

Microdata.no - Safe Access to Register Microdata

Johan Heldal, Svein Johansen (Statistics Norway) and Ørnulf Risnes (Norwegian Centre for Research Data)

johan.heldal@ssb.no, svein.johansen@ssb.no, ornulf.risnes@nsd.no

Abstract and Paper

In 2012 the Norwegian Research Council granted Statistics Norway (SN) and the Norwegian Centre for Research Data (NSD) nearly 4 mill. € (in Norwegian currency) in order to develop a Research Infrastructure for Register Data (RAIRD). The solution should satisfy the following requirements:

1. Online Remote Access
2. Microdata should be invisible for the users
3. Users should be allowed to combine data from different registers
4. All statistical results returned to the users should be confidentially safe

In March 2018 the system was launched under the name microdata.no for testing by trusted researchers in approved research institutions and in May 2018 officially opened for all researchers in these institutions. There are by now (April 26, 2019) more than 30 institutions with more than 220 users connected to the system.

In the paper and presentation for the workshop we will describe the working of microdata.no and how we have solved challenges to create a microdata access that is easy to use as well as confidentially safe. We will also present our ideas for further development.

We wish to give a demonstration of the system in a software session.

Safe Access to Register Microdata – microdata.no¹

Johan Heldal*, Svein Johansen**, Ørnulf Risnes***

* Statistics Norway (SN), Akersveien 26, N-0177 Oslo, Norway, johan.heldal@ssb.no

** Statistics Norway (SN), Oterveien 23, N-2211 Kongsvinger, Norway, svein.johansen@ssb.no

*** Norwegian Centre for Research Data (NSD), Harald Hårfagres gate 29, N-5007 Bergen, Norway, ornulf.risnes@nsd.uib.no

Abstract: Norway has a large number of registers of individuals established for administrative and statistical purposes, covering the entire population or significant subpopulations. The registers can be merged by personal identification numbers to form event history data bases. Access to these register data for research is highly demanded by researchers in Norway, but has been limited by strict rules, lengthy application processes and high costs. In order to get around these difficulties Statistics Norway and the Norwegian Centre for Research Data have established a system for easy remote access to the registers with confidentiality on the fly. The system, called microdata.no, opened for users in 2018. This paper describes the basic features of this system and ideas for further development.

Keywords: micro-data access, architecture, metadata, data protection, disclosure control

1. Introduction

At the UNECE/Eurostat Work Session on SDC in Helsinki 2015 two of us (Johan Heldal and Ørnulf Risnes) presented a project called “[The RAIRD project: Remote Access Infrastructure for Register Data](#)”. The goal of the RAIRD project was to establish a remote access system where researchers and their master- and PhD students could get access to data from administrative and statistical register in a safe and easy way without having to go through a lengthy application process. The project was funded by the Norwegian Research Council and was a teamwork between Statistics Norway (SN) and the Norwegian Centre for Research Data (NSD). Some of the requirements for the final system to be established were that

1. The system should be a Remote Access system. Microdata should not leave SN.
2. Microdata should be invisible to the users.
3. Users should be able to combine data from different registers.
4. The statistical output from the system should be confidentially safe.

The system was developed between 2012 and 2018. It became a reality and opened for users in March 2018 under the name microdata.no. About 300 users from about 30 approved research institutions in Norway are now connected to the system. For the moment SN and NSD runs the system at own cost. We received the approval for new grants for further development of the system (RAIRD II) from the Norwegian Research Council at the deadline of this paper.

During a two years’ start-up phase, the use of microdata.no is free of charge. This start-up phase will end 1. April 2020. After that date the institutions will be charged a fixed subscription fee and variable costs related to their actual use of the service.

Section 2 gives a rough description of the system architecture, the data within the system, metadata, user interface and about working with microdata.no. Section 3 describes data access, present statistical disclosure control measures and plans for how the SDC can be developed further with new funding.

¹ The opinions presented in this paper are those of the authors, not those of SN or NSD

2. The Structure of microdata.no

2.1 System Architecture

The system architecture was developed to provide researchers with rich, intuitive, efficient and flexible tools, in a managed executional context that prevents unwanted data disclosure. The ambition was to create an environment for researchers to work safely with personal data in an informed, autonomous and ergonomic way. The basic implementation is written in Python but the system has its own web-interface with Stata-like commands. Current statistical algorithms are taken from libraries in the Python programming environment like SciPy, NumPy, Pandas and Statmodels.

All end-user results are subject to automated statistical disclosure control before the result are returned to the user. Only methods that will not disclose individual information or whose output can be protected will be allowed in the system. See section 3.

