Norsk Regnesentral
NORWEGIAN COMPUTING CENTER

Note

A brief overview of methods for synthetic data for official statistics

Note no SAMBA/23/20

Authors
Gunnhildur Högnadottir Steinbakk
Øyvind Langsrud
Anders Løland

Date 1st July 2020

Norwegian Computing Center

Norsk Regnesentral (Norwegian Computing Center, NR) is a private, independent, non-profit foundation established in 1952. NR carries out contract research and development projects in information and communication technology and applied statistical-mathematical modelling. The clients include a broad range of industrial, commercial and public service organisations in the national as well as the international market. Our scientific and technical capabilities are further developed in co-operation with The Research Council of Norway and key customers. The results of our projects may take the form of reports, software, prototypes, and short courses. A proof of the confidence and appreciation our clients have in us is given by the fact that most of our new contracts are signed with previous customers.

Norsk Regnesentral
Norwegian Computing Center
Postboks 114, Blindern
NO-0314 Oslo, Norway

Besøksadresse
Office address
Gaustadalléen 23
NO-0373 Oslo, Norway

Telefon · telephone
(+47) 22 85 25 00
Telefaks · telefax
(+47) 22 69 76 60

Internett · internet
www.nr.no
E-post · e-mail
nr@nr.no

Title	**A brief overview of methods for synthetic data for official statistics**

Authors	**Gunnhildur Högnadottir Steinbakk , Øyvind Langsrud** <oyvind.langsrud@ssb.no> **Anders Løland** <anders.loland@nr.no>

Date	1st July 2020
Publication number	SAMBA/23/20

Abstract

The access of quality data is valuable, and there is an increased attention to sharing data within the public sector and to other actors. Synthesising the data can be a way that satisfies both the requirements of preserving information value and privacy, which is necessary to enable efficient and safe data sharing. We give a summary of important references, relevant for Statistics Norway and other national statistics institutes, for generating synthetic data for (essentially) statistical purposes.

Keywords	Statistical disclosure, utility measures, imputation, masking, categorical data, differential privacy, microdata
Target group	To whom it may concern
Availability	Open
Project	Big Insight SSB
Project number	220712
Research field	Technology, industry and administration
Number of pages	22
© Copyright	Norwegian Computing Center

Contents

1 Introduction

The demand and volume of data containing sensible information on persons or enterprises have increased significantly over the last several years. At the same time, privacy protection principles and regulations have imposed restrictions on the access and use of individual data. Synthetic data that mimics characteristics from the original data without exposing the real data can be useful when privacy protection is important. This report is part of the early efforts to build capacity and investigate the usability of synthetic data for statistical and analytical purposes in Statistics Norway (SSB) and in the Norwegian public sector in general. Some theoretical (Langsrud, 2019) and application-oriented (Heldal and Iancu, 2019) work has already been done at Statistics Norway. The latter is about a special form of synthetic data that involves sampling from registers.

Note that synthetic data can be viewed as an alternative or a supplement to remote access systems where the microdata is hidden from the user. Although, synthetic data might also be generated within such a system. Statistics Norway and the Norwegian Centre for Research Data have established a system for easy remote access to some Norwegian registers with confidentiality on the fly. The system is called microdata.no (Heldal et al., 2019).

There has been an increased interest in developing systems that generate synthetic data in the public sector (Kaloskampis et al., 2019). The on-going digitalisation of the public sector requires and expects more data sharing, while there is a need for proper privacy protection. Also, the general developments and improvements in computer science technology have increased the demand for access to data on the micro level.

The national statistical institutes gather valuable information on many different aspects of society. An important role for these institutes is to provide and communicate official statistics related to the society, population and economy. To stimulate to research and developments in official statistics, broad access to data from theses institutes is desirable (see, i.e., Nowok et al., 2016; U.S. Census Bureau, 2018).

Different statistical disclosure control (SDC) methods are used by national statistical institutes to control and limit the risk of re-identification and attribute disclosure of datasets (Elliot and Domingo-Ferrer, 2018). Traditional techniques for protecting confidentiality are, amongst others, perturbations, categorising, noise obfuscation, encryption and masking (Hundepool et al., 2014, 2012; Templ et al., 2015) . However, synthetic data and other SDC methods are only one step in preserving privacy in a service, a system or a database. Other important tasks for privacy preserving can be access control, data curation and management.

The goal of synthetic data is to generate a useful data set whilst protecting data confidentiality. The usefulness of the data depends on what the synthetic data is used for and in which setting these data is applied. Replicating underlying patterns in the real dataset will be necessary if the synthetic dataset is used for analytical purposes, whilst a simpler routine may be used for generating synthetic data when testing code and information systems.

