UNECE Expert Meeting on Statistical Data Confidentiality 2023

The 2023 UNECE Expert meeting on Statistical Data Confidentiality will be held in Wiesbaden, Germany, from 26 to 28 September.

The first information note for this meeting is available here. A second information note, containing practical information for participants is available here. This meeting will be comprised of sessions on the following topics:

- Innovative approaches in granting access to microdata for scientific purposes (new microdata access modes, services and tools);
- Producing useful microdata files; different approaches depending on different types of files;
- Risk assessment: Privacy, confidentiality, and disclosure vs utility;
- Output checking in research data centres;
- Challenges in publishing safe tables and maps;
- SDC communication, education, and training; and
- Other emerging issues.

We ask all participants to register via the link below. (Please note that you may need to create a user account on that website (and activate it via a link sent to you by email) before you can register:-

https://indico.un.org/e/SDC2023

Abstracts and Papers

Please note that the ordering of the papers below does not imply chronological order during the meeting itself. - A timetable for this event will be released when ready.

<table>
<thead>
<tr>
<th>Organisation represented by speaker</th>
<th>Title of contribution</th>
<th>Abstract text</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statistics Norway</td>
<td>Disclosure-Based Framework for Comparing Frequency Table Protection</td>
<td>For the protection of the dissemination tables from the 2021 population census, Eurostat recommended a combined use of the cell-key method and targeted record swapping. As part of a grant awarded to Statistics Norway on multi-grid geographical data, we compared the recommendation to alternative methods (in particular small count rounding) on dissemination of frequency data over multiple grid systems. This was done using Norwegian census data as a use case. In this work, we present the findings of this project, as well as discuss the comparison framework used. This framework is based on a suite of disclosure scenarios that can occur in frequency tables. Using established notions from information retrieval, disclosures are counted and evaluated for each scenario, providing measures of risk. Given an acceptable threshold for risk, methods deemed satisfactory are compared using common utility measures. Of the remaining methods, only those preserving enough utility are considered as viable protection methods.</td>
</tr>
</tbody>
</table>

Announcement

For info: Separate virtual workshop on Governance of Confidential Research Data in Low- and Middle-Income Countries
Taking place on 21, 23 & 25 August 2023
Further details are available here: https://way.office.com/3WXStYp4HeP6wCF?ref=Link&loc=play
In the European projects “Harmonized protection of census data in the ESS” and “Open source tools for perturbative confidentiality methods,” it was suggested to use Targeted Record Swapping and/or the Cell Key Method to protect the census 2021 tables. At the end of 2022 the first tables were sent to Eurostat, containing information at the level of 1km x 1km grid cells. In this paper we will discuss the outcomes of a workshop dedicated to discuss the SDC methods actually used to protect the grid cells, the intended methods to be used to protect the census hypercubes and the consequences for other publications of population statistics.

The past few years have seen an explosion of the volume of geo-referenced data, a trend that can be observed in the world of official statistics: large scale imputation, generalizing survey results to the whole population, is made more and more common thanks to the efficiency and the flexibility of new machine learning algorithms. Official agencies are now capable of providing realistic estimates of population characteristics at lower than ever aggregation levels, but communicating survey results at always finer geographical scales strongly increases privacy risks. Thus, in order to maintain trust between populations and their administrations, official statistical offices must ensure highest levels of confidentiality. In this context, Differential Privacy (DP) has been successfully applied to protect individual’s privacy by addition of properly scaled random noise. We first discuss the specifics of DP applied to regionalized statistics and present a baseline framework minimizing the amount of noise necessary to successfully control disclosure risk when releasing spatial aggregates. The technical readiness of the framework is illustrated through a synthetic case study based on Swiss poverty statistics using the OpenDP Library. Finally, we discuss some limitations of the DP framework when controlling disclosure risk of geo-referenced data and present some ongoing themes of research.

