Sampling Strategy for Agriculture Censuses and Surveys in Developing Countries[1]

I. P. David
Asian Development Bank
Manila, Philippines

Abstract: Full enumeration agriculture censuses have become less feasible financially and operationally in many developing countries. They had not always lived up to their promise to serve as frame for the intercensal surveys and to provide reasonably accurate benchmark data for small areas. Observed large deviations between the survey and census results point to serious non-sampling errors especially in the latter. Some developing countries had resorted to replacing agriculture censuses with sample enumerations that differ from the intercensal surveys only in the choice of a smaller domain, hence a bigger sample. There is need for strategies that link the sample census and the intercensal surveys, in a manner appropriate to the limited technical and operational capabilities of the developing countries. Using the Philippines as example, the paper presents the design of the 1991 sample census of agriculture and outlines a strategy to draw second phase replicated samples directly from it.

Abbreviations: BAS - Bureau of Agriculture Statistics, CA - census of agriculture, CBS - Central Bureau of Statistics, CPH - census of population and housing, CV - coefficient of variation, FAO - Food and Agriculture Organization, ESCAP - Economic and Social Commission for Asia and the Pacific, NSO - National Statistics Office.

1. Introduction

To the public, census means complete count. The specific meaning, according to Webster's Dictionary, is a periodic governmental enumeration of population. From people and property at the time of the Romans, the subject of censuses has expanded in modern times to include agriculture including fishery and forestry, to establishments or economic activities. The idea that censuses should be complete counts has survived, through complete enumeration for a few main items put in a so-called 'short questionnaire' and sampling for the remainder organized in a 'long questionnaire'. Such is usually the case for censuses of population and housing (CPHs). For censuses of agriculture (CAs), villages and higher administrative levels are usually completely covered, with the sampling done at the household or operational holding level, such as complete enumeration of large holdings and sampling of the smaller ones.

The main reasons often cited for continuing to do complete enumeration censuses (as opposed to sample enumerations) are the following:

(a) Only a full census can provide benchmark data to meet the demand for small area data [e.g. ESCAP, 1997]. If the full objectives of a national statistical programme for food and agriculture are to be met, it is desirable to have complete enumeration; only in this way is it possible to obtain statistics for every geographical, agro-ecological or administrative region, irrespective of size [FAO, 1995]. Sample enumerations could not provide minor crops

[1] Paper presented at the International Conference on Agriculture Statistics 2000, 18-20 March 1998, Washington, D.C. Comments and assistance by Abuzar Asra and Dalisay Maligalig of ADB and S. Sugito of the Indonesian Central Bureau of Statistics are greatfully acknowledged. The views expressed in the paper are those of the author and not of the ADB.

acreage for small areas in particular and data for efficient resources allocation as well as for setting up action programmes at the grassroots levels in general [BBS, 1997]. The belief that only complete enumeration can meet the data needs of administrative institutions at all levels was one of three reasons given for China's decision to completely tally 230 million households, 800 thousand administrative villages, 60 thousand towns and over two thousand township enterprises during its 1997 CA [Zhu, 1997].

(b) A full enumeration census serves as a sampling frame for inter-censal surveys. The importance attached to this reason can be absolute, due to the mistaken notion that only complete enumeration can provide sampling frames for regular surveys [Zhu]. Surprisingly, this view is widespread among official statisticians. It was voiced during a recent meeting of a Working Group of Statistical Experts in the ESCAP region, and only after some lively discussion spurred by an intervention by the author did the group acknowledge that while a full census provided an ideal sampling frame for ensuing surveys, a sample (as well as other data sources such as area frames) could also serve that purpose [ESCAP, op cit]. The view may well have been nurtured by standard international recommendations on conducting censuses; e.g., the abovementioned FAO publication states that, when a complete enumeration is possible, a census of agriculture provides a frame for later sample surveys -- with no further clarification that there are other options for constructing frames.

(c) Another reason for a full enumeration census is that it is required by law in some countries. This pertains mainly to population censuses, the results of which are used in redistricting, creation of new administrative areas and budget apportionment. It is rare that agriculture censuses are used for similar purposes. However, the law may specify that full enumeration censuses be done by one agency and sample enumerations by another agency. Nonetheless, laws are man-made and can be changed, as in the transfer of the 1997 US CA from the Bureau of the Census to the Department of Agriculture.

