RBEIS imputation system

Fern Leather
Principal Statistical Methodologist | Office for National Statistics

September 2020

Outline

- What is it?
- Why was it developed?
- Demo

What is RBEIS?

- *Rogers & Berriman E&I System*
- A new more robust system for carrying out imputation of categorical variables in social surveys
- Variant of nearest neighbour donor imputation
- Developed in conjunction with Paul Smith of Southampton University
- Designed to minimise conditional imputation variance when the analytical aim is to produce a single, unique data set (rather than multiple imputed data sets)
- Successfully implemented in the E&I of social surveys at ONS since 2017

Why was it developed?

- ONS had previously used CANCEIS for imputation for a number of social surveys (WAS, EUSilc)
- While the NIM methodology in CANCEIS has many benefits, it is fundamentally designed for the E&I of large data sets such as the Census
- With the generic aim of producing a unique data set with relatively small sample-based social survey data, there has always been the risk of imputation variance having an undesirable impact on survey estimates
- In 2015 ONS had the resources and opportunity to address this problem

CANCEIS

- In CANCEIS, once a donor pool has been finalised for each record, a sampling with replacement is carried out, which results in both a risk of bias and high imputation variance if there are only a small number of records to be imputed

$$ConDistr_{imp} \neq ConDistr_{obs}$$

$$ConDistr_{imp\,1} \neq ConDistr_{imp\,2} \neq ConDistr_{impn}$$

CANCEIS

Final Donor Pool	
Donor	Var1
1	1
2	2
3	4
4	2

Frequency Distribution			
Var1	N	%	Exp(N)
1	1	25	0.5
2	2	50	1
4	1	25	0.5

Note: in this example, there are 2 records to impute with identical characteristics across all auxiliary variables

Imputed Distribution	
Recipient	Var1
A	1 or 2 or 4 (25:50:25 probability)
B	1 or 2 or 4 (25:50:25 probability)

How does RBEIS differ?

- Donor pools are constructed for Imputation Groups (records which share identical values for a set of auxiliary variables) and **NOT** individual records.
- These pre-processed donor pools are transformed into final donor pools which are the same size as the corresponding IGroup.
- The imputed distribution will **always** reflect the observed distribution conditioned on the matching variables

$$ConDistr_{imp} = or \cong ConDistr_{obs}$$

RBEIS

First Donor Pool	
Donor	**Var1**
1	1
2	2
3	4
4	2

Frequency Distribution			
Var1	**N**	**%**	**Exp(N)**
1	1	25	0.5
2	2	50	1
4	1	25	0.5

Note: in this example, there are 2 records to impute with identical characteristics across all auxiliary variables

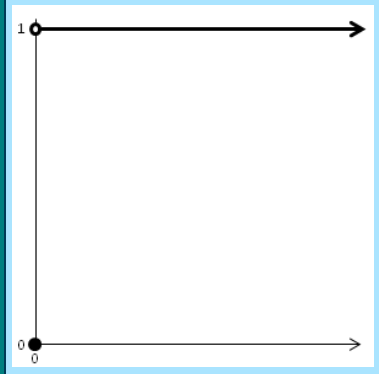
Final Donor Pool	
Recipient	**Var1**
A	2
B	1 or 4 (50:50 probability)

How are the donor pools created?

- Object is to choose records with characteristics which resemble as closely as possible those of the recipient record.
- Six distance functions to choose from.
- Weights assigned by the user to each auxiliary variable.
- Those records with the minimum sum of weighted distances are considered as potential donors and enter the pool.

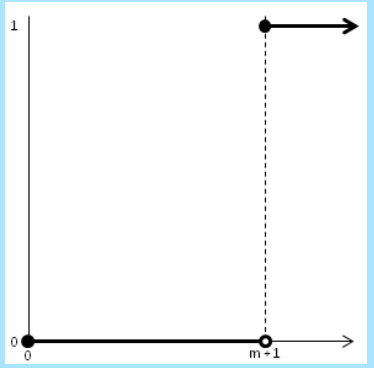
How are distances calculated?

