs. Specifically, the number of adjacencies, t_{jim}, for the m-th ED identified in the i-th cluster selected from the j-th region is replaced by the average of the number of EDs adjacent to the 20 target EDs selected from the j-th region. Similarly, the number of EDs adjacent to both members of the p-th ED pair, t_{jiq}, is replaced by t_j, the average number of EDs adjacent to all the pairs in the j-th region that consist of a target ED and a bounding ED. The advantage of the Ernst estimator is that no additional map work is needed for its application.

The Chapman estimator is a biased estimator that utilizes the duplication data obtained from all of the EDs in the selected clusters. The first term of this estimator, the within-ED duplication term, is identical to that of the multiplicity estimator. With the second term an attempt was made to provide a nearly unbiased estimator of the total number of between-ED duplicates by assigning half of a between-ED duplicate to each of the two EDs involved in a between-ED duplication. The counts of between-ED duplicates for the EDs are weighted-up and summed across the EDs in the cluster.**** In addition to being biased, a disadvantage of the Chapman estimator is that its use requires additional map work to count adjacencies.

Although the estimation formulas given in equations (1)-(4) are in terms of the number of HU duplicate listings, the estimated proportion of duplicate listings will probably be of greater interest. This estimated proportion is of the form:

$$p = x'/y', \tag{5}$$

where x' = one of the four estimators of the number of duplicate HU listings, and
y' = the sample estimate of the number of HUs, excluding conventional areas.

The form of the estimator y' is equivalent to the first term (i.e., the within-ED term) in x', with the number of within-ED duplicate HUs in the formula replaced by the number of HUs in the ED.

The variance of x' will be estimated using the ultimate cluster approach, assuming that the sample of ED clusters constitutes a simple random sample. The variance of p will be estimated using the ultimate cluster approach combined with the standard Taylor series approximation to the variance of a ratio. The details of variance estimation are given in an internal Census Bureau memorandom available from the author.

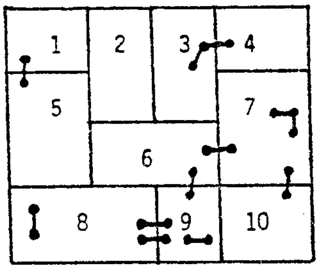
5. Numerical Examples

It is difficult to make general comparisons of the precision of these four estimators because of the uncertainty of how duplicates are distributed within and between EDs across the country. There could be considerable clustering of duplicates or they might be widely scattered. Also, the relative numbers of between and within-ED duplications is not known. Consequently, any attempt to try to develop a model of within and between ED duplications, from which general comparisons could be made, would be of questionable value.

To obtain some comparison of the precision of the four estimators, three hypothetical populations have been developed: one containing only ten EDs and the other two containing 60 EDs each. The population of ten EDs serves to illustrate the method of application of the estimators as well as to compare them. For all three populations the biases, variances, and mean square errors (MSEs) were derived, for various sample sizes, for each of the four estimators of the number of duplicate HU listings. The calculations of these statistics were based on the enumeration of all possible systematic samples of ED clusters.

The hypothetical example of only ten EDs is illustrated in Figure 1. For this population there are twelve duplicate listings as indicated by the short line segments.

Figure 1



Five of the 12 duplicates in this population are within-ED duplicates and seven are between-ED duplicates.

For this example only one ED cluster was selected to estimate the total number of duplicates in the population. For each of the ten possible sample clusters, the four estimators given in the previous section were used to estimate the total number of duplicates in the population. Each of these 40 estimates, broken down by a within-ED duplication component and a between-ED duplication component, is given in Table 1. Based on the ten estimates for each of the estimators, the expected value and mean square error for each of the four estimators were derived and are given in the last two rows of Table 1.

