AI-Based Privacy Preserving Census(like) Data Publication

Johannes Gussenbauer (Statistics Austria) Alexander Kowarik (Statistics Austria)
Klaudius Kalcher (Mostly AI) Michael Platzer (Mostly AI)

Abstract

Sharing data with the public is one of the core principles of Statistics Austria. The access and use of micro data is, however, restricted due to privacy protection principles and regulations, as it contains sensitive information on a highly granular level. Synthetic micro data, in contrast, might be safely shared with the public and scientific community to enable innovative research and support policy making. In this study, we created synthetic micro data of the Austrian population with the aim of sharing the synthetic data with the public and the scientific community. The synthetic data are based on the Rich Frame, a pseudonymized micro data set containing every person registered in the Austrian housing and living register as main residence within private (non-institutional) households. It includes both personal and household specific attributes, as well as income data. The data generation method uses generative deep neural networks to produce a synthetic data set. The utility of the synthetic data set is measured by comparing the synthetic data with the original data on certain marginal distributions. Additionally, the residual privacy risk of the synthetic data set is assessed.

Contents

1 Introduction

Access to micro data greatly supports data driven research, policy making and enables faster scientific progress. Despite these benefits tha acces to micro data for researchers is in general quite limited and often challenging, see National Academies of Sciences and Medicine (2016) and Lugg-Widger et al. (2018). The process to acquire micro survey data from Eurostat can take up from 8 to 14 weeks from data request to data delivery, see https://ec.europa.eu/eurostat/web/microdata. Even if access to micro data is granted it is quite common that data is partially suppressed or censored due to privacy protection laws. Generating synthetic data with very low re-identification risk can help with this issue. The results presented in this paper are part of the FFG-Project “AI-Based Privacy-Preserving Big Data Sharing for Market Research.” Collaborators of this project are the WU (Vienna University of Economics and Business), MOSTLY AI, George Labs by Erste Group Bank AG and Statistics Austria. The aim of this project is to train generative deep learning models on micro data, specifically sequential or longitudinal micro data, to generate synthetic data. A key aspect of the data synthesis are the privacy safeguards maintained during model training as well as privacy evaluations of the resulting dataset. The use case for Statistics Austria is the synthesis of the so called Rich Frame which is a pseudonymized micro data set containing every person registered in the Austrian housing and living register as main residence within private (non-institutional) households, including personal and household specific attributes, as well as income data. The rest of this paper is structured as follows. Section 2 gives a short overview of the methodology of the data synthetization, section 3 discusses the utility of the synthetic micro data by comparing it with the original data input and section 4 assesses the privacy risk of the synthetic data set. This paper is supported by the “ICT of the Future” funding programme of the Austrian Federal Ministry for Climate Action, Environment, Energy, Mobility, Innovation and Technology.

2 Methodology

The synthetic data have been generated using the MOSTLY AI synthetic data platform. On the platform, generative deep learning networks are built and trained using the input data. After this training is complete, the trained model is used to create new data rows that are all independent from the original data as well as from each other. Either one table of independent rows or two tables with a foreign key relationship can be used as input - in the latter case, relationships between the linked rows of the two tables as well as between rows of the second table that are linked to the same row in the primary table are also modeled. Typically, this is used to model sequential data and allows to maintain statistical links between attributes of events in a sequence of events, but the same modeling principle can also be applied to a set of members of a household.

Thus, the dataset at hand was represented as two tables. The first contained the household identifier and the features relating to the household as a whole, specifically: - geographic variables: NUTS2, municipality and urbanity - variables on the type housing: apartment or house, building period of house and tenancy

The second table contained the individual-level data features, and also included the household identifier to link the individuals to the household and to each other. This second table included the following features: - basic socio-demographic variables: sex, age, education, employment status, citizenship and country of birth (ISO 3166-1 numeric) - income variables: yearly income from employment and self employment.

The full dataset corresponds to the housing and living register from the first quarter of 2021 containing 8845691 people living in 4015907 households.

Internally, the MOSTLY AI platform creates one network for each table to be synthesized, each being built specifically for that table. This process is fully automatic - it uses as building blocks different network layers depending on the type of the columns (in principle, those are numeric, categorical, datetime, geoposition and text columns; though only numeric and categorical columns were present in the dataset synthesized for this paper), and can be tuned using 3 hyperparameters: general model size, complexity of the link between the two tables, and complexity of the sequential structure of the linked table. Once the model is trained, the original data are not retained or used for generation in any way. Instead, synthetic data are created purely from a random seed, using only the trained network layers.