On the other hand, the literature is replete with the major drawbacks of complete censuses: very high costs, delayed results, higher non-sampling error not necessarily less than the total error in a sample enumeration. What is not said often enough is that, on account of their sizes, complete enumeration CAs use different, less expensive and less accurate data collection methods than those employed in the intercensal surveys. This often results in significant discrepancies between the census and the regular sample surveys, especially in, but not limited to, countries where different agencies are responsible for these undertakings. Thus, the regular sample surveys are still done during the census year. This practice further aggravates costs, confounds users, puts the national statistical system in bad light and is being increasingly questioned by those in charge of government finance and planning.

The rest of the paper will try to analyze whether or not the main reasons for complete enumeration CAs continue to be valid or strong in the developing countries. Examples of two countries, Indonesia in Southeast Asia and Nepal in South Asia, that have relied on sample censuses, are presented in the next section. Section 3 discusses the circumstances behind and details of the sampling strategy used in the 1991 Philippines CA. Section 4 proposes a strategy that directly links the census (as a first phase sample) to the intercensal surveys (as second phase samples) in a manner that integrates the agricultural data collection system, effectively solves sampling frame problem and narrows the gap between census and survey results.

2. Sampling in Agriculture Censuses in Asian Developing Countries

We discuss here two CAs in Asia which, together with the Philippines' that is presented in the next section, depart significantly from usual international recommendations and challenge the distinction between censuses and surveys.

2.1 Nepal Agriculture Census

Nepal has a population of 22 million, 85% of whom live in rural areas. Its US $200 annual per capita income and 28 % adult literacy rate are lowest in Asia [ADB, 1997]. There were four CAs between 1962 and 1992. Full enumeration had not been considered in any one of them because the Central Bureau of Statistics (CBS) could not possibly get the Government to allocate sufficient budget in the first place, and in the second place there were not enough qualified personnel to be found to effectively do the data collection and successfully complete the task within any reasonably allotted time.

Nepal is divided into five development regions, 15 zones, 75 districts and 37,000 wards. The agriculture censuses were designed to produce statistics for each of the 75 districts. For the 1992 CA, sample wards were chosen in each district with probability proportional to the estimated number of agricultural holdings from the most recent population census. A sample of holdings was selected in each sample ward using simple systematic sampling, with the sampling intervals adjusted so that the holdings within each district have equal weights; however, the weights differed across districts as the assigned ward sampling rates varied depending on the relative agricultural importance of the districts. The overall sampling fractions for the wards and holdings were 14% and 4.5% respectively [Bastola, 1991 and CBS, 1994].

While the CAs are the responsibility of the CBS, the intercensal agriculture statistics until 1992 were produced by the Ministry of Agriculture using what can be best described as subjective ways. There was no link between the two data sets-- for instance, in utilizing the census data as sampling frames for the intercensal data collection activities. With the responsibility for the intercensal agriculture statistics transferred officially to CBS since 1993, CBS' pressing problem was how to design a system of sample surveys that could generate sufficiently useful statistics for the country's 15 zones. One option, which is to draw a subsample from the unequal probability without replacement census sample and getting the estimation method right, will require survey sampling theoreticians not available in CBS (and in most agriculture statistics agencies in developing Asia). What CBS did with the one agriculture survey it managed to do since 1993 was to follow the same procedure as in the 1992 CA, with the 15 zones instead of the 75 districts as domains; i.e., 30 wards per zone were drawn using probability proportional to size (pps) sampling from the frame of wards constructed for the census, and a systematic sample of 15-20 households was selected from each sample ward. But for the common sampling frame, the census and survey were otherwise independent.

2.2. Indonesian Agriculture Census

Indonesia has 200 million people spread in 27 provinces, 298 districts (regencies), 69,000 villages and 178,000 enumeration areas (wilcahs). CAs are done during years ending with 3.

The domains in these censuses are the districts. The extreme fragmentation of landholdings (average farm size is less than 1/2 hectare) and archipelagic nature of the country make agricultural data collection costly, problematic and time consuming; hence, full enumeration had not been attempted in any of the four CAs conducted between 1963 and 1993. In the 1993 CA, 20% of the wilcahs in each district were selected using pps systematic sampling, with number of households as size measure, except in special districts (classified as municipalities and provincial capitals) where there was complete inclusion of all "agriculturally concentrated" wilcahs [CBS, 1995 ; Sukmadi, 1991; Wirosardjono, et al, 1991]. There was complete listing of the households in the sample wilcahs, during which time a short questionnaire on land use by crop type and a few structural variables, was accomplished on each agricultural household. Thus, census estimates on these variables were based on a cluster sample of wilcahs. For the other variables, a long questionnaire was used on a 20% simple systematic sample of the agricultural holdings in the sample wilcahs, or a 4% overall sampling rate.

