Diffuser la culture de la reproductibilité par une formation aux bonnes pratiques: de la qualité d'un projet aux pipelines de données

Lino Galiana* Romain Avouac†

Résumé

L'objet de cette communication est de présenter les choix faits pour diffuser la culture de la reproductibilité à l'INSEE avec `R` à travers une formation ouverte et construite de manière collaborative : https://inseefrlab.github.io/formation-bonnes-pratiques-git-R/

L'objectif de cette formation est de sensibiliser des publics aux niveaux de compétences divers à la culture de la reproductibilité et au partage de projets statistiques s'appuyant sur le langage `R`.

Après avoir proposé des éléments généraux sur la lisibilité d'un code et d'un projet statistique en `R`, cette formation propose une série de choix *opinionated*: quels formats de données privilégier ? comment gérer dans un projet ouvert des éléments de configuration à ne pas partager comme des jetons ? doit-on systématiquement structurer son projet sous forme de package ?

Plutôt que d'insister sur le développement de packages, qui est assez exigeant sur la maintenance de documentation et de tests, cette formation propose plutôt dans le cadre de projets statistiques de privilégier l'apprentissage des environnements virtuels (avec `renv`) et des _pipelines de données (avec `targets`).
Mots-clefs : Bonnes pratiques – Qualité – Packages – renv – Environnements virtuels – Pipelines – targets

Développement

La reproductibilité est une notion centrale pour la recherche scientifique mais aussi dans le cadre d'une organisation amenée à reproduire régulièrement des productions statistiques. Ce concept est néanmoins assez marginal dans beaucoup de formations en statistiques appliquées.

Cette communication propose d'évoquer et discuter certains choix faits dans le cadre d'une formation aux bonnes pratiques pour les projets statistiques construite à l'Insee et dans les services statistiques ministériels.

Cette formation, dont les supports et le code source sont ouverts a été construite de manière collaborative et est disponible sur https://inseefrlab.github.io/formation-bonnes-pratiques-git-R/ (dépôt github : https://github.com/InseeFrLab/formation-bonnes-pratiques-git-R).

Le concept de bonnes pratiques, issu du monde du développement logiciel, est devenu incontournable dans le cadre de projets statistiques du fait de l'évolution de ceux-ci vers une complexification du code, des architectures de traitement ou des livrables attendus des statisticiens et *data scientists*. Ces bonnes pratiques sont désirables puisqu'elles améliorent la communicabilité du code, élément central dans des logiciels open source tout en réduisant la charge de maintenance des projets sans compromettre la qualité de ceux-ci.

Cette formation propose à la fois des éléments généraux et des applications pratiques censées représenter une situation fréquemment rencontrées et assez instructive sur l'intérêt des bonnes pratiques: reprendre et faire évoluer un projet mal structuré est très coûteux.

Après une initiation à `Git` (concepts et cas pratiques dans une démarche collaborative), ce cours propose les éléments suivants:

1. Qualité du code. Evocation des standards de qualité du code (*tidyverse style guide*) et des outils (*linters* et *formaters*) pour les mettre en oeuvre
2. Structure des projets vers la modularité. Les *packages* sont évoqués comme une forme modulaire aboutie mais ne sont pas nécessairement préconisés pour tous les projets

*INSEE, lino.galiana@insee.fr
†INSEE, romain.avouac@insee.fr

3. Format des données. Ce chapitre revient sur quelques formats de données adaptés à la diffusion auprès de statisticiens, comme le format parquet.
4. Environnements reproductibles. Ce chapitre approfondit les implications d'un projet reproductible en évoquant la question du contrôle de l'environnement d'exécution avec le package `renv`
5. Pipelines de données. Ce chapitre évoque l'intérêt de la structuration d'une chaine de traitement sous la forme de pipeline de données et présente la manière dont `targets` permet une gestion très fluide de celle-ci.

En complément du site web du cours, de l'exemple fil rouge de restructuration d'une chaine R pour améliorer sa qualité, les auteurs de la formation ont également enrichi la documentation utilitR de compléments sur les bonnes pratiques, par exemple sur le concept de qualité du code (https://www.book.utilitr.org/02_bonnes_pratiques/01-qualite-code).

Cette formation, parfois `opinionated`, fait quelques préconisations qu'il est intéressant de discuter avec la communauté des praticiens R.