# Evaluation of the HLG-MOS Synthetic Data For National Statistical Organizations: A Starter Guide

Hariolf Merkle, Felix Geyer, Michel Reiffert, Reinhard Tent, Steffen Moritz (DESTATIS)

## 1. Did you find the guide useful to generate synthetic data? If so, what aspects were most useful?

In general, we found the guide provided a good first introduction into the topic. Our team consisted of persons with different technical backgrounds, yet everybody found the guide useful. Depending on the person these were the most useful aspects of the guide:

- ➔ General introduction to the topic
- ➔ Application examples from different national statistics offices
- ➔ Suggestions for R / Python packages
- ➔ Introduction into different Use Cases
- ➔ Discussion of privacy risks and measures
- ➔ Discussion of utility considerations
- ➔ Reference to scientific literature to investigate further

## 2. Did you find the guide useful to evaluate the utility and disclosure of the synthetic data? If so, what aspects were most useful?

The guide provided a good starting point to think about utility and disclosure. Yet, we soon realized that utility and disclosure is very dataset and use case specific. Thus, the guide could only provide us with a starting point and general considerations, which needed to be supplemented with our own research. The aspects we found in Chapter 5 “Utility measures for evaluating synthetic data” most useful were:

- ➔ General introduction to the topic
- ➔ Reference to scientific literature to investigate further
- ➔ Suggested R packages facilitate the generation and utility evaluation of synthetic datasets

The guide gives in Chapter 4 “Disclosure considerations for synthetic data” a thorough overview of the most relevant privacy preserving techniques. Each technique is explained well-understandable and several references are provided.
However, Section “Disclosure risk measures” lacks some important information and hence could be more useful in our eyes. The information given to the reader is rather scarce and in contrast to the other sections, there aren’t any references to literature. That makes it cumbersome for interested readers to understand how the disclosure measures presented could be implemented and used in practice. Especially the description of the method of feature mean scaled variance is fairly short and ambiguous. We found it hard to find additional literature on that specific topic, when we conducted some research of our own.

Additionally, we are missing a hint that not each of the listed methods is suited for every kind of synthetic data and an assessment on which method is appropriate for which kind of data. After all, this is meant to be a “starter guide", but it is not possible for a beginner, without prior knowledge to get started with the given information.

Furthermore Figure 4 cannot, from our point of view, be understood the way it is presented, since axis labels are missing and the information you get from the text are not sufficient to compensate. For a starter guide we would appreciate some instructions and maybe some pseudo code, to help the interested reader maneuver through the already difficult process of evaluating the disclosure risk.

## 3. Were there any parts of the guide that were unclear or misleading?

Overall, we had the feeling guide itself was quite straightforward. Most challenging for us in the creation of the synthetic datasets was that the topic itself is quite comprehensive and the starter guide is not able to mention everything. Especially the realistic real-world ACS dataset provided aspects that were not fully covered in the guide. These were the few parts, we found unclear/misleading:

- ➔ The order of the utility measures in Chapter 5 could be reordered to start from simple (e.g. pearson correlation plots) to more advanced measures.

## 4. Did you encounter an aspect of the synthetic data generation or evaluation process that was missing from the guide that you would like to propose a modification for?

While we found the guide to be quite useful, we made the experience that especially the real-word dataset posed challenges to us that were not mentioned in the guide. Which to some extent is expected, since a starter guide is also supposed to be of limited scope. Yet, some of the following aspects might be worth to be mentioned in the starter guide:

- The guide could emphasize more the issues regarding logical rules. There is a hint of structural zeros but an explanation of the concept of structural and sample zeros is missing.

- The guide might give the reader a hint regarding different types of data and the resulting challenges thereof. Generating synthetic datasets in the presence of clustering, e.g. persons in households or panel data, could be treated in the generation section as well in the sense of user sensitisation.

- The guide itself had no code examples, would have been nice to have a chapter or a document with collected code examples. The examples for the challenge were distributed among several documents, which was to some extend confusing