Synthetic Data For National Statistical Organizations: A Starter Guide

Feedback Workshop, November 17, 2021

Kenza Sallier, Statistics Canada

Gillian Raab, Emeritus Professor, Edinburgh Napier University, Part-time Research Fellow Administrative Data Research Centre – Scotland

Christine Task, Knexus Research

Kate Burnett-Isaacs, Statistics Canada

Today's Agenda

- Welcome
- Overview of the synthetic data
- Overview of the guide
- Use cases
- Methods recommendations
- Health Break
- Utiliy measures
- Next Steps

Slido: Understanding Users & Feedback

Go to sli.do and enter event #034032

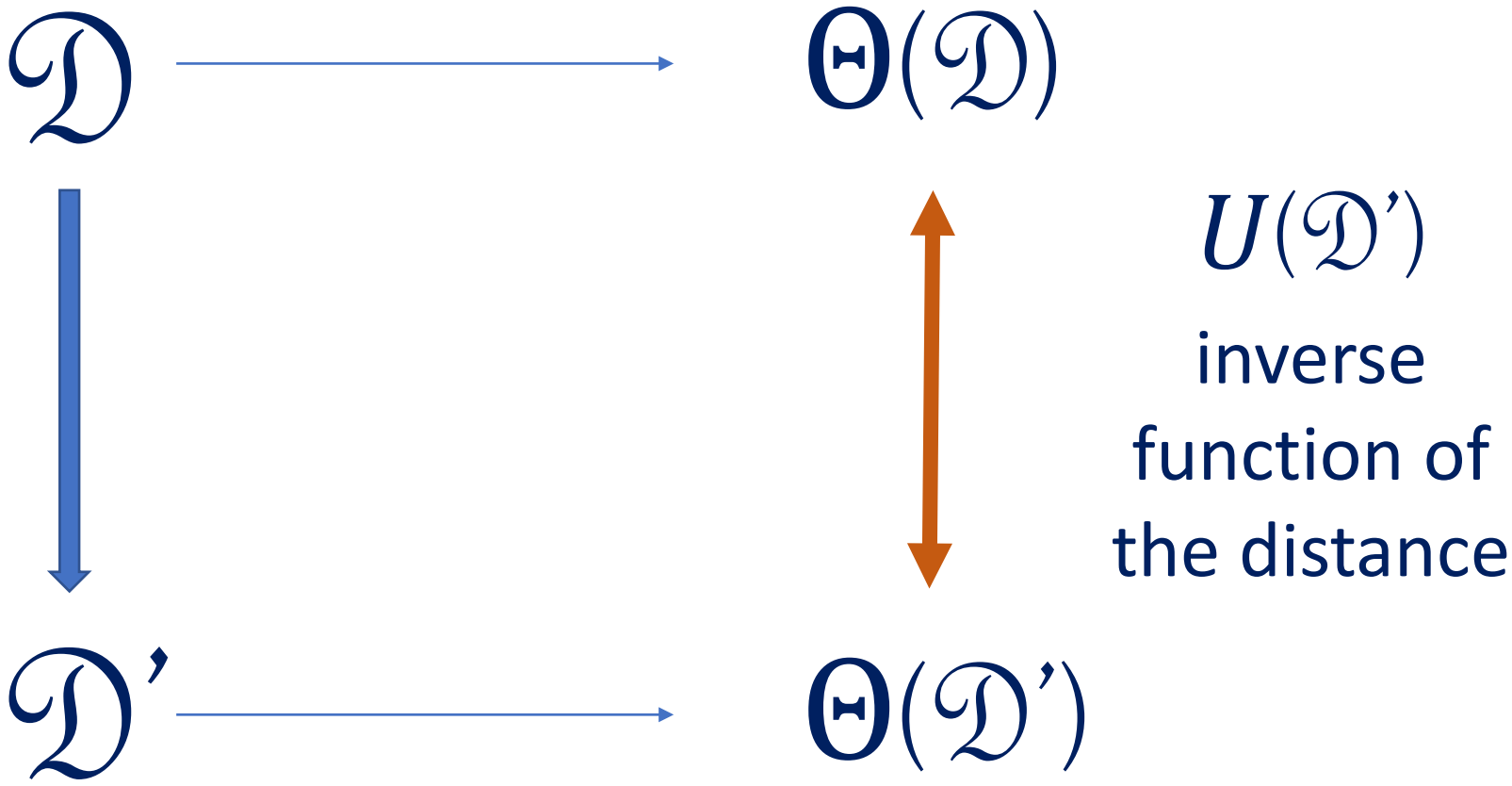
We will be using sli.do with this event number to ask questions

