The Philippine Statistician, 2002
Vol. 51, Nos. 3-4, pp. 11-22

Coverage Error Problems in Establishment Surveys in the Philippines[1]

Arturo Y. Pacificador, Jr.[2]

1. INTRODUCTION

Sample surveys are important sources of data for various purposes. The usefulness of data from such sources are often indicated by its timeliness, relevance and quality - each of which are equally important. Quality in this context is directly proportional to the errors committed in the conduct of surveys. It is well accepted that surveys (including censuses) are not free of any error which in statistical terms is simply defined as the difference between the estimate and the true but unknown value of the parameter. Survey errors are generally classified as sampling (SE) and non-sampling errors (NSE). Sampling error is usually associated with the sampling process and its magnitude is measured in terms of the standard error. It is well known that this error is inversely proportional to the sample size. On the other hand, nonsampling errors arise in all the other aspects of the survey operation from conceptualization to processing and dissemination of results. Unlike SE, NSE is more difficult to measure and usually increases along with increase in the sample size. It is believed, that if NSE is not properly controlled, this type of error will dominate the Total Survey Error and could compromise the usefulness of survey results. Nonsampling errors are usually classified into coverage, non-response, measurement, and processing errors. Onate (1988) developed a conceptual framework on the sources and type of NSE in field surveys. The framework is presented in Figure 1.

This paper focuses on coverage errors and its effect on survey estimates. In particular, some references will be made with regards to such type of errors occurring in establishment surveys in the Philippines.

[1] Paper presented at the Burton T. Onate Memorial Research Conference, August 16, 2002, UP Los Baños
[2] Professor of Statistics and former student of Dr. Oñate

TIME	UNIVERSE	STAGES OF SURVEY OPERATION	SOURCES AND TYPES OF NSE	PARAMETERS
T_1	Target (U_T)	Conceptualization		$Y_T, \bar{Y}_T, N_T, S_T^2$
T_2	Frame (U_F)	Construction/ Development of Sampling Frame		$Y_F, \bar{Y}_F, N_F, S_F^2$
T_3	Survey (U_S)	Actual Collection of Data		$Y_S, \bar{Y}_S, N_S, S_S^2$
T_4	Inference (U_I)	EDP Stage/Tabulation / Estimation Procedure		$Y_I, \bar{Y}_I, N_I, S_I^2$

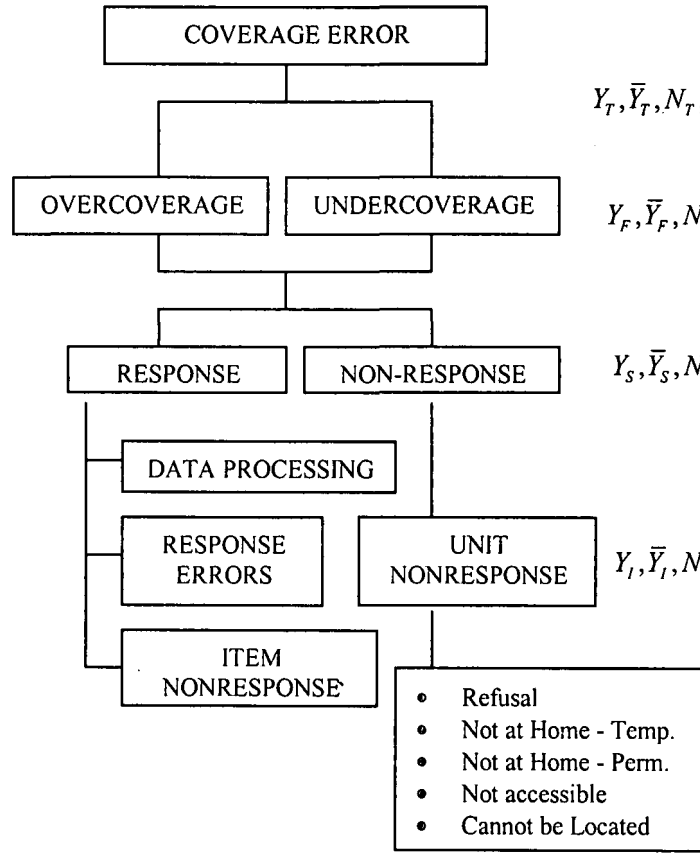

FIGURE 1 – Oñate's Conceptual Framework : Sources and Types of Nonsampling Errors in Field Surveys