However, even for testing information systems, rich, representative and realistic synthetic data might be important in order to take the full advantage of such data. One example is the ongoing modernisation process of The Norwegian National Registry (Behjati et al., 2019; Tan et al., 2019). An important step in this process is building a test environment with a synthetic population, to integrate information-system testing across organisations and consumers of the Norwegian National Registry. The goal is to make a synthetic population representative of the Norwegian population, that takes the dynamics of events that may happen to individuals in the real populations into account. Also simulations of inter-personal relations, such as marital and parental relations, are essential to simulate realistic scenarios. An unrepresentative dataset, with a relatively small number of combinations of personal data-attributes, is likely to fail to find results matching different needs of queries from the approximately 2 000 users of this system (Tan et al., 2019).

As of today, there is no tradition of synthesising data at SSB. Nor are there any guidelines on how to apply it and in which context it is an appropriate methodology. Synthesising the data could be a way that satisfies both the requirements of preserving information value and privacy, which is necessary to enable efficient and safe data sharing. Providing methodologically sound guidance of standards and working solutions for synthetic data may prepare SSB for sharing data and knowledge.

Collecting and recording microdata, in which each record contains several variables concerning a person or an organisation, is an important part of the national statistical institutes activities. Thus, before describing the state of the art for synthetic data, we will briefly discuss microdata in the context of traditionally methods for SDC within official statistics.

1.1 Microdata and disclosure risk

Microdata are individual data on persons, establishments and enterprises, that Statistics Norway and other national statistical agencies are mandated to collect from surveys and registers. The primary objective of this section is to briefly describe the traditional SDC methods most commonly used on microdata and to link synthetic data generation in the context of SDC.

In a typical microdata set, the variables are classified as:

- *Identifiers*: Variables that unambiguously identify the person or organization (e.g. passport number).
- *Key variables (quasi-identifiers)*: Variables which identify the person or organization with some degree of ambiguity (e.g. name, gender, age).
- *Confidential variables*: Variables containing (private or sensitive) information not available in external sources.

Before release of a microdata set, the identifiers are removed. The SDC methods for microdata deals with the key variables and the confidential variables. These two variable types are related to the two main risks of disclosure that need to be controlled.

- *Identity disclosure*: In combination the key variables allow linkage with external information to disclose the identity of some records.
- *Attribute disclosure*: An intruder can determine a confidential variable of a specific person or organization.

It should also be emphasised that SDC is only one step of the process of releasing data. These disclosure risks depend on its intended applications and must be considered within the whole context of the data release framework.

Table 1 summarises commonly used SDC methods applied to microdata, where the main methods are divided into two main categories, non-perturbative and perturbative methods. The non-perterbative methods involve information reduction and coarsening data, while the other change the data. Different SDC methods are used for different types of disclosure control.

Traditionally, the perturbative methods are grouped into two categories:

- Perturbative masking: A modified version of the microdata set is made.
- Synthetic microdata generation: The data are randomly drawn from a statistical model.

In addition, we have hybrid synthetic data generators, which generate data by combining the original and synthetic data.

An overview of classical perturbative masking methods within the above categories is given in Table 1. A formal distinction between synthetic generators and perturbative masking methods is impossible, even under a classical definition of synthetic data (Table 1). More about SDC methods for microdata can be found in official statistics literature, such as Hundepool et al. (2012) and Templ (2017). In some examples in the literature, data synthesis is not viewed as an SDC method. Anyway, the goal is the same: to release useful data whilst maintaining confidentiality.

1.2 Overview

In the following, we give a literature review and a summary of important references, relevant for SSB and other national statistics institutes, that will support the work on establishing guidelines on generating synthetic data for statistical purposes. This includes statistical metrics for measuring similarity between the synthetic and the real dataset and methods for testing the disclosure risk of synthetic dataset. We end this report by discussing some topics related to synthetic data.

Method	Classification of SDC	Description
Sampling	Non-perturbative	A sample is released instead of the whole dataset.
Global recoding	Non-perturbative	Also known as generalisation or coarsening. Numerical values are replaced by intervals or rounded values. For categorical data, the categories are combined.
Top/bottom coding	Non-perturbative	Values above/below a threshold are grouped together.
Local suppression	Non-perturbative	Certain values of certain records are replaced with missing values.
Noise addition	Perturbative: masking	Only applicable for continuous data. Add noise to each record, either uncorrelated/white noise or taking into account covariance structure.
Data swapping	Perturbative: masking	Values are exchanged randomly among individual. records. For continuous data, a variant is rank swapping, which swaps values within a range.
Post randomization (PRAM)	Perturbative: masking	Each categorical value is changed to a new value according to a matrix of transition probabilities (Markov matrix).
Micro-aggregation	Perturbative: masking	A method is first used to group the records into clusters. The aggregates or averages within these clusters are released instead of the individual record values.
Fully synthetic	Perturbative: synthetic data	A dataset where all variables are randomly drawn from a statistical model.
Partial synthetic	Perturbative: synthetic data	Some variables are as in the original data and the rest are randomly drawn from a statistical model (which involves all the variables).
Hybrid data	Perturbative: hybrid data	Data are generated in such a way that the result lies between (partially) synthetic data and original data.