What Problem Would **Synthetic Data** Solve?

- National statistical offices (NSOs) are striving to provide greater transparency and openness
- Need to disseminate quality data sets to support testing, evaluation, education and development purposes
- **Output Privacy**
  Method: Confidentiality remains a top priority
- Synthetic data can be a solution to providing rich data while respecting integrity and confidentiality imperatives.

The concept of data synthesis

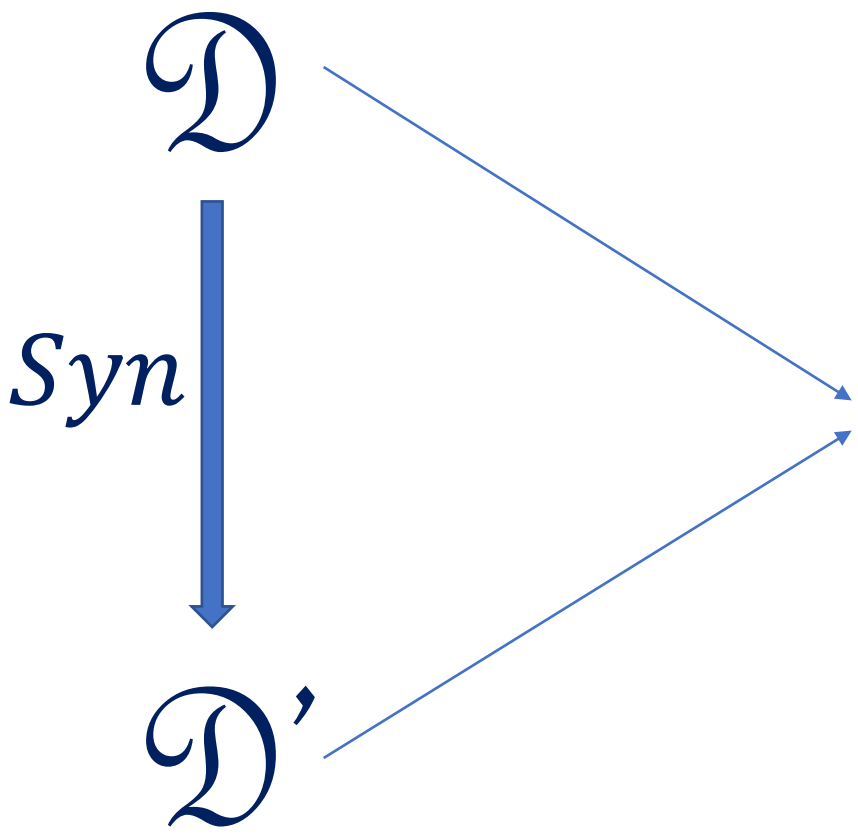
$\mathcal{D}$ the original dataset
$\mathcal{D}'$ the synthetic dataset
Syn Process creating synthetic data
Θ Results of analyses
U the utility



The concept of data synthesis

$\mathcal{D}$ the original dataset
$\mathcal{D}'$ the synthetic dataset
Syn Process creating synthetic data
Θ Results of analyses
U the utility

