FOOD SECURITY RESEARCH PROJECT

RECOMMENDATIONS ON SAMPLE DESIGN FOR POST-HARVEST SURVEYS IN ZAMBIA BASED ON THE 2000 CENSUS

By

David J. Megill

WORKING PAPER No. 11
FOOD SECURITY RESEARCH PROJECT
LUSAKA, ZAMBIA
February 2004
(Downloadable at: **http://www.aec.msu.edu/agecon/fs2/zambia/index.htm** ***)***

RECOMMENDATIONS ON SAMPLE DESIGN FOR POST-HARVEST SURVEYS IN ZAMBIA BASED ON THE 2000 CENSUS

David J. Megill

FSRP Working Paper No. 11

February 2004

ACKNOWLEDGMENTS

The Food Security Research Project is a collaboration between the Agricultural Consultative Forum (ACF), the Ministry of Agriculture, Food and Fisheries (MAFF), and Michigan State University's Department of Agricultural Economics (MSU).

We wish to acknowledge the financial and substantive support of the United States Agency for International Development (USAID) in Lusaka. Research support from the Global Bureau, Office Agriculture and Food Security, and the Africa Bureau, Office of Sustainable Development at USAID/Washington also made it possible for MSU researchers to contribute to this work.

This study has been made possible thanks to the contributions of a number of people and organizations. In particular, the Agriculture and Environment Division, Central Statistical Office (CSO), and the Database Management Unit, Ministry of Agriculture and Cooperatives (MACO), who provided assistance in analyzing and utilizing 2000 Census data. MACO and CSO also actively participated in technical discussions that took place during the sample design process, providing the opportunity for the author, D.J. Megill (sampling consultant for FSRP), to refine the methodologies proposed in FSRP Working Paper No. 2 and present a working version of the new sampling methodology.

Comments and questions should be directed to the In-Country Coordinator, Food Security Research Project, 86 Provident Street, Fairview, Lusaka; tel: 234539; fax: 234559; email: fsrp@coppernet.zm

FOOD SECURITY RESEARCH PROJECT TEAM MEMBERS

The Zambia FSRP field research team is comprised of Jones Govereh, Billy Mwiinga, Jan Nijhoff, Gelson Tembo and Ballard Zulu. MSU-based researchers in the Food Security Research Project are Antony Chapoto, Cynthia Donovan, Thomas Jayne, David Tschirley, and Michael Weber.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ANNEXES

1. BACKGROUND

The Central Statistical Office (CSO) has been conducting the annual Zambia Post-Harvest Survey (PHS) for many years. The current sampling frame for this survey is based on the census supervisory areas (CSAs) and standard enumeration areas (SEAs) defined for the 1990 Zambia Census of Population and Housing. A new listing of households is conducted in the sample SEAs each year for selecting the sample households. Although this is still a representative sample of households, changes in the population distribution during the past decade have made the 1990 sampling frame for the PHS less efficient. Now that the data from the 2000 Zambia Census of Population and Housing are available, it is possible to develop a more effective sampling frame for the PHS. The 2000 census questionnaire included a question on whether the household engaged in the agricultural activities (crop growing, livestock and poultry raising, and fish farming), as well as check items to identify the specific crops grown and animals raised by the household. These data will be very useful for developing an updated sampling frame and more efficient sample design for the PHS surveys in the next decade, beginning with the 2003-04 survey.

In July 2000 the consultant examined the current sample design for the PHS and documented his findings and recommendations in the report on "Review of Sample Design for Post-Harvest Survey (1997/98) and Recommendations for Improving the Sampling Strategy and Estimation Procedures." That report describes the current sample design for the PHS, includes the tabulation of sampling errors for selected survey estimates using the CENVAR software, and examines issues related to the precision of the survey estimates based on the current sample design. It should be used as a companion reference with the current report as part of the PHS methodological documentation. One of the conclusions in the previous report was that the coefficients of variation (relative sampling errors) were fairly high for certain crops that were less frequent or had a limited geographic distribution. The new sample design will attempt to improve the level of precision for these crops by introducing more stratification at the second sampling stage.

The CSO established a working group of statisticians, systems analysts and subject-matter experts to assist in developing the new sample design for the PHS. A list of the staff included in this working group on the PHS sample design is presented in Annex I. The consultant met with them on the first day of his visit, and in a follow-up meeting they established a consensus on recommendations for the major decisions that needed to be made regarding the sampling frame for the PHS. The consultant would like to thank them for their valuable input into the sampling methodology presented here. The collaboration from the staff of the Food Security Research Project (FSRP) was also important in developing the recommendations in this report.

The purpose of this report is to make recommendations on the new sample design for the PHS based on the 2000 Zambia Census of Population and Housing sampling frame. This technical assistance was funded by USAID through the FSRP.

2. SAMPLING FRAME AND UNITS OF ANALYSIS FOR THE NEW PHS

The sampling frame for the new PHS will be based on the information and cartographic materials from the 2000 Zambia Census of Population and Housing. Zambia is divided into nine provinces, which are further divided into 70 districts. Each district is administratively subdivided into constituencies and wards. For the purposes of the 2000 census enumeration, a cartographic operation was conducted to define census supervisory areas (CSAs), which are further divided into standard enumeration areas (SEAs). The SEA is the smallest area with well-defined boundaries identified on census sketch maps; each SEA was covered by an individual enumerator for the census data collection.

In the case of the previous PHS, a stratified three-stage sample design was used. The CSAs were the primary sampling units (PSUs) selected with probability proportional to size (PPS) at the first stage, where the measure of size was based on the total number of households in the CSA. At the second sampling stage one SEA was selected with PPS within each sample CSA. This resulted in a similar dispersion of the sample and probabilities of selection as if the SEAs had been selected directly at the first sampling stage. Within each sample SEA the households were listed and stratified by size for selecting the sample households at the last sampling stage.

A stratified two-stage sample design will be used for the new PHS. The working group recommended defining the PSU as one or more SEAs with a minimum of 30 agricultural households. This sampling approach will be easier to implement and provide more flexibility for the stratification of SEAs by predominant crop. The sample households will be selected at the second stage from the listing stratified by farm size category.

One advantage of defining the CSAs as PSUs in the previous PHS sample design is that it would be possible to rotate the sample SEAs within the larger sample PSU over time, but this was not done for the previous surveys. Given that each sample SEA is uniquely associated with one CSA, it is still possible to consider defining the CSAs containing the sample SEAs as larger area sampling units in the future for possible sample replacement or rotation.

One of the first issues addressed by the working group on the PHS sample design was whether to limit the scope of the survey to include only agricultural households. The recommendation was to exclude from the survey households which are not engaged in agricultural activities. At the same time, it was decided not to set lower limits on the land area or number of livestock and poultry to identify agricultural households. This is similar to the current approach of screening for agricultural households.

During the listing operation, the households will be asked the following questions:

> Was the household engaged in any of the following activities during (reference period):
>
> a. Crop production?
> b. Livestock production?
> c. Poultry production?
> d. Fish farming?

If the answer to all these questions is no, the household would be excluded from the listing frame for the selection of sample agricultural households for the PHS.

The reason for excluding the non-agricultural households is to improve the efficiency of the sampling frame for crop and livestock production and other agricultural characteristics. Although the rural households of landless farm laborers and those engaged in other economic activities are of analytical interest, they can best be studied through other surveys such as the Living Conditions Monitoring Survey.

When the owners do not live on the farm but there is a full-time manager living there, it is recommended to include the household of the farm manager in the listing frame for the PHS in sample SEAs. In this case the manager would be considered the farm operator who can generally provide information on the crop and livestock production of the farm.

Another important issue discussed with the working group was whether to include agricultural households in urban areas in the sampling frame for the PHS. Although it would be ideal to include in the survey agricultural production for households living in urban areas, there is also concern about spending more of the limited survey resources to cover urban agriculture which mostly involves garden plots. As a compromise, the working group recommended to include in the PHS sampling frame urban SEAs in which 70 percent or more of the households are agricultural according to the 2000 Zambia Census.

Table 1 shows the distribution of all SEAs from the 2000 Zambia Census by province, rural and urban, and percentage of agricultural households. A total of 16,746 SEAs were defined for the 2000 Census: 12,202 rural and 4,544 urban. All of the rural SEAs are included in the sampling frame for the PHS. It can be seen in Table 1 that a total of 586 urban SEAs have 70 percent or more agricultural households. Although these urban SEAs in the frame only represent about 12.9 percent of all the urban SEAs, they contain 32.2 percent of the 211,670 urban agricultural households identified in the sampling frame. The 70 percent cut-off for agricultural households in urban SEAs is a compromise to identify SEAs with predominantly agricultural activities. The urban sample will only be about 5 percent of the total, but it will be possible to study the crop and livestock production found in these areas to determine whether the urban sampling frame should be expanded in the future.