2.2 Data

Available for the users is a fixed set of presently 124 register variables on population, education, labour market, income and welfare benefits with event histories on 10.2 million individuals, everyone who have ever been assigned a national ID number in Norway. The variables can either be *fixed* (in time, e.g. country or place of birth, date of birth and sex), *accumulated* (over a year, e.g. incomes and taxes), *cross section* (once a year) or *temporal* (event history variable that can change at any time). The microdata resides in the microdata.no *Data Store* (in Statistics Norway) and users must import the variables of interest to his or her *user workspace* where they can be manipulated and analysed. Variables in data store (i.e. long, narrow data sets containing variable values and encrypted personal IDs) are confidential.

microdata.no does not contain health data. Inclusion of health data is a future goal but will most likely require a distributed data store. The present organization of the data store could have been distributed but that is at the moment not required.

2.3 Metadata

Since microdata is invisible, the interface to microdata.no will be completely through metadata which must have all the qualities needed to meet researchers' needs. Metadata is open for everyone on all variables from the web page microdata.no (figure 1) and can be seen by clicking on Variables (figure 2), but is also integrated in the interface where each of the 142 variables can be clicked on in a separate panel to show their metadata. (Unfortunately, at the time of writing metadata is only in Norwegian but will be translated.)

microdata.no applies the Generic Statistical Information Model (GSIM) for the metadata. Each of the 142 contextualized variables is associated with detailed metadata records according to a metadata model developed during the RAIRD project. The model is informed and inspired by conceptual models from GSIM [2] and DDI [3], both of which have terminologies and constructions that support both informative and technical metadata. Metadata plays several roles in RAIRD/microdata.no.

- It informs users about definitions, data types, temporal nature and codes.
- It drives most technical components; data may only be accessed *through* metadata.
- It is used to assist users interactively when formulating scripts in the user interface.

