Contents of WMS Data Package

List of Tables

The following tables are tabulated by two criteria:
a: socio-economic group
b: region and district

Clarification of Definitions and Calculations
Applied in the Statistical Tables
to the Basic Welfare Monitoring Survey Report

I. Dissaggregation of Results :

A. Dissaggregation by Region and District

The first way in which the data have been disaggregated is **geographical**. The sample design has been stratified in such a way as to be able to give district-level results, and wherever possible results have been broken down by District. A warning must be given that at this level, the sample size becomes quite small, which means that the margin of error due to sampling, is large. District-level results must, therefore, be interpreted with caution.

B. Disaggregation by Socio-economic Group

The second way in which the results are shown is by **socio-economic group** (SEG). The selection of appropriate criteria for defining SEGs was complicated, since it was necessary to ensure that the chosen criteria were accurately represented by the variables in data and that the sample size for each SEG was adequate in order to make meaningful interpretations. Variables which were examined to define the socio-economic groups included: occupation and sex of the head of the household ; land holding size; main sources of income; and area classification of urban/rural.

In the case of the WMS, the occupation of the head of household was found to give erratic and uneven-sized groups, primarily due to the inadequately defined variable for a persons main economic activity. The main economic status was defined in relation to time spent per day on the activity rather than the main source of income and there was potential confusion between the different categories of economic status. To identify a household's main source of income, therefore, it was necessary to use the WMS collected income data directly.

The following criteria were used and combined to define 14 SEGs.

The first classification is by main source of income such that households have been divided between *agricultural households* and *non-agricultural households*.

Agricultural households have then been split into *pasturalists* and *agriculturalists*. The agriculturalists are then separated into *export-crop oriented* (farmers recording sales of export or commercial crops), and *food-crop oriented* (farmers with no export or commercial crop sales). Out of the food crop farmers, a further sub-group, *subsistence farmers*, has been identified and defined as farmers whose consumption of own production is greater in value then their sales of own production. Each of these agriculturalist groups has then been split by *sex of head of household*.

Non-agricultural households are first classified as being either *rural* or *urban*, according to where they live. Then households whose main source of income is either wages or self-employment are classified into *public sector*, *formal sector*, or *informal sector*, using the occupation of the head to identify the sector. Where the head is not a wage earner or self-employed, then the occupation of any other household member who is a wage earner or self-employed is used instead. It is possible, therefore, that a misclassification may be made in cases where the head of the household is one of several members who are wage earners or self-employed , but not the main earner. This is considered to be a relatively rare situation though.

The last group is the non-agricultural households whose main sources of income are neither from wages nor from self-employment. This group has been classified as households with *income from other sources*. These sources may be varied and may include rent, gifts and transfers. This group was not divided into urban and rural due to a small urban sample size for this group.

The final list of SEGs is as follows:

- *Export farmer-male headed*
- *Export farmer-female headed*
- *Food crop farmer-male headed*
- *Food crop farmer-female headed*
- *Subsistence farmer-male headed*
- *Subsistence farmer-female headed*
- *Pastoralist*
- *Public sector urban*
- *Public sector rural*
- *Formal sector urban*
- *Formal sector rural*
- *Informal sector urban*
- *Informal sector rural*
- *Income from other sources*

II. Relative Prices:

Regional (provincial) price deflators, using Nairobi as the base region, were applied to adjust household expenditure and income levels for regional price variations. The deflators are:

Province	**1992/3**
Central	0.918
Coast	0.914
Eastern	0.833

Nyanza	0.783
Rift Valley	0.811
Western	0.818
Kisumu	0.876
Nakuru	0.870
Mombassa	0.916
Nairobi	1.000

These regional price deflators are based on 1992 prices for a basket of food items, and were estimated and supplied by the Ministry of Planning .

III. Adult Equivalents:

The following adult equivalent scale were applied to household members before calculating consumption/expenditure figures per adult equivalent:

Age	Adult equivalent
0-4	0.24
5-14	0.65
15+	1.00

IV. Poverty Line:

Two relative poverty lines were established:

- A relative poverty line set at 66% of mean consumption levels
- A relative poverty line set at 33% of mean consumption levels

V. Poverty Measures:

Poverty was measured using the $P\alpha$ set of measures suggested by Forster, Greer and Thorbecke. These are given as:

Where q = people below the poverty line, z = the poverty line, and y_i = per capita consumption of the ¡th person. When $\alpha = 0$, $P\alpha$ becomes P_0, which gives the **headcount** and measures the ***prevalence*** of poverty . When $\alpha = 1$, $P\alpha$ becomes P_1, which gives the **poverty gap** and measures the *depth* of poverty. When $\alpha = 2$, $P\alpha$ becomes P_2, which gives a measure of the *severity* of poverty.

