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

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)*

Microdata

MicroData

- A microdata file is a dataset that holds information collected on individual units; examples of units include people, households or enterprises.
- For each unit, a set of variables is recorded and available in the dataset.

Statistical Disclosure Control - Concepts

(1) A **microdata file** is a dataset that holds information collected on individual units; examples of units include people, households or enterprises.
For each unit, a set of variables is recorded and available in the dataset.

(2) A **disclosure risk** occurs if an unacceptably narrow estimation of a respondents confidential information is possible or if exact disclosure is possible with a high level of confidence.

Identity disclosure:

- In this case, the intruder associates an individual with a released data record that contains sensitive information, i.e. linkage with ex- ternal available data is possible.
- Identity disclosure is possible through direct identifiers, rare combinations of values in the key variables and exact knowledge of continuous key variable values in external databases.
- For the latter, extreme data values (e.g., extremely high turnover values for an enterprise) lead to high re-identification risks, i.e. it is likely that responends with ex- treme data values are disclosed.

Attribute disclosure:

- In this case, the intruder is able to determine some characteristics of an individual based on information available in the released data.
- For example, if all people aged 56 to GO who identify their race as black in region 12345 are unemployed, the intruder can determine the value of the variable labor status.

Inferential disclosure:

- In this case, the intruder, though with some uncertainty, can predict the value of some characteristics of an individual more accu- rately with the released data.

Statistical Disclosure Control - Concepts

- Removing and encrypting are only acceptable ways for disclosure control of identifier values.
- For another types of attributes we can apply masking methods.
- They can in turn be divided on two groups depending on their effect on the original data .

Perturbative methods. The original microfile is distorted, but in such a way that a difference between values of statistical rates of masking and original data would be acceptable.

Non-perturbative methods. They protect data without altering them. Non-perturbative methods base on principles of suppression and generalization (recoding).

- Suppression is a removing some data from original set.
- Recoding is a data enlargement.
- In case of continuous data (for example, age) non-perturbative methods can transform it to interval, fuzzy or categorical datatype.

Sometimes methods generating synthetic data or combined methods are also applied. Such methods are not acceptable at microdata preparation, because if we dont know the aim of the follow-up analysis, we can not generate adequate set of **synthetic data**.

Types of variables

In accordance with disclosure risks, variables can be classied into three groups, which are not necessarily exclusive:

Direct Identifiers are variables that precisely identify statistical units.

- For example, social insurance numbers, names of companies or persons and addresses are direct identiers
- (Remark: Primary Keys)

Types of variables

Key variables are a set of variables that, when considered together, can be used to identify individual units. For example, it may be possible to identify individuals by using a combination of variables such as gender, age, region and occupation. Other examples of key variables are income, health status, nationality or political preferences.

Types of variables

- Key variables are also called implicit identiers or quasi-identifiers.
- When discussing SDC methods, it is preferable to distinguish between categorical and continuous key variables based on the scale of the corresponding variables.
- Non-identifying variables are variables that are not direct identiers or key variables.

Types of Disclosure

- In general, disclosure occurs when an intruder uses the released data to reveal previously unknown information about a respondent.
- There are three different types of disclosure: *(next set of slides)*

Types of Disclosure

Identity disclosure:

- In this case, the intruder associates an individual with a released data record that contains sensitive information, i.e. linkage with external available data is possible.
- Identity disclosure is possible through direct identfiers, rare combinations of values in the key variables and exact knowledge of continuous key variable values in external databases.
- For the latter, extreme data values lead to high re-identification risks, i.e. it is likely that responends with extreme data values are disclosed.

Types of Disclosure

Attribute disclosure:

- In this case, the intruder is able to determine some characteristics of an individual based on information available in the released data.
- For example, if all people aged 56 to GO who identify their race as black in region 12345 are unemployed, the intruder can determine the value of the variable labor status.
- **Also** Diet - detailed questionnaire on diet can give clues to other aspects of some people's lives.

Types of Disclosure

Inferential disclosure:

- In this case, the intruder, though with some uncertainty, can predict the value of some characteristics of an individual more accurately with the released data.

Linkage

Linkage

- If linkage is successful based on a number of identifiers, the intruder will have access to all of the information related to a specific corresponding unit in the released data.
- This means that a subset of critical variables can be exploited to disclose everything about a unit in the dataset.

sdcMicro

- The R package **sdcMicro** is a well-known collection of microdata protection methods developed by **Statistics Austria**, which is already in use in several national statistics offices.
- **sdcMicro** has become one of the standard tools for **microdata protection** during the last five years.
- The IHSN is supporting the further development of sdcMicro and has partnered with its developers to perform the following tasks (next slide).

