International Journal of Basic and Applied Sciences, 5 (1) (2016) 15-29

International Journal of Basic and Applied Sciences

Website: www.sciencepubco.com/index.php/IJBAS
doi: 10.14419/ijbas.v5i1.5470
Research paper

Clustering and multiple imputation of missing data

Elsiddig Elsadig Mohamed Koko [1]*, Amin Ibrahim Adam Mohamed [2]

[1] *Sudan University of Science & Technology, Faculty of science, Department of Statistics*
[2] *Omdurman Islamic University , Faculty of Economic studies, Department of Statistics*
**Corresponding author E-mail: siddiggt@gmail.com*

Abstract

The present work specifically focuses on the data analysis as the objective is to deal with the missing values in cluster analysis. Two-Step Cluster Analysis is applied in which each participant is classified into one of the identified pattern and the optimal number of classes is determined using SPSS Statistics/IBM. Any observation with missing data is excluded in the Cluster Analysis because like multi-variable statistical techniques. Therefore, before performing the cluster analysis, missing values will be imputed using multiple imputations (SPSS Statistics/IBM). The clustering results will be displayed in tables. Furthermore, goal of analysis is to reduce biases arising from the fact that non-respondents may be different from those who participate and to bring sample data up to the dimensions of the target population totals.

Keywords: *Cluster Analysis; Missing Data; Multiple Imputation; Two-Step Cluster Analysis.*

1. Introduction

During the research of household health data in Sudan, the researchers had to face certain measures problems, because the peoples were not participating in the collection of data. Though, the situation was 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 the analysis of research, so they left a negative effect on the data treatment methods.

1.2. Cluster analysis

In research, cluster analysis provides significant information about the demographic of data. It is a task of analyzing the whole population by making clusters according to the total of the population. It defines examining data withdrawal, analysis of image and bioinformatics. It is not an automatic process, but a repeating method of analyzing the different factors and related 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 [1].

Following three definitions of cluster are found to be best described in the literature reviewed [2]; [3]. Regions of low density of points are separated from regions of relatively high density of points of a multidimensional space and such connected regions are called clusters.

Cluster is such an aggregation of points where distance between the points included in a cluster is less than the distance between the points within the cluster and the points outside the cluster.

Cluster forms set alike entities in a way that the entities across clusters are unlike, Certain properties like separation, shape, dimension, variance, and density are used to compare the clusters even though the cluster is an application dependent concept [3]. In comparison with other areas of space, a region of compact and tight high-density points is called cluster. Small degree of variance or dispersion is meant by the tightness and compactness. The shape of cluster is a priori known, determined by the clustering criteria, and used algorithm. Distance between clusters and the degree of possible cluster overlap is defined through separation.

1.2. Main elements of cluster analysis

The successful completion of tasks presumes a large number of correct choices and decisions from several alternatives despite of the simple idea behind cluster analysis. Before attaining the results, at least nine major elements appear in cluster analysis [4]. The list with strategy of missing data and data presentation, because the current real-world data set contains missing values as well, includes Interpretation of results. Number of clusters, Computer and algorithms implementation (and their reliability, e.g., convergence), Choice of missing data strategy, Choice of clustering criterion (objective function), Choice of (dis)similarity measures, Normalization of variables, what to cluster variables or data units, Choice of variables, Choice of objects, and Data presentation [5], [6]. According to [7], importance of strategies used in cluster validity, normalization, data representation, and data collection is same as that of the cluster strategy itself. Importance of choice of the best (dis)similarity measure is greater than the importance of choice of clustering algorithms [9]. Result interpretation and estimation of the number of clusters are closely related to the validity of resulting cluster solution e.g. a kind of validation technique, visual exploration of the obtained solution [2]; [3].

1.3. Missing data in cluster analysis

In evaluation of different imputation methods on biological data, [10] clearly stated:

Copyright © 2016 Elsiddig Mohamed Koko, Amin Ibrahim Mohamed. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

"However, it is important to exercise caution when drawing critical biological conclusions from data that is partially imputed. [. . .] [E] Stimated data should be flagged where possible [. . .] to avoid drawing unwarranted conclusions." There is no mechanism to indicate the less reliability of imputed values still data imputation is common despite of this warning. [10] Described other sophisticated approaches to handle missing values including inferring the missing values' feature based on observed features and similarities between the missing observation and known observation in the data set; modelling the selected and observed values according to the true distribution; and replacing all the missing values with the actual mean value. [11] Processed data with missing value and presented modification in EM algorithm. Values for the missing features, data cluster assignments, and maximum likelihood parameters are simultaneously estimated in this model. Due to the lack of full reliability, each of these approaches suffers from a disability of discounting imputed values. Marginalisation is also sometimes considered as a better solution because no new data values are created. Supervised methods such as Hidden Markov Models [12] or neural [13] networks are focused in most of the previous work in marginalisation. [14] Proposed a set of hard constraints that were guaranteed satisfied by the output produced by a variant of k-means. Hard constraints show whether the group of certain items should or should not be formed while soft constraints shows the strength of grouping. Both the soft and hard constraints satisfy in the clustering of data with missing values. Later the clustering application became a core method of knowledge discovery and data mining due to the summarizing, descriptive, and unsupervised nature of data clustering. New clustering algorithms developed due to the increasing number of large multidimensional data collections especially during the last decade [15]; [16]; [17].

One of the key assumptions in cluster analysis is the unknown structure of target data set as emphasized in the first definition. This assumption of clustering (unsupervised classification) majorly differs it from classification (supervised classification). The object collections with unknown class labels are focused in cluster analysis unlike classification where a priori knowledge of category structures is available. Process of cluster formation is unaffected of information about the data sources such as class labels, which influence the interpretation of results [8]. During the configuration of correct number of clusters or initial parameters, the understanding of domain is often of great use. Multi-dimensionality of the data objects (records, observations etc.) is stressed in the second and third definitions. The difficulty that a human being faces in grouping of objects that possess three or more variables without automated methods emphasizes the importance of previous notion. The notion of similarity is addressed in most of the aforementioned definitions naturally. Choosing an appropriate measure of similarity is one of the most influential tasks of cluster analysis, as similarity is one of the key issues of cluster analysis. Problem of selecting a measure of similarity depends on the data. [4] used the degree of "natural association" instead of talking about "similarity". Cluster analysis, based on aforementioned definitions, can be described as analysing a multidimensional data set with an unknown structure and choosing a measure of similarity to determine a (small) number of meaningful variables or objects. Here, meaningful refer to the description given by [17].

2. Methods

2.1. Suggestions for analysing survey data

Linear statistics can be used to calculate unbiased linear estimates of population that can provide the inverse of probability of selection for each observational unit and design based weights for the theoretical case of [18] surveys from all sample members with complete response). The observations of non-response that are dropped from the analysis without taking any other action lead to biased estimates of household surveys in practice. The biasness due to non-response is now been reducing using the continually developing techniques. Among the earliest techniques proposed, one simplest one was given by [19] that reduce the difference between the parameters of population for non-respondents and respondents to a negligible value through partitioning the sample into weighting classes. [21] Evolved the calibration method of [20] for weigh adjustment of post-stratification, for non-response, or for both. Calibration methods simultaneously control the weighted sample distribution in several dimensions.

The non-linear estimate is consistent in the trivial sense that is they would exactly equal the value of comparable finite population if the size of sample was increased to the finite size of population but the non-linear estimates are not unbiased for small samples [22]. The sample size can be allowed to increase without limit if it is allowed to consider the population of finite size as arising from a hypothetical population of infinite size. In this case, as the sample size increases, probability of the non-linear estimate converges to the parameter of super-population and thus the consistency of the model can be claimed [23].

2.2. Missing data treatment

A general introduction of incompleteness on only one variable is provided by [38] and [39] and through the application on cancer data set, recent developments were reviewed in a comprehensive fashion by [27]. A comprehensive reading list in form of annotated bibliography is provided online [28].

Here the literature review of the analysis of a data set with missing values is done through updating the prior review of [29]. The discussion will cover the compromises, approaches, assumptions of modelling within current implications. The general focus is on the methods that are used to resolve the issues in analysis that arise when some observations are missing from the data set and to deal with the complications arising in cluster analysis [30]; Robins, [31]; [32].

2.2.1. Multiple imputations

[19] Describes the reasons for using a three-step approach, multiple imputations, in estimation of models with incomplete data. First reason is the uncertainty about the non-response model reflected by the creation of plausible values for missing values. Missing observations are then imputed or filled out by these plausible values. A number of completed data set is created through this process repeatedly. Second reason is the availability of complete data methods for analysing the data sets. The last but not least reason is the handling of uncertainty regarding the imputation allowing by combined results.

