UTRECHT UNIVERSITY

MASTER'S THESIS

Synthetic Data Sets for the Analysis of Private Data

Author:
Stijn van den Broek

First supervisor:
Dr. Gerko Vink

Second supervisor:
Prof. Dr. Stef van Buuren

A thesis submitted in fulfillment of the requirements for the degree of Master of Science

Universiteit Utrecht

July 2, 2021

UTRECHT UNIVERSITY

Abstract

Master of Science

Synthetic Data Sets for the Analysis of Private Data

by Stijn van den Broek

Due to recent controversies surrounding the use of private data for analysis, analysing private data for scientific purposes may prove to be problematic. A solution to this problem is the creation of synthetic data sets. These data sets should maintain identical properties to that of the original data as to result in the same inference that the original data would have resulted in. This thesis researches the quality of synthetic data generated by using the multiple imputation capabilities of the MICE package in R. The results of this thesis show that for both univariate analysis and multivariate analysis, the properties of the original data are preserved well, with minimal bias and confidence valid results when using bootstrapped data to generate the synthetic data.

Keywords
MICE, multiple imputation, private data, synthetic data

Acknowledgements

I thank Dr. Gerko Vink for constructing the outline of this thesis and providing me with knowledge and feedback.

I thank Thom Volker for his prior work on synthetic data and his guidance and feedback during the development of this thesis.

Contents

Chapter 1

Introduction

With the focus of the last decade being on online privacy and anonymity - and violations of the aforementioned - analysing personal data for scientific purposes may become problematic [Drechsler, 2011; Isaak and Hanna, 2018]. A solution to this problem would be to create synthetic data sets for the analysis of private data.

1.1 A brief history of synthetic data

The need for correct private data is not something new, with examples of government agencies in the United States and Germany struggling to provide quality data while preserving confidentiality [Drechsler, 2011]. Data is everywhere and can contain valuable information, which could also benefit research. However, the use of private data is only possible if the privacy of the subjects in the data sets can be ensured, and current forms of anonymisation might not always prove to be as effective as previously thought [Ghinita et al., 2007]. Over the years, there have been various approaches to anonymizing private data, including removing sensitive data from data sets [Mohammed et al., 2011], publishing only aggregated data [Krishnamachari, Estrin, and Wicker, 2002], and partially synthesizing sensitive data [Little, 1993; Drechsler, 2011].

Bound to ensure the privacy of subjects in data sets is statistical disclosure limitation. There are two main branches of statistical disclosure limitation, the first being a reduction of the amount of information of a file by making the data less exact, e.g. converting continuous variables to categorized variables. The second branch is maintaining most of the original data, but changing values on the micro-level, e.g. converting values of above 1000 to "1000+" [Drechsler, 2011]. However, these methods reduce the quality of the data [Winkler, 2007], which is undesirable for the analysis of such data.

Rubin (1993) proposed multiply imputed synthetic data sets [Rubin, 1993]. These data sets consist of a random sample of units from the original population where the missing values are imputed with multiple imputation. Using this approach, multiple synthetic data sets are generated and distributed to the public. However, the quality of the synthetic data depends on the accuracy of the model that is used to impute the data. If there is a wrong distribution or a missing relationship between variables, bias could be introduced to the data [Drechsler, 2011].

In the same year, Little (1993) proposed the creation of partially synthetic data, where variables that have a high risk of disclosure are imputed and the rest of the data remains unchanged [Little, 1993]. Compared to Rubin's (1993) proposal, Little's (1993) method suffers less of lowering data quality, as there are less variables in the data that can have a lowered data quality due to being synthetic. These approaches form the basis for generating synthetic data based on original data and the protection of individual privacy.

1.2 Generating synthetic data

Ultimately, the best form of anonymized data would be data that is the same as the original data, but not containing any original data itself. To achieve this, one would need to generate data with the exact same properties as the original data. Generating data with the same properties of the original data is a procedure that is also found in data imputation [Rubin, 1993; Nowok, Raab, Dibben, et al., 2016]. An example of a well performing data imputation approach is multiple imputation. Multiple imputation aims to fill in missing values with samples from an imputation model, essentially generating data that has the same properties as the original data without missingness [Murray et al., 2018].

