

Travailler avec R

Eric Marcon

02/05/2022



Ce document est réalisé de façon dynamique et reproductible grâce à :

- \LaTeX , dans sa distribution Miktex (<http://miktex.org/>) et la classe memoir (<http://www.ctan.org/pkg/memoir>).
- R (<http://www.r-project.org/>) et RStudio (<http://www.rstudio.com/>)
- bookdown (<http://bookdown.org/>) et memoR (<https://ericmarcon.github.io/memoR/>)

Son code source est sur GitHub : <https://github.com/EricMarcon/travailleR/>.

Le texte mis à jour en continu peut être lu sur <https://ericmarcon.github.io/travailleR/>.

Les versions d'étape sont déposées sur HAL : <https://hal.archives-ouvertes.fr/hal-03022820>.



Photographie en couverture : Hadrien Lalagüe

TABLE DES MATIÈRES

Table des matières	iii
Présentation	ix
Objectifs	ix
Conventions	ix
1 Logiciels	1
1.1 R	1
1.1.1 Installation	1
1.1.2 Rtools	2
1.1.3 Mise à jour	2
1.1.4 Librairies	3
1.2 RStudio	3
1.2.1 Installation	3
1.2.2 Encodage des fichiers	4
1.2.3 Dossier de travail	5
1.2.4 Solution retenue	5
1.3 Packages	6
1.3.1 Installation depuis CRAN	6
1.3.2 Installation depuis GitHub	7
1.3.3 Installation depuis Bioconductor	7
1.3.4 Solution retenue	8
1.4 git et GitHub	9
1.4.1 git	9
1.4.2 GitHub	9
1.4.3 Authentification SSH	10
1.4.4 Obtention d'un jeton d'accès personnel	11
1.5 MiKTeX	11
1.5.1 Installation	12
1.5.2 Mises à jour	12
1.6 Zotero	12
1.7 Go	13
2 Utiliser R	15
2.1 Les langages de R	15
2.1.1 Base	15

TABLE DES MATIÈRES

2.1.2	S3	16
2.1.3	S4	18
2.1.4	RC	19
2.1.5	S6	20
2.1.6	Tidyverse	20
2.2	Environnements	22
2.2.1	Organisation	23
2.2.2	Recherche	24
2.2.3	Espaces de nom des packages	25
2.3	Mesure du temps d'exécution	27
2.3.1	system.time	27
2.3.2	microbenchmark	27
2.3.3	Profilage	29
2.4	Boucles	30
2.4.1	Fonctions vectorielles	30
2.4.2	lapply	31
2.4.3	Boucles for	32
2.4.4	replicate	33
2.4.5	Vectorize	34
2.4.6	Statistiques marginales	34
2.5	Code C++	34
2.6	Paralléliser R	35
2.6.1	mclapply (fork)	35
2.6.2	parLapply (socket)	39
2.6.3	foreach	39
2.7	Etude de cas	41
2.7.1	Création des données	41
2.7.2	Spatstat	41
2.7.3	apply	42
2.7.4	boucle for	44
2.7.5	boucle foreach	45
2.7.6	Rcpp	46
2.7.7	RcppParallel	47
2.7.8	Conclusions sur l'optimisation de la vitesse du code	50
2.8	Flux de travail	50
2.8.1	Principe de fonctionnement	51
2.8.2	Exemple minimal	51
2.8.3	Intérêt pratique	54
3	Git et GitHub	55
3.1	Principes	55
3.1.1	Contrôle de source	55
3.1.2	git et GitHub	56
3.2	Créer un nouveau dépôt	56
3.2.1	A partir d'un projet existant	56
3.2.2	Prendre en compte des fichiers	57

3.2.3	Valider des modifications	57
3.2.4	Créer un dépôt vide sur GitHub	59
3.2.5	Lier git et GitHub	60
3.2.6	Pousser les premières modifications	62
3.2.7	Cloner un dépôt de GitHub	63
3.3	Usage courant	63
3.3.1	Tirer, modifier, valider, pousser	63
3.3.2	Régler les conflits	64
3.3.3	Voir les différences	65
3.3.4	Revenir en arrière	65
3.3.5	Voir l'historique	65
3.4	Branches	66
3.4.1	Créer une nouvelle branche	67
3.4.2	Changer de branche	67
3.4.3	Pousser la nouvelle branche	67
3.4.4	Comportement du système de fichier	67
3.4.5	Fusionner avec <code>merge</code>	68
3.4.6	Fusionner avec une requête de tirage	68
3.5	Usage avancé	69
3.5.1	Commandes de git	69
3.5.2	Taille d'un dépôt	69
3.5.3	Supprimer un dossier	70
3.5.4	Revenir en arrière	72
3.6	Données confidentielles dans un dépôt public	72
3.6.1	Génération d'une paire de clés pour le propriétaire du projet	73
3.6.2	Génération d'une paire de clés pour le projet	73
3.6.3	Création d'un coffre-fort	73
3.6.4	Ajout des utilisateurs	74
3.6.5	Stockage des données	74
3.7	Pages GitHub	75
3.7.1	Activation	75
3.7.2	Badges	76
4	Rédiger	77
4.1	Bloc-note Markdown (R Notebook)	77
4.2	Modèles R Markdown	79
4.3	Articles avec bookdown	80
4.3.1	Ecrire	81
4.3.2	Modèle Simple Article	87
4.3.3	Autres modèles	89
4.4	Présentation Beamer	89
4.5	memoir	90
4.5.1	Créer	90
4.5.2	Ecrire	90
4.5.3	Tricoter	91

TABLE DES MATIÈRES

4.5.4	Finitions	91
4.5.5	Site gitbook	92
4.5.6	Intégration continue	93
4.5.7	Google Analytics	93
4.6	Site web R Markdown	93
4.6.1	Modèle	93
4.6.2	Améliorations	93
4.6.3	Contôle de source	94
4.7	Site web personnel : blogdown	95
4.7.1	Installation des outils	95
4.7.2	Créer	96
4.7.3	Construction du site	97
4.7.4	Site multilingue	98
4.7.5	Paramétrier	98
4.7.6	Ecrire	100
4.7.7	Intégration continue	109
4.7.8	Mises à jour	109
4.8	Exportation de figures	110
4.8.1	Formats vectoriels et raster	110
4.8.2	Fonctions	110
4.8.3	Package ragg	111
4.9	Flux de travail	112
4.9.1	Déclaration du flux	112
4.9.2	Déclaration des cibles	113
4.9.3	Exécution du flux	114
4.9.4	Utilisation des résultats	115
4.9.5	Contrôle de source	115
5	Package	117
5.1	Premier package	118
5.1.1	Création	118
5.1.2	Première fonction	119
5.1.3	Contrôle de source	121
5.1.4	package.R	122
5.2	Organisation du package	122
5.2.1	Fichier DESCRIPTION	122
5.2.2	Fichier NEWS.md	124
5.3	Vignette	124
5.4	pkgdown	125
5.5	Code spécifique aux packages	126
5.5.1	Importation de fonctions	126
5.5.2	Méthodes S3	128
5.5.3	En pratique	130
5.5.4	Code C++	138
5.5.5	Package bien rangé	139
5.6	Bibliographie	139

5.6.1	Préparation	139
5.6.2	Citations	140
5.7	Données	141
5.8	Tests unitaires	141
5.9	Fichier <code>.gitignore</code>	143
5.10	Intégration continue	143
5.11	CRAN	144
5.11.1	Test du package	144
5.11.2	Soumission	144
5.11.3	Maintenance	144
6	Intégration continue	147
6.1	Outils	147
6.1.1	GitHub Actions	147
6.1.2	Codecov	147
6.1.3	GitHub Pages	148
6.2	Principes	148
6.2.1	Obtention d'un jeton d'accès personnel	148
6.2.2	Secrets du projet	148
6.2.3	Activation du dépôt sur CodeCov	149
6.2.4	Scripter les actions de GitHub	149
6.2.5	Données confidentielles dans un dépôt public	155
6.3	Modèles de scripts	155
6.3.1	<code>memoR</code>	156
6.3.2	Projet d'ouvrage	156
6.3.3	Articles et présentations	157
6.3.4	Site web <code>blogdown</code>	158
6.3.5	Packages R	160
6.4	Ajouter des badges	161
7	Shiny	163
7.1	Première application	163
7.2	Application plus élaborée	164
7.2.1	Méthode de travail	164
7.2.2	Exemple	165
7.3	Hébergement	168
8	Enseigner avec R	171
8.1	<code>learnr</code>	171
8.1.1	Premier tutoriel	171
8.1.2	Diffusion	172
8.2	GitHub Classrooms	172
8.2.1	Inscription	172
8.2.2	Organisations	172
8.2.3	Nouvelle salle de classe	173
8.2.4	Préparer un modèle de dépôt	173

TABLE DES MATIÈRES

9 Conclusion	175
Bibliographie	177
Table des figures	179

PRÉSENTATION

Objectifs

Ce document est le support du cours *Travailler avec R*.

Il propose une organisation du travail autour de R et RStudio pour, au-delà des statistiques, rédiger des documents efficacement avec R Markdown, aux formats variés (mémos, articles scientifiques, mémoires d'étudiants, livres, diaporamas), créer son site web et des applications R en ligne (Shiny), produire des packages et utiliser R pour l'enseignement. Il complète *Reproducible Research with R and R Studio* (GANDRUD 2015) par une approche plus concrète, avec des solutions prêtées à l'emploi.

L'optimisation de l'utilisation des nombreux outils disponibles est traitée en détail : **rmarkdown**, **bookdown** et **blogdown** pour la rédaction, **roxygen2**, **test-that** et **pkgdown** pour les packages, le contrôle de source avec git et GitHub, l'intégration continue avec les Actions GitHub etCodecov. Des exemples sont présentés à chaque étape, et le code nécessaire est fourni.

Le chapitre 1 est consacré à l'installation des outils nécessaires : R, git et LaTeX. Le chapitre 2 détaille quelques aspects avancés de l'utilisation de R : les différents langages, les environnements, la performance du code. L'utilisation de base de R n'est pas reprise ici : de bons cours sont suggérés. Le chapitre 3 présente le contrôle de source avec git et GitHub.

Le chapitre 4 montre comment rédiger des documents simples (articles) ou complexes (ouvrages) avec R Markdown, intégrant les données, le code pour les traiter et le texte pour les présenter. Le chapitre 5 présente une méthode pas à pas pour créer efficacement un package. Le chapitre 6 introduit l'utilisation de l'intégration continue pour produire automatiquement des documents, vérifier le code des packages et produire leurs vignettes. Le chapitre 7 présente Shiny, l'outil de mise en ligne d'applications R. Enfin, le chapitre 8 introduit les outils destinés à l'enseignement de et avec R.

Conventions

Les noms des packages sont en gras dans le texte, exemple : **ggplot2**.

L'identifiant utilisé sur GitHub est noté *GitHubID*.

Le signe |> dans le code des exemples indique que la suite du code devrait

TABLE DES MATIÈRES

se trouver sur la même ligne, mais est coupée pour le formatage de ce document. Son usage est limité aux fichiers de configuration YAML, surtout utilisés dans le chapitre 6. Dans tous les autres cas, le code peut être utilisé directement.

LOGICIELS

L'outil central est évidemment R, mais son fonctionnement est aujourd'hui difficilement envisageable sans son environnement de développement RStudio. Pour le contrôle de source, git et GitHub sont de fait les standards. L'ensemble doit être complété par une distribution LaTeX pour la production de documents au format PDF. Un outil de gestion bibliographique est indispensable : Zotero et son extension Better BibTeX sont parfaitement adaptés au cadre de travail présenté ici. Enfin, d'autres logiciels d'usage plus ponctuel peuvent être nécessaires, comme Go.

Leur installation et leur organisation cohérente sont présentées dans ce chapitre.

1.1 R

1.1.1 Installation

R est inclus dans les distributions de Linux : le paquet est nommé `r-base`. Il ne contient pas des outils de développement souvent nécessaires, donc il est préférable d'installer aussi le paquet `r-base-dev`. La version de R est souvent un peu ancienne. Pour disposer de la dernière version, il faut utiliser un miroir de CRAN comme source des paquets : voir la documentation complète pour Ubuntu¹.

Sous Windows ou Mac, installer R après l'avoir téléchargé depuis CRAN².

¹<https://doc.ubuntu-fr.org/r>

²<https://cran.r-project.org/>

1.1.2 Rtools

Sur Mac, l’installation de R est suffisante à partir de la version 4.0.0.

Sous Windows, l’installation doit être complétée par les “Rtools”, qui contiennent les outils de développement dont ceux nécessaires à la compilation des packages contenant du code C++.

