CAMEROON AGRICULTURAL PLANNING AND POLICY PROJECT

Produced with technical assistance from Washington State University
and the Consortium for International Development under
funding from USAID/Cameroon Contract No. 631-0059-C-00-9015-00

REPUBLIQUE DU CAMEROUN
Paix - Travail - Patrie

MINISTERE DE L'AGRICULTURE

REPUBLIC OF CAMEROON
Peace - Work - Fatherland

MINISTRY OF AGRICULTURE

ANALYSIS AND EVALUATION OF CAMEROON AGRICULTURAL SURVEY DESIGN

FINAL REPORT

by

Charles R. Perry (Statistician, NASS/USDA)

March 3 to April 7, 1990

DEPARTMENT OF AGRO-ECONOMIC SURVEYS
AND AGRICULTURAL PLANNING

AGRICULTURAL AND FORESTRY
PLANNING SERVICE

Cameroon Agricultural Policy and Planning Project

Produced with technical assistance from Washington State University and the Consortium for International Development under funding from USAID/Cameroun Contract No. 631-0059-C-00-9015-00

FINAL REPORT MAY 15, 1990

SHORT TERM TECHNICAL ASSISTANCE TRIP REPORT: ANALYSIS AND EVALUATION OF CAMEROON AGRICULTURAL SURVEY DESIGN

BY CHARLES R. PERRY, PH.D.
NATIONAL AGRICULTURAL STATISTICS SERVICE
U.S. DEPARTMENT OF AGRICULTURE
WASHINGTON, D.C. 20250

MARCH 3 - APRIL 7, 1990

CAMEROON AGRICULTURAL PLANNING AND POLICY PROJECT

INTRODUCTION AND PURPOSE OF TRIP

PURPOSE OF TRIP:
The purpose of the trip was specified by the six components of the scope of work:

1. Provide appropriate formulas for estimating the variances under the current sampling procedure and provide guidance as to their application.

2. Supervise the re-calculation of estimates for totals, variances and coefficients of variation for 1985-87 agricultural survey data.

3. Assist in the preparation of procedure and a program for the detection and handling of outliers.

4. Determine the relative importance of adding second stage sampling units as compared to first stage sampling units for the reduction of sampling variance at the provincial and national level. Provide estimates of sample size necessary for achieving 10 and 15 percent coefficients of variation at the national and provincial levels respectively.

5. Provide guidance in determining if there is significant year to year differences in previous survey results -- guidance in appropriate statistical testing procedures.

6. Provide guidance in establishing cut-off point for extreme operators.

RELATIONSHIP OF SCOPE OF WORK TO PROJECT OBJECTIVES:
Items 1,2, 3 and 5 above provide statistical analysis and summary procedures for the 1985-89 agricultural survey necessary to make the results useful for publication and for a national agricultural commodity data bank.

Items 1, 4, and 6 provide the information necessary for the optimal design of a new agricultural survey system.

ACTIVITIES AND ACCOMPLISHMENTS

Listed below are the major activities and accomplishments numbered sequentially as they relate to the scope of work.

1. Provided appropriate formulas along with explanations and guidance for their use in estimating the variances, coefficients of variation, and totals with the existing systematic probability proportional to size sampling design (sys pps). These formulas along with standard references are given in Appendix A.1. Related instructions for drawing the first and second stage samples are given in Appendix A.2.

2. Obtained documentation of existing data summary system and ascertained the nature of its output. Determined that the existing system did not have the capability of producing the necessary statistics for computation of the variance components necessary for the determination of relative importance of adding first and second stage sampling units on variance of the survey items. Moreover, it was not clear from program documentation and output exactly how the estimates for the totals and their variance were computed.

The above insufficiencies made it necessary to develop rather involved summary and analysis programs to accomplish the objectives given in the statement of work. An appropriate summary and analysis system was written in SAS for the existing survey design to obtain the statistics needed for the various components of the scope of work. The SAS programs, which are documented by comments within the code, provide the major components for a complete summary and analysis system for the current and future agricultural surveys.

Copies of these SAS programs were left with the CAPP project. These programs were successful run by Mr. Jim Otto on the 1985 and 1986 data sets. For documentation and reference, printed copies of the updated versions of these programs and examples of their out put have been bound into one volume. This documentation along with a diskette containing the programs is being shipped to the CAPP project office under a separate cover.

