IPSDS INTERNATIONAL PROGRAM IN SURVEY & DATA SCIENCE
SPONSORED BY THE Federal Ministry of Education and Research
AUFSTIEG DURCH BILDUNG OFFENE HOCHSCHULEN

Introduction to Record Linkage with Big Data Applications

SURV 699(667)

2 credits/4 ECTS

Summer 2020

Instructors

Manfred Antoni, PhD, manfred.antoni@iab.de
Prof. Stefan Bender, PhD, stefan.bender@bundesbank.de
Christian Borgs, PhD, christian.borgs@uni-duisburg-essen.de
Prof. Joseph W. Sakshaug, PhD, joe.sakshaug@iab.de

Video lectures by Manfred Antoni, PhD, Prof. Stefan Bender, PhD, Christian Borgs, PhD, Prof. Joseph W. Sakshaug, PhD

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 showing how each of them are performed in practice using R, the course will demonstrate the various benefits of record linkage. Participants will also learn about potential challenges that 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 able to evaluate the success of data linkage
- be able to perform each step in the record linkage process using the R software

Prerequisites

Students should have knowledge of basic statistical concepts. They need to have an intermediate knowledge of R. Familiarity with regular expressions, the R packages ggplot2 and tidyverse is useful 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 prior to participating in 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 (i.e. readings, studying), in-class-activities (i.e. watching videos, participating in online meetings), and follow-up activities (i.e. working on assignments and exams) – as in an on-site course. As a rule of thumb, you can expect to spend approximately 3h/week on in-class-activities and 9 hours per week on out-of-class activities (preparing for class, readings, assignments, projects, studying for quizzes and exams). Therefore, the workload in all courses will be approximately 12h/week. This is a 2-credits/4-ECTS course that runs for 8 weeks. Please note that the actual workload will depend on your personal knowledge.

Mandatory Weekly Online Meetings

Wednesday, 1:00 PM EDT/7:00 PM CEST, starting on June 10

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 Tuesday, 1:00 PM EDT/7:00 PM CEST, i.e. 24 hours before each online meeting; questions not posted in time will not be counted for the grading and may not be answered in 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 in the forum on the course website (deadline: Tuesday, 1:00 PM EST/7:00 PM CEST, i.e. 24 hours before each online meeting)
- 3 homework assignments (worth 50% in total)

A+	*100 - 97*
A	*96 - 93*
A-	*92 - 90*
B+	*89 - 87*
B	*86 - 83*
B-	*82 - 80*
Etc.	

The grading scale is a base scale recommended by the IPSDS. Variations for grading on a scale are at the discretion of the instructor.

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 demand for using different data sets in a “combined way” to analyse research questions is increasing. This is where record linkage comes into play as the common technique to integrate seperate data sets.

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. Participants will also learn about potential challenges record linkage projects may face.

The schedule of the course will follow 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,
- increasing the efficiency of the matching step (different blocking techniques),
- alternative ways of conducting the comparison 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,
- evaluation and visualization of different quality aspects of the linkage result.

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

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 readings will be made available on the course website: *https://www.elms.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

https://www.uni-mannheim.de/en/research/good-research-practice/ (University of Mannheim).

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 Accessibility & Disability Service (ADS) office to register in person for services. Please call the office to set up an appointment to register with an ADS counselor. Contact the ADS office at 301.314.7682; https://www.counseling.umd.edu/ads/.

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

Video lecture (Bender, Sakshaug): available online, Wednesday, June 3, 2020

Online meeting (Antoni, Bender, Borgs, Sakshaug): Wednesday, June 10, 2020, 1:00 PM EDT/7:00 PM CEST

Required readings:

> Christen (2012), chapters 1 (sections 1.1-1.3) and 2.
>
> Bender, S., Jarmin, R., Kreuter, F. and Lane, J. (2017): Privacy and Confidentiality, In: Foster, I., Gahin, R., Jarmin, R. S., Kreuter, F. und Lane, J. (eds.): Big Data and Social Science – A Practical Guide to Methods and Tools, Chapter 12, p. 299-312, Chapman & Hall. https://textbook.coleridgeinitiative.org/chap-privacy.html

Recommended readings:

> Christen, P., and Belacic, D. (2005): Automated Probabilistic Address Standardisation and Verification, Australasian Data Mining Conference Proceedings. http://cs.anu.edu.au/~Peter.Christen/publications/ausdm2005-christen-

address.pdf.

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
Video lecture (Antoni): available online, Wednesday, June 10, 2020

Online meeting (Antoni): Wednesday, June 17, 2020, 1:00 PM EDT/7:00 PM CEST

Online quiz 1: due Friday, June 19, 2020, 1:00 PM EDT/7:00 PM CEST

Required readings:

Christen (2012), chapters 3 (sections 3.1-3.5 and 3.8) and 4 (sections 4.1-4.5, 4.8 and 4.13).

Recommended readings:

Christen, P. (2012): A survey of indexing techniques for scalable record linkage and deduplication. IEEE transactions on knowledge and data engineering, 24(9), 1537-1555. https://pdfs.semanticscholar.org/dd61/496eb7f77c198702ddf7b6bea867faea4a76.pdf.

