Chapter 3

3.1 Methodology

This chapter is divided into two sections. First section presents the methodology of the Sudan Household Health Survey (SHHS) and the second section presents the statistical procedure to be used in current research.

3.2 Sudan Household Health Survey (SHHS)

3.2.1 Sample Design

Design of the sample for the Sudan Household Health Survey (SHHS) provided estimation of a large number of indicators on the basic health situation at the national level and for 25 States (Eastern Equatoria, Central Equatoria, Western Equatoria, Western Bahr El Ghazal, Lakes, Northern Bahr El Ghazal, Warrap, Unity, Upper Nile, Jonglei, South Darfur, West Darfur, North Darfur, South Kordofan, North Kordofan, White Nile, Blue Nile, Gezira, Sinnar, Gadarif, Khartoum, Kassala, , Red Sea, River Nile, Northern). The target universe for the SHHS included the population living in individual households and the nomadic population camping at a location/place at the time of the survey. The units of analysis for the SHHS, therefore, are the individual households and persons within the households. Some questionnaire modules correspond to particular subgroups of the population, such as that for women between the ages of 15 and 49, and children under the age of 5 years. The population living in institutions and group quarters such as hospitals, military bases and prisons, were excluded from the sampling frame. The States were identified as the main sampling domains and a stratified multi-stage sample design was used for the SHHS.

3.2.2 Sampling frame and units of analysis

One of the challenging aspects of planning for the SHHS was compiling a sampling frame with as complete coverage of the Sudan population as possible. This arose because the last Census in Sudan was in 1993, which, for purposes of providing a suitable sampling frame, was considered too far and out of date. Besides, 1993 was a period of armed conflict, and only the garrison towns of Juba, Malakal and Wau and other selected areas were actually enumerated in Southern Sudan. Therefore, no maps and lists actually existed for most of Southern Sudan.

To circumvent the shortcoming, various other sources of geographic information were examined. One of the sources with the best coverage in Southern Sudan was the World Health Organization's list of villages and estimated population developed for the National Immunisation Days (NIDs) campaign. The population estimates were, however, a rough demographic estimation based on the number of under-five children identified by the EPI Programme. The list of villages and estimated population developed for the NIDs campaign was also used for compiling the sampling frame for the three Darfur States. Thus, while for 12 States (South Kordofan, North Kordofan, White Nile, Blue Nile, Sinnar, Gezira, Khartoum, Gadarif, Kassala, Red Sea, River Nile, and Northern), the sampling frame was compiled using the list of villages and estimated population updated by the Central Bureau of Statistics on the basis of the Census enumeration areas, the sampling frames for three Darfur States (South Darfur, West Darfur, and North Darfur) and for all the ten States in Southern Sudan were compiled using the list of villages and estimated population developed for the NIDs campaign.

3.2.3 Stratification

Stratifying the sampling frame into homogeneous areas was one of the most important features of the sample design for the SHHS. Within each stratum,

independent the sample selection was carried out. The domain of the analysis, the available information, and the other important measureable characteristics of survey determined the nature of stratification. Correspondence between the major graphic domains and the first level of stratification was taken into account in SHHS, that is, the 25 States in Sudan. In the case of 12 States, with a town or other relatively large town (for example, with a population of 50,000 or more), it was considered necessary to establish a separate stratum for the towns (urban areas) and for the remainder of the State. The primary sampling units were distributed to rural and urban domains in proportion of the size of rural and urban populations in 12 States including South Kordofan, North Kordofan, White Nile, Blue Nile, Sinnar, Gezira, Khartoum, Gadarif, Kassala, Red Sea, River Nile, the Northern, but in three States in Darfur and all the ten States in Southern Sudan, stratification on the urban and rural level could not be done clusters were distributed directly to the State domain proportional to the size of the primary sampling units (PSUs) directly.

Within each State, the PSUs were ordered geographically by locality/county to ensure a good geographic distribution of the sample through stratifying implicitly after systematic selection of PSUs.

3.2.4 Size and Allocation of Samples

The operational and resource constraints and the required accuracy of estimation for each survey domain was considered in determining the size of sample. The sample size was also determined by the geographic levels at which the survey data were to be tabulated. Since reliable estimates for key indicators were needed for each of the 25 States of Sudan, it was considered necessary to ensure that each State had a sufficient sample size. The survey budget was based on a sample of 25,000 households for Sudan, or about 1,000 households per State,

though an effective sample size of 900 households was considered sufficient for most State-level estimation.

The number of sample PSUs (villages) for the SHHS, and the number of households selected within each sample village/quarter were determined keeping in view the Survey objectives. It was recognized that for estimation at the national level, it would be more efficient to have a proportional allocation of the sample to the States based on their approximate population. However, it was noted that these population estimated were only approximate, and might be over-estimated, and therefore, given the large variability in the population by State, the sample size for the smallest States based on a proportional allocation would be too few to produce reliable results.

Since a similar level of precision was required for the survey results from each State, it was decided to use an equal allocation of 40 sample segments per State. Considering the nature of the survey as well as the logistics, cost of the field operations, and current transportation and communication constraints, it was decided to select 25 households per segment.