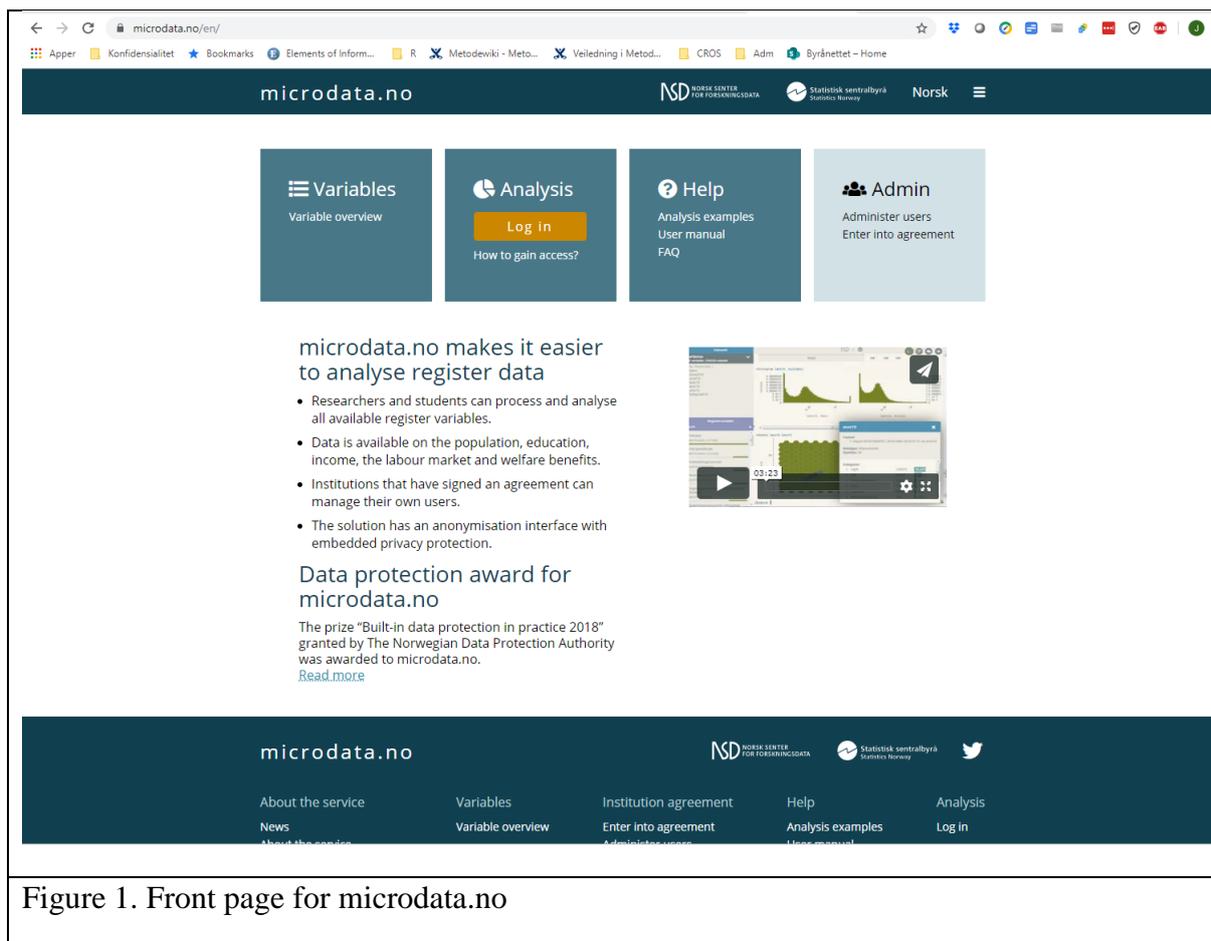


Figure 1. Front page for microdata.no

Furthermore, both GSIM and DDI have developed specialized models that address the relationship between a variable’s measure component (substantive content), its identifier component(s) and its attribute(s) (e.g. temporal attributes).

The models were extended, e.g. to enable support for code lists that evolve over time. Users can inspect code evolutions in various parts of the user interface and use this information to inform their data transformations and analytical interpretation.

Metadata in microdata.no builds upon several of SN’s metadata management systems, including the relatively new classification system KLASS [4].

2.4 User interface

The analytical environment in microdata.no has two main end-user-interfaces, *command line* interface and *script* interface. In the command line interface users can submit command queries one by one, using Stata-like commands, assisted by automatic writing aids for the commands. In script mode several commands can be run as an entity. In command mode there are also two metadata panels, one for the data store and one for the work space. In these panels, users can click on the variable names to see their metadata. In both interfaces, variables can be transformed, new variables created and descriptive statistics (tables, graphs etc.) and analyses can be requested.

