Synthetic data in Official Statistics

Simon Xi Ning Kolb[1]

[1]*Federal Statistical Office of Germany, Germany*

Abstract

While the impact of digitization is still growing at an unprecedented pace, the amount and importance of data has already reached unforeseen dimensions. Data driven decision making has become the paradigm of both the private and public sector alike. As the importance and dependence on high quality data becomes more obvious every day, custodians of official data find themselves in a challenging spot when it comes to making data publicly available.

It is the task of German Federal Statistical Office to provide data to the general public and the independent scientific community. Whenever data is made accessible to the public, one faces the risk of violating data privacy. The standard scenario consists of data collection followed by the publication of aggregates produced from the underlying microdata. It is widely known that aggregation does not offer sufficient protection depending on data granularity and type. Hence, a wide range of so-called post tabular SDC methods have been developed, which modify the aggregated statistics in a way that aims to reduce disclosure risk. These changes will invariably result in reduced data usefulness while leaving the underlying micro data unchanged. However, the federal office of statistics does not confine data publication solely to aggregates. The goal of the Research Data Centre of the Statistical Offices of the Federation and the Federal States is to provide access to microdata for research purposes. This process is strictly governed by the Federal Statistics Law and requires strong data anonymity. So far, achieving thorough microdata anonymization while maintaining statistical usefulness has posed a significant challenge. The rapid advancement in computing power and data availability emphasizes the urgent need for improvements of traditional pre-tabular SDC techniques. In this work, we explore synthetic data as a means to overcome the limitations of traditional SDC methods for both microdata and aggregates. By synthesizing the microdata, post tabular methods can be avoided, as all contributions to a table are synthetic values. Maintaining analytical utility in the synthesized data and reducing disclosure risk is however no trivial task. We investigate the potential of a synthetic data approach with one use case.

Keywords: synthetic data, data privacy, tabular data

1. Introduction

The thorough protection of official statistics is a challenging task. Although the disclosure control department of the German Federal statistics office is armed with a range of tools, straightforward application is the exception rather than the rule. The prevailing method involves suppressing disclosive table cells. To avoid disclosure-by-differencing attacks, secondary suppression or coarsening is additionally applied. Another known method is the Cell Key method[11,22](CKM), which adds controlled noise to cells instead of complete suppression. Occasionally, we are faced with a statistic which does not align well with either approach. The road traffic accident statistics fall into this category. Whenever a traffic accident happens on a public road, the police officer on duty files a report describing the accident. These files are then transferred to the statistical offices of the states where they undergo certain plausibility checks before being forwarded to the Federal Statistical Office for SDC treatment.

This statistic consists of a huge dataset with more than 500 variables, both categorical and numerical. The overwhelming majority of variables are deemed non-sensitive. Sensitive variables are nonetheless included in very coarse aggregations and made public. Additionally, the German Federal and State Statistical offices maintain a geocoded open data file of non-sensitive microdata. The combination of open data file and coarse tables are released before the main body of low level and hence sensitive aggregates. This makes Cell suppression extremely difficult to use in the face of many publicly known marginal sums (i.e. known, because they could be derived from the open data file), which cannot be used in secondary suppression any longer. The Cell Key Method is also unsuited due to the pre-released table margins, since one can only apply noise to inner table cells. Given fixed table margins, an attack on the CKM protected inner cells can be mapped to a constrained optimization problem, which might produce unique “solutions” for the original values of the CKM protected inner cells, and hence reduce protection afforded to those cells.

In this work, we use a synthetic data approach modifying only the sensitive variables and evaluate the data quality and risk via selected tables meant for publication.

Throughout the paper, we refer to aggregated data as a *table*. Whenever we talk about non-aggregated microdata, we will use the term *microdata*.

2. Methods

The idea of using synthetic data in order to limit disclosure risk stems from a study by Rubin[1] in 1993. There he maps the task of disclosure control to an imputation problem. Coming from the context of survey statistics he proposes to treat unsampled units as missing. By doing so, one can then use an imputation model to create synthetic samples which replace the missing data. This approach is what we would call a *fully synthetic dataset* nowadays. The main advantage of this option is the high level of protection it offers, since no original observations are released to the public. However, finding an ideal model which captures all the multivariate relationships between variables remains highly challenging as of today. A less perturbative approach was published by Little[2] (1993). He proposed to limit the synthesis to sensitive variables only, thus leaving the dataset partially unchanged. Because one only changes a subset of all variables, the resulting quality of data is improved, albeit the risk of disclosure will be potentially higher. This led to the naming of *partially synthetic datasets*. The partial synthesis approach can be taken further still by limiting the synthesis to a *subset of data units.* The choice of target units which are to be synthesized might be selected given some user defined criteria or in the simplest case, purely at random.

