COS-D407. Scientific Modeling and Model Validation

Lecturer: Christina Bohk-Ewald

Week 1

University of Helsinki, Finland
26.10.2020–09.12.2020

Brief round of introduction

Who are you? What is your study background?

What are you most interested in with respect to scientific modeling and model validation?

What experiences do you have in this topic and in programming with R?

Course outline

- Course timeline
- Course content: general purpose
- Course content: intended learning outcomes
- Course content by week
- Organization of (lecture and lab) sessions
- What is expected of you
- Feedback, assessment, and grading
- Course learning material on GitHub

Course timeline by week

We go digital with this course using Zoom and meet twice a week:

Week 1	26.10.2020: 10:15–11:45	28.10.2020: 14:15–15:45	Zoom
Week 2	02.11.2020: 10:15–11:45	04.11.2020: 14:15–15:45	Zoom
Week 3	09.11.2020: 10:15–11:45	11.11.2020: 14:15–15:45	Zoom
Week 4	16.11.2020: 10:15–11:45	18.11.2020: 14:15–15:45	Zoom
Week 5	23.11.2020: 10:15–11:45	25.11.2020: 14:15–15:45	Zoom
Week 6	30.11.2020: 10:15–11:45	02.12.2020: 14:15–15:45	Zoom
Week 7	07.12.2020: 10:15–11:45	09.12.2020: 14:15–15:45	Zoom

$\rightarrow$ Meeting-URLs and passwords have been sent to you by email

Course content in general

This course introduces you to the basic ideas and steps of scientific modeling and model validation in the social sciences.

It places the primary focus on reflecting on the scientific method and on how to use and assess statistical models to understand and predict observed phenomena in the real world.

This course covers concepts, methods, and tools used for both selecting and assessing the performance of statistical models, most notably robustness checks, bias-variance tradeoff, and cross-validation.

You will get a practical sense for these techniques through hands-on exercises using real-world data to analyze the spread of the coronavirus pandemic, also using the software R.

Participating in this course will be very useful for you in terms of developing skills in applied data science, implementation of statistical models & assessment techniques, and data visualization.

Course content in general

Application-oriented basic course on scientific modeling and model validation using R.

The applications are based on real-world data for analyzing the spread of the coronavirus pandemic.

You need no prior knowledge in scientific modeling & model validation, nor in analyzing the spread of the COVID-19 pandemic.

Having some experience in programming with R would be an advantage, but it is generally best to learn something by just doing it. All R code will be provided.

Course content: intended learning outcomes

By the end of this course you will be able:

- to explain basic ideas and steps of the scientific method
- to explain basic ideas behind scientific modeling and model validation
- to reflect on the meaning of a model's outcome, its possible limitations and implications
- to explain and adopt concepts, methods, and tools for selecting statistical models (e.g., bias-variance tradeoff) and to assess their performance (e.g., robustness analysis, cross-validation)
- to present and discuss concepts, methods, and tools for selecting and assessing statistical models in the context of the social sciences.

Course content by week

- Weeks 1 & 2: Introduction to science and the scientific method, and to the role of scientific modeling & model validation within the scientific process from a broad (scientific) perspective.

- Weeks 3, 4 & 5: Introduction to a statistical model for estimating COVID-19 infections and to strategies for assessing its outcome (even though *true* values are not available to compare the outcome to).

- Weeks 6 & 7: Introduction to a toolbox of classical concepts and tools for selecting statistical models (e.g., bias-variance tradeoff) and to assess their performance (e.g., cross-validation).

Organization of sessions each week

Sequence of alternating activities including, for example

- Recap of previous material in brief Q & A sessions
- Mini lectures to introduce new topics / content
- Hands-on exercises using real-world data in R
  and reflecting on their possible meaning, limitations, and implications
- Discuss emerging issues and solve them together in class
- Present and interactively discuss findings of hands-on exercises

$\rightarrow$ Up to 3 hours per week allow to flexibly organize course sessions and provide sufficient time for self-study

What you are encouraged to do

For the sake of a safe, respectful, and trustful learning course environment:

- Please show your video. All the time, but at least whenever you speak and also during discussions (exception: technical difficulties).
- What happens during this course, stays in this course. There will be no recordings of any course session.
- Please just ask questions / give comments during this course. Note that a speaker will have a hard time following the chat while speaking. That is why it is important that you just ask your question.
- Please ensure respectful interaction with each other.
- Please constructively criticize concepts, not people.
- Please make others feel welcome to speak their mind, also when it might be in disagreement with yours.

