INFO 7470
Synthetic Data

John M. Abowd and Lars Vilhuber
May 9, 2016
Outline

• Theory and Methods for Synthetic Data
• LODES/OnTheMap Synthetic Data
• SIPP Synthetic Data
• LBD Synthetic Data
THEORY AND METHODS
Alternative Definitions

• In *Statistical Disclosure Limitation*: data published as samples from a predictive distribution estimated using all of the confidential data.

• In *Privacy-preserving Data Analysis*: data queries that have been produced by a formal privacy model then published in the same format as the underlying database.
Other Definitions

• Some official data providers: fake data guaranteed to replicate the database schema and therefore suitable for program development only.

• Many official statistical agencies: “synthetic estimates” are model-based, design-consistent replacements for direct survey estimates.
Synthetic Data Validation

- Estimation of statistical models using the synthetic data that have been subsequently estimated using the confidential data
- Confidential data estimates are released after applying statistical disclosure limitation to the model outputs
- Inference is performed using the confidential data estimates, correcting for the SDL applied to those estimates
Synthetic Data Verification

- Estimation of statistical models using the synthetic data
- Inferenced performed using the synthetic data, adjusting for the statistical disclosure limitation used to create the synthetic data
- Verification of the inference using the confidential data with statistical disclosure limitation applied to the verification statistic
Confidentiality Protection

• Traditional SDL: the protection is due to the absence of any actual respondent information in the synthetic data (or the presence of only limited respondent information for partially synthetic data)

• Formal privacy: the protection is due to the properties of the randomization algorithm used to protect the query answers from which the synthetic data were built
Analytical Validity

• Vs. the sample: are the inferences made using the synthetic data similar to the inferences made using the original confidential data
• Vs. the population: are the inferences made from the synthetic data generalizable to the target population without reference to the original confidential data
The Feedback Cycle

• The next slide is from the December 7, 2004 presentation that John Abowd gave at the Census Bureau introducing the synthetic data research program, supported by an NSF ITR grant, that led to the production of the OnTheMap, SIPP, and LBD synthetic data products

© John M. Abowd and Lars Vilhuber 2016, all rights reserved
The Research – Synthetic Data Feedback Cycle

Confidentiality Protection

Scientific Modeling

Data Synthesis

Analytic Validity
A Brief History


• Fully synthetic data (Rubin 1993): all records and values in the released data are sampled from a predictive joint distribution estimated from the confidential data

• Partially synthetic data (Little 1993): only some records/variables are synthesized; those that are synthetic are sampled from the same predictive joint distribution, conditional on the exact values also released on the record
LODES/ONTHEMIDAS SYNTHETIC DATA
Dirichlet-Multinomial Synthesizer

- $I$ origins
- Model each destination $d$ separately for each demographic segment (age, earnings, industry)
- Sample data $X$ tabulated into $n$
- Synthetic data tabulated into $m$
- Usually $m = n$, but not in the OTM application

\[
n = (n_1, \ldots, n_I), \quad n = \sum n_i
\]
\[
\alpha = (\alpha_1, \ldots, \alpha_I), \quad \alpha_0 = \sum \alpha_i = |\alpha|
\]
\[
\pi = (\pi_1, \ldots, \pi_I)
\]
\[
n \sim M(\pi, n)
\]
\[
\pi \sim D(\alpha), \text{ a priori}
\]
\[
\pi \sim D(\alpha + n), \text{ a posteriori}
\]
\[
m = (m_1, \ldots, m_I), \quad m = \sum m_i
\]
\[
m \sim M(\pi, m), \text{ a posteriori}
\]
Synthetic Data Model

• Likelihood of place of residence (index $i$) conditional on place of work (index $j$) and characteristics (index $k$):

$$p(n_{ijk} \mid \pi_{i|jk}) \propto \prod_{i=1}^{I} \pi_{i|jk}^{n_{ijk}}$$

• The resulting posterior for $\pi$ is Dirichlet with parameter $n_{jk} + \alpha_{jk}$ for each unique workplace and characteristic combination (age, earnings, industry).

