Statistical Disclosure Control

- Official Statistics and Survey Methodology
- Numerous Topics including:
- Martin Templ (Univ. of Vienna)
- The **sdcMicro** package
- (R Conference in Romania)

Official Statistics

> *Official statistics are statistics published by government agencies or other public bodies such as international organizations. They provide quantitative or qualitative information on all major areas of* **citizens' lives**, *such as economic and social development, living conditions,health, education, and the environment.*
> *(Wikipedia)*

- Remark: Are we talking about data about individual people?

Official Statistics

”Personally identifiable information” (PII), as used in US privacy law and information security, is information that can be used on its own or with other information to identify, contact, or locate a single person, or to identify an individual in context.

Official Statistics

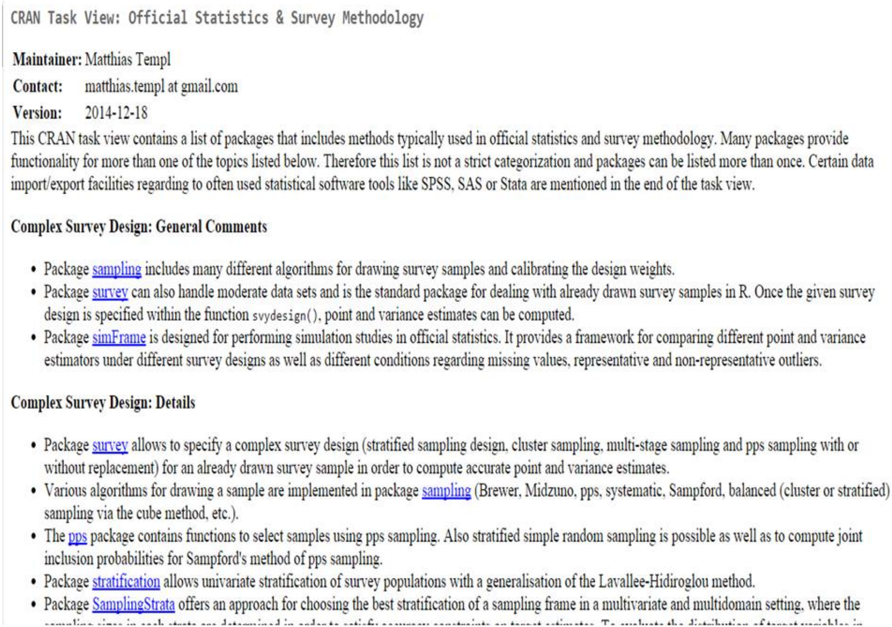
> *PII is "any information about an individual maintained by an agency, including*
>
> *(1) any information that can be used to distinguish or trace an individuals identity, such as name, social security number, date and place of birth, mothers maiden name, or biometric records; and*
>
> *(2) any other information that is linked or linkable to an individual, such as medical, educational, financial, and employment information."*

(NIST Special Publication 800-122)

-
- **De-anonymization** is the reverse process in which anonymous data is cross-referenced with other data sources to re-identify the anonymous data source.

Celebrities since 1950 (may be deceased)

- US, Born in 1935, identical twin, other twin deceased.
- US, Redhaired, 6"4'
- UK, Greek Cypriot Heritage, Convert to Islam
- UK, Female, amputation of left leg below the knee
- Irish, Male, Blond-haired, Mullingar
- USA, Married to UK based Lawyer.
- UK, Same Sex Marriage, 27 years older than husband.
- Mexican, Irish Heritage



