Pourquoi devrions-nous arrêter d'embêter les gens avec la " recherche reproductible " et autres " bonnes pratiques " ?

Christophe Pouzat

STATISTIQUE & SOCIETE

Statistique et société

10 | 1 | 2022
Les chiffres au coeur de la crise du coronavirus (1)

Pourquoi devrions-nous arrêter d'embêter les gens avec la « recherche reproductible » et autres « bonnes pratiques » ?

Why should we stop bothering people with "replicable research" and other "good practices"?

Christophe Pouzat

OpenEdition Journals

Édition électronique
URL : https://journals.openedition.org/statsoc/332
DOI : 10.4000/statsoc.332
ISSN : 2269-0271

Éditeur
Société Française de Statistique (SFdS)

Édition imprimée
Date de publication : 1 mars 2022
Pagination : 53-57

Référence électronique
Christophe Pouzat, « Pourquoi devrions-nous arrêter d'embêter les gens avec la « recherche reproductible » et autres « bonnes pratiques » ? », *Statistique et société* [En ligne], 10 | 1 | 2022, mis en ligne le 01 octobre 2023, consulté le 20 février 2024. URL : http://journals.openedition.org/statsoc/332 ; DOI : https://doi.org/10.4000/statsoc.332

Pourquoi devrions-nous arrêter d'embêter les gens avec la « recherche reproductible » et autres « bonnes pratiques » ?

Christophe POUZAT[1]
IRMA, Université de Strasbourg et CNRS UMR 7501

TITLE
Why should we stop bothering people with "replicable research" and other "good practices"?

RÉSUMÉ
Comment continuer à encourager étudiants et collègues à faciliter l'accès à leurs travaux — en documentant et en rendant publics leurs programmes, en rendant accessibles leurs données, c'est-à-dire en mettant en œuvre la « recherche reproductible » — à la lumière d'un contre-exemple : le rapport d'un groupe d'épidémiologistes de l'Imperial College de Londres ayant eu une importance, semble-t-il déterminante, sur les confinements et donc sur nos vies depuis un an ? Telle est la question posée sous forme de dialogue par ce texte.

Mots-clés **:** *transparence, simulations stochastiques, Covid-19.*

ABSTRACT
How can we keep advocating to students and colleagues the practice of "reproducible research" – code and data documentation sharing, etc. – when we consider the counterexample of the report 9 of the Imperial College (London) epidemiologists? This report seems to have played a key role in the successive lockdown decisions, playing thereby a determinant role on our lives since last year. This is the question addressed by the following text in a dialogue format.

Keywords**:** *transparency, stochastic simulations, Covid-19.*

1. christophe.pouzat@math.unistra.frr

1. Introduction

Le dialogue imaginaire qui suit est une conséquence (de plus) de la Covid 19. Depuis un peu plus de 15 ans, j'« embête » étudiants et collègues pour les convaincre du bien-fondé d'une façon de faire et de présenter un travail scientifique : suivre les principes de la « recherche reproductible » (Claerbout et Karrenbach, 1992 ; Pouzat, Davison et Hinsen, 2015) – je vais les énoncer dans les grandes lignes après cette introduction. Voilà qu'arrivent la Covid 19 et les modélisateurs d'épidémie. Cet événement constitue, à mes yeux, la démonstration la plus flagrante de l'inanité de la démarche que j'ai prônée. Ma conclusion découle du report 9 du groupe du Professeur Ferguson à l'Imperial College de Londres (Ferguson *et al.*, 2020), du rôle qu'a joué ce rapport dans la décision britannique de confiner l'ensemble de la population, du rôle qu'il semble avoir joué dans la même décision chez nous[2] – j'espère que nous en saurons plus sur ce point bientôt – et de ce que nous apprennent les quelques examens maintenant disponibles du modèle et des simulations de ce rapport.

2. Un dialogue (de sourds)

(Ce dialogue qui n'en est pas un au sens strict, peut aussi être vu comme une communication sur la recherche reproductible lors d'une « session flash[3] » d'un congrès scientifique, accompagnée/entrelacée des notes prises par un auditeur visiblement dubitatif.)

- Chers collègues, chers étudiants, nous devrions toujours documenter les programmes que nous développons pour notre travail de recherche (Oliveira et Stewart, 2006 ; Klemens, 2014). C'est la meilleure garantie de pérennité de cette partie de notre travail. Cette documentation nous permettra de réexaminer nos codes si des erreurs sont constatées, même plusieurs mois ou années après leur écriture. Cela nous permettra aussi de développer de nouveaux programmes, basés sur nos anciens et surtout, cela permettra à d'autres, de notre labo ou d'ailleurs (si nous rendons le code accessible) de faire de même (Knuth, 1984).

- Mais le Professeur Ferguson explique à propos du programme utilisé pour le rapport 9[4] : « J'ai écrit ce code (des milliers de lignes de code C *non documentées*[5]) il y a plus de 13 ans pour modéliser une pandémie de grippe… ».