Table 1. Distribution of Sample SEAs in 2000 Zambia Census of Population and Housing Frame by Percentage of Agricultural Households, Province, Rural and Urban

Province, Urban/Rural	Number of Sample SEAs in 2000 Zambia Census Frame							
	Total	0% agric. hhs.	0.1-19.9% agric. hhs.	20-49.9% agric. hhs.	50-69.9% agric. hhs.	70-79.9% agric. hhs.	80-89.9% agric. hhs.	90-100% agric. hhs.
ZAMBIA	16,746	88	1,659	1,957	1,442	1,016	1,863	8,721
Rural	12,202	4	106	433	645	747	1,676	8,591
Urban	4,544	84	1,553	1,524	797	269	187	130
Central	1,754	3	86	200	158	115	235	957
Rural	1,412	0	19	65	71	87	220	950
Urban	342	3	67	135	87	28	15	7
Copperbelt	2,259	4	238	769	443	193	213	399
Rural	642	0	6	26	50	70	145	345
Urban	1,617	4	232	743	393	123	68	54
Eastern	2,480	1	33	77	74	67	223	2,005
Rural	2,314	0	2	14	37	54	210	1,997
Urban	166	1	31	63	37	13	13	8
Luapula	1,504	0	12	88	170	143	240	851
Rural	1,363	0	6	54	127	121	221	834
Urban	141	0	6	34	43	22	19	17
Lusaka	1,836	71	1,039	351	121	64	68	122
Rural	377	1	19	62	68	50	59	118
Urban	1,459	70	1,020	289	53	14	9	4
Northern	2,531	0	44	137	190	166	355	1,639
Rural	2,275	0	14	74	115	137	316	1,619
Urban	256	0	30	63	75	29	39	20
Northwestern	1,122	2	6	46	56	70	162	780
Rural	1,024	1	4	13	20	61	153	772
Urban	98	1	2	33	36	9	9	8
Southern	1,847	5	165	214	169	124	224	946
Rural	1,503	0	26	94	116	109	216	942
Urban	344	5	139	120	53	15	8	4
Western	1,413	2	36	75	61	74	143	1,022
Rural	1,292	2	10	31	41	58	136	1,014
Urban	121	0	26	44	20	16	7	8

Table 2 shows the distribution of the rural and urban SEAs in the PHS sampling frame from the 2000 Census (excluding the urban SEAs with less than 70 percent agricultural households) by the number of agricultural households in the SEA. Although there is considerable variability in the number of agricultural households per SEA, the selection of the sample SEAs with probability proportional to size (PPS) will improve the efficiency of the sampling frame. Given that non-agricultural households will not be included in the PHS, the measure of size will be based on the number of agricultural households in the SEA.

Table 2. Distribution of SEAs in Post-Harvest Survey Frame by Number of Agricultural Households in SEA, and Province, Rural and Urban

Province	Total	0 agric. hhs.	1-9 agric. hhs.	10-29 agric. hhs	30-49 agric. hhs.	50-99 agric. hhs.	100-199 agric. hhs.	200-299 agric. hhs.	300-399 agric. hhs.	400+ agric. hhs.
ZAMBIA	12,788	4	68	412	1,318	6,816	3,897	243	23	7
Rural	12,202	4	66	401	1,282	6,500	3,689	236	19	5
Urban	586	0	2	11	36	316	208	7	4	2
Central	1,462	0	11	60	191	858	332	10	0	0
Rural	1,412	0	10	56	187	832	317	10	0	0
Urban	50	0	1	4	4	26	15	0	0	0
Copperbelt	887	0	6	34	72	416	329	20	7	3
Rural	642	0	5	31	61	274	249	17	4	1
Urban	245	0	1	3	11	142	80	3	3	2
Eastern	2,348	0	2	33	172	1,245	842	52	2	0
Rural	2,314	0	2	33	171	1,227	827	52	2	0
Urban	34	0	0	0	1	18	15	0	0	0
Luapula	1,421	0	4	41	162	717	451	39	5	2
Rural	1,363	0	4	41	155	687	430	39	5	2
Urban	58	0	0	0	7	30	21	0	0	0
Lusaka	404	1	4	44	71	171	101	11	1	0
Rural	377	1	4	42	65	162	92	10	1	0
Urban	27	0	0	2	6	9	9	1	0	0
Northern	2,363	0	12	85	215	1,315	698	37	0	1
Rural	2,275	0	12	83	208	1,270	664	37	0	1
Urban	88	0	0	2	7	45	34	0	0	0
Northwestern	1,050	1	5	29	130	626	242	16	1	0
Rural	1,024	1	5	29	130	609	234	15	1	0
Urban	26	0	0	0	0	17	8	1	0	0
Southern	1,530	0	11	58	204	797	429	25	5	1
Rural	1,503	0	11	58	204	784	417	24	4	1
Urban	27	0	0	0	0	13	12	1	1	0
Western	1,323	2	13	28	101	671	473	33	2	0
Rural	1,292	2	13	28	101	655	459	32	2	0
Urban	31	0	0	0	0	16	14	1	0	0

Since the urban SEAs in the frame are limited to those with at least 70 percent agricultural households, the SEAs with few agricultural households are mostly rural. It can be seen in Table 2 that there are four rural SEAs without any agricultural households. Since the sample SEAs will be selected with PPS within each stratum, these four rural SEAs with no households will have a zero probability of selection. Another 68 SEAs in the frame have 1 to 9 agricultural households, so they would have a very small probability of selection. The working group on the PHS sample design discussed the possibility of establishing a minimum measure of size for such SEAs, but they decided to leave them in the frame with the original measure of size. In the case of any such SEA with few agricultural households which is selected in the PHS sample, it would be combined with an adjacent SEA to form a PSU with a minimum of 30 agricultural households.

Table 3 shows the distribution of all households and agricultural households in the PHS sampling frame by province, urban and rural, with the corresponding averages per SEA, and the percent of agricultural households. The average number of households per SEA is 100 for rural SEAs and 116 for urban SEAs, and the corresponding average number of agricultural households is 89 for rural SEAs and 95 for urban SEAs (those with 70 percent or more agricultural households). The overall percentage of agricultural households is 88.4 percent for rural SEAs and 82.0 percent for urban SEAs in the PHS sampling frame.

Table 3. Distribution of All Households and Agricultural Households in PHS Frame by Province, Rural and Urban, with Corresponding Averages per SEA and Percentage of Agricultural Households

Province, Urban/Rural	All Households		Agricultural Households		Percent Agricultural Households
	Total Number	Average Number per SEA	Total Number	Average Number per SEA	
ZAMBIA	1,292,057	101	1,138,407	89	88.1%
Rural	1,223,874	100	1,082,482	89	88.4%
Urban	68,183	116	55,925	95	82.0%
Central	134,275	92	116,522	80	86.8%
Rural	129,084	91	112,379	80	87.1%
Urban	5,191	104	4,143	83	79.8%
Copperbelt	103,261	116	86,960	98	84.2%
Rural	73,295	114	62,454	97	85.2%
Urban	29,966	122	24,506	100	81.8%
Eastern	235,462	100	223,523	95	94.9%
Rural	231,413	100	220,152	95	95.1%
Urban	4,049	119	3,371	99	83.3%
Luapula	152,349	107	131,068	92	86.0%
Rural	146,134	107	125,870	92	86.1%
Urban	6,215	107	5,198	90	83.6%
Lusaka	46,266	115	31,824	79	68.8%
Rural	43,343	115	29,511	78	68.1%
Urban	2,923	108	2,313	86	79.1%
Northern	232,135	98	206,885	88	89.1%
Rural	222,621	98	198,951	87	89.4%
Urban	9,514	108	7,934	90	83.4%
Northwestern	93,550	89	85,432	81	91.3%
Rural	90,725	89	83,089	81	91.6%
Urban	2,825	109	2,343	90	82.9%
Southern	157,240	103	131,379	86	83.6%
Rural	153,653	102	128,450	85	83.6%
Urban	3,587	133	2,929	108	81.7%
Western	137,519	104	124,814	94	90.8%
Rural	133,606	103	121,626	94	91.0%
Urban	3,913	126	3,188	103	81.5%

The distribution of the SEAs and agricultural households in the new PHS sampling frame by district, urban and rural, is presented in Table 4.

