SPONSORED BY THE
Federal Ministry of Education and Research

Introduction to Record Linkage with Big Data Applications

1 credit

Surv 667

Instructors

Manfred Antoni, manfred.antoni@iab.de
Stefan Bender, stefan.bender@bundesbank.de
Joseph Sakshaug, joe.sakshaug@manchester.ac.uk

Short Course Description

The course will address methods to combine data on given entities (people, households, firms etc.) that are stored in different data sources. By showing the strengths of these methods and by providing numerous practical examples the course will demonstrate the various benefits of record linkage. The participants will also learn about potential pitfalls record linkage projects may face.

Course and Learning Objectives

By the end of the course, students will…

- *be familiar with a host of record linkage applications from different countries or jurisdictions that link a variety of data sources and use different types of linkage*
- *know how to improve the quality of linkage identifiers by applying pre-processing routines*
- *be familiar with different methods of increasing the efficiency of record linkage*
- *be able to understand, select and apply appropriate record linkage methods (e.g., deterministic and probabilistic linkage)*
- *be familiar with packages for record linkage in R*
- *be able to evaluate the success of data linkage*

Prerequisites

Students should have knowledge of basic statistical concepts. They should have a basic understanding of R and should be able to adapt and run R scripts, for example from RStudio. A basic understanding of regular expressions is useful but not strictly 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. Just like in an on-site course, homework will be assigned and graded and there will be short online quizzes during the course.

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 1-credit course that runs for 4 weeks. Hence, the total average workload is about 12 hours per week.

Mandatory Weekly Online Meetings

Monday, 7 pm CET / 1 pm 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 post questions about the materials covered in the videos and readings of the week in the forum "Place to post your questions" on the course page before the meetings (deadline for submitting questions to the forum is Monday, 1:00 pm CET / 7 am EST, i.e. 6 hours before the online meeting).

Grading

Grading will be based on:

- 3 online quizzes (worth 10% each, 30% in total)
- Participation in the weekly online meetings (20% of grade): engagement in discussions during the meetings and submission of questions via e-mail (deadline: Monday, 1:00 pm CET / 7 am EST, i.e. 6 hours before the online meeting)

- 2 homework assignments (worth 25% each, 50% in total)

Dates of when assignments 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

The course will provide an introduction to record linkage: it will address methods to combine data on given entities (people, households, firms etc.) that are stored in different data sources. By showing the strengths of these methods and by providing numerous practical examples ranging from linked survey and administrative data to Big Data applications, the course will demonstrate the various benefits of record linkage. The participants will also learn about potential pitfalls record linkage projects may face.

The schedule of the course will be following a prototypical record linkage process:

- *the need for common identifiers (e.g., names, addresses, birth dates) and the importance of assuring high data quality even during the planning phase of each project.*
- *preparation of these identifiers before the actual linkage.*
- *increase the efficiency of the matching step (different blocking techniques).*
- *alternative ways of conducting the matching step, namely rule-based, distance-based and probabilistic record linkage.*
- *as data protection requirements are an important issue in many applications, methods of privacy preserving record linkage are discussed.*
- *the multitude of suitable software products and their specific capabilities in dealing with record linkage problems.*

Numerous practical examples will give participants an opportunity to create and discuss own ideas for promising record linkage projects. By the end of the course participants *will enable to assess the feasibility of, plan and manage record linkage projects as well as to perform each step along an actual linkage process.*

Readings

Primary readings will be from the following volume:

> Christen, Peter. (2012). Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. Berlin: Springer.

Additional required and recommended readings will be made available on the course website:
jpsmonline.umd.edu

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.htm

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

Unit 1: Introducing record linkage in the age of Big Data

Video lecture (Bender, Sakshaug): available online Monday, January 30, 2017

Online meeting (Antoni, Bender, Sakshaug): February 6, 2017, 7 pm CET / 1 pm EST

Readings:

Christen (2012), Chapters 1 and 2.

Topics:

- Introduction
- What is RL? What is it not?
- Privacy issues (Bender)
- Consent (Sakshaug)
- Process of record linkage

Unit 2: Collecting and pre-processing linkage identifiers & blocking techniques

Video lecture (Antoni): available online Monday, February 6, 2017

Online meeting (Antoni): February 13, 2017, 7 pm CET / 1 pm EST

Online quiz 1: due Wednesday after online meeting, February 15, 2017, 12 pm CET / 6 pm EST

Homework assignment 1: due Friday after online meeting, February 17, 2017, 12 pm CET / 6 pm EST

Readings:

Christen (2012), Chapters 3 and 4.

Topics:

- Identifiers: Which are useful? How to ensure their quality?
- Importance and limitations of preprocessing

- Excursion: regular expressions
- Blocking

Unit 3: Comparison and classification of record pairs

Video lecture (Antoni): available online Monday, February 13, 2017

Online meeting (Antoni): February 20, 2017, 7 pm CET / 1 pm EST

Online quiz 2: due Wednesday after online meeting, February 22, 2017, 12 pm CET / 6 pm EST

Readings:

Christen (2012), Chapters 5, 6 and 8.

Recommended:

Fellegi, I. P. and Sunter, A. B. (1969) A Theory for Record Linkage, Journal of the American Statistical Association, 64, 1183-1210.

Topics:

- Distance-based linkage
- Probabilistic record linkage
- Privacy-preserving record linkage (Antoni)

Unit 4: Advanced topics, software options and literature review

Video lecture (Antoni, Bender, Sakshaug): available online Monday, February 20, 2017

Online meeting (Antoni, Bender, Sakshaug): February 27, 2017, 7 pm CET / 1 pm EST

Online quiz 3: due Wednesday after online meeting, March 1st, 2017, 12 pm CET / 6 pm EST

Homework assignment 2: due Friday after online meeting, March 3, 2017, 12 pm CET / 6 pm EST

Readings:

Christen (2012), Chapter 7.

Recommended:

Schild, C.-J. and Schultz, S. (2016). Linking Deutsche Bundesbank Company Data using Machine-Learning-Based Classification: Extended Abstract. In Proceedings of the Second International Workshop on Data Science for Macro-Modeling (DSMM'16). ACM, New York, Article 10, DOI: http://dx.doi.org/10.1145/2951894.2951896

Schnell, R., Bachteler, T. and Bender, S. (2004). A Toolbox for Record Linkage. Austrian Journal of Statistics, 33(1 & 2), 125-133.

Topics:

- Evaluation (Antoni)
- Advanced Classification Techniques (Bender)
- Software options (Sakshaug)
- Literature overview (Sakshaug)
- Record linkage in the age of Big Data – outlook (Bender)