재미 한인 통계 학자
특 강 교 육 자 료

# Managing Non-Sampling Errors and Small Area Estimation

2001. 8


통 계 기 획 국
조 사 관 리 과

# 목 차

# Managing Non-Sampling Errors

Statistical Research Division
Bureau of the Census

김 종 익 박사

**Managing Non-Sampling Errors**
**- from the Perspective of Sample Redesigns**

Jay Jong-Ik Kim
Statistical Research Division
Bureau of the Census

## I. Introduction

The U.S. Bureau of the Census redesigns demographic samples every ten years after its decennial census to reflect the changes in demography and geography. During the redesign phase, efforts are made to reduce the potential non-sampling errors. For example, if the same household is selected for interviews for more than one survey, the residents of the household may refuse to cooperate with the interviewers for the subsequent surveys. This will result in type A non-response and non-sampling error for the subsequent surveys. When a large number of experienced field representatives (field reps) are replaced with inexperienced new field reps after the redesign, there is no guarantee that the quality of interviews the new field reps provide is as good as that of the outgoing field reps. In this workshop, I will deal with some measures we take which help reduce non-sampling errors, but deal with one interesting procedure, the maximum Primary Sampling Unit (PSU) overlap between 1990 and 2000 surveys extensively.

## II Non-Sampling Errors

Sample estimates are almost never identical to the population values because of sampling and non-sampling errors. Sample is a subset of the target population. The error caused by enumerating a subset rather than whole population is called sampling error. All errors other than the sampling error, which make the sample estimates different from the parameter is called non-sampling errors. The types of non-sampling errors can be summarized as follows.

**Specification Error** - concepts, objectives, data elements;

**Frame Error** - omission, erroneous inclusion, duplication;

**Nonresponse Error** - unit, within unit, item, incomplete data;

**Processing Error** - editing, data entry, coding, weights, tabulation;

**Measurement Error** - interviewer, respondent, survey instrument, mode of data collection, method of data capture and setting.

*The survey instrument* - survey questionnaire and the instructions to the respondents for supplying requested information.

*The respondent* in the business survey may rely on the business information system the company maintains.

*The mode of data collection* can be telephone, face to face, self-administration, Internet, etc.

*The method of data capture* includes paper and pencil, computer keyboard, telephone keypad, etc.

*The interviewer* may be a prerecorded voice over the telephone or may not exist for self-administered surveys or Internet surveys.

III Unduplicated Sample Selection

If a household was visited by a field representative for an interview which lasted four hours and is visited again for another interview within a week, there could be many households which do not want to cooperate with the field representative, resulting in (unit) non-response. Because of this, the Bureau of the Census try to avoid duplicated selection. Thus, the Bureau coordinates its sample selection operations among many surveys under its perview and "unduplicates" the housing units if they are in more than one survey. Note even if PSU's are not selected by the Bureau (CE PSUs are picked by the Bureau of Labor Statistics and National Health Interview Survey PSUs are selected by the National Center for Health Statistics), the unduplication operation is performed during the Within-PSU sample selection. In sampling process, if a county is selected in three surveys, such as CPS, NCVS and CE, then CPS selects its sample first. Selected units are removed from the sampling frame and NCVS selects its sample units from the remaining units and so on. The sample selection is computerized in the unit frame and in the frame software is designed such that housing units which are already in a survey are ignored by the next surveys in their sample selection. In the area frame, the person who selects the sample makes sure that no unit is selected in more than one survey.

However, there are exceptions. The National Health Interview Survey (NHIS) uses all area frame design disregarding whether the selected area is "good address" area (i.e., its house number and street name are available) or "bad address" area (i.e., its house number or both house number and street name are missing. Rural route and P.O. Box addresses are such cases). NHIS is used as a screening devise for their subsequent surveys for the National Center for Health Statistics (NCHS). NHIS gathers information on the respondents' health status. When NCHS wants to survey cancer patients, then they pick cancer patients from among the respondents of NHIS. In order for NCHS to interview the selected, NCHS has to know their addresses. However, the Census Bureau cannot release their addresses because they are protected by Title 13 which authorizes the Bureau to conduct the decennial census. Note the Bureau uses the list of addresses obtained from the Census for selecting their sample units from the address frame. In order for the NCHS to have addresses of the residents of sample housing units, the Census Bureau has to list the selected areas afresh. Because of this, NHIS uses all area frame design. If a "good address" area is selected by NHIS and other surveys, the other surveys had to use area frame design in that area in order to identify duplicated units and unduplicate them among the surveys. In the 1990 design, we used such design. However, in the 2000 redesign, the "other" surveys will let the computer pick sample

units from address from in the good address areas and NHIS will use area frame to pick their sample. Because of this difference, we will allow duplicated selection between NHIS and other surveys.

Also the American Community Survey (ACS) will be allowed to pick sample units disregarding whether the units are in other survey or not. The reason is, ACS is the surrogate for the Census sample and Census is always allowed to visit a household irrespective of whether it is in other survey or not. In addition, there can other types of duplication. For example, the previous decade's sample units of the Bureau's demographic surveys can be duplicated with the current decade's redesigned sample units of the Bureau's demographic surveys. Business and economic surveys for self-employed and the surveys, which other organizations conduct, can also be duplicated with the Bureau's redesigned sample units.

When a PSU is selected by more than one survey, the common area becomes BPC (Basic PSU Component) as mentioned before. Within this area, unduplication operations are performed.

## IV **Maximizing PSU Overlap**

The purpose of maximizing PSU overlap between 1990 and 2000 redesigns is to reduce measurement errors by keeping as many experienced field representatives as possible.

### IV .1 Overview

#### IV.1.1 Demographic Survey Redesign

In 2002, the Demographic Directorate of the Bureau of the Census, coordinating with survey sponsors, will redesign most of the major Demographic surveys. The redesign improves the data collected in those surveys by using the new information collected in the 2000 Census to stratify and sample so as to improve estimates and reduce variances. In the first stage of sample selection, the United States is divided into just over 2,000 geographic areas. Each of these areas will consist of a county or a group of contiguous counties.[1] For each of those surveys, a sample of these areas is selected across the country. These areas are referred to as Primary Sampling Units (PSUs). Sampling for the

[1] The National Health Interview Survey (NHIS) redesign project group is currently examining whether they want to define PSUs at the Census tract level. If they do so, the procedures herein for NHIS will be Census tract based vs. the county based process indicated.

surveys then continues by taking a sample of households from within each of the selected PSUs. We will be selecting PSUs for the following surveys:

The Current Population Survey (CPS)
The Survey of Income and Program Participation
The National Crime Victimization Survey (NCVS)

The following surveys will be included in the final Post PSU Selection files; however, either the PSUs have been selected in the past or the PSUs will be selected outside of this process:

The Consumer Expenditures Surveys (CE)
The American Housing Surveys (AHS)
The National Health Interview Survey (NHIS)

1. *PSU Stratification*

Before we select PSUs for each survey, they are grouped into strata. The strata are first geographically restricted by state or region; CPS and Survey of Income and Program Participation by state and NCVS by region (some groupings of states). Within the state or region, strata are defined separately for each survey being redesigned. Since population is always a variable of interest, certain highly populous PSUs in each state or region are automatically selected for the sample. Each of these PSUs is in a stratum by itself. They are referred to as Self-Representing (SR) strata. The remaining PSUs in the state or region are grouped into strata called Non-Self-Representing (NSR) strata. These strata are defined by population and by other target variables of interest to that survey. PSUs are then selected from each stratum. The surveys select either one or two PSUs per NSR stratum. Completion of stratification by all of the surveys marks the start of this PSU selection process.

2. *Probability Proportional to Size (PPS)*

In the NSR strata, the selection of one or two PSUs, dependent upon the survey, is made using probabilities proportional to its "measure of size" (MOS). The MOS for each survey is some measure of population or housing units. This approach further tends to reduce the variance of survey estimates (Cochran 1977, p. 299).

3. *Maximizing Overlap between 1990 and 2000 Designs*

The new selection of PSUs will inevitably mean that some of the PSUs that were in sample throughout the 1990s will not be in the new sample for the 2000s. Similarly, some PSUs that were not in sample in the 1990s will be in the new sample. These changes usually

force the regional offices to release an experienced interviewer who lives in PSUs that are "dropped" and hire new interviewers in PSUs that are "picked up." After the selection of the PSUs for 2000, the PSUs that were in both the 1990 and 2000 designs are collectively known as the "PSU Overlap."

Since the 1970 redesign, we have been using techniques to increase PSU overlap, while maintaining our PPS design. The conditional probabilities of a PSU's new sample selection given each possible prior sample are altered to increase overlap, while their joint "accumulation", the unconditional probability, is kept fixed.

4. *PSU Selection*

Using the "conditional probabilities" provided by the maximum overlap methodology, we then select the PSUs for the new design PPS in each stratum and forward the Post PSU Selection Files for within-PSU sample selection.

IV.2. Required Information

The following files of information will be input to the PSU selection process:

5. *1990 Design Information*

County/minor civil division (MCD)/Census county division (CCD) level file(s) containing 1990 PSU sampling design and selection information. For all counties/MCDs/CCDs in the United States, it will indicate

1. the PSU in which it is located.
2. the PSU's 1990 stratum for each survey.
3. the PSU's unconditional probability of selection (POS) for each survey.
4. whether the PSU was selected for each survey.

6. *2000 PSU Definitions Information*

County level file(s) containing the 2000 PSU definitions and measures of size for each survey. For each county, it will indicate

1. the county's MOS for each survey
2. the PSU in which it is located for each survey

7. *2000 PSU Stratification Information*

PSU level file(s) indicating each PSUs' 2000 stratum and an MOS for each survey.

8. *Sampling Interval Information*

The respective design branches will provide the strata sampling intervals (SIs) for CPS, Survey of Income and Program Participation, NHIS, and NCVS for the strata developed in the stratification process. These SIs can take the form of a file identifying the SI for each strata, a file of SIs by state or a single national SI.

9. *Pass-along information*

The following information is input to the PSU selection for the sole purpose of passing it along to the Within-PSU stage of the sampling:

1. Metropolitan Area Definitions File
2. Consumer Expenditures Surveys (CE) PSU Selection Information
3. American Housing Surveys (AHS) PSU Selection Information
4. Survey of Income and Program Participation PSU Selection Information

IV.3 PSU Selection Methodology

10. *Unconditional Probability of Selection*

The unconditional probability of selection for each PSU within each survey will be computed within each stratum using the number of PSUs being selected in that stratum times the PSU's MOS divided by the total MOS for the stratum, or

$$P(PSU_i \in S) = \text{Min}\left( (\#PSUs / \text{stratum}) * \frac{MOS_i}{\sum_{k=1}^{n} MOS_k}, 1 \right) .$$

For surveys and strata that will select two PSUs per stratum, the maximum overlap algorithm requires that we calculate an unconditional probability of selection for each pair of PSUs in the stratum. It will be computed using a formula attributable to Brewer (1963) and Durbin (1967):

Let $p_i = \frac{MOS_i}{\sum_{k=1}^{n} MOS_k}$. If $\max\{p_i\} < \frac{1}{2}$,

$$P(PSU_i \in S \text{ and } PSU_j \in S) = \frac{2p_ip_j}{\left(1+\sum_{k=1}^{n}\frac{p_k}{1-2p_k}\right)}\left(\frac{1}{1-2p_i}+\frac{1}{1-2p_j}\right),$$

where $i = 1, 2, \ldots, n$ ; $j = 1, 2, \ldots, n$; $i \neq j$.

If $p_i \geq \frac{1}{2}$ for some i, that PSU will be selected with probability = 1, and the remainder of the stratum will be treated as a one-PSU/stratum selection. A more detailed discussion of on the derivation and properties of these formulas is in attachment A.

11. *Rotating PSUs*

Some PSUs have too few housing units (HUs) to provide sample for even one demographic survey throughout the decade. These PSUs are also geographically large enough that it would be impractical to combine them with neighboring counties to increase the number of HUs in the PSU. We call these PSUs "small PSUs." Within a stratum, these small PSUs will be combined to create a "rotating PSU" cluster. If the small PSUs in the stratum collectively do not have enough HUs to make up the decade's sample, the smallest "large PSU" (by HUs) will be added to the cluster. The PSUs in the cluster will share the sample for the decade by rotating from one of the PSUs to the next.

1. Small PSUs will be defined separately for each survey using a minimum number of HUs required for sample. These "cutoffs" will be calculated by multiplying the number of samples to be taken during the decade times the number of HUs per NSR PSU per sample, dividing by a "buffering" factor like 0.8 or 0.7 and increased to the next 100. Cutoffs will be calculated by the design branches.

2. The probability of selection for the rotating PSU cluster will be the collective size of the cluster using the Measure of Size (MOS) chosen for each survey.

3. The overlap factor in the linear programming model for any new sample rotating PSU will be zero, regardless of whether an overlap exists or not.

4. If a rotating PSU is selected for sample, the Random Arc Method (RAM) will be used to determine the order of rotation and the starting point. The details of this method are in attachment B.

12. *PSU Overlap*

We will use the procedure based on Ernst (1986) to calculate the conditional probabilities that will be used for PSU selection for CPS, NCVS, and Survey of Income and Program Participation. NHIS will not use an overlap procedure, so the unconditional probability of selection will be used for selection. The overlap method is the same procedure used for CPS and NCVS in 1990. Now that Survey of Income and Program Participation is a state-based design, the computer memory restriction that resulted in special procedures for Survey of Income and Program Participation overlap in 1990 will not be necessary in 2000. A complete discussion of the development of the Ernst (1986) algorithm is given in attachment C.

Each survey-NSR stratum combination will run through overlap calculations separately, as listed below. To assist in the discussion, here is some notation:

| | |
|---|---|
| S | will represent the stratum being processed. |
| $k = 1, 2, \ldots, n$ | will represent the n possible selections in S. For 1-PSU/stratum cases, that will just be the n PSUs. For 2-PSU/stratum cases, it will be the n pairs of PSUs. |
| $i = 1, 2, \ldots, r$ | will represent the r strata in the 1990 design that contain at least one PSU that is partially or completely contained in one of the PSUs within S. |
| $T_i$ | will represent the $i^{th}$ stratum of those 1990 strata. |
| $j = 1, 2, \ldots, u_i$ | will represent the $u_i$ possible 1990 selections in $T_i$ in terms of the PSUs in S only. For 1990 1-PSU/stratum cases, $u_i$ will just be the PSUs in $T_i \bullet S$ plus 1 (for other PSUs in $T_i$). For 2-PSU/stratum cases, $u_i$ will be the PSUs in $T_i \bullet S$ plus the pairs of the PSUs in $T_i \bullet S$, plus 1 (for other pairs of PSUs in $T_i$). |

1. Initial Input Data Preparation - For each survey-NSR stratum, S, the following input information will be calculated

    1. n is determined as the number of PSUs in S for 1-PSU/stratum cases. For 2-PSU/stratum cases, n is the number of pairs of PSUs.
    2. $\bullet_k$, $k = 1 \ldots n$ is the unconditional probability of selection (UPOS) for the n possible selections. For 1-PSU/stratum cases it is just the UPOS of each of the PSUs in S. For 2-PSUs/stratum cases it is the joint UPOS for each of the pairs.
    3. r and the $T_i$ are determined by comparing 1990 strata PSUs with the PSUs in S.
    4. $u_i$, $i = 1, 2, \ldots r$ are determined as follows:
       For 1990 1-PSU/stratum cases, $u_i$ will just be the number of PSUs in $T_i \bullet S$ plus 1 (for other PSUs in $T_i$).

For 1990 2-PSU/stratum cases, $u_i$ will be the number of PSUs in $T_i$ • S plus the number of pairs of the PSUs in $T_i$ • S, plus 1 (for other pairs of PSUs in $T_i$).

5. $p_{ij}$, are the 1990 UPOS of the $j^{th}$ possible selection in the i th 1990 stratum, $T_i$.

2. Measure of Overlap - For any combination of the $j^{th}$ possible 1990 selection in the $i^{th}$ 1990 stratum $T_i$ and the $k^{th}$ possible 2000 selection in S, where $i = 1 \ldots r$, $j = 1 \ldots u_i$, and $k = 1 \ldots n$, a measure of overlap, $c_{ijk}$ will be calculated as is given below. Because, in general, any stratum could have one or two PSUs in each possible selection, we will introduce some additional notation:
   1. Compare counties in each of the PSUs (there may be two ) in the $k^{th}$ possible 2000 selection in S to each of the PSUs (there may be two) in the $j^{th}$ possible 1990 selection in the $i^{th}$ 1990 stratum $T_i$. Let $f_{ijtkh}$ = the total Housing Units (HUs) common to the $t^{th}$ PSU in the 1990 set and the $h^{th}$ PSU in the 2000 set, divided by the total HUs in all of the $j^{th}$ possible 1990 selection in the $i^{th}$ 1990 stratum $T_i$. HU counts for both 1990 and 2000 PSUs used in this calculation will be from the 2000 Census.
   2. $$c_{ijk} = \sum_{h=1}^{m_s}\left[1-\left[\prod_{t=1}^{v_i}\left(1-f_{ijtkh}\right)\right]\prod_{\substack{q=1\\q\neq i}}^{r}\left(1-\sum_{w=1}^{u_q}p_{qw}\left[1-\prod_{t=1}^{v_i}\left(1-f_{qwtkh}\right)\right]\right)\right]$$

      where $m_s$ is the number of PSUs being selected in S and $v_i$ is the number of PSUs that were selected in the $i^{th}$ 1990 stratum.
   3. However, if the $k^{th}$ possible 2000 selection in S is a rotating PSU, $c_{ijk} = 0$.
   4. It should be noted that in a simpler case like the CPS, where both the 1990 and 2000 selections were both done 1-PSU/NSR Stratum, the notation would simplify to
      $$c_{ijk} = 1-\left(1-f_{ijk}\right)\prod_{\substack{q=1\\q\neq i}}^{r}\left(1-\sum_{w=1}^{u_q}p_{qw}f_{qwk}\right).$$
      However, the above formula may allow a more general programming to handle any situation.

