International Journal of Advanced Statistics and Probability, 3 (2) (2015) 191-209
www.sciencepubco.com/index.php/IJASP
©Science Publishing Corporation
doi: 10.14419/ijasp.v3i2.5318
Research Paper

Missing data treatment method on cluster analysis

Elsiddig Elsadig Mohamed Koko *, Amin Ibrahim Adam Mohamed

Sudan University of Science & Technology, Faculty of science, Department of Statistics
**Corresponding author E-mail: siddiggt@gmail.com*

Abstract

The missing data in household health survey was challenged for the researcher because of incomplete analysis. The statistical tool cluster analysis methodology implemented in the collected data of Sudan's household health survey in 2006.
Current research 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. However, the risk of over-fitting of the data must be considered because cluster analysis is a multivariable statistical technique. 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. 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.

Keywords: *Cluster Analysis; Missing Data; Multiple Imputation Method; Sudan Household Health Survey (SHHS).*

1. Introduction

It could be very tedious for any researcher to make an analysis of incomplete data. In any research, data plays a significant role from every aspect. When the researchers had performed household health survey in Sudan County, they faced many problems because the people were not interested to provide their health data. Basically, an epidemics diseases had scattered in Sudan, due to which many people had died, which affected by the people. Due to this the household health survey program was commenced for inspecting the affected people and cause of this epidemic disease (see [1]).

1.1. Missing data

The incomplete data was insignificant for the analysis of research, 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 [2].

1.2. Criticism on data collection of household health surveys

Household surveys performed by World Health Organisation (WHO) are often criticized for estimation of missing observations ([4], [5] and [3]) and for the methodology for information collection ([6]; [7]). Biasness in the information is also found in much criticism, including the use of too small data for too many imputations, use of limited number of questions out of a large number of questions for indexing, inadequacy of the sample to represent the population, inherent flaws in method, and majority of key informants being the people of WHO[5].

1.3. Sudan household health survey (SHHS)

The Sudan household health survey was conducted in the year 2006. The household survey executed by Central Bureau of Statistics of Sudan and Federal Health Ministry department representing the two other health-related department. This survey was scientifically and monetary supported by UNICEF and Pan Arab health project organization. The
Sudan household health survey accumulated the hard work of all the health-related agencies to perform a unanimous survey that has fulfill the interests of all stakeholders as it was a mixture of analyzing the multiple factors that were announcing the scarcity 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. 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 [42].
Design of the sample for the Sudan Household Health Survey (SHHS) provided estimates on a large number of indicators on the basic health situation at the national level and for 25 States. 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.

2. Methods

2.1. Missing data treatment

Every real world study frequently faces the complication of missing data. From chances to the design, there are numerous factors that result in missing data. The situations leading to missing data often occur as some participants in a study try to protect confidentiality and purposely excise information; to provide values some subjects may decline, and some variables may not be collected from all subjects. A potentially biased as well as an inefficient method of complete case in which the subjects with missing observations are dropped, is frequently used despite its disadvantages. Many researchers have been trying to find the efficient and appropriate method for analysing the data with missing values. A large number of values with missing data might lead by few missing data points in each covariate thus many real-world settings require the models that incorporate predictors observed partially.
Multiple imputation has been primarily focused in real-world settings in the comprehensive overviews provided by [8], [9], and [5]. Though the work is somewhat dated, maximum likelihood approach of [10] was included in the hierarchy of approaches to deal with the missing predictors described by [11].

2.1.1. “Ad-hoc” methods

Missing data are found to be addressed by a series of “ad-hoc” methods. One approach for continuous involves creation of a new variable that indicates the missing data, recoding missing observations to some common value, and then including the interaction between these variables and the variables themselves in the model. Creation of new variables for missing data holds for categorical variable too. These ad-hoc approaches are not recommended due to the potential induction of biasness ([12]; [13]). Another approach is to drop those subjects from the analysis which lack information for many variables. This approach is also not attractive as it often results in unnecessary large standard errors, consequent biasness, and exclusion of important variables. Two other non-recommended methods that have large variability and potential of inducing biasness are found in the work of [14], [15] and [16]. These methods impute missing values through using the last observed value (also known as last observation carried forward LOCF or last value carried forward LVCF) for longitudinal analysis and the average of observed values that is mean imputation.

2.1.2. Multiple imputations

[17] 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.
A public survey data was the first setting to use the method of multiple imputations. Inclusion of detailed and confidential information in a model can be created as auxiliary information, which is unsuitable to include in the public data set. Hence, in survey data settings, multiple imputation remains ideally suited [18]. Each of that data sets can be analysed through the utilization of existing software provided the complete data sets. However, in a setting where a single person is the imputer and the analyst, multiple imputations is more commonly used [19] and [20].
The potential of bias arises from the misspecification of the model; hence the suitable specification of the model of imputation is the key issue for an analyst. Estimation of multivariate mode only needs the variance-covariance matrix and mean vector. Therefore, this computationally traceable model has been used very often. Biasness in result and

complications in analysis often occur when some of the variables are not Gaussian, in such situation multiple imputation is used [21]. Complication in joint distribution due to missing values in multiple continuous and categorical variables is a salient reason for using multiple imputations. However, the model for analysis must not be richer than the one used for imputation [2]. Following is the description of a number of methods that were found in the literature reviewed.

In addition to the aforementioned imputation methods which replace each missing value with one value, the multiple imputations (MI) by [2] replaces each missing value with a set of plausible values that represent the uncertainty of the correct value.

2.1.3. Conditional Gaussian

[22] Improved the Conditional Gaussian approach of imputation for both discrete and continuous missing values. Cases of continuous variables assume a multivariate normal distribution and cases of discrete variables assume a log-linear model [23]. In real-world of multiple categorical variables, a proliferation of parameters can be led by the fit of this general location model as saturated multinomial with shared covariance and separate means (as pointed in [24]. This resulted in the need of simplified log-linear model in practice. S-Plus missing data library and Schafer's mix program (assuming a form of monotonicity) has been implementing this approach.

2.1.4. Methods of weighting

The approach of weighting methods used to account missing predictor data (as pointed in [25] and [26]. Complete cases use weights in this approach, which are actually the probabilities obtained through fitting a model for the probability of missingness. Software such as SAS, SUDAAN, SPSS, or Stata that allows for weights can be used to fit weighting approaches.

The general formula for a sample design weight is arithmetically very simple; it is 1 divided by the probability of selection due to the survey design. However, these are usually scaled, so we define the weight as proportional to this number. For example, if there are 3 adults in a given household, the resulting sample design weight for the single interviewed adult will be proportional to 1/ (1/3), i.e. proportional to 3. In a one adult household, the weight will be simple proportional to 1/1, i.e. proportional to one. In other words, the influence of the former respondent is being increased threefold relative to the influence of the latter respondent to exactly compensate for the fact the former respondent was three times less likely to be included in the sample.