This study utilizes a standardized “census-like” dataset that is structured uniformly across all participating countries to assess disclosure risk based on grid data. We begin by evaluating and comparing the risk using this approach. Next, we apply spatial SDC methods from the R package sdcSpatial, including kernel density smoothing and quadtree aggregation. We re-evaluate the disclosure risk using these methods and analyze the resulting utility loss. Our analysis will be conducted across multiple countries, allowing for a comprehensive comparison of the utility loss between them.

In numerous countries, perturbative methods are increasingly used as a privacy protection method for official statistics. The U.S. Census Bureau has studied the applicability of perturbation based on differential privacy for official statistics. The U.S. Census Bureau has also empirically investigated the mechanism of differential privacy for the publication of statistical tables created based on data from the 2020 Census. In particular, in order to create and publish statistical tables for smaller geographical areas, the U.S. Census Bureau has examined the applicability of differential privacy for 2010 Census data as a protection against “database reconstruction attacks.”

Several empirical studies on the effectiveness of perturbative statistical methods such as additive noise, data swapping and PRAM for Japanese official microdata were conducted by Ito and Murata (2011), Ito and Hoshino (2012, 2013, 2014), and Ito et al. (2017, 2018). Other studies have investigated the possibility of adapting differential privacy for detailed geographical data from the Japanese Population Census, and examined the potential of differential privacy as an anonymization method for Japanese statistical data (Ito and Terada (2019) and Ito et al. (2020)). When discussing future directions for the creation and publication of statistical tables, it is important to consider the potential of differential privacy. Towards this objective, this paper conducts a comparative analysis of data usability for various differential privacy methods (with PRAM as a traditional disclosure limitation method) for statistical tables at different geographical levels created using individual data from the 2015 Japanese Population Census.

In a recent court filing, the then associate director for research and methodology, and chief scientist, claims that: “Somewhere between 52 and 179 million person who responded to the 2010 [U.S] Census can be correctly re-identified .” If true, this claim speaks poorly of the U.S. Census Bureau’s efforts to protect respondent confidentiality in the 2010 Census. Fortunately, these claims turn out to be exaggerated. We present a careful re-examination of the data to show that the claims of disclosure of identity of the respondents are overstated. By linking the Census Bureau definition of reconstruction and reidentification and the traditional statistical disclosure definitions of value and identity disclosure, we show that these claims are based on a misunderstanding of these definitions.
Research and policy development on the governance of confidential research data is dominated by the work of academics and government agencies based in high-income countries (HICs). This leaves three quarters of the world’s population faced with a corpus of theory and good practised guidelines which, although robust and well-established, makes little or no reference to the specific circumstances of low- and middle-income countries (LMICs). It may be that the data governance models developed in LMICs may be easily transferable to other contexts (there is some evidence, for example, that human-centred training adapts well), but in general there is little or no examination of this issue. There is however a large demand; a recent announcement of a training course in data governance for LMICs was 10x over-subscribed within the first two weeks of launch.

This paper reports on a virtual workshop to be held at the end of August 2023 which will explore this topic with researchers and statistical organisations across LMICs. The topics will include:
- How is research on data governance vary across countries?
- Are there key gaps in education, training, or information resources?
- Are there general lessons that can be applied across cultures and continents?
- Where and how do we develop capacity?
- Are there specific issues relating to LMICs which are poorly covered in HIS (e.g. treatment of indigenous peoples)?
- Potential participants are welcome to contact the organisers at dragon@uwe.ac.uk.

Remote access to European microdata

Until 2021, access to highly detailed European microdata (secure use files; microdata to which no or only very limited methods of statistical disclosure control have been applied) could only be performed by researchers visiting the safe centre located at Eurostat premises in Luxembourg. In 2021, the European Statistical System Committee (ESSC) agreed to enable remote access to secure use files. Eurostat together with representatives of national statistical offices, worked on the legal, organisational and technical aspects of remote access to European microdata. The approach proposed was accepted by the ESSC and the remote access system became operational in January 2022.