Most applications of data synthesis rely on either a joint or sequential modelling framework. The joint modelling approaches[4,5] intend to approximate the fully joint distribution and are increasingly studied in the recent years by the deep learning community through GAN-based models[6]. The sequential approach on the other hand, factorizes the full distribution into many conditional distributions. This way, sensitive variables are synthesized sequentially and potentially conditioned on each other. Using this framework, one can freely choose the model that describes the conditional distributions for each variable. In a partial synthetic setting, this offers the possibility of leveraging the statistical information present in the variables which are to be left unchanged, by using these as predictors for the synthesis model. Possible model choices range from simple linear regression to more complex machine learning models. Reiter[7] further developed this framework and introduced the usage of CART in 2005.

2.1 Application to the traffic accident dataset

In our study, we fully rely on the sequential framework. By using the R package synthpop[8], which offers implementations of many synthesis methods, we get access to the aforementioned CART models. This package has become the first choice for data synthesis experiments as well as a baseline for benchmarking in the scientific community.

In a sequential setting, the researcher has to specify two important things. The first one being the conditional relationships between variables in the form of a predictor matrix. This object specifies for each variable to be synthesized, the corresponding predictor variables. The more informative predictors used for the synthesis of a target variable, the better the fitted model. However, care needs to be taken when blindly allowing for the maximum number of predictors possible. The reason being, that synthesis will slow down drastically for CART models trained on categorical predictors with high feature dimensionality. This is due to the binary splitting procedure underlying CART fitting, which will exhaustively search for the best possible split among all predictor values. In order to alleviate this computational limitation, we performed a feature selection via random forests and additionally removed all predictors with feature dimensionality higher than 20. The second specification which greatly impacts the synthesis is the variable sequence[9]. Due to the sequential character of this framework, variables synthesized later will be potentially less accurate if these use prior synthesized variables as predictors, as fluctuations will inevitably propagate down through the models. Here we followed the recommendations of the package creators.

We additionally stratified the synthesis, by treating the units from every German Federal State as a separate dataset. While this did not improve the data quality much, it did cut down the computation time. The use of the rules feature of synthpop, in order to guarantee plausibility was also necessary due to many deterministic dependencies between variables.

3. Evaluation methods

The synthetic data community has developed a large body of measures for determining data quality. However, since data synthesis is mostly concerned with microdata, the vast majority of *utility metrics* are not designed for tables, which is the main data format used in publications by statistical agencies. Other perturbative disclosure control methods like the cell key method[11,12] are therefore partly guided by rather simple rules like controlling the maximally allowed deviation of a table cell. In order to give a practical example of the potential workflow during the publication process, we restricted our evaluation to measures related to cell count (since we are only producing frequency tables from the traffic accidents statistic) deviations.

The Federal and the Statistical Offices of the German states have a high double-digit number of different tables they tend to publish. We focussed our analysis on a small sample of those, which are representative of the general table structures. These tables tend to include three to six variables, including the Community Identification Number, which acts as a fixed key variable. Aggregation is performed mostly on the three aggregation levels: municipality, district and state. We evaluated the quality of the synthetic data by tabulating and comparing with the original tables. By doing so, we get a quality estimate of the actual data product.

3.1 Utility measures

Although a single measure should never be enough to investigate the overall data quality, we worked on an adequate formulation of a quality measure that tries to capture the spirit of the publication requirements. We used the *averaged frequency of deviates* as a utility measure:

$$u(s,b) = \frac{1}{2}\left(\frac{n_{small}(s)}{N_{small}} + \frac{n_{big}(b)}{N_{big}}\right) \quad .$$

We denote a cell in the synthetic table as a *small cell* when the **original** count is $n \leq 10$. N_{small} is the number of these small cells and N_{big} the number of all bigger cells. Then, $n_{small}(s)$ is the number of *small cells* where the synthetic cell count differs from the original count by more than a user defined number s. To account for the larger count range of bigger cells, $n_{big}(b)$ is defined as the number of cells where the relative error $\frac{|n_{syn,big}-n_{orig,big}|}{n_{orig,big}}$ between synthetic and original cell count is larger than a user defined number b. This measure produces values between 0 and 1, and can be understood as the average of small and big fractions of deviations.