VI. Weighting:

Weighting of the data followed the specified method established for the Kenya's National Sample Survey and Evaluation Programme (NASSEP III). The basic weights, before adjusting for noninterview, are the reciprocals of the probabilities of selection, which is as follows:

Basic cluster weight = 'cluster weight' * 'household weight"
where:
'cluster weight' = 1/cluster probability
'household weight' = 1/ (10/# of households) or (# households/ 10)

The nonresponse adjustment factor is estimated at the district level as follows:

Nonresponse Adjustment Factor = Cj / Ij
where:
Cj = total number of selected households in the district or
(10 times the number of clusters per district)
Ij = Actual number of households in the district which responded *

* Ij is the number of households in the sample used in the analysis (8102)

The final adjusted weight which is applied to the sample used in the analysis is:

Final Adjusted Weight = Basic Cluster Weight * Nonresponse Adjustment Factor

VII. Issues of Data Problems and Outliers:

Errors of the following type were found in the analysis of the data:

1.) Incomplete data -
* Missing records for households
* Missing variables

Households that were missing records could not be used and were dropped from the analysis. Three criteria were applied in determining the basic sample size (i.e. the number of households) used in the analysis:
- a complete set of records (four records in total, i.e. households could not have missing records)
- valid cluster weights (66 households were located in clusters which did not have cluster weights)
- valid information on variables used to assign households to the socio-economic groups

2.) Categorical variables with values out range

* Most categorical variables had a few cases of illegal codes in all four records

Categorical variables indicating a household's district and whether the household was in a urban or rural area were cleaned using the household's cluster number and a computer print of cluster location. It was not possible to correct all other illegally coded categorical variables without referring back to the original questionnaire. These illegally coded variables were set to system missing and excluded from any particular analysis which required that variable.

3.) Continuous variables with values out of range

* Expenditure, price, wage, and income values out of range

Distributions for expenditure values were skewed considerably as a result of a small number of households reporting excessive expenditures. Validation guidlines for out of range values in the basic survey expenditure were supplied from the Central Bureau of Statistics and applied to the WMS data. Consultations with the Ministry of Planning and the Central Bureau of Statistics indicated that errors in reporting or data entry were likely for these cases and a decision was reached to exclude the out of range values from the present analysis. This involved dropping fewer than 30 rural households which reported total expenditures between 500,000 and 1,000,000 per year.

Due to time constraints and the fact the number households reporting outliers was small in comparison to the total sample size, outliers were dropped from the present analysis, however, this issue should be revisited and a more appropriate solution found in any further analysis.

Follow-up to Data Problems:

The Central Bureau of Statistics has been given documentation in order to follow-up on further cleaning of the data:

1.) Detailed list of households with missing records, including the identification of the cluster and household number for households missing records and the identification of specific records missing

2.) Frequency distributions for all categorical variables with illegal codes

3.) Basic descriptive statistics and histograms of expenditure and income variables with out of range values, as well as a listing households with out of range values by cluster number and household number.

The Central Bureau of Statistics is currently reviewing these data problems by referring to the original survey questionnaires. It is hoped that these issues will be resolved before any further analysis is undertaken.

Guidelines for Using Documentation on WMS data

I. SPSS Data System Files

There are four records of information corresponding to the four sections of the survey questionnaire. Each record of information is contained within an spss data system file, identifiable by the file extension .sys. These spss data system files are the following:

* srt193.sys - Record 1, Household Characteristics
* srt293.sys - Record 2, Household Expenditures
* srt393.sys - Record 3, Household Income
* srt493.sys - Record 4, Household Assets, Amenities, & Land Utilization

Additional spss data system files are:

* skeyvar.sys - Key variable data file, contains variables that use variables from more than one record
* weights.sys - Weight variable file, contains basic weight, adjustment factor, and final weight

All of the above files contain only households used in the present analysis. This sample size is 8102. The original files which contain all households in the data, i.e. even households with missing records & missing variables, are available in files withe the same file name except that the first "s" in the filename is dropped. For example, the original file containing all households for record 1 is called rt193.sys.

There are data dictionaries for each of the data system files which provide basic information of variable and value label definitions, and formulas for calculating new variables.

II. SPSS Computer Programs

There are two types of spss computer programs for each of the spss data system files:

1.) SPSS computer programs which translate dbase files into spss data system files, assign variable and value labels, and create all new variables. Filenames for these take the filename extension .prg and are as follows:

* rt193.prg - record 1 computer program
* rt293.prg - record 2 computer program
* rt393.prg - record 3 computer program
* rt493.prg - record 4 computer program
* keyvar.prg - key variable computer program
* weights.prg - weights computer program
* misshh.prg - computer program which identifies households with missing records

2.) SPSS compute programs which generate tables from each of the information in the different records. These programs follow the sample pattern for identifying the record - the filenames are:

* rt1tab.prg
* rt2tab.prg
* rt3tab.prg
* rt4tab.prg
* keytab.prg

Comments on Questionnaire and Survey Procedures with Recommendations forWMS2

The following comments have been prepared during the course of the processing of the WMS. They indicate problems that were encountered and make suggestions as to how these can be avoided and improvements made to subsequent rounds of the WMS.

General:

Sample design: WMS1 used a fixed 10 household take per cluster for all NASSEP III clusters. This is satisfactory, but if the computerization of the NASSEP frame is completed before the next round, consideration should be given to the idea of stratification within clusters. Criteria for such stratification could include, for the rural areas, sex of head and holding size , and for the urban areas, sex and main occupation of the head.