[33] found accountable limiting of the imputed values (e.g. plausible value of years of smoking for non-smokers is only zero). Similarly problem of limiting arise when variables require transformations, or when there is a certain range defined for missing values (e.g. a five point Liker scale may have values between three and four). Calculating the standard errors of the maximum likelihood estimation is another complication. SPSS, the S-Plus missing data library, and LogXact version 7 address these complications and provide the implementations of maximum likelihood [34].

2.2.2. Chained equations

Chain equations are used in an alternative variable-by-variable approach [35]; [33]; [36]. Other variables are involved as predictors in the separate specification of each variable in this imputation model. An imputation is generated for the missing variable at each stage of the algorithm then the next variable is imputed using the previous this imputed value. The process reaches convergence at last after the repetition of the Gibbs sampling procedure to impute the missing values. Multiple imputations are generated using separate chains. Predictive matching (where the value from one the nearest set of observed value in the data set is taken by the imputed variables) or a linear regression model is involved in the model for continuous variables. For categorical variables, poly-

tomous models are needed and logistic regression can be fit for dichotomous variables. AregImpute (for R and S-Plus), IVEware (for SAS or standalone), ICE (for Stata), or MICE library (for R and S-Plus) can provide the implementations of the chained equation approach.
[33] Describes the problem with the approach of chained equation approach as its inability to converge to a sensible stationary distribution where multivariate distributions and separate variables are not compatible though [35] obtained reasonable imputations in a series of studies on simulation even with incompatible separate models. Further establishment of the validity of this approach needs additional work.

2.2.3. Methods for monotone data sets

SAS PROC MI handles data sets with monotone missing structure implementing a number of approaches. A value randomly from such a set of observed values whose values are closest to the predicted value can be imputed using the method of predictive mean matching. Imputation of a categorical variable with more than two levels complicates this method while the method remains straightforward in imputation of a continuous random variable. The observations with different numbers of missing values are processed in this approach in an ascending order of the number of missing values. In application to missing predictor models, biased results were found with the predictive mean matching approaches, which warned the analysts from using this approach [37]. Missing values can be imputed in the similar way using propensity score or regression models.

2.3. 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 using the likelihood measure. The procedure assumes all the variables to be independent. A multinomial distribution is assumed for the categorical variables and Normal (Gaussian) distribution is assumed for the continuous variables.
- 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.

2.3.1. 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 the clustering algorithm. SPSS Statistics/IBM provides the option of "To be Standardized" for those continuous variables that are not standardized.

2.3.2. Data analysis

Clustering variables are assumed independent in 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. Results & discussion

3.1. Knowledge of means of HIV/AIDS of women

Table1 present the percentage knowledge of HIV/AIDS women who Ever heard of HIV or AIDS is 71.6%, only about one-half of women (51.3 per cent) knew that AIDS transmitted from mother to child through breastmilk. 84.0% of women knew Can AIDS be avoided ,about 54.4% believed that a Healthy-looking person to have AIDS , 66.0% of women knew that AIDS from mother to child during pregnancy, and nly 57.3 % of women knew that AIDS from mother to child at delivery.

Table 1: Knowledge of HIV/AIDS Percentage of Woman Year of Birth (1951-1991)

		Ever heard of HIV or AIDS				Total
		Yes	No	DK	Missing	
Year of birth of woman	Count	5303224	2052597	227	48775	7404823
	% of Total	71.6%	27.7%	.0%	.7%	100.0%
		Can AIDS be avoided?				Total
		Yes	No	DK	Missing	
	Count	4497187	171670	620316	63057	5352230
	% of Total	84.0%	3.2%	11.6%	1.2%	100.0%
		Healthy-looking person to have AIDS				Total
		Yes	No	DK	Missing	
	Count	2913588	1353238	988175	97223	5352224
	% of Total	54.4%	25.3%	18.5%	1.8%	100.0%
	Count	AIDS from mother to child during pregnancy				Total
		Yes	No	DK	Missing	
		3530084	653207	1065719	103221	5352231
	% of Total	66.0%	12.2%	19.9%	1.9%	100.0%
		AIDS from mother to child at delivery				Total
		Yes	No	DK	Missing	
	Count	3067909	926956	1220824	136540	5352229
	% of Total	57.3%	17.3%	22.8%	2.6%	100.0%
		AIDS from mother to child through breastmilk				Total
		Yes	No	DK	Missing	
	Count	2746182	1189134	1306943	109964	5352223
	% of Total	51.3%	22.2%	24.4%	2.1%	100.0%

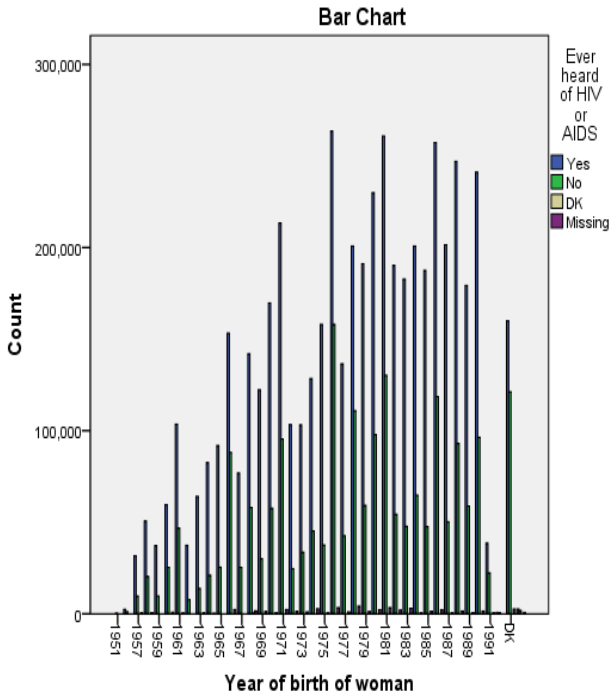


Fig. 1: Ever Heard of HIV or AIDS.

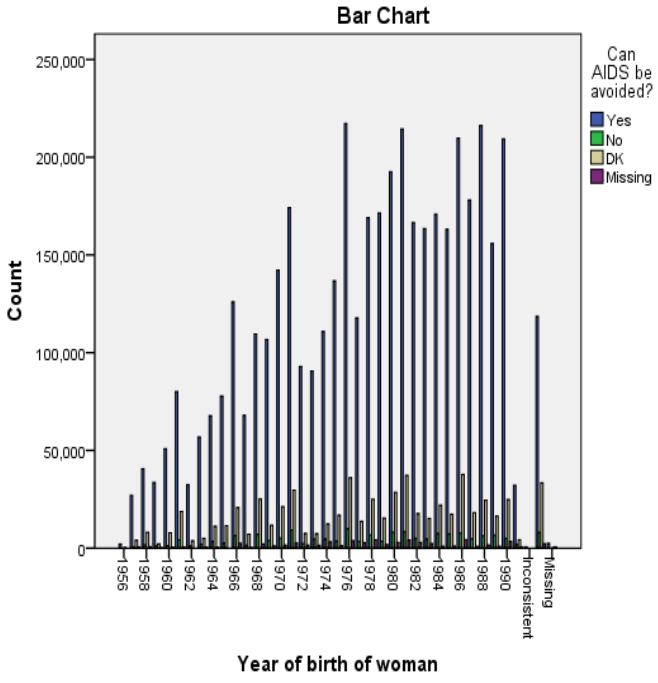


Fig. 2: Can AIDS be Avoid?

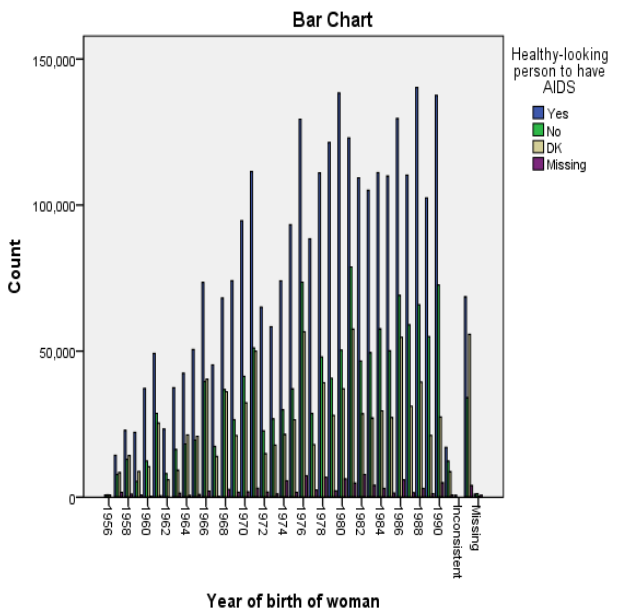


Fig. 3: Ever Heard of HIV or AIDS.