2.1.5. Chained equations

Chain equations are used in an alternative variable-by-variable approach [27] and [28]. 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, polytomous models are needed and logistic regression can be fit for dichotomous variables. Are Impute (for R and S-Plus), IVE ware (for SAS or standalone), ICE (for Stata), or MICE library (for R and S-Plus) can provide the implementations of the chained equation approach.

[28] 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 [27] 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.

2.1.6. Bayesian approaches

Posterior distribution sampling of interest is involved in Bayesian framework. Bayesian methods have been more generally applied while multiple imputations were obtained within a Bayesian framework. The close relationship between MI and ML methods and the Bayesian approach estimates the covariates with a prior distribution as described [29]. Estimation of relationships requires a model with a package like WinBugs and specific coding of prior distributions partly due to the flexibility of these methods.

2.2. Cluster analysis

2.2.1. What is a cluster?

A formal definition of cluster is hard to give despite of the easy visual recognition of clusters from a two-dimensional view. The lack of formal and universal definition of cluster is addressed by many authors with the contribution in the literature of clustering. However, giving one definition is regarded as an intractable problem by the authors [30] and [31]. The weakly defined notion of a cluster depends on the application [32]. The definition is also affected by the goal of cluster analysis. There are different sizes and shapes of clusters depending on the application. Moreover, due to dependency on the resolution, one is looking in the data (global versus local); even the number of inherent clusters in the data is not unambiguous [32].
Typically, strong internal similarities are possessed in data description in terms of clusters yielded from clustering methods [33]. External isolation (separation) and internal cohesion (homogeneity) are often used to define cluster. Hence the definition of cluster is a set of objects dissimilar to the objects in the other clusters but similar to the objects within the same cluster [34].

2.2.2. Missing data in cluster analysis

In describing the two alternative approaches of handling missing values; marginalization were missing values are ignored and imputation were estimated values are used to fill in missing values [12] did not consider imputation as a reliable approach in comparison with actually observed data.
Therefore, there is no universally best algorithm for clustering [20]. The best understanding of data set can be obtained when several cluster algorithms are tried [35]. Contributions made by engineers [37], social scientists [38], statisticians [36], biologists [39] and psychologists [40] show that the development of clustering methods is interdisciplinary.

2.2.3. Two-step cluster 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.

2.2.4. 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 to be 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. Fair robustness of the procedure in case of violation of both the distributional assumption and the assumption of independence is indicated by the empirical internal testing, 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.

2.2.5. Suggestions for analyzing survey data

As pointed in (41) suggested that analysis of survey data is based on assumption if it ignores the weights and the sample design. The weights for estimating relationships between variables, rates, or means might be safely ignores if the sample design is capable of generating sample of equal probability. [41] Called these designs epsem designs and stated that at near final or the final or stage of the design, epsem can be designed even with complex multi-stage samples. Even with the initially epsem design, unequal weights can be created by the adjustments for non-response.

3. Results & discussion

3.1. Characteristics of woman respondents

First, descriptive analysis using frequency tabulation was conducted. The Lists in Table1 indicate that 32,599 women (age 15-49 years) identified in the selected households, 26,923 were successfully interviewed, yielding a response rate of 82.6 percent. It is important to note that while the average response rate for women's were over 90 percent in 11 states, between 80 and 90 percent in five states, between 70 and 80 per cent in two states, between 60 and 70 percent in three states and between 50 and 60 percent in four states, being highest in Gezira at 98.6 per cent and the lowest in Western Bahr El Ghazal at 55.4 per cent. , as indicated in Table1, the response rate for women was low. The response rate for women's questionnaire was less than 60 per cent in four states in Southern Sudan.

Table 1: Number of Women Response Rates

state	Completed	Not at home	Refused	Partly completed	Incapacitated	Other	Total	Response rate %
Northern	1290	54	0	0	16	20	1380	93.5%
River Nile	1408	54	2	0	7	1	1472	95.7%
Red Sea	1139	17	3	1	3	12	1175	96.9%
Kassala	1200	14	0	0	7	20	1241	96.7%
Gadarif	1207	44	5	0	8	26	1290	93.6%
Khartoum	1324	183	13	1	1	34	1556	85.1%
Gezira	1533	13	0	0	4	5	1555	98.6%
Sinnar	1347	21	0	0	1	17	1386	97.2%
Blue Nile	1220	101	5	0	5	6	1337	91.2%
White Nile	1500	23	1	0	6	4	1534	97.8%
North kordofan	1258	55	3	0	8	14	1338	94.0%
South kordofan	905	140	3	0	0	12	1060	85.4%
North Darfur	1055	104	4	0	2	32	1197	88.1%
West Darfur	773	97	6	1	1	24	902	85.7%
South Darfur	1027	39	1	0	5	12	1084	94.7%
Jongolei	887	197	33	0	0	339	1456	60.9%
Upper Nile	612	223	17	0	1	101	954	64.2%
Unity	906	274	38	2	1	92	1313	69.0%
Warab	1046	172	24	0	1	114	1357	77.1%
North Bahr Al_Gazal	837	308	31	3	0	319	1498	55.9%
West Bahr Al_Gazal	717	287	18	1	0	272	1295	55.4%
Lakes	899	352	63	0	1	170	1485	60.5%
West Equatoria	825	303	13	0	1	53	1195	69.0%
Central Equatoria	1067	242	43	17	0	47	1416	75.4%
East Equatoria	941	105	11	0	0	66	1123	83.8%
Total	26923	3422	337	26	79	1812	32599	82.6%

Table2 display the characteristics of female respondents 15-49 years of age. The table includes information on the distribution of women according to age, marital status, motherhood status, education and wealth index quintiles In addition to providing useful information on the background characteristics of women, the table is also show the numbers of observations in each background category. These categories are used in the subsequent tabulations of this work.

Table 2 Women in the age group 25-29 years constituted the largest proportion (21.1 %) of the total number of women followed by women in the age group 20-24 years (18.7 per cent), women in the age group 15-19 years (17.7 per cent), women in the age group 30-34 years (14.9 per cent), and women in the age group 35-39 years (14.1 %). About 8% of the women were in the age group 40-44 years while the lowest proportion of women was in the age group 45-49 years (5.5 per cent). About 65.5 percent were currently married/in union and 28.6 per cent were formerly married/in union while never married/in union women constituted 5.9 percent. Women with no formal education made up 49.8 percent of the total while 41.2 per cent had primary education and 8.9 percent had secondary or higher education. The wealth index quintiles show that about 17.7 percent of women belong to the poorest households while women from the richest households constitute about 23.5 percent.