Table 4. Total Number of SEAs and Agricultural Households in the Sampling Frame for the Post-Harvest Survey by Province and District, Rural and Urban, Based on 2000 Zambia Census of Population and Housing

Province/ District	Total		Rural		Urban	
	No. SEAs	No. Agric. Households	No. SEAs	No. Agric. Households	No. SEAs	No. Agric. Households
ZAMBIA	12,788	1,138,407	12,202	1,082,482	586	55,925
CENTRAL	1,462	116,522	1,412	112,379	50	4,143
Chibombo	426	31,823	424	31,681	2	142
Kabwe Urban	35	2,841	4	235	31	2,606
Kapiri Mposhi	372	27,223	368	26,898	4	325
Mkushi	192	13,954	192	13,954	0	0
Mumbwa	220	20,760	211	20,069	9	691
Serenje	217	19,921	213	19,542	4	379
COPPERBELT	887	86,960	642	62,454	245	24,506
Chililabombwe	48	4,177	31	2,548	17	1,629
Chingola	69	6,307	47	4,336	22	1,971
Kalulushi	43	5,127	27	3,714	16	1,413
Kitwe	74	6,829	28	2,697	46	4,132
Luanshya	98	8,302	64	5,176	34	3,126
Lufwanyana	117	11,658	117	11,658	0	0
Masaiti	176	17,778	176	17,778	0	0
Mpongwe	93	10,364	93	10,364	0	0
Mufulira	90	7,223	59	4,183	31	3040
Ndola Urban	79	9,195	0	0	79	9195
EASTERN	2,348	223,523	2,314	220,152	34	3,371
Chadiza	175	14,987	170	14,606	5	381
Chama	112	13,835	107	13,260	5	575
Chipata	600	53,435	590	52,425	10	1,010
Katete	371	33,814	367	33,366	4	448
Lundazi	380	42,830	379	42,711	1	119
Mambwe	92	8,848	92	8,848	0	0
Nyimba	142	11,756	142	11,756	0	0
Petauke	476	44,018	467	43,180	9	838
LUAPULA	1,421	131,068	1,363	125,870	58	5,198
Chienge	165	12,373	165	12,373	0	0
Kawambwa	198	18,216	175	16,363	23	1,853
Mansa	275	27,757	274	27,652	1	105
Milenge	66	5,484	66	5,484	0	0
Mwense	216	21,424	210	20,818	6	606
Nchelenge	176	15,545	173	15,107	3	438
Samfya	325	30,269	300	28,073	25	2196
LUSAKA	404	31,824	377	29,511	27	2,313
Chongwe	185	16,905	185	16,905	0	0
Kafue	170	10,286	162	9,828	8	458
Luangwa	32	2,877	30	2,778	2	99
Lusaka Urban	17	1,756	0	0	17	1,756

Table 4. Total Number of SEAs and Agricultural Households in the Sampling Frame for the Post-Harvest Survey by Province and District, Rural and Urban, Based on 2000 Zambia Census of Population and Housing (Continued)

Province/ District	Total		Rural		Urban	
	No. SEAs	No. Agric. Households	No. SEAs	No. Agric. Households	No. SEAs	No. Agric. Households
NORTHERN	2,363	206,885	2,275	198,951	88	7,934
Chilubi	127	12,672	124	12,463	3	209
Chinsali	226	21,872	211	20,531	15	1,341
Isoka	191	16,859	184	16,351	7	508
Kaputa	187	13,205	186	13,082	1	123
Kasama	231	22,369	192	18,796	39	3,573
Luwingu	192	14,151	188	13,930	4	221
Mbala	299	25,441	297	25,224	2	217
Mpika	233	22,562	223	21,639	10	923
Mporokoso	130	13,297	129	13,221	1	76
Mpulungu	126	9,284	126	9,284	0	0
Mungwi	269	22,390	267	22,208	2	182
Nakonde	152	12,783	148	12,222	4	561
NORTHWESTERN	1,050	85,432	1,024	83,089	26	2,343
Chavuma	55	5,975	55	5,975	0	0
Kabompo	167	11,981	163	11,653	4	328
Kasempa	90	8,004	89	7,872	1	132
Mufumbwe	74	7,160	71	6,808	3	352
Mwinilunga	275	19,511	269	19,017	6	494
Solwesi	281	22,186	273	21,435	8	751
Zambesi	108	10,615	104	10,329	4	286
SOUTHERN	1,530	131,379	1,503	128,450	27	2,929
Choma	340	22,789	335	22,307	5	482
Gwembe	52	4,743	49	4,510	3	233
Itezi-tezi	76	5,929	73	5,597	3	332
Kalomo	270	22,428	269	22,345	1	83
Kazungula	125	10,550	125	10,550	0	0
Livingstone	17	1,124	12	649	5	475
Mazabuka	191	19,019	186	18,163	5	856
Monze	204	20,211	202	20,058	2	153
Namwala	92	10,942	90	10,708	2	234
Siavonga	66	6,089	66	6,089	0	0
Sinazongwe	97	7,555	96	7,474	1	81
WESTERN	1,323	124,814	1,292	121,626	31	3,188
Kalabo	208	22,525	198	21,414	10	1,111
Kaoma	280	24,710	277	24,389	3	321
Lukulu	108	12,152	106	12,009	2	143
Mongu	271	21,096	265	20,478	6	618
Senanga	181	18,081	180	18,000	1	81
Sesheke	145	13,530	136	12,616	9	914
Shangombo	130	12,720	130	12,720	0	0

3. STRATIFICATION FOR NEW PHS SAMPLE DESIGN

One of the most important features of an efficient sample design is the stratification of the sampling frame into homogeneous areas. The sample selection is carried out independently within each stratum, although it is also desirable to order the PSUs by certain criteria within each stratum to provide further implicit stratification when systematic selection is used. The nature of the stratification depends on the most important characteristics to be measured in the survey, as well as the domains of analysis. The most effective stratification is at the PSU level, although stratification of the listed households in sample SEAs at the second stage is also beneficial to select larger farms and particular crops of interest with a higher probability.

3.1. Stratification of PSUs for New PHS Sampling Frame

The first level of stratification generally corresponds to the major geographic domains defined for the PHS. Although most survey estimates will be made for the nine provinces and at the national level, some estimates may also be produced at the district level. The CSO wants to ensure that each district is allocated a minimum of two sample SEAs. Therefore the sample SEAs will be stratified by district, as in the previous sampling frame for the PHS. Given a certain amount of homogeneity of agricultural characteristics within each district, this should provide a reasonable level of sampling efficiency. It is also possible to introduce further implicit stratification of the sample within each district by ordering the frame by certain criteria prior to the selection of the SEAs systematically with PPS.

Within each district, the frame of SEAs was ordered by certain characteristics to provide further implicit stratification when the sample is selected systematically with PPS. The number of sample SEAs allocated to most districts was too small to establish explicit rural and urban strata, so the rural and urban region was the first sorting variable. Within each district, the urban SEAs appear in the sorted frame following the rural SEAs.

In order to ensure a representative distribution of the new PHS sample for certain crops, a new crop stratification code was introduced. Eight crops were identified to receive special treatment in the new sample design to improve of the precision of the survey estimates of crop area and production: sorghum, rice, cotton, Burley tobacco, Virginia tobacco, sunflower, soybeans and paprika. In the previous surveys the CVs for these important crops was relatively high because of the smaller number of observations or geographical concentration. A crop stratum code was assigned to each SEA based on which of these crops was predominant (excluding sorghum), that is, grown by more households in the SEA. The SEAs where none of the seven crops was predominant was given a “general” stratum code. Table 5 shows the distribution of SEAs in the PHS sampling frame by the crop stratum code and province. It can be seen that at the national level the SEAs are fairly evenly distributed by crop, except for Virginia tobacco (with 318 SEAs in the frame), Burley tobacco (with 931 SEAs) and paprika (with 753 SEAs). The distribution of the SEAs by crop stratum vary considerably by province, given the different cropping patterns.

Table 5. Distribution of SEAs in the PHS Sampling Frame by Province and Predominant Crop Stratum

Province	Total Number of SEAs in PHS Sampling Frame by Predominant Crop Stratum								
	Total	(1) Rice	(2) Cotton	(3) Burley Tobacco	(4) Virginia Tobacco	(5) Sunflower	(6) Soybeans	(7) Paprika	(8) General
ZAMBIA	12,788	1,278	1,962	931	318	2,871	2,027	753	2,648
Central	1,462	14	473	28	34	411	133	86	283
Copperbelt	887	3	4	23	8	46	480	133	190
Eastern	2,348	164	1,099	115	8	515	216	11	220
Luapula	1,421	272	3	205	66	79	359	88	349
Lusaka	404	3	62	6	1	133	38	42	119
Northern	2,363	308	2	219	79	606	531	78	540
Northwestern	1,050	63	5	155	41	277	152	100	257
Southern	1,530	-	290	26	17	774	83	120	220
Western	1,323	451	24	154	64	30	35	95	470

Sorghum is also included in the targeted crop list and was initially included in the crop stratification. However, given that sorghum is grown in about 82 percent of the SEAs in the new sampling frame for the PHS, most of the SEAs were assigned to the sorghum stratum, making the other crop stratification less effective. As a result, sorghum was dropped from the first stage crop stratification. On the other hand, only 22 percent of the agricultural households in the frame grew sorghum, so this crop was integrated into the second stage stratification scheme described in Section 3.2.