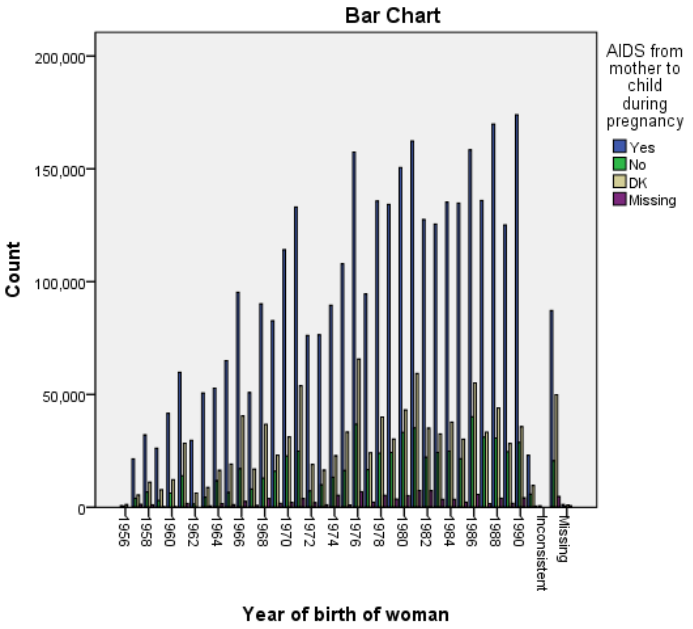


Fig. 4: Can AIDS be Avoid?

The chart in Fig.1, Fig.2, Fig.3, Fig.4 and Fig.5 shows the percentage of HIV/AIDS women year.

Case processing summary of missing 14.5% for year of birth women and 38.2% all other cases of knowledge Hiv/AIDS of women, all cases indicate Number of valid cases is different from the total count in the crosstabulation table because the cell counts have been rounded except year of birth women.

Table 2: Case Processing Summary

	Cases Valid N	Percent	Missing N	Percent	Total N	Percent
Year of birth of woman * Ever heard of HIV or AIDS	7404823	85.5%	1253566.825	14.5%	8658389.825	100.0%
Year of birth of woman * Can AIDS be avoided?	5352230[a]	61.8%	3306159.825	38.2%	8658389.825	100.0%
Year of birth of woman * Healthy-looking person to have AIDS	5352224[a]	61.8%	3306165.825	38.2%	8658389.825	100.0%
Year of birth of woman * AIDS from mother to child during pregnancy	5352231[a]	61.8%	3306158.825	38.2%	8658389.825	100.0%
Year of birth of woman * AIDS from mother to child at delivery	5352229[a]	61.8%	3306160.825	38.2%	8658389.825	100.0%
Year of birth of woman * AIDS from mother to child through breastmilk	5352223[a]	61.8%	3306166.825	38.2%	8658389.825	100.0%

a. Number of valid cases is different from the total count in the crosstabulation table because the cell counts have been rounded.

3.2. Multiple imputations

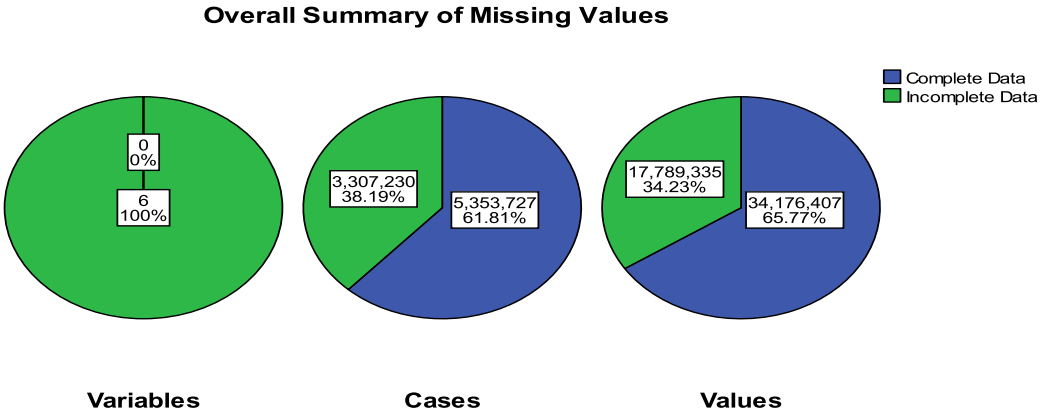


Fig. 5: Summary Missing Values.

Fig.5 shows that:

- The Variables chart shows that each of the six analysis variables has at least one missing value on a case.
- The Cases chart shows that 3,307,230 of the 10,000,000 cases have at least one missing value on a variable.
- The Values chart shows that 17,789,335 of the 50,000,000 values (cases × variables) are missing.
- There are 5353727 (61.81 %) complete cases and 65.77% complete values.

Table 3: Variable Summary

	Missing N	Percent	Valid N	Mean	Std. Deviation
AIDS from mother to child through breastmilk	3307230	38.2%	5353727		
AIDS from mother to child at delivery	3307230	38.2%	5353727		
AIDS from mother to child during pregnancy	3307230	38.2%	5353727		
Healthy-looking person to have AIDS	3307230	38.2%	5353727		
Can AIDS be avoided?	3307230	38.2%	5353727		
Year of birth of woman	1253185	14.5%	7407772	2291.58	1555.874

The variable summary is displayed for variables with at least 10% missing values, and shows the number and percent of missing values for each variable in the table. It also displays the mean and standard deviation for the valid values of scale variables, and the number of valid values for all variables. AIDS from mother to child through breastmilk , AIDS from mother to child at delivery , The patterns chart Fig.6 displays missing value patterns for the analysis variables. Each

Pattern corresponds to a group of cases with the same pattern of incomplete and

Complete data. Pattern 1 represents cases, which have no missing values,

AIDS from mother to child during pregnancy, Healthy-looking person to have AIDS, and Can AIDS be avoided?, have the most missing values, in that order.

The descriptive statistics Table 3 14.5% for (Year of birth of woman) shows means and standard deviations in each set of imputed values, 38.2% for all other variables.

while Pattern 2 represents cases that have missing values on HA9B,HA8C, HA9A,HAB and HA3_x and Pattern 3 represents cases which have missing values on all variables.

This dataset is nonmonotone and there are many values that would need to be imputed in order to achieve monotonicity.

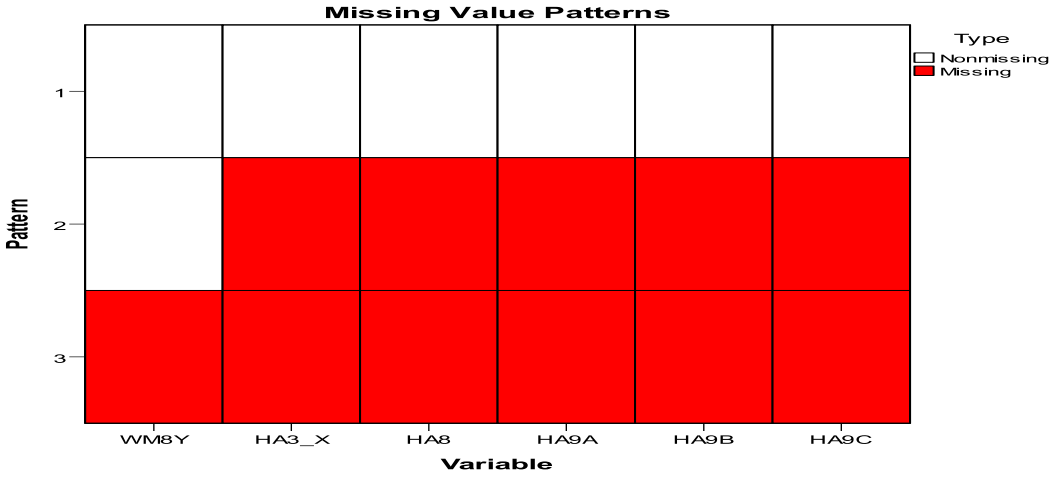


Fig. 6: Missing Value Patterns.

