Recent Advances In Confidentiality Protection

John M. Abowd
April 2005

Synthetic Data

- There are many kinds of synthetic data
- The examples in this handout use a single common definition:
  - Synthetic data is a sample composed of draws from the posterior predictive distribution of the confidential data, given some conventionally disclosure-controlled data
- We use the Sequential Regression Multivariate Imputation algorithm (Ragunathan, et al. 2001) as implemented in Abowd and Woodcock (2001)

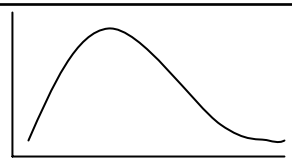
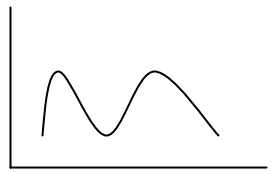
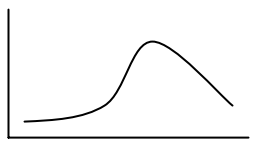
Synthetic Cross-Sectional Data

- Review the Virginia Origin-Destination matrix documentation.
- A non-confidential version is included in today's materials.

Example I

- Universe of employees at a point in time
- Variables:
  - Person Identifier (artificial)
  - Sex
  - County of Residence
  - Date of Birth (year:month:day)
- In this example sex and county of residence are disclosure controlled using conventional methods

Example I: Process

	County A	County B	…
Male	 Date of Birth Distribution	 Date of Birth Distribution	
Female	 Date of Birth Distribution	 Date of Birth Distribution	

Example I: Process

- Form the table: Sex x County of Residence
- Call each cell size n(i,j)
- For cell create n(i,j) records with the appropriate values of Sex and County of Residence
- For each record, treat Date of Birth as an item missing value, given Sex and County of Residence
- Estimate the posterior predictive distribution for Date of Birth, given Sex and County of Residence
- Sample n(i,j) values from this posterior predictive distribution, one for each synthetic record
- Repeat M times for each cell

Example I: Result

- M synthetic samples from the universe of employees
- Record:
  - Person Identifier (artificial)
  - Sex (not drawn from posterior)
  - County of Residence (not drawn from posterior)
  - Date of Birth (imputed for every record)

Example I: Why It Works

- The table Sex x County of Residence has been protected by conventional methods (minimum cell size, collapsing of small cells, etc.)
- The synthetic data set exactly reproduces the Sex x County of Residence table by construction
- Every record in the synthetic data set contains an imputed Date of Birth
  - No record corresponds to a person in the universe
  - The univariate distribution of Date of Birth within Sex x County of Residence cell can be reproduced to an arbitrary level of precision

Example I: Why It Isn't Trivial

- These synthetic data can only be produced inside the firewall
  - The DRB would be justifiably reluctant to release a list of exact birth dates for each cell of the Sex x County of Residence table
- These synthetic data are inference valid
  - The synthetic data contains exactly the same statistical information as the confidential data
- We can measure the effect of disclosure protection on data quality
  - The multiple synthetic data implicates are not identical so the analyst can use the between implicate variation to measure the extent to which confidentiality protection made the inferences less precise

Example II

- Universe of all business establishments observed over multiple consecutive time periods (from begin date to end date, quarterly)
- Record:
  - Business Identifier (artificial)
  - Establishment Identifier (artificial)
  - Industrial Classification (SIC Division)
  - County of Operation
  - Establishment start date (year:quarter)
  - Establishment end date (year:quarter)
  - Employment (as of the 12th day of the first month) for every quarter, including zeros
- In this example county of operation and SIC division are disclosure controlled using conventional methods

Example II: Process

	County A	County B	….
Industry 1	EIN_A1_1 EIN_A1_2 ….	EIN_B1_1 EIN_B1_2	
Industry 2	EIN_A2_1 EIN_A2_2	EIN_B2_1 EIN_B2_2	
Industry 3	EIN_A3_1 EIN_A3_2	EIN_B3_1 EIN_B3_2	

Example II: Process

- At a reference date form the table County of Operation x SIC Division (example done assuming the earliest date is the reference date)
- Call the cell size n(i,j)
- For each cell create n(i,j) records with the appropriate values of County of Operation and SIC Division
- For each record, select one of the longitudinal histories for establishments alive at the reference date or born within the allowable window for birth (four quarters) (Note that actual responses are used here)
- Estimate the posterior predictive distribution for start date, end date, and employment given the longitudinal histories, County of Operation, and SIC Division
- Sample n(i,j) values from this posterior predictive distribution, one for each synthetic record
- Move forward one quarter, repeat the process for establishments that are now in the birth window
- Repeat M times for each cell

Example II: Result

- M synthetic longitudinal establishment samples from the universe of establishments ever operating over the period
- Record:
  - Business Identifier (suppressed)
  - Establishment Identifier (artificial)
  - Industrial Classification (SIC Division, not sampled from posterior)
  - County of Operation (not sampled from posterior)
  - Establishment start date (year:quarter)
  - Establishment end date (year:quarter)
  - Employment (as of the 12th day of the first month) for every quarter, including zeros