3.2 Risk measures

The main disclosure scenario we face here is group disclosure. Example*: "Attacker A and neighbour N live in a small village. Attacker A knows of a car accident of his neighbour N in 2022 on the main road. A knows that N drives a SUV and looks up an appropriate table for his municipality. There he finds out, that in 2022 all traffic accident involving SUVs were due to drunk driving."*

For the risk measure we extracted the cells in the synthetic and original tables which constitute a group disclosure. We then computed the set overlap between these two sets of cells and used the fraction $\frac{common\ disclosive\ cells}{synthetic\ disclosive\ cells}$ as our risk measure[10].

4. Results

In the first step, we fully synthesized the sensitive variables and tabulated accordingly. For the evaluation, the tables made from synthetic data are then compared to the original tables on all aggregation levels. In Fig. 1 we show the distribution of absolute and relative deviations for two selected tables. Both tables are 5 dimensional. Additionally, table A includes one sensitive and table B two sensitive variables. We observe from the left plots that small cells tend to have small deviations and the majority do not change cell counts at all. Although this applies to all aggregation levels, it is more prominent in lower aggregations. Large deviations are observed first of all on state level. We note here, that the fraction of small cells on municipality level is considerably smaller (municipality.: ~99%, state: ~70%) than at state level, and that the distribution of small cell counts at state level is shifted towards the higher end (10). The impact of increasing the number of synthesized variables is barely visible in the longer tales of the histogram for table B. The difference becomes more apparent when we look at relative deviations of large cells, where the peak has clearly moved towards larger relative deviations. We note here that synthetic microdata does lead to rare outliers (deviations extending beyond the x-axis of Fig. 1) in table cells, ranging from 0.0001%-0.002% of cells. The simplest way to address these non-plausible counts is post hoc, by marking outlier cells and providing a short note to inform users that these cell counts are probably unreliable.

In order to reduce deviations, we went on to synthesize random subsets of the whole dataset, followed by the standard tabulation procedure we described above. The smaller the synthesis fraction, the less we perturb the microdata. Less perturbation goes hand in hand with an increase in disclosure risk. In order to monitor the changes in quality and risk, we visualized the changes in both quantities in a Risk-Utility map in Figure 2. The choice of measures follows our risk and utility definitions in section 3. The RU map clearly shows the expected increase in data quality, when the proportion of synthetic units is reduced. Simultaneously, the global risk

