Package 'Scalelink'

January 20, 2025

Type Package

Title Create Scale Linkage Scores

Version 1.0

Date 2019-02-05

Author Chris Charlton [aut, cre],
Harvey Goldstein [aut]

Maintainer Chris Charlton <c.charlton@bristol.ac.uk>

Depends R (>= 2.10.0)

Description
Perform a 'probabilistic' linkage of two data files using a scaling procedure using the methods described in Goldstein, H., Harron, K. and Cortina-Borja, M. (2017) <doi:10.1002/sim.7287>.

License GPL (>= 2)

Imports Rcpp (>= 0.12.9), RcppParallel

LinkingTo Rcpp, RcppParallel

SystemRequirements GNU make

Encoding UTF-8

RoxygenNote 6.1.1

NeedsCompilation yes

Repository CRAN

Date/Publication 2019-02-20 19:40:09 UTC

Contents

buildAstar *buildAstar*

Description

Builds the A* matrix

Usage

```
buildAstar(foinew, ldfnew, grainsize, debug)
```

Arguments

`foinew`	numeric `matrix` representing the file of interest
`ldfnew`	numeric `matrix` representing the linking data file
`grainsize`	integer determining minimum grain size for parallisation
`debug`	Boolean indicating whether to output additional debugging information

Details

`buildAstar` takes a matrix representing the file of interest and a matrix representing the linking data file and creates a matrix that can then be used to generating linking scores. Reporting frequency as this occurs can be specified via the nreport option. This is implemented in C++ to provide a speed increase over implementing it directly in the R equivalent.

calcScores *Calculates linking scores for a file of interest and linkage data file.*

Description

This function calculates a score from two files, the file of interest (FOI) and linkage data file (LDF).

Usage

```
calcScores(FOI, LDF, missing.value = NA, min.parallelblocksize = 1,
  output.varnames = NULL, debug = FALSE)
```

Arguments

`FOI`	A `data.frame` object, `matrix` or `vector` to be used as the file of interest. This must contain only the variables of interest, and these must be in the same order as the LDF.
`LDF`	A `data.frame` object, `matrix` or `vector` to be used as the linkage data file. This must contain only the variables of interest, and these must be in the same order as the FOI.

`missing.value`	Value used to represent missing data; Defaults to NA
`min.parallelblocksize`	The minimum block size when splitting up the data accross processors. You may wish to change this to optimise the allocation of processors. see (https://rcppcore.github.io/RcppParallel/#tuning).
`output.varnames`	Labels to apply to function output; Defaults to column names of the FOI `data.frame`
`debug`	Boolean indicating whether to output additional debugging information

Value

A list containing: An numeric `vector` of scores, one for each of the identifiers of interest and a `matrix` containing A*.

Author(s)

Goldstein H., and Charlton, C.M.J., (2017) Centre for Multilevel Modelling, University of Bristol.

`FOI` *File of interest*

Description

File of interest data data with 7742 records and 5 variables.

Format

A data frame with 7742 observations on the following 5 variables:

`id` Record Identifier (not used for linking).

`Day` Day of Week.

`Month` Month of Year.

`Year` Year.

`Sex` Gender: with codes `1` Male and `2` Female.

Details

The `FOI` dataset is one of the sample datasets provided with this package for demonstration purposes.

Source

Synthetic data created by Harvey Goldstein

Examples

```
data(FOI, package = "Scalelink")
summary(FOI)
```

LDF *Linking data file*

Description

Linking data file data with 10000 records and 5 variables.

Format

A data frame with 10000 observations on the following 5 variables:

`id` Record Identifier (not used for linking).

`Day` Day of Week.

`Month` Month of Year.

`Year` Year.

`Sex` Gender: with codes `1` Male and `2` Female.

Details

The `LDF` dataset is one of the sample datasets provided with this package for demonstration purposes. This version include records with missing data

Source

Synthetic data created by Harvey Goldstein

Examples

```
data(LDF, package = "Scalelink")
summary(LDF)
```

LDFCOMP *Linking data file*

Description

File of interest data data with 8142 records and 5 variables.

Format

A data frame with 8142 observations on the following 5 variables:

`id` Record Identifier (not used for linking).

`Day` Day of Week.

`Month` Month of Year.

`Year` Year.

`Sex` Gender: with codes `1` Male and `2` Female.

Details

The `LDFCOMP` dataset is one of the sample datasets provided with this package for demonstration purposes. This version has records containing missing data removed

Source

Synthetic data created by Harvey Goldstein

Examples

```
data(LDFCOMP, package = "Scalelink")
summary(LDFCOMP)
```

`Scalelink` *Record linkage via scaling algorithm*

Description

Scalelink is an R command to perform 'probabilistic' linkage of two data files using a scaling procedure.

Details

With increasing availability of large data sets derived from administrative and other sources, there is an increasing demand for the successful linking of these to provide rich sources of data for further analysis. Variation in the quality of identifiers used to carry out linkage means that existing approaches are often based upon 'probabilistic' models, which are based on a number of assumptions, and can make heavy computational demands. This package implements the method proposed in Goldstein, H., Harron, K. and Cortina-Borja, M. (2017). In this paper we suggest a new approach to classifying record pairs in linkage, based upon weights (scores) derived using a scaling algorithm. The proposed method does not rely on training data, is computationally fast, requires only moderate amounts of storage and has intuitive appeal.

References

Scalelink: Goldstein, H., Charlton, C.M.J. (2017) Scalelink: A Package to link data via scaling.

Paper: Goldstein, H., Harron, K. and Cortina-Borja, M. (2017). A scaling approach to record linkage. Statistics in Medicine. DOI: 10.1002/sim.7287

Maintainer

Chris Charlton <`c.charlton@bristol.ac.uk`>

Author(s)

Charlton, C.M.J., Goldstein H (2017) Centre for Multilevel Modelling, University of Bristol.

Examples

```
library(Scalelink)

## Set the number of CPU cores to use (omit to use all available)
RcppParallel::setThreadOptions(numThreads = 2)

data(FOI, package = "Scalelink")
data(LDFCOMP, package = "Scalelink")

idcols <- c("Day", "Month", "Year", "Sex")
result <- calcScores(FOI[, idcols], LDFCOMP[, idcols])

print(result$scores)

## Scalelink package provides two examples using synthetic data
## one with complete data and one containing missing values

## Not run:
## For a list of demo titles
demo(package = 'Scalelink')

## To run a demo
demo(Example1)

## Using your own data
##If you had the following files in your working directory:
##FOI:
##A space-delimited file of interest (NFOI x PFOI). NFOI is number of records
##IDENTIFIERS_FOI:
##A space-delimited file containing a row vector length PFOI with a 1 where it is an identifier
##LDF:
##A space-delimited linking data file (NLDF x PLDF). NLDF is number of records
##IDENTIFIERS_LDF:
##A space-delimited file containing a row vector length PLDF with a 1 where it is an identifier

##Then you can calculate scores as follows:
FOI<-read.table("FOI")
LDF<-read.table("LDF")
IDENTIFIERS_FOI<-read.table('IDENTIFIERS_FOI')
IDENTIFIERS_LDF<-read.table('IDENTIFIERS_LDF')
result <- calcScores(FOI[, which(IDENTIFIERS_FOI == 1)], LDF[, which(IDENTIFIERS_LDF == 1)],
missing.value=-9.999e+029)

##To view the scores:
print(round(result$scores, 2))

##To view the A* matrix:
print(result$astar)

## End(Not run)
```

Index