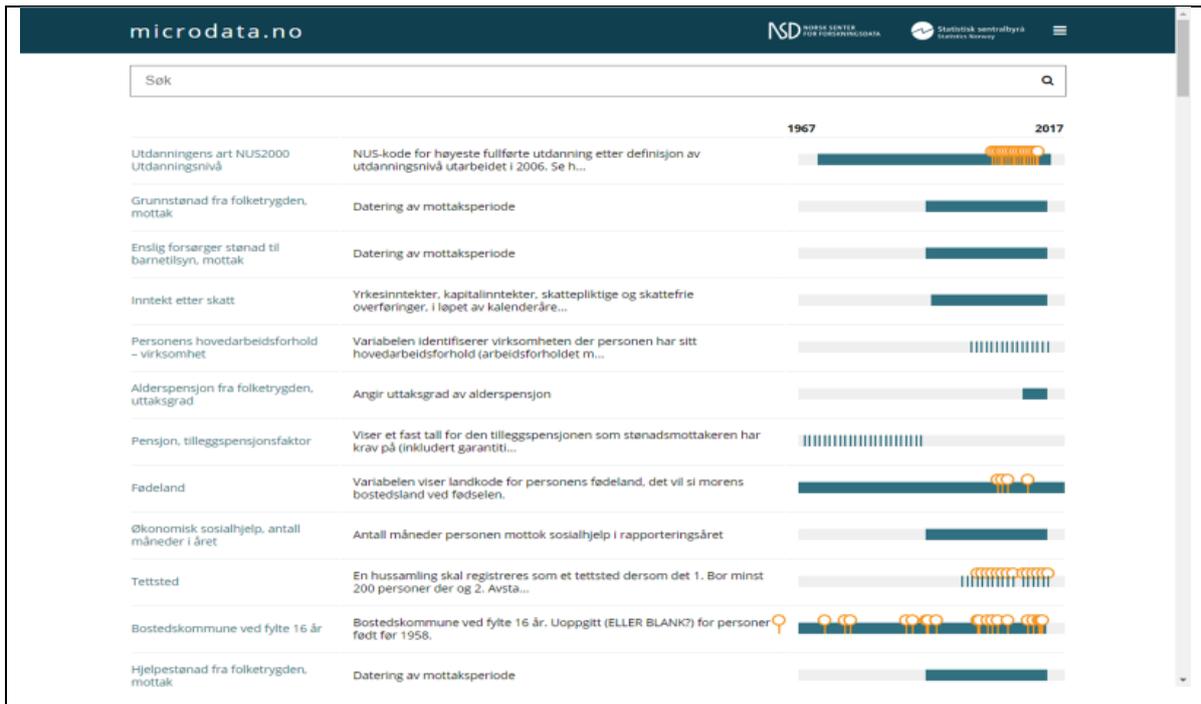


Figure 2. Screenshot from the variable catalogue

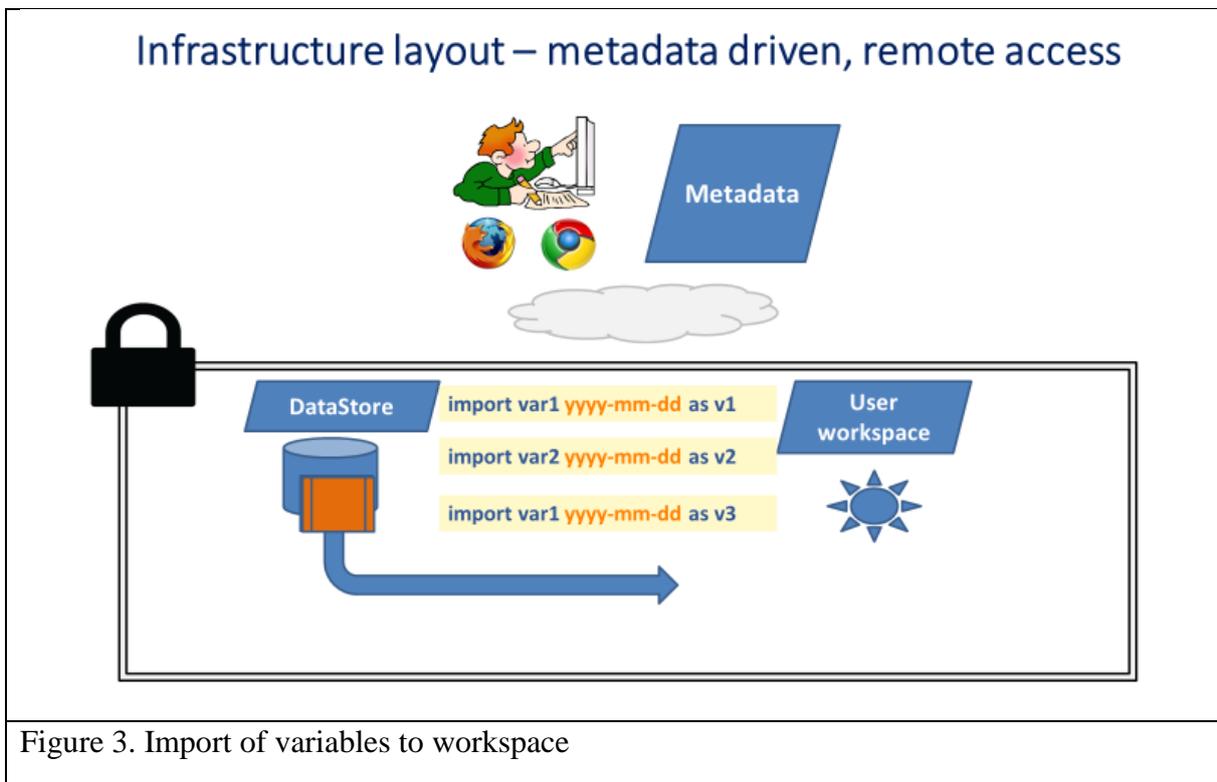


Figure 3. Import of variables to workspace

2.5 Working with microdata.no

A work session always starts with creating an empty workspace with a command like

```
>> create-dataset demographydata
```

Then variables are imported one by one. If the population living in Norway January 1. 2000 is interesting for study this is done with the command

```
>>import POPULATION_REGSTAT 2000-01-01 as regstat00
```

POPULATION_REGSTAT is the variable name in Data Store and regstat00 is the name of own choice in workspace. This command imports variable values for everyone with a value on the variable January 1. 2000, with values 1. Resident, 3. Emigrated, 5. Dead or 9. Unknown. Other variables could have been chosen for the first import, but the choice is not indifferent since only units (persons) with valid values on that variable will be imported. Later import is *left joined*, meaning that only values for individuals already in the dataset can be imported. Using the encrypted national IDs variables are merged when imported to the same work space. If we are only interested in residents, they can be selected with

```
>>keep if regstat00 == '1'
```

The population can be subset further using imported variables in the commands keep if <varname> or drop if <varname> . Analysis examples and user manual (in English) can be found from the microdata.no webpage.

3 Security and Confidentiality

The foundation for security and confidentiality in microdata.no is an interpretation of the philosophy of the Five Safes (safe people, safe settings, safe data, safe projects and safe outputs (e.g. Desai et al. 2016 [17])). The context of microdata.no is rather different from the original context in which the Five Safes were developed. Safe people are defined by the access policy to the system. “Safe settings” are defined by the fact that microdata is invisible for the users, logging of all user activity on the system and our communication with the users. Invisibility of microdata is also an element of “Safe data” along with winsorization which is described in section 3.2. Safe outputs are defined by the SDC-measures (on the fly) described in section 3.2. Safe projects are less important as access to microdata.no does not require a defined project, only an approved affiliation. All in all, we consider these elements as sufficient for the security and confidentiality of the system. The SDC-measures presently used are defined in section 3.2. Section 3.3 presents a critical discussion of them and proposes some changes.