To generate a synthetic data set, a sample of the original data can be fully imputed despite already existing, or a sample of the original data can be used as a reference for imputing the other values that are missing due to only having a sample. Drechsler (2011) calls the former partially synthetic data, while the latter is called fully synthetic data. However, partially synthetic data can also only have synthetic values for data with a high risk of disclosure [Little, 1993].

1.3 Goals of this research

Solving the problem of generating correct synthetic data increases privacy while at the same time making the exchange of sensitive data safer and easier as synthetic data would make it impossible to trace any data back to an existing person [Graham, Young, and Penny, 2009]. It would therefore be of benefit for third parties and the subjects whose personal details can be found in data sets, to have third parties analyse a synthetic version of the original data.

Due to the similar nature of data imputation and synthetic data generation, this thesis aims to further research the potential, and deliver a proof-of-concept, of using multiple imputation for generating synthetic data. Thus, the questions that this thesis tries to answer are:

RQ1. Is it possible to generate synthetic data with multiple imputation that maintains nearly identical properties to that of the original data for statistical inference?

RQ2. Is it possible for a classification algorithm to be able to distinguish between the original data and synthetic data?

To answer these questions, first the method will be discussed. Then, the results will be described, followed by the discussion. Lastly, RQ1 and RQ2 will be answered in the conclusion.

Chapter 2

Method

The complete code and workspace of this thesis can be found in `https://github.com/Stijnvandenbroek/Synthetic-data`.

2.1 Data used

The data used for this research is the publicly available Pima Indians Diabetes Database [Smith et al., 1988]. The data has 768 rows and 9 columns. The columns consist of several predictor variables such as glucose, skin thickness and BMI, and one binary target variable "Outcome" which indicates whether or not the subject has diabetes. The data set contains missing values in the form of zeroes. Due to this, variables containing missing values can be treated as semicontinuous. There has been chosen for an approach of treating the zeroes that indicate missing values as complete data, as well as singly imputing the missing data first to create some form of truth, and then generating the synthetic data.

2.2 Capturing the properties of the data

To analyse the differences between the original data and the synthetic data, the statistical properties need to be captured for univariate analysis, and a prediction model needs to be established for multivariate analysis.

The differences between the original data and the synthetic data will be assessed univariately, on the level of individual variables; and multivariately, in terms of the relationships between variables. Statistical properties of interest include the mean, standard deviation, skewness, kurtosis and standard error of each variable. For the prediction model the "Outcome" variable will be the target variable, with BMI, Glucose and Pregnancies as the three predictor variables. Due to the binary nature of the "Outcome" variable, the prediction model that will be used is a logistic regression. The model is defined as follows:

$$\text{Outcome} = \beta_0 + \beta_1 \text{BMI} + \beta_2 \text{Glucose} + \beta_3 \text{Pregnancies} + \epsilon.$$

2.3 Generating the synthetic data

To generate the synthetic data, the Mice package [van Buuren and Groothuis-Oudshoorn, 2011] in R will be used for multiply imputing the entire non-imputed original data set, imputed original data set and bootstrapped versions of the aforementioned. This is done for 1000 simulations, with 5 imputations per simulation.

A method is needed for imputing each variable in the data. Predictive mean matching has been proven to work well for imputing semicontinuous data [Vink et

al., 2014], but it is not optimal for generating synthetic data as due to the univariate nature of predictive mean matching, higher order interactions between variables are not preserved very well. Instead, the method for synthesizing the data will be classification and regression trees (CART) [Breiman et al., 1984], as CART has shown promise for generating synthetic data [Reiter, 2005].

The base settings when generating the synthetic data will be a minbucket (the minimum number of observations for a terminal node) of 3, a complexity parameter of 1e-4, and a maxit (the number of iterations when imputing) of 1. Furthermore, there will be experimented with different parameter combinations such as a complexity parameter of 1e-32, and a maxit of 5 for the bootstrapped data.