Performed the re-calculation of estimates for totals, variances and coefficients of variation for 1985-87 agricultural survey data using the SAS programs. The re-calculated estimates were done in two ways: by select strata and by summary strata in order to insure the appropriateness of the collapsed summary strata estimates. In most cases the re-calculated estimates were similar to the existing summary system output. In others the re-calculated estimates were quite different from the existing system output.

It was not possible to determine the cause of these differences. However, it seems clear that the existing system suffers from a lack of numerical precision, and that it uses formulas that are only partially appropriate for the existing survey design. In any case, the new SAS summary and analysis procedures are numerically very accurate and based on the correct formulas. The new system provides a wide range of variance component estimates and other statistics that should be helpful in optimizing the efficiency of the existing and future survey designs.

Copies of recalculated estimates (in some cases by select and summary strata) were left with the CAPP project office. A complete set of these estimates, along with other product described below, have been bound and are being shipped to the CAPP project office under a separate cover.

3. Developed procedures for statistically detecting and handling outliers (highly influential farm household item responses -- unusually large expanded values) in the survey data. The approach was as follows: develop fixed cutoff values by province for each survey item (both area and production for 20 agricultural items) that statistically classify a predetermined percent, p, of the expanded data as outliers, summarize the non-outlier and outlier data separately, estimate the current total by adding to the expanded non-outlier portion of the current data the mean of the expanded outliers taken over several years. The cutoff and length of the historical series are determined empirically so that they perform optimally on the available historical data (here 1985-1989 survey agricultural data).

The literature and my past experience indicate the cutoffs set to censor about 5-15 percent of the upper tail of distribution of the expanded values and a 5-year outlier series will provide acceptable performance with current design. The use of this type outlier detection and handling procedure is documented in a USDA/NASS technical report "Estimation of Totals for Skewed Populations in Repeated Agricultural Surveys" by Thomas, Perry, Viroonsri.

Wrote into the new SAS summary and analysis procedures a numerical subroutine that determines the outlier data, divides each observation into its outlier and non-outlier portions, and summarizes the resulting outlier and non-outlier data sets separately. These programs were tested on the 1987 data by arbitrarily setting provincial item cutoff at 2 percent of nation's item total. (Even though just test, the results indicate that the general procedure will significantly dampen the effect of outliers with only minor introduction of bias). A program was written to sort each item of the historical data sets in descending order in preparation for the analysis necessary to determine the optimal cutoff values. The completion of this analysis was limited by computer disk space and time constraints (preliminary versions of the necessary programs were written but only partially debugged due to the extra time required to get around limited computer disk space).

Prelininary versions of robust estimates based on 1985 -- 1987 data were computed after returning to D.C. using 5, 10, 15, 20 and 25 percent provincial level cutoffs. Tables containing the robust estimates along side the original estimates have been bound and are being shipped to the CAPP project office under a separate cover.

4. Wrote a SAS program to compute estimates of the statistics necessary to determine the relative importance of adding first and second stage sampling units: estimates of the total, the total variance, the first stage variance, the second stage variance, etc. Estimates were made of the optimal number of second stage sampling units per first stage sampling unit for each commodity item by province. These estimates assume that it costs five times as much to add a large unit as a small unit. These relative cost figures are roughly those associated with the 1989 survey. Both tabular and graphical presentations of the results were prepared.

The tables show the change in coefficient of variation that is expected to result from adding various combinations of first and second stage sampling units. The graphical output contains four different graphs for each commodity for both area and production: a graph showing the relative increase (decrease) in survey cost required to achieve a specified provincial level coefficient of variation for various subsampling rates; a graph showing the relative increase (decrease) in the firsts stage sample size needed to achieve a specified provincial coefficient of variation for various subsampling rates; a graph showing the first stage sample size required to achieve a specified coefficient of variation for various subsampling rates; and a graph showing the first stage sample size required to achieve a specified coefficient of variation for the current subsampling rates.

These graphs are used as follows: 1. the optimal number of second stage units per first stage unit is determined from the relative cost graphs -- a compromise reached considering the commodities of interest and the cost graphs; then 2. the minimal number of first stage units needed to achieve a specified coefficient of variation is determined from the first stage sample size graph by finding the value of the first stage sample listed on the vertical axis that corresponds to the selected optimal second stage sample size listed on the horizontal axis.

