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

Generating synthetic data using REaLTabFormer, and assessing the probabilistic measure of statistical disclosure risk

Aivin Solatorio (World Bank) and Olivier Dupriez (World Bank)
asolatorio@worldbank.org and odupriez@worldbank.org

Abstract

This paper introduces two implementations of REaLTabFormer, a GPT-based transformer model designed for generating synthetic microdata, both relational and non-relational. We provide a concise overview of the model and evaluate its performance using a benchmark dataset provided by the US National Institute of Standards and Technology (NIST). Furthermore, we utilize training data from diverse sources to create a synthetic census dataset for an imaginary country. This dataset serves to assess the accuracy of probabilistic statistical disclosure risk measures implemented in the sdcMicro and μ-Argus software applications. Finally, we propose an alternative approach to measure the risk, which harnesses a synthetic superpopulation.

Disclaimer

1 Introduction

Synthetic microdata has the potential to serve as a substitute for collected data that cannot be shared due to legal or ethical restrictions (UNECE, 2023). However, the acceptance of synthetic data in the research community relies heavily on the ability of data generation models to accurately capture and reproduce the statistical characteristics of real data. Striking a balance between achieving a high level of statistical similarity and maintaining disclosure risks within acceptable limits remains a formidable challenge, especially for relational data. Existing open-source models for generating relational synthetic data primarily rely on hierarchical modelling algorithms (Patki et al., 2016), which are based on traditional statistical models and Gaussian Copulas. Unfortunately, these models often fail to adequately capture the complex relationships within and across tables (Solatorio and Dupriez, 2023). In this paper, we propose the REaLTabFormer model as an alternative approach. Section 2 provides a brief introduction to the model. Section 3 presents benchmark results using data from the non-relational Diverse Communities Data Excerpts published by the US National Institute of Standards and Technology (NIST). In Section 4, we outline the process of creating a comprehensive synthetic dataset for an imaginary country, leveraging multiple data sources. We employ this dataset to evaluate the probabilistic measure of statistical disclosure risk implemented in sdcMicro and μ-Argus, while also proposing an alternative approach for assessing this risk. Finally, in the conclusion, we suggest potential avenues for further research.

2 REaLTabFormer: a brief overview

REaLTabFormer (Solatorio and Dupriez, 2023) is a generative model designed for producing both relational and non-relational tabular data. The model leverages the transformer-decoder architecture of GPT-2, originally developed for autoregressive tasks, to generate non-relational tabular data, which we refer to as "parent tables." For generating relational data, or "child tables," the model adopts a sequence-to-sequence (Seq2Seq) architecture as introduced by Yun et al. (2019). The encoder network incorporates the weights of the network trained to generate parent tables as input for producing child tables through the decoder network. To ensure data integrity, each column (variable) in the training dataset is independently encoded using a specific token vocabulary. This encoding method, inspired by IBM's TabFormer model (Padhi et al., 2020), enables the model to assign a zero probability to invalid values for any given column.

While REaLTabFormer does not provide a guarantee of differential privacy, it incorporates several privacy safeguards. Firstly, the model utilizes a target masking procedure as a form of regularization, with the objectice of minimizing the likelihood of the generative model "memorizing" and replicating records from the training data. This involves introducing missing (masked) values for a certain proportion of the data. The model then learns to predict the masks instead of the actual masked values. During the generation of synthetic records, the model fills in the masked values with probabilistically determined values, thereby reducing the probability of exact record replication from the training data. In our experiments, we employed a mask rate of 10 percent.

Secondly, the model incorporates an automatic detection and prevention mechanism to address overfitting during training. Overfitting, a common challenge in deep learning models, especially when applied to small datasets, can result in the generation of observations that closely mimic the training data. To mitigate the issue, the model implements an overfitting prevention method based on the analysis of the distribution of the distance to closest record (DCR) measure as proposed by Park et al. (2018). For each observation, the DCR is calculated as the minimum distance between a synthetic data record and the records in the training data. A quantile difference statistic, denoted as Q_d, is derived from the distribution of DCR. REaLTabFormer utilizes Q_d to identify instances where the distance between the synthetic data sample and the training data approaches zero, indicating potential overfitting. The model bootstraps over random samples from the training data to establish a threshold for Q_d, which acts as a signal for overfitting. The measure is regularly estimated during the model training process, and training automatically terminates once the threshold is reached.

3 Benchmarking REaLTabFormer using NIST data and standardized report

We compare the performance of REaLTabFormer with the Maximum Spanning Tree (MST) model by McKenna, Miklau, and Sheldon (2021). The MST model was the winning entry in the NIST Differential Privacy Synthetic Data Competition in 2018. Our comparison involves generating a synthetic or "deidentified" version of the national 2019 partition of the Diverse Communities Data Excerpts published by the US National Institute of Standards and Technology (NIST) under its Collaborative Research Cycle.[1]. These data were extracted from the American Community Survey conducted by the US Census Bureau. The analysis focuses on the ten variables from the Demographic-Focused Subset: SEX (sex), MSP (marital status), RAC1P (race), OWN_RNT (own or rent housing), PINCP_DECILE (person's income discretized as a 10% percentile bin relative to the income distribution in their Public Use Microdata Area), EDU (highest educational attainment), AGEP (age), HOUSING_TYPE (single housing unit or group quarters), DVET (disability due to military service), and DEYE (vision difficulty).