2. COVERAGE ERRORS

Coverage error can be defined as the error in the estimate that results from (1) the failure to include in the frame all units belonging to the target universe or the failure to include specified units in the conduct of the survey, and (2) the erroneous inclusion of some units because of a defective or outdated frame or of specified units more than once in the actual survey. Based on this definition, coverage error can be taught of as a result of imperfections by the sampling frame in mimicking the target population. Hansen, et.al. (1964) identified some reasons for the imperfections in the frame which leads to erroneous coverage. These are:

(1) Frames which contains units not in the target universe (overcoverage);
(2) Frames which do not contain some units in the target universe or (undercoverage);
(3) Frames containing erroneous and incomplete entries;
(4) Frames containing duplicates; and,
(5) Insufficient information to locate the units.

These imperfections result in at least one of the following:

(1) For overcoverage, some units not in the target universe will have non-zero probability of being included in the sample;
(2) For undercoverage, some units in the target universe will have zero probability of being included in the sample because frames provide observational access to these units;
(3) For duplications, some units in the target universe will have larger inclusion probabilities than what should be assigned by the sampling design employed in the selection of the sample;
(4) Incomplete data because of the inability to locate the sample unit as a result of incomplete or insufficient information.

Coverage errors may be a serious problem. Groves (1983) and Murthy (1983) observe that this may even be larger in magnitude than incomplete or missing data due to nonresponse. Thus it is important to understand is causes and its effect on survey estimates.

3. COVERAGE ERRORS IN ESTABLISHMENT SURVEYS

Establishments surveys is one of the major source of data from the formal sector covering diverse topics such as production, employment, labor relations, R&D activities, investment initiatives among others. The major establishment in the Philippines are usually conducted by the National Statistics Office (NSO), Bureau of Labor and Employment Statistics (BLES), Bangko Central ng Pilipinas (BSP), Department of Trade and Industry (DTI), Bureau of Agricultural Statistics (BAS). Many of these surveys employ probability sampling with stratification in terms of size measured by Average Total Employment (ATE), industry classification indicated in the Philippine Standard Industry Classification (PSIC) system, and region. Such surveys often cover the larger establishments (with an ATE of at least 10). As a probability sample, it employs the list of establishment from the Census of Establishment (CE) as sampling frame. The CE is one massive statistical operation done

every five years covering all establishments in the Philippines. However, some problems were encountered in the use of this frame such as:

a) Undercoverage – failure to include in the list some establishments in operation, failure to update the CE during intercensal periods thus missing out new establishments (births), incorrect size entries.
b) Overcoverage – duplications, incorrect size entries, establishments that cannot be located.
c) Shifts in economic activity (PSIC), size (ATE), and location which again is a result of the absence of an ideal updating mechanism. Examples of these type of problem is presented in Table 1 from an earlier study by Onate, et. al. (1989).

Table 1.
Percent distribution of establishments in the 1987 NSO Annual Survey of Establishments which changed ATE levels (Total Sample Size = 16,364)

FROM ATE	TO ATE			TOTAL
	Below 10	10-19	20 and over	
10-19	5.9	-	6.4	12.4
20 and over	0.8	6.7	-	7.5
TOTAL	6.7	6.7	6.4	19.9

The results in table 1 indicate provides an indication of coverage errors as a result of changes in ATE totaling close to 20 percent of the overall sample. Again, this demonstrates the problem of using censuses as source of frame.