The first and second stage variance components estimation procedures were made a part of the standard output from the new SAS summary and analysis procedures since they are very useful to professional statisticians in diagnosing anomalies in survey data. The procedures approximate the optimal number of second stage units per first stage unit and graph the results in terms of the relative cost to achieve a coefficient of variation of 10 and 15 percent at the provincial and national levels respectively.

Samples of the graphical and tabular output from these procedures were left with the CAPP project office. Three complete set of these graphs and tables (one set based on the original 1985-87, one set based on the 1985-87 data and 5 percent provincial level cutoffs, and one set based on the 1985-87 data and adjustment for a properly drawn second stage sample) along with documentation describing how to use these products in

the redesign of the Cameroon Agricultural Survey have been bound and are being shipped to the CAPP project office under a separate cover.

5. Discussed at length the appropriateness (inappropriateness, actually) of the current sampling design and data collection scheme for estimating year to year changes in crop area and production. The current data collection scheme uses a new second stage sample each year, therefore it has very limited power to detect year to year level changes. Its primary power is in determining the average levels over several years, not year to year, changes in level. To overcome this deficiency any new survey design should incorporate a replicated rotational data collection scheme, i.e. data should be collected on a specified portion of the sample for two or more consecutive years.

6. Discussed establishing the cut-off point for extreme operators. In general the extreme operator list should be as large as can be efficiently managed with existing staff and data base crosschecking ability. The development of an extreme operator list is absolutely imperative for improving the efficiency of the agricultural survey. However, adequate quality control checks and procedures must be developed and institutionalized to insure that the cross checking between the extreme operator list and the households listed in sampled segments is accurate. Otherwise error in the cross checks will likely introduce nonsampling errors which offset the sampling gain in sampling efficiency associated with the use of the extreme operator list.

MAJOR FINDINGS AND CONCLUSIONS

1. The application of the standard variance estimation formulas for probability proportional size with replacement sampling (pps wr) to the current probability proportional size systematic sample design (pps sys) using minimal size summary strata provides relatively accurate variance estimates. The estimates are guaranteed by theory to have a positive bias. A comparison of the variance estimates computed with the new SAS summary system using the minimal size summary strata are on the average only six percent smaller then those derived with the system using the much larger select strata. Since the decrease normally is an exponential function of the strata size one can safely assume that the bias in the variance estimates based on summary strata is positive but much less than six percent. Hence, these computationally efficient formulas can be used with safety in estimating the variances for the various item totals in the agricultural survey.

2. The trial computer runs with the crude 2 percent outlier cutoff and further examination of the expansion factors associated with the outliers (highly influential farm data) show conclusively that the current sampling procedure which takes a fixed number of second stage units (farm households) from each selected first stage unit (segment) is highly inefficient and that this fixed secondary sample size is one of the root cause of outliers in the survey data. The trial computer runs also show that the second root cause of most outliers is the inclusion of extremely large operations in the area frame of census enumeration district. The correction of the outlier problem in the agricultural survey is thus three fold: a. Allocate the total number of second stage sampling units to the first stage units so that the sample becomes self weighting at least across the rural strata with in each province, b. Develop an complementary extreme operator list for the area frame that covers as many as possible of the large operations, and c. Determine optimal cutoffs for the statistical detection of outliers and apply the estimation procedures outlined earlier.

RECOMMENDATIONS

1. That the first stage sample of segments in the new agricultural survey be divided into either four or eight replicates and that each first stage unit be drawn probability proportional to size with one unit per substratum -- a design analogous to the standard randomized block experimental design. This type design will permit rotating in two new replicates each year with only minimal impact on the year to year level estimates. Further more, it will permit unbiased estimation of all variance components and other advantages associated with replicated designs.

2. That the second stage sample of farm households be drawn in two replicates using simple random samples and that the replicates be rotated on a two year basis. This will permit a 37.5 percent year to year overlap of the farm households in the survey. This rotational second stage sampling scheme will permit much more accurate estimates to be made of the year to year level changes to be made than currently can be made with the existing sample design.

