Introduction to Computational Social Science

Session 7: Research design & workflows

Maximilian Haag Constantin Kaplaner

Geschwister-Scholl-Institute for Political Science
LMU Munich

23.01.2023
Room B U103, Tue 14:00–18:00 (bi-weekly)

Today’s session

Lecture

1. Presentations + term papers
2. Research design basics
3. Research in CSS
4. CSS Workflows: best practices & possibilities

Presentations & term papers

Term papers: timeline

2023-01-23 (this session) Research design & CSS workflow session

2023-02-06 (in 2 weeks) Student presentations of term paper ideas

…

2023-03-13 Term paper submission deadline

Grading

Per the study regulations, your final grade is composed of your **presentation grade** (50%) + **term paper grade** (50%).

Term papers i

You can choose to either

1. write a **research design** for a CSS paper → theory + concept focused
2. conduct an **analysis** using the techniques covered in the course → code focused

Feel free to make use of RStudio cloud → will stay open until the end of March 2023

Term papers ii

Option: Research design

Write up a research design for a potential study in the realm of CSS. This includes an introduction, research question, embedding of the question into the literature, theory and research design part.

Requirements

- ~ 20,000 characters (incl. spaces and references)
- PDF format

→ See research design paper template!

Option: Analysis / Coding project

Create a CSS coding project using (one of) the techniques covered in the course. This may include a data collection effort, an analysis or similar (see later slides). Describe your efforts and present your results in a short text.

Requirements

- ~ 10,000 characters (incl. spaces and references) + code
- PDF format write-up and `R` code
- Make sure to produce reproducible and commented code

→ See coding project template!

There are no special formatting requirements for the papers but please make them readable (e.g. 11pt font size, 1.4 spacing, serif font) and use a consistent citation style.

Make sure you visit the office hours to discuss your term paper with us!

Presentations

Task

Present the idea for your term paper.

Requirements

~ 10 min. presentation with slides

- Focus on identifying a clear research question / the goal of your paper
- Describe the methodology / data you intend to use
- Present your efforts so far
- Make sure to include your open questions and issues as well

Presentations

Task

Present the idea for your term paper.

Requirements

~ 10 min. presentation with slides

- Focus on identifying a clear research question / the goal of your paper
- Describe the methodology / data you intend to use
- Present your efforts so far
- Make sure to include your open questions and issues as well

Feel free to visit the office hours before your presentation and / or send us a proposal for your term paper to receive feedback in advance!

Office hours

Make use of the office hours to receive feedback and get help if you get stuck with your paper! This can include but is not limited to

- coding problems
- research design questions
- theoretical questions
- methodological questions
- presentation of your results

→ *Write us an e-mail describing your problem / general topic of inquiry and we will schedule an in-person meeting / Zoom call or, depending on the issue, handle your problem via e-mail.*

Research design basics

Types of research in CSS

We can roughly distinguish 3 types of research designs:

- Exploratory
- Descriptive
- Confirmatory

Types of research in CSS

Given you would like to write about political polarization, how would the different approaches look like?

Types of research in CSS

Exploratory: RQ: How is political polarization written about in the media in the US? RD: Collect a sample of newspaper articles from the New York Times that mention polarization, try to infer something from your exploration (e.g. a new measurement how to find polarization in news paper articels, a new theory of when polarization is mentioned)

Descriptive:

RQ: How has the salience of political polarization evolved over the last 30 years in the US? RD: Collect all newspaper from the New York times, identify articles with a connection to polarization, map out and describe the temporal developments / trends.

Confirmatory:

RQ: Do Violent Protests Affect Expressions of Party Identity? Evidence from the Capitol Insurrection (Eady et al., 2022)

RD: Collected daily twitter data of 3.4 million twitter users located in the US, identified their political affiliation, test if Republicans changed their affiliation more compared to Democrats (Difference-in-Differences) after Janaury 6th

https://www.cambridge.org/core/journals/american-political-science-review/article/do-violent-protests-affect-expressions-of-party-identity-evidence-from-the-capitol-insurrection/76D0502E7C5A83D3900DE42C5FDCA8EE

Types of analyses in Political Science

→ When working with newly obtained or unstructured data, it is often helpful to engage in descriptive and exploratory analysis first to get a feel for the data and the ways in which you might be able to use them

→ Most Political Science research is focused on confirmatory analysis; however *good exploratory and descriptive analysis* is often helpful in making data accessible and generating ideas for yourself and other researchers!

→ *Data collection or methodological application efforts* are just as important as explanatory research

Research process

Finding a topic for your paper

Sources of inspiration

- interest in a particular technique
- interest in a particular substantive topic
- existing research papers (→ use Google scholar and tools like elicit.org)

Possible ways of interacting with existing research

- replicate an approach in another context
- how can an existing argument be tested using CSS data / techniques?
- apply new techniques to existing data (→ take look at data repositories like Harvard Dataverse)

Define a research question

RQ:

- once you have a rough topic you should think about your research question
- the research question is the central part of your paper!
- it lays out what you are trying to answer
- it also defines what *type* of research you are conducting!

ask yourself:

- what might be missing in the literature about a topic?
- how can I add to that or how can I fill the gap?

Time for theory

once you have a research question you can start working on theory

Exploratory analysis

- When doing exploratory analysis we often lack an existing theory or framework
- Show why we are missing that or why current theory might not apply
- Descripe what we might have to do gain a better insight

Descriptive analysis:

- What concept am I trying to describe?
- What *is* the thing I am studying?