Table 2: Women's Characteristics

			Number of woman	
			weighted	unweighted
Age	15-19	Count	1529508	4677
		% of Total	17.7%	17.4%
	20-24	Count	1611527	5005
		% of Total	18.6%	18.6%
	25-29	Count	1835955	5847
		% of Total	21.2%	21.7%
	30-34	Count	1291155	4037
		% of Total	14.9%	15.0%
	35-39	Count	1217325	3778
		% of Total	14.1%	14.0%
	40-44	Count	696905	2099
		% of Total	8.0%	7.8%
	45-49	Count	475590	1479
		% of Total	5.5%	5.5%
Total		Count	8657965	8657965
		% of Total	100.0%	100.0%
Marital/Union status	Currently married/in union	Count	5435614	17216
		% of Total	66.1%	67.8%
	Formerly married/in union	Count	2292572	6688
		% of Total	27.9%	26.3%
	Never married/in union	Count	495020	1487
		% of Total	6.0%	5.9%
Total		Count	8223206	25391
		% of Total	100.0%	100.0%
Motherhood status Ever given birth	Yes	Count	5615186	17882
		% of Total	64.9%	66.4%
	No	Count	3041795	9034
		% of Total	33.6%	35.1%
Total		Count	8656981	26916
		% of Total	100.0%	100.0%
Education	None	Count	4353377	14716
		% of Total	50.3%	54.7%
	Primary	Count	3508224	10383
		% of Total	40.5%	38.6%
	Secondary +	Count	784808	1776
		% of Total	9.1%	6.6%
	Missing/DK	Count	11981	48
		% of Total	.1%	.2%
Total		Count	8658390	26923
		% of Total	100.0%	100.0%
Wealth index quintiles	Poorest	Count	1611387	5067
		% of Total	21.4%	21.1%
	Second	Count	1497565	4720
		% of Total	19.9%	19.6%
	Middle	Count	1357048	4329
		% of Total	18.0%	18.0%
	Fourth	Count	1051533	3342
		% of Total	14.0%	13.9%
	Richest	Count	700768	2282
		% of Total	9.3%	9.5%
Total		Count	6218301	19740
		% of Total	82.6%	82.1%

3.2. Describing the pattern of missing data

Table 3: Univariate Statistics Pattern of Missing Data

	N	Mean	Std. Deviation	Missing		No. of Extremes[a]	
				Count	Percent	Low	High
Marital/Union status	8210555	1.40	.600	2182830	21.0	0	1487
Wealth index quintiles	7513617	2.66	1.333	2879768	27.7	0	0
Education	8645015	1.60	.707	1748370	16.8	0	48
Ever given birth	8643611	1.35	.477	1749774	16.8	0	0
Age of Woman	8645015			1748370	16.8		
a. Number of cases outside the range (Mean - 2*SD, Mean + 2*SD).							

Table3 Indicate that withl5 (Wealth index quintile) has the greatest number of cases with missing values (27.7%), while age (Age of woman), melevel (level of education) and cm1 (ever given birth) has the least (16.8%). Marital/Union status has the greatest number of extreme values.

Table 4: Separate Variance T Testsa Pattern of Missing Data

		Marital/Union status	Wealth index quintiles	Education	Ever given birth
Marital/Union status	t	.	-74.5-	217.8	432.8
	df	.	2572921.1	483505.4	532098.8
	# Present	8210555	5887772	8210555	8209802
	# Missing	0	1625845	434460	433809
	Mean(Present)	1.40	2.64	1.61	1.36
	Mean(Missing)	.	2.72	1.37	1.13
Wealth index quintiles	t	131.5	.	-182.1-	-129.6-
	df	4420726.6	.	4362590.7	4344345.9
	# Present	5887772	7513617	6208445	6207041
	# Missing	2322783	0	2436570	2436570
	Mean(Present)	1.42	2.66	1.57	1.34
	Mean(Missing)	1.36	.	1.67	1.39
Education	t	.	-88.1-	.	.
	df	.	1873585.8	.	.
	# Present	8210555	6208445	8645015	8643611
	# Missing	0	1305172	0	0
	Mean(Present)	1.40	2.64	1.60	1.35
	Mean(Missing)	.	2.75	.	.
Ever given birth	t	1905.8	-88.0-	23.4	.
	df	8209817.1	1876894.5	1404.1	.
	# Present	8209802	6207041	8643611	8643611
	# Missing	753	1306576	1404	0
	Mean(Present)	1.40	2.64	1.60	1.35
	Mean(Missing)	1.00	2.75	1.31	.
AGE OF WOMAN	t	.	-88.1-	.	.
	df	.	1873585.8	.	.
	# Present	8210555	6208445	8645015	8643611
	# Missing	0	1305172	0	0
	Mean(Present)	1.40	2.64	1.60	1.35
	Mean(Missing)	.	2.75	.	.
For each quantitative variable, pairs of groups are formed by indicator variables (present, missing).					
a. Indicator variables with less than 5% missing are not displayed.					

Table4 show that when wealth is missing, the mean education is 1.57, compared to 1.67 when wealth is no missing. In fact, the missingness of wealth seems to affect the means of several of the quantitative (scale) variables. This is one indication that the data may not be missing completely at random.

Table 8: Mstatus (Marital Status) Pattern of Missing Data

			Total	Currently married/in union	Formerly married/in union	Never married/in union	Missing
							SysMis
WM9	Present	Count	8645015	5427278	2289044	494233	434460
		Percent	83.2	100.0	100.0	100.0	19.9
	Missing	% SysMis	16.8	.0	.0	.0	80.1
CM1	Present	Count	8643611	5426525	2289044	494233	433809
		Percent	83.2	100.0	100.0	100.0	19.9
	Missing	% SysMis	16.8	.0	.0	.0	80.1
melevel	Present	Count	8645015	5427278	2289044	494233	434460
		Percent	83.2	100.0	100.0	100.0	19.9
	Missing	% SysMis	16.8	.0	.0	.0	80.1
wlthind5	Present	Count	7513617	3804983	1715323	367466	1625845
		Percent	72.3	70.1	74.9	74.4	74.5
	Missing	% SysMis	27.7	29.9	25.1	25.6	25.5
Indicator variables with less than 5% missing are not displayed.							

Looking at the Table 8 for melevel (Marital status), the number of missing values in the indicator variables does not appear to vary much between melevel (marital status) categories. Unmarried people reported wm9 (Age of woman) 100.0% of the time, and married people reported the same variable 100.0% of the time. The difference is none.

Table 9: Melevel (Education) Pattern of Missing Data

			Total	None	Primary	Secondary +	Missing/DK	Missing
								SysMis
WM9	Present	Count	8645015	4346614	3502454	783983	11964	0
		Percent	83.2	100.0	100.0	100.0	100.0	.0
	Missing	% SysMis	16.8	.0	.0	.0	.0	100.0
CM1	Present	Count	8643611	4345644	3502020	783983	11964	0
		Percent	83.2	100.0	100.0	100.0	100.0	.0
	Missing	% SysMis	16.8	.0	.0	.0	.0	100.0
mstatus	Present	Count	8210555	4048820	3384147	766910	10678	0
		Percent	79.0	93.1	96.6	97.8	89.3	.0
	Missing	% SysMis	21.0	6.9	3.4	2.2	10.7	100.0
wlthind5	Present	Count	7513617	3248990	2432898	517820	8737	1305172
		Percent	72.3	74.7	69.5	66.0	73.0	74.7
	Missing	% SysMis	27.7	25.3	30.5	34.0	27.0	25.3
Indicator variables with less than 5% missing are not displayed.								

