Package 'Rmtcd'

October 14, 2015

Type Package

Title Synthesizing Truncated Count Data

Version 1.0

Date 2015-07-26

Description To maintain confidentiality, national statistical agencies traditionally do not include small counts in publicly released tabular data products. They typically delete these small counts, or combine them with counts in adjacent table cells to preserve the totals at higher levels of aggregation. In some cases these suppression procedures result in too much loss of information. To increase data utility and make more data publicly available, Hawala et. al. created methods and software (in C++) to generate synthetic values for the small counts from a Bayesian hierarchical model. The software generates synthetic data and computes several measures of disclosure risk, and was applied by the Census Bureau in synthesizing small county-to-county migration counts. This package provides an R-interface for implementing the code.
The original source C++ model is left unchanged aside from minor additions for R-compatibility. All inputs
and outputs from the model are R objects, as described in the documentation for [link here].

License Creative Commons Attribution-NonCommercial-ShareAlike (>= 4) | file LICENSE

Imports Rcpp (>= 0.11.6)

LinkingTo Rcpp

R topics documented:

`mtcdForEntireDataFrame` *Create Synthetic Dataset*

Description

Returns synthetic datasets from a Bayesian hierarchical model. This function wraps C++ code proposed in "Synthesizing Truncated Count Data for Confidentiality," and developed by Sam Hawala, Jerry Reiter and Quanli Wang. References can be found in the package description.

Usage

```
mtcdForEntireDataFrame(df, seed = 0, niters = 1e+05, burnin = 30000,
  stride = 500, numModel = 10)
```

Arguments

`df`	A dataframe with three columns, the last column containing count to be synthesized, the first two must be numeric, numbers that uniquely identify counties
`seed`	positive integer for random number generation
`niters`	positive integer, number of iterations for parameter generation
`burnin`	a positive integer specifying the number of burnins(iterations of parameters will not be saved)
`stride`	the model will save a model(set of parameters) for every stride number of models
`numModel`	a positive integer specifying the number of synthetic datasets to be generated

Value

A dataframe containing the synthetic datasets and disclosure risk measures

Details

The first three columns of output are the original dataset. The next numModel columns are the synthetic datasets. The next two colunms give the 95 percent confidence intevals estimated using all saved models. The following columns give the dt, Rall and Runq risk measurements.

See Also

`Rmtcd`

Examples

```
setwd("00_pkg_src/Rmtcd/test")
read.table("data.txt") -> dataFrame
mtcdForEntireDataFrame(dataFrame, numModel = 5)
```

`mtcdForSmallCount` *Create Synthetic Dataset*

Description

Returns synthetic datasets from a Bayesian hierarchical model. This function wraps C++ code proposed in "Synthesizing Truncated Count Data for Confidentiality," and developed by Sam Hawala, Jerry Reiter and Quanli Wang. References can be found in the package description.

Usage

```
mtcdForSmallCount(df, seed = 0, niters = 1e+05, burnin = 30000,
  stride = 500, numModel = 10, upperLimit = 9)
```

Arguments

`df` A dataframe with three columns, the last column containing count to be synthesized, the first two must be numeric, numbers that uniquely identify counties

`seed` positive integer for random number generation

`niters` positive integer, number of iterations for parameter generation

`burnin` a positive integer specifying the number of burnins(iterations of parameters will not be saved)

`stride` the model will save a model(set of parameters) for every stride number of models

`numModel` a positive integer specifying the number of synthetic datasets to be generated

`upperLimit` a positive defining the "small count" threshold

Value

A dataframe containing the synthetic datasets and disclosure risk measures

Details

The first three columns of output are the original dataset. The next numModel columns are the synthetic datasets(note that for large counts, the synthetic data is just the original data). The next two colunms give the 95 percent confidence intevals estimated using all saved models. The following columns give the dt, Rall and Runq risk measurements. (for large counts, columns after the synthetic datasets do not exist)

Examples

```
read.table("data.txt") -> dataFrame
mtcdForSmallCount(dataFrame, numModel = 5, upperLimit = 15)
```

`Rmtcd` *Synthesizing Truncated Count Data*

Description

This package provides an R wrapper function for the statistical model proposed in "Synthesizing Truncated Count Data for Confidentiality," originally developed by Sam Hawala, Jerry Reiter and Quanli Wang (2014). Their C++ code served as the basis, and can be found at http://sites.duke.edu/tcrn/research-projects/downloadable-software/

Background

To maintain confidentiality, national statistical agencies traditionally do not include small counts in publicly released tabular data products. They typically delete these small counts, or combine them with counts in adjacent table cells to preserve the totals at higher levels of aggregation. In some cases these suppression procedures result in too much loss of information. To increase data utility and make more data publicly available, Hawala et. al. created methods and software (in C++) to generate synthetic values for the small counts from a Bayesian hierarchical model. The software generates synthetic data and computes several measures of disclosure risk, and was applied by the Census Bureau in synthesizing small county-to-county migration counts.

This package provides an R-interface for implementing the code. The original source C++ model is left unchanged aside from minor additions for R-compatibility. All inputs and outputs from the model are R objects, as described in the documentation for [link here].

Functions

This package contains two functions: mtcdForEntireDataFrame and mtcdForSmallCount. mtcdForEntireDataFrame takes in a data frame and generate synthetic data for each row in the data frame, while mtcdForSmallCount allows users to specify an upper limit that defines "small count", and only synthesize counts that are smaller than or equal to the specified number.

Author(s)

In order of contribution: Yuxin (Charley) Chen (<`yc769@cornell.edu`>), Hautahi Kingi (<`hrk55@cornell.edu`>), Alice Chou (<`aec247@cornell.edu`>), Lars Vilhuber (<`lars.vilhuber@cornell.edu`>). Please direct queries to the Labor Dynamics Institute (<`ldi@cornell.edu`>)

References

The original source code in C++ is available at `http://sites.duke.edu/tcrn/research-projects/downloadable-software/` This R package is maintained at `https://github.com/ncrncornell/Rmtcd`

Development of the original code was supported under the NSF grant SES-1131897 (NCRN Duke-NISS, `http://sites.duke.edu/tcrn/`). Development of the R wrapper was supported by the NSF Census Research Network grant #1131848 to Cornell University (`https://www.ncrn.cornell.edu/`)

Index