Improvement of Sample Design by Using Results from the Agricultural Census

Michael Steiner
Jeffrey Bailey
Jeffrey Beranek
Joseph Parsons

National Agricultural Statistics Service, U.S. Department of Agriculture

Abstract: An overview of current Chinese agricultural data collection and sampling methods will be contrasted with a pilot sample survey program that is being developed using the results from the recent Chinese agricultural census. Under the current design, separate samples of rural households are often utilized to meet national, provincial, and local statistical data needs. The pilot sample survey program will utilize a replicated, multi-stage, multivariate probability proportional to size (MPPS) sample design. The design uses the Census of Agriculture data to focus samples in villages where high concentrations of the items of interest existed at the time of the census. The utilization of multiple variables to create measures of size allows specific crops, livestock and economic items to be targeted for sample selection.

1. Introduction

The Chinese Census of Agriculture has provided a wealth of detailed crop, livestock, and rural household statistics. In addition to their use as measures of agricultural and rural economic activity, these data may be used to efficiently and effectively focus future sample surveys. Typically, governments conduct periodic censuses of all population units creating a baseline data set to facilitate development of future sample surveys. Information collected during the census can also be used to reconcile the differences between official statistics and census results and as control or auxiliary data for sample surveys conducted between census base periods. Specific details are described by Vogel (2000). As more information or auxiliary data about population units ***are*** available for use in developing a sample design strategy, the more likely that estimates with acceptable levels of precision can be achieved with the use of a sample survey.

Since 1996, the National Bureau of Statistics (NBS) and the National Agricultural Statistics Service (NASS) of the United States Department of Agriculture (USDA) have had a cooperative agreement to exchange information and statistical methodology. It is under this arrangement that work has been done in development of a new sample design procedure for agricultural surveys in China. A sample design for rural households has been developed for use in a pilot survey. Data from the provinces of Guangdong and Zhejiang were utilized in analysis for sample design.

2. National Bureau of Statistics Survey Program

2.1 Scope and Content

NBS is currently conducting two types of surveys. The first type is an agricultural yield survey. Crop yield samples are selected and surveys conducted for winter wheat, corn, rice, sorghum, soybeans, cotton, and sweet potatoes. Specific details are described by Gong Jian Yao and Chen Jixen (1989), Li Ting Jun (1995) and Vogel (1999). The second type of survey is a multi-purpose household survey. The Household Survey utilizes twelve questionnaires, collecting information

on income and expenditures, food consumption, food production, house, land and investment, inventory of personal belongings, and household characteristics. Specific details are described in a recent publication from the National Bureau of Statistics (Food and Agricultural Statistics Centre, 2000).

2.2 Current Sampling Process and Design

In 1984, NBS (then SSB) set up a General Organization of Rural Socio-Economic Survey (GORSES) to conduct rural surveys. Sample counties were selected in 1984, and survey teams were established within all sample counties. Both the crop yield samples and the household samples were selected in several stages. The first stage of selection was a sample of 857 counties out of the over 2,500 total counties and cities at the county level (Zhu, 1998). In the selected counties, 9,000 administrative villages were selected for the Household Survey, and then 70,000 rural households were selected. The crop yield survey is conducted in 18,000 administrative villages within the 857 selected counties. The current samples were selected using two data variables as criteria, total grain production per capita income.

2.3 Improving Sample Design

The NBS indicated that the Census of Agriculture will be used to set up a new rural statistical survey system, and that sampling surveys will be promoted, using data from the Census Of Agriculture (Food and Agricultural Statistics Centre, 1999). NBS has identified a number of major commodities to be included in future survey programs. The new commodities have not been part of the existing survey program. Currently, samples selected for different levels of government (national, provincial, prefecture, and county) are not additive. Provincial level survey results are currently obtained from one sample of villages. Prefecture-level results are currently obtained from a second sample of villages. County-level survey results are currently obtained from a third sample of villages. The three samples are independent and data from the different samples ARE NOT combined, even though data collected from the different samples are the same. The objectives of a new sample design are: **(1) to effectively utilize data from the Census of Agriculture, (2) to select samples using multiple variables which is necessary to support the proposed expanded survey program, and (3) to integrate the statistical needs for different levels of government.**

This paper focuses on obtaining a sample of villages for use in all types of agricultural surveys (both for surveys in the existing program and for new surveys being proposed). After selection of a village sample, some type of a selection of households or fields will be necessary unless data are obtained for the entire sample village. This paper does not investigate selection of households or fields within a village.

