Chapter One : The Introduction

1.1 preface

It could be very tedious for any researcher to make an analysis of incomplete data. In any research, data plays a significant role in every aspect. When the researchers had performed household health survey in Sudan, they faced many problems because the people were not interesting to provide their health data. epidemic diseases had scattered in Sudan, which let many people had been died who affected by diseases. for this, the household health survey program was commenced for inspecting the affected people and cause of this epidemic disease (Henderson & Sundaresan, 1982).

1.2 The research problem

The research of household health data in Sudan, the researchers were facing certain main problems, because peoples were not participating in data collect. However, the situation very perilous for the investigator during the survey, but they had to complete their study about the epidemic diseases that commonly spread in Sudan. The data was missing, such as demographic data, which is mandatory for the researcher, and the data related to the factors of the epidemics diseases. The incomplete data was insignificant for research analysis, so they left a negative effect on the data treatment methods.

Since different missing data patterns may require different imputation methods, we studied the missing pattern of the datasets before selecting an appropriate imputation method. As first introduced by Little and Rubin (1987), there are three major patterns: 1) missing completely at random (MCAR); 2) missing at random (MAR); and 3) nonignorable missing data (MNAR). MCAR occurs when the

missing values on variable Y are independent of all other observed variables and the values of Y itself, is very strong assumption and can be impractical for real-life data (Muthén et al. (1987)). MAR assumes that the probability that an observation is missing on variable Y depends on other observed variables but not on the values of Y itself, which is more plausible than MCAR. Under MNAR a missing value no longer occurs “at random” since the probability that an observation is missing on variable Y depends on other unobserved variables.

1.2.1 Statement of the Problem

Effect of missing data treatment methods on cluster analysis performed on Sudan Household Health survey data. This is the statement of the research proposal. The main theme is to identify the effects of the missing data on cluster analysis.

1.3 Methodology

Statistical methods implement in this survey, which was accurate and authentic to find out the factors of diseases in Sudan. The data was completely collected due to various circumstances or respondent’s interest, which effect bit in analyzing the whole research.

1.4 Sudan Household Health Survey (SHHS)

The Sudan household health survey was conducted in the year 2006. The household survey had been executed by Central Bureau of Statistics of Sudan and Federal Health Ministry there arethe two other branches. This survey was scientifically and monetary supported by UNICEF and Pan Arab health project organization. The Sudan household Health Survey ,2006 accumulated the hard work of all the health related agencies to perform a unanimous survey that has been fulfill the interests of all stakeholders as it was a mixture of analyzing the

multiple factors that were being announcement in Sudan such as, food safety, medicines and other health related factors. The strategy and the execution of SHHS such as technical, working group, steering and coordination body. These are all the basic and strong structures, which the organization has must followed, and work according to it.

The Sudan household health survey provides a quality work under this organization and the precious data on the condition of women and children data, which are mostly affected to this epidemic disease. It was started with the targeted objective, to complete all the related information or statistics about the Sudan people affected data. It could be beneficial for the agencies, which were prevailing under the health issues (Rose, et.al, 2006).

1.4.1 Data Sources

The data had been collected in SHHS were used to collect the data related to women and children were the analyzing key pointers that allow the country to observe development towards the organization objectives such as MDGs , which had a main focus on realizing the child diseases and other worldwide concurred upon obligations.

1.5 Objective of the study

The main objective of Sudan the study were as follows:

- To provide every data for helping the condition of women and children in Sudan.
- To provide data for monitoring progress for the achiementof MDGs (Millennium Development Goals).

- To donate the development of information and observing methods in Sudan and to reinforce technological proficiency in the plan, completion, and investigation of such schemes.
- To reinforce and construct the institutional capability of government associates for the imminent 2007 market research and large level assessments.
- To compensate for differential probabilities of selection among subgroups (age-sex-race/ethnicity subdomains; persons living in different geographic strata sampled at different rates);
- To reduce biases arising from the fact that nonrespondents may be

Different from those who participate;

- To bring sample data up to the dimensions of the target population totals;
- To compensate, to the extent possible, for inadequacies in the sampling

Frame (resulting from omissions of some housing units in the listing of

Area segments, omissions of persons with no fixed address, etc.); and

- To reduce variances in the estimation procedure by using auxiliary Information that is known with a high degree of accuracy.
- To determine whether the additional information generated by imputation improves the predictive modeling results.

The objective of this research proposal is to identify the effect of the missing data on statistical technique cluster analysis by using the source of household health survey data that conducted in 2006 for analyzing the generated factors due to which the epidemic had enhanced in women and child of Sudan (Kim, et.al, 2007).

1.6 Importance of the study

The importance of this research is to examine the factors that involved in this epidemic disease that affected both children and women of Sudan country. To accomplish these objectives, the organization had to follow the strategies in systematic approach. All health related agencies, which were collaborating in this health survey that supported both technically and financially had interested to scrutinize the factors by using the cluster analysis. The main importance of this survey had to inspect the issues and problems that faced by the women and children of Sudan in epidemics diseases.

1.7 Questionnaires

Sudan Household Health survey used five sets of questionnaire. The food security questionnairewhich will not be analysed in here. Community leaders were addressed by the Community Questionnaire. The caretakers or mothers of all the children under 5 years of age were addressed by Under-five Questionnaire in each household. 15-19 years age women in each household were adreesed by Women questionnaire. All the households and all de jure members of households were addressed to collect information through Household Questionnaire.

