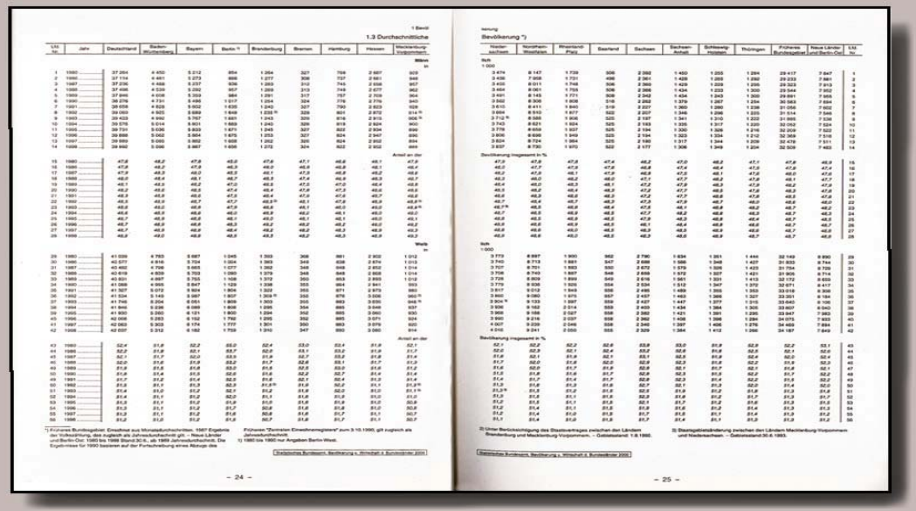
Statistical Disclosure Control of Census Frequency Tabular Data



Overview of methods

Sarah Giessing, Destatis

Frequency Tabular Data – Basic Terminology

- „Table cells“ refer to groups of units (households, persons), defined by cross-combination of qualitative grouping variables (Geography, Age, Sex,…)
- Table cells = Hypercube cells
- „cell value“ = „cell count“
  = Number of resp. units
  = Sum of a (0,1) indicator variable

Note: each unit contributes to numerous cells of the Hypercubes

Statistical Disclosure Control of Census Frequency Tabular Data

- **Disclosure risks**
  - **Identification**
  - **Attribute Disclosure**
  - **Disclosure by Differencing and/or Linking**
- **Protection methods**
  - **Post-tabular**
  - **Pre-tabular**

Statistical Disclosure Control of Census Frequency Tabular Data

- **Disclosure risks**
  - **Identification**
  - **Attribute Disclosure**
  - **Disclosure by Differencing and/or Linking**
- **Protection methods**
  - **Non-perturbative**
  - **Perturbative**

Statistical Disclosure Control of Census Frequency Tabular Data

- **Disclosure risks**
  - **Identification**
  - **Attribute Disclosure**
  - **Disclosure by Differencing and/or Linking**
- **Protection methods**
  - **Post-tabular**
  - **Pre-tabular**
- **Information Loss**

Risk of Identification

…for units with rare combinations / uniques (=1‘s, 2‘s)

- No direct disclosure risk
- But Imagine:
  - External knowledge may identify uniques
  - Then: linking to other tables (of other data sources)
    => disclose sensitive attributes of the unique individual
- Example: Male, immigrant, age 80+, divorced (=unique)
  - Receives social security
    - through linking to social security statistics
- Critical: rare combinations of identifying variables
  - …like
    - Geography (low level)
    - Sex
    - broad categories of age or nationality etc.

Risk of Attribute Disclosure

Can concern individuals or groups of individuals

- **Examples:**
  - **All inhabitants of a municipality are Roman-catholic (= group attribute disclosure)**
    - **Or: all except for one**
  - **All immigrants are muslims**
    - **there is only one immigrant (individual attribute disclosure)**
  - **All ‚high income' households fall into the same income size class category**

Problem: distribution / pattern of 0's across the categories of a sensitive variable

Risk of Disclosure by Differencing

Examples:

- **Two tables:**
  a) All Pensioners
  b) Pensioners at home
  Differencing => c) Pensioners in residential homes
  - **c) refers to small population:**
    - **likely to contain critical patterns of 0‘s, 1‘s, 2‘s**
- **Two tables with slightly different age-band, defined by identical classifications otherwise:**
  a) 0-20;20-40;others
  b) 0-25;25-60;others
  Differencing => c) 20-24
  - **c) refers to small population:**
    - **likely to contain critical patterns of 0‘s, 1‘s, 2‘s**

Risk of Disclosure by Linking

Example

- Two tables:
  a) non-immigrants by age (1-year age band) and sex
  b) 5-20 year olds by sex and type of school (all !)
  - Summing up from b) (across school types) =>
    c) 5-20 year olds by sex (all!)
  - Summing up from a) (across age years) =>
    d) 5-20 year old non-immigrants by sex
  - Differencing c) and d) =>
    e) 5-20 year old immigrants by sex
  - e) refers to small population:
    - likely to contain critical patterns of 0‘s, 1‘s, 2‘s

Statistical Disclosure Control of Census Frequency Tabular Data

- Disclosure risks
  - Identification
  - Attribute Disclosure
  - Disclosure by Differencing and/or Linking
- **Protection methods**
  - **Post-tabular**
  - **Pre-tabular**

Post-tabular Protection Methods

- **Non-Perturbative Methods**
  - **Recoding**
  - **Cell Suppression**
- Perturbative Methods
  - Rounding
  - Noise

Cell Suppression

Primary Suppression

- Small cells (1‘s, 2‘s)? All?
- How about 0‘s?
- Is group disclosure to be avoided? Always? (Many trivial cases!)

