Regularizing approaches to data harmonization, cleaning using a library of processes

We noted that in the Fair by Role spreadsheet we need a column for Data Producer

Platform independent representations of what follows are important. These:

• Can be ingested by code generation tools

• Can generate human readable instructions and checklists

We discussed general categories of data processing:

• Cleaning

• Editing

• Imputation

• Recoding

• Harmonization

• Subsetting

(dropping variables

 or observations/cases/units)

• Adding variables via transformations

• Adding observations/cases/units

A question is how many of the above operations are specific to the social sciences?

We made a distinction between **purpose specific cleaning** vs **generic cleaning**. Examples of the latter might be checking that values fall into the defined value domain for a variable. Examples of the latter might be generating dummy variables from a categorical variable having a nominal domain for the purpose of a specific analysis.

Cleaning a dataset for users in one domain might be different than “cleaning” a dataset for the purpose of harmonizing it with a dataset from a different domain.

In working with data is there a way to automatize cleaning and harmonize data? This work can take 60% to 90% of the work time in working with the data.

We looked at as an example, the Demographic and Health Survey (DHS ):

3000 variables for a woman, 100K individuals in a dataset.

For a machine learning application the algorithims need numeric data. Input data may be categorical and need to be transformed. In the specific example we looked at the data were transformed from Stata to Python. Many categorical variables had text strings as their values. Many if statements transformed text values to numeric values.

We noted that DHS is DDI coded (see IPUMS).

We need to describe processes used to clean these data.

Steve mentioined another example: the Index of Social Surveys (ISS) program (GSS in US)

We questioned - why do countries do their own surveys? (access to microdata, not familiar with relevant standards) As a result there are multiple approaches to measuring the same concept.

Collecting the data consistently would be the best solution, but sometimes there are preexisting approaches (e.g. ISS) having had an evolutionary history.

We noted that it might be possible to use external classifications as input to a transformation process.

Secondary users of data and collection agencies both may have needs to do cleaning, but with different requirements.

Levels at which these operations (Cleaning etc.) may be expressed:

* Descriptions of tasks to be done
* Code tools.

### Can the Variable Cascade help here?

The relevant conceptual value domain is described at the ConceptualVariable level

Representation comes in at the RepresentedVariable level

Example Variable

For Question : What is your sex ?

Conceptual Value Domain: Male, Female

Represented value Domain: 1=Male 2=Female

Represented value Domain M=Male, F=Female

Variable with question What is your Gender? Is this the same value domain ? Depends on what concepts are intended (biological, self identified, legal)

Issues in our specific use case (cleaning DHS):

Python import of Stata sometimes fails.

Variables with more than xx% missing are dropped.

Variables with constant values are dropped

Variables with all values duplicating another variable are dropped.

Cases with specific values (substantative or sentinal) are dropped

Mapping different text codes with identical meaning are harmonized (e.g. a case mismatch) One canonical value is selected.

Mapping nuanced sentinel values to common value (R, Python NaN or NA). An example would be if Stata used .D for don’t know and .R for refused. Python cannot have differing values of missing. If the differences need to be preserved perhaps a flag variable needs to be created (like tripleS does).

Variables may need to be renamed (e.g. v003 becomes Sex)

Dummy variable coding. Example a Color variable with values Red, Yellow and Blue could be represented by three binary variables Red, Yellow, or Blue, with 1 if that color was the value in the original variable. This is often done for analyses (like regression) where a variable with a nominal level of representation cannot be used.

Dummy coding may have a limit on the number of categories that can be used.

Dealing with top and bottom coding

Ordinal variables into interval variables

Collapsing categories (e.g. « catholic, roman Catholic)  « Catholic »

 For this example is there an international classification of religion that could drive the collapse of lower levels into higher levels of the classification?

Documenting mappings between value domains might be done with a DDI correspondence table (with rationale))

Complex mapping rules may be needed (e.g. if there are fewer than five variables with anNA value recode NA =.1 but if all variables are NA, all are left as NA)

Dropping variables (e.g. survey paradata)

Drop variables with low variance? This is a choice. For some purposes these may need to be left in, in other cases they can skew results. An example might be if only one person of a certain characteristic was sampled, but the expected number in the sample should have been much higher, tabluations against that characteristic could misrepresent the actual proportions in the population,

Correlations among variables may need to be addressed. Different measures need to be used for different categories of value domains.

* Numerical (Pearson)
* Dummy (??)
* Mixed (eta ?)
* Character (non parametric, e.g. Spearman rank?)

Top coded extreme values can be a problem (example: Number of injections in the last 12 months =”90+ “)

Dealing with outliers is the other side of that coin

### Imputation

Imputation is replacing missing values with estimates based on models from other variables. There is a literature on this subject involving the reason for the missing value (“missing completely at random” etc.). Which variables are used as inputs to the estimation process may depend on the intended analyses so as to not get self-fulfilling results.

There is a similar process in other domains not using the same terminology. In image processing, for example missing pixels may be estimated by pixels in a surrounding region. Values may be changed to “sharpen” an image.

What is the relevance of the above to non-survey cleaning activities? Some of these may be quite generic and some specific to this domain.

Where in the FAIR framework does this all fit? These seem to b “I” and “R” issues but it’s not clear that any of the sub categories in our spreadsheet quite work.

Guidelines

High End GuideLines

Practictioner Guidelines

Best practice documents

Library of Descriptions of operations

Library of machine actionable operations

 platform specific

 open standard representations like VTL (SDMX), SDTL (C2 metadata)

Technical Examples

Our Python code examples?

Structured English

Machine actionable generic representations like VTL (SDMX), SDTL (C2 metadata)?