BUREAU OF THE CENSUS
STATISTICAL RESEARCH DIVISION
Statistical Research Report Series
No. RR91/11
*Previously TN91/01

Modification of the Reduced-Size Transportation Problem for Maximizing Overlap when Primary Sampling Units are Redefined in the New Design

by

Lawrence R. Ernst
Michael M. Ikeda

U. S. Bureau of the Census
Statistical Research Division
Washington D.C. 20233

Report Issued 6/10/92

1. INTRODUCTION

A maximum overlap method for two-PSU-per-stratum designs is described in Ernst (1989, Sec. 3.1). The procedure requires identical primary sampling unit (PSU) definitions in the initial and final designs. This note outlines two modifications to the Ernst procedure to account for different PSU definitions in the two designs.

The first modification is a procedure for establishing a one-to-one correspondence between the PSUs in the initial and final designs when PSU definitions are different in the two designs. The maximum overlap algorithm includes an ordering procedure for PSUs and pairs of PSUs that requires setting up a one-to-one correspondence between the initial and final PSUs. This correspondence can include dummy PSUs (artificial PSUs that have a zero probability of selection).

The second modification is a change in the calculation of the cost matrix to account for the possibility of several initial PSUs intersecting with one final PSU.

This note was motivated by the planned overlap of the 1990's Survey of Income and Program Participation (SIPP) sample PSUs with the 1980s SIPP sample PSUs, which will use the procedure of Ernst (1989) with these modifications. The procedure described first in Section 2 and 3 is the one that was actually programmed, due to the fact that it required the fewest changes to the program that had previously been written to implement the procedure of Ernst without these modifications. At the end of these sections changes need to implement an alternative approach that would avoid the need for dummy PSUs are detailed. A program that did not use dummy PSUs would require less computer memory.

2. CREATION OF A ONE-TO-ONE CORRESPONDENCE

The procedure to create the one-to-one correspondence between final and initial PSUs is as follows:

A. Sort all final PSUs in the design in descending order by final measure of size. (In practice for SIPP this is actually done separately in each region.)

B. Match each final PSU to the initial PSU that makes up the largest portion (in final measure of size) of the given final PSU. If that initial PSU has already been matched, use the initial PSU that makes up the 2nd, 3rd, ..., etc. largest portion. Do not match to any initial PSU that has already been matched.

C. Match remaining final PSUs to dummy initial PSUs. The initial stratum code assigned to the dummy initial PSUs rotates among the initial stratum codes.

D. Match remaining initial PSUs to dummy final PSUs. The final stratum code assigned to the dummy final PSUs rotates among the final stratum codes.

Note that step B will match identical final and initial PSUs to each other. The assignment of stratum codes to dummy initial and dummy final PSUs is arbitrary and does not affect the final results.

The ordering of the PSUs after creation of the one-to-one correspondence proceeds identically as described in Section 3.1 of the Ernst paper, except that ratios with 0 in the denominator (which can occur because of the dummy PSUs), are defined to be a very small positive constant, which ensures that pairs containing a dummy PSU appear at the end of each subordering within the main ordering.

The alternative approach, mentioned in the Introduction, that would avoid the need for dummy PSUs would proceed to obtain an ordering of pairs by first performing steps A and B, described above, but would not match unmatched real PSUs with dummy PSUs as in steps C and D. If n′ of the n final PSUs in a stratum are matched to an initial PSU it is only pairs of PSUs from these n′ PSUs that are ordered. The ordering is done precisely as in Ernst (1989) with n′ replacing n. This alternative approach together with corresponding changes in Section 3 would yield the identical overlap to the procedure actually implemented.

An additional possible change in the approach in A-D above would be to match each final stratum independently, instead of matching all PSUs in the design simultaneously. This would allow PSUs in different final strata to be matched to the same initial PSU. This change would generally yield a different overlap than with the approach in A-D. In particular, with this change, a pair of final PSUs in different strata that are matched to the same initial PSU would generally have a higher probability of joint selection in the final sample when the initial PSU was in the initial sample. Such an overlap situation may or may not be considered desirable. In any case, this change was not considered because of the additional programming modifications involved.