Secondary Suppression

- Software τ-ARGUS
- Problem: Concept
  - Need to pre-plan tables and variables. Otherwise:
    - non-optimal suppression patterns
    - co-ordination of suppressions across tables infeasible

Cell Suppression „good“ for

- Sample survey data (ambiguity about 0‘s)
- Not too many primary suppressions (1‘s, 2‘s)
- If tables are:
  - Not too many, not too large, not too many variables per table

Post-tabular Protection Methods

- Non-Perturbative Methods
  - Recoding
  - Cell Suppression
- **Perturbative Methods**
  - **Rounding**
  - **Noise**

Rounding

- **Several Variants**
  - **Conventional Rounding**
  - Probabilistic Rounding
  - Controlled Rounding in τ-ARGUS
- Several alternative approaches
  - Round all frequencies vs. Round only small frequencies
  - Round margins and inner cells independently vs. add rounded inner cells
- Consistency problems between linked tables

Conventional Rounding

- Replace each cell value by the closest multiple of a rounding base
- Does not preserve additivity
- Disclosure risk

Example:
Conventional Rounding to Base 5

Original	28	7	7	7	7
Rounded	30	5	5	5	5

Rounding

- Several Variants
  - Conventional Rounding
  - Probabilistic Rounding
  - Controlled Rounding in τ-ARGUS
- Several alternative approaches
  - Round all frequencies vs. Round only small frequencies
  - Round margins and inner cells independently vs. add rounded inner cells
- Consistency problems between linked tables

Random noise

Define p_{ij} =ℙ(noisy count = j | true count =i)

- **ABS method (Fraser, Wooton, 2005)**
  - **Step 1: produce non-additive tables: Fixed matrix P**
    - **random mechanism based on on micro-data keys => exact between tables consistency of the perturbation**
  - **Step 2: Compute closest additive table to the one resulting from step 1 (Note: step 2 introduces inconsistency between tables)**

Statistical Disclosure Control of Census Frequency Tabular Data

- Disclosure risks
  - Identification
  - Attribute Disclosure
  - Disclosure by Differencing and/or Linking
- **Protection methods**
  - Post-tabular
  - **Pre-tabular**

Pre-tabular Protection Methods

…are all perturbative

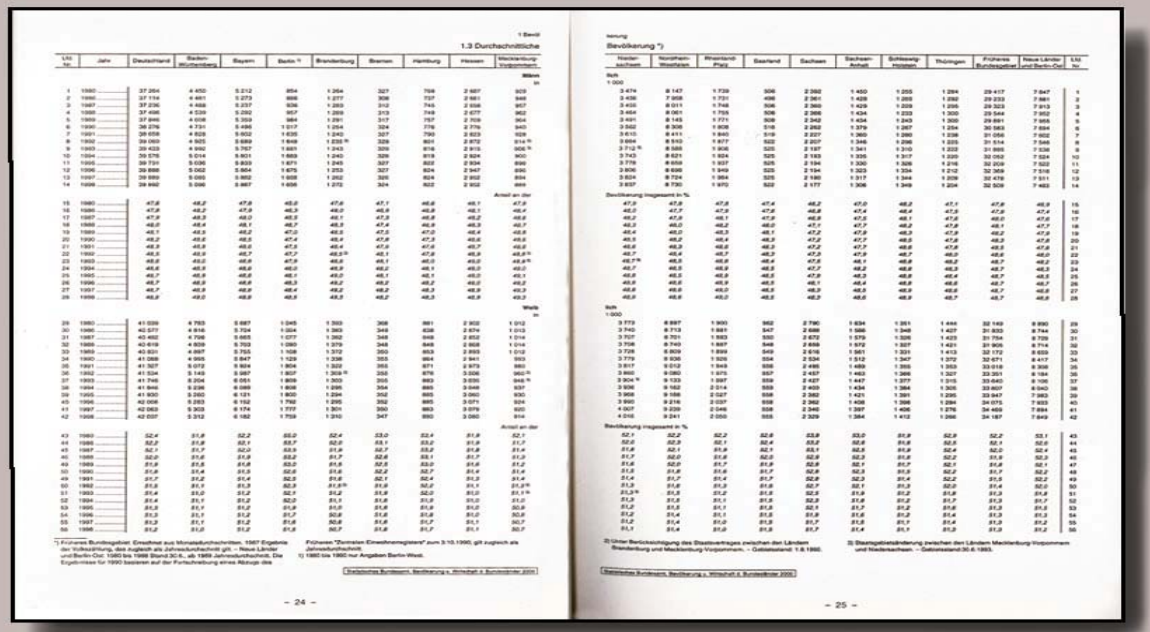
- Record Swapping
  - Swap values of variables between households within the same group
  - Grouping variables (typically):
    - Geo (rough), size of household, age x sex distribution of persons in the household
  - Swapping variable (typically): Geo (fine)
- SAFE
  - A kind of ‚controlled', ‚optimal' swapping
  - Removes all 1's and 2's

Statistical Disclosure Control of Census Frequency Tabular Data

- Disclosure risks
  - Identification
  - Attribute Disclosure
  - Disclosure by Differencing and/or Linking
- Protection methods
  - Post-tabular
  - Pre-tabular
- **Information Loss**

Information Loss

- For Cell Suppression
  - Obvious: Number of suppressed cells
    - … by ‚level‘ of aggregates
- For perturbative methods
  - Table level measure: distribution of (relative) deviations |true – perturbed|
  - Cell level measure?
  - More suggestions: see handbook



Sarah Giessing
Destatis, Mathematical-Statistical Methods
Phone: +49 (0)611 / 75-2701
E-Mail: sarah.giessing@destatis.de