Table 1. Statistical disclosure control (SDC) methods for microdata with a classical definition of synthetic data.

2 A literature review of methods for synthetic data

Originally, Rubin (1993) viewed the synthesised data generation as an example of multiple imputation (MI), where the synthetic populations were generated from their posterior predictive distribution. As opposed to fully synthetic data, the term partially synthetic data is used when only some parts of the data set, usually the sensitive variables, are synthesised (Little, 1993; Reiter, 2003). I addition, the term hybrid synthetic data is used when data are generated in such a way that the result lies between (partially) synthetic data and original data. The first approaches for generating synthetic data were based on parametric modelling. During the 2000s, non-parametric machine-learning methods, such as tree based methods and support vector machines, became more popular (see, e.g. Drechsler and Reiter (2011)).

Recently, more involved machine-learning techniques, such as deep-learning, has been used. In a recent work, a generative adverserial neural network (GAN) was applied for synthesising a clinical trial dataset (Beaulieu-Jones et al., 2019). A research team from the Data Science Campus has used several deep-learning techniques such as GAN, autoencoders and synthetic minority over-sampling (Joshi et al., 2019; Kaloskampis et al., 2019). Other work using deep-learning techniques to generate synthetic data includes Xu et al. (2019) and Behjati et al. (2019).

In addition, Goncalves et al. (2020) compare systematically several methods for generating synthetic patient data that handles multivariate categorical data, such as Bayesian network, mixture of multinomial products, categorical latent Gaussian process and (modified versions of) GAN. For evaluating the synthetic data, they use different criteria that addresses accuracy on individual sample and population level, in addition to metrics measuring disclosure risk. Note that this study focuses on categorical variables and does not include continuous or ordinal variables.

The interest in developing systems for generating synthetic data across the public sector is increasing. The Office for National Statistic (ONS) in the UK published their findings on a pilot study regarding demands and requirements for synthetic data. They included advantages and disadvantages for some open-source available software for generating synthetic data, applied on microdata from the Labour Force Survey (Bates et al., 2018). According to Kaloskampis et al. (2019), the best hope is probably to generate customised synthetic datasets in the future and not relying on a universal design tool, due to the complexity and varied nature of the different datasets.

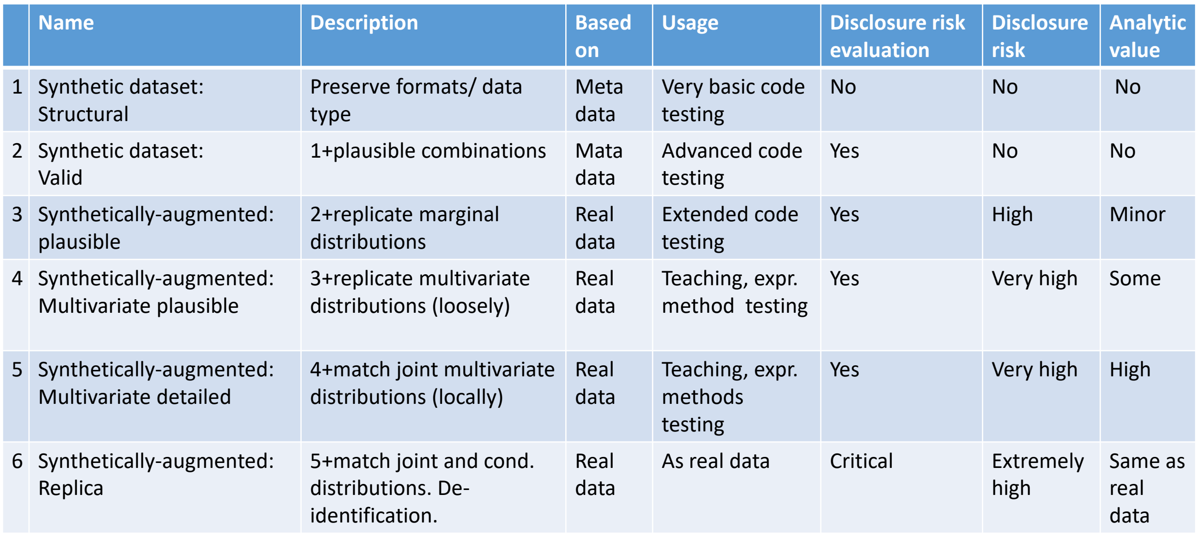
Making correct and realistic inference based on synthetic data might be very difficult, or even impossible, as the model producing the synthetic data will not be the true mechanism generating the true data (Bates et al., 2018). One approach is to release synthetic products for use in an initial and exploratory analysis of the data, while the final analysis is carried out on the real data set in a safe setting controlled by the "data owner institution" or national statistics agencies (see, e.g., Jarmin et al., 2014; U.S. Census Bureau, 2018). Within such a setting, the synthetic data need to be useful and as close to the real data as possible, but they will never be used in the final analysis.

