Z_Annex_1

1. Introduction

The goal of this appendix is to detail the steps taken to mitigate the impact of missing values found in the DHS surveys.

Missing values are a real problem for multiple reasons. First, our decision to use PCA meant that missing data was not allowed as complete datasets are required for the algorithm to work. While there are some related techniques that do work with latent variables, the purpose of our study was also to compare with previous research that used PCA.

1.2. Description of methods

2. Methods

2.1. Indicators for which we can find an alternate source

The raw DHS data set sourced from STATCompiler ("STATcompiler" 2020) contained 37 columns, each corresponding to one of our 46 chosen indicators, and 83 rows, each corresponding to a specific survey (there 83 total surveys for our 29 SSA countries between the years of 2000 and 2018). The 9 other indicators that were included in our study were sourced from alternative sources as they were either already known to be unreliable or were simply not included in the DHS surveys. These 9 indicators are male circumcision rates ("Men.circumcised" sourced from World Bank [@]), ART coverage ("ART" sourced from World Bank [@]), data pertaining to rurality ("rural" sourced from World Bank [@]), gini wealth index ("Wealth.index.Gini" sourced from World Bank [@]), and data pertaining to religion (5 indicators overall: "Christian," "Muslim," "Folk.Religion,"

"Unaffiliated.Religion," and "Other.Religion" sourced from Correlates of War [@]) and had no missing values.

2.2. Identifying missing values from data set

The first step is of course to analyze the data set and get a sense of the impact missing values may have on our plans for analysis. The goal here is twofold, to see how many missing variables there are and where they are missing. As stated above, the raw data contained 46 columns, each corresponding to one of our 46 chosen indicators, and 83 rows, each corresponding to a specific survey. Each latent variable thus has a row and column position corresponding to the survey they belong to and which indicator they represent. It is of interest to list the missing values per country (Table 1).

Country	**Survey**	# Missing Values	Percentage missing
Angola	2015	0	0.00
Benin	2001	8	17.39
	2006	3	6.52
	2011	0	0.00
	2017	0	0.00
Burkina Faso	2003	8	17.39
	2010	0	0.00
Burundi	2010	0	0.00
	2016	0	0.00
Cameroon	2004	4	8.70
	2011	0	0.00
	2018	0	0.00
Chad	2004	6	13.04
	2014	0	0.00
Congo	2005	6	13.04
	2011	0	0.00
Congo Democratic Republic	2007	3	6.52
	2013	0	0.00
Cote d'Ivoire	2011	0	0.00
Ethiopia	2000	13	28.26
	2005	3	6.52
	2011	0	0.00
	2016	0	0.00
Gabon	2000	20	43.48
	2012	0	0.00
Gambia	2013	0	0.00
Ghana	2003	4	8.70
	2008	1	2.17
	2014	0	0.00
Kenya	2003	2	4.35

	2008	3	6.52
	2014	0	0.00
Lesotho	2004	5	10.87
	2009	4	8.70
	2014	0	0.00
Liberia	2007	2	4.35
	2013	0	0.00
Malawi	2000	7	15.22
	2004	3	6.52
	2010	0	0.00
	2015	0	0.00
Mali	2001	9	19.57
	2006	3	6.52
	2012	0	0.00
	2018	0	0.00
Mozambique	2003	2	4.35
	2011	0	0.00
Namibia	2000	9	19.57
	2006	1	2.17
	2013	0	0.00
Niger	2006	3	6.52
	2012	0	0.00
Nigeria	2003	4	8.70
	2008	1	2.17
	2013	0	0.00
	2018	2	4.35
Rwanda	2000	6	13.04
	2005	3	6.52
	2007	27	58.70
	2010	0	0.00
	2014	0	0.00
Senegal	2005	5	10.87
	2010	0	0.00
	2012	19	41.30
	2014	0	0.00
	2015	0	0.00
	2016	0	0.00
	2017	0	0.00
	2018	8	17.39
Sierra Leone	2008	1	2.17
	2013	0	0.00
Togo	2013	0	0.00
Uganda	2000	6	13.04
	2006	3	6.52
	2011	0	0.00

