Scientific Reproducibility in Earth Observation Data Analytics

Hannah Augustin[1], **Martin Sudmanns**[1]
[1] Department of Geoinformatics – Z_GIS, University of Salzburg, Austria

Motivation

Reproducibility is one of the fundamental meta-concepts in science, the basis of the scientific method. It is not just about reproducing an experiment or a result, but also increasing the *transparency* of the process leading to its outcome. Scientific reproducibility in Earth observation (EO) data analytics has not been comprehensively studied. This poster attempts to start a conversation about what reproducibility means in the context of EO data analytics, and *facilitate skill-sharing* of some practices and tools that researchers can use to improve the reproducibility of their work.

What tools or practices do you use to make your work reproducible?

Terms

Across various scientific disciplines and in everyday language, there has long been *confusion about what reproducibility means*. Plesser (2018) and Nüst et al. (2018) recently compared a few definitions from different fields. In the context of EO data analytics, we understand the following terms:

Term	Researchers	Data	Methods	Results
replicability	different	independently collected	different	same
reproducibility	**different**	**same**	**same**	**same**
repeatability	same	same	same	same
transferability	any	different	same	comparable

Goodman et al. (2016) clarified reproducibility by breaking down different aspects within research. Space and time are fundamental to EO, since every spatio-temporal location of an observation is unique and cannot be repeated. This differs from some other scientific disciplines, where data to test a hypothesis can generally be independently collected at any time or place given the necessary materials and knowledge. Here is how we understand specific kinds of reproducibility in EO data analytics based on Goodman et al. (2016), recognising that independent EO-data collection to test hypotheses is often not possible (e.g. historical land cover change detection), depending on the spatio-temporal location, scale and context of the events, processes, states, objects, etc. in question:

methods reproducibility	ability to exactly repeat the **same methods** by providing sufficient detail about procedures and data
results reproducibility	ability to obtain the **same results** through the **same methods** using the **same data**
inferential reproducibility	ability to draw quantiatively **similar conclusions** from results obtained by using **different methods**, independently collected or **different data**, or by **reproducing the original study**

Skill-sharing to Increase Reproducibility

Here is some brainstorming related to EO-data analytics concerning points of *variation* that could impact reproducibility or transparency, identified practices that could improve reproducibility and a few available tools that could facilitate implementing those practices.

General

Hypothesis Creation

- develop clear hypotheses to test before conducting analysis
- disclose reasons a hypothesis is being tested

Make Assumptions Explicit

- communicate assumptions used in the creation of the hypothesis
  - e.g. in bi-temporal change detection, that the images are representative of changes you are hypothesizing and why

Reproduce Published Work

- try to reproduce existing work
- publish or present these outcomes at conferences
- include reproducing existing work in project proposals

Materials

Data
sensor, access point, data selection ...

- use free and open access data
- offer access to data upon request
- explain *why* this data was selected
  - e.g. climate, cloud cover, time of year
- be honest and clear about known bias in the data
  - e.g. fitness-for-use measure

Hardware
CPU, RAM, network ...

- disclose anything that could impact any performance benchmarks or limit implementation elsewhere

Funding
amount, duration, source ...

- disclose funding information
- disclose relevant partners and potential conflicts of interest

Environments

Computational Environment
operating system, dependencies, language ...

- virtual environments
- containerisation

conda, Docker

Data Storage
format, structure, metadata, permanence ...

- use non-proprietary format
- use open-source storage solutions
- version data
- use FAIR principle for metadata handling, including data lineage
- use established backup scenario and long-term storage after the end of the project
- store analysis ready data (e.g. pre-processed) in addition to raw data

PostgreSQL, Open Data Cube, "Community Editions" of rasdaman or SciDB

Software
version, access, dependencies ...

- use open-source software
- disclose what version (including dependences) used

Free and Open-Source Software (FOSS) (e.g. QGIS, GDAL/OGR, SAGA GIS, GRASS GIS, Orfeo Toolbox, InterIMAGE, OSSIM, pktools, GeoDMA), R packages, Python libraries

Methods

Data Pre-Processing
computing environment, algorithms, software, order of actions ...

- use open-source software and algorithms
- provide access to raw data and pre-processing chain
- automate as much as possible and provide commented scripts
  - e.g. using Python, R, etc.
- disclose when manual steps were taken and why

Free and Open-Source Software (FOSS) (e.g. GDAL, Orfeo Toolbox ...)

Analysis
computing environment, algorithms, software, order of actions ...

- use unbiased methods only if the data is also unbiased
- communicate any necessary prior knowledge
- disclose assumptions made throughout analysis
- use versioning of code
- meaningfully name intermediate results, code functions, etc.
- be clear which result was produced with which software, including version
- avoid manual analysis steps
- explain sample selection methods and provide samples used
- use process quality indicators such as number of parameters and human interactions
- if processing big EO data, offer and document analysis of a subset as an example for others to reproduce

Free and Open-Source Software (FOSS), Jupyter Notebooks, Git (for transparent versioning and collaboration)

Results/Dissemination

Results
format, level of interaction, interpretation ...

- publish with code using persistent links
- non-proprietary format
- open-access with clear licence for further use
- enable reprocessing at any time
- meaningfully name results

Documentation
accessibility, level of detail ...

- create clear workflow diagrams
- keep up-to-date, but also older versions, if practical
- clearly comment any code
- record any and all analysis parameters
- meaningfully name functions and variables in code
- document who conducted what analysis, contributed code or wrote text for publication
- archive data, results, etc. with associated Digital Object Identifier (DOI)

Open Science Framework (for documentation), Git (for transparent versioning and collaboration), Zenodo (for long-term storage and DOI)

Publication
review process, access, cost ...

- publish early at appropriate places
- publish data with DOI and clear licence
- publish null findings (e.g. unsupported hypothesis)
- publish code with clear licence
- publish reproduced work
- facilitate others to comment on the work

Please contribute any of your own considerations, tips, comments, suggestions, etc. using the clipboards provided!

Selected References

Goodman, S. N., Fanelli, D., & Ioannidis, J. P. A. (2016). What does research reproducibility mean? *Science Translational Medicine*, *8*(341), 341ps12-341ps12. https://doi.org/10.1126/scitranslmed.aaf5027

Nüst, D., Granell, C., Hofer, B., Konkol, M., Ostermann, F. O., Sileryte, R., & Cerutti, V. (n.d.). Reproducible research and GIScience: an evaluation using AGILE conference papers. https://doi.org/10.7287/peerj.preprints.26561v1

Peng, R. D. (2011). Reproducible Research in Computational Science. *Science (New York, N.Y.)*, *334*(6060), 1226–1227. https://doi.org/10.1126/science.1213847

Plesser, H. E. (2018). Reproducibility vs. Replicability: A Brief History of a Confused Terminology. *Frontiers in Neuroinformatics*, *11*. https://doi.org/10.3389/fninf.2017.00076

Challenges

- EO measurements and acquisitions may be repeated, but might no longer represent the targeted events, processes, states, objects, etc. relevant to the study
- reproducing others work is currently rarely if ever incentivised
- null findings or attempts at reproducing work are rarely if ever published
- interdisciplinary research relies on inferential reproducibility, but may evade reproducibility of methods or results

Contact:
Hannah Augustin
Department of Geoinformatics – Z_GIS
University of Salzburg, Austria
hannah.augustin@stud.sbg.ac.at