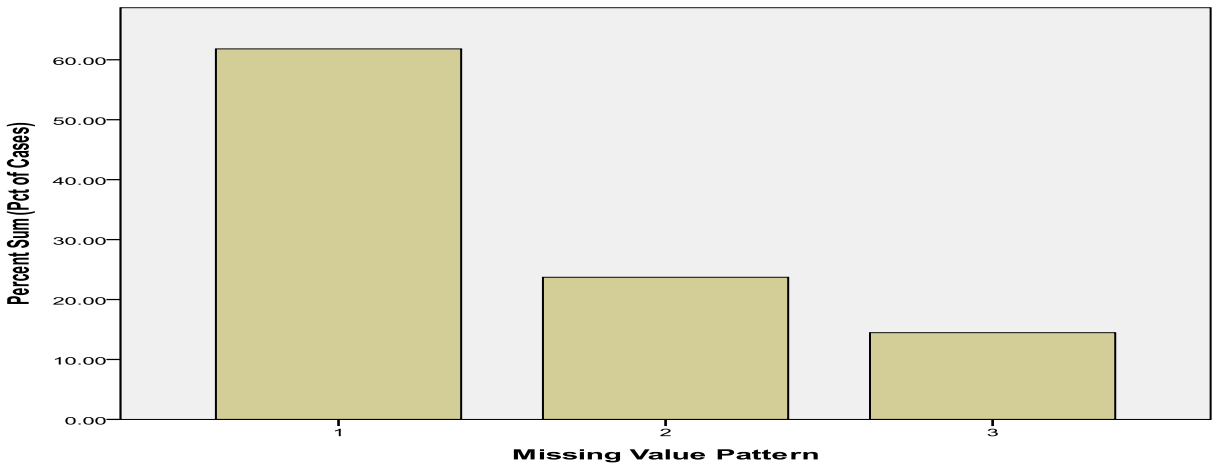


Fig. 7: Missing Value Pattern.

The bar chart in Fig.7 displays the percentage of cases for each pattern. This shows that over half of the cases in the dataset have Pattern 1, and the missing value patterns chart shows that this is the pattern for cases with no missing values. Pattern 2 represents cases with a missing value on, HA8C, HA9A, HAB and HA3_x, Pattern

Three represents cases with a missing value on, HA8C, HA9A, HAB, HA3_x and WM8Y.

Table 4: Imputation Specifications

Imputation Method	Fully Conditional Specification
Number of Imputations	5
Model for Scale Variables	Linear Regression
Interactions Included in Models	(none)
Maximum Percentage of Missing Values	100.0%
Maximum Number of Parameters in Imputation Model	100
Replication Weight Variable	hhweight

The imputation specifications in table4 is a useful review to confirm that the specifications were correct , Imputation Method is Fully Conditional Specification , Number of Imputations is 5 and Model for Scale Variables is Linear Regression , table 5 display that Imputation Results of Fully Conditional Specification Method Iterations is 10 and Dependent Variables Imputed HA3_X,HA8,HA9A,HA9B,HA9C.

Table 5: Imputation Results

Imputation Method		Fully Conditional Specification
Fully Conditional Specification Method Iterations		10
Dependent Variables	Imputed	HA3_X,HA8,HA9A,HA9B,HA9C
	Not Imputed(Too Many Missing Values)	
	Not Imputed(No Missing Values)	WM8Y
Imputation Sequence		WM8Y,HA3_X,HA8,HA9A,HA9B,HA9C

Table 6: Imputation Models

	Model			
	Type	Effects	Missing Values	Imputed Values
Can AIDS be avoided?	Logistic Regression	HA8,HA9A,HA9B,HA9C,WM8Y	2054045	10270225
Healthy-looking person to have AIDS	Logistic Regression	HA3_X,HA9A,HA9B,HA9C,WM8Y	2054045	10270225
AIDS from mother to child during pregnancy	Logistic Regression	HA3_X,HA8,HA9B,HA9C, WM8Y	2054045	10270225
AIDS from mother to child at delivery	Logistic Regression	HA3_X,HA8,HA9A,HA9C, WM8Y	2054045	10270225
AIDS from mother to child through breastmilk	Logistic Regression	HA3_X,HA8,HA9A,HA9B, WM8Y	2054045	10270225

The imputation models in table6 gives details about variable was Imputed. Note in particular that:

- All categorical variables modeled with a logistic regression.
- Each model uses all other variables as main effects.
- The number of missing values for each variable is reported, along with the total number of values imputed for that variable (number missing × number of imputations) for example (2054045×5=10270225).

3.3. Descriptive Statistics knowledge HIV/AIDS

Table 7: HA3_X (Can AIDS Be Avoided?)

Data	Imputation	Category	N	Percent
Original Data		1	4498332	84.0
		2	171720	3.2
		8	620571	11.6
		9	63104	1.2
Imputed Values	1	1	1093635	53.2
		2	71756	3.5
		8	432127	21.0
		9	456527	22.2
	2	1	971065	47.3
		2	46581	2.3
		8	411349	20.0
		9	625050	30.4
	3	1	991824	48.3
		2	48031	2.3
		8	428699	20.9
		9	585491	28.5
	4	1	987850	48.1
		2	59633	2.9
		8	412233	20.1
		9	594329	28.9
	5	1	991901	48.3
		2	47845	2.3
		8	422065	20.5
		9	592234	28.8

The table7 for HA3_X (Can AIDS be avoided?) now has an imputation (5) whose distribution is more in line
with the original data, but the majority are still showing a greater proportion of the cases estimated as being avoided than in the original data. This could be due to random variation, but might require further study of the data to determine whether these values are not missing at random (MAR). We will not pursue this further here.

Table 8: HA9A (AIDS From Mother to Child during Pregnancy)

Data	Imputation	Category	N	Percent
Original Data		1	3530910	66.0
		2	653426	12.2
		8	1066113	19.9
		9	103278	1.9
Imputed Values	1	1	457387	22.3
		2	216996	10.6
		8	757612	36.9
		9	622050	30.3
	2	1	471058	22.9
		2	207205	10.1
		8	732895	35.7
		9	642887	31.3
	3	1	472529	23.0
		2	218044	10.6
		8	761608	37.1
		9	601864	29.3
	4	1	473524	23.1
		2	210061	10.2
		8	761403	37.1
		9	609057	29.7
	5	1	469251	22.8
		2	208048	10.1
		8	755345	36.8
		9	621401	30.3

The table8 for HA9A(AIDS from mother to child during pregnancy) has an interesting result in that, for the imputed values, a greater proportion of the cases are estimated as being AIDS during pregnancy than in the original data. This could be due to random variation; alternatively, the chance of being missing may be related to value of this variable.

Table 9: HA9B (AIDS From Mother to Child at Delivery)

Data	Imputation	Category	N	Percent
Original Data		1	3068673	57.3
		2	927199	17.3
		8	1221241	22.8
		9	136614	2.6
Imputed Values	1	1	409625	19.9
		2	247480	12.0
		8	788967	38.4
		9	607973	29.6
	2	1	414009	20.2
		2	848814	41.3
		8	751484	36.6
		9	39738	1.9
	3	1	411984	20.1
		2	833175	40.6
		8	776338	37.8
		9	32548	1.6
	4	1	408847	19.9
		2	829529	40.4
		8	781710	38.1
		9	33959	1.7
	5	1	872933	42.5
		2	245323	11.9
		8	783093	38.1
		9	152696	7.4

3.4. Checking FCS convergence

When using the conditional specification method, it is a good idea to check plots of the means and standard deviations by iteration and imputation for each scale dependent variable for which values are imputed in order to help assess model convergence.

