Synthetic Data for National Statistical Organisations: A Starter Guide

Methods and Recommendations

HLG-MOS Synthetic Data Webinar

November 17, 2021

Kenza Sallier- Statistics Canada

Outline

- Context
- Methods, tools and recommendations
- Methods decision tree and lessons learned

Context

- Data synthesis is not new (Rubin, 1993), in theory but in practice it is (less then 10 years)
- The data revolution had (notably) two impacts that explains National Statistical Organizations (NSOs) interest around synthetic data:
  1. NSOs want to take the lead in providing data about their country: open data initiative and being more user-centric
  2. Advancement in technology and computer capacity made it possible to implement methods and develop tools
- Chapter 3 of the ***Synthetic Data for National Statistical Organisations: A Starter Guide***
- Collaborative work since January 2020
- Project had 50 participants from 15 NSOs, one academia institute and 3 private sector participants

Context

- Many methods exist: Focus is made on methods that can be **implemented** within the infrastructure of a NSO
- Selection was based on reference, if the method was truly implemented either in a academic or NSO context.
- The goal is to highlight the **applicability** of each of these methods in the practice of statistical organizations -> **provide an overview of the method, pros and cons, tools and references**
- Also provide **recommendations** on the use cases based on the methods selected: indications on how to select the right method for a given project

Things to consider when picking a method

- It is important to start by identifying the type of synthetic data required and in what context they will be used:
  - ➢ Desired analytical value to be preserved
  - ➢ Release strategy

- Complexity of the dataset, volume, software knowledge and computational capacity

- 3 main categories of methods
  1. Sequential modelling
  2. Simulated data
  3. Deep learning

Sequential modeling

- Sequential modeling: or conditional modeling -> we condition on specific variables to generate others.

- 2 methods:
  1. The fully conditional specification (FCS)
  2. Information Preserving Statistical Obfuscation (IPSO)

Sequential Modeling
Fully conditional Specification

- ✓ Stems from imputation (Van Buuren et al. 2006)
- ✓ The goal, in theory, is to preserve all relationships between variables
- ✓ Assumption that the analytical value is contained in the joint distribution of all variables
- ✓ Aims at approximating the joint distribution and generate new data points from the estimated joint distribution
- ✓ The FCS uses Bayes' Theorem to express the joint distribution of the variables as

$$\boxed{f_{X_1,X_2,\ldots,X_p} = f_{X_1} \times f_{X_2|X_1} \times \cdots \times f_{X_p|X_1,X_2,\ldots,X_{p-1}}}$$

- ✓ Now, the statistical problem is solved by approximating each of the univariate distributions

UNECE

Sequential Modeling
Fully conditional Specification

$$\boxed{f_{X_1,X_2,\dots,X_p} = f_{X_1} \times f_{X_2|X_1} \times \cdots \times f_{X_p|X_1,X_2,\dots,X_{p-1}}}$$

1. Model the univariate distribution f_{X_1} based on the original data
2. Generate values from the non conditional model in order to obtain synthetic X_1 values
3. Model the conditional distribution $f_{X_2|X_1}$ based on the original data
4. Generate values from the model using $X_{1,syn}$ values as input to obtain synthetic X_2 values
5. Repeat 3 and 4 until the last variable X_p

UNECE

Sequential Modeling
Fully conditional Specification

Tools for FCS:

- The R package Synthpop is a tool for generating synthetic datasets.
- The main method available to produce synthetic datasets is the FCS (Nowok et al., 2015).
- For more information visit www.synthpop.org.uk.

Sequential Modeling
Fully conditional Specification

Pros	Cons
This method is easy to understand and easy to explain. Because the target is the joint distribution of the dataset, this method aims at preserving (in theory) all relationships between all variables. Relationships of interest are not required to be known prior to the creation process. Furthermore, because it stems from imputation, this approach naturally bears a strong resemblance operationally speaking with its well-established data-editing sibling.	For skewed data (such as business or economic data), the presence of outliers remains a challenge in terms of disclosure or perceived disclosure control. With many variables the process can become time-consuming.