The information system (hereinafter “KIOSK”) where the secure use files are stored, can be remotely accessed from accredited access points hosted by research entities in the European Union countries, in Iceland, Lichtenstein, Norway and Switzerland, and in some other countries. KIOSK offers: (i) continuous availability of data and metadata to researchers; (ii) software tools for data analysis; (iii) tools suitable to the data protection and statistical disclosure control (SDC). Besides several software platforms and tools popular among the researchers, KIOSK also includes an output checking tool named “ACRO” (Automatic Checking of Research Outputs).

The covid19 pandemic assisted the acceleration of routine access to medical records for research. In the UK platforms including OpenSatellite and NHSDigital alongside emerging trust hospital based TRES demonstrate the utility and need for medical researchers to access and use microdata safely and securely. Whilst many employ traditional principles based SDC standards to statistical outputs, complexity arises when considering structured data - for example genome, medical imaging. MESI EEG data where the output is associated often includes reference to individual observations. Present imagining libraries and databases have demonstrated awareness and need for metadata standards however consideration of these is challenging. This approach can assist with understanding and help development of data disclosure control in these challenging contexts. For the moment it works in STATTA and deals with primary confidentiality only.

Federated data synthesis

Federated Learning (FL) is a decentralized approach to training statistical models, where training is performed across multiple clients, producing one global model. This approach can be used in scenarios where multiple clients have data available, but do not have the computing power or access the data held by another client. FL can be used to train models in a way that each client is only responsible for the data they own, maintaining privacy and security.

In general, research participants are welcome to contact the organisers at dragon@uwe.ac.uk.
The risk of identity disclosure is determined by two factors: the probability of disclosure conditional on a certain scenario of attack and the probability of that scenario actually taking place. Most SDC policies assume the worst case scenario by setting the second factor equal to one. However, in the case of newtypes of datasets it is interesting to investigate what scenarios are relevant to focus on.

Statistics Netherlands has recently developed a population-scale network data where nodes are persons and links represent various real-world connections including family, household, work, school, and geographical connections[1]. In this context we have developed an anonymity measure where it is assumed that an attacker has certain prior knowledge about the network structure surrounding a node[2,3].

In order to gain some insight into how likely it is that an attacker actually obtains such knowledge, a hackathon was organized where students were challenged to discover real-world connections surrounding a selection of persons that volunteered to participate. In a time span of about four hours, a group of 22 students found more than 5000 typed links surrounding 26 volunteers by searching or scraping the web. Students were asked to judge the reliability of link and link type and register the source of information. The results of the hackathon were partly checked by the volunteers.

Analyses of this data set provides anecdotal evidence for differences between the ease with which different link types can be found online. Although perceived relatively unreliable, social media links are extremely easy to obtain while links related to geographical vicinity and household sharing appear difficult to find. We also find differences between reliability estimates by the hackathon participants and the reliability indicated by the volunteers, depending on link type and source of information.


A Case Study of Output Checking in Japan

In Japan, the National Statistics Center has been responsible for checking the output of on-site use of official statistical microdata since its launch in 2019. We have accumulated experience in output checking as we examine how to apply checking rules to various outputs produced by researchers in different research fields. We have also identified the need for adding new rules. In this paper, we present our experiences in output checking as a case study in Japan and describe the new rules for quantities, which we plan to introduce.

Checking Data Outputs from Research Works: a Mixed Method with AI and Human Control

CASD is a public organization that can be likened to a research data center: it hosts a large amount of highly detailed data from various administrations (tax, data health, industrial data, etc.). These data pertain to both private and public people and institutions. The aim of CASD is to make such data available for research purposes, duly authorized by the data producers. CASD developed a specific remote access system which enables the researcher to visualize data, to interactively process any calculations for their research, with a wide range of scientific and data analysis software tools (SAS, R, Python...), and from the available data. The CASD data platform enforces a strict security policy, including the use of an output checking procedure: even if each researcher is accountable for their research work, especially when it comes to the compliance requirements if the outputs retrieved from their virtual research environment with GDPR or various confidentiality regulations, additional protection measures are in place. In particular, most of the outputs are manually controlled by CASD staff according to the rules data producers have established for this purpose.