For the differentially private MST model, we adopt the same settings as the authors, running it with an epsilon=10 DP budget and a pre-processor_epsilon=1. We run REaLTabFormer with its default parameters, including a masking rate of 10 percent. Both models are instructed to generate a synthetic dataset of the same size as the training data. To assess and compare the output of the models, we employ the SDNist Deidentified Data Report Generator [2], a tool that provides metrics for evaluating and reporting the utility and privacy of synthetic data generators (Task et al., 2023). The quality metrics presented below have been extracted from the standardized SDNist reports. In terms of the evaluation metrics, both models achieve a high k-marginal score[3], with MST holding a slight advantage. However, REaLTabFormer outperforms MST in several other measures, including propensity mean square error and the number of inconsistencies (Table 1), propensities distribution, structure of principal components, Pearson correlations, and linear regression models.

Table 1. Summary measures from the SDNist Deidentified Data Report Generator, MST and REaLTabFormer

	k-marginal score	Propensity mean square error	Number of inconsistencies	Unique target data records exactly matched in deidentified data*	Number of target data records exactly matched in deidentified data on quasi-identifiers**
MST	**966**	0.010	1017	**7.41%**	101 (0.37%)
ReaLTabFormer	957	**0.003**	**18**	9.48%	**90 (0.33%)**

* Refers to sample uniques (considering all variables) that are present in the synthetic data.

** RAC1P, OWN_RENT, MSP, SEX, EDU

Propensities distribution

High-quality synthetic data should not be easily distinguishable from the training data. The SDNist produces a chart (Figure 1) that displays the distribution of data samples over 100 propensity bins to assess how easily it is to distinguish training data from synthetic ("deidentified") data. For synthetic data of high quality, the two lines should align. Both models perform well, although with an advantage to REaLTabFormer.

Pearson correlation coefficient difference

The Pearson pairwise correlations between variables are calculated for the two synthetic data, and the difference with the training data are reported in the charts in Figure 2. A perfect synthetic data would result in

[1] The Diverse Communities Data Excerpts consist of a selection of 24 variables drawn from the American Community Survey. They are provided for three geographic partitions, including a national partition with 27,253 observations. The data dictionary is available at https://github.com/usnistgov/SDNist/tree/main/nist%20diverse%20communities%20data%20excerpts

[2] We used version 2.2 available at https://github.com/usnistgov/SDNist/releases/tag/v2.2.0

[3] "*The k-marginal metric checks how far the shape of the deidentified data distribution has shifted away from the target data distribution. It does this using many 3-dimensional snapshots of the data, averaging the density differences across all snapshots. It was developed by Sergey Pogodin as an efficient scoring mechanism for the NIST Temporal Data Challenges, and can be applied to measure the distance between any two data distributions. A score of 0 means two distributions have zero overlap, while a score of 1000 means the two distributions match identically.*" (Extracted from SDNist report)

an all-white chart, as lighter colours indicate that correlations are better preserved. REaLTabFormer outperforms the MST model in this aspect. The same conclusion is reached with the Kendall coefficients, also reported by the SDNist report (not shown).

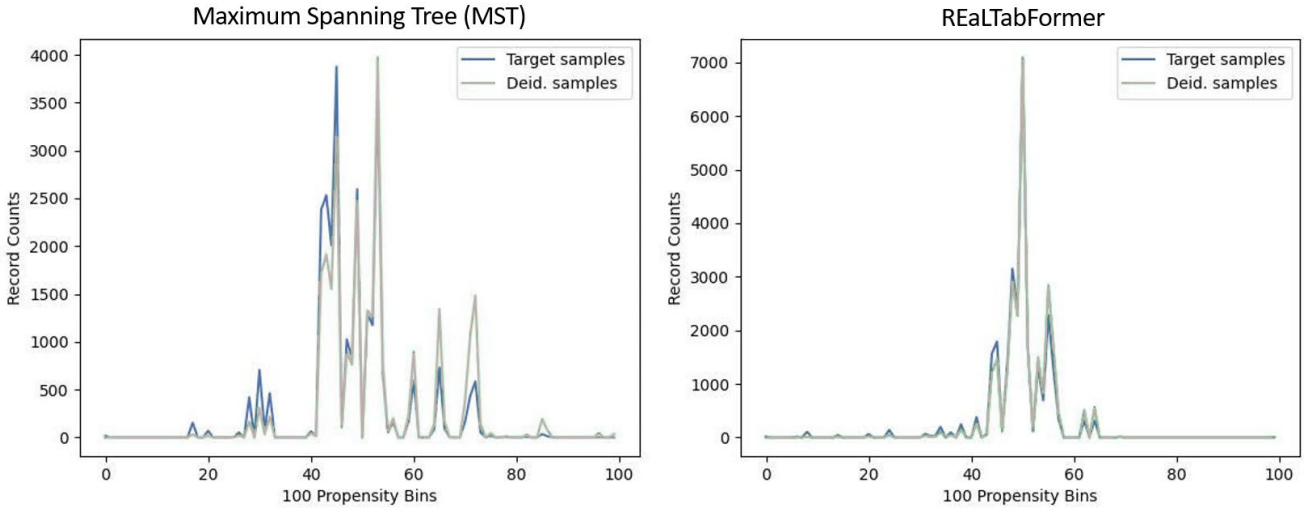

Fig. 1. Distribution of data samples over 100 propensity bins, MST (left) and REaLTabFormer (right)