2.3 Some Remarks and Observations

The CAs of Nepal and Indonesia can be aptly called large-scale stratified two-stage sample surveys, with districts as strata or domains, wards/wilcahs as first-stage units and holdings as second-stage units. Size is the remaining attribute that distinguish them from the intercensal surveys. Previously, *sample census* was a term endorsed by FAO and adopted by countries to describe censuses that were not full enumeration. Beginning with the year 2000 round of censuses, FAO [op cit, chapter 4] proposes *sample enumeration* instead, reserving *census* to a full enumeration. The future will tell if the countries will accept this suggestion; for one, national statistical agencies in developing countries already find it difficult to obtain special appropriations for an agriculture census, and they could find it much more so for a sample enumeration.

FAO advises that a well designed sample enumeration requires less resources, is more manageable and could produce equal or better quality data than a complete census at the national or large geographic levels, but cautions that the former requires sampling experts that are, as mentioned above, in short supply in developing countries in general, and in agriculture statistics agencies in particular. It should be added that the level of theoretical sophistication required gets higher when the sample census is used to design the intercensal sample surveys; and the difficulty escalates with the level of complexity of the census sampling scheme. In Indonesia, for example, the sample wilcahs from the census were (post)stratified and a subsample was drawn for the intercensal surveys. In this case, developing a theoretically sound estimation procedure has to take into account the fact that the census wilcahs were selected with varying probabilities without replacement. In our experience, full awareness of such theoretical nuances is not always present in developing country statistical offices; expatriate consultants too are not always cognizant of it, as well as of the developing country capabilities, and would often advise using sophisticated sampling schemes for the main objective of optimizing the precision of estimates from the sample census.

Linking the CA and the intercensal surveys offer added advantages, including enhancing the chances that the discrepancies in the results from these two sources will attract the attention they deserve, and possibly be reduced. Using the census as sampling frame, as concomitant information in ratio or regression estimation, or directly by drawing a subsample from it are three ways of providing the link. The discrepancy, however, may persist in situations where different methods of collecting the basic data are used. In Indonesia, for example, the CAs were the only occasions when production and area harvested by type of crop were obtained by interviewing the farm households. On the other hand, the intercensal surveys used crop cutting to estimate yields, which were multiplied by the estimates of total harvested area obtained from the village heads to estimate production. The significantly lower estimates from the census compared to the surveys has been a long-standing problem as will be further explored in the next section.

3. Designing the 1991 Philippines Census of Agriculture

3.1 Background

In the Philippines, which has a decentralized statistical system, the law assigns all censuses to the National Statistics Office (NSO) that is attached to the Planning Department, and all other agricultural data collection activities to the Bureau of Agriculture Statistics (BAS) of the Agriculture Department. The country is divided into 13 regions including the National Capital Region, 80 provinces, 1600 towns and over 40,000 villages. The 1960, 1971 and 1980 CAs covered all the villages and the large operational holdings; i.e. sampling was limited to the farm households with small holdings (e.g. systematically sample one in twenty of those that are less than five hectares in size). However, the results were not published at the village level, but only at the town level. Thus, the reality was that the definition of a small area was the town . And this meant that sampling of the villages in a town could have been considered, provided that all the towns were covered. This observation led to the treatment of the towns as domains and sampling of the villages in the 1991 census, as explained in the next subsection.

The major BAS surveys use the provinces as domains, the villages as primary sampling units and the households as ultimate sampling units. However, because of the high sampling errors of the provincial estimates, only regional and national level estimates are publicly released. This raises a number of interesting unanswered questions: (a) Since the nonavailability of provincial level estimates has not triggered any strong protests by the central and provincial government planners nor censure of the statistical system, does it mean that the surveys could otherwise be redesigned to have the regions as domains and consequently, small area in the agriculture censuses can be the provinces?[2] (b) Given the level of resources available to BAS, can the high province level sampling errors be due to the inherently high variability of agricultural variables, or are they due largely to the inefficiency of the sampling strategy being used in the surveys?

The BAS surveys need village level frames in each province, with the censuses the ideal sources. However, for a number of reasons, the (unpublished) village level censuses data had not been available to the BAS, so that the latter had been resorting to creating village frames from information gathered independently from the village heads[3]. Moreover, while only the censuses

---

[2] Indeed, the statistical system compiles economic accounts for the regions and the country, but not for the provinces.

[3] Villages are called Barangays. From time to time, in an operation called Barangay Screening Survey (although it is a complete enumeration), BAS obtains frame information such as number of households, crop areas and livestock numbers from all the 40,000+ Barangay heads.

were able to provide town level data, the significant discrepancies between the censuses and BAS survey results observed even at the regional and national levels cast some doubt on the trustworthiness of the censuses as sources of small area statistics. This can be seen in Table 1 for corn and coconut, the two most important crops after rice. The relative differences are very high and go in either direction at the regional levels. The survey estimate of corn area in the country is 30% higher than that found in the census. The divergences for smaller areas and less prevalent crops will be even bigger.