2.4 Analysing the synthetic data

To analyse the generated synthetic data, the estimates of the synthetic data need to be pooled in order to determine whether or not the estimates are confidence valid. Subsequently, the confidence validity of the pooled data is calculated.

2.4.1 Pooling rules

For each simulation, the estimates of the predictor variables are pooled. To pool the estimates, two pooling functions are used to compare the different results. The first pooling function is the standard pooling function of MICE [van Buuren and Groothuis-Oudshoorn, 2011] which is in accordance with Rubin's rules [Rubin, 1987]. The second pooling function is a pooling function where the pooling rules are designed for partially synthetic data [Reiter, 2003]. The key difference between the standard pooling function according to Rubin's rules and the pooling function for partially synthetic data is that the standard pooling function uses an extra component for calculating the variance compared to the pooling function for partially synthetic data. The variance of the standard pooling function according to Rubin's rules is calculated as:

$$T = \bar{U} + B + \frac{B}{m}.$$

With T being the variance of a given variable, $\bar{U}$ being the mean within imputation variance, B being the variance between estimates of the imputations, and m being the number of imputations. On the other hand, within the pooling function for partially synthetic data, the variance is calculated as:

$$T = \bar{U} + \frac{B}{m}.$$

As a result, using the pooling function in accordance with Rubin's rules creates a greater confidence interval width, which in turn causes a higher coverage rate when calculating the the 95% coverage rate of a variable.

2.4.2 Confidence validity

The resulting pooled data will be analysed by comparing the averages of the statistics and parameters of the synthetic data with that of the original data by calculating the bias of the synthetic data. In addition, the 95% coverage rates of the synthetic estimates will be calculated for both the original and imputed data, as well as the bootstrapped version of each of the aforementioned. The differences in results from

different complexity parameters and numbers of iterations when generating synthetic data will also be compared.

2.5 Classification of mixed data

Finally, to classify the data on whether or not it is synthetic, half of the data of each of the 1000×5 imputations will be sampled and merged with a sample of half the true data. A logistic regression model will then be used to classify the mixed data on whether or not it is synthetic. The reason for using a logistic regression is due to the fact that the variable indicating whether or not the data is synthetic, is binary. An optimal outcome of this classification would yield an accuracy of around 0.5, meaning that the model is unable to perform better than random classification and thus indicating that it is unable to distinguish between true data and synthetic data.

Chapter 3

Results

3.1 Confidence validity

Looking at both the confidence validity of the synthetic data based on the non-imputed original data in table 3.1 and the synthetic data based on the imputed original data in table 3.2, where the standard pooling function in accordance with Rubin's rules is used, it is evident that apart from the intercept showing slight undercoverage, all variables are confidence valid. Furthermore, the synthetic data that is based on data that is not bootstrapped has coverage rates nearing 1.00 for both the imputed and non-imputed original data, which is to be expected as in this case the sampling variance should not be considered [Vink and van Buuren, 2014]. The synthetic data based on bootstrapped data has coverage rates equal to, or slightly exceeding 0.95.

Table 3.1: The bias, confidence interval width and coverage rate of the 95% confidence interval of the synthetic data based on the non-imputed original data, pooled with pooling rules according to Rubin's rules. Depicted are the results for two different complexity parameters for CART.

		CP = 1e-4			CP = 1e-32		
	Term	Bias	CIW	Coverage	Bias	CIW	Coverage
No bootstrap Maxit = 1	(Intercept)	0.97	14.14	0.99	0.98	13.58	0.97
	BMI	-0.01	0.31	1.00	-0.01	0.29	1.00
	Glucose	-0.01	0.08	0.98	-0.01	0.07	0.98
	Pregnancies	-0.03	0.47	1.00	-0.03	0.45	1.00
Bootstrap Maxit = 1	(Intercept)	0.60	11.12	0.94	0.60	11.12	0.94
	BMI	-0.01	0.23	0.96	-0.01	0.23	0.97
	Glucose	-0.00	0.06	0.95	-0.00	0.06	0.95
	Pregnancies	-0.01	0.38	0.97	-0.01	0.36	0.97
Bootstrap Maxit = 5	(Intercept)	0.60	11.12	0.94	0.60	11.12	0.94
	BMI	-0.01	0.23	0.96	-0.01	0.23	0.97
	Glucose	-0.00	0.06	0.95	-0.00	0.06	0.95
	Pregnancies	-0.01	0.38	0.97	-0.01	0.36	0.97

