2021

Synthetic Data

FOR NATIONAL STATISTICAL ORGNIZATIONS

A STARTER GUIDE

Contents

List of Figures

List of Tables

Acknowledgments

The authors of the chapters are as follows:

1. Introduction: Kate Burnett-Isaacs
2. Uses of synthetic data: Kate Burnett-Isaacs
3. Methods for creating synthetic data: Kenza Sallier
4. Disclosure considerations for synthetic data: Kate Burnett-Isaacs, Claude Girard, Alistair Ramsden, and Manel Slokom
5. Utility measures for evaluating: Gillian Raab

Thank you to Gail Bouchette, Jia Xin Chang, Héloïse Gauvin, Claude Girard, Ioannis Kaloskampis, Alistair Ramsden, Rolando Rodriguez, Manel Slokom, Christine Task and Steven Thomas for their contributions to this guide.

In addition, thank you to those who presented content during the HLG-MOS Synthetic Data Project over the course of 2020-2021. Their content has been included in this guide: Geoffrey Brent, Joseph Chien, Joerg Drechsler, Nicolas Grislain, Ioannis Kaloskampis, Ovyind Langsrud, Gillian Raab, Rolando Rodriguez, Kenza Sallier and Christine Task.

Chapter 1: Introduction

Data are a valuable commodity, providing critical input for statisticians, economists, and data scientists to generate timely and granular insights to respond to the information needs of a broad range of stakeholders. In a world where increasingly large volumes of data coming from an increasing number of providers, National Statistical Offices (NSOs) are using innovation to ensure data standards and definitions, good privacy and confidentiality management systems, and responsible data-sharing.

NSOs have a leadership role to play in establishing safe and transparent ways to share data, expertise, and best practices to support the use of data in testing, evaluation, education, and development purposes. With data integrity and confidentiality at the forefront, NSOs are well-positioned to provide the tools, methods and approaches to promote responsible data-sharing to meet a growing number of stakeholders needs in this ever-changing and fast-paced data ecosystem.

NSOs recognize that the call for greater openness and transparency of data must be met while simultaneously remaining steadfastly committed to protecting the confidentiality and privacy embedded in their data holdings. The dual mission of NSOs is nicely conveyed by Duncan et al. (2011) who write on page 12:

> *Data stewardship organization is serving two masters* [providing high quality information and protecting confidentiality] – *each with conflicting interests and concerns*.

It is widely acknowledged that releasing useful *and completely* safe information cannot be achieved in full ─ see for instance page 135 of El Emam (2013), item 4 on page 5 of Elliott et al. (2016) and the Office of the Privacy Commissioner of Canada who writes[1]:

> *It should be noted that there is no such thing as zero* [disclosure] *risk when releasing data.*

Also, it is widely acknowledged that 'safety' is relative, not absolute. For instance, Desai et al. (2016) write

> *'Safety' is a measure, not a state. For example, 'safe data'* […] *does not mean that the data is non-disclosive. 'Safe data' could be classified using a statistical mode of re-identification risk, or a much more subjective scale, from 'very low' to 'very high'. The point is that the user has some idea of 'more safe data' and 'less safe data'.*

It is within this context that NSOs must establish measures for the relative importance of the utility of the statistical information released and the protection for the person-level (or business level) information gathered from which it is derived.

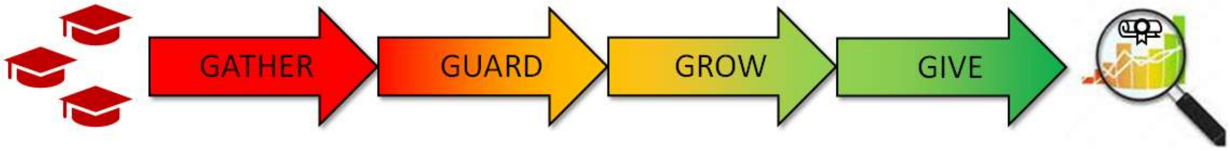
With the emergence of synthetic data, NSOs now have a promising output disclosure option that responds to the call to expand the usefulness of data holdings while maintaining the confidentiality

[1] See paragraph 130 of https://www.priv.gc.ca/en/opc-actions-and-decisions/investigations/investigations-into-federal-institutions/2018-19/pa_20191209_sc/

of record-level information. This presents the opportunity to move toward standard approaches with this *Synthetic Data in NSOs, a starter guide,* a compendium of theoretical methods to create synthetic data and a consensus of practical applications and best practices to promote consistency, transparency and comparability within statistical agencies, as well as with users in academia and the private sector. This synthetic data guide is intended to provide practical and direct guidance to decision-makers working in NSOs to determine if synthetic data is the right solution for responsible data-sharing. This guide opens with the most common scenarios within the scope of NSOs where synthetic data would be a suitable solution (Chapter 2). Chapter 3 discusses methods used to generate synthetic data as well as recommendations for determining the appropriate application of each. Chapter 4 highlights important considerations when disclosing synthetic data including privacy preserving techniques and measures to assess disclosure risk. Finally, Chapter 5 presents utility measures that can be used to assess how well the synthetic data meets the analytical needs of users.

Key concepts for synthetic data

Any discussion on synthetic data must involve an understanding of utility as well as privacy, sensitivity, security and confidentiality. Any decision on synthetic data creation and use involves the balance of these concepts. The underlying notions of privacy, sensitivity, security, utility and confidentiality merit further consideration, especially since 'privacy' and 'confidentiality' are often used interchangeably *in official statistics*. This simple and linear narrative illustrates how these notions all come together; it is inspired from the 4G framework[2] to describe the life cycle of data holdings (Rancourt, 2019).



Privacy [3] is associated with *how personal data are gathered* by a NSO, involving such considerations as the right of individuals to be free from observation, the NSO's entitlement to ask for and to obtain information pertaining to individuals, and individuals' *consent to share their information* with a NSO *only under agreed-upon terms*.

Once information has been entrusted to a NSO *through a sharing agreement*, it exists in a highly identifiable person-level form and must therefore be guarded against any unauthorized access. This is where the various data-storing options, access protocols and security measures available to NSOs come into play. Then utility is grown by *transforming* person-level data into statistical information, a form better suited for release purposes. This is where finite population estimates are calculated, complex statistical analyses are performed and analytical datasets are produced. For example, data gathered from graduates could be transformed into enrollment numbers by major fields of study. Finally, confidentiality issues relate to unwarranted disclosure, as per the

[2] Some representations include a fifth 'G' with Governance overseeing the other four Gs described here

[3] No consensus exists on the precise meaning of 'privacy': not only definitions may differ from one field to another (e.g., computer science vs official statistics), but they may also vary within a given field (e.g., between NSOs).

information-sharing agreement, of the *personal data entrusted to a NSO* that may occur when *statistical information* is released.

Thus, the 4Gs imply that privacy and confidentiality are distinct notions as they arise at *opposite* ends of the data life cycle - Gather and Give. This is not to say these notions are independent, but rather that concerns one may have with regard to each deserve separate considerations.

In the event of unwarranted disclosure, it is reasonable to expect that the harm done to an individual will be greater the more sensitive-in-nature the information revealed. However, it does not mean that information less sensitive in nature is any less confidential. In fact, reiterating the point made above, a piece of information is confidential because it is covered by the terms and conditions of the sharing agreement, not because of how sensitive it is deemed to be by its custodian.

To help manage confidential data, a NSO can rely on the Five Safes framework[4] developed at the Office of National Statistics in the United Kingdom, notably when thinking of data access solutions e.g., Desai et al. (2016). One of the framework's five dimensions is Safe Data, which examines the disclosure risk posed by the data itself. As access solutions, Public-Use (Microdata) Files (PUFs or PUMFs) and synthetic datasets are made to score high on the Safe Data scale through the use of appropriate disclosure control methods.

Extending the discussion on confidentiality, are the concepts of disclosure and disclosure risk. Disclosure risk is the risk or possibility of inappropriate release of data or attribute information (OECD, 2003). Disclosure risk applies to the dissemination of any microdata set or aggregate statistic and synthetic data is no exception. In fact, there is often a balance when creating synthetic data between the utility and the disclosure risk as the more closely synthetic data results emulates those of the original real data, the higher the risk that confidential information in the original real data could be disclosed.

National Statistical Organizations' data access options

Typically, much of the data collected and stored by NSOs is sensitive in nature and can only be accessed by trusted researchers and employees working in a secure environment under strict conditions. Original data files are not good candidates for public release because the information they contain can often be attributed to a respondent, constituting a violation of the requirements of statistical legislation, regulations, policies, standards, and relevant ethical guidelines. NSOs have a number of alternatives for safe data-sharing, ranging from dummy files to Research Data Centres, however there remains a gap in sustainable access solutions that balances utility with confidentiality.

At a minimum, NSOs have used dummy files for both internal and external data access solutions. Dummy files are data sets where almost none of the original dataset's analytical value is preserved. The focus is rather on maintaining the structure and the logical rules of the original file. The

[4] See www.fivesafes.org for details

problem with these files as a suitable disclosure option is that they provide very little analytical value and so are not useful to many users.

Of course, NSOs release real data. The core business of NSOs is to disseminate aggregate statistical information. This information is very valuable to the public, as well as industry and political decision makers, however, due to confidentiality reasons, aggregate information does not provide the level of detail in the data that users of official statistics are increasingly looking for.

NSOs do have Public Use Microdata Files (PUMF) at their disposal to provide users with more granular data. PUMFs are anonymized microdata files containing information relating to a sample of individual units from a survey, census or administrative file. The anonymization process is one where identifying information is modified or suppressed to ensure that unique entities cannot be identified on a datafile. Though PUMFs provide more granularity, the anonymization process can limit their analytical value.

To provide the utility many users are looking for while maintaining confidentiality, NSOs have created physical Research Data Centers (RDCs) to facilitate research projects that draw on survey data or administrative microdata files. RDCs provide direct access to a wide range of anonymized but not fully confidentialized microdata to accredited researchers under strict conditions. These physical location requirements can pose constraints for users. As a result, some NSOs like Statistics Canada, Statistics New Zealand, Australian Bureau of Statistics and Statistics Netherlands, have introduced real-time remote access or remote execution solutions, to enable users to quickly obtain a full range of descriptive statistics without the physical location requirement. This access option is typically limited to accredited researchers.

The traditional options for output disclosures used by NSOs are depicted in figure 1, with dummy files on the far left as the dissemination option that carries the least disclosure risk, however with the least utility, to the original data on the top right, as a point of reference where the utility is the greatest, however so is the disclosure risk.

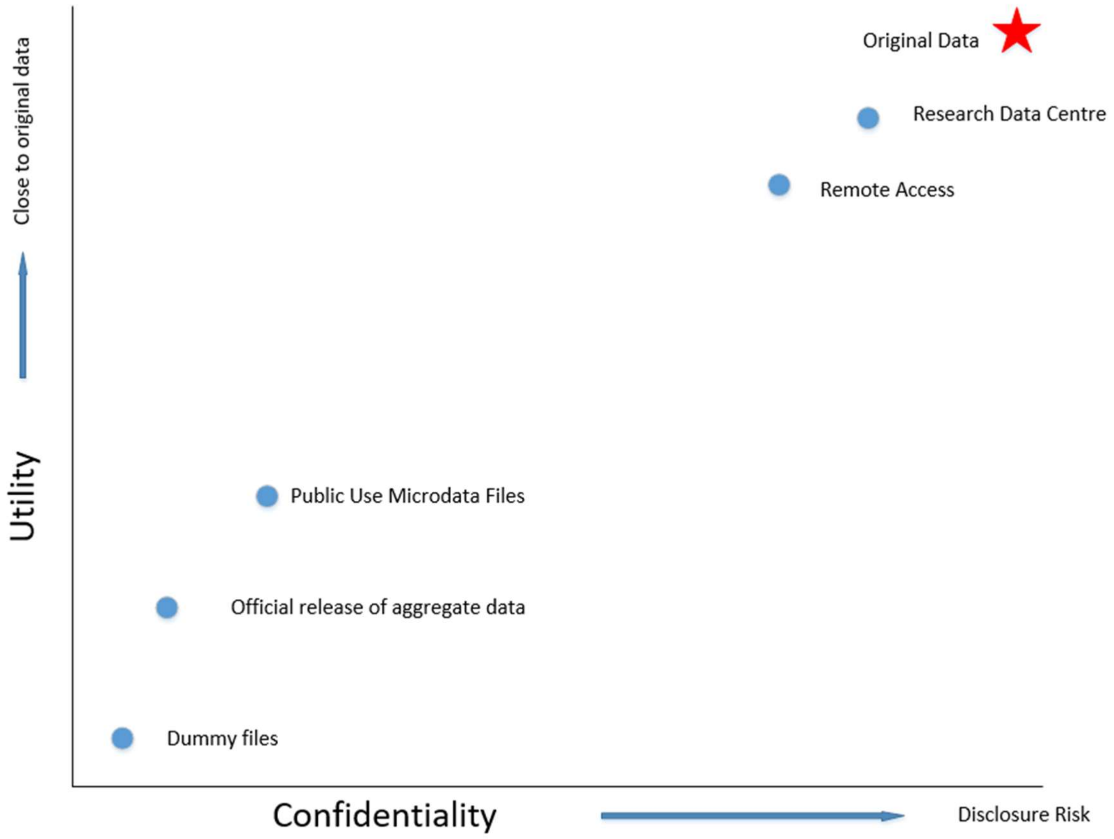


Figure 1: Confidentiality versus utility of current output disclosure mechanisms

As shown in figure 1, particularly between PUMF and remote access disclosure options, NSOs face a gap in utility and confidentiality options that needs to be filled in order to be more transparent, and to ensure data holdings are made more accessible to users. NSOs want to achieve more utility that each data set brings for their particular use, and with the same level of statistical disclosure control or confidentiality that traditional methods offer. Synthetic data is a viable alternative data dissemination strategy that can facilitate data access, especially in cases involving highly sensitive data.

A brief introduction to synthetic data

Synthetic data is defined as stochastically generated data with analytical value, while maintaining high levels of disclosure control. Synthetic data has roots in edit and imputation methods, and has experienced greater usefulness with recent advancements in computing and data science methods as well as the drive by NSOs for more open and transparent data-sharing. Generating synthetic data involves a modelling or generation process that targets both the preservation of analytical value and confidentiality.

The advantage that synthetic data brings to the suite of disclosure options for NSOs is that it breaks the direct chain from collected information to analytical data through a modelling or generation process. The stochastically generated synthetic data transforms personal (or business) information and provides users with faceless data that cannot be traced back to the individuals because the data was *generated* as opposed to directly *collected*.

At a high level, the goal of data synthesis is to take the original, confidential dataset (D) that outputs results (Θ(D)), synthesize the data (D’) so the confidentiality of the records is maintained, while also ensuring that the results of the synthetic data (Θ(D’)) match as closely as possible the results obtained based on the original data.

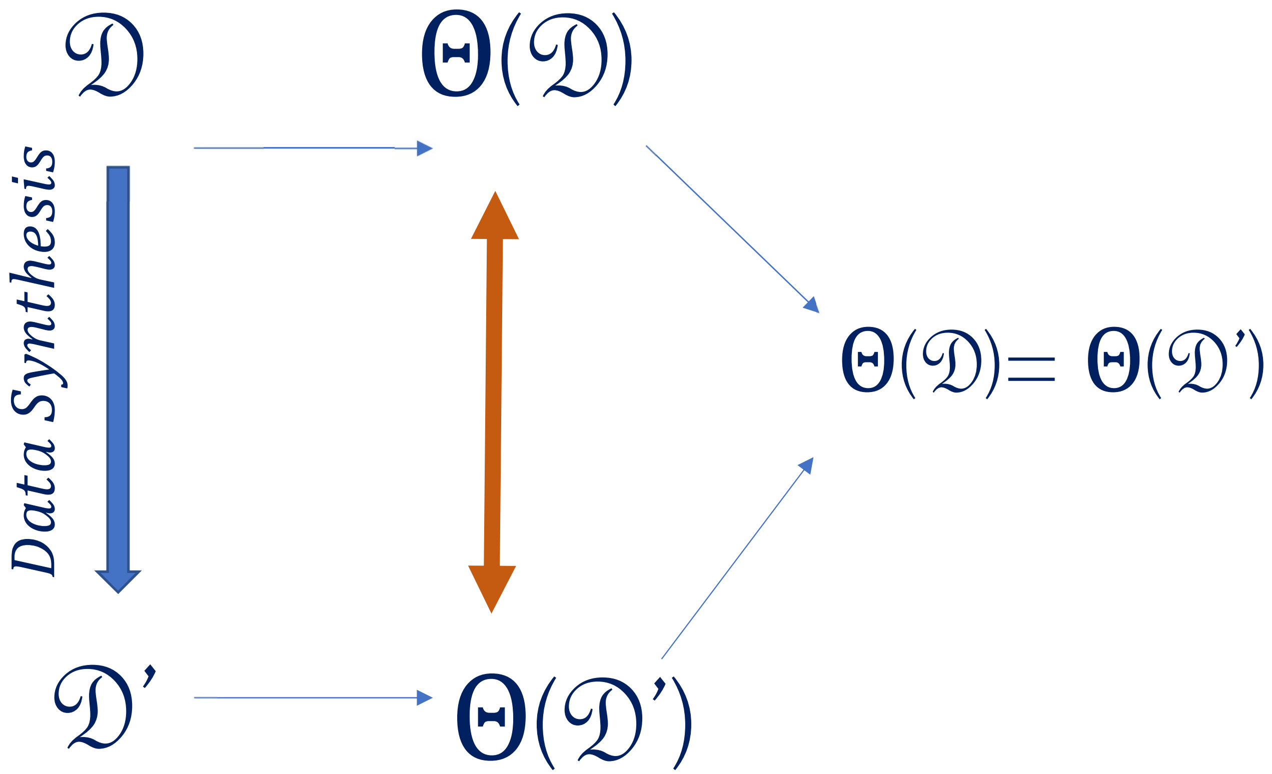


Figure 2: Illustration of the synthetic data generation process (Sallier, 2020)

Types of synthetic files

There are various types of synthetic files and elements of each should be taken into consideration when determining which one is best fit for purpose.

Dummy files

Dummy files are data sets where almost none of the analytical value is preserved. The focus is on maintaining the structure and the logical rules of the original file. These files are often used to test programs and processes, and to provide remote access to structurally similar but not inferentially valid data sets. Since there is no analytical value to these files, there is also almost no disclosure risk. This type of file is well known and in wide use, so is not a focus of this guide.

Fully synthetic files

In fully synthetic data files, all the variables are synthesized. The goal is to preserve significant levels of relevant analytical value compared to the original dataset in order to meet the needs of the user. This can be done by preserving univariate distributions for the variables from the original real data to the synthetic data, or preserving one or more multivariate or joint distributions. Variables can be generated in order to preserve only particular statistics (e.g. margins, mean, etc), or entire sets of relevant descriptive statistics for relevant distributions.

These files have the same use as PUMFs, but present greater analytical value when the joint distribution of the original data is preserved. These files strive to present low disclosure risk, however they could present actual inferential or merely perceived disclosure risk. A balance is needed as when their utility increases, so too can their potential disclosure risk.

Partially synthetic files

In partially synthetic files, only some of the variables are synthesized. The goal is the same as the approach with fully synthetic files but the approach typically focuses on a subset of variables in the data set. For example, synthesizing the most sensitive variables and leaving all of the other variables untouched.

Chapter 2: Uses of synthetic data

Synthetic data can solve statistical disclosure problems faced by NSOs, but the value of synthetic data varies with the type of problem. This section explains the main uses of synthetic data in NSOs, with a discussion of utility, disclosure risk requirements, and risk mitigation.

Releasing synthetic microdata to the public

High utility and high confidentiality

Traditional output disclosure measures used by NSOs can limit users' options to high quality microdata. With transparency and data access becoming increasingly important, NSOs are trying synthetic data as a new output disclosure option.

With public dissemination of synthetic microdata, NSOs cannot know or control how the data are used. There is no prior knowledge of what distributions, variables, or relationships need to be preserved in the synthetic data. As many as possible of the original data's relationships should be preserved to maximize the utility of the released file.

Of course, this increased utility increases disclosure risk. Since the audience is the public, there are no other controls or vetting processes over access and use of the data. Therefore, the confidentiality requirements is of utmost importance.

Example: Statistics New Zealand's synthetic unit record files

Statistics New Zealand (Stats NZ) releasing data at increased granularity. One way they are doing this is through Synthetic Unit Record Files (SURFs). SURFs are generated by a mathematical model, based on, but not the same as, original data. Stats NZ has released a few such files, including one based on the 2007 NZ Income Survey and a 'Census for Schools' SURF based on the 2019 NZ Household Savings Survey and NZ Census.

These files are semi-realistic representations of a sample of the New Zealand population, but respecting only the distributions, variables, and relationships that are preserved by the synthesis method. Published data are typically perturbed (noise-added) representations of population data. Hence, privacy of the data is preserved; even if some synthetic data looks like the original data, an attacker arguably cannot be sure which data are the same, which data are different, and whose data is whose.

For example, the Stats NZ SURF files released in 2007 were based on the Income Survey from the second quarter of 2003. Each file contains over 11,000 records. One hundred of these files were released in 2007 representing 100 samples of the New Zealand population between the ages of 25 and 64, who participated in paid work. The variables included are age, sex, ethnicity, highest educational qualification, weekly hours worked, and weekly income.

These SURFs were released with the intention and clear communication that they could be used for teaching or learning purposes, developing analytical methods or processes, or some level of statistical inference (Stats NZ, 2011). Stats NZ policy is to release SURF data with appropriate metadata about their methodology, inferential validity, and safety (disclosure risk).

Testing analysis

High utility and high confidentiality

Some NSOs grant confidential microdata access to trusted parties, remotely or at physical Research Data Centres (RDCs). But security checks, vetting, and approvals can greatly delay important research and analysis projects. Synthetic data could allow researchers to more easily develop and test their models, algorithms or analyses, and potentially conduct exploratory data analysis and/or determine initial hypotheses or conclusions, while they wait for access to the original real data. The original data would be required only to complete their research.

