Using Principal Component Scores as Stratification Variable: An Alternative to Multiple Frame Sampling Methodology

Erniel B. Barrios[1] and Kevin Carl P. Santos[1,2]
[1]School of Statistics, University of Philippines, PHILIPPINES
[2]Corresponding Author: Kevin Carl P. Santos,
e-mail: kevincarlsantos@gmail.com

Abstract

In the Philippines, several studies pointed out that samples from Rice and Corn Production Survey no longer suffice to provide "acceptable" estimates of Livestock and Poultry statistics. Simulation studies considered the use of inventories of different animals as stratification variables or size measures. However, the results showed that the use of a particular animal type would not necessarily yield efficient estimates for other animal inventories. This leads to the use of multiple frame sampling methodology as an ideal data collection method which entails high costs. This paper proposes the use Principal Component (PC)-based scores as the stratification variable as an alternative to multiple frame sampling to lower the costs because a PC summarizes information contained in a set of auxiliary variables (e.g. inventory for different animal types); nevertheless, PC's also give premium to variables with large variability. Thus, the relatively "rare" animals, i.e. those present in very few *barangays* only, can potentially sway the PC to their advantage at the expense of the more "common" and, perhaps, the more important animal types. Hence, the authors recommend that only the inventories of the more important animals are included in the generation of the PC scores that will be used to stratify the population of interest. Various sampling experiments are performed using *barangay* level data on different animal inventories in order to determine the efficiency of the estimates using the proposed method.

Keywords: principal component analysis, livestock and poultry statistics, stratified sampling

1. Introduction

The Philippine Rice and Corn Production Survey (RCPS) is based on a sampling design developed using the 1991 Census of Agriculture and Fisheries (CAF) as the sampling frame. The current Backyard Livestock and Poultry Survey (BLPS) uses one of the four replicates in RCPS based on a two-stage sampling with the *barangays* as the primary sampling unit (PSU) and the farming household as the secondary sampling unit (SSU).

PSU selection in RCPS is dependent of the classification of the provinces based on *palay* of corn concentration. For pure *palay* and pure corn provinces (those provinces whose region produce are either *palay* only or corn only), Replicate 1 consisting of ten (10) sample *barangays* are covered for BLPS. For other provinces (neither corn nor *palay* is the major produce), only five (5) sample *barangays* are drawn for the BLPS.

In the selection of sample households (SSU), the BLPS incorporates non-farming household, in addition to farming household of the RCPS. A non-farming household is defined as any household in which no member operates a farm; or has a member operating agricultural land that does not qualify as a farm; or has given up operating its farm and nobody from the same *barangay* has taken over.

There is cost advantage in using rice and corn production area-based sampling design for poultry and livestock surveys. However, the correlation between rice and corn production areas and livestock and poultry inventory is no longer supported by the 2002 Census of Agriculture. Thus, rice and corn production areas may no longer relevant as auxiliary variables for poultry and livestock surveys (Barrios, 2009).

Aside from presence/absence of some animals in rice- and corn-producing *barangays,* the coefficient of variations (CVs) further provide evidence on the inappropriateness of rice and corn sample *barangays* for poultry and livestock surveys. The coefficient of variation at the *barangay* level for quail for meat inventory, the highest among all poultry and livestock, is 269% higher than the CV for rice production area. Chicken for eggs inventory and carabao inventory also have much larger CV, about 230% times the CV of rice production areas (at the aggregate *barangay* totals). This means that the sample selected for an indicator that is less variable is also used to estimate another indicator with much larger heterogeneity. From the concept of design effect, quail for meat inventory, chicken for eggs inventory, carabao inventory, among others, should be taken in more sample *barangays* than what is needed to estimate rice production (Barrios, 2009).

Rice and corn production areas are no longer correlated with animal inventories as it used to be. Barrios (2009) noted that the validity of the use of rice and corn samples as source of information for animal inventory is rather doubtful. At the least, this is easily exposed to selection bias and assurance of underestimation since many animal-raising *barangays* are different from those who plants rice and corn. Furthermore, as the weights for rice and corn samples are applied to inflate animal inventories, this will not only yield tremendous sampling error and bias, but also an erroneous distribution across the provinces and the regions since it will be the rice and corn production areas that will be reflected and not those of the different animals.