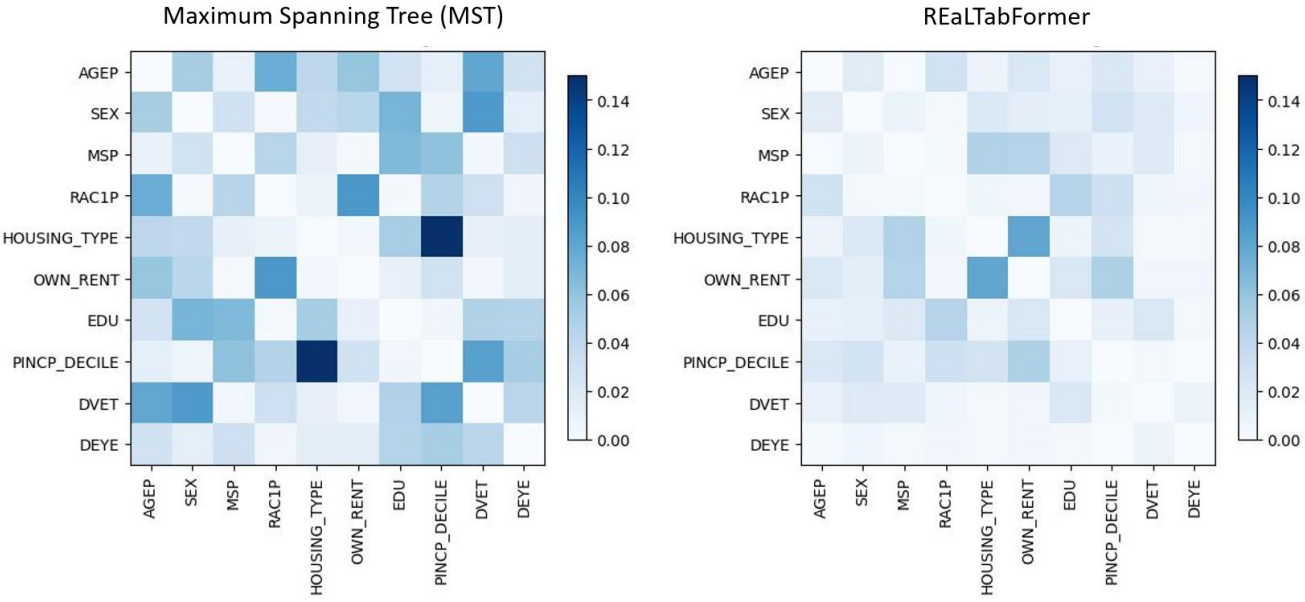

Fig. 2. Pearson Correlation Coefficient Difference, MST (left) and REaLTabFormer (right)

Principal components

The quality of the model is also assessed by analysing the contribution of features in each principal component (PC). Zooming in on the PC-0 and PC-1 pair panel (Figure 3) and highlighting in red the individuals that satisfy a given constraint (in this case MSP = "N" for AGEP < 15, i.e., individuals who are unmarried because they are children) provides further evidence of the capability of REaLTabFormer to generate high-utility data. The features that contribute to the PC and their contribution ratio are as follows:

- PC-0: RAC1P (0.16), OWN_RENT (0.07), DEYE (0.04), HOUSING_TYPE (0.03), SEX (0.01)
- PC-1: HOUSING_TYPE (0.71), MSP (0.23), DEYE (-0.01), EDU (-0.01), DVET (-0.02)

The highlighted regions in Figure 3 (right) for ReaLTabFormer exhibit greater similarity to the highlighted regions in the target data (Figure 3, center). This indicates that the synthetic data generated by ReaLTabFormer more effectively preserves the structure and feature correlations of the target data compared to the MST results (Figure 3, left).

Fig. 3. Principal Component Analysis queries

Regression models

To further assess the output of the models, a linear regression model is fit to predict PINCP_DECILE based on EDU for the adult population (AGEP > 15). Figure 4 shows the results for the full adult population (left), and for the two smallest population groups, respectively the American Indian, Alaskan Native and Native Hawaiians (AIANNH) women (middle) and AIANNH men (right).[4]