3.1 Data Access

microdata.no is a service provided to master-/PhD-students and researchers at universities and approved research institutions in Norway. Each institution enters into an agreement of use with NSD and SN, giving them the right to sign in students and research fellows. Both the contract and the user administration are digital and self-administered by the institution.

Accredited users access through Norway’s national digital log in solution to public services; *ID-porten* [5] which requires a permanent or temporary Norwegian national identity number and uses three-factor authentication. At first log in, the users accept the terms of use, e.g. not to use the service for other purposes than research and not in any way try to attack the privacy of the system.

Registered users have access to all variables in microdata.no and need no approved project. Need to know does not apply.

3.2 Present Disclosure Control Measures

In official statistics the layout and content of all output to be published is fixed and controlled by the statistical office. Users of official statistics cannot see the microdata laying behind. Never the less statistical disclosure control is required. Users of microdata.no will be able to edit and manipulate populations and microdata in different ways and there is no fixed output. Thus, even though individual records are invisible for the users of microdata.no, there are many ways for a user to disclose individual information if nothing is done to prevent it. The primary methods for disclosure control used for the moment are

Minimum size of populations to be analysed (at least 1000 individuals). The limit 1000 is rather ad hoc, but there must be a lower limit. It cannot be 1.

Noise addition on counts (max ± 5). The noise is constant in the sense that repeating the same count within the same population frame (after subsetting) will lead to the same perturbed result whatever context. To achieve this, we use record keys and cell keys inspired by the Australian Table Builder and Data Analyser [6]. To avoid some differencing attacks through subsetting by only a few, perhaps only one individual, the cell key for the population total is added to the basic cell key. This will cause all cells to change if a table is requested before and after such subsetting.

The noise X on a count n satisfies some restrictions:

- X is a random integer with expectation 0 ($EX = 0$).
- $-5 \leq X \leq 5$.
- The perturbed count, $Y = n + X \geq 0$. Negative perturbed counts are not allowed.
- Zeros are not perturbed
- Perturbed counts Y in the range $\{1,2,3,4\}$ are not allowed.
- The noise distribution $p_n(x) = P(X = x|n)$ must maximize the entropy

$$\mathcal{E}(p_n) = -E(\log p_n(X)) = -\sum_{x=-5}^5 p_n(x) \log(p_n(x))$$

The resulting distribution is represented as a p -table and has been used to generate a random look-up table for the noise. The p -table is public and is shown in the user manual found from [microdata.no](#). The look-up table is confidential. With the above restrictions $Var(Y|n)$ will depend on n for $n < 10$ but will be constant equal to 10 for $n \geq 10$.

Although the noise on counts is unbiased this will not be the case for estimates based on functions of perturbed counts. Enderle et al. [15] have studied the biases and variances of ratios of counts and shown that variances can be significant unless all involved counts are large.

The encrypted national IDs, which effectively can be considered as pseudo-random numbers, are used as record keys. The cell keys are constructed by using bitwise XOR addition of the binary versions of the record keys. This amounts to using bitwise addition without using carry-over and generates cell keys with the same pseudo-random properties as the record keys.

Since perturbations of inner cells and marginals in frequency count tables are independent, the perturbed tables are not additive. This is a drawback both from a user perspective and an SDC perspective and an additivity module will be built on top of the tabulations. But to our own astonishment, and concern, so far there have been no user complaints about the non-additivity.

At all imports and subsettings with keep if or drop if commands all numerical variables are *automatically winsorized 2 percent*, 1 percent at each end of their distribution. This is a kind of dynamic top- and bottom-coding and replaces direct perturbations on their totals. The win-

sorization affects all descriptive statistics and analyses involving numeric variables and its current implementation has some drawbacks. For a discussion of this, see section 3.3.2.

Winsorized magnitude totals are perturbed proportionally to the perturbation of their underlying counts, meaning that (winsorized) averages are preserved unless the underlying counts have been perturbed to zero and so are ratios between numeric averages.

All activity on microdata.no is logged. This enables NSD/SN to reconstruct all user activity. On suspicion of misconduct the user will be asked to explain his/her activity. If adverse activity is revealed the user will have his/her access closed and the entire research institution can be shut out as well.

Spread sheets are potentially disclosive. If a user knows the value of one of the variables in a spreadsheet, the value of the other variables can sometimes be read with a too large accuracy. To avoid this, spreadsheets are smoothed with hexbin plots, like in the Australian Data Analyser. Hexbin plots partition the graphics area in hexagons with different colour shades indicating an interval for the number of points in it. Hexagons with zero or too few counts are kept white.