3.2.5 Sample selection procedures

The sample selection methodology for the SHHS was based on a stratified multi-stage sample design. The steps involved in the sample selection included the following: Selection of Sample Primary Sampling Units (Villages: For the first stage of selection of the sample for the SHHS, a frame of primary sampling units (PSUs) which covered as much of the population as possible was established. The PSU was defined as the smallest area or administrative unit, which could be identified in the field with commonly recognised boundaries. Any areas that could not be included in the survey because of problems of security or accessibility were excluded from the frame before the first stage selection of sample PSUs. The

villages or quarters constituted the PSUs for the SHHS. Therefore, the list of villages was used as the most effective sampling frame of PSUs for the first stage of sampling. For some States, the list of villages appeared to be complete, and population estimates were available for all villages, so this frame was used for the first stage selection of villages with PPS. In the case of these States, at the first stage of sampling, the sample PSUs (villages) within each State were selected with probability proportional to size (PPS) for each stratum, where the measure of size was based on the estimated total population.

State had population estimated but figures were missing for some villages, an average measure of size was imputed for these villages; in a way such villages had an equal probability of selection in the frame. In other words, the sampling frame of villages was compiled separately for each State based on the best available sources. When the estimated population was not available, an average measure of size was imputed; in this way, such villages had an equal probability of selection in the frame.

In case of a few States, where the sampling frame did not include population estimated, it was decided to select the sample villages with equal probability. There were four States in South Sudan (Upper Nile, Jonglei, Unity and Lakes) which did not have population measures in the frame. In these four States, the sample villages were selected systematically with equal probability. The same type of sample selection spreadsheet was used for these States, but each village was assigned a measure of size of one. In cases where a selected village could not be found in the field or could not be reached because of security or access problems, a neighbouring village in the sampling frame replaced it. All 40 villages within the sampled segments in each State were fully covered with the exception of only 12 segments in two States in Southern Sudan (7 segments in Upper Nile and 5 in Western Bahr El Ghazal States) that had to be substituted for insecurity,

influencing accessibility during the fieldwork period. Segmenting of large sample villages: Some of the villages in the frame had five hundred (200) or more households. In case of a sample village with a large number of households (for example, greater than 200), the village was subdivided into smaller segments of similar size (with about 80 to 120 households each) with clear defined boundaries in order to facilitate the listing process and avoid coverage problems. Following this, one sample segment was selected by random with equal probability for the listing of households at the second sampling stage.

The List of households in sample villages or segments: A list of the households was undertaken in each sample segment prior to the SHHS data collection in order to enumerate all housing units and households within the boundaries of each sample village or segment. At the last sampling stage, the households were selected systematically with a random start from this household listing for each sample segment. The supervisor was responsible for verifying the boundaries of the sample village or segment in order to ensure good coverage of the sample households.

Selection of sample households within sample village or segment: At the last sampling stage, a sample of 25 households was selected systematically for enumeration with a random start from the household list for each sample village or segment. If a village had less than 25 households, all of them were selected. Once the list was completed, the supervisor referred to the sample selective table to find the row corresponding to the total number of households listed; this row identified the 25 household numbers selected. This table was generated with an Excel spreadsheet.

3.3 Estimation and weighting procedures

For reporting national level results, and to obtain unbiased estimation from the SHHS data, appropriate weights were applied to the sample data based on the probabilities of selection. Measures of sampling variability for key survey estimates were also calculated. Sample in the Sudan Health Survey was not self-weighted. Since the regions varied in size, each region essentially had different sampling fractions allocated equal number of houses. Hence, subsequent analysis of survey data used calculated simple weights for this reason.

3.4 Data analysis

Current research specifically focuses on the data analysis, as the objective is dealing with the missing values in cluster analysis. Two-Step Cluster Analysis is be applied in which each participant is be classified into one of the identified pattern and the optimal number of classes is determined using SPSS Statistics/IBM. However, the risk of over-fitting of the data must be considered because cluster analysis is a multivariable statistical technique.

When there is limited generalisability outside of the available sample, the available data is excessively fit in an analysis and over-fitting occurs. Classification over-fitting can occur because there present an excessive number of ‘noise’ variables, or because the sample size is inadequate relative to the number of variables, or because the participants lack representativeness. Cluster analysis often faces the inadequate consensus about appropriate sample size ratios and considerable debate about over-fitting in statistical classification. However, authors have argued that each independent variable have a minimum of ten events to avoid over-fitting in other forms of multivariable analysis.

Any observation with missing data is excluded in the Cluster Analysis because like multi-variable statistical techniques, Cluster Analysis also does not

tolerate missing values. Therefore, before performing the cluster analysis, missing values will be imputed by using multiple imputations *(SPSS Statistics/IBM)*.

Prior to cluster analysis, log transformation will approximate normality in the data because household data does not follow the strict assumption of Two-Step Cluster Analysis that is the interval data have normal distribution. However, determination of interquartile ranges and median does not require the data to follow normal distribution hence raw data is applicable for obtaining these statistics.