The crop stratum was the second ordering variable for the sampling frame of SEAs within each district. The implicit stratification of SEAs by predominant crop will ensure a representative sample for each crop, with a proportional allocation of the sample SEAs by crop. Given the first level stratification by district, the sample is too small to establish explicit crop strata within each district.

Following the ordering of the frame by rural/urban and crop stratum codes, the SEAs in the frame for each district were sorted by all the hierarchical geographic codes below the district level: constituency, ward, CSA and SEA. This will ensure that the geographical distribution of the sample SEAs is representative. This implicit geographical stratification should also improve the efficiency of the sample for agricultural characteristics, given the similarity of cropping patterns and animal raising in neighboring areas.

3.2. Stratification of Households at the Second Sampling Stage

The listing of households will be used to stratify the households by farm size, number of livestock and the growing of special crops at the second sampling stage within each sample SEA. The previous sample design included stratification of households listed in sample SEAs by two farm size categories: Category A with 0 to 4.99 hectares (has.) and Category B with 5 to 19.99 has. The previous report examined the distribution of households by farm size and concluded that it would be more efficient to subdivide the first category into two

categories. As a result, the working group recommended establishing the following farm size categories for the stratification of households listed in sample SEAs:

Category A - 0 - 1.99 has.
Category B - 2.00 - 4.99 has.
Category C - 5 - 19.99 has.

In order to simplify the selection and estimation procedures for the livestock and crop stratification at the second sampling stage, the working group decided to integrate this stratification with Categories A, B and C based on farm size. The Category C households will generally be included in the sample with certainty (up to 10 households), and the Category B households will be selected with a higher probability than the Category A households. During the listing operation, it will be necessary to collect information on farm size, similar to the current procedures, as well as the number of livestock and poultry, and the presence of particular targeted crops.

Any farms with a large number of livestock or poultry will be added to Category C (if they do not qualify based on land area). The following minimum number of animals will be used to assign listed households to Category C:

Cattle - 50
Pigs - 20
Goats - 30
Poultry - 50

In addition, the same eight targeted crops identified previously (sorghum, rice, cotton, Burley tobacco, Virginia tobacco, sunflower, soybeans and paprika) were identified for special treatment at the second sampling stage. Within each sample SEA, the households will first be stratified into Categories A, B and C according to the farm size and number of livestock and poultry. Then households may be added to Categories B and C based on the special crops, using the following criteria:

(1) If the sample SEA only has 1 or 2 households with any of these individual crops, these households should be assigned to Category C (in case they do not qualify based on land area and animals).

(2) If condition (1) does not apply, but the sample SEA has only 3 to 5 households with any of these individual crops, such households should be assigned to Category B (if they were previously assigned to Category A based on land area and livestock).

The allocation of 20 sample households by category within each sample SEA is described in Section 4 on Sample Size and Allocation.

The farm size specified for Category C (5-19.99 has.) does not include farms with 20 or more has., because these are supposed to be included in the special frame for large commercial farms which is supposed to be completely enumerated in a special survey for these farms. Given the large contribution of the farms with 20 or more has. to the production for certain crops, it is recommended to integrate these surveys as much as possible. In this case a

multiple frame would be used for the integrated survey: (1) a list frame for the large farms with 20 or more has., which would continue to be included in the sample with certainty; and (2) an area frame of SEAs to cover the remaining agricultural households. When any household listed in the sample SEAs is found to have 20 or more has., it would be necessary to verify that it is included in the list frame of large farms. If it is missing from the list frame, it should be included in the sample for the integrated survey with certainty at the second stage, and would receive the same weight (expansion factor) as the sample SEA (generally this will be the weight for the Category C households).

The list frame of large commercial farms should receive special treatment for the data collection and estimation procedures, given that many of them are unique and have a relatively high contribution to the total production of certain crops. A strong effort should be made to collect the data for these large farms. An effective outreach program should be designed to obtain their cooperation, which sometimes may require contact by higher-level CSO officials. In cases where very large farms cannot be interviewed, it is recommended to impute the missing information based on historical data or independent sources, such as administrative information from farm associations or government records.

4. SAMPLE SIZE AND ALLOCATION

The sample size for a particular survey is determined by the accuracy required for the survey estimates for each domain, as well as by the resource and operational constraints. The accuracy of the survey results depends on both the sampling error, which can be measured through variance estimation, and the nonsampling error, which can only partially be measured through expensive re-interview or validation studies. The sampling error is inversely proportional to the sample size. On the other hand, the nonsampling error may increase with the sample size, since it is more difficult to control the quality of a larger operation. It is therefore important that the overall sample size be manageable for quality and operational purposes.

Given the two-stage sample design for the PHS, it is important to examine the allocation of first stage and second stage sampling units. The previous PHS sampling methodology of selecting 20 households per sample SEA was based on cost and operational considerations, and it is reasonable to continue with this sampling strategy. However, the new sample design provides more stratification at the second sampling stage based on the listing information, so it is possible to improve the efficiency of the allocation of the 20 sample households within each sample SEA. The sample allocation is described separately for first and second stage sampling units.

4.1. Number and Distribution of Sample SEAs

The report on "Review of Sample Design for Post-Harvest Survey (1997/98) and Recommendations for Improving the Sampling Strategy and Estimation Procedures" includes tables on the measures of precision (standard errors, coefficients of variation and 95 percent confidence intervals) for selected estimates from the 1997/98 PHS, calculated using the CENVAR software. These tables also show the design effect for each survey estimate, which mostly measures the clustering effect from the multi-stage sample design. It can been seen in those tables that the CVs for total crop production at the national level are fairly high for most crops, and vary by the number of sample households growing the crop. Only four crops have CVs lower than 10 percent: maize, millet, groundnut and cassava. On the other hand, the CV for the estimate of total production is higher than 20 percent for seven crops: rice, sunflower, soybeans, Irish potatoes, Virginia tobacco, Burley tobacco and cowpeas. The crop with the highest CV was soybeans (50.8 percent). In order to decrease these CVs substantially, it would be necessary to increase the number of sample SEAs. Given that the limited resources will only permit a very small increase in the number of sample SEAs, the new sample design attempts to improve the efficiency of the stratification of households at the second sampling stage to obtain lower CVs for the most important crops.

Given the level of resources available for conducting the PHS each year, the CSO decided that the maximum total number of sample SEAs which can be enumerated is 410. There are approximately 207 enumerators available for the PHS data collection, so each enumerator would cover an average of two sample SEAs. This represents a very slight increase from the previous sample of 405 sample SEAs. Since the CSO did not replace the missing sample SEAs in the previous surveys, the effective sample size was actually considerably less; for example, only 383 sample SEAs were enumerated for the 1997/98 PHS. For the new PHS, it

is recommended to replace any sample SEA which cannot be enumerated, in order to maintain the effective sample size. This should result in a modest improvement in the precision of the survey estimates, although there would still be a corresponding bias when the original sample SEAs cannot be enumerated.

For national-level estimates from the PHS data, it is efficient to allocate the sample SEAs to each province and stratum approximately proportionally to the number of agricultural households in the frame. However, some PHS estimates will also be tabulated at the provincial level, so it is necessary to establish a minimum number of sample SEAs for the smallest provinces. In determining the sample allocation scheme, we first examined the proportional allocation of 400 SEAs by province, as shown in Table 6. The 10 additional sample SEAs were allocated later to the districts which only received one sample SEA based on the proportional allocation. The proportional allocation of sample SEAs by province was then adjusted by establishing a minimum of 24 sample SEAs for the smallest province (Lusaka), and a maximum of 72 sample SEAs for the largest provinces (Eastern and Northern). This adjusted proportional allocation should be efficient for both national and provincial-level estimates. It is similar to the previous distribution of the sample SEAs, which is also shown in Table 6, although the number of sample SEAs for Lusaka was increased from 14 to 24, and the maximum sample size was decreased to 72 sample SEAs for the largest province. It is interesting to note that in the case of the two largest provinces, the highest proportional allocation changed from Northern to Eastern Province, indicating a shift in the proportion of households. Also, the previous frame used the total number of households for allocating the sample, while the new sampling frame is based on the number of agricultural households. In the 2000 Zambia Census, Eastern Province had a total of 229,902 households compared to 216,791 for Northern Province. Eastern Province also had a higher percent of agricultural households (94.9 percent) compared to Northern Province (89.1 percent).