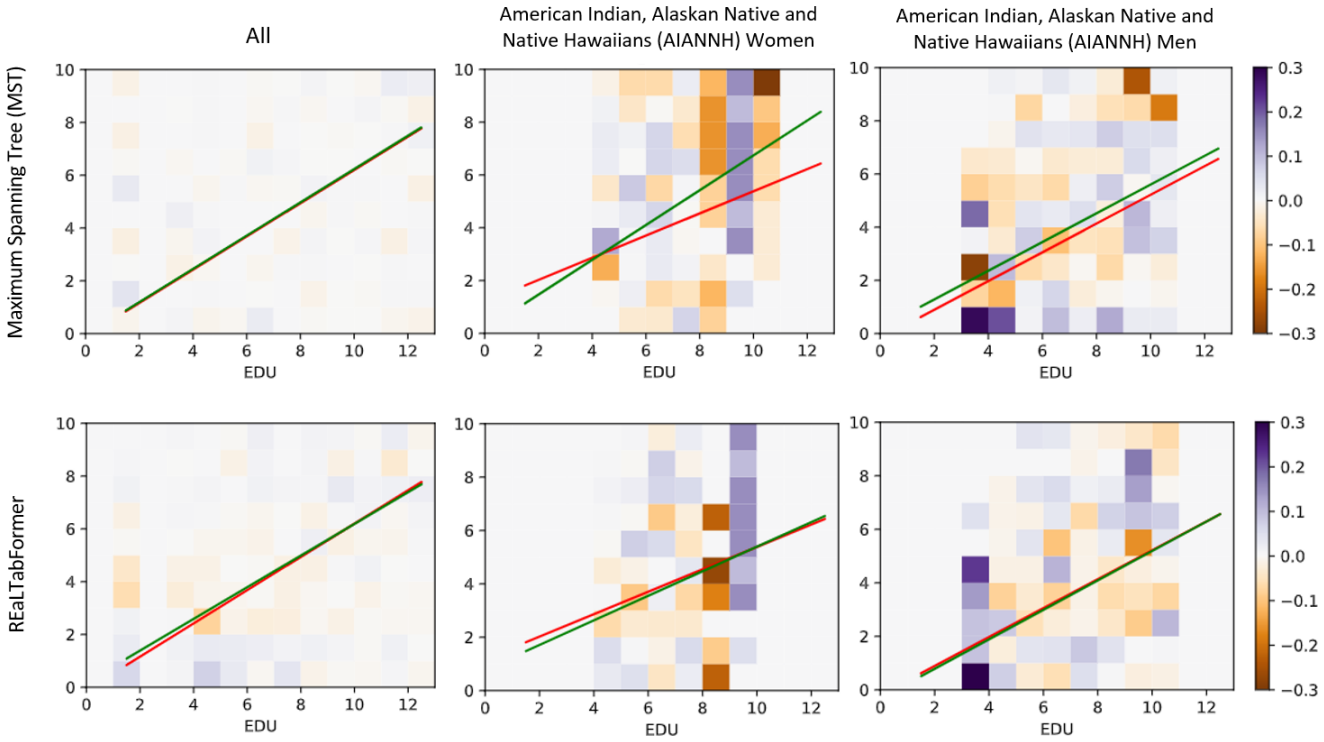

Fig. 4. Linear regression models on synthetic data generated by MST (top) and REaLTabFormer (bottom)

4 Synthetic data for an imaginary country, and disclosure risk assessment

In this second application of GPT and Seq2Seq-based generative models[5], we aim to create a "census dataset" comprising 10 million individuals, designed for simulation and training purposes. Our objective is to generate a realistic statistical portrayal of an imaginary middle-income country's entire population. The dataset includes

[4] The number of observations in the groups, respectively in the target data and in the MST and REaLTabFormer deidentified datasets, are 23,006 / 23,010 / 23,333 (All), 376 / 367 / 404 (AIANNH men), 395 / 429 / 433 (AIANNH women).

[5] For this application, we used GPT and Seq2Seq models, which are fundamental components of the REaLTabFormer model. REaLTabFormer was created after we produced the synthetic data for the imaginary country, by packaging these components.

various household and individual-level variables typically obtained from population censuses and household surveys. These variables include demographics, education, labour, housing, household expenditure, assets ownership, child anthropometry, and more. The intended applications of this data encompass training in sampling techniques. Therefore, the census data must incorporate a realistic enumeration area variable, which serves as a primary sample unit in household survey design. Here, we provide an overview of the process, while the detailed description can be found in the technical document accompanying the data publication.

The generation of the full-population dataset followed a 4-step modelling approach.[6] Initially, we trained a parent model focusing on household composition. Subsequently, a series of seq2seq models were trained to capture the hierarchical structure of the data. Specifically, the following seq2seq models were employed: (i) mapping household composition to household variables, (ii) utilizing combined household composition and household variables as input to generate head of household variables, and (iii) incorporating all previous variables to generate the remaining household members' attributes. Additionally, we created the enumeration area and geographic variables utilizing data from multiple sources. To complete the dataset, anthropometric and consumption variables and a few others were added through cross-dataset imputation.

We utilized the generative models in conjunction with data from the IPUMS International program to generate the core dataset. For training the model, we selected 43 IPUMS census datasets from 30 different countries. These datasets provided a total of 236 million observations, from which we randomly extracted a sample of 6.4 million observations. After the generative models were trained, we generated a raw dataset consisting of 5 million households, equivalent to approximately 20 million individuals. As none of the selected datasets encompassed all the variables of interest, this training dataset contained missing values. However, we configured the models to prevent the generation of missing values during the process. After subjecting the raw synthetic data to validators, the dataset was reduced to roughly 4.4 million households, corresponding to around 17.7 million individuals. This reduction implies an efficiency rate of approximately 88% in the generation process. From this pool, we extracted the necessary number of observations and allocated them to geographic regions based on a target distribution.

The IPUMS datasets do not provide all variables of interest. To add data on child anthropometrics (height and weight of children aged 0 to 5 years), main source of drinking water, type of toilet used by the household, ownership of a bicycle and motorcycle, and bank account ownership, we incorporated data from 15 datasets published by the Demographic and Health Survey (DHS) program. These selected DHS datasets were recoded to ensure a set of consistent overlapping variables that could be utilized as predictors. Subsequently, we employed a random forests regression model for the imputation of these variables.