Descriptive statistics have been given priority in SDC since they are the easiest to use for someone who wants to challenge the confidentiality of the system. Analyses such as regressions, logistic regressions etc. are only affected by the winsorization, so far not by perturbation. This is partly an issue of capacity, but also because we want to avoid measures that can have adverse effects on the output of analyses. We will look into this now that we have got more funding, but we are not particularly worried that malicious users will try to take advantage. Scenarios for attacks using regression models or generalized regression have been described in the literature. We have seen them as less likely to occur and protection against them have not been given priority.

3.3 Experiences and plans for further development of SDC

Since microdata.no was launched we have realized that the present SDC measures have some weaknesses, both from an SDC point of view and from some user points of view. We will in this subsection give an evaluation of the present SDC measures and discuss some possible ideas for changes and innovations.

3.3.1 Noise on counts

When frequency count tables are seen in isolation many zeros in the same row or column are sometimes more serious from an SDC point of view than small counts since they exclude some categories and make others more likely or even certain in an SDC scenario. The present noise on counts protects by turning some small counts into false zeros that cast doubt on whether zeros are really zeros and so make disclosures uncertain. The present noise on counts in microdata.no is finite integer noise and since the noisy tables are non-additive we can sometimes see combinations of perturbed cell values that can only arise from one or a few true values. Such disclosure of counts will not necessarily cause disclosure related to specific persons but is never the less undesirable. Rinott et al. [9] also points out that maximum entropy noise, although intuitively appealing, lacks theoretical foundation. They suggest that the ABS Table Builder reconsiders its use of maximum entropy noise. That is something we will do as well, and some ideas are proposed below.

Use of Differential Privacy noise ($DP(\epsilon)$), i.e. Laplace distribution, eventually with some cut-off ($DP(\epsilon, \delta)$) does not mean that we have the intention to turn microdata.no into a completely

differentially private system. Complete Differential Privacy requires a confidentiality budget to each user or the entire user community. microdata.no is not a finite dataset but a dynamic data structure whose data will expand both in time scope and scope of variables. We are by no means able to handle a confidentiality budget for the users. In a DP system it should not be possible to decide whether an eligible person participates in a dataset or opted out. microdata.no comprises the entire population over time. Individuals belonging to a selected study population cannot opt out. Privacy can only be a question of an individual's contribution to the statistics. But we want to consider which advantages Laplace noise can offer compared to the maximum entropy noise presently used or other entropy-based noise structures.

DP-noise raises some issues where we either have to make decisions or find solutions. Some of them are related to perturbation of zeros.

1. Classical Laplace DP-noise generates negative values for counts or otherwise estimates outside the possible range of the statistic that is perturbed.
2. DP-noise perturbs zeros.
3. Laplace noise is unbiased ($EX = 0$)? If we perturb zeros but do not accept negative perturbed values some bias must be accepted.
4. Should noise variance be kept constant (independent of n)? In many applications today that is the philosophy, but we already deviate from this in microdata.no.
5. What should the privacy parameter ϵ be?
6. DP-noise generates non-integer values for counts. Is that acceptable?
7. We wish to keep the cell-key system for constant noise. But if zeros are to be perturbed we have to find a way to generate unique cell-keys for empty cells.
8. Structural zeros should not be perturbed.

Our attitude is that negative perturbed counts are not acceptable. Such values create problems for users, in particular if tables are to be used as input to other analyses. So, if zeros are to be perturbed we have to accept some degree of noise bias. But then it is also difficult to require constant noise variance. When noise variance depends on n it cannot be published without disclosing n .

We don't see non-integer counts as a problem. We have discussed this with researchers who say that having perturbed counts represented with decimals is no problem for them. In a more general context we find reason to ask if perturbed tables should look like perturbed. Integer perturbed counts do not always look perturbed for an untrained eye and can be taken face value which they should not. We have already seen users who neither care about the noise nor the non-additivity of the tables. Another aspect of this is that an additivity module will generate non-integer perturbed counts even if the original noise is integer. However, in dissemination of official statistics there is reason to believe that non-integer counts would raise eyebrows.

We consider ways to establish cell keys for non-empty cells. In a tabular context an option is to generate cell keys for empty internal cells from those of their one-way marginal (non-empty) cells, but we have not concluded on this.

Structural zeros may occur in so many ways. We may perhaps be able to control some of the most obvious situations, but far from all combinations that can occur in microdata.no. Users of the system will often have more information relating to their particular tables, and a possible solution can be to let them optionally define some known zero-cells as structural zeros a priori.