4. EFFECTS OF COVERAGE ERROR

Generally, bias in estimation is introduced by different types of nonsampling error. Figure 2 present an illustration of coverage error and will be used later on to derive the extent of bias in estimation brought about by coverage errors.

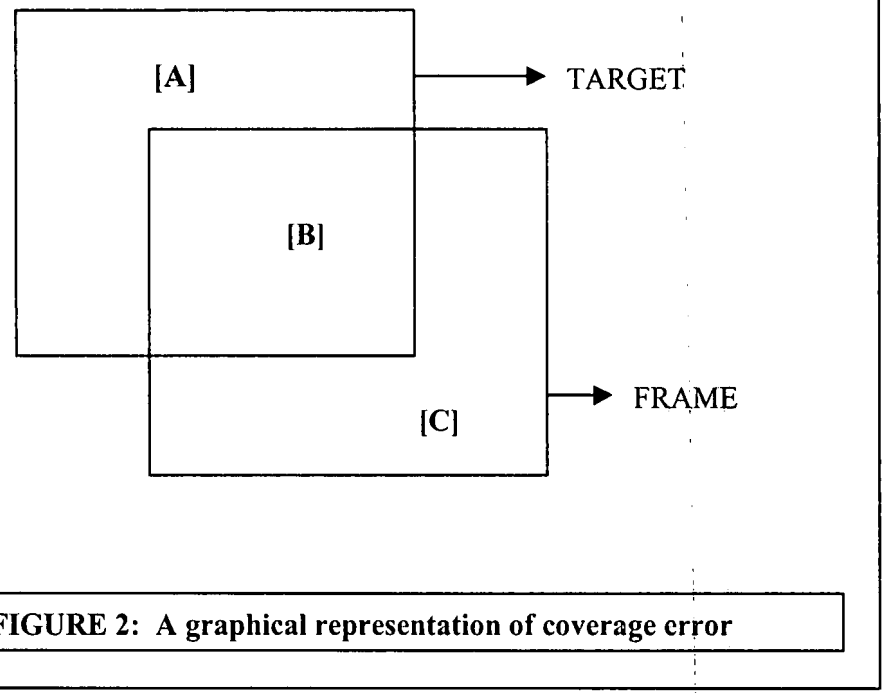

FIGURE 2: A graphical representation of coverage error

Region [A] represents all units in the target universe but were not included in the frame universe. This illustrates the case of undercoverage. In the case of establishment surveys, this may take the form of failure of the frame to account for births or incorrect size entries and therefore totally missed out in the survey. Region [B] consists of all units in both the target and frame universes (correct coverage) while region [C] consist of all units in the frame universe but are not part of the target universe (overcoverage). The figure also implies that under and over coverage can exist simultaneously in any survey.

To illustrate the effects of coverage errors on survey estimates, we assume that the object of estimation is the parameter (say population total) of the target population. We illustrate this for the case of estimating the population totals.

Let

$N_{[A]}$ Total number of units in region [A]. That is, the total number of elements of the target population not included in the sampling frame.

$N_{[B]}$ Total number of units in region [B] - the elements in the sampling frame that belongs to the target universe.

$N_{[C]}$ Total number of elements in the sampling frame but is not part of the target universe.

$N_{[T]}$ Total number of elements in the target universe.

$N_{[F]}$ Total number of elements in the frame universe (sampling frame).

$Y_{[\,]}$ Total of all units in region [ ].

Note that

$$N_{[T]} = N_{[A]} + N_{[B]} \quad (1)$$
$$N_{[F]} = N_{[B]} + N_{[C]} \quad (2)$$
$$Y_{[F]} = Y_{[B]} + Y_{[C]} \quad (3)$$
$$Y_{[F]} = Y_{[B]} + Y_{[C]} \quad (4)$$

Also, $N_{[A]}, N_{[B]}, N_{[C]}, N_{[T]}$ are unknown quantities and so are $Y_{[A]}, Y_{[B]}, Y_{[C]}, Y_{[F]}, Y_{[T]}$. In this case, suppose that we are interested in estimating $Y_{[T]}$.