(a) The estimates of the true data are -8.12 for the intercept, 0.08 for BMI, 0.03 for glucose and 0.14 for Pregnancies.

Table 3.2: The bias, confidence interval width and coverage rate of the 95% confidence interval of the synthetic data based on the imputed original data, pooled with pooling rules according to Rubin's rules. Depicted are the results for two different complexity parameters for CART.

		CP = 1e-4			CP = 1e-32		
	Term	Bias	CIW	Coverage	Bias	CIW	Coverage
No bootstrap Maxit = 1	(Intercept)	1.40	12.20	0.84	1.41	12.52	0.85
	BMI	-0.02	0.26	0.98	-0.02	0.28	0.98
	Glucose	-0.00	0.06	1.00	-0.00	0.06	0.99
	Pregnancies	-0.03	0.48	1.00	-0.03	0.50	1.00
Bootstrap Maxit = 1	(Intercept)	0.66	10.03	0.93	0.67	10.00	0.93
	BMI	-0.01	0.19	0.96	-0.01	0.20	0.96
	Glucose	-0.00	0.05	0.96	-0.00	0.05	0.96
	Pregnancies	-0.01	0.37	0.96	-0.02	0.38	0.96
Bootstrap Maxit = 5	(Intercept)	0.66	10.03	0.93	0.67	10.00	0.93
	BMI	-0.01	0.19	0.96	-0.01	0.20	0.96
	Glucose	-0.00	0.05	0.96	-0.00	0.05	0.96
	Pregnancies	-0.01	0.37	0.96	-0.02	0.38	0.96

(a) The estimates of the true data are -8.12 for the intercept, 0.08 for BMI, 0.03 for glucose and 0.14 for Pregnancies. Depicted are the results for two different complexity parameters for CART.

On the other hand, when looking at the confidence validity of the synthetic data based on the non-imputed original data in table 3.3 and the synthetic data based on the imputed original data in table 3.4, where the pooling function for partially synthetic data is used, it is evident that not all variables are confidence valid. In this case, the confidence interval widths (CIWs) are substantially smaller compared to the confidence interval widths of tables 3.1 and 3.2, while having the same bias.

Table 3.3: The bias, confidence interval width and coverage rate of the 95% confidence interval of the synthetic data based on the non-imputed original data, pooled with pooling rules for partially synthetic data. Depicted are the results for two different complexity parameters for CART.

		CP = 1e-4			CP = 1e-32		
	Term	Bias	CIW	Coverage	Bias	CIW	Coverage
No bootstrap Maxit = 1	(Intercept)	0.97	2.61	0.89	0.98	2.59	0.90
	BMI	-0.01	0.06	1.00	-0.01	0.06	1.00
	Glucose	-0.01	0.01	0.90	-0.01	0.01	0.88
	Pregnancies	-0.03	0.11	0.99	-0.03	0.11	0.99
Bootstrap Maxit = 1	(Intercept)	0.60	2.61	0.82	0.60	2.61	0.83
	BMI	-0.01	0.06	0.89	-0.01	0.06	0.89
	Glucose	-0.00	0.01	0.85	-0.00	0.01	0.85
	Pregnancies	-0.01	0.11	0.91	-0.01	0.11	0.92
Bootstrap Maxit = 5	(Intercept)	0.60	2.61	0.82	0.60	2.61	0.83
	BMI	-0.01	0.06	0.89	-0.01	0.06	0.89
	Glucose	-0.00	0.01	0.85	-0.00	0.01	0.85
	Pregnancies	-0.01	0.11	0.91	-0.01	0.11	0.92

(a) The estimates of the true data are -8.12 for the intercept, 0.08 for BMI, 0.03 for glucose and 0.14 for Pregnancies.