2.4 Evaluation of Selected Counties in the Guangdong and Zhejiang Provinces

Agricultural census data from Guangdong and Zhejiang were analyzed to evaluate the current sample design used by NBS for sample surveys. Tables 1 through 4 evaluate county level data for Guangdong and Zhejiang provinces. A distribution is shown for four major data items, showing data for counties in the current NBS survey program versus counties *not* in the NBS

survey program. As the tables indicate, sample counties tend to have much larger quantities of items than non-sample counties. For example, for area of cultivated land in Guangdong Province, 83% of the sample counties have values greater than the mean amount per county, and 90% of the sample counties have values greater than the median amount per county. For area of cultivated land in Zhejiang Province, 76% of the sample counties have values greater than the mean amount per county, and 84% of the sample counties have values greater than the median amount per county.

Table 5 shows the proportion of the provincial total for major commodities accounted for by the counties in the existing Household Survey program and the counties *not* in the existing Household Survey program. For example, in Guangdong Province, 26 % of the counties are sample counties in the Household Survey, accounting for 45% of the cultivated area in the province, 45% of the grain area, 51% of the peanuts and 44% of the vegetables. In Guangdong Province, sample counties account for 53% of the total cattle, 42% of the hogs and 37% of the poultry. In Zhejiang Province, 31% of the total counties are sample counties in the Household Survey, accounting for 45% of the cultivated area in the province, 45% of the grain area, 48% of the wheat area, and 33% of the rapeseed area. In Zhejiang Province, sample counties account for 38% of the cattle, 58% of the sheep and 33% of the hogs. *Table 6* shows the average (mean) amount per county.

Sampling a higher proportion of the large counties has desirable features, such as accounting for a high proportion of agricultural production. However, samples need to be representative of the entire population. Smaller counties may be under-represented with the current sample of counties used for the Household Survey. The variation in the percent of coverage is an issue. For example, in Zhejiang Province, the counties in the existing Household Survey represent 48 % of the wheat produced in the province, while only 33 % of the rapeseed. In Guangdong Province, the counties on the Household Survey account for 53 % of the cattle and only 37 % of the poultry.

This variation in percent raises concerns and questions in the estimation process regarding how data from sample counties are weighted to represent the entire population of counties. NBS does not currently utilize sample selection probabilities in calculating weights for summarization of data from the Household Surveys. Data from the Agricultural Census could be used to calibrate data for the sample counties, in order to address the size distribution situation. However, recent events in China such as major changes in growing patterns could reduce efficiency of a sample selected only in certain counties. For example there has been a recent shift from grain to vegetables in certain areas of eastern China. Does the current sample of counties efficiently capture this change in cropping patterns? Another example is for hogs, which are an important commodity in Zhejiang Province (hog data has been identified as a key variable in future sample selection in Zhejiang Province). Currently, none of the five top hog counties (accounting for an inventory of nearly 2 million hogs) are counties in the Household Survey Program. What effect does changes in hog prices and the omission of the top five hog producing counties have on results such as income from the Household Survey?

