h_da
HOCHSCHULE DARMSTADT
UNIVERSITY OF APPLIED SCIENCES

The Potential of Synthetic Data for Georeferenced Official Microdata

Hendrik Gressmann

Supervisor: Prof. Dr. Antje Jahn, Prof. Dr. Timo Schürg

The conflict between data utility and data privacy

In Germany, the Federal Statistical Office is based on §16 BStatG (Bundesstatistikgesetz) legally required to provide their official data sources to independent research institutions for research purposes. Scientific research may profit heavily from official microdata products with precise geospatial attributes. Moreover, such data can be linked to auxiliary data using a matching process that relies on the congruence of their geographic links, which enables precise geospatial analysis in various scientific fields.

Official microdata products comprise, among data from other sources, also private data of survey participants. Thus, data privacy regulations require sensitive attributes, like geospatial information on the participant's housing locations, to be encrypted, aggregated, removed, or anonymized using various relocation and perturbation processes. This severely limits the value of these microdata products for research using georeferenced analysis.

Basic Approach: Data Synthesis

In order to solve that conflict, scientists came up with a promising trade-off: They release synthetic versions of the original data. Since data synthesis reduces the data potential for analysis that does not rely on precise geospatial information, [1] suggested publishing two data sets, instead of just one synthetic data set:

- The original data with removed low-level geospatial identifiers.
- A synthetic data set, where all attributes except for the geospatial attributes are synthesized.

The objective of this thesis was to compose an assessment on the potential of synthetic data for German official statistics, under consideration of the extensive legal requirements for official sensitive microdata. We applied five data synthesis methods to the German Census Data from 2011 and evaluated relevant metrics regarding data utility and data privacy.

Synthetic Data

The main purpose of synthetic data is to mask or anonymize sensitive attributes due to data privacy reasons, mostly in person-related data sources, while still providing a data basis for generating sufficiently accurate results from data analysis. Data synthesis is carried out by training mathematical models on the original data and generating synthetic observations from those models, which is performed via sampling with conditioning on the attributes that should not be synthesized.
The quality of synthetic data is assessed under consideration of the following two aspects.

- **Data Utility:** Quantification of the usefulness - the overall distribution of the synthetic data should be close to the distribution of the original data
- **Data Privacy:** Quantification of the risk of a potential attacker being able to learn some information on individuals in the original data, based on the synthetic data and other prior knowledge. In reality, we can only quantify the re-identification risk for a predefined attacking scenario.

For the data synthesis models there exists not only one optimization criterion. Various metrics on utility and privacy evaluation focusing different characteristics need to be considered, which eventually allows a reasonable tradeoff between data utility and data privacy.

Data Synthesis Methods

Copula Synthesis with Frequency Encoding

- *Procedure*: Estimating a multivariate normal distribution function, Sampling the synthetic observations from this function with conditioning on the geospatial attributes
- *Weaknesses*: Parametric assumptions too restrictive, Frequency encoded attributes follow uniform distribution (not normal distribution)

Copula Synthesis with One-Hot Encoding

- *Procedure*: Estimating a multivariate normal distribution function, Sampling the synthetic observations from this function with conditioning on the geospatial attributes.
- *Strengths*: Ability to model the relations between different classes of two categorical attributes
- *Weaknesses*: Parametric assumptions partially too restrictive, Reversing the One-Hot Encoding produces strong biases in the synthetic data

CART Synthesizer

- *Procedure*: For every attribute that should be synthesized: Training a of a CART Model with the respective attribute as target variable and all already synthesized attributes and the geospatial attributes as predictors, Sampling synthetic values from the final leaves of the tree
- *Strengths*: Low computational costs, No parametric assumptions, Complex distributions

Random Forest Synthesizer

- *Procedure*: For every attribute that should be synthesized: Training a of a Random Forest with the respective attribute as target variable and all already synthesized attributes and the geospatial attributes as predictors, Prediction of the synthetic values only using the Out-Of-Bag Trees.
- *Strengths*: No parametric assumptions, Can model even more complex distributions, Produces especially strong results regarding data privacy
- *Weaknesses*: High computational costs, weaker that CART Synthesizers regarding data utility

Geomasking

This is a classical data anonymization procedure and technically does not include data synthesis. In this thesis, Geomasking is used as baseline method and is carried out by spatial perturbation of the spatial identifiers.

Data Privacy Evaluation

The **Population Uniqueness** is the proportion of individuals who's identity can uniquely be re-identified via identical and unique attribute values in the synthetic and original data sets. *Result*: Re-Identification risk negligible for all synthesis methods (except for the Geomasking)

For the **Privacy Attack**, a complex ML model (e.g. Random Forest, Neural Network) is trained on the synthetic data with the geospatial attributes as target variables. Then, the true geospatial attributes are predicted with the original data as test data. A low accuracy implicates a low re-identification risk. *Result*: Re-Identification risk negligible for all Synthesizers

For all Synthesizers (except for the Geomasking), no significant data privacy risks could be detected. Therefore, the comparison and the final assessment are mostly based on the results of the data utility analysis.

Data Utility Evaluation

The Propensity Mean Squared Error (pMSE) is a metric describing the **Global Utility**, thus, how well the distribution of the original data is preserved during data synthesis. This metric is obtained from the propensity scores of a model that tries to discriminate between synthetic and original observations. A low pMSE indicates a low discrimination power, and therefore a high utility.

Copula (Frequency Enc.)	Copula (One-Hot Enc.)	CART	Random Forest	Geomasking
0.24999	0.242	0.198	0.219	0.100

Table 1. pMSE, obtained from comparison of entire data sets

The preservation of the **Univariate and Bivariate Distributions** was analyzed using different test statistics for homogeneity, like the **G-Test Statistic** (similar to the Pearson's χ^2-Test Statistic) and the **VW-Test Statistic** (a Likelihood-Ration Test Statistic). Low test statistic values indicate a high utility.

Outcome Specific Metrics are obtained by the calculation of use cases on both data sets and a comparison of the results. In Figure 2, a visual comparison of the unemployment rates calculated separately for small grid cells of up to 100 km² is to be found, here displayed for a randomly selected county.

All data utility metrics show the markedly best results for the CART Synthesizer. The worst results are obtained for the two Copula Synthesi.

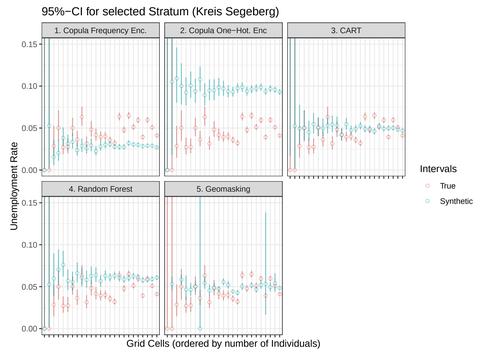

Figure 1. Confidence Intervals of the proportion of a small class in both data sets for every synthesis method

Final Assessment

Providing controlled access to synthetic official microdata for independent research institutions would be conceivable. A condition for enabling such access would be to technically ensure that researchers can only access one of the data sets at a time.

References

[1] T. Schmid T. Koebe, A. Arias-Salazar. Releasing survey microdata with exact cluster locations and additional privacy safeguards, 2023. URL `https://doi.org/10.1057/s41599-023-01694-y`.