Table 1. 1980 Survey(S) and Census(C) Estimates of
Corn and Coconut Areas ('000 Hectares) [a]

Regions/	Corn			Coconut		
Country	S	C	100x(S-C)/C	S	C	100x(S-C)/C
1	55	38	46	[c]	[c]	[c]
2	283	134	112	[c]	[c]	[c]
3	8	10	-12	[c]	[c]	[c]
4 [b]	225	95	149	543	633	-14
5	179	114	58	351	519	-32
6	71	95	-25	94	81	16
7	473	370	28	158	139	14
8	187	64	190	334	407	-18
9	252	235	7	411	352	17
10	285	414	-31	368	258	43
11	738	498	48	561	430	30
12	441	401	10	260	166	56
Philippines	3199	2467	30	3103[d]	2996[d]	4

[a] Totals and percentages may not compute due to rounding.
[b] Estimates for the National Capital Region are included in Region No. 4.
[c] Coconut is a minor crop in these regions.
[d] Includes smaller areas in regions 1, 2 and 3.
Sources: David, et al [1990a, 1990b].

3.2 Events Leading to the 1991 Agriculture Census

For the 1990 CPH, NSO got a $9 million special appropriation from the Philippine Congress. Then NSO asked for a slightly smaller amount for the CA -- but did not get it. There are other reasons for this rejection besides reducing the public deficit. The need for a CA is not as clear and known as for a CPH, especially to politicians whose backgrounds are rarely agrarian. With continuing fragmentation of farms, the proportion of farm households to all households in developing countries is usually above 50 %, while 2-6 % (and declining) is typical in a developed country[4]. Thus, the escalating costs needed to continue covering an increasing number of farm households could bring full enumeration agriculture censuses beyond financial and political feasibility in many developing countries.

After some assurance that roughly one third of the originally requested amount could be raised from alternative sources, a group of concerned statisticians and agriculture policy makers discussed ways of saving the CA. The consensus that emerged included combining NSO and BAS field resources for the field work, drastically restricting the questionnaire to structural

[4] Compare, for example, the 4.8 million estimated number of farm households in the Philippines in 1991 to the 2.1 million farms enumerated during the 1987 US Census of Agriculture.

variables such as land use and inventories of farm assets, employing more extensive sampling than in previous CAs, and postponing the reference year from 1990 to 1991 to buy some time for the necessary changes. In particular, we had six months to come up with the sample before the start of the field work.

3.3. Sampling Scheme

It was decided that the 1991 CA should provide separate estimates for the 1600 towns. This was deemed singularly important since the CAs are the only sources of town level statistics on a national scale. The sampling strategy, however, should deliver certain levels of precision, e.g. CV less than 10% for the majority and less than 20 % for eighty percent of the towns, at least for the major structural variables like total farm area and areas under the principal crops. Moreover, the census sample should be able to serve as frame for the ensuing quarterly agricultural sample surveys; and if no other data sources were to be used for that purpose, it is natural to adapt two-phase sampling, with census and surveys as first and second phase samples, respectively. This was the approach followed, another advantage of which, as mentioned previously, is its potential to reduce the discrepancies between census and surveys results.

With complete village level data files from the 1980 CA and new village lists and advance household and population estimates from the 1990 CPH as basis, comparative experiments were done on different strategies; i.e., for every candidate sampling scheme and estimation procedure, draw (by computer) all samples and compute the "exact" error rates (CVs), using the 1980 CA data as proxy for the 1991 target data. The computations were carried out on five important variables, namely total farm area (TFA) and total areas under field crops, rice, corn and coconut. The relative efficiencies of the different strategies were compared through the cumulative distribution functions of the town CVs.

Based on the results of the experiments, the following sampling procedure was adopted for the 1991 Philippines CA [David, 1993a]: in each town, arrange the villages in descending size of TFA in 1980, draw the biggest village with certainty, take half of the remaining villages using simple systematic sampling, and enumerate completely all the farm households within. The large estates and corporate farms were not included in this scheme, as these were enumerated completely.

