# Pourquoi devrions nous arrêter d'embêter les gens avec la « recherche reproductible » ?

Christophe Pouzat

*<2020-12-06 dim.>*

## 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 » – 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*[1] du groupe du Professeur Fergusson à l'Imperial College de Londres, 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 – j'espère que nous en saurons plus sur ce point bientôt – et de ce qui nous apprennent les quelques examens maintenant disponibles du modèle et des simulations de ce rapport.

## Un dialogue

— Chers collègues, chers étudiants, nous devrions toujours documenter les programmes que nous développons pour notre travail de recherche. C'est la meilleure garantie de pérennité de cette partie de notre travail. Cette documentation nous permettra de ré-examiner 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.

1. `https://www.imperial.ac.uk/mrc-global-infectious-disease-analysis/covid-19/report-9-impact-of-npis-on-covid-19/`

— (Un étudiant) Mais le Professeur Ferguson explique à propos du programme utilisé pour le rapport 9[2] : « J'ai écrit ce code (des milliers de lignes de code `C` *non documentées*[3]) 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. Rappelez vous qu'il a fallut 8 ans pour que Don Knuth déclare TEX[4], son programme de composition de documents, *bug free*. 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.
— (Un étudiant) Mais le Professeur Ferguson n'a pas rendu son programme public. Une version ré-écrite (par qui ?) a été rendue public[5] en avril dernier, alors que ce « même » programme était utilisé depuis plus de 13 ans.
— Nous devrions documenter nos données comme les paramètres utilisés par nos programmes, c'est la seul façon de pouvoir réutiliser, vérifier, partager cette partie de notre travail.
— (Un étudiant) Mais Monsieur, je me répète, il n'y a pas traces de cela dans l'abondante production du Professeur Fergusson 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[6] :
  — (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. […]
  — (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, qu'elles 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 paratger d'une façon raisonnable.
— Nous devrions rendre nos données comme les paramètres utilisés par nos programmes publiques pour des raisons identiques à celles évoquées pour le partage des codes.
— (Un étudiant) Et je vous fais la même réplique que pour votre dernier

2. `https://twitter.com/neil_ferguson/status/1241835454707699713`
3. C'est moi qui souligne.
4. `https://fr.wikipedia.org/wiki/TeX`
5. `https://github.com/mrc-ide/covid-sim`
6. `https://github.com/mrc-ide/covid-sim/issues/144`

point.

— Partager programmes, paramètres et données ne suffit pas, nous devons aussi expliquer, dans un « document reproductible » [7], 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.

— (L'étudiant décidément têtu) Pourquoi s'embêter ainsi, même le « rapport 10 » [8], 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 [9] par exemple), nous devons toujours faire beaucoup (entre 500 et 1000) simulations pour une collection de paramètres donnée, puis caractériser la distribution des quantités d'intérêts par la moyenne et l'écart type (voir plus, boîtes à moustaches, etc).

— (Toujours le même étudiant) Mais Monsieur, 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 »[8], 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 Stepehn Egeln [10] et c'est discuté dans l'évaluation demandée par la *Royal Society* à Edling et ses collaborateurs [11]. Pourquoi alors devrais-je être « tatillon » comme vous l'êtes ?

— ...

— (L'étudiant 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 chance d'avoir un poste, peut-être même qu'un jour, qui sait, je me retrouverais à siéger dans une commission chargée d'évaluer le travail de ce fêlé...

7. http://publications-sfds.fr/index.php/stat_soc/article/view/448/422
8. https://github.com/mrc-ide/covid-sim/tree/master/report9
9. https://fr.wikipedia.org/wiki/M%C3%A9thode_de_Monte-Carlo
10. https://zenodo.org/record/3865491#.XuPW_y-ZPGI
11. https://www.researchsquare.com/article/rs-82122/v3