Now consider the cross tabulation Table 9 for melevel (Level of education). If a respondent has at least some secondary+ education, a response for marital status is more to be missing. At least 93.1% of the respondents with none education reported marital status. On the other hand, only 97.8% of those with a secondary + reported marital status. The number is even lower for those with none education,

Table 10: Wlthind5 (Wealth) Pattern of Missing Data

			Total	Poorest	Second	Middle	Fourth	Richest	Missing
									SysMis
WM9	Present	Count	8645015	1608916	1495252	699573	1049835	699573	2436570
		Percent	83.2	84.4	82.4	79.2	82.6	79.2	84.6
	Missing	% SysMis	16.8	15.6	17.6	20.8	17.4	20.8	15.4
CM1	Present	Count	8643611	1608916	1494420	699573	1049618	699573	2436570
		Percent	83.2	84.4	82.4	79.2	82.6	79.2	84.6
	Missing	% SysMis	16.8	15.6	17.6	20.8	17.4	20.8	15.4
mstatus	Present	Count	8210555	1529072	1412702	664627	998739	664627	2322783
		Percent	79.0	80.2	77.9	75.3	78.6	75.3	80.7
	Missing	% SysMis	21.0	19.8	22.1	24.7	21.4	24.7	19.3
melevel	Present	Count	8645015	1608916	1495252	699573	1049835	699573	2436570
		Percent	83.2	84.4	82.4	79.2	82.6	79.2	84.6
	Missing	% SysMis	16.8	15.6	17.6	20.8	17.4	20.8	15.4
Indicator variables with less than 5% missing are not displayed.									

Now consider the cross tabulation Table 10 for wlthind5 (wealth). If a respondent has at least some wealth, a response for melevel (education level) is more to be missing. At least 84.4% of the respondents with poorest wealth reported melevel (education). On the other hand, only 82.6% of those with Middle reported melevel (education level). The number is even lowering for those with richest.

Table 11: EM Estimated Statistics

EM Means[a]	
WM9(age of woman)	
28.20	
a. Little's MCAR test: Chi-Square = .001, DF = 0, Sig. =.	
EM Covariances[a]	
	WM9
WM9	73.123
a. Little's MCAR test: Chi-Square = .001, DF = 0, Sig. =.	
EM Correlations[a]	
	WM9
WM9	1
a. Little's MCAR test: Chi-Square = .001, DF = 0, Sig. =.	

Table 11 describe that the null hypothesis for Little's MCAR test is that the data are missing completely at random (MCAR). Because the significance value is less than 0.05 in our work, we can conclude that the data are not missing completely at random. This confirms the conclusion we drew from the descriptive statistics and tabulated patterns.

3.3. Using multiple imputations to complete and analyze a dataset

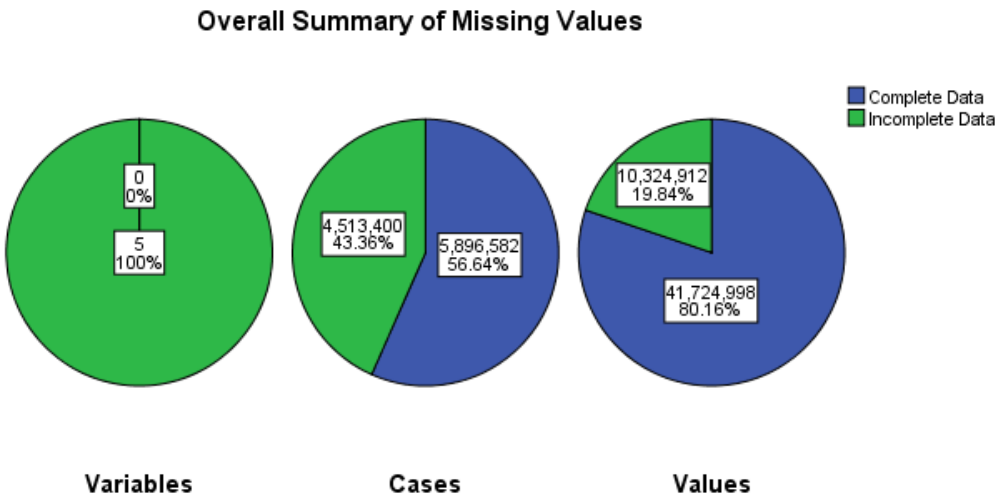

Fig. 1: Shows That:

- The Variables chart shows that each of the 5 analysis variables has at least one missing value on a case.
- The Cases chart shows that 4,513,400 of the 10,000,000 cases have at least one missing value on a variable.
- The Values chart shows that 10,324,912 of the 50,000,000 values (cases × variables) are missing.
- There are 5896582 (56.64 %) complete cases and 80.16% complete values.

3.4. Imputation models

Table 13: Imputation Specifications

Imputation Method	Automatic
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	wmweight

Table 14: Imputation Results

Imputation Method		Fully Conditional Specification
Fully Conditional Specification Method Iterations		10
Dependent Variables	Imputed	WM9,CM1,mstatus,melevel,wlthind5
	Not Imputed(Too Many Missing Values)	
	Not Imputed(No Missing Values)	
Imputation Sequence		WM9,melevel,CM1,mstatus,wlthind5

Table 15: Imputation Models

	Model		Missing Values	Imputed Values
	Type	Effects		
Age of woman	Linear Regression	melevel,CM1,mstatus,wlthind5	1307138	6535690
Education	Logistic Regression	CM1,mstatus,wlthind5,WM9	1307138	6535690
Ever given birth	Logistic Regression	melevel,mstatus,wlthind5,WM9	1308549	6542745
Marital/Union status	Logistic Regression	melevel,CM1,wlthind5,WM9	1742478	8712390
Wealth index quintiles	Logistic Regression	melevel,CM1,mstatus,WM9	2440309	12201545

Table 16: WM9 (Age of Woman) Imputed Values

Data	Imputation	N	Mean	Std. Deviation	Minimum	Maximum
Original Data		8658984	28.37	8.636	15.00	49.00
Imputed Values	1	1307138	28.76	8.528	-4.47-	56.87
	2	1307138	28.39	8.702	-7.61-	57.23
	3	1307138	28.59	8.441	-.81-	57.55
	4	1307138	28.90	8.564	-.41-	57.32
	5	1307138	28.94	8.599	-3.80-	57.55
Complete Data After Imputation	1	9966122	28.42	8.623	-4.47-	56.87
	2	9966122	28.37	8.645	-7.61-	57.23
	3	9966122	28.40	8.611	-.81-	57.55
	4	9966122	28.44	8.628	-.41-	57.32
	5	9966122	28.44	8.633	-3.80-	57.55

The descriptive statistics Table 16 for wm9 (Age of woman) shows means and standard deviations in each set of imputed values roughly equal to those in the original data; however, an immediate problem presents itself when you look at the minimum and see that negative values for age have been imputed. We will need to run a custom model with constraints on certain variables. However, age shows other potential problems. The mean values for each imputation are considerably higher than for the original data, and the maximum values for each imputation are considerably lower than for the original data. The distribution of age tends to be highly right-skew, so this could be the source of the problem.