3. CALCULATION OF THE COST MATRIX

The cost matrix to be computed in this section uses the following guidelines that were agreed upon for the SIPP overlap. A PSU A selected in the final sample is considered successfully overlapped with the initial sample if there exists an initial sample PSU B for which $A \cap B \neq 0$. (The concept of partial successes, discussed in Ernst (1986, Sec. 5) is not considered here. All nonempty intersections are considered a complete successes.) Furthermore if the pair of PSUs A_1, A_2 are selected in the final sample and each of these PSUs overlap with an initial sample PSU, then this is calculated as two successful overlaps even if A_1 and A_2 overlap only with the same initial sample PSU.

Notation not defined below is as defined in Ernst (1989, Sec. 3.1).

Let S be the given final stratum.

Let $A_1,...,A_n$ be the final PSUs in the given final stratum, including dummy final PSUs.

Let $B_1,...,B_m$ be the initial PSUs in one-to-one correspondence with the final PSUs plus all other initial PSUs that have a nonempty intersection with some final PSU in final stratum S. $B_1,...,B_n$ are the initial PSUs in one-to-one correspondence with $A_1,...,A_n$.

$B_1,...,B_m$ are in initial strata 1,...,r.

Let T_{kl}^* be {1,...,n} except for {f(1),...,f(k-1)} and $\{g_k(1),...,g_k(\ell-1)\}$.

Let $H_t = \{j: B_j \cap A_t \neq \varnothing, j=1,...,m\}$, t=1,...,n.

Let $T_{kl}^{**} = T_{kl}^* \cup \{n+1,...,m\}$.

Let $H_{t\delta} = H_t \cap F_\delta \cap \{T_{kl}^{**} \sim \{f(k), g_k(\ell)\}\}$, $\delta=1,...,r$, t=1,...,n.

The f and g arrays define an ordering of pairs of final PSUs.
A pair $\{f(k), g_k(\ell)\}$ comes before another pair $\{f(k'), g_{k'}(\ell')\}$ if and only if either $k<k'$ or $k=k'$ and $\ell<\ell'$.

F_δ, $\delta=1,...,r$, consists of the PSUs in initial stratum δ which intersect with PSUs in final stratum S.

Now consider $b_{it} = P((\cup H_t) \cap (\cup I) \neq \varnothing \mid I_i)$, t=1,...,n, $i=1,...,\binom{n}{2}+n+1$

where I is the set of initial PSUs in the initial sample and I_i is the ith associated set of initial PSUs.

Let

$p_{ij} = P(i,j \in I)$, i,j=1,...,m,

$p'_\alpha(T) = P(I \cap F_\alpha \subset T)$, $T \subset \{1,...,m\}$, $\alpha=1,...,r$,

$p''_{i\alpha}(T) = P(i \in I \text{ and } I \cap F_\alpha \subset T)$, $T \subset \{1,...,m\}$, $\alpha=1,...,r$, $i \in F_\alpha \cap T$.

Then if $I_i = \varnothing$,

$$b_{it} = 0 \text{ if } H_t \cap T' = \varnothing,\ T' = \{n+1,\ldots,m\}$$

$$= 1 - \prod_{\delta=1}^{r} (1-b_{it\delta}) \text{ otherwise,}$$

where

$$b_{it\delta} = \sum_{j \in H'_{it\delta}} [p''_{j\delta}(T')/p'_{\delta}(T')] - \sum_{\substack{h,j \in H'_{it\delta} \\ h<j}} [p_{hj}/p'_{\delta}(T')],$$