Sequential Modeling
Fully conditional Specification

Recommandations

Releasing synthetic microdata to the public & Testing analyses	Education	Testing Technology
Recommended	Can be used. If analyses conducted and statistical conclusions are pre-determined it might be too time-consuming in comparison to other methods.	Can be used but might be too advanced in comparison to the real analytical need

Sequential Modeling

Information Preserving Statistical Obfuscation (IPSO)

✓ The goal is to preserve **specific** statistics and statistical conclusions **related to linear regression**

X				Y			
X_1	X_2	…	X_p	Y_1	Y_2	…	Y_L

Non-confidential
Independent

Confidential
Dependent

- Assume multivariate normality distributions
- Assume a linear regression model

$$Y = X \cdot \beta + \Sigma$$

$$\hat{\beta}_{original} = \hat{\beta}_{synthetic}$$
$$\hat{\Sigma}_{original} = \hat{\Sigma}_{synthetic}$$

Sequential Modeling
IPSO

1. We adjust the model $Y_{original} = X \cdot \beta + \Sigma$
2. Once β and Σ estimated, we use $\hat{Y}$ ($\hat{Y} = X \cdot \hat{\beta} + \hat{\Sigma}$) as a baseline
3. We add a normally distributed noise to $\hat{Y}$ to obtain synthetic values

We cannot stop here because we need to ensure :

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{original}} = \widehat{\boldsymbol{\beta}}_{\boldsymbol{synthetic}} \text{ and } \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{original}} = \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{synthetic}}$$

4. Modify the values of $\hat{Y}$ and/or of X in order to force the equality

The synthesizer could decide to preserve other specific parameters or sufficient statistics derived from the regression model

Sequential Modeling
IPSO

Tools for IPSO:

- Mu-Argus, Implementation of Domingo-Ferrer and Gonzalez-Nicolas (2010), https://github.com/sdcTools/muargus
- R package sdcMicro, An implementation of Ting et al. (2008) is included as a noise addition method, https://cran.r-project.org/package=sdcMicro
- R package RegSDC, Implementation of all methods described in Langsrud (2019), https://CRAN.R-project.org/package=RegSDC

Sequential Modeling
IPSO

Pros	Cons
Like the FCS, the method is easy to understand and to explain. With this method, it is possible to preserve exactly some pre-identified parameters and sufficient statistics. Thus, any analysis relying on (multivariate) normality will produce the exact same results in the original and synthetic data. IPSO can be implemented as part of another method or process, to generate synthetic datasets. These hybrid methods may be used alleviate the normal distributions assumptions	Normal distribution for all variables is a strong assumption that is seldom true.

Sequential Modeling
IPSO

Recommandations

Releasing synthetic microdata to the public & Testing analyses	Education	Testing Technology
Recommended if the analyses are all related to linear regressions, otherwise not recommended.	Recommended if the analyses are all related to linear regressions otherwise not recommended	Can be used but might be too advanced in comparison to the real analytical need

Simulated Data

- Simulations are often used in statistics to generate artificial data in order to conduct empirical analyses

- Thus, a new perspective on simulations is that simulation processes can be used to create artificial data as ***synthetic data.***

- 2 methods:
  1. From dummy files to more analytically advanced synthetic files
  2. Pseudo likelihood

Simulated Data

From dummy files to more analytically advanced synthetic files

X_1	X_2	…	X_p

$$\overrightarrow{X_i} \sim \mathrm{N}(\mu, \sigma), i.i.d., \forall\, i = 1, \dots p$$

Chosen without using any real data

Perfectly safe

No analytical value and considered a dummy file

$\neq$

Not useful

UNECE

Simulated Data

From dummy files to more analytically advanced synthetic files

The synthesizer could also decide to use information from the original data, in the generation process, to ensure that some of the analytical value is preserved.

X_1	X_2	…	X_p

$$\overrightarrow{X_1} \sim \mathrm{N}(\widehat{\mu_1}, \widehat{\sigma_1})$$
$$\overrightarrow{X_2} \sim \mathrm{N}(\widehat{\mu_2}, \widehat{\sigma_2})$$
$$\vdots$$
$$\overrightarrow{X_p} \sim \mathrm{N}(\widehat{\mu_p}, \widehat{\sigma_p})$$

Estimates based on the original data

Simulated Data
From dummy files to more analytically advanced synthetic files

- The simulation process can be refined by using more information from the original microdata to preserve more statistical properties

- The Fleshman-Vale-Maurelli method (Fleishman, 1978 and Maurelli, 1983) can generate multivariate non-normal distributions with the following features preserved:
  - ✓ means
  - ✓ variances
  - ✓ Skews
  - ✓ Intercorrelation between variables

→ It is possible to preserve pre-identified statistics

UNECE

Simulated Data
From dummy files to more analytically advanced synthetic files

Tools:

- In general, simulation processes can be programmed easily enough using any software.

