microdata.no
System for Access to Register Microdata for Researchers

UNECE/Eurostat Work Session on Statistical Confidentiality
The Hague, Netherlands
29-31 October 2019
Authors

Dr. Johan Heldal
Senior Adviser | Division for Methods
Department for Digitalisation and shared business services | Statistics Norway
johan.heldal@ssb.no

Svein Johansen
Senior Adviser | Division for Access to Microdata
Department of Social Statistics | Statistics Norway
svein.johansen@ssb.no

Ørnulf Risnes
Head of Digital Development
Norwegian Centre for Research Data (NSD)
ornulf.risnes@nsd.no
microdata.no

• Launched in March 2018.
• 30 institutions with some 300 users are connected to the system.
• Present content: 142 register variables on population, education, labour market, income, and welfare benefits.
• Event history data on 10,2 million individuals.
  • All national IDs ever issued in Norway.
• Remote web-access. *Invisible* microdata.
• Developed 2012–2018 jointly by Statistics Norway and NSD – Norwegian Centre for Research data.
  • Funded by The Norwegian Research Council (NRC).
  • New funding from NRC from 2020 to go further.
• Own interface with Stata-like commands.
microdata.no makes it easier to analyse register data

- Researchers and students can process and analyse all available register variables.
- Data is available on the population, education, income, the labour market and welfare benefits.
- Institutions that have signed an agreement can manage their own users.
- The solution has an anonymisation interface with embedded privacy protection.

Data protection award for microdata.no

The prize “Built-in data protection in practice 2018” granted by The Norwegian Data Protection Authority was awarded to microdata.no.
Read more
Infrastructure layout – metadata driven, remote access

- import var1 yyyy-mm-dd as v1
- import var2 yyyy-mm-dd as v2
- import var1 yyyy-mm-dd as v3
Confidentially safe output

Subsetting user workspace to study population

drop if ...
keep if ...

[Diagram showing data analysis and subset selection]
Temporality structure, an example.

import SIVSTANDFDT_SIVSTAND 2002-01-01 as marstat02

import-event SIVSTANDFDT_SIVSTAND 2005-01-01 to 2015-12-31 as marstat05_15
Two Interfaces: Command Interface and Script Interface
Hidden slides describe them.
Statistical Disclosure Control - Foundation

• An interpretation of the Five Safes model (e.g. Desai, Ritchie & Welpton, 2016) into our context.

• Safe people, achieved through access policy
  • Only researchers in approved research institutions can have access.

• Safe settings
  • Safe login (three factor authentication), Invisible microdata, logging of all activity, communication with users.
    • Norwegian national ID required for access.

• Safe Data
  • Invisible microdata, automatic winsorization of numerical variables.

• Safe outputs
  • Statistical Output Disclosure Control, on the fly.

• Safe projects
  • Less relevant. Approval of projects not needed.
**Statistical Disclosure Control (SDC)**

- Invisibility of microdata is not enough!

**Challenges:**

- Users can manipulate population definitions and variables quite freely.
- Which is essential to do interesting research.
- But what kinds of smart attacks can a smart adverse user hit on?
- We have to play Devil’s Advocate to imagine.

- But we have to weigh risks against utility.

- SN and NSD have near 40 years experience giving access to microdata on contract.

- We have never seen intentional breaches.

- Detected breaches will hit researchers institution.
Present SDC measures – Output SDC

• Minimum size of populations to be analysed: At least 1,000
• **Constant** unbiased restricted maximum entropy noise ($X$) on counts.
  ($|X \leq 5|$, $X$ whole number)
  • Applies Cell Key method. (Same cell, same perturbation within same population).
  • Unbiased noise, $EX = 0$. Zeros not perturbed.
  • No negative perturbed counts and no perturbed counts 1-4.
• Automatic 2% winsorization of all numerical variables when subsetting and importing (1% at each end).
  • A dynamic form of top and bottom coding. Hides extreme values
  • Winsorized magnitude totals and adjusted proportionally to perturbed counts.
• Scatter plots are smoothed with hexbin plots.
• **Along with new funding, the SDC measures will be evaluated and reconsidered based on experiences and new knowledge.**
Some understandings

