

**9<sup>th</sup> National Convention on Statistics (NCS)**  
EDSA Shangri-La Hotel  
October 4-5, 2004

**The Development of the 2003 Master Sample (MS)  
for Philippines Household Surveys**

by:  
Marissa L. Barcenas

For additional information, please contact:

Author's name:	Marissa L. Barcenas
Designation:	Head
Agency:	Statistical Methodology Unit National Statistics Office
Address:	Sta. Mesa, Manila

# **The Development of the 2003 Master Sample (MS) for Philippines Household Surveys**

by:  
Marissa L. Barcenas

## **ABSTRACT**

Since the 1950's, the Philippines, particularly the National Statistics Office (NSO) have employed the concept of a master sample (MS) in the design and conduct of its household surveys. A master sample as used in this context is defined as a sample from which subsamples can be selected to serve the needs of more than one survey or survey rounds (UN-NHSCP).

The master sample consists of a randomly selected set of geographical areas, termed primary sampling units (PSUs), which are taken to be barangays or combinations of barangays. The samples of households and persons for all of the surveys are then selected within that set of PSUs. The use of a master sample of PSUs has the benefits of using the same set of interviewers for each of the surveys and of sharing the costs of updating listings of housing units within the PSUs across several surveys.

This paper documents all the important processes that were followed in designing the 2003 Master Sample (MS). In particular, this paper provides a description of the choice of the primary sampling units, determination of sample size in terms of the number of housing units/households, the choice of domain, selection procedures adopted, weighting and estimation procedures, and rotation of samples.

## **I. Introduction**

Since the 1950's, the Philippines, particularly the National Statistics Office (NSO) have employed the concept of a master sample (MS) in the design and conduct of its household surveys. A master sample as used in this context is defined as a sample from which subsamples can be selected to serve the needs of more than one survey or survey rounds (UN-NHSCP). With several intercensal household surveys to be conducted for a given period, using the MS approach rather than designing surveys independently, is more cost effective in the sense that common resources are utilized in the preparation of sampling frames, sample design and selection. Moreover, under a MS setup, it is a lot easier to link different surveys as a result of using standard concepts and definitions (Verma, 1991).

The 2003 MS intends to represent the total household population of the Philippines. It excludes persons living in institutions and the homeless. In addition, due to operational considerations, some 0.4 percent of the total population living in least accessible barangays has been excluded from the sampling frame.

The PSUs in the 2003 MS were selected within a set of strata using probability proportional to estimated size (PPES) sampling, where the measure of size was the number of households in the PSU according to the 2000 Census of Population and Housing (CPH). The primary strata were the 17 regions of the country. Within each region, further stratification was performed using geographic groupings such as provinces, highly urbanized cities (HUCs), and independent component cities (ICCs). Within each of these substrata formed within regions, the PSUs were further stratified, to the extent possible, using the proportion of strong houses, the proportion of households in agriculture, and a measure of per capita income as stratification factors. Regional stratification was of major importance because regions are domains of study for which separate estimates are required from many of the surveys. To meet this requirement, higher than average sampling fractions were applied in regions with smaller populations in order to produce adequate sample sizes in those regions.

The full 2003 MS consists of a sample of 2,835 PSUs (some of which were selected with certainty). This number was chosen to be large enough to satisfy the needs of the large surveys such as the Labor Force Survey (LFS) and Family Income and Expenditure Survey (FIES). This number is, however, larger than desirable for smaller surveys. The full MS has therefore been designed as a combination of four replicates, each of approximately 709 PSUs, with each replicate being a national sample design. Smaller surveys can thus be confined to one, two, or three of the replicate samples as desired.

This paper provides description of the important processes that were followed in coming up with the 2003 Master Sample (MS) design briefly described above. In particular, this paper describes briefly the activities done on the development of the frame, sample allocation procedures, the choice of the ultimate sampling units, the choice of domain, selection procedures adopted, and rotation of samples. Discussions are heavily lifted from the 2003 MS Documentation prepared by the former Research and Development Unit (RDU) of NSO under the technical assistance granted by the Asian Development Bank (ADB) on Improving Poverty Monitoring Surveys. Development of the 2003 MS design and its documentation were done under the technical guidance of Dr. Graham Kalton, Senior Statistician and Senior Vice President of Westat; Dr. Arturo Y. Pacificador of the University of the Philippines at Los Baños; and Dr. Dalisay S. Maligalig, Statistician, Asian Development Bank (ADB).

