UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS
Expert Meeting on Statistical Data Confidentiality
26-28 September 2023, Wiesbaden

Making Attribute Information of Synthetic Data Interpretable With the Aggregation Equivalence Level

Lotte Pater (Dienst Uitvoering Onderwijs, Ministry of Education, Culture and Science, the Netherlands);
Sanne C. Smid (Dienst Uitvoering Onderwijs, Ministry of Education, Culture and Science, the Netherlands)
lotte.pater@duo.nl; sanne.smid@duo.nl

Abstract

Interest in synthetic data techniques, including for official statistics, has been rising in recent years. This is in large part because synthetic data is very strong in preventing the identification of specific individuals. At the same time, it is known that synthetic data can contain probabilistic information about characteristics of individuals in the real data (sometimes known as attribute information). If non-statisticians want to make well-considered decisions on the privacy impact of a synthetic dataset, it is essential that they have interpretable estimates of the privacy impact for attribute information specifically. Many organizations already publish aggregated datasets, where attribute information is also relevant. In this paper, we propose the Aggregation Equivalence Level (AEL), which puts the attribute information of synthetic data in the context of attribute information of aggregated data, as measured using the Differential Correct Attribution Probability (DCAP). We also provide a management summary to communicate the key message of this paper to privacy officers, lawyers and managers.

Keywords: synthetic data, aggregated data, disclosure, attribute information, DCAP

Online data archive and supplemental files: https://osf.io/rdpab/

1. Introduction

Synthetic data has risen in prominence as one possible way to make the trade-off between data privacy and data dissemination less strict (Emam et al., 2020). Synthetic tabular datasets are created by fitting an algorithm on a real tabular dataset. This allows one to – ideally – generate a synthetic dataset that is similar enough on the structural level to the real dataset to be fit for information purposes, yet dissimilar enough on the individual level to not contain confidential personal information. For an introduction into the synthetic data literature, we refer to El Emam (2020), Drechsler and Haensch (2023) and Jordon et al. (2022). In the private sector, the number of companies offering synthetic data services globally has grown from 13 in January 2017 to 58 in October 2021 and 99 in February 2023 (Devaux, 2021, 2023). On the public side, synthetic data has also garnered interest for official statistics. For example, the United Nations Economics commission for Europe (UNECE) published a starter guide "Synthetic Data for Official Statistics" (UNECE, 2022).

The statistical disclosure literature distinguishes between different types of information disclosure. Two of the most important types are re-identification and attribute information (Elliot, 2014; Emam et al., 2020; Taub et al., 2018). Disclosure through re-identification refers to the situation where available information about an individual can lead to identification of this individual in the dataset. Attribute disclosure refers to disclosure of some characteristics of a group in the data (sometimes combined with re-identification), with varying levels of certainty (see e.g., Emam et al., 2020; Taub et al., 2018). For example, if a dataset shows that all females between 40-50 years in geographical region X have breast cancer, and you know a female between 40 and 50 in region X, then you know that she has breast cancer. Another example: if a dataset shows that 90% of the girls at school Y have failed their school exams, and you know a girl at school Y, then you know there is a *high probability* that she has failed her exam.

In the statistical disclosure literature, terms such as 'attribute information', 'attribute disclosure' and 'group information' are used with varying exact definitions. In this paper, we use 'attribute information' as a catch-all term to describe all situations where datasets can disclose (probabilistic) information about characteristics of individuals in reality.

The value of synthetic data relies on the assumption that any given synthetic dataset contains so little personal information that it is not problematic to share (Emam et al., 2020). This is assumption is often take for granted, in large part because re-identification risk is "no longer meaningful" (Taub et al., 2018, p. 122) for any fully synthetic dataset, "because it breaks the link between the data subjects and the data" (Taub et al., 2018, p. 122). This leaves attribute information as the type of disclosure most relevant for synthetic data.[1] After all, synthetic data can contain probabilistic information about characteristics of real individuals just as much as real data can.

To quantify the amount of attribute information in synthetic data, Taub et al. (2018) proposed the Differential Correct Attribution Probability (DCAP), based on the work of Elliot (2014). The DCAP is discussed in more detail in Section 2 of this paper. Conceptually, this metric quantifies the amount of new information that can be obtained from a synthetic dataset relative to a the univariate baseline. Using the DCAP, one is able to quantify attribute information for synthetic data. However, this metric is always relative to the specific dataset and context. Consequently, the interpretation of the DCAP is difficult, which makes it very complex for lawyers, privacy officers and managers to weigh the privacy implications of any synthetic dataset using the DCAP.

