ASSIGNING PERMANENT RANDOM NUMBERS TO THE BUREAU OF LABOR STATISTICS LONGITUDINAL (UNIVERSE) DATA BASE

Shail Butani, Kenneth W. Robertson, Kirk Mueller
Shail Butani, BLS, 2 Massachusetts Ave., NE Room 4985, Washington, DC 20212-0001

KEY WORDS: Sampling Frame, Sampling Units, Permanent Random Numbers, Collocated Permanent Random Numbers

I. Introduction

The U.S. Bureau of Labor Statistics (BLS) routinely conducts several large-scale establishment surveys in a Federal/State cooperative environment. The major ones are: the Current Employment Statistics (CES) Survey, commonly known as the monthly payroll survey; the Occupational Employment Statistics (OES) Survey; and the Occupational Safety and Health (OSH) Survey. At present, the CES and OES Surveys are going through major redesigns. A major feature of the redesigns calls for centralized processing of the sample allocation and selection in order to achieve efficiency and to spread the response burden more evenly among the businesses; the OSH Survey already has centralized processing. Additionally, the BLS is in the process of designing a longitudinal database (LDB) for its Business Establishment List (BEL). These opportunities presented an ideal time to begin the use of permanent random numbers (PRN) at BLS.

We first describe the BEL that serves as a sampling frame for BLS' business surveys. Next, we discuss PRNs and collocated PRNs. An outline of some of the complexities of assigning PRNs when surveys have different designs with respect to sampling units, stratification, allocation and selection cells is also given in this section. In Section IV, we give a brief summary of the sample designs for the CES, OES, and OSH Surveys. The methodology of assigning PRNs and collocated PRNs to the initial frame is discussed in Section V. This is followed by a discussion of assigning collocated PRNs to new businesses in a manner that will ensure proper representation for ongoing surveys. In Section VI, we present some data as to how well the collocated PRNs technique is working. Finally, we conclude with a brief summary.

II. Business Establishment List (BEL)

The primary source of the BLS' BEL is the quarterly contributions reports filed by employers for each unemployment insurance (U.I.) account with their state unemployment insurance agency. The data for both private and public sector workers are delivered to BLS after they go through several stages of refinement by the employment security agencies of 50 states and the District of Columbia as part of the Covered Employment and Wages, or ES-202, Program. For the purposes of this paper, geography is restricted to these jurisdictions. Employment covered under the U.I. laws provides a virtual census (98 percent) of employees on nonfarm payrolls. This rich and comprehensive database has about seven million records. Among other data elements, each record has a Standard Industrial Classification (SIC) code, a state code, a county code that can be mapped into a Metropolitan Statistical Area (MSA), employment for each month of the quarter, and total quarterly wages. A record usually represents the worksite or establishment level for each U.I. account.

BEL is also currently going through a major revision. Among other things, this revision involves improved methodology for linking records both within and across quarters. This allows us to better track data on job creations, destruction, and wages, etc. at the establishment as well as at the U.I. account level over time. The new frame called the longitudinal database (LDB) will for each multi-establishment employer have a record at the U.I. account level as well as a record associated with each worksite. (The current universe database has records only at the worksite level.)

III. Permanent Random Numbers

Objective--Primary purposes for using permanent random numbers (PRNs) are: (1) to achieve the amount of sample overlap desired within a survey from one time period to another; and (2) to minimize the overlap in samples between different surveys. Because the business population is dynamic, it is also important to update the sample in order to take into account changes to the frame due to births (new businesses), deaths (out-of-businesses), mergers, acquisitions, and changes in employment size, geography, and industry of a unit.

PRN Methodology—A detailed description on various variations of the PRNs methodology is given in Ohlsson (1995). In this paper, we describe a simplistic version of PRNs and collocated PRNs. In its very basic form, we assign PRNs that are uniformly distributed over the interval [0, 1) to every

record on the frame; we carry out the PRN to 12 decimal places in order to minimize the number of ties that could occur in seven million records. Let X_i denote the random number for unit i in a stratum; next in each selection stratum, we order the population in ascending order of the X_i's. Then, selecting the first n units on the list should constitute the desired simple random sample without replacement.

Births are assigned new PRNs independent of existing units and PRNs; deaths are removed from the frame together with their PRNs; and units that move from one stratum to another due to changes in MSA, SIC, size, etc. ("strata jumpers") keep their previously assigned PRNs. (It is worth noting that survey data should not be used to update the frame as it may lead to a biased sample.) Thus, we are drawing the sample for the next survey round from an up-to-date frame.