Table 1: Cultivated Land, by County, Guangdong and Zhejiang Provinces, 1996

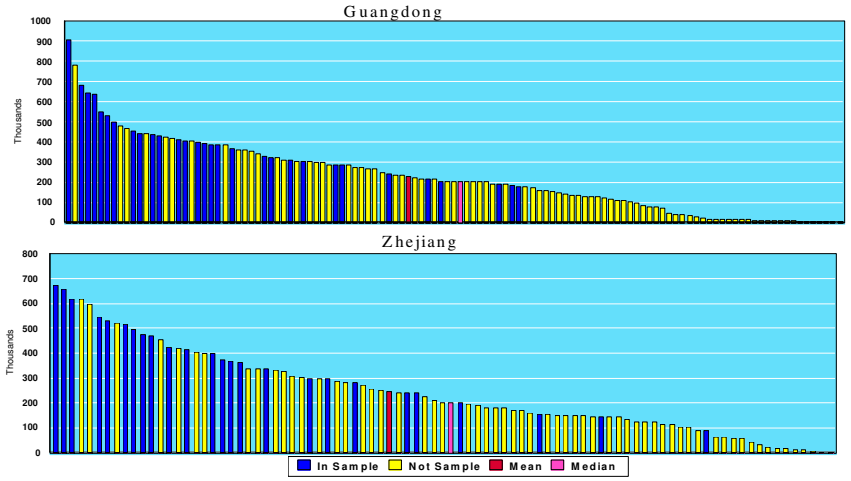


Table 2: Grain Crops: Area Planted, by County, Guangdong and Zhejiang Provinces, 1996

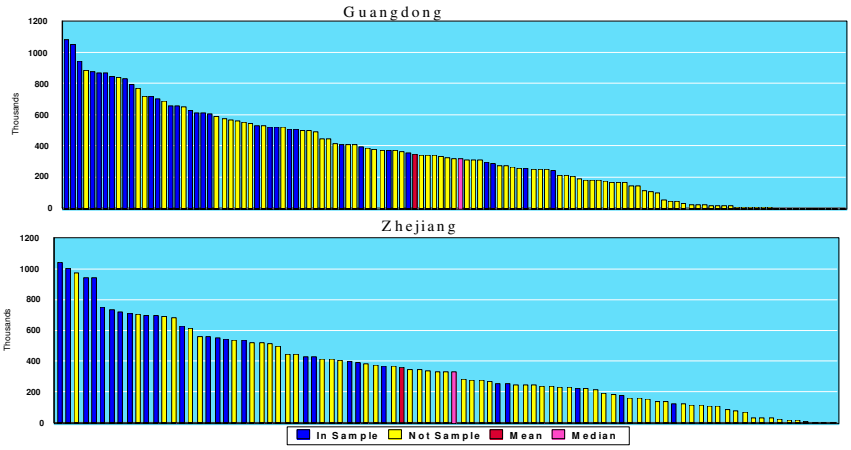


Table 3: Hog Inventory: Number of Head on Hand, December 31, 1996, by County, Guangdong and Zhejiang Provinces

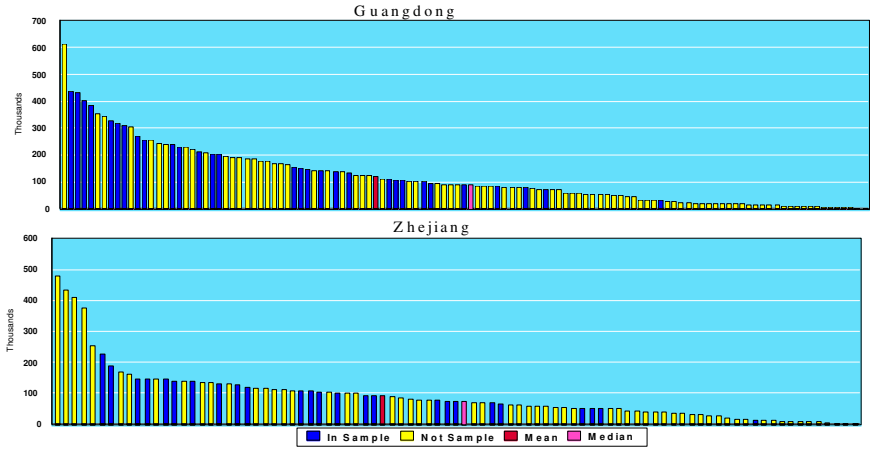


Table 4: Cattle Inventory: Number of Head on Hand, December 31, 1996, by County, Guangdong and Zhejiang Provinces