In some ways, this use is easier to accommodate, because NSOs typically know the types of analyses that researchers conduct. NSOs can generate synthetic data that preserve specific distributions, variables and relationships of interest to the researchers. The utility of those relationships in question take top priority. Variables or relationships that are not of interest need not be preserved, allowing for more flexibility in synthetic data model choices.

These datasets often involve extensive work for NSOs to provide custom synthetic data files for users.

Example: Statistics Canada synthetic census-based data

Starting in June 2021, one of the ongoing projects of Statistics Canada has been the creation of a synthetic version of a census-based database. The objective is to test and run the new dynamic microsimulation model of the Canadian retirement and income system, built for Employment and Social Development Canada (ESDC). A desired feature of the model is non-confidentiality, allowing the model to be used at any location. This would increase operational flexibility and the potential for external collaboration and broader use. Different options were explored and synthetic data was chosen as it seems likely to provide results closest to the original data.

The original database represents a part of the Canadian population in 2011 with some basic cross-sectional characteristics of the starting population. Based off of this starting population, the dynamic microsimulation can be thought of as experimenting with a virtual society of millions of individuals whose lives evolve over time. The microsimulation model will allow academics, researchers and government policy makers to model changes to the Canada Pension Plan (CPP), enabling research in public pensions and, more broadly, income security in retirement. The public (synthetic) database will support model development as well as preliminary program assessment, policy analysis and research. For final analysis and publication, the microsimulation model will be run on the original data housed in the Research Data Centers.

Example: Provision of synthetic data for users of the Scottish Longitudinal Study

The Scottish Longitudinal Study (SLS) is a source of linked data. At the core of the SLS is a 5% sample from the Census data for Scotland, where individuals (SLS members) are linked over time between Censuses. The data also include information on all household members of each SLS member. Further data sets are permanently linked to the SLS, including births, deaths, marriages and school-record data. Other data sets can be linked to the SLS for specific projects including hospital admissions, cancer registrations and many more. Further details can be found at https://sls.lscs.ac.uk/.

Currently, an extract of the data is prepared for each user with an approved project. The extract can only be viewed and analysed under supervision in the National Records of Scotland Offices in Edinburgh, which incurs a travel burden on researchers. To reduce this burden, researchers can request a synthetic data extract at the time of applying for access to the real data. To receive synthetic data, researchers and other members of the research team must complete safe-researcher training and agree to comply with conditions on the storage and use of synthetic data. The synthetic extract is then supplied to the researcher to analyse on their own computer. Results from synthetic data can only be shared among members of the research team and no findings from the synthetic data can be published. Final analyses for publication must be run on the original data in the Edinburgh office. In exceptional cases, SLS staff can run analyses remotely.

Example: Using synthetic data to test machine learning algorithms at the Australian Bureau of Statistics

Machine learning (ML) and artificial intelligence techniques have become more prevalent in both producing and gleaning insights from official statistics. At the Australian Bureau of Statistics (ABS), data scientists needed non-confidential data to test ML methods. The requirements for the synthetic data was that there was no original microdata file and that all information used to generate the data was public. The key goal of the synthetic data was to represent entities and relationships of interest for the ML model, such as persons, households, regions, industries and business units.

Using a microsimulation model, the ABS is able to create a synthetic data set, using only publicly available information that provides the detail and relationships appropriate for testing ML models.

Education

High utility and medium confidentiality

High quality data is needed in order for students, academics, and users in general to learn new concepts and methods related to a variety of topics such as data science, statistics, data analysis and even technology. The more complex the methods (such as machine learning or complex statistics), the more important it is that the data yield realistic results.

In providing data for education purposes, a NSO may know the specific method or topic that is being studied and, as in the testing analysis use case, may preserve only the distributions of interest. Alternatively, synthetic data made for another use can be repurposed for education, provided the educational use has similar utility and disclosure risk requirements to the original purpose.

In many other use cases, the high utility required puts pressure on disclosure risk. For education and training, where the audience is not security-vetted, disclosure risk becomes very important; setting lower inferential validity requirements can mitigate disclosure risk.

Use Case: Statistics Canada Canadian Health Measures Survey

Many universities in Canada are developing undergraduate programs focussing on faculty-mentored research to accelerated professional development. These programs aim to develop capacities in selecting and evaluating literature, working with large datasets, critical thinking and problem solving, collaboration, communication, and project management.

The Canadian Health Measures Survey (CHMS) at Statistics Canada includes a comprehensive dataset of questionnaire, physical and laboratory data. The data are currently accessible only via Research Data Centres (RDCs), which require stringent security checks that can be slow to complete. For undergraduate research programs to work with CHMS data, it is proposed to create a scientific use files (SUF) whereby some of the variables and survey weights are synthesized for open access. It is important to note that publishing based on the SUF would be prohibited; users would require access the original microdata for publication purposes. This dataset would allow students to develop analytical skills on data derived from a complex survey design.

Testing technology

Medium utility and medium confidentiality

When testing new software and technology, dummy data are often used: the file layout and error rates are representative of real data, but there is no analytical value. However, as complex technologies such as artificial intelligence and machine learning become more prevalent, testing will require more analytically realistic data. In these cases, synthetic data with some inferential validity can be beneficial.

The utility of the original real data needs to be preserved to some level, so that the results of the system can be assessed and verified – even though the conclusions drawn from the results have minimal value.

Example: UK Office of National Statistics (ONS) Census systems testing

The ONS used synthetic data in testing the census processing system prior to the 2021 census. Census processing comprises several phases, each dependent on the previous, with various quality 'gate checks'. While many system tests work on 'dummy' data, other tests require distributions representative of the population: realistic data is needed to verify that the checks work as expected and to ensure sufficient system resources under realistic workloads. Specifications on required test data distributions ran to roughly 60 pages.

Multiple synthetic datasets were generated to test load balancing and various functions used in the processing pipeline. Some synthetic variables previously not captured in the UK population were modelled and generated.

Chapter 3: Methods for creating synthetic data

There are many methods to generate synthetic data and to determine what method to use, it is important to start by identifying the type of synthetic data required and in what context they will be used. Namely, when creating a synthetic dataset, the synthesizer needs to take into account the desired analytical value to be preserved, as well as the acceptable level of disclosure risk, which itself mainly depends on level of accessibility of the synthetic dataset produced (public release, restricted release, etc.). In regards to the analytical value to be preserved, the spectrum of options is quite wide. Indeed, some projects require to only preserve specific statistics and statistical conclusions, that would have been identified prior to the synthesis exercise, and, at the opposite, some other projects require to maintain as much as possible all relationship between variables without selecting any specific statistics to be preserved in advance.

Lately, recent advancements in computer and software technology have made possible the implementation of different methods, more or less complex, in order to generate synthetic datasets of various natures. This work package aims to provide an overview of the methods available for producing synthetic data and establish consensus or recommendations on the most appropriate methods to use. The methods presented have been classified in 3 categories of methods: sequential modeling, simulated data and deep learning methods. The chapter finishes with a discussion on how to integrate survey features in the data synthesis process. This last section notably presents the last method described in this chapter. The focus is made on methods that can be implemented within the infrastructure of a NSO. The goal is to highlight the applicability of each of these methods in the practice of statistical organizations as well as the pros and the cons of the methods used and the synthetic datasets produced. In addition to the methods, this section also presents some of the tools that can help creating synthetic data as well as their respective limitations or considerations as well as recommendations on which methods to use for the use cases presented in chapter 2.

Sequential Modeling

The Fully Conditional Specification (FCS) method

The Fully Conditional Specification (FCS) method was originally developed in an imputation context (Van Buuren et al. 2006). Given that data synthesis can be seen as a massive imputation process on a dataset, the FCS can also be used to create synthetic data. Conceptually, analytical characteristics of the original file comes from the joint distribution of all variables of the dataset of interest. In practice, the joint distribution is not known and must be estimated using modelling. Experience has shown us that modelling the joint distribution in one step is usually too difficult in practice. Another option, more realistic, is offered by the FCS, which is a divide-to-conquer type of approach. More specifically, FCS decomposes the multidimensional joint distribution into a series of conditional and *univariate* distributions:

$$f_{X_1,X_2,\dots,X_p} = f_{X_1} \times f_{X_2|X_1} \times \dots \times \; f_{X_p|X_1,X_2,\dots,\, X_{p-1}} \qquad (1)$$

In other words, instead of trying to explain at once all relationships between the variables that exist in the dataset, the synthesizer proceeds step by step, by modelling and generating one variable at the time, conditionally to the previous ones.

Data synthesis using the FCS can be implemented as a two-step process. First, the FCS is used to model the joint distribution by carefully model each variable: the original dataset is used to estimate each of the conditional distributions represented in the right-hand side of (1). The second step consists in generating synthetic values for a given variable using the estimated model for this variable, using as input the synthetic values already produced for the previous variables (Drechsler, 2011). Because the goal is to preserve the joint distribution as a whole, we could say that, ***theoretically***, the FCS aims at preserving all distributions and statistical conclusions and not specific statistics pre-identified.

This method leads to 2 methodological questions to consider during the implementation: the order in which the variables will be synthesized and generated, and the models used for each of the variables. Notably, the analytical quality of the synthetic data produced depends directly on the quality of the input (original) data, the order of the variables and the models chosen. There is no known standard to select the order of the variables, subject matter expertise is probably required to ensure a logical order of the variables (for example, synthesizing age and level education prior to synthesizing income). Models should be chosen carefully, depending on the nature of the targeted variable. However, it is importance to notice this lets some flexibility to the synthesizer as each of the variables can be modelled differently from the others. Thus, some variables could be modelled through parametric models and some other through non-parametric or mixed models. In regard to that, the Classification and Regression Tree (CART) machine learning model is often seen as a standard (Drechsler and Reiter, 2011). Indeed, it can be more easily implemented and more adapted to data with irregular distributions (Reiter 2005). CART can notably capture non-linear relationships between variables which may not have been properly taken into account with parametric modelling. This is important in attempting to retain analytical value in the synthetic dataset.

Pros, cons and considerations

Table 1: Pros and cons of Fully Conditional Specification

Pros	Cons
This method is easy to understand and easy to explain. Because the target is the joint distribution of the dataset, this method aims at preserving all relationships between all variables. Relationships of interest are not required to be known prior to the creation process. Furthermore, because it stems from imputation, this approach naturally bears a strong resemblance operationally speaking with its well-established data-editing sibling.	For skewed data (such as business or economic data), the presence of outliers remains a challenge in terms of disclosure or perceived disclosure control. With many variables the process can become time-consuming.

Tools for FCS

The R package Synthpop is a tool for generating synthetic datasets. The main method available to produce synthetic datasets is the FCS (Nowok et al., 2015). For more information visit www.synthpop.org.uk.

In practice: Statistics Canada

As of today, Statistics Canada has released two synthetic data of high analytical value for public use. In both cases the mandate was to provide synthetic datasets of high analytical value to participants of hackathons. High analytical value means that statistical conclusions from the synthetic file should be as close as possible as the ones from the original file, independently of the analysis performed.

The first experience was in 2018 during which Statistics Canada's Health Analysis Division sponsored a hackathon as part of the 5th International Population Data Linkage Network conference (Banff, Canada). The goal was to have participants compete in a time-limited and team-based analysis using a synthetic dataset mimicking a real linked dataset that combines some socio-demographic variables and mortality indicators. The dataset to be synthetized was the result of the linkage of the 2006 Canadian population Census Long Form to the 2015 Canadian Mortality Registry.

The second experience was very similar to the first one. In 2019, the Canadian Partnership Against Cancer (CPAC) in collaboration with Statistics Canada's Centre for Population Health Data (CPHD) invited researchers from across Canada to participate in a hackathon taking place during the 2019 Canadian Cancer Research Conference (Ottawa, Canada). Similarly to the first project, the participants had to conduct analyses on relationships between cancer incidence, treatment and sociodemographic characteristics. Here, the original data to be synthesized was the result of the linkage of the 2006 Census Long Form to the Canadian Cancer Registry (1992-2015) and to the Canadian Vital Statistics Death Database (1992-2014).

More detailed information in terms of the method, the implementation and the evaluation can be found in Sallier (2020).

The Information Preserving Statistical Obfuscation (IPSO) method

The goal with the Information Preserving Statistical Obfuscation (IPSO) method is to generate new synthetic data values while preserving specific statistics and statistical conclusions (Cano and Torra, 2009).

With this method, we consider the original data as being made of two subsets of variables: the matrix $\boldsymbol{Y}$ and the matrix $\boldsymbol{X}$. The matrix $\boldsymbol{X}$ is made of non-confidential variables and the matric Y is made of confidential variables. This method works under the assumption multivariate normality distributions and of a linear regression model where the variables in the matrix $\boldsymbol{X}$ are considered independent and those of the matrix $\boldsymbol{Y}$ are considered dependent. Usually, the synthetic dataset to be released is made of a synthetic version of $\boldsymbol{Y}$ ($\boldsymbol{Y'}$) (and in this case if fully synthetic) or is made of the junction of $\boldsymbol{Y'}$ and $\boldsymbol{X}$ (and in this case, is partially synthetic).

The main goal of the IPSO method is that, when adjusting the regression model, $\boldsymbol{Y} = \beta\boldsymbol{X} + \varepsilon$, the synthetic values $\boldsymbol{Y'}$ gives the same (or very close) estimates of parameters and standard errors (and covariance matrices) as the original $\boldsymbol{Y}$. There are multiple vintages of the IPSO method but they all start with the same two first steps: the multiple regression model $\boldsymbol{Y} = \beta\boldsymbol{X} + \varepsilon$ is adjusted on the original data and the fitted values $\widehat{\boldsymbol{Y}}$ are used as a base to construct the synthetic values $\boldsymbol{Y'}$. Then, a normally distributed noise is added to $\widehat{\boldsymbol{Y}}$ to obtain the synthetic values $\boldsymbol{Y'}$. We could stop with these two first steps but there would be the following inconvenient: if a user fit the multiple linear regression model $\boldsymbol{Y'} = \beta\boldsymbol{X} + \varepsilon$, the estimates of $\hat{\beta}$ and $\hat{\Sigma}$ would highly likely be different from the estimates obtain with the model fitted to the original data : model, $\boldsymbol{Y} = \beta\boldsymbol{X} + \varepsilon$.

Thus, some versions of IPSO consist in adding extra steps to force the equality $\hat{\beta}_{original} = \hat{\beta}_{synthetic}$ whereas some other versions are even stricter and force both equalities $\hat{\beta}_{original} = \hat{\beta}_{synthetic}$ and $\hat{\Sigma}_{original} = \hat{\Sigma}_{synthetic}$. This can be achieved by either modifying the values of $\boldsymbol{Y'}$ or values of X in such a way that equalities are respected. In the same vein, the synthesizer could decide in advance what specific parameters or sufficient statistic derived from the regression model they would like to preserve.

Hybrid methods can be obtained by completing existing methods with IPSO. There are also some other methods having IPSO as a special case. Domingo-Ferrer and Gonzalez-Nicolas (2010) combines microaggregation with generation of synthetic data: the IPSO procedure is run separately within each microaggregation cluster. The statistics preserved by ordinary IPSO are also preserved by this method. Using the method in Muralidhar and Sarathy (2008), it is possible at the variable level to select the degree of similarity to the original data. There is also the Random orthogonal matrix masking (Ting et al., 2008) that controls the relationship with the original data by a single parameter. By using one of these hybrid methods, the problem of non-normal data can be reduced which help decreasing the need for strong assumptions. More generally, Langsrud (2019) describes all the above methods under a common framework and develops improved algorithms and generalized methods within the framework.

Pros, cons and considerations

Table 2: Pros and cons of Information Preserving Statistical Obfuscation

Pros	Cons
Like the FCS, the method is easy to understand and to explain. With this method, it is possible to preserve exactly some pre-identified parameters and sufficient statistics. Thus, any analysis relying on (multivariate) normality will produce the exact same results in original and synthetic data. IPSO can be implemented as part of another method or process, to generate synthetic datasets. These hybrid methods may be used alleviate the normal distributions assumptions	Normal distribution for all variables is a strong assumption that is seldom true.

Tools to apply IPSO

Examples and R packages for IPSO can be found at the following resources:

Mu-Argus, Implementation of Domingo-Ferrer and Gonzalez-Nicolas (2010), https://github.com/sdcTools/muargus

R package sdcMicro, An implementation of Ting et al. (2008) is included as a noise addition method, https://cran.r-project.org/package=sdcMicro

R package RegSDC, Implementation of all methods described in Langsrud (2019), https://CRAN.R-project.org/package=RegSDC

Simulated Data

From dummy files to more analytically advanced synthetic files

Simulations are often used in statistics to generate artificial data in order to conduct empirical analyses such as testing out a hypothesis or estimating specific statistics. Indeed, it is not always possible or it could be difficult to obtain results and conclusion using algebra, when the distribution of the data analyzed is not known or when statistics of interest are complex. Thus, a way to overcome this challenge is to use computer experiments relying on a large number of repeated random sampling processes to obtain numerical values and results. Monte-Carlo experiments for density estimation (L'Écuyer and Puchhammer, 2021) and Bootstraps procedures for variance estimation (Efron, 1979) are concrete examples where simulations are powerful tools when algebra and equations become too complex to be solved.

Thus, a new perspective on simulations is that simulation processes can be used to create artificial data as synthetic data. For example, we could generate p independent vectors $X_1, X_2, \ldots, X_p$ of size N using a normal distribution generator process to obtain a synthetic data made of N synthetic units and p synthetic variables. In this example, the synthesizer could fix the values of the parameters to be used in the normal distribution generator process (i.e. the mean and the variance) without using any of real data. In other words, synthetic data can be generated from 'scratch' without any disclosure risks. Here the analytical value is considered to be null in the sense that no attributes of the original data has been preserved in the synthetic data. Actually, this type of synthetic data are often called "dummy" files within NSOs. However, it is important to realize that these dummy files can nonetheless be useful, depending on the users' needs. For example, if the goal is to test processes without any regards to the values or relationships between variables, existing in the original data. This type of simulation processes is easy to implement and is not time-consuming and, of course, other types of statistical distributions could be used.

The synthesizer could also decide to use information from the original data, in the generation process, to ensure that some of the analytical value is preserved. For example, if we use our previous example, the synthesizer could have decided to generate the data is such a way that the parameters used to generate each of the p variables are actually the estimates observed in the original sample for each of the original variables. Thus, we would generate one synthetic variable per original variable using the estimated mean and variance observed for this original variable. In that case, the shape (or distribution) of the original variables might not be preserved (if the original variables do not follow a normal distribution) but the estimated means and variances for each of the synthetic variables would be the same in the original and synthetic datasets. Also, because the variables would have been generated independently, the relationships between them would not be preserved.

In fact, the idea is that simulation processes can be refined by using more or less in depth information extracted from the original microdata data to preserve more or less detailed statistical properties. For instance, the Fleishman-Vale-Maurelli method derived from Fleishman (1978) and Vale and Maurelli (1983) is an approach that uses information from the original data to generate multivariate nonnormal distributions with specific features preserved such as intercorrelation between variables and marginal (univariate) means, variances, skews and kurtoses. This method is well suited to capture correlations between continuous variables. Options for capturing relationships between categorical variables include drawing values from the estimated multinomial equation or classification and regression tree approaches to create safe categorical variables.

Table 3: Pros and cons of simulated data

Pros	Cons
Simulation processes are often easy to understand and can create completely safe data when no information pertaining to the original data is used. For more advanced types of simulations, some analytical value can be preserved.	Usually, does not allow to meet complex analytical needs.

Tools to apply the method

In general, simulation processes can be programmed easily enough using any software. For a more complex type of simulation, the R package semTools (https://CRAN.R-project.org/package=semTools) simulates microdata using the co-variance matrix, skewness and kurtosis from the original sample data (Jorgensen et al. 2019).

In practice: Australian Bureau of Statistics

In 2017 the Australian Bureau of Statistics (ABS) has started using a serverless architecture (cloud) provided by the Amazon Web Server (AWS) for some of their projects. In order to explore emerging tools on their AWS, the ABS have generated completely safe synthetic datasets, using simulations, that can be send out to the cloud to explore ML methods before obtaining approvals to data assets.

Deep learning

Deep learning is a subset of machine learning and is a growing genre in the Data Science and Artificial Intelligence arenas. These methods are becoming more popular in the field of synthetic data because synthesizers are dealing more and more with large, unstructured data. At the writing of this guide, the deep learning method to generate synthetic data used in practice by NSOs is Generative Adversarial Methods.

Generative Adversarial Methods (GAN)

With improvements in technology and computational capacity, implementation of machine learning processes have become easier and more accessible. Thus, it is natural that machine learning approaches have been more and more employed to generate synthetic datasets. More specifically, the use of deep learning models has become appealing because of their capacity to extract from big datasets very powerful predicting model.

The generative adversarial network (GAN) (Goodfellow, et al., 2014) is a prominent generative model used for synthetic data generation. The model tries to learn the underlying structure of the original data by generating new data (more specifically, new samples) from the same statistical distribution as the original data. Because the theory and implementation processes related to deep learning and neural networks can be technically challenging, we will mainly explain the overall concepts, as more information can be found in the references.

The main idea is to use GAN as a game-theoretic approach to synthesising new data and the model can be seen as it learns to generate from training distribution through a two-player game. The main idea behind a GAN is to have two competing neural network models. One is called the Generator and takes noise (or random values) as input and generates samples. The other model, the Discriminator, receives samples from both the generator and the training data, and attempts to distinguish between the two sources. In other words, the Discriminator is similar to a binary classifier that would take as input real (or original) data but also generated (or synthetic) data and would compute a pseudo-probability value that would be compared to a fixed threshold value. Under the threshold the data is considered being generated and above the threshold the data is considered being real. The training process is an iterative one where the two networks play a continuous game during which the Generator is learning to produce more and more realistic samples, and the Discriminator is learning to get better and better at distinguishing generated data from real data. This co-operation between the two networks is expected for the success of GAN where they both learn at the expense of one another and attain an equilibrium over time.