Distance Function 1



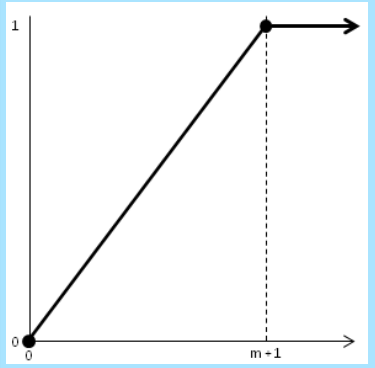
$$D_{IG,Donor,k} = \begin{cases} 0 & if \ V_{IG,k} = V_{Donor,k} \\ 1 & otherwise \end{cases}$$

Distance Function 2



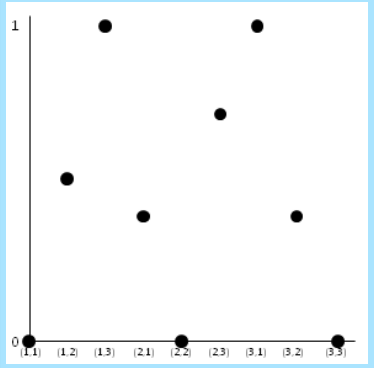
$$D_{IG,Donor,k} = \begin{cases} 0 & if \ |V_{IG,k} - V_{Donor,k}| \leq m \\ 1 & otherwise \end{cases}$$

Distance Function 3



$$D_{IG,Donor,k} = \begin{cases} \frac{|V_{IG,k} - V_{Donor,k}|}{m+1} & if \ |V_{IG,k} - V_{Donor,k}| \leq m \\ 1 & otherwise \end{cases}$$

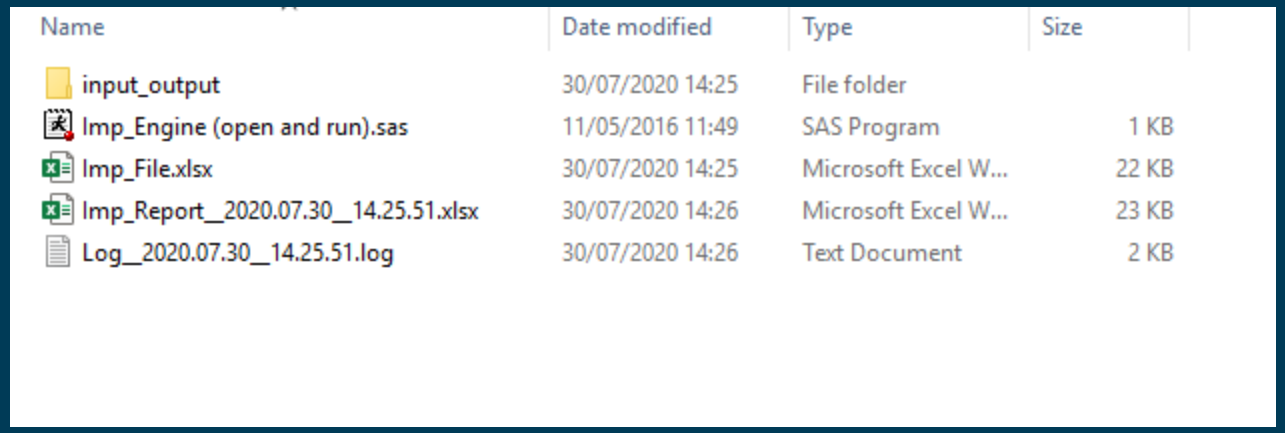
Distance Function 4/5/6



User-defined.

If relationship not specified:
DF4 → DF1, DF5 → DF2, DF6 → DF3

Demo



Name	Date modified	Type	Size
input_output	30/07/2020 14:25	File folder	
Imp_Engine (open and run).sas	11/05/2016 11:49	SAS Program	1 KB
Imp_File.xlsx	30/07/2020 14:25	Microsoft Excel W...	22 KB
Imp_Report__2020.07.30__14.25.51.xlsx	30/07/2020 14:26	Microsoft Excel W...	23 KB
Log__2020.07.30__14.25.51.log	30/07/2020 14:26	Text Document	2 KB

Demo

VIEWTABLE: TMP2.pre_imputed

	Player_ID	Team_ID	Position	League	Age	Nat_ID
1	1	1	1	1	33	35
2	2	1	2	1	31	10
3	3	1	2	1	30	-10
4	4	1	2	1	21	2
5	5	1	2	1	30	-10
6	6	1	3	1	23	37
7	7	1	3	1	20	19
8	8	1	3	1	25	9
9	9	1	3	1	27	10
10	10	1	-7	1	27	31
11	11	1	4	1	29	3
12	12	2	1	3	31	20
13	13	2	2	3	28	3
14	14	2	2	3	33	1
15	15	2	2	3	27	1
16	16	2	2	3	31	15
17	17	2	-7	3	24	-10
18	18	2	-7	3	30	32
19	19	2	-7	3	26	1
20	20	2	3	3	26	8