Therefore, the goal of our study is to provide non-statisticians with guidance to make well-considered decisions on attribute information of a synthetic dataset. We build on the work of Taub et al. (2018), and propose the Aggregation Equivalence Level (AEL). The AEL puts attribute information of synthetic data in the context of attribute information of (censored) aggregated data, as measured by the Differential Correct Attribution Probability (DCAP). Many organizations, especially government organizations, already release aggregated datasets, where attribute information is also relevant. Lawyers, privacy officers and managers are already used to making decisions about privacy implications of aggregated datasets, and often have policy on how to deal

[1] If attribute information with 100% certainty is regarded as its own form of information disclosure, as in e.g. (Emam et al., 2020), it can be argued that this doesn't apply to synthetic data either. However, it should be noted that this depends on the specifics of the synthesis technique.

with this in different contexts and for different types of data. Expressing attribute information of synthetic data in the context of attribute information of aggregated data concretizes attribute information, helps to interpret attribute information of synthetic data, and supports privacy officers, lawyers and managers with decisions on privacy implications of synthetic data.

By proposing the Aggregation Equivalence Level (AEL), we also build on the work of Little et al. (2022). They compared the utility and disclosure risk of synthetic data and samples of microdata, to increase the understanding of disclosure risk of synthetic data. This is especially helpful for organizations that are used to release randomly selected samples of the original data as a Statistical Disclosure Control technique. In this paper, we show how to use the AEL, and how it increases the interpretability of attribute information of a synthetic dataset by putting it in the context of attribute information of aggregated datasets. This is especially useful for organizations that are used to publish aggregated datasets.

The remainder of this paper is organised as follows: in Section 2, we discuss the CAP and DCAP. In Section 3, we propose the Aggregation Equivalence Level, which puts attribute information of synthetic data in the context of attribute information of aggregated data. In Section 4, we discuss an empirical example to illustrate the use of the proposed AEL and show how this increases the interpretability of attribute information of a synthetic dataset. We end with a conclusion and discussion in Section 5. On the Open Science Framework (OSF, https://osf.io/rdpab/) we provide annotated R-code to reproduce our example, as well as a management summary (in English and Dutch) to communicate the key message of this paper to non-statisticians.

2. Differential Correct Attribution Probability (DCAP)

The *Correct Attribution Probability* (CAP) was first proposed by Elliot (2014) and further developed by Taub et al. (2018), specifically to measure attribute information for synthetic datasets. Conceptually, the CAP measures the probability that an intruder guesses an attribute about any specific person right. This specific attribute is called the *target variable* or *target attribute*. For example, imagine an intruder is attempting to guess whether a specific student passed or failed their final high school exam and assume that the total pass rate of 84% is published. With no other information, the most reasonable approach for the intruder is to hedge their bets using the total pass rate. In other words, for every student they guess that that student passed with a probability of 84% and failed with a probability of 16%. In that case the CAP equals 84% for the students who passed and 16% for the students who failed. To calculate the CAP for the entire dataset, we simply take the average over all students. It is easy to see that in this case $CAP = 0.84 \cdot 0.84 + 0.16 \cdot 0.16 = 0.73$. We call the CAP calculated using just the univariate distribution of a variable the *baseline CAP* for that variable.[2]

Note that the Correct Attribution Probability is a probabilistic concept. That is, the attacker does not make a single guess, but rather a probabilistic one. This can also be seen as expressing a measure of belief or certainty. Also note that more diversity in the target variable leads to a lower baseline CAP. If the pass rate was 50%, we would get a CAP of $0.5 \cdot 0.5 + 0.5 \cdot 0.5 = 0.5$, instead of 0.73. This makes sense conceptually, as it is easier to guess an attribute when it has a monotone distribution. One consequence of this is that the CAP for a target variable based on a (synthetic) dataset is not interpretable if you do not know what the corresponding baseline CAP is. A CAP of 0.74 would represent a small increase of 0.01 with a total pass rate of 84% but a rather large increase of 0.24 with a total pass rate of 50%.

[2] A variant metric for continuous variables has been proposed (Elliot, 2014), but is outside the scope of this paper.

	Passed	Failed	Baseline Guess	Baseline CAP	Table-based guess	Table-based CAP	DCAP
School A	0	1			{0%, 100%}	1	0.27
School B	6	2			{75%, 25%}	0.63	-0.11
School C	9	1			{90%, 10%}	0.82	0.09
School D	6	0			{100%, 0%}	1	0.27
Total	21	4	{84%, 16%}	0.73		0.81	0.08

Table 1: Example of the DCAP calculations for a aggregated, non-censored dataset.

