# Readme / Social Housing Code Review

Benjamin Vandenbroucke / May 2017

This document aims at detailing the code flow and the considered assumptions for the reviewed code.

The SAS project *SH\_flow.egp* includes all the used sas programs and macros (in ./sasprog/ and ./sasauto/ folders). 2 intermediate steps have to be done with R (in ./rprogs).

The 12 steps process has to be run in 4 parts, switching between SAS and R:

- SAS: sh\_main.sas from 1 to the step 9 (create train/test dataset)

- R: main\_part1.R (propensity model)

- SAS: sh\_main.sas from 10 to step 11 (import scores and create cost table)

- R: main\_part2.R (cost analysis)

The 12 steps of the main program *sh\_main.sas* are detailed hereafter:

1. SET UP VARIABLES AND MACROS

In this step, some parameters has to be defined before running:

* The path of the ./*sasprog*/ folder
* Setup the libraries
* Macros variables to be defined in the *sh\_si\_setup()* macro:
  + Schema of the project
  + Name of the population cohort
  + Windowing parameters (profile and forecast periods, discounting rate)

2. DEFINE THE COHORT (2005/2006 HNZ APPLICATIONS)

This step creates the household and the individual tables for the applications of the considered period (2005/2006 in the base version, but can be changed:

* sand.hnz\_ind\_newapps\_0506 (50206 individuals)
* sand.hnz\_hh\_newapps\_0506 (21888 applications)

The following assumptions are considered to build the cohort of applications:

* All applications in the period = 34188
* Drop the ones having a multiple and distinct exit status = 34188-4089 = 30099
* Only the first application considered if several per individuals = 28605
* Removing duplicates snz\_uid = 26883
* Removing apps with a snz\_uid not linked in the concordance table = 26883 – 4995 = 21888
* Final number of applications = 21888

3. PERSONAL CHARACTERISTICS

This step creates the master characteristics table for snz\_uid included in the applications:

* Sand.pop\_master\_char

In cases where primary applicant is not part of the actual application, we use the eldest member of the application as the primary applicant:

* 99 applications : 99 primary\_snz\_uid updated in hnz\_hh\_newapps\_0506
* 252 rows updated in hnz\_ind\_newapps\_0506

4. ALIGN SIAL EVENTS TO THE PERIODS

This script creates the events tables from the SIAL events tables aligned to the profile and forecast periods:

* sand.XXX\_YYY\_events

This version points to a specific ACC SIAL table from the 20160715 refresh due to data availability issue in the latest refresh (data only from 2014). snz\_uid are updated to be consistent with last refresh.

IRD events are split into cost and revenue datasets.

5. APPLY DISCOUNTING TO THE EVENTS

This script creates the discounted events tables:

* work.XXX\_YYY\_events\_disc

The discounting rate is applied to the cost variable agencies (applied also on the revenue variable for IRD)

6. ROLLUP EVENTS

This script creates the rolled-up events tables:

* work.XXX\_YYY\_rollup(l/w)

MSD T2 events are split in 2 outputs:

* T2 events excluding WFF and AS
* T2 AS events

IRD costs are rolled up only for 'BEN','CLM','PEN','PPL','STU'

IRD revenues are split in 2 outputs:

* IRD\_wages for 'W&S' only
* IRD\_taxes = 13.39% \* 'W&S'

7. CREATE INDIVIDUAL LEVEL DATASETS

This script creates individual level variables datasets:

* hnz\_apps\_ind\_2005\_06\_cohort
* individual\_variables
* all\_ind\_variables\_cohort\_0506

8. CREATE HOUSEHOLD LEVEL DATASETS

This script creates household level variables datasets:

* hnz\_apps\_hhld\_2005\_06\_cohort
* household\_variables
* all\_hh\_variables\_cohort\_0506

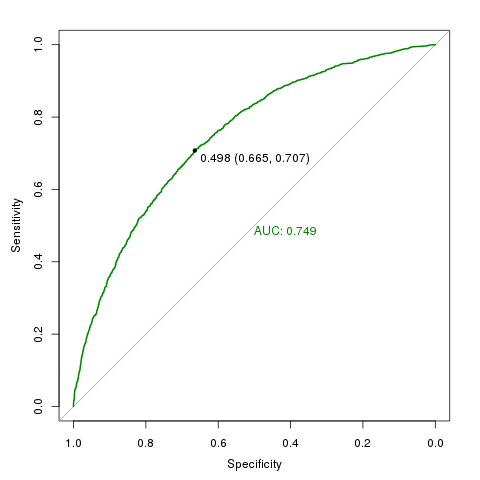
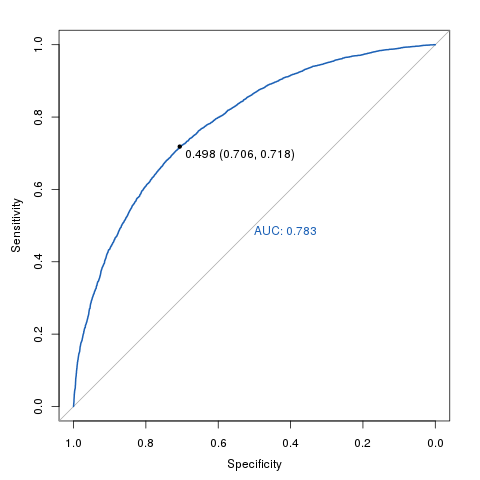
9. CREATE TRAIN/TEST DATASET

This script creates train / test partition of the cohort and adds a few variables (age band, appl. year/quarter...).

10. PROPENSITY MODEL \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

At this point, run the R file 'main\_part1.R including the following steps:

* Load libraries (packages versions are given in comments, might not work properly with previous/latest versions )
* 1. extract\_dataset : extract the training/validation set from the database
* 2. premodelling\_tasks: cleaning, visualisations and summaries
* 3. check\_correlations: dropping correlated variables
* 4. xgb\_model : grid search and model training
* 5. postmodelling\_tasks: final model training and scoring



Training and validation ROC

* 6. ps\_weighting\_and\_balancing: weights based on propensity scores

The folder ./output/ contains summary statistics and plots for the different stages of the modelling.

Back to SAS: Import the scored dataset generated through R with propensity weights.

11. CREATE COST MASTER TABLE

Create the cost master table including scaled weights:

* sand.hh\_cost\_master

12.COST ESTIMATES / BOOTSTRAP

At this point, run the R file 'main\_part2.R'

* 7.cost\_estimate\_boostrap: bootstrap algorithm to estimate the cost impact by agency(7) and subject area (21)

The following assumptions are considered:

* Considered T1 benefit is the IRD/BEN cost, not the MSD/T1
* T2 AS is excluded of the total MSD cost
* IRD TAX is excluded of the IRD cost
* NNP, PHA and PRM are excluded of the MOH costs

Final results, cost tables and plots are outputted in the folder ./output/

Z:\MAA2016-15 Supporting the Social Investment Unit\github_social_housing\QAd - Ben\output\plots\cost_per_sa.tiff