Table 3.4: The bias, confidence interval width and coverage rate of the 95% confidence interval of the synthetic data based on the imputed original data, pooled with pooling rules for partially synthetic data. Depicted are the results for two different complexity parameters for CART.

		CP = 1e-4			CP = 1e-32		
	Term	Bias	CIW	Coverage	Bias	CIW	Coverage
No bootstrap Maxit = 1	(Intercept)	1.40	2.65	0.38	1.41	2.64	0.35
	BMI	-0.02	0.06	0.91	-0.02	0.06	0.91
	Glucose	-0.00	0.01	0.98	-0.00	0.01	0.98
	Pregnancies	-0.03	0.11	1.00	-0.03	0.11	1.00
Bootstrap Maxit = 1	(Intercept)	0.66	2.76	0.83	0.67	2.75	0.84
	BMI	-0.01	0.06	0.89	-0.01	0.06	0.89
	Glucose	-0.00	0.01	0.90	-0.00	0.01	0.89
	Pregnancies	-0.01	0.11	0.91	-0.02	0.11	0.91
Bootstrap Maxit = 5	(Intercept)	0.66	2.76	0.83	0.67	2.75	0.84
	BMI	-0.01	0.06	0.89	-0.01	0.06	0.89
	Glucose	-0.00	0.01	0.90	-0.00	0.01	0.89
	Pregnancies	-0.01	0.11	0.91	-0.02	0.11	0.91

(a) The estimates of the true data are -8.91 for the intercept, 0.09 for BMI, 0.04 for glucose and 0.14 for Pregnancies.

When comparing the bias to the true estimates, the bias looks to be very low with the bootstrapped data having a lower bias than its counterpart without bootstrapping. When comparing the results of the different complexity parameters, it seems

as though the different complexity parameters have a very small impact on the results, with minor fluctuations in bias, confidence interval width and coverage rate. In addition, the different numbers of iterations for the bootstrapped data do not seem to have had a noticeable impact on the results. The difference in results for the synthetic data based on non-imputed original data and the synthetic data based on imputed original data shows that the synthetic data based on the imputed original data has a higher bias compared to the synthetic data based on the non-imputed original data.

3.2 Univariate statistics

As is shown in table 3.5, the bias of the synthetic data based on the non-imputed original data is relatively low, indicating that the univariate properties of the original data are preserved quite well.

Table 3.5: The statistics of the original non-imputed data and the bias of the corresponding synthetic data with a complexity parameter of 1e-4 and no bootstrapping.

	Term	Mean	Standard Deviation	Skewness	Kurtosis	Standard Error
True values	Pregnancies	3.85	3.37	0.90	0.14	0.12
	Glucose	120.89	31.97	0.17	0.62	1.15
	BloodPressure	69.11	19.36	-1.84	5.12	0.70
	SkinThickness	20.54	15.95	0.11	-0.53	0.58
	Insulin	79.80	115.24	2.26	7.13	4.16
	BMI	31.99	7.88	-0.43	3.24	0.28
	DiabetesPedigreeFunction	0.47	0.33	1.91	5.53	0.01
	Age	33.24	11.76	1.13	0.62	0.42
	Outcome	1.35	0.48	0.63	-1.60	0.02
Bias of synthetic values	Pregnancies	0.00	-0.00	-0.00	-0.01	-0.00
	Glucose	0.00	-0.02	0.01	-0.03	-0.00
	BloodPressure	-0.01	-0.01	0.00	0.00	-0.00
	SkinThickness	0.00	-0.00	-0.01	-0.03	-0.00
	Insulin	-0.01	-0.11	-0.02	-0.19	-0.00
	BMI	-0.00	-0.00	0.01	-0.05	-0.00
	DiabetesPedigreeFunction	-0.00	-0.00	-0.02	-0.16	-0.00
	Age	0.00	-0.00	-0.00	-0.00	-0.00
	Outcome	0.00	-0.00	-0.00	0.00	-0.00