• Synthesize residence counts by sampling from the posterior predictive distributions conditional on already protected (and published) destination employment counts, $m_{jk}$
Final Privacy Settings for OnTheMap and LODES V3-11

• Unadjusted $\varepsilon = 4.6$
• Probability of failure $\delta = 0.000001$
• Minimum retention probability $\text{min}_p = 0.025$
• Adjusted-$\varepsilon = \varepsilon + \min(\alpha)\ln 2 + \ln(1/\text{min}_p) = 8.9$
• Expected-$\varepsilon = (1-\delta) \text{ Adjusted-}\varepsilon + \delta 25$
• Jensen-Shannon and Mean Integrated Squared Error loss functions used to set tradeoffs
• Multinomial-Dirichlet Posterior sampled for every workplace block in the U.S. (about 1.4 million)
Analytical Validity Measures (JSD)

- The divergence between posterior (or synthetic data) and likelihood for a population is measured by the Jensen-Shannon Divergence (JSD) over a 29 point grid defined by the cross product of:
  - 8 commute distance categories (in miles: 0, (0-1), [1-4), [4-10), [10-25), [25-100), [100,500), [500+])
  - 5 commute direction categories (NW, NE, SW, SE, “N/A”)
- Measure aggregated over the stratifications
- Measure transformed so that it is a metric (square root JSD) and translated so that it is increasing in data quality
- \( D_{JSD} = 1 \) if identical; \( D_{JSD} = 0 \) if no overlap

\[
D_{JSD}(P, L) = 1 - \sqrt{0.5 \sum_i L(i) \log_2 \left( \frac{L(i)}{0.5P(i) + 0.5L(i)} \right)} + 0.5 \sum_i P(i) \log_2 \left( \frac{P(i)}{0.5P(i) + 0.5L(i)} \right)
\]
LODES Quality Measured by Jensen-Shannon Divergence

- Posterior-Likelihood
- Synthetic-Likelihood

Expected Adjusted Epsilon vs. 1-\sqrt{\text{JSD}}
LODES Quality Measured by Jensen-Shannon Divergence (zoomed)
LODES Quality Measured by Jensen-Shannon Divergence (small tracts)

- Posterior-Likelihood
- Synthetic-Likelihood

Expected Adjusted Epsilon

May 9, 2016

© John M. Abowd and Lars Vilhuber 2016, all rights reserved
SURVEY OF INCOME AND PROGRAM PARTICIPATION (SIPP) SYNTHETIC DATA
Survey of Income and Program Participation (SIPP)

• Goal of SIPP: accurate info about income and program participation of individuals and households and its principal determinants

• Information:
  – Cash and noncash income on a sub-annual basis.
  – Taxes, assets, liabilities
  – Participation in government transfer programs

http://www.census.gov/sipp/intro.html
Background

• In 2001, a new regulation authorized the Census Bureau and SSA to link SIPP and CPS data to SSA and IRS administrative data for research purposes
• Idea for a public use file was motivated by a desire to allow outside access to long administrative record histories of earnings and benefits linked to household demographic data
• These data allow detailed statistical and simulation study of retirement and disability programs
• Census Bureau, Social Security Administration, Internal Revenue Service, and Congressional Budget Office all participated in development
SSB Basic Methodology

- Experiment using “synthetic data”
- In fact: partially synthetic data with multiple imputation of missing items
- Partially synthetic data:
  - Some (at least one) variables are actual responses
  - Other variables are replaced by values sampled from the posterior predictive distribution for that record, conditional on all of the confidential data
History of the SSB

- 2003-2005: Creation, but not release, of three versions of the “SIPP/SSA/IRS-PUF” (SSB)
- 2006: Release to limited public access of SSB V4.2
  - Access to general public only at Cornell-hosted Virtual RDC (SSB server: restricted-access setup)
    - With promise of evaluation of Virtual RDC-run programs on internal Gold Standard
  - Ongoing SSA evaluation
  - Ongoing evaluation at Census (in RDC)
- 2010: Release of SSB V5 at Census and on the Virtual RDC (codebook: http://www.census.gov/sipp/SSB_Codebook.pdf)
  - Restructured to vastly improve analytical validity of SIPP variables
- 2013: Release of SSB V5.1 at Census and on the Virtual RDC (documentation in preparation)
  - First user-initiated variables
- 2015: Release of SSV V6.0
Common Structure to Multiple Imputation and Synthesis

• Hierarchical tree of variable relationships (parent-child relationship, accounting for structure)
• At each node, independent SRMI is used
  – Statistical model is estimated for each of the variables at the same level
  – Statistical models are estimated separately for groups of individuals
  – Then, a proper posterior predictive distribution is estimated
  – Given a PPD, each variable is imputed/synthesized, conditional on all values of all other variables for that record
• The next node is processed
MI and Synthesis

• Initial iterations for missing data imputation, keeping all observed values where available
• Final iteration is for data synthesis (replacing all observed values, see exceptions)
Multiple Imputation for Confidentiality Protection

