A Vision for the Future of Data Access

Jerry Reiter
Department of Statistical Science
Duke University
PI: Duke/NISS NCRN node

Acknowledgements

- Research ideas in this talk supported by
  - National Science Foundation
    - SES-11-31897 , ACI 14-43014, CNS-10-12141
  - National Institutes of Health: R21-AG032458
  - US Bureau of the Census

An argument for public use data

- Record-level data are enormously beneficial for society
  - Facilitates research and policy-making
  - Trains students at skills of data analysis
  - Enables development of new analysis methods
  - Helps citizens understand their communities

- Even in a world where analysis is brought to the data

But….

- Data stewards are ethically and often legally obligated to protect confidentiality
- Releasing record level data can be risky for data subjects (and for data stewards!)
- Big data even riskier to release
  - Data often from administrative sources or social media, hence available to others
  - Large number of variables for matching
  - With many variables everyone is a population unique

What can be done?

- Many data stewards alter data before releasing them
  - Aggregate data (coarsen geography, top-code, collapse categories)
  - Suppress data
  - Swap variables across records
  - Add random noise

- These methods can be ineffective
  - Low intensity perturbations may not be protective
  - High intensity perturbations destroy quality

A potential path forward

- An integrated system including
  - unrestricted access to (fully) synthetic data, coupled with
  - means for approved researchers to access the confidential data via remote access solutions, glued together by
  - verification servers that allow users to assess the quality of their inferences with the synthetic data so as to be more efficient with their use (if necessary) of the remote access to the confidential data.

Synergies of integrated system

- Use synthetic data to develop code, explore data, determine right questions to ask
- User saves time and resources if synthetic data good enough for her purpose (and so does steward!)
- If not, user can apply for special access to data
- This user has not wasted time
  - Exploration with synthetic data results in more efficient use of the real data
  - Explorations done offline free resources (cycles and staff) for final analyses

Synthetic data research of Duke/NISS node

- Methods for generating synthetic data
  - Synthesize individuals nested within households
  - Synthesize data that originally had faulty values
  - Synthesize data with high resolution geography
  - Synthesize (economic) data to sum to published totals
  - Replace swapping with partial synthesis in county-to-county migration flow data
- Applications
  - Synthetic version of Census of Manufactures
  - Synthetic version of Longitudinal Business Database

Synthetic categorical data nested within households

- Categorical data challenging to model well
  - What interactions to choose in log-linear models?
- Nesting within households adds complications
  - Household level and individual level variables
  - Within-household associations
  - Structural zeros abound

Synthetic categorical data nested within households

- Model data using two levels of latent classes
  - Each household is member of a data-estimated class (e.g., households of size 3 where everyone is same race)
  - Within household classes, each individual is a member of a data-estimated class (e.g., individuals who tend to be female HH heads of a particular race)
- Restrict support to feasible combinations of individuals
- Paper posted on arXiv.org

Results from ACS simulation

Disclosure risks: generally low

	Original	NDPMPM1
All same race		
$n_i - 2$	(.968, .976)	(.945, .958)
$n_i - 3$	(.952, .968)	(.910, .931)
$n_i - 4$	(.927, .947)	(.871, .900)
Spouse present	(.681, .699)	(.668, .695)
Spouse & white HHH	(.566, .586)	(.555, .579)
Spouse & black HHH	(.090, .102)	(.086, .101)
White couple	(.556, .576)	(.541, .566)
White couple own	(.496, .516)	(.470, .495)
Same race couple	(.659, .677)	(.633, .662)
Only mother w/ children	(.169, .183)	(.160, .179)
One parent w/ children	(.204, .220)	(.218, .243)
At least one child	(.494, .514)	(.497, .520)
At least one parent	(.022, .028)	(.027, .036)
At least one sibling	(.021, .027)	(.027, .034)
At least one grandchild	(.054, .064)	(.062, .076)
Three-generation family	(.061, .071)	(.073, .086)
Non-white couple own	(.081, .091)	(.057, .069)

Simultaneous edit-imputation and data synthesis

- Economic data (e.g., ASM) often reported with errors
  - Ratio of two variables falls outside plausible ranges
  - Balance equation does not hold
- Edit imputation for economic data
  - Change minimum number of fields to impute (MFI) to make plausible record
  - Blank and impute selected fields
- Potential shortcomings
  - Underestimate uncertainty
  - Error localization does not fully utilize information in data

Simultaneous edit-imputation and data synthesis

- Stochastic edit-imputation
  - Assume underlying data follow mixture of multivariate normal distributions, constrained to space of feasible values
  - Posit stochastic model for locations of fields to change
  - Posit measurement error model
  - Run MCMC to create multiple imputations
  - Paper to appear in *JASA*. R package to CRAN almost.

- Model readily adapted to data synthesizer
  - Run edit imputation routine to create completed dataset(s)
  - Run model on completed dataset(s) to create synthetic data

Simultaneous edit-imputation and data synthesis

- Empirical application on pre-edited data from two industries in the 2007 Census of Manufactures (can't share results yet – not approved by DRB)

- Similar distributions as the edited data
- Stochastic edit-imputation results in less attenuated correlations compared to MFI
- All values adhere to edit constraints
- Generally low risks

More information

- Duke/NISS NCRN node: sites.duke.edu/tcrn/
  - Papers on data confidentiality (as well as edit-imputation, missing data, data integration)
  - Software implementing (some of) the methods
- The NCRN network: ncrn.info