3. Linear Programming Solution - Given the input available above, the following linear programming (LP) model will be solved by LP software producing the joint probability values $x_{ijk}$ and the $y_i$ "pseudo-probabilities":
   1. Maximize $Z = \sum_{i=1}^{r}\sum_{j=1}^{u_i}\sum_{k=1}^{n}c_{ijk}x_{ijk}$

2. Subject to constraints $\sum_{k=1}^{n} x_{ijk} = p_{ij} y_i$ and $\sum_{i=1}^{r} \sum_{j=1}^{u_i} x_{ijk} = \pi_k$.

4. Conditional Probabilities - Using the output from the LP, $y_i$ and $x_{ijk}$, $i = 1 \ldots r$, $j = 1 \ldots u_i$, and $k = 1 \ldots n$, compute the conditional probabilities,
P(the $k^{th}$ possible 2000 selection in S is selected given the 1990 sample that was selected) = $P\left(N = N_k \middle| I_1 = I_{1j_1}, \ldots, I_r = I_{rj_r}\right) = \sum_{i=1}^{r} \frac{x_{ij_i k}}{p_{ij_i}}$, where the $I_i = I_{ij_i}$ represent the actual selections in the r 1990 strata.

13. *PSU Selection*

PSU selection will be performed for each survey's NSR strata using a PPS procedure, using the conditional probabilities discussed above for CPS, Survey of Income and Program Participation, and NCVS, and using unconditional probabilities of selection for NHIS.

## 2. Sampling Intervals

1. *Definition*

The sampling interval (SI) at any level is the inverse of the sampling fraction. The term "sampling interval" comes from systematic sampling, where if you were selecting every $25^{th}$ unit for the sample, the sampling interval would be 25.

2. *National, State, and Stratum SI*

The survey design branch for CPS, Survey of Income and Program Participation, NHIS, and NCVS will provide the sampling intervals (SIs) for the strata developed in the stratification process. These SIs can take the form of a file identifying the SI for each strata, a file of SIs by state, or a single national SI.

3. *Calculating Within PSU SIs*

The Within-PSU SI for each PSU in sample will be the stratum SI times the unconditional probability of selection of the PSU.

5. each survey-PSU's stratum identifier
6. each survey-PSU's measure of size (MOS)
7. each survey-PSU's unconditional probability of selection (POS).
8. each survey-PSU's sampling interval information.
9. a rotating PSU indicator for each survey-PSU
10. an SR/NSR flag
11. miscellaneous information and flags needed by the surveys, particularly the CE and AHS surveys.

A second tract-level PPSF will be created for NHIS. It will contain the same information related to NHIS only. However, the BPC codes places in this PPSF will reflect the county-level BPCs developed for the other surveys.

2. *Design Branch PSU Files*

Files will be created for each of the surveys separating that surveys information from the PPSF. These files will be given to the survey's design branch. A copy of the CPS PSU file will also be given to the Manufacturing and Construction Division for their PSU selection process.

## 5. Sample Verification

In order to minimize the sample verification time after PSU Selection and before the sample is forwarded to within PSU sampling, we will add functions to our software to compile statistics and summaries to aid the design branches in their review. The details of this part of the process will be defined prior to completion of the specification.

## 6. Survey Needs Satisfaction

1. *Current Population Survey (CPS)*

    1. CPS will stratify NSR PSUs by state.
    2. CPS will select one PSU per NSR stratum.
    3. CPS will use an MOS of civilian noninstitutionalized population aged 16 or older from the 2000 Census.
    4. CPS will overlap on its 1990 sample PSUs. However, it may decide not to overlap in some states.

2. *Survey of Income and Program Participation*

    1. Survey of Income and Program Participation will stratify NSR PSUs by state.

## 3. Basic PSU Components (BPCs)

### 1. *Definition*

One final component we will define for output is the BPC code. In the output we will delineate the PSUs that were selected for CPS, Survey of Income and Program Participation, NCVS, CE surveys, and AHS surveys. PSUs will be further broken down by counties since the definitions of the PSUs are different across the surveys. Due to these differences, a single county within a multi-county CPS PSU may be in a CE PSU while the others are not, as illustrated below. So, a BPC is defined as each PSU, sub-PSU county group, or single county within a state that has a different combination of selected survey-PSUs from the other BPCs in the state. In the simple example below, three BPCs would be defined: The single common county would be one BPC, the county group with CPS alone would be a second, and the county or county group with CE alone would be a third.

### 2. *Methodology*

BPCs will be defined and handled the same way as in the 1990 redesign, except for the special NHIS handling described below. BPCs will be defined in relation to three of the surveys which selected PSUs within this process, CPS, Survey of Income and Program Participation, and NCVS, and the surveys that were selected elsewhere and provide input to this process, the CE surveys and the AHS survey. NHIS will not be considered in the development of BPCs. A five-digit BPC code will be assigned to each county that was selected for at least one of the surveys. The codes will be made up a two-digit state code followed by a three-digit sequence number. Each sequence number will define the largest group of counties in which the same combination of Survey PSUs has been selected for sample. A separate sequence will begin from 001 in each state.

## 4. Output Information

### 1. *Post PSU Selection File (PPSF)*

County level file(s) containing the 2000 PSU sampling design and selection information for CPS, Survey of Income and Program Participation, NCVS, the CE surveys, and the AHS surveys. For each county, it will indicate

1. the county definition by region, state, and county codes and names
2. the BPC code
3. each survey that has selected in that county
4. the PSU identifier for each survey that selected in that county

2. Survey of Income and Program Participation will select two PSUs per NSR stratum in some states and one PSU per NSR stratum in others.
3. Survey of Income and Program Participation will use an MOS of housing units from the 2000 Census, which should be limited by excluding institutional group quarters.
4. Survey of Income and Program Participation will overlap on its 1990 sample PSUs.

3. *National Crime Victimization Survey (NCVS)*

1. NCVS will stratify NSR PSUs by region (some groupings of states). However, it may stratify within states that do not have SR PSUs.
2. NCVS will select one PSU per NSR stratum.
3. NCVS will use a MOS of housing units plus housing unit equivalents from military, "street people", and noninstitutional Group Quarters, all from the 2000 Census.
4. NCVS will overlap on its 1990 sample PSUs.

4. *Survey of Construction (SOC)*

1. SOC will require a file identifying the CPS PSUs selected for sample. At present, this file will either be the full PPSF or the CPS design branch file.
2. SOC may want the use of some of our code to perform similar processes in their overlap efforts.

5. *Consumer Expenditure Surveys (CE)*

The final PPSF will have CE selection information from the selections done by the Bureau of Labor Statistics to pass on to within-PSU selection. No other processes are required for CE surveys.

6. *American Housing Survey (AHS)*

The final PPSF will have AHS- National and AHS-Metropolitan Survey PSU selection information provided by the Longitudinal Surveys Branch of DSMD to pass on to within-PSU selection. No other processes are required for AHS surveys.

## 7. References

Alexander, Charles H., Ernst, Lawrence R., and Hass, Michael E., "A System for Replacing Primary Sampling Units When the Units Have Been Exhausted," Proceedings of the Survey Research Methods Section, American Statistical Association, 1982, pp. 211 - 216.

Biemer, P. P. and Fecso, R.S. (1995), "Evaluating and Controlling Measurement Error in Business Surveys," *Business Survey Methods*, pp. 257 - 281.

Brewer, K. R. W., (1963) "A model of systematic sampling with unequal probabilities," *Australian Journal of Statistics*, 5, pp. 5 - 13.

Causey, Beverly D., Cox, Lawrence H., and Ernst, Lawrence R., (1985) "Applications of Transportation Theory to Statistical Problems," *Journal of the American Statistical Association*, Vol. 80, No. 392, pp. 903 - 909.

Cochran, William G., *Sampling Techniques*, Wiley, New York, 1977.

Durbin, J., (1967) "Design of Multi-Stage Surveys for the Estimation of Sampling Error," *Applied Statistics*, Vol. XVI, No. 2, pp. 152 - 164.

Ernst, Lawrence R., (1986) "Maximizing the Overlap between Surveys when Information is Incomplete," *European Journal of Operational Research*, Vol. 27, No. 2, pp. 192-200.

Keyfitz, Nathan, (1954) "Sampling with Probabilities," *Journal of the American Statistical Association*, Vol. 46, pp. 105-109.

Kim, J (2000) "Equivalency of Two Cost Formulae," Internal Census Bureau Memorandum, August 25, 2000, pp 1-6.

Kim, J (2000) "Costs in 2 PSU/Stratum Design," Internal Census Bureau Memorandum, March 24, 2000, pp 1-3.

Kim, J (2000) "Calculation of Joint Probability of Selection for Ernst's 86 Algorithm," Internal Census Bureau Memorandum, September 15, 1999, pp 1-3.

Kim, J (2001) "Probability of Duplicated Selection and Its Effect on Nonresponse, Bias and Variance." To be presented at the 53rd Session of the International Statistical Institute, Seoul, Korea

Yates, F. and Grundy, P. M., (1953) "Selection Without Replacement from Within Strata with Probability Proportional to Size," Journal of the Royal Statistical Society, Series B, Vol. 15, pp. 253-261.

Attachments (3)

# 2000 Redesign - Calculating the Joint Probability of Selection for Strata Where We Select Two Primary Sampling Units

## 1. Background

For the surveys in the 2000 Redesign, we divide the United States into geographical areas which we call primary sampling units (PSUs). Each is made up of one or a group of counties. Highly populous PSUs are automatically selected for sample. They are called Self-Representing (SR) PSUs. The remaining Non-Self Representing (NSR) PSUs are put into strata. Within each of those strata, either one PSU is selected or two PSUs are selected. To reduce variance, we perform the selection using probability proportional to size (PPS) with a measure of size (MOS) based on either the population or the number of housing units within each PSU relative to the other PSUs in the stratum. If we are selecting only one PSU within an NSR stratum, the probability of selection for each is just the PSU's MOS divided by the total MOS for the stratum, or $p_i = \frac{MOS_i}{\sum_{MOS_k \in S} MOS_k}$. However, if we are selecting two PSUs in the stratum, the formula for the joint probabilities of selection for each pair is not that straightforward.

## 2. Requirements

We will be selecting two-PSU/stratum samples by assigning joint probabilities, $\pi_{ij}$, to each pair of PSUs and then selecting one of the pairs. We will not be selecting them sequentially, without replacement. So, what properties do we want the joint probability formula to have?

### a. *Probability Axioms*

Obviously, the formula must comply with the axioms of probability, $0 \le \pi_{ij} \le 1$, $\sum_{\text{all } (i,j)} \pi_{ij} = 1$.

b. *Maintain PSU PPS*

To maintain the PPS design, the probability of selection for each PSU must still be proportional to its MOS relative to the other PSUs in the stratum, or

$$\sum_{\text{all } j} \pi_{ij} = \text{constant} * \frac{MOS_i}{\sum_{MOS_k \in S} MOS_k}.$$

3. **Formula #1 - Yates-Grundy Formula**

The first, most obvious, approach is to derive a formula based on selecting the first PSU using $p_i = \frac{MOS_i}{\sum_{MOS_k \in S} MOS_k}$, and selecting the second PSU using $\frac{MOS_j}{\sum_{k \neq i} MOS_k} = \frac{p_j}{1-p_i}$. This would produce an unordered probability of selection of $\pi_{ij} = p_i p_j \left( \frac{1}{1-p_i} + \frac{1}{1-p_j} \right)$.

a. *Probability Axioms*

As long as there are two PSUs with non-zero MOS's in the stratum, $0 \leq \pi_{ij} \leq 1$ and

$$\sum_{\text{all }(i,j)} \pi_{ij} = \frac{1}{2} \sum_i \sum_{j \neq i} \left[ p_i p_j \left( \frac{1}{1-p_i} + \frac{1}{1-p_j} \right) \right] = \sum_i \left[ p_i + p_i \sum_{j \neq i} \frac{p_j}{1-p_j} \right]$$

$$= \frac{1}{2} \left[ 1 + \sum_{\text{all } k} \frac{p_k}{1-p_k} - \sum_i \frac{p_i^2}{1-p_i} \right] = 1.$$

b. *Maintain PSU PPS*

However, $P(PSU_i \text{ is selected}) = \sum_{j \neq i} \pi_{ij} = p_i + \sum_{j \neq i} \frac{p_j}{1-p_j} \neq \text{constant} * p_i$.

4. **Formula #2 - Durbin-Brewer Formula**

A joint probability formula developed by Brewer (1963) and Durbin (1967) for the joint probability of selection is $\pi_{ij} = \dfrac{2p_i p_j}{1+\sum\limits_{\text{all k}} \dfrac{p_k}{1-2p_k}}\left[\dfrac{1}{1-2p_i}+\dfrac{1}{1-2p_j}\right]$. It is restricted to values of $p < 0.5$. If any one of the values of p is • 0.5, that PSU is converted to SR and only one of the remaining PSUs is selected thereafter.

a. *Probability Axioms*

Under the assumption that each $0 < p_i < 0.5$, • i, $0 \le \pi_{ij} \le 1$ and

$$\sum_{\text{all (i,j)}} \pi_{ij} = \frac{1}{2}\sum_i \sum_{j\neq i} \frac{2p_i p_j}{1+\sum\limits_{\text{all k}} \dfrac{p_k}{1-2p_k}}\left[\frac{1}{1-2p_i}+\frac{1}{1-2p_j}\right].$$

$$= \frac{1}{1+\sum\limits_{\text{all k}} \dfrac{p_k}{1-2p_k}} \sum_i p_i\left[\frac{1-p_i}{1-2p_i}+\sum_{\text{all k}}\frac{p_k}{1-2p_k}-\frac{p_i}{1-2p_i}\right] = 1.$$

b. *Maintain PSU PPS*

P($PSU_i$ is selected) =

$$\sum_{j\neq i} \pi_{ij} = \frac{2p_i}{1+\sum\limits_{\text{all k}} \dfrac{p_k}{1-2p_k}}\left[\frac{1-p_i}{1-2p_i}+\sum_{\text{all k}}\frac{p_k}{1-2p_k}-\frac{p_i}{1-2p_i}\right] = 2p_i$$ which maintains PSU PPS.

# Rotating Primary Sampling Units

1. **Background**

For each of the surveys, there are some Primary Sampling Units (PSUs) that do not have enough Housing Units (HUs) to provide samples for the entire decade. Yet, these same PSUs are physically large enough that combining it with additional adjacent counties to enlarge its sample is not reasonable. The method we use to include these PSUs in our sample selection without being potentially forced to sample the same HU twice, is the "Rotating PSU."

2. **"Small PSUs"**

A PSU is regarded as small for a particular survey if 70% or 80% (to be determined) times the number of housing units in the PSU is less than the number of HUs that will be sampled throughout the use of the 2000 design (or about 10 years). This figure varies from survey to survey. These will be recomputed for the 2000 redesign, but for 1990 the cutoffs were:

- CPS 1,600 HUs
- Survey of Income and Program Participation 1,200 HUs
- NCVS 2,100 HUs
- NHIS 1,500 HUs

3. **Rotating PSU Clusters**

For each Non-Self-Representing (NSR) stratum and each survey, all small PSUs will be combined into a rotating PSU cluster. If the total HUs for the cluster are still below the small PSU cutoff, the next smallest PSU in the stratum will be added to the cluster. Thus, across all of the surveys in the cluster there is enough sample for the survey over the decade.

4. **Sample Selection**

In the 1990 redesign, sample selection for the rotating PSU clusters was handled in two ways. In 2000, we will be using only the second method.

a. *One PSU/Stratum.* For the survey and stratum in question, if we were selecting only one PSU, all of the PSUs in the rotating cluster were selected as separate PSUs. If any one of them was selected for sample, the rotating cluster was selected for sample.

b. *Two PSUs/Stratum.* If we were selecting two PSUs for that stratum, the above method could not be used as it would be possible to select two of the PSUs in the cluster. So, the

entire cluster was treated as one PSU with its cumulative measure of size used to calculate its probability of selection.

c. *2000 Redesign Method.* Both methods above are effectively identical. However, only the second method is useable in both situations. In order to minimize the programming effort, we will use the second method.

## 5. Rotation

If a rotating cluster is selected for sample for a survey, the samples through the decade will be rotated from one of the PSUs in the cluster to another to produce all of the samples necessary. The next step in the process will be deciding the order of rotation and the number of samples from each PSU before it rotates out.

## 6. Rotation Schedule

The method used for determining the rotation schedule is called the Random Arc Method (RAM). The method was originally proposed and shown to maintain unbiasness in Alexander, et al. (1982). For the survey and rotating cluster in question, perform the following steps:

a. *Step 1* => Sort the PSUs in the rotation cluster in random order.

b. *Step 2* => For each PSU, $A_i$, let $\bullet(A_i)$ = the number of complete samples that could be taken from $A_i$ for this survey. Let $\bullet(c) = \bullet\ \bullet(A_i)$. Conceptually, create a series of line segments proportional to each of the $\bullet(A_i)$ and connect them end to end in the order determined in step 1.

c. *Step 3* => Bend the line around to form a circle of circumference $\bullet(c)$ with clockwise orientation.

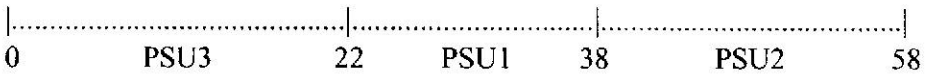
d. *Step 4* => Randomly select a starting point X on the circle.

e. *Step 5* => The arc which includes the point X represents the initial PSU in the decade of samples in the new design. Beginning at point X, number the samples and identify the sample numbers where a rotation from one PSU to the next PSU in rotation occurs.

