SURV 667 Introduction to Record Linkage with Big Data Applications
1 credit/2 ECTS
Winter 2018/2019

Instructor(s)
Manfred Antoni, manfred.antoni@iab.de
Stefan Bender, stefan.bender@bundesbank.de
Joseph Sakshaug, joesaks@uni-mannheim.de

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 an advanced knowledge of R or Stata. 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
Tuesday, 1PM (EST)/7PM (CET), starting on February 5

Meetings will be held online through Zoom. Follow the link to the meeting sessions on the course website on https://www.elms.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 on the course page before the meetings (no later than Monday, 1PM EST/7PM CET, i.e. 24 hours before the online meeting; questions not posted in time will not be counted for the grading and may not be answered in the online meeting, time permitting).

Students have the opportunity to use the Zoom meeting room set up for this course to connect with peers outside the scheduled weekly online meetings (e.g., for study groups). Students are encouraged to post the times that they will be using the room to the course website forum to avoid scheduling conflicts. Students are not required to use Zoom and can of course use other online meeting platforms such as Google Hangout or Skype.
**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 in the forum on the course website (deadline: Monday, 1PM (EST)/7PM (CET), i.e. 24 hours before the online meeting)
- 2 homework assignments (worth 25% each, 50% in total)

Dates of when assignment will be due are indicated in the syllabus. Extensions will be granted sparingly and only with prior arrangement with the instructors.

**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:


Additional required and recommended readings will be made available on the course website: http://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

https://www.president.umd.edu/sites/president.umd.edu/files/documents/policies/III-100A.pdf (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/.

Students at the University of Mannheim should contact the Commissioner and Counsellor for Disabled Students and Students with Chronic Illnesses at http://www.uni-mannheim.de/studienbueros/english/counselling/disabled_persons_and_persons_with_chronic_illnesses/.

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: Introducing record linkage in the age of Big Data

Online meeting (Antoni, Bender, Sakshaug): Tuesday, February 5, 2019, 1PM (EST)/7PM (CET)

Video lecture (Bender, Sakshaug): available online, Tuesday, January 29, 2019

Readings:
Christen (2012), Chapters 1 and 2.

Recommended:

Topics:
- Introduction
- What is RL? What is it not?
- Privacy issues
- Consent
- Process of record linkage
Unit 2: Collecting and pre-processing linkage identifiers & blocking techniques

Online meeting (Antoni): Tuesday, February 12, 2019, 1PM (EST)/7PM (CET)

Online quiz 1: due Thursday, February 14, 2018, 1PM EST/7PM CET

Video lecture (Antoni): available online, Tuesday, February 5, 2019

Homework assignment 1: due Monday, February 18, 2019, 1PM EST/7PM CET

Readings:
Christen (2012), Chapters 3 and 4.

Recommended:

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

Online meeting (Antoni): Tuesday, February 19, 2019, 1PM (EST)/7PM (CET)

Online quiz 2: due Thursday, February 21, 2018, 1PM EST/7PM CET

Video lecture (Antoni): available online, Tuesday, February 12, 2019

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

Recommended:


Topics:
- Distance-based linkage
- Probabilistic record linkage
- Privacy-preserving record linkage

Unit 4: Advanced topics, software options and literature review

Online meeting (Antoni, Bender, Sakshaug): Tuesday, February 26, 2019, 1PM (EST)/7PM (CET)

Online quiz 3: due Thursday, February 28, 2018, 1PM EST/7PM CET

Video lecture (Antoni, Bender, Sakshaug): available online, Tuesday, February 19, 2019

Homework assignment 2: due Monday, March 4, 2019, 1PM EST/7PM CET

Readings:
  Christen (2012), Chapter 7.

Recommended:


Topics:
- Evaluation
- Advanced Classification Techniques
- Software options
- Literature overview
- Record linkage in the age of Big Data – outlook

Note: Student access to the course website will be revoked two weeks after the final exam.