Institute for Employment Research

The Research Institute of the Federal Employment Agency

A short introduction to synthetic data for statistical disclosure control

European Data Access Forum

Luxembourg, 24.03.2015

Jörg Drechsler

The Idea Behind Synthetic Data

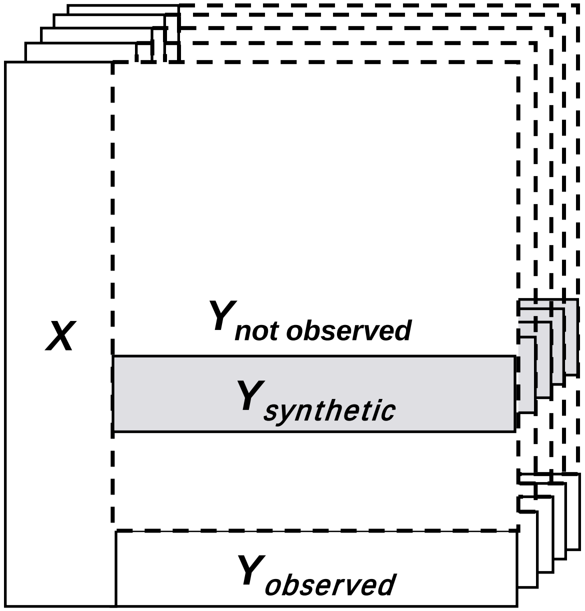
- SDL technique for data dissemination
- idea is closely related to multiple imputation for nonresponse
- generate synthetic datasets by drawing from a model fitted to the original data
- not the missing values but the sensitive values are replaced with a set of plausible values given the original data
- generate multiple draws to be able to obtain valid variance estimates from the synthetic data

The Idea Behind Synthetic Data

- three steps necessary for data release:
  - Fit model to the original data
  - Repeatedly draw from that model to generate multiple synthetic datasets
  - Release these datasets to the public

- over the years different designs for generating synthetic data evolved
- two main approaches: fully synthetic datasets and partially synthetic datasets

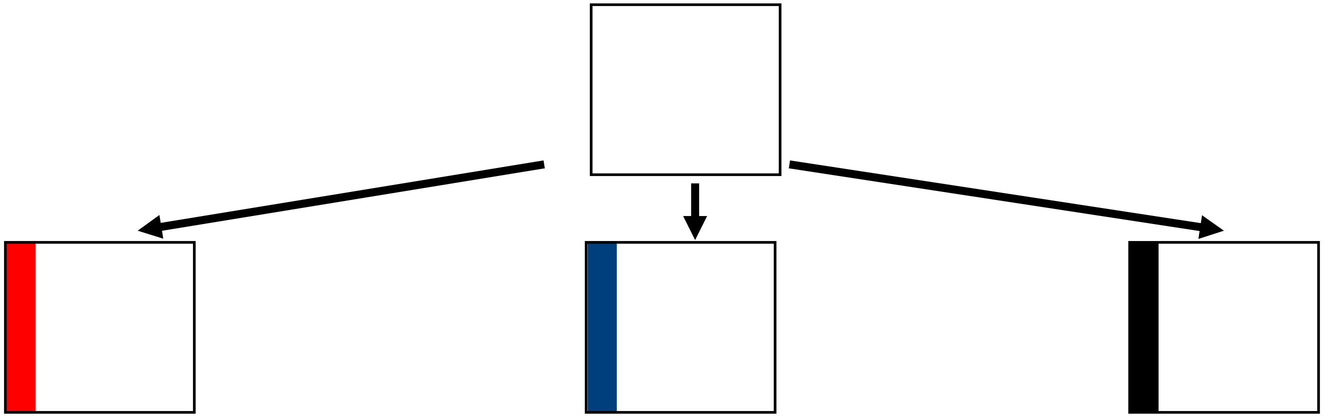
Fully Synthetic Data (Rubin, 1993)



- advantages: - data are fully synthetic
  - re-identification of single units almost impossible
  - all variables are still fully available
- disadvantages: - strong dependence on the imputation model
  - setting up a model might be difficult/impossible

