Quick reference for CKM in µ-Argus 5.1.7

Date: January 2023

Author: Peter-Paul de Wolf

Abstract:

This Quick Reference is an addition to the manual of µ-ARGUS version 5.1.3, describing the new functionality in µ-ARGUS version 5.1.6 that can be used to apply Targeted Record Swapping.

Note that some functionality is still under development.

Content

Introduction and background

Since version 5.1.5 a new protection method is included in μ-ARGUS: Targeted Record Swapping (TRS for short). For a description of this method, see deliverables 2.1 and 2.2 of the SGA "Open Source tools for perturbative confidentiality methods", partly funded by Eurostat (Contract N° 2018.0108 under FPA N° 11112.2014.005-2014.533). For more information on the implemented Targeted Record Swapping method itself, see https://github.com/sdcTools/recordSwapping/tree/master/docs.

In the subdirectory `hhdata` a small testdataset is provided named `Samhh.asc` with the corresponding `Samhh.rda` to be able to read it into μ-argus. This dataset contains microdata on members of households, with a household identifier to be able to detect the households.

Note that, for using household data in μ-argus it is essential that the records within the microdata file are grouped by the households. The easiest way to accomplish this is to sort the microdata file on the household identifying variable.

Prerequisites

TRS is applied targeting hierarchical regional variables. That is, at each level of the regional variable records may be swapped.

In order to be able to apply TRS, the microdata set needs to include the regional variable at each level as a separate variable. I.e., in case the hierarchy is COUNTY – DISTRICT – TOWN (going from coarse to more detailed), each record should contain three variables defining the county, the district and the town respectively.

TRS is applied on the level of households, so the microdata set needs to contain a household identifier. Moreover, the dataset needs to be sorted on that household identifier.

The variables that you want to use for TRS should not contain missing values and should be integer valued (category labels should be integers).

Known issues/restrictions:

- Currently only *k*-anonymity on individual records is used as risk model.
- You will get an error message when changing between risk models in the specify Combinations window (this is a known general μ-argus bug). This is filed as Issue #99 on https://github.com/sdcTools/UserSupport.
- There are no checks on validity of the used variables. I.e., it is not guaranteed that you will get error messages when the used variables are of incorrect format.
- In case non-integer variables are used in the TRS procedure, all values might be changed to zero's in the output (protected file).
- The manual is not yet updated.

Arguments

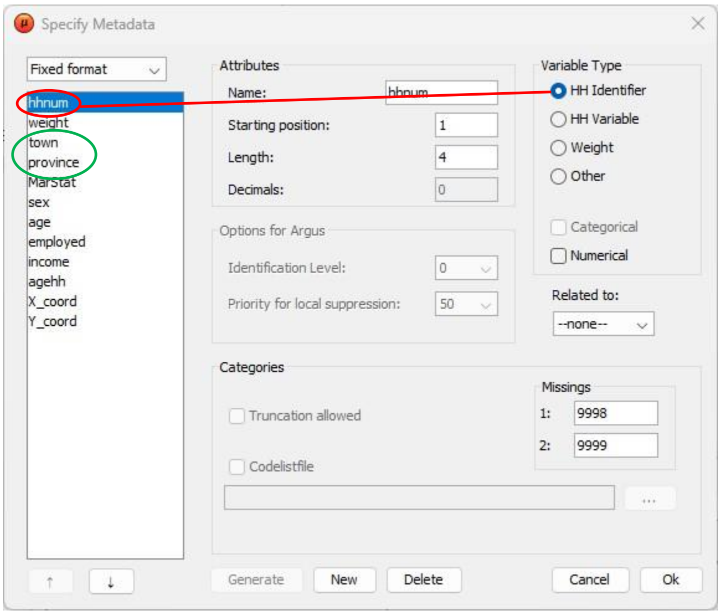
The variables that can be used in TRS are of four types:

1. (Similar) Variables that define which households are to be considered "similar" enough to be swapped over region.

2. (Hierarchy) Variables that define the regional hierarchy to be used.
3. (Risk) Variables that define the individual risk for each record.
4. (Carry) Variables that need to be swapped along with the region, e.g. variables that are connected to the regional variables like grid squares or X/Y-coordinates.

