

Discussion: The Promise & Peril of Synthetic & Integrated Data

Alexander W Blocker

Harvard University
Department of Statistics

ICSA 2010
21 June, 2010

Outline

- 1 Modeling, inference, & congeniality
- 2 Defining & evaluating privacy
- 3 Closing remarks

What is congeniality?

What is congeniality?

- Compatibility of imputer's and analyst's models

What is congeniality?

- Compatibility of imputer's and analyst's models
- In Bayesian terms, equivalent posterior predictive distributions:

$$f(Y_{mis}|X, Y_{obs}, I) = g(Y_{mis}|X, Y_{obs}, I, A)$$

where f is analyst's model, g is imputer's model, and A corresponds to imputer's additional information

What is congeniality?

- Compatibility of imputer's and analyst's models
- In Bayesian terms, equivalent posterior predictive distributions:

$$f(Y_{mis}|X, Y_{obs}, I) = g(Y_{mis}|X, Y_{obs}, I, A)$$

where f is analyst's model, g is imputer's model, and A corresponds to imputer's additional information

- More general definition and discussion in Meng (1994)

Why does it matter?

Why does it matter?

- If models are congenial, multiple imputation combining rules provide valid inferences

Why does it matter?

- If models are congenial, multiple imputation combining rules provide valid inferences
- Without congeniality, variances are not accurate

Why does it matter?

- If models are congenial, multiple imputation combining rules provide valid inferences
- Without congeniality, variances are not accurate
- Point estimates suffer to varying degrees

Why does it matter?

- If models are congenial, multiple imputation combining rules provide valid inferences
- Without congeniality, variances are not accurate
- Point estimates suffer to varying degrees
- Impact depends upon fraction of missing information and nature of uncongeniality

Considerations with sythetic data

Considerations with sythetic data

- Extreme case with all observations missing (for fully and some partially sythetic data)

Considerations with sythetic data

- Extreme case with all observations missing (for fully and some partially synthetic data)
- Greater potential for congeniality issues to affect analysis

Considerations with sythetic data

- Extreme case with all observations missing (for fully and some partially synthetic data)
- Greater potential for congeniality issues to affect analysis
- Not possible for users to perform their own imputation

Considerations with sythetic data

- Extreme case with all observations missing (for fully and some partially synthetic data)
- Greater potential for congeniality issues to affect analysis
- Not possible for users to perform their own imputation
- Imputer's choices have greater overall effect on analysis

When should we worry?

When should we worry?

- With missing data imputation, typically believe statistical agency understands nonresponse issues best; hence, best position to handle missing data modeling

When should we worry?

- With missing data imputation, typically believe statistical agency understands nonresponse issues best; hence, best position to handle missing data modeling
- Similar situation with synthetic data; agency should have best available knowledge and resources to properly synthesize

When should we worry?

- With missing data imputation, typically believe statistical agency understands nonresponse issues best; hence, best position to handle missing data modeling
- Similar situation with synthetic data; agency should have best available knowledge and resources to properly synthesize
 - Even with issues, sythetic data provides framework for dealing with disclosure limitation

When should we worry?

- With missing data imputation, typically believe statistical agency understands nonresponse issues best; hence, best position to handle missing data modeling
- Similar situation with synthetic data; agency should have best available knowledge and resources to properly synthesize
 - Even with issues, sythetic data provides framework for dealing with disclosure limitation
 - Easier to understand & handle modeling issues than issues from less principled methods (noise infusion, swapping, etc.)

When should we worry?

- With missing data imputation, typically believe statistical agency understands nonresponse issues best; hence, best position to handle missing data modeling
- Similar situation with synthetic data; agency should have best available knowledge and resources to properly synthesize
 - Even with issues, sythetic data provides framework for dealing with disclosure limitation
 - Easier to understand & handle modeling issues than issues from less principled methods (noise infusion, swapping, etc.)
- Largest concern is repurposing of sythetic & imputed data

When should we worry?

- With missing data imputation, typically believe statistical agency understands nonresponse issues best; hence, best position to handle missing data modeling
- Similar situation with synthetic data; agency should have best available knowledge and resources to properly synthesize
 - Even with issues, sythetic data provides framework for dealing with disclosure limitation
 - Easier to understand & handle modeling issues than issues from less principled methods (noise infusion, swapping, etc.)
- Largest concern is repurposing of sythetic & imputed data
 - Models must be carefully reevaluated for new applications

When should we worry?

- With missing data imputation, typically believe statistical agency understands nonresponse issues best; hence, best position to handle missing data modeling
- Similar situation with synthetic data; agency should have best available knowledge and resources to properly synthesize
 - Even with issues, sythetic data provides framework for dealing with disclosure limitation
 - Easier to understand & handle modeling issues than issues from less principled methods (noise infusion, swapping, etc.)
- Largest concern is repurposing of sythetic & imputed data
 - Models must be carefully reevaluated for new applications
 - Formerly innocuous assumptions can become major issues

Types of disclosure

Types of disclosure

Identity Reidentification of subject; major issue for microdata

Types of disclosure

- Identity** Reidentification of subject; major issue for microdata
- Attribute** Reveal or provide accurate value of confidential value; main issue for magnitude data

Types of disclosure

- Identity** Reidentification of subject; major issue for microdata
- Attribute** Reveal or provide accurate value of confidential value; main issue for magnitude data
- Inferential** Attributes predictable from statistical properties of data

Fully vs. partially synthetic data

Fully vs. partially synthetic data

- Fully synthetic data only has inferential disclosure risk

Fully vs. partially synthetic data

- Fully synthetic data only has inferential disclosure risk
 - Never observe sensitive data for released units