It may be useful to cite briefly some other reasons behind certain elements of the scheme. (a) We examined the biases in simple systematic sampling (using 1980 data as proxy), found the levels trifle, and came to the decision that the simplicity of the scheme outweighs the complications of further reducing/eliminating bias, e.g. by substituting circular systematic sampling. (b) Why not sample the farm households? The average size of a farm holding in the Philippines is just a little over two hectares. In the previous censuses, households with less than one-tenth hectare or less than a prescribed number of livestock and poultry (with weights assigned to types) were not covered. Moreover, only large farms (defined as at least five or more hectares in 1980) were completely enumerated and a one in twenty sample was drawn from the small farms. The 1980s and the present decade saw an increased concern for reducing poverty, and along with it came the realization that statistical information on the poor is weak. The policy decision to enumerate in the sample villages all households that were engaged in some agricultural activity during 1991 was an attempt to improve information on the poorest of the rural poor. (c) Why not pps sampling? The main reason was to avoid complications in the ensuing sample surveys that will be the responsibility of BAS. Also, NSO management reaction to the suggestion to use pps sampling in the CA was a preference to be somewhat economically inefficient rather than complicate operations and data processing. Systematic sampling of an ordered population gives up some efficiency to pps sampling in exchange for robustness. (d) One systematic sample will not give a valid estimate of the sampling error. This was resolved partly by a decision to enumerate the complements to the census samples in four pilot provinces that the NSO had been working on to test the instruments and procedures being developed before the disapproval of the proposed census budget. Actual sampling errors can be observed in these provinces.

3.4 Insights from the Four Completely Enumerated Provinces

Past information is routinely used to plan future data collection, expecting that correlations between the two sets continue to be strong. In developing countries, the correlations often turn out weaker than anticipated, probably because changes in agriculture are faster than generally perceived, and definitely because of the upsurge of non-sampling errors in massive operations. The village level correlations between 1980 total farm area and five characteristics from the 1991 CA in the four full enumeration provinces are shown in Table 2. The correlation coefficient of total farm area against itself 11 years after does not go beyond 0.7. The lower correlations with the other characteristics mean that the precisions of the estimates for these characteristics will not live up to expectations.

Table 2. Correlations Between Village Values of 1980 Total Farm Area
and 1991 Characteristics, by Province

Province 1991 Characteristics	Bukidnon (N=451)	Isabela (N=1054)	Laguna (N=670)	Marinduque (N=218)
Total Farm Area	0.61	0.65	0.72	0.70
Rice Area	0.30	0.53	0.52	0.36
Corn Area	0.49	0.30	0.11	0.49
Banana Hills	0.34	0.07	0.29	0.55
Coconut Trees	0.13	0.09	0.39	0.49

Since census results take too long to process, it is standard procedure in many developing countries to release advance tabulations of main structural variables based on summary tables from the field. In the Philippines 1991 CA, village totals were tabulated by field staff from the main questionnaires (Form 2) and entered in another document (Form 4). In the interest of time, the latter was not subjected to rigorous editing and validation. Table 2 was constructed from Form 4, hence it is entirely plausible that the correlation values were made lower by the undetected errors in the input data; unfortunately, for reasons of easy and early availability, it is usually Form 4 type data that are used to design intercensal surveys.

Forms 2 and 4 data are supposed to be identical. A comparison, however, revealed many discrepancies mostly due to addition and transcription errors from the former to the latter, but also encoding errors in both forms. An example is shown in Table 3. The total farm area in one province (Laguna, Table 3a) showed a 9.8% difference, which was traced to one town (Santa Maria, Table 3b), and then to an encoding error in Form 2 of one village (Tungkod, Table 3c). Some of the more serious response or recording errors that escape validation checks of the encoded data may still be detected during the processing stage: a serendipitous plus for a sample census is that large sampling errors (when computed) indicate the presence of large individual values that often turn out to be non-sampling errors instead of true outliers. So many gross errors in the basic data were detected this way in the four provinces, to the extent that, after correction, the town level estimates for the main structural (area) variables would go from unacceptable to generally acceptable.

Town statistics on the minor crops exhibited high sampling errors, so that only estimates at the provincial or regional level should be released to the public. Of course, what is minor in one location could be major elsewhere (see e.g. corn in Table 3), hence a practical publication decision can be made only after the review of the initial tabulations. Minor crops are characterized by being unevenly distributed, found in a few large tracts among multitudes of very small ones, or are 'rare events'; this inherent very high variability often implies that nothing short of complete enumeration will ensure that sampling error will be controlled. Thus, in the future, NSOs in consultation with users may do well to ponder whether to continue complicating CAs by including minor crops in all questionnaires, or whether these could be better dealt with through specialized studies in specific areas.

