Methodology: Data cleaning

1 Data cleaning

All data sources potentially include errors and missing values – data cleaning addresses these anomalies. Not cleaning data can lead to a range of problems, including linking errors, model mis-specification, errors in parameter estimation and incorrect analysis leading users to draw false conclusions.

The impact of these problems is magnified in the S-DWH environment[1] due to the planned re-use of data: if the data contain untreated anomalies, the problems will repeat. The other key data cleaning requirement in a S-DWH is storage of data before cleaning and after every stage of cleaning, and complete metadata on any data cleaning actions applied to the data.

The main data cleaning processes are editing, validation and imputation. Editing and validation are sometimes used synonymously – in this manual we distinguish them as editing describing the identification of errors, and validation their correction. The remaining process, imputation, is the replacement of missing values.

Different data types have distinct issues regards data cleaning, so data-specific processing needs to be built into a S-DWH.

- Census data – although census data do not usually contain a high percentage of anomalies, the sheer volume of responses, allied with the number of questions, so data cleaning needs to be automatic wherever possible
- Sample survey data – business surveys generally have less responses, more variables, and more anomalies than social surveys – and are more complex due to the continuous nature of the variables (compared to categorical variables for social surveys) – so data cleaning needs to be very differently defined for business and social surveys
- Administrative data – traditional data cleaning techniques do not work for administrative data due to the size of the datasets and the underlying data collection (which legally and/or practically precludes recontact to validate responses), so data cleaning needs to be automatic wherever possible

2 Editing

Data editing can take place at different levels, and use different methods – the choice is known as the data editing strategy. Different data editing strategies are needed for each data type – there is no "one size fits all" solution for a S-DWH.

Macro- and micro-editing

Editing can be at the micro level, editing individual records, or the macro level, editing (aggregate) outputs.

- Macro-editing is generally subjective – *eye-balling* the output, in isolation and/or relative to similar outputs/previous time periods, or calculating measures of growth and applying *rules of thumb* to decide whether they are realistic or not. This type of editing would not suit the S-DWH

[1] The option of cleaning the data outside the S-DWH, using legacy (or newly built systems), and then combining cleaned data in the S-DWH is not recommended here – due to additional costs and lack of consistency/coherence – but the basic theory is the same wherever data cleaning is performed.

environment, as outputs are separated by two layers from inputs, and given the philosophy of re-use of data it would be difficult to define a process where "the needs of the one (output) outweigh the needs of the many". Hence nothing more is said about these methods.
- Micro-editing methods are numerous and well-established, and are appropriate for a S-DWH where editing should only take place in the sources layer. Hence these are the focus here.

Hard and soft edits

Editing methods – known as rules – detect errors, but once a response fails the treatment varies dependent on the rule type.

- Hard edits (*some* validity, consistency, logical and statistical) do not require validation and can be treated automatically – see below.
- Soft edits (all remaining) require external validation – see section 3.

Automatic editing

Automatic editing, mentioned in section 1 as a key option for census data, is also commonly used for business survey data as a cost- and burden-saving measure when responses fail hard edits. Given the high costs associated with development of a S-DWH, automatic editing should be implemented wherever possible – at least during initial development. However, another advantage of automatic editing applies both during development and beyond – it will lead to more timely data, as there will be less time spent validating failures, which will benefit all dependent outputs.

Selective editing

Selective (also significance) editing, like automatic editing, is a cost- and burden-saving measure. It reduces the amount of overall validation required by automatically treating the least important edit rule failures *as if they were not failures* – the remaining edit rule failures are sent for validation.

The decision whether to validate or not is driven by the selective editing score – all failures with scores above the threshold are sent for validation, all those with scores below are not validated.

The selective editing score is based on the actual return in period t (y_t), the expected return $E(y_t)$ (usually the return y_{t-1} in the previous period, but can also be based on administrative data), the weight in period t (w_t) – which is 1 for census and administrative data – and the estimated domain total in the previous period (Y_{t-1}):

$$\frac{\boldsymbol{w_t}|\boldsymbol{y_t} - \boldsymbol{E}(\boldsymbol{y_t})|}{\boldsymbol{Y_{t-1}}}$$

The selective editing threshold is set subjectively to balance cost versus quality: the higher the threshold, the better the savings but the worse the quality. In a S-DWH context, as responses can be used for multiple outputs, it is impossible to quantify the quality impact, so selective editing is of questionable utility. It should definitely be out of scope for all data in the backbone of the S-DWH.

3 Validation

Data validation takes place once responses fail edit rules, and are not treated automatically. The process involves human intervention to decide on the most appropriate treatment for each failure – based on three sources of information (in priority order):

- primary – answer given during a telephone call querying the response, or additional written information (*eg the respondent verified the response when recontacted*)
- secondary – previous responses from the same respondent (*eg if the current response, although a failure, follows the same pattern as previous responses then the response would be confirmed*)
- tertiary – current responses from similar respondents (*eg if there are more than one respondents in a household, their information could explain the response that failed the edit rule*)

In addition to these objective sources of information, there is also a valuable subjective source – the experience of the staff validating the data (*eg historical knowledge of the reasons for failures*).

In a S-DWH environment, the requirement for clean data needs to be balanced against the demand for timely data. This is a motivation for automatic editing, and is also a consideration for failures that cannot be automatically treated. The process would be more objective than outside a S-DWH, as the experience of staff working on a particular data source – the subjective information source for validation – would be lost given generic teams would validate all sources. This lack of experience could also mean that the secondary information source for validation – recognition of patterns over time – would also be less likely to be effective. This means that in a S-DWH, validation would be more likely to depend on the primary and tertiary sources of information – direct contact with respondents, and proxy information provided by similar respondents (or provided by the same respondent to another survey or administrative source).

4 Imputation

The final stage of data cleaning is imputation for partial missing response (item non-response) – the solution for total missing response (unit non-response) is estimation (see 1.3). To determine what imputation method to use requires understanding of the nature of the missing data.

Types of missingness

Missing data can be characterized as 3 types:

- MCAR (missing completely at random) – the missing responses are a random subsample of the overall sample
- MAR (missing at random) – the rate of missingness varies between identifiable groups, but within these groups the missing responses are MCAR
- NMAR (not missing at random) – the rate of missingness varies between identifiable groups, and within these groups the probability of being missing depends on the outcome variable

In a S-DWH environment, the ability to determine the type of missingness is in theory diminished due to the multiple groups and outcome variables the data could be used for, but in practice the type of missingness should be determined in terms of the primary purpose of the data source, as again it is impossible to predict all secondary uses.

Imputation methods

There is an intrinsic link between imputation and automatic editing: imputation methods define how to automatically replace a missing response based on an imputation rule; automatic editing defines how to automatically impute for a response failing an edit rule. Thus imputation methods are akin to automatic editing treatments, but the names are different.

There are a huge number of possible imputation methods – the choice is based on:

- the type of missingness – *generally* deterministic for MCAR, stochastic for MAR, deductive for NMAR
- testing each method against the truth – achieved by imputing existing responses, and measuring how close they imputed response is to the real response

In a S-DWH environment, the choice of imputation method should be determined based on the primary purpose of the data source – in concordance with the type of missingness. This chosen method, and its associated variance, must form part of the detailed metadata for each imputed response to ensure proper inference from all subsequent uses.