Since 2007, the US Census Bureau has published a publicly available synthetic version of the dataset Survey Income and Program Participation (SIPP), called SIPP Synthetic Beta (U.S. Census Bureau, 2018). This synthetic dataset has been updated regularly, mainly by adding new, synthesised variables and modelling improvements since its first release. All variables in the dataset are synthesised sequentially by drawing variables from their conditional marginal posterior predictive distributions, using normal linear regression, logistic regression or Bayesian boostrap models depending on the type of variables. Their approach also take familial linkages into account.

synthpop is a freely available R package (Nowok et al., 2016). It has been developed by three longitudinal studies in UK (Raab et al., 2018). synthpop also synthesise the dataset sequentially (Drechsler, 2011), preserving their conditional distributions approximately, using non-parametric (classification and regression trees) or parametric methods (regression methods such as linear or logistic) in addition to random sampling with replacement (Nowok et al., 2016). To prevent synthesising nonsense data, specific rules can be implemented. synthpop has an option for applying statistical disclosure control to the synthesised data, such as top and bottom coding and removing any unique cases that are identical to unique individuals in the real data (with the `sdc()` function).

The Office of National Statistics in the UK (ONS) categories the synthetic data into six data spectrums according to their disclosure risk and analytic value (Bates et al., 2018). The more the synthetic data resembles the real data, the higher the analytic value and the disclosure risk of the synthetic data are. The findings of Bates et al. on data spectrums are summarised in Figure 1.

Microdata often include time to event data (i.e., longitudinal studies), where both the event itself and the time of that event occurred are of interest. Thus, the synthesised data might have to resemble realistic realisations of events and the time aspect of these events, including censoring of events that occurs when subjects or records do not experience the event of interest during the follow-up time. To deal with time to event data, Raab et al. (2018) synthesised an event indicator first and then the follow-up



	Name	Description	Based on	Usage	Disclosure risk evaluation	Disclosure risk	Analytic value
1	Synthetic dataset: Structural	Preserve formats/ data type	Meta data	Very basic code testing	No	No	No
2	Synthetic dataset: Valid	1+plausible combinations	Mata data	Advanced code testing	Yes	No	No
3	Synthetically-augmented: plausible	2+replicate marginal distributions	Real data	Extended code testing	Yes	High	Minor
4	Synthetically-augmented: Multivariate plausible	3+replicate multivariate distributions (loosely)	Real data	Teaching, expr. method testing	Yes	Very high	Some
5	Synthetically-augmented: Multivariate detailed	4+match joint multivariate distributions (locally)	Real data	Teaching, expr. methods testing	Yes	Very high	High
6	Synthetically-augmented: Replica	5+match joint and cond. distributions. De-identification.	Real data	As real data	Critical	Extremely high	Same as real data

Figure 1. Description data spectrum by the Office of National statistics in the UK.

time was synthesised separately for each type of event.

How to preserve string or text data is, however, less well documented than numeric data (Bates et al., 2018). Such information may be of interest to those whose aims include data linkage, although preserving specific names or addresses in itself is not the main focus.

Data sets in general usually consist of a mix of different types (i.e., categorical, ordinal and continuous). Whether the data are continuous or categorical has important implications for the generation of synthetic data. For some models it is easy to generate synthetic data that preserves all sufficient statistics exactly. For categorical data, this is generally more problematic (Burridge, 2003) . For synthetic continuous data, all numbers generated will be different (before rounding) and none of the numbers will be equal to any numbers in the original data. Measures of risk must therefore be calculated in a different way. For example, risk measures based on exact matches are meaningless for continuous data. Thus, distance measurements and intervals must be used.

Bates et al. (2018) mention verification servers as a solution for the analyst to get feedback on the accuracy of the synthetic dataset. The verification servers have access to both the real and synthetic data and can verify specific results requested by the analyst of the synthetic data. However, such systems can have a potential to leak confidential information. One way of limiting risk of disclosure, is to put restrictions on type of queries and giving only coarse responses to user queries (Bates et al., 2018).