## **II. The development of the sampling frame of PSUs**

Development of the frame of PSUs involved making a decision on the PSU and its formation. The choice of PSUs for the 2003 MS was based on the following criteria: a PSU should be (1) well defined with clear and stable boundaries; (2) Information should be available on the estimated sizes (to be used in probability proportional to size sampling) and on characteristics (to be used for stratification); (3) there should be a large number of them from which to select the sample; and (4) it should be large enough in terms of households and population to support all the household surveys for which the MS will be used. Since the NSO intends to use the MS for a minimum of ten years and a quarterly

rotation of samples of approximately 20 per PSU with fifty percent returning samples in two consecutive years is planned, a PSU should have at least 400 households.

Alternative choices that were evaluated include Census Enumeration Areas (EAs), barangays, and municipalities. However, EAs are too small to serve as PSUs for the MS while municipalities are too few to allow flexibility in the design. Barangays are more suitable in terms of numbers; there are in total 41,942 of them. However, most barangays did not meet the minimum size requirement and were therefore combined with other barangays to form a PSU.

To allow some flexibility, a 2003 MS PSU was defined to be a barangay with at least 500 households or a combination of barangays that together have at least 500 households.

The sampling frame of PSUs was built using the Enumeration Area Reference File (EARF) of the 2000 CPH. The EARF contains the distribution of the target population, which is the household population, by geographic area.

The household population comprises all persons living in private households, where a **household** is defined to be an aggregate of persons, generally but not necessarily bound by ties of kinship, who live together under the same roof and eat together or share in common the household food. Household membership consists of the head of the household, relatives living within him, and other persons who share the community life for reasons of work or other consideration. A person who lives alone is considered a separate household.

Prior to combining barangays to form PSUs, some components of the target population were excluded or given special consideration. These include the least accessible barangays (LABs) which by definition are barangays that are accessible by regular means of transportation for less than three times a week or only by costly means. For operational consideration and in view of the fact that LABS contained only 0.4 percent of the total household population, LABS were excluded from the frame. Second, was the case of barangays with peace and order problem (POPs). Due to considerable cases of POPs barangays, a total of 1,049 barangays, POPS were not excluded but combined with non-POPs barangays to facilitate interview of households from the PSU. The assumption is that peace and order problem is temporary in nature and that effective strategies for conducting field operations in these barangays can be explored as was done in the past. Third is the case of very large barangays in highly urbanized areas. These large barangays were considered as certainty PSUs.

The general approach for combining barangays was to group small barangays that are contiguous based on their relative positions in the municipal map. Using this procedure and setting the minimum size to 500, a total of 16,586 PSUs was formed for the master sample frame.

### III. Determination of Sample Size and Sample Allocation

Determination of sample PSUs to select for the MS and the allocation of the sampled PSUs across geographical divisions of the country required the decision on the domain or geographical division for which the sample was designed to produce individual estimates of adequate precision. To do this, the 1996 MS domains<sup>1</sup> were evaluated as to precision of estimates of frequently gathered indicators such as unemployment, poverty incidence, and contraceptive prevalence rates at different geographical divisions (region, province, and 1996 MS domains) against the desired level of precision. For illustration of the general results, Table 1 presents a summary distribution for poverty incidence estimates. Results show that only regional level estimates are guaranteed to meet the desired precision of 5 percent Coefficient of Variation<sup>2</sup> (CV). The first row of results in Table 1 shows that about 71 percent of the 127 domain estimates have CVs of more than 10 percent. To obtain estimates with at most 10 percent CV, the sample size for these domains would have to be quadrupled. Even if the cities/municipalities are excluded, over half of the provinces have CVs of greater than 10 percent.