One proposal for a noise structure that complies to our conclusions to items 1-5 above is to add Laplace noise and take the absolute value, that is

$$Y = |n + X| \geq 0.$$

Then all zero counts will be positively perturbed and $DP(\epsilon)$ gives

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

where ϵ is the DP privacy parameter. Since the absolute value is just a transformation of an ordinary DP perturbed value, Y still satisfies DP. If ϵ is not very small, the bias will quickly approach zero as n increases. Variance will not be constant but

$$\frac{1}{\epsilon^2} = \text{Var}(Y|0, \epsilon) < \text{Var}(Y|n, \epsilon) \uparrow \frac{2}{\epsilon^2} \text{ as } n \uparrow \infty.$$

and

$$\lim_{n \rightarrow \infty} \text{MSE}(Y|n, \epsilon) = \text{MSE}(Y|0, \epsilon) = \frac{2}{\epsilon^2} \geq \text{MSE}(Y|n, \epsilon) = \frac{2}{\epsilon^2} - \frac{2n}{\epsilon} e^{-\epsilon n} \geq \frac{2}{\epsilon^2} (1 - e^{-1}).$$

Rather than expressing the uncertainty in terms of noise variance or MSE it can be expressed in terms of likelihood. The true n can be considered an unknown parameter in the density $f_Y(y|n, \epsilon)$ and the likelihood function $L(n|y, \epsilon)$ can be easily calculated for integer values of n in the vicinity of y . This likelihood can be expressed as

$$L(n|y, \epsilon) = \frac{\epsilon}{2} (e^{-\epsilon|y-n|} + e^{-\epsilon(y+n)}); y, n \geq 0.$$

$L(n|y, \epsilon)$ is not a function of the unknown n and, if desired, is fairly easy to calculate for some integer values of n in the vicinity of y .

We will have to establish an additivity module on top of the perturbed table. This can be done with Iterative Proportional Fitting “top down”, meaning that first the (perturbed) one-way marginals are fitted to the (perturbed) total for the table, then all two-way marginals are fitted to the fitted one-way marginals and so on. An additivity module may reduce impact of noise on inner cells in a table. A correct likelihood function like the above will depend on the noise on all cells and become more complicated.

3.3.2 Winsorization

All imported and derived numerical variables and winsorized 2 percent at import and subsetting. In a sense this can be seen as a kind of replacement for the sensitivity parameter $S(f) = \max_{x, x' \in D} |f(x) - f(x')|$ which is so important for the Laplace noise in differential privacy.

The winsorization, as implemented, has some undesirable effects on some variables. For instance, year and month of birth is winsorized. A consequence is that when calculating age in a given year, age is also winsorized and no one appears to be older than 89. This prohibits studies of the elderly population which should be legitimate. Instead of winsorizing at import and subsetting winsorizing will in the future take place when certain statistics are requested and not for all numerical variables.

Winsorization causes some bias when calculating means and totals of numerical variables with skew distributions like income and wealth. If relevant, a correction for this bias can be estimated based on models for the tail behavior for the variable, e.g. Pareto distributions for income and wealth. Such corrections should be optional for the user and based on the distribution of the

observations that are not winsorized only. The cumulative tail distribution of a variable X with Pareto tail is

$$F_X(x|X > x_m, \alpha) = 1 - \left(\frac{x_m}{x}\right)^\alpha, x > x_m.$$

x_p is the cut-off for winsorization. Then $E(X|X > x_p) = \frac{\alpha}{\alpha-1}x_p$ if $\alpha > 1$. α must be estimated from the non-winsorized observations $\{X: x_m < X \leq x_p\}$ which can be done. This can be used as a model-based correction for winsorization bias in relevant variables.

We carried out tests of regression analyses with highly skewed variables like income as dependent and wealth as one of the independent variables without and with winsorization on one or both variables. The tests show significant impact of the winsorization on the parameter estimates. But if income and wealth are log-transformed, giving a more symmetric distributions the effect almost disappears.

To avoid undesirable consequences, we consider changing the winsorization to take effect only when specific descriptive commands are run, like tabulate, summarize, histogram and so on. We will then have better control and can winsorize separately within cells or subgroups.

3.3.3 Protection against analysis attacks

O’Keefe and Chipperfield [12] have given a review of attack methods and confidentiality protection measures relevant for a system like microdata.no. One way to protect results from being used for potential disclosure is to add noise on the results directly, e.g. DP-noise. Adding noise to results, for instance in regression, is simpler and can be accounted for in variance estimates and test statistics.

An alternative, as proposed in Rinott et. Al. [9] section 6 and references therein, is to take the known perturbation distribution into consideration in the analysis, for instance the likelihood function. This requires modifications of the analysis software but can be feasible in analyses where the input can be reduced to a smaller sufficient statistic like tables for log-linear or logistic regression or cross-product matrices. Then noisy sufficient statistics created in advance can be given as input and valid results consistent with the input can be calculated.