*Example:*

Suppose a survey requires 50 HUs per sample in stratum S with 25 samples to be taken over the next 10 years. Therefore, each PSU or rotating cluster requires at least 1,250 HUs in

stratum S. Suppose a rotating cluster of three small PSUs is selected for sample in stratum S. PSU1 has 824 HUs, PSU2 has1,012 HUs, and PSU3 has 1,111HUs. Therefore, • (PSU1) = 16, • (PSU2) = 20, • (PSU3) = 22, and • (c) = 58. The random sort in step 1 produces the order: PSU3 first, PSU1 second, and PSU2 third. The line of length 58 is created, made up of segments of length 22, 16, and 20 in that order:




The line is bent into a circle, and the point X between 0 and 58 is randomly selected with a uniform probability distribution. Suppose the point is selected is 46.

Then, PSU2 will be the first PSU in the new redesign. It will get sample 1 - 13 (46 through 58 on the line above). Then it will rotate to PSU3, which will get samples 14 - 25 (1 through 12 on the line above). And, PSU1 will not rotate into the sample.

## 7. More than One Survey in a Rotating PSU

It is possible, though extremely unlikely, that two or more surveys could select a rotating PSU, or one survey could select a rotating PSU and a second survey could select one of the PSUs in the rotating PSU. Following the procedures used in 1990, we will not make allowances for these possibilities. Adjustments will be made as part of the Within-PSU processing.

# Primary Sampling Unit Overlap

## 1. Background

a. *What is Primary Sampling Unit (PSU) Overlap?*

For a given survey, all PSUs (or portions of PSUs) that will be selected in the 2000 redesign which were also in the 1990 design are known collectively as PSU Overlap. Some of this overlap would happen by design (2000 Self-Representing (SR) PSUs overlapping 1990 SRs) and some by random selection (the same Non-SR (NSR) PSU selected by chance in 2000) without altering the design process. Starting in the 1970 redesign, the Current Population Survey (CPS) implemented a design change to increase the expected PSU overlap, while maintaining each PSU's unconditional probability of selection.

b. *Why Maximize Overlap?*

If you look at one of our maps showing the Sample PSUs for the CPS, you can see the 750+ red areas indicating the sample PSUs where we conduct CPS interviews and the 1,250+ white areas where we do not. Those sample PSUs were selected in the early 1990s. In the early 2000s, we will select new PSUs using updated information from the 2000 Census. For every case where we drop an existing CPS PSU and pick up a new sample PSU, we may have to "release" a trained and experienced interviewer from the old PSU and hire a new interviewer in the new PSU. We will then have to train the new interviewer, suffer from the lower response rates, and experience lower data quality that usually seen with a less experienced interviewer. Maximizing overlap in 1990 prevented about 100 of these transitions from happening in CPS alone.

c. *What is the "Downside" of PSU Overlap?*

Our methods of maximizing PSU Overlap maintain the desired unconditional probability of selection (POS) for each PSU. Armed with this information, unbiased weighted estimators can be constructed. The POS for any given PSU was selected based on a measure of size (often population or # of housing units) relative to the other PSUs in the same stratum to take advantage of the benefits of PPS[2] sampling. The PSUs were assigned to strata using a number of stratification variables, grouping PSUs with similar values to take advantage of the

[2] Probability proportional to size

variance reduction that can be realized from a stratified design. So, if maximizing overlap doesn't disturb these design considerations, what does it affect?

An assumption of the stratified design is stratum-to-stratum independence. The PSU Overlap process produces a dependence of NSR strata within the state or region. This dependence will increase variance, though to date there has been no research into the extent of the impact.

d. *1990 Overlap Methodology*

- CPS & the National Crime Survey (NCS)

  In 1990, CPS and NCS used the Ernst(1986) method. Both surveys selected one PSU per stratum.

- Survey of Income and Program Participation

  In 1990, the Survey of Income and Program Participation maximized overlap for the first time, and selected two PSUs per stratum. Survey of Income and Program Participation could not use a methods similar to those used for CPS and NCS, due to the large number of PSUs in some of Survey of Income and Program Participation's regional strata. The algorithms caused a computer memory size problem, even in a mainframe computer. So, a "workaround" algorithm was developed. Even using that algorithm, a few of Survey of Income and Program Participation's largest strata still had to be selected independently due to size restrictions.

e. *2000 Methodology*

Fortunately, the size problems with the Survey of Income and Program Participation are a nonissue in 2000 for several reasons:

- Computers are exponentially faster and have more memory.
- Linear programming algorithms are much more efficient.
- Survey of Income and Program Participation will be stratifying by state so the sizes of the strata are greatly reduced.

So, with the size problem eliminated, we can use Ernst(1986) for all of the surveys that are maximizing overlap: CPS, the Survey of Income and Program Participation and the National Crime Victimization Survey (NCVS).

## 2. Maximizing Overlap Overview

a. *General Idea*

The idea of maximizing overlap between the 1990 design and the 2000 design, without disturbing the fundamentals of our survey design (too much), is to first regard the two sample selections as one big sampling experiment in which the unconditional probabilities for each design are fixed, but the joint probabilities of selection are adjusted to improve the chance of overlap.

b. *Simple PSU Overlap Maximization Example - Keyfitz's (1954) PSU Overlap Method*

The simplest case for overlap would exist if, for both the old and new designs, you only selected one PSU per stratum, the PSU definitions did not change (i.e., each PSU was made up of the same geographical area in both designs), and stratum definitions did not change (i.e., each stratum contains the same PSUs in both designs).

! Given an NSR stratum in CPS, we'll do the same thing we did in the example above: Set up a matrix with the probabilities of selection for the initial sample as the marginal sums of the rows and the probabilities of selection for the new sample as the marginal sums of the columns. This will eventually be the matrix of Joint probabilities for old and new PSU selections. Example[3]:
  - " Suppose we have a simple stratum with three PSUs.
  - " Let the unconditional probabilities of selection in the initial sample be 0.36, 0.24, and 0.40 for PSU1, PSU2, and PSU3, respectively.
  - " Let the unconditional probabilities of selection in the new sample be 0.50, 0.30, and 0.20 for PSU1, PSU2, and PSU3, respectively.
  - " So the matrix would look like:

**Table 1. Joint Probabilities**

| | | New PSU | | |
|---|---|---|---|---|
| | **Old PSU** | 1 | 2 | 3 |
| 1 | 0.36 | | | |
| 2 | 0.24 | | | |
| 3 | 0.40 | | | |
| | | 0.50 | 0.30 | 0.20 |

[3] Example is taken from Draft Technical Paper 63, Appendix A.

! From there, it is fairly easy to arrange for the maximum probabilities to be on the diagonal, where overlap will occur, and yet still have the rows and columns sum to their required marginals.

**Table 2. Joint Probabilities**

| | | New PSU | | |
|---|---|---|---|---|
| | **Old PSU** | 1 | 2 | 3 |
| 1 | 0.36 | 0.36 | 0 | 0 |
| 2 | 0.24 | 0 | 0.24 | 0 |
| 3 | 0.40 | 0.14 | 0.06 | 0.20 |
| | | 0.50 | 0.30 | 0.20 |

! Finally, we compute the conditional probabilities for each new PSU being selected given that a specific old PSU was in sample by dividing by the marginal probability for that old PSU. In the example:

**Table 3. Conditional Probabilities**

| | | New PSU | | |
|---|---|---|---|---|
| | **Old PSU** | 1 | 2 | 3 |
| 1 | 0.36 | 1 | 0 | 0 |
| 2 | 0.24 | 0 | 1 | 0 |
| 3 | 0.40 | 0.35 | 0.15 | 0.5 |
| | | 0.50 | 0.30 | 0.20 |

! We then note which PSU was actually selected in the initial sample and then use the conditional probabilities in that row to select the new sample. So, in the example, if old PSU #3 was selected in the initial sample, the 3rd row of probabilities (shaded above) will be used to select the new PSU.

c. *The Linear Programming General Solution - One PSU Per Stratum*

Unfortunately, some surveys select two PSUs per stratum and stratifications do change (a lot), so the simple method above needed to be generalized. That is where the Causey et

al.(1985) method came in. It is a linear programming solution to the above problem only generalized so it is applicable to selection of more than one PSU and to the case where stratifications change.

This method is similar to the technique used above by Keyfitz (1954), only it is generalized to cover situations with fewer assumptions. As with Keyfitz, it sets up a matrix with initial sample selection cases as rows and new sample selection cases as columns. However, it does so in a more general approach using linear programming to solve for the joint probabilities in the matrix.

For one specific survey, focusing on a single stratum, S, in the new sample and the PSUs in that stratum:

- Create a joint probability matrix. Each row is a possible combination of those PSUs that could have been selected in the old sample. Each column is a possible combination of those PSUs that could be selected in the new sample.

- Enter the known marginal probabilities for each row and each column.

- Set up a "transportation problem" to be solved using transportation problem-specific software or a general linear programming program.

- $Z = \sum_{i=1}^{m^*} \sum_{j=1}^{n^*} c_{ij} x_{ij}$ that we want to maximize, subject to constraints,

  $\sum_{i=1}^{m^*} x_{ij} = p_j$ J = 1, 2, . . . ,n* (columns sum to the marginals); and

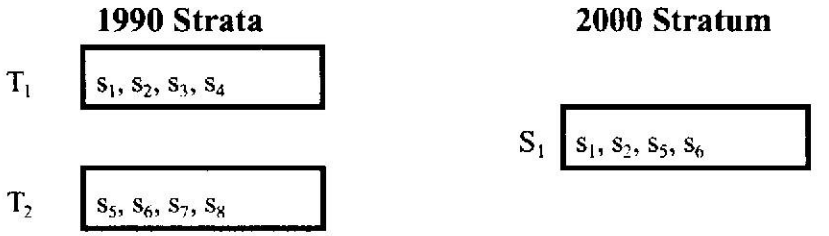
  $\sum_{j=1}^{n^*} x_{ij} = \pi_j$, i = 1, 2, . . . , m* (rows sum to the marginals),

  where $x_{i\,j\,k}$ are the unknown joint probabilities that the software will solve for us, and $c_{i\,j\,k}$ = # of PSUs that overlap between old sample selection j and new sample selection k, so Z is the expected number of overlaps.

- Calculate the conditional probabilities for each row by dividing by the marginal total.

- Then, use the row of conditional probabilities that correspond to the combination that actually was selected in the old sample to select for the new sample.

Example: Consider the simple




Assuming this is a two PSU per stratum design in 1990 and 2000, then it is possible that $s_1$, $s_2$, $s_5$, and $s_6$ were selected in 1990, because of the shift in strata definitions and contents.

So, the possible sets of PSUs in stratum $S_1$ that were selected for sample in 1990 would be, $\varnothing, \{s_1\}, \{s_2\}, \{s_5\}, \{s_6\}, \{s_1, s_2\}, \{s_1, s_5\}, \{s_1, s_6\}, \{s_2, s_5\}, \{s_2, s_6\}, \{s_5, s_6\}, \{s_1, s_2, s_5\}, \{s_1, s_2, s_6\}, \{s_1, s_5, s_6\}, \{s_2, s_5, s_6\}, \{s_1, s_2, s_5, s_6\}$, making 16 possible sets.

And, the possible sets that could be selected in 2000 would be, $\{s_1, s_2\}, \{s_1, s_5\}, \{s_1, s_6\}, \{s_2, s_5\}, \{s_2, s_6\}, \{s_5, s_6\}$.

! Construct a matrix for joint probabilities with the 1990 selection choices as the rows and the 2000 selection choices as the columns, noting the overlap values:

**Table 4. Joint Probabilities**

| | | | 2000 PSU Selections | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | 1 | 2 | 3 | 4 | 5 | 6 |
| | 1990 PSU Selections | | $\{s_1, s_2\}$ | $\{s_1, s_5\}$ | $\{s_1, s_6\}$ | $\{s_2, s_5\}$ | $\{s_2, s_6\}$ | $\{s_5, s_6\}$ |
| 1 | $\varnothing$ | $p_1$ | $x_{11}$ | $x_{12}$ | $x_{13}$ | $x_{14}$ | $x_{15}$ | $x_{16}$ |
| | | | $c_{11}=0$ | $c_{12}=0$ | $c_{13}=0$ | $c_{14}=0$ | $c_{15}=0$ | $c_{16}=0$ |
| 2 | $\{s_1\}$ | $p_2$ | $x_{21}$ | $x_{22}$ | $x_{23}$ | $x_{24}$ | $x_{25}$ | $x_{26}$ |
| | | | $c_{21}=1$ | $c_{21}=1$ | $c_{21}=1$ | $c_{21}=0$ | $c_{21}=0$ | $c_{21}=0$ |
| 3 | $\{s_2\}$ | $p_3$ | $x_{31}$ | $x_{32}$ | $x_{33}$ | $x_{34}$ | $x_{35}$ | $x_{36}$ |
| | | | $c_{31}=1$ | $c_{32}=0$ | $c_{33}=0$ | $c_{34}=1$ | $c_{35}=1$ | $c_{36}=0$ |
| 4 | $\{s_5\}$ | $p_4$ | $x_{41}$ | $x_{42}$ | $x_{43}$ | $x_{44}$ | $x_{45}$ | $x_{46}$ |

| | 1990 PSU Selections | | $\{s_1, s_2\}$ | $\{s_1, s_5\}$ | $\{s_1, s_6\}$ | $\{s_2, s_5\}$ | $\{s_2, s_6\}$ | $\{s_5, s_6\}$ |
|---|---|---|---|---|---|---|---|---|
| | | | $c_{41}=0$ | $c_{42}=1$ | $c_{43}=0$ | $c_{44}=1$ | $c_{45}=0$ | $c_{46}=1$ |
| 5 | $\{s_6\}$ | $p_5$ | $x_{51}$ | $x_{52}$ | $x_{53}$ | $x_{54}$ | $x_{55}$ | $x_{56}$ |
| | | | $c_{51}=0$ | $c_{52}=0$ | $c_{53}=1$ | $c_{54}=0$ | $c_{55}=1$ | $c_{56}=1$ |
| 6 | $\{s_1, s_2\}$ | $p_6$ | $x_{61}$ | $x_{62}$ | $x_{63}$ | $x_{64}$ | $x_{65}$ | $x_{66}$ |
| | | | $c_{61}=2$ | $c_{62}=1$ | $c_{63}=1$ | $c_{64}=1$ | $c_{65}=1$ | $c_{66}=0$ |
| 7 | $\{s_1, s_5\}$ | $p_7$ | $x_{71}$ | $x_{72}$ | $x_{73}$ | $x_{74}$ | $x_{75}$ | $x_{76}$ |
| | | | $c_{71}=1$ | $c_{72}=2$ | $c_{73}=1$ | $c_{74}=1$ | $c_{75}=0$ | $c_{76}=1$ |
| 8 | $\{s_1, s_6\}$ | $p_8$ | $x_{81}$ | $x_{82}$ | $x_{83}$ | $x_{84}$ | $x_{85}$ | $x_{86}$ |
| | | | $c_{81}=1$ | $c_{82}=1$ | $c_{83}=2$ | $c_{84}=0$ | $c_{85}=1$ | $c_{86}=1$ |
| 9 | $\{s_2, s_5\}$ | $p_9$ | $x_{91}$ | $x_{92}$ | $x_{93}$ | $x_{94}$ | $x_{95}$ | $x_{96}$ |
| | | | $c_{91}=1$ | $c_{92}=1$ | $c_{93}=0$ | $c_{94}=2$ | $c_{95}=1$ | $c_{96}=1$ |
| 10 | $\{s_2, s_6\}$ | $p_{10}$ | $x_{101}$ | $x_{102}$ | $x_{103}$ | $x_{104}$ | $x_{105}$ | $x_{106}$ |
| | | | $c_{101}=1$ | $c_{102}=0$ | $c_{103}=1$ | $c_{104}=1$ | $c_{105}=2$ | $c_{101}=1$ |
| 11 | $\{s_5, s_6\}$ | $p_{11}$ | $x_{111}$ | $x_{112}$ | $x_{113}$ | $x_{114}$ | $x_{115}$ | $x_{116}$ |
| | | | $c_{111}=0$ | $c_{112}=1$ | $c_{113}=1$ | $c_{114}=1$ | $c_{115}=1$ | $c_{116}=2$ |
| 12 | $\{s_1, s_2, s_5\}$ | $p_{12}$ | $x_{121}$ | $x_{122}$ | $x_{123}$ | $x_{124}$ | $x_{125}$ | $x_{126}$ |
| | | | $c_{121}=2$ | $c_{122}=2$ | $c_{123}=1$ | $c_{124}=2$ | $c_{125}=1$ | $c_{126}=1$ |
| 13 | $\{s_1, s_2, s_6\}$ | $p_{13}$ | $x_{131}$ | $x_{132}$ | $x_{133}$ | $x_{134}$ | $x_{135}$ | $x_{136}$ |
| | | | $c_{131}=2$ | $c_{132}=1$ | $c_{133}=2$ | $c_{134}=1$ | $c_{135}=2$ | $c_{136}=1$ |
| 14 | $\{s_1, s_5, s_6\}$ | $p_{14}$ | $x_{141}$ | $x_{142}$ | $x_{143}$ | $x_{144}$ | $x_{145}$ | $x_{146}$ |
| | | | $c_{141}=1$ | $c_{142}=2$ | $c_{143}=2$ | $c_{144}=1$ | $c_{145}=1$ | $c_{146}=2$ |
| 15 | $\{s_2, s_5, s_6\}$ | $p_{15}$ | $x_{151}$ | $x_{152}$ | $x_{153}$ | $x_{154}$ | $x_{155}$ | $x_{156}$ |
| | | | $c_{151}=1$ | $c_{152}=1$ | $c_{153}=1$ | $c_{154}=2$ | $c_{155}=2$ | $c_{156}=2$ |
| 16 | $\{s_1, s_2, s_5, s_6\}$ | $p_{16}$ | $x_{161}$ | $x_{162}$ | $x_{163}$ | $x_{164}$ | $x_{165}$ | $x_{166}$ |
| | | | $c_{161}=2$ | $c_{162}=2$ | $c_{163}=2$ | $c_{164}=2$ | $c_{165}=2$ | $c_{166}=2$ |

| 1990 PSU Selections | $\{s_1, s_2\}$ | $\{s_1, s_5\}$ | $\{s_1, s_6\}$ | $\{s_2, s_5\}$ | $\{s_2, s_6\}$ | $\{s_5, s_6\}$ |
|---|---|---|---|---|---|---|
| | $\bullet_1$ | $\bullet_2$ | $\bullet_3$ | $\bullet_4$ | $\bullet_5$ | $\bullet_6$ |

! Now we have to fill in the marginal probabilities. It's here that stratum-to-stratum independence comes in. The way we've selected PSUs, we know the probabilities of the selections in a single stratum in 1990. However, to compute the marginal probability, $p_1$, we have to compute the probability that $s_3$ and $s_4$ were selected in $T_1$ and $s_7$ and $s_8$ in $T_2$. If we have stratum to stratum independence, we just multiply the probabilities. If not, we may not be able to compute $p_1$.