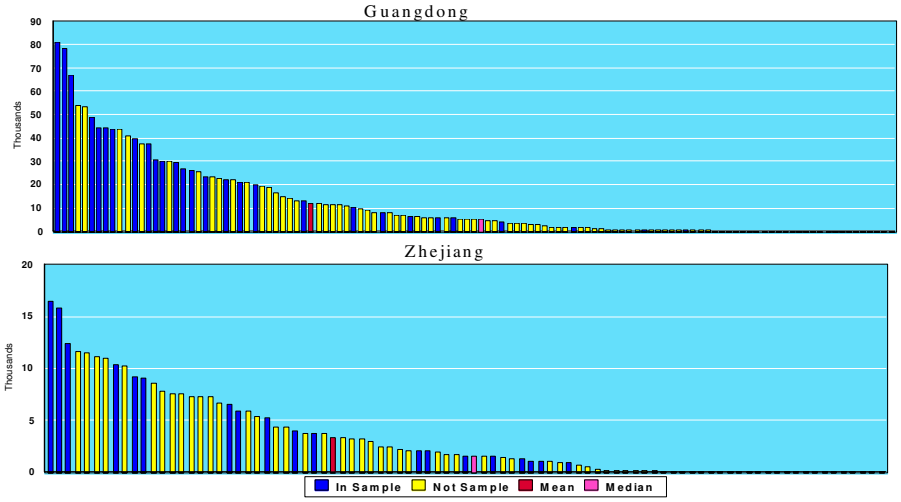


Table 5: Agricultural Census Data – Proportion of the Provincial Total accounted for by Counties in Household Survey Program- Guangdong and Shejiang Provinces

Commodity Code	Name	Counties in Household Survey	Counties Not in Household Survey	TOTAL for All Counties
Guangdong Province				
	Counties	26%	74%	100%
101	Household	44%	56%	100%
301	Cultivated Area	45%	55%	100%
402	Area for Grain	45%	55%	100%
410	Peanuts	51%	49%	100%
416	Vegetables	44%	56%	100%
501	Cattle	53%	47%	100%
512	Hogs	42%	58%	100%
514	Poultry	37%	63%	100%
Zhejiang Province				
	Counties	31%	69%	100%
101	Household	44%	56%	100%
301	Cultivated Area	45%	55%	100%
402	Area for Grain	45%	55%	100%
403	Rice	46%	54%	100%
404	Wheat	48%	52%	100%
409	Rapeseed	33%	67%	100%
416	Vegetables	42%	58%	100%
501	Cattle	38%	62%	100%
511	Sheep	58%	42%	100%
512	Hogs	33%	67%	100%

Table 6: Agricultural Census Data- Average Amounts per County, by Counties in Household Survey Program, Guangdong and Zhejiang Provinces

Commodity Code	Name	Counties in Household Survey	Counties Not in Household Survey	TOTAL for All Counties
Guangdong Province				
101	Household	715,826	310,588	414,495
301	Cultivated Area	402,020	170,413	229,799
402	Area for Grain	620,755	258,244	351,196
410	Peanuts (mu)	47,479	15,855	23,964
416	Vegetables (mu)	89,396	39,156	52,038
501	Cattle (head)	25,828	7,778	12,406
512	Hogs (head)	199,939	95,120	121,997
514	Poultry (head)	2,762,604	1,601,697	1,899,366
Zheijiang Province				
101	Household	601,689	302,599	388,544
301	Cultivated Area	385,427	193,166	248,413
402	Area for Grain	566,926	278,250	361,203
403	Rice (mu)	483,685	226,927	300,708
404	Wheat (mu)	37,063	16,029	22,073
409	Rapeseed (mu)	38,988	31,318	33,522
416	Vegetables (mu)	46,157	25,265	31,269
501	Cattle (head)	4,445	2,910	3,351
511	Sheep (head)	21,300	6,286	10,601
512	Hogs (head)	106,366	88,948	93,953

1 mu = 1/15 hectare.