3. That the second stage sample size (number of farm households) for each first stage sample unit (segment) be made self-weighting by making the second stage sample sizes proportional to the number of listed households in the segment and inversely proportional to the segment's probability of selection. This procedure for allocating the second stage sample is much more efficient than the current procedure of taking approximately the same number of units from each selected segment. This method of allocating the second stage sample will also eliminate many of difficulties associated with outliers in the current agricultural survey since many of the outliers arise from unusually large expansion factors and not from unusually large farming operation.

4. That the basic SAS programs, the new SAS summary system, be refined and integrated to give an operational summary and analysis package for the agricultural surveys. This will require purchasing a SAS site license and some short term technical assistance. Adopting SAS as the standard programming language used for analysis of agricultural survey data will guarantee computational accuracy. SAS contains a wide range of statistical routines (multivariate regression, non-

linear regression and optimization, time series analysis, etc.) that will facilitate economic analysis of the basic survey statistics. SAS routines are numerically very accurate (the internal computations are normally carried out to 15 or more decimal place accuracy). Many internal checks are routinely made that flag missing data items etc. and that improve survey data handling. Data editing routines are easier to develop when the summary system is written in SAS. SAS is sufficient for all data handling needs associated with the agricultural survey. As an example, during this TDY a basic routines for summarizing, computing variance, displaying result, optimizing the sampling etc. was developed.

4. That in service training be provided for a certain number of Cameroonians that will provide an understanding of the new agricultural sample design and the management of the new summary and analysis procedures, including the outlier detection and handling procedures. The short term objectives of the training will be to provide an in depth understanding of the underlying theoretical concepts, the programing of these concepts, the operational use of the developed programs, and the ability to modify and adjust the programs as needed. The long term objectives of the training will be to institutionalize these procedures into the agricultural survey, nurture the professional expertise of the participant, and nurture the ability of the participant to apply the theoretical concepts developed to the current and future survey and their redesign. If desirable, consideration could be given to promoting contact with Cameroonian universities as a means of developing local university awareness of the statistical training needs of the Cameroon Agricultural Survey.

Charles R. Perry, Jr.

Charles R. Perry, Jr., Ph.D.

May 15, 1990

Date

APPENDIX A.1

SUMMARY AND ESTIMATION FORMULAS FOR THE CAMEROON AGRICULTURAL SURVEY

CHARLES R. PERRY, JR, PH.D.
MAY 1, 1990

Basic Stratum Level Sampling Procedure:

Assume that at the first stage of sampling n segments (PSU's) are drawn independently from a stratum so that the expected number of times the ith segment is selected is:

$$\pi_i = np_i$$

where p_i is the relative size of the ith segment. That is:

$$p_i = \frac{X_i}{\sum_{i=1}^{n} X_i}$$

where X_i is the size of the ith segment.

Assume that at the second stage of sampling m_i farm households (SSU's) are drawn with equal probability and without replacement from the M_i farm households listed in the ith segment.

Basic Stratum Level Summary and Estimation Formulas:

Formulas are given below for estimating the stratum total and its variance under the assumptions outlined above. Formulas are also given for estimating the components of variance associated with the first and second stage sampling. The statistical properties of these estimates are summarized.

Unbiased estimates for the ith segment total and its variance are given by:

$$\hat{Y}_i = \frac{M_i}{m_i} \sum_{j=1}^{m_i} y_{ij}$$

$$\hat{V}ar_2(\hat{Y}_i) = \frac{M_i(M_i-m_i)}{m_i(m_i-1)}\left(\sum_{j=1}^{m_i} y_{ij}^2 - \frac{\left(\sum_{j=1}^{m_1} y_{ij}\right)^2}{m_i}\right)$$

where y_{ij} is the commodity value for the jth farm household from the ith segment, M_i is the number of listed farm households and m_i is the number of sampled farm households .