Because the GAN relies on neural network, that means that the approach can be used to generate discrete, continuous or text synthetic data.

Table 4: Pro and cons of GANs

Pros	Cons
GAN has been used in NSOs to generate continuous, discrete but also text datasets, while ensuring that the underlying distribution and patterns of the original data are preserved. Furthermore, recent research has been focused on the generation of free-text data which can be convenient in some specific situation where models need to be developed to classify text data.	GAN can be seen as complex to understand, explain or implement when there is only a minimal knowledge of neural networks. There is often a criticism associated to neural networks as lacking in transparency. The method is time consuming and has a high demand for computational resources. GAN tends to suffer from mode collapse and lack of diversity. Modelling discrete data can be difficult for GAN models.

With the drawbacks of GAN methods, autoregressive models to generate synthetic data have become more popular in the literature, though have not yet been utilized by NSOs. Autoregressive models have seen advantages in performance over GAN models, particularly for tabular data. They have seen to provide more flexibility, allowing the synthetic data generator to be more easily training on data with missing values or missing columns. Autoregressive models are more advantageous than GAN models because distribution models so they can give the likelihood of new observations (Leduc and Gislain, 2021).

Tools to apply the method

There is no specific tool per se to apply the method. However, Kaloskampis et al (2020) provides detailed information on the method and how to implement it in the NSO context. GAN methods for synthetic data has been explored in the literature, with many authors proposing solutions to some of the GAN drawbacks. Choi et al. (2017) propose an autoencoder and then uses the decoder as the final part of the generator. Park et al. (2018) introduces convolutions in the generator, Jordan et al. (2018) uses the PATE framework to generate differentially private synthetic data while Xu et al. (2019) introduce a node specific normalization and conditional sampling to tackle mode collapse. There also exists a useful benchmark framework called SDGym (https://github.com/sdv-dev/SDGym) for comparing tabular data generative models.

In Practice: Data Science Campus, Office for National Statistics

The ONS Data Science Campus is exploring using synthetic data to replace sensitive real data in the case of testing their Census 2021 system. To communicate their work, Kaloskampis et al (2020) published a study on generating synthetic data sets based on the US Census Income dataset available on the UCI repository (https://archive.ics.uci.edu/ml/datasets/adult). The dataset contains numerical and categorical variables including socio-demographic information and variables related to income (such as working status and income itself). GAN was used in a binary classification model context in order to generate new synthetic data. More specifically, the idea was to train a GAN algorithm to predict whether the income of an individual exceeds $50,000 per year based on some of the variables available in the original dataset. Here, Income is the target variable and the Generator provides synthetic values at the end of the process.

Note on data synthesis including survey features

As of today, most of the research related to data synthesis has been focused on census datasets and data products taken from administrative data. However, lately, a new direction has been taken in terms of research to notably explore how to synthesize samples drawn from a finite population. In other terms, with NSOs collecting data via many projects relying on probabilistic surveys, a natural question which arises is: how do we include survey design features in the data synthesis process.

Most of the other methods presented in this chapter work under the assumption that the original data cover the entire population of interest or is drawn via un non-informative sampling process from a finite population (such as a simple random sampling procedure). In other words. If the original data consists of sample from a finite population, the methods usually aim at synthesizing the original data without regard for the finite population from which it was drawn. Therefore, statistical conclusions obtained via the synthetic file can only be comparable to the ones obtained in the original *sample* and not necessarily the original *population*, especially when the sampling process follows an informative design (Lavallée and Beaumont, 2015). Thus, a way to address this would be to incorporate information of the sampling process in the data synthesis procedure in order to obtain a synthetic dataset from which we can estimate characteristics of the original *population*. This section aims at summarizing some of previous discussions with experts that have proposed some strategies for generating synthetic surveys weights or incorporating weights in the data synthesis process.

First of all, it was a consensus that it was unwise to release *original* sampling weights, even if the rest of the variables were to be synthesized. Indeed, weights can often be a fingerprint of the information that went into producing them notably because if all units in a given demographic slice has the same distinct weight, then synthetic records with that weight must have originated from data in that demographic group. In that case, if an attacker knows the mapping between weights and groups, they can track back to the original information. In this case providing synthetic weights or including weight in the data synthesis process can help prevent this problem by removing the exact fingerprint mapping between the weights and the original demographic group. Alternatively, weights can be inherently very distinct, this is more commonly an issue with business data, where large businesses may have weights that look unlike every other business. In this case, there is a risk that even the synthetic weights may be used to reidentify outlier businesses assuming the synthetic weights resemble the original weights (and so the synthetic versions of the weights are also identifying outliers). In this case, differential privacy might be a solution or, potentially, this could be a good case for including weights directly in the data synthesis process in such a way that no weights would be released to users. In either case, it is important to realize that disclosure risks of releasing synthetic weights can be estimated the same way we estimate the disclosure risk of any other variable.

Then two types of strategies were raised:

1. Generate and provide synthetic weights to users
    a. Treat the weights as a variable to be synthesized among all the others
    b. Synthesize the design variables and recalculate synthetic sampling weights based on them and the original sampling design
2. Use weighted models to approximate distributions from the original *population*, and generate values from them (in this case no weights would be provided to users)

1. Generate and provide synthetic weights to users

Some users might want to have synthetic weights made available with the rest of the synthetic variables, especially if the goal is to explore the synthetic dataset while waiting access for the original dataset that is expected to include weights.

a. Treat the weights as a variable to be synthesized among all the others

The idea was to consider survey weights as any other variables to be synthesized. This approach was the most unexplored one by the researchers who collaborated for the creation of this guide. One of the main concerns was to ensure that the synthesized weights need to be coherent with the synthesized design variables as well as all the other variables.

b. Synthesize the design variables and recalculate synthetic sampling weights based on them and the original sampling design

Here, the goal would be to synthesize all variables involved in the sampling design and recalculate synthetic weights in the same way that they were calculated at the design stage of the survey. However that covers mainly the calculation of *basic* sampling weights. Usually, in practice, basic weights are then being processed to take into account calibration, non-response and even smoothing in some occasions. For example, for calibration, it could be possible to adjust the synthetic weights to ensure that some weighted totals of the synthetic file correspond to known totals from the original population. Also, for non-response adjustment, a solution could be to re-weigh for total non-response with the synthetic weights. In that case, it would be important to synthesize total non-response indicator variables in the synthesis process and then use those to re-weight synthetic weights with traditional methods.

In general, providing synthetic weights means that a lot of attention should be paid to the original survey design and where those weights are coming from. In addition to that it is important to realize that it also limits the set of synthesis approaches that should be used. Indeed, recalculating synthetic sampling weights might be more complicated with more complex survey designs. Thus, it is important to understand users' needs, to ensure that the effort put in the synthesis process truly matches the analytical needs of the requirement.

2. Use weighted models to approximate distributions from the original *population*, and generate values from them (in this case no weights would be provided to users)

The pseudo likelihood method was suggested in this case. This method can be considered as an example of an advanced simulation process that generate data of high analytical value. The idea is to preserve as much all links between variables and univariates statistics as they exist in the original *population*.

Pseudo likelihood

The pseudo likelihood method generates synthetic populations by incorporating survey weights into the models based on the pseudo likelihood approach (Kim et al, 2020). The idea is estimate the distribution of the finite population. Once that the finite population density is estimated, the synthesizer can generate fully synthetic populations by drawing values repeatedly from it. This notably requires to derive the full conditional distributions of the Markov chain Monte Carlo (MCM) algorithm for posterior inference by using the pseudo likelihood function.

Table 5: Pros and cons of Pseudo Likelihood Method

Pros	Cons
Addresses informative sampling. When generating synthetic populations, the sampling process is already accounted for; thus, the uncertainty introduced by sampling process is also accounted for. Providing synthetic populations can be better than providing synthetic samples and can be more convenient (no need for original survey weights and no need to estimate sampling variance).	There are potential challenges with the choice of prior distribution for the MCM algorithm.

Tools to apply the method

There is no known tool per se to apply the method. However, section 2.1 and 2.2 of Kim et al (2020) provides detailed information on the method and how to implement it.

In Practice: Division of Statistics and Data Science, University of Cincinnati

In the study conducted by Kim et al (2020), synthetic datasets were generated using the 2012 Economic Census, which is the US Government's official five-year measure of American business and economy. More specifically, synthetic data were generated from data of three different industries, each representing a specific economic sector. The Economic Census data collection process was conducted using a stratified systematic sample in which all large establishments were included and other establishments were sampled within strata defined by their NAICS industry and the state.

Recommendations

This section provides recommendations on synthetic data generation methods depending on the use case as well as the requirements for the synthetic data. Table 6 links the recommended methods with the use cases presented in chapter 2. Figure 3 illustrates a method decision tree to help practitioners identify the method most adapted to their project.

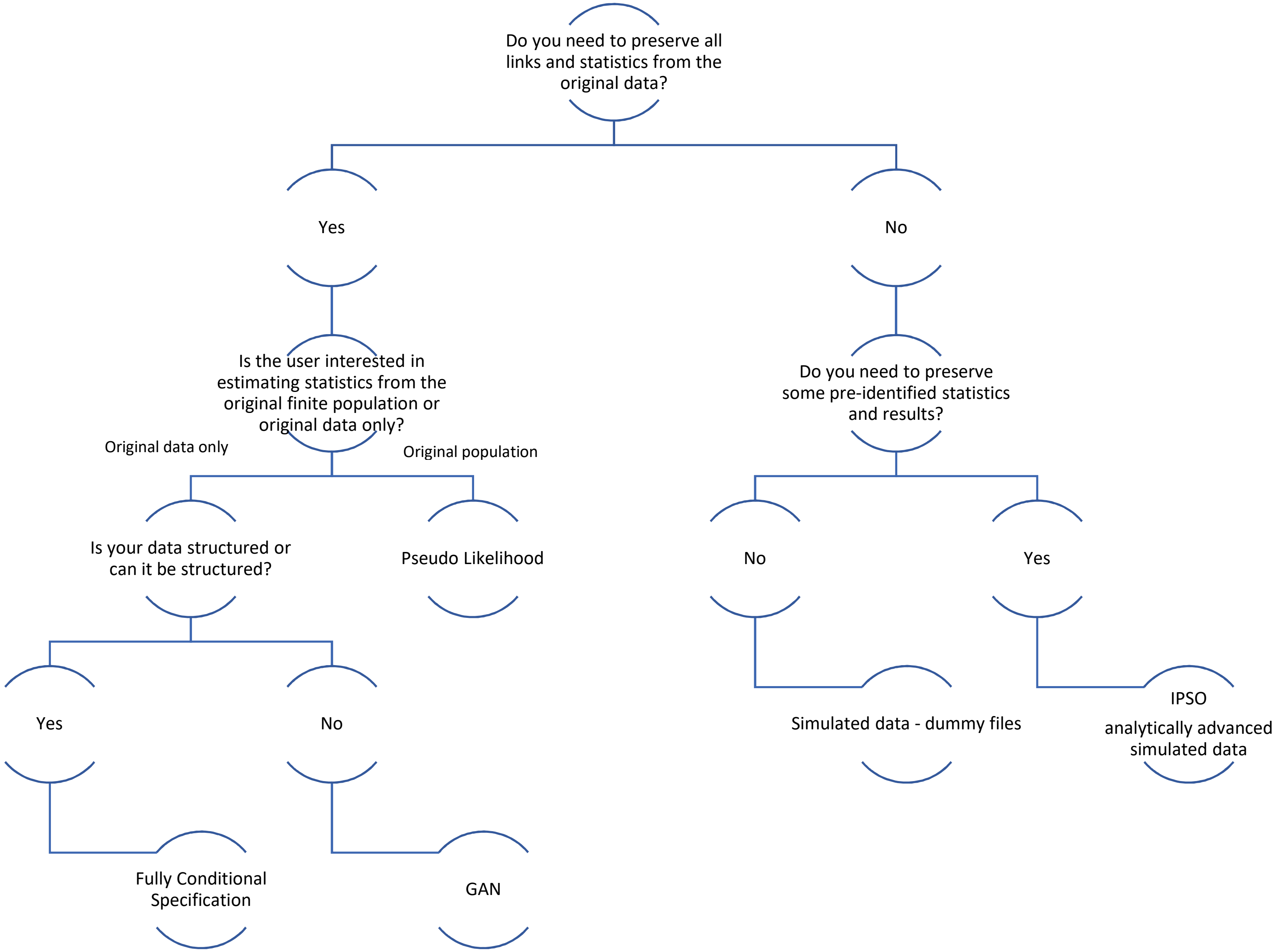
Table 6: Method recommendations by use case

		Use Cases			
Methods		**Releasing synthetic microdata to the public & Testing analysis**	**Education**	**Testing Technology**	**Comments**
Sequential Modelling	Fully Conditional Specification	Recommended	Can be used. If analyses conducted and statistical conclusions are pre-determined it might be too time-consuming in comparison to other methods.	Can be used but might be too advanced in comparison to the real analytical need	This method aims, in theory, at preserving all links between variables from the original data. Disclosure risk and analytical value need to be evaluated according to the release process.
	IPSO	Recommended if the analyses are all related to linear regressions this method, otherwise not recommended.	Can be used. Highly depends on the context: Recommended if the analyses are all related to linear regressions this method is recommended, otherwise not recommended.	Can be used but might be too advanced in comparison to the real analytical need	This method preserves results and statistics related to linear regressions specifically. Disclosure risk and analytical value need to be evaluated according to the release process.

Simulated Data	Dummy files	Not recommended	Can be used if training does not require analytical value in the data	Recommended	This method does not preserve any analytical value from the original data but is easy and quick to implement. Data is totally safe.
	Analytically advanced simulated data	Recommended if analyses conducted are related to the pre-identified results that needed to be preserved in the synthesis process. Otherwise, not recommended.	Can be used. Recommended if analyses conducted are related to the pre-identified results that needed to be preserved in the synthesis process. Otherwise, not recommended.	Can be used but might be too advanced in comparison to the real analytical need	This method only preserves pre-identified results and statistics. Disclosure risk and analytical value need to be evaluated according to the release process.
	Pseudo Likelihood	Strongly recommended if user wants to estimate statistics from the original finite population. However not recommended if user is expecting synthetic weights.	Can be used. If analyses conducted and statistical conclusions are pre-determined it might be too time-consuming in comparison to other methods.	Can be used but might be too advanced in comparison to the real analytical need	This method aims, in theory, at preserving all links between variables from the original population. Disclosure risk and analytical value need to be evaluated according to the release process.

Deep Learning	Generative Adversarial Network	Recommended especially in presence of text or unstructured data.	Can be used. If analyses conducted and statistical conclusions are pre-determined it might be too time-consuming in comparison to other methods.	Can be used but might be too advanced in comparison to the real analytical need	This method aims, in theory, at preserving all links between variables from the original data. Only method that handles unstructured and text data. Disclosure risk and analytical value need to be evaluated according to the release process.

Figure 3: Method decision tree


Do you need to preserve all links and statistics from the original data?
Yes
No
Is the user interested in estimating statistics from the original finite population or original data only?
Original data only
Original population
Is your data structured or can it be structured?
Pseudo Likelihood
Yes
No
Fully Conditional Specification
GAN
Do you need to preserve some pre-identified statistics and results?
No
Yes
Simulated data - dummy files
IPSO
analytically advanced simulated data

Chapter 4: Disclosure considerations for synthetic data

The synthetic data generation process breaks the chain between the data disseminated and the information collected. Due to this transformation, disclosure control of synthetic data is still being explored, as the need to add disclosure protection on top of synthetic data comes into question.

According to the Organization of Economic Cooperation and Development (OECD, 2003), disclosure relates to the inappropriate release of data or attribute information of an individual or an organization. Disclosure risk is the risk or possibility of disclosure occurring. There are two main types of disclosure: identification disclosure and attribution disclosure. Attribute disclosure can occur when an intruder may infer from mere knowledge that a population unit is represented in a data set, however the intruder does not possess the combination of attributes for the observation in question. Identification disclosure or identity disclosure can occur if an intruder identifies at least one respondent in the disseminated microdata. A small subset of variables in the data are used to make the linkage and once the linkage is successful, the intruder has access to all other information. This identification may lead to the disclosure of (sensitive) information about the respondent.

Although no record in a (fully) synthetic data file corresponds to a real person or household, there is concern that attribute and identification disclosure risk could still be present. Attribute disclosure could still be present if synthetic data can be linked to commercial or public data sources that then can provide more information on specific individuals. Identification disclosure could be present if unique observations found in the population are present in the synthetic data (Drechsler and Reiter, 2009). These situations could result in loss of reputation for the data holders and put at risk respondents' willingness to participate in surveys, census or provide their information by other means. Therefore NSOs may still decide to use additional disclosure controls in addition to synthetic data.

Privacy definitions and guarantees are context dependent (Dwork, 2011). NSOs should choose whether or not to implement additional disclosure controls on their synthetic data, as well as any specific privacy preserving techniques based on their own legislative and operational frameworks. Common privacy preserving techniques can be used on synthetic as well as real data. This chapter presents overviews of privacy preserving techniques that fall into two categories: indistinguishability based techniques such as k-anonymity, l-diversity, t-closeness and differential privacy. The purpose of this chapter is to present the disclosure control options available to NSOs and their synthesizers.

K-anonymity

K-Anonymity is one of the most known privacy preserving techniques. K-anonymity method gives K-level of anonymity to data, which means the information for each record contained in the release cannot be distinguished from at least k-1 other records whose information are in the data. Records can be associated to each other by certain identifying attributes, such as age, gender and location, in the case of say census or medical records. Breaches or attacks can occur when these attributes, called quasi-identifiers, can be linked with external data to identify unique records in the population (Machanavajjhala et al. 2007). K-Anonymity model distorts quasi-identifier values so that no record is uniquely identifiable from a group of k records. The Parameter k indicates the degree of anonymity (Sweeney 2002).

There are two main methods to achieve k-anonymity: suppression and generalization. Suppression is a method of ensuring privacy by selectively hiding the confidential information before disclosure. Cell suppression is a main method of data suppression. Under this methodology all sensitive cells are suppressed from publication, sometimes including non-sensitive cells as complementary suppression to obscure the values of the sensitive cells.

Generalization coarsens an attribute to a more general value (Lefevre et al, 2006). This creates groups of individuals that share the same generalized attribute value. There are two types of generalization that can be done, full domain vs local generalization. Full domain generalize all values of an attribute to the same level. Local generalization generalize values of an attribute to different levels.

For an example a k- anonymity, consider Table 7: Example of records with sensitive medical record Table 7, where there are 12 individuals with a record of their medical condition. The medical condition of each individual is considered sensitive, meaning that an adversary must not be allowed to discover its value. Whereas, neighbourhood (as defined by the first three digits of a postal code called Forward Sortation Area or FSA), age and occupation are considered non-sensitive. In this example, the quasi-identifier is the combination of Postal Code, Age and Occupation attributes.

Table 7: Example of records with sensitive medical record information

	Non-sensitive			**Sensitive**
	FSA	Age	Occupation	Medical Condition
1	A1A	27	Teacher	Heart Disease
2	A1B	28	Electrician	Diabetes
3	A1C	29	Teacher	Cancer
4	A1D	24	Doctor	Cancer
5	C3E	35	Teacher	Cancer
6	C3E	37	Electrician	Diabetes
7	C3R	40	Doctor	Heart Disease
8	C3O	40	Teacher	Diabetes
9	C2R	50	Electrician	Cancer
10	C4M	48	Doctor	Heart Disease
11	C8S	49	Doctor	Heart Disease
12	C8Z	50	Teacher	Cancer

Table 8 contains an example of k-anonymity achieved by utilizing generalization and suppression (denoted by *) techniques. For instance, we aim to achieve 4-anonymity. This means that the values for FSA, age, and occupation of the individual records should be generalized in such a way that we can form equivalent classes with at least four records. The quasi-identifiers of these records should be indistinguishable from each other.

Table 8: 4-anonymous version of Table 7

	Non-sensitive			**Sensitive**
	FSA	Age	Occupation	Medical Condition
1	A1*	2*	*	Heart Disease
2	A1*	2*	*	Diabetes
3	A1*	2*	*	Cancer
4	A1*	2*	*	Cancer
5	C3*	≤40	*	Cancer
6	C3*	≤40	*	Diabetes
7	C3*	≤40	*	Heart Disease
8	C3*	≤40	*	Diabetes
9	C**	≤50	*	Cancer
10	C**	≤50	*	Heart Disease
11	C**	≤50	*	Heart Disease
12	C**	≤50	*	Cancer

The main strengths of K-anonymity are its simplicity and potential to protect against re-identification attacks. Re-identification attacks can either happen through linking records in datasets or through multiple queries to the same database to obtain relational inferences.

However, K-anonymity assumes that each record in a dataset represents a unique individual. If this is not the case, an equivalence class of K records does not necessarily link to K individuals with K-anonymity (Mendes et al. 2017).

Both attribute and identification disclosure are still at risk. There are three types of attacks that can be used against k-anonymity: unsorted matching attack, complementary release attack, and temporal attack (Sweeney 2002).

Both attribute and identification disclosures are still at risk. There are three types of attacks that can be used against k-anonymity: unsorted matching attack, complementary release attack, and temporal attack (Sweeney 2002).

Unsorted matching attacks is based on the order in which the groups appear in the released tables. For instance, if the released tables have the same order of the generalized groups, then a direct matching of groups across tables position can reveal sensitive information. This can be prevented by randomly sorting the order of the groups for the released tables.

Complementary release attack is based on finding quasi-identifiable attributes that are a subset of the attributes in complementary datasets that have been released. By obtaining quasi-identifiers through multiple datasets that are published, re-identification attack is possible. This can be prevented by doing a thorough inspection of external information.