- Nous devrions aussi rendre nos programmes publics, nous sommes financés par des fonds publics et, comme la plupart de nos programmes sont peu utilisés, cela augmente les chances que les inévitables *bugs* soient trouvés (Hoaglin et Andrews, 1975 ; Eglen *et al.*, 2017). Rappelez-vous qu'il a fallu 8 ans pour que Don Knuth déclare TeX[6], son programme de composition de documents, *bug free* (Knuth, 1988, Préface). Or TeX était un programme ouvert, très utilisé par des gens qui savaient ce qu'ils faisaient et pour lequel les erreurs étaient très visibles.

- Mais le Professeur Ferguson n'a pas rendu son programme public. Une version réécrite (par qui ?) a été rendue publique[7] en avril dernier, alors que ce « même » programme était utilisé depuis plus de 13 ans.

2. Avis du Conseil Scientifique du 12 mars 2020, page 2, deuxième point de la sous-liste à puces (COVID-19 2020).
3. La version pour « adultes » de ce que les jeunes sont invités à faire avec « Ma thèse en 180 secondes ».
4. https://twitter.com/neil_ferguson/status/1241835454707699713
5. C'est moi qui souligne.
6. https://www.tug.org/whatis.html
7. https://github.com/mrc-ide/covid-sim

- Nous devrions documenter nos données comme les paramètres utilisés par nos programmes ; c'est la seule façon de pouvoir réutiliser, vérifier, partager cette partie de notre travail (Hoaglin et Andrews, 1975).

- Mais, je me répète, il n'y a pas trace de cela dans l'abondante production du Professeur Ferguson depuis son travail sur la « vache folle » au milieu des années 90. C'est en tout cas ce que suggère l'échange suivant sur le « problème 144 » du site GitHub[8] :
  - (Wes Hinsley, un membre du labo de Ferguson) [...] Plusieurs dizaines de milliers de simulations ont été utilisées pour modéliser la propagation de l'épidémie décrite dans le rapport 9[9]. [...]
  - (Franck Ch. Eigler) « Plusieurs dizaines de milliers de simulations... » Y a-t-il une trace écrite [dans un fichier d'ordinateur] de celles-ci ? Si oui, quelles sont les raisons pour ne pas simplement les partager ? Si non... ce serait très fâcheux.
  - (Wes Hinsley) Seulement qu'il y a plusieurs dizaines de milliers de simulations. Comme je l'ai écrit, nous explorons des stratégies pour les partager d'une façon raisonnable.

- Nous devrions rendre nos données, comme les paramètres utilisés par nos programmes, publics pour des raisons identiques à celles évoquées pour le partage des codes (Pouzat, Davison et Hinsen, 2015).

- Cela m'inspire la même réplique que pour le dernier point.

- Partager programmes, paramètres et données ne suffit pas, nous devons aussi expliquer, dans un « document reproductible » (Claerbout et Karrenbach, 1992 ; Fomel et Hennenfent, 2007 ; Pouzat, Davison et Hinsen, 2015), comment les programmes et les paramètres sont appliqués aux données pour obtenir les résultats (tables, figures) de nos articles ; puis partager ce « document ». Cela rend la détection et la correction des inévitables erreurs beaucoup plus efficaces ; cela permet à d'autres de critiquer notre travail et de construire sur celui-ci.

- Pourquoi s'embêter ainsi, même le « rapport 10 »[10], réplique du 9 avec la version publique du programme, ne satisfait pas à ces critères !

- Enfin, mais j'ai presque honte de vous rappeler des principes méthodologiques aussi élémentaires : lorsque notre travail fait intervenir des modèles intrinsèquement aléatoires (du fait d'emploi de méthodes de Monte-Carlo[11] par exemple), nous devons toujours faire beaucoup (entre 500 et 1000) de simulations pour une collection de paramètres donnée, puis caractériser la distribution des quantités d'intérêt par la moyenne et l'écart type (voire plus, boîtes à moustaches, etc.) (Hammersley et Handscomb, 1967 ; Ripley, 1987 ; Asmussen et Glynn, 2007 ; Graham et Talay, 2011 ; Ross, 2017).

- Mais je ne comprends pas, l'équipe du Professeur Ferguson n'a effectué qu'*une seule simulation* par jeu de paramètres ; ils l'expliquent eux-mêmes dans l'introduction au « rapport 10 »[12], on le voit dans le troisième point de la réplication d'une partie des simulations du « rapport 9 » (avec la version publique du programme) par Stephen Eglen (Eglen, 2020) et c'est discuté dans l'évaluation de Edeling et ses collaborateurs (Edeling

8. https://github.com/mrc-ide/covid-sim/issues/144
9. Le modèle de Ferguson comporte plus de 900 paramètres ; la plupart d'entre eux sont définis sur un intervalle de **R**. Si nous avions affaire à 900 paramètres binaires (prenant deux valeurs possibles), il y aurait déja $2^{900} > 10^{896}$ combinaisons à explorer.
10. https://github.com/mrc-ide/covid-sim/tree/master/report9
11. https://fr.wikipedia.org/wiki/M%C3%A9thode_de_Monte-Carlo
12. https://github.com/mrc-ide/covid-sim/tree/master/report9

et al., 2020), membres du groupe *Rapid Assistance in Modelling the Pandemic*[13] de la Royal Society. Pourquoi alors devrais-je être aussi « tatillon » ?

