



The Future for Synthetic Data

Jerry Reiter

Department of Statistical Science

Duke University



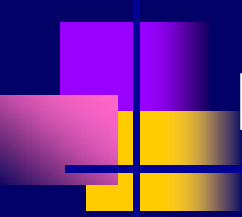
Some key benefits of approach

- Higher intensity of SDL possible (not necessary).
- Handle missing and synthetic data at same time.
- Preserve tails of distributions and geographies.
- Analysts use standard methods.
- Agency can reveal nature of SDL to public.



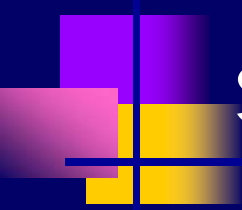
Existing methodology

- Fully and partially synthetic data.
- Missing and confidential data simultaneously.
- Different numbers of imputations for different variables.
- Some flexible synthesis methods.
- Risk measures for partial synthesis.
- Inferences for scalars.
- Tests of multi-component hypotheses.



Broad classification of future research aims.

- Partial synthesis designs.
- Better synthesis methods.
- Increased range of analyses.
- Proof that synthetic data works.



Partial synthesis: How do we select values to synthesize?

- Replace records deemed at high risk of identification or attribute disclosure, but want to have high utility as well.
- Requires measurements of risk at record level that should account for:
 - multiple copies of data,
 - intruder knowledge and behavior,
 - information released about models.
- Requires measurements of utility of data sets.



Issues for selection

- May be sufficient from risk perspective to replace only some (not all) identifiers.

Example:

Person is unique on age, race, sex, marital status, county.

Not unique if either (i) county is not released exactly, (ii) age is not released exactly, or (iii) sex and race both are not released exactly.



Issues for selection

- Replacing values for some records could impact risks for other records.

Example: Intruder knows that record 1 has larger income than record 2, and could identify these records from released incomes.

Synthesizing income for either record might reduce risk for both records.



Issues for selection

- Some variables have greater impact on data usefulness than others.

Example:

X is nearly uncorrelated with variables of interest to analysts. Synthesizing X may have low impact on overall data utility.

Record has high leverage for a regression, and leverage is attributable to one variable X. Altering that record's X has greater impact on coefficients than altering other values.



Partial synthesis: Selecting synthetic datasets

- Toss out synthetic datasets that give implausible results or that are too risky.
- What metrics do we use for these decisions?
- What is the effect on inferences?



Flexible imputation models

Synthetic data models should preserve as many relationships as possible.

Implications?

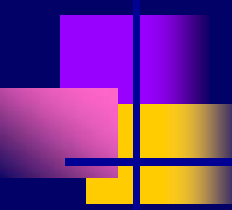
- accept variance to avoid bias
- use informative prior distributions
- sacrifice full Bayesian simulation
- need to worry about disclosure risk
- use semi- and non-parametric methods



Increase range of analyses

Techniques for valid inferences only for simple estimands. How about:

- Multivariate analysis
- Hierarchical Bayesian models
- Automated methods of model selection
- Diagnostics from multiple datasets.



Trusting synthetic data: Interactive data quality feedback via verification servers

- Analyst computes $Q(M)$ on synthetic data, M .
- Analyst sends request for quality of $Q(M)$ to a server that has M and the original data, O .
- Server computes $FM(Q(M), Q(O))$, which measures the similarity of $Q(M)$ and $Q(O)$.
- Server reports FM to analyst without $Q(O)$.
 FM leaks information about values in O .



Concluding remarks

- Many unsolved problems, but widespread investigation of synthetic data only in last decade.
- Worth pursuing this agenda because:
 - risks may grow so much as to require large amounts of data alteration.
 - traditional SDL would distort data utility too much.
 - synthetic approaches could be one of the last remaining ways to provide microdata.



Inference with partially synthetic datasets (no missing data)

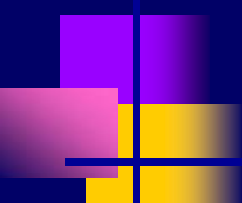
Reiter (2003, *Survey Methodology*)

- Estimand: $Q = Q(X, Y)$

- In each synthetic dataset d_i

$$q_i = Q(d_i) \quad u_i = U(d_i)$$

Quantities needed for inference (no missing data)


$$\bar{q}_m = \sum_{i=1}^m q_i / m$$

$$b_m = \sum_{i=1}^m (q_i - \bar{q}_m)^2 / (m - 1)$$

$$\bar{u}_m = \sum_{i=1}^m u_i / m$$



Inference with partially synthetic data (no missing data)

- Estimate of Q : \bar{q}_m

- Estimate of variance is

$$T_p = \bar{u}_m + b_m / m$$

- For large n and m , use normal based inference for Q :

$$\bar{q}_m \pm 1.96 \sqrt{T_p}$$