**Impudon Python Redesign**

***Methods Documentation***

This is a high-level documentation outlining two imputation methods integrated into our Python code-base: donor imputation and historical trend imputation. This document will provide an overview of these techniques, their implementation and utilization within our system. The donor imputation method leverages imputation class variables to identify the nearest neighbour, serving as the donor for missing data imputation. On the other hand, historical trend imputation employs a straightforward approach, utilizing the ratio of current and previous averages to modulate the previously reported value, thereby imputing the present reported value with enhanced accuracy that accounts for a previous to current wave trend.

Note that the code itself is documented differently through docstrings found right in the Python scripts and also in the /docs directory where API documentation in the form of HTML files can be found; the HTML files can be auto-generated using pdoc (see documentation). These documentation files could be used to render a static API documentation website in the future. The historical trend imputation is implemented from scratch with native Python and Pandas whereas the donor imputation is implemented using Scikit-learn. We recommend that the interested technical user goes through the code and traces every function to understand the inter-mechanics of the implementation.

**Donor Imputation**

**Processing Steps**

1. **Data preparation**
   * Bucket values of **variables to be imputed**
   * Standardize missing values for imputation: there are several values that can denote values to impute. *Verify with clients if they want to impute for refused/don't know responses.*
   * Having a standard format for specifying variables to impute (done through config file)
   * Make one copy of data prior to donor imputation, and add column for donor use counter.
2. **For applicable variables only, do historical imputation**
   * Use functions found in **src** directory of the codebase.
3. **Find donor**
   * Use function **nearest\_neighbour\_imputation** function to get donor value and its index.
   * Exclude donors with over 5 uses.

Based on StatCan’s earlier advice, RHS uses ratio or trend imputation method to do historical imputation

**Historical Imputation in RHS**

1. Based on StatCan’s earlier advice, RHS uses ratio or trend imputation method to do historical imputation for the work income received by main respondent and value of outstanding loan undertaken by main respondent. These two variables are selected as the values are expected to exhibit trends as time passes and they are likely to follow either an upward or a downward global measure of trend over time. In addition, trend imputation is undertaken only when the following criteria is met:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Specific response | | |
| Criteria | **Work income (including bonus and net earnings/profits)** | **Housing loan value** | |
| Past waves’ values | Non-imputed | | |
| Same records of characteristics in current and past waves | Did not change job across waves and have only one job in both waves – this is to ensure that the job reported in both waves are the same. | Reported the same housing in both waves.  For imputation of values for those respondents who reported the outstanding loan value for all owned housings, the number of owned housings should also be less than or equal to the previous wave. | |
| Sufficient records based on imputation parameter(s) | Sample size is at least 30\* | | |
| Imputation parameter | Based on industrial coding of main respondent’s current job – this is a derived variable based on Singapore Standard Industrial Classification | | If main respondent only owns one housing, donors are selected by public or private housing as price varies greatly between the two in Singapore. Otherwise, there is no imputation parameter if a respondent owns multiple housings. |

\*With reference to Statistics Canada’s quality level guidelines

1. This is done by calculating the change observed in selected donors over the reference time period through using the ratio of average specific responses in Wave t over Wave t-1 (for upward trend) and multiplied by the Wave t-1 response reported by the respondent to fill in the missing value in Wave t:

**X Reported value in past wave**

1. For downward trend exhibited by housing loan, this will be based on the ratio of average loan values in Wave t-1 over Wave r and multiplied by the Wave t-1 response reported by the respondent to fill in the missing value in Wave t.

**Post-imputation’s reasonableness check**

1. For work income, donor imputation using recipients from current wave data is applied if the trend imputed value is more than two times the value reported in past wave.

**Upcoming works**

1. With more waves’ information on the value of outstanding loan on commercial properties (the information was collected starting from Wave 3), we intend to study whether to apply trend imputation on this if we could identify whether the commercial properties owned across waves are similar.