Reproducibility in Practice : Loosing your Balance on the Shoulder of Giants

Working Paper

Juste Raimbault

Monday 9th March

Abstract

As scientific reproducibility is an essential requirement for any study, its practice seems to be increasing [Stodden, 2010] and technical means to achieve it are always more developed (as e.g. ways to make data openly available, or to be transparent on the research process such as `git` [Ram, 2013], or to integrate document creation and data analysis such as `knitr`), at least in the field of numerical modeling and simulation. However, the devil is indeed in the details and obstacles judged at first sight as minor become rapidly a burden for reproducing and using results obtained in some previous researches. We describe two cases studies where models of simulation are apparently highly reproducible but unveil as puzzles on which research-time balance is significantly under zero.

1 On the Need to Explicit the Model

A current myth is that providing entire source code and data will be a sufficient condition for reproducibility. It will work if the objective is to produce exactly same plots or statistical analysis, assuming that code provided is the one which was indeed used to produce the given results. It is however not the nature of reproducibility. First, results must be as much implementation-independent as possible for clear robustness purposes. Then, in relation with the precedent point, one of the purposes of reproducibility is the reuse of methods or results as basis or modules for further research (what includes implementation in another language or adaptation of the method), in the sense that reproducibility is not replicability as it must be adaptable [Drummond, 2009].

Our first case study fits exactly that scheme, as it was undoubtedly aimed to be shared with and used by the community since it is a model of simulation provided with the Agent-Based simulation platform NetLogo [Wilensky, 1999]. The model is also available online [De Leon et al., 2007] and is presented as a tool to simulate socio-economic dynamics of low-income residents in a city based on a synthetic urban environment, generated to be close in stylized facts from the real town of Tijuana, Mexico. Beside providing the source code, the model appears to be poorly documented in the literature or in comments and description of the implementation. Comments made thereafter are based on the study of the urban morphogenesis part of the model (setup for the "residential dynamics" component) as it is our field of study [Raimbault, 2014]

Rigorous Formalization

Transparent Implementation

Expected Model Behavior

2 On the Need of Exactitude in Model Implementation

Conclusion

References

[De Leon et al., 2007] De Leon, F., Felsen, M., and Wilensky, U. (2007). Netlogo urban suite-tijuana bordertowns model. *Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL*.

[Drummond, 2009] Drummond, C. (2009). Replicability is not reproducibility: nor is it good science.

[Raimbault, 2014] Raimbault, J. (October 2014). Vers des modèles couplant développement urbain et croissance des réseaux de transports, phd project description. Technical report, Géographie-Cités UMR CNRS 8504.

[Ram, 2013] Ram, K. (2013). Git can facilitate greater reproducibility and increased transparency in science. *Source code for biology and medicine*, 8(1):7.

[Stodden, 2010] Stodden, V. (2010). The scientific method in practice: Reproducibility in the computational sciences.

[Wilensky, 1999] Wilensky, U. (1999). Netlogo.