Unbiased estimates for the stratum total and its variance are given by:

$$\hat{Y} = \sum_{i=1}^{n} \frac{\hat{Y}_i}{\pi_i} = \frac{1}{n}\sum_{i=1}^{n} \frac{\hat{Y}_i}{p_i}$$

$$\hat{V}ar(\hat{Y}) = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{\hat{Y}_i}{\pi_i} - \frac{1}{n}\hat{Y}\right)^2$$

$$= \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{\hat{Y}_i}{p_i} - \frac{1}{n}\sum_{i=1}^{n}\frac{\hat{Y}_i}{p_i}\right)^2$$

Unbiased estimates for the first and second stage variance components are given by:

$$\hat{V}ar_2(\hat{Y}) = \frac{1}{n^2}\sum_{i=1}^{n}\frac{\hat{V}ar_2(\hat{Y}_i)}{p_i^2}$$

$$\hat{V}ar_1(\hat{Y}) = \hat{V}ar(\hat{Y}) - \hat{V}ar_2(\hat{Y})$$

These variance estimation formulas are applied in practice to sample designs that do not strictly meet the segment selection criteria outlined above. The standard applied procedure is to first subdividing the original select strata into small contiguous summary strata and then apply the formulas. This was the procedure followed with the 1985-1987 data from Cameroon Agricultural Survey

which used a systematic probability proportional to size sample design. That is, to the extent possible adjacent segments were paired to form summary strata (implicit strata). Then the variance formulas were applied at the summary strata level and the results accumulated to the select strata, District, Provincial and National levels. Theoretical considerations show that the variance estimates derived by this procedure always have a positive bias. The biases in the current application, as discussed in my March - April trip report, are probability much less than 10 percent and hence insignificant relative to obvious non-sampling errors and biases.

Determining the Size of the Second Stage Sample for Segment i:

To help reduce the number and effect of outliers, which result in the 1985 -1987 data primarily from the wide variability of expansion factors within strata, the sample should be made self weighting -- the first and second stage samples should be drawn in such a way that all the resulting expansion factors within a stratum are essentially the same. The general procedural rule to follow is: make the samples self weighting over the largest possible subdivisions that do not result in major loss of sampling efficiency. Following this procedure will minimize the number of outliers and facilitate implementing a formal statistical procedures to detect and handle the outliers and influential data values that remain. For example, making the sampling in the urban strata self weighting over all urban strata will permit treating outlier from these strata, which are generally always a problem, in a uniform manner. Making the sampling in the rural strata of provinces that having similar variance characteristics self weighting will facilitate setting realistic outlier truncation limits and smoothing out of outlier effects over larger political subdivisions.

Moreover, in practice, drawing the subsample of farm household so that the sampling becomes self weighting is also the optimal way to allocate a subsample to the various segments. Here, "the optimal way to draw the subsample" means that there is no other way to allocate the total subsample of farm households to the segments that will result in a smaller variance. Or equivalently, "the optimal way to draw the subsample" means that, if the total subsample size is to be a fixed multiple of the number of segments, then the smallest number of segments will be required to achieve a specified variance when the subsampling of farm households is allocated to the segments so that the sample becomes self weighting.

The sampling becomes self weighting and the subsampling allocation to the segments optimal when the number of farm households drawn from the ith segment is proportional to the number of farm

households listed in the segment and inversely proportional to segment's probability of selection. The formula for determining the number of farm household to draw from a segment is given below.

The number of farm households to be drawn from the ith segment is:

$$m_i = k \frac{M_i}{p_i}$$

where M_i is the number of listed farm households, p_i is the segment's probability of selection and k is given by:

$$k = \frac{n f_o}{\sum_{i=1}^{n} \frac{M_i}{p_i}}$$

where f_o is the average number of farm households selected from the n segments.
These formula means that if n segments are to be sampled and if average number of farm households to be sampled per segment is f_o, then the total number of farm households to be sampled, n times f_o, is optimally allocated to the segments by the formula:

$$m_i = \frac{\left(\frac{M_i}{p_i}\right)}{\left(\sum_{i=1}^{n} \frac{M_i}{p_i}\right)} (n f_o)$$

The last formula for sample size in the ith segment can be rewritten in terms of the segment sizes X_i as:

$$m_i = \frac{\left(\frac{M_i}{X_i}\right)}{\left(\sum_{i=1}^{n} \frac{M_i}{X_i}\right)} (n f_o)$$

To apply the last set of formula at the Provincial level:

1. Determine an optimal average subsample size, f_o, from the relative cost graphs provided. This will require your expert judgement in striking a compromise among the estimated optimal average subsample sizes for the various items of interest.

2. Determine for the selected average sample size, f_o, the number of segments, n, required to achieve the desired 15 percent CV from the SSU sample size graphs provided. Here again your expert judgement will be required in reaching a compromise over the items of interest between the sample sizes required and what is possible.