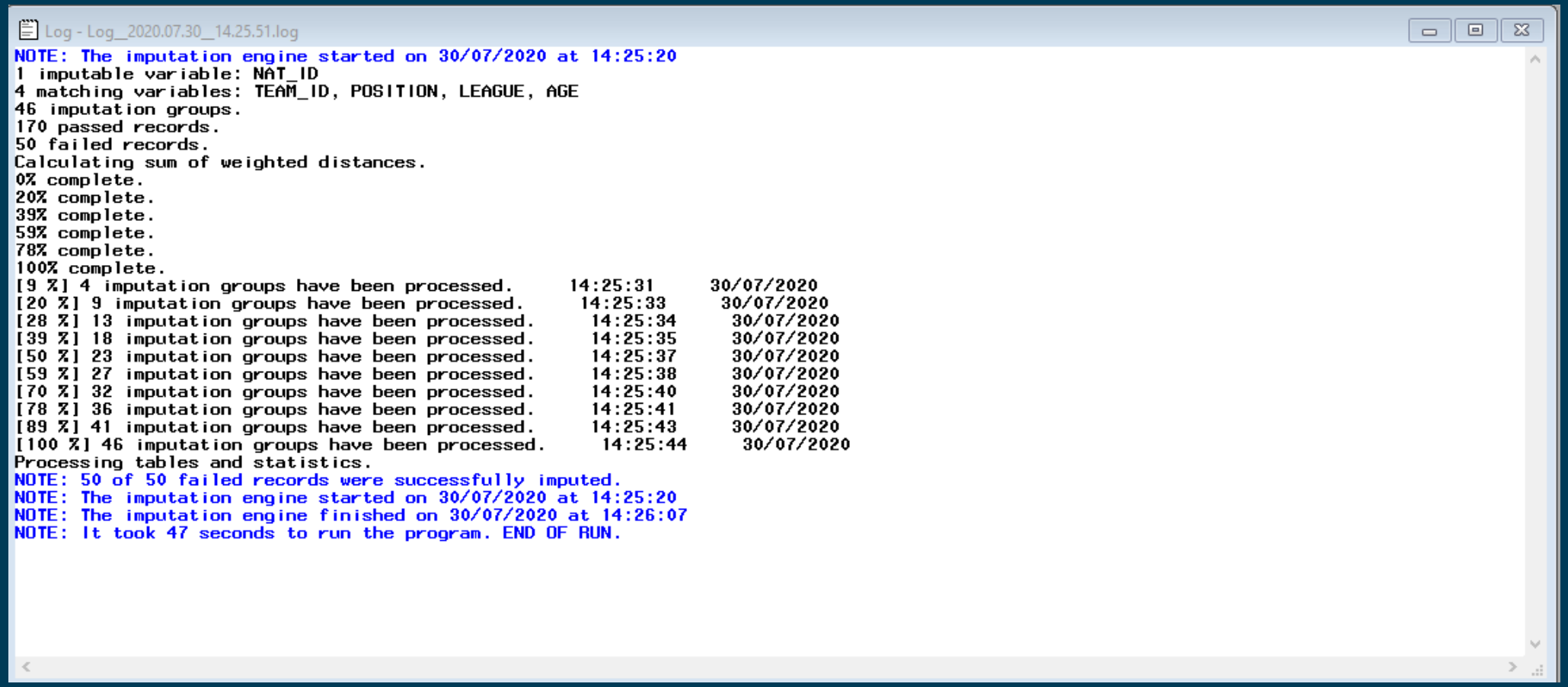
Demo

	A	B	C	D	E	F	G	H	I	J
1	**Variable**	**VarType**	**Range**	**Imputable**	**Weight**	**DisFun**	**MaxDif**	**DFMImp1**	**DFMDon1**	**DFMDis1**
2	Team_ID	1	1:20	0	1	1	0			
3	Position	1	-7,1:4	0	5	4	0	-7	-7,1:4	0
4	League	1	1:3	0	4	1	0			
5	Age	1	18:40	0	5	3	5			
6	Nat_ID	1	1:49	1	0	1	0			
7										
8										
9										
10										

FilNam | **ImpPar** | DonExc | ConRul | SysPar

Imp file - Contains the information relevant to the variable(s) the user wishes to impute, including: file paths, data set names, details of imputable variables, details of matching variables, donor exclusion rules and system parameters.