The *Differential Correct Attribute Probability* (DCAP) accounts for this. The DCAP is defined as the CAP for a target variable based on a dataset, minus the baseline CAP. Take, for example, the situation where we also know the amount of students who passed and failed on every school, as in the example in table 1. In that case an attacker is able to tailor their guesses based on the school a student attends (the table-based guesses). We call this taking the school as a *key value* for the DCAP. In table 1, the correct attribution probability rises to 0.81 with this new information and the DCAP equals $0.81 - 0.73 = 0.08$. Note that for some students (those attending School A and School D) the attacker can know whether they passed or failed with 100% certainty. The DCAP does not treat this situation as special, so other checks are necessary in situations where this is a specific concern.[3]

One disadvantage of the DCAP is the range of values it can take in practice. Intuitively, one might expect the minimal possible value of the DCAP to be 0 and the maximal possible value to be the value when the full dataset is known (0.08 in this case). Yet, this is not what happens in reality. For example, consider the case where the total pass rate is 84% but an attacker mistakenly beliefs is to be 100%. Their Correct Attribute Probability would be $0.84 \cdot 1 + 0.16 \cdot 0 = 0.84$ and the DCAP would be $0.84 - 0.73 = 0.11$. This is a higher DCAP than if they would have the full dataset, despite the attacker having incorrect information. This usually seems to arise when the attackers guesses do not conform to the univariate distribution of the target variable. Negative DCAP values arise in similar cases, and are also possible when a synthetic dataset has especially low utility (Taub et al., 2018, p. 126).

	Passed	Failed	Baseline Guess	Baseline CAP	Table-based guess	Table-based CAP	DCAP
School A	**<7**	**<7**			**{86%, 14%}**	**0.14**	**-0.59**
School B	6	2			{75%, 25%}	0.63	-0.11
School C	9	1			{90%, 10%}	0.82	0.09
School D	**<7**	**<7**			**{86%, 14%}**	0.86	0.13
Total	21	4	{84%, 16%}	0.73		**0.74**	**0.01**

Table 2: Example of the DCAP calculations for a aggregated dataset censored at n<7. Bolded cells indicate differences from table 1.

The Differential Correct Attributed Probability (DCAP) really shines when the attacker has some but not all information, such as when there is a censored or synthetic dataset available. Take for example table 2, which depicts the same dataset, but has schools with less than 7 students censored. An attacker would not be able to distinguish between school A and school D here and is forced to group them together. In particular, their guess is based on the total passed and failed students among all schools with censored totals. These are $21 - 6 - 9 = 6$ and $4 - 2 - 1 = 1$ respectively, so they will guess a student on any school with a censored total passed with a 86% chance and failed with a 14% chance. In this example, the CAP is 0.74 and the

[3] As mentioned earlier, this is unlikely to happen for synthetic datasets.

corresponding DCAP is just 0.01, which is significantly less than when the attacker has the full dataset. Note that this definition assumes the attacker knows the exact total amounts of passed and failed students in the real dataset. This is a slightly stricter assumption than normally for the DCAP, where the assumption is just that the attacker knows the percentages of the univariate distribution.

	Passed	Failed	Baseline Guess	Baseline CAP	Table-based guess	Table-based CAP	DCAP
School A	**1**	**1**			**{50%, 50%}**	**0.5**	**-0.23**
School B	**8**	**2**			**{80%, 20%}**	**0.65**	**-0.08**
School C	9	1			{90%, 10%}	0.82	0.10
School D	**3**	0			{100%, 0%}	1	0.27
Total	21	4	{84%, 16%}	0.73		**0.80**	**0.06**

Table 3: Example of the DCAP calculations for a synthetic dataset, based on the real dataset in table 1. Bolded cells indicate differences from table 1.

Finally, the Differential Correct Attribute Probability for a synthetic dataset is computed very similarly. Here, the distributions for every school in the synthetic data are used by the attacker in calculating their guesses. Note that the underlying reality does not change. In the example in table 3, this happens to lead to significantly worse guesses for students in Schools A and B, but better in Schools C and D. In total, this dataset has a CAP of 0.80 and a corresponding DCAP of 0.06. This means it contains significantly more attribute information than the aggregated dataset censored at $k<7$, but less than the full aggregated dataset.