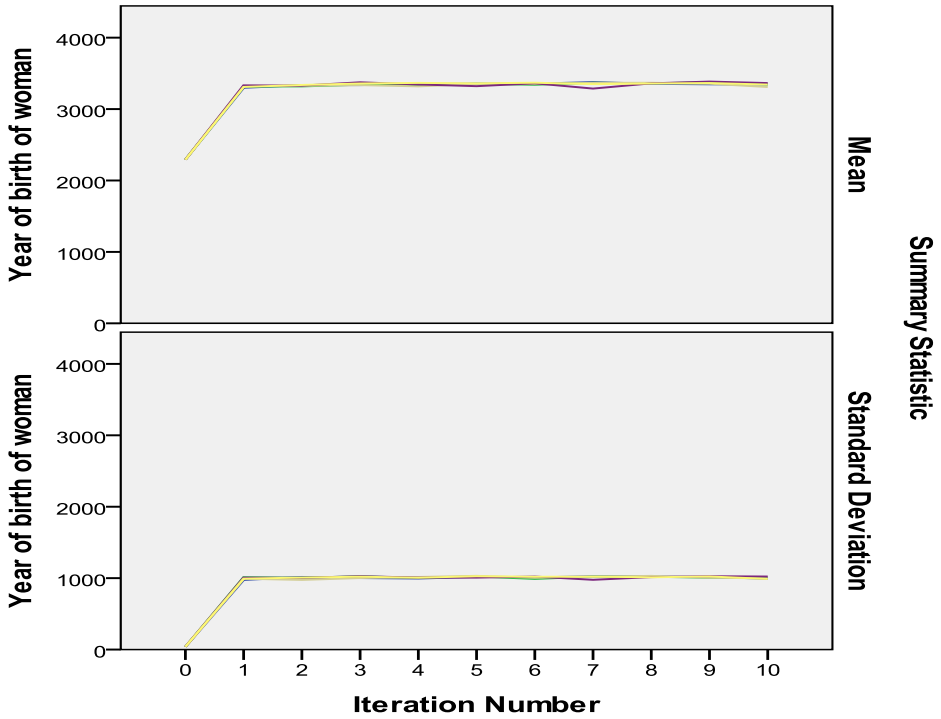


Fig. 8: FCS Iteration Number.

You have created a pair of multiple line charts, showing the mean and standard deviation of the imputed values of year of birth women at each iteration of the FCS imputation method for each of the five requested imputations. The purpose of this plot is to look for patterns in the lines. There should not be any, and these look suitably “random”. You can create similar plots for the other scale variables, and note that those plots also show no discernable patterns.

3.5. Two-step cluster analysis

Model Summary and Cluster Quality

- The model summary in Fig.9 and Fig.24 indicates that tow clusters were found based on the six input features (fields) selected.
- The model summary in Fig.12, Fig.18 and Fig.21 indicates that four clusters were found based on the six input features (fields) selected.
- The model summary in Fig.15 indicates that three clusters were found based on the six input features (fields) selected.
- The model summary of cluster quality chart in Fig.9, Fig.12, Fig.15, Fig.18, Fig.21, Fig.24 indicates that the overall model quality is "Fair".

Cluster Distribution

The Cluster Sizes view in Fig.10 shows the frequency of each cluster. The pie chart assigned to the cluster, 41.3% of the records were assigned to the first cluster and 58.7% to the second. while Fig.13 , Fig.19 and Fig.22 shows 4 cluster size , 11.5% size of smallest cluster and 43.6% size of largest cluster, Fig.15 indicate that 3 cluster size 11.8% size of smallest cluster and 56.9% size of largest cluster, only 2 cluster size in Fig.25 indicate 34.0% for first cluster and 66.0% for the second cluster.

Fig.11 clusters are sorted from smallest to largest by cluster size, so they are currently ordered 1, 2.

Fig.14 clusters are sorted from smallest to largest by cluster size, so they are currently ordered 2, 1, 3.

Fig.17 clusters are sorted from smallest to largest by cluster size, so they are currently ordered 3, 1, 2.
Fig.20 clusters are sorted from smallest to largest by cluster size, so they are currently ordered 1, 2, 3.
The cluster means suggest that the clusters are well separated.
The cluster means (for continuous fields) and modes (for categorical fields) are useful, but only give information about the cluster centers. In order to get a visualization of the distribution of values for each field by cluster.

3.5.1. Imputation number = original data

Fig. 9: Imputation Original Model Summary.

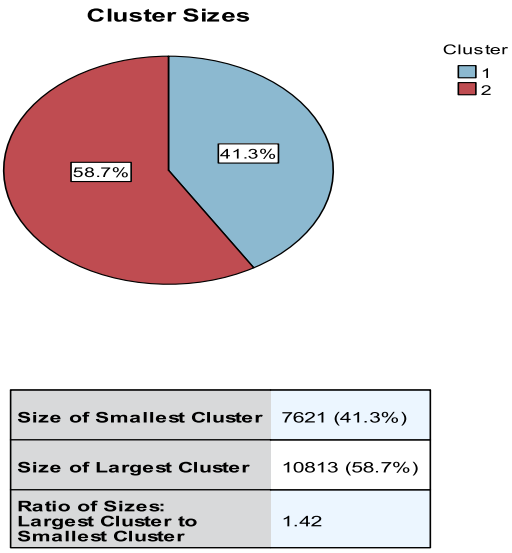


Fig. 10: Imputation Original Cluster Size.

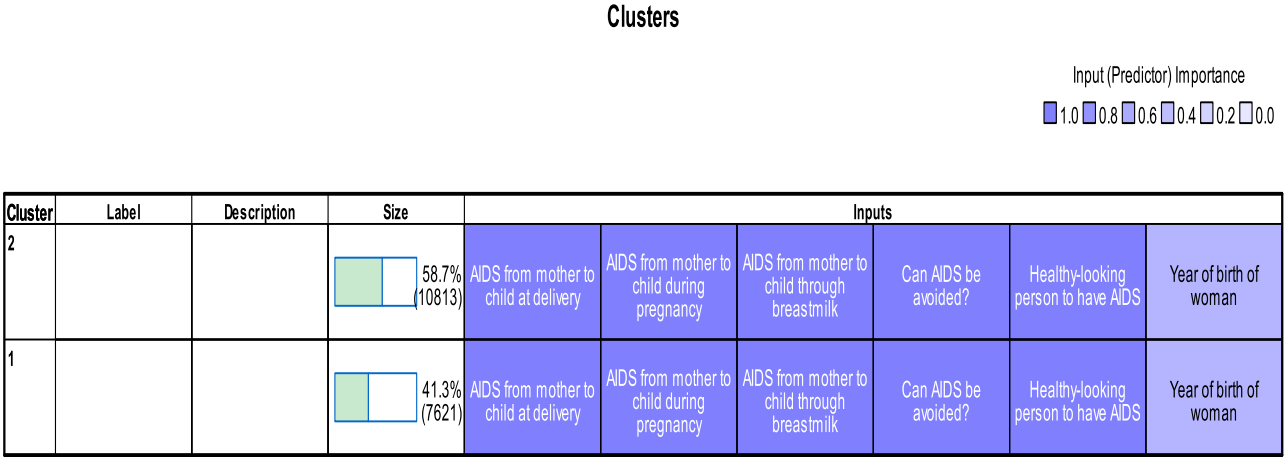


Fig. 11: Imputation Original Data Custers

3.5.2. Imputation number = 1

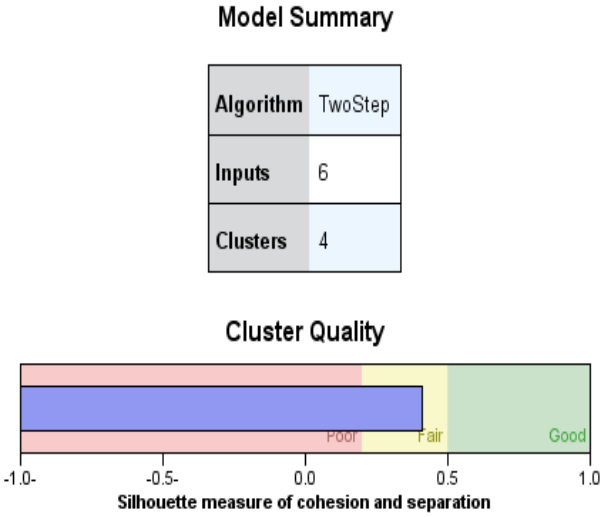


Fig. 12: Imputation Number 1 Model Summary.

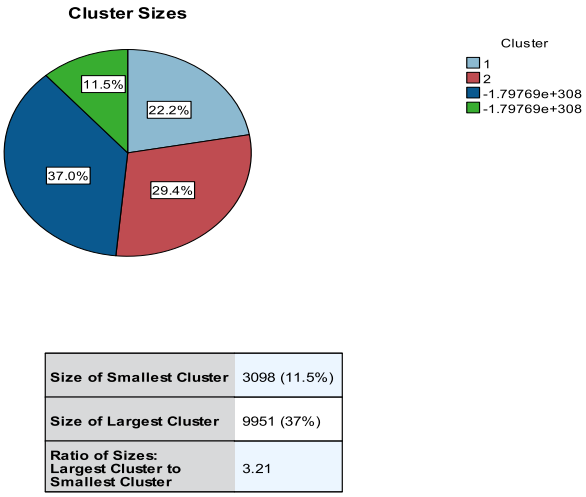


Fig. 13: Imputation Number 1 Cluster Size.