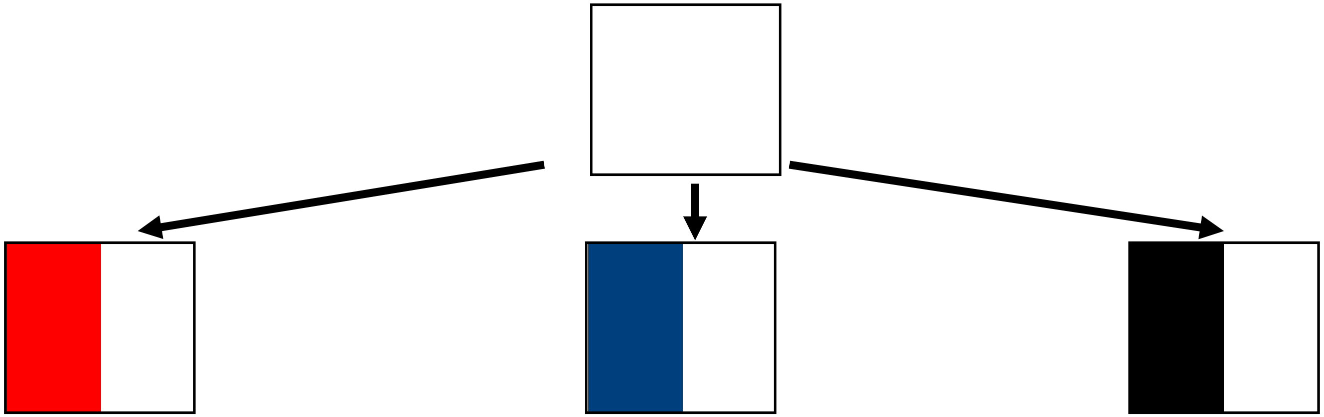
Partially Synthetic Data (Little, 1993)

- only potentially identifying or sensitive variables are replaced



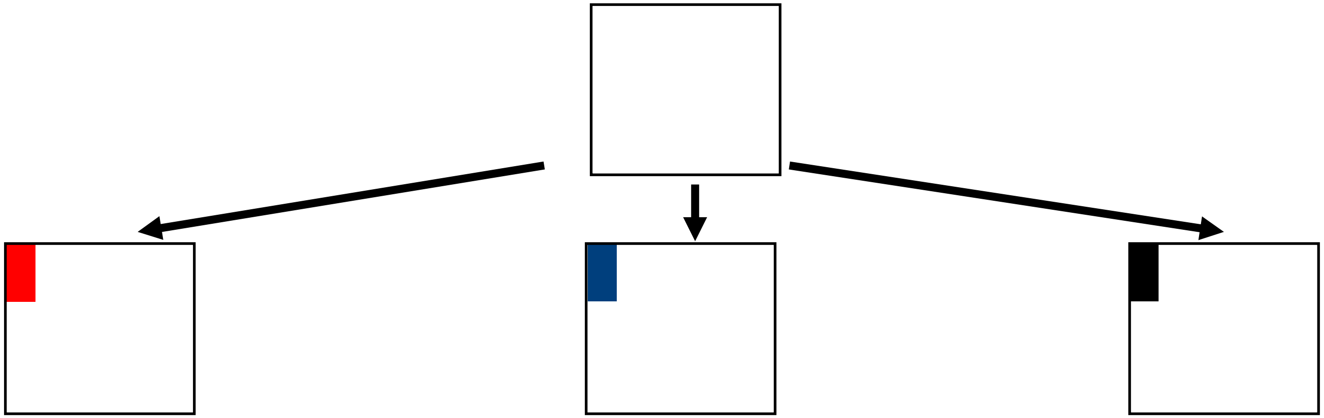
Partially Synthetic Data (Little, 1993)

- only potentially identifying or sensitive variables are replaced



Partially Synthetic Data (Little, 1993)

- only potentially identifying or sensitive variables are replaced



- advantages: - model dependence decreases
  - models are easier to set up
- disadvantages: - true values remain in the dataset
  - disclosure might still be possible

Analysis of Synthetic Data

- analysis based on the synthetic data is straight forward for the user
- analyse each synthetic dataset separately
- combine the results from the different datasets to obtain final estimates
- comparable to combining procedures for multiple imputation for nonresponse
- combining procedures for the estimated variance of the estimated parameter of interest differs

Synthetic Data Compared to Other SDC Techniques

Advantages

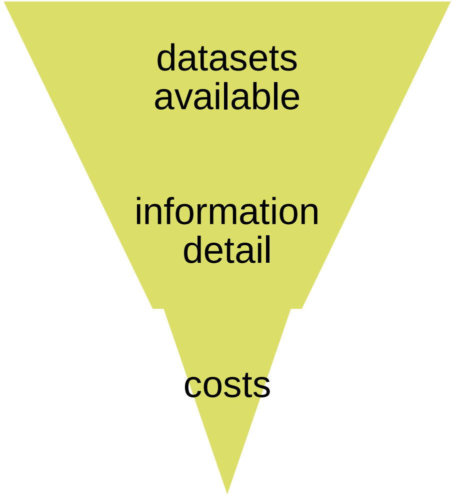
- tries to preserve the multivariate relationship between the variables and not only specific statistics
- users can analyze each dataset using standard software
- suitable for any variable type
- can address some of the problems often encountered in practice
  - item nonresponse
  - skip patterns
  - logical constraints