3. Importance of the Census of Agriculture in China

3.1 Benchmark Data

In 1997, China successfully conducted its first national agricultural census accounting for approximately 214 million rural households. The census is critical in China because it provides the primary source of consistent, detailed data about agriculture at the local level. It gives benchmark data on agricultural and economic activity, the farm population and its infrastructure upon which to build a sound, ongoing statistical program.

3.2 Strengthen National Agricultural Statistics Information System

While there has always been a need for reliable data on which to base decisions, this need is much stronger in today's world economy. Government policy makers can only manage food and agriculture policy effectively if they have an efficient information and forecast system. With the rapidly changing global economy, an up-to-date information system is required to provide the agricultural community with information necessary to maintain competitiveness. One vital public service is information on which policy makers can make informed decisions.

Currently in China, many programs rely on a small number of surveys. The quality of all surveys depend on the frames from which their sampling is based. The existing sampling frame is inflexible and limits efficiencies, and has been targeted for improvement (Zhu, 1998). The agricultural census in China can be used to improve the quality of the list sampling frame so that subsequent samples can be representative of the population of rural households in China, and provide greater flexibility for the changing needs of agriculture.

4. Multivariate Probability Proportional To Size Sampling (MPPS)

This design is referred to as Multivariate Probability Proportional To Size (MPPS) because the relative size of each unit is determined using more than one item of interest. The design takes advantage of the efficiencies of a Probability Proportional To Size (PPS) design, while adding the dimension of utilizing multiple variables in the sample allocation. The MPPS sampling design is an ideal design for a multipurpose survey where the sample units of the population are specialized and/or only have a subset of the items of interest. The design makes it easy to target a sample size for every item. This ability to target samples makes the design significantly easier to construct and more flexible to accommodate changing needs. Under the common stratified design, creating strata for one item is simple.

Even for two items there are methods to define the best strata, but for more than two items the task becomes difficult and subjective. Additionally, the MPPS design is more efficient than a stratified design. With the MPPS design, units have a unique and optimal selection probability.

The MPPS design does use and rely heavily upon auxiliary or control data. Any potential gains in efficiency are dependent upon the quality of the control data. The design, since it has no strata, does not have 'built in' nonresponse weighting groups. These will need to be created when adjusting for nonresponse is necessary. Also, because this is a new design, methods for allocating and determining samples are still being developed.

In analysis of Chinese data, we concentrated on sampling villages, although the ideas can be applied to any primary sampling units that have several candidate size measures. Each village will have a unique probability of selection, B_i, determined by the available auxiliary data and a desired item-level sample size. Let:

$$\pi_i = \min\left(1,\ \max\left(n_1 \frac{x_{1,i}^{3/4}}{\sum_{i=1}^{N} x_{1,i}^{3/4}}, \ldots, n_K \frac{x_{k,i}^{3/4}}{\sum_{i=1}^{N} x_{k,i}^{3/4}}\right)\right) \qquad \textbf{(1)}$$

where $x_{k,i}$ is the size measure of the k^{th} item of interest for the i^{th} village; n_k is the desired sample size for the k^{th} item (that is, we want n_k villages in the sample to possess the k^{th} item); K is the total number of items of interest in the survey and N is the total number of villages in the sampling population (in the province, for example). The auxiliary data is raised to the 3/4 power because the data are several years old and we do not want very large villages to have extremely large values of B_i. This will reduce the number of villages where B_i is truncated to 1. The sample can be selected systematically from a randomly sorted list using an interval of 1. Note that we are placing a ceiling of 1 on the probabilities of selection so that in a systematic sample

no village is selected more than once. In a two stage design(selecting villages at the first stage, and selecting households at the second stage), there may be no need to adopt this approach if we allow a village to be selected more than once and subsample households independently each time such a multiple selection occurs.