- For a more complex type of simulation, the R package semTools (https://CRAN.R-project.org/package=semTools) simulates microdata using the co-variance matrix, skewness and kurtosis from the original sample data (Jorgensen et al. 2019).

Simulated Data
From dummy files to more analytically advanced synthetic files

Pros	Cons
Simulation processes are easy to understand and can create completely safe data when no information pertaining to the original data is used. Can generate fully synthetic files. For more advanced types of simulations, some analytical value can be preserved.	Usually, does not allow to meet complex analytical needs.

UNECE

Simulated Data
Dummy Files

Recommandations

Releasing synthetic microdata to the public & Testing analyses	Education	Testing Technology
Not recommended	Can be used if training does not require analytical value in the data	Recommended

Simulated Data

Analytically Advanced Simulated Data

modernstats by HLG - MOS

Recommandations

Releasing synthetic microdata to the public & Testing analyses	Education	Testing Technology
Recommended if analyses conducted are related to the pre-identified results that needed to be preserved in the synthesis process. Otherwise, not recommended.	Recommended if analyses conducted are related to the pre-identified results that needed to be preserved in the synthesis process. Otherwise, not recommended.	Can be used but might be too advanced in comparison to the real analytical need

UNECE

Simulated Data
Pseudo likelihood

- Most of the research related to data synthesis has been focused on census datasets and administrative data

- Other methods presented work under the assumption that the original data covers the entire population of interest

- Therefore, statistical conclusions obtained via the synthetic file can only be comparable to the ones obtained in the original sample and **not necessarily the original population**.

- Issue when the sampling process follows an **informative design** (Lavallée and Beaumont, 2015)

- With NSOs collecting data via many projects relying on probabilistic surveys, a natural question which arises is: how do we include survey design features in the data synthesis process?

Simulated Data
Pseudo likelihood

More detailed discussion on synthetizing weights in the guide

- Goal: incorporate information of the sampling process in the data synthesis in order to obtain a synthetic dataset from which we can estimate characteristics of the original **population**
- An option : providing synthetic weights with the synthetic *sample*
- Another option: use weighted models to approximate distributions from the original population, and generate values from them
- The pseudo likelihood method was suggested in this case.
- Example of an advanced simulation process that generates data of high analytical value.
- The idea is to preserve as much all links between variables and univariates statistics as they exist in the original **population**

Simulated Data
Pseudo likelihood

- The pseudo likelihood method generates synthetic populations by incorporating survey weights into the models based on the pseudo likelihood approach (Kim et al, 2020).
- The idea is to build to estimate the distributions of the finite population.
- Once that the finite population density is estimated, the synthesizer can generate fully synthetic populations by drawing values repeatedly from it.
- No explicit distributions so we cannot rely on more regular models
- Requires to derive the full conditional distributions of the Markov Chain Monte Carlo (MCMC) algorithm for posterior inference by using the pseudo likelihood function.

Simulated Data
Pseudo likelihood

Tools:

- There is no known tool per se to apply the method. However, section 2.1 and 2.2 of Kim et al (2020) provides detailed information on the method and how to implement it.

Simulated Data
Pseudo likelihood

Pros	Cons
Addresses informative sampling. When generating synthetic populations, the sampling process is already accounted for; thus, the uncertainty introduced by sampling process is also accounted for. Users can estimate parameters from the original ***population***. Providing synthetic populations can be better than providing synthetic samples convenience -> no need to estimate sampling variance nor to provide synthetic weights	There are potential challenges with the choice of prior distributions in the MCMC algorithm

Simulated Data
Pseudo likelihood

Recommandations

Releasing synthetic microdata to the public & Testing analyses	Education	Testing Technology
Strongly recommended if users want to estimate statistics from the original finite population.	Can be used. If analyses conducted and statistical are pre-determined it might be too time-consuming in comparison to other methods.	Can be used but might be too advanced in comparison to the real analytical need

UNECE

Deep Learning
GAN

- With improvements in technology and computational capacity, implementation of machine learning processes have become easier and more accessible.

