

Synthetic and Hybrid Data in ESSNET-SDC

Josep Domingo-Ferrer¹, Jörg Drechsler², Silvia Polettini³
and Úrsula González-Nicolás¹

¹ Universitat Rovira i Virgili, Catalonia

² IAB, Germany

³ Università di Napoli, Italy

- 1 Introduction
- 2 The ESSNET-SDC Deliverable on Synthetic Data
 - The introduction of the deliverable
 - Information-preserving synthetic data
 - Synthetic datasets based on multiple imputation
 - Pros and cons of the different approaches
 - Suggestions for Eurostat
- 3 "R" implementation of Muralidhar-Sarathy's hybrid generator
- 4 Microaggregation-based hybrid data
- 5 "R" implementation of the microaggregation-based generator

Delivered items

- ESSNET-SDC Deliverable: Report on Synthetic Data Files, by J. Domingo-Ferrer, J. Drechsler and S. Polettini.
- "R" implementation of Muralidhar-Sarathy's hybrid data (for μ -Argus), by Ú. González-Nicolás.
- Description of microaggregation-based hybrid data, by J. Domingo-Ferrer.
- "R" implementation of microaggregation-based hybrid data, by Ú. González Nicolas.

The ESSNET-SDC Deliverable on Synthetic Data

Final version delivered on Jan. 6, 2009 with table of contents:

- 1 Introduction
- 2 Information-preserving synthetic data
- 3 Synthetic datasets based on multiple imputation
- 4 Pros and cons of the different approaches
- 5 Suggestions for Eurostat

The introduction of the deliverable

The following methods in the previous literature on synthetic data:

- Data distortion by probability distribution
- Synthetic data by multiple imputation
- Synthetic data by bootstrap
- Synthetic data by Latin Hypercube Sampling
- Partially synthetic and hybrid microdata
- Data shuffling

Information-preserving synthetic data

- The Information-Preserving Statistical Obfuscation (IPSO) procedure by Burrige (2003) is first described.
- A procedure to generate numerical hybrid data mixing original data with IPSO-generated data and preserving means and covariances of the original data, by Muralidhar and Sarathy (2008), is then reviewed.
- The parameter choice in Muralidhar-Sarathy (2008) may be quite challenging for the non-expert.

Fully synthetic datasets based on multiple imputation

- The Rubin (1993) approach to synthetic data generation by multiple imputation (MI) is described.
- The MI approach applies to any type of data, numerical or categorical.
- The disclosure risk of synthetic, MI-generated data is very low.
- The joint distribution of the original data does not differ from the joint distribution of the MI-generated data.
- An example application in the German IAB Establishment Survey is discussed.

Partially synthetic datasets based on multiple imputation

- In this approach, only observed values that bear a high disclosure risk (key variables) or very sensitive values are replaced with synthetic values.
- Imputed values are drawn from the posterior predictive distribution $f(Y|X)$, where Y indicates variables that need to be modified to avoid disclosure and X are variables that remain unchanged.
- Disclosure risk is higher than for fully synthetic microdata.
- Both data utility *and* disclosure risk should be checked.
- An example application to the U.S. Survey of Income and Program Participation (SIPP) is discussed.



Pros and cons of the different approaches

- A choice must be made between fully synthetic or partially synthetic data.
- Fully synthetic data offer very low disclosure risk but also limited data utility, which may, however be more than just preservation of the model they were simulated for.
- Partially synthetic and hybrid data are two different approaches to mixing original and synthetic data in quest of more utility.

Pros and cons of the different approaches

- Model-based synthetic data generation, such as multiple imputation, can preserve constraints inherent to the data semantics, which in general cannot be achieved by Muralidhar-Sarathy hybrid data.
- Offering utility for arbitrarily user-defined subdomains is a challenge not met by the vast majority of synthetic data generators.
- Multiple imputation and model-based simulation methods are relatively transparent to the data user: metadata can be published and the user can assess how much distortion there was.
- This is less true for Muralidhar-Sarathy hybrid data.



Suggestions for Eurostat

- Simple methods like those in the Introduction and Muralidhar-Sarathy's are rather limited as to the statistics and models they preserve.
- MI-based synthetic data are more flexible.
- Recommendations are given for MI-based fully or partially synthetic data generation.
- It is suggested that Eurostat, like other major statistical offices such as U. S. Census Bureau, should fund research and development of synthetic or hybrid datasets for their most important surveys.
- Investment would be also needed to train European NSIs in understanding and using such complex methods.



"R" implementation of Muralidhar-Sarathy's hybrid generator

- Written as an "R" package by Ú. González-Nicolás (URV), and sent to Anco for possible inclusion in μ -Argus.



Microaggregation-based hybrid data

- Proposed by Domingo-Ferrer (2009), to solve some of the above shortcomings of hybrid data
- It consists of
 - 1 Running microaggregation on an original data set, with minimum group size k
 - 2 Within each of the groups obtained from microaggregation, run IPSO and replace the original data in the group with the IPSO-generated data.
 - 3 The resulting data are hybrid data.

On parameter k

The method parameterization is very simple:

- The smaller k , the more similar are the output hybrid data to the original data.
- The larger k , the more synthetic are the output hybrid data (for k equal to the data set size, they are completely synthetic).

Data utility

- The output hybrid data exactly preserve the means and the covariances of the original data.
- The output hybrid data approximately preserve higher-order moments, provided that k is small and the microaggregation technique is good enough (yields homogeneous group).
- For arbitrary user-defined subdomains, means, covariances and higher-order moments are preserved, provided the same conditions above.

"R" implementation of the microaggregation-based generator

- Written as an "R" package by Ú. González-Nicolás (URV), and sent to Anco for possible inclusion in μ -Argus.