3 Utility

From the trained model we generated 4015907 households containing 8942421 number of people, thus the number of people in the synthetic data set is 1.0935% higher than in the input data set. We will showcase the utility of the synthetic data by looking at the consistency between synthetic variables and comparing the marginal distribution between synthetic and input data for selected variables. Especially important is the comparison of marginal distributions from derived variables which depend on the composition of the household members.

3.1 Consistency of output

When generating synthetic data it is not only important that the marginal distributions in the input data set are well reconstructed but also that the data records itself are consistent. We checked this requirement on a selected number of criteria. First we looked at the oldest persons inside each household. Since the a household cannot consist of only minors, e.g. age smaller than 15 years, there should be no such cases in the synthetic data. We observe however 535 such households in the data. Next we checked some variables which directly depend on each other. For instance the variables NUTS2 and municipality, which is a 5 digits code starting with the NUTS2 region. The synthetic data thus should generate only variables combinations of NUTS2 regions the municipalities which are actually possible. In the synthetic data we find 352 households which do not follow this condition. Another example for direct dependency between variables are the variables for citizenship and country of birth which were generated in 3 different granularities. The synthetic data contains 2074 and 245 persons where these three different granularities contain impossible variable combinations for citizenship and country of birth respectively. It should be noted that using the same variables with different levels of granularities helps train the model on one hand and if inconsistencies occur they can easily be fixed since variables with less values can easily created from the variable with the highest granularity.

Overall the synthetic data, when observing these specific variables, seems to be consistent for almost all records.

3.2 Marginal distribution of selected variables

Comparing the marginal distribution between the synthetic data set (`Synthetic`) and the input data set (`Target`) of selected variables we see that they agree quite well with each other. Figure 1 shows the distributions of age by NUTS2 regions and degree of urbanization. Although the age distributions do differ to some extend it is clear to see that the differences in shape across NUTS2 region and urbanisation are preserved in the synthetic data. For instance in the NUTS2 region AT13 (Vienna) there is a higher density of younger people compared to rural areas in NUTS regions AT21 (Carinthia), AT22 (Styria) or AT31 (Upper Austria).

Similar conclusion can be drawn from Figure 2 where the number of people by type of employment, sex and urbanisation are compared. We again see some differences between the distributions but important structures, like a rising number of unemployed persons in cities, are captured in the synthetic data set as well.

Figure 3 shows the income distribution by education and sex between input (dashed) and synthetic data (solid). The distribution don't agree on all parts, for instance the higher peak in income for the category `<15 years or education unknown`, but again we see that characteristic shapes between educational levels for each sex are well preserved overall.

3.3 Marginal distribution of derived variables

An important aspect for the synthesis with the data at hand is how well the structure of the households, e.g. the composition of age, education, employment, income, ect... , are preserved in the synthetic data. Since the household structure can be quite complex and difficult to model it is important to compare the marginal distribution on derived variables which depend on the composition of the household members. For this purpose we are looking at the distribution of personal attributes of household members with the highest income inside each household. Figure 4 shows, similar to figure 2, the distribution of people with the highest income per household by type of employment, sex and urbanisation. When multiple people have the highest income then one is chosen randomly. It is easy to see that the synthetic data reproduces the input data quite well, as seen for example in the gender ratio of highest earners per household. The higher percentage of males is reproduced in the synthetic data, as well as the effect that this imbalance is more pronounced in rural areas compared to cities. Important to note is that this distribution is quite different from figure 2, thus indicating that the composition of household members is considered during data synthesis.

An important indicator when talking about yearly income and households is the so called equivalized household income which is used to determine the poverty threshold in the EU-SILC survey. A definition of the equivalized household income can be found here: https://ec.europa.eu/eurostat/statistics-explained/index.php?title