In practice, one selects a simple random sample of size $n_{[F]}$ from $N_{[F]}$. Define $\hat{Y}_{[F]} = N_{[F]}\bar{y}_{[F]}$ where $\bar{y}_{[F]}$ is the sample mean. With the scenario presented in figure 2, possible strategies can be used in estimating $Y_{[T]}$. These are:

Strategy 1: Assume that there in no overcoverage or treating overcoverage units as part of the target universe. In this case, the estimator for the population total $Y_{[T]}$ is

$$\hat{Y}_{[1]} = N_{[F]}\bar{y}_{[F]} = \hat{Y}_{[F]} \quad (5)$$

We note that $\hat{Y}_{[F]}$ is an unbiased estimator of $Y_{[F]}$ but is biased for $Y_{[T]}$. The bias in using $\hat{Y}_{[1]}$ as an estimator of $Y_{[T]}$ is given as follows:

$$\begin{aligned} Bias(\hat{Y}_{[1]}) &= E(Y_{[1]}) - Y_{[T]} \\ &= Y_{[F]} - Y_{[T]} \\ &= (Y_{[B]} + Y_{[C]}) - (Y_{[A]} + Y_{[B]}) \\ &= Y_{[C]} - Y_{[A]} \end{aligned} \quad (6)$$

From the bias expression given in (6), it can easily be noted that bias is simply the difference between the totals of the units in the frame but are not in the target universe (units which should have not been enumerated) and the units in the target universe not covered in the sampling frame (and hence has no chance of inclusion to the sample) . It is also independent of the sample size $n_{[F]}$ indicating that the magnitude of the bias remains unchanged even if the sample size is increased. With size as strata and for surveys covering only larger establishments, then $Y_{[A]}$ represents the total of all establishments that was not covered.

Strategy 2: Another strategy in estimating $Y_{[T]}$ which many practitioners are inclined to adopt in practice is to ignore units in [C] appearing in the sample of $n_{[F]}$ from $N_{[F]}$ or treating these units as "nonresponse" and making the necessary adjustment in the estimate. That is, an estimator of $Y_{[T]}$ in this situation is defined as:

$$\hat{Y}_{[2]} = N_{[F]}\left(\frac{n_{[B]}}{n_{[F]}}\right)\overline{y}_{[B]} \tag{7}$$

where $n_{[B]}$ Is the number of units in the sample that belongs to [B] – correctly classified units belonging to the sample.

$\overline{y}_{[B]}$ Mean of units in the sample belonging to [B].

It must be noted that the quantity $N_{[F]}\left(n_{[B]}/n_{[F]}\right)$ is an unbiased estimate of $N_{[B]}$ under simple random sampling. Thus, (7) is in effect a postratified estimator of the total of all units belonging in [B].

The bias in using $\hat{Y}_{[2]}$ as an estimator of $Y_{[T]}$ is given as:

$$\begin{aligned} Bias(\hat{Y}_{[2]}) &= E(\hat{Y}_{[2]}) - Y_{[T]} \\ &= E_{n_{[F]}} E\left\{ N_{[F]}\left(\frac{n_{[B]}}{n_{[F]}}\right)\overline{y}_{[B]} \mid n_{[F]} \right\} - Y_{[T]} \\ &= E_{n_{[F]}}\left\{ N_{[F]}\left(\frac{n_{[B]}}{n_{[F]}}\right)\overline{Y}_{[B]} \right\} - (Y_{[A]} + Y_{[B]}) \\ &= N_{[B]}\overline{Y}_{[B]} - (Y_{[A]} + Y_{[B]}) \\ &= Y_{[B]} - (Y_{[A]} + Y_{[B]}) = -Y_{[A]} \end{aligned} \tag{8}$$

From (8), $\hat{Y}_{[2]}$ underestimates $Y_{[T]}$ by the total of all units in the target universe but was not included in the frame (undercoverage). As in (6), the magnitude of the bias depends on the extent of undercoverage and the contribution of these units to the population total. It is also independent of the sample size.