Confirmatory analysis:

- How might x influence y?
- What is the causal path?
- What *causes* x to influence y?
- Formulate hypothesis

Specify your research design

After defining your research design and gaining an understanding of the underlying theory it is time to define your research design. Questions to answer:

- **Case selection:** What case am I looking at and why? (e.g. which country, which organisation…)
- **Data:** What steps do I take to collect my data? What is included what not (be precise!)
- **Operationalization/Measurement:** How do I translate my concept into concrete variables? How can I measure those?
- **Model specification:** (In confirmatory analysis) what model do I use to test my hypothesis?

Do it

Now you should collect your data and conduct your analysis!

Present results

After doing the analysis present your results. You should include:

- Descriptive data
- Results from your analysis (e.g. plots of time trends, regression results …)
- Describe what results you retrieved

Conclusion / Discussion

- Briefly summarize what you found
- What are the limitations of your research?
- How do these results fit into the broader research landscape?
- What are avenues for future research?

CSS workflows: Best practices & possibilities

Overview

In CSS and research, there a range of tools an techniques that can help us simplify and standardize certain processes, this includes

- writing **reproducible code** so that your analysis can be replicated by others (see also session #02)
- writing **markup languages** to write your paper in an easily reproducible and interactive documents
- using **literature management** software to keep track of your literature and citations
- using **version control** to track changes to your code
- using servers for resource intense **computation**

Reproducbility: core principles (P. Ball 2016b, 2016a; John McLevey, Pierson Browne, and Tyler Crick 2022)

1. **Transparency:** Analysis parts are complete and sufficiently documents
2. **Auditability:** Analysis can be executed by other researchers or on different platforms
3. **Reproducbility:** Results are the same for anyone running the analysis using the same code and data
4. *Scalability*: Code can handle other inputs and outputs than those used in the specific project

Reproducibility: What does this mean for us? i

Share your full code and data **+**

→ **Transparency**

- write and store code in a way that is easily understandable for others (within reason)
- include all parts of code (incl. for figures, tables)
- document files, code and your data collection process

→ **Auditability**

- include every step of the process (package dependencies, environment settings) in your code
- make sure the analysis can be run in a newly set up environment

→ **Reproducibility**

taking **transparency** and **auditability** into account, make sure your results remain stable when re-running your analysis

→ **Scalability** (within reason and your own ability)

try to write generalized rather than highly specific code

Reproducible & interactive documents i

Markup languages separate text and formatting. This way you can write your paper without constantly worrying what it will look like. Popular choices in research include

- LaTeX → Overleaf
- `pandoc` / Rmarkdown (→ check out the lab files ending in `.Rmd` from our sessions)

While markup languages have many advantages (e.g. figure and table numbering / referencing, easy and consistent citing and much more), they also require some time to get used to and have a bit of a learning curve. *However, it can be useful to get now to be prepared for later challenges (e.g. your Bachelor thesis).*

Reproducible & interactive documents ii

There are many resources to get you started and Google and StackOverflow are your friend!

LaTeX

- Overleaf LaTeX guide
- latex-tutorial.com
- Blog post: LaTeX for the humanities

`pandoc` / Rmarkdown

- RStduio introduction to Rmarkdown
- Getting started with `pandoc`

Fun fact: These slides are written in Markdown and typeset in LaTeX using `pandoc` / Rmarkdown.

Literature management

Keep track of the literature you need for your research using literature management software. It can help you

- organize and sort through the literature you have saved
- save papers straight from your browser
- easily cite research in your own paper using various citation styles

Popular software includes

- Zotero (free & open source!)
- EndNote

Wikipedia also has a large list of reference mangement software.

Version control using `git` and GitHub

`git`

- Version control system (VCS): a system used to manage and track changes to your documents
- allows for easy collaboration on coding projects

It may take some time to work with a VCS, but there are `git` quickstart tutorials on the internet to help you ease in.

GitHub

- an online version of `git`
- you can easily make your projects public
- and invite others to contribute to your project

As a student, you can set up a free GitHub account with all the features of a pro account!

Computation

Problem

Computational analysis or data collection often takes time and resources → can render your computer unusable for a couple hours/days/weeks

Solution

Rstudio Server can be installed on server instances of cloud computing operators (AWS, Google Cloud).

Google also offers Google Colab (https://colab.research.google.com). While focused on Python, you can also run R code in the colab environment (https://towardsdatascience.com/how-to-use-r-in-google-colab-b6e02d736497)

Next session

Next session 06 Feb 2023: Presentation of your term paper idea

→ Don’t hesitate to make use of the office hours before and after the session!

References i

John McLevey, Pierson Browne, and Tyler Crick. 2022. “Reproducbility and Principled Data Processing.” In *Handbook of Computational Social Science: Data Science, Statistical Modelling, and Machine Learning*, edited by Uwe Engel, Anabel Quan-Haase, Sunny Liu, and Lars Lyberg. Vol. 2. European Association of Methodology Series. New York: Routledge.

P. Ball. 2016a. “Principled Data Processing. Data & Society Talks: Small Group Session.”

———. 2016b. “The Task Is a Quantum of Workflow. Human Rights Data Analysis Group.”

Appendix i