Temporal attacks is based on temporal inference. Since a k-anonymity solution of a dataset at time *t* has no requirement to respect the k-anonymity solution of the same dataset at time *t+1*, then joining the two k-anonymity solution datasets can release sensitive information. This is prevented by doing a thorough inspection of external information and adjust for potential quasi-identifiers as well.

ℓ-Diversity

As mentioned above, there is still the potential for an attacker to identify information, even if k-anonymity is met. ℓ-diversity is an extension of K-anonymity. It requires every equivalence class to abide by the ℓ-diversity principle. An equivalence class is ℓ-diverse if at least ℓ "well-represented" values exist for the sensitive attributes (Machanavajjhala et al. 2007).

Continuing the example from the previous section, **Error! Reference source not found.** reflects a 3-diverse version of table 8.

Table 9: 3-diverse table of medical condition data

	Non-sensitive			**Sensitive**
	Postal Code	Age	Occupation	Medical Condition
1	A1*	≤30	*	Heart Disease
2	A1*	≤30	*	Diabetes
3	A1*	≤30	*	Cancer
4	A1*	≤30	*	Cancer
5	C**	≥30	*	Cancer
6	C**	≥30	*	Diabetes
7	C**	≥30	*	Heart Disease
8	C**	≥30	*	Cancer
9	C**	≤50	*	Heart Disease
10	C**	≤50	*	Heart Disease
11	C**	≤50	*	Cancer
12	C**	≤50	*	Diabetes

Distinct 3-diversity means that each equivalence class should contain at least three distinct values for the sensitive variable "medical condition": Heart disease, Cancer, Diabetes. Thus, there should at least be three records in each equivalence class.

ℓ-Diversity increases privacy protection compared to K-anonymity (Li et al. 2007). It protects against attribute disclosure and addresses the vulnerability to unsorted matching attacks and background attacks (Machanavajjhala et al. 2007). However, there are downsides to ℓ -diversity. For starters, it may be difficult and not necessary, say if the sensitive attribute is binary like in the case of sex – male and female. The problem is exacerbated when there is a small number of one of the binary sensitive attributes (Li et al. 2007). In addition, ℓ -diversity may not be enough to prevent attribute disclosure in cases where the distribution of the sensitive variables is skewed or the sensitive variables are very similar to each other e.g., income values.

t-Closeness

t-closeness provides privacy in cases where K-anonymity and ℓ-diversity fail to do so (Li et al. 2007). t -closeness requires the distribution of the sensitive values in each equivalence class to be "close" to the corresponding distribution in the original table.

In other words, as presented by Li et al. (2007), t-closeness is based on the premise that a user has a prior knowledge of the sensitive attributes of a record. By using the prior knowledge about the individual record's sensitive attributes (β_0), along with prior knowledge of the distribution of those sensitive attributes in the population, the user can form a belief for that individual record (β_1). Then, once the user has access to the released table, they can use their knowledge to identify the corresponding class the record is in and to learn more about the distribution of those sensitive attributes in that class. This provides the user more information on the individual record (β_2).

t-closeness aims to reduce the difference between β_1 and β_2. To do so, it is assumed that the distribution of the sensitive attributes in the population is in fact a public knowledge. With t-closeness, information is released in such a way that a user can learn very little additional information about an individual record. This means that the goal of t-closeness, is to have the distribution of sensitive information in the population and that of any class, as close as possible.

Li et al. (2007) state that the t-closeness principle states is:

> *"An equivalence class is said to have t-closeness if the distance between the distribution of a sensitive attribute in this class and the distribution of the attribute in the whole table is no more than a threshold t. A table is said to have t-closeness if all equivalence classes have t-closeness."*

Methods to measure the difference in distribution include variational distance, Kullback-Leibler distance, and Earth Mover's (EMD) distance.

t-closeness takes as input a table or dataset $T(A_1, A_2 \dots A_N)$, a parameter $\mathcal{K}$ specifying the minimum cluster or group size and a value for t. The output is a table T', a set of clusters/groups satisfying K-anonymity and t -closeness. It works as follows (example from Soria et al. (2015) the Standard Microaggregation and Merging algorithm measures the distance using EMD):

1. Apply microaggregation to t with minimum cluster size $\mathcal{K}$, store the output in T'.
2. While distance between $T'$and T is larger than t:
    a. Choose the cluster in T' with the greatest distance w.r.t. t and store this cluster in $\mathcal{C}$.
    b. Choose the cluster in T'closest to $\mathcal{C}$ in terms of key variables, store this cluster in $\mathcal{C}'$
    c. Merge $\mathcal{C}$ and $\mathcal{C}'$ in T'.

Continuing with the same example but add more categories to the sensitive variable Disease.

Table 10: Table of medical condition data that has 0.25 -closeness with respect to Disease

	Non-sensitive			**Sensitive**
	Postal Code	Age	Occupation	Medical Condition
1	A1*	≤30	*	Heart Disease
2	A1*	≤30	*	Diabetes
3	A1*	≤30	*	Pneumonia
4	A1*	≤30	*	Bronchitis
5	C**	≥30	*	Cancer
6	C**	≥30	*	Diabetes
7	C**	≥30	*	Heart Disease
8	C**	≥30	*	Pneumonia
9	C**	≤50	*	Heart Disease
10	C**	≤50	*	Bronchitis
11	C**	≤50	*	Cancer
12	C**	≤50	*	Diabetes

Consider the sensitive variable of Table 10. The values of Disease are given as:

$Q' = \{Heart\ Disease, Diabetes, Pneumonia, Bronchitis, Cancer, Diabetes, Heart\ Disease,$
$Pneumonia, Heart\ Disease, Bronchitis, Cancer, Diabetes\}$.

In the set Q, there is three instances of *heart disease*, three instances of *Diabetes*, two instances of *Pneumonia*, two instances of *Bronchitis,* and finally two instances of *cancer.* Accordingly, the distribution of these categories over Table 10 is $Q = \{\frac{3}{12}\ HD, \frac{3}{12}\ D, \frac{2}{12}\ P, \frac{2}{12}\ B, \frac{2}{12}\ C\}$. In the first equivalence class of Table 10: $P_1\{Heart\ Disease, Diabetes, Pneumonia, Bronchitis\}$ there is one occurrence of each of "*heart disease", "Diabetes", "Pneumonia" and "Bronchitis"* yielding a distribution of $P_1 = \{\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, 0\}$. Given that the Disease is a categorical variable, we will apply the EMD distance. We recall that EMD stands for Earth Mover's Distance: $E\ (\boldsymbol{P}, \boldsymbol{Q}) = \frac{1}{2} \sum_{i=1}^{m} |\ p_i - \ q_i|$, a measure that is equal to one-half of the Manhattan distance.

So, *t*-closeness for the Disease is calculated as follows:

$E\ (P_1, \boldsymbol{Q}) = \frac{1}{2}\ [\left|\frac{1}{4} - \frac{3}{12}\right| + \left|\frac{1}{4} - \frac{3}{12}\right| + \left|\frac{1}{4} - \frac{2}{12}\right| + \left|\frac{1}{4} - \frac{2}{12}\right| + \left|0 - \frac{2}{12}\right|\quad] \approx 0.166$. Then, $P_2 = \{\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, 0, \frac{1}{4}\}$ and $E\ (P_2, \boldsymbol{Q}) =$ $\frac{1}{2}\ [\left|\frac{1}{4} - \frac{3}{12}\right| + \left|\frac{1}{4} - \frac{3}{12}\right| + \left|\frac{1}{4} - \frac{2}{12}\right| + \left|0 - \frac{2}{12}\right| + \left|\frac{1}{4} - \frac{2}{12}\right|\quad] \approx 0.166$. Finally, $P_3 =$ $\{\frac{1}{4}, 0, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\}$ and $E\ (P_3, \boldsymbol{Q}) =$ $\frac{1}{2}\ [\left|\frac{1}{4} - \frac{3}{12}\right| + \left|0 - \frac{3}{12}\right| + \left|\frac{1}{4} - \frac{2}{12}\right| + \left|\frac{1}{4} - \frac{2}{12}\right| + \left|\frac{1}{4} - \frac{2}{12}\right|\quad] \approx 0.25$. Thus, $t =$ $\max(0.166, 0.166, 0.25) = \boldsymbol{0.25}$.

For more detail on *t*-closeness, see Dosselmann et al. (2019).

Differential Privacy

Differential Privacy (DP) was introduced in 2006 in computer science by Cynthia Dwork, Frank McSherry, Kobbi Nissim and Adam Smith. It is only recently that DP has made its way to official statistics and DP-compliant methods are playing an increasingly important role in today's digital world. While DP's terminology is well established – making use of computer science terms such as databases, mechanisms and queries – in this document we adopt a more statistical language to convey the same notions.

Contrary to popular belief, DP is neither a method nor an algorithm but a definition supporting a mathematical disclosure-control framework. Thus, despite its name, the intended use of DP in official statistics is to prevent disclosure when *releasing statistical information* rather than to address privacy concerns when *gathering personal data* from individuals. As such, DP presents itself as an alternative to the traditional disclosure control framework which has been used by NSOs for decades now.

The two main vintages of DP are: the one-parameter ε-DP and the weaker two-parameter (ε–δ) DP. In this note we focus on the more stringent ε-DP, which is the form people usually have in mind when referring to DP. Where applicable, DP can be used by a data custodian to provide explicit and mathematically provable disclosure protection guarantees *when releasing statistical information derived from personal data*. DP is the first disclosure control framework explicitly stating the type and degree of protection it is offering.

What is a DP_ε -compliant Method?

Any method applied to a dataset *D* for the purpose of releasing statistical information – either in the form of tables or datasets (including those of a synthetic kind) – is at risk of disclosing some of the personal information *D* contains. Such a method *M* is called (ε-)differentially private or DP_ε-compliant when it meets DP's mathematical requirements which impose restrictions on the *type* of disclosure that may occur.

Furthermore, the privacy parameter ε – whose value in practice is set by the data custodian – determines the *degree* of disclosure protection a DP_ε-compliant method *M* is offering by means of the upper limit it imposes on the amount of person-level information *M might* be disclosing. More explicitly, the data custodian controls through ε the (average) amount of suitable random noise used by *M* to produce its outputs, which incidentally also impacts their utility: the degrees of protection and utility are closely linked. Indeed, the larger (smaller) the value set to ε the less (more) noisy M's outputs become and, generally speaking, the greater (lesser) the disclosure risk they pose and the greater (lesser) their utility[5].

[5] ε is often referred to as the "privacy loss parameter", with small values leading to less loss of privacy, and larger values leading to greater losses of privacy.

Why the name, Differential Privacy?

To better understand DP we need to state the problem it has been introduced to solve. In preamble suppose you were told that John’s height is 10 centimetres (or 4 inches) more than the Canadian average for men. Now should you learn from a dataset of national heights, say, that the Canadian average is 1m78 (or 5 feet 10 inches) then you would correctly conclude that John is 1m88 (or 6 feet 2 inches) tall.

Note that, and this is key, John’s height may have become known to you *without him ever having participated* in the data collection of the national heights dataset. In the situation where indeed John did not contribute his information to the dataset, then its custodian can hardly prevent disclosure from happening since it is directly due to *external* factors such as what people (like you) already know about John. After all, while it is certainly John’s prerogative to not contribute information to this dataset, he cannot prevent *others* from contributing their own information which may bring the dataset to play an accessory role in a disclosure event involving him.

By thinking of examples such as this one, the authors of DP posited that disclosure protection guarantees cannot be absolute in nature: simply put, a method ought not to be held responsible for a disclosure event that *could have occurred without* the concerned individual ever having contributed to the dataset itself. Instead, they proposed assessing a method’s role in a disclosure event differentially *by comparing* what it can reveal about an individual *when their information is present* in the dataset *to when their information is not included*.

In effect, this differential assessment limits *M*’s role in a disclosure event to what it can reveal *as a result* of the participation (or absence thereof) of the concerned individual to the dataset. Hence, if *M*’s outputs were about the same regardless of whether John had participated to the dataset or not, then *M* would hardly be at risk of disclosing something specific about John since his contributions have little influence on *M*’s outputs. This will become clearer later when examples of a DP$_\varepsilon$-compliant method are presented.

What are exactly the mathematical requirements underlying Differential Privacy?

The concerns just expressed over the role any one piece of information might play in determining *M*’s output lead to DP’s mathematical requirements. But first, those concerns must be rephrased in the language of datasets that *M* actually understands. To that end, consider *two* datasets that are identical *except* for a single piece of information one has and the other not; in DP’s parlance these are called adjacent datasets. Then, *for a given* output of the method *M and a given pair* of adjacent datasets *D* and *D*’, determine *how more likely* is this output value to occur when *M* is applied to *D* instead of *D*’.

The relative likelihood assessment just described involves a *specific* combination of an output *and* a pair of adjacent input datasets. DP-compliance requires doing the same, but this time for an *arbitrary* combination of an output value and a pair of adjacent datasets. More specifically, *M* is DP$_\varepsilon$-compliant *if* an output is no more likely to occur than the limit set by the data custodian through the privacy parameter ε when *M* is applied to *D* rather than to *D*', *irrespective of which one output and which one pair* of adjacent datasets *D* and *D*' are considered.

This is DP's way of sizing up the influence *any one individual in the dataset* might have on *any one* of *M*'s outputs, as their personal information may be the only thing separating *D* from *D*'. This is not to say that the upper limit associated with a DP_ε-compliant method holds for *all* outputs *and all* pairs *simultaneously*; it rather means that it is holding for *any one arbitrary* output-pair combination, as opposed to holding just for one *cherry-picked* combination.

Furthermore, the upper limit calculated as part of DP's likelihood assessment has to hold *even if* the user already knows everything about the dataset's contents *except* for just one individual's contribution *and* knows the inner working of *M*. Thus, DP has what can be called a no-secrecy policy: the validity of its protection guarantee does not rely on users having little knowledge of both the dataset and the disclosure control method used, an assumption usually made for the protection offered to be most effective. On the contrary, DP's likelihood assessment is to be performed by assuming users actually have an understanding of *D* and *M* comparable to that of the data custodian. We will discuss this further below when examples of a DP-compliant method are presented. DP's no-secrecy policy is a major selling point and explains to a large extent practitioners' interest in DP as a disclosure control framework.

What does it mean to have formal disclosure protection guarantees?

The explicit and mathematically provable protection guarantees DP offers characterize this framework. While such formal guarantees are a novelty in a disclosure control context, they are a familiar occurrence in probability sampling: statistical bias and variance are examples of formal *quality* guarantees. For probability sampling and DP, formal guarantees stem from the mathematical requirements underlying each framework.

In contrast, it is usually *not* possible to formally assign bias and variance measures to estimates derived from nonprobability samples, which were en vogue in the 1930s before probability sampling was introduced and have recently been making a comeback. While a carefully designed nonprobability sample might allow for sound inferences to be made, it has traditionally been lacking the mathematical underpinnings needed to support formal quality guarantees such as bias and variance.

What does a DP_ε-compliant method look like?

Just like a probability sample *looks* the same as a nonprobability sample (both are just subsets of the population), a DP_ε-compliant method *looks* no different than any other method. The difference rather lies in how the samples and methods are *conceived*: both a probability sample and a DP_ε-compliant method result from a process designed to meet well-defined mathematical requirements from which they each inherit their formal guarantees. While some of the dissemination methods exploited under the traditional disclosure control framework already are DP_ε-compliant, most are not (although some can be modified to become differentially private).

Applications: two examples of a DP_ε-compliant method

In this section two examples of a DP_ε-compliant method are given for illustrative purposes; the first releases a numerical value and the other a dataset. In each case, the DP_ε-compliant method is derived from a non-compliant one to showcase its distinctive features.

First, consider the situation in which a data custodian is looking to release a count extracted from *D* after rounding it to the *nearest* 5, say. Hence, raw counts of 14 and 12 would get released as 15 and 10, respectively. This method is *not* differentially private. Indeed, a key theorem of DP (i.e., *a verifiable consequence of DP's own mathematical requirements*) states that no deterministic method can be differentially private. And rounding to the *nearest* 5 is such a method since, for instance, it *always* returns 15 when the raw count is 14 (and *always* returns 10 when the raw count is 12). Note that increasing the rounding base to 10 or 100 would not change the outcome: it is the deterministic 'nearest-to' feature of the rounding that poses the problem here, not the actual base value.

It is possible here (and instructive) to establish non-compliance directly, that is without appealing to DP's theorem. Consider a dataset *D* containing 13 yeses, which gets rounded *up* to 15, along with its adjacent *D*' obtained from *D* by dropping one of those thirteen records. The ensuing count of 12 yeses for *D*' would get rounded *down* to 10, allowing one to conclude *with certainty* that the dropped record was a 'yes'. The same argument, but expressed this time in DP's parlance, shows that the output value of 15 occurs with *probability one* when the rounding is applied to *D* (with its 13 yeses) but with *probability zero* when the rounding is applied to *D*' (with its 12 yeses). This can be paraphrased as saying that *M*'s output of 15 is "infinitely" more likely to occur from *D* rather than from its neighbor *D*', which is something DP does not allow to happen as "infinity" is larger than any pre-set limit.

From this we conclude that a method must have a random component to be differentially private. However, this is not a sufficient condition. For example, rounding a raw count to the multiple of 5 directly above or below it based on the flip of a coin (e.g., rounding 14 to either 10 or 15, each with probability ½) has a random component but is *not* differentially private. Indeed, consider *D* leading to a count of interest of 15. Since it is already a multiple of 5, it would get released as 15. Now consider *D*' such that its count is 15-1=14. Under the current random rounding scheme this count has equal chances of getting released as either 10 or 15. Because the output value 10 is *possible* for *D*' but *impossible* for *D*, their relative occurrence likelihood is "infinite" which again is something DP does not allow to happen.

In this situation a DP_ε-compliant method is Laplace's mechanism, which is obtained by adding noise generated from Laplace's *continuous* distribution to the raw count. For example, 15.199385… and 13.836519… are just two of the *countless* possibilities for Laplace's output to the same input value of 14. (Later we discuss a way of using Laplace's mechanism to release *whole* numbers, which is what users would naturally expect to get as released counts.)

Laplace's mechanism[6] was the first example ever given of a DP_ε-compliant method and it possibly remains the simplest one to this day. The variance of the Laplace distribution determines the (average) amount of noise that gets generated. Not surprisingly then, the privacy parameter ε of Laplace's mechanism is directly tied to the variance of the underlying Laplace distribution, which drives the degree of protection (and utility) conferred by the data custodian to Laplace's outputs.

Considering users' dislike of noise-adding methods, a welcome consequence of DP's no-secrecy policy is that a data custodian *can safely divulge* the variance of Laplace's distribution used (or equivalently, the value actually set to the privacy parameter ε). This is a departure from traditional disclosure control practices which call for such information to be kept secret in order for the protection offered to be most effective. Thus, under DP users are given the means to assess the statistical significance of the conclusions they are drawing by factoring in the (average) amount of noise that has gone into the outputs analyzed. It is important to note that in practice, accounting for the noise in the output is extremely difficult. To date there is a knowledge gap between NSOs and users on interpreting noisy results. Un-sophisticated users are often not comfortable at looking at negative counts. Some NSOs, for example the case of the Australian Bureau of Statistics, have found that there are significant investments in educating users so that they can be comfortable with non-additive tables.

For the second example of a DP_ε-compliant method, suppose a Yes-No question is administered to individuals using a Randomized Response (RR) method as a means of reducing the response bias due to the sensitive nature of the information gathered – see for instance Section 12.5 in Lohr (1999) for a short discussion of RR methodology. A RR method introduces plausible deniability[7] by altering any given response with probability p (which is known to the data custodian) before recording it in a dataset. Also, a RR method is to keep no trace of the reported answers nor of which ones were actually altered.

It can be shown that knowing p allows one to draw meaningful statistical conclusions about the true proportion of yeses from the *recorded* values alone. However, in the traditional disclosure control setting only the data custodian would be allowed to know p. In contrast, in the case of a DP_ε-compliant RR method the value for p could be safely divulged in accordance with DP's no-secrecy policy allowing users to make valid inferences as well.

We can see such a RR method as producing partially synthetic data, albeit of a very limited analytical kind. Indeed, not only are cross-relations among variables not captured by the method discussed here, but there is also legitimate ground to question its efficacy in addressing the initial bias concerns.

[6] While it is tempting to use the better-known Gaussian or normal distribution instead of Laplace's, this strategy does not *quite* lead to a DP_ε-compliant method since it only meets the requirements of the weaker two-parameter (ε–δ) form of DP.

[7] The notion that respondents can deny that the recorded answers are truly those they initially expressed.

For the sake of further illustrating how DP-compliance works, suppose the data custodian felt[8] that only yeses were sensitive and therefore needed to be protected. To this end a RR method is used whereby only some of the yeses are randomly altered: all reported noes are directly recorded in the dataset.

While economical in the amount of noise it uses, this RR method is *not* DP-compliant. To see why, consider a dataset as output of this method in which a 'yes' has been recorded for a certain individual. The custodian might argue that without knowing how the method works, a user cannot say *for certain* whether this value points to a reported 'no' instead of a reported 'yes'. However, this line of argumentation does not comply with DP's no-secrecy policy which requires assuming the user *does know* how the method works. And a user knowing that a reported 'no' cannot possibly be recorded as a 'yes' in the output dataset by this RR method would conclude *with certainty* – which is something DP does not allow to happen – that the individual had to have answered 'yes' to the question. In this situation DP-compliance requires randomly altering some of the noes as well *to complete* the masking of the yeses undertaken.

Differentially Private Data Synthesis

How can Differential Privacy be applied in the context of synthetic data? The simplest method of creating a DP synthetic data set is sometimes called the "histogram" method. It proceeds as follows:

- group all variables in the data into categories
- produce a cross-tabulation of all combinations (the histogram)
- Add Laplace noise to each cell of the table
- Round the results to whole numbers and recreate the original data

This is DP because an individual can only appear in one cell of this table. Unless there are a very small number of variables in the data set this gives synthetic data with very poor utility. For more variables, the histogram will contain many zero cells, particular if a decision is made to set all negative counts to zero, and can produce a data set with many more records than the original.