Walk through

1. Open microdata file and specify corresponding metadata (e.g. via the `.rda`-file).
   Note that in the example of the next figure, the hierarchy is defined in two levels using two variables: province and town (green selection). The household identifying variable is defined as well (red selection).



Specify Metadata

Fixed format

hhnum
weight
town
province
MarStat
sex
age
employed
income
agehh
X_coord
Y_coord

Attributes
Name: hhnum
Starting position: 1
Length: 4
Decimals: 0

Options for Argus
Identification Level: 0
Priority for local suppression: 50

Variable Type
● HH Identifier
○ HH Variable
○ Weight
○ Other

□ Categorical
□ Numerical

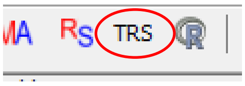
Related to: --none--

Categories
□ Truncation allowed
□ Codelistfile

Missings
1: 9998
2: 9999

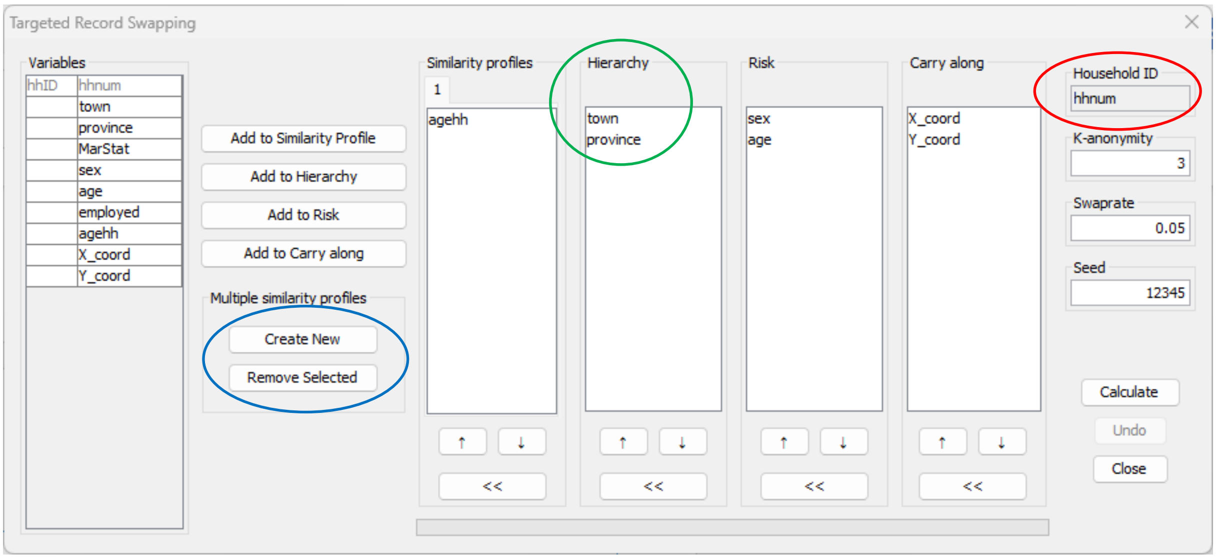
2. Specify Combinations. For the moment, just specify something. In a later version we will relate the risk-model used in the targeted record swapping to the risk model chosen in this window.

3. Use the menu-item “Modify/Targeted Record Swapping” or the button TRS



to open the dialog window for specifying the Targeted Record Swapping.

4. You will then get the following window:



Targeted Record Swapping

Variables

hhID	hhnum
	town
	province
	MarStat
	sex
	age
	employed
	agehh
	X_coord
	Y_coord

Similarity profiles: 1 — agehh

Hierarchy: town, province

Risk: sex, age

Carry along: X_coord, Y_coord

Household ID: hhnum

K-anonymity: 3

Swaprate: 0.05

Seed: 12345

Note that the household identifying variable is detected automatically (red selection in figure), since this was already defined in the metadata. It is not possible to change this selection of the household identifier.

Use the buttons so select the variables for Similar (used to define similar households, i.e., households that may be swapped), for Hierarchy (used to define the hierarchy in e.g. the regions when selecting records to swap with), for Risk (used to calculate the *k*-anonymity on, at individual-level) and for Carry along.