The weights for a MPPS sampling design ($w_i=1/B_i$) are defined from multiple measures of size, so that the sum of the weights will not estimate anything meaningful. Therefore, it is recommended to adjust the weights using ratio estimation. Let $\lceil_{k,I}$ be the estimate for item k in village i. The ratio $N / 3^n w_i$ is an adjustment to the estimate $3^n w_i \lceil_{k,i}$ that accounts for the number of villages in the population. This is easier to see if we define an adjusted weight, $w_i^a = w_i \cdot N / 3^n w_i$. Then the sum of the w_i^a weights for the sampled villages is N. Another adjustment uses the population total of the measures of size. In equations (2) and (3) below, the $x_{k,i}$ are the auxiliary data for item k used in equation (1) to determine the selection probability for the i^{th} unit. The adjustment for the k^{th} item is $3^N x_{k,j} / 3^n w_i x_{k,i}$, which accounts for the population total of the variable X_k. Note that if we set $x_{k,i}$ to 1 for all k and i in the population then we have the adjustment based on the number of villages, N. Estimates for the population total and variance of the estimates can be computed as follows:

$$Y_K = \sum_{j=1}^{N} x_{kl} \frac{\sum_{i=1}^{n} w_i Y_{k,i}}{\sum_{i=1}^{n} w_i x_{k,i}} \qquad (2)$$

$$\hat{V}(\hat{Y}_k) = \frac{\left(\sum_{j=1}^{N} x_{kj}\right)^2}{\left(\sum_{i=1}^{n} w_i x_{ki}\right)^2} \sum_{i=1}^{n} w_i^2 \hat{e}_{ki}^2 \quad \text{where } \hat{e}_{ki} = \hat{Y}_{ki} - x_{ki} \frac{\sum_{i=1}^{n} w_I \hat{Y}_{kI}}{\sum_{I=1}^{n} w_I x_{kI}} \quad \text{and } w_i = \frac{1}{\pi_i} \qquad (3)$$

Note that if we calculate an adjusted weight, w_k^a, using the ratio of the summed auxiliary data to the summed sample weights, this will create a set of K weights for each unit in the sample. Alternatively, we can store only the original weights ($w_i=1/B_i$) and use K ratio adjustments during estimation. Kott and Bailey (2000) describe recent experiences using MPPS of the National Agricultural Statistics Service, United States Department of Agriculture. Amrhein and Bailey (1998) describe research concerning MPPS sample designs for the Romanian National Statistics Agency.

5. Improving the NBS Statistical Information System

5.1 Using Census Data to Analyze Sampling Strategies

Village-level census data were used to test various sampling plans via simulation. Three sampling strategies were analyzed. Alternative 1 selects samples from villages within all counties in the province. Alternative 2 selects villages within a new sample of counties (Counties were selected using MPPS procedures). Alternative 3 selects villages within the

counties currently utilized in the Rural Household survey. The results presented demonstrate the process of simulation and provide important measures of efficiencies for determining an appropriate sampling strategy.

Sampling was replicated 1000 times for each of the three sample design alternatives and estimates of totals were calculated after each replicate. The estimates were stored after each replicate and the root mean squared errors (RMSE) were calculated for each item where RMSE for item k is calculated by

$$RMSE_k = \sqrt{\frac{\sum_{j=1}^{1000} (\hat{Y}_{k,j} - Y_k)^2}{1000}}$$

where $\hat{Y}_{k,j}$ is the estimate for the j^{th} replicate and Y_k is the known population total for the simulated survey response for item k.

Calculating the ratio of RMSEs provides an indication of which sampling strategy performs better. *Table 7* shows that Alternative 1 (sampling villages), provides superior efficiency. The ratio of RMSE for Alternative 3 (old counties, villages), is over five times larger than Alternative 1. This example demonstrates the potential gains in efficiencies from sampling villages within all counties compared to the current approach of sampling villages within a sample of counties. Alternative 2 (a new sample of counties selected using a MPPS design) would result in significant gains in efficiencies over Alternative 3. It is important to remember that the original sample of counties was selected in 1984 using grain production and per capita income. Since 1984, there have been major changes in the agricultural economy of China. This example demonstrates the process of analysis to provide results for deciding on a sample design.