- Denote confidential data by $Y$ and non-confidential data by $X$
- $Y = (Y_{mis}, Y_{obs})$ and $X$ has no missing data
- PPD: $p(Y_{mis} | Y_{obs}, X)$
- Complete data: $Y^{\ell}_{mis}$ from $p(Y_{mis} | Y_{obs}, X)$
  \[ D^{\ell} = (Y^{\ell}_{mis}, Y_{obs}, X) \]
- Synthetic data: $Y^{k,\ell}$ from $p(., | Y^{\ell}_{mis}, Y_{obs}, X)$
  \[ D^{k,\ell} = (Y^{k,\ell}, X) \]
- Major emphasis is to find a good estimate of the PPD
Mathematical Background

• Combining missing data and confidentiality protection using the same PPD
• Original application Kennickell (1997) was to the Survey of Consumer Finances
• SIPP application is based on Reiter (2004)
Completed Data Calculations

• Notation
  – Script \( \ell \) is index for missing data implicate
  – \( m \) is total number of missing data implicates

• Estimate from one completed implicate
  \[ q^\ell = q(D^\ell) \]

• Average of statistic across implicates
  \[ \bar{q}_m = \frac{1}{m} \sum_{\ell=1}^{m} q^\ell \]
Total Variance and Between Variance

• Total variance of average statistic

\[ T_m = \bar{u}_m + \left( 1 + \frac{1}{m} \right) b_m \]

• Variance of the statistic across implicates: between variance

\[ b_m = \sum_{\ell=1}^{m} \frac{(q^\ell - \bar{q}_m)(q^\ell - \bar{q}_m)^T}{m - 1} \]
Within Variance

• Variance of the statistic from each completed implicate

\[ u^{\ell} = u(D^{\ell}) \]

• Average variance of statistic: within variance

\[ \bar{u}_m = \sum_{\ell=1}^{m} \frac{u^{\ell}}{m} \]
Synthetic and Completed Data

- Notation
  - Script $\ell$ is index for missing data implicate
  - Script $k$ is index for synthetic data implicate
  - $m$ is total number of missing data implicates
  - $r$ is total number of synthetic implicates per missing data implicate

- Estimate from one synthetic implicate
  $$q^{k,\ell} = q\left(D^{k,\ell}\right)$$

- Average of statistic across synthetic implicates
  $$\bar{q}^{\ell} = \Sigma_{k=1}^{r} \frac{q^{k,\ell}}{r}$$
Grand Mean and Overall Variance

• Average of statistic across all implicates

\[
\bar{q}_M = \sum_{\ell=1}^{m} \sum_{k=1}^{r} \frac{q^{k,\ell}}{mr} = \sum_{\ell=1}^{m} \frac{\bar{q}^{\ell}}{m}
\]

• Total variance of average statistic

\[
T_M = \left(1 + \frac{1}{m}\right) B_M - \frac{b_M}{r} + \bar{u}_M
\]
Between Variances

- Variance of the statistic across missing data implicates: between \( m \) implicate variance

\[
B_M = \sum_{\ell=1}^{m} \frac{(\bar{q}^\ell - \bar{q}_M)(\bar{q}^\ell - \bar{q}_M)^T}{m - 1}
\]

- Variance of the statistic across synthetic data implicates: between \( r \) implicate variance

\[
b_M = \sum_{\ell=1}^{m} \sum_{k=1}^{r} \frac{(q^{k,\ell} - \bar{q}^\ell)(q^{k,\ell} - \bar{q}^\ell)^T}{m(r - 1)} = \sum_{\ell=1}^{m} \frac{b^\ell}{m}
\]
Formulae: Within Variances

• Variance of the statistic on each implicate
  \[ u^{k,\ell} = u(D^{k,\ell}) \]

• Average variance of statistic: within variance

\[ \bar{u}_M = \sum_{\ell=1}^{m} \sum_{k=1}^{r} \frac{u^{k,\ell}}{mr} = \sum_{\ell=1}^{m} \frac{\bar{u}^{\ell}}{m} \]

• Source: Reiter, Survey Methodology (2004): 235-42.
Example: Average AIME/AMW

• Average Indexed Monthly Earnings
• Estimate average on each of synthetic implicates
  – AvgAIME(1,1), AvgAIME(1,2), AvgAIME(1,3), AvgAIME(1,4),
  – AvgAIME(2,1), AvgAIME(2,2), AvgAIME(2,3), AvgAIME(2,4),
  – AvgAIME(3,1), AvgAIME(3,2), AvgAIME(3,3), AvgAIME(3,4),
  – AvgAIME(4,1), AvgAIME(4,2), AvgAIME(4,3), AvgAIME(4,4)
• Estimate mean for each set of synthetic implicates that correspond to one completed implicate
  – AvgAIMEAVG(1), AvgAIMEAVG(2), AvgAIMEAVG(3), AvgAIMEAVG(4)
• Estimate grand mean of all implicates
  – AvgAIMEGRANDAVG
Example (cont.)