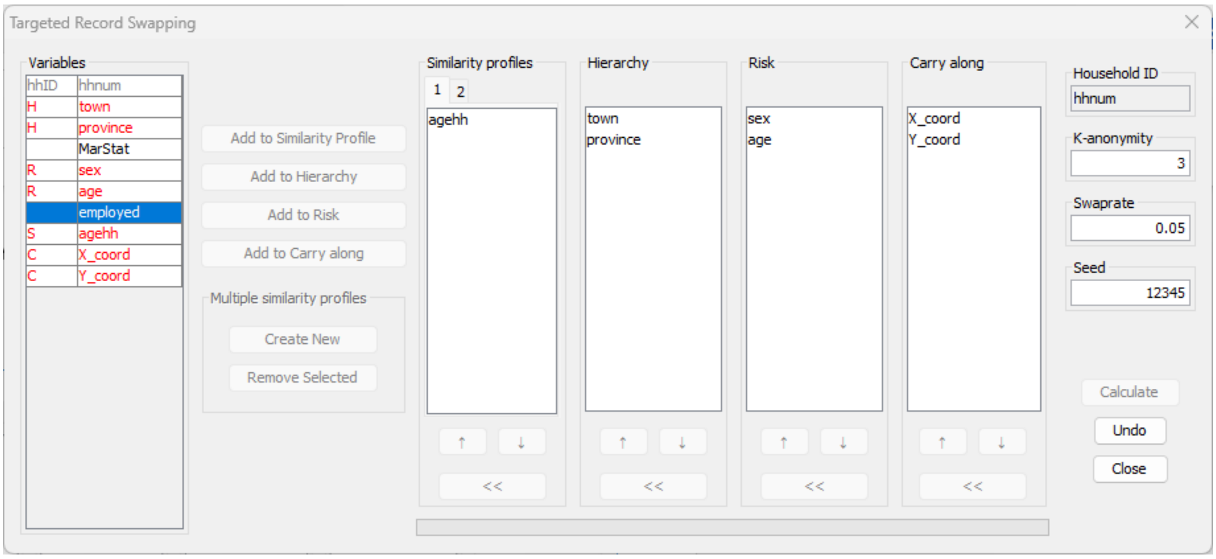
You can remove variables from a list, by selecting them and then using the respective << buttons.

The order in the variables for Hierarchy should be top down, i.e., the first variable should be the highest level and the last variable should be the most detailed level. In the picture: a province is a collection of towns, hence the top variable in the green selection in the previous figure is province, the last variable is town. You can change the order of these variables using the ↑ ↓ buttons.

You can specify multiple similarity profiles. You can add or remove similarity profiles by using the buttons in the blue circle in the above picture. If more than one profile is specified, TRS will look for donor households in the first profile. In case no donor households can be found, it will try the second similarity profile, and so on.

5. Specify the parameters to be used in the targeted record swapping: the threshold for the *k*-anonymity and the swaprate (a number between 0 and 1). The swaprate is the minimum fraction of records that will be swapped. You can also specify the seed to be used in the random number generator used to select records to swap with.

   Finally click “Calculate”. You will then see the following window (note that two similarity profiles are present):



Variables

hhID	hhnum
H	town
H	province
	MarStat
R	sex
R	age
	employed
S	agehh
C	X_coord
C	Y_coord

Similarity profiles (1, 2): agehh

Hierarchy: town, province

Risk: sex, age

Carry along: X_coord, Y_coord

Household ID: hhnum

K-anonymity: 3

Swaprate: 0.05

Seed: 12345

In the list of variables, the ones that are used are coloured red and have Info S if they are used as a Similar variable, H if used as a Hierarchy variable, R if used as a Risk variable and C if used as a Carry variable.
Then click OK to close the window.

6. Save the file using the menu item “Output/Make protected file”. When saving a file protected with the Targeted Record Swapping method, it is logical to select “No Suppression”.
7. In the report file you will find information on which variables are used in the Targeted Record Swapping method.

Features to be included in next versions:

- Allowing different risk models
    - Detecting the risk model from the Specify Combination window