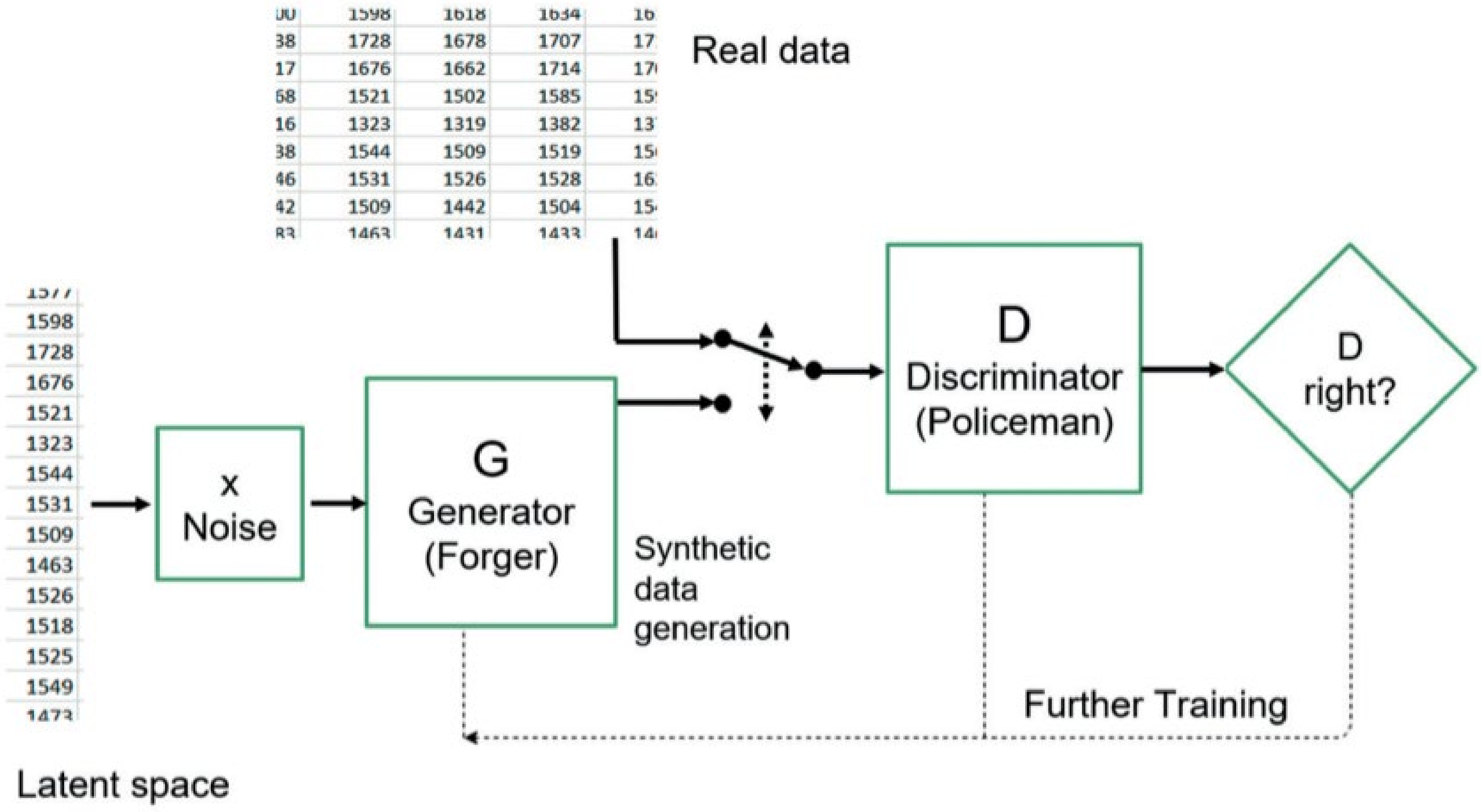
- Machine learning approaches have been more and more employed to generate synthetic datasets.

- More specifically, the use of deep learning models has become appealing because of their capacity to extract from big datasets very powerful predicting model.

- The generative adversarial network (GAN) (Goodfellow, et al., 2014) is a prominent generative model used for synthetic data generation.

Deep Learning
GAN

Theory and implementation processes can be technically challenging, we will mainly explain the overall concepts, as more information can be found in the references.



Source: https://doi.org/10.1111/1740-9713.01466

Deep Learning
GAN

Tools:

- There is no known tool per se to apply the method. However, Kaloskampis et al (2020) provides detailed information on the method and how to implement it.

