SPONSORED BY THE
Federal Ministry of Education and Research

Data Confidentiality and Statistical Disclosure Control

Problems with traditional approaches and alternatives based on synthetic data

SURV 735

2 credit/4ECTS

Spring 2017

Instructor

Jörg Drechsler, joerg.drechsler@iab.de

Short Course Description

This course will provide a gentle introduction to statistical disclosure control with a focus on generating synthetic data for maintaining the confidentiality of the survey respondents. The first part of the course will introduce several traditional approaches for data protection that are widely used at statistical agencies. Some limitations of these approaches will also be discussed. The second part of the course will introduce synthetic data as a possible alternative. This part of the course will discuss different approaches to generating synthetic datasets in detail. Possible modeling strategies and analytical validity evaluations will be assessed and potential measures to quantify the remaining risk of disclosure will be presented. To provide the participants with hands on experience, all steps will be illustrated using simulated and real data examples in R.

Course and Learning Objectives

By the end of the course, students will…

- know which measures are typically taken by statistical agencies to guarantee confidentiality for the survey respondents if data are disseminated to the public.
- be aware of potential limitations of these measures.
- have a practical understanding of the concept of synthetic data.
- be able to judge in which situations the approach could be useful.
- know how to generate synthetic data from their own data.
- have a number of tools available to evaluate the analytical validity of the synthetic datasets.
- know how to assess the disclosure risk of the generated data.

Prerequisites

The students should be familiar with the statistical software R. Some background regarding general linear modelling is expected. Familiarity with the concept of Bayesian statistics is helpful but not required.

Class Structure and Course Concept

This is an online course using a flipped classroom design. It covers the same material and content as an on-site course but runs differently. In this course, you are responsible for watching video recorded lectures and reading the required literature for each unit and then “attending” mandatory weekly one-hour online meetings where students have the chance to discuss the materials from a unit with the instructor.

Although this is an online course where students have more freedom in when they engage with the course materials, students are expected to spend the same amount of time overall on all activities in the course – including preparatory activities (readings, studying), in-class-activities (watching videos, participating in online meetings), and follow-up activities (working on assignments and exams) – as in an on-site course. As a rule of thumb, for each credit offered by a course, students can expect to spend one hour per week on in-class activities and three hours per week on out-of-class activities over the span of a full 12-week term. This is a 2-credit course that runs for 8 weeks. Hence, the total average workload is about 12 hours per week.

Mandatory Weekly Online Meetings

Thursday, 5pm – 5:50pm (CET)/ 11 am – 11:50pm (EST)
Meetings will be held online through BlueJeans. Follow the link to the meeting sessions on the course website on jpsmonline.umd.edu. If video participation via Internet is not possible, arrangements can be made for students to dial in and join the meetings via telephone.

In preparation for the weekly online meetings, students are expected to watch the lecture videos and read the assigned literature before the start of the meeting. In addition, students are encouraged to e-mail questions about the materials covered in the videos and readings of the week to the instructor (joerg.drechsler@iab.de) before the meetings (deadline for sending questions via e-mail is Thursday, 6am CET/Wednesday, midnight EST).

Grading

Grading will be based on:

- 2 quizzes (worth 15% total)
- Participation in the weekly online meetings, engagement in discussions during the meetings and/or submission of questions via e-mail (10% of grade)
- Three homework assignments (45%)
- A final online exam (30% of grade)

Dates of when assignment will be due are indicated in the syllabus. Late assignments will not be accepted without prior arrangement with the instructor.

Technical Equipment Needs

The learning experience in this course will mainly rely on the online interaction between students and the instructor during the weekly online meetings. Therefore we encourage all students in this course to use a web camera and a headset. Decent quality headsets and web cams are available for less than $20 each. We ask students to refrain from using built-in web cams and speakers on their desktops or laptops. We know from our experience in previous online courses that this will reduce the quality of video and audio transmission and therefore will decrease the overall learning experience for all students in the course. In addition, we suggest that students use a wire connection (LAN), if available, when connecting to the online meetings. Wireless connections (WLAN) are usually less stable and might be dropped.

Long Course Description

Statistical agencies and other data collecting institutions constantly face the dilemma between providing broad access to their data and maintaining the confidentiality of the individuals included in the collected data. To address this trade-off various statistical disclosure control (SDC) methods have been developed which help to ensure that no sensitive information can be disclosed based on the disseminated data. However, applying these methods usually comes at the price of information loss or potentially biased inferences based on the protected data.