Le chemin des Rtools (avant la version 4.2) doit être déclaré à R, en exécutant dans la console de RStudio la commande suivante (adaptée à la version 4.0 des Rtools) :

```
# Rtools : déclaration du chemin,
# nécessite de redémarrer RStudio
writeLines('PATH="${RTOOLS40_HOME}\\\usr\\\bin;${PATH}"',
           con = "~/Renviron")
```

Depuis la version 4.2, cette action est inutile.

Les Rtools doivent être complétés par quelques utilitaires manquants, à installer quand le besoin apparaît (en général, un avertissement de R indiquant que le logiciel n’est pas installé).

La vérification des packages renvoie un avertissement si *qpdf*³ n’est pas installé. Télécharger le fichier zip et coller tout le contenu du dossier bin dans le dossier *usr/bin* de Rtools (*C:\Rtools42\usr\bin* pour la version 4.2).

Un autre avertissement est renvoyé en absence de *Ghostscript*⁴ à télécharger et installer. Copier ensuite le contenu du dossier bin dans le dossier *usr/bin* de Rtools.

1.1.3 Mise à jour

Il est conseillé d’utiliser la dernière version mineure de R : par exemple, 4.0.x jusqu’à la sortie de la version 4.1. Il est obligatoire d’utiliser la toute dernière version pour préparer un package soumis à CRAN.

Des changements importants ont lieu entre les versions majeures (la version 4 ne permet pas d’utiliser un package compilé pour la version 3) mais aussi parfois entre versions mineures (un fichier de données binaires .rda enregistré sous la version 3.3 ne peut pas être lu par la version 3.6). Il est donc utile de mettre R à jour régulièrement.

L’installation d’une nouvelle version ne désinstalle pas automatiquement les versions anciennes, ce qui permet d’en utiliser plusieurs en cas de besoin (par exemple, si un package ancien et indispensable n’est plus disponible). En usage courant, il est préférable de désinstaller manuellement les anciennes versions après l’installation d’une nouvelle.

³<https://sourceforge.net/projects/qpdf/>

⁴<https://www.ghostscript.com/>

1.1.4 Librairies

Les packages de R se trouvent dans deux dossiers :

- la bibliothèque système (*System Library*) contient les packages fournis avec R : **base**, **utils**, **graphics** par exemple. Elle se trouve dans un sous-répertoire du programme d'installation (C:\Program Files\R\R-4.1.0\library pour R version 4.1.0 sous Windows 10).
- La bibliothèque utilisateur (*User Library*) contient ceux installés par l'utilisateur. Jusqu'à la version 4.1, elle se trouve dans le dossier personnel de l'utilisateur, dans un sous-dossier R\win-library\4.1\). Depuis la version 4.2, ce dossier se trouve dans les paramètres locaux de l'utilisateur dont l'emplacement du dossier se trouve dans la variable d'environnement %LOCALAPPDATA%.

Jusqu'à la version 4.1, si le dossier personnel de l'utilisateur est sauvegardé (par exemple, s'il est répliqué dans le cloud par OneDrive sous Windows), il n'est pas optimal d'y placer les packages : le trafic généré par leur sauvegarde serait lourd et inutile. Pour que les packages soient installés automatiquement dans la bibliothèque système, il suffit que l'utilisateur y ait le droit d'écrire. Sous Windows, donner le droit "Modifier" au groupe des utilisateurs de l'ordinateur sur le dossier de la bibliothèque, en plus du droit de lecture par défaut 1.1). A partir de la version 4.2, il n'y a plus de raison de modifier le fonctionnement par défaut : les paramètres locaux ne sont pas sauvegardés.

Si la bibliothèque utilisateur est retenue, il faut penser à vider le dossier correspondant à l'ancienne version de R en cas changement de version mineure.

L'emplacement des librairies est donné par la fonction `.libPaths()` :

```
.libPaths()

## [1] "/Users/runner/work/_temp/Library"
## [2] "/Library/Frameworks/R.framework/Versions/4.2/Resources/library"
```

1.2 RStudio

RStudio est une interface graphique pour R et bien plus : il est conçu pour simplifier la gestion des projets, faciliter la rédaction et la production de documents et intégrer le contrôle de source par exemple.

1.2.1 Installation

Installer la dernière version de *RStudio Desktop* à partir du site de RStudio⁵.

Une commande est disponible dans le menu "Help" de RStudio pour vérifier l'existence d'une version plus récente, à installer.

⁵<https://rstudio.com/products/rstudio/download/>

1. LOGICIELS

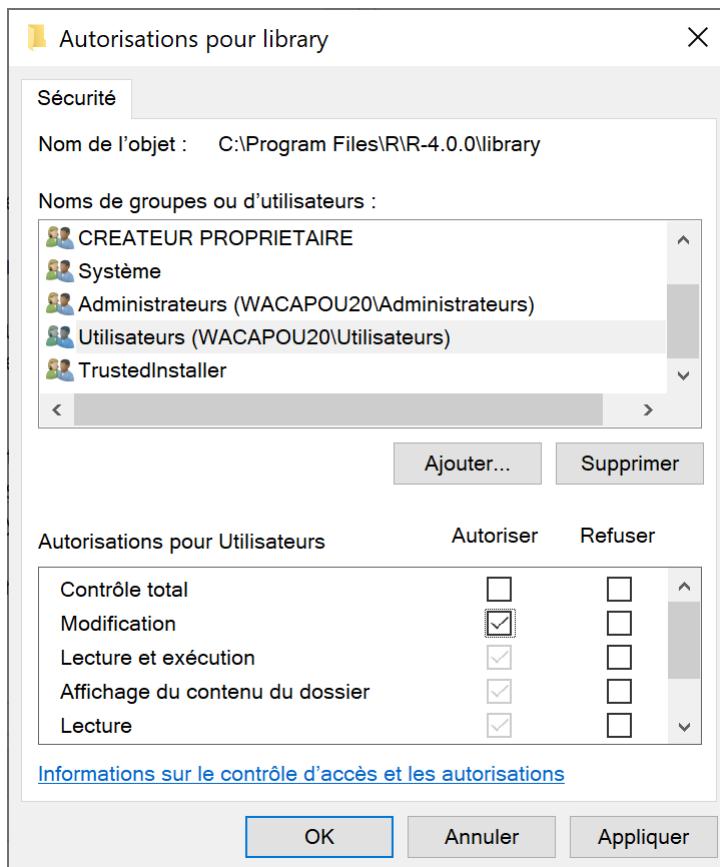


FIG. 1.1 : Activation du droit de modifier la bibliothèque système sous Windows.

1.2.2 Encodage des fichiers

Les fichiers manipulés dans R sont très majoritairement des fichiers texte. Les caractères spéciaux, notamment les accents, peuvent être codés de diverses façons mais la déclaration du codage n'est pas intégrée aux fichiers. Le codage par défaut dépend du système d'exploitation, ce qui pose régulièrement des problèmes de lisibilité des fichiers partagés. Le codage UTF8 est devenu le standard parce qu'il est universellement reconnu et supporte tous les alphabets sans ambiguïtés.

Dès la première utilisation de RStudio, créer un nouveau fichier R (menu “File > New File > R Script”), l'enregistrer au format UTF8 (“File > Save with Encoding...”), choisir UTF8 dans la liste des formats et cocher la case “Set as default encoding for source files”. Supprimer le fichier après l'avoir enregistré.

Les nouveaux fichiers seront codés au format UTF8. Les fichiers codés sous un autre format ne s'afficheront pas correctement : ils pourront être réouverts avec leur codage d'origine (“File > Reopen with Encoding...”), en essayant éventuellement plusieurs codages jusqu'à obtenir un affichage correct, et sauvegardés au format UTF8 ensuite.

1.2.3 Dossier de travail

Le dossier de travail par défaut est le dossier personnel de l'utilisateur, appelé ~ par RStudio :

```
Sys.getenv("R_USER")
```

```
## [1] "
```

- Mes Documents sous Windows ;
- Home sous Mac ou Linux.

Il faut systématiquement travailler dans des sous-dossiers de ~, par exemple : ~\Formation.

Pour le bon fonctionnement des *RTools*, le nom complet du répertoire de travail ne doit pas contenir d'espace (utiliser les tirets bas _) ni de caractère spécial. Le dossier de travail en cours est obtenu par la commande `getwd()`.

```
getwd()
```

```
## [1] "/Users/runner/work/travailleur/travailleur"
```

L'utilisation du contrôle de source (voir chapitre 3) crée de nombreux fichiers de travail. Les projets sous contrôle de source ne devraient pas se trouver dans un dossier déjà sauvegardé par un autre moyen, comme un lecteur OneDrive sous Windows, sous peine d'une utilisation excessive des ressources : chaque validation de modifications engendre la sauvegarde des fichiers modifiés, mais aussi des fichiers de contrôle qui peuvent être de grande taille.

1.2.4 Solution retenue

L'organisation de l'environnement travail est une affaire personnelle, qui dépend des préférences de chacun. L'organisation proposée ici n'est qu'une possibilité, à adapter à ses propres choix, mais en respectant les contraintes mentionnées.

Sous Windows, une organisation optimale est la suivante :

- Dans son dossier personnel (Mes Documents, ~ pour R), un dossier R est utilisé pour les projets simples, sans contrôle de source. La sauvegarde de ce dossier est gérée par ailleurs.
- Un dossier hors du dossier personnel est utilisé pour les projets sous contrôle de source. L'utilisateur doit y avoir le droit d'écrire. Dans l'organisation de Windows, le dossier correspondant à ces critères est %LOCALAPPDATA%, typiquement C:\Users\NomUtilisateur\AppData\Local. Le dossier sera donc %LOCALAPPDATA%\ProjetsR à créer : exécuter `md %LOCALAPPDATA%\ProjetsR` dans une invite de commande.

1. LOGICIELS

Epingler ce dossier à l'accès rapide de l'explorateur de fichiers (figure 1.2) : coller %LOCALAPPDATA%\ProjetsR dans la barre d'adresse de l'explorateur de fichiers, valider, puis faire un clic droit sur “Accès Rapide” et épinglez le dossier.

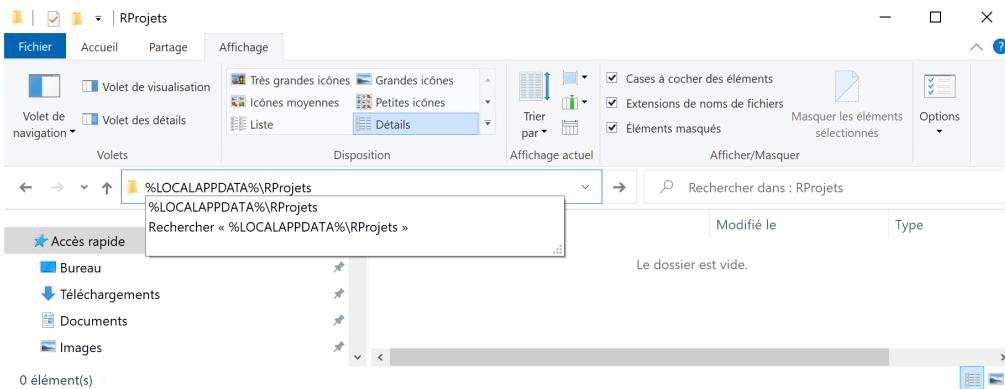


FIG. 1.2 : Dossier pour les projets sous contrôle de source, sous Windows.

1.3 Packages

1.3.1 Installation depuis CRAN

L'installation classique des packages fait appel à CRAN. Un bouton “Install” se trouve dans la fenêtre *Packages* de RStudio.

Les packages sont déposés sur CRAN par leurs auteurs sous forme de code source, compressé dans une fichier .tar.gz. Ils sont disponibles pour le téléchargement dès leur validation. Ils doivent ensuite être mis au format binaire pour Windows (dans un fichier .zip), ce qui prend un peu de temps.

A la demande de l'installation d'un package sous Windows, CRAN propose la version source plutôt que la version binaire si elle est plus récente 1.3).

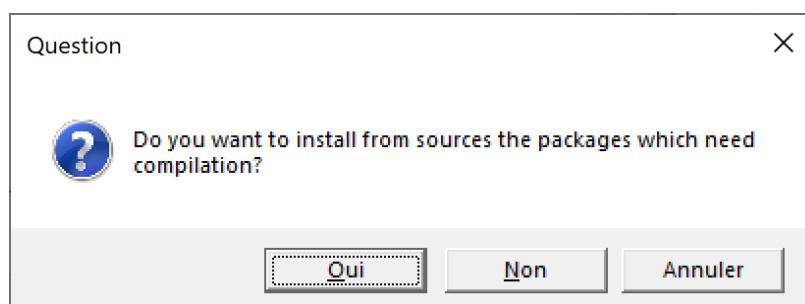


FIG. 1.3 : Activation du droit de modifier la bibliothèque système sous Windows.