Deep Learning
GAN

Pros	Cons
GAN can be used to generate continuous, discrete but also text datasets, while ensuring that the underlying distribution and patterns of the original data are preserved. Can generate fully synthetic datasets. Aims at preserving all relationships between variables. Can handle unstructured data.	GAN can be seen as complex to understand, explain or implement when there is only a minimal knowledge of neural networks. There is often a criticism associated to neural networks as lacking of transparency or being a black box. The method is time consuming and has a high demand for computational resources.

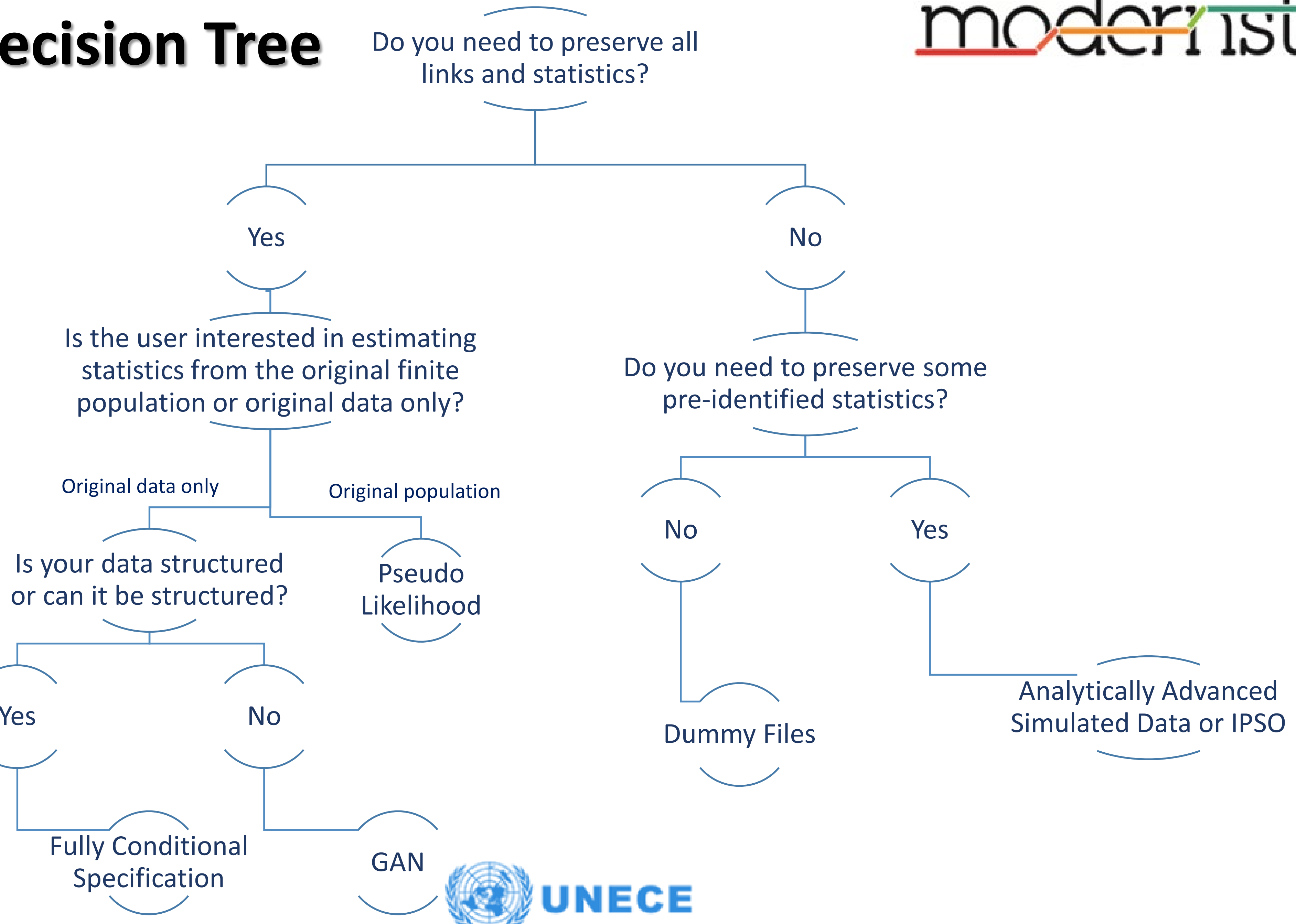
Deep Learning
GAN

Recommandations

Releasing synthetic microdata to the public & Testing analyses	Education	Testing Technology
Recommended especially in presence of text or unstructured data	Can be used. If analyses conducted and statistical conclusions are pre-determined it might be too time-consuming in comparison to other methods.	Can be used but might be too advanced in comparison to the real analytical need