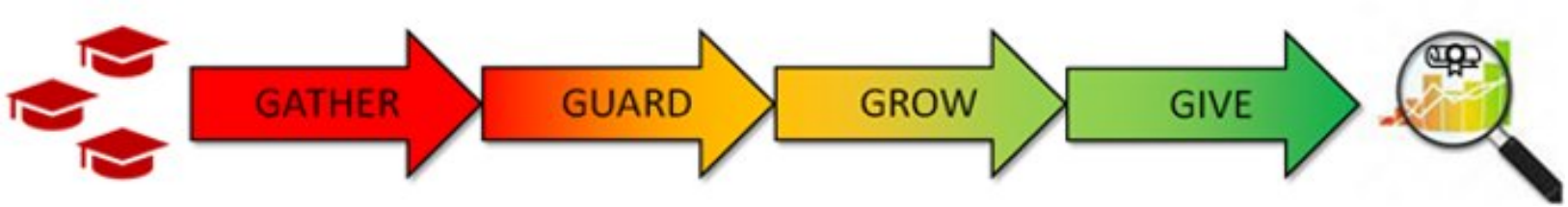
Θ should not be known in advance



Minimal distance, maximum utility

Key Concepts

- Privacy
- Sensitivity
- Security
- Utility
- Confidentiality
- Disclosure risk



What is the Synthetic Data for NSOs Starter Guide?

Synthetic Data for NSOs Starter Guide

Contents:

Purpose

Present **theoretical methods to create synthetic data** and provide an international **consensus on practical applications and best practices** to promote **consistency, transparency and comparability within** and across statistical **agencies, as well as** among **users in academia and the private sector.**

Provide coherent guidance to **decision makers** working at **any level in NSOs** so that they can **determine if synthetic data is the right solution to their data disclosure problem.**

What is Synthetic Data Suitable For?

Use Case in the Starter Guide

Disseminating to the Public

High Utility and High Confidentiality

- Want to provide microdata with high analytical value to all users
- Challenge:
- No knowledge of the type of analysis being conducted
- High need for confidentiality

Example: Statistics New Zealand Synthetic Unit Record File

- Statistics New Zealand is looking to expand the granularity of data they release through Synthetic Unit Record Files (SURFs)
- SURFs are mathematical model generated datasets, based on, but not the same as, original data.
- Stats NZ has released a few of these files in the past, including a SURF based on the NZ Income Survey in 2007, and a ‘Census for Schools’ SURF based on the NZ Household Savings Survey and NZ Census in 2019.

Testing Analysis

High Utility and High Confidentiality
Provide synthetic data to researchers or other users while they wait to get access to real data

PRIOR KNOWLEDGE OF ANALYSIS BEING CONDUCTED AND VARIABLES OF INTEREST

RESEARCHERS MAY ALREADY OF SOME LEVEL OF SECURITY CLEARANCE

Statistics Canada synthetic census-based data

- Statistics Canada is creating a synthetic version of a census-modified database in order to make the data accessible to a broader audience outside of the traditional Research Data Centers.
- The target of the synthetic dataset is to test and run the New Dynamic Microsimulation Model of Retirement Income to provide preliminary results

Education

High Utility and Medium Confidentiality

High quality data is needed in order for students, academic and users in general to learn new concepts and methods.

The more complex the methods, the more important it is that the data used in this training can provide realistic results and emulate what students will be facing in the real world.

Example: Scottish Centre for Administrative Research

- Synthetic data provided for a course on the use of administrative data for social and health research
- Original data from the linked Census and administrative records on youth employment and school attendance
- This allowed students on course to get exposure to real data and their problems.

Testing Systems

Medium Utility and Medium Confidentiality

Traditionally use dummy files

However, more and more systems will need to be tested where the outputs and analysis of those outputs need to mirror real life.

UK Office of National Statistics (ONS) Census systems testing

The ONS Census team was developing the processing platform for the 2021 UK Census Data Science Campus made a synthetic version of the previous Census to test the 2021 platform

The synthetic data were initially generated within a secure environment for use within the organisation but is being expanded with the inclusion of privacy preserving guarantees.

Feedback Question

DOES YOUR DISCLOSURE PROBLEM FIT INTO ONE OF THESE FOUR CATEGORIES?

IF NOT, PLEASE SHARE YOUR SITUATION WITH US.