There is another kind of non-sampling error, the source of which can be traced to variations in the field interpretation of concepts, which in turn is influenced by the type of statistical system found in the country. The deviations between census and survey results illustrated in Table 1 are partly of this kind, and as mentioned before, partly due to differences in the methods used. Another example is banana counts from the 1991 CA which in some provinces were much higher than the estimates from the intercensal surveys. It turned out some CA field personnel were recording the number of banana plants or suckers instead of hills. These problems are prone to happen in decentralized statistical systems where there is bound to be a difference in the depth of understanding of agriculture between an NSO that deals with the subject during a census only and a statistics bureau in the agriculture ministry that has to update the country's agriculture database continuously.

Table 3. Comparison of Otherwise Identical Form 2 and Form 4 Data

a. Province Totals

Province/Form		Area in ('000) Hectares Total	Rice	Corn
Bukidnon	Form 2	332.2	69.6	256.9
	Form 4	331.8	71.7	260.0
	% Difference	1.3	-3.0	-1.2
Isabela	Form 2	261.1	227.3	160.8
	Form 4	263.9	232.3	161.9
	% Difference	-1.1	-2.2	-0.7
Laguna	Form 2	96.8	36.9	1.2
	Form 4	87.3	36.7	1.0
	% Difference	9.8	0.4	19.9
Marinduque	Form 2	43.6	10.4	1.8
	Form 4	43.0	10.2	1.8
	% Difference	1.4	1.5	-0.4

b. Laguna Province

Town	Form 2	Form 4
Alaminos	4,038	3,936
Bay	2,601	2,638
Binan	1,266	1,237
Cabuyao	2,297	2,247
Calamba	6,141	6,209
Calauan	3,869	4,190
Cavinti	5,774	5,988
Famy	1,799	1,878
Kalayaan	3,147	3,132
Liliw	2,364	2,364
Los Banos	1,148	1,136
Luisiana	3,527	3,525
Lumban	1,567	1,563
Mabitac	1,746	1,747
Magdalena	1,946	1,945
Majayjay	2,318	2,335
Nagcarlan	5,605	5,596
Paete	957	949
Pagsanjan	1,876	1,880
Pakil	1,783	1,841
Pangil	1,827	1,817
Pila	1,794	1,741
Rizal	556	573
San Pablo City	10,962	10,782
San Pedro	979	978
Santa Cruz	1,862	2,111
Santa Maria	16,201	6,137
Santa Rosa	2,223	2,250
Siniloan	3,310	3,333
Victoria	1,294	1,298
Laguna	96,779	87,319

c. Santa Maria Town

Village No.	Form 2	Form 4
1	98	98
2	328	322
3	325	326
4	208	208
5	146	146
6	170	171
7	209	209
8	374	371
9	446	449
10	79	80
11	255	255
12	205	205
13	220	220
14	76	76
15	260	260
16	177	111
17	249	249
18	141	142
19	300	293
20	266	266
21	325	325
22	74	74
23	697	699
24	321	321
25 (Tungkod)	10,251	261

4. A Proposed Strategy for the Intercensal Surveys

4.1. Current and Proposed Designs

The BAS updates the Philippines agriculture database through a system of surveys , some of them quarterly, each survey covering a crop or group of crops. These surveys share a common sampling frame (see footnote 3) and sampling procedure, namely, stratify the villages in each province, draw a sample without replacement with probability proportional to size, and select a simple systematic sample of farm households to interview. However, the stratification and sample villages varied between surveys, e.g. between the rice survey and corn survey. The estimation procedure is design-based and straightforward, with no attempt to use auxiliary variables or composite estimators to link current and past rounds. Despite best intentions, plans to partially replace sample villages had not been implemented; instead a wholesale change happened when the entire village sampling frame was updated or when there was a design change.

In 1993, we proposed a redesign that utilized the 1991 CA exclusively and kept the provinces as domains: (a) Stratify the CA sample villages by putting the largest (that was drawn with certainty) in each town into the first, and divide the remaining into *H-1* strata. (b) Draw *R* replicated samples each of size *H*. This can be done either by choosing one village in each stratum using a specified selection mechanism, replacing and repeating the process *R* times, or equivalently, drawing *R* villages with replacement in each stratum. The result is a second phase sample from the (first phase) census sample. Sampling with replacement ensures independence of the replicate samples with respect to the second phase.

4.2 A Generalized Estimation Procedure

The estimation outlined below is adaptable to any selection probabilities used to draw the second-phase sample villages and to any sampling scheme used at the household and lower levels. This is an important advantage when different procedures and samples have to be used for the rice and corn surveys, coconut surveys, livestock surveys, etc. It is assumed, however, that the first-phase psus were drawn with equal probabilities, except the largest ones that were drawn with certainty and which were grouped into the first stratum at the second-phase. Details of the estimation theory have been developed in David and Maligalig [1993].