Disadvantages

- lot of work
- depends heavily on the quality of the imputation models

Synthetic Data as a Data Access Concept

- synthetic data should not be the only piece in the data dissemination puzzle
- there is nothing like **the** data user
- current situation: three different data access channels
  - on-site access
  - remote analysis servers / remote execution
  - data dissemination



- all channels have there advantages and disadvantages
- remote access interesting alternative but with open questions

Situations in Which Synthetic Data Can be Useful

- type of data
  - highly sensitive data
  - data with very skewed distributions
- for descriptive analysis and policy information
- as a source of information to guide decision whether application for access to the confidential data at the RDC is worth the effort
- to develop analysis code to be run on the original data
- probably not to obtain results to be published in scientific journals

Synthetic Data in Practice

- several synthetic data products have been released
- U.S. Census Bureau the main driving force
  - the SIPP/SSA/IRS Public Use File Project
  - information on individuals living in group quarters in the ACS
  - synthetic LBD
  - OntheMap
  - more products in the development stage
- agencies around the world conduct research on synthetic data

A Brief Practical Example – The Synthesis of the IAB Establishment Panel

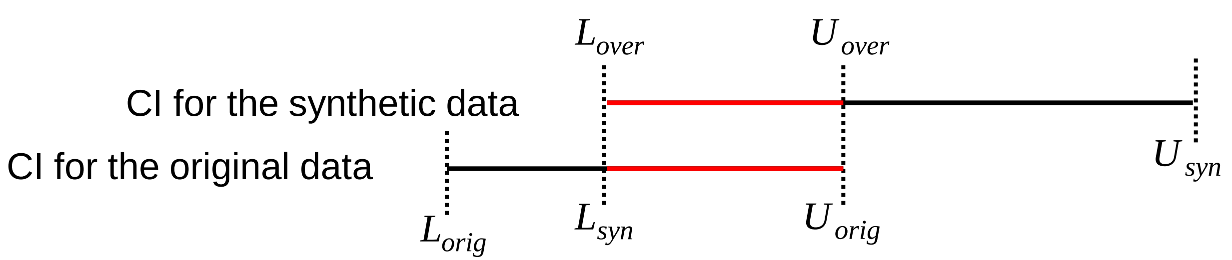
- Partially synthetic data to leave as much data unchanged as possible
- Restriction to one wave (2007)
- Almost all continuous variables are synthesized
- Combination of variables that could be used for re-identification purposes (e.g. region, industry, establishment size) and sensitive variables (e.g. turnover, subsidies)
- All records are synthesized for each variable

- Two regression analyses to illustrate analytical validity of the synthetic data

A Useful Analytical Validity Measure – The confidence interval overlap

- Suggested by Karr et al. (2006)
- Measure the overlap of CIs from the original data and CIs from the synthetic data
- The higher the overlap, the higher the data utility
- Compute the average relative CI overlap for any estimate of interest

$$J_k = \frac{1}{2}\left[\frac{U_{over,k} - L_{over,k}}{U_{orig,k} - L_{orig,k}} + \frac{U_{over,k} - L_{over,k}}{U_{syn,k} - L_{syn,k}}\right]$$