PAPFAM and MIC3 models of questionnaires were followed in making first three questionnaires. Translation and wording of questionnaires were modified after pre-testing the questionnaires. Questionnaires included following modules.

Mothers of children age under-five were addressed by the Under-five Children Questionnaire. A primary caretaker of under-five children were interviewed in such cases where the household list did not include the mother. Modules in the questionnaire for children under five included Anthropometry, Malaria, Immunisation, Care of Illness, Breastfeeding, Vitamin A, and Birth Registration. Modules in the questionnaire for individual women included HIV

knowledge, Contraception, Marriage and Union, Maternal and Newborn Health, Tetanus Toxoid, Child Birth History, and Child Mortality. Modules in household questionnaire included Maternal Mortality, Salt Iodization, Malaria, Household income and resources, Household characteristics, Water and Sanitation, Education, and Household list. Other than administering the questionnaires, fieldwork teams additionally measured heights and weights of children under age of five years and be tested the iodine content in the salt used for cooking purpose in the households.

1.7.1 Questionnaires Sample

Table 1.1 Demographic Data	
Variables	**Category**
Child gender	Female/Male
Child age	1-10 years
Child education	Depends if they attend school or not because of poverty and hunger issues
Women age	Varies from 20-38 years
Women education	Depend if they every attend school or not. If they attend school, so at what level she continues her education.
Area	Urban/Rural
Household ducation	Overall family education
Household wealth	Head of the family earning
	Diseases Data
Type of Diseases	It varies from person to person.
Factors of Diseases	Main issues face by the child/women during disease
Treatment	If ever treat or consult any doctor
Duration	From how long the child/women encounter from this disease

1.8 Cluster Analysis

In research, cluster analysis provides significant information about the demographic of data. It is a task for analyzing the whole population by making clusters according to the total of the population.

It defines testing data withdrawal, analysis of image and bioinformatics. It is not an automatic process, but a repeating method of analyzing the different factors and related to information that are related to the task. In Sudan, when the researchers partially analyzed the population, because the data was incomplete and beyond the expectation of the researchers who collected household health survey data of Sudan. The health researcher made a cluster of different samples and analyzed according to it (Ngondi, et.al, 2007).

However, in practice, it may not always be possible to cluster huge datasets by using clustering algorithms successfully, for weakness of most existing automated clustering algorithms on dealing with arbitrarily shaped data distribution of the datasets.

As Abul et al pointed out "In high dimensional space, traditional clustering algorithms tend to break down in terms of efficiency as well as accuracy because data do not cluster well anymore" (Abul, et al , 2003). In addition, the very high computational cost of statistics-based cluster validation methods directly affects the efficiency of cluster validation (Huang, 2001).

1.9 Scope and Limitations

In the Sudan Household Health Survey (SHHS), after obtaining the outcomes from different clusters samples, the researchers scrutinized all the clusters of household's health surveys.

The problem faced by the researchers because of missing data and incomplete information. most data had not available from in South Sudan because of controversial issues. The respondents did not take any interest to provide their demographic diseases complete information for lack of awareness.

The researchers adopted the confined analysis through which they applied to the results of the clusters samples from the overall selected or targeted population. Initially, they faced turmoil situations because of missing factors in the data, but they analyzed by using the statistical tool that is cluster analysis. Through all the possible data, which they obtained from different parts of the Sudan, they comprehend that data related to child and women ratio in epidemic diseases.

The possibility of the prevailing factors had enhanced, if the preventive action had not adopted. The questionnaires helped the researcher to aware from the factors through which the diseases had increased in both women and children.

This survey was partially beneficial because of the missing data, but the researchers had completed their imaginary analysis by taking the whole population. The cluster sampling helped them for analyzing the effects that facing by the victims through this epidemic diseases.

Through this survey, the health committee had successful for founding the factors of diseases. Now they can take any precautions or preventive action for reducing this epidemic in Sudanese people (Habbani, 2006).

It is not an easy task to reduce the epidemic diseases in Sudan, but the health organizations had to take any certain measures steps for reducing the effects of this disease. Certain limitations, which can observe through this survey such as, the organizations had to implement more cost to stop this disease.

It includes cost, strong strategy that implement by all the organizations or the workers. These are the main limitations, which can analyze after all this health

survey in Sudan. To cover the entire Sudan is not an easy; the results are somehow hypothetical because of the missing data, which affects the entire analysis.

The health organizations had taken any possible measures steps for reducing the level of disease in the entire Sudan, however it needs more cost because the country had faced the situation of poverty and hunger for which this epidemic diseases easily grasped them (Geltman,et.al,2005).

1.10 Previous Studies

Under this household health survey in Sudan, the health agencies, which were, pool resources this basic and prominent mission performed a significant survey for collecting the health disease data in women and children. it was the basic agenda of the health agencies; besides that, the research team also set their new target to find the factors that enhance the diseases factors in the women and child.

This had achieved by using the statistical sampling technique such as cluster analysis (Obermeyer, et.al, 2008).

1.11 Organization of The Study

The rest of the thesis is organized as follows: An overview of Data Collection of Household Health Surveys, Suggestions for Analyzing Survey Data and missing data treatment methods is presented in Chapter 2. Methodology of the Sudan Household Health Survey (SHHS) and the statistical procedure to be used in current research is presented in Chapter 3. Chapter 4 discusses in detail the methodology of clustering the pre-processed data, focusing on cluster analysis,

cluster validation, and consensus clustering. Experimental results and analysis are provided in Chapter 5.