BUREAU OF THE CENSUS
STATISTICAL RESEARCH DIVISION
TECHNICAL NOTE SERIES
No. TN-92/01

SUMMARY OF THE PERFORMANCE OF THE MAXIMUM OVERLAP ALGORITHMS FOR THE 1990'S REDESIGN OF THE DEMOGRAPHIC SURVEYS

by

Lawrence R. Ernst
Michael M. Ikeda
Bureau of the Census
Statistical Research Division
Washington, D.C. 20233

Report issued: November 2, 1992

Two different maximum overlap procedures were used in the 1990's redesign to overlap with the 1980's design. The first procedure, for designs which select one primary sampling unit (PSU) per stratum, was used for the Current Population Survey (CPS) and the National Crime Victimization Survey (NCVS). This procedure is detailed in Ernst (1986). For CPS, the procedure resulted in an average increase in expected overlap of .26 PSU/stratum and for NCVS the procedure resulted in an average increase in expected overlap of .30 PSU/stratum. The second procedure, designed for two-PSU-per-stratum designs, was used for the Survey of Income and Program Participation (SIPP). The procedure resulted in an average increase in expected overlap of .94 PSU/stratum. This procedure is detailed in Section 3.1 of Ernst (1989). Modifications allowing for different PSU definitions in the 1980's and 1990's designs are given in Ernst and Ikeda (1992). The modified procedure was the procedure that was implemented. Two 1990's strata were not overlapped because they contained too many PSUs. The overlap procedure used for SIPP employs a methodology that is more effective in increasing overlap than the procedure used for CPS and NCVS. However, the SIPP procedure requires the assumption of stratum-to-stratum independence in the initial design, an assumption only met by SIPP among these three surveys.

The one-per-stratum overlap procedure was used to overlap 375 CPS final (1990's design) strata. The average expected overlap was 0.621 PSUs/stratum, compared to an expected overlap of 0.363 under independent selection.

For NCVS, the one-per-stratum procedure was used to overlap 152 final strata. The average expected overlap was 0.508 PSUs/stratum compared to an expected overlap of 0.207 under independent selection.

The two-per-stratum procedure was used to overlap 103 final strata in SIPP. The average expected overlap was 1.523 PSUs/stratum compared to 0.582 for independent selection. Two strata (with 69 and 72 PSUs) contained too many PSUs to be overlapped during production. The cutoff was 57 PSUs. Note that strata with up to 68 PSUs have been successfully run (without the modifications in Ernst and Ikeda (1992)) on SRD's Solbourne Model 5/605 computer.

An "upper bound" for the expected overlap was calculated for each SIPP final stratum that was overlapped. The average "upper bound" was 1.616 PSUs/stratum, reasonably close to the average expected overlap of 1.523 using the overlap procedure. This "upper bound" was calculated since the two-PSU-per-stratum procedure is not an optimal procedure and we wanted to estimate how close to optimal the results are in practice. The optimal procedure would have been impossible to implement for most final strata because of the large size of

the required linear programming problem. Details on the "upper bound" are given in the appendix.

The maximum clock times during production for the one-PSU-per-stratum overlap program were fairly short. For CPS, the maximum clock time for a state (20 strata) was less than 1 minute on the Demographic Statistical Methods Division's (DSMD's) VAX Model 6420 computer. For NCVS, the maximum clock time for a region (44 strata) was about 2.5 minutes.

The maximum clock time during production for the two-PSU-per-stratum overlap program was longer. Most strata still took a short time to run. However, the maximum clock time for a region (44 strata, the largest contained 46 PSUs) was 1 hour and 40 minutes on DSMD's VAX Model 6420 computer. The two-PSU-per-stratum program takes longer because it requires many more variables than the one-PSU-per-stratum program to overlap a stratum with the same number of PSUs.

To obtain clock and computer times for various sized strata, an earlier version of the two-PSU-per-stratum overlap program was run on "final" strata containing different numbers of PSUs. The "final" strata came from two "final" stratifications obtained by stratifying the 1980's design SIPP non-certainty PSUs in the Midwest region using 1970 data.

The earlier version of the two-PSU-per-stratum program did not allow for different initial and final PSU definitions. However, assuming that the most time-consuming part of the program is the linear programming portion, the earlier version should take about the same amount of time to run as the final version if they are both run (on the same computer) on final strata with the same number of PSUs.

The clock and CPU times for final strata with different numbers of PSUs are given below. The earlier version of the two-PSU-per-stratum program was run on SRD's Solbourne Model 5/605 computer. The times are given in the format hrs:min:sec. Thus, the 65 PSU stratum took 2 hours, 19 minutes and 1.4 seconds of CPU time. There were two "final" stratifications, each with 31 strata. The median number of PSUs in a stratum (for the entire group of 62 strata) was 17 PSUs. The 37 PSU stratum had the 6th largest number of PSUs. The 68 PSU stratum was the largest stratum.

Number of PSUs	CPU Time	Clock Time
18	0:35.7	0:50.0
37	5:43.6	5:58.7
49	24:05.0	24:28.0
65	2:19:01.4	2:24:49.8
68	2:23:42.8	2:31:05.3

Note that the 65 and 68 PSU strata ran for almost identical times even though total array size goes up roughly as the fourth power of the number of PSUs. The explanation for this is not known.