What is expected of you

Active participation in class to deeply understand principles and practices in scientifc modeling and model validation.

Feedback, assessment, and grading

...are in alignment with key learning contents and learning activities:

- Formative feedback during course to discover and close learning gaps
- Summative assessment to generate course grade
  - Actively participating in class, and presenting and interactively discussing findings of your and other course participants' hands-on exercises. (40%)
  - Report that covers core topics of this course: scientific method, scientific modeling, and model validation. Report should briefly summarize the hands-on exercises and put them into a broader context; it should be approximately 3000 words long (not counting references, figures, tables, and R-code). You are supposed to write your report during this course, however, it will be due on December 21, 2020. (60%). $\rightarrow$ more information in week 7

What is expected of you

If you cannot attend a lecture or lab session, please give me a note by email.

Course learning materials

Course learning materials on GitHub:

`https://github.com/christina-bohk-ewald/2020-COS-D407-scientific-modeling-and-model-validation`

Contact

Email: christina.bohk-ewald@helsinki.fi

Office: Unioninkatu 35, room 202

Appointments: arrangement by email and personal communication

First week's class:

Introduction to science in general

- What is (the purpose of) science?
- How to produce valuable scientific outcome?
- How are methodology, applications, and validation interconnected in scientific research?
- Where to place scientific modeling and model validation within scientific process?
- What to say about scientific progress and open science?

First week's class in the lab: Brief introduction to R & Reflection about the meaning of science for you

- Programming in R: `https://www.r-project.org/`
  - Very first steps in R
  - Finding help online (R documentation, online tutorials, and user platforms such as stack overflow)
    - ★ E. Paradis (2005). R for Beginners. `https://cran.r-project.org/doc/contrib/Paradis-rdebuts_en.pdf`
    - ★ Grolemund and Wickham (2017). R for Data Science. `https://r4ds.had.co.nz/`
- Reflect on what science means to you and what could make scientific outcome (in)valuable

Central questions

What is (the purpose of) science?

How to produce valuable scientific outcome?

$\longrightarrow$ Addresses basic principles of science and its responsibility for societies

Central questions

Before we start:

What is your understanding of this matter?

What is (the purpose of) science?

What is (the purpose of) science?

Science is crucial for human development through generating knowledge in various areas / fields.

- Science (Latin word *scientia*) means knowledge.
- Science is about creating and organizing knowledge about the real world in a systematic manner.
- Science is about finding (testable) explanations and predictions for any kind of real world phenomenon with, e.g., oppenness, creativity, skepticism, and utter honesty.

$\rightarrow$ Many different definitions of what science is

Sagan (1997)

What is (the purpose of) science?

- Scientific findings are testable in the way that they can be falsified. Anything that cannot be falsified belongs to the realm of beliefs and opinions (Karl Popper).
- Dynamic nature of science and scientific findings. Scientific hypotheses and models hold until they get replaced with hypotheses and models that have more explanatory or predictive power for particular real world phenomena.
- Science can also be regarded as being about finding the truth (if that is at all possible) and identifying (or distinguishing it from) *bullshit* (Harry Frankfurt).

Sagan (1997)

What is (the purpose of) science?

"Claims that cannot be tested,

assertions immune to disproof are veridically worthless,

whatever value they may have

in inspiring us or in exciting our sense of wonder."

Carl Sagan (1997)

What is (the purpose of) science?

- Science is about openness towards (new) ideas & carefully testing their validity with respect to power / capability of explaining or predicting real world phenomena.
- Creating scientific findings requires:
  - Adequate reasoning
  - Coherent argumentation
  - Rigorous standards of evidence
  - Utter honesty

$\rightarrow$ All of these requirements need to be met in order to solve research questions (and to address societal challenges like the covid-19 pandemic).

Sagan (1997)

How to produce valuable scientific outcome?

- Scientific findings are well-documented in papers and also in various (social) media outlets and platforms.
- But less documented appears to be the scientific method / process (which often is a long and exhausting endeavor with many ups and downs) that scientists went through in order to generate these findings in the first place.
- And this can be a problem, because the scientific process is at least equally important as the scientific findings; also in order to assess the quality of these scientific findings.

Sagan (1997)
Feynman (1974)

How to produce valuable scientific outcome?