This course will introduce the data protection strategies that are commonly used by statistical agencies and discuss their advantages and limitations. We will also briefly look at the computer science perspective on data privacy. We will discuss the differences to the SDC perspective and what the SDC community could learn from the approaches developed in computer science.

The main part of the course will focus on a relatively new approach to statistical disclosure control that has been implemented successfully for some data products recently: Generating synthetic data. With this approach statistical models are fitted to the original data and draws from these models are released instead of the original data. If the synthesis models are selected carefully, most of the relationships found in the original data are preserved.

You will learn about the general idea of synthetic data and the two main approaches for generating synthetic datasets. The close relationship to multiple imputation for nonresponse will also be discussed.

The quality of the synthetic data crucially depends on the quality of the models used for generating the data. Thus, the course will present various parametric and nonparametric modeling strategies in great detail.

The quality needs to be evaluated in two dimensions: (i) How well is the analytical validity preserved, i.e. how close are analysis results based on the synthetic data to results obtained from the original data? (ii) What is the remaining risk of disclosure for

the released data?
Several strategies to measure these two dimensions will be introduced. All steps of the synthesis process from generating the data, over analyzing the data, to evaluating the analytical validity and disclosure risk will be illustrated using simulated and real data examples in R.

Readings

All readings will be made available on the course website: jpsmonline.umd.edu

Interested students might find the following additional recommended book helpful in preparing for the course:

> Drechsler, J. (2011). Synthetic datasets for statistical disclosure control. Theory and implementation. Lecture notes in statistics, 201, New York: Springer

Academic Conduct

Clear definitions of the forms of academic misconduct, including cheating and plagiarism, as well as information about disciplinary sanctions for academic misconduct may be found at

http://www.graduate.umaryland.edu/policies/misconduct.html (University of Maryland) and

Knowledge of these rules is the responsibility of the student and ignorance of them does not excuse misconduct. The student is expected to be familiar with these guidelines before submitting any written work or taking any exams in this course. Lack of familiarity with these rules in no way constitutes an excuse for acts of misconduct. Charges of plagiarism and other forms of academic misconduct will be dealt with very seriously and may result in oral or written reprimands, a lower or failing grade on the assignment, a lower or failing grade for the course, suspension, and/or, in some cases, expulsion from the university.

Accommodations for Students with Disabilities

In order to receive services, students at the University of Maryland must contact the Disability Support Services (DSS) office to register in person for services. Please call the office to set up an appointment to register with a DSS counselor. Contact the DSS office at 301.314.7682; http://www.counseling.umd.edu/DSS/.

Course Evaluation

In an effort to improve the learning experience for students in our online courses, students will be invited to participate in an online course evaluation at the end of the course (in addition to the standard university evaluation survey). Participation is entirely voluntary and highly appreciated.

Class Schedule

Please note that assignments and dates are subject to change. Information (e.g., articles and assignments) posted to the course website supersedes the information noted here.

Unit 1: A Brief History of Data Confidentiality & Traditional Approaches for Data Protection

Video lecture (Drechsler): available online Thursday, March 23, 2017

Online meeting (Drechsler): Thursday, March 30, 2017, 5pm (CET)/11am (EST)

Readings:

Reiter, J. P. (2012). Statistical approaches to protecting confidentiality for microdata and their effects on the quality of statistical inferences. *Public Opinion Quarterly* **76**, 163–181.

Recommended (optional):

Winkler, W. E. (2007). Examples of easy-to-implement, widely used methods of masking for which analytic properties are not justified. Technical report, Statistical Research Division, U.S. Bureau of the Census, Washington, DC.

Unit 2: The Computer Science Perspective on Data Privacy & Introduction to Multiply Imputed Synthetic Datasets

Video lecture (Drechsler): available online Thursday, March 30, 2017

Online meeting (Drechsler): Thursday, April 6, 2017, 5pm (CET)/11am (EST)

First homework assignment: due Thursday, April 11, 2017, midnight (CET)/6pm (EST)

Readings:

Kinney, S. K., Reiter, J. P., Reznek, A. P., Miranda, J., Jarmin, R. S., and Abowd, J. M. (2011). Towards unrestricted public use business microdata: The synthetic Longitudinal Business Database. *International Statistical Review* **79**, 363–384.

Rubin, D. B. (1993). Discussion: Statistical disclosure limitation. *Journal of Official Statistics* **9**, 462–468

Recommended (optional):

Little, R. J. A. (1993). Statistical analysis of masked data. *Journal of Official Statistics* **9**, 407–426.

