Statistical disclosure control and micro data

Methods

Statistics Netherlands

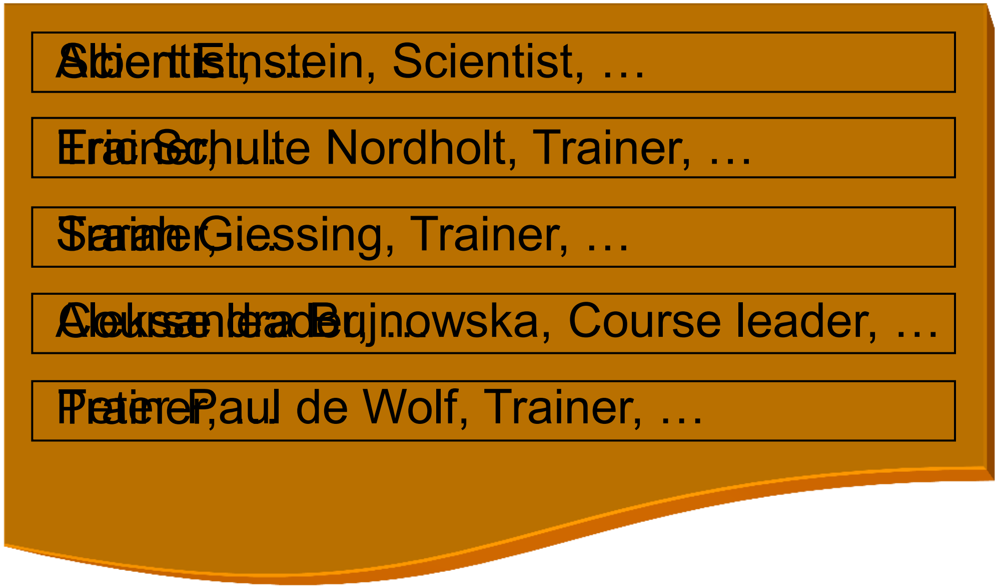
Contents

- SDC-methods
  - (sub)sampling
  - data swapping
  - Categorical variables
    - recoding, suppression, PRAM
  - Continuous variables
    - top/bottom coding, micro-aggregation, noise addition
- Miscellaneous topics

SDC methods

First thing to do

Remove direct/formal identifiers



Albert Einstein, Scientist, …

Eric Schulte Nordholt, Trainer, …

Sarah Giessing, Trainer, …

Aleksandra Bujnowska, Course leader, …

Peter-Paul de Wolf, Trainer, …

Scientist, …

Trainer, …

Trainer, …

Course leader, …

Trainer, …

SDC methods

(Sub)sampling

Release only a sample of the records

- Reduces the effect of response knowledge on original sample
- Extreme local suppression: suppress all values in certain (stochastic) set of records

NB:

effect on sampling weights?

SDC methods

data swapping

Eric, Course leader, Stat. Neth., …
Eric, Trainer, Stat. Netherlands, …

⋮

Aleksandra, Course leader, Eurostat, …
Aleksandra, Trainer, Eurostat, …

Select two records *i* and *j*

Interchange (‘swap’) scores on variable(s) of records *i* and *j*

Several *selection schemes* possible

- (Approximately) preserving certain statistics (e.g., up to *p*-th order interactions)
- Random (SRS, STSI, …)
- Rank swapping (in μ-ARGUS)

SDC methods

data swapping (example)

records consisting of three parts:

x : defines geographic area

y : household characteristics

(relating number of persons in household, race, age)

z : all other variables

Assumption:

x and z conditionally independent, given y

Swap households with same y between areas

SDC methods

Categorical (recoding)

Usually *global* recoding

- combine certain categories to new category
- apply this to entire data set

Goal:

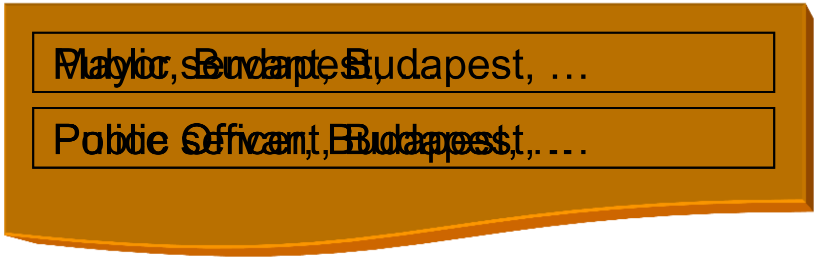
- increase (population) frequency

Example:

Occupations ‘Mayor’ and ‘Police Officer’

recoded into

Occupation ‘Public servant’



Mayor, Budapest, …

Public servant, Budapest, …

Police Officer, Budapest, …

Public servant, Budapest, …

SDC methods

Categorical (suppression)

Local suppression

- Replace score by ‘missing’
- Applied to one record at a time

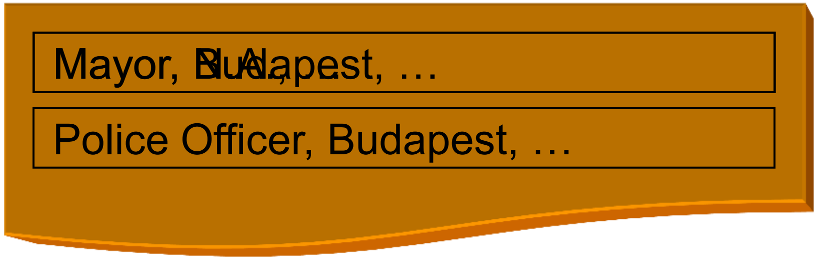
Example:

record 1: Mayor, Budapest

Not Available (Missing)

record 2: Police Officer, Budapest

Officer, Budapest



Mayor, Budapest, …

N.A.,

Police Officer, Budapest, …

SDC methods

Categorical (suppression)

How to choose variables to be suppressed?