La liste des packages concernés est affichée dans la console, par exemple :

```
There are binary versions available but the source
versions are later:
      binary    source needs_compilation
boot       1.3-24   1.3-25        FALSE
class      7.3-16   7.3-17        TRUE
```

Certains packages nécessitent une compilation (colonne `needs_compilation`), en général parce qu'ils contiennent du code C++. Ils ne pourront être installés que par les *Rtools*.

L'installation des packages en version source est beaucoup plus longue qu'en version binaire. Sauf si une version précise d'un package est nécessaire, il est donc préférable de refuser l'installation des versions source.

Les packages peuvent être mis à jour un peu plus tard, après leur compilation par CRAN.

Le bouton “Update” dans la fenêtre *Packages* de RStudio permet de mettre à jour tous les packages installés.

1.3.2 Installation depuis GitHub

Certains packages ne sont pas disponibles sur CRAN mais seulement sur GitHub parce qu'ils sont encore en développement ou parce qu'ils ne sont pas destinés à un large usage par la communauté des utilisateurs de R. Il peut aussi être utile d'installer une version de développement d'un package publié sur CRAN pour un usage ponctuel comme le test de nouvelles fonctionnalités.

L'installation est gérée par le package **remotes**. L'argument `build_vignettes` est nécessaire pour créer les vignettes du package.

```
remotes::install_github("EricMarcon/memoiR", build_vignettes = TRUE)
```

Le nom du package est entré sous la forme “GitHubID/NomduPackage”. L'installation est faite à partir du code source et nécessite donc les *Rtools* si une compilation est nécessaire. `install_github()` vérifie que la version sur GitHub est plus récente que l'éventuelle version installée sur le poste de travail et ne fait rien si elles sont identiques.

1.3.3 Installation depuis Bioconductor

Bioconductor est une plateforme complémentaire de CRAN qui héberge des packages spécialisés dans la génomique. L'installation des packages de Bioconductor nécessite le package **BiocManager** pour sa fonction `install()`. Le premier argument de la fonction est un vecteur de caractères contenant les noms des packages à installer, par exemple :

```
BiocManager::install(c("GenomicFeatures", "AnnotationDbi"))
```

La fonction `install()` appelée sans arguments met à jour les packages.

1.3.4 Solution retenue

A chaque mise à jour mineure de R, tous les packages doivent être réinstallés. La façon la plus efficace de le faire est de créer un script Packages.R à placer dans ~\R. Il contient une fonction qui vérifie si chaque package est déjà installé pour ne pas le refaire inutilement.

```
# Installation des packages de R #####
# Installer les packages si nécessaire #####
InstallPackages <- function(Packages) {
  sapply(Packages, function(Package)
    if (!Package %in% installed.packages() [, 1])
      {install.packages(Package)})
}

# Outils de développement #####
InstallPackages(c(
  # Outils de développement. Importe remotes, etc.
  "devtools",
  # Exécution de Check par RStudio
  "rcmdcheck",
  # Formatage du code R (utilisé par knitr)
  "formatR",
  # Documentation des packages dans /docs sur GitHub
  "pkgdown",
  # Bibliographie avec roxygen
  "Rdpack",
  # Mesure des performances
  "rbenchmark",
  # Documentation automatique des packages
  "roxygen2",
  # Tests des packages
  "testthat"
))

# Markdown #####
InstallPackages(c(
  # Tricot
  "knitr",
  # Documents markdown complexes
  "bookdown",
  # Sites web
  "blogdown",
  # Modèles de documents
  "memoir"
))

# Tidyverse #####
InstallPackages("tidyverse")
```

La dernière partie du script est à compléter avec les packages utilisés régulièrement.

Ce script est à exécuter à chaque mise à jour de R, après avoir éventuellement activé le droit d'écriture dans la librairie système (voir section 1.1.4).

1.4 git et GitHub

1.4.1 git

git est le logiciel de contrôle de source utilisé ici. Son utilisation est détaillée dans le chapitre 3.

Pour Windows et Mac, l'installation a lieu à partir du site web de git⁶.

git est intégré dans les distributions Linux. Pour Ubuntu, le package apt est git-all.

git est installé sans interface graphique, fournie par RStudio.

Dans RStudio, modifier les options globales (menu “Tools > Global Options...”). Sélectionner *Terminal* et l'option *New Terminals open with* : GitBash.

Vérifier la bonne installation de git en tapant la commande git -h dans le terminal de RStudio : l'aide doit s'afficher.

Après l'installation de git, il est possible que le terminal de RStudio ne fonctionne plus correctement et renvoie un message d'erreur contenant les éléments suivants :

```
*** fatal error - cygheap base mismatch detected
This problem is probably due to using incompatible
versions of the cygwin DLL.
```

Le message d'erreur est imprécis : la librairie qui ne doit exister qu'en un seul exemplaire n'est pas cygwin1.dll mais msys-2.0.dll. Rechercher ce fichier dans les dossier d'installation de git et de Rtools. Ils se trouvent normalement dans usr/bin. Remplacer celui de git par celui de Rtools : la version des deux fichiers doit être identique.

Entrer ses informations d'identification en exécutant les commandes suivantes dans le terminal :

```
git config user.name
git config user.email
```

Le nom d'utilisateur est libre, de préférence “Prénom Nom”.

1.4.2 GitHub

GitHub est la plateforme accessible par un [site web](#) qui permet de partager le contenu des dépôts *git*. Pour l'utiliser, il suffit d'ouvrir un compte avec la même adresse de messagerie que celle enregistrée dans git.

Le nom du compte GitHub est noté ici *GitHubID*. Chaque compte GitHub permet d'héberger des dépôts (un dépôt contient les fichiers d'un projet) à l'adresse <https://github.com/GitHubID/NomDuDepot>⁷. Chaque dépôt peut

⁶<https://git-scm.com/>

⁷Exemple : <https://github.com/EricMarcon/travailleR>

1. LOGICIELS

disposer d'un site web à l'adresse <https://GitHubID.github.io/NomDuDepot/>⁸. Enfin, un site web global est prévu pour chaque utilisateur à l'adresse <https://GitHubID.github.io/>⁹.

1.4.3 Authentification SSH

La communication entre git (installé sur l'ordinateur local) et GitHub (plate-forme en ligne) nécessite de s'authentifier.

Deux méthodes sont disponibles : HTTPS (aussi appelée SSL) et SSH. SSH est la plus robuste mais nécessite la création d'une clé privée.

Dans le terminal de RStudio, exécuter :

```
ssh-keygen -t ed25519 -C "user.email"
```

L'adresse de messagerie (qui remplace “user.email”) doit être celle enregistrée dans la configuration de git et le compte GitHub. La clé est enregistrée dans le dossier .ssh du répertoire personnel de l'utilisateur. Il est possible d'ajouter un mot de passe (*passphrase*) à la clé, qui devra être tapé à la première utilisation de chaque session de travail. Si l'ordinateur est correctement sécurisé (pas d'accès physique par des tiers), la laisser vide permet de gagner en fluidité.

Attention : la clé privée est strictement confidentielle et ne doit être copiée nulle part où elle pourrait être lue par un tiers (attention aux sauvegardes automatiques notamment). Elle n'a pas besoin d'être bien sauvegardée : en cas de perte, elle sera remplacée facilement.

Les clés sont normalement stockées dans le dossier ~/.ssh, quel que soit le système d'exploitation, mais l'emplacement du dossier personnel ~ est ambiguë sous Windows : pour R, c'est le dossier Documents, mais pour d'autres logiciels, c'est le dossier racine de l'utilisateur, parent de Documents.

Dans le terminal de RStudio, vérifier le bon fonctionnement de la clé :

```
ssh -T git@github.com
```

Si un message d'erreur indique qu'aucune clé n'est trouvée, deux solutions sont possibles :

- Dupliquer le dossier .ssh (avec l'explorateur de fichiers) dans Documents ;
- Dupliquer le dossier .ssh dans le dossier du programme RStudio (généralement C:\Program Files\RStudio\), dans resources\terminal\bash\.

En cas de succès, un message indique que l'authenticité du serveur GitHub ne peut pas être vérifiée : un contrôle manuel est nécessaire pour la première

⁸Exemple : <https://EricMarcon.github.io/travailleR/>

⁹Exemple : <https://EricMarcon.github.io/>

connexion. Vérifier auprès de GitHub que l’empreinte du serveur est correcte¹⁰ et taper `yes`. Le serveur est ajouté automatiquement à la liste des serveurs connus, dans le fichier `known_hosts`.

Dans le dossier `.ssh`, deux fichiers sont créés : l’un contient la clé privée, l’autre, avec l’extension `.pub`, la clé publique correspondante. Ouvrir le second avec un éditeur de texte et copier la clé publique dans le presse-papier. Sur GitHub, afficher les réglages de son compte (menu “Settings”), sélectionner “SSH and GPG Keys”, cliquer sur “New SSH Key” et coller la clé dans le champ “Key”. Donner un nom à la clé dans le champ “Title”. Le nom peut être celui de l’ordinateur sur lequel la clé a été créée. La clé ne doit pas être copiée sur plusieurs ordinateurs : en cas de besoin, créer une nouvelle clé sur chaque poste de travail utilisé.

Si la clé est compromise (perte ou prêt de l’ordinateur qui la contient), la supprimer sur GitHub et en créer une nouvelle.

1.4.4 Obtention d’un jeton d’accès personnel

L’authentification HTTPS est l’alternative à l’authentification SSL : il faut choisir une méthode et s’y tenir par la suite. Pour utiliser l’authentification HTTPS, la création d’un jeton d’accès personnel est nécessaire.

Les jetons sont créés sur GitHub, dans les paramètres de son compte d’utilisateur, dans “Developer Settings > Personal Access Tokens”¹¹.

Générer un nouveau jeton, le décrire en tant que “git-RStudio” et lui donner l’autorisation “repo”, c’est-à-dire modifier *tous* les dépôts (il n’est pas possible de limiter l’accès à un dépôt particulier). Le jeton est une chaîne de caractère qui ne pourra pas être relue plus tard : elle doit être sauvegardée comme un mot de passe.

1.5 MiKTeX

Pour produire des documents au format PDF, une distribution Latex est nécessaire. La solution légère consiste à installer le package **tinytex** : sa fonction `install_tinytex()` installe une distribution LaTeX optimisée pour R Markdown.

Une distribution complète est préférée ici parce qu’elle permet l’utilisation de LaTeX au-delà de RStudio. MiKTeX¹² est une très bonne solution pour Windows et Mac.

¹⁰<https://docs.github.com/en/github/authenticating-to-github/githubs-ssh-key-fingerprints>

¹¹<https://help.github.com/en/github/authenticating-to-github/creating-a-personal-access-token-for-the-command-line>

¹²<https://miktex.org/download>

1.5.1 Installation

Télécharger le fichier d’installation et l’exécuter. Plusieurs choix sont à faire pendant l’installation :

- Installer le programme pour tous les utilisateurs (avec des droits d’administrateur) ;
- Le format par défaut du papier : choisir A4 ;
- Le mode d’installation des packages manquants : choisir “Always Install” pour qu’ils soient téléchargés automatiquement en cas de besoin.

Pour Linux, suivre les instructions sur le site de MiKTeX.

1.5.2 Mises à jour

MiKTeX est installé avec les packages LaTeX les plus utilisés. Si un document nécessite un package manquant, il est chargé automatiquement. Les mises à jour de packages doivent être faites périodiquement avec la console MiKTeX, accessible dans le menu Démarrer.

Quand elle est lancée sans élévation des privilèges, la console propose de passer en mode administrateur. Cliquer sur “Switch to Administrator mode”.

Dans les paramètres (*Settings*), vérifier que les packages s’installent toujours automatiquement et que le format du papier est bien A4.

Dans le menu des mises à jour (*Updates*), cliquer sur “Check for updates” puis “Update now”.

Si l’installation automatique est défaillante, il est possible d’installer manuellement un package dans le menu “Packages”.

1.6 Zotero

Zotero¹³ est le logiciel de gestion bibliographique le plus utilisé. Ses extensions permettent de compléter ses fonctionnalités selon les besoins de chacun. Better BibTeX permet d’exporter et de maintenir à jour une sélection des références bibliographiques (une collection de Zotero) sous la forme d’un fichier BibTeX dans un projet R, où il pourra être utilisé dans la rédaction de documents ou la documentation de packages.

Télécharger le fichier d’installation et l’exécuter. Créer un compte utilisateur sur le site web de Zotero. Lier l’installation locale au compte : dans le menu “Edition > Préférences”, sélectionner “Synchronisation > Paramètres” et s’authentifier dans la zone “Synchronisation des données”. Cocher ensuite la case “Synchroniser automatiquement” mais pas “Synchroniser le texte intégral des pièces jointes indexées” parce que la taille totale des textes intégraux synchronisés de cette manière entre le compte Zotero en ligne et le poste de travail est limitée à 300 Mo.

