Using Big Data for social science research

Mihály Fazekas
School of Public Policy
Central European University
Winter semester 2020-21 (2 credits)
Class times: 15.30-17.10, Mondays (weekly)
Office hours: By appointment
Teaching Assistant: Daniel Kovarek
Version: 20/11/2020

Introduction

The course is an introduction to state-of-the-art methods to use Big Data in social sciences research. It is a hands-on course requiring students to bring their own research problems and ideas for independent research. The course will review three main topics making Big Data research unique:

1. New and emerging data sources such social media or government administrative data;
2. Innovative data collection techniques such as web scraping; and
3. Data analysis techniques typical of Big Data analysis such as machine learning.

Big Data means that both the speed and frequency of data created are increasing at an accelerating pace virtually covering the full spectrum of social life in ever greater detail. Moreover, much of this data is more and more readily available making real-time data analysis feasible.

During the course students will acquaint themselves with different concepts, methodological approaches, and empirical results revolving around the use of Big Data in social sciences. As this domain of knowledge is rapidly evolving and already vast, the course can only engender basic literacy skills for understanding Big Data and its novel uses. Students will be encouraged to use acquired skills in their own research throughout the course and continue engaging with new methods.

Learning outcomes

Students will be acquainted with basic concepts and methods of Big Data and their use for social sciences research. They will gain first-hand experience with applying such methods to real-life research problems. The acquired knowledge will enable students to use Big Data methods in their individual research on various topics of political science, economics, and sociology.

Teaching format

The course consists of 12 sessions, one each week. Each session lasts for 100 minutes.

Pre-requisites

- Elementary proficiency in quantitative methods and familiarity with R. Please send an example R script to demonstrate meeting this requirement.
- Enrolment in MA or PhD course.

Requirements

- Students are required to attend classes regularly, familiarize themselves with each session's reading list and to participate actively in course discussions, in particular providing constructive feedback on other students' presentations.
- Students will pick a data source and research question at the beginning of the course which they will have to regularly work on and report to the class. The methods and approaches learnt in each session will have to be applied to the selected source and research question.
- Students will have to write individual final papers and submit their database and codes which they produced throughout the whole course. The final paper will be short, not longer than 3000 words, describing and critically assessing the data source, data collection method, and analytical tools used in light of the selected research question and relevant prior literature. Great emphasis will be given to the submitted database and annotated codes.

Assessment

Attendance and class-room participation 10 %
Data abstract 15%
In-class presentations 30 % (data and codes also to be submitted)
Individual student project & final paper 45% (example paper and grading criteria will be shared)

Deadlines

Data abstract due on the 20th of January 2020

Final papers due on the 11th of April 2020 (Each week of delay will result in a reduction of the final grade by one 'step', for example from a B+ to a B, then from a B to a B- etc.)

Core readings

- Mihály Fazekas (2014), The Use of 'Big Data' for Social Sciences Research: An Application to Corruption Research. SAGE Research Methods Case, see: http://srmo.sagepub.com/view/methods-case-studies-2014/n283.xml?rskey=eFkV0g&row=12
  Short videos on the paper:
  http://methods.sagepub.com/video/srmpromo/0Vt2p3/introduction-to-big-data-for-social-science-research and
  http://methods.sagepub.com/video/srmpromo/WHQehe/using-big-data-to-measure-formidable-concepts-the-case-of-government-contra
- Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani (2015) An Introduction to Statistical Learning: With Applications in R. 6th edition, Springer, London. For data and R codes see: http://www-bcf.usc.edu/~gareth/ISL/book.html

Optional introductory reading to R

- Robert I. Kabacoff (2011) R in Action. Data Analysis and Graphics with R. Manning Publications, New York https://www.manning.com/books/r-in-action

- Alain F. Zuur, Elena N. Ieno, and Erik H.W.G. Meesters (2009) A Beginner's Guide to R. Springer, London.
- Garrett Grolemund and Hadley Wickham (2016) R for Data Science. O'Reilly Media, Sebastopol, CA. See: http://r4ds.had.co.nz/