Table 6. Distribution of Agricultural Households in New Sampling Frame for Post-Harvest Survey by Province, Proportional Allocation of Sample SEAs, and Initial Adjusted Sample Allocation by Province

Province	Total Number of Agricultural Households in Sampling Frame	Percent of Agricultural Households	Proportional Allocation of 400 Sample SEAs	Initial Adjusted Allocation of Sample SEAs	Previous Allocation of Sample SEAs
ZAMBIA	1,138,414	100.0%	400	400	405
Central	116,522	10.2%	41	40	40
Copperbelt	86,960	7.6%	31	30	24
Eastern	223,523	19.6%	79	72	72
Luapula	131,068	11.5%	46	44	49
Lusaka	31,831	2.8%	11	24	14
Northern	206,885	18.2%	73	72	80
Northwestern	85,432	7.5%	30	30	30
Southern	131,379	11.5%	46	44	50
Western	124,814	11.0%	44	44	46

After determining the adjusted allocation of 400 sample SEAs by province specified in Table 6, these SEAs were allocated to districts within each province proportionally to the number of agricultural households. Table 7 shows the percent of the agricultural households in the PHS sampling frame by district within each province, with the corresponding proportional allocation of the sample SEAs by district. It can be seen that a few districts were proportionally allocated only one sample SEA. Given that each district is a stratum requiring a minimum of two sample SEAs, this proportional allocation was adjusted by increasing the number of sample SEAs to a minimum of two per district. In rounding the number of sample SEAs allocated to each district to an integer, it was found that sometimes the total number of SEAs for the province increased or decreased by one, so it was necessary to examine the decimals in the allocation of sample SEAs to adjust the final sample size for the province. The final number of sample SEAs allocated to each province was also rounded up to an even number. Table 7 shows the proportional allocation of the sample SEAs by district within each province, and the final adjusted allocation of sample SEAs. The sample of SEAs selected for the new PHS was based on this adjusted allocation specified in Table 7.

Table 7. Percent Distribution of Agricultural Households in the Sampling Frame by Province and District, and Proposed Allocation of Sample SEAs for Post-Harvest Survey

Province/ District	Total No. Agric. Households	Percent Households Within Province	Initial Allocation of Sample SEAs Proportionally within Province	Adjusted (Final) Sample Allocation of Sample SEAs
ZAMBIA	1,138,407		400	410
CENTRAL	116,522	100.0%	40	42
Chibombo	31,823	27.3%	11	11
Kabwe Urban	2,841	2.4%	1	2
Kapiri Mposhi	27,223	23.4%	9	10
Mkushi	13,954	12.0%	5	5
Mumbwa	20,760	17.8%	7	7
Serenje	19,921	17.1%	7	7
COPPERBELT	86,960	100.0%	30	32
Chililabombwe	4,177	4.8%	1	2
Chingola	6,307	7.3%	2	2
Kalulushi	5,127	5.9%	2	2
Kitwe	6,829	7.9%	2	3
Luanshya	8,302	9.5%	3	3
Lufwanyana	11,658	13.4%	4	4
Masaiti	17,778	20.4%	6	6
Mpongwe	10,364	11.9%	4	4
Mufulira	7,223	8.3%	2	3
Ndola Urban	9,195	10.6%	3	3
EASTERN	223,523	100.0%	72	72
Chadiza	14,987	6.7%	5	5
Chama	13,835	6.2%	4	4
Chipata	53,435	23.9%	17	17
Katete	33,814	15.1%	11	11
Lundazi	42,830	19.2%	14	14
Mambwe	8,848	4.0%	3	3
Nyimba	11,756	5.3%	4	4
Petauke	44,018	19.7%	14	14
LUAPULA	131,068	100.0%	44	44
Chienge	12,373	9.4%	4	5
Kawambwa	18,216	13.9%	6	6
Mansa	27,757	21.2%	9	9
Milenge	5,484	4.2%	2	2
Mwense	21,424	16.3%	7	7
Nchelenge	15,545	11.9%	5	5
Samfya	30,269	23.1%	10	10
LUSAKA	31,824	100.0%	24	26
Chongwe	16,905	53.1%	13	13
Kafue	10,286	32.3%	8	8
Luangwa	2,877	9.0%	2	3
Lusaka Urban	1,756	5.5%	1	2

Table 7. Percent Distribution of Agricultural Households in the Sampling Frame by Province and District, and Proposed Allocation of Sample SEAs for Post-Harvest Survey (Continued)

Province/ District	Total No. Agric. Households	Percent Households Within Province	Initial Allocation of Sample SEAs Proportionally within Province	Adjusted (Final) Sample Allocation of Sample SEAs
NORTHERN	206,885	100.0%	72	72
Chilubi	12,672	6.1%	4	4
Chinsali	21,872	10.6%	8	7
Isoka	16,859	8.1%	6	6
Kaputa	13,205	6.4%	5	5
Kasama	22,369	10.8%	8	8
Luwingu	14,151	6.8%	5	5
Mbala	25,441	12.3%	9	9
Mpika	22,562	10.9%	8	8
Mporokoso	13,297	6.4%	5	5
Mpulungu	9,284	4.5%	3	3
Mungwi	22,390	10.8%	8	8
Nakonde	12,783	6.2%	4	4
NORTHWESTERN	85,432	100.0%	30	32
Chavuma	5,975	7.0%	2	2
Kabompo	11,981	14.0%	4	5
Kasempa	8,004	9.4%	3	3
Mufumbwe	7,160	8.4%	3	3
Mwinilunga	19,511	22.8%	7	7
Solwesi	22,186	26.0%	8	8
Zambesi	10,615	12.4%	4	4
SOUTHERN	131,379	100.0%	44	46
Choma	22,789	17.3%	8	8
Gwembe	4,743	3.6%	2	2
Itezi-tezi	5,929	4.5%	2	2
Kalomo	22,428	17.1%	8	7
Kazungula	10,550	8.0%	4	3
Livingstone	1,124	0.9%	0	2
Mazabuka	19,019	14.5%	6	6
Monze	20,211	15.4%	7	7
Namwala	10,942	8.3%	4	4
Siavonga	6,089	4.6%	2	2
Sinazongwe	7,555	5.8%	3	3
WESTERN	124,814	100.0%	44	44
Kalabo	22,525	18.0%	8	8
Kaoma	24,710	19.8%	9	9
Lukulu	12,152	9.7%	4	4
Mongu	21,096	16.9%	7	7
Senanga	18,081	14.5%	6	6
Sesheke	13,530	10.8%	5	5
Shangombo	12,720	10.2%	4	5

4.2. Allocation of Sample Households within SEA

The new PHS sample design includes more stratification at the second sampling stage in order to improve the sampling efficiency in a cost-effective manner, as described in Section 3. The stratification by three farm size categories was integrated with the stratification for livestock and special crops in order to simplify the sample selection and estimation procedures.

In order to specify the selection and estimation procedures, the following terms are defined:

N = total number of households listed in the sample SEA
N_A = number of households listed in category A within the sample SEA
N_B = number of households listed in category B within the sample SEA
N_C = number of households listed in category C within the sample SEA
n_A = number of sample households selected in category A within the sample SEA
n_B = number of sample households selected in category B within the sample SEA
n_C = number of sample households selected in category C within the sample SEA

The following steps are recommended to allocate the 20 sample households by category within each sample SEA:

(1) If N_C is less than or equal to 10, select all the N_C households in Category C with certainty at the second sampling stage (that is, $n_C = N_C$).

(2) If N_C is greater than 10, select 10 households in Category C (systematically with a random start) at the second sampling stage (that is, $n_C = 10$).

(3) After determining the number of sample households in Category C (n_C), divide the remaining number of sample households in the SEA ($20 - n_C$) by 2, and round up. This will be the number of sample households to be selected in Category B (n_B) if it is less than or equal to N_B; otherwise, $n_B = N_B$.

(4) The number of sample households in Category A (n_A) will be determined as the remainder: $n_A = 20 - n_B - n_C$

Using this procedure, there will be a minimum of five sample households selected in Category B when there are five or more households listed in this category. In cases where there are 10 households selected in Category C, there would be five sample households in Category B and five sample households in Category A.

5. SAMPLE SELECTION PROCEDURES

The sample selection methodology for the new PHS is based on a stratified two-stage sample design. The procedures used for each sampling stage are described separately here.

5.1. First Stage Selection of Sample SEAs

At the first sampling stage the sample SEAs were selected within each stratum (district) systematically with PPS from the ordered list of SEAs in the PHS sampling frame. The measure of size for each SEA is based on the number of agricultural households identified in the 2000 Zambia Census. The sorting of the frame of SEAs within each district provides further implicit stratification by the specified criteria. The following first stage sample selection procedures were used:

(1) Sort the SEAs within each district by the following codes: region (rural/urban), crop stratum, constituency, ward, CSA and SEA.

