Workshop | Reproducibility of Scientific Results

BPLIM – Banco de Portugal

December 15, 2021

– Please type here your thoughts on the discussion

CHAT notes:

Sylverie Hebert, Banque de France – "**Implementing Reproducibility and Transparency in Central Banks**"

[https://sylverieherbert.github.io/]

Orozco et al JES 2020: https://onlinelibrary.wiley.com/doi/epdf/10.1111/joes.12389

-> Thibaut Lamadon: 2:20 PM

I have this big picture question which is what would it mean if one hard coded the estimate in the code. I can elaborate during discussion!

-> Lars Vilhuber: 2:21 PM

No guarantees that "computational reproducibility" checks would catch that, but they would expose that code to (cheap) scrutiny, and thus it should be found post-publication (and sometimes pre)

-> Miguel Portela: 2:29 PM

Adding to this project structure one can run a Git repository that would be made available when the paper is submitted

-> Miguel Portela: 2:31 PM

The repository would "enforce" reproducibility and neat organization of the project since the beginning

-> Luiza Cardoso De Andrade: 2:31 PM

Yes, using a git repository will allow you to share the whole folder structure.

-> Thibaut Lamadon: 2:37 PM

Conda is amazing and very portable. Conda export tends to freeze packages version almost too precisely, which makes it sometime hard to port to a different platform. In my experience writing the environment file by hand is a good practice. I am happy to elaborate!

-> Miguel Portela: 2:38 PM

The git solution, with these requirement files discussed in the previous slide, running on binder would make the whole process smooth and auditable

-> Matthias Gomolka Deutsche Bundesbank: 2:40 PM

In R, you can use the package "renv" to encapsulate your environment. It's simple and works great.

-> Nikolay Iskrev: 2:40 PM

adding "--from history" flag (e.g. "conda env export --from-history") alleviates/solves that problem

-> Miguel Portela: 2:42 PM

hashing the data be relevant

-> Matthias Gomolka Deutsche Bundesbank: 2:43 PM

For advanced automation in R, check out https://docs.ropensci.org/targets/. "targets" let's you create reproducible data pipelines and always keeps track of all elements of your pipeline. That means, when you re-run your pipeline, it skips all steps which are already up-to-date.

-> Luiza Cardoso De Andrade: 2:48 PM

My team has put together some guidelines on how to approach the creation of a flowchart and other documentation for a project: https://dimewiki.worldbank.org/Data_Map

Luiza Cardoso de Andrade, World Bank – "**Frequent reproducibility mistakes in Stata and how to avoid them**"

[https://luizaandrade.github.io/]

DimeWiki: https://dimewiki.worldbank.org/Main_Page

https://social-science-data-editors.github.io/template_README/

Discussion on using '\' for PATH: see https://twitter.com/AeaData/status/1396209749029531655

– Slides: https://osf.io/wd5ah/

from Lars Vilhuber to everyone: 3:14 PM

Rmarkdown (via pandoc) can output LaTeX, and can be adapted (via pandoc) to use certain journal styles. It will get you there 90% of the way

from Sergio Correia to everyone: 3:14 PM

@Sebastian, I understand RMarkdown creates PDF files via Pandoc, so you should also be able to keep the temporary .tex files and submit those (that's what I do but from plain markdown)

from Sebastian Buhai to everyone: 3:15 PM

I think the 10% above Lars's 90% is the annoying part :-). But if you folks say it works most of the time, I should give it a try.

from Lars Vilhuber to everyone: 3:15 PM

We can talk about other solutions that get you there 100% after Julia's talk.

from Sergio Correia to everyone: 3:16 PM

Related to the problem Luiza is discussing (Installing packages, and installing specific versions), what I do I -require- (experimental-ish command): https://github.com/sergiocorreia/stata-require

from Sergio Correia to everyone: 3:17 PM

require >= 0.9.4, from(ssc)

require ftools >= 2.48.0, from("https://github.com/sergiocorreia/ftools/raw/master/src/")

from Luiza Cardoso De Andrade to everyone: 3:50 PM

There are some microeconometrics methods that are implemented in Stata packages, but not yet in R. R and Python (and Julia) are much more general languages than Stata, and although most of the time this is an advantage, sometimes it's not. Many on the people who develop econometrics packages only develop them in Stata

Thibaut Lamadon, University of Chicago – “**Reproducibility in economics with administrative data: Making the most of existing techniques**”

[https://www.lamadon.com/]

State of the art: https://github.com/tlamadon/blm-replicate

“DVC is built to make ML models shareable and reproducible. It is designed to handle large files, data sets, machine learning models, and metrics as well as code.”

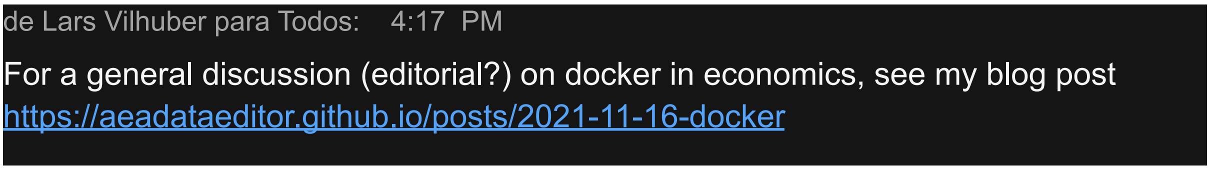
https://dvc.org/

Thibaut’s continuous reproducibility

https://github.com/tlamadon/continuous-reproducibility

-> go to the link and press ‘.’ on your keyboard

Containers
Singularity: you can build your singularity container here https://cloud.sylabs.io/home



de Lars Vilhuber para Todos: 4:17 PM

For a general discussion (editorial?) on docker in economics, see my blog post https://aeadataeditor.github.io/posts/2021-11-16-docker