Synthetic data are often used for testing it systems within production-like test environments. Patki et al. (2016) proposed a general synthetic data generator, which they named the Synthetic Data Vault, to synthesise complete tables of a relational database. With the aim to resemble the data both statistically and structurally, they used a combination of parametric statistical models and a Gaussian Copula.

A relational database with a synthetic Norwegian population is important in the work on modernisation of the Norwegian National Registry (Behjati et al., 2019; Tan et al., 2019). Their approach for generating the synthetic population is based on techniques from natural language processing, by training Recurrents Neural Network (RNNs) to predict the next state or event for a synthetic person (Behjati et al., 2019).

Finally, generating synthetic populations has been used for simulating scenarios of future populations under varying constraints and profiles, such as within transport modelling (see, i.e., Borysov et al 2019) or for simulating tax policies in Luxembourg (Soltana et al, 2019).

3 Measuring the quality of synthetic datasets

The users of synthetic data would in many cases like the synthetic data to be as close to the real data as possible. However, there will always be a balance between the disclosure risk and the need for realistic synthetic data that have the same characteristics and properties as the original data. A proper evaluation of synthetic data should include a combined assessment of quantitative measures of both utility and disclosure risk (Raab et al., 2018).

3.1 Utility measures

According to Raab et al. (2018), analysis and inference based on the synthetic dataset will never produce exactly the same results as those found using the real data. Although there are exceptions (Burridge, 2003; Langsrud, 2019), it is generally important to quantify how well the synthetic data resembles the original data.

Bates et al. (2018) recommend to evaluate the validity of the synthetic data first, and then to compare the key statistical properties of the synthetic data to the original data to determine how well they are preserved. Here, "validity " means that format and data types are preserved and that the combinations of variables are plausible according to the original data (see the data spectrum 2 in Figure 1). As a minimum, before releasing synthetic data, Raab et al. (2018) recommend to visually inspect the marginal distribution of the synthetic data against the observed data.

The synthetic data are said to have high utility if inference on the synthetic and original data agrees. Sometimes synthetic data utility measures have been classified into *general* and *specific* measures of utility (see, e.g. Snoke et al. (2018)). The latter, also noted as narrow measures, compare differences in specific models between original and released data, such as comparing data summaries and the estimates of models fitted to synthetic data with those from the original data. The general or broad measures compare general differences between the original and synthesised data, such as Propensity Score Measures (Snoke et al., 2018) or distances between distributions (e.g., Kullback-Leibler or Hellinger distance). The specific type measures are more commonly used in the literature than the general ones.

Simple methods of evaluating synthetic data against original data include visual comparisons of univariate distributions of each variable, for example by comparing their histograms, cumulative frequency functions or box-plots. Also, summarising the distribution of continuous variables by its mean, median, different quantiles, outliers or extremes (such as top or bottom 5 % of values) and measures of variability, may be used for comparing synthetic and original data, for example by computing the relative or absolute differences between the pairs of these summarising measures. For categorical data, inspections of relevant tabular data such as cell counts may be applied (Raab et al., 2018). An alternative approach is to examine if confidence intervals of points estimates (for example of the total population in a survey) based on synthetic data agrees with the corresponding confidence intervals based on real data.

In many cases it might be important to preserve correlations and important multivariate structures in the synthetic data. Such relations can be examined by a pairwise scatter plot between the variables in the synthetic dataset (such as gender and age), where the same plot is repeated for the original data, to visualise the difference between the real and the synthetic datasets (Kaloskampis et al., 2019). Also, the pairwise Pearson correlation in the real and synthetic data may be used to investigate the differences between those two correlation matrices, for example by plotting a heatmap or computing an overall mean of the correlation differences (Beaulieu-Jones et al., 2019).

A model-based approach for checking multivariate distributions can be applied to both categorical (such as logistic regression) and continuous variables (such as multiple regression). The regression coefficients estimated based on synthetic and real data can be compared by inspecting (absolute or relative) differences or degree of overlapping confidence intervals. Bates et al. (2018) point out that only the important variables of interest should be included for large data sets, due to the risk of overfitting and computational difficulties if the entire multivariate structure should retain.

Data reduction techniques can be used to determine whether the same important features and patterns in the original and synthetic data have been preserved. Popular data reduction techniques are principal component analysis (PCA) (Abdi and Williams, 2010), t-Distribution Stochastic Neighbor Embedding (t-SNE) (Maaten and Hinton, 2008) and multiple correspondence analysis (MCA) (Greenacre and Blasius, 2006). MCA applies to categorical data and allows one to analyse the pattern of relationships of several categorical dependent variables.