Example II: Why It Works

- It's not a disclosure - the year x quarter x SIC Division x County of Operation table (similar to county business patterns)
- Company structure is not disclosed – (the Business Identifier is suppressed)
- Establishment births and deaths are synthetic - because establishment start date and establishment end date are imputed for every longitudinal history all of the histories do not correspond to the birth or death of any actual establishment
- The Employment levels are synthetic - because employment as of the 12th day of the first month of the quarter is imputed for every quarter of every history
- BUT the synthetic histories are very accurate statistical summaries of the underlying confidential data: moments, covariation, and unusual outcomes (outliers) all occur with conditional probabilities comparable to the underlying data

Example II: Why It Works

- Imputing the establishment characteristics one at a time (given the values of all the other characteristics) is an application of Rubin's multiple imputation missing data method
- The sequential regression multivariate imputation (Ragunathan, et al. 1998) provides a very accurate method for estimating and sampling from the posterior predictive distribution when there are many variables
- The M implicates provide that analyst with information about the sensitivity of the analysis to the confidentiality protection system

Example II: Pitfalls

- The synthetic establishments are created by using a confidential longitudinal history to condition the posterior predictive distribution used to generate the M synthetic samples (we call these “source records” in the synthetic data creation process)
- There is, consequently, some source-record re-identification risk
  - Mitigated by the imputation of start and end dates
  - Existence of the establishment is a matching variable not a blocking variable in the re-identification record linking
  - Further mitigated by the imputation of all the employment levels, which are also matching variables not blocking variables
  - Can be detected and controlled during the creation of the synthetic data
  - Bootstrap sampling can be used instead of direct selection of the source records from the universe file

Example II: Why It Isn't Trivial

- As in the first example, the implicates are produced by sampling from the posterior predictive distribution, given the actual values of the confidential data. Again, this is an "inside the firewall" method. The implicate for quarter 3 employment, for example, is produced conditional on the observed, confidential values of employment in all other quarters
- Because every data item except SIC Division and County of Operation are imputed from a sophisticated missing data model, the analyst will encounter many data anomalies, a feature of the confidential data that the multiple imputation method reliably reproduces
- This is not the same as releasing the sufficient statistics for an elaborate simulation model of the establishments because the simulation of a particular variable, say employment in quarter 3, cannot be conditioned on the actual values of employment in all the other quarters

Example III

- Universe of employment spells (jobs) observed over multiple consecutive time periods from begin date to end date, quarterly
- Record:
  - Person Identifier (artificial)
  - Business Identifier (artificial)
  - Establishment Identifier (artificial, missing for multi-units)
  - Industrial Classification (SIC Division, from example II)
  - Employment as of 12th day of quarter (from example II)
  - County of Operation (from example II)
  - Employment start date (year:quarter)
  - Employment end date (year:quarter)
  - Quarterly earnings, for every quarter including zeros
- In this example county of operation and SIC division are carried over from the establishment data file and are disclosure controlled by conventional methods; establishment employment is carried over from the establishment file but has been disclosure controlled via synthetic data

Example III: Process A

- There is no information on the establishment identifier for multi-units
  - The establishment identifier for multi-units is multiply imputed based on a probability model developed from the entire LEHD infrastructure
  - County of operation and SIC Division are, thus, imputed for every employment spell from at a multi-unit
- Some individuals never work in a multi-unit
  - Create artificial multi-units within County of Operation and SIC Division
  - Use the same probability model as for real multi-units to impute the establishment identifier for these artificial multi-units
- Verify that every individual has at least one employment spell at a multi-unit or an artificial multi-unit

Example III: Process B

- Each employment spell is a source record for the synthetic data
- Estimate the posterior predictive distribution for employment start date, employment end date, and quarterly earnings given the longitudinal employment histories, SIC Division, County of Operation, and establishment employment.
- For each source record, create one synthetic record by drawing from the posterior predictive distribution of employment start date, employment end date, and quarterly earnings given all the other confidential values.
- Repeat M times

Example III: Result

- M synthetic longitudinal establishment samples from the universe of establishments ever operating over the period
- Record:
  - Person Identifier (artificial)
  - Business Identifier (suppressed)
  - Establishment Identifier (from posterior for multi-units and artificial multi-units)
  - Employment start date (year:quarter)
  - Employment end date (year:quarter)
  - Quarterly earnings, for every quarter including zeros

Example III: Why It Works

- For each individual, the complete employment history (all employment spells) is synthetic
- Although we started with a source record that was confidential, the establishment identifier (for multi-units and for artificial multi-units) is drawn from the predictive posterior. Every individual has at least one imputed establishment identifier
- As with the employer example II, the business identifier is suppressed to avoid disclosing company structure information
- Because the start date, end date and quarterly earnings are all imputed, there is no disclosure of the employment spell characteristics. In each implicate the employment spells have different start and end dates and different values of quarterly earnings, all consistent with the posterior predictive distribution