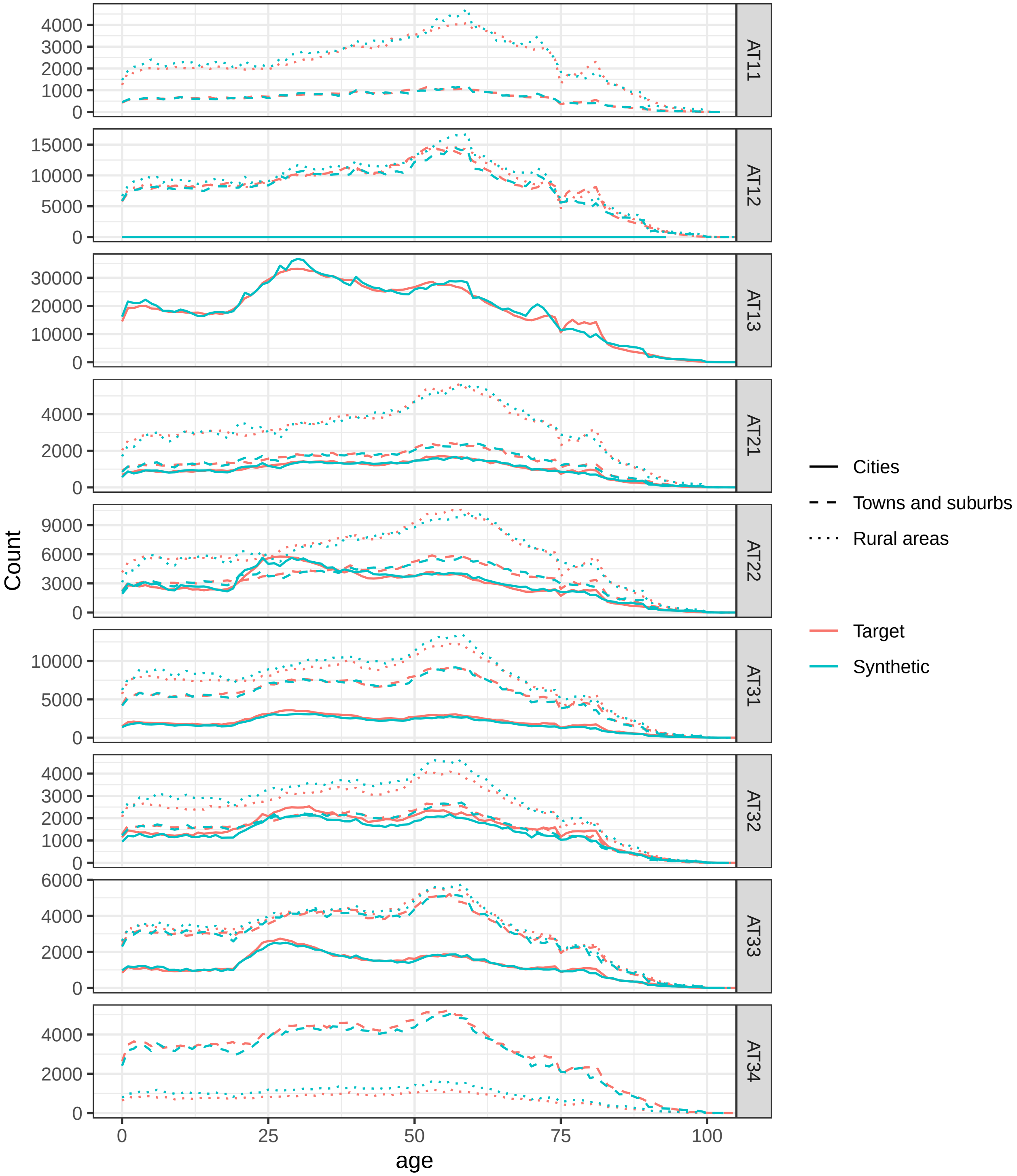

Figure 1: Number of people by age, NUTS2 regions and degree of urbanisation.