REFERENCES

Ernst, L.R. (1986), "Maximizing the Overlap Between Surveys When Information is Incomplete," European Journal of Operational Research, 27, 192-200.

Ernst, L.R. (1989), "Further Applications of Linear Programming to Sampling Problems," SRD Research Report Series, No. RR-89/05, Bureau of the Census, Statistical Research Division.

Ernst, L.R., and Ikeda, M. (1992), "Modification of Reduced-Size Transportation Problem for Maximizing Overlap When Primary Sampling Units are Redefined in the New Design," SRD Technical Note Series, No. TN-91/01, Bureau of the Census, Statistical Research Division.

APPENDIX

The "Upper Bound" Calculation in the Two-PSU-Per-Stratum Overlap Program

The PSU definitions are not identical in the 1980's and 1990's SIPP designs. One of the preliminary steps in the two-per-stratum algorithm is a one-to-one matching of initial (1980's) and final (1990's) PSUs. Some artificial PSUs may be used as placeholders in the matching. These artificial PSUs are assigned zero probability and do not affect the results.

Now let S be the given final stratum containing the n final PSUs $A_1,...,A_n$ (possibly including some artificial PSUs). Let $B_1,...,B_m$ $m \geq n$ be the initial PSUs in the one-to-one matching plus all other PSUs which have a nonempty intersection with some PSU in final stratum S. $B_1,...,B_n$ are the initial PSUs in the one-to-one matching.

The program calculates, for each possible initial set, the expected overlap with each possible final pair in stratum S. An initial set can be a pair of PSUs, a single PSU, or the null set. Only the PSUs $B_1,...,B_n$ can be in an initial set corresponding to the given final stratum. If the actual initial set is a pair of PSUs, then at least that pair must have been in the actual initial sample. For the actual initial set to be a single PSU B_k ($1 \leq k \leq n$), B_k must have been the only PSU among $B_1,...,B_n$ in the actual initial sample. For the actual initial set to be the null set, none of $B_1,...,B_n$ can be in the actual initial sample.

For any given initial set, there is some final pair that has the largest expected overlap, conditional on this initial set. The "upper bound" assigns that final pair to sample whenever the given initial set is the actual initial set. The expected overlap calculation, however, does not use information about whether $B_{n+1},...,B_m$ were in sample. It is possible that a procedure that does use this information could produce a higher expected overlap than the "upper bound".

Consider first the case where there is a real one-to-one correspondence between initial and final PSUs for the given final stratum, that is when m=n. If the initial set is a pair of PSUs, then the largest possible expected overlap is 2 (this would be true in any case). If the initial set is a single PSU, then the largest possible expected overlap is 1, and if the initial set is the null set, then the largest possible expected overlap is 0. It is clear that, in this case, the "upper bound" will be a true upper bound.

The "upper bound" does not take into account the constraints on final PSU and final pairwise probabilities. It may, therefore, overestimate the optimal expected overlap. For example, suppose the initial set $\{B_1,B_2\}$ has a probability of .4. Suppose the final pair $\{A_1,A_2\}$ only

has a probability of .3. The "upper bound" will assume $\{A_1,A_2\}$ is the final pair whenever $\{B_1,B_2\}$ is the initial set. Since $\{A_1,A_2\}$ can only be in sample 75% of the time that $\{B_1,B_2\}$ is the initial set, the "upper bound" is overestimating the optimal overlap.

Now consider the case where there is not a real one-to-one correspondence. The following example illustrates that the computed "upper bound" may not truly be an upper bound in this case. Suppose that we have a final stratum with final PSUs A_1, A_2, A_3. We have initial PSUs B_1, B_2, B_3 in the one-to-one matching, initial PSU B_4 that intersects final PSU A_1, and initial PSU B_5 that intersects final PSU A_2. Let B_4 have a probability of .4 of being in sample and B_5 a probability of .3 of being in sample. Assume B_3, B_4, and B_5 were in different initial strata. Suppose the initial set consists only of PSU B_3. Then the "upper bound" assumes $\{A_1,A_3\}$ is in the final sample. The assigned expected overlap (conditional on B_3) for the "upper bound" is 1.4 (.4 for A_1 plus 1 for A_3). If we used information about whether B_4 and B_5 were in sample then the actual expected overlap, conditional on the initial set $\{B_3\}$, might be higher. If the new joint probabilities were large enough, we could assign $\{A_1,A_3\}$ as the final pair whenever B_4 was in sample and $\{A_2,A_3\}$ as the final pair when B_5 (but not B_4) was in sample. This would produce an overlap of 2 PSUs whenever either B_4 or B_5 was in sample, an event with a probability of .58. Therefore the actual expected overlap (conditional on B_3) would be 1.58, which is higher than the "upper bound" formula assigns in this case. Recall, however, that the "upper bound" does not take into account the constraints on final PSU and final pairwise probabilities. Because of this, the "upper bound" is perhaps more likely in general, to overestimate the optimal expected overlap than it is to underestimate it.