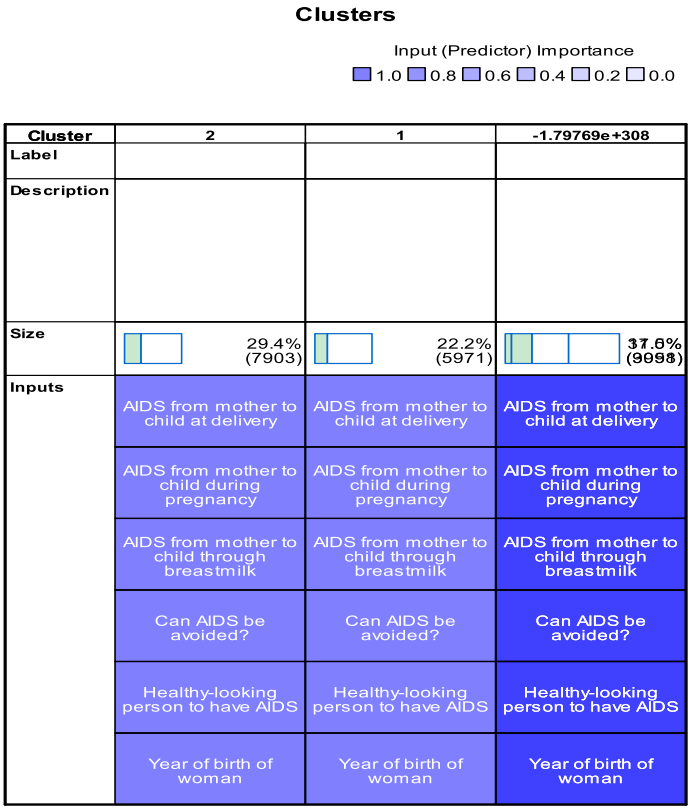


Fig. 14: Imputation Number 1 Clusters.

3.5.3. Imputation number = 2

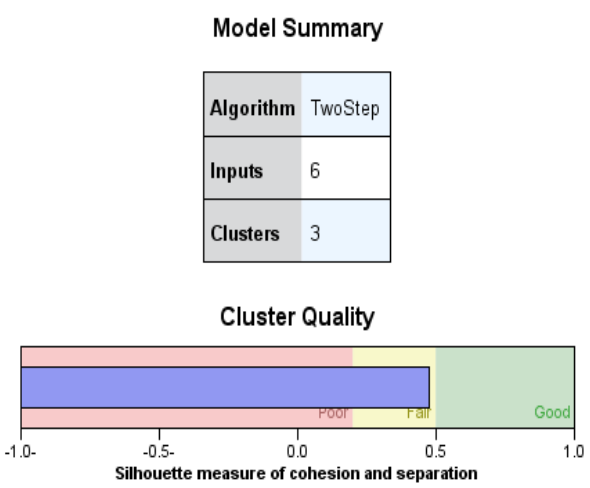


Fig. 15: Imputation Number 2 Model Summary.

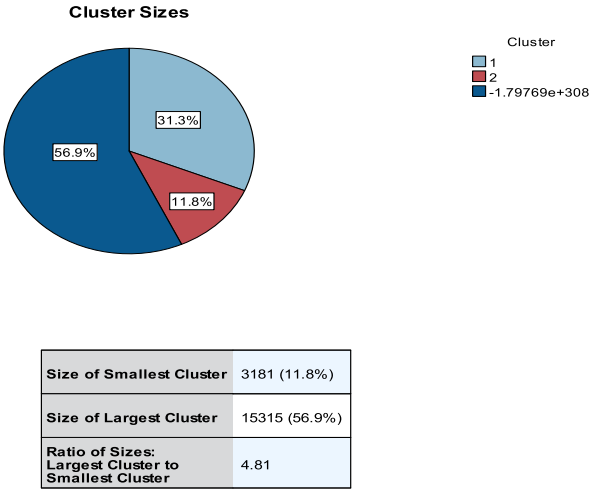


Fig. 16: Imputation Number 2 Cluster Size.

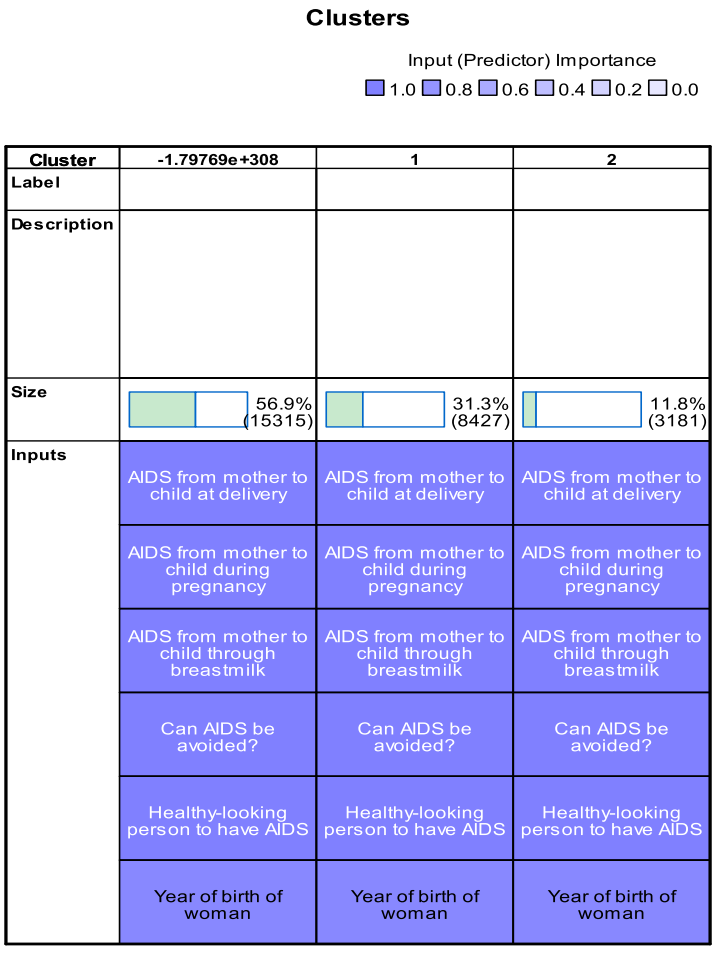


Fig. 17: Imputation Number 2 Clusters.

3.5.4. Imputation number = 3

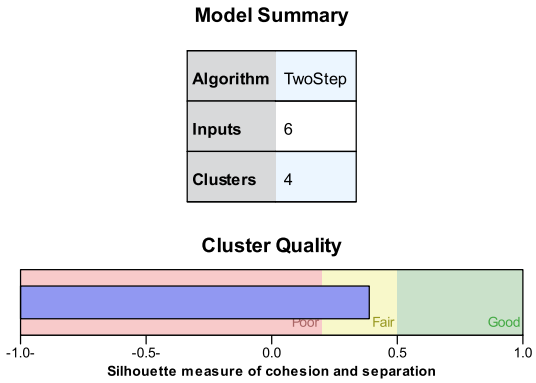


Fig. 18: Imputation Number 3 Model Summary.

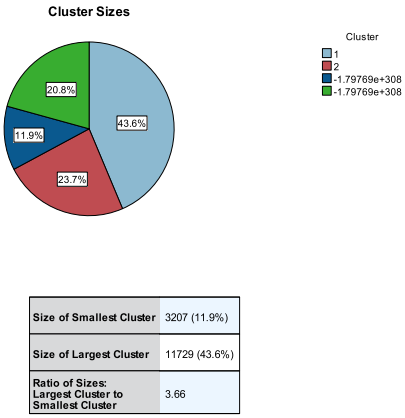


Fig. 19: Imputation Number 3 Cluster Size.