A more useful approach to creating DP synthetic data is to create DP sufficient statistics for the model used to generate the synthetic data. The composition theorem shows that the ε for the synthetic data is then the sum of the εs for all of the statistics used to define the model fit. The method in this class that has been used most often to create synthetic data is to define the model from a set of margins, each of which has noise added to make it DP. Additional computation is required in this case to make the margins consistent with each other. This is justified by the post-processing theorem (Bowen and Snoke, 2021) which states that once a DP result has been released, any subsequent transformation of these outputs as also DP. An important limitation of this method is that any exploration of the data to define the model to use for the synthesis must contribute to

[8] As argued above, deciding which information is to be protected ought to be a matter of the agreement passed between parties and not a matter of opinion as is the case here.

the privacy budget. If the model is selected from another similar data set, say from a previous year's version of the data release, then this information can be used freely without contributing to the privacy budget,

Another approach to creating synthetic data involves selecting a model for the synthetic data from a series of iterative steps, each of which is taken based on DP information. The idea is to improve the agreement between the synthetic data and the original at each step. These methods include the multiplicative weights algorithm (MWEM) and various methods based on generative adversarial networks (GANs). The ε from these procedures is the sum of those from each step. Unlike non DP iterative procedures the number of steps must be fixed in advance so as to maintain the DP property

Is a DP-compliant method always better than a non-compliant one?

The short answer is no, not always. Just like an ill-designed probability sample can lead to nonsensical conclusions (as exemplified by the famous tale of Basu's elephants9), an ill-devised differentially private method can do a poor job of preserving the personal information used as input. If the method is not well designed or the data is too difficult, the added noise can overwhelm the input, rendering the released outputs all but useless.

Previously we discussed the basic approach for added Laplacian noise to counts of individuals. Data synthesis is often performed by combining these privatized counts to parameterize or train a model which can generate new records in the original schema. Simple histogram models, probabilistic graphical models (PGM)s and generative adversarial networks (GAN) have all been used for this (Bowen and Snoke, 2021). Intuitively, for high utility we want the added noise to be small in comparison to the original count values, so the relative shift between the input and output counts is not large. More accurate counts lead to more realistic data models. But, due to differential privacy composition, the more counts we take over a given individual, the more noise we need to add to maintain the same level of privacy.

This means that challenging cases occur if a given data schema has a large number of variables (which require many counts for the model to capture), or the data has few records (meaning many counts are small), or if the variables have many possible values (meaning the data is spread sparsely across different options and many counts are small or zero). These conditions can pose significant problems for utility that researchers are actively working to overcome (Bowen and Snoke, 2021). However, when a data set has a large number of records, a small number of variables, and does not spread the data out too sparsely, existing DP synthetic data generators can produce very high quality data with robustly protected privacy. Preprocessing data can often improve performance by eliminating variables and reducing granularity on variables with large numbers of possible values.

[9] First told in Basu (1971), numerous other accounts of the story are widely available on the Web.

In addition, DP methods can fall short on protecting privacy as well. If the privacy parameter is excessively large and the method does minimal additional processing to the data, the provided protection can deteriorate to the point of being meaningless.

Also, a commonly-held belief has DP providing *unfailing* protection against disclosure of personal information, which it does not. For instance, DP does not prevent the effective disclosure protection from decreasing when multiple outputs involving the same individual are released. For example, the random noise used in Laplace's mechanism tends to cancel out when several of its outputs are averaged out to form a single result. On the bright side, not only does DP's Compositional Theorem warn data custodians of the compounded privacy loss incurred by making repeated use of a DP_ε-compliant method to release statistical information, it also *quantifies* that loss. It states that the composite privacy loss ε is the sum of the privacy losses for each output. This has given rise in practice to the notion of privacy budget which is used to closely monitor the situation: once the cumulative privacy loss incurred from a series of releases made from a dataset reaches the budgeted value set by the data custodian, access to the dataset is closed.

Some Implementation Considerations

A successful implementation of any new methodology to a specific context promises to be a challenge, as various practical constraints will bring issues not directly addressed by the theory. And looking to release differentially-private synthetic data is not an exception. However, attention to the following details will certainly help with the implementation of DP in such a context.

With respect to implementing data synthesis within a national statistical agency, Sallier and Girard (2018) recommend decomposing the required tasks into pre-processing, synthesis and post-processing steps. In their experience (which did not involve DP considerations), a successful implementation of data synthesis hinges on making informed, forward-looking decisions at the pre-processing stage of the project. Indeed, while it is tempting to rush straight into synthesizing a dataset, it is important to first reflect on the balance to be struck between utility and confidentiality. By avoiding making decisions based on utility alone, one prevents putting too much strain on synthesis to protect confidentiality, whether this is attempted under DP or not.

To illustrate, consider the following example. A dataset *D* contains information about families, including the number of siblings each contains. Clearly the participation of a family with ten siblings will have a greater impact on the *total number of siblings* than on the *total count of families*. Indeed, the latter statistic inherently offers greater anonymity since a family with ten siblings only contributes a value of 1 to the tally, as all other families do. Pre-processing involves deciding here whether information about family size remains in the dataset or not; and if it does, then cutting off its tail by resorting to an open ended category such as "more than 5 siblings" would reduce the burden put onto synthesis.

When designing a DP_ε-compliant method to meet practical needs, two important consequences of DP's mathematical requirements may prove very useful: the Compositional Theorem and the Post-Processing Theorem. We saw previously that the Compositional Theorem warns data custodians against the compounded privacy loss incurred by repeated use of a DP$_\varepsilon$-compliant method on a dataset. But the same theorem can also be used to manufacture a nontrivial DP$_\varepsilon$-compliant method by putting together two existing and simpler DP$_{\varepsilon/2}$-compliant methods. Thus, a complex DP$_\varepsilon$-compliant method need not be created from scratch, but it can rather be built using available or easier-to-design DP-compliant pieces.

We alluded to the Post-Processing Theorem before (although not by name) when evoking a way of releasing Laplace's outputs as integers; loosely put, it states that any transformation of the output of a DP$_\varepsilon$-compliant method is itself DP$_\varepsilon$-compliant *provided it is operating independently of the input dataset*. Thus, to get an integer out of Laplace's mechanism one simply needs to round Laplace's output to the *nearest* integer. In the same vein, the Post-Processing Theorem supports the practice of setting a negative Laplace's output to 0 prior to being released without upsetting its DP-compliance status. However, if one were to round Laplace's output to the nearest integer *only when* the raw count is smaller than 15, say, then this would violate the theorem's premise. Indeed, since Laplace's output itself does not reveal whether the original raw count was smaller than 15 or not, this rounding rule can only be implemented by first revisiting the dataset. But then because Laplace's mechanism is no longer the only way statistical information is getting released from the dataset, its previously-earned DP-compliance status gets revoked. (This is not a statement of non-compliance per se – it merely implies that DP-compliance of the combined Laplace-with-conditional-rounding method must be examined anew.)

But wait, why is it rounding Laplace's output to the *nearest* integer leads here to DP-compliance when we previously established that rounding a raw count to the *nearest* 5 (or 10, 100, etc.) is not itself a DP$_\varepsilon$-compliant method? The apparent contradiction is resolved by paying close attention to *what* gets rounded in each case. While DP says that rounding *a raw count* to the nearest integer is not effective at preventing disclosure of personal information from happening, rounding *Laplace's output* to the nearest integer poses no issue when it is solely done for practical reasons. Indeed, in this case Laplace is the one method responsible for protecting the personal data that has gone into the output, not the subsequent rounding performed. The concern here rather becomes whether the extra rounding performed can *undo* the protection already provided by Laplace's mechanism. DP says this will not happen *as long as* Laplace's mechanism remains the only way statistical information gets released from the dataset.

Why are DP-compliant methods not more widely used in official statistics?

DP is the first framework capable of addressing formal disclosure protection guarantees. However, the underlying mathematical requirements are often difficult to satisfy in practice and many practitioners actually find them too stringent to begin with. More specifically, they see the disclosure scenario underlying DP as too severe: a user will simply not know just about everything that is contained in the dataset as DP requires the custodian to assume through its mathematical requirements. As a result, they claim, DP's requirements unduly undermine the utility that can ever be attained. DP_ε -compliant methods may in fact not provide the required utility under different NSO's requirements. The DP framework was developed under cryptography literature and the main objective is to protect privacy. Protecting privacy is an important consideration for NSOs, but protection needs to be balanced with providing useful statistical information to inform decision making. Deployment of the DP-compliant methods will be context dependent to maximise NSOs ability to provide useful statistical information.

While DP's requirements are indeed strong, there is more to them than getting data custodians to assume the existence of some super-user. First and foremost, DP's requirements exist to ensure that the guarantees offered by custodians do not depend on what they have assumed users *do not know*. For instance, when designing their disclosure control strategies data custodians may very well underestimate the extent of the information already available to users prior to making their own releases. And if they do, then the protection guarantees offered will be weakened by what users actually do know.

Even when data custodians do have a good sense of what users already know in terms of a-priori information, their protection strategy could be compromised by a critical piece of information that comes to light after they make their releases. Thus, DP's requirements exist to also provide protection against unforeseen risk factors rather than just against some conjured omniscient user.

Without knowing what the future holds, the beginnings of DP are reminiscent of those of probability sampling back in the first half of the 20th Century. As Olkin (1987) points out from a conversation held with survey pioneer Morris Hansen, survey practitioners initially found it difficult to comply with the requirements of probability sampling. The first probability sample designs often were too rudimentary to meet realistic survey needs and there were many issues (e.g., how to devise unbiased estimators, how to assess their variance and how to deal with out-of-scope units, domain estimation and nonresponse) for which adequate answers only came later. Thus, we can expect a body of best practices to emerge from the lessons learned as DP is applied to specific contexts e.g., Hawes (2020).

Currently, complex survey features such as clustering and hierarchical structures (e.g., family-related information on records pertaining to individuals) make it challenging to find DP_{ε}-compliant methods of any practical use in a survey context. Also, not only can it be difficult in certain circumstances to devise a useful non-trivial DP-compliant method but formally establishing its compliancy can prove to be a daunting task.

One can expect more flexible mathematical requirements to be proposed in the years to come allowing for formal disclosure protection guarantees to be used in a wider array of practical situations than what is currently possible under DP. For example, while we know of ways to implement a DP-compliant method to release a count, it is not quite clear how to proceed for the total of a non-dichotomous variable such as income. The wider the range of values a variable can take, the larger DP's upper limit tends to become in order for it to hold for all possible outputs which can render a DP-compliant method all but useless in practice. Also, further guidance will be needed on how best to handle survey-specific features such as sample design information including survey weights. The flurry of DP-related papers published these last few years is a testament to the efforts being deployed to put forward formal disclosure protection guarantees beyond what the pioneering work of Cynthia Dwork, Frank McSherry, Kobbi Nissim and Adam Smith has provided already through DP.

Finally, even without seeking to achieve compliance, NSOs can still benefit from DP by reviewing their current disclosure control practices in light of its principles to identify and close gaps in the protection offered. For example, we know from a theorem of DP that a method needs to have a random component to be DP$_\varepsilon$-compliant. This suggests that a NSO rounding a count to the nearest pre-set base value would gain from using some *random* rounding method instead, thereby making its practices presumably less prone to disclosure than they were before.

Disclosure risk measures

Privacy preserving techniques provide NSOs with the tools they need to disseminate their data with various levels of privacy protection and guarantees. Whether or not NSOs decide to implement additional disclosure controls on top of synthetic data, there are methods that can be used to asses the disclosure risk present in the synthetic data. A variety of techniques are available to determine if attribute or identification disclosure risk is present in synthetic data.

Peer Review

Peer review of disclosure methods, such as rounding, generalization, suppression, or even differential private methods such as Laplacian mechanism, is a fit for purpose exercise to determine and demonstrate whether or not confidentiality methods or privacy-preserving techniques are fit for use based on a NSO's own legislative and operational frameworks.

For example, Statistics New Zealand's disclosure control practices are based on New Zealand statute, international and national best practices. Disclosure risk according to Statistics New Zealand is based on these objectives:

- maintaining privacy via confidentiality, by outputting zero 'sufficiently accurate' disclosures of individuals, or of particulars relating to individuals
- maintaining data quality via confidentiality, by using methods which introduces zero bias, or as little bias as possible, into the original data, and the original data's means and other dataset measures, where these means and measures are estimated using published 'confidentialized' data.

In practice, Statistics New Zealand endeavours to have zero tolerance for publishing accurate or discernible disclosures of counts of 1 or 2, or particulars relating to one or two unit records. This includes also potentially not disclosing counts of 0, and/or counts of 3, 4, or 5, as 'coverage' or 'protection' for counts of 1 or 2, etc.

Some typical methods to achieve this end, either used by or under investigation for use at Stats NZ, and which demonstrably introduce limited bias, include:

A. Random rounding to base 3 (RR3).
B. Fixed, or consistent, random rounding to base 3 (RR3).
C. P% rule based suppression and aggregation.
D. Noised Counts and Magnitudes (NCM), which is also substantially a differentially private method, and hence could also be measured via its differential privacy parameters.
E. R-Synthpop Classification and Regression Tree (CART) based non-1:1-mapped synthetic data, preserving univariate and bivariate inferential validity, which is also

substantially a differentially private method, and hence could also be measured via its differential privacy parameters.

In summary, in terms of measuring disclosure risk, confidentiality methods which do not accurately disclose counts of 1 or 2, or particulars relating to one or two unit records, and which do not introduce bias, are fit for purpose.

International best practices for peer review can be found in the European Statistical System Peer Reviews Guide for NSIs and Other National Authorities (Eurostat) or European Statistical System Code of Practice Peer Reviews: The National Statistical Institute's guide, Version 1.3 (Eurostat, 2007).

Feature mean Scaled Variance

A measure of disclosure risk suitable for data with a one to one mapping between the original and synthetic data is call *feature mean scaled variance*. This method is only suitable and fit-for-purpose for data with a 1:1 mapping between original data and synthetic data; and, which is all ordinal data, e.g. Boolean variables, ordinal numeric variables, ordinal categoric variables.

In this measure, all variances between mapped original and synthetic data points are feature scaled to the range 0 and 1. These feature scaled variances substantively report combinable inaccuracy values for the synthetic data compared to the original data. Value of around 0 refer to identical data. Values of around 0.5 or more refer to highly non identical data.

The threshold for the difference between too-accurate-to-publish data and sufficiently inaccurate-to-publish data is between 0.05 and 0.2. These figures are an approximation of the two sigma confidence interval tail, and the smaller pareto value, respectively. But the likelihood that the appropriate threshold boundary is somewhere between these two figures can also be examined graphically by an informed observer.

Synthetic data which is in the inaccuracy values range (0, 0.05) when examined graphically seems too-accurate-to-publish, and synthetic data which is in the inaccuracy values range (0.2, 1) when examined graphically seems sufficiently inaccurate-to-publish.

The actual threshold for publication purposes is proposed to be in the inaccuracy values range (0.05, 0.2). The exact threshold is a matter of risk appetite and unit record quantitative need. By analogy, Abowd (2016) suggests approximately 90% accuracy and 10% inaccuracy could be an appropriate threshold.

As an example in the following figures, each figure visualises a de-identified single unit record’s set of feature scaled variances between mapped original and synthetic data points. The x-axis ‘index’ is the list of ordinal variables in the unit record, excluding identifying variables. Hence in these graphs, the first three identifying variables are excluded, and the remaining 954 ordinal variables are included.

The y-axis is the range of possible feature scaled variance values for each unit record variable, each such value being a floating point number no less than 0 and no more than 1.

Three selected example de-identified unit records are depicted, sequentially as unsorted variable results for that unit record, sorted variable results for that unit record, and a histogram of variable results for that unit record. Figure 4 to 6 visualise a unit record with mean feature scaled variance = 0.05, roughly 5% inaccuracy, when comparing the original record to the synthetic record.

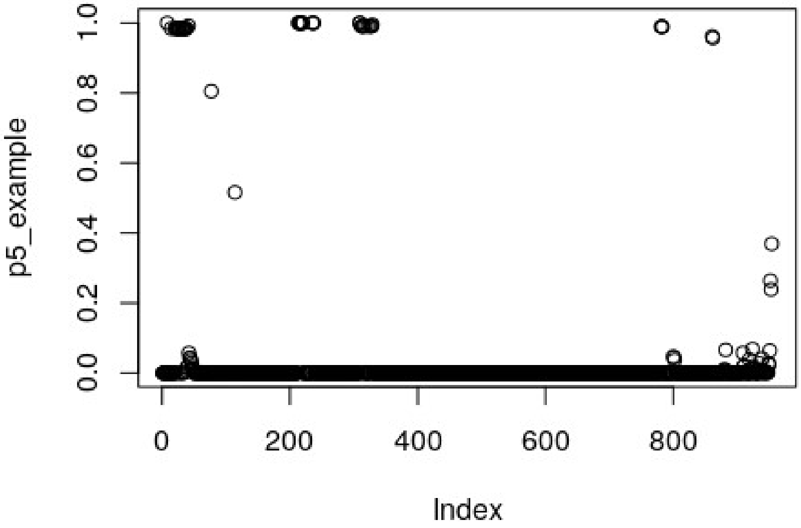


Figure 4: Visual of 5% feature mean scaled variance inaccuracy compared with its original unit record, unsorted

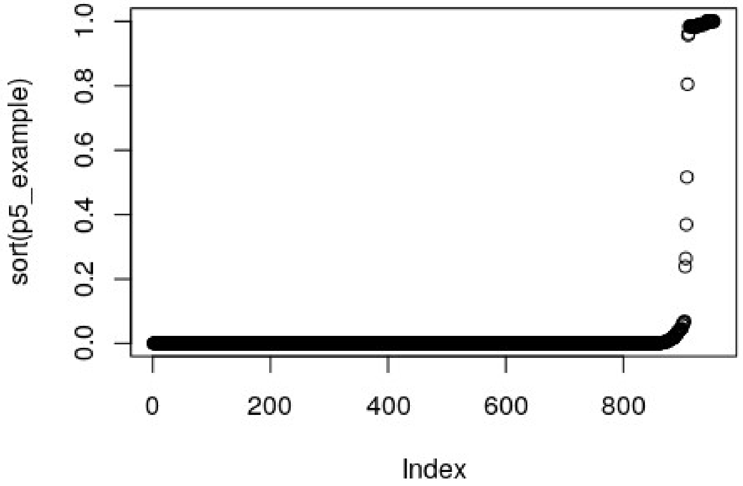


Figure 5: Visual of 5% feature mean scaled variance inaccuracy compared with its original unit record, sorted

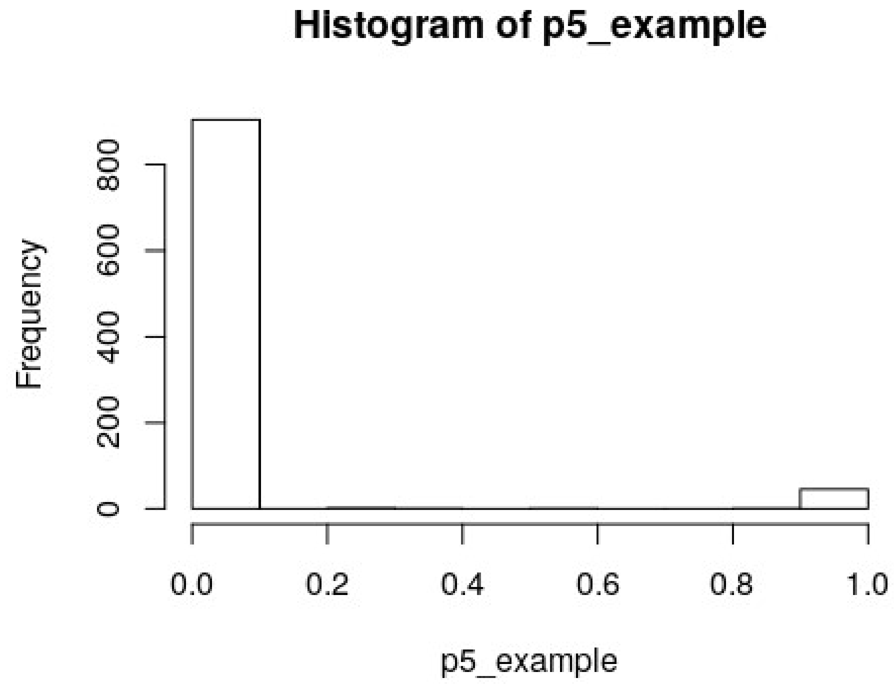


Figure 6: Histogram of 5% feature mean scaled variance inaccuracy compared with its original unit record

Figure 7 to Figure 9 visualise a unit record with mean feature scaled variance = 0.10, around 10% inaccuracy, when comparing the original record to the synthetic record.