3. Aggregation Equivalence Level (AEL)

The concept of the Aggregation Equivalence Level (AEL) is to find the level k where the aggregated dataset contains the same amount of attribute information as the synthetic dataset, or slightly more than that.

In Box 1, the steps of calculating the Aggregation Equivalence Level are presented. First, create a synthetic data set.

Secondly, create multiple aggregated datasets with varying levels of aggregation. Use aggregation levels that cover a range of possible values. For example, start with $k = \{1,2,\ldots,20\}$. A level of $k = 5$ here means that all cells where less than 5 people are included are suppressed by the text "n < 5".

Then, compute the average DCAP for each aggregated dataset and the synthetic dataset. In the fourth and final step, the DCAPs are compared. Choose the level k that has the same DCAP as the DCAP for the synthetic dataset, or – if the levels are not exactly the same - choose the level k such that the DCAP for the synthetic datasets is between the DCAPs aggregated at level k and level $k+1$, to be more conservative. This level k is the Aggregation Equivalence of the synthetic data set.

Box 1: Steps to compute the AEL

1. Create a synthetic data set of the observed data.
2. Create multiple aggregated datasets with varying levels of censoring. k, e.g. $k = \{1,2,\ldots,20\}$.
3. Compute the average DCAP for each aggregated dataset and the synthetic dataset.
4. Compare the DCAPs and choose the level k such that the DCAP for the synthetic dataset is between the DCAPs for datasets aggregated at level k and level $k+1$.

4. Empirical Example: School Exam Results in the Netherlands

We will illustrate the use of the Aggregation Equivalence Level (AEL) by discussing an empirical example. We use freely available online open data from DUO.[4] The data and our example code can be found at the OSF via https://osf.io/rdpab/. The data contains information about all secondary schools in the Netherlands, the number of examination candidates at that school and the number of students that passed the exam. It is further split out among some other categories: the student's gender (male or female), the level of examinations that they are taking and their specialisation profile (e.g. "culture and society" or "maritime and technology"). There are also some variables that are higher level, such as the educational region.

We selected data from the school year 2021-2022, which totals 184.077 students on 1.156 schools. We included 7 variables: one indicating whether the student passed or failed, two with information about the examination level, two with information about the student's specialisation, one with a code corresponding to the school and one with the school region. We used the exam result as the target variable and the other six variables as keys.

School	Level & type		Passed	Failed	Baseline Guess	Baseline CAP	Table-based guess	Table-based CAP	DCAP
00AH00	HAVO	CM	3	0			{100%, 0%}	1	0.10
00AH00	HAVO	EM	35	4			{90%, 10%}	0.82	-0.08
...	...	...	...	...			...	...	...
16PI01	VWO	EM	13	1			{93%, 7%}	0.87	-0.04
Total			174499	9578	{95%, 5%}	0.90		0.91	0.01

Table 4: Part of the DCAP calculation for the aggregated, non-censored dataset of Dutch school exams. The 6 variables used have been condensed into 3 for readability.

We follow the steps to compute the Aggregation Equivalence Level as presented in Box 1.

- Step 1: To create the synthetic dataset, we have used the package *synthpop* (Nowok et al., 2016, version 1.7-0) in R. The data was synthetized using Classification and Regression Trees (CART) (Reiter, 2005). We stratified the synthesis on the school region and used default settings otherwise.

- Step 2: We created 20 censored aggregated datasets based on the open data. The aggregation levels used are $k \in \{1, 2, \ldots, 20\}$. Code to replicate this, and all other output of this paper can be found on the OSF via https://osf.io/rdpab/. We used R with the *tidyverse* packages (Wickham et al., 2019).

- Step 3: We computed the DCAP for the 20 aggregated data sets and the synthetic data set. The code to compute the DCAP in R is partly based on the DCAP code in Python of Taub et al. (2018).

- Step 4: To compare the DCAPs, we present the results in a plot. Figure 1 shows the DCAP for the various aggregation levels *k*, and a constant line representing the DCAP of the synthetic dataset. It can be seen that the average DCAP of the synthetic dataset is 0.0081. The average DCAPs of the aggregated datasets vary between 0.0048, when the dataset is aggregated at $k = 20$ and 0.0117, when the dataset is "aggregated' at $k = 1$ (i.e. not aggregated at all). We see that the DCAP is not exactly equal to an aggregated DCAP value, it is between the DCAPs for aggregated datasets censored at $k = 9$ and $k = 10$. Therefore, we choose the highest level *k* where DCAP value is higher than the synthetic DCAP. In our example, this would be aggregation level $k = 9$. This means that the synthetic dataset contains the same amount of or less information in terms of attribute information, as the aggregated dataset where the aggregated data is censored at $n < 9$ (i.e., $k = 9$).