- …
- (L'auditeur pour lui-même) Ce type est fou à lier ! Pendant qu'il va perdre son temps à documenter, rendre public, simuler à outrance, moi je vais publier beaucoup plus de papiers, j'aurai plus de chances d'avoir mes réponses aux appels d'offre acceptées, j'aurai plus de chances d'avoir un poste. Peut-être même qu'un jour, qui sait, je me retrouverai à siéger dans une commission chargée d'évaluer le travail de ce fêlé…

Références

Asmussen S. and P. W. Glynn (2007), *Stochastic Simulation: Algorithms and Analysis*, Stochastic Modelling et Applied Probability, Springer-Verlag New-York.

Claerbout J. and M. Karrenbach (1992), « Electronic Documents Give Reproducible Research a New Meaning », *in* Proceedings of the 62nd Annual Meeting of the Society of Exploration Geophysics, pp. 601-604, http://sepwww.stanford.edu/doku.php?id=sep:research:reproducible:seg92.

COVID-19, Conseil Scientifique (2020), « Avis du Conseil scientifique Covid-19 du 12 mars 2020 », https://solidarites-sante.gouv.fr/IMG/pdf/avis_conseil_scientifique_12_mars_2020.pdf.

Edeling W., A. Hamid, R. Sinclair, D. Suleimenova, K. Gopalakrishnan, B. Bosak, D. Groen, I. Mahmood, D. Crommelin, and P. Coveney (2020), « Model uncertainty and decision making: Predicting the Impact of COVID-19 Using the CovidSim Epidemiological Code », *Research Square*, https://doi.org/10.21203/rs.3.rs-82122/v3.

Eglen S. J. (2020), « CODECHECK certificate 2020-010 », *Zenodo*, https://doi.org/10.5281/zenodo.3865491.

Eglen S. J., B. Marwick, Y. O. Halchenko, M. Hanke, Sh. Sufi, P. Gleeson, R. A. Silver, A. P. Davison, L. Lanyon, M. Abrams, T. Wachtler, D. Willshaw, Chr. Pouzat, and J.-B. Poline (2017), « Toward standard practices for sharing computer code and programs in neuroscience », *Nature Neuroscience*, vol. 20, n° 6, pp. 770-773, https://doi.org/10.1038/nn.4550.

Ferguson N. M., D. Laydon, G. Nedjati-Gilani *et al.* (2020), « Report 9 - Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand », Imperial College London, https://www.imperial.ac.uk/mrc-global-infectious-disease-analysis/covid-19/report-9-impact-of-npis-on-covid-19/.

Fomel S. and G. Hennenfent (2007), « Reproducible Computational Experiments Using Scons », *in Proc. IEEE Int'l Conf. Acoustics, Speech and Signal Processing*, 1257-60, p. 4.

Graham C. et D. Talay (2011), *Simulation stochastique et méthodes de Monte-Carlo*, Les Éditons de l'École Polytechnique.

13. https://royalsociety.org/topics-policy/Health%20and%20wellbeing/ramp/

Hammersley J. M. et D. C. Handscomb (1967), *Les méthodes de Monte-Carlo*, Monographies Dunod, Dunod.

Hoaglin D. C. and D. F. Andrews (1975), « The Reporting of Computation-Based Results in Statistics », *The American Statistician*, vol. 29, n° 3, pp. 122-26, https://doi.org/10.1080/00031305.1975.10477393.

Klemens B. (2014), *21st Century C. C tips from the new school*, Second Edition, O'Reilly.

Knuth D. E. (1984), « Literate Programming », *The Computer Journal*, vol. 27, n° 2, pp. 97-111.

Knuth D. E. (1988), *TEX: The Program*, Computers & typesetting, Addison Wesley Publishing Company.

Oliveira S. and D. E. Stewart (2006), *Writing Scientific Software*, Cambridge University Press.

Pouzat Chr., A. Davison et K. Hinsen (2015), « La recherche reproductible : une communication scientifique explicite », *Statistique et Société*, vol. 3, n° 1, pp. 35-38, http://statistique-et-societe.fr/article/view/448.

Ripley B. D. (1987), « Stochastic Simulation », *Wiley Series in Probability and Statistics*, janvier.

Ross S. M. (2017), *Simulation*, 5th edition, Academic Press.