¹³<https://www.zotero.org/>

Télécharger l’extension Better BibTeX¹⁴ et l’installer avec le menu “Outils > Extensions” : cliquer sur le bouton des paramètres en haut à droite de la fenêtre, puis “Install Add-on From File...” et sélectionner le fichier qui vient d’être téléchargé.

Paramétrier Better BibTeX à partir du menu “Edition > Préférences > Better BibTeX”. Les options à modifier sont les suivantes :

- “Clés de citation > Format de clé” : [auth:capitalize] [year] pour que les citations disposent d’un identifiant unique de la forme “Nom2021”;
- “Clés de citation > Conserver les clés de citation unique dans” : “Toutes les collections” pour que les identifiants des citations ne soient pas ambigus.
- “Exportation > Gestion des Champs > Champs à exclure de l’exportation” : “abstract, file” pour ne pas générer des fichiers bibliographiques surchargés d’informations inutiles dans les projets R.

Il est conseillé d’utiliser l’extension ZotFile¹⁵ pour mieux contrôler l’emplacement du texte intégral (les fichiers PDF liés au références bibliographiques). L’installer puis la paramétrier dans “Outils > Préférences de ZotFile...> Paramètres généraux > Emplacement des fichiers > Dossier personnalisé” : choisir le dossier de stockage des textes intégraux. Si le dossier personnel de l’utilisateur est sauvegardé (par exemple, s’il est répliqué dans le cloud par OneDrive sous Windows), y placer ce dossier de stockage permet à la fois de sauvegarder les textes intégraux mais aussi d’y accéder à partir de plusieurs postes de travail ou directement en ligne. Cette solution est bien plus efficace que la synchronisation par défaut de Zotero, limitée en volume.

1.7 Go

Go¹⁶ n’est utilisé que par le générateur de sites web Hugo (voir section 4.7).

Télécharger le fichier d’installation et l’exécuter. A la fin de l’installation, exécuter la commande `go version` dans un terminal pour vérifier son bon fonctionnement.

Les mises à jour se font en installant la nouvelle version par dessus la précédente.

¹⁴<https://retorque.re/zotero-better-bibtex/installation/>

¹⁵<http://zotfile.com/>

¹⁶<https://golang.org/>

UTILISER R

La documentation consacrée à l'apprentissage de R est florissante. Les ouvrages suivants sont une sélection arbitraire mais utile pour progresser :

- L'[Introduction à R et au tidyverse](#) (BARNIER 2020) est un excellent cours de prise en main.
- [Advanced R](#) (WICKHAM 2014) est la référence pour maîtriser les subtilités du langage et bien comprendre le fonctionnement de R.
- [R for Data Science](#) (WICKHAM et GROLEMUND 2016) présente une méthode de travail complète, cohérente avec le tidyverse.
- Enfin, [Efficient R programming](#) (GILLESPIE et LOVELACE 2016) traite de l'optimisation du code.

Quelques aspects avancés du codage sont vus ici. Des précisions sur les différents langages de R sont utiles pour la création de packages. Les environnements sont présentés ensuite, pour la bonne compréhension de la recherche des objets appelés par le code. Enfin, l'optimisation des performances du code est traitée en détail (boucles, code C++ et parallélisation) et illustrée par une étude de cas.

2.1 Les langages de R

R comprend plusieurs langages de programmation. Le plus courant est S3, mais ce n'est pas le seul¹.

2.1.1 Base

Le cœur de R est constitué des fonctions primitives et structures de données de base comme la fonction `sum` et les données de type `matrix` :

¹<https://adv-r.had.co.nz/OO-essentials.html>

2. UTILISER R

```
pryr::otype(sum)  
  
## [1] "base"  
  
typeof(sum)  
  
## [1] "builtin"  
  
pryr::otype(matrix(1))  
  
## [1] "base"  
  
typeof(matrix(1))  
  
## [1] "double"
```

Le package **pryr** permet d'afficher le langage dans lequel des objets sont définis. La fonction `typeof()` affiche le type de stockage interne des objets :

- la fonction `sum()` appartient au langage de base de R et est une fonction primitive (*builtin*) ;
- les éléments de la matrice numérique contenant un seul 1 sont des réels à double précision, et la matrice elle-même est définie dans le langage de base.

Les fonctions primitives sont codées en C et sont très rapides. Elles sont toujours disponibles, quels que soient les packages chargés. Leur usage est donc à privilégier.

2.1.2 S3

S3 est le langage le plus utilisé, souvent le seul connu par les utilisateurs de R.

C'est un langage orienté objet dans lequel les classes, c'est-à-dire le type des objets, sont déclaratives.

```
MonPrenom <- "Eric"  
class(MonPrenom) <- "Prenom"
```

La variable `MonPrenom` est ici de classe “Prenom” par une simple déclaration.

Contrairement au fonctionnement d'un langage orienté objet classique², les méthodes S3 sont liées aux fonctions, pas aux objets.

²<https://www.troispointzero.fr/le-blog/introduction-a-la-programmation-orientee-objet-poo/>

```
# Affichage par défaut
MonPrenom

## [1] "Eric"
## attr(,"class")
## [1] "Prenom"

print.Prenom <- function(x) cat("Le prénom est", x)
# Affichage modifié
MonPrenom

## Le prénom est Eric
```

Dans cet exemple, la méthode `print()` appliquée à la classe “Prenom” est modifiée. Dans un langage orienté objet classique, la méthode serait définie dans la classe `Prenom`. Dans R, les méthodes sont définies à partir de méthodes génériques.

`print` est une méthode générique (“un générique”) déclaré dans `base`.

```
pryr::otype(print)

## [1] "base"
```

Son code se résume à une déclaration `UseMethod("print")` :

```
print

## function (x, ...)
## UseMethod("print")
## <bytecode: 0x7fc287b0ee08>
## <environment: namespace:base>
```

Il existe beaucoup de méthodes S3 pour `print` :

```
head(methods("print"))

## [1] "print.acf"      "print.AES"       "print.all_vars"
## [4] "print.anova"    "print.any_vars"  "print.aov"
```

Chacune s’applique à une classe. `print.default` est utilisée en dernier ressort et s’appuie sur le type de l’objet, pas sur sa classe S3.

```
typeof(MonPrenom)
```

```
## [1] "character"
```

```
pryr::otype(MonPrenom)
```

```
## [1] "S3"
```

2. UTILISER R

Un objet peut appartenir à plusieurs classes, ce qui permet une forme d'héritage des méthodes. Dans un langage orienté objet classique, l'héritage permet de définir des classes plus précises ("PrenomFrancais") qui héritent de classes plus générales ("Prenom") et bénéficient de cette façon de leurs méthodes sans avoir à les redéfinir. Dans R, l'héritage consiste simplement à déclarer un vecteur de classes de plus en plus larges pour un objet :

```
# Définition des classes par un vecteur
class(MonPrenom) <- c("PrenomFrancais", "Prenom")
# Ecriture alternative, avec inherits()
inherits(MonPrenom, what = "PrenomFrancais")
```

```
## [1] TRUE
```

```
inherits(MonPrenom, what = "Prenom")
```

```
## [1] TRUE
```

Le générique cherche une méthode pour chaque classe, dans l'ordre de leur déclaration.

```
print.PrenomFrancais <- function(x) cat("Prénom français:", 
  x)
MonPrenom
```

```
## Prénom français: Eric
```

En résumé, S3 est le langage courant de R. Presque tous les packages sont écrits en S3. Les génériques sont partout mais passent inaperçus, par exemple dans des packages :

```
library("entropart")
.S3methods(class = "SpeciesDistribution")

## [1] autoplot plot
## see '?methods' for accessing help and source code
```

La fonction `.S3methods()` permet d'afficher toutes les méthodes disponibles pour une classe, par opposition à `methods()` qui affiche toutes les classes pour lesquelles la méthode passée en argument est définie.

De nombreuses fonctions primitives de R sont des méthodes génériques. Utiliser l'aide `help(InternalMethods)` pour les découvrir.

2.1.3 S4

S4 est une évolution de S3 qui structure les classes pour se rapprocher d'un langage orienté objet classique :

- les classes doivent être définies explicitement, pas simplement déclarées ;

- les attributs (c'est-à-dire les variables décrivant les objets), appelés *slots*, sont déclarés explicitement ;
- le constructeur, c'est-à-dire la méthode qui crée un nouvelle instance d'une classe (c'est-à-dire une variable contenant un objet de la classe), est explicite.

En reprenant l'exemple précédent, la syntaxe S4 est la suivante :

```
# Définition de la classe Personne, avec ses slots
setClass("Personne",
         slots = list(Nom = "character", Prenom = "character"))
# Construction d'une instance
Moi <- new("Personne", Nom = "Marcon", Prenom = "Eric")
# Langage
pryr::otype(Moi)

## [1] "S4"
```

Les méthodes appartiennent toujours aux fonctions. Elles sont déclarées par la fonction `setMethod()` :

```
setMethod("print", signature = "Personne", function(x, ...) {
  cat("La personne est:", x@Prenom, x@Nom)
})
print(Moi)

## La personne est: Eric Marcon
```

Les attributs sont appelés par la syntaxe `variable@slot`.

En résumé, S4 est plus rigoureux que S3. Quelques packages sur CRAN : **Matrix**, **sp**, **odbc**... et beaucoup sur Bioconductor sont écrits en S4 mais le langage est maintenant clairement délaissé au profit de S3, notamment à cause du succès du **tidyverse**.

2.1.4 RC

RC a été introduit dans R 2.12 (2010) avec le package **methods**.

Les méthodes appartiennent aux classes, comme en C++ : elles sont déclarées dans la classe et appelées à partir des objets.

```
library("methods")
# Déclaration de la classe
PersonneRC <- setRefClass("PersonneRC",
  fields = list(Nom = "character", Prenom = "character"),
  methods = list(print = function() cat(Prenom, Nom)))
# Constructeur
MoiRC <- new("PersonneRC", Nom = "Marcon", Prenom ="Eric")
# Langage
pryr::otype(MoiRC)

## [1] "RC"
```

2. UTILISER R

```
# Appel de la méthode print  
MoiRC$print()
```

```
## Eric Marcon
```

RC est un langage confidentiel, bien que ce soit le premier “vrai” langage orienté objet de R.

2.1.5 S6

S6³ perfectionne RC mais n'est pas inclus dans R : il nécessite d'installer son package.

Les attributs et les méthodes peuvent être publics ou privés. Une méthode `initialize()` est utilisée comme constructeur.

```
library(R6)  
PersonneR6 <- R6Class("PersonneR6", public = list(Nom = "character",  
                                                 Prenom = "character", initialize = function(Nom = NA, Prenom = NA) {  
                                                 self$Nom <- Nom  
                                                 self$Prenom <- Prenom  
                                                 }, print = function() cat(self$Prenom, self$Nom)))  
MoiR6 <- PersonneR6$new(Nom = "Marcon", Prenom = "Eric")  
MoiR6$print()
```

```
## Eric Marcon
```

S6 permet de programmer rigoureusement en objet mais est très peu utilisé. Les performances de S6 sont bien supérieures à celles de RC mais sont inférieures à celles de S3⁴.

La non-inclusion de R6 à R est montrée par `pryr` :

```
pryr::otype(MoiR6)
```

```
## [1] "S3"
```

2.1.6 Tidyverse

Le tidyverse est un ensemble de packages cohérents qui ont fait évoluer la façon de programmer R. L'ensemble des packages indispensables peut être chargé par le package **tidyverse** qui n'a pas d'autre utilité :

```
library("tidyverse")
```

Il ne s'agit pas d'un nouveau langage à proprement parler mais plutôt d'une extension de S3, avec de profondes modifications techniques, notamment l'évaluation non conventionnelle des expressions⁵, qu'il n'est pas essentiel de maîtriser en détail.

³<https://r6.r-lib.org/>

⁴<https://r6.r-lib.org/articles/Performance.html>

⁵<https://dplyr.tidyverse.org/articles/programming.html>

Ses principes sont inscrits dans un manifeste⁶. Son apport le plus visible pour l'utilisateur sont l'enchaînement des commandes dans un flux (pipeline de code).

En programmation standard, l'enchaînement des fonctions s'écrit par emboîtements successifs, ce qui en rend la lecture difficile, surtout quand des arguments sont nécessaires :

```
# Logarithme de base 2 de la moyenne de 100 tirages
# aléatoires dans une loi uniforme
log(mean(runif(100)), base = 2)
```

```
## [1] -1.127903
```

Dans le tidyverse, les fonctions s'enchaînent, ce qui correspond souvent mieux à la réflexion du programmeur sur le traitement des données :

```
# 100 tirages aléatoires dans une loi uniforme
runif(100) %>%
  # Moyenne
  mean %>%
  # Logarithme
  log(base=2)
```

```
## [1] -0.9772102
```

Le tuyau `%>%` est un opérateur qui appelle la fonction suivante en lui passant comme premier argument le résultat de la fonction précédente. Les arguments supplémentaires sont passés normalement : pour la lisibilité du code, il est indispensable de les nommer. La plupart des fonctions de R sont utilisables sans difficultés dans le tidyverse bien qu'elles n'aient pas été prévues pour cela : il suffit que leur premier argument soit les données à traiter.