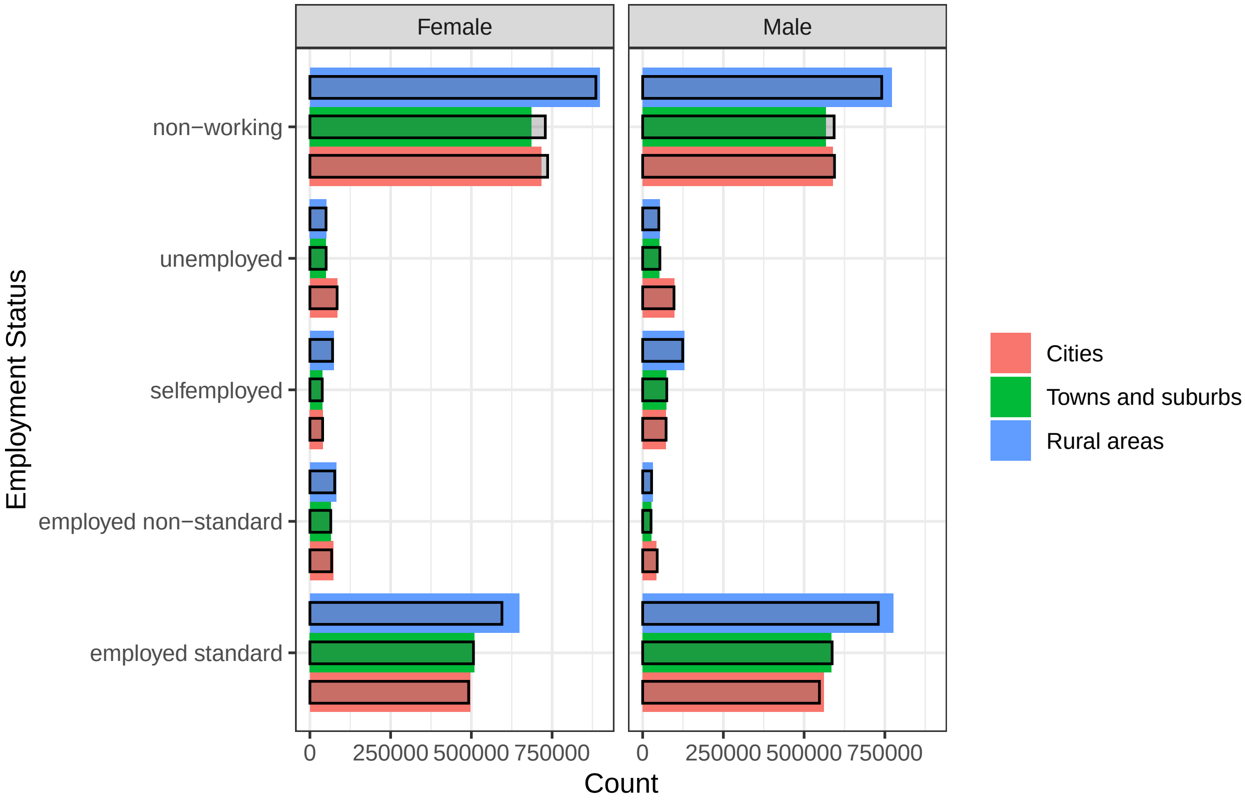

Figure 2: Number of people by type of employment, sex and urbanisation. The grey bars with black borders indicator the number of people observed in the input data.