However, some projects hosted by the CASD infrastructure can ask for ‘automatic outputs’ where manual checking is not systematic. This requires the authorization of every data producer involved. Usually, some outputs are forbidden in order to prevent massive disclosure (size and number of outputs are limited, encrypted files are forbidden). Manual checks are conducted randomly for other outputs, by CASD staff or by producers directly.

In order to make the system more efficient, CASD has developed a new tool, based on a Deep Neural Network (DNN) model. The model was trained using approximately 11k previous results of manual output checks: ok or refused, and the reason for elements rejected. A list of controls (the details of which are confidential), for metadata elements (elements of context) as well as for content of files, is implemented for each output. It is important to point out that the algorithm is able to open more than fifty file format types (csv, sas...) and to check their results of manual output checks: ok or refused, and the reason for elements rejected. A list of controls (the details of which are confidential), for metadata elements (elements of context) as well as for content of files, is implemented for each output. It is important to point out that the algorithm is able to open more than fifty file format types (csv, sas...) and to check their content, especially for detecting individual data patterns. Once all this has been carried out, the algorithm provides a ‘risk score’. If the score is too high, the output is manually checked by a CASD employee. The final check result is used to train the DNN model to continuously improve its accuracy. First training runs of the model are based on manual checks, then continuously increased with semi-automatic checks.

Many disclosure attempts result from misunderstandings of confidentiality rules, despite all researchers having received security training before they can gain access to any data through CASD. To avoid potential bias from the model, a few other automatic outputs, which present a lower risk score, are randomly selected and also checked. The results are then fed into the DNN model. This also allows the tool to detect which additional variables would be useful for the model: this kind of model is not static and evolves over time.

In order to make the system more efficient, CASD has developed a new tool, based on a Deep Neural Network (DNN) model. The model was trained using approximately 11k previous results of manual output checks: ok or refused, and the reason for elements rejected. A list of controls (the details of which are confidential), for metadata elements (elements of context) as well as for content of files, is implemented for each output. It is important to point out that the algorithm is able to open more than fifty file format types (csv, sas...) and to check their results of manual output checks: ok or refused, and the reason for elements rejected. A list of controls (the details of which are confidential), for metadata elements (elements of context) as well as for content of files, is implemented for each output. It is important to point out that the algorithm is able to open more than fifty file format types (csv, sas...) and to check their content, especially for detecting individual data patterns. Once all this has been carried out, the algorithm provides a ‘risk score’. If the score is too high, the output is manually checked by a CASD employee. The final check result is used to train the DNN model to continuously improve its accuracy. First training runs of the model are based on manual checks, then continuously increased with semi-automatic checks.

This tool is regularly refined: the aim is not to control every automatic output but only those which seem to be high risk. Tests are conducted to reach the best level of efficiency regarding the risks of disclosure. CASD also has a sanction system which ensures the legal and personal accountability of users. After investigation, all breaches are subject to actions in close collaboration with the data producers or competent authorities.

Remote Access for Scientific Use – A New Pathway for German Official Statistics Microdata Access

The basic objective of the Research Data Centres of the Statistical Offices of the Federation and the Federal States (RDC) is not only to enable the access to official statistics microdata but to continuously improve it and to adapt it to the changing needs of science.

The RDC offer different access paths through which differently anonymised data products can be analysed. All access paths differ both in terms of the anonymity degree of the provided microdata as well as in the access way of data provision. This range will be enlarged by a new remote access system including new data products. The analysis potential of the data provided therein will classify within the scientific use files transmitted to the scientific institutions and the data provided for on-site analysis at the RDC safe centres. Providing that the ongoing evaluation phase will turn out positive, this data access option will become part of the RCD’s portfolio from 2024 onwards.