! We now set up a classic transportation problem:

$Z = \sum_{i=1}^{16} \sum_{j=1}^{6} c_{ij} x_{ij}$ that we want to maximize, subject to constraints,

$\sum_{i=1}^{16} x_{ij} = p_i$, i = 1, 2, . . . , 6 (columns sum to the marginals); and

$\sum_{j=1}^{6} x_{ij} = t_j$, j = 1, 2, . . . , 16 (rows sum to the marginals).

And, solve it to compute the values of $x_{ij}$.

! So, the solution to the transportation problem will produce each of the $x_{ij}$ in the joint probability matrix. As with the Keyfitz procedure, we then compute the conditional probabilities by dividing by the row totals:

## Table 5. Conditional Probabilities

| | | | 2000 PSU Selections | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | 1 | 2 | 3 | 4 | 5 | 6 |
| | 1990 PSU Selections | | $\{s_1, s_2\}$ | $\{s_1, s_5\}$ | $\{s_1, s_6\}$ | $\{s_2, s_5\}$ | $\{s_2, s_6\}$ | $\{s_5, s_6\}$ |
| 1 | $\varnothing$ | $p_1$ | $x_{11}/p_1$ | $x_{12}/p_1$ | $x_{13}/p_1$ | $x_{14}/p_1$ | $x_{15}/p_1$ | $x_{16}/p_1$ |
| 2 | $\{s_1\}$ | $p_2$ | $x_{21}/p_2$ | $x_{22}/p_2$ | $x_{23}/p_2$ | $x_{24}/p_2$ | $x_{25}/p_2$ | $x_{26}/p_2$ |
| 3 | $\{s_2\}$ | $p_3$ | $x_{31}/p_3$ | $x_{32}/p_3$ | $x_{33}/p_3$ | $x_{34}/p_3$ | $x_{35}/p_3$ | $x_{36}/p_3$ |
| 4 | $\{s_5\}$ | $p_4$ | $x_{41}/p_4$ | $x_{42}/p_4$ | $x_{43}/p_4$ | $x_{44}/p_4$ | $x_{45}/p_4$ | $x_{46}/p_4$ |
| 5 | $\{s_6\}$ | $p_5$ | $x_{51}/p_5$ | $x_{52}/p_5$ | $x_{53}/p_5$ | $x_{54}/p_5$ | $x_{55}/p_5$ | $x_{56}/p_5$ |
| 6 | $\{s_1, s_2\}$ | $p_6$ | $x_{61}/p_6$ | $x_{62}/p_6$ | $x_{63}/p_6$ | $x_{64}/p_6$ | $x_{65}/p_6$ | $x_{66}/p_6$ |
| 7 | $\{s_1, s_5\}$ | $p_7$ | $x_{71}/p_7$ | $x_{72}/p_7$ | $x_{73}/p_7$ | $x_{74}/p_7$ | $x_{75}/p_7$ | $x_{76}/p_7$ |
| 8 | $\{s_1, s_6\}$ | $p_8$ | $x_{81}/p_8$ | $x_{82}/p_8$ | $x_{83}/p_8$ | $x_{84}/p_8$ | $x_{85}/p_8$ | $x_{86}/p_8$ |
| 9 | $\{s_2, s_5\}$ | $p_9$ | $x_{91}/p_9$ | $x_{92}/p_9$ | $x_{93}/p_9$ | $x_{94}/p_9$ | $x_{95}/p_9$ | $x_{96}/p_9$ |
| 10 | $\{s_2, s_6\}$ | $p_{10}$ | $x_{101}/p_{10}$ | $x_{102}/p_{10}$ | $x_{103}/p_{10}$ | $x_{104}/p_{10}$ | $x_{105}/p_{10}$ | $x_{106}/p_{10}$ |
| 11 | $\{s_5, s_6\}$ | $p_{11}$ | $x_{111}/p_{11}$ | $x_{112}/p_{11}$ | $x_{113}/p_{11}$ | $x_{114}/p_{11}$ | $x_{115}/p_{11}$ | $x_{116}/p_{11}$ |
| 12 | $\{s_1, s_2, s_5\}$ | $p_{12}$ | $x_{121}/p_{12}$ | $x_{122}/p_{12}$ | $x_{123}/p_{12}$ | $x_{124}/p_{12}$ | $x_{125}/p_{12}$ | $x_{126}/p_{12}$ |
| 13 | $\{s_1, s_2, s_6\}$ | $p_{13}$ | $x_{131}/p_{13}$ | $x_{132}/p_{13}$ | $x_{133}/p_{13}$ | $x_{134}/p_{13}$ | $x_{135}/p_{13}$ | $x_{136}/p_{13}$ |
| 14 | $\{s_1, s_5, s_6\}$ | $p_{14}$ | $x_{141}/p_{14}$ | $x_{142}/p_{14}$ | $x_{143}/p_{14}$ | $x_{144}/p_{14}$ | $x_{145}/p_{14}$ | $x_{146}/p_{14}$ |
| 15 | $\{s_2, s_5, s_6\}$ | $p_{15}$ | $x_{151}/p_{15}$ | $x_{152}/p_{15}$ | $x_{153}/p_{15}$ | $x_{154}/p_{15}$ | $x_{155}/p_{15}$ | $x_{156}/p_{15}$ |
| 16 | $\{s_1, s_2, s_5, s_6\}$ | $p_{16}$ | $x_{161}/p_{16}$ | $x_{162}/p_{16}$ | $x_{163}/p_{16}$ | $x_{164}/p_{16}$ | $x_{165}/p_{16}$ | $x_{166}/p_{16}$ |
| | | | $\bullet_1$ | $\bullet_2$ | $\bullet_3$ | $\bullet_4$ | $\bullet_5$ | $\bullet_6$ |

Let's say that we check the old sample and find that $\{s_3, s_4\}$ were selected from $T_1$ and $\{s_5, s_6\}$ were selected from $T_2$. This corresponds to row #11, so we would select the new sample with those conditional probabilities. If there is an entry of 1 in the row, we pick the combination of

PSUs corresponding to the column. If there is more than one non-zero entry, then we have to generate a random number to pick one of the combinations of PSUs. This suggests we do not necessarily pick the combination which has the highest conditional probability.

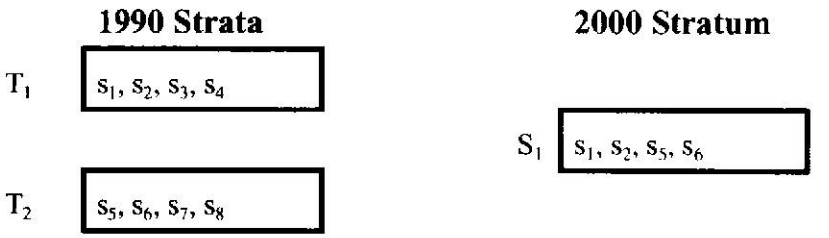
This method is optimal and easy to implement, but the independence assumption was not met by CPS in almost any redesign. So, a variation had to be created that did not need the independence assumption, Ernst (1986).

## 3. Ernst (1986) Overlap Algorithm

### d. *The Problem with the Causey et al. (1985) Method*

Ernst(1986) method sacrifices some of the optimal overlap characteristic in order to be able to deal with the independence problem. The lack of independence in the initial sample makes it at best extremely difficult to compute joint probabilities across strata and at worst impossible. This method attempts to maximize overlap with less information: the unconditional sample probabilities within each of the initial sample's strata.

Suppose we had the simplistic case illustrated below.




Further suppose that we selected one PSU per stratum in 1990 and we are selecting one PSU per stratum in 2000. The possible combinations of the PSUs from the 2000 stratum, $\{s_1, s_2, s_5, s_6\}$, that could have been in sample in 1990 are • , $\{s_1\}$, $\{s_2\}$, $\{s_5\}$, $\{s_6\}$, $\{s_1, s_5\}$, $\{s_1, s_6\}$, $\{s_2, s_5\}$, and $\{s_2, s_6\}$. If we wanted to use the straightforward Causey, et al (1985) method to maximize overlap, we would have to know joint probabilities of selection across the 1990 strata, $T_1$ and $T_2$, for each of them. For example, the probability that $\{s_1\}$ is in the initial sample is the joint probability that $s_1$ is selected from $T_1$ and neither $\{s_5\}$ nor $\{s_6\}$ is selected from $T_2$. If the 1990 sample selection were performed such that each stratum was independently selected, then these joint probabilities could be calculated easily by multiplying the probabilities. However, the sampling of $T_1$ and $T_2$ in 1990 was not performed in a manner making the stratum independent of each other. The only information that is available to us are the probabilities of selection within each 1990 stratum.

e. *Partitioning The Joint Space*

Since we know the probabilities of selection within each individual stratum in the initial sample, the solution presented in Ernst (1986) is to partition the initial sample space by stratum. To do that, he sets up a new step to the overall joint process: the random selection of one of the stratum from the initial sample. So, the simple 1 PSU/stratum case above would become the following 3-dimensional table of joint probabilities:

**Table 6. Joint Probabilities**

| | | | | | 2000 | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | 1 | 2 | 3 | 4 |
| Stratum | | | | | $\{s_1\}$ | $\{s_2\}$ | $\{s_5\}$ | $\{s_6\}$ |
| 1990 | $T_1$ | 11 | • | $y_1p_{11}$ | $x_{111}$ | $x_{112}$ | $x_{113}$ | $x_{114}$ |
| | | 12 | $\{s_1\}$ | $y_1p_{12}$ | $x_{121}$ | $x_{122}$ | $x_{123}$ | $x_{124}$ |
| | | 13 | $\{s_2\}$ | $y_1p_{13}$ | $x_{131}$ | $x_{132}$ | $x_{133}$ | $x_{134}$ |
| | $T_2$ | 21 | • | $y_2p_{21}$ | $x_{211}$ | $x_{212}$ | $x_{213}$ | $x_{214}$ |
| | | 22 | $\{s_5\}$ | $y_2p_{22}$ | $x_{221}$ | $x_{222}$ | $x_{223}$ | $x_{224}$ |
| | | 23 | $\{s_6\}$ | $y_2p_{23}$ | $x_{231}$ | $x_{232}$ | $x_{233}$ | $x_{234}$ |
| | | | | | $\bullet_1$ | $\bullet_2$ | $\bullet_3$ | $\bullet_4$ |

where $y_1$ and $y_2$ are the probabilities of selection for the two stratum, and $p_{ij}$ are the probabilities of selection for each of the PSU possibilities given the selection of the stratum.

f. *Assumptions*

As with Causey, Cox, and Ernst (1985), we focus on each of the new sample's strata, one at a time. Let's call it S. To construct the process, we define our selection process as three steps:

- We note which strata in the old sample (1990 in our case) contain one or more PSUs that are also in the new sample stratum. Call them $T_i$. Then we will randomly select one of those strata. Each of the stratum in the old sample that contains one or more

PSUs in the new sample stratum will be assigned a probability, as determined by our linear programming solution. Call the selected stratum T.

- We then note which PSU or combination of PSUs was selected from stratum T.
- We then select our sample PSU or PSUs from the new sample stratum, S, conditioned on the results of the first two steps.

g. *Formulation*

Our linear programming problem now becomes:

- Maximize $Z = \sum_{i=1}^{r} \sum_{j=1}^{u_i} \sum_{k=1}^{n} c_{ijk} x_{ijk}$ under constraints

$$\sum_{i=1}^{r} \sum_{j=1}^{u_i} x_{ijk} = \pi_k, \; k = 1, 2, \ldots, n \quad \text{and}$$

$$\sum_{k=1}^{n} x_{ijk} = p_{ij} y_i, \; i = 1, 2, \ldots, r, \; j = 1, 2, \ldots, u_i$$

- The index i represents each of the r strata in the old sample that have one or more PSUs in the new sample stratum, S.
- The index j represents each of the $u_i$ sets of PSUs that are in both the new sample stratum and the old sample stratum, $T_i$ (one of the r strata discussed above). For 1990 1-PSU/stratum cases, $u_i$ will just be the PSUs in $T_i$ • S plus 1 (for other PSUs in $T_i$). For 1990 2-PSU/stratum cases, $u_i$ will be the PSUs in $T_i$ • S plus the pairs of the PSUs in $T_i$ • S, plus 1 (for other pairs of PSUs in $T_i$).
- The index k represents each of the possible selection sets of PSUs that could be chosen for the new sample.
- As before, $\pi_k$ is the unconditional probability that the $k^{th}$ possible new sample selection set of PSUs is, in fact, selected.
- $y_i$ is the probability of selecting the $i^{th}$ stratum in the old sample.
- The definition of $c_{ijk}$ gets a little more complicated. As before it is the expected number of overlaps given that the $i^{th}$ old sample stratum was chosen, that the $j^{th}$ PSU set in that stratum was selected for sample, and that the $k^{th}$ new sample PSU set is chosen for the new sample.
- Using notation where $N_{kh}$ is the $h^{th}$ PSU in the $N_k$ PSU set, $h = 1, \ldots, m^{\bullet}$, then

$$c_{ijk} = \sum_{h=1}^{m'} P\left(N_{kh} \in I \mid T = T_i, I_i = I_{ij}, N = N_k\right)$$

- Those probabilities come in three forms:

if $N_{kh} \bullet I_{ij}$, $P\left(N_{kh} \in I \mid T = T_i, I_i = I_{ij}, N = N_k\right) = 1$,

if $N_{kh} \in T_i \cap I_{ij}^c$, $P(N_{kh} \in I \mid T = T_i, I_i = I_{ij}, N = N_k) = 0$

if $N_{kh} \notin T_i$, $P(N_{kh} \in I \mid T = T_i, I_i = I_{ij}, N = N_k) = P(N_{kh} \in I)$

- In the third one above is where we may lose optimality if there is a lack of independence in the initial sample. If $(N_{k\,h} \bullet I)$ is correlated with $(T = T_i, I_i = I_{i\,j})$, the probability we use in our object function will not be accurate. This means that we may not be optimal in maximizing overlap.
- BUT, we do maintain our constraint probabilities, so there is no bias due to inaccurate weights, computed from inaccurate probabilities of selection.

h. *Example* Consider the same example as before.

Assuming this is a two PSU per stratum design in 1990 and 2000, then it is possible that $s_1$, $s_2$, $s_5$, and $s_6$ were selected in 1990, because of the shift in strata definitions and contents.

So, the possible sets of PSUs in stratum $S_1$ that were selected for sample in 1990 would be, from $T_1$ $\varnothing$, $\{s_1\}$, $\{s_2\}$, $\{s_1, s_2\}$ and from $T_2$ $\varnothing$, $\{s_5\}$, $\{s_6\}$, $\{s_5, s_6\}$.