Enumerators Manual: As it currently stands this is not a particularly useful document, neither for the field staff nor for analysts and users of the survey data. The document should contain specific definitions of all key concepts used in the survey, such as: household, holding, occupational status etc. It is recommended that the document be extensively reviewed and updated.

Data collection: The processing of the data unearthed some major data collection problems. The most important of these was the fact that a very large number of household questionnaires were returned incomplete (one of more forms left blank or missing). This is not permissible since every household must have all four forms. This mistake points to a **weakness in training and/or supervision** and must be watched very carefully at the next round.

Data entry:

Errors of the following types were found in the analysis of the data:

- Incomplete data,
  - Missing records
  - Missing variables
- Coded variables with illegal codes
- Continuous variables out of range

Households that are **missing records,** as noted above, cannot be used in the analysis and had to be dropped. This problem should be detected in the field but, if that fails, it can be detected at the data entry stage by close monitoring and control at the data entry stage. The problem of **missing variables** generally results from the enumerator leaving cells blank, rather than coding zero. The issue needs to be addressed at the training stage and rules established regarding when to leave items blank and when to code zero. (The problem can however be rectified at the processing stage - but this is not good practice). The problem of **illegal codes** should be picked up at the time of data entry if a proper data entry program is used. **Out-of-range**

values can also be picked up at the time of data entry it range checks are specified in advance and built into the data entry program. It is strongly recommend that a data entry package such as IMPS is used for entering the data so that non-legitimate codes and range checks can be run at the time of data entry.

Questionnaire:

The questionnaire is one of the shortest, yet potentially most useful household survey questionnaires that the CBS has produced. It is difficult, however, to know what reliability to place on annual income and expenditure data which is collected from a single interview. A unique opportunity to test this presents itself however this year due to the fact that the Urban Household Budget Survey and the WMS will both be conducted in the same clusters at the same time. It is suggested that a research proposal should be elaborated under which the findings of the two surveys could be compared. It is possible that separate reseach funding could be found for this exercise.

Record 1

1. Main economic status: This is a key variable through which socioeconomic groups (SEGs) can be derived. The current codes however appear ambiguous - especially those relating to agriculture. We suggest that the question be changed to Main Economic Activity during the last 12 months, and that the following codes are used:

1 Agriculture - own or household holding/farm. Note, this will include unpaid family workers, but not those whose primary activity is working on the holdings of others. Further subclassification may then be made using such other household variables as holding size and income

2 Agriculture - pasturalist. Note: it is important to make a distinction here between a pasturalist (eg Maasai herdsman) and a livestock farmer. It is the former that is wanted here (the livestock farmer should be coded as 1)

3. Agriculture - working on the holding/farm of others. This group will include landless labourers

4 Public sector. eg. any government employee, including teachers, doctors etc.

4 Private sector, formal - wage earner. Note: this excludes all those working in the informal (Jua-Kali) sector.

5 Self-employed - private sector, formal.(Not Informal sector)

6 Informal sector - Self-employed and employees.

7 Student

8 Inactive or unemployed

2. A column should be added to indicate whether the household member is absent or present, the absence or presence of the head of household is particularly important for gender analysis and for differentiating between different types of female-headed households. The following codes should be applied.

1 = Present

2 = Temporarily absent
3 = Regularly absent for periods up to 1 week
4 = Regularly absent for 1 - 4 weeks
5 = Regularly absent for periods over 4 weeks
Note: If the member is permanently absent then he/she should not be a household member

Record 2

3. The validity of expenditure estimates that are based on a single interview needs to be tested against the UHBS and a separate research project is proposed.

4. In the meantime, it is probable that better results will be obtained if a normative question is asked i.e. " How much maize do you normally consume in a month. This will reduce the effects of seasonal variability.

5. Also, in order to avoid the bias that is introduced as a result of wages being usually paid at the end of the month, it would be better to use the one-month reference period, for food and other regularly consumed items, rather than the week. This also reduces the bias that results from zero purchases being made during the reference period in question. It is suggested that the idea of using two reference periods (weeks and months) be dropped since there is no real way of deciding which to use.

6. House rent should not include imputed rent, since what is sought is actual expenditure.

7. Questions on agricultural costs are asked for last season and for last year. They should either be asked for both seasons (long and short) or for the past year.

Record 3

8. Livestock products. A separation should be made between the sale and consumption of livestock products.
9. Transfers. The accuracy of the reporting could be improved by including separate questions for transfers in kind and cash transfers.

10. Rents. Should not included imputed rent.

11. Crops: As currently designed, it is impossible to know which are the main crops that have been produced and consumed this is an **important gap.** Columns must be added for "crop codes.

12. Employment of head and of spouse. In addition to the information on current main occupations of the head and spouse it would be very useful to include

information on previous main occupation, as this will make it possible to monitor patterns of occupational change occurring during the period of economic adjustment.

Record 4

13 No immediate comments