Optional advanced reading

- Christen, Peter (2012) Data Matching. Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. Springer, London.
- Trevor Hastie, Robert Tibshirani, Jerome Friedman (2013), The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd edition, Springer. For data and R codes see: http://statweb.stanford.edu/~tibs/ElemStatLearn/

Course Schedule

#	Date	Topic
1	11/1/2020	Introduction
2	18/1/2020	Potential data sources and how to assess them
3	25/1/2020	Visual arguments
4	1/2/2020	Web scraping, APIs, and parsing I
5	8/2/2020	Web scraping, APIs, and parsing II
6	15/2/2020	Model evaluation and significance testing
7	22/2/2020	Student presentations: data collection
8	1/3/2020	Analytical overview, non-linear regressions, caveats
9	8/3/2020	Unsupervised learning: Introduction to clustering and text mining
10	15/3/2020	Text mining continued, data matching, and deduplication
11	22/3/2020	Supervised learning: decision trees and random forests
12	29/3/2020	Student presentations: analytical results

Detailed course program

Session 1: Introduction

Session 1: Course overview, planning student projects (scoping student interest, selection of topics), introduction to what Big Data means and getting started with R

Easy introductory readings:

- *Dutcher, Jenna. (2014). What is Big Data? UC Berkeley Data Science Blog. See: https://datascience.berkeley.edu/what-is-big-data/*
- *Chris Anderson. The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. Wired Magazine, vol 16 no 7. June 2008. See: https://www.wired.com/2008/06/pb-theory/*
- *Introduction to R:*
  - *Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani (2015) An Introduction to Statistical Learning: With Applications in R. 6th edition, Springer. Chapter 2.3*

Sessions 2-3: Identifying, understanding, structuring, and critically assessing new data sources

Session 2: Potential data sources and how to assess them (e.g. social media data, government administrative data, internet analytics (e.g. google trends), smartphone data) and getting started with R

- *Mihály Fazekas (2014), The Use of 'Big Data' for Social Sciences Research: An Application to Corruption Research. SAGE Research Methods Case.*
  Short videos on the paper:
  http://methods.sagepub.com/video/srmpromo/0Vt2p3/introduction-to-big-data-for-social-science-research and
  http://methods.sagepub.com/video/srmpromo/WHQehe/using-big-data-to-measure-formidable-concepts-the-case-of-government-contra
- *Advanced introduction to R:*
  - *Luis Torgo (2011) Data Mining with R: Learning with Case Studies. CRC Press. Chapter 1.*
  - *Atz, Ulrich. (2013). 11 Tips on How to Handle Big Data in R. Open Data Institute Blog. See:* http://labs.theodi.org/blog/2013/07/18/fig-data-11-tips-how-handle-big-data-r-and-1-bad-pun/

Session 3: Visual arguments: principles of good data visualisation, data visualisation practice using R and Tableau

- *Edward Tufte (2001) The Visual Display of Quantitative Information. 2nd edition, Graphics Press. Chapter 2.*
- *R Shiny introduction (Start Your first Shiny app): http://shiny.rstudio.com/articles/#first-app*
- *Tableau introductory video (1. Tableau Public Overview): https://public.tableau.com/en-us/s/resources*

Further readings

- *Alberto Cairo (2019) How Charts Lie: Getting Smarter About Visual Information. W W Norton*

Sessions 4-5: Understanding and using new data collection and management techniques and assessing their strengths and weaknesses

Session 4: Web scraping, APIs, and parsing I

Session 5: Web scraping, APIs, and parsing II

Combined readings for sessions 4-5:

- *Conceptual overview: Simon Munzert, Christian Rubba, Peter Meissner, Dominic Nyhuis (2015) Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining. Wiley. Chapter 9.*
- *Documented practical examples:*
  - *http://stat4701.github.io/edav/2015/04/02/rvest_tutorial/ (scraping and parsing)*
  - *http://www.columbia.edu/~cjd11/charles_dimaggio/DIRE/styled-4/code-13/ (scraping and API)*
  - *https://sites.google.com/a/stanford.edu/rcpedia/screen-scraping/web-scraping-with-r (scraping and parsing)*
  - *http://bogdanrau.com/blog/collecting-tweets-using-r-and-the-twitter-search-api/ (API)*
  - *https://blog.predictiveheuristics.com/2014/10/28/a-primer-on-web-scraping-with-r/*
  - *https://github.com/pablobarbera/social-media-workshop*

Further readings for sessions 4-5:

- *Challenges of "found data" – methods to process data originally collected for other purposes:*
  - *Karimi, Fariba, et al. "Inferring gender from names on the web: A comparative evaluation of gender detection methods." Proceedings of the 25th International Conference Companion on World Wide Web. International World Wide Web Conferences Steering Committee, 2016.*
  - *Inferring gender and race from facial image data: Face++. https://github.com/FacePlusPlus/detect-demo*

Sessions 6-12: Data analytic techniques

Session 6: Model evaluation and significance testing

- *Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani (2015) An Introduction to Statistical Learning: With Applications in R. 6th edition, Springer. Chapter 2, 5.1*
- *Phillip I. Good (2006) Resampling Methods. A Practical Guide to Data Analysis. 3rd edition, Birkhauser, Boston. Chapter 3.*

Session 7: Student presentation of data collection results and data clinic

Session 8: Introduction to the new analytical repertoire, non-linear regression methods (e.g. regression splines), and caveats

- *Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani (2015) An Introduction to Statistical Learning: With Applications in R. 6th edition, Springer.*
  - *Introduction: Ch 2.1*
  - *Chapter 7*
- *Lazer et al. 2014. The Parable of Google Flu: Traps in Big Data Analysis. Science*

Further readings

- *Ginsberg et al. 2009. Detecting influenza epidemics using search engine query data. Nature.*

Session 9: Unsupervised learning: Introduction to clustering and text mining (main empirical examples from text mining)

- *Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani (2015) An Introduction to Statistical Learning: With Applications in R. 6th edition, Springer. Chapter 10.*
- *Justin Grimmer and Brandon M. Stewart. 2013. Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. Political Analysis, Vol. 21, No. 3, pp 267-297.*

Further useful compendium of data and software packages

- *https://quanttext.com/*

Session 10: Text mining continued, data matching, and deduplication

- *Roberts et al. 2014. Structural topic models for open-ended survey responses. American Journal of Political Science*
- *Dusetzina SB, Tyree S, Meyer AM, et al. (2014) Linking Data for Health Services Research. A Framework and Instructional Guide. Rockville (MD): Agency for Healthcare Research and Quality (US). Ch. 4.: https://www.ncbi.nlm.nih.gov/books/NBK253312/*

Further readings

- *John Wilkerson and Andreu Casas (2017), Large-Scale Computerized Text Analysis in Political Science: Opportunities and Challenges. Annu. Rev. Polit. Sci. 2017. 20:529–44*
- *https://cran.r-project.org/web/packages/RecordLinkage/RecordLinkage.pdf*
- *https://www.r-bloggers.com/fuzzy-string-matching-a-survival-skill-to-tackle-unstructured-information/*
- *Christen, Peter (2012) Data Matching. Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. Springer, London.*

Session 11: Supervised learning: decision trees and random forests

- *Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani (2015) An Introduction to Statistical Learning: With Applications in R. 6th edition, Springer. Chapter 8.*

Neural networks with R (if time permits)

- *Collection of resources: https://github.com/rdrr1990/kerasformula/blob/master/short_course/APSA_readme.md*
- *Trevor Hastie, Robert Tibshirani, Jerome Friedman (2013), The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd edition, Springer. Chpater 11*
- *François Chollet and JJ Allaire. Deep Learning with R. Manning Publications Co., 2018*

Session 12: Student projects' final presentation and discussion