Le pipeline ne permet de passer qu'une seule valeur à la fonction suivante, ce qui interdit les fonctions multidimensionnelles, de type `f(x, y)`. La structure de données préférée est le *tibble*, qui est un dataframe amélioré : sa méthode `print()` est plus lisible, et il corrige quelques comportements non-intuitifs des dataframes, comme la conversion automatique en vecteurs des dataframes à une seule colonne. Les colonnes du dataframe ou du tibble permettent de passer autant de données que nécessaire.

Enfin, la visualisation des données est prise en charge par **ggplot2** qui s'appuie sur une grammaire des graphiques (WICKHAM 2010) solide sur le plan théorique. Schématiquement, un graphique est construit selon le modèle suivant :

```
ggplot(data = <DATA>) +
  <GEOM_FUNCTION>(
    mapping = aes(<MAPPINGS>),
    stat = <STAT>,
    position = <POSITION>
  ) +
  <COORDINATE_FUNCTION> +
  <FACET_FUNCTION>
```

⁶<https://cran.r-project.org/web/packages/tidyverse/vignettes manifesto.html>

2. UTILISER R

- les données sont obligatoirement un dataframe ;
- la géométrie est le type de graphique choisi (points, lignes, histogrammes ou autre) ;
- l'esthétique (fonction `aes()`) désigne ce qui est représenté : c'est la correspondance entre les colonnes du dataframe et les éléments nécessaires à la géométrie ;
- la statistique est le traitement appliqué aux données avant de les transmettre à la géométrie (souvent “identity”, c'est-à-dire aucune transformation mais “boxplot” pour une boîte à moustache). Les données peuvent être transformées par une fonction d'échelle, comme `scale_y_log10()` ;
- la position est l'emplacement des objets sur le graphique (souvent “identity” ; “stack” pour un histogramme empilé, “jitter” pour déplacer légèrement les points superposés dans un `geom_point()`) ;
- les coordonnées définissent l'affichage du graphique (`coord_fixed()` pour ne pas déformer une carte par exemple) ;
- enfin, les facettes offrent la possibilité d'afficher plusieurs aspects des mêmes données en produisant un graphique par modalité d'une variable.

L'ensemble formé par le pipeline et **ggplot2** permet des traitements complexes dans un code lisible. La figure 2.1 montre le résultat du code suivant :

```
# Données sur les diamants fournies par ggplot2
diamonds %>%
  # Ne conserver que les diamants de plus d'un demi-carat
  filter(carat > 0.5) %>%
  # Graphique : prix en fonction du poids
  ggplot(aes(x = carat, y = price)) +
    # Nuage de points
    geom_point() +
    # Echelle logarithmique
    scale_x_log10() +
    scale_y_log10() +
    # Régression linéaire
    geom_smooth(method = "lm")
```

Dans cette figure, deux géométries (nuage de points et régression linéaire) partagent la même esthétique (prix en fonction du poids en carats) qui est donc déclarée en amont, dans la fonction `ggplot()`.

Le tidyverse est documenté en détail dans WICKHAM et GROLEMUND (2016) et **ggplot2** dans WICKHAM (2017).

2.2 Environnements

Les objets de R, données et fonctions, sont nommés. Comme R est modulaire, avec la possibilité de lui ajouter un nombre indéterminé de packages, il est très probable que des conflits de nom apparaissent. Pour les régler, R dispose d'un système rigoureux de précédence des noms : le code s'exécute dans un environnement défini, héritant d'environnements parents.

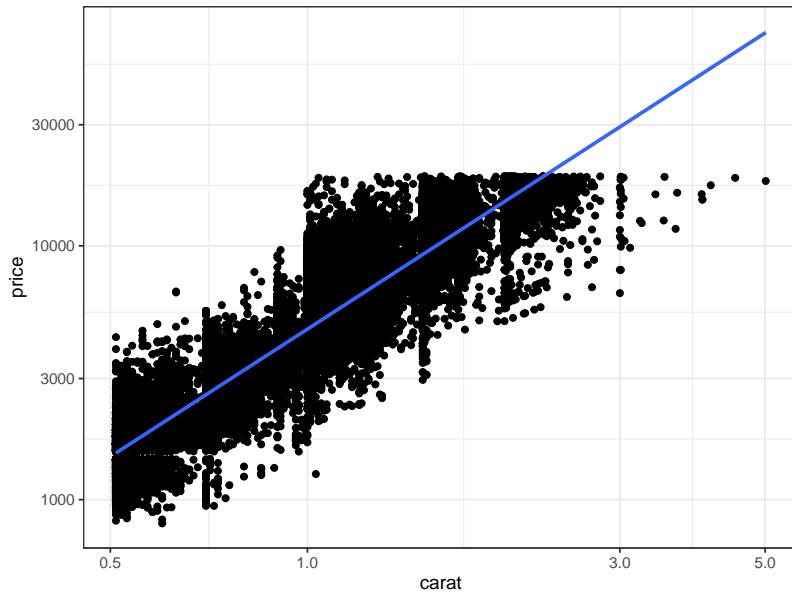


FIG. 2.1 : Prix des diamants en fonction de leur poids. Démonstration du code de **ggplot2** combiné au traitement de données du tidyverse.

2.2.1 Organisation

R démarre dans un environnement vide. Chaque package chargé crée un environnement fils pour former une pile des environnements, dont chaque nouvel élément est appelé “fils” du précédent, qui est son “parent”.

La console se trouve dans l’environnement global, fils du dernier package chargé.

```
search()

## [1] ".GlobalEnv"      "package:R6"
## [3] "package:entropart" "package:forcats"
## [5] "package:stringr"   "package:dplyr"
## [7] "package:purrr"     "package:readr"
## [9] "package:tidyverse" "package:tibble"
## [11] "package:ggplot2"   "package:tidyverse"
## [13] "package:kableExtra" "package:stats"
## [15] "package:graphics"   "package:grDevices"
## [17] "package:utils"      "package:datasets"
## [19] "package:methods"    "Autoloads"
## [21] "package:base"
```

Le code d’une fonction appelée de la console s’exécute dans un environnement fils de l’environnement global :

```
# Environnement actuel
environment()
```

```
## <environment: R_GlobalEnv>
```

2. UTILISER R

```
# La fonction f affiche son environnement
f <- function() environment()
# Affichage de l'environnement de la fonction
f()
```

```
## <environment: 0x7fc294d64650>
```

```
# Environnement parent de celui de la fonction
parent.env(f())
```

```
## <environment: R_GlobalEnv>
```

2.2.2 Recherche

La recherche des objets commence dans l'environnement local. S'il n'est pas trouvé, il est cherché dans l'environnement parent, puis dans le parent du parent, jusqu'à l'épuisement des environnements qui génère une erreur indiquant que l'objet n'a pas été trouvé.

Exemple :

```
# Variable q définie dans l'environnement global
q <- "GlobalEnv"
# Fonction définissant q dans son environnement
qLocalFonction <- function() {
  q <- "Fonction"
  return(q)
}
# La variable locale est retournée
qLocalFonction()
```

```
## [1] "Fonction"
```

```
# Fonction (mal écrite) utilisant une variable qu'elle ne
# définit pas
qGlobalEnv <- function() {
  return(q)
}
# La variable de l'environnement global est retournée
qGlobalEnv()
```

```
## [1] "GlobalEnv"
```

```
# Suppression de cette variable
rm(q)
# La fonction base::q est retournée
qGlobalEnv()
```

```
## function (save = "default", status = 0, runLast = TRUE)
## .Internal(qt(save, status, runLast))
## <bytecode: 0x7fc276863a18>
## <environment: namespace:base>
```

La variable `q` est définie dans l'environnement global. La fonction `qLocalFonction` définit sa propre variable `q`. L'appel de la fonction retourne la valeur locale de la fonction parce qu'elle se trouve dans l'environnement de la fonction.

La fonction `qGlobalEnv` retourne la variable `q` qu'elle ne définit pas localement. Elle la recherche donc dans son environnement parent et trouve la variable définie dans l'environnement global. En supprimant la variable de l'environnement global par `rm(q)`, la fonction `qGlobalEnv()` parcourt la pile des environnements jusqu'à trouver un objet nommé `q` dans le package `base`, qui est la fonction permettant de quitter R. Elle aurait pu trouver un autre objet si un package contenant un objet `q` avait été chargé.

Pour éviter ce comportement erratique, une fonction ne doit *jamais* appeler un objet non défini dans son propre environnement.

2.2.3 Espaces de nom des packages

Il est temps de définir précisément ce que les packages rendent visible. Les packages contiennent des objets (fonctions et données) qu'ils *exportent* ou non. Ils sont habituellement appelés par la fonction `library()` qui effectue deux opérations :

- elle *charge* le package en mémoire, ce qui permet d'accéder à tous ses objets avec la syntaxe `package::objet` pour les objets exportés et `package:::objet` pour ceux qui ne le sont pas ;
- elle *attache* ensuite le package, ce qui place son environnement en haut de la pile.

Il est possible de détacher un package avec la fonction `unloadNamespace()` pour le retirer de la pile des environnements. Exemple :

```
# entropart chargé et attaché
library("entropart")
# Est-il attaché ?
isNamespaceLoaded("entropart")

## [1] TRUE

# Pile des environnements
search()

## [1] ".GlobalEnv"           "package:R6"
## [3] "package:entropart"    "package:forcats"
## [5] "package:stringr"       "package:dplyr"
## [7] "package:purrr"         "package:readr"
## [9] "package:tidyverse"      "package:tibble"
## [11] "package:ggplot2"        "package:tidyverse"
## [13] "package:kableExtra"     "package:stats"
## [15] "package:graphics"       "package:grDevices"
## [17] "package:utils"          "package:datasets"
## [19] "package:methods"        "Autoloads"
## [21] "package:base"
```

2. UTILISER R

```
# Diversity(), une fonction exportée par entropart est
# trouvée
Diversity(1, CheckArguments = FALSE)

## None
##      1

# Détailler et décharger entropart
unloadNamespace("entropart")
# Est-il attaché ?
isNamespaceLoaded("entropart")

## [1] FALSE

# Pile des environnements, sans entropart
search()

## [1] ".GlobalEnv"          "package:R6"
## [3] "package:forcats"     "package:stringr"
## [5] "package:dplyr"        "package:purrr"
## [7] "package:readr"         "package:tidyverse"
## [9] "package:tibble"        "package:ggplot2"
## [11] "package:tidyverse"     "package:kableExtra"
## [13] "package:stats"         "package:graphics"
## [15] "package:grDevices"     "package:utils"
## [17] "package:datasets"      "package:methods"
## [19] "Autoloads"             "package:base"

# Diversity() est introuvable
tryCatch(Diversity(1), error = function(e) print(e))

## <simpleError in Diversity(1): could not find function "Diversity">

# mais peut être appelée avec son nom complet
entropart::Diversity(1, CheckArguments = FALSE)

## None
##      1
```

L'appel de `entropart::Diversity()` charge le package (c'est-à-dire, exécute implicitement `loadNamespace("entropart")`) mais ne l'attache pas.

En pratique, il faut limiter le nombre de package attachés pour limiter le risque d'appeler une fonction non désirée, homonyme de la fonction recherchée. Dans les cas critiques, il faut utiliser le nom complet de la fonction : `package::fonction()`.

Un problème fréquent concerne la `filter()` de `dplyr` homonyme de celle de `stats`. Le package `stats` est habituellement chargé avant `dplyr`, un package du tidyverse. `stats::filter()` doit donc être appelée explicitement.

Cependant, le package `dplyr` ou `tidyverse` (qui attache tous les packages du tidyverse) peut être chargé systématiquement en créant un fichier `.RProfile` à la racine du projet contenant la commande :

```
library("tidyverse")
```

Dans ce cas, **dplyr** est chargé *avant stats* et c'est sa fonction qui est inaccessible.

2.3 Mesure du temps d'exécution

Le temps d'exécution d'un code long peut être mesuré très simplement par la commande `system.time`. Pour des temps d'exécution très courts, il est nécessaire de répéter la mesure : c'est l'objet du package **microbenchmark**.

2.3.1 system.time

La fonction retourne le temps d'exécution du code.

```
# Ecart absolu moyen de 1000 valeurs dans une loi uniforme,
# répété 100 fois
system.time(for (i in 1:100) mad(runif(1000)))
```

```
##    user  system elapsed
##  0.023   0.000   0.023
```

2.3.2 microbenchmark

Le package **microbenchmark** est le plus avancé.