Let $y_{h(r)}$ be the design unbiased estimate of the r^{th} sample village total in stratum h and $f_{h(r)}$ the second phase selection probability of the same village. Then $y_{1(r)} / f_{1(r)}$ is an unbiased estimate of the stratum *1* total and $y_{h(r)} / \tfrac{1}{2} f_{h(r)}, \quad h = 2, \ldots, H$ are the corresponding estimates for the remaining strata, with the factor $\tfrac{1}{2}$ reflecting the first phase inclusion probability. An estimate of the population total from the r^{th} replicate sample is

$$y_{(r)} = \frac{y_{1(r)}}{f_{1(r)}} + \sum_{h=2}^{H} \frac{y_{h(r)}}{\frac{1}{2} f_{h(r)}} \qquad (1)$$

The average of the *R* replicate samples

$$y = \frac{1}{R} \sum_{r=1}^{R} y_{(r)} \qquad (2)$$

is also unbiased and the usual estimator of the variance is

$$\mathrm{var}(y) = \frac{1}{R(R-1)} \sum_{r=1}^{R} (y_{(r)} - y)^2 \qquad (3)$$

which is known to be unbiased when the replicate samples are truly independent. However, the independence is lost under two phase sampling [Wolter, 1985]. For the specific sampling scheme proposed here, the variance of y consists of three terms,

$$\mathrm{Var}(y) = \frac{1}{R^2} \sum_{r=1}^{R} \sum_{h=1}^{H} \left\{ \sum_{i=1}^{N_h} \frac{Y_{hi}^2}{w_{hi(r)}} - Y_h^2 \right\} - \frac{1}{R} \sum_{h=2}^{H} \left(Y_{h[1]} - Y_{h[2]} \right)^2 + \left(Y'_{[1]} - Y'_{[2]} \right)^2 \qquad (4)$$

where N_h denotes the size of stratum *h*; Y_{hi} the *i*th village total in stratum $h; Y_h$ the stratum *h* total;

$w_{hi} = f_{1i(r)}$ for stratum *1* and $\frac{1}{2} f_{hi(r)}$ otherwise; $Y_{h[1]}$ is the total of the first phase sample villages in stratum *h* and $Y_{h[2]}$, the corresponding total of the complement sample; $Y'_{[1]} = \sum_{h=2}^{H} Y_{h[1]}$ and $Y'_{[2]} = \sum_{h=2}^{H} Y_{h[2]}$. The third term, which is the squared difference between the first phase systematic sample total and its complement, is also the covariance between estimates from the replicate samples. It is a consequence of the non-independence of the replicate samples under two phase sampling. Its negative value turns out to be the bias in the variance estimate (4), since it can be shown that

$$\mathrm{E}(\mathrm{var}(y)) = \mathrm{Var}(y) \; - \; \left(Y'_{[1]} - Y'_{[2]} \right)^2 . \qquad (5)$$

4.3 Practical Problems of Adoption Despite Advantages for Developing Countries

As described above, Indonesia and Nepal are two countries that had ran CAs that were stratified multistage sample surveys. In the former the intercensal surveys were subsamples of the CA, with no indication, however, that these are being treated as second phase samples. In the latter the intercensal surveys were drawn independently of the CA from the same so-called census sampling frame. Hence, one would hope that adapting the sampling strategy proposed here to these countries would not meet serious difficulty since no change in fundamental principle nor additional resources is required.

In the Philippines, however, while the 1991 CA was conducted as suggested, the BAS has yet to use the results to redesign the intercensal surveys as proposed above. Hindsight points to the major impediments: a budget earmarked mainly for staff salaries and continuation of the regular surveys, no permanent research and development staff or activity, lack of technical capacity especially in survey sampling theory, concern for consequences of a new method leading to results that are very different from the historical time series, and natural inertia towards significant change. These situations tend to be present in other Asian developing countries as well; hence, often an external financial and technical assistance becomes the strongest catalyst to change.

The proposed strategy provides economical and practical solutions to the enduring problems confronting developing country statistical systems. It solves the problem of sampling frames for agricultural surveys. We may see more error estimates being computed which few countries are able to do at the moment. A rotation or replacement of $r < R$ replicates from one survey round to the next does not unduly complicate the field and estimation procedures. Continuous updating of weights in between censuses can be done through relisting of units within villages (primary sampling units) in a subset of the replicates samples, with the subset changing from round to round. Important applied research, such as comparing data collection or training methods, can be easily built into the survey operation. Phased processing of the data in subgroups of replicates will help meet varied user requirements, like when the agriculture minister wants advance estimates within one month from the survey, the planning minister three months as inputs to the quarterly national accounts, and final estimates six months after. Likewise, in the event that only a portion of the survey appropriation is actually released, which is not that uncommon in developing countries, reducing the sample to $r < R$ randomly chosen replicates preserves the distribution of work loads and changes only the accuracies and stabilities of estimates but not the estimators.