3.5. Custom imputation model

wm9(age of woman's) is highly right-skew, and further analysis will likely use the logarithm of age, so it seems sensible to impute the log-age directly see Table 21.

Table 21: Logage

Data	Imputation	N	Mean	Std. Deviation	Minimum	Maximum
Original Data		8658984	3.2982	.30925	2.7081	3.8918
Imputed Values	1	1307138	3.3007	.30916	2.2083	4.3542
	2	1307138	3.3056	.31072	2.1610	4.3468
	3	1307138	3.3097	.30939	2.2814	4.3984
	4	1307138	3.2919	.30790	2.2108	4.3906
	5	1307138	3.2924	.31190	2.2033	4.4032
Complete Data After Imputation	1	9966122	3.2986	.30924	2.2083	4.3542
	2	9966122	3.2992	.30945	2.1610	4.3468
	3	9966122	3.2997	.30929	2.2814	4.3984
	4	9966122	3.2974	.30908	2.2108	4.3906
	5	9966122	3.2975	.30961	2.2033	4.4032

The descriptive statistics in Table 21 for logage (age of woman) under the custom imputation model with constraints shows that the problem of negative imputed values for tenure has been solved.

3.6. Nominal regression

Table 22: Case Processing Summary

		N	Marginal Percentage
Result of women 's interview	Completed	153204	87.7%
	Not at home	12880	7.4%
	Refused	1230	.7%
	Partly completed	85	.0%
	Incapacitated	295	.2%
	Other	7040	4.0%
Ever given birth	Yes	115766	66.3%
	No	58968	33.7%
Marital/Union status	Currently married/in union	117017	67.0%
	Formerly married/in union	47510	27.2%
	Never married/in union	10207	5.8%
Education	None	93235	53.4%
	Primary	66909	38.3%
	Secondary +	11895	6.8%
	Missing/DK	2695	1.5%
Wealth index quintiles	Poorest	44191	25.3%
	Second	41833	23.9%
	Middle	38289	21.9%
	Fourth	29543	16.9%
	Richest	20878	11.9%
Valid		174734	100.0%
Missing		20860	
Total		195594	
Subpopulation		24001[a]	
a. The dependent variable has only one value observed in 24001 (100.0%) subpopulations.			

Table 23: Model Fitting Information

Model	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	169938.358			
Final	160575.209	9363.149	55	.000

Table 24: Pseudo R-Square

Cox and Snell	.052
Nagelkerke	.084
McFadden	.055

Table 25: Likelihood Ratio Tests

Effect	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood of Reduced Model	Chi-Square	df	Sig.
Intercept	160575.209[a]	.000	0	.
logage	160576.806	1.597	5	.902
melevel	169534.249	8959.040	15	.000
CM1	160645.603	70.394	5	.000
mstatus	160618.077	42.868	10	.000
wlthind5	160840.481	265.272	20	.000

The chi-square statistic is the difference in -2 log-likelihoods between the final model and a reduced model. The reduced model is formed by omitting an effect from the final model. The null hypothesis is that all parameters of that effect are 0.

a. This reduced model is equivalent to the final mode l because omitting the effect does not increase the degrees of freedom.

Table 26: Model Fitting Information

Model	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	169938.358			
Final	160575.209	9363.149	55	.000

Table 28: Pseudo R-Square

Cox and Snell	.052
Nagelkerke	.084
McFadden	.055

Table 27: Likelihood Ratio Tests

Effect	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood of Reduced Model	Chi-Square	df	Sig.
Intercept	160575.209[a]	.000	0	.
logage	160576.806	1.597	5	.902
melevel	169534.249	8959.040	15	.000
CM1	160645.603	70.394	5	.000
mstatus	160618.077	42.868	10	.000
wlthind5	160840.481	265.272	20	.000

The chi-square statistic is the difference in -2 log-likelihoods between the final model and a reduced model. The reduced model is formed by omitting an effect from the final model. The null hypothesis is that all parameters of that effect are 0.
a. This reduced model is equivalent to the final model because omitting the effect does not increase the degrees of freedom.

3.7. Two-step cluster analysis

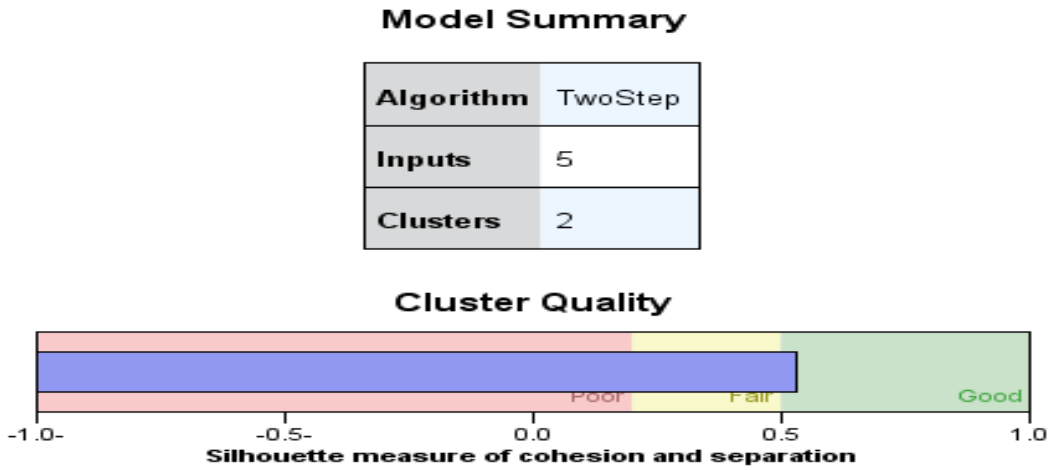

Fig. 1: Model Summary

- The model summary table in Fig.2 indicates that tow clusters were found based on the seven input features (fields) selected.
- The cluster quality chart in Fig. 2, Fig. 5 and Fig. 10 indicates that the model summary quality is "Good" while quality chart in Fig. 6, Fig. 8 and Fig. 12 indicates that the model summary quality is "Fair".