- Version 4.4.0 of the sdcMicro package is available on the Comprehensive R Archive Network (CRAN).
- Existing guidelines and a user guide for **sdcMicro** are being updated.
- A specific tutorial is being developed to show how to implement these concepts and algorithms on real datasets (see Vignettes)

R-Package sdcMicro and sdcMicroGUI

- SDC methods introduced in this guideline can be implemented by the R-Package sdcMicro.
- Users who are not familiar with the native R command line interface can use sdcMicroGUI, an easy-to-use and interactive application.

SDC Methods

Mathematical Foundation of SDC

- In general, SDC methods borrow techniques from other elds.
- For instance, multivariate (robust) statistics are used to modify or simulate continuous variables and to quantify information loss.
- Distribution-fitting methods are used to quantify disclosure risks.
- Statistical modeling methods form the basis of perturbation algorithms, to simulate synthetic data, to quantify risk and information loss.
- Linear programming is used to modify data but minimize the impact on data quality.

SDC Methods

- Problems and challenges arise from large datasets and the need for efcient algorithms and implementations.
  (**Remark** 1000 variables+)
- Another layer of complexity is produced by complex structures of hierarchical, multidimensional data sampled with complex survey designs.
  (**Remark** Rooted Questions)
- Missing values are a challenge, especially for computation time issues; structural Zeros (values that are by denition Zero) also have impact on the application of SDC methods.

SDC Methods

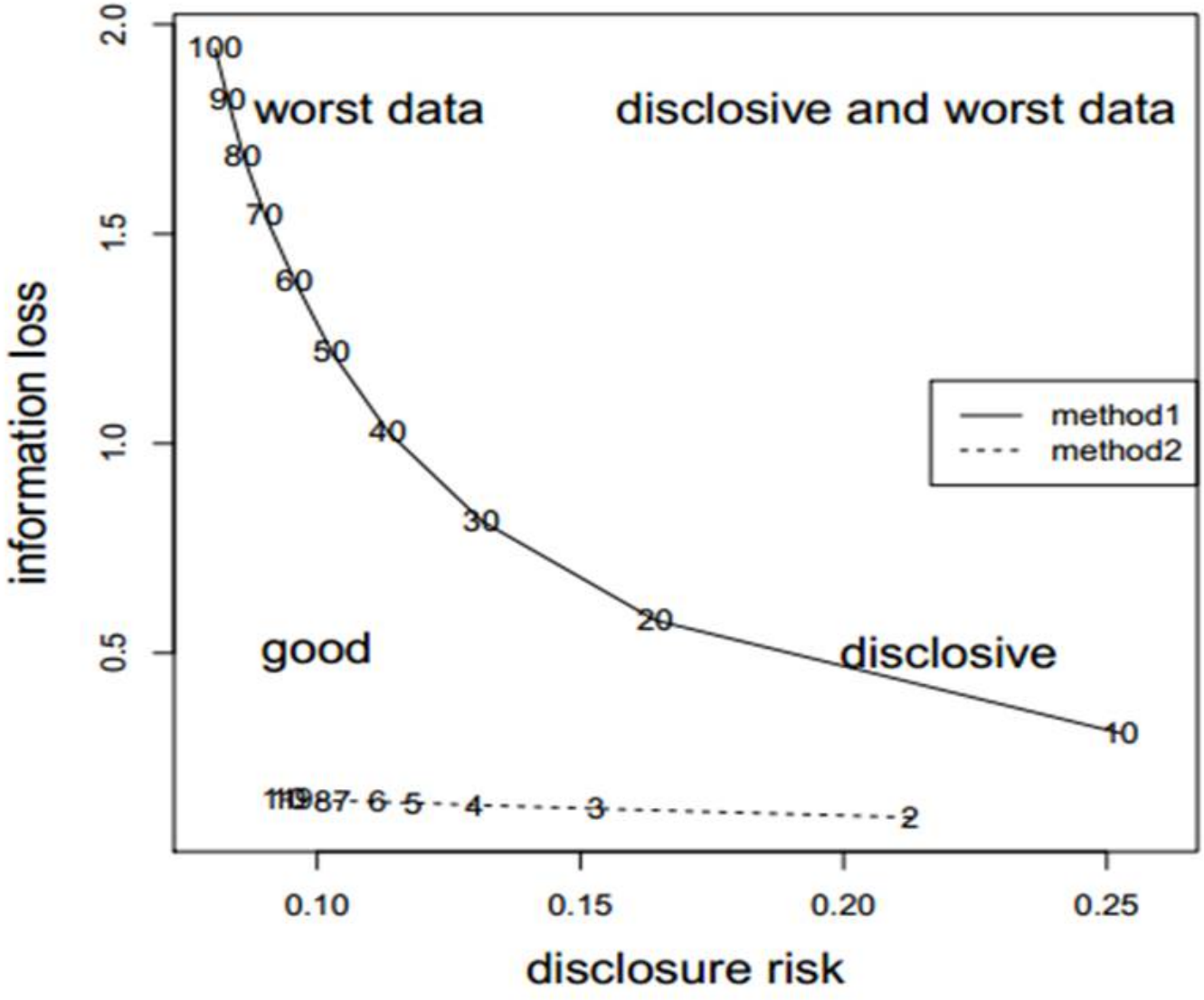
Furthermore, the compositional nature of many components should always be considered, but adds even more complexity.