**Table 1. Percent distribution of CVs of estimated poverty incidence by different geographic subgroups using 2000 FIES**

Geographic subgroup	Number of subgroups	CV Values					
		<5%	5%-	10%-	15%-	20%-	25%+
1996 Sampling Domains	127	2.4	26.8	28.3	16.5	7.9	18.1
Province	82	3.7	43.9	40.2	9.8	1.2	1.2
Region	17	35.3	64.7				

Source of basic data: 2000 FIES, National Statistics Office

Next step is the determination of total sample size. Total sample size was determined based on a desired CV of 5 percent in all regions but NCR (with a low poverty rate and CV of 10 percent was considered acceptable). Taking both precision levels and available resources into account, it was decided that a sample of about 44,000 responding households would be adopted for a FIES and an LFS using the 2003 MS. This sample size is slightly greater than that for the FIES and LFS with the previous MS design.

To allocate the total sample size of 44,000 households across the region, two classes of estimates were considered: 1) Estimates at the national level for the total sample and for subgroups that cut across regions (e.g. numbers of crop farmers and female headed households, proportions of persons in poverty and of persons in the labor force who are unemployed etc.), and estimates of the differences between subgroups; and 2) Estimates at the regional level (e.g. unemployment rate or poverty incidence for each region), and estimates of

<sup>1</sup> In the 1996 MS, the sampling domains were provinces and cities/municipalities with population of 150,000 or more

<sup>2</sup> The CV is the ratio of the standard error of an estimate to the estimate itself. The standard error is computed in a manner that takes the survey's complex sample design into account.

differences between regions. Proportional allocation suits class (1) estimates, while equal allocation suits class (2) estimates (Kish 1988). The difference between these two allocations is substantial when the regions differ markedly in population size, as is the case in practice. A compromise allocation that is suboptimal for both classes of estimates but that performs reasonably well for both is the Kish allocation. Kish procedure allocates the sample size  $n$  using the

formula

$$n_d = n \frac{\sqrt{L^2 + IW_d^2}}{\sum_d \sqrt{L^2 + IW_d^2}} . \quad (1)$$

where  $n_d$  is the sample size allocated to region  $d$ ,  $n$  is the total sample size,  $L$  is the number of regions ( $L=17$ ),  $N_d$  and  $N$  are the total number of households in region  $d$  and for the entire country,  $W_d = N_d / N$  is the proportion of the population in region  $d$ , and  $I$  is an index denoting the relative importance assigned to estimates of class (1) as compared to those of class (2). With  $I=0$ , the Kish allocation reduces to the equal allocation. With  $I \rightarrow \infty$ , the Kish allocation tends to the proportional allocation. Evaluation of this procedure using different values of  $I$  resulted to using  $I=1$ .

To determine the optimum subsample size or the number of households to be selected in a PSU, the simple cost model given below was used:

$$c_{opt} = \sqrt{\frac{C_1 (1-rho)}{C_2 \quad rho}} \quad (2)$$

where  $C_1$  is the cost of adding an additional PSU into the sample,  $C_2$  is the cost of an additional interview, and  $rho$  is a measure of the within-stratum homogeneity of the PSUs for the survey variable of interest (Kish, 1965). Rough estimates of  $rho$  and  $C_1 / C_2$  were obtained from past survey. Given the simplified nature of the cost model, the estimate of  $c_{opt}$  was treated as just an indication of the approximate magnitude of the PSU sample size. The final PSU sample size was decided upon in consideration of fieldwork logistics. It was also noted that  $C_1 / C_2$  can vary across different parts of the population and it is much smaller in large urban areas than elsewhere. For this reason the subsample size for the NCR has been set lower than for other parts of the country. Based on the considerations given above a subsample size of 12 households per PSU has been chosen for the NCR and one of 16 households has been chosen for all other regions.

The number of PSUs to be sampled in each region is then computed by dividing the allocated sample size by the desired subsample size per PSU—12 for the NCR and 16 all other regions. Note that the PSUs vary considerably in size. In order to achieve an equal probability sample of households within a region and at the same time have some control on the subsample size in each sampled PSU, the MS PSUs were selected by systematic sampling with probabilities proportional to measure of sizes based on their number of households in the 2000 CPH. To avoid multiple selection, very large PSUs were removed from the systematic selection process and treated as certainty PSUs.

Also removed from the systematic sampling procedure were PSUs whose selection probabilities based on their sizes were 0.7 or larger. These PSUs were similarly included in the MS with certainty and are strata rather than PSUs. Once the initial set of certainty PSUs had been removed, another iteration was conducted to identify other PSUs that now satisfied the condition to be certainty selections. Certainty PSUs and non-certainty PSUs are also often termed self-representing and non-self-representing PSUs, respectively, or SR and NSR PSUs for short.