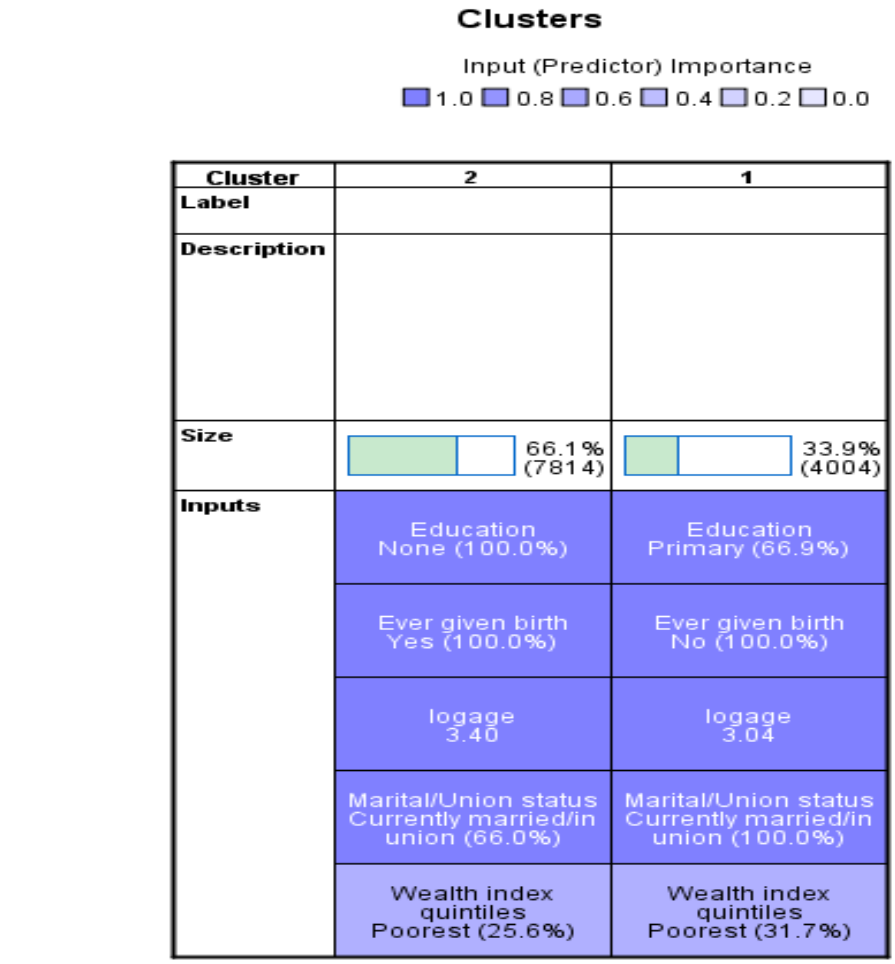

Fig. 2: Custer

The Cluster Sizes view in Fig. 3 shows the frequency of each cluster. Hovering over a slice in the pie chart reveals the number of records assigned to the cluster. 33.9% (4004) of the records were assigned to the first cluster and 66.1% (7814) to the second.

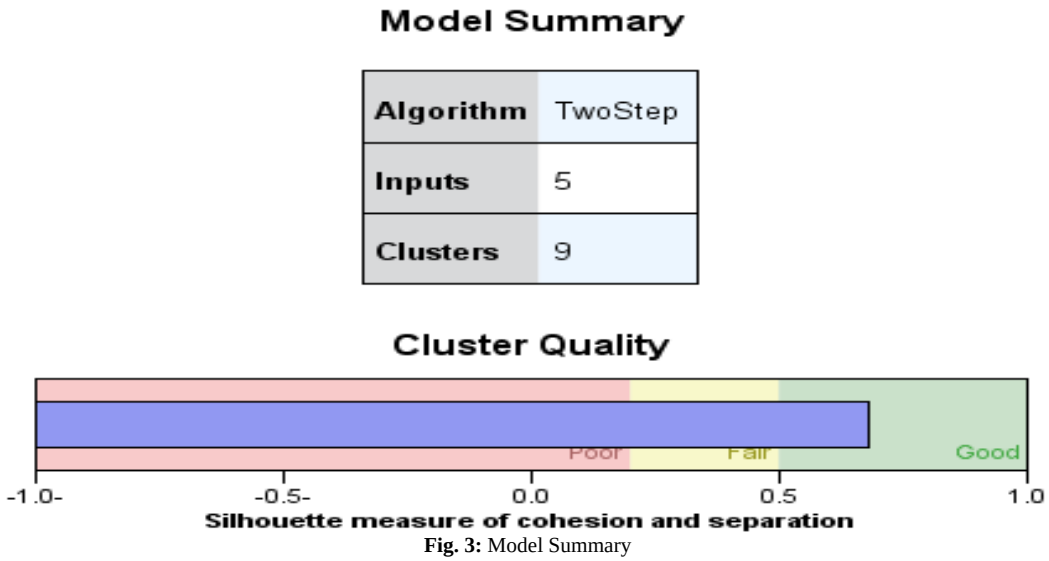

Fig. 3: Model Summary

Model Summary

Algorithm	TwoStep
Inputs	5
Clusters	4

Cluster Quality

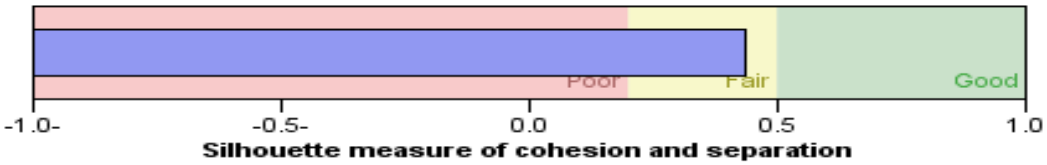

Fig. 4: Model Summary

Clusters

Input (Predictor) Importance

1.0 0.8 0.6 0.4 0.2 0.0

Cluster	-1.79769e+308	2	1
Label			
Description			
Size	36.6% (8805)	23.4% (4356)	20.1% (3749)
Inputs	Education Primary (67.0%)	Education None (100.0%)	Education None (100.0%)
	Ever given birth No (100.0%)	Ever given birth Yes (100.0%)	Ever given birth Yes (100.0%)
	logage 3.05	logage 3.39	logage 3.44
	Marital/Union status Currently married/in union (100.0%)	Marital/Union status Currently married/in union (100.0%)	Marital/Union status Formerly married/in union (100.0%)
	Wealth index quintiles Poorest (31.1%)	Wealth index quintiles Poorest (46.6%)	Wealth index quintiles Second (29.3%)