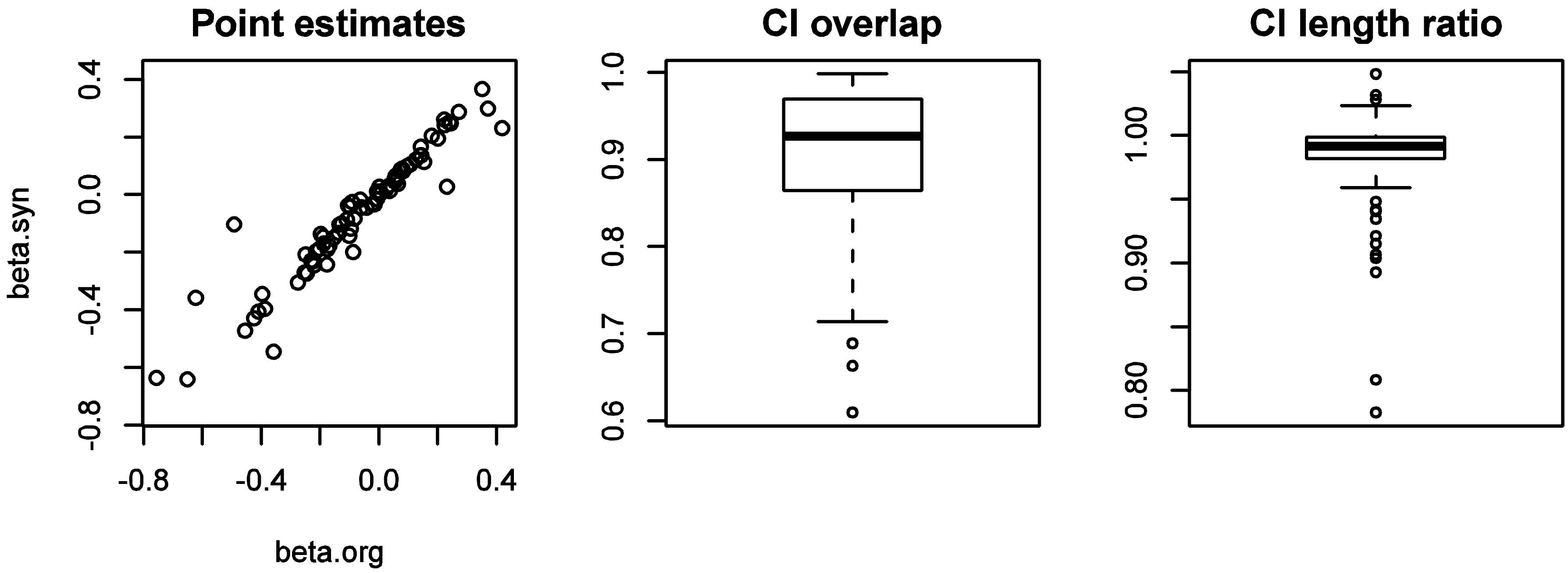


Results from the first regression

	beta org.	*beta syn.*	*J.k.beta*	*z-score org.*	*z-score syn.*	*CI length ratio*
Intercept	-0.809	-0.752	0.87	-7.23	-6.85	0.99
5-10 employees	0.443	0.437	0.97	8.52	7.99	1.06
10-20 employees	0.658	0.636	0.90	11.03	10.88	0.98
20-50 employees	0.797	0.785	0.95	13.02	12.36	1.04
100-200 employees	0.892	0.908	0.96	9.23	9.48	0.99
200-500 employees	1.131	1.125	0.99	9.99	9.87	1.01
>500 employees	1.668	1.641	0.97	8.22	8.33	0.97
growth in employment exp.	0.010	0.006	0.98	0.18	0.12	0.99
decrease in emp. expected	0.087	0.100	0.96	1.11	1.27	1.00
share of female workers	1.449	1.366	0.73	17.63	18.71	0.89
share of employees with university degree	0.319	0.368	0.91	2.18	2.59	0.97
share of low qualified workers	1.123	1.148	0.93	12.17	11.87	1.05
share of temporary employees	-0.327	-0.138	0.75	-1.74	-0.71	1.05
share of agency workers	-0.746	-0.856	0.88	-3.09	-4.24	0.84
employment in the last 6 month	0.394	0.369	0.87	8.33	7.82	1.00
dismissal in the last 6 months	0.294	0.279	0.92	6.38	6.03	1.00
foreign ownership	-0.113	-0.117	0.99	-1.33	-1.38	0.99
good or very good profitability	0.029	0.033	0.98	0.72	0.82	0.99
salary above collective wage agreement	0.020	0.031	0.95	0.35	0.54	0.99
collective wage agreement	0.016	0.007	0.95	0.31	0.13	0.97

- Average CI overlap: 0.92

Results from the second regression



- Average CI overlap: 0.91 Minimum CI overlap: 0.61

Major Concerns

generating useful synthetic data is too much work

- will always be work intensive
- practical implementations only for ten years
- everything needed to be developed from scratch
- a lot has been learned since then
- new projects can build on this
- nonparametric modeling tools such as CART can further simplify the modeling task
- software for generating synthetic data is now available

Major Concerns

synthetic data will never be accepted by the users

- new ideas are always confronted with skepticism
- agencies have several options to build trust in the synthetic data results
- conduct many analytical validity checks and publish results
- provide details regarding the models that have been used to generate the data
- give guarantee to run final results on the original data
- establish verification servers

Conclusions

- disseminating disclosure protected data that will provide valid results for any possible query is impossible
- useful SDC method should provide the user with information, which analyses might provide valid results
- synthetic data most promising approach for data dissemination
- should only be one pillar of a general data access strategy
- on-site/remote access useful other pillars
- synthetic data still helpful to provide useful test data
- step from theoretical concept to practical implementation managed successfully

IAB

Thank you for your attention

Jörg Drechsler
joerg.drechsler@iab.de

Major Concerns

quality of the synthetic data strongly depends on the quality of the models

- certainly true
- especially critical in the design based world of official statistics
- agencies can release information on the models used to generate the synthetic data
- might be possible to generate customized synthetic datasets in the future
- generate synthetic data on the fly