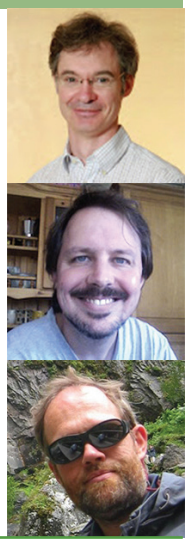
La recherche reproductible : une communication scientifique explicite



Christophe POUZAT[1], Andrew DAVISON[2], Konrad HINSEN[3]

Chargés de recherche du CNRS en statistique, neurosciences computationnelles, et physio-chimie computationnelle

Dans les articles de recherche ordinaires, tout n'est pas écrit, loin de là : des connaissances sont présupposées, des détails techniques sont omis. Depuis quelques années, des chercheurs ont entrepris de publier des articles qui, en plus du contenu scientifique classique, contiennent toute l'information nécessaire à la reproduction de celui-ci, une fois les données acquises. Des logiciels spécifiques ont été mis au point pour rendre aisée la production de tels articles. Les chercheurs ont intérêt à en profiter, car les revues scientifiques et les agences de financement de la recherche leur demandent de plus en plus d'adopter ce genre de pratiques.

La locution « recherche reproductible » apparaît de plus en plus fréquemment dans des articles scientifiques, des forums ou des blogs voire dans les préoccupations de grandes agences de recherche scientifique comme le National Institute of Health -NIH- (Collins, 2014). Ce phénomène a probablement de quoi surprendre auteurs et lecteurs de littérature scientifique pour qui le qualificatif de « scientifique » entraîne, implicitement au moins, la notion de reproductibilité. Tout lecteur, prenant un peu de recul, va néanmoins très vite réaliser que ce que Michael Polanyi (1998) désignait par « connaissance tacite » joue un rôle – nécessairement – considérable dans la vie quotidienne du scientifique, comme de toute personne d'ailleurs. En statistique, nous pourrions mettre les « conditions de régularité » des théorèmes que nous employons dans la catégorie des « connaissances tacites » ; en substance, ces dernières nous permettent de communiquer de façon concise avec nos collègues ; elles nous fournissent une certaine économie de pensée. La concision qu'elles apportent devient par contre un handicap lorsque nous devons communiquer avec des scientifiques d'un autre domaine, mais aussi avec des non scientifiques : deux situations fréquentes pour les statisticiens. Des difficultés peuvent apparaître également, en interne à un domaine de recherche, lorsque nous essayons d'accéder directement à l'« ancienne » littérature le concernant, puisque dans ce cas, la connaissance tacite peut avoir dérivé au cours du temps. Enfin, avec le recours de plus en plus fréquent à des moyens informatiques importants dans tous les aspects de la recherche, les résultats scientifiques dépendent souvent d'un grand nombre de détails techniques d'un protocole de calcul, qui restent également dans le domaine de la connaissance tacite parce que considérés comme détails techniques (Collins, 2014).

1. MAP5 Université Paris-Descartes et CNRS UMR 8145 , 45, rue des Saints-Pères 75006 Paris , christophe.pouzat@parisdescartes.fr; b.; c.
2. Unité de Neurosciences, Information et Complexité (UNIC; FRE3693 CNRS) 1, avenue de la Terrasse 91198 Gif sur Yvette, andrew.davison@unic.cnrs-gif.fr
3. Centre de Biophysique Moléculaire (UPR4301 CNRS) Rue Charles Sadron, 45071 Orléans Cédex 2, konrad.hinsen@cnrs-orleans.fr

L'objectif

La « recherche reproductible » peut être vue comme une méthode de réduction de l'implicite dans une partie de notre communication. Elle va résulter en un « document dynamique » ou « article actif » (*active paper*) (Hinsen 2014), c'est-à-dire un document qui, en plus de l'article scientifique classique, comportera *toute l'information requise à la reproduction de celui-ci, une fois les données acquises*. Dans la pratique, ce qui est donc entendu par « reproduction » est tout ce qui vient après la collecte des données ; mais comme l'approche requiert un *accès libre* à celles-ci, elles deviennent critiquables et comparables, constituant ainsi un maillon vers une reproductibilité de l'ensemble du processus, qui prend toute son importance dans une période où la production de connaissance repose de plus en plus sur l'utilisation de bases et banques de données collectées par des tiers et rendues accessibles. Plus explicitement, un document dynamique va donner accès à son lecteur, d'une part, à l'ensemble des données brutes sur lesquelles reposent les résultats présentés, d'autre part, à l'ensemble des codes sources développés spécifiquement pour analyser les données et à une description de nature algorithmique de la façon dont les « codes » ont été appliqués aux données. Tout lecteur, s'il le souhaite, pourra alors régénérer l'ensemble des figures et des tables contenues dans l'article, sous réserve qu'il dispose du même environnement logiciel que les auteurs de l'article[4].