Fig. 5: Clusterd

Model Summary

Algorithm	TwoStep
Inputs	5
Clusters	4

Cluster Quality

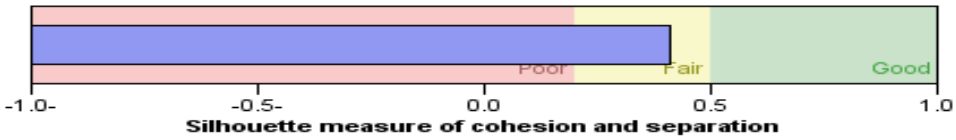

Fig. 6: Model Summary Imputation

Clusters

Input (Predictor) Importance

1.0 0.8 0.6 0.4 0.2 0.0

Cluster	-1.79769e+308	2	1
Label			
Description			
Size	22.9% (3048)	26.9% (5167)	9.3% (1786)
Inputs	Education Primary (67.0%)	Education None (77.5%)	Education None (100.0%)
	Ever given birth No (100.0%)	Ever given birth Yes (100.0%)	Ever given birth Yes (100.0%)
	logage 3.06	logage 3.39	logage 3.44
	Marital/Union status Currently married/in union (100.0%)	Marital/Union status Currently married/in union (100.0%)	Marital/Union status Formerly married/in union (100.0%)
	Wealth index quintiles Poorest (25.2%)	Wealth index quintiles Poorest (51.2%)	Wealth index quintiles Poorest (61.1%)

Fig. 7: Clusters

Imputation Number = 4

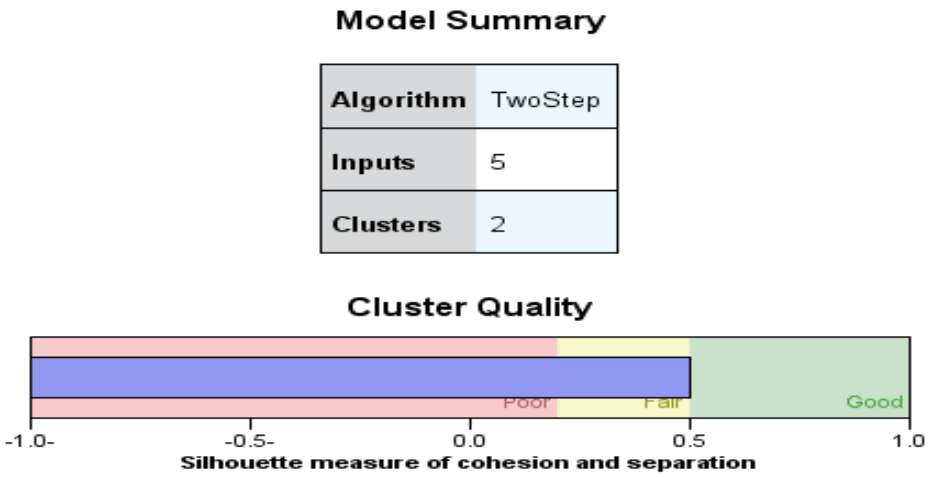

Fig. 8: Model Summary Imputation

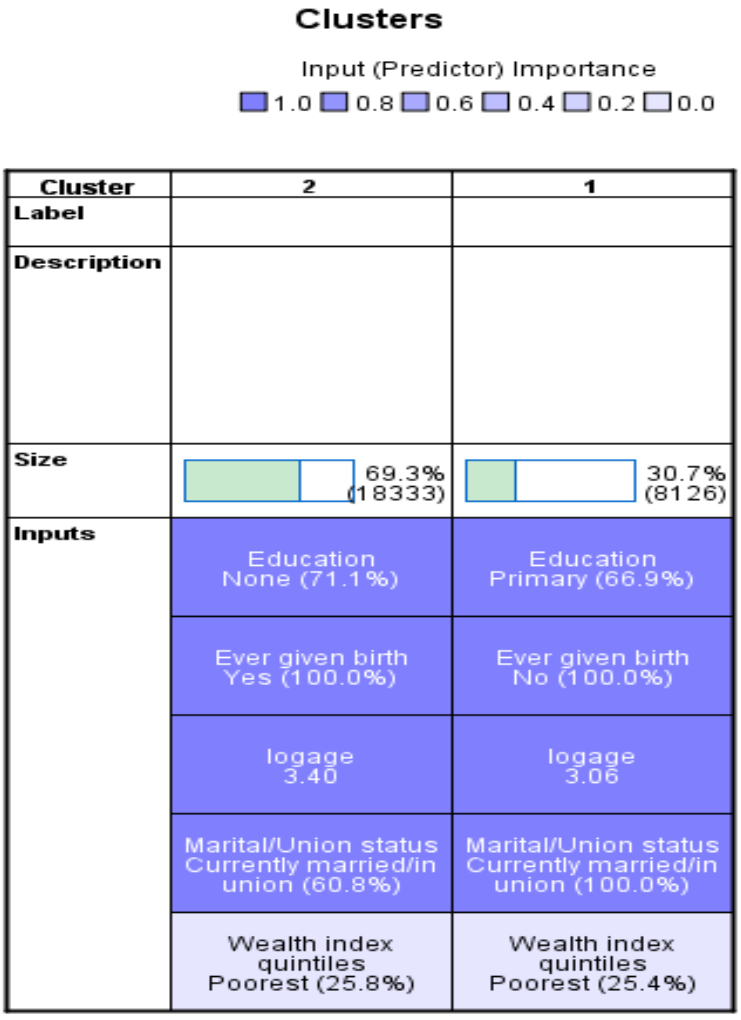

Fig. 9: Cluster

Imputation Number = 5

Model Summary

Algorithm	TwoStep
Inputs	5
Clusters	4

Cluster Quality

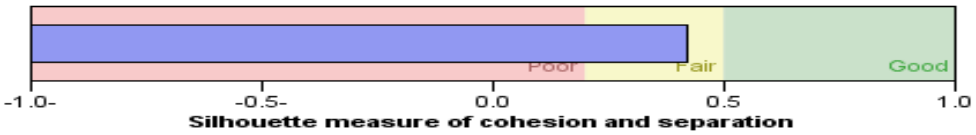

Fig. 10: Model Summary Imputation

Clusters

Input (Predictor) Importance

1.0 0.8 0.6 0.4 0.2 0.0

Cluster	1	-1.79769e+308	2
Label			
Description			
Size	27.5% (6361)	39.7% (9185)	17.1% (3964)
Inputs	Education None (61.8%)	Education None (100.0%)	Education None (100.0%)
	Ever given birth Yes (100.0%)	Ever given birth Yes (100.0%)	Ever given birth Yes (100.0%)
	logage 3.42	logage 3.39	logage 3.39
	Marital/Union status Formerly married/in union (100.0%)	Marital/Union status Currently married/in union (100.0%)	Marital/Union status Currently married/in union (100.0%)
	Wealth index quintiles Second (28.5%)	Wealth index quintiles Second (44.7%)	Wealth index quintiles Poorest (53.2%)

Fig. 11: Cluster

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. Using the multiple imputation procedures. Two-Step Cluster Analysis is chosen over a wide range of approaches of statistical pattern-recognition available for clustering household health data. 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 their 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. Prior to cluster analysis, log transformation will be 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 assumed to be 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 (co linearity) 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 co linearity. Specific diagnostic techniques for different techniques of cluster analysis are required for conditional correlation while calculation for global correlation is easy.