Clustering variables are be assumed independent in latent class analysis, Two-Step cluster analysis, and many other diverse traditional clustering techniques and analysis. The variables that form clusters thus have a low correlation (collinearity) between each other. Conditional correlation (conditional on membership in one or more clusters) and global correlation (between the variables entered into the analysis) are the possible forms of this collinearity. Specific diagnostic techniques for different techniques of cluster analysis are required for conditional correlation while calculation for global correlation is easy. Construction of Pearson correlation matrices is necessary because collinearity is very likely to occur in household health data. Reporting the range, standard deviations, and mean of these correlations will describe the global collinearity in these data.

SPSS Statistics version 19.0.0 (IBM, Chicago IL, USA) will be used for correlations, cluster analysis, and multiple imputations. Excel 2008 for Mac version 12.2.8 (Microsoft Corporation, Redmond, WA, USA) will be used to perform all other analyses.

3.4.1 Two-Step Cluster Analysis

Two-Step Cluster Analysis is chosen over a wide range of approaches of statistical pattern-recognition available for clustering household health data including neural networks, classical cluster analysis, and probabilistic data-mining and latent class analysis. Reasons for choosing Two-Step Cluster Analysis are the shorter learning curve of Two-Step Cluster Analysis than the alternative approaches method of this analysis readily available in the basic version of SPSS base on the probability. However, method selection is also guided by some head-to-head comparisons of these approaches of cluster analysis. The natural groupings (or clusters) that are usually not apparent will be revealed by the design of the exploratory tool and procedure of Two-Step Cluster Analysis. The algorithm employed in current research differentiate from other clustering techniques due to the following several desirable features.

- The continuous and categorical variables are assumed independent and hence it is possible to place a joint multinomial-normal distribution.
- Across different clustering solutions, the values of a model-choice criterion can be compared automatically determine the optimal number of clusters.
- The records can be summarized in a cluster features (CF) tree the Two-Step algorithm construct therefore, the researcher will be able to analyse large data files.
- The procedure will produce descriptive statistics by cluster for the final clustering, cluster frequencies for the final clustering, and information criteria (the Bayesian Information Criterion (BIC) or the Akaike Information Criterion (AIC)) by numbers of clusters in the solution.

- The procedure will produce variable importance charts, pie charts of cluster frequencies, and bar charts of cluster frequencies.
- A probability distribution will be placed on the variables by using the likelihood measure. The procedure assumes all the variables to be independent. A multinomial distribution is be assumed for categorical variables and Normal (Gaussian) distribution is assumed for continuous variables.
- For all the continuous variables, “straight line” distance between two clusters will be measured through the Euclidean measure.
- The procedure provides a summary of the continuous variable standardization specifications.
- Either the Akaike Information Criterion (AIC) or the Bayesian Information Criterion (BIC) will be specified to determine the number of clusters through automatic clustering algorithm.

3.4.2 Assumptions of Data in Two-Step Cluster Analysis

Both categorical and continuous variables can be analysed through this procedure. Clustering is based on attributes that are represented by variables while objects to be clustered are presented by cases. Variables in the cluster model are assumed independent likelihood distance measure. The procedure also assumes that each categorical variable follows a multinomial distribution while each continuous variable follows a normal distribution known as Gaussian distribution. The empirical internal testing indicates fair robustness of the procedure in case of violation of both the distributional assumption and the assumption of independence but the researcher must be well aware whether these assumptions are met or not. Standardized continuous variables are applicable for clustering algorithm. SPSS

Statistics/IBM provides the option of “To be Standardized” for those continuous variables that are not standardized.

3.4.3 Two-Step Cluster Analysis Plots

Charts showing the within-cluster variation of each variable will be obtained. For each categorical variable, a clustered bar chart will be produced, showing the category frequency by cluster ID. For each continuous variable, an error bar chart is produced, showing error bars by cluster ID. A pie chart showing the percentage and counts of observations within each cluster will be obtained.

Several different charts showing the importance of each variable within each cluster will be obtained. The output will be sorted by the importance rank of each variable.

Measure of variable importance to plot will use Chi-square or t-test of significance. Chi-Square will report a Pearson chi-square statistic as the importance of a categorical variable and a *t* statistic will report the importance of a continuous variable. Significance reports one minus the *p* value for the test of equality of means for a continuous variable and the expected frequency with the overall data set for a categorical variable.

The confidence level for the test of equality of a variable's distribution within a cluster versus the variable's overall distribution will be set. The value of the confidence level will be shown as a vertical line in the variable importance plots, if the plots are created by variable or if the significance measure is plotted. Variables those are not significant at the specified confidence level will not be displayed in the variable importance plots.

3.4.4 Two-Step Cluster Analysis Output

The clustering results will be displayed in tables. The descriptive statistics and cluster frequencies will be produced for the final cluster model, while the information criterion table will display results for a range of cluster solutions.

Two tables will describe the variables in each cluster. In one table, descriptive statistics will be reported for continuous variables by cluster including mean and standard deviation. The other table will report frequencies of categorical variables by cluster. A table will report the number of observations in each cluster. A table will contain the values of the AIC or BIC, depending on the criterion chosen in the main dialog box, for different numbers of clusters.