Demo


Log - Log_2020.07.30_14.25.51.log
NOTE: The imputation engine started on 30/07/2020 at 14:25:20
1 imputable variable: NAT_ID
4 matching variables: TEAM_ID, POSITION, LEAGUE, AGE
46 imputation groups.
170 passed records.
50 failed records.
Calculating sum of weighted distances.
0% complete.
20% complete.
39% complete.
59% complete.
78% complete.
100% complete.
[9 %] 4 imputation groups have been processed. 14:25:31 30/07/2020
[20 %] 9 imputation groups have been processed. 14:25:33 30/07/2020
[28 %] 13 imputation groups have been processed. 14:25:34 30/07/2020
[39 %] 18 imputation groups have been processed. 14:25:35 30/07/2020
[50 %] 23 imputation groups have been processed. 14:25:37 30/07/2020
[59 %] 27 imputation groups have been processed. 14:25:38 30/07/2020
[70 %] 32 imputation groups have been processed. 14:25:40 30/07/2020
[78 %] 36 imputation groups have been processed. 14:25:41 30/07/2020
[89 %] 41 imputation groups have been processed. 14:25:43 30/07/2020
[100 %] 46 imputation groups have been processed. 14:25:44 30/07/2020
Processing tables and statistics.
NOTE: 50 of 50 failed records were successfully imputed.
NOTE: The imputation engine started on 30/07/2020 at 14:25:20
NOTE: The imputation engine finished on 30/07/2020 at 14:26:07
NOTE: It took 47 seconds to run the program. END OF RUN.

Demo

Imp report

Sheet	Includes	
Main	Set of MVs, minimum sum of weighted distances for each group, distributions of the pre-processed donor pools and number of records in IGroup.	
Main_Full	Main with no spaces between IGroups.	
Min_Dis	Distribution of the minimum sum of weighted distances. Frequencies per IGroup and per record are given.	
Dif_Pcnt	Distribution of Dif_Pcnt (the difference between observed and imputed percentages for each value of our IV set).	How much variability is there in donors when conditioning on MVs?
C_Stat	Distribution of C_Stat (the 'average' sum of the squares of observed percentages for the values in each pre-processed donor group).	How informative is our MV set?