Example III: Why It Works

- The essential statistical information in the employment history is the association of the individual with a sequence of employers
- By using a confidential history as the source record for a complete set of employment spells for the individual, virtually all of the statistical information is preserved even though all of the data records are synthetic

Example III: Pitfalls

- For an individual the exact pattern of start dates, end dates, and number of simultaneous employers in the confidential data is essentially unique
  - Hence, there is some re-identification risk that must be assessed and controlled
  - Imputing start and end dates makes each implicate of the history unique and eliminates re-identification across implicates
  - Insuring that every individual's set of employment histories contains at least one multi-unit or artificial multi-unit ensures that the synthetic histories never correspond to a real individual (place of work is imputed for at least one history for each individual)
- Employment counts at establishments from the synthetic histories do not match the employment counts from the establishment micro data
  - This is also a feature of the underlying confidential data

Example III: Why It Isn't Trivial

- The DRB would be justifiably reluctant to release a sample of the employment histories because of the uniqueness of the pattern of employment and number of employers per period. The re-identification risk (by an individual with access to the underlying confidential micro data, provided by a non-Census source) is unacceptably high
- Each synthetic employment history corresponds to a draw from a posterior predictive distribution that is conditioned on confidential micro data for that spell (the complete earnings history) and for that employer (the establishment characteristics)
- Releasing the sufficient statistics for the posterior predictive distribution would not be sufficient to reproduce the statistical features of the confidential data
- The multiple implicates permit the researcher to assess the sensitivity of the models to the confidentiality protection of the microdata

Example IV

- The LEHD Infrastructure Files
  - Individual
  - Employer (establishment)
  - Employment history (jobs)
- Record structure (as in Examples I-III)
- Sex, County of Residence, County of Operation (work), and Employer SIC Division have been disclosure controlled by conventional methods

Example IV: Process A

- The individual frame is the universe of persons who ever worked in a given state from begin date to end date
- The establishment frame is the universe of establishments that operated in a given state from begin date to end date
- The employment history frame is the universe of employment spells (person-business) in a given state from begin date to end date

Example IV: Process B

- Variables on the individual file are those listed in example I, county of residence and sex are not drawn from the posterior distribution
- Variables on the establishment file are those listed in example II, county of operation and SIC division are not drawn from the posterior distribution
- Variables on the employment history file are those listed in example III, at least one employment history for each individual comes from a multi-unit or artificial multi-unit
- The person, business and establishment identifiers are common to all three files

Example IV: Process C

- Estimate the posterior predictive distribution for date of birth conditional on all the information in that persons individual file, employment histories, and establishment records (including past, present and future values of longitudinal variables in the employment history and establishment files)
- Estimate the posterior predictive distribution for establishment birth date, establishment death date, and employment as of the 12th day of the quarter conditional on the history information in the establishment file, summary information from all of the employee histories (distribution summaries like means and percentiles including past, current and future), and summary information about the demographic characteristics of the employees from the individual file
- Estimate the posterior predictive distribution for employment start date, employment end date, and quarterly earnings conditional on all the variables in the individual's employment history (past, current, and future employment periods of the source record spell; past, current, and future data from other spells for the same individual; past, current, and future data from all employers)

Example IV: Process D

- Using each record in the confidential files as a source record, create M implicates of all the variables for which posterior predictive distributions were estimated, conditioning on the complete confidential record for the individuals and establishments
- For each of the M implicate files generate a unique set of artificial individual and establishment identifiers so that individuals and establishments can be linked within an implicate but not across implicates

Process IV: Result

- Individual record: as in example I
- Establishment record: as in example II
- Employment history record: as in example III

Example IV: Why It Works

- The posterior predictive distributions estimated from the fully linked files summarize information about all three universes using the information connecting employers and employees
- The essential statistical information in the integrated data is the connection between employers and employees
- The synthetic data protects the confidentiality of the observed connection between employers and employees by generating a synthetic individual connected to a synthetic establishment via a synthetic employment history
- None of the pieces of the synthetic data correspond to confidential entities: neither the identity of the respondents nor the actual attributes of the respondents appear in the synthetic data sets
- The resulting data are inference valid because they have been constructed from a reliable estimate of the posterior predictive distribution

Example IV: Pitfalls

- Because the source records are the actual confidential data records, the re-identification risk outlined in examples II and III applies to the integrated example
  - The re-identification risk can be assessed and controlled
  - All variables implicit in the creation of linked, longitudinal histories (individual and establishment) have been imputed
  - For each individual, at least one identifier linking that individual to an employing establishment has been imputed

Example IV: Why It Isn't Trivial

- Because every synthetic record contains variables that have been imputed conditional on information in many of the actual confidential records of the individual, establishment and work history files, this type of synthetic data can only be produced inside the firewall
- Release of all the sufficient statistics for all of the posterior predictive distributions used to create these synthetic data would not be equivalent and would result in synthetic data that are analytically less useful

Formulae

- Review Abowd Primer
- Included with today's materials