- You can assess the quality of scientific findings through, e.g., discovering potential sources of error & limitations in the underlying scientific process used to generate them
  - What might impair scientific findings?
  - What else could explain the presented outcome?
  - How large is the error / uncertainty of the results?
  - What do we not know about scientific setup?
- Feynman (1974): "Your first job is not to fool yourself — and you are the easiest person to fool"

→ Requires *scientific skepticism* and *scientific integrity* along the way

---

Sagan (1997)
Feynman (1974)

How to produce valuable scientific outcome?

"Whenever a theory appears to you as the only possible one,

take this as a sign

that you have neither understood the theory nor the problem

which it was intended to solve."

Karl Popper

How to produce valuable scientific outcome?

"[...] **scientific integrity**, a principle of scientific thought that corresponds to a kind of utter honesty [...]

For example, if you're doing an experiment, you should report everything that you think might make it invalid—not only what you think is right about it:

other causes that could possibly explain your results; and things you thought of that you've eliminated by some other experiment, and how they worked—to make sure the other fellow can tell they have been eliminated"

Richard P Feynman (1974)

How to produce valuable scientific outcome?

Interconnectedness of selecting suitable **methodology**,

its careful **validation**, and reasonable **application**

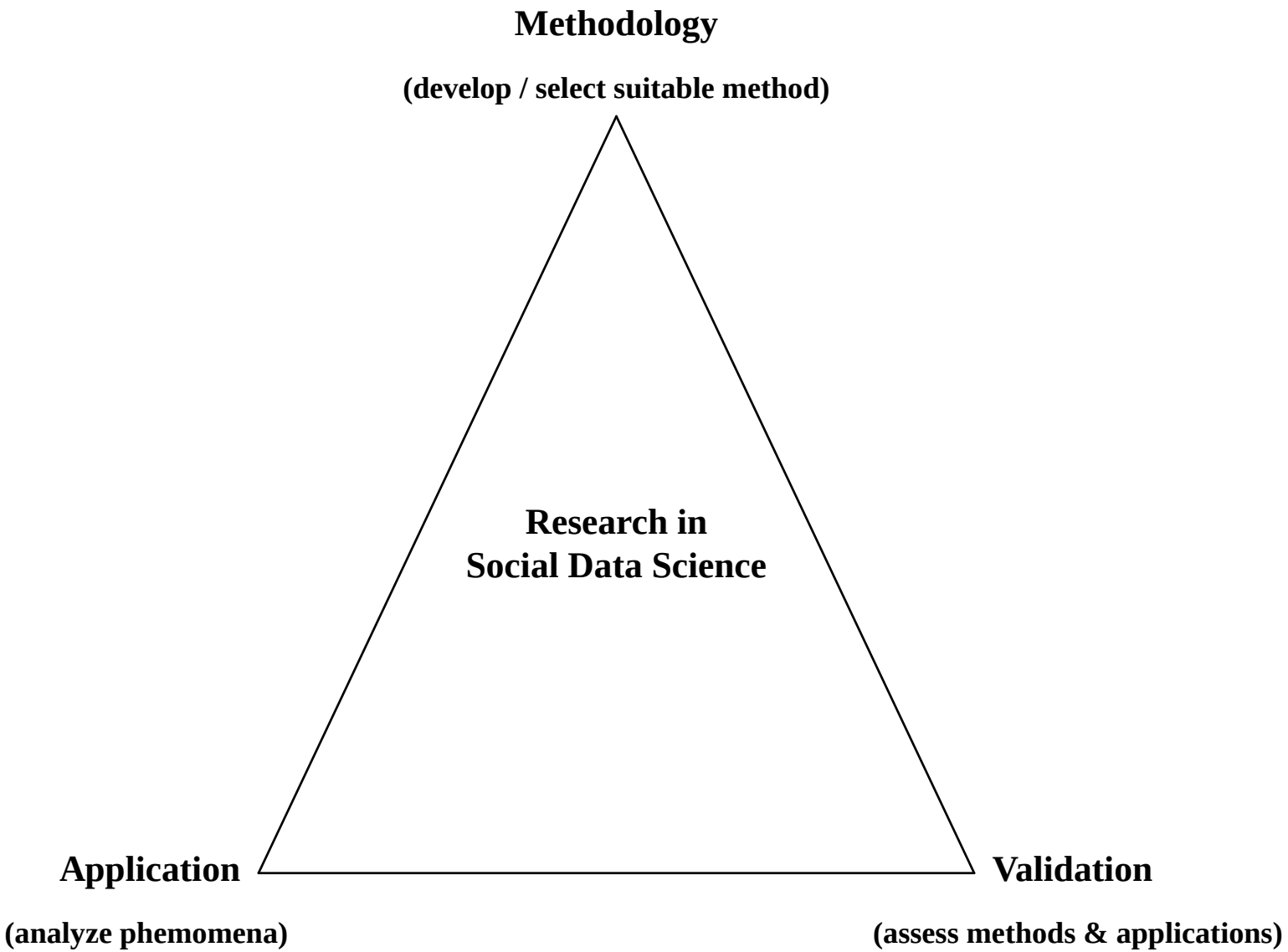
in order to produce (valuable) research within social (data) science