Authors conducted simulation studies that consider inventory of different animals as stratification variable/size measure/auxiliary variable. The results are pointing out that the use of inventory for a particular animal type will not necessarily produce efficient estimates for another animal type. The implication is that the ideal design in collecting poultry and livestock statistics should use multiple frames, i.e., different frame for different animal types. This will of course entail geometric progression in cost. To mitigate the cost-

escalation induced by multiple frames, principal component of auxiliary variables from the different frames can be used as a stratification variable. A principal component should summarize information contained in a set of auxiliary variables (e.g., inventory for different animal types). However, principal components also give premium for variables with larger variation (Jolliffe, 2002). Thus, for relatively "rare" animals, i.e., those present in very few *barangays* only, they can potentially sway the component to their advantage, at the expense of the more "common" and perhaps the more important animal types. In using principal component as an auxiliary variable, it is recommended that only the inventory of the most important animals will be included.

2. Results and Discussion

Taking into account the issue of relevance of the rice and corn production survey design in the estimation of livestock and poultry inventory, a new Backyard Livestock and Poultry Survey (BLPS) design will be developed. Sampling experiments were done using data from two provinces to assess the statistical viability of the different sampling strategies. Pangasinan is chosen among the top poultry and livestock producers, while Davao del Norte is chosen to represent low livestock and poultry producing provinces. Two possible sampling frames were considered namely: 2002 Census of Agriculture, and Avian Population Survey (APS) and Livestock and Poultry Survey (LPS) Databases. The sampling rate is fixed at 5% or 10% of the *barangays* within each stratum.

2.1 Sampling Strategy 1

This sampling strategy uses the LPS and APS databases as the sampling frame. The *barangays* within the province (treated as the domain) are stratified using the following as stratification variables, such as Cattle Inventory, Swine Inventory, Chicken Inventory and Principal Component (PC)-based scores of the inventory of the "more important" animals. Only the "more important" animal types were included in the construction of the component (hog, chicken, carabao, and goat) since other animals that are less common exhibit very high variances that could dominate in the extraction of the PC (Jolliffe, 2002). Inclusion of high variance in PC extraction put "too much" importance on those variables hence, the stratification could possibly address only the heterogeneity of these "less important" variables, at the expense of the "more important", i.e. more common variables.

The sample *barangays* are chosen with each stratum using either Simple Random Sampling (SRS) or sampling with Probability Proportional to Size (PPS). In case PPS is used, the stratification variable is also used as the size measure. Three stata were formed (small, medium, large-inventory *barangays* of poultry and livestock). From each stratum,

sample *barangays* were drawn using probability proportional to size sampling with PC-based scores as the size measure in each stratum. This will ensure that the *barangays* who are most likely producing a variety of livestock and poultry types will be included.

In Pangasinan, with cattle inventory as a stratification variable, there is a very high bias for chicken inventory in both SRS and PPS. There is however lower bias in the inventories of cattle, swine, carabao, goat, and surprisingly in duck inventory, especially in PPS. The CVs are mostly <15% when 10% sampling rate is used. Using chicken inventory as the stratification variable, CV for chicken is low, but bias is quite high. Furthermore, SRS is comparable to PPS. With swine as stratification variable, bias for chicken is very high, although biases for other animal inventories are low. CVs are low in PPS. Using the PC-based scores as the stratification variable and PPS as the sampling scheme, the percent bias is low for most animals except for sheep and duck and CV is generally low for all animal inventories except for sheep as shown in Table 1.

In Davao del Norte, a province producing less poultry and livestock, cattle inventory as a stratification variable yield high bias especially for duck and sheep, and CVs are generally high. With chicken inventory as the stratification variable, biases are within acceptable range, but CVs are high. Using swine inventory as stratification variable, bias are low for SRS, but there are many high CVs. Table 1 presents the case using PC as stratification variable which yielded low bias in PPS except for sheep inventory and CVs are high only for some animal inventories such as duck and sheep.

Table 1. Percent Bias and CV of the Estimated Animal Inventory using PC-based scores as Size Measure (Sampling Rate = 10%)

	Pangasinan ("Producing")				Davao del Norte ("Non-producing")			
Animal	Inventory	Estimate	% Bias	CV	Inventory	Estimate	% Bias	CV
Cattle Inventory	131723	136617	3.72	9.70%	9753	10731	10.03	17.70%
Swine Inventory	217820	204492	6.12	10.07%	92770	75508	18.61	9.70%
Chicken Inventory	2468450	2565270	3.92	7.10%	826459	736228	10.92	15.80%
Carabao Inventory	85164	83758	1.65	9.49%	27102	34295	26.54	9.20%
Duck Inventory	229098	143376	37.42	12.19%	67143	61931	7.76	26.80%
Goat Inventory	209977	221461	5.47	7.08%	44511	43394	2.51	8.70%
Sheep Inventory	940	224.792	76.09	52.41%	654	1073.215	64.1	63.50%

2.2 Sampling Strategy 2

This sampling strategy uses the LPS and APS databases as the sampling frame. The

barangays within the province (treated as the domain) are drawn directly using PPS using the following size measures, such as Cattle Inventory, Swine Inventory, Chicken Inventory, and Principal Component (PC)-based score of all inventory.