• Some differencing attacks are possible because noise is constant.
  • We have already done some adjustments.
• Winsorization.
  • Has large impact on some studies with age, income and wealth variables.
• In tables: Zeroes are disclosive as much as small counts.
  • Our noise creates more zeroes and uncertainty, but is it enough?
  • Should zeroes be perturbed? Differential privacy noise does.
• Presently: Tables are not additive
  • We have to introduce an additivity module for tables.
• Some users take results face value, disregarding noise uncertainty.
  • and disregarding present nonadditivity of tables.
• We must learn to tackle temporality in the data.
Some topics

- Should zeros be perturbed?
  - Yes, but what then about structural zeros?
- Can we accept perturbations to somewhere out of range of the data/parameter?
  - Can we accept negative counts?
  - No. Users will not accept them. Difficult if input to new analyses.
- Can we accept biased perturbations?
  - If zeros are perturbed and non-negative counts are not accepted we have to.
- Do perturbed counts have to be integers?
  - No why? Perturbed counts should look perturbed not to be taken face value.
- Is Differential Privacy a solution?
  - DP-noise (Laplace noise) can do these things.
  - Rinott et al. (2018) suggests that Australian Table Builder should consider DP.
  - Microdata.no and TB have much in common.
Differential Privacy and microdata.no

• Microdata.no is not a closed data universe.
  • Interactive situation, not only a TB, but also an analysis tool.
  • Users select some part of the data in time and scope to analyse
  • microdata.no will expand in time, units and variables.
• DP noise alone does not make microdata.no DP system.
  • DP requires a budget ($\epsilon$) for consumption of privacy.
  • We are not in a position to establish or handle a privacy budget.
  • What should the budget $\epsilon$ be?
• DP is at opposite to the Five Safes and the ideas for data access that microdata.no are built on.
• DP is to some extent also at odds with constant noise.
• But DP (Laplace) noise can be an option/inspiration when looking for better noise.
Some further proposals

• Wish to keep constant noise (same cell same perturbation)
  • Constant noise limits consumption of privacy.
  • Required for user trust in results.
  • When sharing a script with a colleague or a peer reviewer with access, the two should see the same result.

• Make non-structural zeros invisible by perturbing them to positive numbers.

• Let *noise on counts be continuous*, not integer
  • Then no users will take the perturbed counts face value.
  • Additivity model will generate non-integer counts anyway.

• To prevent computer intensive attacks: Set lower bound on response time or a maximum CPU-time for a query or script. Trusted users who need it can have exemption.
Some ideas for noise – comments are welcome

• $n$ true count
• $X$ Laplace noise $\sim f(x) = \frac{\epsilon}{2} e^{-\epsilon |x|}$ (eventually with bounded $x$)
• $Y = |n + X| \geq 0$

Then the noise will have some bias that quickly approaches zero as $n$ increases

$$E(Y|n) = n + \frac{1}{\epsilon} e^{-\epsilon n}$$

$$\lim_{n \to \infty} MSE(Y|n, \epsilon) = MSE(Y|0, \epsilon) = \frac{2}{\epsilon^2} \geq MSE(Y|n, \epsilon)$$

$$= \frac{2}{\epsilon^2} - \frac{2}{\epsilon} e^{-\epsilon n} \geq \frac{2}{\epsilon^2} (1 - e^{-1})$$
Uncertainty

• Uncertainty can be expressed for integer $n$s by the likelihood function

$$L(n|Y, \epsilon) = \frac{\epsilon}{2} \left(e^{-\epsilon|Y-n|} + e^{-\epsilon(Y-n)}\right)$$

which does not depend on the true $n$.