SDC techniques can be divided into three broad topics:

- Measuring disclosure risk
- Methods to anonymize micro-data
- Comparing original and modied data (information loss)

Risk Versus Data Utility

The goal of SDC is always to release a safe micro dataset with high data utility and a low risk of linking condential information to individual respondents. Figure 1 shows the trade-off between disclosure risk and data utility.


information loss
2.0
1.5
1.0
0.5
100
90
80
70
60
50
40
30
20
10
worst data
disclosive and worst data
good
disclosive
method1
method2
0.10
0.15
0.20
0.25
disclosure risk

Risk Versus Data Utility

- For Method 1 (in this example adding noise), the parameter varies between 10 (small perturbation) to 100 (perturbation is 10 times higher).
- When the parameter value is 100, the disclosure risk is low since the data are heavily perturbed, but the information loss is very high, which also corresponds to very low data utility.

Risk Versus Data Utility

- When only low perturbation is applied to a dataset, both risk and data utility are high.
- It is easy to see that data anonymized with Method 2 (we used microaggregation with different aggregation levels) have considerably lower risk; therefore, this method is preferable.

Risk Versus Data Utility

- In addition, information loss increases only slightly if the parameter value increases; therefore, Method 2 with parameter value of approximately 7 would be a good choice in this case since this provides both, low disclosure risk and low information loss.

Risk Versus Data Utility

- For higher values, the perturbation is higher but the gain is only minimal, lower values reports higher disclosure risk, Method 1 should not be chosen since the disclosure risk and the information loss is higher than for method 2.
- However, if for some reasons method 1 is chosen, the parameter for perturbation might be chosen around 40 if 0.1 risk is already considered to be safe.
- For data sets concerning very sensible information (like cancer) the might be, however, to high risk and a perturbation value of 100 or above should then be chosen for method 1 and a parameter value above 10 might be chosen for method

Considerations

- In real-world examples, things are often not as clear, so data anonymization specialists should base their decisions regarding risk and data utility on the following considerations:
- What is the legal situation regarding data privacy? Laws on data privacy vary between countries; some have quite restrictive laws, some dont, and laws often differ for different kinds of data (e.g., business statistics, labor force statistics, social statistics, and medical data).

- How sensitive is the data information and who has access to the anonymized data file? Usually, laws consider two kinds of data users: users from universities and other research organizations, and general users, i.e., the public. In the rst case, special contracts are often made between data users and data producers.
- Usually these contracts restrict the usage of the data to very specic purposes, and allow data saving only within safe work environments. For these users, anonymized microdata les are called scientic use les, whereas data for the public are called public use les. Of course, the disclosure risk of a public use ile needs to be very low, much lower than the corresponding risks in scientic use les. For scientic use les, data utility is typically considerably higher than data utility of public use les.

- Another aspect that must be considered is the sensitivity of the dataset. Data 011 individuals medical treatments are more sensitive than an establishments turnover values and number of employees. If the data contains very sensitive information, the microdata should have greater security than data that only contain information that is not likely to be attacked by intruders.

- Which method is suitable for which purpose? Methods for Statistical Disclosure Control always imply to remove or to modify selected variables.
- **Key** The data utility is reduced in exchange of more protection.
- While the application of some specific methods results in low disclosure risk and large information loss, other methods may provide data with acceptable, low disclosure risks.

Recommendations (or lack thereof)

- General recomendations can not be given here since the strenghtness and weakness of methods depends on the underlying data set used.
- Decisions on which variables will be modied and which method is to be used result are partly arbitrary and partly result from a prior knowledge of what the users will do with the data.

- Generally, when having only few categorical key variables in the data set, recoding and local suppression to achieve low disclosure risk for categorical key variables is recommended.
- In addition, in case of continous scaled key variables, microaggregation is easy to apply and to understand and gives good results.
- For more experienced users, shuffling may often give the best results as long a strong relationship between the key variables to other variables in the data set is present.

- In case of many categorical key variables, postrandomization might be applied to several of these variables. Still methods, such as postrandomization (PRAM), may provide high or low disclosure risks and data utility, depending on the specic choice of parameter values, e.g. the swapping rate.

- Beside these recommendations, in any case, data holders should always estimate the disclosure risk for their original datasets as well as the disclosure risks and data utility for anonyrnized versions of the data.
- To achieve good results (i.e., low disclosure risk, high data utility), it is necessary to anonyrnize in an explanatory manner by applying different methods using different parameter settings until a suitable trade-off between risk and data utility has been achieved.