Table 7: Comparisons of Sample Design Strategies - Zhejiang Province - Ratio of RMSEs (Root Mean Square Errors)

Stages of Sampling	Total Area Sown (401)	Total Area for Grain (402)	Total Area for Rice (403)
(1) Villages	1	1	1
(2) New Counties, Villages	1.8	3.6	3.1
(3) Old Counties, Villages	5.4	5.6	5.4

A sample size of 250 villages was used in each case. Zhejiang Province data were used in the simulations. Zhejiang Province had 25 agricultural counties participate in the Household Survey, with 10 villages selected per county.

5.2 Replication in Sample Selection

Even in the best circumstances, drawing a good, representative sample for a country the size of China poses a huge challenge. In addition to meeting statistical needs for the country level and for the provinces, municipalities and autonomous regions, there are over 550 prefectures and cities at prefectural level, and over 2,500 counties and cities at county level. Such a large number of different administrative areas complicates sample selection. The use of a replicated sample selection can have an important impact on the quality of the results from a survey. Replication

allows for the design of efficient samples for a range of uses. The replicated sample allows the flexibility to change the sample size by adding or deleting *replicates* (that is, smaller independent samples). Replicated sampling essentially selects multiple samples from the survey population according to the same design.

China currently conducts surveys at the Provincial, Prefecture, and County level. The samples are funded by different administrative areas and are independent, and are not additive. Replicated sample selection provides the means to conduct a large-scale, agricultural survey using a group of small, independent samples, or replicates, that will, when added, provide more stable and reliable results at all levels.

6. Proposed Strategy for New Sample Survey Design in China

6.1 Scope of Pilot Survey

NBS is preparing to conduct a pilot survey using a MPPS sample design. The pilot survey will be conducted in Guangdong Province collecting annual crop production data and livestock statistics. The sample addresses the three issues outlined earlier in section 2.3: (1) to effectively utilize data from the Census of Agriculture, (2) to select samples using multiple variables which is necessary to support the proposed expanded survey program, and (3) to integrate the statistical needs for different levels of government. In addition, the pilot survey data will be analyzed and weighted using data from the Agricultural Census.

6.2 Sample Design Guidelines

Data for major crop and livestock commodities from the Chinese Census of Agriculture will be utilized in selecting the sample for the pilot survey. A sample a villages will be selected from **ALL** counties in the province. Data from the traditional sample counties will be analyzed and compared to the traditional non-sample counties, and with data from all counties. The sample will be selected in pieces or replicates, identifying portions of the sample that will be funded by the three different levels of government (provincial, prefecture, and county). Samples funded by the three different levels of government **WILL** be additive in the summarization process.

6.3 Sample Allocation

Different combinations of variables are being analyzed for use in the sample selection procedure. One possible combination is illustrated below.

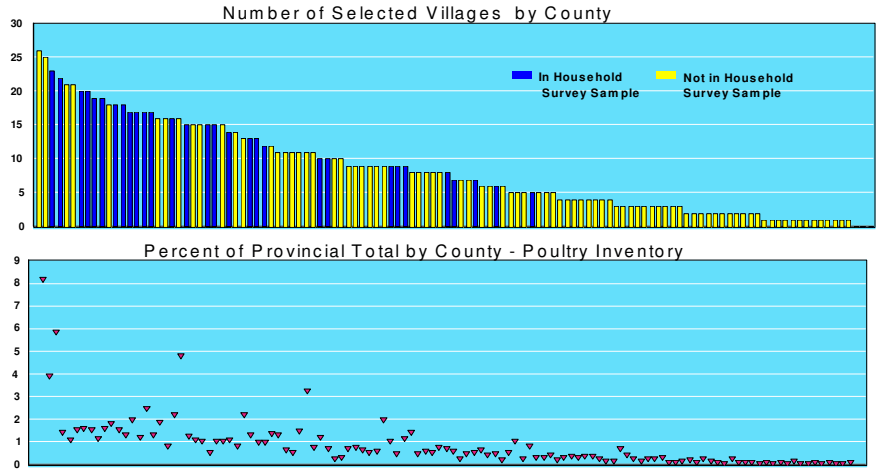
Variables	Targeted Sample Size (Provincial-level) Number of Villages
Number of Households	400
Household Members	400
Cultivated Land Area	600
Vegetables Area	600
Hog Inventory	600
Poultry	400