How to produce valuable scientific outcome?

Be aware of the interconnectedness of

- Developing or selecting **methodology** to analyze phenomena and to validate its suitability to do so.
- Assessing suitability of selected methodology and its application to analyze phenomena through careful **validation**.
- Analyzing phenomena through reasonable **application** of selected methododoly that has been assessed to be suitable for the task at hand.

How to produce valuable scientific outcome?



Scientific modeling & model validation in scientific process?

- **Scientific modeling.** Related to selecting suitable methodology for analyzing phenomena ($\rightarrow$ *triangle*). We focus on, e.g., concepts, methods, and tools for selecting suitable methodology to explain or predict phenomena.

- **Model validation.** Related to assessing the suitability of selected methodology to analyze phenomena in various settings ($\rightarrow$ *triangle*). We focus on, e.g., concepts, methods, and tools for assessing performance, flexibility, and accuracy of selected methodology with respect to explaining or predicting phenomena.

Scientific work

Scientific work requires, e.g.,
openness towards (new) ideas & creativity,
reasonable & careful planning and conduct of analysis,
rigorous & enduring testing,
and inherent skepticism & integrity
in order to produce valid and reliable outcome.

Scientific progress and open science

- Scientific debates fuel scientific progress through questioning and evaluating your own work and the work of other scholars.
- To allow for assessing scientific work it needs to be, e.g., testable, transparent, and reproducible; accessible.
- **Open science** is broadly understood as, e.g., providing open access to scientific work and its related publications.

$\rightarrow$ Some scientific journals appear to follow those principles, also as they introduce new categories for publishing different kinds of scientific work.

$\rightarrow$ Encouragements for publishing programming code and for reproducing previous scientific findings also put again more emphasis on scientific process. ($\rightarrow$ also important for you wrt seminar, master, and doctoral theses.)

Bijak (2019)

What you have learned today about the scientific method in general

- Describe the general purpose of science.
- Describe and explain what it takes to produce valuable scientific outcome.
- Describe and explain how methodology, application, and validation are interconnected in scientific process.
- Describe the position and function of scientific modeling and model validation within the scientific process.
- Describe the relationship between scientific progress and open science.

Course learning materials

Course learning materials on GitHub:

`https://github.com/christina-bohk-ewald/2020-COS-D407-scientific-modeling-and-model-validation`

Recommended learning material for today's class

- **Richard P Feynman (1974)**
  Cargo Cult Science. Some remarks on science, pseudoscience, and learning how to not fool yourself.
  Caltech's 1974 commencement address.
  `http://calteches.library.caltech.edu/51/2/CargoCult.htm`
- **Carl Sagan (1997)**
  The Demon-Haunted World: Science as a Candle in the Dark.
  Ballantine Books.
- **Jakub Bijak (2019)**
  Editorial: P-values, theory, replicability, and rigour.
  Demograhic Research 41(32): 949–952.
  `https://www.demographic-research.org/volumes/vol41/32/`
- **Grolemund and Wickham (2017)**
  R for Data Science.
  O'Reilly Media.
  `https://r4ds.had.co.nz/`

Thank you for your attention!

christina.bohk-ewald@helsinki.fi

First week's class in the lab: Brief introduction to R & Reflection about the meaning of science for you

- Programming in R: `https://www.r-project.org/`
  - Very first steps in R
  - Finding help online (R documentation, online tutorials, and user platforms such as stack overflow)
    - ★ E. Paradis (2005). R for Beginners. `https://cran.r-project.org/doc/contrib/Paradis-rdebuts_en.pdf`
    - ★ Grolemund and Wickham (2017). R for Data Science. `https://r4ds.had.co.nz/`
- Reflect on what science means to you and what could make scientific outcome (in)valuable