CRAN Task View: Official Statistics & Survey Methodology

Maintainer: Matthias Templ

Contact: matthias.templ at gmail.com

Version: 2014-12-18

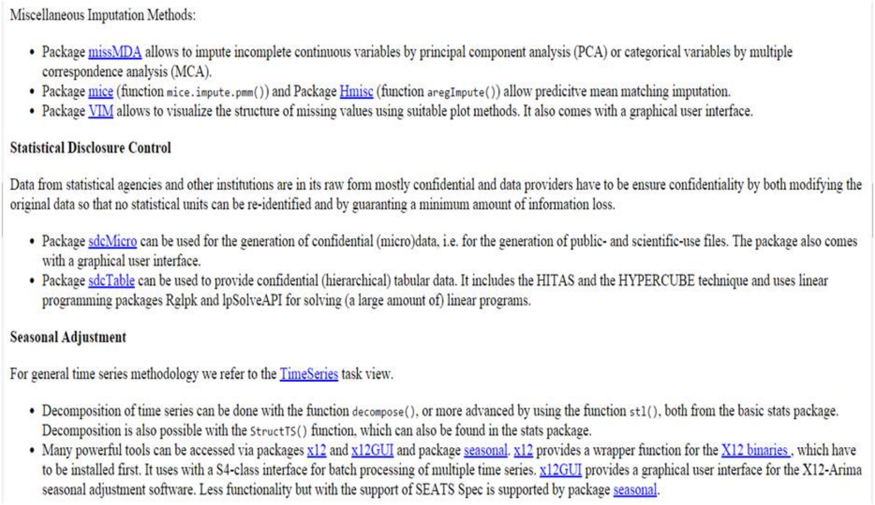
This CRAN task view contains a list of packages that includes methods typically used in official statistics and survey methodology. Many packages provide functionality for more than one of the topics listed below. Therefore this list is not a strict categorization and packages can be listed more than once. Certain data import/export facilities regarding to often used statistical software tools like SPSS, SAS or Stata are mentioned in the end of the task view.

Complex Survey Design: General Comments

- Package sampling includes many different algorithms for drawing survey samples and calibrating the design weights.
- Package survey can also handle moderate data sets and is the standard package for dealing with already drawn survey samples in R. Once the given survey design is specified within the function `svydesign()`, point and variance estimates can be computed.
- Package simFrame is designed for performing simulation studies in official statistics. It provides a framework for comparing different point and variance estimators under different survey designs as well as different conditions regarding missing values, representative and non-representative outliers.

Complex Survey Design: Details

- Package survey allows to specify a complex survey design (stratified sampling design, cluster sampling, multi-stage sampling and pps sampling with or without replacement) for an already drawn survey sample in order to compute accurate point and variance estimates.
- Various algorithms for drawing a sample are implemented in package sampling (Brewer, Midzuno, pps, systematic, Sampford, balanced (cluster or stratified) sampling via the cube method, etc.).
- The pps package contains functions to select samples using pps sampling. Also stratified simple random sampling is possible as well as to compute joint inclusion probabilities for Sampford's method of pps sampling.
- Package stratification allows univariate stratification of survey populations with a generalisation of the Lavallee-Hidiroglou method.
- Package SamplingStrata offers an approach for choosing the best stratification of a sampling frame in a multivariate and multidomain setting, where the sampling sizes in each strata are determined in order to satisfy accuracy constraints on target estimates. To evaluate the distribution of target variables in



Miscellaneous Imputation Methods:

- Package missMDA allows to impute incomplete continuous variables by principal component analysis (PCA) or categorical variables by multiple correspondence analysis (MCA).
- Package mice (function `mice.impute.pmm()`) and Package Hmisc (function `aregImpute()`) allow predicitve mean matching imputation.
- Package VIM allows to visualize the structure of missing values using suitable plot methods. It also comes with a graphical user interface.