At BLS, the use of PRNs originated with the CES redesign. In the various sample design simulations that were conducted, we used the PRNs methodology to: update the probabilities of selection; include new U.I. accounts (business births); remove old U.I. accounts (deaths); handle units that have changed size class, MSA, or SIC (strata jumpers); and perform sample rotation. The main reasons for using PRNs in CES are to achieve the desired sample overlap from previous survey year to the next and to properly represent births in the sample.

Overlap between surveys--Let us assume that the CES, OES, and OSH Surveys have essentially the same sample designs and sample sizes. Let us further assume that we want to minimize the overlap between the three surveys. Then, we could divide the interval [0,1) into three equal parts, say [0.00, 0.33), [0.33, 0.66), and [0.66, 1.00), and take the first n units in each interval.

Overlap within a survey--In the above example, suppose we desire to rotate one-third of the sample in each of the three surveys. To achieve this, in each sampling stratum, we select the $\{(n/3) + 1\}$th through the $(4n/3)$th units in each of the three intervals in [0, 1); where, n is the desired sample size for that stratum. It should be noted that the realized overlap in the sample is approximately equal to the desired one. This is because births and deaths rarely offset each other exactly and because of "strata jumpers".

Complications in use of PRNs--In the real world, things are rarely as simple as in the above example. In our case, the sampling unit is not the same for the three surveys; the allocation and selection cells vary across the surveys; the amount of overlap desired from one period to the next within each survey also differs.

The major issue with the assignment of PRNs pertains to over-stratification leading to small numbers of population units in most allocation and selection cells. Although there are about seven million establishments on the frame, allocation and selection cells became thin when one considers stratifying by geography (state or MSA), industry (major industry division, 2- or 3-digit SIC), and employment size class. If the number of units were large, then simply assigning permanent random numbers would suffice. Another complexity involved proper representation of births. Births tend to be small relative to the continuous population in the stratum. To alleviate the problems, collocated permanent numbers are used to distribute the units evenly on the interval [0,1). Collocation, however, raises many other issues. Do we collocate at the U.I. account or at the worksite level; what level of geography, industry to use; which definition of size class? These and other complexities are discussed in the section following the sample design.

Collocated PRN Methodology-- Collocated PRNs, CX_i's, are computed as $\{(R_i - \varepsilon)/N\}$, where, R_i is the rank of X_i in the sorted list for the stratum, N is the number of units in the frame for that stratum, and ε is a random number that is uniformly distributed on the interval [0,1]. Note that the same value of ε is used for every unit in that stratum. With this transformation, the CX_i's are evenly spaced in [0,1). As with PRNs, ε and collocated PRNs were carried out to 12 decimal places.

Again, let us suppose that there are only three units on the frame in a selection cell and we desire to take one unit for each of the three surveys. If the PRNs are not collocated, then there is a 1/27 chance that all three units fall in the interval [0.00, 0.33). Hence, the first unit in this interval will be selected for all three surveys since the second and third intervals have no units (sequential selection wraps-around). Similarly, there is a 1/27 likelihood that all three units are in the second or third interval; overall, the probability is 3/27 that any one of the three unit will be selected in all three surveys. Collocation of PRNs ensures that three different units will be selected for the three surveys. In a similar manner, collocation ensures that each quarter's births are evenly dispersed within each stratum. In the interval [0,1) collocation of PRNs for

birth units is performed independently of the continuous units.

IV. Sample Designs for CES, OES, and OSH Surveys

CES Survey Redesign--The CES is a monthly survey designed to produce estimates of employment, payroll, and hours worked by various levels of industrial details for the Nation, 50 States plus D.C., and MSAs. The primary advantage of the CES estimates is their timeliness. Because the primary purpose of CES is to measure the monthly change, both the current and the new samples are designed to have over 90 percent sample overlap between two consecutive months. A detailed description of the redesign with respect to goals, sample design parameters, features, and characteristics are described in Butani, Stamas and Brick (1997).

OES Survey Redesign--The OES is an annual survey; the redesign calls for the survey to measure employment and wages for over 750 occupations by various levels of industrial activity for the Nation, 50 States plus D.C. and for over 350 Metropolitan areas. A detailed description of the OES sample redesign is given in an internal document (BLS, 1998).

OSH Survey--Like the OES, the OSH Survey is an annual survey; it is designed to estimate the number and frequency of work-related injuries and illnesses by detailed industry for the Nation and for Sates participating in the survey. For a detailed description, see the BLS Handbook of methods.

Summary of the three designs--The important features of the three sample designs are summarized below. Note: Some surveys include Puerto Rico, Virgin Islands, and other U.S. Territories. For the purposes of this paper, we will concentrate only on fifty States plus the District of Columbia.