[4] Data downloaded from https://duo.nl/open_onderwijsdata/voortgezet-onderwijs/aantal-leerlingen/examens.jsp#examenkandidaten-en-geslaagden

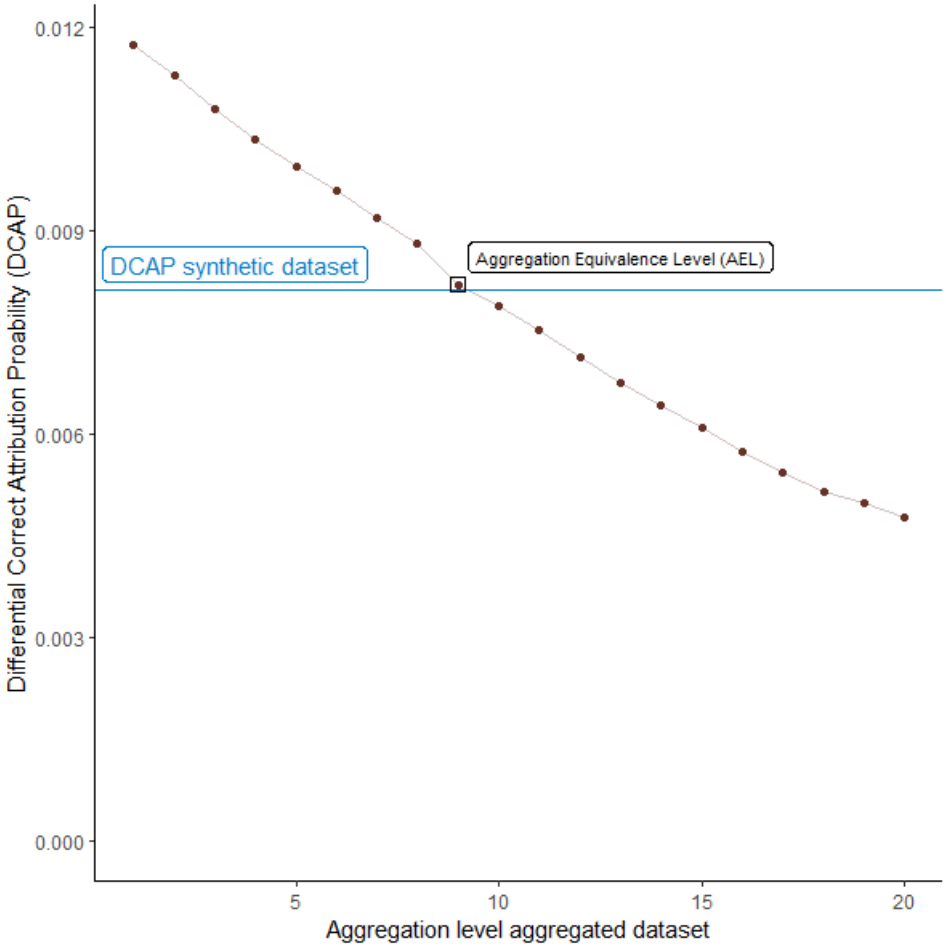

Figure 1: Picking the Aggregation Equivalence Level (AEL) by comparing the DCAPs of a synthetic dataset with aggregated datasets censored at different levels.

5. Discussion

In this paper, we propose the Aggregation Equivalence Level (AEL) to put attribute information of synthetic data in the context of attribute information of aggregated data. We discuss an empirical example and show how to use the AEL in practice.

We believe that the AEL is a promising metric to make decisions about privacy implications of synthetic data easier. However, there are also some limitations that need to be further investigated. First, we considered comparing the targeted CAP (TCAP) (Taub et al., 2018) alongside the DCAP. This metric quantifies the increased risk of attribute information relative to the baseline specifically for individuals unique in the synthetic data. We consider the TCAP a valuable metric for synthetic data and feel that is would be helpful to put this metric in context of aggregated data as well. However, the TCAP cannot be meaningfully computed for aggregated datasets, as single cases are by definition suppressed in the aggregated data. Secondly, as mentioned in Section 2, it is possible that the DCAP has a minimum value lower than the baseline CAP, and a maximum value higher than the full observed data. This can complicate the interpretation.