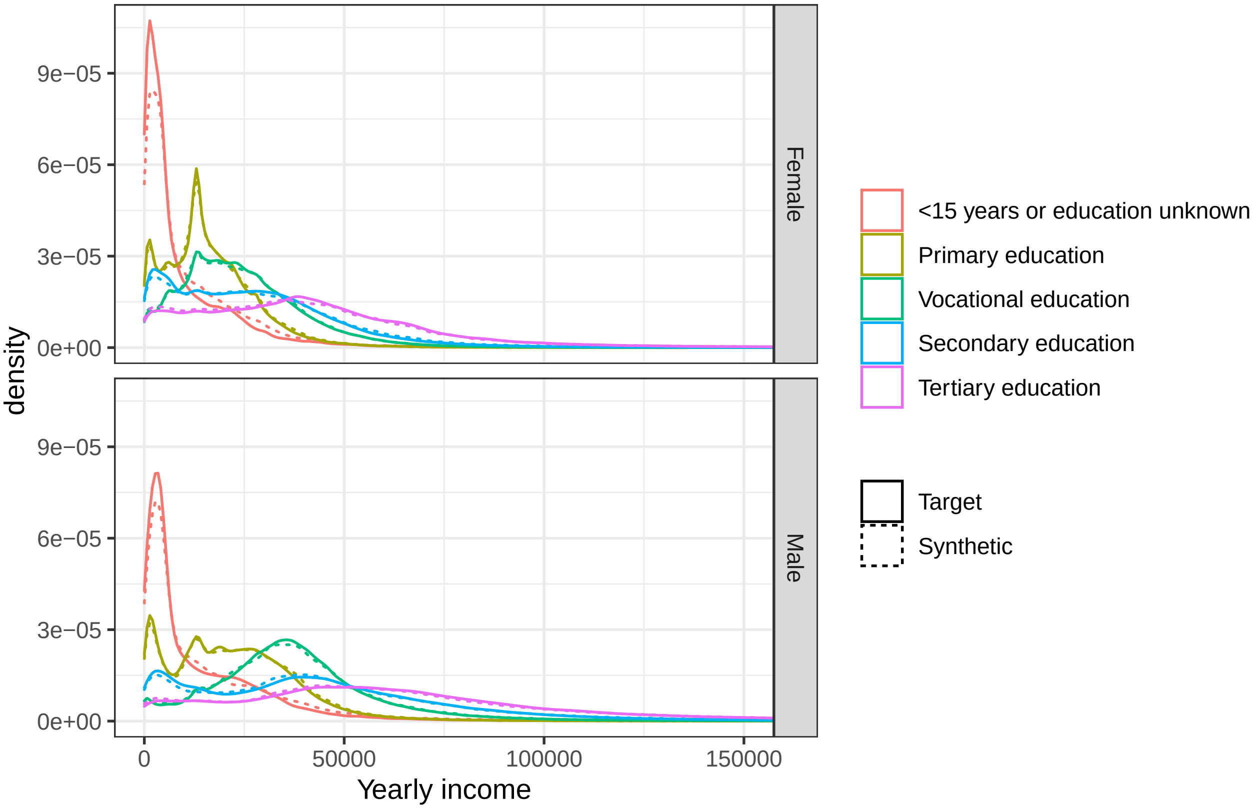

Figure 3: Distribution of yearly income by education and sex between input (dashed) and synthetic data (solid). People without yearly income are excluded.