Sample Design for CES, OES, and OSH Surveys

Feature	CES-R	OES-R	OSH
Frequency	Monthly	Annual	Annual
Sampling Unit	U.I. account	Worksite	Worksite
Sample Size	250,000 U.I. accounts or 650,000 worksites	400,000 worksites	200,000 worksites
Allocation Cells	ST/Major Industry Divisions/8 Size Classes	MSA/3-digit SIC/7 Size Classes	ST/Varying Industry/ 5 Size Classes
Selection Cells	Major Industry Divisions/8 Size Classes/MSA within a State 11 X 8 X 350	MSA/3-digit SIC/ 7 Size Classes 350 X 377 X 7	ST/Varying Industry/ 5 Size Classes 51 X 100 –300 X 5
Desired Annual Overlap	Over 90%	Zero over 3 years	Independent Samples
Treatment of Births	Quarterly Update	Annual	Annual
Frame Maintenance	Annual	Annual	Annual

Employment Size Class Definitions

CES-R	OES-R	OSH
1-9	1-4 (Beginning with 98 Survey)	
10-19	5-9	1-10
20-49	10-19	11-49
50-99	20-49	
100-249	50-99	
250-499	100-249	50-249
500-999	250 + (Certainty)	250-999
1000+ (Certainty)		1000 +

V. Assignment of PRNs and Collocated PRNs

Initial Assignment of Collocated PRNs--We begin by assigning collocated PRNs to each worksite (7 million records) on the 1995 second quarter frame, and collocating PRNs within the MSA/3-digit SIC/OES size class strata. For initialization purposes, we chose OES selection cells because OES has the largest sample size, rotates the fastest, the industry level is fixed and is at a low level (3-digit SIC).

To handle the CES survey, we aggregated all worksites belonging to a multi-establishment employer to their U.I. account level. These multi-establishment U.I. accounts were then assigned a PRN and were collocated separately from the single establishment U.I. accounts at the MSA/3-digit SIC/OES size class level. In the CES sample, this procedure was beneficial in distributing the multi-establishment and single establishment accounts according to their proportion in the population.

Assignment of Collocated PRNs to Birth Units--Birth units are most important to the CES Survey because of the contribution they make to over-the-year change in employment. For this reason, with each quarterly update of the frame, the new units are first assigned PRNs at the U.I. account level and collocated at the ST/Major Industry Division/ 8 CES size classes level (CES allocation cells). Generally, there are few birth units each quarter and the allocation and selection cell levels are the same for the supplemental sample of births each quarter. Additionally, the vast majority of birth units are single establishments; thus, the worksite and U.I. account levels are the same for these units. PRNs are also assigned to the worksites for the occasional birth units that are multi-establishment accounts and to the expansion worksites for multi-establishment accounts; these PRNs are also collocated separately at the CES allocation cell level.

Starting Points for Each Survey-- Another issue is what should be the starting points for each survey? Given the above sample designs, there is really no clean way to determine the optimal starting points for the three surveys. The determination of starting points is complicated since the sampling fractions vary by allocation cells and the allocation cells themselves are different for each survey. Based on the rotation period for each survey and on several approximations including sampling fractions, the proposed starting points for each survey and size class as determined by the Bureau's Office of Survey Methods Research (BLS, 1996) are given in the table below.

Starting Points

Size Class	CES-R	OSH	OES-R
1-9	0	0.20	0.25
10-49	0	0.45	0.55
50-249	0	0.40	0.50
250-999	0	0.66	0
1000+	0	0	0

VI. Distribution of Universe and Samples

At present, CES and OES have begun using collocated PRNs in selecting the samples; OSH is expected to start using them with the implementation of the LDB in 1999. The universe and sample distribution for CES pertaining to multi-establishment accounts and birth units are summarized in the first two tables below, while the overlap between the OES and CES samples is given in Table 3.

Table 1. CES Universe and Sample Distributions--1st Quarter 1997 Frame

Distribution of UI Accounts, by Size and UI Type,

CES - National - Wholesale Trade

		Multis	Singles	Total UIs
Size 1	% Popl'n	0.099	99.901	415,707
Employment 1-9	% Sample	0.067	99.933	5,997
Size 2	% Popl'n	0.922	99.078	74,376
Employment 10-19	% Sample	0.925	93.075	2,270
Size 3	% Popl'n	6.622	93.378	49,499
Employment 20-49	% Sample	6.875	93.125	2,720
Size 4	% Popl'n	21.893	78.107	15,320
Employment 50-99	% Sample	23.050	76.950	1,436
Size 5	% Popl'n	37.078	62.921	7,530
Employment 100-249	% Sample	36.227	63.773	1,129
Size 6	% Popl'n	51.671	48.329	1,736
Employment 250-499	% Sample	48.134	51.876	453
Size 7	% Popl'n	65.411	34.589	584
Employment 500-999	% Sample	64.706	35.294	204
Size 8	% Popl'n	80.258	19.742	233
Employment 1000 +	% Sample	80.258	19.742	233