The number of non-certainty PSUs to be sampled in each region was initially determined by subtracting the number of certainty selections from the desired number of sample PSUs. This number was allocated between the provinces and HUC/ICCs in the region in proportion to their number of households in non-certainty PSUs in the 2000 CPH. The number of non-certainty PSUs to be sampled in each province and HUC/ICC was then adjusted to be a multiple of four to facilitate the formation of four replicates for use when a particular survey requires only a sample of the full set of MS PSUs. Table 2 presents the initial household sample size allocated to each region using the Kish allocation with  $l=1$  and also the numbers of certainty (self-representing or SR) PSUs, non-certainty (non-self-representing or NSR) PSUs, and total PSUs in the 2003 MS.

**Table 2. Sample size and number of PSUs by region**

Region	Number of households ('000)	$W_d$	Sample size	Number of MS PSUs		
				SR PSUs	NSR PSUs	Total
Region I - Ilocos Region	837	0.05	2,408	0	148	148
Region II - Cagayan Valley	568	0.04	2,085	0	132	132
Region III - Central Luzon	1,676	0.11	3,726	7	224	231
Region IVA - CALABARZON	1,936	0.13	4,181	32	232	264
Region IVB - MIMAROPA	452	0.03	1,974	2	124	126
Region V - Bicol Region	892	0.06	2,493	0	156	156
Region VI - Western Visayas	1,220	0.08	2,970	7	176	183
Region VII - Central Visayas	1,142	0.07	2,848	9	168	177
Region VIII - Eastern Visayas	712	0.05	2,249	0	140	140
Region IX - Zamboanga Peninsula	547	0.04	2,064	11	120	131
Region X - Northern Mindanao	698	0.05	2,232	12	128	140
Region XI - Davao Region	754	0.05	2,300	23	120	143
Region XII - SOCCSKSARGEN	646	0.04	2,171	17	120	137
National Capital Region	2,066	0.13	4,413	193	164	357
Cordillera Administrative Region	265	0.02	1,838	7	108	115
Autonomous Region in Muslim Mindanao	496	0.03	2,013	2	124	126
Caraga	397	0.03	1,928	8	112	120
<b>Philippines</b>	<b>15,312</b>	<b>1.00</b>	<b>43,882</b>	<b>330</b>	<b>2,496</b>	<b>2826</b>

#### IV. Sample Selection

Although the MS PSUs are barangays and combinations of barangays, the sample of households for any one survey is confined to a single EA within each sampled PSU. The EA is introduced as an extra stage of sampling in order to reduce travel time for interviewers. Also, this step limits the updating needed for the housing unit list to a single EA in each PSU.

The samples of EAs and their associated PSUs are selected with PPES. The measures of size used in the PPES selection were based on the number of households that the EAs contained in the 2000 CPH. However, these numbers are adjusted to facilitate the sampling of households within the selected EAs and to compensate for non-coverage and population growth. The sample of EAs is selected by systematic sampling from a list of all EAs in a stratum ordered by PSU, with the PSUs ordered by the implicit stratification variable.

The ultimate sampling units for household surveys are households and the persons living in them. However, the listings that are created in the sampled EAs are often housing units rather than households. Table 4 shows that, based on the 2000 CPH, a housing unit contains a single household in almost all cases across all regions. Some housing units contain two households and a few contain more than two households but the average number of households per housing unit is only 1.03. Based on these numbers, the proposed procedure of selecting all households in sampled housing units seems generally acceptable.

Up to three households are sampled within the selected housing units. In the few cases where a housing unit have more than three households, a sample of three households is selected with equal probability.