Randall, S. M., Ferrante, A. M., Boyd, J. H. and Semmens, J. B. (2013): The effect of data cleaning on record linkage quality. BMC medical informatics and decision making, 13(1), 64, https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/1472-6947-13-64

Topics:

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

Unit 3: Data preprocessing and core concepts of data quality for linking
Video lecture (Antoni, Borgs): available online, Wednesday, June 17, 2020

Online meeting (Antoni, Borgs): Wednesday, June 24, 2020, 1:00 PM EDT/7:00 PM CEST

Homework assignment 1: due Tuesday, June 30, 2020, 1:00 PM EDT/7:00 PM CEST

Recommended readings:

> Spolsky, J. (2013): The Absolute Minimum Every Software Developer Absolutely, Positively Must Know About Unicode and Character Sets (No Excuses!). https://www.joelonsoftware.com/2003/10/08/the-absolute-minimum-every-software-developer-absolutely-positively-must-know-about-unicode-and-character-sets-no-excuses
>
> Wickham, H., Grolemund, G. (2017): R for Data Science. Bejing: O'Reilly. Chapter 11. https://r4ds.had.co.nz

Topics:

- Basics of dealing with strings in R
- Potential pitfalls with character encoding
- Regular expressions
- Lookup tables

Unit 4: Comparison and classification of record pairs
Video lecture (Antoni): available online, Wednesday, June 24, 2020

Online meeting (Antoni): Wednesday, July 1, 2020, 1:00 PM EDT/7:00 PM CEST

Online quiz 2: due Friday, July 3, 2020, 1:00 PM EDT/7:00 PM CEST

Required readings:

> Christen (2012), chapters 5 (sections 5.1-5.5, 5.9 and 5.17), 6 (sections 6.1-6.3) and 8 (sections 8.1-8.3).

Recommended readings:

> Doidge, J. C. and Harron, K. (2018): Demystifying probabilistic linkage. International Journal for Population Data Science, 3(1). https://doi.org/10.23889/ijpds.v3i1.410.
>
> Draisbach, U. and Naumann, F. (2013): On choosing thresholds for duplicate detection. In Proceedings of the 18th International Conference on Information Quality (ICIQ). https://hpi.de/fileadmin/user_upload/fachgebiete/naumann/publications/2013/O

n_Choosing_Thresholds_for_Duplicate_Detection.pdf.

Fellegi, I. P. and Sunter, A. B. (1969): A Theory for Record Linkage, Journal of the American Statistical Association, 64, 1183-1210. https://courses.cs.washington.edu/courses/cse590q/04au/papers/Felligi69.pdf

Topics:

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

Unit 5: Probabilistic record linkage and blocking (application)
Video lecture (Antoni, Borgs): available online, Wednesday, July 1, 2020

Online meeting (Antoni, Borgs): Wednesday, July 8, 2020, 1:00 PM EDT/7:00 PM CEST

Homework assignment 2: due Tuesday, July 14, 2020, 1:00 PM EDT/7:00 PM CEST

Topics:

- Implementation of blocking
- Comparison of record pairs
- Classification of matches

Unit 6: Advanced topics, software options and literature review
Video lecture (Antoni, Bender, Sakshaug): available online, Wednesday, July 8, 2020

Online meeting (Antoni, Bender, Sakshaug): Wednesday, July 15, 2020, 1:00 PM EDT/7:00 PM CEST

Online quiz 3: due Friday, July 17, 2020, 1:00 PM EDT/7:00 PM CEST

Required readings:

Christen (2012), chapter 7 (sections 7.1-7.3).

Ghani, R. and Schierholz, M. (2017): Machine Learning, In: Foster, I., Gahin, R., Jarmin, R. S., Kreuter, F. and Lane, J. (eds.): Big Data and Social Science – A Practical Guide to Methods and Tools, Chapter 7, p. 147-186, Chapman & Hall. https://textbook.coleridgeinitiative.org/chap-ml.html

Recommended readings:

Schild, C.-J., Schultz, S. and Wieser, F. (2017): Linking Deutsche Bundesbank Company

Data using Machine-Learning-Based Classification. Technical Report 2017-01, Deutsche Bundesbank Research Data and Service Centre. https://www.bundesbank.de/Redaktion/EN/Downloads/Bundesbank/Research_Centre/research_data_company_data.pdf?__blob=publicationFile.

Schnell, R., Bachteler, T. and Bender, S. (2004): A Toolbox for Record Linkage. Austrian Journal of Statistics, 33(1 & 2), 125-133. http://www.stat.tugraz.at/AJS/ausg041+2/041+2Schnell.pdf

Topics:

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

Unit 7: Privacy-preserving record linkage using R
Video lecture (Antoni, Borgs): available online, Wednesday, July 15, 2020

Online meeting (Antoni, Borgs): Wednesday, July 22, 2020, 1:00 PM EDT/7:00 PM CEST

Required readings:

Schnell, R., Rukasz, D. (2019): PPRL: Privacy Preserving Record Linkage. https://cran.r-project.org/web/packages/PPRL/index.html.

Schnell, R., Bachteler, T. and Reiher, J. (2009): Privacy-preserving Record Linkage Using Bloom Filters. BMC Medical Informatics and Decision Making 9 (41).

Vatsalan D., Christen P. and Verykios, V. S. (2013): A taxonomy of privacy-preserving record linkage techniques. Journal of Information Systems.

Topics:

- Encryption of linkage identifiers using Bloom filters
- Blocking on Bloom filters using Multibit-Trees
- Comparison and classification

Unit 8: Evaluation and visualization of linkage quality
Video lecture (Antoni, Borgs): available online, Wednesday, July 22, 2020

Online meeting (Antoni, Borgs): Wednesday, July 29, 2020, 1:00 PM EDT/7:00 PM CEST

Homework assignment 3: due Tuesday, August 4, 2020, 1:00 PM EDT/7:00 PM CEST

Recommended readings:

Wickham, H (2016): ggplot2. Elegant Graphics for Data Analysis. Second edition. Cham: Springer. (draft version of third edition available for free here: https://ggplot2-book.org)

Topics:

- Calculating measures of linkage quality
- Visualization of linkage quality