L'objectif est de comparer la vitesse du calcul du carré d'un vecteur (ou d'un nombre) en le multipliant par lui-même ($x \times x$) ou en l'élevant à la puissance 2 (x^2).

```
# Fonctions à tester
f1 <- function(x) x * x
f2 <- function(x) x^2
f3 <- function(x) x^2.1
f4 <- function(x) x^3
# Initialisation
X <- rnorm(10000)
# Test
library("microbenchmark")
(mb <- microbenchmark(f1(X), f2(X), f3(X), f4(X)))
```

```
## Unit: microseconds
##   expr     min      lq      mean    median      uq      max neval
##   f1(X) 38.321 41.6150 57.95106 43.6350 46.2780 1286.237   100
##   f2(X) 47.596 49.8990 67.90378 51.4305 56.2845 1420.475   100
##   f3(X) 276.623 281.2775 304.26661 285.8795 292.2925 1851.762   100
##   f4(X) 411.144 415.0140 443.44620 421.8570 433.9420 1993.775   100
```

2. UTILISER R

Le tableau retourné contient les temps minimum, médian, moyen, max et les premiers et troisièmes quartiles, ainsi que le nombre de répétitions. La valeur médiane est à comparer. Le nombre de répétition est par défaut de 100, à moduler (argument `times`) en fonction de la complexité du calcul.

Le résultat du test, un objet de type `microbenchmark`, est un tableau brut des temps d'exécution. L'analyse statistique est faite par les méthodes `print` et `summary`. Pour choisir les colonnes à afficher, utiliser la syntaxe suivante :

```
summary(mb) [, c("expr", "median")]
```

```
##     expr    median
## 1 f1(X) 43.6350
## 2 f2(X) 51.4305
## 3 f3(X) 285.8795
## 4 f4(X) 421.8570
```

Pour faire des calculs sur ces résultats, il faut les stocker dans une variable. Pour empêcher l'affichage dans la console, la solution la plus simple est d'utiliser la fonction `capture.output` en affectant son résultat à une variable.

```
dummy <- capture.output(mbs <- summary(mb))
```

Le test précédent est affiché à nouveau.

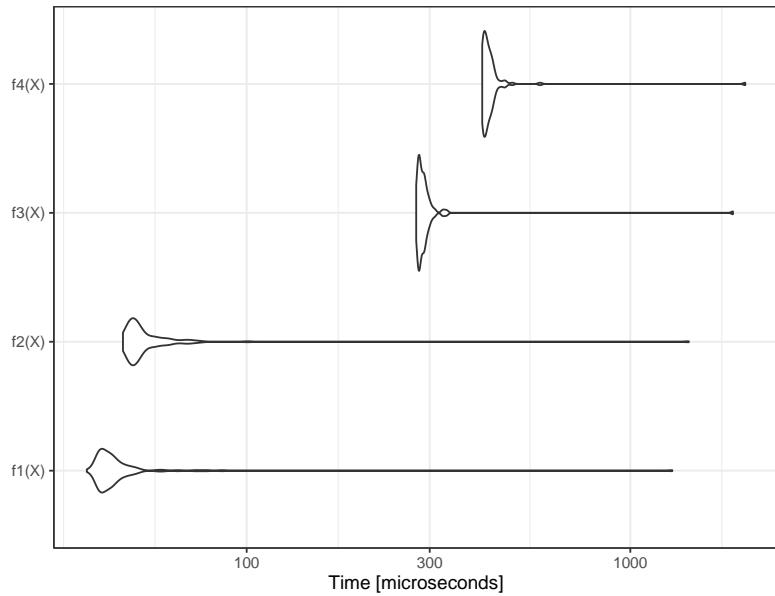
```
summary(mb) [, c("expr", "median")]
```

```
##     expr    median
## 1 f1(X) 43.6350
## 2 f2(X) 51.4305
## 3 f3(X) 285.8795
## 4 f4(X) 421.8570
```

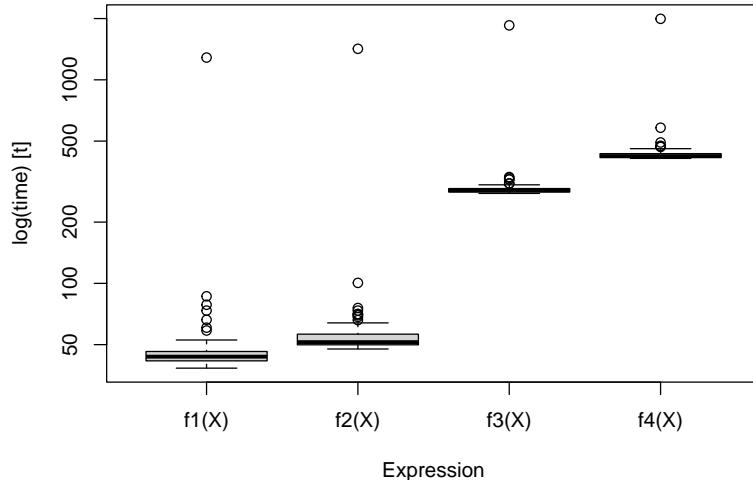
Le temps de calcul est à peu près identique entre $x \times x$ et x^2 . Le calcul de puissance est nettement plus long, surtout si la puissance n'est pas entière, parce qu'il nécessite un calcul de logarithme. Le calcul de la puissance 2 est donc optimisé par R pour éviter l'usage du log.

Deux représentations graphiques sont disponibles : les violons représentent la densité de probabilité du temps d'exécution ; les boîtes à moustache sont classiques.

```
library("ggplot2")
autoplot(mb)
```



```
boxplot(mb)
```



2.3.3 Profilage

`profvis` est l'outil de profilage de RStudio.

Il permet de suivre le temps d'exécution de chaque ligne de code et la mémoire utilisée. L'objectif est de détecter les portions de code lentes, à améliorer.

```
library(profvis)
p <- profvis({
  # Calculs de cosinus
  cos(runif(10^7))
  # 1/2 seconde de pause
```

2. UTILISER R

```
    pause(1/2)
})
htmlwidgets::saveWidget(p, "docs/profile.html")
```

Le résultat est un fichier HTML contenant le rapport de profilage⁷. On peut observer que le temps de tirage des nombres aléatoires est similaire à celui du calcul des cosinus.

Lire la documentation complète⁸ sur le site de RStudio.

2.4 Boucles

Le cas le plus fréquent de code long à exécuter est celui des boucles : le même code est répété un grand nombre de fois.

2.4.1 Fonctions vectorielles

La plupart des fonctions de R sont vectorielles : les boucles sont traitées de façon interne, extrêmement rapide. Il faut donc raisonner en termes de vecteurs plutôt que de scalaires.

```
# Tirage de deux vecteurs de trois nombres aléatoires entre
# 0 et 1
x1 <- runif(3)
x2 <- runif(3)
# Racine carrée des trois nombres de x1
sqrt(x1)
```

```
## [1] 0.9427738 0.8665204 0.4586981
```

```
# Sommes respective des trois nombres de x1 et x2
x1 + x2
```

```
## [1] 1.6262539 1.6881583 0.9063973
```

Il faut aussi écrire des fonctions vectorielles sur leur premier argument. La fonction `lnq` du package **entropart** retourne le logarithme déformé d'ordre q d'un nombre x .

```
# Code de la fonction
entropart::lnq
```

```
## function (x, q)
## {
##   if (q == 1) {
##     return(log(x))
##   }
##   else {
##     Log <- (x^(1 - q) - 1)/(1 - q)
```

⁷<https://EricMarcon.github.io/travailleR/profile.html>

⁸<https://rstudio.github.io/profvis/>

```

##           Log[x < 0] <- NA
##           return(Log)
##       }
## }
## <bytecode: 0x7fc276ac29e8>
## <environment: namespace:entropart>

```

Pour qu'une fonction soit vectorielle, chaque ligne de son code doit permettre que le premier argument soit traité comme un vecteur. Ici : `log(x)` et `x^` sont une fonction et un opérateur vectoriels et la condition `[x < 0]` retourne aussi un vecteur.

2.4.2 lapply

Les codes qui ne peuvent pas être écrits comme une fonction vectorielle nécessitent des boucles.

`lapply()` applique une fonction à chaque élément d'une liste. Elle est déclinée sous plusieurs versions :

- `lapply()` renvoie une liste (économise le temps de leur réorganisation dans un tableau) ;
- `sapply()` renvoie un dataframe en rassemblant les listes par `simplify2array()` ;
- `vapply()` est presque identique mais demande que le type de données du résultat soit fourni.

```

# Tirage de 1000 valeurs dans une loi uniforme
x1 <- runif(1000)
# La racine carrée peut être calculée pour le vecteur ou
# chaque valeur
identical(sqrt(x1), sapply(x1, FUN = sqrt))

## [1] TRUE

mb <- microbenchmark(sqrt(x1), lapply(x1, FUN = sqrt), sapply(x1,
  FUN = sqrt), vapply(x1, FUN = sqrt, FUN.VALUE = 0))
summary(mb)[, c("expr", "median")]

##                                     expr   median
## 1                         sqrt(x1) 4.6025
## 2                  lapply(x1, FUN = sqrt) 305.9085
## 3                  sapply(x1, FUN = sqrt) 366.0535
## 4 vapply(x1, FUN = sqrt, FUN.VALUE = 0) 297.8595

```

`lapply()` est beaucoup plus lent qu'une fonction vectorielle. `sapply()` nécessite plus de temps pour `simplify2array()`, qui doit détecter comment rassembler les résultats. Enfin, `vapply()` économise le temps de détermination du type de données du résultat et permet d'accélérer le calcul avec peu d'efforts.

2.4.3 Boucles for

Les boucles sont gérées par la fonction `for`. Elles ont la réputation d'être lentes dans R parce que le code à l'intérieur de la boucle doit être interprété à chaque exécution. Ce n'est plus le cas depuis la version 3.5 de R : les boucles sont compilées systématiquement avant leur exécution. Le comportement du compilateur "juste à temps" est défini par la fonction `enableJIT`. Le niveau par défaut est 3 : les fonctions sont toutes compilées, et les boucles dans le code aussi.

Pour évaluer le gain de performance, le code suivant supprime toute compilation automatique, et compare la même boucle compilée ou non.

```
library("compiler")
# Pas de compilation automatique
enableJIT(level = 0)

## [1] 3

# Boucle pour calculer la racine carrée d'un vecteur
Boucle <- function(x) {
  # Initialisation du vecteur de résultat, indispensable
  Racine <- vector("numeric", length = length(x))
  # Boucle
  for (i in 1:length(x)) Racine[i] <- sqrt(x[i])
  return(Racine)
}
# Version compilée
Boucle2 <- cmpfun(Boucle)
# Comparaison
mb <- microbenchmark(Boucle(x1), Boucle2(x1))
(mbs <- summary(mb)[, c("expr", "median")])
```

```
##           expr   median
## 1  Boucle(x1) 851.7015
## 2 Boucle2(x1)  78.6525
```

```
# Compilation automatique par défaut depuis la version 3.5
enableJIT(level = 3)
```

```
## [1] 0
```

Le gain est considérable : de 1 à 11.

Les boucles `for` sont maintenant nettement plus rapides que `vapply`.

```
# Test
mb <- microbenchmark(vapply(x1, FUN = sqrt, 0), Boucle(x1))
summary(mb)[, c("expr", "median")]

##           expr   median
## 1 vapply(x1, FUN = sqrt, 0) 303.924
## 2      Boucle(x1)    76.568
```

Attention, le test de performance peut être trompeur :

```
# Préparation du vecteur de résultat
Racine <- vector("numeric", length = length(x1))
# Test
mb <- microbenchmark(vapply(x1, FUN = sqrt, 0),
                      for(i in 1:length(x1))
                        Racine[i] <- sqrt(x1[i]))
summary(mb) [, c("expr", "median")]

##                                     expr     median
## 1           vapply(x1, FUN = sqrt, 0) 308.978
## 2 for (i in 1:length(x1)) Racine[i] <- sqrt(x1[i]) 3610.893
```

Dans ce code, la boucle for n'est pas compilée donc elle est beaucoup plus lente que dans le cadre normal de son utilisation (dans une fonction ou au niveau supérieur du code).