L'intérêt

Indépendamment de la justification philosophique qui met l'accent sur la plus grande adéquation entre un idéal scientifique[5] et une pratique quotidienne, il y a d'excellentes raisons, plus banales, pour adopter une pratique « reproductible », tant au niveau individuel qu'au niveau d'un laboratoire[6]. La première raison touche au problème mentionné ci-dessus de la difficulté d'accès à l'ancienne littérature ; en matière d'analyse de données, une période de six mois peut déjà faire office de temps long et toute personne, à l'exception des plus méticuleuses dans la tenue de leur cahier de laboratoire, sait que reproduire une des *ses propres* figures après un tel délai peut parfois relever du casse-tête. La recherche reproductible ne va pas forcément faire disparaître instantanément les problèmes rencontrés dans ces circonstances, mais elle va permettre d'identifier leurs éventuelles sources – un changement de version d'un logiciel par exemple – de façon beaucoup plus rapide. Notre expérience d'une dizaine d'années avec ce type d'approches montre qu'elles apportent une bien plus grande pérennité au travail du chercheur. Ce qui vaut pour le chercheur « s'observant lui-même » à quelques mois ou années d'écart, vaut d'autant plus pour l'étudiant ou le stagiaire post-doctoral poursuivant le travail d'un de ces prédécesseurs, surtout si celui-ci a déjà quitté le laboratoire. Ainsi la recherche reproductible va automatiquement entraîner une conservation des savoir-faire et, par-là, faciliter leur transmission au sein d'une équipe, d'un laboratoire comme d'un institut.
Convaincu de l'intérêt de la recherche reproductible, le lecteur se demande sans doute comment la mettre en pratique. La recherche reproductible est depuis quelques années un domaine en plein développement et, comme tout domaine en pleine croissance, il se présente au novice, à travers la littérature, sous un jour assez chaotique. Le but de cet article, après avoir brièvement présenté le développement historique de la recherche reproductible, est de fournir une boussole, et une cartographie minimale, utiles au lecteur qui voudrait aller plus loin.

Une brève histoire de la recherche reproductible et de ses outils

La première tentative concrète de mise en œuvre d'« approches reproductibles », *au niveau des*

4. Ce qui implique de décrire cet environnement de façon suffisamment explicite.
5. Idéal qui nécessite – dans une certaine mesure au moins – la reproductibilité.
6. Si nous militons pour une « version forte » de la recherche reproductible comme mode de partage par défaut au sein d'une communauté scientifique, il nous semble important de souligner qu'un chercheur pourrait vouloir publier « comme avant » et néanmoins trouver une approche et des outils intéressants dans le présent article.

publications, est apparue en économie au début des années quatre-vingt (Dewald, Thursby, and Anderson 1986). Le *Journal of Money, Credit and Banking* a alors adopté une politique éditoriale demandant aux auteurs les programmes et données utilisés dans leurs articles « empiriques », ainsi que la mise à disposition de ceux-ci à tout chercheur sur simple demande. Cette mise à disposition s'est néanmoins faite de manière informelle par dépôt des codes et données dans un répertoire (d'ordinateur). Les approches reproductibles proposées par la suite peuvent être vues, en quelque sorte, comme le détournement (ou l'adaptation) d'outils créés dans un but assez différent. Ces outils sont ceux forgés par les informaticiens pour développer des logiciels fiables, bien documentés, faciles à faire évoluer et modifiables par d'autres personnes que leur auteur. Le premier outil est un moteur de production dont l'archétype dans le monde Unix est make : un logiciel, programmé par un langage de script, qui permet d'automatiser et d'ordonnancer la construction / compilation de logiciels « complexes » à partir de fichiers sources. Il est assez simple de remplacer le produit final, un logiciel complexe, par un article au format PDF (via LaTeX) et les compilations intermédiaires par des appels à des logiciels d'analyse de données en mode *batch* (non-interactif) – le résultat de tels appels étant par exemple la génération des figures de l'article. C'est l'idée utilisée par les géophysiciens du *Stanford Exploration Project*[7] (Claerbout and Karrenbach 1992). Au début des années 2000, des statisticiens, Friedrich Leisch et Tony Rossini (Leisch 2002b; Leisch 2002a; Leisch 2003; Rossini and Leisch 2003), se sont inspirés de la « programmation lettrée », proposée par Don Knuth lorsqu'il développait TeX (Knuth 1984). Ils ont ainsi créé la fonction Sweave du logiciel R qui traite un fichier au format texte (ASCII ou UTF-8) où le texte d'un article, écrit avec LaTeX, est mélangé aux lignes de code R qui génèrent les figures et les tables de l'article[8].