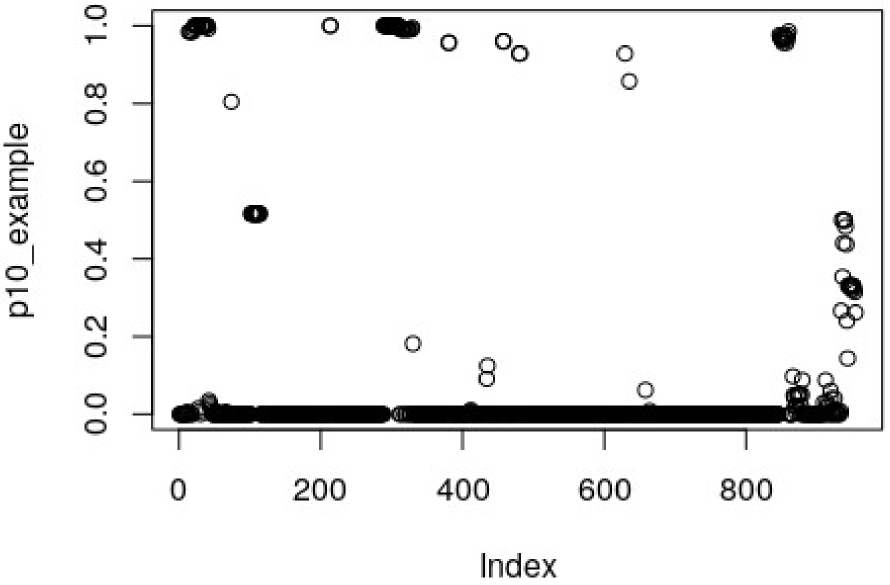


Figure 7: Visual of 10% feature mean scaled variance inaccuracy compared with its original unit record, unsorted

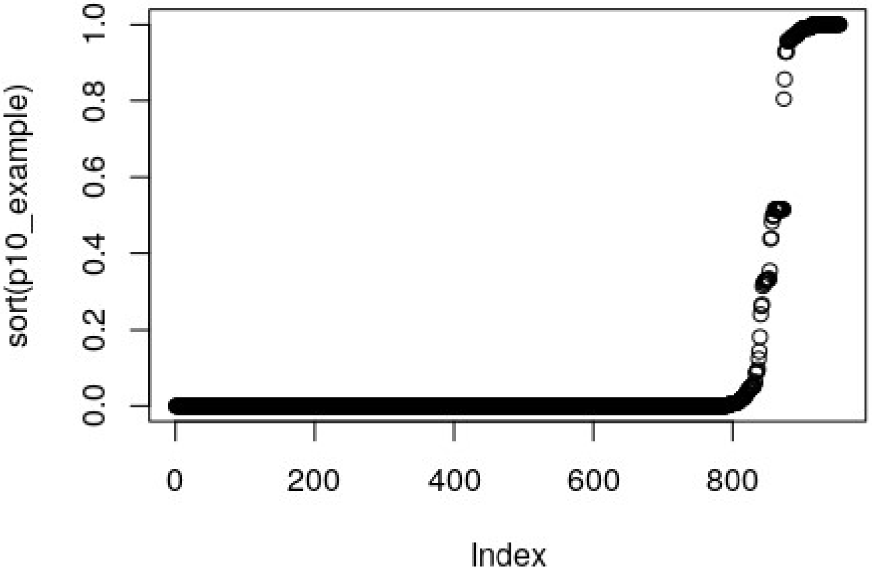


Figure 8: Visual of 10% feature mean scaled variance inaccuracy compared with its original unit record, sorted

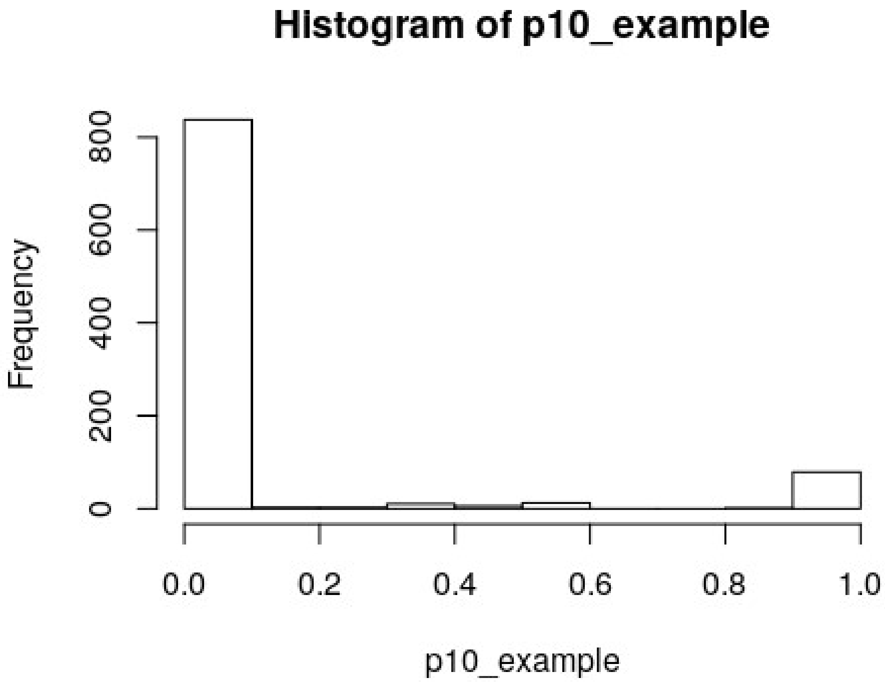


Figure 9: Histogram of 10% feature mean scaled variance inaccuracy compared with its original unit record

Figure 10 to 12 visualise a unit record with mean feature scaled variance = 0.20, around 20% inaccuracy, when comparing the original record to the synthetic record.

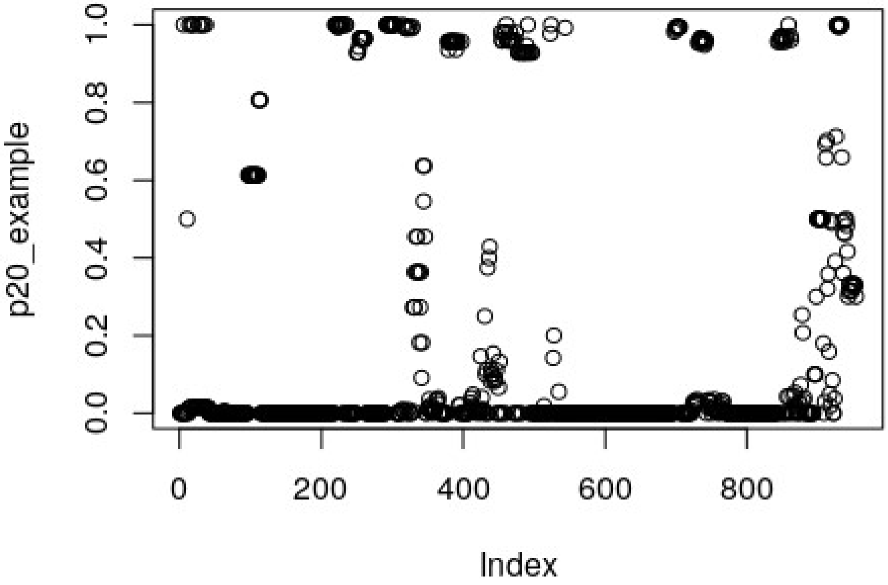


Figure 10: Visual of 20% feature mean scaled variance inaccuracy compared with its original unit record, unsorted

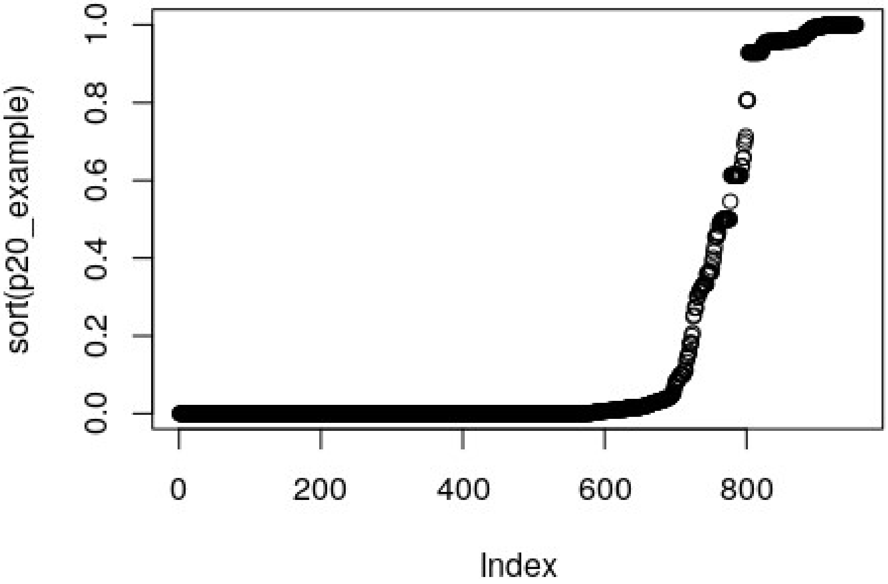


Figure 11: Visual of 20% feature mean scaled variance inaccuracy compared with its original unit record, sorted

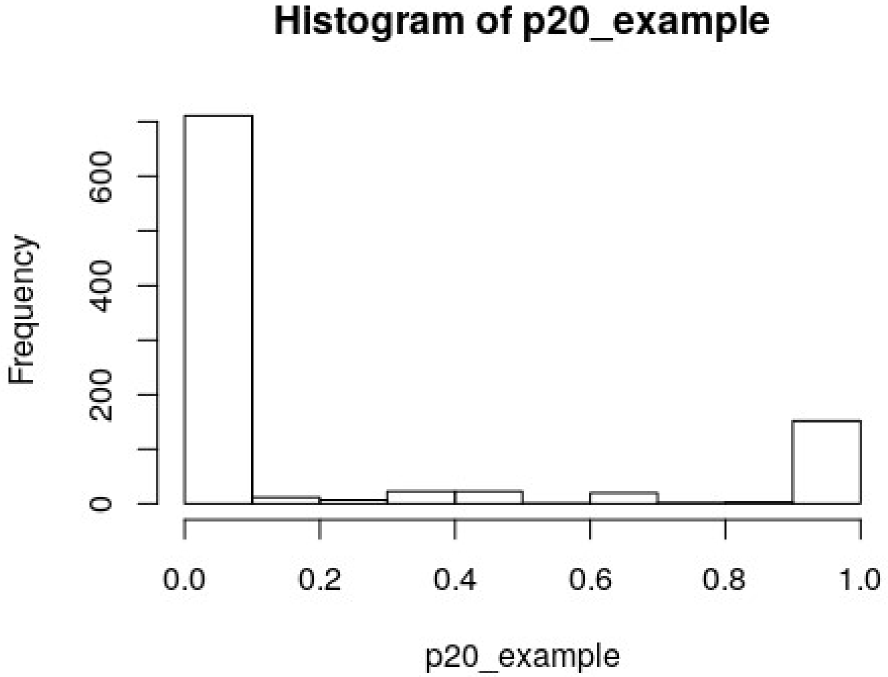


Figure 12: Histogram of 20% feature mean scaled variance inaccuracy compared with its original unit record

Rates related to database reconstruction

Identity disclosure can be more difficult to measure and *rates related to database reconstruction* looks to assess this concern. These rates refer to the percentage of matches one gets from reconstructed data to the real data, with the intent to determine how easily the real data can be generated with the information available to the public. As an example, a subset of the Summary File 1 (SF1) of the 2010 US Census was taken to review. The SF1 contains data from the questions asked of US citizens and about every American housing unit. The SF1 has the following data protections: the housing units in the file were swapped at some unpublished rate; the group quarters in the file were protected using synthetic data techniques and the tables themselves are an information reduction. The data for the variables age, sex, race, Hispanic/Latino ethnicity and census block (geography) variables were reconstructed using a subset of the SF1 through a system of equations that when solved, converted to microdata. These records were then matched with data from commercial holdings. In this example, the US Census Bureau had a specific series of commercial data holdings. More generally, the exercise could be conducted with any additional datasets that contain identifiers, like name and address, and a subset of variables to match with the subset under review (for example age and sex). In the US Census example, after combining the reconstructed with the commercial data, the five variables plus a unique identifier where linked exactly to the same variables in the Census Edited Files (confidential data). There are two main rates to consider: putative (suspected match) and confirmation linkage, or confirmed match.

The matching process works as follows:

1. Reconstructed data says there is 1 male white Hispanic person aged 52 in a block
2. The commercial data shows that the only 52 year old male in that block is Bob at X address (putative match)
3. Attach white/Hispanic to the putative match
4. Look for Bob/X address/age 52/male/white/Hispanic in the CEF
5. Find that record in the CEF (confirmed match)

This exercise can be scaled to rather large datasets, as demonstrated by the exercise conducted by the US Census Bureau. Table 11 summarizes the putative, confirmed and precision rates of the Census exercise. Precision is defined as confirmed divided by putative. The precision variable in Table 11 shows how often we are right when we think we are.

Table 11: Disclosure Risk Assessment of Population Uniques by Block Population Size (Abowd 2021b)

Block Population Bin	Putative Re-identifications (Source: Commercial Data)	Confirmed Re-identifications (Source : Commercial Data)	Precision (source : Commercial Data)	Putative Re-identifications (Source: CEF)	Confirmed Re-identifications (Source : CEF)	Precision (source : CEF)
Total	137,709,807	52,038,366	37.79%	238,175,305	178,958,726	75.14%
0						
1-9	1,921,418	1,387,962	72.24%	4,220,571	4,093,151	96.98%
10-49	25,148,298	13,481,700	53.61%	47,352,910	43,415,168	91.68%
50-99	30,567,157	12,781,790	41.82%	51,846,547	42,515,756	82.00%
100-249	38,306,957	13,225,998	34.53%	63,258,561	45,807,270	72.41%
250-499	21,789,931	6,408,814	29.41%	35,454,412	22,902,054	64.60%
500-999	13,803,283	3,460,118	25.07%	23,280,718	13,514,134	58.05%
1000+	6,172,763	1,291,984	20.93%	12,761,586	6,711,193	52.59%

Chapter 5: Utility measures for evaluating

The utility, or value, of a synthetic data set reflects how useful that data set is to the purpose or the use case for the data. As discussed in Chapter 2, synthetic data is often used either instead of the original data or as a preliminary analysis to guide the final results which will be run on the original data. In both cases, the utility of synthetic data is to give the same conclusions as would have been arrived at from the original confidential data. This is equally important in the second case because preliminary analysis on the synthetic data will guide the final models used.

This guide recommends in Chapter 3, methods of creating synthetic data depend on the use case, what the synthesizer wants to preserve and the type of original data. Once the synthetic data sets have been created, the utility can be evaluated.

There are two broad categories of utility measures: "broad", "global" or "general" utility measures, as opposed to "narrow" or "specific" measures. In this guide, we will use the terms general and specific measures. Specific measures are useful when evaluating a specified analysis, however they are not useful for tuning –modifying synthesis methods to improve the utility – as the synthesizer often does not know the analysis that will be conducted using the synthetic data when tuning has to take place.

According to Raab et al. (2021), there are two main reasons we might wish to evaluate the general utility of synthetic data:

1. To compare different synthesis methods for the same data set in order to generate the most useful synthetic data set for the user .
2. To diagnose where the original and synthetic data distributions differ and thus tune the synthesis methods to improve the utility of the synthetic data.

For the first of these a number of measures have been proposed that summarise the utility of the data, or sometimes of a subset of the data, by a single number. Two main methods have been proposed to compute these measures. The first, proposed by Karr et al. (2006) and Woo et al. (2009), is to combine the two data sets (original and synthetic) and to use the information in the records to predict their source. Several measures can be calculated from this approach some of which are calculated from the propensity score, the probability that a record is from the synthetic data. The second method used to compute a single comparative measure is to compare tables created from the synthetic data with those from the original. These two methods are related, as we will discuss below.

A single measure does not provide guidance as to what aspects of the synthetic data differ from the original. Thus we need different strategies to fulfil the second of Raab's requirements. Several methods have been suggested for this, sometimes making use of the utility measures discussed above for subsets of variables in the records.

Table 12: Summary of utility measures

Method	**Measure or procedure**
Specific	Confidence interval overlap
	Mahalanobis distance ratio
	Task accuracy (Kaloskampis et al. 2020)
General Single measure from propensity score	propensity score mean squared error (pMSE) Karr et al. (2006) and Woo et al. (2009)
	Kolmogorov- Smirnov Statistic comparing propensity scores for original and synthetic data (SPECKS) (Bowen et al., 2021)
	Other comparisons of propensity scores for original and synthetic data , e.g Wilcoxon signed rank statistic (U)
	Percentage over 50% of combined records correctly predicted by the propensity score (PO50)
General Single measure from tables	Voas-Williamson statistic (VW) (Voas and Williamson 2001)
	Freeman-Tukey (FT) (Voas and Williamson 2001)
	Likelihood ratio statistic from tables (G) and other members of the divergence family (Voas and Williamson 2001)
	Jensen-Shannon Divergence (JSD) (Fuglede and Topsoe, 2004)
	Bhattacharyya metric (BhattD) Bhattacharyya (1943)
	Mean absolute difference in densities (MabsDD) (Ridgeway et al. 2021)
	Difference of correlation matrices (Kaloskampis et al. 2019)
	Weighted mean absolute difference in densities (WMabsDD) Raab (2021)
Methods for exploring utility	Comparing histograms by visualisation and summary measures (www.synthpop.org.uk and Kaloskampis et al. 2020)
	Comparing cross-tabulations from marginal distributions (Raab 2011 and NIST, 2021)
	Comparing and visualising other summary statistics (e.g. Pearson Correlations) (Beaulieu-Jones et al. 2019, Kaloskampis et al. 2020)

Table 12 provides a summary of the methods discussed in this chapter classified as methods for specific utility, as single measures of general utility, by two methods, or as methods for exploring, summarizing or visualising utility.

Specific utility measures

Specific utility measures compare the results of statistical models fitted to the synthetic and the original data. The most widely used of these is the confidence interval overlap that provides both a summary measure and a visualisation of the results from a statistical model from the two data sources. Other summary measures include various graphical comparisons as well as standardised differences in coefficients and an overall lack of fit measure that can be computed from the variance matrix of the coefficients.

Confidence interval overlap

As discussing in Chapter 2, one of the most popular use cases for synthetic data is testing analysis, where often a user of synthetic data is testing a linear or generalized linear regression. This activity not only produces coefficients but also confidence intervals. A measure to assess the utility of synthetic data for testing analysis is to evaluate how the confidence intervals of an estimate differ between the real and synthetic data (Karr et al. 2006). The confidence interval overlap measure is such a measure. Karr et. Al. (2006) suggest using the percentage overlap of confidence intervals (IO), defined for each coefficient β_i as $IO_i = 0.5\left[\frac{\min(u_o,u_s)-\max(l_o,l_s)}{u_o-l_o} + \frac{\min(u_o,u_s)-\max(l_o,l_s)}{u_s-l_s}\right]$, where the confidence interval for the original data is $(u_o; l_o)$ and for the synthesised data $(u_s; l_s)$. The numerators in each of the terms in this equation is the overlap of the intervals which becomes negative when the intervals are disjoint. The average of the overlaps can then be used as a summary measure of utility. The IO measure takes a maximum value of 1 when the intervals are the same length, but lower when they are different lengths.

Confidence Interval overlaps are most often used to compare results for fitting statistical models to the original and synthetic data. It is recommended that a graphical display that compares the fit of the models should be the first step in evaluating the fits from the two sources. An example is Figure 13 which shows output from the *synthpop* package for R (www.synthpop.org.uk).

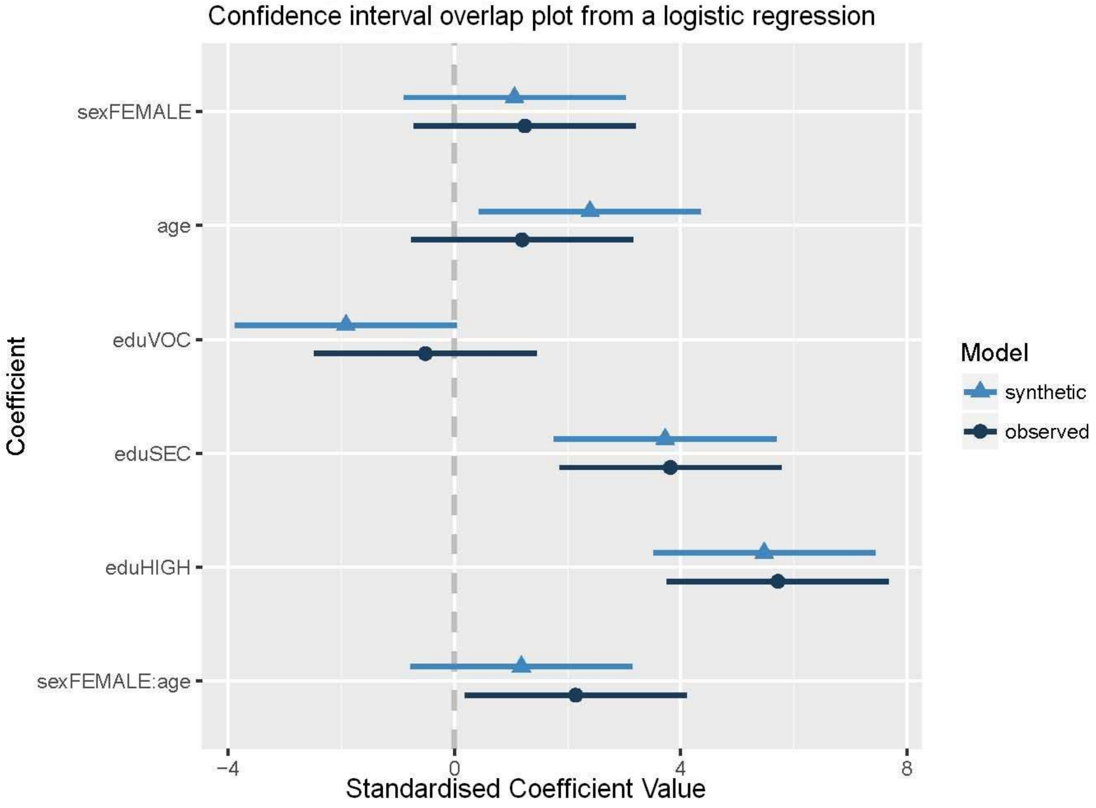


Figure 13: illustration of confidence interval overlap from Raab and Nowok (2017).