To incorporate variables related to household expenditures, both total and categorized by product or service, we integrated 58 datasets from the World Bank Global Consumption Database (GCD). To ensure consistency, we converted the provided values, which were in local currency and for different survey years, into 2020 $ purchasing power parities (PPP). For this purpose, we utilized consumption growth data and PPP conversion factors from the World Bank's World Development Indicators database. Additionally, we scaled the expenditure values proportionally to establish an annual mean per capita expenditure of $3,500 PPP for each survey. The resulting data file presents the consumption profiles of 1,207,951 households, displaying a quasi-lognormal distribution of per capita expenditure.

The imputation process for the consumption variables was divided into two tasks. Firstly, variables available in the GCD datasets were recoded as relevant, ensuring a consistent set of variables shared with the core synthetic dataset. Next, using these variables as predictors, we employed a random forest regression model to impute the

[6] The full population dataset and data for a sample of 8,000 households are available as open data from the World Bank Microdata Library, where a more detailed technical description of the synthetic data generation process is provided. The data and data dictionary are available in English (DOIs: https://doi.org/10.48529/78M1-AE09 and https://doi.org/10.48529/MC1F-QH23) and in French (DOIs: https://doi.org/10.48529/X5BG-SD13 and https://doi.org/10.48529/42QP-VB86). For information on the process, see also the Github repository at https://github.com/avsolatorio/synthetic-pop

total household expenditure for each synthetic household record. Subsequently, a transformer-based model was trained to generate the proportions of each consumption category for every household.

Creating a realistic enumeration area variable proved to be a more complex task. We opted for a hierarchical probabilistic generative model, informed by the empirical data at hand, to generate enumeration areas and allocate households accordingly. Utilizing variables shared between the IPUMS and DHS datasets, we applied K-Means clustering to the enumeration area data from 15 DHS surveys, aiming for high granularity with K=50 clusters. This analysis effectively captured the clustering effect within enumeration areas, a crucial aspect we sought to replicate in the synthetic data.

In order to inform the probabilistic model on how to distribute households within enumeration areas, we analysed the empirical characteristics of DHS enumeration areas. Specifically, we examined the frequency distribution of households within these areas, observing a distribution that displayed a truncation effect in the higher tail. Of particular interest to us was the distribution of the number of distinct clusters of households belonging to the same enumeration area. While not a perfect fit, we found that this distribution could be approximated by a Poisson distribution, which we employed in our probabilistic model. Additionally, we utilized a negative binomial distribution to model the number of households per enumeration area.

This same process was applied to both urban and rural areas, albeit with different sets of parameters. For urban areas, the mean number of households per enumeration area was set at 500, while for rural areas, a mean of 350 households was used. In both cases, a standard deviation of 100 was implemented. Furthermore, the clusters were parameterized based on their overall similarity, with variations introduced between urban and rural areas. It was assumed that urban areas would exhibit greater diversity compared to their rural counterparts.

Finally, these enumeration areas were employed to distribute households according to geography. To achieve the desired population distribution by region and urban/rural classification, we utilized a target table. Iteratively sampling from the available enumeration areas, we ensured that the expected population specified in the target table was met for each stratum.

The final dataset is a fully synthetic dataset that does not contain any missing values. It was generated through a process that involved sampling, recoding, and integrating training datasets that had previously undergone anonymization procedures. Precautions were also taken to prevent overfitting and data copying when applying the models. As a result, the synthetic dataset is free from any potential risks associated with identity or attribute disclosure. Consequently, it was released as open data.

Assessment of individual disclosure risk measures, and an alternative approach

We utilized this synthetic population dataset to evaluate the effectiveness of the statistical disclosure risk estimation method employed in sdcMicro and μ-Argus. This method is thoroughly described in the μ-Argus manual (Statistics Netherlands, 2014), as well as in Benedetti and Franconi (1998) and Franconi and Polettini (2004).

The individual risk of disclosure refers to the maximum probability that an observation can be correctly re-identified. The assumption is made that an intruder possesses an error-free dataset encompassing a direct identifier of respondents, covering the entire population. The disseminated data, also assumed to be error-free, represents a sample from this same population. Both the intruder's dataset and the sample dataset share a common set of indirect identifiers or key variables. The combination of these key variables yields a key *k* for each observation, which can be used to match the two data files. The frequency of combination *k* in the full population data is denoted as F_k, whereas the frequency of combination *k* in the sample data is denoted as f_k. If F_k is known, the probability of correct re-identification is $1/ F_k$. However, in practical scenarios, F_k is unknown and must be estimated based on f_k, the distribution of key frequencies in the sample. Benedetti and Franconi (1998) addressed the uncertainty surrounding F_k using a Bayesian approach and a superpopulation framework. The method was tested by Benedetti, Capobianchi and Franconi and (2003) using data from the Italian Household Consumption Survey (HCS) in 1997. Additionally, Seri, Di Consiglio, and Franconi (2003) conducted simulations with a dataset consisting of 15 million observations extracted from the Italian 1991

population census dataset. These assessments concluded that the method provides satisfactory approximations of the true risk, albeit with potential over- or under-estimations under varying circumstances.