Within the presentation, existing and firmly established data access paths are outlined and their contractual respectively legal conditions are explained. Subsequently, the newly installed ongoing evaluation phase will turn out positive, this data access option will become part of the RCD’s portfolio from 2024 onwards.

Providing that the ongoing evaluation phase will turn out positive, this data access option will become part of the RCD’s portfolio from 2024 onwards.

SACRO: semi-automated output checking

Output checking can require significant resources, acting as a barrier to scaling up the research use of confidential data. We report on a proposal, SACRO, to develop a general-purpose, semi-automatic output checking systems that works across the range of restricted research environments. SACRO is designed to:

• Automate checking of most common statistics, using best practice principles-based modelling
• Support researchers using the major analytical languages (R, Python and Stata), with minimal changes, by exploiting the ‘wrapper’ approach successfully trialled already
• Support secure environments with different operating models and output checking workflows, through a process of co-design to maximise usability

SACRO builds on previous work: (ACRO, funded by Eurostat and reported in the 2021 Workshop) to establish the proof-of-concept (monopet) and Py Acro which showed how a software-independent tool might be developed. It differs from those earlier projects in terms of a wider range of statistics covered, and a requirement to achieve general applicability. To do this, the project draws on our extensive networks of practitioners. A series of workshops and ‘hands-on’ evaluations ensure the design frameworks support buy-in from a wide range of prospective users across health and social sciences, and from the public and private sectors.
Towards a comprehensive theory and practice of output SDC

In 2000, the statistical disclosure control of outputs (OSDC) was largely limited to models of table protection developed by and intended for national statistical institutes (NSIs), as a particular branch of general SDC theory. However, in this century OSDC as a field of enquiry has expanded significantly, reflecting the important of secure research environments run by NSIs and others. OSDC is still a relatively under-developed field compared to SDC for tables or microdata. There are a small number of practitioner guides, and some theoretical articles, but this is a diffuse literature.

In the UK, the spectrum of universities and data providers is collaborating to provide an integrated analysis of output checking including:
- Key theoretical and practical concepts (eg safe statistics, principles-based OSDC)
- A comprehensive listing of statistics, associated risks, and mitigation measures as well as various practical element to support output checking.

A key element of this is a theory-driven classification which enables us to have that comprehensive listing whilst still limiting the dimensionality of OSDC guidelines to a manageable number of rules. This paper explains this model and how it has been co-developed with RDCs and others, and considers whether this provides a sustainable model for future development of the OSDC field.

In this work, we aim to implement a semi-automated output checking algorithm. A semi-automated output checking algorithm aims to facilitate the task for humans (or checkers) to check output data, to determine whether a particular output result or output is safe to be published to public or unsafe. By semi-automated we mean that human (or developer) is part of the system (called human-in-the-loop).

Few existing work looked at automated output checking. For instance, Domingo et al., (2021) proposed an approach that leverages machine learning to assist human checkers in output checking. First, Domingo et al., (2021) created synthetic output checking log files based on subset of rules (called 14 rules of thumb). Next, they trained a neural network model on each synthetic log file. Then, they tested how well the rules used to generate log files have been learned and how well the rules that were not used for training have also been captured and learned. In this paper, we will extend the work of Domingo et al., (2021) in threefold. First, we propose to test our predictions on real log file data. This will help us to approximate how good are our predictions when the model is trained on synthetic data. Second, we will replace the synthetic training log files by real log files. Finally, we will incorporate human-in-the-loop. This is an important step because it will allow our model to be updated during training taking into account human feedback.

An overview of data protection strategies for individual-level geocoded data

In response to a growing need for small-scale geographic information in various research areas, data-collecting institutions are increasingly georeferencing individual-level data. However, due to confidentiality concerns, external researchers typically have very limited access to these data if at all, resulting in a substantial loss of informational value. A growing body of literature on data protection strategies for geocoded data attempts to find solutions for the tradeoff between privacy protection and utility preservation of the individual-level data. The purpose of this paper is to systematically collect and review the literature in the field and to offer a classification of existing methods. Various strategies for estimating the utility and the remaining risk of disclosure for the protected data are also discussed.