Strategy 3: One solution to minimize if not eliminate the effect of coverage error is the use of multiple frame. The whole idea is to find another list of establishments, preferably constructed independently of the CE and perhaps at different points in time. The two (or more) lists are then compared. The following table shows the possible scenario when such comparison is made.

Table 2.
Possible scenario in the comparison of two independent list of establishments.

	Number of Establishments IN List 2	Number of Establishments NOT in List 2	TOTAL
Number of Establishments IN List 1 (CE)	$N_{[11]}$	$N_{[12]}$	$N_{[1+]}$
Number of Establishments NOT in List 1 (CE)	$N_{[21]}$	$N_{[22]}$	$N_{[2+]}$
TOTAL	$N_{[+1]}$	$N_{[+2]}$	$N_{[++]}$

If we assume that both lists collectively is able to list all establishments in the target universe, then $N_{[21]}$ represents the establishments not listed in the CE. If the second list was constructed at a later time or the reference period is closer to the survey, then it may be broken down as units not listed at the time of the CE and "births" in the establishment frame, i.e. establishments that started to operate after the CE period. It must also be noted that in the comparison of the lists, only $N_{[11]}, N_{[12]}$, and $N_{[21]}$ can be observed. Thus, the quantity $N_{[22]}$ is unobservable and is a conceptual quantity indicating that coverage errors may be present in both lists. In relation to this, only $N_{[1+]}$ and $N_{[+1]}$ can be computed and their sum represents the total number of distinct establishments enumerated in both lists assuming of course that duplication is not present. Thus, with two lists, a graphical representation of coverage errors is presented in Figure 3.