(2) Cumulate the measures of size (number of households) down the ordered list of SEAs within the district. The final cumulated measure of size will be the total number of agricultural households in the frame for the district (M_h).

(3) To obtain the sampling interval for district h (I_h), divide M_h by the total number of SEAs to be selected in district h (n_h) specified in Table 7: $I_h = M_h/n_h$.

(4) Select a random number (R_h) between 0 and I_h. The sample SEAs in district h will be identified by the following selection numbers:

$S_{hi} = R_h + [\, I_h \, x(i-1)]$, rounded up,
where i = 1, 2, ..., n_h

The i-th selected SEA is the one with a cumulated measure of size closest to S_{hi} but not less than S_{hi}.

The Excel software was used for selecting the sample of 410 sample SEAs for the PHS following these procedures, based on the allocation of the sample SEAs specified in Table 7. The Excel file has a separate spreadsheet for each province, showing the ordered frame of SEAs with the corresponding 2000 Zambia Census information. It documents the first stage systematic selection of sample SEAs with PPS for each district within the province. The file has a summary spreadsheet with the frame information for the 410 sample SEAs, which was used for calculating the weights by category, as described in Section 6 on Estimation Procedures. Table 8 presents a summary of the distribution of the sample 410 sample SEAs by province, rural and urban. Given that frame of sample SEAs within each district was sorted by region (rural and urban) for the systematic PPS selection at the first sampling stage, the final number of rural and urban sample SEAs within each district is based on proportional allocation.

Table 8. Final Distribution of New Sample of SEAs and Households for Post-Harvest Survey by Province, Rural and Urban

Province	Total		Rural		Urban	
	No. Sample SEAs	No. Sample Households	No. Sample SEAs	No. Sample Households	No. Sample SEAs	No. Sample Households
ZAMBIA	410	8,200	388	7,760	22	440
Central	42	840	39	780	3	60
Copperbelt	32	640	25	500	7	140
Eastern	72	1,440	72	1,440	-	-
Luapula	44	880	42	840	2	40
Lusaka	26	520	23	460	3	60
Northern	72	1,440	70	1,400	2	40
Northwestern	32	640	30	600	2	40
Southern	46	920	45	900	1	20
Western	44	880	42	840	2	40

In examining the distribution of agricultural households from the 2000 Census data for the 410 sample SEAs, it was found that the minimum number of agricultural households in a sample SEA is 27. Since this would be sufficient for selecting the sample households for the PHS, it will not be necessary to combine any small sample SEA with an adjacent SEA to form a larger PSU.

5.2. Listing of Households in Sample SEAs

A listing operation will be conducted in each sample SEA to provide an updated frame of households for the second sampling stage. In order to implement the recommended stratification of the households at the second sampling stage, it will be necessary to develop a more comprehensive listing sheet to identify agricultural households and collect data on farm size, number of livestock and the growing of specific crops. Each household identified within the boundaries of the sample SEAs will be listed. The agricultural households listed in each sample SEA will assigned to one of the three categories A, B or C depending on the farm size, number of livestock and growing of special crops, based on the criteria defined in Section 3.2.

5.3. Second Stage Selection of Households in Sample SEAs

At the second sampling stage the households within each stratification category (A, B and C) will be selected separately. First it will be necessary to allocate the 20 households to the three categories using the procedures specified in Section 4.2, in order to determine the number of sample households to be selected in each category (n_A, n_B and n_C). Then the following steps will be used to select the sample households in each category within a sample SEA for the PHS:

(1) The listed agricultural households assigned to each category will be maintained in the same order in which they were listed, in order to obtain a representative sample throughout the SEA using systematic random sampling. One way to organize the listed agricultural households for the sample selection would be to add a column to the listing sheet for the ordering number. The households within each category in a sample SEA can be assigned serial numbers preceded by the letter of the category. For example, the households in Category A would be assigned serial numbers A1, A2, A3, ..., A(N_A).

(2) For each category in an SEA, the specified number of sample agricultural households will be selected systematically with a random start. If the agricultural households in Category C are included in the sample with certainty (that is, $N_C = n_C$), they will all be identified as sample households. This also applies to the Category B agricultural households in any sample SEA where $N_B = n_B$.

(3) For each noncertainty category S in the SEA, the sampling interval (I_{2s}) is defined as the inverse of the sampling rate. The sampling intervals will be calculated as follows:

$$I_{2A} = \frac{N_A}{n_A};\ I_{2B} = \frac{N_B}{n_B};\ I_{2C} = \frac{N_C}{n_C}$$

(3) For each noncertainty category S in the SEA, select a random number (R_{2s}) with two decimal places, between 0.01 and I_{2s}. The sample agricultural households within category s in the sample SEA will be identified by the following selection numbers:

$S_{2si} = R_{2s} + [\, I_{2s} \times (i - 1)]$, rounded up,
where i = 1, 2, 3,..., n_s (the number of agricultural households to be selected in Category s in the sample SEA).

The i-th selected household is the one with a serial number equal to S_{2si}.

A spreadsheet was developed for calculating the sampling interval, generating the random start and identifying the systematic selection of households for each category in a sample SEA. The Excel file includes a separate spreadsheet for each category. Following the listing operation it will be necessary to enter in the selection spreadsheets the total number of agricultural households listed in Categories A, B and C within the sample SEA. This spreadsheet also determines the number of sample households to be selected in each category based on the second stage sample allocation procedures specified in Section 4.2. In the case of the 2003/4 PHS, it is possible that the selection of sample households may have to be completed in the field immediately following the listing operation, given the timing of the survey. However, it may be possible to use the sample selection spreadsheet when a computer can be used in the provincial office. This spreadsheet will facilitate the selection of the sample agricultural households and document the sample selection.

6. REVIEW OF DISTRIBUTION OF CROPS IN 410 SAMPLE SEAS SELECTED FOR THE 2003 PHS

After selecting the sample of 410 sample SEAs for the new PHS, the 2000 Zambia Census data on crops and livestock for these sample SEAs was examined. First a spreadsheet with the census frame information for the 410 sample SEAs was used to estimate the approximate number of households with each crop and type of livestock that can be expected in the new PHS sample, based on a simple random sample of 20 households in each sample SEA. In the of case of the eight special crops included in the second stage stratification scheme, the additional number of households with these crops in Category C included in the sample with certainty at the second sampling stage was also estimated, in order to determine the effect of this sampling strategy. These results are presented in Table 9, which also shows the percent of SEAs and households in the frame with these crops.

It can be seen in this table that the increase in the estimated number of sample households varies from 4.6 percent for sorghum to 66.0 percent for Virginia tobacco. The level of increase in the number of sample households with each crop depends on the number of households with the crop in the sample SEAs. For example, in the case of Virginia tobacco there are apparently many SEAs with only one or two households with this crop.

The approximate increase in the number of sample households with particular crops presented in Table 9 does not include the effect of the sampling strategy for Category B households, which will also increase the number of sample households with the targeted crops. When only 3 to 5 households in the sample SEA grow one of these special crops, these households will be included in Category B (unless they are already in Category C), which will have a higher sampling rate than Category A households. This sampling strategy for Category B will probably especially increase the number of sample households for sorghum, which did not benefit as much from the sampling procedure for Category C as the other crops.

Table 9. Comparison of Number of Sample Households with Eight Targeted Crops Based on Second Stage Sampling Strategy for Category C with Corresponding Random Selection of Households in 410 Sample SEAs

Crop	Percent SEAs with Crop in Frame	Percent Agricultural Households with Crop in Frame	Estimated No. Sample Households with Crop, Selected at Random in Sample SEAs	Estimated No. Sample Households with Crop, after Including with Category C with Certainty	Estimated Percent Increase in No. of Sample Households with Crop
Sorghum	81.9%	22.4%	1,811	1,894	4.6%
Rice	47.9%	6.6%	480	578	20.5%
Cotton	45.0%	7.9%	592	676	14.2%
Burley Tobacco	48.7%	3.3%	289	382	32.0%
Virginia Tobacco	35.5%	1.5%	150	249	66.0%
Sunflower	65.3%	8.6%	671	768	14.5%
Soybeans	65.0%	5.0%	381	495	30.1%
Paprika	47.8%	2.5%	193	307	59.5%

7. SELECTION PROCEDURES FOR REPLACING MISSING SAMPLE SEAS

For the previous PHS each year there were a few sample SEAs which could not be enumerated because they were inaccessible. For example, in the case of the 1996/97 PHS, 22 out of the 405 sample SEAs were not covered by the survey because of inaccessibility. Some of the missing sample SEAs were found to be in swampy areas which were difficult to reach. When sample SEAs are not enumerated there will be a corresponding bias in the survey results, and the effective number of sample SEAs and households in the survey data will be reduced, thus increasing the sampling errors. Given that the SEAs were selected systematically (with PPS) within each district, a missing sample SEA means that a part of the district is not represented in the survey. This is especially important for districts with only a few sample SEAs. Although this may only have a small effect on the national-level estimates, the provincial-level estimates would be more affected. In some districts the missing SEAs were in flooded rice-growing areas, so the survey estimates for the area and production of rice would suffer a corresponding bias.