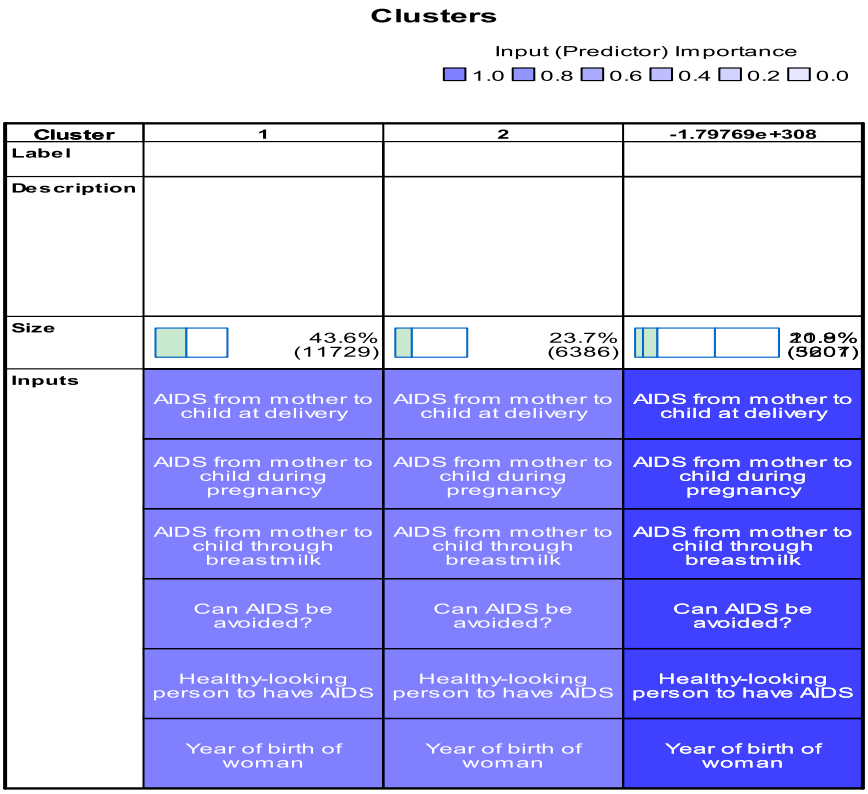


Fig. 20: Imputation Number 3 Clusters

3.5.5. Imputation number = 4

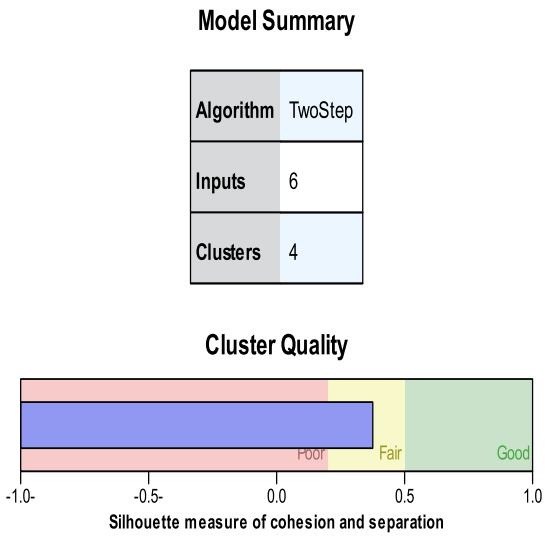
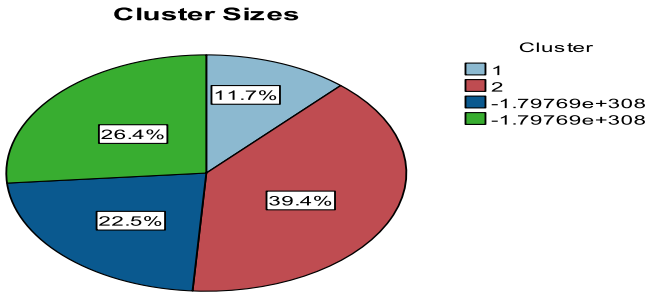
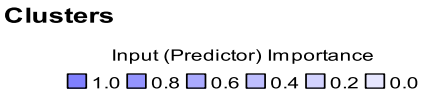


Fig. 21: Imputation Number 4 Model Summary.



Size of Smallest Cluster	3157 (11.7%)
Size of Largest Cluster	10600 (39.4%)
Ratio of Sizes: Largest Cluster to Smallest Cluster	3.36

Fig. 22: Imputation Number 4 Cluster Size.



Cluster	2	-1.79769e+308	1
Label			
Description			
Size	39.4% (10600)	26.5% (7096)	11.7% (3157)
Inputs	AIDS from mother to child at delivery	AIDS from mother to child at delivery	AIDS from mother to child at delivery
	AIDS from mother to child during pregnancy	AIDS from mother to child during pregnancy	AIDS from mother to child during pregnancy
	AIDS from mother to child through breastmilk	AIDS from mother to child through breastmilk	AIDS from mother to child through breastmilk
	Can AIDS be avoided?	Can AIDS be avoided?	Can AIDS be avoided?
	Healthy-looking person to have AIDS	Healthy-looking person to have AIDS	Healthy-looking person to have AIDS
	Year of birth of woman	Year of birth of woman	Year of birth of woman

Fig. 23: Imputation Number 4 Clusters.

3.5.6. Imputation number = 5

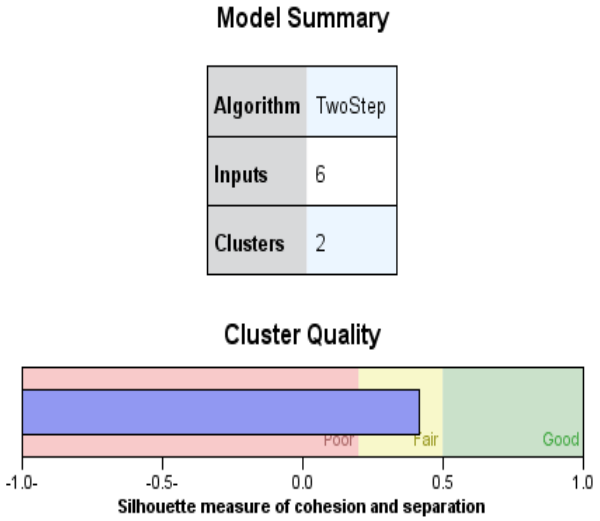


Fig. 24: Imputation Number 5 Model Summary

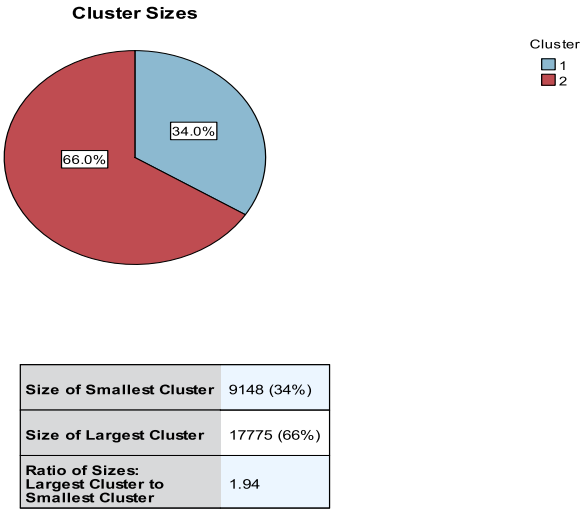


Fig. 25: Imputation Number 5 Cluster Size

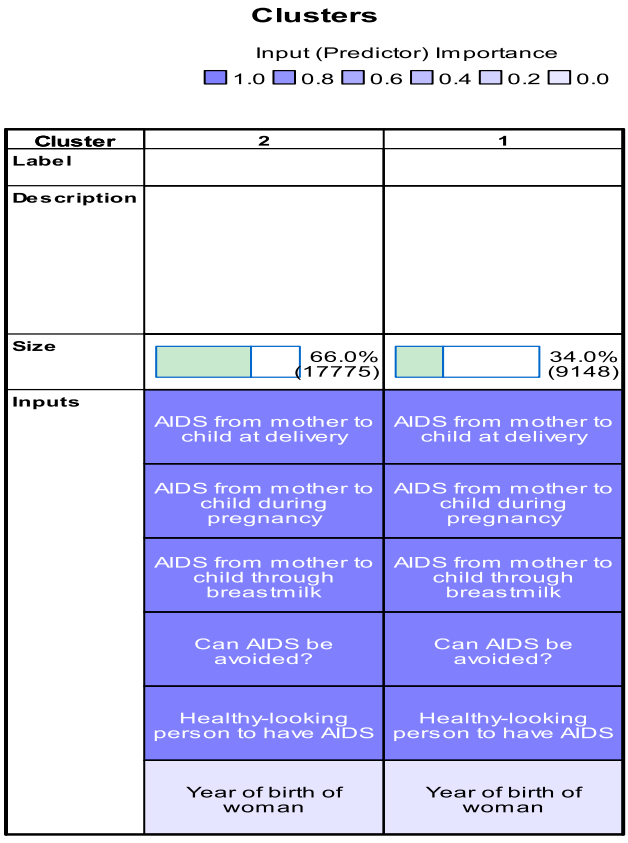


Fig. 26: Fig. 15 Imputation Number 5 Clusters.