Fully vs. partially synthetic data

- Fully synthetic data only has inferential disclosure risk
 - Never observe sensitive data for released units
 - Does not help adversary that knows true population model

Fully vs. partially synthetic data

- Fully synthetic data only has inferential disclosure risk
 - Never observe sensitive data for released units
 - Does not help adversary that knows true population model
- Partially synthetic data can have other risks

Fully vs. partially synthetic data

- Fully synthetic data only has inferential disclosure risk
 - Never observe sensitive data for released units
 - Does not help adversary that knows true population model
- Partially synthetic data can have other risks
 - Reidentification and attribute disclosure are possible

Fully vs. partially synthetic data

- Fully synthetic data only has inferential disclosure risk
 - Never observe sensitive data for released units
 - Does not help adversary that knows true population model
- Partially synthetic data can have other risks
 - Reidentification and attribute disclosure are possible
 - Larger issue in sparse settings

Fully vs. partially synthetic data

- Fully synthetic data only has inferential disclosure risk
 - Never observe sensitive data for released units
 - Does not help adversary that knows true population model
- Partially synthetic data can have other risks
 - Reidentification and attribute disclosure are possible
 - Larger issue in sparse settings
 - Need for more careful privacy analysis

Differential privacy

Differential privacy

Differential Privacy Suppose we draw datasets Z from $P(Z|X, Y)$.

Our method is ϵ -differentially private if, for all (Z_1, Z_2) differing by at most one row,

$$\log\left(\frac{P(Z_1|X, Y)}{P(Z_2|X, Y)}\right) \leq \epsilon$$

Differential privacy

Differential Privacy Suppose we draw datasets Z from $P(Z|X, Y)$.

Our method is ϵ -differentially private if, for all (Z_1, Z_2) differing by at most one row,

$$\log\left(\frac{P(Z_1|X, Y)}{P(Z_2|X, Y)}\right) \leq \epsilon$$

- From Dwork (2006); major area of research in computer science

Differential privacy

Differential Privacy Suppose we draw datasets Z from $P(Z|X, Y)$.

Our method is ϵ -differentially private if, for all (Z_1, Z_2) differing by at most one row,

$$\log\left(\frac{P(Z_1|X, Y)}{P(Z_2|X, Y)}\right) \leq \epsilon$$

- From Dwork (2006); major area of research in computer science
- Property of mechanism, not particular released dataset

Differential privacy

Differential Privacy Suppose we draw datasets Z from $P(Z|X, Y)$.

Our method is ϵ -differentially private if, for all (Z_1, Z_2) differing by at most one row,

$$\log\left(\frac{P(Z_1|X, Y)}{P(Z_2|X, Y)}\right) \leq \epsilon$$

- From Dwork (2006); major area of research in computer science
- Property of mechanism, not particular released dataset
- Insensitive to inferential disclosure; idea is that given individual does have large effect on released data

Formal privacy for synthetic data

Formal privacy for synthetic data

- Desire for strong guarantees of privacy from partially synthetic data

Formal privacy for synthetic data

- Desire for strong guarantees of privacy from partially synthetic data
- Integration of synthetic data and differential privacy by OnTheMap

Formal privacy for synthetic data

- Desire for strong guarantees of privacy from partially synthetic data
- Integration of synthetic data and differential privacy by OnTheMap
- A posteriori identification probabilities from Reiter (2005)

Formal privacy for synthetic data

- Desire for strong guarantees of privacy from partially synthetic data
- Integration of synthetic data and differential privacy by OnTheMap
- A posteriori identification probabilities from Reiter (2005)
- Both types of evaluations are useful, but all need to account for prior information of adversaries

Formal privacy for synthetic data

- Desire for strong guarantees of privacy from partially synthetic data
- Integration of synthetic data and differential privacy by OnTheMap
- A posteriori identification probabilities from Reiter (2005)
- Both types of evaluations are useful, but all need to account for prior information of adversaries
 - Automatic with differential privacy, can be difficult with other approaches

Formal privacy for synthetic data

- Desire for strong guarantees of privacy from partially synthetic data
- Integration of synthetic data and differential privacy by OnTheMap
- A posteriori identification probabilities from Reiter (2005)
- Both types of evaluations are useful, but all need to account for prior information of adversaries
 - Automatic with differential privacy, can be difficult with other approaches
- Handling of repeated, longitudinal releases is major area for further developments

Administrative records: the good, the bad, and the ugly

Administrative records: the good, the bad, and the ugly

- Significant challenges in adapting administrative data to statistical use

Administrative records: the good, the bad, and the ugly

- Significant challenges in adapting administrative data to statistical use
- Information not collected or designed for further analyses

Administrative records: the good, the bad, and the ugly

- Significant challenges in adapting administrative data to statistical use
- Information not collected or designed for further analyses
- High coverage and detail allow new types of analysis, but require new tools

Administrative records: the good, the bad, and the ugly

- Significant challenges in adapting administrative data to statistical use
- Information not collected or designed for further analyses
- High coverage and detail allow new types of analysis, but require new tools
 - Increased need to manage privacy concerns

Administrative records: the good, the bad, and the ugly

- Significant challenges in adapting administrative data to statistical use
- Information not collected or designed for further analyses
- High coverage and detail allow new types of analysis, but require new tools
 - Increased need to manage privacy concerns
 - Questions of data quality and validity for different analyses

Challenges & future work

Challenges & future work

- Developing methods to handle non-random sampling; combining large-scale records with small random samples

Challenges & future work

- Developing methods to handle non-random sampling; combining large-scale records with small random samples
- Allowing for exploratory analysis and model-building in protected data

Challenges & future work

- Developing methods to handle non-random sampling; combining large-scale records with small random samples
- Allowing for exploratory analysis and model-building in protected data
- Measuring the quality of complex integrated data