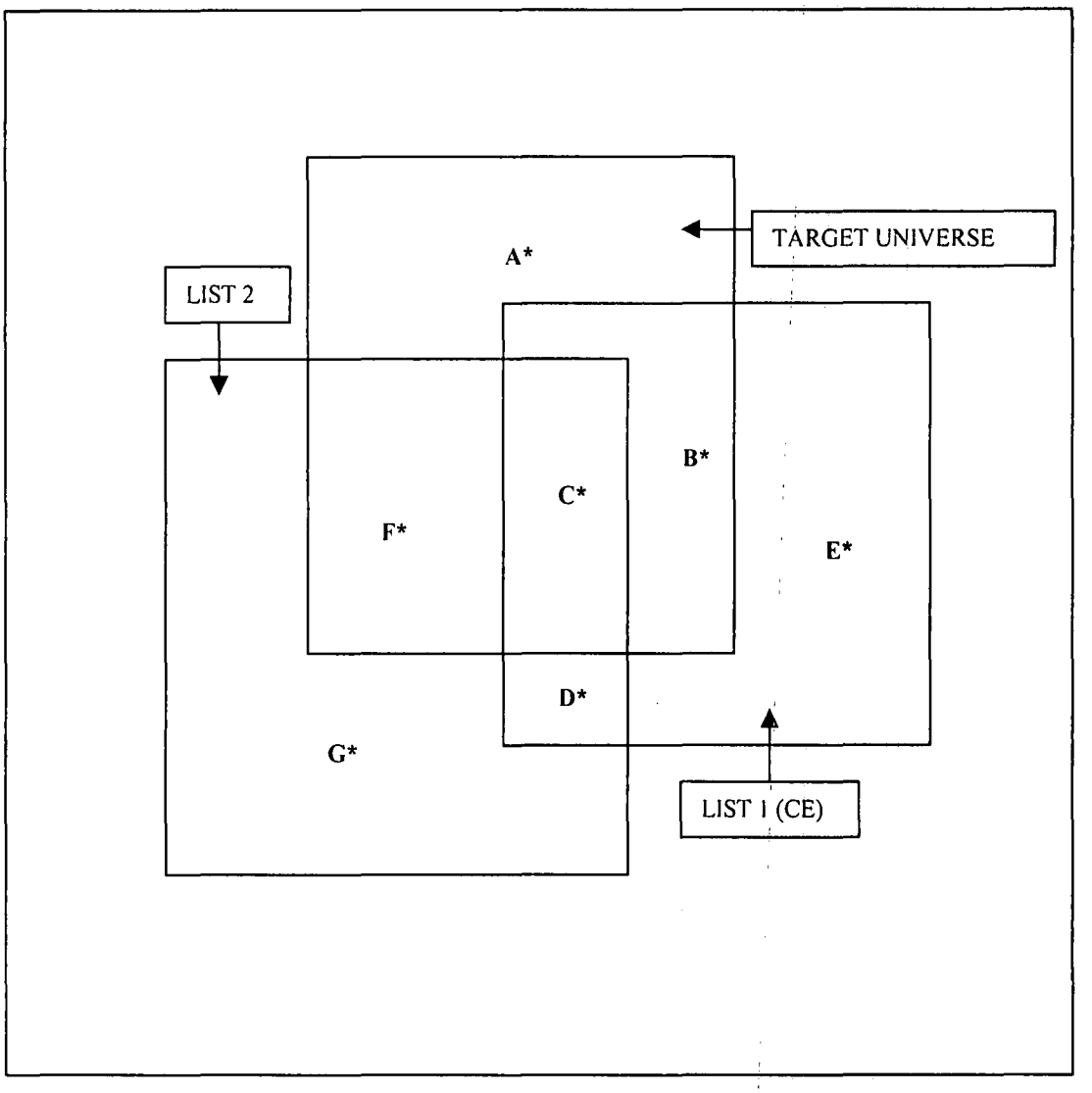

Figure 3: Coverage Errors with two frames.

Similar to earlier notations, let $N_{[\,]}$ and $Y_{[\,]}$ be the number and population total of elements falling in region [ ]. Thus, the total number of elements in the target population is:

$$N_{[T]} = N_{[A^*]} + N_{[B^*]} + N_{[C^*]} + N_{[F^*]} \tag{9}$$

The population total of the elements in the target population, which is the object of estimation, is defined as:

$$Y_{[T]} = Y_{[A^*]} + Y_{[B^*]} + Y_{[C^*]} + Y_{[F^*]} \tag{10}$$

With two frames, the strategy in estimating (10) calls for the selection of a sample of size $n_{[1^*]}$ from List 1 and an independent sample of size $n_{[2^*]}$ from among units in List 2 but

not in List 1. Note, the total number of elements in List 2 but not in List 1 can be determined or is observable and in terms of Figure 3, this is defined as:

$$N_{[2^*]} = N_{[F^*]} + N_{[G^*]} \tag{11}$$

Conceptually, the sample sizes $n_{[1^*]}$ and $n_{[2^*]}$ can be broken down as:

$$\begin{aligned} n_{[1^*]} &= n_{[B^*]} + n_{[C^*]} + n_{[D^*]} + n_{[E^*]} \\ n_{[2^*]} &= n_{[F^*]} + n_{[G^*]} \end{aligned} \tag{12}$$

Further, let

$$Y_{[1^*]} = Y_{[B^*]} + Y_{[C^*]} + Y_{[D^*]} + Y_{[E^*]} \tag{13}$$

be the population total of all elements in List 1. Similarly, let

$$Y_{[2^*]} = Y_{[F^*]} + Y_{[G^*]} \tag{14}$$

be the population total of all elements in List 2 but not in List 1.