Patterns in categorical variables may be inspected by frequency tables or contingency coefficients, and by evaluating the (relative or absolute) differences in the frequency estimates in each cell (pairs or groups) between the original or synthetic data. Alternatively, we may test the differences in the contingency coefficients using z-scores based on the relative size of a given category between the observed and synthetic data. A final measure of fit by computing the sum of squared z-scores gives a chi-squared measure of fit for the contingency table. Such calculations can verify whether the relationship between variables have been maintained in the synthetic dataset (Bates et al., 2018).

Alternatively, the quality of synthetic data can be quantified by training a set of classifier on the synthetically generated dataset and testing their performance on a test set from the real dataset (Joshi et al., 2019). The results can be summarised by some proper dimension reduction plot (such as t-SNE or PCA) or by computing the average classification accuracy of the classifiers. Similarly, the performance of the chosen set of classifiers can be trained on the real dataset and validated on the synthetic test dataset.

3.2 Measures for disclosure risk

The disclosure risk associated with synthetic data, as with all other data released from statistical agencies, must be taken seriously. There is most likely a disclosure risk, in particular if the synthetic data are constructed with accurate representation of relationships and structure in the original data. Traditionally, the agencies have released their data in the form of cross-tabulation or by other summaries and aggregations. The literature for assessing disclosure risk for such data is vast, however, the disclosure risk measures for synthetic data and other microdata in general, are less developed to handle complexities of a real dataset.

Traditionally, synthetic data are viewed as an alternative to controlling confidentiality risk through SDC methods. Therefore, the risk is often neglected. But the fact that the data are synthetic does not mean that it is not possible to reveal private information from the data.

Reiter and Mitra (2009) consider identification disclosure risks in partially synthetic data, by computing the probabilities of identification conditional on the released data. When generating fully synthetic dataset, the disclosure risk is different than a partial synthetic dataset, as the records do not relate to original record in terms of 1-to-1 correspondence. However, attribute disclosure can happen by for example by using prior knowledge and matching relevant variables, without uniquely identifying the records of a target person in the data sets.

Recently, there have been some work on measuring attribute disclosure risks for categorical data. Elliot and Taub (2019) introduces an adaptation based on previous literature called *targeted correct attribution probability*. A similar measure is introduced by Raab (2019) and is called *correct relative attribution probability*. Hittmeir et al. (2020) describe a generalized method based on the concept of *correct attribution probability*.

4 Discussion

4.1 Synthetic methods vs masking methods

As mentioned in Section 1.1, perturbative methods are grouped into masking methods and synthetic methods. Classical masking methods, such as swapping and PRAM, can make a huge change to original data. On the other hand, it is possible to generate synthetic data in such a way that they are very similar to the original data. This depends on how complicated the model is. A broad class of masking methods falls into the category of matrix masking. As described by Langsrud (2019), the classical synthetic method described by Burridge (2003) can be written as matrix masking. Thus, the distinction between masking and synthesising disappears. Another example that makes the definition of synthetic data problematic, is the method synthetic reconstruction implemented in the R package simPop (Templ et al., 2017). A part of the method is to ensure closeness to expected frequencies from the log-linear model. However, classical synthetic data is about drawing data from the model so that the variation around expected frequencies follows the statistic laws. Such a log-linear method is implemented in the R package synthpop (Nowok et al., 2016).

In the modern machine learning literature, synthetic data is also defined without mentioning sampling from a statistical model. Denman et al. (2020) defines synthetic data like this: *Synthetic datasets should replicate original datasets in a private yet coherent manner such that the synthetic datasets preserve the privacy of the users but retain the crucial aspects of the original dataset*. Such a loose definition is necessary for general machine learning methods to be included. The conclusion from this is that it is not possible to make a precise definition of synthetic data which at the same time distinguishes this from perturbative masking methods. What matters is the properties of the method in terms of utility and privacy and not the name of the method.

4.2 Statistical inference from synthetic data

With a classical definition, synthetic data needs to be drawn according to a model. Under a parametric model, parameters can be estimated from the original data. The parameter estimates are functions of the sufficient statistics and this gives rise to two approaches to data generation. 1) One approach is exact conditional simulation. Then, the entire dataset is simulated conditioned on the sufficient statistics so that these are preserved exactly. With regression models and under normality assumptions, such simulations can be performed relatively easily (Burridge, 2003; Langsrud, 2019). This approach ensures that statistical inference from original and synthetic data gives the same results. 2) With other models, data records are generated instead as independent random draws using estimated parameters as true parameters. Then, the sufficient statistics will not be preserved exactly and the statistical inference will produce divergent results. Multiple simulation (or multiple imputation) can be used for more precise inference. This approach is implemented in the R package synthpop (Nowok et al., 2016). With advanced models, it can be problematic to draw entire records jointly from a distribution.

