Targeted Rank Swapping in C++

## Function Definitions

### *targetRankSwap*()

Applies targeted rank swapping on data set.

#### Input

* Data Input: Personfile as 2-dim array or numeric matrix containing households identifier, geographic regions and risk variables. Each column contains only numeric values, character columns need to be coded to integer. (this would make the C++ code easier, if we deal with this recoding outside of the actual function).
* Geographic Hierarchy: integer vector containing column indices for gegraphic hierarchy stored in data input. Order of indices defines gegraphic hierarchy. Is read from left to right meaning that the most left indices corresponds to the highest hierarchy level and the most right corresponds to the lowest hierarchy level.
* Risk Variables: integer vector containing column indices for risk variables stored in data input.
* Household ID: integer, index of column containing household identifier in data input.
* Swapping Rate: double between 0 and 1 determinig the proportion of swaps in each region.
* Seed: integer, set seed for random number generator

#### Output

* Data containing household identifier and **NEW** geographic levels. Household identifier is unique in the output. For example:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| HHID | GEO 1 | GEO 2 | GEO 3 | GEO 4 |
| 1 | 1 | 3 | 4 | 10 |
| 2 | 5 | 2 | 2 | 8 |
|  |  |  |  |  |
| n | 6 | 4 | 5 | 1 |

#### Question?

* Should we make this more generic, so not geographic variables, but something like to be swapped variables and work in the cells defined by these variables, which would lead to the same result for hierachical geographic variables?
* As an improvement the household ID step could be optional to be more generic, do we want to reflect that in the input right now already?

### *diagnostics*()

Runs diagnostics after targeted rank swapping (this might also be done as a default or optionally directly in the *targetRankSwap* function.

#### Input

* original Data Input: Personfile containing households identifier and geographic regions
* rank swapped Data: output of *targetRankSwap*().

#### Output

* Rectangular Table containing the complete geographic hierarchy as well as the number of households and persons and number of households and persons swapped. For example:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| GEO 1 | GEO 2 | GEO 3 | GEO 4 | N Households | N Persons | Households swapped | Persons swapped |
| 1 | 1 | 1 | 1 | 1000 | 4000 | 100 | 380 |
| 1 | 1 | 1 | 2 | 5000 | 10000 | 500 | 1200 |
| 1 | 1 | 1 | 3 | 20000 | 75000 | 2000 | 6200 |
|  |  |  |  |  |  |  |  |
| 6 | 10 | 15 | 20 | 2100 | 3800 | 210 | 350 |