• Between $m$ implicate variance

$$B_M = \sum_{i=1}^{4} \frac{(\text{avgAIME}_{avg}^{(i)} - \text{avgAIME}_{\text{rand avg}})(\text{avgAIME}_{avg}^{(i)} - \text{avgAIME}_{\text{rand avg}})}{3}.$$  

• Between $r$ implicate variance

$$b_M = \sum_{i=1}^{4} \sum_{k=1}^{4} \frac{(\text{avgAIME}_{(i,k)}^{(i)} - \text{avgAIME}_{avg}^{(i)})(\text{avgAIME}_{(i,k)}^{(i)} - \text{avgAIME}_{avg}^{(i)})}{4(3)}.$$
Example (cont.)

• Variance of mean from each implicate
  – \( \text{VAR}[\text{AvgAIME}^{(1,1)}] \), \( \text{VAR}[\text{AvgAIME}^{(1,2)}] \), \( \text{VAR}[\text{AvgAIME}^{(1,3)}] \), \( \text{VAR}[\text{AvgAIME}^{(1,4)}] \)
  – \( \text{VAR}[\text{AvgAIME}^{(2,1)}] \), \( \text{VAR}[\text{AvgAIME}^{(2,2)}] \), \( \text{VAR}[\text{AvgAIME}^{(2,3)}] \), \( \text{VAR}[\text{AvgAIME}^{(2,4)}] \)
  – \( \text{VAR}[\text{AvgAIME}^{(3,1)}] \), \( \text{VAR}[\text{AvgAIME}^{(3,2)}] \), \( \text{VAR}[\text{AvgAIME}^{(3,3)}] \), \( \text{VAR}[\text{AvgAIME}^{(3,4)}] \)
  – \( \text{VAR}[\text{AvgAIME}^{(4,1)}] \), \( \text{VAR}[\text{AvgAIME}^{(4,2)}] \), \( \text{VAR}[\text{AvgAIME}^{(4,3)}] \), \( \text{VAR}[\text{AvgAIME}^{(4,4)}] \)

• Within variance

\[
\bar{u}_M = \sum_{i=1}^{4} \sum_{k=1}^{4} \frac{\text{Var}[\text{avgAIME}^{(i,k)}]}{4(4)}
\]
Example (cont.)

- Total Variance

\[ T_M = \left( 1 + \frac{1}{4} \right) B_M - \frac{b_M}{4} + \bar{u}_M. \]

- Use AvgAIMEGRANDAVG and Total Variance to calculate confidence intervals and compare to estimate from completed data
Results: Average AIME

<table>
<thead>
<tr>
<th></th>
<th>AVG STAT</th>
<th>Total VAR</th>
<th>Betw. M Var</th>
<th>Betw. R Var</th>
<th>Betw. Var</th>
<th>Within Var</th>
<th>confidence interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>synthetic</td>
<td>1094.2</td>
<td>91.8</td>
<td>59.3</td>
<td>13.3</td>
<td>21.1</td>
<td></td>
<td>1074.5 1113.9</td>
</tr>
<tr>
<td>completed</td>
<td>1142.5</td>
<td>52.8</td>
<td></td>
<td></td>
<td>23.4</td>
<td>23.7</td>
<td>1129.3 1155.7</td>
</tr>
</tbody>
</table>

*All individuals with TOB_2000=1

Average of AIME (Average Indexed Monthly Earnings)/AMW(Average Monthly Wage)
Methods for Estimating the PPD

• Sequential Regression Multivariate Imputation (SRMI) is a parametric method where PPD is defined as

\[ p(\tilde{Y} | Y_{obs}, X_{obs}) = \int p(\tilde{Y} | Y_{obs}, X_{obs}, \theta) p(\theta | Y_{obs}, X_{obs}) d\theta \]

• The BB is a non-parametric method of taking draws from the posterior predictive distribution of a group of variables that allows for uncertainty in the sample CDF

• We use BB for a few groups of variables with particularly complex relationships and use SRMI for all other variables
SRMI Method Details

• Assume a joint density $p(Y, X, \theta)$ that defines parametric relationships between all observed variables.