The two hypothetical universes of 60 EDs used to make comparisons of the four estimators have substantially different numbers and patterns of within-ED and between-ED duplications. Population I has a total of 56 duplicates: 20 within-ED duplicates and 36 between-ED duplicates. Population II has 120 duplicates: 80 within-ED duplicates and 40 between-ED duplicates. Even though the second population has more total duplicates, it has more EDs that have no duplication errors than does Population I (20 vs. 14). Also, the number of EDs that have a substantial number of duplicates (e.g. more than 4) is much higher for Population II.

Population I was constructed prior to the collection of most of the survey data. Population II was constructed after a substantial portion of the survey data was collected. In constructing Popu-

Table 1 Comparison of Estimators of the Number of Duplicate HU Listings for the Population of 12 EDs Illustrated in Figure 1

	Simple Estimator			Multiplicity			Ernst Estimator			Chapman Estimator		
Target ED	Within	Between	x'_0	Within	Between	x'_1	Within	Between	x'_2	Within	Between	x'_3
1	0	5	5	0	3.33	3.33	0	3.33	3.33	0	3.67	3.67
2	0	0	0	2.00	3.33	5.33	2.00	2.86	4.86	2.00	3.00	5.00
3	10	5	15	5.33	5.33	10.67	6.00	5.71	11.71	5.33	5.51	10.84
4	0	5	5	5.33	5.33	8.67	10.00	3.33	13.33	5.33	3.67	9.00
5	0	5	5	2.50	3.33	5.83	2.00	2.86	4.86	2.50	2.67	5.17
6	0	10	10	9.83	13.17	23.00	6.25	11.67	17.92	9.83	10.08	19.91
7	20	10	30	7.33	9.83	17.17	6.67	10.00	16.67	7.33	9.44	16.77
8	10	10	20	4.50	8.67	13.17	5.00	9.00	14.00	4.50	9.96	14.46
9	10	15	25	7.83	13.17	21.00	8.00	12.50	20.50	7.83	12.97	20.80
10	0	5	5	5.33	6.50	11.83	7.50	7.50	15.00	5.33	8.28	13.61
Mean	5	7	12	5.0	7.0	12.0	5.3	6.9	12.2	5.0	6.9	11.9
MSE			91			40.06			32.1			34.6

Total number of duplications in population = 12

Number of within duplications = 5
Number of between duplications = 7

lation II an attempt was made to represent the duplication patterns that were found in the survey. Consequently, the comparisons of the four estimators based on Population II should be more meaningful than those based on Population I.

The biases, variances, and MSEs for the four estimators for various sample sizes are given in Table 2 for Population I and in Table 3 for Population II.

Table 2. Comparison of the Four Estimators for 60-ED Hypothetical Population I

Number of within-ED duplicates = 20
Number of between-ED duplicates = 36

	Estimators			
	Simple	Mult.	Ernst	Chapman
n=1 cluster				
Bias	0	0	4.00	3.58
Variance	3044.0	901.28	1103.66	968.19
MSE	3044.0	901.28	1119.68	981.00
n=2 clusters				
Bias	0	0	3.82	3.18
Variance	1064.0	347.30	463.09	378.02
MSE	1064.0	347.30	477.70	388.10
n=4 clusters				
Bias	0	0	3.37	1.69
Variance	404.0	171.62	218.92	190.70
MSE	404.0	171.62	230.25	193.55
n=6 clusters				
Bias	0	0	3.56	2.46
Variance	489.0	129.46	173.82	104.20
MSE	489.0	129.46	186.50	110.28
n=10 clusters				
Bias	0	0	3.67	2.12
Variance	101.0	17.93	16.69	11.11
MSE	101.0	17.93	30.16	15.61