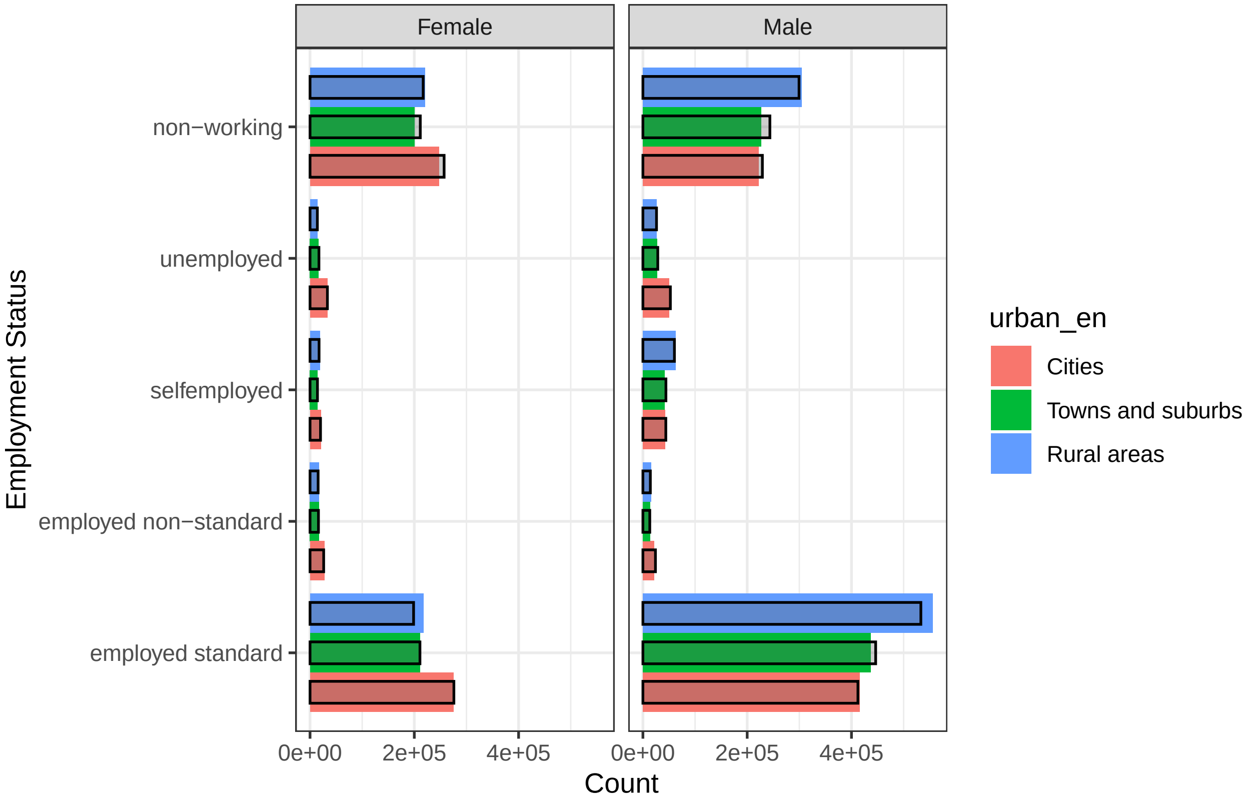

Figure 4: Distribution highest income earners per household by type of employment, sex and urbanisation. The grey bars with black borders indicator the number of people observed in the input data.

=Glossary:Equivalised_income. Figure 5 shows the distribution of equivalised household income by degree of urbanisation and household size (colours) with household size top-coded at 5. Again, we see that the distribution of the variable seem to be quite well replicated in the synthetic data. Characteristic peaks and specific structures - like lower incomes for larger households but higher larger incomes for single or two-person households in cities compared to rural areas - are preserved.

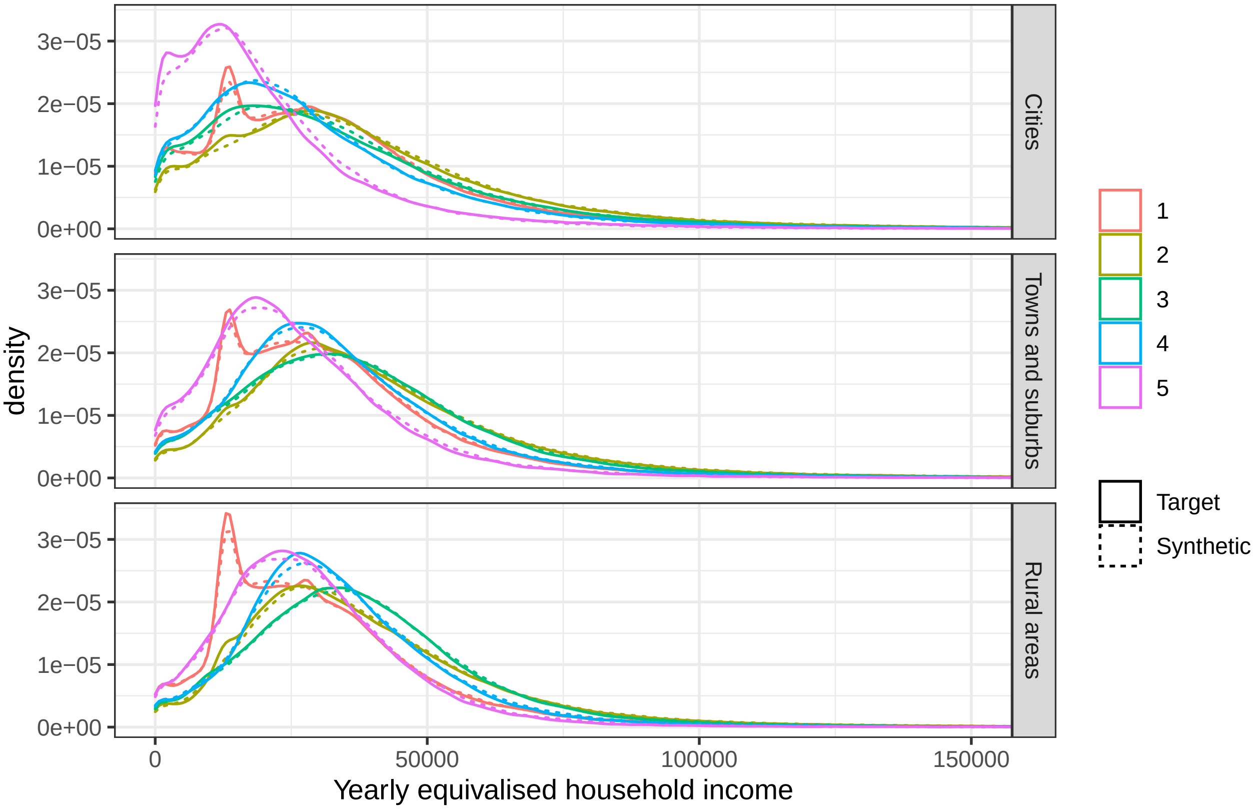

Figure 5: Distribution of equivalised household income per household size, top coded by 5 and urbanisation.

