ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS 6 June 2023

Seventy-first plenary session
Geneva, 22-23 June 2023
Item 2 of the provisional agenda
Work of the High-level Group for the Modernisation of Official Statistics

SYNTHETIC DATA FOR OFFICIAL STATISTICS - A STARTER GUIDE

Note by the Secretariat

This document introduces the "Synthetic Data for Official Statistics - A Starter Guide", published in 2023 based on the High-Level Group for the Modernisation of Official Statistics (HLG-MOS) Synthetic Data Project (2020-2021) and its Synthetic Data Challenge (2022).

The note is presented to the Conference for information. The full publication is available at https://unece.org/statistics/publications/synthetic-data-official-statistics-starter-guide.

I. BACKGROUND

1. National Statistical Offices (NSOs) have always faced a tension in deciding how much information to release to users, and how much emphasis to place on minimising the risk of disclosure of the details from a specific record within an underlying microdata set. This tension has increased in recent years, as NSOs have faced greater pressure to release more detailed data, and faster than ever before. Traditionally, NSOs have managed these risks either via the public dissemination of only tabulated aggregates of the underlying micro-datasets, and/or by authorising certain groups of users (e.g., accredited researchers) access to some of their microdata (i.e., accessing individual records to perform sophisticated analyses). However, these are not ideal solutions, because:

- Tabulated aggregates will often not satisfy the information needs of many users, who may demand additional breakdowns of tabulated data. Each new breakdown provided makes it harder to suppress the risk of disclosing information about a specific individual record, and requires more resources to monitor and manage the production of such tables.
- In the case of statistical microdata access, especially stringent measures are required to manage which users can access such datasets. The process and procedures to manage access can be cumbersome, bureaucratic, time-consuming and not without risks that could potentially have serious consequences in the event of a disclosure.

2. Synthetic data provides an alternative option for managing the release of data by NSOs, which may be more convenient for certain use cases, for which synthesized micro-level records may be sufficiently realistic to satisfy the analytical requirements of the users of such data, while posing a substantially reduced risk of disclosing information about the original data.

II. DEVELOPMENT OF THE GUIDE

3. Each year, the HLG-MOS chooses two projects focusing on emerging technologies and innovative ideas to drive the modernisation of official statistics. In November 2020, the "Synthetic Data for Official Statistics" project was proposed as a candidate for the HLG-MOS projects and later selected in early 2021. The project team has developed the *Synthetic Data for Official Statistics - A Starter Guide* based on their work in 2021, the Synthetic Data Challenge (2022), and previous efforts by the HLG-MOS Blue Skies Thinking Network. The Guide was approved by HLG-MOS and released as official UNECE publication in January 2023.

III. OVERVIEW OF "SYNTHETIC DATA FOR OFFICIAL STATISTICS – A STARTER GUIDE"

4. The Guide is for those working in NSOs who are involved in managing access to statistical data, and who wish to explore the possibility of using synthetic data as a possible method for users to access it. The Guide includes the following chapters:

- Chapter 1 – Introduction
- Chapter 2 – Use of synthetic data
- Chapter 3 – Methods for creating synthetic data
- Chapter 4 – Disclosure considerations for synthetic data
- Chapter 5 – Utility measures for evaluating synthetic data

5. The Guide highlights some recent successful applications of synthetic data by a number of different NSOs, and introduces some of the different approaches that can be taken to creating synthetic data, including recommendations on which approaches to use in different situations, as well as practical tips and resources for getting started for practitioners. The Guide also includes chapters dedicated to disclosure risk considerations when releasing synthetic data (including privacy preserving techniques, and measures to assess disclosure risk), and on utility measures that can be used to assess how well the synthetic data meets the analytical needs of users.

6. The Guide is available at https://unece.org/statistics/publications/synthetic-data-official-statistics-starter-guide.