However, the problem of drawing from a k-variate distribution can be replaced by drawing k times from conditionally specified univariate distributions. This is a practical and popular approach to synthetic data (Drechsler, 2011) and is also implemented in synthpop. Precise inference may not be the main aim of synthetic data. It may be that some statistics are preserved or that the data set is very similar to the original data in other ways. The measure of utility can be tailored as needed and the data can be generated in such a way that this purpose is achieved as far as possible. Machine learning methods are well suited for such tailoring.

4.3 Differential Privacy

A disadvantage of traditional risk assessment is that assumptions about an intruder's knowledge are made. An alternative that avoids this is differential privacy (Dwork and Roth, 2014). The aim of differential privacy is to make sure that, by looking at the released statistics/data, it is not possible to see if any individual's data was included in the original dataset or not. Formally, the privacy guarantee is quite technical and formulated in terms of probabilities. To fulfil the privacy guarantee, random noise must be added to all published statistics. The US Census Bureau is now aiming for differential privacy and they call it the new gold standard in data privacy protection (U.S. Census Bureau, 2020). Differentially private data products from the 2020 Census will be released in March 2021.

It is possible to generate synthetic data in a differentially private way. Then, a differential privacy mechanism must be integrated into the algorithm. One way to achieve this is to add noise to the estimated sufficient statistics (Raab, 2019). Competitions have been arranged to create new methods for differentially private synthetic data (PSCR, 2018).

References

Abdi, H. and Williams, L. J. (2010). Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459.

Bates, A. G., Špakulová, I., Dove, I., and Mealor, A. (2018). Synthetic data pilot. In: Working paper series, Office for National Statistics.

Beaulieu-Jones, B., Wu, Z., Williams, C., Lee, R., Bhavnani, S., Byrd, J., and Greene, C. (2019). Privacy-preserving generative deep neural networks support clinical data sharing. *Circulation: Cardiovascular Quality and Outcomes*, 12.

Behjati, R., Arisholm, E., Bedregal, M., and Tan, C. (2019). Synthetic test data generation using recurrent neural networks: a position paper. In *2019 IEEE/ACM 7th International Workshop on Realizing Artificial Intelligence Synergies in Software Engineering (RAISE)*, pages 22–27. IEEE.

Burridge, J. (2003). Information preserving statistical obfuscation. *Statistics and Computing*, 13(4):321–327.

Denman, Q., Ponsen, M., and Offermans, M. (2020). An investigation of deep learning techniques to generate synthetic data. Technical report, Maastricht University, Department of Knowledge Engineering and Statistics Netherlands (CBS).

Drechsler, J. (2011). *Synthetic Datasets for Statistical Disclosure Control*. Springer, New York.

Drechsler, J. and Reiter, J. P. (2011). An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets. *Computational Statistics & Data Analysis*, 55(12):3232–3243. Available from: http://www.sciencedirect.com/science/article/pii/S0167947311002076.

Dwork, C. and Roth, A. (2014). The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3–4):211–407. Available from: http://dx.doi.org/10.1561/0400000042.

Elliot, M. and Domingo-Ferrer, J. (2018). The future of statistical disclosure control. In: Privacy and data confidentiality methods – a national statistician's quality review, Office for National Statistics.

Elliot, M. and Taub, J. (2019). The Synthetic Data Challenge. In *Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality*. the Hague, the Netherlands, October 29-31 , 2019. Available from: https://statswiki.unece.

org/download/attachments/225935487/SDC2019_S3_UK_Synthethic%20Data%20Challenge_Elliot_AD.pdf.

Goncalves, A., Ray, P., Soper, B., Stevens, J., Coyle, L., and Sales, A. P. (2020). Generation and evaluation of synthetic patient data. *BMC Medical Research Methodology*, 20:1–40.

Greenacre, M. and Blasius, J. (2006). *Multiple correspondence analysis and related methods*. CRC press.

Heldal, J. and Iancu, D.-C. (2019). Synthetic data generation for anonymization purposes. Application on the Norwegian Survey on living conditions/EHIS. In *Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality*. the Hague, the Netherlands, October 29-31 , 2019. Available from: https://statswiki.unece.org/download/attachments/225935487/SDC2019_S1_Norway_Heldal_Iancu_AD.pdf.

Heldal, J., Johansen, S., and Risnes, Ø. (2019). Microdata.no – Safe Access to Register Microdata. In *Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality*. the Hague, the Netherlands, October 29-31 , 2019. Available from: https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2019/mtg1/SDC2019_S7_Norway_Heldal_AD.pdf.