**Table 3. Percent distribution of housing units by number of households and regions**

Region	Number of housing units	Percent distribution of housing units by number of households (hhs)					Mean number of hhs
		1 hh	2 hhs	3 hhs	4 hhs	≥ 5 hhs	
Region I - Ilocos Region	808,126	97.5	2.1	0.3	0.1	-	1.03
Region II - Cagayan Valley	544,524	98.4	1.4	0.1	-	-	1.02
Region III - Central Luzon	1,601,018	98.4	1.4	0.2	-	-	1.02
Region IV – Southern Tagalog	2,369,425	98.4	1.3	0.2	-	-	1.02
Region V - Bicol Region	883,175	98.9	1.0	0.1	-	-	1.01
Region VI – Western Visayas	1,192,185	98.6	1.3	0.1	-	-	1.02
Region VII - Central Visayas	1,117,462	98.8	1.0	0.1	-	-	1.01
Region VIII – Eastern Visayas	707,560	99.1	0.8	0.1	-	-	1.01
Region IX - Zamboanga Peninsula	687,329	98.8	1.1	0.1	-	-	1.01
Region X - Northern Mindanao	532,849	98.6	1.2	0.2	-	-	1.02
Region XI - Davao Region	1,044,609	98.3	1.5	0.2	-	-	1.02
Region XII - SOCCSKSARGEN	485,313	97.5	1.9	0.4	0.1	0.1	1.03
National Capital Region	2,001,679	95.3	3.7	0.6	0.2	0.2	1.06
Cordillera Administrative Region	259,890	98.7	1.1	0.1	-	-	1.02

Autonomous Region in Muslim Mindanao	366,291	94.9	3.8	0.9	0.3	0.2	1.07
Caraga	386,283	98.4	1.4	0.2	-	-	1.02
Disputed Area <sup>a</sup>	3,396	92.3	6.2	0.9	0.3	0.4	1.10
<b>Philippines</b>	<b>14,891,114</b>	<b>97.9</b>	<b>1.7</b>	<b>0.2</b>	<b>0.1</b>	<b>-<sup>b</sup></b>	<b>1.03</b>

<sup>a</sup> a disputed area is claimed by two or more political units

<sup>b</sup> (-) less than 0.1%

## Sample Rotation

The sample rotation scheme adopted for the 2003 MS balances two concerns: 1) estimation of year-on-year changes; and 2) increasing the sample size for provincial estimation, particularly for labor force estimates. This rotation scheme retains housing units in the sample for only two rounds, one year apart. That is, a fifty percent overlap is effected for two rounds one year apart with quarterly rotation of the full sample within a year (during non-FIES years). The quarterly rotation aimed to increase the sample size for annual provincial estimation although some calculations performed using the data from the 2001 LFS showed that the increase in sample size would not be in the same magnitude as the actual fourfold increase. The reason for the lesser increase in effective sample size is that, because of fieldwork constraints, the separate quarterly samples of housing units must all be selected in the same set of MS PSUs. A fourfold increase in effective sample size would be achieved if the four quarterly samples were selected in different PSUs.

Although the increase in effective sample size for average annual estimates might not be great, the increase may be valuable for the production of provincial level estimates and estimates for other smaller subgroups of the population.

Figure 1 illustrates a simple rotation design in which housing units remain in sample for two consecutive years. This design has different samples for each quarter of the year and each sample is generally interviewed on two occasions one year apart. Exceptions occur at start-up, when the 2004 samples for the B rotation cluster drop out in 2005, and the reorganization of the interview schedule in 2007 to accommodate the FIES requirements.

**Figure 1. A simple sample rotation design from 2004 to 2008**

Year	Quarter	Sample/Rotation Cluster <sup>a</sup>	
2004	January	A1	B1
	April	A2	B2
	July	A3	B3
	October	A4	B4
2005	January	A1	B5
	April	A2	B6
	July	A3	B7
	October	A4	B8
2006	January	A5	B5
	April	A6	B6
	July	A7	B7
	October	A8	B8
2007	January	A7	B7
	April	A6	B9
	July	A5	B10
	October	A8	B11
2008	January	A9	B12
	April	A10	B9
	July	A11	B10
	October	A12	B11

<sup>a</sup> Numbers represent rotation groups formed for the housing units within the sampled EAs and letters represent rotation clusters. Rotation cluster A includes replicates one and two while rotation cluster B includes replicates 3 and 4.

## REFERENCES

- Kish, L [1965] Survey Sampling. New York: John Wiley & Sons, 268-269.
- Kish L. [1988] "Multipurpose Sample Designs" Survey Methodology. Statistics Canada, Vol. 14 No.1, 19-32
- Kish L. [1987] Statistical Design for Research. New York: John Wiley & Sons, 113-115.