UNECE

Methods Decision Tree



Some lessons learned

- Many methods exist
- More complex methods are not always the best option (ex: testing technology)
- It’s really important to understand how the synthetic datasets will be used

References

- Cano, Isaac and Torra, Vicenç. (2009), Generation of Synthetic Data by means of fuzzy c-Regression. 1145-1150. 10.1109/FUZZY.2009.5277074.
- Domingo-Ferrer, J., Gonzalez-Nicolas, U. (2010), Hybrid microdata using microaggregation, Information Sciences, 180(15), 2834-2844.
- Drechsler J. (2011), Synthetic Datasets for Statistical Disclosure Control. New York: Springer-Verlag.
- Drechsler, J. and J.P. Reiter. (2011), An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets. Computational Statistics and Data Analysis, 55, 3232-3243.
- Efron, B. (1979), Bootstrap Methods: Another Look at the Jackknife." Ann. Statist. 7 (1) 1 – 26, January.
- Fleishman, Allen I. (1978), A Method for Simulating Non-Normal Distributions.
- Goodfellow, I. Pouget-Abadie, J. Mirza, M. Xu, B. Warde-Farley, D. & Ozair, S. Courville, A. and Bengio, Y. (2014), Generative Adversarial Networks. Advances in Neural Information Processing Systems. 3. 10.1145/3422622.
- Jorgensen, T. D., Pornprasertmanit, S. Schoemann, A. M. and Rosseel, Y. (2019), semTools : Useful Tools for Structural Equation Modeling.

References

- Kim, H. J. Drechsler, J. and Thompson, K. J. (2021), Synthetic microdata for establishment surveys under informative sampling, Journal of the Royal Statistical Society Series A, Royal Statistical Society, vol. 184(1), pages 255-281, January.
- Kaloskampis, I., Joshi, C., Cheung, C., Pugh, D. and Nolan, L. (2020), Synthetic data in the civil service. Significance, 17: 18-23. https://doi.org/10.1111/1740-9713.01466
- Langsrud, Ø. (2019), Information Preserving Regression-based Tools for Statistical Disclosure Control, Statistics and Computing, 29, 965–976.
- Lavallée, P. and Beaumont, J.-F. (2015), Why We Should Put Some Weight on Weights. Survey Insights: Methods from the Field, Weighting: Practical Issues and 'How to' Approach, Invited article, Retrieved from https://surveyinsights.org/?p=6255
- L'Ecuyer, P. and Puchhammer, F. (2021), Density Estimation by Monte Carlo and Quasi-Monte Carlo, Monte Carlo and Quasi-Monte Carlo Methods.
- Muralidhar, K., Sarathy, R. (2008), Generating Sufficiency-based Non-synthetic Perturbed Data, Transactions on Data Privacy, 1(1), 17-33
- Nowok, B., Raab, G.M. and Dibben, C. (2015), synthpop: Bespoke creation of synthetic data in R.

References

- Rubin, B., (1993),Discussion: Statistical disclosure limitation. Journal of Official Statistics 9(2), pp 462-468.
- Sallier, K. (2020), Toward More User-centric Data Access Solutions: Producing Synthetic Data of High Analytical Value by Data Synthesis’. Statistical Journal of the IAOS, vol. 36, no. 4, pp. 1059-1066, 2020
- Ting, D., Fienberg, S.E., Trottini, M. (2008), Random orthogonal matrix masking methodology for microdata release, International Journal of Information and Computer Security, 2(1), 86-105.
- Van Buuren, S., Brand, J. P. L. Groothuis-Oudshoorn, C. G. M. and Rubin, D. B. (2006), Fully Conditional Specification in Multivariate Imputation. Journal of Statistical Computation and Simulation 76 (12): 1049–64.
- Vale, C. D., and Maurellim V. A. (1983), Simulating Multivariate Nonnormal Distributions. Psychometrika 48 (3): 465–71.

Thank you!
Questions?

Kenza.sallier@statcan.gc.ca