Hittmeir, M., Mayer, R., and Ekelhart, A. (2020). A baseline for attribute disclosure risk in synthetic data. In *Proceedings of the Tenth ACM Conference on Data and Application Security and Privacy*, pages 133–143.

Hundepool, A., de Wolf, P.-P., Bakker, J., Reedijk, A., Franconi, L., Polettini, S., Capobianchi, A., and Domingo, J. (2014). mu-ARGUS user's manual, version 5.1. Technical report, Statistics Netherlands.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E. S., Spicer, K., and de Wolf, P.-P. (2012). *Statistical Disclosure Control*. John Wiley & Sons, Ltd.

Jarmin, R. S., Louis, T. A., and Miranda, J. (2014). Expanding the role of synthetic data at the U.S. Census Bureau. *Statistical Journal of the IAOS*, 30(1-3):117–121.

Joshi, C., Kaloskampis, I., and Nolan, L. (2019). Generative adversial networks (GANs) for synthetic data generation with binary classes. Report, Data Science Campus.

Kaloskampis, I., Pugh, D., Joshi, C., and Nolan, L. (2019). Synthetic data for public good. Report, Data Science Campus.

Langsrud, Ø. (2019). Information preserving regression-based tools for statistical disclosure control. *Statistics and Computing*, 29(5):965–976. Available from: https://doi.org/10.1007/s11222-018-9848-9.

Little, R. J. (1993). Statistical analysis of masked data. *Journal of Official Statistics*, 9:407–426.

Maaten, L. v. d. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605.

Nowok, B., Raab, G., and Dibben, C. (2016). synthpop: Bespoke creation of synthetic data in r. *Journal of Statistical Software*, 74.

Patki, N., Wedge, R., and Veeramachaneni, K. (2016). The synthetic data vault. In *2016 IEEE International Conference on Data Science and Advanced Analytics, DSAA 2016, Montreal, QC, Canada, October 17-19, 2016*, pages 399–410. Available from: https://doi.org/10.1109/DSAA.2016.49.

PSCR (2018). 2018 Differential Privacy Synthetic Data Challenge. The Public Safety Communications Research (PSCR) Division. Available from: https://www.nist.gov/ctl/pscr/open-innovation-prize-challenges/past-prize-challenges/2018-differential-privacy-synthetic.

Raab, G. (2019). Practical Experience with Making Synthetic Data Differentially Private. Simons Workshop on "Data Privacy: From Foundations to Applications", March 6th, 2019, Berkeley, USA. Available from: https://simons.berkeley.edu/talks/tba-47.

Raab, G., Nowok, B., and Dibben, C. (2018). Practical data synthesis for large samples. *Journal of Privacy and Confidentiality*, 7(3):67–97. Available from: https://journalprivacyconfidentiality.org/index.php/jpc/article/view/407.

Reiter, J. (2003). Inference for partially synthetic, public use microdata sets. *Survey Methodology*, 29:181–189.

Reiter, J. P. and Mitra, R. (2009). Estimating risks of identification disclosure in partially synthetic data. *Journal of Privacy and Confidentiality*, 1(1). Available from: https://journalprivacyconfidentiality.org/index.php/jpc/article/view/567.

Rubin, D. (1993). Discussion: Statistical disclosure limitation. *Journal of Official Statistics*, 9(2):461–468.

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(3):663–688. Available from: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssa.12358.

Tan, C., Behjati, R., and Arisholm, E. (2019). A model-based approach to generate dynamic synthetic test data: A conceptual model. In *2019 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW)*, pages 11–14. IEEE.

Templ, M. (2017). *Statistical disclosure control for microdata: methods and applications in R*. Springer, Cham, Switzerland.

Templ, M., Kowarik, A., and Meindl, B. (2015). Statistical Disclosure Control for Micro-Data Using the R Package sdcMicro. *Journal of Statistical Software*, 67(4).

Templ, M., Meindl, B., Kowarik, A., and Dupriez, O. (2017). Simulation of synthetic complex data: The R package simPop. *Journal of Statistical Software*, 79(10):1–38.

U.S. Census Bureau (2018). *Synthetic SIPP Data. Version 7.0*. Survey of Income and Probram Participation.

U.S. Census Bureau (2020). Disclosure Avoidance and the 2020 Census. Webpage with regular updates. Available from: https://www.census.gov/about/policies/privacy/statistical_safeguards/disclosure-avoidance-2020-census.html.

Xu, L., Skoularidou, M., Cuesta-Infante, A., and Veeramachaneni, K. (2019). Modeling tabular data using conditional GAN. In *Advances in Neural Information Processing Systems*, pages 7335–7345.