In microdata.no it is possible to copy statistical output to clipboard and put the results into further analysis with other kinds of software. Such analyses will not take the statistical structure of perturbation into account and can lead to wrong statistical conclusions. If we can offer analysis software taking the noise into considerations, this can be avoided.

4 Future Development

Statistics Norway and the Norwegian Centre for Research Data have developed a system for safe access to invisible register data for research for researchers with Norwegian ID. Only statistical output is visible and is also confidential on the fly. The system is a frame that could be adapted to survey data as well and organized in a distributed manner.

More variables will be included in the microdata.no data store in the near future. In a slightly longer perspective we hope to be able to incorporate health data as well. This will probably require distributed data storage and encrypted transfers of microdata to workspace and possibly modifications of present SDC.

With new funding the system will be extended with more functionality, both with respect to statistical methods, better output and improved SDC.

We also want to give international access, but this requires a log-on solution without a Norwegian ID.

REFERENCES

- [1] Linnerud, Jenny SSB, Ørnulf Risnes NSD, Arofan Gregory MTNA (2014) RAIRD Information Model RIM v1_0
https://statswiki.unece.org/display/gsim/RAIRD+Information+Model+RIM+v1_0
- [2] Hamilton, Alistair, Choi, InKyung (2018) Generic Statistical Information Model
<https://statswiki.unece.org/display/gsim/Generic+Statistical+Information+Model>
- [3] DDI - Data Documentation Initiative. <http://www.ddialliance.org/>
- [4] Statistics Norway (2018) Statistical Classifications and Codelists
<https://www.ssb.no/en/klass/>
- [5] Agency for Public Management and eGovernment (Difi) (2018) ID-porten
<http://eid.difi.no/en/id-porten>
- [6] Thompson, G. et al (2013) Methodology for the Automatic Confidentialisation of Statistical Outputs from Remote Servers at the Australian Bureau of Statistics, *UNECE Work Session on Statistical Confidentiality, Ottawa, Canada, 20-23 October 2013*
- [7] Dwork, C. (2006) Differential Privacy. In *Proceedings of the 33rd International Symposium on Automata, Languages and Programming (ICALP)*, 2:1-12
- [8] Dwork, C., McSherry, F., Nissim, K., and Smith, A.: Calibrating Noise to Sensitivity in Private Data Analysis.
- [9] Rinott, Y., O’Keefe, C.M., Shlomo, N. and Skinner, C. Confidentiality and Differential Privacy in the Dissemination of Frequency Tables. *Statistical Science Vol 33, No. 3 2018, pp 358-385.*
- [10] Wood, A. et al. Differential Privacy: A Primer for the Non-Technical Audience.
- [11] Gaboardi, M. et al: PSI (Ψ): a Private data Sharing Interface (working paper 2018)
- [12] O’Keefe, C.M. and Chipperfield, J.O.: A Summary of Attack Methods and Confidentiality Protection Measures for Fully Automated Remote Analysis Systems. *International Statistical Review (2013), 81,3, pp426-455*
- [13] Chipperfield, J.O.: Disclosure -Protected Inference with Linked Microdata Using a Remote Analysis Server. *Journal of Official Statistics Vol. 30, No. 1, 2014, pp 123-146.*
- [14] Gomatam, S., Karr, A.F., Reiter, J.P. and Sanil, A.P.: Data Dissemination and Disclosure Limitation in a World Without Microdata: A Risk-Utility Framework for Remote Access Analysis Servers. *Statistical Science 2005, Vol. 20 no. 2, pp 163-177.*

- [15] Enderle, T., Giessing, S. and Tent, R.: Design Confidentiality on the Fly Methodology – Three Aspects. In *PSD'2018 Privacy in Statistical Databases (J. Domingo-Ferrer and F. Montes, eds.)* pp. 28-42. Springer Verlag
- [16] Dwork, C. and Smith, A.: Differential Privacy for Statistics: What we Know and What we Want to Learn. *J. Privacy and Confidentiality* (2009) 1. No. 2 pp 135-154
- [17] Desai, T., Ritchie, F. and Welpton, R.: Five Safes: designing data access for research. *University of West England, Bristol, Economics Working Paper Series no. 1601* (2016).
- [18] Chipperfield, J., Goe, D. and Loong, B.: The Australian Bureau of Statistics and releasing frequency tables via a remote server. *Statistical Journal of IAOS*, 32 (2016), pp 53-64.
- [19] Shlomo, N., Antal, L. and Elliot, M.: Measuring Disclosure Risk and Data Utility for Flexible Table Generators. *Journal of Official Statistics*, Vol. 31, No. 2, 2015, pp i305-324.