Unit 3: Analyzing Synthetic Datasets & Relationship to Multiple Imputation for Nonresponse

Video lecture (Drechsler): available online Thursday, April 6, 2017

Online meeting (Drechsler): Thursday, April 13, 2017, 5pm (CET)/11am (EST)

Quiz 1: due Thursday, April 20, 2017, 5pm (CET)/9am (EST)

Readings:

Raghunathan, T. E., Reiter, J. P., and Rubin, D. B. (2003). Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics* **19**, 1–16.

Reiter, J. P. (2003). Inference for partially synthetic, public use microdata sets. *Survey Methodology* **29**, 181–189.

Recommended (optional):

Rässler, S., Rubin, D. B., Zell, E. R (2007). Incomplete data in epidemology and medical statistics. In: C. R. Rao, J. Miller, D.C. Rao eds., *Handbook of Statistics*, 27, Elsevier, 569–601.

Unit 4: Synthesis Models Part I (Univariate and Linear Regression Models)

Video lecture (Drechsler): available online Thursday, April 13, 2017

Online meeting (Drechsler): Thursday, April 20, 2017, 5pm (CET)/11am (EST)

Second homework assignment: due Monday, April 24, 2017, midnight (CET)/6pm (EST)

Readings:

Raghunathan, T.E., Lepkowski, J.M., van Hoewyk, J., and Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a series of regression models. *Survey Methodology* **27**, 85–96.

Recommended (optional):

Rubin, D. B. (1981). The Bayesian bootstrap, *The Annals of Statistics* **9**, 130–134.

Unit 5: Synthesis Models Part II (Models for Categorical Variables and Nonparametric Models) & Modeling Strategies

Video lecture (Drechsler): available online Thursday, April 20, 2017

Online meeting (Drechsler): Thursday, April 27, 2017, 5pm (CET)/11am (EST)

Quiz 2: due Monday, May 1, 2017, midnight (CET)/6pm (EST)

Readings:

Reiter, J. P. (2005). Using CART to generate partially synthetic, public use microdata. *Journal of Official Statistics* **21**, 441–462.

Schenker, N., Raghunathan, T. E., Chiu, P. L., Makuc, D. M., Zhang, G., and Cohen, A. J. (2006). Multiple imputation of missing income data in the National Health Interview Survey. *Journal of the American Statistical Association* **101**, 924–933.

Recommended (optional):

Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input (disc: P558-573). *Statistical Science* **9**, 538–558.

Unit 6: Analytical Validity & Disclosure Risk Part I (Theory)

Video lecture (Drechsler): available online Thursday, April 27, 2017

Online meeting (Drechsler): Thursday, May 4, 2017, 5pm (CET)/11am (EST)

Third homework assignment: due Monday, May 15, 2017, midnight (CET)/6pm (EST)

Readings:

Karr, A. F., Kohnen, C. N., Oganian, A., Reiter, J. P., and Sanil, A. P. (2006). A framework for evaluating the utility of data altered to protect confidentiality. *The American Statistician* **60**, 224–232.

Abayomi, K., Gelman, A., and Levy, M. (2008). Diagnostics for multivariate imputations. *Journal of the Royal Statistical Society, Series C* **57**, 273–291

Recommended (optional):

Woo, M. J., Reiter, J. P., Oganian, A., and Karr, A. F. (2009). Global measures of data utility for microdata masked for disclosure limitation. *Journal of Privacy and Confidentiality* **1**, 111–124.

Unit 7: Disclosure Risk Part II (Examples in R) & Discussion of the Chances and Obstacles of the Synthetic Data Approach

Video lecture (Drechsler): available online Thursday, May 4, 2017

Online meeting (Drechsler): Thursday, May 11, 2017, 5pm (CET)/11am (EST)

Readings:

Drechsler, J. and Reiter, J. P. (2008). Accounting for intruder uncertainty due to sampling when estimating identification disclosure risks in partially synthetic data. In J. Domingo-Ferrer and Y. Saygin, eds., *Privacy in Statistical Databases*, 227–238. New York: Springer

Reiter, J. P. and Mitra, R. (2009). Estimating risks of identification disclosure in partially synthetic data. *Journal of Privacy and Confidentiality*, **1(1)**, Article 6.

Recommended (optional):

Reiter, J. P., Wang, Q., and Zhang, B. (2014), Bayesian estimation of disclosure risks in multiply imputed, synthetic data, *Journal of Privacy and Confidentiality* **6(1)**, Article 2.

Unit 8: Discussion of the Third Homework Assignment

Video lecture: no video lecture

Online meeting (Drechsler): Thursday, May 18, 2017, 5pm (CET)/11am (EST)

Final Exam Take Home

Final Exam due Thursday, May 25, 2017, midnight (CET)/6pm (EST)