The availability of openly accessible synthetic population datasets opens up possibilities for conducting more diverse simulations, utilizing a wider range of key variables and samples of varying sizes and designs. In this report, we provide a brief assessment of the individual risk measures implemented in μ-Argus and sdcMicro, using our synthetic country population dataset. The assessment assumes six key variables: *geo1* (geography level 1, 10 categories), *geo2* (geography level 2, 61 categories), *urbrur* (urban/rural, 2 categories), *sex* (2 categories), *age_fix* (age in completed years, 101 categories), and *marstat* (marital status, 4 categories). In total, there are 985,760 possible combinations, although many of them are either unrealistic or not present in the synthetic dataset. We drew samples of different sizes, ranging from 1,000 to 100,000 (only three samples are shown here), and estimated the individual disclosure risks using sdcMicro. We then compared these estimates with the true risks, which are known when the full population dataset is available. The sdcMicro approach leverages the frequencies found in the sample and the sample weights to estimate the risk. As the proportion of existing keys found in a sample increases with the sample size (Figure 5, left), the reliability of the modelled risk estimates also increases with the sample's diversity and size (Figure 5, centre and right).

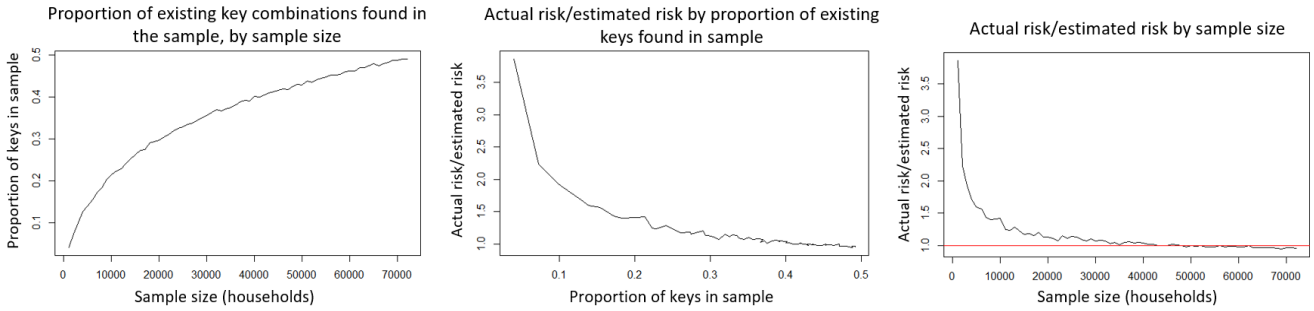

Fig. 5. Impact of sample size and coverage of existing keys on the reliability of the reidentification risk estimates

In our simulation, we observed that the risk measures tend to be underestimated when the sample size is small. However, as the sample size increases, the risk estimates improve and approach the true measure as expected. Figure 6 displays the plots for stratified samples of varying sizes (1,000, 5,000, and 25,000 households) out of a sample frame containing 2.5 million households. The diagonal line represents a match between the true and modelled estimates. Points above the line indicate instances where sdcMicro overestimated the risk, while points below the line represent underestimations of the risk. Larger samples offer a better representation of the key variables' diversity, leading to the estimates converging towards the diagonal line.

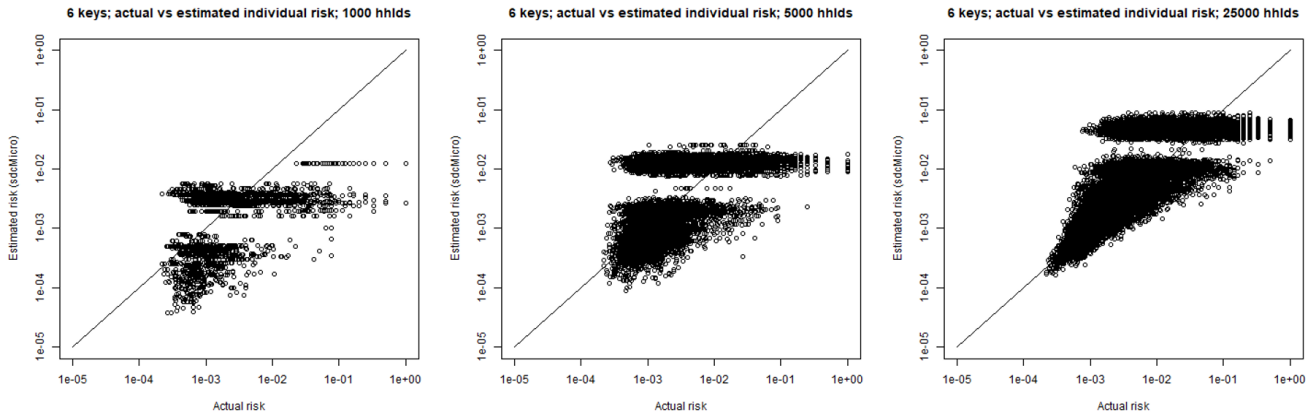

Fig. 6. Actual vs estimated individual risk, stratified samples of 1,000, 5,000 and 10,000 households

The precise value of the risk estimate for each observation is not the primary concern for organizations that release anonymized datasets. What truly matters is whether the risk estimate falls below or above the threshold of "acceptable risk." For observations with risk estimates significantly below the threshold, a small error in the estimate holds no consequence. The observations that require attention are twofold: those with a

reidentification risk above the threshold but not flagged as unsafe, and those without any risk but erroneously identified as risky, leading to unnecessary utility loss in the anonymization process. To evaluate this issue, we calculate both the true and modelled reidentification risk estimates for each observation in the sample dataset. Subsequently, we categorize each observation as "At risk" when the risk value exceeds a predefined threshold (e.g., >= 0.01) and as "Not at risk" otherwise, using both the true and modelled risk estimates. This classification allows us to generate a confusion matrix for each version of the sample dataset, where false positives and false negatives indicate prediction errors. False positives correspond to safe records that are mistakenly flagged as at risk, while false negatives represent observations at risk but wrongly reported as safe by the modelled estimate.