Overview of the project and potentials of dataset synthesisation for official statistics and research

Statistical Disclosure Control for integrated and georeferenced data is a new challenge for statistical institutes. New digital data in combination with traditional data offer many new analysis possibilities. Moreover, these complex datasets are usually georeferenced in a very fine-grained way. Traditional confidentiality procedures reach their limits here. Destatis, the German Federal Statistical Office, is working together with various universities on the further development of existing procedures in order to ensure the protection of individuals even for complex data. The lecture will present the first results of the project “Anonymization for integrated and georeferenced Data” (AnigeD) funded by the German Ministry of Research. As part of AnigeD, Destatis further deals with dataset synthesisation of population as well as economic statistics to examine probable potentials for data provision to the research community as an answer to the growing scientific interest in official data. Confidentiality issues lead to several measures to ensure confidentiality of respondent units. Consequently, there is a diametrical relationship between level of anonymity and analytic potential of provided datasets which may not completely satisfy the needs of the scientific community. Research on data synthesisation is currently suspecting synthetic datasets to be a probable solution to this problem due to their artificial nature. With the following research an official statistics dataset will be synthesised and evaluated regarding analytical utility as well as the level of confidentiality. Furthermore, an evaluation regarding the ease of use of certain provision methods of synthetic datasets will be presented.

Smoothing the way for secure data access using synthetic data

In the UK, sensitive and potentially disclosive data (including survey and government-owned administrative data) are kept securely and safely in de-identified form and are only released for research and official statistics purposes, researchers can use it for exploratory analysis to determine if the real data is what they need; use it to improve their applications for funding and for access; and develop and test their code while they are waiting for access to the real data; and they can continue to develop their code outside of the SDE, therefore minimising the time and resources spent inside the environment.

In the UK, only a small number of data services provide access to synthetic data services, despite the development of numerous methods for creating synthetic data in the last decade or so. Administrative Data Research UK (ADR UK) and the Economic and Social Research Council (ESRC), the UK funding council for social and economic research in the UK, have invested in a programme of work to support the creation and routine operationalisation to supply low-fidelity synthetic data to support data access for research and improve the efficiency of SDEs. They have:
- Conducted in-depth study of the concerns and myths held by government data owners surrounding synthetic data production and use;
- Funded the creation of a Python Notebook tool to create synthetic data easily, at low cost and minimal risk which has been tested and approved by government departments;
- Focused on the position statement across its UK network taking the vision for synthetic data within its remit and mission;
- Embarked on a significant project to explore the utility and use cases of different approaches to synthetic data creation and to evaluate the efficacy of different models to provide recommendations for how synthetic data production can be achieved at scale whilst still acceptable to data owners;
- Developed a public dialogue on the acceptability of synthetic data, and public understanding of it and its uses to increase trust and confidence in its development for research for public good.

This session will describe the secure data landscape within which synthetic data sits in the UK and explain the approach taken by ADR UK and ESRC to utilise it as a catalyst for better and smoother research. We will demonstrate the effectiveness of provisioning access to low-fidelity data by describing how it makes the researcher journey for accessing data and use of data in a SDE more productive, while simultaneously reducing the burden for data custodians and maintaining confidentiality.
Microdata files are a particularly interesting product for the research community, since they offer greater flexibility when designing the different analyses and interpretation of results. At Eustat we offer this information for some statistical surveys and administrative records, generally in the area of population. All microdata files are protected prior to publication, that is, they do not include direct identifiers and have been treated to make the disclosure from indirect identifiers extremely difficult.