Les boucles longues permettent un suivi de leur progression par une barre de texte, ce qui est un autre avantage. La fonction suivante exécute des pauses d'un dixième de seconde pendant le temps passé en paramètre (en secondes).

```
BoucleSuivie <- function(duree = 1) {
  # Barre de progression
  pgb <- txtProgressBar(min = 0, max = duree * 10)
  # Boucle
  for (i in 1:(duree * 10)) {
    # Pause d'un dixième de seconde
    Sys.sleep(1/10)
    # Suivi de la progression
    setTxtProgressBar(pgb, i)
  }
}
BoucleSuivie()
```

```
## =====
```

2.4.4 replicate

`replicate()` répète une instruction.

```
replicate(3, runif(1))

## [1] 0.9453453 0.5262818 0.7233425
```

Ce code est équivalent à `runif(3)`, avec des performances similaires à celles de `vapply` : de 50 à 100 fois plus lent qu'une fonction vectorielle.

```
mb <- microbenchmark(replicate(1000, runif(1)), runif(1000))
summary(mb) [, c("expr", "median")]
```

```
##                                     expr     median
## 1 replicate(1000, runif(1)) 4297.887
## 2             runif(1000)   33.781
```

2.4.5 Vectorize

`Vectorize()` rend vectorielle une fonction qui ne l'est pas, par des boucles.
Ecrire plutôt les boucles.

2.4.6 Statistiques marginales

`apply` applique une fonction aux lignes ou colonnes d'un objet en deux dimensions.

`colSums` et ses semblables (`rowSums`, `colMeans`, `rowMeans`) sont optimisées.

```
# Somme des colonnes numériques du jeu de données diamonds de ggplot()
# Boucle identique à l'action de apply(, 2, )
BoucleSomme <- function(Table) {
  Somme <- vector("numeric", length = ncol(Table))
  for (i in 1:ncol(Table)) Somme[i] <- sum(Table[, i])
  return(Somme)
}
mb <- microbenchmark(BoucleSomme(diamonds[-(2:4)]),
                      apply(diamonds[-(2:4)], 2, sum),
                      colSums(diamonds[-(2:4)]))
summary(mb) [, c("expr", "median")]

##                                     expr     median
## 1   BoucleSomme(diamonds[-(2:4)]) 3.824472
## 2   apply(diamonds[-(2:4)], 2, sum) 9.798455
## 3       colSums(diamonds[-(2:4)]) 2.578881
```

`apply` clarifie le code mais est plus lent que la boucle, qui est à peine plus lente que `colSums`.

2.5 Code C++

L'intégration de code C++ dans R est largement simplifiée par le package **Rcpp** mais reste difficile à déboguer et donc à réserver à du code très simple (pour éviter toute erreur) et répété un grand nombre de fois (pour mériter l'effort). La préparation des données et leur vérification doivent être exécutées sous R, de même que le traitement et la présentation des résultats.

L'utilisation habituelle est l'inclusion de code C++ dans un package, mais l'utilisation hors package est possible :

- Le code C++ peut être inclus dans un document C++ (fichier avec l'extension `.cpp`) : il est compilé par la commande `sourceCpp()` qui crée les fonctions R permettant d'appeler le code C++.
- Dans un document RMarkdown, des bouts de code Rcpp peuvent être créés pour y insérer le code C++ : ils sont compilés et interfacés pour R au moment du tricotage.

L'exemple suivant montre comment créer une fonction C++ pour calculer le double d'un vecteur numérique.

```
#include <Rcpp.h>
using namespace Rcpp;

// [[Rcpp::export]]
NumericVector timesTwo(NumericVector x) {
    return x * 2;
}
```

Une fonction R du même nom que la fonction C++ est maintenant disponible.

```
timesTwo(1:5)

## [1] 2 4 6 8 10
```

Les performances sont deux ordres de grandeur plus rapides que le code R (voir l'étude de cas, section 2.7).

2.6 Paralléliser R

Lorsque des calculs longs peuvent être découpés en tâches indépendantes, l'exécution simultanée (*parallèle*) de ces tâches permet de réduire le temps de calcul total à celui de la tâche la plus longue, auquel s'ajoute le coût de la mise en place de la parallélisation (création des tâches, récupération des résultats...).

Lire l'excellente introduction de Josh Errickson⁹ qui détaille les enjeux et les contraintes de la parallélisation.

Deux mécanismes sont disponibles pour l'exécution de code en parallèle :

- *fork* : le processus en cours d'exécution est dupliqué sur plusieurs coeurs du processeur de l'ordinateur de calcul. C'est la méthode la plus simple mais elle ne fonctionne pas sous Windows (limite du système d'exploitation).
- *socket* : un cluster est constitué, soit physiquement (un ensemble d'ordinateurs exécutant R est nécessaire) soit logiquement (une instance de R sur chaque cœur de l'ordinateur utilisé). Les membres du cluster communiquent par le réseau (le réseau interne de l'ordinateur utilisé pour un cluster logique).

Différents packages de R permettent de mettre en œuvre ces mécanismes.

2.6.1 mclapply (fork)

La fonction `mclapply` du package **parallel** a la même syntaxe que `lapply` mais parallélise l'exécution des boucles. Sous Windows, elle n'a aucun effet puisque le système ne permet pas les *fork* : elle appelle simplement `lapply`.

⁹<http://dept.stat.lsa.umich.edu/~jerrick/courses/stat701/notes/parallel.html>

2. UTILISER R

Cependant, un contournement existe pour émuler `mclapply` sous Windows en appelant `parLapply`, qui utilise un cluster.

```
##  
## mclapply.hack.R  
##  
## Nathan VanHoudnos  
## nathanvan AT northwestern FULL STOP edu  
## July 14, 2014  
##  
## A script to implement a hackish version of  
## parallel:mclapply() on Windows machines.  
## On Linux or Mac, the script has no effect  
## beyond loading the parallel library.  
  
require(parallel)  
  
## Loading required package: parallel  
  
## Define the hack  
# mc.cores argument added: Eric Marcon  
mclapply.hack <- function(..., mc.cores=detectCores()) {  
  ## Create a cluster  
  size.of.list <- length(list(...)[[1]])  
  cl <- makeCluster( min(size.of.list, mc.cores) )  
  
  ## Find out the names of the loaded packages  
  loaded.package.names <- c(  
    ## Base packages  
    sessionInfo()$basePkgs,  
    ## Additional packages  
    names( sessionInfo()$otherPkgs ))  
  
  tryCatch( {  
  
    ## Copy over all of the objects within scope to  
    ## all clusters.  
    this.env <- environment()  
    while( identical( this.env, globalenv() ) == FALSE ) {  
      clusterExport(cl,  
                    ls(all.names=TRUE, env=this.env),  
                    envir=this.env)  
      this.env <- parent.env(environment())  
    }  
    clusterExport(cl,  
                  ls(all.names=TRUE, env=globalenv()),  
                  envir=globalenv())  
  
    ## Load the libraries on all the clusters  
    ## N.B. length(cl) returns the number of clusters  
    parLapply( cl, 1:length(cl), function(xx){  
      lapply(loaded.package.names, function(yy) {  
        require(yy , character.only=TRUE)})  
    })  
  
    ## Run the lapply in parallel  
    return( parLapply( cl, ... ) )  
  }, finally = {  
    ## Stop the cluster  
    stopCluster(cl)  
  })  
}  
  
## Warn the user if they are using Windows
```

```

if( Sys.info() [['sysname']] == 'Windows' ){
  message(paste(
    "\n",
    "  *** Microsoft Windows detected ***\n",
    "  \n",
    "  For technical reasons, the MS Windows version of mclapply()\n",
    "  is implemented as a serial function instead of a parallel\n",
    "  function.",
    "  \n\n",
    "  As a quick hack, we replace this serial version of mclapply()\n",
    "  with a wrapper to parLapply() for this R session. Please see\n",
    "  http://www.stat.cmu.edu/~nmv/2014/07/14/\n",
    "  implementing-mclapply-on-windows \n\n",
    "  for details.\n\n"))
}

## If the OS is Windows, set mclapply to the
## the hackish version. Otherwise, leave the
## definition alone.
mclapply <- switch( Sys.info() [['sysname']],
  Windows = {mclapply.hack},
  Linux   = {mclapply},
  Darwin  = {mclapply})

## end mclapply.hack.R

```

Le code suivant teste la parallélisation d'une fonction qui renvoie son argument inchangé après une pause d'un quart de seconde. Ce document est tricoté avec 3 cœurs, qui sont tous utilisés sauf un pour ne pas saturer le système.

```

f <- function(x, time = 0.25) {
  Sys.sleep(time)
  return(x)
}
# Laisser un cœur libre pour le système
nbCoeurs <- detectCores() - 1
# Série : temps théorique = nbCoeurs/4 secondes
(tserie <- system.time(lapply(1:nbCoeurs, f))

##      user  system elapsed
##  0.003   0.000   0.762

# Parallèle : temps théorique = 1/4 seconde
(tparallele <- system.time(mclapply(1:nbCoeurs, f, mc.cores = nbCoeurs)))

##      user  system elapsed
##  0.006   0.013   0.395

```

La mise en place de la parallélisation a un coût d'environ 0.14 secondes ici. Le temps d'exécution est bien plus long en parallèle sous Windows parce que la mise en place du cluster prend bien plus de temps que la parallélisation n'en fait gagner. La parallélisation est intéressante pour des tâches plus longues, comme une pause d'un seconde.

```

# Série
system.time(lapply(1:nbCoeurs, f, time = 1))

```

2. UTILISER R

```
##    user  system elapsed
##    0.00    0.00   2.24

# Parallèle
system.time(mclapply(1:nbCoeurs, f, time = 1, mc.cores = nbCoeurs))

##    user  system elapsed
##  0.005  0.011  1.125
```

Le temps additionnel nécessaire pour l'exécution parallèle du nouveau code est relativement plus faible : les coûts deviennent inférieurs à l'économie quand le temps de chaque tâche s'allonge.

Si le nombre de tâches parallèles dépasse le nombre de cœurs utilisés, les performances s'effondrent parce que la tâche supplémentaire doit être exécutée après les premières.

```
system.time(mclapply(1:nbCoeurs, f, time = 1, mc.cores = nbCoeurs))
```

```
##    user  system elapsed
##  0.001  0.003  1.141
```

```
system.time(mclapply(1:(nbCoeurs + 1), f, time = 1, mc.cores = nbCoeurs))
```

```
##    user  system elapsed
##  0.004  0.012  2.200
```

Le temps reste ensuite stable jusqu'au double du nombre de cœurs. La figure 2.2 montre l'évolution du temps de calcul en fonction du nombre de tâches.

```
Taches <- 1:(2 * nbCoeurs+1)
Temps <- sapply(Taches, function(nbTaches) {
  system.time(mclapply(1:nbTaches, f, time=1, mc.cores=nbCoeurs))
})
library("tidyverse")
tibble(Taches, Temps=Temps[["elapsed", ]]) %>%
  ggplot +
  geom_line(aes(x = Taches, y = Temps)) +
  geom_vline(xintercept = nbCoeurs, col = "red", lty = 2) +
  geom_vline(xintercept = 2 * nbCoeurs, col = "red", lty = 2)
```

La forme théorique de cette courbe est la suivante :

- pour une tâche, le temps est égal à une seconde plus le temps de mise en place de la parallélisation ;
- le temps devrait rester stable jusqu'au nombre de cœurs utilisés ;
- quand les cœurs sont tous utilisés (pointillés rouges), le temps devrait augmenter d'une seconde puis rester stable jusqu'à la limite suivante.

En pratique, le temps de calcul est déterminé par d'autres facteurs difficilement prévisibles. La bonne pratique est d'adapter le nombre de tâches au nombre de cœurs sous peine de perte de performance.

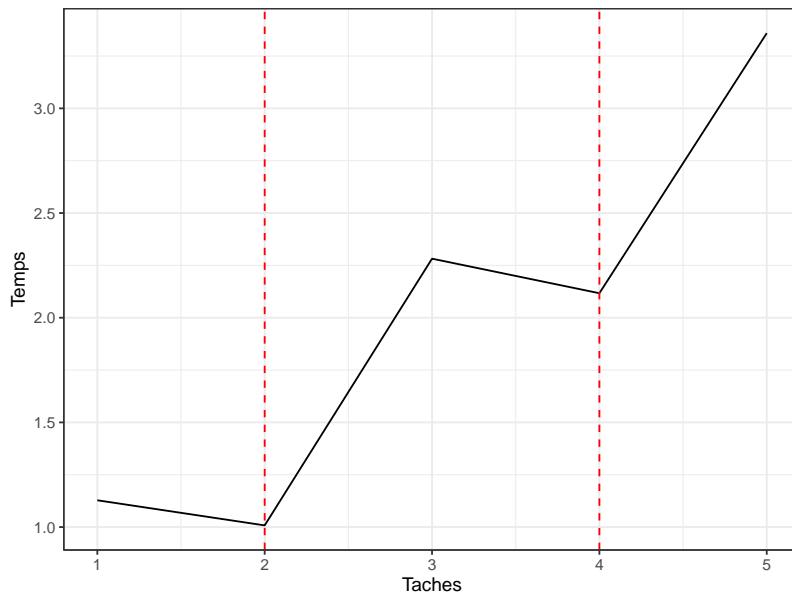


FIG. 2.2 : Temps d'exécution en parallèle de tâches nécessitant une seconde (chaque tâche est une pause d'une seconde). Le nombre de tâches varie de 1 à deux fois le nombre de coeurs utilisés (égal à 2) plus une.