		Based on the model-estimated risk	
		At risk	*Not at risk*
Based on the true risk	*At risk*	True positive (TP)	False negative (FN)
	Not at risk	False positive (FP)	True negative (TN)

The following summary measures can then be derived from the confusion matrix:[7]

- **Accuracy:** ACC = TP+TN / TP+TN+FN+FP
- **Error rate:** ERR = FP+FN / TP+TN+FN+FP
- **False positive rate:** FPR = FP / TN+FP
- **Sensitivity**, also called **recall** or **true positive rate:** SN = TP / TP+FN
- **Precision**, also called **positive predictive value:** PREC = TP / TP+FP
- **Specificity**, also referred to as the **true negative rate:** SP = TN / TN+FP

All indicators in this study have values ranging from 0 to 1, with 1 representing the ideal value for accuracy, sensitivity, precision, and specificity. Conversely, for the error rate and false positive rate, the ideal value is 0. To assess the level and variation of these measures, we generated 1,000 random samples of 10,000 households and produced confusion matrices using sdcMicro to obtain model-based estimates. A risk threshold of 0.01 was set, meaning that an observation is considered 'At risk' if it has a 1% or greater chance of being re-identified. The first row of Table 2 presents the mean values of the summary measures.

Table 2. Summary measures (means) of indicators derived from confusion matrices

Approach	Accuracy	Error rate	False positive rate	Sensitivity	Precision	Specificity
sdcMicro	0.877	0.123	0.119	0.838	0.451	0.881
Synthetic	**0.965**	**0.035**	**0.108**	**0.973**	**0.987**	**0.892**

An alternative superpopulation approach

One weakness of the individual risk measurement approach utilized in sdcMicro and μ-Argus is its inability to effectively model a superpopulation that encompasses the diversity of keys when dealing with small sample sizes or fractions. To address this limitation, we propose an alternative approach that involves using REaLTabFormer to generate a superpopulation that more accurately estimates the true frequencies of the keys, leveraging the information available in the sample. We create a synthetic superpopulation consisting solely of the key variables, with a number of observations equivalent to the size of the extrapolated sample population. This method generates a more diverse superpopulation compared to the Bayesian approach, and the risk estimates for “high risk” keys (F_k <= 3) are better captured as evidenced by the relative risk metric (Figure 7). Another notable property of the synthetic superpopulation method also shown in the figure is that it captures the distribution of the risk better than the Bayesian method used in the sdcMicro.

We generated confusion matrices for 1,000 random samples of 10,000 households and show the mean values of the summary measures in the second row of Table 2. All values obtained using our new approach surpass those

[7] See https://classeval.wordpress.com/introduction/basic-evaluation-measures/

obtained with the sdcMicro measures.[8] More importantly, the sensitivity metric which informs how well risky observations are correctly identified has significantly improved using the synthetic superpopulation method. We should note, however, that this improvement comes at the expense of substantial computational requirements.

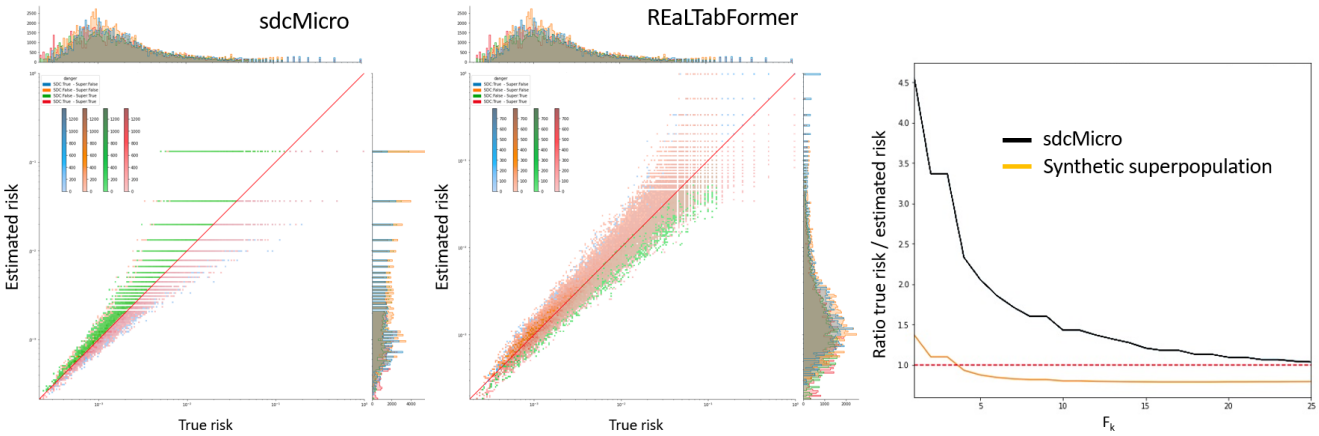

Fig. 7. Estimated and actual risk values using the sdcMicro (left) and REaLTabFormer (center) superpopulation models, and relative risk trends for both estimation strategies (right). Risk estimates based on six key variables and a sample of 200,000 households.