In order to reduce this bias and maintain the effective sample size, it is recommended to select a replacement sample SEA for each original sample SEA which cannot be covered by the survey. Sometimes it may be possible to select a new sample SEA within the same sample CSA, although in some cases the entire sample CSA may be inaccessible. In this case an alternative would be to use sampling procedures similar to those used for selecting the original sample SEAs in selecting the replacement sample SEAs. This will also ensure that each new replacement SEA is selected from the same part of the frame within the district as the original sample SEA which it is replacing. Although some of the SEAs in this part of the frame may also be inaccessible, the replacement SEA should be as close as possible to the original sample SEA.

One procedure which can be used to select the replacement sample SEAs would be to check the original systematic selection of SEAs in the spreadsheet with the sampling frame. The information for all the SEAs in the frame within half of the sampling interval before and after the sample SEA being replaced can be copied into a separate spreadsheet in order to select a replacement sample SEA with PPS. It may first be necessary to determine whether these SEAs are accessible, and eliminate from the list those which are not accessible. The measures of size (number of agricultural households) for this list of SEAs from the frame should be cumulated in order to select the replacement sample SEA with PPS.

An example is presented here based on the actual PHS sampling frame to illustrate this procedure. Let us assume that the second sample SEA selected in Chibombo District of Central Province became inaccessible and needs to be replaced. The original sample SEA to be replaced is identified as follows:

Province:	01
District:	101
Constituency:	2
Ward:	12
Region:	1 (Rural)
CSA:	2
SEA:	3
Crop Stratum:	2 (Cotton)

The cumulated measure of size for this SEA in the sampling frame is 4,588, and the sampling interval for Chibombo District is 2,893.

Table 10. Example of Selection of Replacement Sample SEA with PPS from PHS Sampling Frame for Chibombo District, Central Province

Province	District	Constituency	Ward	Region	CSA	SEA	Crop Stratum	M.S. (No. Agric. Hhs.)	Original Cum. M.S.	New Cum. M.S.	Selected
1	101	1	19	1	5	3	2	71	3131	71	
1	101	1	19	1	6	1	2	77	3208	148	
1	101	1	19	1	6	2	2	37	3245	185	
1	101	1	19	1	6	3	2	40	3285	225	
1	101	1	19	1	6	4	2	46	3331	271	
1	101	1	19	1	7	1	2	31	3362	302	
1	101	1	19	1	7	2	2	44	3406	346	
1	101	1	19	1	7	3	2	42	3448	388	
1	101	1	19	1	8	1	2	35	3483	423	
1	101	1	19	1	8	2	2	35	3518	458	
1	101	1	19	1	8	3	2	53	3571	511	
1	101	1	19	1	8	4	2	71	3642	582	
1	101	1	19	1	9	1	2	109	3751	691	
1	101	1	19	1	9	2	2	81	3832	772	
1	101	1	19	1	9	3	2	85	3917	857	
1	101	1	19	1	10	1	2	102	4019	959	*
1	101	1	19	1	10	2	2	70	4089	1029	
1	101	1	19	1	10	3	2	90	4179	1119	
1	101	1	19	1	11	1	2	47	4226	1166	
1	101	1	19	1	12	2	2	37	4263	1203	
1	101	1	19	1	13	1	2	41	4304	1244	
1	101	1	19	1	13	2	2	37	4341	1281	
1	101	1	19	1	13	3	2	73	4414	1354	
1	101	1	19	1	13	4	2	62	4476	1416	
1	101	2	12	1	4	2	2	115	4703	1531	
1	101	2	12	1	5	1	2	104	4807	1635	
1	101	2	14	1	2	1	2	55	4862	1690	
1	101	2	14	1	3	2	2	93	4955	1783	
1	101	2	14	1	3	3	2	156	5111	1939	
1	101	2	14	1	4	1	2	77	5188	2016	
1	101	2	14	1	4	2	2	54	5242	2070	
1	101	2	14	1	4	4	2	61	5303	2131	
1	101	2	14	1	5	2	2	55	5358	2186	
1	101	2	14	1	5	4	2	96	5454	2282	
1	101	2	14	1	7	2	2	76	5530	2358	
1	101	2	14	1	10	1	2	109	5639	2467	
1	101	2	15	1	1	1	2	123	5762	2590	
1	101	2	15	1	1	2	2	127	5889	2717	
1	101	2	15	1	1	3	2	77	5966	2794	
1	101	2	15	1	2	1	2	75	6041	2869	

In order to identify the range of SEAs in the frame before and after this SEA, we first divide the sampling interval by 2 and obtain 1,447. The frame for selecting the replacement SEA will include the SEAs with an original cumulated measure of size within the following range:

Lower limit = 4,588 - 1,447 = 3,141
Upper limit = 4,588 + 1,447 = 6,035

The SEAs included in this range from the original cumulated measures of size in the PHS sampling frame (excluding the sample SEA being replaced) are presented in Table 10, which also shows the new cumulated measure of size for the listed SEAs. The new total cumulated measure of size for the listed SEAs is 2,869 (close to one sampling interval).

In the Excel spreadsheet a random number between 1 and 2,869 was generated, 953, identifying the new sample SEA selected for replacement with PPS.

8. ESTIMATION PROCEDURES

8.1. Weighting Procedures

The CSO staff has experience in using appropriate weighting procedures for the previous PHS. In order for the sample estimates from a particular survey to be representative of the population, it is necessary to multiply the data by a sampling weight, or expansion factor. The basic weight for each sample household would be equal to the inverse of its probability of selection (calculated by multiplying the probabilities at each sampling stage).

Based on the current sample design for the PHS, the probability of selection within each SEA is different for the households listed in each category. The probability of selection for sample households in each category within a sample SEA can be generalized as follows:

$$p_{shi} = \frac{m_h \times N_{hi}}{N_h} \times \frac{n_{shi}}{N_{shi}},$$

where:

p_{shi} = probability of selection for the sample households in Category s (that is, A, B or C) within the i-th sample SEA in district (stratum) h

m_h = number of sample SEAs selected in district h

N_{hi} = total number of agricultural households in the frame for the i-th sample SEA in district h

N_h = total number of agricultural households in the frame for district h

n_{shi} = number of sample agricultural households selected in Category s from the listing for the i-th sample SEA in district h

N_{shi} = total number of households in Category s from the listing for the i-th sample SEA in district h

The two terms in p_{Shi} correspond to the first and second stage probabilities of selection; at the first stage the SEAs were selected with PPS, and at the second stage the households were selected with equal probability within each stratification category.

Based on the current sampling procedures, in most sample SEAs the households in Category C will be selected with certainty at the second sampling stage (that is, $n_{Shi} = N_{Shi}$), in which case these households will have the same probability of selection as the sample SEA.

The basic sampling weight is equal to the inverse of the probability of selection. Therefore the corresponding basic weight for the sample households in stratification Category S would be calculated as follows:

$$W_{shi} = \frac{N_h}{m_h \times N_{hi}} \times \frac{N_{shi}}{n_{shi}},$$

where:

W_{shi} = basic weight for the sample households in Category s within the i-th sample SEA in district h

It should be noted that the sample households selected in each stratification category keep the specified weight, even if it is found later that the farm size was misclassified according to the survey data.

It is also important to adjust the weights to take into account the noninterviews in each stratification category within a sample SEA. The numerator of this adjustment factor would be the total number of households selected in the particular category within the sample SEA; the denominator would be the number of completed household questionnaires. The final weight adjusted for noninterviews would be calculated as follows:

$$W_{shi}' = \frac{N_h}{m_h \times N_{hi}} \times \frac{N_{shi}}{n_{shi}} \times \frac{n_{shi}}{n_{shi}}, = \frac{N_h}{m_h \times N_{hi}} \times \frac{N_{shi}}{n_{shi}},$$

where:

W'_{shi} = weight adjusted for noninterviews for the sample households in Category s within the i-th sample SEA in district h

n'_{shij} = number of sample households in Category s with completed interviews within the i-th sample SEA in district h

The CSO has been implementing a similar procedure for adjusting the weights for noninterviews in previous surveys. Instead of first calculating the basic weight, the final weight is calculated directly by substituting the value of n_{shi} with that for n'_{shi} in the formula for the basic weight specified previously.