And, the possible sets that could be selected in 2000 would be, $\{s_1, s_2\}$, $\{s_1, s_5\}$, $\{s_1, s_6\}$, $\{s_2, s_5\}$, $\{s_2, s_6\}$, $\{s_5, s_6\}$. We define

$$p_{kh} = P(N_{kh} \mid T = T_i, I = I_{ij})$$

**Table 7. Costs for stratum T1**

| | | 2000 PSU Selections | | | | | |
|---|---|---|---|---|---|---|---|
| | | $k=1$ | $k=2$ | $k=3$ | $k=4$ | $k=5$ | $k=6$ |
| | 1990 PSU Selections | $\{s_1, s_2\}$ | $\{s_1, s_5\}$ | $\{s_1, s_6\}$ | $\{s_2, s_5\}$ | $\{s_2, s_6\}$ | $\{s_5, s_6\}$ |
| 1 | $\{s_1\}$ | Cell 111 | Cell 112 | Cell 113 | Cell 114 | Cell 115 | Cell 116 |
| | | $c_{111}=1$ | $c_{112}=1+p_{21}$ | $c_{113}=1+p_{22}$ | $c_{114}=p_{21}$ | $c_{115}=p_{22}$ | $c_{116}=p_{21}+p_{22}$ |
| 2 | $\{s_2\}$ | Cell 121 | 122 | 123 | 124 | 125 | 126 |
| | | $c_{121}=1$ | $c_{122}=p_{21}$ | $c_{123}=p_{22}$ | $c_{124}=1+p_{21}$ | $c_{125}=1+p_{22}$ | $c_{126}=p_{21}+p_{22}$ |
| 3 | $\{s_1, s_2\}$ | 131 | 132 | 133 | 134 | 135 | 136 |
| | | $c_{131}=2$ | $c_{132}=1+p_{21}$ | $c_{133}=1+p_{22}$ | $c_{134}=1+p_{21}$ | $c_{135}=1+p_{22}$ | $c_{136}=p_{21}+p_{22}$ |
| 4 | $\{\varnothing\}$ | 141 | 142 | 143 | 144 | 145 | 146 |
| | | $c_{141}=0$ | $c_{142}=p_{21}$ | $c_{143}=p_{22}$ | $c_{144}=p_{21}$ | $c_{145}=p_{22}$ | $c_{146}=p_{21}+p_{22}$ |

**Table 8. Costs for stratumT2**

| | | 2000 PSU Selections | | | | | |
|---|---|---|---|---|---|---|---|
| | | $k=1$ | $k=2$ | $k=3$ | $k=4$ | $k=5$ | $k=6$ |
| | 1990 PSU Selections | $\{s_1, s_2\}$ | $\{s_1, s_5\}$ | $\{s_1, s_6\}$ | $\{s_2, s_5\}$ | $\{s_2, s_6\}$ | $\{s_5, s_6\}$ |
| 1 | $\{s_5\}$ | $Cell_{211}$ | Cell 212 | Cell 213 | Cell 214 | Cell 215 | Cell 216 |
| | | $c_{111}=p_{11}+p_{12}$ | $c_{112}=1+p_{11}$ | $c_{113}=p_{11}$ | $c_{114}=1+p_{12}$ | $c_{115}=p_{12}$ | $c_{116}=1$ |
| 2 | $\{s_6\}$ | Cell 221 | 222 | 223 | 224 | 225 | 226 |
| | | $c_{121}=p_{11}+p_{12}$ | $c_{122}=p_{11}$ | $c_{123}=1+p_{11}$ | $c_{124}=p_{12}$ | $c_{125}=1+p_{12}$ | $c_{126}=1$ |
| 3 | $\{s_5, s_6\}$ | 231 | 232 | 233 | 234 | 235 | 236 |
| | | $c_{131}=p_{11}+p_{12}$ | $c_{132}=1+p_{11}$ | $c_{133}=1+p_{11}$ | $c_{134}=1+p_{12}$ | $c_{135}=1+p_{12}$ | $c_{136}=2$ |
| 4 | $\{\varnothing\}$ | 241 | 242 | 243 | 244 | 245 | 246 |
| | | $c_{141}=p_{11}+p_{12}$ | $c_{142}=p_{11}$ | $c_{143}=p_{11}$ | $c_{144}=p_{12}$ | $c_{145}=p_{12}$ | $c_{146}=0$ |

- Construct a matrix for joint probabilities with the 1990 selection choices as the rows and the 2000 selection choices as the columns, noting the overlap values:

- Now we have to fill in the marginal probabilities. It's here that stratum-to-stratum independence comes in. The way we've selected PSUs, we know the probabilities of the selections in a single stratum in 1990. However, we do not have stratum to stratum independence

- We now set up a classic transportation problem:

$Z = \sum_{i=1}^{2} \sum_{j=1}^{4} \sum_{k=1}^{6} c_{ijk} x_{ijk}$ that we want to maximize, subject to constraints,

$$\sum_{k=1}^{6} x_{ijk} = p_{ij} y_i,\ i = 1, 2;\ j=1, 2, 4;\ \text{and}$$

$$\sum_{i=1}^{2} \sum_{j=1}^{4} x_{ijk} = \pi_k,\ j = 1, 2, 3, .. ,6\ \text{(column sums)}.$$

Note $y_i = P(T = T_i)$, solve it to compute the values of $x_{i\,j\,k}$.

! So, the solution to the transportation problem will produce each of the $x_{i\,j}$ in the joint probability matrix. As with the Keyfitz procedure, we then compute the conditional probabilities by dividing by the row totals:

## Table 9. $X_{i\,j\,k}$'s which Maximize Z

| | (i,j) | k=1 | k=2 | k=3 | k=4 | k=5 | k=6 |
|---|---|---|---|---|---|---|---|
| 1 | (1,1)=$\{\varepsilon_1\}$ | $x_{111}'$ | | | | | |
| 2 | (1,2)=$\{\varepsilon_2\}$ | | | | | | |
| 3 | (1,3)=$\{\varepsilon_1,\varepsilon_2\}$ | | | | | | |
| 4 | (1,4)=$\{\bullet\}$ | | | $x_{143}'$ | | | |
| 5 | (2,1)=$\{\varepsilon_5\}$ | | | | | | |
| 6 | (2,2)=$\{\varepsilon_6\}$ | | | | | | |
| 7 | (2,3)=$\{\varepsilon_5,\varepsilon_6\}$ | | | | | $x_{235}'$ | |
| 8 | (2,4)=$\{\bullet\}$ | | | | | | |

In the above $x_{ijk}'$ are specific values maximizing Z, objective function.

Conditional probabilities are calculated using the following formula.

$$P(N = N_k \mid I_1 = I_{1j_1}, I_2 = I_{2j_2}, \ldots I_r = I_{rj_r}) = \sum_{i=1}^{s} \frac{x_{ij_2k}}{p_{ij_2}}$$

Note the above can be proved using the Laplace's rule of succession, which is

$$P(F \mid F_n) = \sum_{i=0}^{k} P(F \mid F_n E_i)\, P(E_i \mid F_n)$$

In Table 1, $\{s_1\}$ is a combination of $\{s_1, s_3\}$ and $\{s_1, s_4\}$. This is because it is a two PSUs per stratum design. In this Table, $\{\bullet\}$ means $\{s_3, s_4\}$. Similarly, in Table 2, $\{s_5\}$ is a combination of $\{s_5, s_7\}$ and $\{s_1, s_8\}$, and $\{\bullet\}$ is $\{s_7, s_8\}$.

In calculating conditional probabilities, all possible realizations in 1990 design should be taken into account. That is, in Table 3, (row 1, row 5), (row 1, row 6), (row 1, row 7), (row 1, row 8), (row 2, row 5), (row 2, row 6), (row 2, row 7), (row 2, row 8), (row 3, row 5), (row 3, row 6), (row 3, row 7), (row 3, row 8), (row 4, row 5), (row 4, row 6), (row 4, row 7) and (row 4, row 8).

Suppose in the old sample $\{s_3, s_4\}$ was selected from $T_1$ and $\{s_5, s_6\}$ was selected from $T_2$. This corresponds to {row 4, row 7}, so we would select the new sample accordingly. Row 4 correspondents to i = 1 and $j_1 = 4$ and row 7 to i=2 and $j_1$=3, the conditional probability will be, when k=1,

$$P(N = N_1 \mid I_1 = I_{14_1}, I_2 = I_{23_1}) = \frac{x_{14_1 1}}{p_{14_1}} + \frac{x_{23_1 1}}{p_{23_1}}$$

For all k, the above probabilities are computed. If there is an entry of 1, we pick that k. If there is more than one non-zero entry, then we have to generate a random number to pick one of the k.

Note this method is not optimal because the stratum to stratum independent selection assumption is not met by any redesign. So, a variation had to be created that did not need the independence assumption, Ernst (1986).

i. *After the Linear Programming Solution*

Once the above linear programming solution is derived, we have the values for the x array and the y vector. The next step is to compute the conditional probability associated with the 1990 actual sample selection. That is done with the following formula:

$$P\left(N = N_k \middle| I_1 = I_{1j_1}, \ldots, I_r = I_{rj_r}\right) = \sum_{i=1}^{r} \frac{x_{ij_ik}}{p_{ij_i}}$$

## 3. Changes in PSU Definitions

a. *Overview*

The final adjustment that must be made to the algorithms is to account for changes in PSU definitions between 1990 and 2000 designs. The PSUs are defined in terms of counties (and in some cases minor civil division or Census county division). Those definitions change for a variety of reasons:

- The survey goes from a regional design to a state design (e.g., Survey of Income and Program Participation)
- A PSU is found to be poorly defined from a logistical sense
- States change county definitions

Because of these changes, the algorithm for the measure of overlap, $c_{ijk}$, is no longer as simple as the definition above. Now we have to change the definition to account for partial overlap.

b. *New Overlap Definition*

The term "PSU overlap" was originally defined to be the expected number of NSR PSUs that are in both the new and old designs. Since we now have to account for partial overlaps, we have to decide what measure we will use to compute the amount of overlap.

The original purpose for increasing overlap was to minimize interviewer changes. In a partial overlap, a reasonable measure would be the probability that the existing interviewer in the 1990 design PSU is also in the 2000 design PSU. To estimate that probability, we will use the number of Housing Units (HUs) in the partial overlap area divided by the HUs in the 1990 PSU. These counts will be from the 2000 Census.

c. *Procedure*

Unfortunately, the formulas are not that simple due to the fact that we could have two PSUs per stratum. Here's the development:

For any combination of the $j^{th}$ possible 1990 selection in the $i^{th}$ 1990 stratum $T_i$ and the $k^{th}$ possible 2000 selection in S, where $i = 1 \ldots r$, $j = 1 \ldots u_i$, and $k = 1 \ldots n$, a measure of overlap, $c_{ijk}$ will be calculated as follows. Because, in general, any stratum could have one or two PSUs in each possible selection, we will introduce some additional notation:

- Compare counties in each of the PSUs (there may be two ) in the $k^{th}$ possible 2000 selection in S to each of the PSUs (there may be two) in the $j^{th}$ possible 1990 selection in the $i^{th}$ 1990 stratum $T_i$. Let $f_{ijtkh}$ = the total Housing Units (HUs) common to the $t^{th}$ PSU in the 1990 set and the $h^{th}$ PSU in the 2000 set, divided by the total HUs in all of the $j^{th}$ possible 1990 selection in the $i^{th}$ 1990 stratum $T_i$. HU counts for both 1990 and 2000 PSUs used in this calculation will be from the 2000 Census.
- $$c_{ijk} = \sum_{h=1}^{m_S} \left[ 1 - \left[ \prod_{t=1}^{v_i} \left( 1 - f_{ijtkh} \right) \right] \prod_{\substack{q=1 \\ q \neq i}}^{r} \left( 1 - \sum_{w=1}^{u_q} p_{qw} \left[ 1 - \prod_{t=1}^{v_i} \left( 1 - f_{qwtkh} \right) \right] \right) \right]$$

  where $m_S$ is the number of PSUs being selected in S and $v_i$ is the number of PSUs that were selected in the $i^{th}$ 1990 stratum.
- However, if the $k^{th}$ possible 2000 selection in S is a rotating PSU, $c_{ijk} = 0$.

d. *Formula Breakdown*

- $\prod_{t=1}^{v_i} \left( 1 - f_{ijtkh} \right)$ is the conditional probability given $T = T_i$ and $I_i = I_{ij}$ that no 1990 sample interviewer from a PSU in $T_i$ resides in $N_{kh}$.
- That is multiplied by $\prod_{\substack{q=1 \\ q \neq i}}^{r} \left( 1 - \sum_{w=1}^{u_q} p_{qw} \left[ 1 - \prod_{t=1}^{v_i} \left( 1 - f_{qwtkh} \right) \right] \right)$, which is the unconditional probability that no 1990 sample interviewer in any PSU in the other r - 1 strata besides the $i^{th}$ one resides in $N_{kh}$.
- So, one minus those two multiplied together gives the probability that at least one interviewer from a 1990 PSU resides in $N_{kh}$, given that $T = T_i$ and $I_i = I_{ij}$.
- Finally, that probability for each of the $N_{kh}$ s (there could be two) is added together to produce the overlap figure.

## V Equivalency of Two Cost Formulae

### V.1 Introduction

Ernst's algorithm for maximizing PSU overlap between two redesigns involves linear programming and thus costs. In the case of 2 PSUs/stratum, the cost can be calculated using joint probability of selection in light of 2 PSU per stratum design. Or it can be computed as if it is a 1 PSU/stratum design but incorporating the fact that 2 PSUs are picked from the stratum. The former formula is quite involved because it uses Durbin's formula for joint probability of selection. However, the latter formula is simple as it does not use Durbin's or any other complex
formula. This memorandum shows that they are in general equivalent.

Maximizing PSU overlap between redesigns involves maximizing

$$\sum_{i=1}^{I}\sum_{j=1}^{J_i}\sum_{k=1}^{K} a_{ijk}\pi_{ijk} \qquad \ldots\ldots \qquad \ldots\ldots \quad (1)$$

where $\pi_{ijk} = P(T = T_i, I_i = I_{ij}, N = N_k)$,
and

$$a_{ijk} = \sum_{h=1}^{m} p_{ijkh}''$$

where

$p_{ijkh}''$ is the conditional probability that $N_{kh}$ was in the 1990 sample given that T= $T_i$ and $I_i = I_{ij}$,

h denotes the PSUs in $N_{kh}$,

m is the number of PSUs in $N_k$, $k^{th}$ possible outcome for the set of new sample PSUs in S.

$p_{ijkh}''$ can be evaluated as follows;

$$p_{ijkh}'' \begin{cases} 1, & \text{if } N_{kh} \in I_{ij} \\ 0, & \text{if } N_{kh} \in T_i - I_{ij} \\ p_{kh}', & \text{otherwise.} \end{cases} \qquad \ldots\ldots\ldots (2)$$

Note $p_{kh}'$ is the unconditional probability that $N_{kh}$ was in the 1990 sample.

Note this formula can be used for deriving costs for both 1 PSU/stratum and 2 PSUs/stratum design.

## V.2 Two Approaches of Computing Costs for 2 PSUs/Stratum Design

We will develop cost formula for 2 PSUs/Stratum design. Two approaches can be taken for computing costs.

Approach 1.

Let $S = (s_1, s_2, \ldots, s_k)$, where $s_i$ is a single PSU in the stratum. Let $p_i = P(s_i)$, which is the probability of selecting one PSU with the probability proportional to size (PPS). That is,

$$p_i = \frac{MOS(s_i)}{\sum_i MOS(s_i)}.$$

Thus $p_{kh}'$ in equation (2) becomes $2p_{kh}$, where $p_{kh} = \frac{MOS(s_{kh})}{\sum_h MOS(s_{kh})}$. The other elements in equation (2) remain the same.

Approach 2.

This approach takes into account the fact that we are dealing with 2 PSUs/stratum design from the beginning. In this approach, $p_{kh}'$ is expressed as

$$p_{kh}' = P(kh') + P(kh j_{h'j}) \qquad \ldots\ldots\ldots (3)$$

h' in $p(kh')$ is a joint event between h and another PSU other than PSU j which was in the same initial stratum k but not in the initial sample. In equation (3) we assume two PSUs (h and j) were in the initial sample and in $T_i$. If there is only one PSU in a joint set $S \cap T_i$, then equation (3) would become $2p_{kh}$, just like in Approach 1.

Theorem. The Costs Derived from Both Approaches Are Identical.

There could be two scenarios. That is, the number of PSUs in $S \cap T_i$, i.e., $n\{S \cap T_i\}$ is 1 or 2[4]

Case 1. $n\{S \cap T_i\} = 2$ for a stratum.
In this case, $P(kh') = 2p_{kh} - P(kh j_{h'j})$. Thus following the equation (3), $p_{kh}'$ reduces to $2p_h$.

---

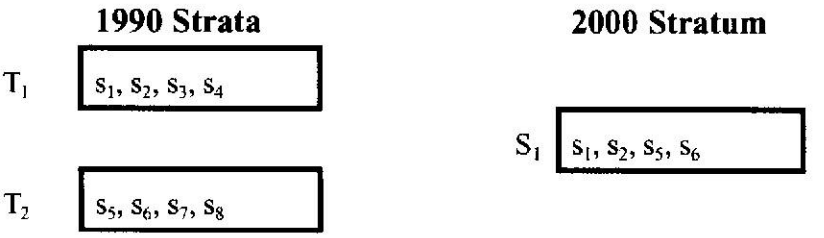
[4] When $n\{S \cap T_i\} = 0$ all strata, PSUs will be selected independently from1990 design. Thus no cost is involved.

This gives the same cost as Approach 1.

Case 2. $n\{S \cap T_i\} = 1$ for a stratum.

In this case $P(kh') = 2p_{kh}$. Thus again this gives the same cost as Approach 1.

Example 1.




1990 strata $T_1$ and $T_2$ have two sample PSUs each. Thus this example belongs to Case 1.

The costs are in Tables below, ignoring prime and double prime. They are derived as follows.

For convenience we define $p_i = 2P(s_i)$ for i=1,2,5,6. Note in this definition of $p_i$, the fact that we are dealing with 2 PSUs/stratum is incorporated. Thus the multiplier 2 is involved. Using approach 1, we have the following.

**Approach 1.**

**Table 10. Costs for stratum T1**

| | | 2000 PSU Selections | | | | | |
|---|---|---|---|---|---|---|---|
| | | k=1 | k=2 | k=3 | k=4 | k=5 | k=6 |
| | 1990 | $\{s_1, s_2\}$ | $\{s_1, s_5\}$ | $\{s_1, s_6\}$ | $\{s_2, s_5\}$ | $\{s_2, s_6\}$ | $\{s_5, s_6\}$ |
| 1 | $\{s_1\}$ | Cell 111 | Cell 112 | Cell 113 | Cell 114 | Cell 115 | Cell 116 |
| | | $c_{111}=1$ | $c_{112}= 1+p_5$ | $c_{113}= 1+p_6$ | $c_{114}= p_5$ | $c_{115}= p_6$ | $c_{116}=p_5+p_6$ |
| 2 | $\{s_2\}$ | Cell 121 | 122 | 123 | 124 | 125 | 126 |
| | | $c_{121}=1$ | $c_{122}= p_5$ | $c_{123}= p_6$ | $c_{124}= + p_5$ | $c_{125}= 1+p_6$ | $c_{126}=p_5+p_6$ |

| | 1990 | $\{s_1, s_2\}$ | $\{s_1, s_5\}$ | $\{s_1, s_6\}$ | $\{s_2, s_5\}$ | $\{s_2, s_6\}$ | $\{s_5, s_6\}$ |
|---|---|---|---|---|---|---|---|
| 3 | $\{s_1, s_2\}$ | 131 | 132 | 133 | 134 | 135 | 136 |
| | | $c_{131}=2$ | $c_{132}=1+p_5$ | $c_{133}=1+p_6$ | $c_{134}=1+p_5$ | $c_{135}=1+p_6$ | $c_{136}=p_5+p_6$ |
| 4 | $\{\varnothing\}$ | 141 | 142 | 143 | 144 | 145 | 146 |
| | | $c_{141}=0$ | $c_{142}=p_5$ | $c_{143}=p_6$ | $c_{144}=p_5$ | $c_{145}=p_6$ | $c_{146}=p_5+p_6$ |

**Table 11. Costs for stratumT2**

| | | 2000 PSU Selections | | | | | |
|---|---|---|---|---|---|---|---|
| | | k=1 | k=2 | k=3 | k=4 | k=5 | k=6 |
| | 1990 PSU Selections | $\{s_5, s_6\}$ | $\{s_5, s_1\}$ | $\{s_5, s_2\}$ | $\{s_6, s_1\}$ | $\{s_6, s_2\}$ | $\{s_1, s_2\}$ |
| 1 | $\{s_5\}$ | Cell 211 | Cell 212 | Cell 213 | Cell 214 | Cell 215 | Cell 216 |
| | | $c_{111}=1$ | $c_{112}=1+p_1$ | $c_{113}=1+p_2$ | $c_{114}=p_1$ | $c_{115}=p_2$ | $c_{116}=p_1+p_2$ |
| 2 | $\{s_6\}$ | Cell 221 | 222 | 223 | 224 | 225 | 226 |
| | | $c_{121}=1$ | $c_{122}=p_1$ | $c_{123}=p_2$ | $c_{124}=1+p_1$ | $c_{125}=1+p_2$ | $c_{126}=p_1+p_2$ |
| 3 | $\{s_5, s_6\}$ | 231 | 232 | 233 | 234 | 235 | 236 |
| | | $c_{131}=2$ | $c_{132}=1+p_1$ | $c_{133}=1+p_2$ | $c_{134}=1+p_1$ | $c_{135}=1+p_2$ | $c_{136}=p_1+p_2$ |
| 4 | $\{\varnothing\}$ | 241 | 242 | 243 | 244 | 245 | 246 |
| | | $c_{141}=0$ | $c_{142}=p_1$ | $c_{143}=p_2$ | $c_{144}=p_1$ | $c_{145}=p_2$ | $c_{146}=p_1+p_2$ |

**Approach 2. 2 PSUs/stratum approach**

Using the same notation as in equation (3), we have the following.