with $H'_{it\delta} = H_t \cap \{j: j \in F_\delta \cap T'\}$

If $I_i = \{v\}$, $v \in \{1,\ldots,n\}$, v in initial stratum α, then

$$b_{it} = 0 \text{ if } H_t \cap T'' = \varnothing,\ T'' = T' \cup \{v\},$$

$$= 1 \text{ if } v \in H_t,$$

$$= 1 - \prod_{\delta=1}^{r} (1-b_{it\delta}) \text{ otherwise,}$$

where

$$b_{it\delta} = \sum_{j \in H''_{it\delta}} [p_{vj}/p''_{v\alpha}(T'')] \text{ if } \alpha=\delta,$$

$$= \sum_{j \in H''_{it\delta}} [p''_{j\delta}(T'')/p'_{\delta}(T'')] - \sum_{\substack{h,j \in H''_{it\delta} \\ h<j}} [p_{hj}/p'_{\delta}(T'')] \text{ if } \alpha \neq \delta$$

with $H''_{it\delta} = H_t \cap \{j: j \in F_\delta \cap T'\}$.

Now consider $I_i = \{f(k), g_k(\ell)\}$ with f(k) in initial stratum α, $g_k(\ell)$ in initial stratum β.

Then

$b_{it} = 1$ if $f(k) \in H_t$ or $g_k(\ell) \in H_t$,

$= 0$ if $H_t \cap T_{k\ell}^{**} = \varnothing$,

$= 1 - \prod_{\delta=1}^{r} (1-b_{it\delta})$ otherwise,

where

$b_{it\delta} = 0$ if $\alpha=\beta=\delta$,

$= \sum_{j \in H_{t\delta}} [p_{f(k),j} / p''_{f(k),\alpha}(T_{k\ell}^{**})]$ if $\alpha=\delta\neq\beta$,

$= \sum_{j \in H_{t\delta}} [p_{g_k(\ell),j} / p''_{g_k(\ell),\beta}(T_{k\ell}^{**})]$ if $\beta=\delta\neq\alpha$

$= \sum_{j \in H_{t\delta}} [p''_{j\delta}(T_{k\ell}^{**}) / p'_{\delta}(T_{k\ell}^{**})] - \sum_{\substack{h,j \in H_{t\delta} \\ h<j}} [p_{hj} / p'_{\delta}(T_{k\ell}^{**})]$ if $\alpha\neq\delta, \beta\neq\delta$

Ratios with 0 in the denominator can occur, due to dummy PSUs, and are defined to be 0.

Finally, let c_{ij} be the entry in the cost matrix corresponding to the initial associated set I_i and the final outcome S_j, $S_j=\{s,t\}$, $s\neq t$, $s,t=1,...,n$. Then $c_{ij}=b_{is}+b_{it}$. The transportation problem to solve for the modified procedure is the same as for the original procedure except for the cost matrix.

If the alternative approach not requiring dummy PSUs, described in the next-to-last paragraph of Section 2, is used then this section would be modified as follows:

Replace the fourth paragraph of this section by: Let $A_1,...,A_n$ be the final PSU in the given final stratum, of which $A_1,...,A_{n'}$ are matched to initial PSUs.

Replace n by n′ in the fifth paragraph, in the definitions of $T_{k\ell}^{*}$, $T_{k\ell}^{**}$, and T', and in the range of v.

Finally, for the alternative approach, the associated set I_i with a set of initial PSUs is defined slightly differently with n' replacing n. Thus, if I includes at least two integers in $\{1,...,n'\}$ then $I_i = \{f(k), g_k(\ell)\}$, where $\{f(k), g_k(\ell)\}$ is the first pair in the ordering for which $\{f(k), g_k(\ell)\} \subset I$. If I includes exactly one integer, v, in $\{1,...,n'\}$, then $I_i = \{v\}$. If $I \cap \{1,...,n'\} = \varnothing$, then $I_i = \varnothing$.

REFERENCES

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

_____ (1989), "Further Applications of Linear Programming to Sampling Problems," SRD Research Report Series, No. RR-89/05, Bureau of the Census, Statistical Research Division, Washington, D.C. 20233.