This paper describes the process for obtaining a safe microdata file from survey data. The first step is to assess which records can be easily identified and the second step is to implement the necessary protection measures. The objective is to maintain a balance between the risk of identification and the usefulness of the information provided to the user. To illustrate this process, the analysis carried out for the Labour Force Survey in the Basque Country will be shown. The final product consists of a safe microdata file with its associated metadata.

**KEY WORDS:** Microdata, Confidentiality, Dissemination, Metadata

---

### Assessing the utility of synthetic data: A density ratio perspective

Synthetic data can be a solution to reduce disclosure risks that arise when disseminating research data to the public. However, for the synthetic data to be useful for general inferential purposes, it is paramount that its distribution is similar to the distribution of the observed data. Often, data disseminators consider multiple synthetic data models and make refinements in an iterative fashion. After each adjustment, it is crucial to evaluate whether the quality of the synthetic data has actually improved. Although many methods exist to provide such an evaluation, their results are often incomplete or even misleading. To improve the evaluation strategy for synthetic data, and thereby the quality of synthetic data itself, we propose to use the density ratio estimation framework. Using techniques from this field, we show how an interpretable utility measure can be obtained from the ratio of the observed and synthetic data densities. We show how the density ratio estimation framework bridges the gap between fit-for-purpose and global utility measures, and discuss how it can also be used to evaluate analysis-specific utility. Using empirical examples, we show that density ratio estimation improves on existing (global) utility measures by providing higher statistical power and offering a fine-grained view of discrepancies between the observed and synthetic data. Moreover, we describe several additional advantages of the approach, such as providing a measure of utility on the level of individual synthetic data points, automatic model selection without requiring user specification, and readily available high-dimensional extensions. We conclude that density ratio estimation provides a promising framework in synthetic data generation workflows and present an R-package with functionality to implement the approach.

---

### Comparing attribute disclosure in synthetic and aggregated data

Interest in synthetic data techniques, including for official statistics, has been rising in recent years. This is in large part because synthetic data is very strong in preventing the identification of specific individuals. At the same time, it is known that synthetic data can contain probabilistic information about characteristics of individuals in the real data (sometimes known as attribute disclosure). If privacy officers or managers want to make well-considered decisions on the privacy impact of a synthetic dataset, it is important that they have interpretable estimates of the privacy impact for attribute disclosure specifically. Many organizations already publish aggregated datasets, where attribute disclosure is also relevant. This paper puts the attribute disclosure of synthetic data in the context of attribute disclosure of aggregated data. The attribute disclosure is measured using the targeted differential correct attribution probability (TCAP). This paper specifically uses open data published by the Dutch government, and also summarizes our experience in using this method in communication.

---

### Synthetic data and its distribution

The possibility of mounting reconstruction attacks against census outputs has been given as an argument to protect those outputs using differential privacy. However, several authors (Ruggles and Van Riper, 2022; Muralidhar and Domingo-Ferrer, 2023) have shown that reconstruction is different from reidentification and that there are no privacy risks when many possible reconstructions exist. Dick et al. (2023) have recently presented a method to reconstruct record prototypes from aggregate query statistics of the US Decennial Census data, where we call prototypes a record present with some multiplicity (i.e., number of repetitions) in the original data. The proposed method ranks the reconstructed prototypes by how frequently they appear in multiple reconstructions. The authors take this ranking as confidence metric telling how likely it is for a reconstructed record to be actually present in the original data. They interpret this as a privacy threat and an indication of reidentifiability.

We show that their ranking does not properly consider the multiplicity with which a record appears in the original data, and tends to award higher ranks to the most repeated records (which are intrinsically protected against re-identification by the repetition), whereas it misses outliers (which are those records that are really at risk of re-identification). Thus, their confidence-ranked reconstruction is ineffective at assessing the privacy risks, at detecting the most privacy-sensitive records (outliers), and at guiding re-identification attacks.

---

### Synthetic data for public use

Confidence-ranked reconstruction of census records does not reflect privacy risks or reidentifiability.