Un exemple

Nous avons préparé quelques versions – disponibles sur un dépôt GitHub associé à cet article – d'un cas concret de recherche reproductible dans un contexte qui devrait être proche d'un travail « quotidien » de statistique appliqué : téléchargement de données, chargement de celles-ci dans le logiciel d'analyse, vérifications de la fidélité de l'importation des données par génération de « résumés numériques », construction d'un graphe. Nous avons choisi la reproduction d'un graphe de William Playfair comme illustration. Nous mettons particulièrement l'accent sur la version combinant R et son extension R Markdown ainsi que sur la version combinant Python et le «carnet de notes» (notebook) IPython.

Conclusions

La recherche reproductible est bien plus qu'un *buzzword*, c'est une façon un peu différente de faire ce que le scientifique faisait déjà ; une façon de communiquer plus explicitement et de préserver son travail de façon plus « rationnelle » et plus systématique. C'est une approche encore incomplète dans la mesure où elle ne recouvre pas la phase de génération / collecte des données ; mais si ces dernières sont considérées comme « fixées » – ce qui est le contexte de travail typique des praticiens de la fouille de données –, elles doivent être partagées et deviennent ainsi critiquables et comparables. Il n'y a de cela que quelques années, dix ans tout au plus, pratiquer la recherche reproductible demandait un travail supplémentaire non négligeable au chercheur. Aujourd'hui, avec l'émergence de langages de programmation interactifs comme R et Python, avec des environnements de travail comme RStudio et IPython, avec la généralisation des langages de balisage légers comme pandoc markdown et des logiciels de gestion de version, changer ses habitudes pour rendre son travail quotidien reproductible ne demande guère plus qu'une demi-journée d'auto-formation. Les journaux scientifiques et

7. Le logiciel de recherche reproductible de ce groupe, Madagscar (Fomel and Hennenfent 2007), est maintenant basé sur le moteur de production scons.
8. Sweave recopie la partie texte du fichier d'entrée, telle quelle, dans un nouveau fichier LaTeX, exécute les lignes de codes puis inclut leurs résultats (figures et tables) dans le nouveau fichier.

les agences de financement sont, de leur coté, de plus en plus sensibles aux questions de partage des données et des codes ; les chercheurs ont donc tout intérêt à adopter les pratiques que nous venons d'exposer ; à tel point qu'il serait approprié, nous semble-t-il, de les inclure tôt dans les cursus universitaires. Alors ne ratez pas le train, il démarre maintenant et le prix de la place n'est vraiment pas élevé !

Références

Claerbout, Jon, and Martin Karrenbach. 1992. "Electronic Documents Give Reproducible Research a New Meaning." In *Proceedings of the 62nd Annual Meeting of the Society of Exploration Geophysics*, 601–4.

Collins, Francis S, and Tabak, Lawrence A. 2014 "NIH plans to enhance reproductibility" *Nature* 505 (7485): 612-613

Dewald, William G., Jerry G. Thursby, and Richard G. Anderson. 1986. "Replication in Empirical Economics: The Journal of Money, Credit, and Banking Project." *American Economic Review* 76 (4): 587–603.

Fomel, S., and G. Hennenfent. 2007. "Reproducible Computational Experiments Using Scons." In *Proc. IEEE Int'l Conf. Acoustics, Speech and Signal Processing*, 1257–60. 4.

Hinsen, Konrad. 2014. "Platforms for Publishing and Archiving Computer-Aided Research." F1000Research 3: 289. doi:10.12688/f1000research.5773.1.

Knuth, Donald E. 1984. "Literate Programming." *The Computer Journal* 27 (2): 97–111.

Leisch, Friedrich. 2002a. "Sweave: Dynamic Generation of Statistical Reports Using Literate Data Analysis." In *Compstat* 2002 — *Proceedings in Computational Statistics*, edited by Wolfgang Härdle and Bernd Rönz, 575–80. Physica Verlag, Heidelberg.

Leisch, Friedrich. 2002b. "Sweave, Part I: Mixing R and LaTeX." *R News* 2 (3): 28–31.

Leisch, Friedrich 2003. "Sweave, Part II: Package Vignettes." *R News* 3 (2): 21–24.

Polanyi, Michael. 1998. Personal Knowledge: Towards a Post-Critical Philosophy. Routledge.

Rossini, Anthony, and Friedrich Leisch. 2003. *Literate Statistical Practice*. UW Biostatistics Working Paper Series 194. University of Washington.