Table 3. Comparison of the Four Estimators for 60-ED Hypothetical Population II

Number of within-ED duplicates = 80
Number of between-ED duplicates = 40

	Estimators			
	Simple	Mult.	Ernst	Chapman
n=1 cluster				
Bias	0	0	6.45	2.59
Variance	21870.0	3627.26	3342.78	3729.94
MSE	21870.0	3627.26	3384.37	3736.64
n=2 clusters				
Bias	0	0	4.43	2.82
Variance	9255.0	1507.15	1199.31	1477.42
MSE	9255.0	1507.15	1218.98	1485.37
n=4 clusters				
Bias	0	0	4.57	2.60
Variance	3960.0	718.99	469.67	669.24
MSE	3960.0	718.99	490.51	676.01
n=6 clusters				
Bias	0	0	4.52	1.42
Variance	3680.0	513.07	436.68	473.40
MSE	3680.0	513.07	457.08	475.42
n=10 clusters				
Bias	0	0	4.61	1.39
Variance	2193.0	321.23	380.37	238.53
MSE	2193.0	321.23	401.64	240.45

6. Conclusions and Recommendations

An inspection of Tables 1-3 indicates that the MSE for the simple estimator is considerably higher than the MSEs for the other estimators in all cases. For the ten-ED population and the first 60-ED population the MSE for the simple estimator is typically two or three times larger than the MSEs for the other three estimators. For the second 60-ED population the MSE for the simple estimator is typically more than five times larger than the MSE for each of the other estimators. These discrepancies in MSEs are not surprising since the simple estimator is based on only those duplicates involving a target ED. The total sample of 568 EDs is about seven times larger than the number of target EDs, 80.

The MSEs for the three full-sample estimators are roughly equal. The MSE for the multiplicity estimator is generally the lowest for estimates for the first 60-ED population (Table 2), while the MSE for the Ernst Estimator is generally the lowest for estimates for the second 60-ED population (Table 3). The bias of the Ernst Estimator, which is roughly 5%-6%, is always higher than that of the Chapman Estimator.

Of the two unbiased estimators, simple and multiplicity, the multiplicity estimator has a considerably lower variance. However, the multiplicity estimator requires additional map work which would involve a substantial amount of additional time and expense. The Ernst Estimator has about as low a MSE for these hypothetical populations as does the multiplicity estimator. In fact for the second 60-ED population--the one which better represents the actual duplication patterns in the 1980 Census--the MSE for the Ernst Estimator is less than the MSE of the multiplicity estimator for all sample sizes except n=10 (target EDs).

The major concern regarding the Ernst Estimator is its bias. A bias of 5%-6% is not much of a problem for an estimator of a duplication rate of 1%-2%. However, the bias could presumably be considerably larger for the Ernst Estimator when applied to the actual survey data. This would be true if the number of within-ED (or between-ED) duplicates was highly correlated to the number of EDs adjacent to a single ED (or to a pair of adjacent EDs). However, even if some correlation exists, it seems unlikely that the bias of the Ernst Estimator would exceed 10%.

Because of its relatively low MSE, especially for the second 60-ED population, and since it does not require additional map work, the Ernst Estimator is recommended for estimating the number of duplicate HU listings. Methods of modifying the estimator slightly to reduce the bias are being explored. One such possibility would be to replace the ratio of ED counts, $N_j/\sum_i n_{j_i}$, in equation (3) by an analogous ratio of houseing unit counts. This could be especially helpful if there is a high correlation between the number of cuplications and the number of HUs within an ED.

Of course, if the additional time and expense involved in doing the map work required for the multiplicity estimator do not turn out to be significant factors, the multiplicity estimator would be preferred since it is unbiased and appears to have an MSE that is comparable to that of the other full-sample estimators.

FOOTNOTES

*The choice of the sample size of 80 EDs was based on some rough precision estimates for the estimated proportion of duplicate listings in a region. These precision estimates are discussed in an internal Census Bureau document that is available form the author.

**Additional terms could be added to accomodate duplication cases involving more than two EDs. However, no cases of this type occurred in the sample.

***This estimator was suggested by Larry Ernst of the Statistical Research Division of the Bureau of the Census.

****The weight-up factor t_{jim}/r_{jim} is needed because the between-ED duplicate count for an ED will only be based on the duplicates identified between that ED and the other EDs in the sample cluster.