3.3 Classification performance

As depicted in table 3.6, the results for the mixed data consisting of non-imputed original data and its corresponding synthetic data, and the mixed data consisting of imputed original data and its corresponding synthetic data, are identical. The accuracy and balanced accuracy of a logistic regression model when classifying whether or not mixed data is synthetic are 0.50 for both the non-imputed data mixed with the corresponding synthetic data and the imputed data mixed with the corresponding synthetic data. Furthermore, the Kappa of 0.00 confirms that there is no agreement between the classification and truth values, and that the logistic regression model

does not perform better than the expected accuracy of 0.50 for random classification. Interesting to note is that the sensitivity and specificity are not 0.50, indicating that the logistic regression model has not assigned an equal amount of each class to the data. Overall, table 3.6 shows that the logistic regression model has not been able to distinguish between synthetic data and true data.

Table 3.6: The classification performance of a logistic regression model on mixed data of non-imputed original data and synthetic data, and imputed original data and synthetic data.

Type	Accuracy	Kappa	Sensitivity	Specificity	Balanced Accuracy
Non-imputed	0.50	0.00	0.73	0.27	0.50
Imputed	0.50	0.00	0.73	0.27	0.50

Chapter 4

Discussion

The results in chapter 3 show promise for the possibility of generating synthetic data with MICE. For the amount of iterations when synthesizing, the results show that between a maxit of 1 and a maxit of 5, there is no noticeable difference. This is due to only observed data being used when synthesizing, which causes all iterations to be more or less the same. Therefore, the amount of iterations for the maxit parameter in MICE can be left at a value of 1. Furthermore, the complexity parameter also shows negligible difference in results when using the default value of 1e-4 and a value very close to zero such as 1e-32.

In addition, the coverage rates of the variables of the synthetic data based on the non-bootstrapped, non-imputed original data all are very close to 1. This can be explained by the fact that in that case, the sample happens to be the population. The pooling rules assume the population to be infinite, and this results in over-covered population estimates [Vink and van Buuren, 2014]. Bootstrapping when generating the synthetic data shows that the coverage rates of the variables are more in line with a coverage rate of 0.95 for a 95% confidence interval.

As for the results of the univariate statistics, the bias of the statistics are all relatively low, indicating that for univariate analysis the synthetic data is of a good quality.

When combining half of the non-imputed original data with half of the corresponding synthetic data and using a logistic regression model to classify whether or not the data is synthetic, the resulting accuracy is near 0.5, indicating that it is unable to perform better than random classification.

Therefore, based on the aforementioned results, it is advised to use the standard pooling method in accordance with Rubin's rules as this results in a valid inference.

4.1 Limitations

This thesis focuses on a single data set. Therefore, the performance of similar approaches to many other data sets has not yet been measured. Furthermore, different data sets might require methods of imputation other than CART for better results, which may an influence the performance.

As for distinguishing synthetic data from true data, the classification has only been done with a logistic regression model and not with any other types of models. As it stands, there is a possibility of other models performing better when it comes to identifying synthetic data.

This thesis has not performed an analysis on the synthetic data on the micro-level. Therefore, there is a possibility of synthetic values closely resembling the true data on the micro-level. However, this is unlikely to be the case for the entire data set, therefore not compromising the anonymity of the complete record information for all subjects.

4.2 Implications

The results of this thesis imply that there is a foundation for the use of MICE to generate valid synthetic data with multiple imputation. In the context of maintaining privacy of individuals, this would mean that generating synthetic data with MICE is an accessible alternative to other anonymisation methods currently available. Generating synthetic data with MICE for distribution to third parties would guarantee the privacy of individuals. This can have the effect of data quality increasing for data that is gathered through surveys, as there are less worries of respondents being identified through the distributed data. Furthermore, more sensitive data can be published without censorship as it no longer compromises the privacy of individuals. It would be interesting to see the performance of the approach of this thesis on other data, as well as the performance difference between different pooling rules and imputation methods.