Mahalanobis distance ratio

For synthetic data the IO does not have straightforward null distributions. A better summary is the Mahalanobis distance between the coefficients. This is calculated from the difference in the estimates and their estimated variance-covariance matrix, $\hat{V}$, gives a lack-of-fit measure: $(\hat{\beta}_s - \hat{\beta}_o)'V(\hat{\beta}_s - \hat{\beta}_o)$.

Unlike the mean IO measure it makes allowance for the correlation between the estimates and will have a null distribution that is asymptotically χ_p^2, where p is the number of parameters in the fitted model. This measure it can be thought of as a distance between the two estimate vectors in the space of the normalised eigen-vectors of the coefficients.

Task accuracy

This method involves identifying tasks relevant to the data set, for example, classification, and compare the task accuracy between the real and synthetic data. This measure is considered specific because it evaluates a specific and not generic task. A task accuracy measure is a useful measure because there are many cases that the measures chosen indicates sufficient utility, but if the synthetic data is built for a specific purpose the results may still not be the same with the real data.

For example, take the US Adult Income data set from UCI repository that has 32,500 records with variables that include age, working status, education and income (Kohavi and Becker, 1996). With this data, the task at hand is to see if income is above or below a certain amount, say make a prediction if an income is above or below $50,000. The first step is to build a model on the real data and determine the accuracy of the model. Next, conduct the same task with synthetic data. The same exercise on the real and synthetic data may result in different outcomes, which provides the synthesizer an assessment of synthetic data accuracy. Figure 14 illustrates such a classification task with the original data (red) and with synthetic data generated using GANs and three values of privacy loss (blue) (Kaloskampis et al. 2020). As shown in Figure 14, the synthesizer or user can now assess if the task using the synthetic has the utility necessary for their use case.

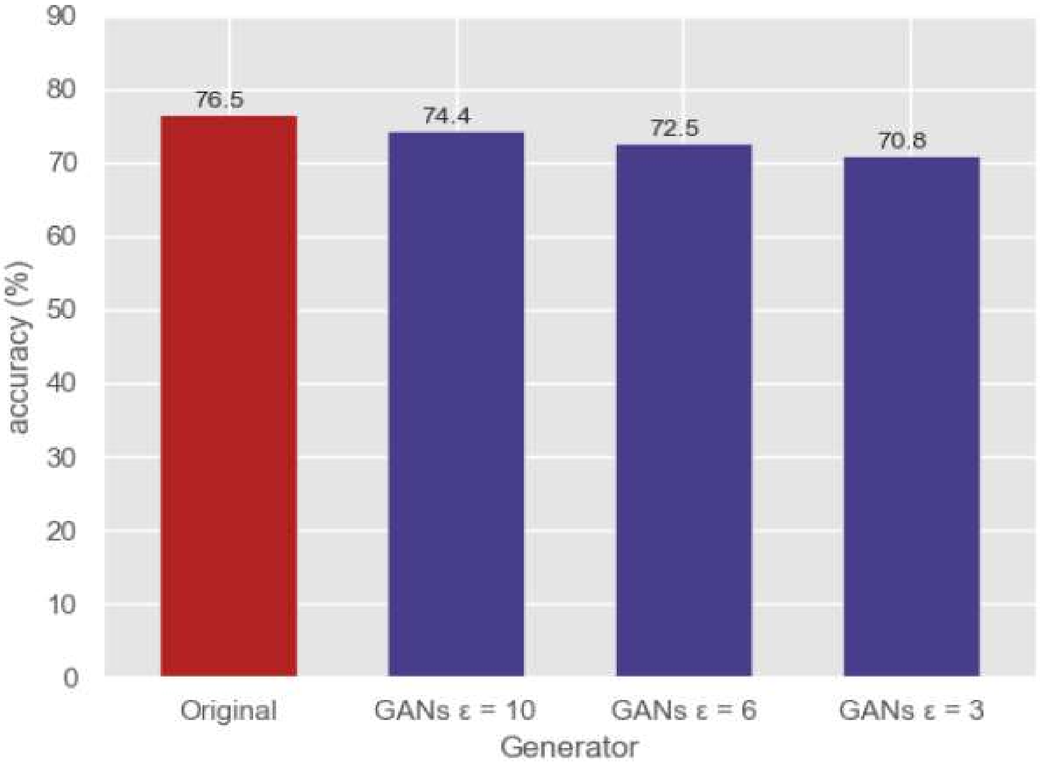


Figure 14: Classification accuracy trained on original US Adult Income data set and synthetic data sets generated with GANs, with different values of privacy loss ε (Kaloskampis et al. 2020).

General utility measures giving a single measure

The statistical analyses for which the synthetic data will be used are typically not known when the synthesis is being carried out. Measures that compare the whole distribution of the synthetic data to that of the original data are referred to as general utility measures. As we mentioned in the introduction to this chapter, these are mostly based on two methods, firstly combining the original and synthetic data and calculating a propensity score, the probability that any record is synthetic and secondly by comparing tables of original and synthetic data. As we will show below these two methods can be considered to be the same thing, but comparing tables allows some extra measures that are not available from the propensity score method.

Measures from the propensity score

Karr et al. (2006) and Woo et al. (2009) evaluate the utility of synthetic data by combining the records of the real and synthetic data and measuring how well the data values predict the source of the records as real or synthetic. An indicator, say x, is assigned a value of 1 for the synthesized data and 0 for the original. A method such as logistic regression or any non-parametric predictive method can be used to attempt to derive the propensity score, $\hat{p}$. The propensity score is the probability that x = 1, meaning that the record was from the synthesised data. If the distributions of the real and synthetic data are indistinguishable then all propensity scores are expected to be close to the proportion of synthetic records in the combined set; 0.5 if the two data sets have the same number of records. Several measures can be computed from $\hat{p}$, four of which are listed in Table 1. The most commonly used is the propensity score mean-squared error (pMSE) (Woo et. al. (2009)) proportional to $\sum(\hat{p} - 0.5)^2$ where the summation is over all rows of the combined data.

Measures from tables

The other source of general utility measures are those that can be obtained from tables, as proposed by Voas and Williamson (2001) for summarising differences between synthetic data and the original. They adapted measures used in computing chi-squared tests for tables. In particular they suggested what we refer to as the Voas-Williamson statistic for comparing tables. It is similar to the usual Pearson chi-squared statistic (*X^2).* We can write the counts for any table with k categories as y_i (i = 1; 2; ..k) and the corresponding synthetic counts as s_i (i = 1; 2; ..k). If the total counts in each table are the same: $\sum y_i = \sum s_i = n$ then

$X^2 = \sum_{i=1}^{k} \frac{(s_i - y_i)^2}{y_i}$. A practical problem with this chi-squared statistic is that the contribution from a cell where the original data has a zero count, but is not a structural zero, is not defined.

Voas and Williamson propose the modification of this statistic by replacing y^i with the mean of y^i and s^i Other statistics in the power-divergence family (Read and Cressie, 1988) could also be used, such as the deviance or the Freeman-Tukey measure. Another related measure is the Jensen-Shannon Divergence which can be considered as a modification of the likelihood ratio statistic to allow for zero values of y_i. Further measures computed from tables include the histogram overlap measure (Bhattacharyya, 1943) and the mean absolute difference in density, over all cells in the table used in evaluating the NIST challenges (Ridgeway et al. 2021), calculated as $MabsDD = \frac{\sum_{i=1}^{k} |\frac{s_i}{n} - \frac{y_i}{n}|}{k}$, in this case. All of these measures can be generalized to the case where the synthetic data has a different number of records from the original, see Raab et al. (2021) for details.

Relationships between the measures

Comparing tables can also be framed as a prediction measure where the propensity score in the case of equal sample sizes is just $\widehat{p_i} = s_i/(s_i + y_i)$. Calculating this propensity score for an n-way table is identical to what would be obtained by using a logistic model with all (n-1)-way interactions. Several of the measures in Table 12 are linearly related. In particular *VW* and *pMSE* are the same measure, as are the three measures *SPECKS, PO50* and *MabsDD* and also $dBatth \propto \sqrt{FT}$. Thus, there are fewer independent utility measures than Table 12 would suggest. Empirical investigations suggest that all of the measures are correlated with one another when compared for different synthetic data sets. For some subgroups, e.g. *VW/pMSE, FT, JSD* the correlations are so high as to suggest that they are essentially the same measure.

A small investigation of the ability of these measures to differentiate a poor synthesis from a good one suggested that *VW/pMSE, FT, JSD* performed slightly better than the other measures discussed here, but almost all gave satisfactory discrimination. These findings are based on recent empirical work by Raab et. al (2021), that would benefit from further development by other groups.

Scaling of utility measures

It is helpful if the utility measures can be on a scale that makes them easy to interpret. For all the measures described here a large value indicates lack-of utility. One method could involve scaling the measures by the maximum value they could take. For example JSD is scaled in this way since its maximum value is 1.0. Other measures have an interpretation that helps to understand them, for example it is easy to think of the percentage correctly predicted for PO50, and Dbhatt has an immediate interpretation as the overlap of matching histograms.

Another approach to scaling utility measures is to express them relative to the value that would be expected if the model used to synthesize the data was the “correct” model. The expected value for the “correct” model can be termed the Null expectation. This approach can also be considered as scaling the measures compared to the expected stochastic error of the distribution. The target value for measures scaled in this way is 1.0. All the measures derived from the various chi-squared tests have known Null expectations, Snoke et al. (2018) derived the Null distribution of this quantity for the pMSE when it is estimated from a model with a fixed number of parameters. They also propose methods for obtaining the Null expectation for any of these measures by replication methods. A modification of one of the methods to use for DP synthetic data has been proposed by Bowen et al. (2021). This scaling by the Null expectation differs from the others in that it defines a target that a good synthesis should achieve. Although an absolute target would be 1.0, synthetic data that have proved to be useful can have values in the range from 3 to 10. Values above 10 signal potential problems with some part of the distribution being evaluated. This is reasonable as we do not believe that real data are generated exactly from a statistical model.

Models for the propensity score.

The choice of model for the propensity score is crucial to its performance and a more important choice than that of the utility measure. Any method that can predict group ownership could be used. Those that have been used in practice are logistic models, which includes the special case of models that define tables, and classification and regression tree (CART) models. The models that can be fitted are limited by the complexity that it is possible to fit from a finite sample of data. Logistic models with a large number of parameters may fail to converge and, even if convergence is achieved, will have many parameters that cannot be estimated from lack of information (aliased parameters). Similarly, comparison of tables with more than a small number of variables will yield large tables, with most of their cells having zero counts, that may lead to computational problems. CART models, that select a partition of the data to describe the distribution, can cope with data sets with more variables. But such models also have computational limits, especially when dealing with categorical variables with many possible levels, as are often found in data from NSOs. A simple model may give an assurance of a good fit although only a very limited aspect of the distribution differences has been assessed.

One number is not enough to describe utility

A person creating synthetic data needs more than a single number to assess the utility of the data they have produced. If the utility appears unsatisfactory they need to know which aspects of the distribution are causing the problem. Many strategies can be devised to explore these differences. Some of these are described in the next section, some using the utility measures described here for subsets of variables or for the partitioning of large synthetic data sets into smaller strata, often defined by geographic areas.

Methods to explore aspects of utility

Univariate comparisons

The starting point of any evaluation of synthetic data is to examine how well the synthetic data reproduce the univariate distribution of each variable. Bar charts of categorical variables or histograms of numeric variables are the obvious first step. These may be accompanied by utility measures computed from the tabulation of each variable. The function to produce plots of each variable in the synthpop package can be accompanied with a table of a variety of utility measures. Kaloskampis et al. (2020) produce similar plots where they display the Bhattacharyya metric alongside the histograms, see Figure 15.

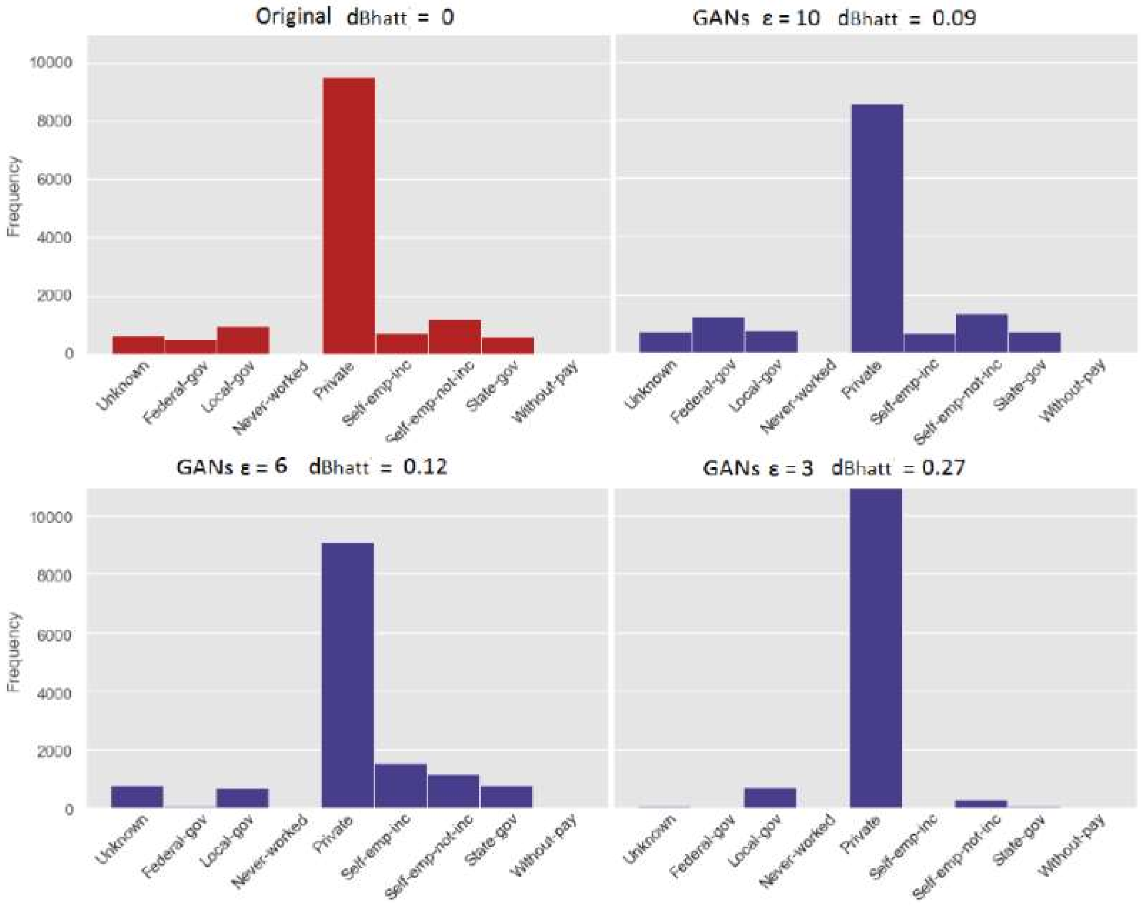


Figure 15: Comparison of histograms of the workclass variable of the US Adult Income data set between original (red) and synthetic data sets (blue) generated with GANs, using different values of privacy loss ε. We denote the Bhattacharyya metric by dBhat

Marginal comparisons

Several methods have been proposed for comparing low-level marginals between the original and synthetic data. Any of the utility metrics that can be computed from tables can be used after first forming categories from any continuous variables. To summarise results from marginals Raab et. al. (2020) propose the following steps.

1. Examine histograms that compare the original and synthetic data and also check the *pMSE-ratio* for the univariate comparison of each variable.
2. Once these are satisfactory, continue by visualising the utility of all two-way relationships between variables.
3. If there is one variable of particular interest, for example an outcome variable in an epidemiological study, then it might be worth checking and visualising all three way relationships that involve that variable.

Figure 16 illustrates the output from step 2 for four syntheses of the same original data. Plot (a) is from a default parametric synthesis, showing that there was a problem with the variable "weight": this had already been noted at step 1.. Plots (b) and (c) show how reordering and stratifying the synthesis improves the utility. Plot (d) shows that synthesizing from a CART model, with no adjustment, gives better utility than any parametric model.

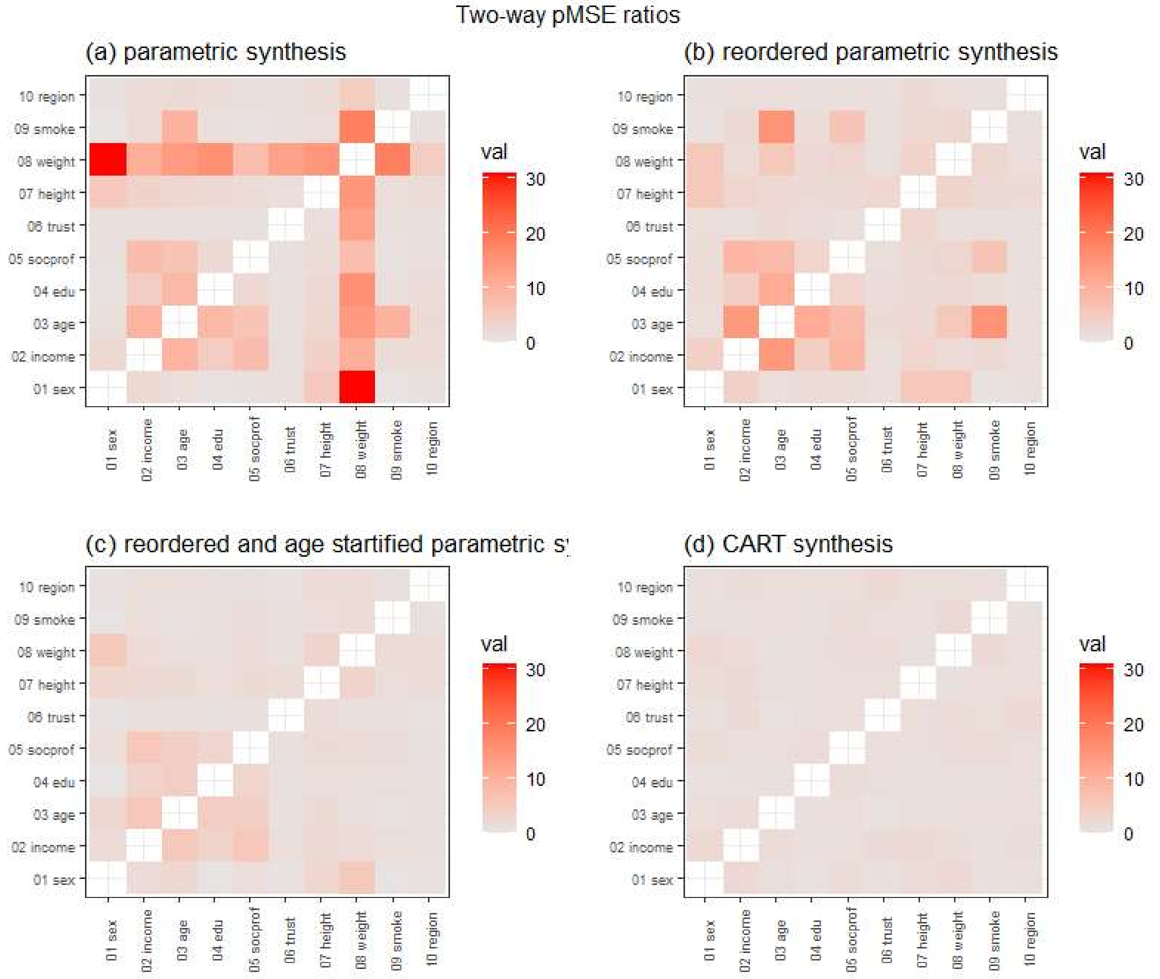


Figure 16: Visualizations of the utility of all two-way relationships between variables. From Raab et al. 2021)

The evaluation of the NIST Deidl-2 synthetic data challenge developed a method based on three-way marginals to evaluate utility. Marginal distribution metrics work well for discretized data, since one can easily consider all possible k-way margins of the full cross-classification of discrete variables. Numeric (integer or floating point) data can function under these metrics as well by discretization. One implementation of a k-way marginal metric is to consider the total absolute deviation across cells of a marginal table based on two "versions" of a dataset. After normalizing the total absolute deviation by dividing the total of the table, one obtains a metric on how close the chosen margin is between the two data versions. This method is quite flexible and allows variations such as fixing certain dimensions to more finely assess differences for certain variables or constructs.

Diving into an example, we consider creating synthetic data for a subset of the 1940 Census Demonstration Data (Ruggles et al., 2018) and assessing the synthetic data using k-way marginals. We take the subset of records for the District of Columbia, and synthesize the following variables:

- Binary sex (SEX)
- Age (AGE)[10]
- Major race category (RACE)
- Ethnicity (HISPAN)

We use two models for synthesis: a naïve model that simply permutes the values of each of the columns above (PERM), and a model based upon sequentially fit classification trees (TREE). The effect of both models is to produce a new dataset that has the same scheme as the original but with potentially modified entries in each row. We then use marginal metrics to assess the relative quality of PERM and TREE in reproducing the original data.

A natural starting point, and indeed often the ending point for much exploratory data analysis, is to consider the one-way marginal metrics for each column. The margins for sex under each data model are:

	Original 1940	PERM	TREE
Male	318,269	318,269	317,917
Female	346,595	346,595	346,947

To get the sex marginal score under TREE, we take the absolute differences in the number of males and females under TREE, sum them, and divide by the total number of persons:

$$\frac{|317917 - 318269| + |346947 - 346595|}{318269 + 346595} = 0.00106$$

To get the overall 1-marginal score under TREE, we perform the same computation for the other three 1-way margins and then take the average:

Sex	1.06E-03
Age	8.00E-04
Race	3.55E-04
Ethnicity	1.99E-04
Average	**6.03E-04**

[10] Though we synthesize the full range of ages, the marginal evaluations are on a recode of age into three bins for children (0-17), adults (18-64), and older adults (65+)

Since scores are more sensible when larger is better, we can use the transform[11]:

((2 - raw_score)/2) x 1000

To map the raw scores to (1000, 0), so that now a score of 1000 means perfect recreation of the margin, and a score of 0 means the margin is maximally perturbed[12].