5 Conclusion and further work

In order for synthetic datasets to gain acceptance as viable substitutes for real microdata, data producers require stronger evidence of their safety, while the research community demands further proof of their utility. This paper successfully demonstrates the capability of the REaLTabFormer model in generating synthetic non-relational and relational data that is both safe and realistic. It also showcases the model's ability to produce improved estimates of individual statistical disclosure risk in microdata.

Moving forward, additional research and development efforts should be pursued in several areas. Firstly, the evaluation of statistical disclosure risk measures and disclosure limitation methods can be expanded to include evaluating the mosaic effect, analysing the effectiveness of reverse-engineering of anonymization procedures, examining the influence of sample design on disclosure risk, and exploring the impact of data inaccuracies on risk assessments (considering scenarios beyond worst-case assumptions). The existence of openly accessible synthetic datasets presents valuable opportunities to conduct simulations that can significantly contribute to advancing this research.

Secondly, it is worth considering the incorporation of differentially private mechanisms into transformer-based models. This would provide a more robust and formally guaranteed approach to ensuring the safety and integrity of the synthetic data.

Thirdly, there is a need for a comprehensive evaluation and enhancement of REaLTabFormer and other synthetic data production models. Our primary focus will be to leverage REaLTabFormer to generate sample datasets with high utility, utilizing actual country survey datasets as input. This process will combine techniques of synthetic data modelling and sample calibration. The entire process will be meticulously documented and made replicable, to serve as a potential model for organizations interested in generating synthetic data for the purpose of creating public-use microdata files.

[8] The key variables and base population used here are the same, but the 1,000 samples are drawn independently. The summary indicators based on 1,000 draws are however expected to provide central measures that can be compared.

References

1. Benedetti, R. and Franconi, L. (1998). Statistical and technological solutions for controlled data dissemination, Pre-proceedings of New Techniques and Technologies for Statistics, 1, 225-232.
2. Benedetti, R., Capobianchi, A., and Franconi, L. (2003). Individual Risk of Disclosure Using Sampling Design Information. *Contributi Istat* n.14-03.
3. Franconi, L. and Polettini, S. (2004). Individual risk estimation in-Argus: a review, in Domingo-Ferrer, J. and Torra, V. (Eds.) Privacy in Statistical Databases, Berlin: Springer-Verlag, 262-272.
4. McKenna, R., Miklau, G., and Sheldon, D. (2021). Winning the NIST Contest: A scalable and general approach to differentially private synthetic data https://arxiv.org/abs/2108.04978
5. Padhi, I., Schiff, Y., Melnyk, I., Rigotti, M., Mroueh, Y., Dognin, P., Ross, J., Nair, R., and Altman, E. (2020). Tabular Transformers for Modeling Multivariate Time Series. https://arxiv.org/abs/2011.01843
6. Park,N., Mohammadi,M., Gorde,K., Jajodia,S., Park, H., and Kim,Y. (2018). Data synthesis based on generative adversarial networks. arXivpreprintarXiv:1806.03384.
7. Patki, N., Wedge, R., and Veeramachaneni, K. (2016). The Synthetic Data Vault. In 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA), pp. 399410, 2016. doi:10.1109/DSAA.2016.49.
8. Seri, G., Di Consiglio, L., and Franconi, L. (2003). Individual Risk Assessment of Social Microdata.
9. Solatorio, A. and Dupriez, O. (2023). REaLTabFormer: Generating Realistic Relational and Tabular Data using Transformers.
10. Statistics Netherlands. (2014). μ-Argus 5.1 User's Manual.
11. Task, C., Bhagat, K., and Howarth, G.S. (2023). SDNist v2: Deidentified Data Report Tool, National Institute of Standards and Technology, https://doi.org/10.18434/mds2-2943
12. United Nations Economic Commission for Europe (UNECE). (2023). Synthetic Data for Official Statistics - A Starter Guide.
13. Yun, C., Bhojanapalli, S., Rawat, A.S., Reddi, S.J., and Kumar, S. (2019). Are transformers universal approximators of sequence-to-sequence functions? arXiv preprint arXiv:1912.10077, 2019.

GitHub and code repositories

1. NIST Diverse Communities Data Excerpts Github repository: https://github.com/usnistgov/SDNist/tree/main/nist%20diverse%20communities%20data%20excerpts
2. SDNist standardized report: Task C., Bhagat K., and Howarth G.S. (2023), SDNist v2: Deidentified Data Report Tool, National Institute of Standards and Technology, https://doi.org/10.18434/mds2-2943
3. Maximum Spanning Tree (MST) model: https://docs.smartnoise.org/synth/synthesizers/mst.html
4. REaLTabFormer model, Github repository: https://github.com/worldbank/REaLTabFormer
5. NIST experiments, Github repository: https://github.com/avsolatorio/REaLTabFormer-NIST
6. Imaginary country synthetic data Github repository: https://github.com/avsolatorio/synthetic-pop