For licenses (Stata) in Docker - https://github.com/AEADataEditor/docker-stata

from Miguel Portela to everyone: 4:16 PM

@Thibaut do you know if we can be sure 100% that next time we build the container we always get the same package version?

from Lars Vilhuber to everyone: 4:17 PM

For a general discussion (editorial?) on docker in economics, see my blog post https://aeadataeditor.github.io/posts/2021-11-16-docker

from Miguel Portela to everyone: 4:19 PM

@Thibaut should we store an image of the container in DockerHub or Sylab at the moment of submission to be absolutely sure about replication?

from Sergio Correia to everyone: 4:19 PM

@Thibaut Is polyglot git-friendly?

from Miguel Portela to everyone: 4:22 PM you can see here a rough/very low-level example on how to use different statistical packages, namely Stata, R, Julia and Python, in the same RMarkdown document here::https://github.com/reisportela/prjs

from Julia Schulte-Cloos to everyone: 4:25 PM

with R you can fix the package versions by relying on an MRAN snapshot

from Miguel Portela to everyone: 4:26 PM

@Lars thank you, so we should keep the container image?

Julia Schulte-Cloos, LMU Munich – "**A tool-kit for reproducible report generation with RMarkdown, Pandoc, and Lua**"

[https://jschultecloos.github.io/]

reproduceR: https://jschultecloos.github.io/reproducr

from Thibaut Lamadon to everyone: 5:08 PM

@julia, have you tried to cache cells?

from Lars Vilhuber to everyone: 5:09 PM

@Thibaut you can cache by default (setting a parameter), or individually.

from Oliver Hahn to everyone: 5:09 PM

@Lars: just writing it in the rmarkdown file?

from Lars Vilhuber to everyone: 5:09 PM

@oliver correct. β will render correctly.

from Sebastian Buhai to everyone: 5:21 PM

@miguel, thibaut: those sort of concerns, whether very justified or less so, motivate data providers/ administrators (particularly for official registers like those in Sweden, or Denmark, or Norway; Portugal a big, great, exception!) to be very reluctant in allowing any recent softwares installed (we're far away from containers: we are not even talking free new programming tools...). I need to send a CD with any new program, even if available online, if I want Denmark Statistics to install it on the server -- might get approved or not, in 3-4 months...

from Thibaut Lamadon to everyone: 5:22 PM

@sebastien, yes, I think it makes sense for these institutions to be concerns. They need to run checks. Often they do on what you take out.

from Lars Vilhuber to everyone:	5:53 PM

https://hdsr.mitpress.mit.edu/pub/fgpmpj1I/release/3

Marianne Saam, ZBW Leibniz Centre for Economics – “**Dawning of a new age? Replications Practices and Journal Data Policies**”

[https://www.zbw.eu/en/marianne-saam]

https://www.jcr-econ.org/

https://i4replication.org/

Also https://www.socialsciencereproduction.org/ as a possible platform, and the venerable https://replication.uni-goettingen.de/

https://bitss.github.io/ACRE/comunications.html

https://aeadataeditor.github.io/posts/2021-11-16-docker

Restud: https://restud.github.io/data-editor/

EJ and Econometrics Journal are both under Joan Llull now
https://ejdataeditor.github.io/index.htm

AEA is at https://aeadataeditor.github.io/

Thibaut Lamadon:	6:10 PM
my slides are here: https://tlamadon.github.io/code4research/

GESIS Notebooks: https://notebooks.gesis.org/

Lars Vilhuber, Cornell University and American Economic Association – **Overview of reproducibility and replicability in economics, with a side trip to provenance**

[https://www.vilhuber.com/lars/]

https://pollev.com/larsvilhuber238

– **Distill**: Communicating with Interactive Articles
https://doi.org/10.23915/distill.00028

– **Stencila**: Executable document pipelines. Author, collaborate, and publish beautiful interactive documents on an open source web platform.
https://stenci.la/

– **DOI for data**, Joana Ferreira Pimentel:
@ Thibaut, it is very easy. You just need to be registered in one DOI registration Agency.
https://www.doi.org/registration_agencies.html

-> Data DOI: https://www.da-ra.de/

Sebastian Buhai:
@Thibaut Zenodo can also create DOIs for software or data https://zenodo.org/

– **Statapackagesearch**
https://github.com/AEADataEditor/Statapackagesearch

– "Reproducibility and Replicability in Economics", by Lars Vilhuber, Dec 21, 2020
https://hdsr.mitpress.mit.edu/pub/fgpmpj1l/release/3

– cascad: https://www.cascad.tech/what-is-a-certification/

– AEA docker & Stata: https://github.com/AEADataEditor/stata-project-with-docker

– AEA, Use of Docker for Reproducibility in Economics:
https://aeadataeditor.github.io/posts/2021-11-16-docker

– Catherine Grace Schenck: 7:09 PM
Also Project TIER's website posts resources for training - including faculty development on how to teach reproducibility:
https://www.projecttier.org/fellowships-and-workshops/fall-2019-faculty-development-workshop/#about-project-tier-faculty-development-workshops