Acknowledgement

I would take this opportunity to thank my research supervisor, family and friends for their support and guidance without which this research would not have been possible.

References

[1] R. H. Henderson, T. Sundaresan, Cluster sampling to assess immunization coverage: a review of experience with a simplified sampling method, *Bulletin of the World Health Organization* 60(2) (1982) 253-260.
[2] R. J. Little, D. B. Rubin, Statistical analysis with missing data, John Wiley & Sons, (1987).
[3] A. Williams, Science or marketing at Who? A Commentary on 'World Health 2000', Health Economics, 10(2) (2000)93-100. http://dx.doi.org/10.1002/hec.594.
[4] A. M. Aalto, U. Häkkinenm, E.Ollila, Measuring the responsiveness of health care system in the World Health Report 2000. In Eds The World Health Report 2000: What does it tell us about health systems? Analyses by Finnish Experts. Helsinki, Finland: National Research and Development Centre for Welfare and Health (STAKES). [http://www.stakes.fi/english/publicati/Publications.htm]. (2000)
[5] R. Little, D. Rubin, Statistical Analysis With Missing Data (2nd ed.), New York: Wiley, (2002). http://dx.doi.org/10.1002/9781119013563.
[6] R. Blendon, M. Kim, and J. M. Benson, The public versus the World Health Organization on health system performance. Health Affairs, 20(3) (2001)10-20. http://dx.doi.org/10.1377/hlthaff.20.3.10.
[7] V. Navarro, World Health Report 2000: Response to Murray and Frenk. Lancet, 357(9269) (2001)1701-1702. http://dx.doi.org/10.1016/S0140-6736(00)04827-3.
[8] P. D. Allison, Missing Data, SAGE University Papers (2002).
[9] J. L. Schafer Analysis of Incomplete Multivariate Data, New York: Chapman & Hall, (1997).
[10] J. G. Ibrahim, "Incomplete Data in Generalized Linear Models," Journal of the American Statistical Association, 85(1990) 765–769. http://dx.doi.org/10.1080/01621459.1990.10474938.
[11] R. J. A. Little, "Regression with Missing X's: A Review," Journal of the American Statistical Association, 87(1992)1227–1237. http://dx.doi.org/10.2307/2290664.
[12] S. Greenland, W. D. Finkle, "A Critical Look at Methods for Handling Missing Covariates in Epidemiologic Regression Analyses," American Journal of Epidemiology, 142 (1995) 1255–1264.
[13] M. Jones, "Indicator and Stratification Methods for Missing Explanatory Variables in Multiple Linear Regression," Journal of the American Statistical Association, 91 (1996) 222–230. http://dx.doi.org/10.1080/01621459.1996.10476680.
[14] 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.
[15] R. J. Cook, L. Zeng, G. Y. Yi, "Marginal Analysis of Incomplete Longitudinal Binary Data: A Cautionary Note on LOCF Imputation," Biometrics, 60 (2004) 820–828. http://dx.doi.org/10.1111/j.0006-341X.2004.00234.x.
[16] J. Carpenter, M. Kenward, S. Evans, "Last Observation Carry-Forward and Last Observation Analysis," Statistics in Medicine, 23 (2004) 3241–3244. http://dx.doi.org/10.1002/sim.1891.
[17] 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.
[18] D. B. Rubin, "Multiple Imputation after 18+Years," Journal of the American Statistical (1996).
[19] J. Barnard, X. L. Meng, "Applications of Multiple Imputation in Medical Studies: From AIDS to NHANES," Statistical Methods in Medical Research, 8(1999) 17–36. http://dx.doi.org/10.1191/096228099666230705.
[20] P. D. Allison, "Imputation of Categorical Variables with PROC MI", Available online at http:// www2.sas.com/ proceedings/sugi30/ 113-30.pdf [accessed July 30, 2006]. Multiple Imputation," The American Statistician, 57 (2005) 229–232.
[21] 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.
[22] J. L. Schafer, Analysis of Incomplete Multivariate Data, New York: Chapman & Hall (1997). http://dx.doi.org/10.1201/9781439821862.
[23] Y. Bishop, S. Fienberg, P. Holland, Discrete Multivariate Analyses (1975).
[24] I. Olkin, R. F. Tate, "Multivariate Correlation Models With Mixed Discrete and Continuous Variables," The Annals of Mathematical Statistics, Theory and Practice, Cambridge, MA: MIT Press 32 (1961) 448–465. http://dx.doi.org/10.1214/aoms/1177705052.
[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] F. Xie, M. C. Paik, "Generalized Estimating Equation Model for Binary Outcomes With Missing Covariates," Biometrics, 90 Statistical Software Reviews 53 (1997) 1458–1466.
[27] S. van Buuren, H. C. Boshuizen, D. L. Knook, "Multiple Imputation of Missing, (1999).
[28] 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.
[29] 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.
[30] M. S. Aldenderfer, R. K. Blashfield, Cluster analysis, Sage Publications, London, England,

[31] R. M. Cormack, A review of classification, Journal of the Royal Statistical Society, Series A (General), 134 (1984) 321–367. http://dx.doi.org/10.2307/2344237.
[32] P. Tan, M. Steinbach, V. Kumar, Introduction to data mining, Addison-Wesley, Networks, 16 (2005) 645–678.
[33] R. Duda , P. Hart , Pattern Classification and Scene analysis, John Wiley & Sons, Inc, NY, (1973).
[34] J. Han, M. Kamber, Data mining: concepts and techniques, Morgan Kaufmann Publishers, Inc., (2001).
[35] 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.
[36] C. Fraley, A. Raftery, Model-based clustering, discriminant analysis, and density estimation, Journal of the American Statistical Association, 97(2002) 611–631. http://dx.doi.org/10.1198/016214502760047131.
[37] A. K. Jain, R. Dubes, Algorithms for clustering data, Prentice-Hall, Inc., Upper Saddle River, NJ, USA: (1988).
[38] K. D. Bailey, Cluster analysis, Sociological Methodology, 6(1975) 59–128. http://dx.doi.org/10.2307/270894.
[39] D. Jiang, C. Tang, A. Zhang, Cluster analysis for gene expression data: A survey, IEEE Transactions on Knowledge and Data Engineering, 16 (2004) 1370–1386. http://dx.doi.org/10.1109/TKDE.2004.68.
[40] G. Milligan, M. Cooper, An examination of procedures for determining the number of clusters in a data set, Psychometrika, 50(1985) 159–179. http://dx.doi.org/10.1007/BF02294245.
[41] Kish, Leslie, *Survey Sampling*, New York: John Wiley & Sons, Inc, (1965).
[42] A. Rose, R. F. Grais & H. Ritter. A comparison of cluster and systematic sampling methods for measuring crude mortality. Bulletin of the World Health Organization, 84(2006) 290-296. http://dx.doi.org/10.2471/BLT.05.029181.