Demo

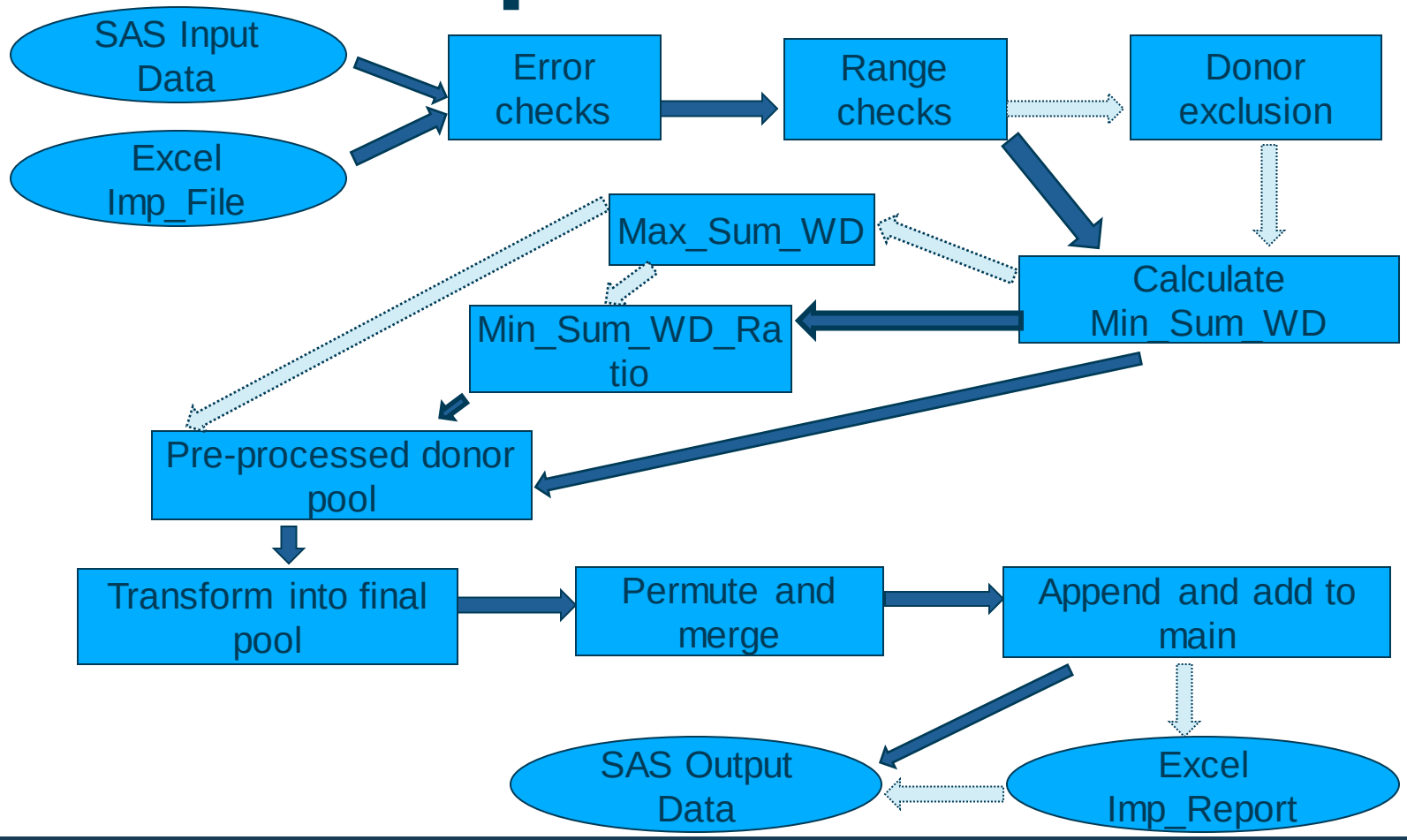
Group	MV_Set	Min_Dis	IV_Set	Obs_n	Obs_Pcnt	Exp_n	Imp_n	Imp_Pcnt	Dif_Pcnt
1	10,4,1,20	1.666667	1	1	100	1	1	100	0
2	11,2,3,21	1.833	20	1	100	1	1	100	0
3	6,3,2,22	1.833	3	1	100	1	1	100	0
4	11,2,3,22	1	20	1	100	1	1	100	0
5	8,2,1,23	0.833	1	1	100	1	1	100	0
6	20,3,1,23	1	1	1	50	0.5	0	0	-50
6	20,3,1,23	1	37	1	50	0.5	1	100	50

Demo

VIEWTABLE: TMP1.post_imputed

	Player_ID	Team_ID	Position	League	Age	Nat_ID	RBEIS_ID	RBEIS_IGroup	RBEIS_Fail_Imp
1	1	1	1	1	33	35	1	.	.
2	2	1	2	1	31	10	2	.	.
3	3	1	2	1	30	10	3	34	.
4	4	1	2	1	21	2	4	.	.
5	5	1	2	1	30	10	5	34	.
6	6	1	3	1	23	37	6	.	.
7	7	1	3	1	20	19	7	.	.
8	8	1	3	1	25	9	8	.	.
9	9	1	3	1	27	10	9	.	.
10	10	1	-7	1	27	31	10	.	.
11	11	1	4	1	29	3	11	.	.
12	12	2	1	3	31	20	12	.	.
13	13	2	2	3	28	3	13	.	.
14	14	2	2	3	33	1	14	.	.
15	15	2	2	3	27	1	15	.	.
16	16	2	2	3	31	15	16	.	.
17	17	2	-7	3	24	6	17	10	.
18	18	2	-7	3	30	32	18	.	.
19	19	2	-7	3	26	1	19	.	.
20	20	2	3	3	26	8	20		

Process map


SAS Input Data
Excel Imp_File
Error checks
Range checks
Donor exclusion
Max_Sum_WD
Calculate Min_Sum_WD
Min_Sum_WD_Ratio
Pre-processed donor pool
Transform into final pool
Permute and merge
Append and add to main
SAS Output Data
Excel Imp_Report

Drawbacks

- Risk of over-specifying the imputation model by including too many matching variables -> I Groups of only 1 record which match to only 1 donor
- Not possible to trace an imputation action back to the exact record that produced it which is not useful from an audit perspective (this is something that could be addressed in updates to RBEIS however).
- Not efficient for larger datasets (i.e. large number of I Groups or matching variables)
- Only limited functionality for edit rules
- Categorical matching variables only

Next steps

- Implement way to trace back donors
- Recode in open source language
- Imputation variance study
- Continuous/ordinal variables

Questions?