Directions for future research involve investigating the behaviour of the DCAP to find out under which circumstances it performs well or poorly. Also, we want to carry out a simulation study to investigate the impact on the AEL of varying the key variables in the DCAPs, as well as the sample size, and the number of categories within the target variable.

Note that when a synthetic and an aggregated data set are both published online, information from both datasets could be combined. This would logically lead to a higher amount of attribute information than the attribute information to be gained from either the aggregated or the synthetic dataset. This also applies for all other relevant data available online. When other data can be linked to published synthetic or aggregated data, the amount of attribute information available will increase.

Another point we would like to stress is the importance of clear communication about attribute information to non-statisticians, especially when the interpretation of privacy risk is not clear-cut. We provide a management summary (in English and Dutch) at the OSF (https://osf.io/rdpab), to communicate the main message of this paper to privacy officers, lawyers and managers. We advise against making statements about certain AEL values which are always "safe" or "unsafe", just as it is not recommended that certain aggregation levels are always considered "safe" or "unsafe". We recommended researchers, privacy officers, lawyers and managers, to think about the specific synthetic data set, what kind of information could be disclosed and how sensitive this information is. The desired aggregation level will also vary depending on the context of the data. One can imagine that a synthetic dataset containing information on the number of students that follow a certain specialization within a school contains less confidential information than a dataset about criminal records. The acceptable amount of attribute information of a synthetic dataset is a consideration that should be made in light of the context of the specific dataset.

It is our hope that putting attribute information of synthetic data in the context of aggregated data helps to concretise attribute information, eases the interpretation of attribute information of synthetic data and makes decisions about privacy implications of synthetic data easier. We hope that this study is a starting point for future research to further investigate the behaviour of the DCAP and the use of the Aggregation Equivalence Level.

References

Devaux, E. (2021). The list of synthetic data companies—2021. *Medium*. https://elise-deux.medium.com/the-list-of-synthetic-data-companies-2021-5aa246265b42

Devaux, E. (2023). *Synthetic Data Directory*. Synthetic Data Directory. https://syntheticdata.carrd.co

Drechsler, J., & Haensch, A.-C. (2023). *30 Years of Synthetic Data* (arXiv:2304.02107). arXiv. http://arxiv.org/abs/2304.02107

Elliot, M. (2014). *Final Report on the Disclosure Risk Associated with the Synthetic Data Produced by the SYLLS Team*. https://hummedia.manchester.ac.uk/institutes/cmist/archive-publications/reports/2015-02%20-Report%20on%20disclosure%20risk%20analysis%20of%20synthpop%20synthetic%20versions%20of%20LCF_%20final.pdf

Emam, K. el, Mosquera, L., & Hoptroff, R. (2020). *Practical synthetic data generation: Balancing privacy and the broad availability of data* (First edition). O'Reilly.

Jordon, J., Szpruch, L., Houssiau, F., Bottarelli, M., Cherubin, G., Maple, C., Cohen, S. N., & Weller, A. (2022). *Synthetic Data—What, why and how?* (arXiv:2205.03257). arXiv. http://arxiv.org/abs/2205.03257

Little, C., Elliot, M., & Allmendinger, R. (2022). Comparing the Utility and Disclosure Risk of Synthetic Data with Samples of Microdata. In J. Domingo-Ferrer & M. Laurent (Eds.), *Privacy in Statistical Databases* (Vol. 13463, pp. 234–249). Springer International Publishing. https://doi.org/10.1007/978-3-031-13945-1_17

Nowok, B., Raab, G. M., & Dibben, C. (2016). **synthpop**: Bespoke Creation of Synthetic Data in *R. Journal of Statistical Software*, *74*(11). https://doi.org/10.18637/jss.v074.i11

Reiter, J. P. (2005). Using CART to Generate Partially Synthetic Public Use Microdata. *Journal of Official Statistics*, *21*(3), 441–462.

Taub, J., Elliot, M., Pampaka, M., & Smith, D. (2018). Differential Correct Attribution Probability for Synthetic Data: An Exploration. In J. Domingo-Ferrer & F. Montes (Eds.), *Privacy in Statistical Databases* (Vol. 11126, pp. 122–137). Springer International Publishing. https://doi.org/10.1007/978-3-319-99771-1_9

UNECE. (2022). *Synthetic Data for Official Statistics—A Starter Guide*. https://unece.org/sites/default/files/2022-11/ECECESSTAT20226.pdf

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., … Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software, 4*(43), 1686. https://doi.org/10.21105/joss.01686