We can then compare these overall 1-way marginal scores across our two models:

Model	Raw Score	Adjusted Score
TREE	6.03E-04	999.7
PERM	0.00E+00	1000

The PERM method exactly recreates the 1-way margins since it samples the columns without replacement. But we see that TREE comes close, which becomes important when we consider marginals beyond one dimension.

We now consider the 3-way marginal metric. This is calculated exactly as above, except now we consider the deviations of the cells for all 3-dimensional margins. Here we see a change:

Model	Raw Score	Adjusted Score
TREE	2.29E-03	998.85
PERM	3.61E-02	981.93

The TREE model now outperforms PERM. In this simple case, where the variables may not have especially strong dependencies, a naïve model such as PERM can still perform well, but as dependencies and dimensionality increase, this will not occur, and the job of the data synthesizer becomes a more delicate task.

We can extend marginal metrics to give finer detail. For instance, we can restrict margins to only those containing at set of variables (e.g. all tables that contain "AGE" as a margin). We can also consider the cell differences between different categories of a variable (e.g. compare scores restricted to cells associated with males versus those associated with females). In this way we can build up a picture of where two datasets differ the most. This is especially helpful in assessing synthetic data when certain use cases may need preservation.

In the NIST Deidl-2 synthetic data challenge example, all three way marginals from their large data set would have been too many to compute, so a subsample of all possible marginals was used. Each was evaluated by calculating the MabsDD, that was then rescaled to give a "human-readable NIST score" defined as $1000\ \left(1 - \frac{MabsDD}{2}\right)$ that ranges from 0 to 1000, with 1000

[11] Derived from the worst case: changing a table with counts of form [0 N] to [N 0] for a raw score of 2/N

[12] This score was used for the NIST 2018 Differential Privacy Synthetic Data Challenge: https://www.nist.gov/ctl/pscr/open-innovation-prize-challenges/past-prize-challenges/2018-differential-privacy-synthetic

representing exact agreement between the tables. The utilities can be examined to identify which variables contribute most often to the tables with low scores. This score method was also used to assess the utility of geographic subsets of the data.

Comparing other statistics

Many other statistics could be compared between the synthetic and original data. In their evaluation of the NIST synthetic data Challenges Bowen and Snoke (2021) propose a range of such measures as well as some of the other utility measures discussed here. They also propose methods of combining and visualising these different measures.

Beaulieu-Jones et al. (2019) have used a range of utility measures to evaluate DP synthetic data created from a clinical trial data set. They compare a number of relevant outcomes and present results graphically. In particular, they present the Pearson correlations between variables as heatmaps, Kaloskampis et al. (2020) has used the same method to present the correlations , see Figure 17. As the visual comparison is often impractical, Kaloskampis et al., (2019) proposed a quantitative measure stemming from these visualisations, based on the difference of the underlying correlation matrices. This method could be useful, for example, in the process of hyperparameter optimisation of a synthetic data generation algorithm.

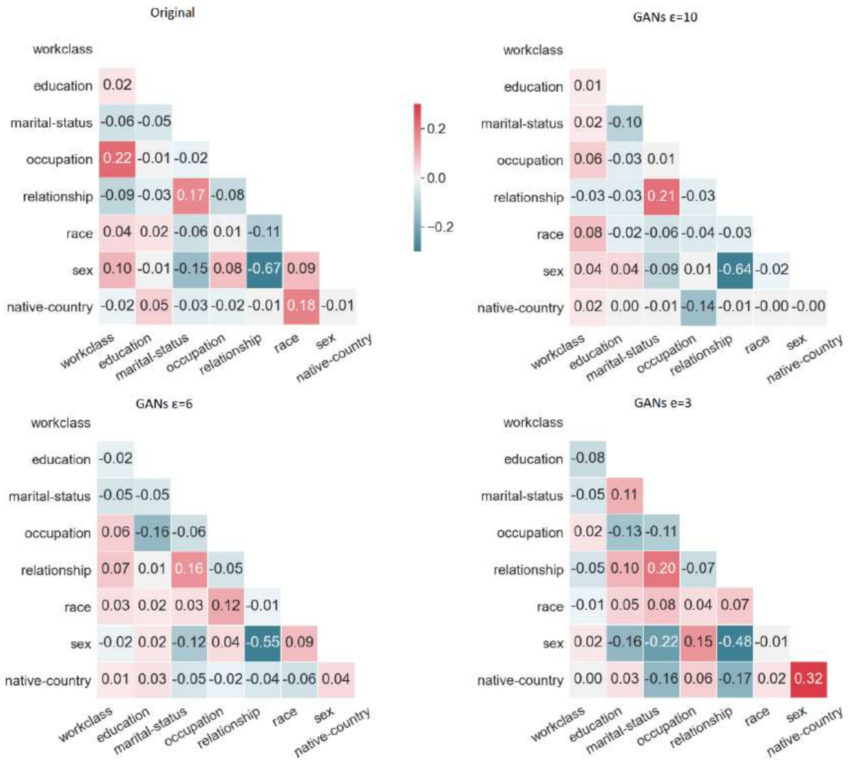


Figure 17: Pairwise Pearson correlation heatmaps for original US Adult Income data set and synthetic data sets generated with GANs, with different values of privacy loss ε (Kaloskampis et al. 2020).

This synthesis uses a differentially private method where the parameter ε determines the degree of privacy loss that increases as ε decreases. This synthesis uses a differentially private method where the parameter ε determines the degree of privacy loss that increases as ε decreases. This method provides a means to quantify the comparison . This could be an alternative to examining two-way marginals. The marginals have the disadvantage of ignoring the ordering of variables in the utility measure. However marginals have the advantage of being defined for categorical data as well as continuous or ordered variables.

Conclusion

We have reviewed methods of evaluating how well synthetic data reproduces features of the original. The methods we have included have either been used by data originating from NSOs, or have the potential for such use. We differentiate between specific and general utility measures and argue that the latter are more useful for comparing synthesis methods and for tuning a synthesis in respect of deficiencies identified. Although a large number of general utility measures have been suggested, some are equivalent to each other and all appear to be highly correlated when compared across different data sets. For measures derived from discriminating between the synthetic and the original data, via a propensity score, the method of discrimination is more important than the utility measure chosen. We argue that a utility measure that provides only a single number is not useful in tuning the synthesis method to improve utility. We give examples of sets of measures and visualizations that can be used to guide synthesis methods.

References

Abowd, J. (2016). How Will Statistical Agencies Operate When All Data Are Private? Washington Statistical Society Julius Shiskin Memorial Award Seminar, USA.

Abowd, J. (2021a). 2010 Declaration of John Abowd, State of Alabama v. United States Department of Commerce. Case No. 3:21-CV-211-RAH-ECM-KCN. (2021)

Abowd, J. (2021b). 2010 Supplemental Declaration of John M. Abowd, State of Alabama v. United States Department of Commerce. Case No. 3:21-CV-211-RAH-ECM-KCN. (2021)

Basu, D. (1971). An essay on the logical foundations of survey sampling, part 1. In Foundations of Statistical Inference. Edited by V.P. Godambe and D.A. Sprott. 203-242. Toronto: Holt, Rinehart and Winston.

Beaulieu-Jones, B., Wu, S., Williams, C., Lee, R., Bhavnani, S., Byrd, J., and Greene, C. (2019) Circulation: Cardiovascular Quality and Outcomes. Volume 12, Issue 7, July 2019 https://doi.org/10.1161/CIRCOUTCOMES.118.005122

Bhattacharyya A (1943). “On a measure of divergence between two statistical populations defined by their probability distributions.” Calcutta Mathematical Society, 35, 99–109.

Bowen, CM., and Lui, F., Su, B. (2021). “Differentially private data release via statistical election to partition sequentially.” METRON, 79(1), 1–31. URL https://doi.org/10.1007/

s40300-021-00201-0.

Bowen,C., and Snoke, J. (2021) Comparative Study of Differentially Private Synthetic Data Algorithms from the NIST PSCR Differential Privacy Synthetic Data Challenge Journal of Privacy and Confidentiality 11 (1)

Cano, Isaac and Torra, Vicenç. (2009), Generation of Synthetic Data by means of fuzzy c-Regression. 1145-1150. 10.1109/FUZZY.2009.5277074.

Choi, E., Biswal, S., Malin, B, Duke, J. Stewart, W., and Sun, J. (2017). Generating multi-label discrete patient records using generative adversarial networks. In Machine learning for healthcare conference, pages 286–305. PMLR.

Cox, L. H. (1980). Suppression methodology and statistical disclosure control. Journal of the American Statistical Association, 75(370), 377-385. https://doi.org/10.1080/01621459.1980.10477481

Desai, T, Ritchie, F. and Welpton, R. (2016). Fives Safes: Designing Data Access for Research. University of the West England Research Repository

Domingo-Ferrer, J., Gonzalez-Nicolas, U. (2010), Hybrid microdata using microaggregation, Information Sciences, 180(15), 2834-2844.

Dosselmann, R., Sadeqi, M., and Hamilton, H. J. (2019). A Tutorial on Computing $ t $-Closeness. *arXiv preprint arXiv:1911.11212*.

Drechsler, J. (2011). Synthetic Data Sets for Statistical Disclosure Control. Springer, New York.

Drechsler, J. and Reiter, J.P (2009), Disclosure Risk and Data Utility for Partially Synthetic Data: An Empirical Study Using the German IAB Establishment Survey. Journal of Official Statistics, Vol. 25, No. 4, pp 589-603.

Drechsler, J. and Reiter, J.P. (2011), An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets. Computational Statistics and Data Analysis, 55, 3232-3243.

Duncan, G., Elliot, M., Salazar-Gonzalez, J. (2011). Statistical Confidentiality: Principles and practice. Springer, New-York

Dwork, C., McSherry, F., Nissim, K. and Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. Proceedings of the 3rd Theory of Cryptography Conference, 265-284.

Dwork, C., McSherry, F., Nissim, K. and Smith, A. (2011). Differential Privacy: A Primer for the Perplexed. Joint UNECE/Eurostat work session on statistical data confidentiality

Efron, B. (1979), Bootstrap Methods: Another Look at the Jackknife." Ann. Statist. 7 (1) 1 – 26, January.

El Emam, K. (2013), Guide to the De-Identification of Personal Health Information. CRC Press

Elliot, M., Mackey, E., and O'Hara., K. (2016), The Anonymization Decision Making Framework. UK Anonymization Network, Manchester

Eurostat (2007), European Statistical System Code of Practice Peer Reviews: The National Statistical Institute's guide, Luxembourg. https://unstats.un.org/unsd/dnss/docs-nqaf/Eurostat-Peer%20review%20NSI%20guide.pdf

Eurostat. *European Statistical System Peer Reviews Guide for NSIs and Other National Authorities*, Retrieved November 12, 2021, from https://ec.europa.eu/eurostat/documents/64157/4372828/5-Guide-for_NSIs-and-ONAs.pdf/57cf3841-b210-48c7-9e5b-1575ce2646c2

Fleishman, A. (1978), A Method for Simulating Non-Normal Distributions.

Fuglede, B.; Topsoe, F. (2004). "Jensen-Shannon divergence and Hilbert space embedding" (PDF). *Proceedings of the International Symposium on Information Theory, 2004*. IEEE. p. 30. doi:10.1109/ISIT.2004.1365067. ISBN 978-0-7803-8280-0. S2CID 7891037.

Goodfellow, I. Pouget-Abadie, J. Mirza, M. Xu, B. Warde-Farley, D. & Ozair, S. Courville, A. and Bengio, Y. (2014), Generative Adversarial Networks. Advances in Neural Information Processing Systems. 3. 10.1145/3422622.

Hardt, M., Ligett, K., and McSherry, F. (2012). A simple and practical algorithm for differentially private data release. In Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 2, NIPS'12, pages 2339{2347, USA. Curran Associates Inc.

Hawes, M.B. (2020). Implementing Differential Privacy: Seven Lessons From the 2020 United States Census.

Hintoglu, A. A., & Saygin, Y. (2010). Suppressing microdata to prevent classification based inference. The VLDB Journal, 19(3), 385-410. https://doi.org/10.1007/s00778-009-0170-1

Jordon, J., Yoon, J., and Van Der Schaar, M. (2018). Pate-gan: Generating synthetic data with differential privacy guarantees. In International Conference on Learning Representations.

Jorgensen, T. D., Pornprasertmanit, S. Schoemann, A. M. and Rosseel, Y. (2019), semTools : Useful Tools for Structural Equation Modeling.

Kim, H. J. Drechsler, J. and Thompson, K. J. (2021), Synthetic microdata for establishment surveys under informative sampling, Journal of the Royal Statistical Society Series A, Royal Statistical Society, vol. 184(1), pages 255-281, January.

Kaloskampis, I., Pugh, D., Joshi, C., Nolan, L., Synthetic data for public good, ONS Data Science Campus blog, 2019. https://datasciencecampus.ons.gov.uk/projects/synthetic-data-for-public-good/

Kaloskampis, I., Joshi, C., Cheung, C., Pugh, D. and Nolan, L. (2020), Synthetic data in the civil service. Significance, 17: 18-23. https://doi.org/10.1111/1740-9713.01466

Karr, A. F., Kohnen, C. N., Oganian, A., Reiter, J. P., and Sanil, A. P. (2006). A framework for evaluating the utility of data altered to protect confidentiality. The American statistician, 60(3):224{232.

Kohavi, R. and Becker, B. (1996) US Adult Income dataset. UCI Machine Learning Repository. bit.ly/3dcnUmZ

Langsrud, Ø. (2019), Information Preserving Regression-based Tools for Statistical Disclosure Control, Statistics and Computing, 29, 965–976.

Lavallée, P. and Beaumont, J.-F. (2015), Why We Should Put Some Weight on Weights. Survey Insights: Methods from the Field, Weighting: Practical Issues and 'How to' Approach, Invited article, Retrieved from https://surveyinsights.org/?p=6255

L'Ecuyer, P. and Puchhammer, F. (2021), Density Estimation by Monte Carlo and Quasi-Monte Carlo, Monte Carlo and Quasi-Monte Carlo Methods.

Leduc, J., and Grislain, N. (2021). Composable Generative Models. arXiv: 2102.09249v1 https://arxiv.org/pdf/2102.09249.pdf

LeFevre, K., DeWitt, D., & Ramakrishnan, R. (2005). Incognito: Efficient full-domain K-anonymity. Paper presented at the 49-60. https://doi.org/10.1145/1066157.1066164

Li, Ninghui; Li, Tiancheng; Venkatasubramanian, Suresh (2007). "T-Closeness: Privacy Beyond k-Anonymity and l-Diversity". *t-Closeness: Privacy beyond* k*-anonymity and* l*-diversity* (PDF). pp. 106–115. doi:10.1109/ICDE.2007.367856

Liu, F. (2016). Model-based differentially private data synthesis. arXiv: Methodology.

Lohr, S. (1999). Sampling: Design and Analysis. Duxbury Press.

Sweeney, L. (2002). k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness, and Knowledge-Based Systems, 10(5), 557-570. https://doi.org/10.1142/S0218488502001648

Machanavajjhala, A., M. Hay and X. He. (2017). Differential Privacy in the Wild: A Tutorial on Current Practices and Open Challenges. Proceedings of the 2017 ACM SIGMOD International Conference on Management of Data. 1727-1730.

Machanavajjhala, A., Kifer, D., Gehrke, J., and Venkitasubramaniam, M. (2007). -Diversity: Privacy beyond k-anonymity. ACM Trans. Knowl. Discov. Data 1, 1, Article 3 (March 2007), 52 pages. DOI = 10.1145/1217299.1217302 http://doi.acm.org/10.1145/1217299.1217302

McMahan, H. B. and Andrew, G. (2018). A general approach to adding differential privacy to iterative training procedures. ArXiv, abs/1812.06210.

Mendes, R., & Vilela, J. P. (2017). Privacy-preserving data mining: methods, metrics, and applications. *IEEE Access*, *5*, 10562-10582.

Muralidhar, K., Sarathy, R. (2008), Generating Sufficiency-based Non-synthetic Perturbed Data, Transactions on Data Privacy, 1(1), 17-33

Nowok, B., Raab, G. M., and Dibben, C. (2015). synthpop : Bespoke creation of synthetic data in R. Package vignette http://cran.r-project.org/web/packages/synthpop/vignettes/synthpop.pdf. Accessed: 2015-02-26.

OECD. (2003). *The OECD Glossary of Statistical Terms*. Https://Stats.Oecd.Org/Glossary/. https://stats.oecd.org/glossary/

Olkin, I. (1987). A Conversation with Morris Hansen. Statistical Science. Vol.2, No.2, pp. 162-179

Park, N., Mohammadi, M., Gorde, K., Jajodia, S., Park, and Kim, Y. (2018). Data synthesis based on generative adversarial networks. Proc. VLDB Endow., page 1071–1083.

Sallier, K and C. Girard. (2018). Toward a Successful Implementation of Synthesis in a National Statistical Agency: A Model for Cooperation. Privacy in Statistical Databases, conference held in Valencia, Spain

NIST. (2021, January 7). *2018 Differential Privacy Synthetic Data Challenge*. NIST. Retrieved October 20, 2021, from https://www.nist.gov/ctl/pscr/open-innovation-prize challenges/past-prize-challenges/2018-differential-privacy-synthetic.

R Core Team (2016) R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project. org

Raab GM, Nowok B (2017) Inference from fitted models in synthpop. R Vignette,

URL https://cran.r-project.org/web/packages/synthpop/vignettes/inference.pdf

Raab, G. M., Nowok, B., Dibben, C. (2021). “Assessing, visualizing and improving the utility of synthetic data.” Paper submitted to UNECE Work Session on Statistical Data Confidentiality 2021' Available from https://arxiv.org/pdf/2109.12717.pdf.

Rancourt, E. (2019) The scientific approach as a transparency enabler throughout the data life-cycle. Statistical Journal of the IAOS 35, 549-558.

Read T, Cressie RC (1988). *Goodness-of-Fit Statistics for Discrete Multivariate Data*. NAC, Springer, Berlin.

Ridgeway, D. , Theofanos, M. , Manley, T. and Task, C. (2021), Challenge Design and Lessons Learned from the 2018 Differential Privacy Challenges, Technical Note (NIST TN), National Institute of Standards and Technology, Gaithersburg, MD, [online], https://doi.org/10.6028/NIST.TN.2151, https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=931343 (Accessed October 20, 2021)

Ruggles, S., Flood, S., Goeken, R., Grover, J., Meyer, E., Pacas, J., and Sobek, M. "PUMS USA: Version 8.0 Extract of 1940 Census for U.S. Census Bureau Disclosure Avoidance Research [dataset]," Minneapolis, 2018.

Ruiz N., Muralidhar K., Domingo-Ferrer J. (2018) On the Privacy Guarantees of Synthetic Data: A Reassessment from the Maximum-Knowledge Attacker Perspective. In: Domingo-Ferrer J., Montes F. (eds) Privacy in Statistical Databases. PSD 2018. Lecture Notes in Computer Science, vol 11126. Springer, Cham. https://doi.org/10.1007/978-3-319-99771-1_5

Sala, C., Xu, L., Xia, K., Hofmann, F., Campo, M., Kolev, P., Montanez, A., Skoularidou, M., Pérez, J., Brugman, S., Zhang, K., and Brenninkmeijer, B. (November 8, 2021). SDGym. Retrieved on November 12, 2021 from https://github.com/sdv-dev/SDGym.

Sallier, K. (2020) 'Toward More User-centric Data Access Solutions: Producing Synthetic Data of High Analytical Value by Data Synthesis'. Statistical Journal of the IAOS, vol. 36, no. 4, pp. 1059-1066.

Sallier, K and C. Girard. (2018). Toward a Successful Implementation of Synthesis in a National Statistical Agency: A Model for Cooperation. Privacy in Statistical Databases, conference held in Valencia, Spain

Snoke, J., Raab, G. M., Nowok, B., Dibben, C., and Slavkovic, A. (2018). General and specific utility measures for synthetic data. Journal of the Royal Statistical Society. Series A: Statistics in Society., 181:663{668.

Soria-Comas, J., Domingo-Ferrer, J., Sanchez, D., & Martinez, S. (2015). t-closeness through microaggregation: Strict privacy with enhanced utility preservation. IEEE Transactions on Knowledge and Data Engineering, 27(11), 3098-3110.

Ting, D., Fienberg, S.E., Trottini, M. (2008), Random orthogonal matrix masking methodology for microdata release, International Journal of Information and Computer Security, 2(1), 86-105.

Van Buuren, S., Brand, J. P. L. Groothuis-Oudshoorn, C. G. M. and Rubin, D. B. (2006), Fully Conditional Specification in Multivariate Imputation. Journal of Statistical Computation and Simulation 76 (12): 1049–64.

Vale, C. D., and Maurellim V. A. (1983), Simulating Multivariate Nonnormal Distributions. Psychometrika 48 (3): 465–71.

Voas, D. and Williamson, P. (2001). Evaluating goodness-of-fit measures for synthetic microdata. Geographical and Environmental Modelling, 5(2).

Woo MJ, Reiter JP, Oganian A, Karr AF (2009). "Global Measures of Data Utility for Microdata Masked for Disclosure Limitation." *Journal of Privacy and Confidentiality*, **1**, 111–124.

Wood, A., Altman, M., Bembenek, A., Bun, M., Gaboardi, M., Honaker, J., Nissim, K., O'Brien, D., Steinke, T. and Vadhan, S. (2018). Differential Privacy: A Primer for a Non-Technical Audience. Vanderbilt Journal of Entertainment and Technology Law 21 (1) 209.

Xu, L., Skoularidou, M., Cuesta-Infante, A., and Veeramachaneni, K. (2019). Modeling tabular data using conditional gan. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alch´e-Buc, E. Fox, and R. Garnett, editors, Advances in Neural Information Processing Systems, volume 32, pages 7335–7345. Curran Associates, Inc.