• It should be possible to see $L(n|Y, \epsilon)$ by clicking on the perturbed counts on the screen.
Winsorization

• Hides extreme values. Represents a kind of dynamic top and bottom coding.

• Serves in a way as a replacement for the DP sensitivity parameter

\[ \Delta = \max_{x \sim x' \in D} |f(x) - f(x')|, \quad x \sim x' \] datasets that differ by one unit only.

• In its present version it takes effect for all numeric variables in user work space, even on dummy variables.

• But must be modified to take effect only on relevant variables or when applied in certain queries.

• Creates biases for skewed variables.

• Correct downward biases with models for tail behaviour, e.g. pareto tails.
Analyses

• Microdata.no is not just a table builder. One can manipulate data and do analyses.
  • Until now primarily regressions and log-linear regressions.
• Winsorization is so far the only SDC measure that affects analyses.
  • Takes effect on both numerical regressors and dependent variables
  • Winsorization and lower bound for \( n \) (1000) prevents some of the scenarios mentioned in the literature.
  • But can be replaced by e.g. leverage restrictions.
  • OLS can be replaced with robust regression.
• O’Keefe & Chipperfield (2013) summarizes some attack scenarios with regressions and proposes measures.
  • We will consider some of their proposals.
• We wish to take noise into account in some analyses (if feasible).
Some challenges

• Constant perturbations of zeros require cell keys for empty cells
  • Can be done by means of cell keys for marginal cells
• Structural zeros.
  • **If noise on zeros:** Impossible to keep track of all possible structural zeros in crosstabulations in the system
  • **Possible solution:** Give users the option to set cells to zero apriori.
• Additivity module
  • Iterative Proportional Fitting to perturbed marginals.
  • IPF stepwise top down since marginals are inconsistent at all levels.
  • $L(n|Y, \epsilon)$ will be more complex.
• Winsorization: Now used on all numerical variables. Too much.
Summary

- microdata.no is an extensive system for data access
  - Contains temporal variables for an entire population over time
  - Users can freely select variables as they like, manipulate data, produce descriptive statistics and do analyses.
  - And technically: Try to attack the system.

- Present SDC in microdata.no is basically founded on our interpretation of the Five Safes (FS).
  - Present SDC measures must be improved and extended.
  - We wish to harvest from both FS and DP although opposite standards.
  - Retain constant noise (although at odds with DP).
  - We do not imagine a fully DP system with a budget.

- New funding from 2020 will make improvements possible.
With new funding from 2020 we want to

• Improve the SDC methods.
• Include variables from new sources
  • E.g. health registers (pilot)
  • Requires distributed Data Store and encrypted import to Work Space.
• Make more statistical methods (safely) available.
• Give safe access to new groups, e.g. public administration.
• International access to researchers
  • Requires log-on solution without Norwegian ID
• Cooperate with other stakeholders home or abroad on use with other kinds of data.
Thank you
Appendix: 1. Record keys

- Microdata.no operates with encrypted person IDs.
- They are 64 bits random binary numbers.
- As a record key, $U_i$ for individual no. $i$, we take 8 of the 64 digits
- With decimal representation, $0 \leq U_i \leq 255$. 
Appendix: 2. Cell keys

- $c$ is a cell, or equivalently: a set of individuals in microdata.no.
- $U_i$ is a record key, an encrypted national ID. 64 bits.
- Cell key: $U_c = \sum_{i \in c} U_i$, is used in a look-up table.
- Bitwise XOR addition: Addition of binary integers without using carry over digits.
- An example with 8 bits.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>Dec</th>
</tr>
</thead>
<tbody>
<tr>
<td>$U_1$</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>203</td>
</tr>
<tr>
<td>$U_2$</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>102</td>
</tr>
<tr>
<td>$U_3$</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>61</td>
</tr>
<tr>
<td>$U_4$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>239</td>
</tr>
<tr>
<td>$U_c$</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>115</td>
</tr>
</tbody>
</table>