• Approximate the joint density by a sequence of conditional densities defined by generalized linear models.

• Same process for completing and synthesizing data

• Synthetic values of some $y_k \in \bar{Y}$ are draws from:

$$
p_k(\tilde{y}_k | Y^m, X^m) = \int p_k(\tilde{y}_k | Y^m_k, X^m, \theta)p_k(\theta | Y^m, X^m)d\theta
$$

where $Y^m, X^m$ are completed data, and densities $p_k$ are defined by an appropriate generalized linear model and prior
SRMI Details: KDE Transforms

• The SRMI models for continuous variables assume that they are conditionally normal.
• This assumption is relaxed by performing a KDE-based transform of groups of related variables.
• All variables in the group are transformed to normality, then the PPD is estimated.
• The sampled values from PPD are inverse transformed back to the original distribution using the inverse cumulative distribution.
SRMI Example: Synthesizing Date of Birth

• Divide individuals into homogeneous groups using stratification variables
  – example: male, black, age categories, education categories, marital status
  – example: decile of lifetime earnings distribution, decile of lifetime years worked distribution, worked previous year, worked current year

• For each group, estimate an independent linear regression of date of birth on other variables (not used for stratification) that are strongly related
SRMI Example: Synthesizing Date of Birth

• Synthetic date of birth is a random variable
• Before analysis, it is transformed to normal using the KDE-based procedure
• Distribution has two sources of variation:
  – variation in error term in regression model
  – variation in estimated parameters: $\beta$’s and $\sigma^2$
• Synthetic values are draws from this distribution
• Synthetic values are inverse transformed back to the original distribution using the inverse cumulative distribution
SYNTHETIC LONGITUDINAL BUSINESS DATABASE
Generic Structure

• Gold standard: given by internal LBD (already completed)

• Partially synthetic:
  – Unsynthesized:
    • County (but not released!) [x1]
    • SIC [x2]
  – Synthesized
    • Birth [y1] and death [y2] year:
    • Multi-unit status [y3]
    • Employment (March 12) [y4]
    • Payroll [y5]
Synthesis: General Approach

- \( Y = [y_1 | y_2 | y_3 | y_4 | y_5] \)
- \( X = [x_1 | x_2] \)
- Generate joint distribution of \( Y | X \) by sampling from conditionals
  - \( f(y_1, y_2, y_3 | X) = f(y_1 | X) \cdot f(y_2 | y_1, X) \cdot f(y_3 | y_1, y_2, X) \)
- Use SIC as “by group”
General Approach to Synthesis

• Drawing from $f(y_k|X,y_1,\ldots,y_{k-1})$
  – Fit model using observed data
  – Draw new values of parameters from posterior distributions
  – Use new parameters to predict $y_k$ from $X$ and synthetic values of $y_1,\ldots,y_{k-1}$
The Sequential Regression Multivariate Imputation (SRMI) Approach

• Calendar:
  – Step 1: Impute $y_1 \mid X$
  – Step 2: Impute $y_2 \mid [y_1 \mid f(X)]$
    • Where $f(X)$ uses state $[x_1']$ instead of county $[x_1]$
• Type of firm
  – Step 3: Impute $y_3 \mid [y_1 \mid y_2 \mid X]$
• Characteristics
  – Step 4: Impute $y_4(t) \mid [y_1 \mid y_2 \mid y_3 \mid y_4(t-1) \mid x_2]$
  – Step 5: Impute $y_5(t) \mid [y_1 \mid y_2 \mid y_3 \mid y_4(t) \mid y_5(t-1) \mid x_2]$
Analytical Validity Tests

- Compare observed data and synthetic data for whole LBD
- Job creation and destruction
- Employment volatility
- Gross employment levels
Job Creation from Births and Expansions: LBD and Implicates by Year

Thousands

Year

LBD
Implicate 1
Implicate 2
Implicate (Mean)
Net Job Creation Rates: LBD v Implicates

Net Job Creation LBD
Net Job Creation Implicate 1
Net Job Creation Implicate 2

© John M. Abowd and Lars Vilhuber 2016, all rights reserved
Figure 3: Share of Employment by Industry Sector and Year, 1976-2000
Figure 2: Share of Establishments by Industry Sector and Year, 1976-2000.
Figure 4: Share of Payroll by Industry Sector and Year, 1976-2000
\[ EMP_i = \alpha + \beta EMP_{i-1} + \delta PAY_i + \theta IND_i + \psi STATE_i + \vartheta AGE_i + \gamma MU_i + \epsilon \]

**Figure 11: Regression Coefficients, LBD vs Synthetic**