T. Dick, C. Dwork, M. Kosar, T. Liu, A. Roth, G. Vietri, and Z. S. Wu (2023) Confidence-ranked reconstruction of census microdata from published statistics. PNAS 120(8):a2218665120.
In this paper we explore the use of the differential privacy model for disseminating microdata. First, we elaborate on the limitations of the k-anonymity model in practice using a synthetic population dataset to assess the performance in terms of confidentiality and information loss. Next, we apply different methods to produce DP-compliant microdata and compare these methods in terms of confidentiality and information loss. Finally, we conclude by discussing the feasibility of DP as a privacy model for releasing microdata.
Generating Synthetic Microdata and Assessing Statistical Disclosure Risk Measures

This paper introduces two implementations of REALTabFormer, a GPT-based transformer model designed for generating synthetic microdata, both relational and non-relational. We provide a concise overview of the model and evaluate its performance using a benchmark dataset provided by the US National Institute of Standards and Technology (NIST). Furthermore, we utilize training data from diverse sources to create a synthetic census dataset for an imaginary country. This dataset serves to assess the accuracy of probabilistic statistical disclosure risk measures implemented in the sdcMicro and Argus software applications. Finally, we propose an alternative approach to measure the risk, which harnesses a synthetic superpopulation.

Intruder testing for Census 2021 England and Wales—checking risk and utility in Build Your Own system

By law, the Office for National Statistics (ONS) must protect the confidentiality of respondents to Census 2021. We protected the confidentiality of individuals’ data in three ways: swapping records between areas, applying a cell key method to each table, and applying disclosure rules in deciding which tables could be published. To assess the effectiveness of these methods and provide assurance, an intruder test was performed on Census 2021 data using a secure version of the outputs system. 51 intruders were recruited to attempt to identify individuals in the planned data outputs. 30 intruders took part, 81 claims were made, and more than half of these claims (41/81) were incorrect. Further steps were taken to reduce the risks identified by the test, making the data the majority of these claims were made from no longer possible to access through the Create a Custom Dataset system. This gave the Office for National Statistics evidence there was sufficient uncertainty in the data to meet the standard required by legal guidance and we would meet our ethical duty to protect confidentiality.

The case of bounds in noisy protection methods: Selected risks and utility perspectives from official population statistics

Noise-based approaches to protecting statistical confidentiality have become increasingly popular over the past decade, including for official population statistics. Many different concepts and practical methods exist meanwhile and have been studied at length. There are some generic risk/utility aspects shared by many of them, for instance the particular effects of bounding the maximum noise magnitude by a fixed value (or not). We focus on such effects of noise bounds in tabular population statistics outputs, showing on the one hand that the additional disclosure risks related to bounding noise can be controlled and on the other hand that there are important specific utility benefits of bounding noise in such outputs.

SDC in statistical education - the Polish experience

In addition to the development of the principles and tools of statistical disclosure control, it is also important to raise public awareness of why and how these methods are applied. In particular, the public should be educated about the use of SDC to protect privacy on the one hand and to maximize the amount of publicly available information, on the other. Educational efforts in this regard should obviously start with the NSI staff responsible for efficient data protection, and then this knowledge should be disseminated among data users and taught to people who handle statistics and data managers in various institutions and economic entities. In the paper we present some activities in this area that have been undertaken in Poland recently. They include specialized training workshops for members of the team responsible for SDC methods, including current and future experts in this field at Statistics Poland as well as SDC topics covered in the preparatory training for new employees of statistical offices in Poland. We will also give an overview of a monograph on SDC methods, which is about to be published by Pozna University of Economics and Business (PUEB), which is addressed to all those interested in ways of handling sensitive data or applying for access to microdata or tabular data from various sources. We will also present the syllabus of a new course entitled “Data confidentiality protection methods”, which is going to be launched at PUEB. Other educational initiatives, such as a tutorial on data protection for students writing their diploma theses will be mentioned.