	2016	0	0.00
Zambia	2001	7	15.22
	2007	1	2.17
	2013	0	0.00
	2018	0	0.00
Zimbabwe	2005	3	6.52
	2010	2	4.35
	2015	0	0.00

Table 1: Missing values per country and survey

Doing so allows us to identify 2 surveys in particular:

- Senegal/2018 has 8 missing values while Senegal/2017 has none
- Senegal/2012 has 19 missing values while Senegal/2010 has none

Both Senegal/2018 and Senegal/2012 can thus be entirely discarded as Senegal will have reliable in close time proximity.

Table 1 also allows us to identify Rwanda/2007 and Rwanda/2005 which have 27 and 3 missing values respectively. Combining them is unfortunately unhelpful as the 3 missing values in Rwanda/2005 are also missing in Rwanda/2007, which can thus be discarded entirely as well.

We are now left with the same 46 columns but only 80 rows in which we can easily identify problematic (in the sense that they have many missing values) indicators by summing the number of missing values column-wise (Table 2), or surveys by summing the number of missing values row-wise (Table 1 again).

Indicator	**# Missing Values**	**Percentage missing**
Mean.number.of.sexual.partners.W.Normalized	22	26.51
Mean.number.of.sexual.partners.M.Normalized	21	25.30
Justified.condom.if.husband.has.STI.M	20	24.10
Ever.paid.for.sex	19	22.89
Justified.condom.if.husband.has.STI.W	15	18.07
Knowledge.about.AIDS.M	12	14.46
Buy.from.shopkeeper.with.AIDS.M	10	12.05
Buy.from.shopkeeper.with.AIDS.W	10	12.05
Wife.beating.justified.M	8	9.64
Married.women.participating.in.decisions	6	7.23
Ever.receiving.HIV.test.W	5	6.02
Married.women.who.disagree.with.wife.beating	5	6.02
Knowledge.about.AIDS.W	4	4.82
Wife.beating.justified.W	3	3.61
Ever.receiving.HIV.test.M	2	2.41
Number.of.co.wives.1	2	2.41

Unprotected.paid.sex	2	2.41
Number.of.co.wives.0	2	2.41
Number.of.co.wives.2	2	2.41
Literate.M	1	1.20
Access.to.media.W	1	1.20
Access.to.media.M	1	1.20
Number.of.wives.2	1	1.20
Number.of.wives.1	1	1.20
Literate.W	1	1.20

Table 2: Missing values per indicator

2.3. Imputation strategy

It is important to note that no country has all missing values for any particular indicator, or in other words, every country has at least one non-latent value for each of the 46 indicators between 2000 and 2018 – this is critical as it substantially improves our imputation results.

We devised an imputation strategy for the missing values above as follows:

1. Indicators that have 3 or more missing values and that have related indicators in the DHS surveys are imputed individually, these indicators are:
    - Wife.beating.justified.[W/M]
    - Knowledge.about.AIDS.[W/M]
    - Buy.from.shopkeeper.with.AIDS.[W/M]
    - Justified.condom.if.husband.has.STI.[W/M]
    - Mean.number.of.sexual.partners.[W/M]
    - Married.women.participating.in.decisions
    - Married.women.who.disagree.with.wife.beating
    - Ever.paid.for.sex
2. The remaining indicators that have only 1 or 2 missing values, or indicators that do not have related indicators in the DHS surveys are then imputed using the entire data set, these indicators are:
    - Literate.[W/M]
    - Access.to.media.[W/M]
    - Ever.receiving.HIV.test.[W/M]
    - Number.of.wives.1
    - Number.of.wives.2
    - Number.of.co.wives.0
    - Number.of.co.wives.1
    - Number.of.co.wives.2
    - Unprotected.paid.sex

2.3.1. Categorical imputation

For those indicators of the first group (3 or more missing values)

3. Results

"STATcompiler." 2020. https://www.statcompiler.com/en.