Statistical Disclosure Control

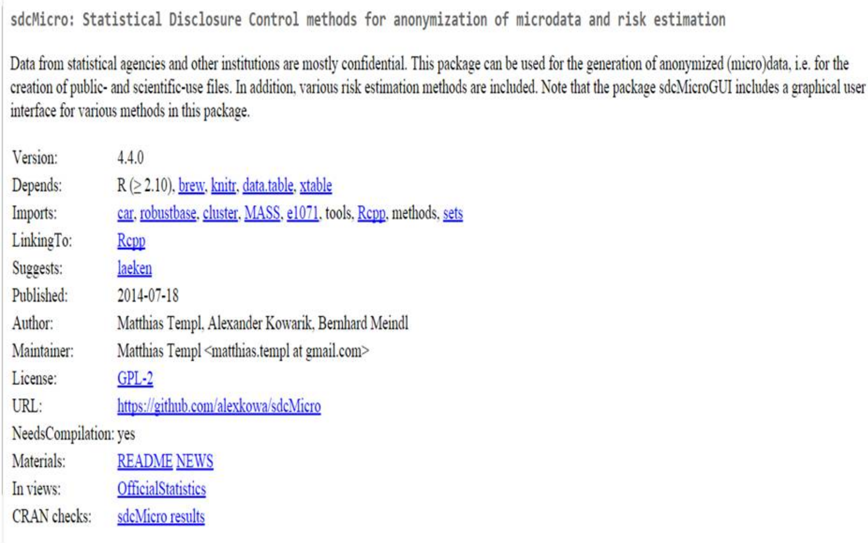
Data from statistical agencies and other institutions are in its raw form mostly confidential and data providers have to be ensure confidentiality by both modifying the original data so that no statistical units can be re-identified and by guaranting a minimum amount of information loss.

- Package sdcMicro can be used for the generation of confidential (micro)data, i.e. for the generation of public- and scientific-use files. The package also comes with a graphical user interface.
- Package sdcTable can be used to provide confidential (hierarchical) tabular data. It includes the HITAS and the HYPERCUBE technique and uses linear programming packages Rglpk and lpSolveAPI for solving (a large amount of) linear programs.

Seasonal Adjustment

For general time series methodology we refer to the TimeSeries task view.

- Decomposition of time series can be done with the function `decompose()`, or more advanced by using the function `stl()`, both from the basic stats package. Decomposition is also possible with the `StructTS()` function, which can also be found in the stats package.
- Many powerful tools can be accessed via packages x12 and x12GUI and package seasonal. x12 provides a wrapper function for the X12 binaries , which have to be installed first. It uses with a S4-class interface for batch processing of multiple time series. x12GUI provides a graphical user interface for the X12-Arima seasonal adjustment software. Less functionality but with the support of SEATS Spec is supported by package seasonal.



sdcMicro: Statistical Disclosure Control methods for anonymization of microdata and risk estimation

Data from statistical agencies and other institutions are mostly confidential. This package can be used for the generation of anonymized (micro)data, i.e. for the creation of public- and scientific-use files. In addition, various risk estimation methods are included. Note that the package sdcMicroGUI includes a graphical user interface for various methods in this package.

Version:	4.4.0
Depends:	R (≥ 2.10), brew, knitr, data.table, xtable
Imports:	car, robustbase, cluster, MASS, e1071, tools, Rcpp, methods, sets
LinkingTo:	Rcpp
Suggests:	laeken
Published:	2014-07-18
Author:	Matthias Templ, Alexander Kowarik, Bernhard Meindl
Maintainer:	Matthias Templ <matthias.templ at gmail.com>
License:	GPL-2
URL:	https://github.com/alexkowa/sdcMicro
NeedsCompilation:	yes
Materials:	README NEWS
In views:	OfficialStatistics
CRAN checks:	sdcMicro results

Mag. Bernhard Meindl
DI Alexander Kowarik
Priv.-Doz. Dr. Matthias Templ

office@data-analysis.at

data-analysis OG

Introduction to Statistical Disclosure Control (SDC)

Authors:

Matthias Templ, Bernhard Meindl and Alexander Kowarik
http://www.data-analysis.at

Vienna, July 18, 2014

Acknowledgement: International Household Survey Network
(IHSN)*

Statistical Disclosure Control