**Table 12. Costs for stratum T3**

| | | 2000 PSU Selections | | | | | |
|---|---|---|---|---|---|---|---|
| | | k=1 | k=2 | k=3 | k=4 | k=5 | k=6 |
| | 1990 | $\{s_1, s_2\}$ | $\{s_1, s_5\}$ | $\{s_1, s_6\}$ | $\{s_2, s_5\}$ | $\{s_2, s_6\}$ | $\{s_5, s_6\}$ |
| 1 | $\{s_1\}$ | Cell 111 | Cell 112 | Cell 113 | Cell 114 | Cell 115 | Cell 116 |
| | | $c_{111}=1$ | $c_{112}= 1+ p_{25'}+p_{256}$ | $c_{113}= 1+ p_{26'}+p_{256}$ | $c_{114}= p_{25'}+ p_{256}$ | $c_{115}= p_{26'} + p_{256}$ | $c_{116}=p_{25'}+ p_{26'}+2 p_{256}$ |
| 2 | $\{s_2\}$ | Cell 121 | 122 | 123 | 124 | 125 | 126 |
| | | $c_{121}=1$ | $c_{122}= p_{25'}+ p_{256}$ | $c_{123}= p_{26'}+ p_{256}$ | $c_{124}= 1+ p_{25'}+p_{256}$ | $c_{125}= 1+ p_{26'}+p_{256}$ | $c_{126}=p_{25'}+ p_{26'}+2 p_{256}$ |
| 3 | $\{s_1, s_2\}$ | 131 | 132 | 133 | 134 | 135 | 136 |
| | | $c_{131}=2$ | $c_{132}=1+p_{25'}+p_{256}$ | $c_{133}=1+p_{26'}+p_{256}$ | $c_{134}=1+p_{25'}+p_{256}$ | $c_{135}=1+p_{26'}+p_{256}$ | $c_{136}=p_{25'}+ p_{26'}+2 p_{256}$ |
| 4 | $\{\emptyset\}$ | 141 | 142 | 143 | 144 | 145 | 146 |
| | | $c_{141}=0$ | $c_{142}=p_{25'}+ p_{256}$ | $c_{143}=p_{26'}+ p_{256}$ | $c_{144}=p_{25'}+ p_{256}$ | $c_{145}=p_{26'}+ p_{256}$ | $c_{146}=p_{25'}+ p_{26'}+2 p_{256}$ |

Note in the above table, $p_{25}'$ is the probability of selecting $s_5$ (note the second subscript 5 in $p_{25}'$ is the same as the subscript in $s_5$. The first subscript 2 represents the fact that it is from the second initial stratum in the old design) in the context of 2 PSUs/stratum. Also $p_{256}$ is the probability of selecting $s_5$ and $s_6$, jointly. Using the definition of $p_5$ in the statement right after the tables which defines $S_1$, $T_1$ and $T_2$, $p_{25}' = p_5 - 2p_{256}$. In cell 112, the cost thus will become $1 + p_5 - 2p_{256} + p_{256} = 1 + p_5$. This is the same as the cost in cell 112 in Table 1. Similarly in cell 126, $p_{26}' = p_6 - 2p_{256}$. Thus $c_{126} = p_{25}' + p_{26}' + 2p_{256}$
$= p_5 - p_{256} + p_6 - p_{256} + 2p_{256} = p_5 + p_6$, which is again the same as the one in Table 1.

Table 4 equivalent to Table 2 can be created similar to Table 3.

# Small Area Estimation

Statistics Canada 근무
Westat 선임연구원(현)

이 현 식 박사

# Small Area (Domain) Estimation: with Emphasis on Estimation of the Mean Squared Error

Hyunshik Lee

## 1. Introduction

Most sample surveys are designed to produce reliable estimates for the whole population and for some large subpopulations. However, users of the survey data often wish to have reasonable estimates at much smaller subpopulations (domains). These small subpopulations may be small geographic areas (e.g., small cities, Goon) or small demographic groups (e.g., 30 years old male with computer engineer's degree) for household surveys or small sectors of industries (e.g., detailed industrial classifications) for business surveys. Techniques of small area or domain estimation have been motivated for estimation of demographic variables for actually small geographic areas. However, demand for such estimates have been expanded to general small subpopulations, which include both small areas as well as small domains, which are not necessarily related to small areas. Nonetheless, the term small area estimation is still in use to mean the estimation problem for any small domain regardless whether the small domain is related to a small geographic area or not. In this note, a small area and a small domain are used interchangeably.

The "smallness" is a relative term in the small area estimation. A large area can be a small area for estimation if the sample at hand does not support the usual estimates that would be used when there is a large sample base. The flip side is also true; a small area can be a large area when there is a large enough sample base for reliable estimation for the domain. Therefore, a small area estimation problem arises when the sample design was not intended to produce reliable estimate for a certain domain but a request for such estimate comes after the fact. If such request is not temporary but will be repeated in a continuous survey (e.g., labor force survey), then it would be advantageous to take that

into consideration in the redesign of the survey as shown by Singh et al. (1992). If a small fraction of the sample size is redistributed to the planned small areas, the ability to produce good small area estimates can be greatly increased without a substantial loss of precision for the large areas the survey originally designed for. Therefore, if small area estimation is planned in advance to produce on a continuous basis, a serious consideration should be given to incorporate such plan in the design or a redesign of the survey.

Small area estimation is one area in survey sampling, where the classical model based statistical theory and traditional design-based methodology intertwine. This is so because the purely design-based methods of estimation, which are sometimes called "direct estimation methods" fail to produce reasonably acceptable estimates, simply because there is no sample base large enough to produce such estimate or sometimes no sample data at all for some small areas. Therefore, it is necessary to borrow "strength" from outside of the small area. Borrowing strength is usually achieved by using some model. The strength borrowed may be in the form of prior knowledge about the small area and thus, Bayesian theory plays a prominent role in small area estimation in order to incorporate the prior knowledge in estimation. Good quality auxiliary data provide a large amount of strength, in which a relatively simple model coupled with such quality auxiliary data can produce good small area estimates. Administrative data collected for administrative purposes, census data although may be old but available at the very small area level, other area data collected in the same survey for the same characteristic, and data collected for the same variable by a different survey, all are potentially useful sources of data for small area estimation. It is not an overstatement to say that the more available useful auxiliary data are, the better job for small area estimation can be done.

## 2. Small Area Estimation Models

Because of the heavy use of auxiliary data in small area estimation, it is not a coincidence that most models used for small area estimation are special cases of the

generalized linear regression model (Marker, 1999). Small area estimation requires pooling of data from outside of the small area. The first idea of doing this was to use a large area regression model assuming that a large area regression model holds for the small area as well, and then a small area estimate is obtained by applying the large area regression equation estimated using the large area data to the small area auxiliary data. That is, the small area estimate is obtained by synthetically combining the large area regression equation and small area auxiliary data. Early small area estimators were mostly of this form.

Let the underlying regression model is expressed as

$$y_{ij} = \mathbf{x}_{ij}^T \beta + e_{ij}, \tag{2.1}$$

where $y_{ij}$ is the $j$-th unit in $i$-th small area, $j = 1, 2, \ldots, N_i$, $i = 1, 2, \ldots, m$, $\mathbf{x}_{ij}$ is the $ij$-th auxiliary vector of $p$ dimension, and $e_{ij}$ is the error term with $E(e_{ij}) = 0$, $V(e_{ij}) = \sigma^2$. Suppose that we are interested in estimation of small area population totals. The synthetic estimator for the total, $Y_i$ of small area $i$ is then given by

$$\hat{Y}_i^S = \mathbf{x}_i^T \hat{\beta} \tag{2.2}$$

where $\mathbf{x}_i = \sum_{j=1}^{N_i} \mathbf{x}_{ij}$ and $\hat{\beta}$ is the usual estimated regression coefficient obtained from the large area. Model (2.1) is given at the element level, which is applicable when the auxiliary variable is available at the element level. However, the model can be written as an area level model as follows:

$$Y_i = \mathbf{x}_i^T \beta + e_i, \tag{2.3a}$$

which is applicable when the auxiliary variables are available at the small area level. Under the superpopulation framework, $e_i$'s are independent and with zero mean and a variance, which approaches to zero as $n_i \to N_i$. Under the finite population context, $Y_i$ is the true $i$-th small area population total without error and thus, we may write

$$Y_i = \mathbf{x}_i^T \beta \quad \text{and} \quad \hat{Y}_i = Y_i + e_i = \mathbf{x}_i^T \beta + e_i, \tag{2.3b}$$

where $\hat{Y}_i$ is the direct estimator and $e_i$ is the sampling error with mean zero and sampling variance.

The estimator given in (2.2) is based on the assumption that model (2.1) or (2.3a or b) holds for all small areas. This assumption, however, does not hold almost always and the synthetic estimator given in (2.2) can be severely biased. The bias can be even dominating in the mean square error (MSE) of the synthetic estimator. The synthetic estimator brings the stability but also the bias, while the direct estimator is unbiased but too unstable. Therefore, a natural way of addressing the bias problem is to compromise between bias and variance, in which, to reduce the bias of the synthetic estimator, the unbiased but highly variable direct estimator is brought back to form a convex linear combination of the synthetic and direct estimators. More about this estimator called the composite estimator, will be discussed later.

Another way, which is more flexible to incorporate small area differences, is to include an area-specific random effect term in the model. Fay and Herriot (1979) first considered the use of such random effect model for small area estimation, by which they opened a new chapter in small area estimation. The random effect model is given by

$$Y_i = \mathbf{x}_i^T \beta + z_i b_i, \quad i = 1, 2, \ldots, m, \tag{2.4}$$

where the $z_i$'s are known constants and $b_i$'s are independently and identically distributed (iid) random variables having $E(v_i) = 0$ and $V(b_i) = \sigma_b^2$. Also, $b_i$'s are often assumed to

follow the normal distribution $N(0, \sigma_b^2)$. Through this model, it is possible to incorporate the between-small area variance in the estimation.

If there is a direct estimator $\hat{Y}_i$ available, it can be written as

$$\hat{Y}_i = Y_i + e_i, \tag{2.5}$$

where the $e_i$'s are the sampling errors independent of $b_i$'s and with $E(e_i \mid Y_i) = 0$ and $V(e_i \mid Y_i) = \sigma_{ei}^2$. Combining (2.4) and (2.5), the small area level random effect model is given by

$$\hat{Y}_i = \mathbf{x}_i^T \beta + z_i b_i + e_i, \quad i = 1, 2, \ldots, m. \tag{2.6}$$

The element level random effect model is usually given in the following form:

$$y_{ij} = \mathbf{x}_{ij}^T \beta + b_i + e_{ij}, \quad i = 1, 2, \ldots, m, \quad j = 1, 2, \ldots, N_i, \tag{2.7}$$

where the $b_i$'s are the same as before and $e_{ij}$'s are iid random variables that are independent of the $b_i$'s and have $E(e_{ij}) = 0$ and $V(e_{ij}) = \sigma_{eij}^2 = c_{ij}\sigma_e^2$. This model is a nested regression model, which was employed by Battese et al. (1988).

## 3. Small Area Estimators

The composite estimator mentioned above is defined as a linear (convex) combination of the synthetic estimator and the direct estimator as shown below:

$$\hat{Y}_i^C = \gamma_i \hat{Y}_i + (1 - \gamma_i)\hat{Y}_i^S, \tag{3.1}$$

where $0 \le \gamma_i \le 1$. The optimal weighting factor $\gamma_i$ is given by

$$\gamma_i^O = \frac{\mathrm{MSE}(\hat{Y}_i^S)}{\mathrm{MSE}(\hat{Y}_i^S) + V(\hat{Y}_i)}, \tag{3.2}$$

which can be estimated by

$$\hat{\gamma}_i^O = \frac{(\hat{Y}_i^S - \hat{Y}_i)^2 + \hat{V}(\hat{Y}_i)}{(\hat{Y}_i^S - \hat{Y}_i)^2}. \tag{3.3}$$

However, it is not stable and there are various methods proposed to get a stable factor (e.g., Shaible, 1978; Purcell and Kish, 1979), some of which lead to the James-Stein estimator.

Simpler factors that depend on the sample size have also been proposed and used in actual surveys. For example, Drew et al. (1982) proposed the following sample size dependent weighting factor:

$$\hat{\gamma}_i^D = \begin{cases} 1, & \text{if } \hat{N}_i \ge \delta N_i \\ \hat{N}_i / (\delta N_i), & \text{otherwise} \end{cases}, \tag{3.3}$$

where $\delta$ is a constant subjectively chosen to control the weighting factor. A value of 2/3 has been used for the Canadian Labor Force Survey. Särndal and Hidiroglou (1989) proposed a different sample size dependent factor.

The optimal factor $\gamma_i$ can be obtained using the theory for the random effect model given in (2.6) (or (2.7)). For the formulation of a small area estimator using the random effect models given in Section 2, there are three basic approaches. The first approach, which Fay and Herriot (1979) first employed is the empirical best linear unbiased prediction (EBLUP). The name EBLUP was not coined by Fay and Herriot but

by Harville (1991). This approach is also called the variance components approach. The second is the empirical Bayes approach, which is based on the theory developed by Morris (1983). The third approach is the hierarchical Bayes approach (HB). These will be discussed in more detail in the following.

### 3.1 Empirical Best Linear Unbiased Prediction (EBLUP) Approach

The model given in (2.6) or (2.7) contains fixed effects as well as random effects. The small area parameter of interest (e.g., population total) can be expressed as linear combinations of the fixed and random effects (i.e., $\mathbf{x}_i^T\beta + z_i b_i$), which can be estimated by the best linear unbiased predictor (BLUP) as derived by Henderson (1950). Under the model (2.6) the BLUP estimator for the *i*-th small area is given by

$$\tilde{Y}_i^H = \gamma_i \hat{Y}_i + (1-\gamma_i)\mathbf{x}_i^T\tilde{\beta} \tag{3.1}$$

where $\hat{Y}_i$ is the direct estimator as before and $\mathbf{x}_i^T\tilde{\beta}$ is the synthetic estimator with the best linear unbiased estimator (BLUE) $\tilde{\beta}$ given by

$$\tilde{\beta} = \left(\sum_{i=1}^{m} \frac{\mathbf{x}_i\mathbf{x}_i^T}{\sigma_b^2 z_i^2 + \sigma_{ei}^2}\right)^{-1}\left(\sum_{i=1}^{m} \frac{\mathbf{x}_i\hat{Y}_i}{\sigma_b^2 z_i^2 + \sigma_{ei}^2}\right) \tag{3.2}$$

and

$$\gamma_i = \sigma_b^2 z_i^2 / (\sigma_b^2 z_i^2 + \sigma_{ei}^2). \tag{3.3}$$

This factor determines the weights given to the synthetic part and the direct estimator. If the between-area variance (the modeling variance) $\sigma_b^2 z_i^2$ is large relative to the total

variance $\sigma_b^2 z_i^2 + \sigma_{ei}^2$, then more weight is given to the direct estimator. Otherwise, more weight is given to the synthetic estimator. This estimator is design consistent since $\gamma_i \to 1$ as the sampling variance $\sigma_{ei}^2 \to 0$ and applicable for general sampling designs because it is modeled at the small area level not at the element level. An excellent summary and applications of the BLUP theory was given by Robinson (1991).

The BLUP estimator is *linear* in $y$-values; *unbiased* in the sense that its expected value is equal to the expected value of $Y_i$; *best* since it has minimum MSE among all linear unbiased predictors. It has become customary to say that fixed effects are estimated and random effects are predicted. Since the small area parameter contains the random effect, it is predicted rather than estimated as the distinction goes.