The Excel spreadsheet with the sampling frame information for the 410 sample SEAs will be used for calculating the final weights for the sample agricultural households in Categories A, B and C in each sample SEA. The formulas specified above are included in the spreadsheet, so it will only be necessary to enter the total number of agricultural households listed for each category in the sample SEA and the number of completed questionnaires for each category, and the weights will be calculated automatically.

Whenever an original sample SEA is replaced, it will be necessary to update the spreadsheet for calculating the weights with the sampling frame information for the replacement SEA. The weight for the sample households in the replacement SEA will be based on its measure of size.

8.2. Types of Survey Estimates

The most common survey estimates to be calculated from the PHS are in the form of totals and ratios. The survey estimate of a total can be expressed as follows:

$$\hat{Y} = \sum_h \sum_i \sum_s \sum_j W_{shi}' y_{shij} \ ,$$

where:

W'_{shi} = final weight for the sample households in Category s within the i-th sample SEA in district h

y_{shij} = value of variable y for the j-th sample household in Category s within the i-th sample SEA in district h

The survey estimate of a ratio is defined as follows:

$\hat{R} = \frac{\hat{Y}}{\hat{X}}$, where $\hat{Y}$ and $\hat{X}$ are estimates of totals for variables y and x, respectively, calculated as specified previously.

In the case of multi-stage sampling, means and proportions are special types of ratios. In the case of the mean, the variable X, in the denominator of the ratio, is defined to equal 1 for each element so that the denominator is the sum of the weights. In the case of a proportion, the variable X in the denominator is also defined to equal 1 for all elements; the variable Y in the numerator is binomial and is defined to equal either 0 or 1, depending on the absence or presence, respectively, of a specified characteristic in the unit observed.

8.3. Ratio Estimation for Particular Crops

In the case of particular crops which have a high level of sampling error because they are rare or grown in limited geographic areas, it may be possible to improve the survey estimates through ratio estimation, assuming that independent data for the crop are available from other sources such as frames maintained by the Ministry of Agriculture or farming associations.

Ratio estimation involves the use of independent information for a survey variable such as area planted for a particular crop. For example, it can be used to estimate total crop production when the total area planted for the crop is known from another source. In this case, the average crop yield would be estimated from the survey data and then multiplied by the total area planted, as follows:

$$\hat{P}_C = \frac{\left(\sum_h \sum_i \sum_s \sum_j W'_{shi} \times y_{Cshij}\right)}{\left(\sum_h \sum_i \sum_s \sum_j W'_{shi} \times x_{Cshij}\right)} \times X_C \text{ ,}$$

where:

y_{Cshij} = production of crop C for the j-th sample household in Category S within the i-th sample SEA in district h

x_{Cshij} = area planted for crop C for the j-th sample household in Category s within the i-th sample SEA in district h

X_C = good estimate of total area planted in crop C from independent source

The first term represents the survey estimate of the average crop yield per hectare. Of course, one limitation of this ratio estimation procedure is the availability of accurate information on the total area planted for the particular crop. However, such data may be available for particular crops such as tobacco which may have farmer associations or special arrangements with a factory.

In other cases such as cotton, an accurate figure for crop production may be available from a processing or marketing company. In this case the total production of cotton from the independent source can be divided by the survey estimate of the average yield for cotton in order to estimate the total area planted in cotton.

8.4. Calculation of Variances

In the publication of the results from each survey it is important to include a statement on the accuracy of the survey data. In addition to presenting tables with calculated sampling errors for the most important survey estimates, the different sources of nonsampling error should be described.

The standard error, or square root of the variance, is used to measure the sampling error, although it may also include a small part of the nonsampling error. The variance estimator should take into account the different aspects of the sample design, such as the stratification and clustering. In order to avoid the time and effort it would require to develop custom variance programs, it is ideal to use an available software package to tabulate the variances. One such program available for calculating the variances for survey data from stratified multi-stage sample designs such as that for the PHS is CENVAR, a component of the Integrated Microcomputer Processing System (IMPS). CENVAR is menu-driven and user-friendly. It uses the data dictionary defined in the DATADICT component of IMPS. It can be used to calculate the variances of totals, means, proportions and other ratios. It produces subpopulation estimates for each category of a classification variable, and these variables can be cross-classified. For each estimate, CENVAR calculates the standard error, coefficient of

variation (CV), 95 percent confidence interval and the design effect (DEFF). This software package uses an ultimate cluster variance estimator.

The report on “Review of Sample Design for Post-Harvest Survey (1997/98) and Recommendations for Improving the Sampling Strategy and Estimation Procedures” includes CENVAR tables for estimates of total area and production of major crops from the 1997/98 PHS data. That CENVAR application can be used as a prototype for future surveys. The CSO has a copy of the IMPS software which includes CENVAR. A short training course in CENVAR was given during the previous visit, but the CSO staff needs to develop experience in using CENVAR by tabulating the standard errors for each survey.

In order to tabulate estimates of standard errors using CENVAR, it is generally necessary to produce a new data input file from the original survey data. Since the CENVAR package will only accept one record type, it is necessary to generate one record for each unit of analysis in the CENVAR data input file. For example, in the case of the estimates by household, such as the average farm size per household, the CENVAR input file should have one record for each in-scope sample household. Each record in the CENVAR data input file should include fields for the stratum, cluster and weight, in addition to the classification and analysis variables which are required for the particular CENVAR analyses. The classification variables are used to produce subpopulation estimates for all their respective categories. The analysis variables are generally continuous variables, such as crop area and production, or count variables, which are equal to 1 if the unit has a certain characteristic and 0 otherwise. CENVAR automatically creates a count variable named INTERCEPT, which is equal to 1 for each record. The INTERCEPT variable can be used to obtain the estimate of the weighted total number of units (for example, the total number of households), or it can be used in the denominator of a ratio in order to obtain a mean or proportion; it can also be used as a classification variable to obtain estimates at the national level.

CENVAR does not accept any blanks in the file. In the case of classification variables, any record with a blank should be imputed with a special code to identify "missing" or "not applicable." The CENVAR output will include estimates for these categories, which can be deleted from the tabulations which will be published. For analysis variables, CENVAR assumes that any missing values are imputed. Once the file is zero-filled, CENVAR will treat any missing value as 0, thus introducing a downward bias in the estimates of means when there are missing values. One way to resolve this problem is to generate an indicator variable for each variable which has missing values. This indicator variable would then be crossed with each classification variable in the subpopulation analyses in order to produce separate estimates for the records with valid data for that variable. The subpopulation estimates for the missing value categories can later be deleted from the CENVAR output tables. This procedure was used for the CENVAR application developed for the 1997/98 PHS data, described in the previous report.

The ultimate cluster variance estimator for a total used by CENVAR can be expressed as follows:

Variance Estimator of a Total

$$V(\hat{Y}) = \sum_{h=1}^{L} \left[\frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} \left(\hat{Y}_{hi} - \frac{\hat{Y}_h}{n_h} \right)^2 \right],$$

where:

$\hat{Y}_{hi} = \sum_s \sum_k W_{shi}' y_{hij}$ = weighted SEA total for variable y

$\hat{Y}_h = \sum_{i=1}^{n_h} \hat{Y}_{hi}$ = weighted district total for variable y

The variance estimator of a ratio used by CENVAR can be expressed as follows:

Variance Estimator of a Ratio

$$V(\hat{R}) = \frac{1}{\hat{X}^2} \left[V(\hat{Y}) + \hat{R}^2 V(\hat{X}) - 2 \hat{R} COV(\hat{X},\hat{Y}) \right],$$

where:

$$COV(\hat{X},\hat{Y}) = \sum_{h=1}^{L} \left[\frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} \left(\hat{X}_{hi} - \frac{\hat{X}_h}{n_h} \right) \left(\hat{Y}_{hi} - \frac{\hat{Y}_h}{n_h} \right) \right]$$

$V(\hat{Y})$ and $V(\hat{X})$ are calculated according to the formula for the variance of a total.

ANNEX I

Working Group Attending Meetings on Post-Harvest Survey Sample Design

Name	Title [1/]
John Kalumbi	Deputy Director, Agriculture and Environment Division
M. Sooka	Statistician
Colby S. Nyasulu	Senior Statistical Officer
Batista Chilopa	Senior Statistician
Doreen Tembo	Statistician
Crispin Sapele	Principal Systems Analyst
Shambulo Kabangu	Systems Analyst
Joseph V. Chanda	Systems Analyst
George S. Namasiku	Systems Analyst
Aaron Phiri	Cartographer (GIS)
Obed C. Kawonga	Nutritionist/Statistician
Dingiswayo Banda	Economist - Ministry of Agriculture and Cooperatives
Nicholas Mwale	Statistician - Ministry of Agriculture and Cooperatives

[1/] All working group members are from CSO, unless otherwise indicated.