Pangasinan, a poultry and livestock-producing province yield different characteristics of the estimates from Davao del Norte, a "non-producing" province. In Pangasinan, PC as a stratification variable yield low biases and low CVs for all animal inventories, except for chicken inventory where CV is high. On the other hand, in Davao del Norte, with PC-based scores as a stratification variable, both bias and CVs are high.

Stratifying by cattle inventory in Pangasinan yield high bias for chicken and sheep inventories but CVs are low in PPS. In Davao del Norte, low bias (except in sheep inventory) and relatively higher CVs are observed. Stratification by swine inventory yield better estimates in Pangasinan, except for some animals, but in Davao del Norte, biases are higher for many animal inventories and CVs are generally higher. With chicken inventory as stratification variable, both bias and CVs are higher for all animals in both provinces.

2.3 Sampling Strategy 3

For this strategy, household data from 2007 Census on Population is used. The *barangays* within the province are stratified as in Strategy 1 using the following stratification variables such as Total Household and Agricultural Household. The *barangays* are then selected using either SRS or PPS (with size measure the same as the stratification variable).

In Pangasinan, a livestock and poultry producing province, lower bias (except in chicken and sheep inventories) and lower CVs (except in sheep inventory) are observed using either total households or agricultural households as the stratification variable. SRS and PPS are better at 10% sampling rate. Estimates of inventories in Davao del Norte yield higher bias and CVs in many animals whether total or agricultural households are used. Furthermore, SRS is better than PPS.

2.4 Sampling Strategy 4

Strategy 4 draws from the recommendations of the Barrios (2009) study. The recommendation was to adopt multiple frame survey designs for three groups of animals. Group 1 includes cattle, carabao and horse with cattle inventory as the stratification variable and the size measure in unequal probability of selection. Group 2 includes hog, goat, and other livestock with hog inventory as the auxiliary variable. Group 3 includes chicken, duck, and quail with chicken for meat inventory as the auxiliary variable. In all cases, Census of Agriculture in 2002 was used as the sampling frame.

The survey domain remains at the provincial level. The design is two-stage, with stratification in the first stage units (*barangays*). Farming households constitutes the secondary sampling units (SSU).

In the first stage, the *barangays* within the province are stratified (3-4 strata) based on the animal inventory (cattle for Group 1, hog for Group 2, and chicken for Group 3). Within each stratum, *barangays* are selected with probability proportional to total animal inventory (PPS). This will ensure that the sample will include those *barangays* with the most number of animal inventories, hence, better access to poultry and livestock information. In rice area based sampling design, there are many sample *barangays* without any household raising these animals; hence, the distribution is highly skewed towards zero, considerably inflating the bias and variances of the estimates.

Once the sample *barangay* is chosen, farming households are selected at random within the sample *barangay*. For Groups 1 and 2, 5 farming households per *barangay* are recommended. For Group 3, 10 farming households should be selected per *barangay* (Barrios, 2009).

Since quail-raising is relatively "rare" compared to other animals, samples in *barangays* under Group 3 can be expanded whenever some quail farmers are observed. The adaptive samples can enhance the estimates of quail inventory.

For animals requiring quarterly estimate, a survey with last quarter as reference can be done once a year. The inventory for other quarters of the year can be imputed from the panel data models (model-based estimation). For animals where semestral estimates are needed, a survey can be conducted in one semester, estimate for the other semester can be imputed from the estimates generated from the semester where a survey was done.

3. Conclusions

The use of principal component score to abate the problem of multiple frames is proven to be effective in livestock and poultry survey. While it is ideal to use individual animal inventory as a stratification variable/size measure, this may blow up cost of survey operations since producing *barangays* for different animals may not be similar. The animal inventory to be included in the computation of principal component score should be chosen among the more important ones and has lower variation. The principal component loadings are generally higher among variables exhibiting large variance.

References

Barrios, E. (2009). Livestock and Poultry Survey: Assessment and Recommendations, Unpublished Manuscript. [funded by Philippine Bureau of Agricultural Statistics]

Jolliffe, I. (2002). *Principal Component Analysis*, New York: Springer Verlag. 2nd ed.