Apprendre R et les statistiques… grâce à R

Philippe Grosjean 1* Guyliann Engels 2†

Résumé

Un apprenant est confronté à plusieurs difficultés lorsqu'il débute dans l'analyse de ses données avec R. Il doit maîtriser un environnement logiciel, un langage de programmation, en même temps que les concepts statistiques. Il doit également apprendre à bien formuler ses questions et à interpréter les résultats obtenus.

Dans le cadre du cours de Science des Données biologiques à l'Université de Mons en Belgique, nous avons développé une approche graduelle qui met en œuvre des outils pédagogiques variés existants (comme {learnr}, {gradethis}, {quarto} ou {ghclass}/GitHub Classroom), mais aussi originaux : {learnitdown} et {learnitgrid}. Nous verrons comment ces différents packages R contribuent à créer un matériel pédagogique interactif et évolutif. Il permet un voyage initiatique qui débute dans un cours en ligne (https://wp.sciviews.org), se poursuit avec des tutoriels {learnr}, pour finalement aborder des projets GitHub/Quarto cadrés avec {learnitgrid}. Au bout du voyage, les étudiants sont capables d'analyser leurs données de manière autonome et de bien communiquer leurs résultats.

Mots-clefs : Apprentissage hybride - Statistique - Science des données - Outils pédagogiques - Cours en ligne interactif

Développement

La pédagogie universitaire classique sous forme de cours en amphithéâtre où les étudiants écoutent passivement suivis de séances d'exercices pratiques a montré ses limites lors des périodes de confinement imposées par le Covid-19. Un enseignement à distance et des classes inversées ou hybrides (présentiel et distanciel à part à peu près égale) en utilisant du matériel pédagogique préenregistré et des exercices interactifs ont alors été développés. Ces approches différentes ont montré leur efficacité et leur intérêt pour les étudiants. Elles ont également permis de réfléchir à une autre pédagogie, notamment dans l'enseignement de la science des données biologiques à l'Université de Mons en Belgique (UMONS).

Nous détaillons dans cette présentation notre approche pédagogique en mode hybride dont l'originalité tient en la combinaison d'exercices en quatre niveaux de difficulté croissante. Ces exercices se basent sur des packages R qui fournissent les ressources nécessaires pour les mettre en œuvre tels {learnr}, {gradethis} et {shiny}. Nous en avons écrit d'autres pour les compléter : {learnitdown}, {learnitdashboard}, {learnitgrid} et {BioDataScience|1|2|3}.

Les quatre cours de science des données à l'UMONS se distribuent de la seconde année universitaire à la cinquième et dernière année du cursus de biologie, donc à cheval sur le Bachelier (la licence en France) et le Master pour un total de 17 crédits ECTS. La matière est découpée en 30 modules qui correspondent à un travail s'étalant sur deux semaines à chaque fois. L'apprentissage est actif et progressif selon quatre niveaux de difficulté croissante.

- **Niveau 1 :** Les étudiants préparent la matière du cours en ligne à leur rythme chez eux et disposent d'exercices H5P (https://h5p.org) et d'applications Shiny pour vérifier leur compréhension des concepts abordés.
- **Niveau 2 :** Ils testent ensuite leurs acquis à l'aide de tutoriels {learnr} et commencent à coder en R dans ces tutoriels interactifs, toujours avant les séances en présentiel. La correction est automatisée à l'aide de {gradethis} de sorte qu'ils ont un retour immédiat et des suggestions pour corriger par eux-mêmes leur code et leurs réponses.

*Service d'écologie numérique, Institut Complexys & Infortech, Université de Mons, Belgique, philippe.grosjean@umons.ac.be
†Service d'écologie numérique, Institut Complexys & Infortech, Université de Mons, Belgique, guyliann.engels@umons.ac.be

Pour les exercices de niveau 1 et 2, l'activité des étudiants est enregistrée dans une base de données de "learning analytics" MongoDB grâce au package R {learnitdown}. La progression dans les exercices est gratifiée de 5% des points sur la note finale. Ce "bonus" insite fortement les étudiants à réaliser tous les exercices chez eux et nous enregistrons un taux de participation de l'ordre de 98%. Bien sûr, cela les incite également à étudier la matière, prérequis obligatoire pour être capable de réaliser ces exercices. Les étudiants arrivent en classe avec leurs questions sur la matière : nous insistons bien sur le fait qu'il est normal qu'ils n'aient pas tout compris et nous travaillons ensemble en présentiel les points à éclaircir.