If six separate samples were selected, the total sample size would be 3000 villages.

A MPPS sample of 1,000 villages would meet the sample sizes targeted in section 6.3. If a survey only seeking provincial-level estimates were conducted, this is the recommended sample size for the province. If prefecture and county level estimates were desired, then a sample of 5000 villages would be utilized, and the sample by county would be approximately five times the numbers outlined in Table 8.

Table 8 shows the number of sample villages by county, and designates whether or not the county is currently a county in the Household Survey program. The bottom part of Table 8 shows the percent of total household members in Guangdong Province for each county. With the proposed MPPS sample of 1000 villages, the largest sample size for a county is 26 villages. Table 8 indicates that the county with 26 samples has over 8 percent of the poultry for the province. The county with 26 samples is not a sample county in the existing Household Survey program.

Table 8: Poultry Inventory, Guangdong Province- Sample of 1000 MPPS



7. Conclusion

The MPPS sampling approach being considered will utilize Agricultural Census data for variables designated as important by NBS and provincial statistical bureaus. Multiple variables can be utilized in the sample allocation, which facilitates the inclusion of new commodities into the NBS survey program. The sample can be selected in pieces or replicates, allowing all samples to be aggregated and offering an alternative to the current situation where separate samples exist for different governmental units that are not additive. Conducting a pilot survey with the proposed MPPS sample design will provide valuable data for estimation research for analyzing the effect of generating statistics from a sample of counties versus a sample from all counties.

References

Amrhein, John F., Bailey, J. (1998). *Sampling Villages for Multipurpose Surveys*. Paper presented at the IASS/IAOS Conference on Statistics for Economics and Social Development, Aguascalientes, Mexico.

Gong Jian Yao and Chen Jixen (1989). *China's Crop Yield Survey and Its Methodology*. Invited paper presented at the 47th session of the International Statistical Institute, Paris, France.

Kott, Phillip, Bailey, J. (2000). *The Theory and Practice of Maximal Brewer Selection with Poisson PRN Sampling*. Paper presented at the International Conference on Establishment Surveys - II, Buffalo, New York.

Li Ting Jun (1995). *Practice and Research on Crop Cutting Surveys*. Paper presented at the 50th session of the International Statistical Institute, Beijing, PRC.

National Bureau of Statistics (1995). *A Guide to the Statistical System in China*. National Bureau of Statistics, Beijing, PRC.

National Bureau of Statistics (2000). *Introduction on Rural Household Survey in China.* National Bureau of Statistics, Beijing, PRC.

Vogel, F.A. (1999). *Review of Chinese Crop Production Forecasting and Estimation Methodology*. United States Department of Agriculture Miscellaneous Publication No. 1556.

Vogel, F.A. (2000). *Reconciling Official Statistics and Census Results*. Paper presented at International Seminar on China Agricultural Census Results, Beijing, PRC.

Zhu Xiangdong (1998). *Agricultural Census and Improvement of Rural Statistics in China*. Paper presented at Agricultural Statistics 2000, an International Conference on Agricultural Statistics, Washington, D.C. USA.

Acknowledgment

We take this opportunity to thank Mr. Phil Kott for his outstanding work in development of MPPS applications for agricultural surveys, and to thank Ms. Barbara Rater for editorial contributions on this paper.