Multiple unsafe combinations in one record

E.g., Mayor × Budapest (work)

and

Budapest (work) × Mayor's residence

'Entropy' (number of categories/information loss):

suppress Budapest

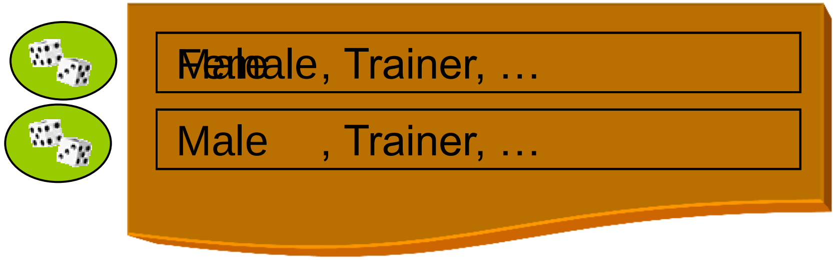
'Priority/weight':

suppress Mayor *and* Mayor's residence

SDC methods

Categorical (PRAM)

Post Randomisation Method



Female, Trainer, …

Male , Trainer, …

Categorical variable ξ with categories 1, …, K

Define transition probabilities $p_{kl} = P(X = l \mid \xi = k)$

I.e., Markov matrix P with p_{kl} as entries.

PRAM: the score m on ξ is replaced by a score drawn from the distribution p_{m1}, …, p_{mK}.

(for each record independently)

SDC methods

Categorical (PRAM)

Since P is known, correction is possible

Compare Randomised Response or Misclassification

E.g., T_ξ is original frequency table of ξ,

T_X is frequency table after PRAM

Then

$$E(T_X \mid \xi) = P^t T_\xi$$

I.e.,

$$(P^{-1})^t T_X$$

is (conditionally) unbiased estimator of T_ξ

SDC methods

Categorical (PRAM)

Variable Gender (male = 1, female = 2)

$p(1,1) = p(2,2) = 0.9$

$p(1,2) = p(2,1) = 0.1$

Original file: $T_{\xi} = (110, 90)$

Perturbed file: $T_X = (107, 93)$ (in expectation: (108, 92))

Unbiased estimate: $(P^{-1})^t T_X = (108.75, 91.25)$

rounded: (109, 91)

SDC methods

Categorical (PRAM)

How to choose *P* ?

- Try to preserve certain statistics (in expectation)
- Exclude illogical changes

  e.g., set transition probability

  unmarried + age < 5

  to

  married + age < 5

  equal to 0
- Make sure that perturbed file is ‘safe’

SDC methods

Categorical (PRAM)

Remarks:

- Every application of PRAM produces different file
- Possible to adjust analyses (burden to user)
- In μ-ARGUS only limited possibilities for *P*
  - Off-diagonal all equal
  - Band-matrix

SDC methods

Continuous

Exact values (usually) not known to attacker

Partition variable into classes

Treat partitioned variable as categorical

SDC methods

Continuous

Example

Age: exact age not known

partition into 5-years classes
if 5-years class occurs often enough: Safe
if not: Not safe and hence ‘do something’ (suppress)

SDC methods

Continuous (top/bottom coding)

Extreme scores may be identifying

Example: income

Possible method:

- Replace all scores above/below certain threshold with that threshold

SDC methods

Continuous (top/bottom coding)

(Top coding)

Estimate the probability of occurrence of a value above a certain threshold

Deduce the expected number $\hat{N}$ of occurrences (in population) above that threshold

Choose threshold such that $\hat{N}$ is ‘large’ enough

SDC methods

Continuous (micro-aggregation)

Univariate:

- order the data set according to variable *X*

  $x_1 < x_2 < \ldots < x_N$
- form groups of consecutive values
  - fixed group size
  - variable group sizes (e.g., use within group variability, in μ-ARGUS)
- replace each score with group average

SDC methods

Continuous (micro-aggregation)

Note:

- Preserves totals

But:

- (Re-)grouping ‘similar’ records (households, …)

Forming groups:

- smaller groups ⇒ less loss of information
- smaller groups ⇒ less protection

SDC methods

Continuous (micro-aggregation)

Extensions:

- Multivariate case (clustering)
- Using other value than group mean
  (possible loss of preservation of total)

Note:

- Dependence *between* records

SDC methods

Continuous (noise addition)

Use a model to add noise to scores
(one record at a time)

E.g., additive noise:

replace score *y* with $y + \varepsilon$

where ε is drawn from a certain distribution

E.g., multiplicative noise:

replace score *y* with λy
where λ is drawn from a certain distribution

Miscellaneous topics

Data perturbing techniques on

- Identifying variables
- Sensitive variables

(e.g., PRAM, noise addition, micro-aggregation)

Rounding

- Continuous variables
- Sort of micro-aggregation/noise addition
- Aesthetic?

Miscellaneous topics

Sampling weights:

Noise addition

- ‘Enough’ different weights
- Overlapping intervals
- Preserving goal of weight inclusion