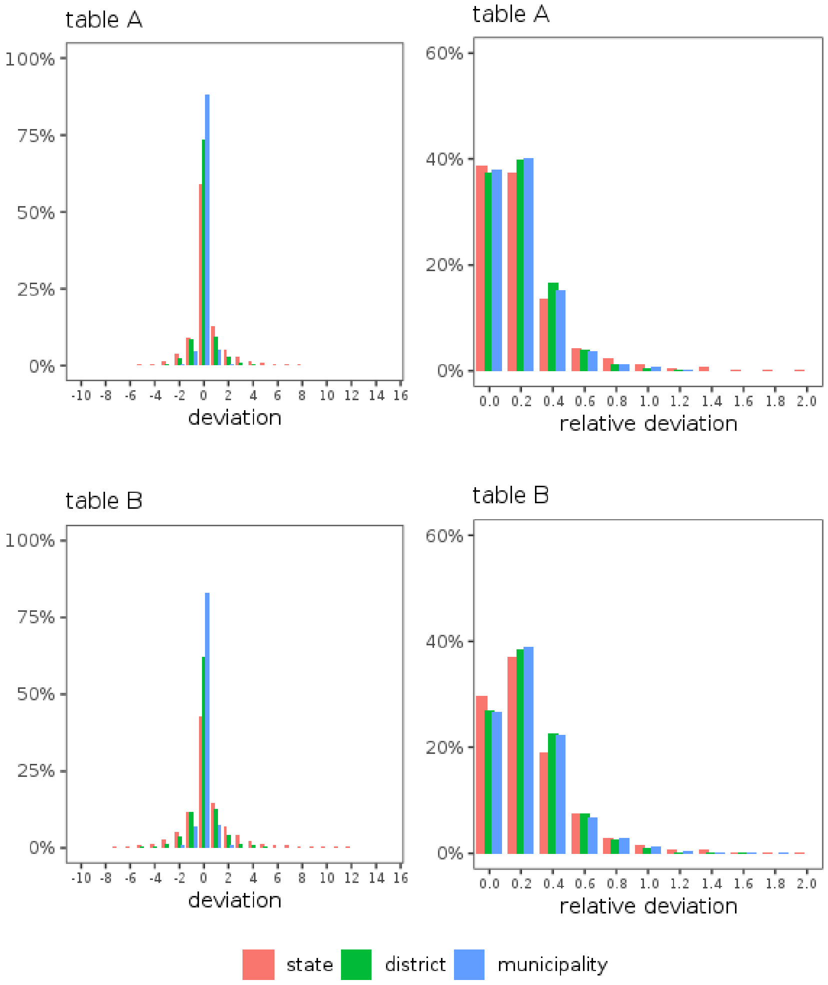

Figure 1: Relative frequency of absolute and relative deviation for two selected tables. Both tablesHistogram bins are centred on the ticks. We only considered cells with original value <= 10 for the absolute deviation plot. For relative deviation, cells with original value > 10 were considered. Rare outliers were cut off visually for better interpretability.

increases, as less perturbation leads to an increased number of disclosive cells being left unchanged. We observe further, that the difference in risk and utility between the two tables grows smaller as we reduce the ratio of synthetic units. This is due to the randomness of subsample selection, which makes selecting units involved in group disclosure unlikely, because these constitute a minority. By increasing the synthetic ratio and changing more disclosive units, the impact of synthesizing two variables will be stronger, as the noise of sequential synthesis tends to grow with the number of variables. This noise becomes more apparent in the tables as more units change, affecting risk positively and the utility negatively. We also see, that the synthetic approach still allows for less protection on lower aggregates, as these tables tend to be much larger and more sparsely populated.