Table 2. CES Universe and Sample Distributions–1st Quarter 1997 Frame

Distribution of UI accounts, by Size and Age, CES

		Original PRNDATE	96-2 PRNDATE	96-3 PRNDATE	96-4 PRNDATE	97-1 PRNDATE	Total
Size	% Popl'n	86.42	3.18	3.14	3.39	3.87	3,897,244
1-9	% Sample	86.43	3.23	3.21	3.34	3.80	78,818
Size	% Popl'n	94.05	1.72	1.56	1.35	1.31	773,131
10-19	% Sample	94.28	1.67	1.50	1.30	1.26	31,083
Size	% Popl'n	95.26	1.38	1.25	1.08	1.04	507,497
20-49	% Sample	95.23	1.42	1.30	1.00	1.05	38,172
Size	% Popl'n	96.15	1.09	1.01	0.82	0.93	175,164
50-99	% Sample	96.14	1.09	1.02	0.73	1.02	23,703
Size	% Popl'n	96.93	0.86	0.82	0.61	0.78	102,828
100-249	% Sample	96.83	0.90	0.81	0.64	0.82	24,023
Size	% Popl'n	97.75	0.59	0.62	0.38	0.66	31,418
250-249	% Sample	97.64	0.61	0.67	0.42	0.67	12,137
Size	% Popl'n	98.04	0.46	0.39	0.39	0.71	14,519
500-999	% Sample	97.71	0.51	0.46	0.46	0.86	8,014
Size	% Popl'n	98.45	0.40	0.35	0.18	0.62	10,646
1000+	% Sample	98.45	0.40	0.35	0.18	0.62	10,646

Table 3. Overlap between OES and CES sample

National

		Size 1	Size 2	Size 3	Size 4	Size 5	Size 6	Size 7	Size 8	Total
Mining	% overlap	26.51	37.56	45.44	55.78	64.37	74.48	89.47	100.00	42.04
	OES sample	2674	1898	2207	1011	640	243	95	33	8801
Construction	% overlap	5.70	5.43	10.57	22.21	38.64	58.11	86.25	100.00	11.33
	OES sample	29666	20042	21824	10285	4963	776	211	30	87797
Dur/Non/Durable	% overlap	6.95	9.51	12.04	20.00	32.96	56.05	76.48	100.00	19.07
	OES sample	37462	29533	37290	24431	22098	8601	3615	1794	164824
TPU	% overlap	11.57	17.63	25.24	36.59	50.63	66.63	80.71	100.00	26.19
	OES sample	18967	14137	17150	9535	6642	1876	783	477	69567
Wholesale	% overlap	5.05	9.21	13.35	20.46	30.87	50.38	71.72	100.00	12.49
	OES sample	31722	25566	29678	12821	6368	1298	336	74	107863
Fire	% overlap	16.31	28.22	30.58	36.30	44.96	60.32	75.59	100.00	27.79
	OES sample	29703	17111	16262	8475	5304	1681	799	464	79799
Services	% overlap	6.90	10.98	16.04	25.47	40.72	61.17	79.81	100.00	17.89
	OES sample	106749	74578	74632	37317	30058	9546	3934	2499	339313
Total	% overlap	8.99	13.63	19.17	29.12	43.94	62.13	78.61	100.00	20.27

VII. Summary

From the above data, it appears that the collocated PRNs technique is working reasonably well in achieving the desired overlap within a survey while maintaining a representative sample selected from an up-to-date frame. It is also performing reasonably well in reducing overlap between the two surveys.

In the future, the BLS would evaluate the use of U.I. account as the sampling unit for OES and OSH Surveys. This should somewhat simplify the collocation process and coordination between surveys.

Any opinions expressed in this paper are those of the authors and do not constitute policy of the Bureau of Labor Statistics.

References

1. Butani, S., G. Stamas and M. Brick (1997), "Sample Redesign for the Current Employment Statistics Survey", Proceedings of the Section on Survey Research Methods, American Statistical Association.

2. Bureau of Labor Statistics, Office of Survey Methods Research, 1996. "Collocated Random Number Assignment". BLS internal report, December.

3. Bureau of Labor Statistics (1997), Handbook of Methods, Washington, D.C., U.S. Department of Labor, Bulletin 2490, pp. 70-79.

4. Bureau of Labor Statistics, Statistical Methods Division, 1998. "Survey Method and Reliability Statement for the 1996 OES Survey". BLS internal report, March.

5. Ohlsson, E. (1995), "Coordination of Samples Using Permanent Random Numbers", In B.G. Cox et al., Business Survey Methods, New York: Wiley, pp. 153-183.