3. Determine the total over all subsample size, n times f_o.

4. Determine the optimal allocation of the total subsample, n times f_o, to the n segments by using the formulas above. Estimating k from historical data would permit drawing a self weighting sample in the field. However, it would not permit determining before the segments are listed the exact total subsample size that will result from the estimated k. Since k will generally be an integer, it should be statistically rounded to an integer; it should not be simply rounded or truncated.

APPENDIX A.2

DRAWING THE FIRST AND SECOND STAGE SAMPLE

CHARLES R. PERRY, JR, PH.D.
USDA/NASS
APRIL 14, 1989

THE STEP BELOW OUTLINE THE PROCEDURES TO BE FOLLOWED IN DRAWING THE FIRST AND SECOND STAGE SAMPLES. THE FIRST STAGE SAMPLE (PSUs) ARE DRAWN PROBABILITY PROPORTIONAL TO SIZE ONE SAMPLE UNIT PER ONE-POINT STRATUM. THE SECOND STAGE SAMPLE (SSUs) ARE DRAWN AS TWO SIMPLE RANDOM SAMPLED ROTATIONAL GROUPS. THE STEPS OUTLINED ARE TO BE APPLIED BY STRATUM TO ALL PROVINCES.

1. Make a form listing the census enumeration districts (PSUs) in serpentine order with the first two columns containing the identifier and size (number of households) respectively.

2. In the third column create the cumulative sum of the sizes from column two.

3. Divide the cumulative total, T, by the stratum sample size, n, to get the "average" one-point stratum size.

4. Divide the serpentinely ordered enumeration districts into n one-point strata with sizes as near T/n as possible. These one point strata should be within a few percent of the "average" one point stratum size. To do this may require splitting (and/or combining) some enumeration districts.

Note 1: Montie, up to this point the above is very similar to what you describe in your PIADIC Area Frame Probability Samples booklet -- only the words have changed: segments (PSUs) are now enumeration districts, area is number of households (or number of farm households), and paper strata are one-point strata. Even the drawing of the samples, which is described below, is only sightly different to what you described in PIADIC.

5. Choose one enumeration district, PSU, from each one-point stratum with probability proportional to size. To do this draw a random number between 1 and size of the one-point stratum (the total of the PSU sizes accumulated over the one-point stratum). The enumeration district (PSU) corresponding to the random number drawn is the selected PSU. Repeat this process for every one-point strata to select the stratum sample.

15

Note 2: To permit a 25 percent rotation of PSUs each year, the sample size, n, in every stratum must be a multiple of four. Multiples of eight would be desirable from the stand point of variance estimation. But, the collapsed strata (implicit strata) method of variance estimation provides a satisfactory variance estimate. Therefore, let simplicity and sample size considerations drive your choice of four or eight. You may wish to combine the urban strata over several provinces, sample the combined urban stratum in multiples of four, and then prorate the estimated commodity totals for the combined urban strata back to the urban province level.

Note 3: The PSUs for the rotational group replacements should be drawn with replacement. That is, a PSU being rotated out is eligible to be selected the replacement PSU. However, if the same PSU is drawn as a replacement PSU, then the old second stage sample (SSUs) may be considered ineligible for selection in the new second stage sample (when SSUs from the old SSU sample are drawn just ignore them and redraw). The problem with drawing the PSUs pps without replacement is that biases will be introduced into the estimates because of correlations between the PSU sizes and commodity proportions.

6. Assuming you make the stratum sample sizes multiples of four, the serpentinely ordered one-point strata are numbered 1, 2, 3, 4, 1, 2, 3, 4,..., 1, 2, 3, 4 to form the four replicates. This is the simple way. What I would prefer is to randomize each group of four into the four replicates just like you would do in a randomized block experimental design. However, only do the randomization if it does not lead to confusion (errors) -- other wise be a good engineer and remember KISS (keep it simple and straight forward).

7. The estimated commodity totals are obtained by expanding the PSU estimates to the one point stratum level and summing to the province and national levels.

PS. Montie, if this is not all clear and you need some examples attached, or what ever, just let me know. Please give my regard to your wife Jo, Frank and Cecile, and the others at CAPP and thanks again for making my TDY successful and pleasant.

16