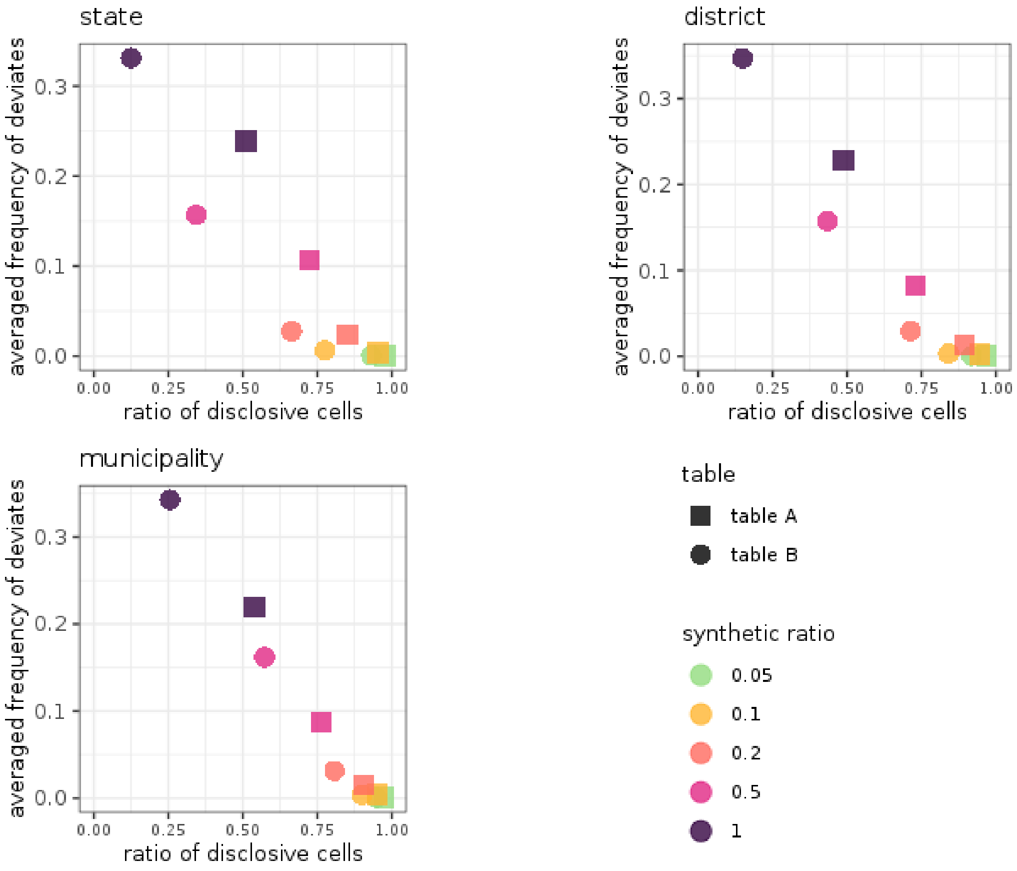

Figure 2: Random subset synthesis in RU-map format for two tables. Disclosure risk on x axis and utility on the y axis with parameter choice (s=3, b=0.3, see section 3.1).

4. Conclusions

We explored the application of a synthetic data approach for one specific statistic. By making use of a flexible synthesis approach, the data provider can tune the synthetic microdata to fit his/her requirements. Even when synthesizing only a subset of the data, a potential attacker cannot infer with certainty, if a disclosive cell is in fact a disclosure case or merely an artefact of the data synthesis. By keeping the synthesis ratio confidential, additional uncertainty is introduced. Additionally, we provided a simple and modifiable utility metric, that is closely related to standard measures of tabular deviations, allowing the data provider to incorporate custom criteria in terms of tolerable deviations.

In this work, we studied the data quality based on a selection of tables. While this produces valuable insights with respect to this selection, we are unable to predict the accuracy of unknown custom tables generated by the statistical agencies. Although we made sure to apply our analysis to a representative selection, given the large number of variables in the dataset, one could cross-tabulate arbitrary combinations of variables. This means that we cannot guarantee a specific level of tabular quality or risk, if the table structures are not known prior to publication. Further research needs to be done, in order to give a prior estimate of risk and utility.

References

[1] Lawn, M., & Nóvoa, A. (2013). The European Educational Space: New Fabrications. *Sisyphus – Journal of Education*, *1*(1), 11-17. https://doi.org/10.25749/sis.2827

[2] Rubin, D. B. (1993). Discussion: Statistical disclosure limitation. *Journal of Official Statistics 9*, 462–468.

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

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

[5] Hu, J., J. P. Reiter, and Q.Wang (2014). Disclosure risk evaluation for fully synthetic categorical data. J. Domingo-Ferrer (Ed.), *Privacy in Statistical Databases,* Number 8744 in Lecture Notes in Computer Science, pp. 185-199. Heidelberg: Springer

[6] Camino, R., C. Hammerschmidt, and R. State (2018). Generating multi-categorical samples with generative adversarial networks. *arXiv:1807.01202* [cs, stat].

[7] Reiter, J. P. (2005d). Using CART to generate partially synthetic, public use microdata. *Journal of Official Statistics 21*, 441–462.

[8] Nowok, B., G. M. Raab, and C. Dibben (2016). synthpop: Bespoke creation of synthetic data in R. *Journal of statistical software 74*, 1–26.

[9] Raab, Nowok and Dibben (2017). Guidelines for producing Useful Synthetic Data. *arXiv:1721.04078v1* [stat.AP].

[10] Geyer, Tent, Reiffert and Giessing (2022). Perspectives for Tabular Data Protection – How About Synthetic Data? *International Conference on Privacy in Statistical Databases* pp 77-91. Springer.

[11] Fraser, B., Wooton, J. (2006): A proposed method for confidentialising tabular output to protect against differencing. *Monographs of Official Statistics*. Work session on Statistical Data Confidentiality, Eurostat-Office for Official Publications of the European Communities, Luxembourg, pp. 299–302

[12] Giessing, S., Tent, R. (2019): Concepts for generalising tools implementing the cell key method the case of continuous variables. *Joint UNECE/EurostatWork Session on Statistical Data Confidentiality,*The Hague, 29–31 October 2019. http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2019/mtg1/SDC2019_S2_Germany_Giessing_Tent_AD.