Last but not least, there is the constant dilemma of choosing between a more economical multisubject survey or more efficient specific surveys. Should it be an integrated crops and livestock survey or separate crops and livestock surveys, a rice and corn survey versus a rice survey and a corn survey, and so on. The first approach leads to less precise estimates; however, a developing country agriculture statistics office can never have enough resources to run too many surveys. The proposed sampling strategy tips the balance towards multisubject surveys, in the sense that some of the potential precision lost at the design stage could be recovered at the estimation stage, by using variance reducing techniques like ratio or regression estimation in the replicate estimator (1). The estimators (2) and (3) still apply, hence the estimation process remains sufficiently simple for the limited capability in the statistics office.

5. Concluding Remarks

Some developing countries had not done complete enumeration agriculture censuses, having found them financially or technically infeasible. We predict that more countries will increasingly find themselves in the same situation. In particular, the present economic and financial turmoil in East and Southeast Asia can put the 2000 round of CAs in these countries at risk, and only drastically reduced costs through extensive sampling may be able to save them from being postponed or canceled.

There have been significant deviations between the reasons for doing massive CAs and the actual uses of the results. They had not always been used to design the ensuing surveys, which when coupled with the use of different data collection methods, tended to lead to conflicting results between the censuses and the surveys. There were cases where the low accuracy of the small area data from CAs and the inordinate delay in their release limited their usefulness.

In this paper we propose replacing a massive or complete CA with a first phase sample, from which second phase subsamples are drawn directly for the ensuing regular surveys. The first phase sample should meet the demand for small area data, keeping in mind, however, that a review of past experience in planning and decision making in the developing country could lead to redefining small area to a higher administrative level. The sampling procedure should match the technical expertise available in the national statistical agency. The sampling strategy at the second phase should likewise be appropriately simple yet reasonably efficient. It should also be sufficiently robust to serve the needs of individual modules in a system of agricultural surveys and survive budgetary cuts and similar uncertainties faced by statistical systems in developing countries.

References

Asian Development Bank (1997), *Key Indicators of Developing Asian and Pacific Countries*, Vol. 28, Oxford University Press.

Bangladesh Bureau of Statistics (1997), "Preliminary Report: Agriculture Census 1996," Dhaka, Bangladesh.

Bastola, T.S. (1991), "Frames Used for the Different Sample Censuses of Agriculture in Nepal", Presented at the International Expert Consultation on Agricultural Censuses, 4-9 October 1991, Beijing, China.

Central Bureau of Statistics, Nepal (1994), "National Sample Census of Agriculture of Nepal 1991/92: Reliability, of Results", Kathmandu.

Central Bureau of Statistics, Indonesia (1995), "Metode Sampling Sensus Pertanian 1993", Jakarta.

David, I.P., A.N. C. Mendoza and D.C. Besa (1990a), "Evaluation of Philippine Corn Statistics" *Journal of Philippine Development*, Vol. XVII, No. 1, pp. 33-66.

__________________________ (1990b), "The Quality of Philippine Coconut Statistics" *Journal of Philippine Development*, Vol. XVII, No. 2, pp. 249-294.

__________ (1993a), "Sampling Strategy for the 1991 Philippines Census of Agriculture", Unpublished Notes.

__________ and D.S. Maligalig (1993b), "A General Purpose Agriculture Survey Design Using the 1991 Agriculture Census as Sampling Frame, Parts I and II", Unpublished Notes.

Economic and Social Commission for Asia and Pacific (1997), Report of Working Group of Statistical Experts Meeting, 11-14 November 1997, Bangkok, Thailand.

FAO (1995), "*Programme for the World Census of Agriculture 2000*", Statistical Development Series 5, Rome.

Sukmadi, B. (1991), "Agriculture Census 1993: Census Frame and Cartography". Presented at the International Expert Consultation on Agricultural Census, 4-9 October 1991, Beijing.

Wirosardjono, S., K. Hardjopertomo and T. Sugito (1991), "Review of Field Operations and Organization of Agricultural Census in Indonesia", Central Bureau of Statistics, Jakarta.

Wolter, K. (1985), *Introduction to Variance Estimation*, Springer-Verlag.

Zhu Xiangdong (1997), "Strategy and Method Selection for the First National Census of Agriculture in China", Presented at the International Consultation on China's Agriculture Census, January 1997, Beijing.