- **Niveau 3 :** Après la séance de questions-réponses, les étudiants attaquent en présentiel des projets GitHub gérés à l'aide de GitHub Classroom et {ghclass}. Ils doivent analyser des données biologiques et rédiger un rapport en Quarto. À ce niveau, les projets sont individuels et guidés. Cela signifie que les différentes étapes de l'analyse sont suggérées dans le template du rapport Quarto et l'interprétation se fait en sélectionnant les phrases qui conviennent dans une liste à choix multiple. Ces projets sont corrigés de manière semi-automatique avec {learnitgrid}. Ce dernier gère aussi une batterie de tests écrits avec {testthat} que les étudiants peuvent utiliser pour vérifier leurs réponses directement. Un test incorrect est assorti d'un lien cliquable vers des suggestions pour corriger l'erreur. Ces projets comptent pour 10% des points.

- **Niveau 4 :** Des projets GitHub similaires à ceux du niveau 3, mais *non* guidés et réalisés par groupes de deux à quatre étudiants terminent la formation. Les instructions sont ici minimales et aucune batterie de tests ne vient épauler les étudiants. Ils sont donc placés en situation "réelle" à devoir analyser des données biologiques par eux-mêmes. Pour certains projets, ils doivent également choisir des données ouvertes qu'ils souhaitent traiter à partir de sites comme Zenodo (https://zenodo.org), Dryad (https://datadryad.org)… 20% des points sont attribués à ces projets de groupe.

Les projets GitHub sont corrigés par grille critériée avec des commentaires expliquant ce qui peut être amélioré. Pour chaque module, les étudiants travaillent dans les projets pendant deux séances totalisant 6h en présentiel et les complètent ensuite à domicile. À l'issue de ce travail, une interrogation écrite ou un exercice pratique sous forme de challenge (contre la montre ou compétition pour le meilleur modèle, par exemple) permet une évaluation individuelle de la progression et complète la note attribuée pour le module.

Le travail se fait dans RStudio sur le cloud (https://saturncloud.io) de sorte que les étudiants ont tous accès strictement aux mêmes ressources matérielles et à la même configuration logicielle sous Linux, et ce, qu'ils soient en classe ou chez eux. Ils posent leurs questions dans les "issues" des dépôts GitHub de leurs projets et ils ont accès à tout moment à leur progression dans les exercices et les projets (rapport de progression à la volée sous forme d'une application Shiny). L'apprentissage selon cette méthode a fait l'objet d'une publication qui détaille les résultats obtenus sur trois années successives (Engels, Grosjean, and Artus 2023).

Tout le matériel didactique développé dans le cadre de ces cours (y compris des centaines d'exercices et les templates d'une quarantaine de projets) est disponible dans l'organisation GitHub 'BioDataScience-Course' (https://github.com/BioDataScience-Course). Il est distribué sous license MIT ou Attribution-NonCommercial-ShareAlike 4.0 International selon le type de contenu. Le cours en ligne donne accès également à ce matériel dans son contexte à partir de https://wp.sciviews.org. Les packages R cités ici sont accessibles depuis CRAN (https://cran.r-project.org) ou depuis les organisations GitHub 'SciViews' ou 'BioDataScience-Course'.

Référence

Engels, Guyliann, Philippe Grosjean, and Frédérique Artus. 2023. "Teaching Data Science to Students in Biology Using r, RStudio and Learnr: Analysis of Three Years Data." *Foundations of Data Science* 5 (2): 266–85. https://doi.org/10.3934/fods.2022022.