Thus, the unbiased estimators of $Y_{[1^*]}$ and $Y_{[2^*]}$ are $\hat{Y}_{[1^*]}$ and $\hat{Y}_{[2^*]}$ respectively and are defined as:

$$\hat{Y}_{[1^*]} = N_{[1^*]}\bar{y}_{[1^*]}, \quad \hat{Y}_{[2^*]} = N_{[2^*]}\bar{y}_{[2^*]} \tag{15}$$

Thus, if overcoverage units are treated as part of the target population under the framework presented in Figure 3, then the estimator of $Y_{[T]}$ is computed as:

$$\hat{Y}_{[T]} = \hat{Y}_{[1^*]} + \hat{Y}_{[2^*]} \tag{16}$$

The bias introduced by coverage error in estimation is:

$$
\begin{aligned}
Bias(\hat{Y}_{[T]}) &= E(\hat{Y}_{[T]}) - Y_{[T]} \\
&= E(\hat{Y}_{[1^*]} + \hat{Y}_{[2^*]}) - (Y_{[A^*]} + Y_{[B^*]} + Y_{[C^*]} + Y_{[F^*]}) \\
&= Y_{[1^*]} + Y_{[2^*]} - (Y_{[A^*]} + Y_{[B^*]} + Y_{[C^*]} + Y_{[F^*]}) \\
&= (Y_{[B^*]} + Y_{[C^*]} + Y_{[D^*]} + Y_{[E^*]}) + (Y_{[F^*]} + Y_{[G^*]}) - (Y_{[A^*]} + Y_{[B^*]} + Y_{[C^*]} + Y_{[F^*]}) \\
&= (Y_{[D^*]} + Y_{[E^*]} + Y_{[G^*]}) - Y_{[A^*]}
\end{aligned}
\tag{17}
$$

The term $Y_{[E^*]} + Y_{[D^*]} + Y_{[G^*]}$ is the population totals of all units included in both lists but are not included in the target population (e.g. smaller establishments). Again, the bias is the difference in the totals of all "out-of-scope" elements and elements in the target population which were not included in the frame. As in earlier results, (17) is again independent of the sample size.

5. CONCLUSION

The paper presented that forms of the biases as a result of coverage errors. An important result is that the bias of the estimates considered does not diminish with increase in sample size and its magnitude depends on the differences in characteristics (e.g. population totals) between units in the target population but are not included in the frame (undercoverage) and "out-of-scope" units. It is also difficult to ascertain whether such biases are negligible unless perhaps the extent of coverage error is very small.

In establishments surveys, such bias may be large even if the extent of units missed out is small especially if the missed unit is one of the major industry player and hence its contribution to the population total maybe large.

Therefore it is important that efforts should be made to minimize the incidence of coverage errors. Among the steps that can be taken to address this concern are:

(a) Ensure quality assurance procedures in the conduct of the Census of Establishments.
(b) The use of multiple frames.
(c) A system of frame updates (especially in the aspect of capturing "births") should be put in place. Perhaps, it is about time that the sampling design of establishments' inquiries be assessed. In the Philippines, it is estimated that about 85% of all large establishments are located in first class municipalities and cities. Such towns number only around 120. Thus, one can concentrate the updating procedure in such towns by using business permits, which by law is renewed annually.

The concepts presented in here are by no means extensive as far as trying to understand the effect of coverage errors in estimation. The frameworks presented can further be refined. One such refinement/extension is presented in Pacificador (1991).

References

HANSEN, M.H., W.N. HURWITZ, AND, T.B. JABINE. 1964. "The use of imperfect lists for probability sampling at the US Bureau of the Census". *Bull. Int'l. Stat. Inst.* pp497-517.

GROVES, R.M. 1989. *Survey Errors and Survey Costs.* John Wiley, U.S.A.

ONATE, B.T. 1988. *Sources, Types, Measurement, Control and Evaluation of Nonsampling Errors: Philippine Experience.* Oarland Publishing, Los Baños, Laguna.

_____., A.Y.PACIFICADOR, JR., AND L.T.HABACON. 1989. "Some Nonsampling Errors in the 1987 Annual Survey of Establishments". Unpublished Research Report. SRTC, Manila, Philippines.

PACIFICADOR, A.Y. JR. 1991. "Some solutions to Nonsampling Errors (NSE) in the Annual Survey of Establishments (ASE)". Ph.D. dissertation, UP Los Baños.