To obtain the MSE of the BLUP estimator, write

$$\begin{aligned}
\tilde{Y}_i^H - Y_i &= \gamma_i \hat{Y}_i + (1-\gamma_i)\mathbf{x}_i^T \tilde{\beta} - Y_i \\
&= \gamma_i \hat{Y}_i + (1-\gamma_i)\mathbf{x}_i^T \beta - Y_i + (1-\gamma_i)\mathbf{x}_i^T (\tilde{\beta} - \beta) \\
&= \gamma_i (\mathbf{x}_i^T \beta + z_i b_i + e_i) + (1-\lambda_i)\mathbf{x}_i^T \beta - (\mathbf{x}_i^T \beta + z_i b_i) + (1-\gamma_i)\mathbf{x}_i^T (\tilde{\beta} - \beta) \\
&= \gamma_i e_i + (\gamma_i - 1) z_i b_i + (1-\gamma_i)\mathbf{x}_i^T (\tilde{\beta} - \beta)
\end{aligned} \tag{3.4}$$

and note that

$$\begin{aligned}
&\mathrm{Cov}\left\{\gamma_i e_i + (\gamma_i - 1) z_i b_i, \mathbf{x}_i^T (\tilde{\beta} - \beta)\right\} = \mathrm{Cov}\left\{\gamma_i e_i + (\gamma_i - 1) z_i b_i, \mathbf{x}_i^T \tilde{\beta}\right\} \\
&= \mathrm{Cov}\left\{\gamma_i e_i + (\gamma_i - 1) z_i b_i, \mathbf{x}_i^T \left(\sum_{i=1}^m \frac{\mathbf{x}_i \mathbf{x}_i^T}{\sigma_b^2 z_i^2 + \sigma_{ei}^2}\right)^{-1} \left(\sum_{i=1}^m \frac{\mathbf{x}_i \hat{Y}_i}{\sigma_b^2 z_i^2 + \sigma_{ei}^2}\right)\right\} \\
&= \left\{\gamma_i \mathbf{x}_i^T \left(\sum_{i\,1}^m \frac{\mathbf{x}_i \mathbf{x}_i^T}{\sigma_b^2 z_i^2 + \sigma_{ei}^2}\right)^{-1} \mathbf{x}_i \sigma_{ei}^2 - (1-\gamma_i)\mathbf{x}_i^T \left(\sum_{i=1}^m \frac{\mathbf{x}_i \mathbf{x}_i^T}{\sigma_b^2 z_i^2 + \sigma_{ei}^2}\right)^{-1} \mathbf{x}_i \sigma_b^2 z_i^2\right\} \Big/ \left(\sigma_b^2 z_i^2 + \sigma_{ei}^2\right) \\
&= 0.
\end{aligned} \tag{3.5}$$

Then, using (3.4) and (3.5), the MSE of the BLUP estimator can be written as in Prasad and Rao (1990) as

$$M_{1i}(\theta) = E(\tilde{Y}_i^{II} - Y_i)^2 = g_{1i}(\theta) + g_{2i}(\theta) \tag{3.6}$$

where $\theta = (\sigma_b^2, \sigma_{ei}^2)$,

$$g_{1i}(\theta) = E[\gamma_i e_i + (\gamma_i - 1) z_i b_i]^2 = \sigma_b^2 z_i^2 \frac{\sigma_{ei}^2}{\sigma_b^2 z_i^2 + \sigma_{ei}^2} = \gamma_i \sigma_{ei}^2, \tag{3.7}$$

and

$$g_{2i}(\theta) = (1-\gamma_i)^2 \mathbf{x}_i^T \left( \sum_{i=1}^{m} \frac{\mathbf{x}_i \mathbf{x}_i^T}{\sigma_b^2 z_i^2 + \sigma_{ei}^2} \right)^1 \mathbf{x}_i . \tag{3.8}$$

Note that

$$E(\tilde{\beta} - \beta)^2 = \left( \sum_{i=1}^{m} \frac{\mathbf{x}_i \mathbf{x}_i^T}{\sigma_b^2 z_i^2 + \sigma_{ei}^2} \right)^{-1}, \tag{3.9}$$

which was used to derive $g_{2i}(\theta)$ in (3.8). It is usually assumed that the sampling variance component $\sigma_{ei}^2$ is known, which can be too restrictive for some applications. For now we assume that $\sigma_{ei}^2$ is known but $\sigma_b^2$ is unknown and thus, only $\sigma_b^2$ has to be estimated. Various methods have been proposed, among which are the maximum likelihood (ML) estimator and the restricted maximum likelihood (REML) estimator assuming the normality, the method of fitting constants, and the method of moment. For example, using the fact that $E\left[\sum_{i}^{m}{}_1 (\hat{Y}_i - \mathbf{x}_i^T \tilde{\beta})^2 / (\sigma_b^2 z_i^2 + \sigma_{ei}^2)\right] = m - p$, Fay and Herriot (1979) used the method of moments estimator $\hat{\sigma}_b^2$, which can be obtained by solving the equation

$$\sum_{i=1}^{m}(\hat{Y}_i - \mathbf{x}_i^T\tilde{\beta})^2/(\sigma_b^2 z_i^2 + \sigma_{ei}^2) = m - p \qquad (3.10)$$

iteratively with $\tilde{\beta}$ given in (3.2). If there is no positive solution, then we set $\hat{\sigma}_b^2 = 0$. These estimators are asymptotically consistent.

Let $\hat{\gamma}_i$ be obtained by plugging in (3.3) one of those asymptotically consistent estimator $\hat{\sigma}_b^2$ mentioned above. With this $\hat{\gamma}_i$, we get a small area estimator for $Y_i$ given by

$$\hat{Y}_i^H = \hat{\gamma}_i\hat{Y}_i + (1-\hat{\gamma}_i)\mathbf{x}_i^T\tilde{\beta}. \qquad (3.11)$$

This estimator is called the empirical BLUP or EBLUP estimator (Harville, 1991) and Kackar and Harville (1984) showed that it is unbiased if $\hat{\sigma}_b^2$ is an even function of $\hat{Y}_i$'s, translation-invariant, and the distributions of $b_i$ and $e_i$ are both symmetric (not necessarily normal). This assumption of symmetry is critical and is not always true. However, if $b_i$ and $e_i$ are not only symmetric but also normal, Kackar and Harville (1984) also showed that the MSE of the EBLUP estimator can be written as

$$M_{2i}(\theta) \equiv E(\hat{Y}_i^H - Y_i)^2 = E(\tilde{Y}_i^H - Y_i)^2 + E(\hat{Y}_i^H - \tilde{Y}_i^H)^2 = M_{1i}(\theta) + E(\hat{Y}_i^H - \tilde{Y}_i^H)^2 . \qquad (3.12)$$

The second term is in general not tractable and using Taylor series expansion, they obtained a second order approximation $(\hat{Y}_i^H - \tilde{Y}_i^H)^2 = \left[\tilde{Y}_i^H(\hat{\theta}) - \tilde{Y}_i^H(\theta)\right]^2 \approx \left[d(\theta)^T(\hat{\theta}-\theta)\right]^2$ where $d(\theta)' = \partial\tilde{Y}_i^H(\theta)/\partial\theta$ and based on this approximation they proposed

$$E\left[d(\theta)^T(\hat{\theta}-\theta)\right]^2 \approx \mathrm{tr}\left[A(\theta)(\hat{\theta}-\theta)(\hat{\theta}-\theta)^T\right], \qquad (3.13)$$

where $A(\theta)$ is the covariance matrix of $d(\theta)$. Prasad and Rao (1990) obtained a further approximation to (3.13) and gave an approximate expression for the second term in (3.12) as follows:

$$g_{3i}(\sigma_b^2) = \sigma_{ei}^4 z_i^4 (\sigma_b^2 z_i^2 + \sigma_{ei}^2)^{-3} V(\hat{\sigma}_b^2), \tag{3.14}$$

where $V(\hat{\sigma}_b^2)$ is the asymptotic variance of $\hat{\sigma}_b^2$, which is given by

$$V(\hat{\sigma}_b^2) = \frac{2}{m} \sum_{i=1}^{m} (\sigma_b^2 + \sigma_{ei}^2 / z_i^2)^2 . \tag{3.15}$$

Recently, other tighter expressions of $V(\hat{\sigma}_b^2)$ have been proposed (Rao, 2001).

The MSE of the EBLUP estimator is often estimated by $M_{1i}(\hat{\theta})$ (i.e., $M_{1i}(\hat{\sigma}_b^2)$ since $\sigma_{ei}^2$ is assumed known), which is given by (3.6) with $\sigma_b^2$ replaced by $\hat{\sigma}_b^2$. However, it can severely underestimate the true MSE and the bias is approximately $g_{3i}(\sigma_b^2)$ in (3.14). Prasad and Rao (1990) also showed that

$$E\left[g_{1i}(\hat{\theta})\right] = g_{1i}(\theta) - g_{3i}(\theta) + o(m^{-1}), \tag{3.16}$$

and from this, they obtained an approximately unbiased estimator of the MSE of the EBLUP estimator $\hat{Y}_i^H$ with expectation correct to $o(m^{-1})$ as given by

$$\text{mse}(\hat{Y}_i^H) = g_{1i}(\hat{\theta}) + g_{2i}(\hat{\theta}) + 2g_{3i}(\hat{\theta}). \tag{3.17}$$

Lahiri and Rao (1995) showed that this estimator is robust against violation of the normality assumption of $b_i$'s in model (2.6).

### 3.2 Empirical Bayes Approach

This approach is based on the Bayesian theory. Assuming first that the model parameters are known, the posterior distribution of the small area parameter given the sample data is obtained and then estimated model parameters are used to obtain an estimated posterior distribution.

Under the model (2.6) with the normality assumption of the errors, the posterior distribution of $Y_i$'s given $\hat{Y}_i$, $\beta$, and $\sigma_b^2$, is normal with mean $Y_i^B$ and variance $g_{1i}(\theta)$ as given in (3.7), where

$$Y_i^B = \gamma_i \hat{Y}_i + (1-\gamma_i)\mathbf{x}_i^T \beta, \quad i = 1, 2, ..., m, \tag{3.18}$$

and $\gamma_i$ is given in (3.3). $Y_i^B$ is the Bayes estimator of $Y_i$ under quadratic loss. Replacing the unknown parameters $\beta$ and $\sigma_b^2$ by their estimates (from marginal distributions), we obtain the empirical Bayes (EB) estimator

$$\hat{Y}_i^{EB} = \hat{\gamma}_i \hat{Y}_i + (1-\hat{\gamma}_i)\mathbf{x}_i^T \tilde{\beta}, \quad i = 1, 2, ..., m. \tag{3.19}$$

This is identical with the EBLUP estimator $\hat{Y}_i^H$ when the same estimates for $\beta$ and $\sigma_b^2$ are used. Moreover, if the frequentist framework is used to measure the uncertainty of the EB estimator, the MSE is the natural choice and then the EBLUP and EB estimators are essentially the same. Therefore, all discussion for the EBLUP concerning the MSE applies to the EB estimator.

However, from Bayesian perspective, inferences are made using the posterior distribution of $Y_i$'s given $\hat{Y}_i$'s with $E(Y_i \mid \hat{Y})$ and $V(Y_i \mid \hat{Y})$ where $\hat{Y} = (\hat{Y}_1, ..., \hat{Y}_m)^T$. When estimates for $\beta$ and $\sigma_b^2$ are used, then $\hat{Y}_i^H = E(Y_i \mid \hat{Y}, \hat{\beta}, \hat{\theta})$ is a reasonable

approximation to $Y_i^B$ but $g_{1i}(\hat{\beta},\hat{\theta}) = V(Y_i \mid \hat{Y},\hat{\beta},\hat{\theta})$ underestimates $V(Y_i \mid \hat{Y})$. This becomes clear if we write

$$V(Y_i \mid \hat{Y}) = E_{\beta,\theta}[V(Y_i \mid \hat{Y},\beta,\theta)] + V_{\beta,\theta}[E(Y_i \mid \hat{Y},\beta,\theta)]. \tag{3.20}$$

The problem is, however, that this expression in (3.20) cannot be evaluated without having the prior distribution of $\beta$ and $\theta$, which is not specified in the EB approach. Note that $g_{1i}(\hat{\beta},\hat{\theta}) = V(Y_i \mid \hat{Y},\hat{\beta},\hat{\theta})$ is a good approximation to the first term of $V(Y_i \mid \hat{Y})$ only in (3.20) and thus, the naïve MSE estimator can be severely biased by missing the second term. Some methods are available to overcome this problem; the bootstrap method of Laird and Louis (1987) and the method using an asymptotic approximation of $V(Y_i \mid \hat{Y})$ proposed by Kass and Steffey (1989). Another natural solution to this problem is to specify the prior distribution of $\beta$ and $\theta$, which will be discussed in the next subsection.

Morris (1983) provided an excellent account of the EB approach and its applications.

### 3.3 Hierarchical Bayes (HB) Approach

As explained in subsection 3.2, by specifying the prior distribution of $\beta$ and $\sigma_b^2$ (still assuming that $\sigma_{ei}^2$ is known), a small area estimate are obtained from the posterior distribution of $Y_i$ given $\hat{Y}$ as its mean and its uncertainty is measured by the posterior variance. The approach often involves high dimensional integration and thus computationally intensive. However, modern advances on computational equipments and techniques make the implementation of the approach easier.

Besides the usual EB assumption of normal errors under model (2.6), we assume first that the unknown $\beta$ has a noninformative uniform prior over $R^p$ to reflect the absence of prior information on $\beta$ and also that $\sigma_b^2$ is known. Then we can get the joint

probability density function $f(\hat{Y},Y,\beta)$ of $\hat{Y}^T=(\hat{Y}_1,\hat{Y}_2,...,\hat{Y}_m)$, $Y^T=(Y_1,Y_2,...,Y_m)$, and $\beta$. Integrating this density function over $\beta$, we obtain the joint density function $f(\hat{Y},Y)$ and the posterior distribution $f(Y\mid\hat{Y})$. Under this setup, the HB estimator for the $i$-th small area and its variance are given by the posterior mean and posterior variance as follows:

$$E(Y_i\mid\hat{Y})=\gamma_i\hat{Y}_i+(1-\gamma_i)\mathbf{x}_i^{\mathrm{T}}\hat{\beta}=\tilde{Y}_i^H,\quad V(Y_i\mid\hat{Y})=M_{i1}=\mathrm{MSE}(\tilde{Y}_i^H-Y_i). \tag{3.21}$$

Therefore, the HB estimator with the non-informative prior on $\beta$ and known $\theta$ is identical with the BLUP estimator.

When $\sigma_b^2$ is unknown but $\sigma_{ei}^2$'s are known, then the posterior distribution is conditional on $\sigma_b^2$ as well, that is, $f(Y\mid\hat{Y},\sigma_b^2)$. Ghosh (1992) derived a closed form posterior distribution $f(\sigma_b^2\mid\hat{Y})$ assuming a non-informative prior on $\sigma_b^2$ over $(0,\infty)$. With this, we obtain

$$f(Y_i\mid\hat{Y})=\int f(Y_i\mid\hat{Y},\sigma_b^2)f(\sigma_b^2\mid\hat{Y})d\sigma_b^2. \tag{3.22}$$

More generally when $\theta=(\sigma_b^2,\sigma_{ei}^2)$ is unknown, the posterior distribution of $Y_i$ given $\hat{Y}$ is given in the following form:

$$f(Y_i\mid\hat{Y})=\int f(Y_i\mid\hat{Y},\theta)f(\theta\mid\hat{Y})d\theta. \tag{3.23}$$

When we compare the HB approach with the EB approach, the small area point estimators are similar but their variance estimators are markedly different because the naïve EB estimator of the MSE does not account the uncertainty of $\hat{\sigma}_b^2$ used to define the estimator. The HB approach has a clear edge on this regard. However, the posterior

distribution $f(\theta \mid \hat{Y})$ does not have a closed form in many cases, for which we can use a numerical method but it is computationally intensive. The computational difficulties can be overcome by using the Gibbs sampler (Gelfand and Smith, 1990), which is based on the popular Markov Chain Monte Carlo (MCMC) simulation method. This is also computationally intensive but can be applied to a general problem and once a computer system is setup, we can proceed routinely. To illustrate the Gibbs sampling idea, let $\theta = (\beta, Y_1, Y_2, \ldots, Y_m, \sigma_b^2)^T = (\theta_1, \theta_2, \ldots, \theta_{m+2})^T$. Starting with arbitrary initial values $\theta^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)}, \ldots, \theta_{m+2}^{(0)})^T$, we generate the $l$-th simulated sample $\theta^{(l)} = (\theta_1^{(l)}, \ldots, \theta_{m+2}^{(l)})^T$, $\theta_1^{(l)}$ from $f(\theta_1 \mid \hat{Y}, \theta_2^{(l-1)}, \ldots, \theta_{m-2}^{(l-1)})$ and $\theta_2^{(l)}$ from $f(\theta_2 \mid \hat{Y}, \theta_1^{(l-1)}, \theta_3^{(l-1)}, \ldots, \theta_{m-2}^{(l-1)})$, and so on. Under some regularity conditions, the simulated sample converges to $f(\theta \mid \hat{Y})$ in distribution as $l \to \infty$. Thus, after a large enough burn-in period, we obtain a sample of $\theta$'s of an appropriate size (say, $J$) from an approximate distribution of $f(\theta \mid \hat{Y})$. We can use the sample to calculate the posterior mean and variance to get the HB small area estimator and its variance estimator as follows:

$$\begin{aligned} \hat{Y}_i^{HB} &= E(Y_i \mid \hat{Y}) \approx \frac{1}{J}\sum_{j=1}^{J} \hat{Y}_i(j) = \tilde{Y}_i^{HB}, \\ V(Y_i \mid \hat{Y}) &\approx \frac{1}{J}\sum_{j=1}^{J}\left[g_{1i}(\theta(j)) + g_{2i}(\theta(j))\right] + \frac{1}{J}\sum_{j=1}^{J}\left[\hat{Y}_i(j) - \tilde{Y}_i^{HB}\right]^2, \end{aligned} \tag{3.24}$$

where $(\theta(1), \theta(2), \ldots, \theta(J))$ is the Gibbs sampling set, $\hat{Y}_i(j)$ is $\tilde{Y}_i^{H}$ in (3.21) with $\theta(j)$, and $g_{1i}(\theta(j))$ and $g_{2i}(\theta(j))$ are given in (3.7) and (3.8), respectively, with $\theta(j)$.