4 Privacy Risk

To assess the privacy risk we follow the suggestions from Platzer and Reutterer (2021) and look at the Identical Match Share (IMS), Distance to Closest Record (DCR) and Nearest Neighbour Distance Ratio (NNDR).

- **IMS**: Calculating the share of identical records between the synthetic data and input data and comparing this value with the share of identical records inside the input data.
- **DCR**: Calculating, for each synthetic data record, the distance to the closest record in the input data and comparing this value with the distribution of the original datapoints distances to their respective nearest neighbor within the input data.
- **NNDR**: Calculating, for each synthetic data record, the ratio between the distances to the nearest and second-nearest neighbors in the input data and comparing this value with the distance ratio calculated from within the input data.

To calcualte the indicators a holout data set is drawn before the training of the model and the indicators are computed between the synthetic data as well as the holdout data and input data. The results are shown in the Table 1.

In terms of these risk measures that households in the synthetic data are not closer to the input data than

Table 1: Privacy measures of synthetic and holdout data set.

	IMS	DCR 5th percentile	NNDR 5th percentile
Holdout Data	0.01%	0.02	0.39
Synthetic Data	0.01%	0.02	0.41

what can be expected when analysing the household in the input data.

5 Conclusion

Micro data is particularly important for data driven research and policy making. However due to privacy regulations it is in general very difficult or not possible give access to such data. Even if access can be granted to researches the procedures to acquire the data can be quite long and the data itself might be very censored. One way to overcome these issues can be the use of synthetic micro data. In this work we presented the results from synthetically generating the Austrian population using deep Generative Adversarial Networks (GANs). The core challenge for this problem is that the data synthesis should not only generate synthetic people but also group them together realistically into households. Comparing the synthetic with the input data it is clearly visible that the distributions for selected variables are mostly captured well. The synthetic data performs especially well on the distribution of more complex derived variables. Since the confidentiality is respected during model training the synthetic data seems to contain acceptable privacy risks. A downside of the presented method is that some distributions do not match too well like the distribution on age or employment status, where other existing software, like Templ et al. (2017), can achieve a better fit. We were also not able to improve on this for our output. On one hand this is because the model is more a black-box and it is difficult to tune the parameters or how to adjust the input to achieve better results. On the other hand a single run for the whole data generation took multiple days, so it is not feasible to run the data synthesis with a large number of different parameters. Another issue, although more minor, is that some of the synthetic data was not consistent, meaning that certain combinations of different variables which should not appear in reality did occur in the synthetic data. For our analysis we checked for municipalities which were in the wrong NUTS2 region, variables for country of birth and citizenship with varying granularity and households which contained only minors. All these issues did however only appear in a very small portion of the synthetic data.

To conclude the synthetic data from the generative deep neural network does replicate structure of the input data fairly well, especially for complex dependency between variables. The results presented here do however only paint a small picture of the overall performance of the data synthesis and a deeper analysis both in terms of utility and privacy risk is needed to draw more final conclusions.

References

Lugg-Widger, Fiona, Lianna Angel, Rebecca Cannings-John, Kerenza Hood, Kathryn Hughes, Gwenllian Moody, and Michael Robling. 2018. "Challenges in Accessing Routinely Collected Data from Multiple Providers in the UK for Primary Studies: Managing the Morass." *International Journal of Population Data Science* 3 (September). https://doi.org/10.23889/ijpds.v3i3.432.

National Academies of Sciences, Engineering, and Medicine. 2016. *Principles and Obstacles for Sharing Data from Environmental Health Research: Workshop Summary.* Edited by Robert Pool and Erin Rusch. Washington, DC: The National Academies Press. https://doi.org/10.17226/21703.

Platzer, Michael, and Thomas Reutterer. 2021. "Holdout-Based Fidelity and Privacy Assessment of Mixed-Type Synthetic Data." https://arxiv.org/abs/2104.00635.

Templ, Matthias, Bernhard Meindl, Alexander Kowarik, and Olivier Dupriez. 2017. "Simulation of Synthetic Complex Data: The R Package simPop." *Journal of Statistical Software* 79 (10): 1–38. https://doi.org/10.18637/jss.v079.i10.