Chapter 5

Conclusion

In conclusion, despite the limitations described in chapter 4, the results of this thesis point towards the possibility of generating synthetic data being a feasible endeavour. The results show that the generated data is highly similar on a univariate level. Furthermore, when using bootstrapped data to generate the synthetic data, the estimates of the variables are confidence valid, indicating that on a multivariate level the synthetic data is highly similar as well. So, to answer RQ1, the results indicate a possibility of generating synthetic data with multiple imputation that maintains nearly identical properties to that of the original data.

In addition, when using a logistic regression model to distinguish between synthetic and true data in a data set consisting of mixed data, the results show that the logistic regression model fails to perform better than random classification. Therefore, RQ2 can be answered with that the results point towards the possibility of the synthetic data being similar enough to the true data that it is impossible for a classification algorithm (in this case a logistic regression model) to be able to distinguish between synthetic and true data.

Bibliography

Breiman, Leo et al. (1984). *Classification and regression trees*. CRC press.

Drechsler, Jörg (2011). *Synthetic datasets for statistical disclosure control: theory and implementation*. Vol. 201. Springer Science & Business Media, pp. 1–5.

Ghinita, Gabriel et al. (2007). "Fast data anonymization with low information loss". In: *Proceedings of the 33rd international conference on Very large data bases*, pp. 758–769.

Graham, Patrick, Jim Young, and Richard Penny (2009). "Multiply imputed synthetic data: Evaluation of hierarchical Bayesian imputation models". In: *Journal of Official Statistics* 25.2, p. 245.

Isaak, Jim and Mina J Hanna (2018). "User data privacy: Facebook, Cambridge Analytica, and privacy protection". In: *Computer* 51.8, pp. 56–59.

Krishnamachari, L, Deborah Estrin, and Stephen Wicker (2002). "The impact of data aggregation in wireless sensor networks". In: *Proceedings 22nd international conference on distributed computing systems workshops*. IEEE, pp. 575–578.

Little, Roderick JA (1993). "Statistical analysis of masked data". In: *Journal of Official statistics* 9.2, p. 407.

Mohammed, Noman et al. (2011). "Differentially private data release for data mining". In: *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 493–501.

Murray, Jared S et al. (2018). "Multiple imputation: a review of practical and theoretical findings". In: *Statistical Science* 33.2, pp. 142–159.

Nowok, Beata, Gillian M Raab, Chris Dibben, et al. (2016). "synthpop: Bespoke creation of synthetic data in R". In: *Journal of statistical software* 74.11, pp. 1–26.

Reiter, Jerome P (2003). "Inference for partially synthetic, public use microdata sets". In: *Survey Methodology* 29.2, pp. 181–188.

— (2005). "Using CART to generate partially synthetic public use microdata". In: *Journal of Official Statistics* 21.3, p. 441.

Rubin, Donald B (1987). *Multiple imputation for nonresponse in surveys*. John Wiley & Sons, p. 76.

— (1993). "Statistical disclosure limitation". In: *Journal of official Statistics* 9.2, pp. 461–468.

Smith, Jack W et al. (1988). "Using the ADAP learning algorithm to forecast the onset of diabetes mellitus". In: *Proceedings of the annual symposium on computer application in medical care*. American Medical Informatics Association, p. 261.

van Buuren, Stef and Karin Groothuis-Oudshoorn (2011). "mice: Multivariate Imputation by Chained Equations in R". In: *Journal of Statistical Software* 45.3, pp. 1–67. URL: `https://www.jstatsoft.org/v45/i03/`.

Vink, Gerko and Stef van Buuren (2014). "Pooling multiple imputations when the sample happens to be the population". In: *arXiv preprint arXiv:1409.8542*, p. 2.

Vink, Gerko et al. (2014). "Predictive mean matching imputation of semicontinuous variables". In: *Statistica Neerlandica* 68.1, pp. 61–90.

Winkler, William E (2007). *Examples of easy-to-implement, widely used methods of masking for which analytic properties are not justified*. Citeseer.