### 3.4 Jackknife MSE Estimate for the EBLUP Estimator

Jiang and Lahiri (1999) proposed the jackknife MSE estimate for the EBLUP estimator. The advantage of the jackknife method and other replication/resampling method is that we do not need to derive tediously a closed form formula. Jackknife

replicates are created by deleting one small area at a time and $m$ replicates are created each corresponding to $m$ small areas. With the full sample and $m$ replicates, the model parameters, $\sigma_b^2$ and $\beta$, are estimated using an appropriate method (e.g., the fitting the constant method, the method of moments, ML, etc.) . Here again $\sigma_{ei}^2$ is assumed to be known.

Step 1. Calculate $\hat{\sigma}_b^2$ and $\hat{\beta}$ with the full sample, and $\hat{\sigma}_b^2(k)$ and $\hat{\beta}(k)$ deleting $k$-th area data, $k = 1, 2, \ldots, m$;

Step 2. Let $\hat{\gamma}_i = z_i^2\hat{\sigma}_b^2(z_i^2\hat{\sigma}_b^2 + \sigma_{ei}^2)^{-1}$, $\hat{\gamma}_i(k) = \hat{\sigma}_b^2(k)z_i^2(\hat{\sigma}_b^2(k)z_i^2 + \sigma_{ei}^2)^{-1}$. Calculate $\hat{Y}_i^H = \hat{\gamma}_i\hat{Y}_i + (1-\hat{\gamma}_i)\mathbf{x}_i^T\hat{\beta}$ and $\hat{Y}_i^H(k) = \hat{\gamma}_i(k)\hat{Y}_i + (1-\hat{\gamma}_i(k))\mathbf{x}_i^T\hat{\beta}(k)$;

Step 3. Calculate $m_{J1i} = g_{1i}(\hat{\sigma}_b^2) - (m-1)\sum_{i=1}^{m}\left[g_{1i}(\hat{\sigma}_b^2(k)) - g_{1i}(\hat{\sigma}_b^2)\right]^2/m$.

Step 4. Calculate $m_{J2i} = (m-1)\sum_{i=1}^{m}\left[\hat{Y}_i^H(k) - \hat{Y}_i^H\right]^2/m$.

Step 5. Let the jackknife MSE estimate be $\mathrm{mse}_J(\hat{Y}_i^H) = m_{J1i} + m_{J2i}$.

### 3.4 Bootstrap Method

Recently, Pfefferman and Tiller (2001) proposed a bootstrap MSE estimation method for a state space time series model, which is also applicable to the small area estimation. In their paper they particularly dealt with the problem of MSE estimation using the bootstrap method, which proceeds as follows:

Step 1. Generate a large number ($B$) of bootstrap samples $\left\{\hat{Y}_i^{(b)} \mid i = 1,2,\ldots,m\right.$ for $b = 1, 2, \ldots, B$ from the model (2.6) with an estimated unknown model parameters $\hat{\theta}$;

Step 2. For each bootstrap sample, re-estimate the unknown model parameters using the same method used for $\hat{\theta}$ with the original sample. Let the $b$-the bootstrap sample parameter estimate be $\hat{\theta}^{(b)}$;

Step 3. Calculate an EBLUP estimator $\hat{Y}_i^H(b)$ using the $b$-th set of estimated model parameters from Step 2 and then calculate

$$m_{2i}^B = \frac{1}{B}\sum_{b=1}^{B}\left[\hat{Y}_i^H(b) - \hat{Y}_i^H\right]^2$$
$$\overline{m}_{1i} = \frac{1}{B}\sum_{b=1}^{B} m_{1i}(\hat{\theta}^{(b)}) \qquad (3.25)$$

where $m_{1i}(\hat{\theta}^{(b)})$ is the naïve MSE estimator obtained from (3.6) with the replacement of $\theta$ by $\hat{\theta}^{(b)}$.

Step 4. Estimate the MSE of the EBLUP estimator by

$$\text{mse}_B = m_{1i}(\hat{\theta}) + m_{2i}^B - \overline{m}_{1i}. \qquad (3.26)$$

The bias of this estimator is $O(m^{-2})$.

### 3.5 When the Sampling Variance Is Unknown

So far, it has been assumed that the sampling variance $\sigma_{ei}^2$ is known but often in reality it is not known. One approach to handle the problem is to model the sampling variance as done by Arora, Lahiri, and Mukherjee (1997) using the gamma distribution. Such approach was also used by the U.S. National Resources Inventory survey (Nusser and Goebel, 1997). However, Wang (2000) found that the modeling approach is not always satisfactory and proposed an EBLUP estimator with directly estimated sampling error and a corresponding MSE estimator under model (2.6) with normal errors.

He proposed estimators for the unknown model parameters, which are different from usual estimators, except the sampling error, which is estimated individually and directly from the sample. However, this is possible only when the small area sample sizes are not too small. Without loss of generality it is assumed that $z_i = 1$. He further assumed that

$$d_i \frac{\hat{\sigma}_{ei}^2}{\sigma_{ei}^2} \sim \chi_{d_i}^2, \tag{3.27}$$

for some degrees of freedom $d_i$. Other model parameters are estimated as follows:

$$\begin{aligned} \breve{\beta} &= \left( \sum_{i=1}^{m} w_{im} \mathbf{x}_i \mathbf{x}_i^T \right)^{-1} \sum_{i=1}^{m} w_{im} \mathbf{x}_i \hat{Y}_i \\ \breve{\sigma}_b^2 &= \max\left( 0, \sum_{i=}^{m} c_{im} \left[ (\hat{Y}_i - \mathbf{x}_i^T \breve{\beta})^2 - \hat{\sigma}_{ei}^2 \right] \right) \end{aligned} \tag{3.28}$$

with suitably chosen weights $w_{im}$'s and positive constant $c_{im}$'s such that $\sum_{i=1}^{m} c_{im} = 1$. Note that $\breve{\sigma}_b^2$ is a truncated estimator, and thus, it is positively biased. With these estimated model parameters, the EBLUP estimator is defined in the usual way, namely,

$$\breve{Y}_i^{H} = \breve{\gamma}_i \hat{Y}_i + (1 - \breve{\gamma}_i) \mathbf{x}_i^T \breve{\beta} \tag{3.29}$$

with

$$\breve{\gamma}_i = \breve{\sigma}_b^2 / (\breve{\sigma}_b^2 + \hat{\sigma}_{ei}^2). \tag{3.30}$$

He derived a large sample MSE expression and its estimator. The approximate MSE is given by

$$\begin{aligned}\mathrm{MSE}(\bar{Y}_i^H - Y_i) &= E(\bar{Y}_i^H - Y_i)^2 \\ &= \gamma_i \sigma_{ei}^2 + (1-\gamma_i)^2 \mathbf{x}_i^T V(\tilde{\beta}) \mathbf{x}_i + (\sigma_b^2 + \sigma_{ei}^2) V(\tilde{\gamma}_i) \\ &= g_{1i} + g_{2i} + g_{3i}\end{aligned} \tag{3.31}$$

where

$$\begin{aligned}V(\tilde{\beta}) &= \left(\sum_{i=1}^{m} w_{im} \mathbf{x}_i \mathbf{x}_i^T\right)^{-1} \sum_{i=1}^{m} w_{im}^2 (\sigma_b^2 + \sigma_{ei}^2) \mathbf{x}_i \mathbf{x}_i^T \left(\sum_{i\ 1}^{m} w_{im} \mathbf{x}_i \mathbf{x}_i^T\right)^{-1} \\ V(\tilde{\gamma}_i) &= (\sigma_b^2 + \sigma_{ei}^2)^{-4} \sigma_{ei}^4 \left[2\sigma_b^2 / d_i + V(\tilde{\sigma}_b^2)\right], \\ V(\tilde{\sigma}_b^2) &= 2\sum_{i=1}^{m} c_{im}^2 (\sigma_b^2 + \sigma_{ei}^2)^2.\end{aligned} \tag{3.32}$$

For large $\sigma_b^2$, the naive estimator $\tilde{\gamma}_i \hat{\sigma}_{ei}^2$ for $g_{1i} = \gamma_i \sigma_b^2$ underestimates and Taylor correction gives

$$\hat{g}_{1i} = \tilde{\gamma}_i \hat{\sigma}_{ei}^2 + (\tilde{\sigma}_b^2 + \hat{\sigma}_{ei}^2) \hat{V}(\tilde{\gamma}_i). \tag{3.33}$$

The second term $g_{2i}$ in (3.32) is estimated by simply substituting unknown parameters by their estimates as given by

$$\hat{g}_{2i} = (1-\tilde{\gamma}_i)^2 \mathbf{x}_i^T \hat{V}(\tilde{\beta}) \mathbf{x}_i \tag{3.34}$$

and similarly the third term by

$$\hat{g}_{3i} = (\tilde{\sigma}_b^2 + \hat{\sigma}_{ei}^2) \hat{V}(\tilde{\gamma}_i). \tag{3.35}$$

The resulting MSE estimator is then given by

$$\begin{aligned}\mathrm{mse}_1(\tilde{Y}_i^H - Y_i) &= \hat{g}_{1i} + \hat{g}_{2i} + \hat{g}_{3i} \\ &= \tilde{\gamma}_i\hat{\sigma}_{ei}^2 + (1-\tilde{\gamma}_i)^2\mathbf{x}_i^T\hat{V}(\tilde{\beta})\mathbf{x}_i + 2(\tilde{\sigma}_b^2 + \hat{\sigma}_{ei}^2)\hat{V}(\tilde{\gamma}_i).\end{aligned} \tag{3.36}$$

The $V(\tilde{\sigma}_b^2)$ that appears in $V(\tilde{\gamma}_i)$ and given in (3.32) is better estimated by

$$\hat{V}(\tilde{\sigma}_b^2) = \sum_{i=1}^{m} c_{im}^2\left[(\hat{e}_i^2 - \sigma_{ei}^2) - \sigma_b^2\right]^2, \quad \hat{e}_i = \frac{m}{m-k}\left(\hat{Y}_i - \mathbf{x}_i^T\tilde{\beta}\right)^2, \tag{3.37}$$

than the one with direct substitution of estimated parameters. The drawback of the MSE estimator given in (3.36) is that it overestimates sometimes severely when $\sigma_b^2$ is small because the first and the third terms ($g_{1i}$ and $g_{3i}$) are overestimated. Wang (2000) provided an improved MSE estimator by using better estimators for these terms. He used the following estimator for $g_{1i}$:

$$\tilde{g}_{1i} = \frac{(\tilde{\sigma}_b^2 + \hat{\sigma}_{ei}^2)\tilde{\sigma}_b^2\hat{\sigma}_{ei}^2 + \tilde{\sigma}_{ei}^2\hat{V}^*(\tilde{\sigma}_b^2) + \tilde{\sigma}_b^2\hat{V}(\hat{\sigma}_{ei}^2)}{(\tilde{\sigma}_b^2 + \hat{\sigma}_{ei}^2)^2 + \hat{V}^*(\tilde{\sigma}_b^2) + \hat{V}(\hat{\sigma}_{ei}^2)}. \tag{3.38}$$

Noting that $E\left[(\tilde{\gamma}_i - \gamma_i)^2(b_i + e_i)^2\right] \approx (\sigma_b^2 + \sigma_{ei}^2)V(\tilde{\gamma}_i)$, he obtained a better estimator for the third term by a numerical integration $\tilde{g}_{3i}$ that approximates $E\left[(\tilde{\gamma}_i - \gamma_i)^2(b_i + e_i)^2 \mid \sigma_b^2, \sigma_{ei}^2\right]$. Then the improved MSE estimator is given by the sum of these terms as follows:

$$\mathrm{mse}_2(\tilde{Y}_i^H - Y_i) = \tilde{g}_{1i} + \hat{g}_{2i} + \tilde{g}_{3i}. \tag{3.39}$$

He showed by simulation that the improved estimator is superior to the MSE estimator given in (3.36) especially when $\sigma_b^2$ is small.

## 4. Concluding Remarks

Small area estimation is a growing area of research. In this paper, we briefly reviewed some small area estimators that are formulated under the random effect area level model with emphasis on the estimation of mean squared error of a small area estimator. Although mentioned briefly at the beginning, we did not cover small area estimation methods based element level models. This review is far from complete and is no way to compete with excellent review papers by Ghosh and Rao (1994) and Rao (1999). It is simply intended to provide some forum for discussion and to bring out some newest developments such as Pfefferman and Tiller (2001) and Wang (2000).

Wang (2000) provides an EBLUP estimator and its MSE estimator, which can be used for the most general setting in the sense that all three model parameters, $\beta$, $\sigma_b^2$, and $\sigma_{ei}^2$ are unknown. Wang's proposed improved MSE estimator is a closed form formula, although fairly complex, and shown to be superior to the less complex one he also derived, especially when $\sigma_b^2$ is small relatively to $\sigma_{ei}^2$. The overestimation problem under this situation is common to all other MSE estimators for the EBLUP estimators and Wang's estimator provides a significant improvement even though it still somewhat overestimates. He also provides a unifying theory for benchmarking of the small area estimates to the large area direct estimate, which is sometimes required and/or helpful to reduce the bias problem of small area estimates.

Replication and resampling methods for the MSE estimation are also promising techniques. The jackknife technique has proved to be a really versatile tool for many statistical problems and the MSE estimation problem for small area estimates is not an exception as shown by Jiang and Lahiri (1999). As for another resampling method, there are several bootstrap estimators proposed, among which the one by Pfefferman and Tiller (2001) is the newest and seems the most promising since it has a lower order of bias than other proposals. Although computer intensive, these methods are general and can be implemented easily in a computer system, which can then be routinely used. However,

more research is needed to understand their properties and performance in various estimation environments.

**References**


Arora, V., Lahiri, P., and Mukherjee, K. (1997). Empirical Bayes estimation of finite population means from complex survey. *Journal of American Statistical Association*, 92, 1555-1562.

Battese, G.F., Harter, R.M., and Fuller, W.A. (1988). An error component model for rpediction of county crop area using survey and satellite data. *Journal of American Statistical Association*, 83, 28-36.

Drew, D., Singh, M.P., and Choudhry, G.H. (1982). Evaluation of small area estimation techniques for the Canadian Labour Force Survey, *Survey Methodology*, 8, 17-47.

Fay, R.E., and Harriot, R.A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of American Statistical Association*, 74, 269-277.

Gelfand, A., and Smith, A.F.M. (1990). Sampling based approaches to calculating marginal densities. *Journal of American Statistical Association*, 85, 398-409.

Ghosh, M. (1992). Hierarchical and empirical Bayes multivariate estimation. In *Current Issues in Statistical Inference: Essays in Honor of D. Basu*, eds. M. Ghosh and P.K. Pathak, pp. 151-177. Hayward, CA: IMS.

Henderson, C.R. (1950). Estimation of genetic parameters (abstract). *Annals of Mathematical Statistics*, 21, 309-310.

Kass, R.E., and Steffey, D. (1989). Approximate Bayesian inference in conditionally independent hierarchical modles (parametric empirical Bayes models). *Journal of American Statistical Association*, 84, 717-726.

Harville, D.A. (1991). Comment on, "That BLUP is a good thing: the estimation of random effects" by G.K. Robinson. *Statistical Science*, 6, 35-39.

Kackar, R.N., and Harville, D.A. (1984). Approximations for standard errors of estimators for fixed and random effects in mixed models. *Journal of American Statistical Association*, 79, 853-862.

Lahiri, P., and Rao, J.N.K. (1995). Robust estimation of mean square error of small area estimators. *Journal of American Statistical Association*, 90, 758-766.

Laird, N.M., and Louis, T.A. (1987). Empirical Bayes confidence intervals based on bootstrap samples. *Journal of American Statistical Association*, 82, 739-750.

Marker, D.A. (1999). Organization of small area estimators using a generalized linear regression framework. *Journal of Official Statistics*, 15, 1-24.

Morris, C. (1983). Parametric empirical Bayes inference: Theory and applications (with discussion). *Journal of American Statistical Association*, 78, 47-65.

Nusser, S.M., and Geobel, J.J. (1997). The National Resources Inventory: a long-term multi-resource monitoring programme. *Environmental and Ecological Statistics*, 2 181-204.

Pfeffermann, D., and Tiller, R. (2001). Bootstrap approximation to prediction MSE for state-space models with estimated parameters. Technical report, U.S. Bureau of Labor Statistics.

Prasad, N.G.N., and Rao, J.N.K. (1990). The estimation of mean squared errors of small area estimators. *Journal of American Statistical Association*, 85, 163-171.

Prucell, N.J., and Kish, L. (1979). Postcensal estimates for local areas (or domain). *International Statistical Review*, 48, 3-18.

Robinson, G.K. (1991). That BLUP is a good thing: the estimation of random effects (with discussion). *Statistical Science*, 6, 15-51.

Rao, J.N.K. (1999). Some Recent Advances in Model-Based Small Area Estimation. *Survey Methodology*, 25, 175-186.

Rao, J.N.K. (2001). Measuring Uncertainty of Small Area Estimation. Presented at the Small Area Conference held in Potomac, Maryland.

Särndal, C.E., and Hidiroglou, M.A. (1989). Small domain estimation: a conditional analysis. *Journal of American Statistical Association*, 84, 266-275.

Schaible, W. L. (1978). Choosing weights for composite estimators for small area statistics. In *Proceedings of the Survey Research Methods Section*, pp.741-746. Washington, DC: American Statistical Association.

Singh, M.P., Gambino, J., and Matel, H. (1992). Issues and options in the provision of small area data. In *Small Area Statistics and Survey Designs* (G. Kalton, J. Kordos, and R. Platek, eds.), Vol.1, pp.37-75. Central Statistical Office, Warsaw.

Wang, J. (2000). Topics in small area estimation with applications to the National Resources Inventory. Ph. D. Dissertation, Iowa State University.