4. Conclusions

This study focuses on missing data treatment on cluster performed on Sudan Household survey. Initially, missing data mechanism and treatment rules are presented. Two-Step Cluster Analysis is chosen over a wide range of approaches of statistical pattern-recognition available for clustering household health data.

Using multiple imputation, we analyzed patterns of missing values and found that much of the information is likely to be lost if the use of simple listwise deletion. After an initial automatic application of multiple imputations, we found that there was a need to maintain restrictions on the estimated values. To implement the program with the constraints produced good values, and there is no direct evidence that the FCS method does not converge. applying the "complete" data with imputation, and we fit multiple logistic regression to the border and display data pooled regression estimates and also discovered that this model fit the final will, in fact, it is not possible using listwise deletion on the original data.

Based on these results, I suggest the following conclusions:

1) Linear Regression imputation with attribution rounding should never be used. It is usually inferior and superior to never linear regression imputation without rounding, which is calculatedly simpler.
2) To estimate ratios and the main benefit of the imputation is to reduce bias when data are MAR but not MCAR.
3) To estimate the independent variable coefficient, which has no missing data, calculated with a linear regression, imputation is about as good as logistics.

Acknowledgement

I would take this opportunity to thank my research supervisor Dr. Amin Ibrahim, family and friends (Dr. Mohammed suleman Gibreel, Dr. Zakria Mohamed Salih) for their support and guidance without which this research would not have been possible.

References

[1] Ngondi, J., Matthews, F., Reacher, M., Onsarigo, A., Matende, I., Baba, S., & Emerson, P. (2007). Prevalence of risk factors and severity of active trachoma in southern Sudan: an ordinal analysis. *American Journal of Tropical Medicine and Hygiene*, *77*(1), 126.

[2] Jain A. K. and Dubes R. C. (1988). Algorithms for clustering data, Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

[3] M. S. Aldenderfer, R. K. Blashfield, Cluster analysis, Sage Publications, London, England.

[4] Anderberg M. R. (1973). Cluster analysis for applications, Academic Press, Inc., London, and ASR: An integrated study. In Proc. of Eurospeech '99, 2407–2410.

[5] Karkka T., Inen and Ayramo, S., (2004).Robust clustering methods for incomplete and erroneous data, in Proceedings of the Fifth Conference on Data Mining,, pp. 101–112.

[6] R. J. Little, D. B. Rubin, Statistical analysis with missing data, John Wiley & Sons, (1987).

[7] Jain, A. K., Duin, R. P.W. and Mao, J. (2000) Statistical pattern recognition: A review, IEEE Trans. Pattern Anal. Mach. Intell., 22, pp. 4–37. http://dx.doi.org/10.1109/34.824819.

[8] A. Jain, M. Murty, P. Flynn, Data clustering: a review, ACM Computing Surveys, 31 (1999) 264–323. http://dx.doi.org/10.1145/331499.331504.

[9] Hastie, T., Tibshirani, R. and Friedman, J. (2001). The elements of statistical learning: Data mining, inference and prediction, Springer-Verlag. http://dx.doi.org/10.1007/978-0-387-21606-5.

[10] Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D. and Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. Bioinformatics 17 520–525. http://dx.doi.org/10.1093/bioinformatics/17.6.520.

[11] Ghahramani, Z. and Jordan, M. I. (1994). Learning from incomplete data. Tech. Rep., Massachusetts Inst. of Technology Artificial Intelligence Lab.

[12] Vizinho, A., Green, P., Cooke, M. and Josifovski, L. (1999). Missing data theory, spectral subtraction and signal-to-noise estimation for robust.

[13] Tresp, V., Neunier, R. and Ahmad, S. (1995). Efficient methods for dealing with missing data in supervised learning. In Advances in Neural Info Proc. Sys. 7.

[14] Wagstaff, K., Cardie, C., Rogers, S. and Schroedl, S. (2001). Constrained k-means clustering with background knowledge. In Proc. of the 18th Intl. Conf. on Machine Learning, 577–584.

[15] J. Han, M. Kamber, Data mining: concepts and techniques, Morgan Kaufmann Publishers, Inc., (2001).

[16] Hand D., Mannila, H. and Smyth P.,Principles of Data Mining, MIT Press,(2001).

[17] P. Tan, M. Steinbach, V. Kumar, Introduction to data mining, Addison-Wesley, Networks, 16 (2005) 645–678.

[18] Horvitz, D. G., and D.J. Thompson, (1952). "A generalization of sampling without replacement from a finite universe." The Journal of the American Statistical Association 47:663-685.

[19] D. B Rubin, "Inference and Missing Data," Biometrika, 63(1987)581–590. Multiple Imputations for Nonresponsive in Surveys, New York: Wiley. 8(1987) 3–15.Association, 91 (1976) 473–489.

[20] Deville, J.C. and C.E. Sarndal, (1992). "Calibration Estimating in Survey Sampling." Journal of the American Statistical Association 87:376-382. http://dx.doi.org/10.1080/01621459.1992.10475217.

[21] Folsom, R. E. and A.C. Singh, (2000). "The General Exponential Model for Sampling Weight Calibration for Extreme Values, Nonresponse, and Post-stratification." in Proceedings of the Survey Research Methods Section, American Statistical Association. Indianapolis, Indiana.

[22] Cochran, W. G., (1977). Sampling Techniques, Third Edition. New York: John Wiley & Sons.

[23] Skinner, C.J., D. Holt and T.M.F. Smith. Editors, (1989). Analysis of Complex Surveys. Wiley, New York.

[24] J. G. Ibrahim, M.-H. Chen, S. R. Lipsitz, "Missing- Data Methods for Generalized Linear Models: Comparative Review," Journal of the American Statistical Association, 100(2005) 332–346. http://dx.doi.org/10.1198/016214504000001844.

[25] J. Carpenter, "Annotated Bibliography on Missing Data", Available online at http://www.lshtm.ac.uk/ msu/ missingdata/biblio.html [accessed July 30, 2006].

[26] Horton, N. J., and Lipsitz, S. R. (2001), "Multiple Imputation in Practice: Comparison of Software Packages for Regression Models With Missing Variables," The American Statistician, 55, 244–254. http://dx.doi.org/10.1198/000313001317098266.

[27] I. Jansen, C. Bounces, G. Molenberghs, "Analyzing Incomplete Discrete Longitudinal Clinical Trial Data," Statistical Science, 21(2006) 52–69. http://dx.doi.org/10.1214/088342305000000322.

[28] Robins, J. M., Rotnitzky, A., and Zhao, L. P. , "Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data," Journal of the American Statistical Association, 90, (1995)106–121. http://dx.doi.org/10.1080/01621459.1995.10476493.

[29] Laird, N. M., "Missing Data in Longitudinal Studies," Statistics in Medicine, 7, (1988) 305–315. http://dx.doi.org/10.1002/sim.4780070131.

[30] T. E. Raghunathan , J. M. Lepkowski, P. Solenberger, "A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models," Survey Methodology, 27(2001) 85–95.

[31] Von Hippel, P., "Biases in SPSS 12.0 Missing Value Analysis," The American Statistician, 58, (2004), 160–164. http://dx.doi.org/10.1198/0003130043204.

[32] Van Buuren, S. (2006), Multiple Imputation Online [accessed August 19, 2015]. Available online at http://www.multiple-imputation.com. (In press), "Creating Multiple Imputations in Discrete and Continuous Data by Fully Conditional Specification," Statistical Methods in Medical Research.

[33] S. van Buuren, H. C. Boshuizen, D. L. Knook, "Multiple Imputation of Missing, (1999).

[34] P. D. Allison, "Multiple Imputation for Missing Data: A Cautionary Tale," Sociological Methods and Research, 28(2000) 301–309. http://dx.doi.org/10.1177/0049124100028003003.

[35] Raghunathan TE. What do we do with missing data? Some options for analysis of incomplete data.Annual Review of Public Health. 2004; 25:99–117. http://dx.doi.org/10.1146/annurev.publhealth.25.102802.124410.

[36] Meng XL. Missing data: dial M for??? Journal of the American Statistical Association.2000; 95(452):1325–1330. http://dx.doi.org/10.1080/01621459.2000.10474341.