2.6.2 `parLapply` (socket)

`parLapply` nécessite de créer un cluster, exporter les variables utiles sur chaque noeud, charger les packages nécessaires sur chaque noeud, exécuter le code et enfin arrêter le cluster. Le code de chaque étape se trouve dans la fonction `mclapply.hack` ci-dessus.

Pour un usage courant, `mclapply` est plus rapide, sauf sous Windows, et plus simple (y compris sous Windows grâce au contournement ci-dessus.)

2.6.3 `foreach`

Fonctionnement

Le package **foreach** permet un usage avancé de la parallélisation. Lire ses vignettes.

```
# Manuel
vignette("foreach", "foreach")
# Boucles imbriquées
vignette("nested", "foreach")
```

Indépendamment de la parallélisation, **foreach** redéfinit les boucles *for*.

```
for (i in 1:3) {
  f(i)
}
# devient
library("foreach")
```

2. UTILISER R

```
##  
## Attaching package: 'foreach'  
  
## The following objects are masked from 'package:purrr':  
##  
##     accumulate, when  
  
foreach(i = 1:3) %do% {  
  f(i)  
}  
  
## [[1]]  
## [1] 1  
##  
## [[2]]  
## [1] 2  
##  
## [[3]]  
## [1] 3
```

La fonction `foreach` retourne une liste contenant les résultats de chaque boucle. Les éléments de la liste peuvent être combinés par une fonction quelconque, comme `c`.

```
foreach(i = 1:3, .combine = "c") %do% {  
  f(i)  
}  
  
## [1] 1 2 3
```

La fonction `foreach` est capable d'utiliser des itérateurs, c'est-à-dire des fonctions qui ne passent à la boucle que les données dont elle a besoin sans charger les autres en mémoire. Ici, l'itérateur `i` passe les valeurs 1, 2 et 3 individuellement, sans charger le vecteur `1:3` en mémoire.

```
library(" iterators")  
foreach(i = icount(3), .combine = "c") %do% {  
  f(i)  
}  
  
## [1] 1 2 3
```

Elle est donc très utile quand chaque objet de la boucle utilise une grande quantité de mémoire.

Parallélisation

Remplacer l'opérateur `%do%` par `%dopar%` parallélise les boucles, à condition qu'un adaptateur, c'est-à-dire un package intermédiaire entre `foreach` et un package chargé de l'implémentation de la parallélisation, soit chargé. `doParallel` est un adaptateur pour utiliser le package `parallel` livré avec R.

```

library(doParallel)
registerDoParallel(cores = nbCoeurs)
# Série
system.time(foreach(i = icount(nbCoeurs), .combine = "c") %do%
{
  f(i)
})

##    user  system elapsed
##  0.003   0.000   0.643

# Parallèle
system.time(foreach(i = icount(nbCoeurs), .combine = "c") %dopar%
{
  f(i)
})

##    user  system elapsed
##  0.007   0.016   0.310

```

Le coût fixe de la parallélisation est faible.

2.7 Etude de cas

Cette étude de cas permet de tester les différentes techniques vues plus haut pour résoudre un problème concret. L'objectif est de calculer la distance moyenne entre deux points d'un semis aléatoire de 1000 points dans une fenêtre carrée de côté 1.

Son espérance est calculable¹⁰. Elle est égale à $\frac{2+\sqrt{2}+5 \ln(1+\sqrt{2})}{15} \approx 0,5214$.

2.7.1 Crédation des données

Le semis de points est créé avec le package **spatstat**.

```

NbPoints <- 1000
library("spatstat")
X <- runifpoint(NbPoints)

```

2.7.2 Spatstat

La fonction `pairdist()` de **spatstat** retourne la matrice des distances entre les points. La distance moyenne est calculée en divisant la somme par le nombre de paires de points distincts.

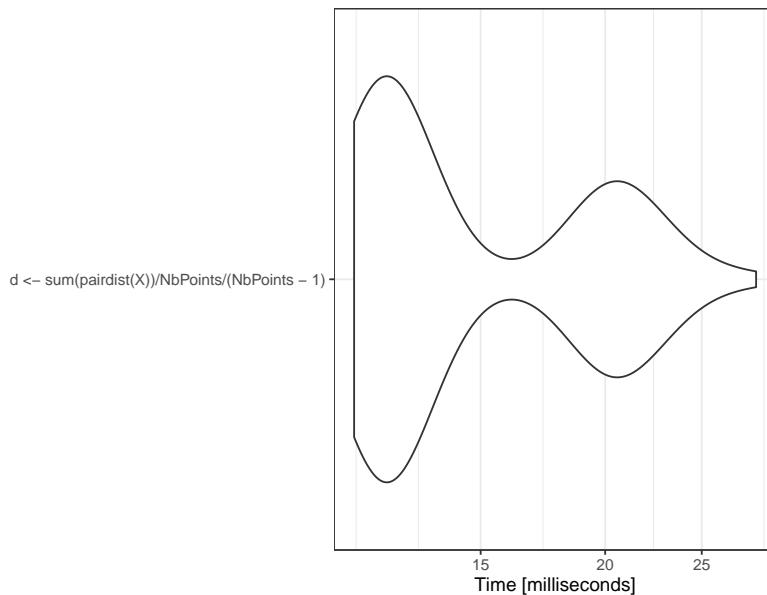
```

mb <- microbenchmark(d <- sum(pairdist(X))/NbPoints/(NbPoints -
  1))
# suppress messages pour éliminer les messages superflus
suppressMessages(autoplot(mb))

```

¹⁰<https://mindyourdecisions.com/blog/2016/07/03/distance-between-two-random-points-in-a-square-sunday-puzzle/>

2. UTILISER R



```
d
```

```
## [1] 0.5154062
```

La fonction est rapide parce qu'elle est codée en langage C dans le package **spatstat** pour le cœur de ses calculs.

2.7.3 apply

La distance peut être calculée par deux sapply() imbriqués.

```
fsapply1 <- function() {  
  distances <- sapply(1:NbPoints, function(i) sapply(1:NbPoints,  
    function(j) sqrt((X$x[i] - X$x[j])^2 + (X$y[i] - X$y[j])^2)))  
  return(sum(distances)/NbPoints/(NbPoints - 1))  
}  
system.time(d <- fsapply1())
```

```
##      user  system elapsed  
##  5.447   0.032  5.514
```

```
d
```

```
## [1] 0.5154062
```

Un peu de temps peut être gagné en remplaçant sapply par vapply : le format des résultats n'a pas à être déterminé par la fonction. Le gain est négligeable sur un long calcul comme celui-ci mais important pour des calculs courts.

```

fsapply2 <- function() {
  distances <- vapply(1:NbPoints, function(i) vapply(1:NbPoints,
    function(j) sqrt((X$x[i] - X$x[j])^2 + (X$y[i] - X$y[j])^2),
    0), 1:1000 + 0)
  return(sum(distances)/NbPoints/(NbPoints - 1))
}
system.time(d <- fsapply2())

##      user  system elapsed
##  5.315   0.033  5.448

d

## [1] 0.5154062

```

Le format de sortie n'est pas toujours évident à écrire :

- il doit respecter la taille des données : un vecteur de taille 1000 pour la boucle externe, un scalaire pour la boucle interne.
- il doit respecter leur type : 0 pour un entier, 0.0 pour un réel. Dans la boucle externe, l'ajout de 0.0 au vecteur d'entiers le transforme en vecteur de réels.

Une amélioration plus significative consiste à ne calculer les racines carrées qu'à la fin de la boucle, pour profiter de la vectorisation de la fonction.

```

fsapply3 <- function() {
  distances <- vapply(1:NbPoints, function(i) vapply(1:NbPoints,
    function(j) (X$x[i] - X$x[j])^2 + (X$y[i] - X$y[j])^2,
    0), 1:1000 + 0)
  return(sum(sqrt(distances))/NbPoints/(NbPoints - 1))
}
system.time(d <- fsapply3())

##      user  system elapsed
##  5.209   0.024  5.258

d

## [1] 0.5154062

```

Les calculs sont effectués deux fois (distance entre les points i et j , mais aussi entre les points j et i) : un test sur les indices permet de diviser presque le temps par 2 (pas tout à fait parce que les boucles sans calcul, qui retournent 0, prennent du temps).

```

fsapply4 <- function() {
  distances <- vapply(1:NbPoints, function(i) {
    vapply(1:NbPoints, function(j) {
      if (j > i) {
        (X$x[i] - X$x[j])^2 + (X$y[i] - X$y[j])^2
      } else {
        0
      }
    })
  })
  return(sum(distances)/NbPoints/(NbPoints - 1))
}
system.time(d <- fsapply4())

```

2. UTILISER R

```
        }
    }, 0)
}, 1:1000 + 0)
return(sum(sqrt(distances))/NbPoints/(NbPoints - 1) * 2)
}
system.time(d <- fsapply4())

##   user   system elapsed
##  3.058   0.029   3.208

d

## [1] 0.5154062
```

En parallèle, le temps de calcul n'est pas amélioré sous Windows parce que les tâches individuelles sont trop courtes. Sous MacOS ou Linux, le calcul est accéléré.

```
fsapply5 <- function() {
  distances <- mclapply(1:NbPoints, function(i) {
    vapply(1:NbPoints, function(j) {
      if (j > i) {
        (X$x[i] - X$x[j])^2 + (X$y[i] - X$y[j])^2
      } else {
        0
      }
    }, 0)
  })
  return(sum(sqrt(simplify2array(distances)))/NbPoints/(NbPoints -
  1) * 2)
}
system.time(d <- fsapply5())

##   user   system elapsed
##  3.391   0.401   2.000

d

## [1] 0.5154062
```

2.7.4 boucle for

Une boucle for est plus rapide et consomme moins de mémoire parce qu'elle ne stocke pas la matrice de distances.

```
distance <- 0
ffor <- function() {
  for (i in 1:(NbPoints - 1)) {
    for (j in (i + 1):NbPoints) {
      distance <- distance + sqrt((X$x[i] - X$x[j])^2 +
      (X$y[i] - X$y[j])^2)
    }
  }
  return(distance/NbPoints/(NbPoints - 1) * 2)
}
# Temps de calcul, mémorisé
(for_time <- system.time(d <- ffor()))
```

```
##    user  system elapsed
##  1.974   0.013   2.010
```

```
d
```

```
## [1] 0.5154062
```

C'est la façon la plus simple et efficace d'écrire ce code sans parallélisation et en se limitant au langage de R.

2.7.5 boucle foreach

Deux boucles foreach imbriquées sont nécessaires ici : elles sont extrêmement lentes en comparaison d'une boucle simple. Le test est lancé ici avec 10 fois moins de points, donc 100 fois moins de distances à calculer.

```
NbPointsReduit <- 100
Y <- runifpoint(NbPointsReduit)
fforeach1 <- function(Y) {
  distances <- foreach(i = 1:NbPointsReduit, .combine = "cbind") %:%
    foreach(j = 1:NbPointsReduit, .combine = "c") %do% {
      if (j > i) {
        (Y$x[i] - Y$x[j])^2 + (Y$y[i] - Y$y[j])^2
      } else {
        0
      }
    }
  return(sum(sqrt(distances))/NbPointsReduit/(NbPointsReduit -
    1) * 2)
}
system.time(d <- fforeach1(Y))
```

```
##    user  system elapsed
##  2.761   0.024   2.800
```

```
d
```

```
## [1] 0.5181951
```

Les boucles foreach imbriquées sont à réserver à des tâches très longues (plusieurs secondes au moins) pour amortir les coûts fixes de leur mise en place.

La parallélisation est efficace dans le code ci-dessous, notamment parce qu'elle permet d'éviter les boucles foreach imbriquées. En revanche, les distances sont calculées deux fois. La performance reste très inférieure à celle d'une simple boucle for (rappel : 100 fois moins de distances sont calculées).

```
registerDoParallel(cores = detectCores())
fforeach3 <- function(Y) {
  distances <-
    foreach(i=icount(NbPointsReduit),
           .combine='+') %dopar% {
      distance <- 0
      for (j in 1:Y$n) {
        distance <- distance +
```

2. UTILISER R

```
        sqrt((Y$x[i]-Y$x[j])^2 + (Y$y[i]-Y$y[j])^2)
    }
    distance
}
return(distances/NbPointsReduit/(NbPointsReduit-1))
}
system.time(d <- fforeach3(Y))
```

```
##      user  system elapsed
##  0.129   0.063   0.121
```

```
d
```

```
## [1] 0.5181951
```

foreach dispose d'adaptateurs optimisés permettant d'utiliser des clusters physiques par exemple. Son intérêt est limité avec le package **parallel**.

2.7.6 RCpp

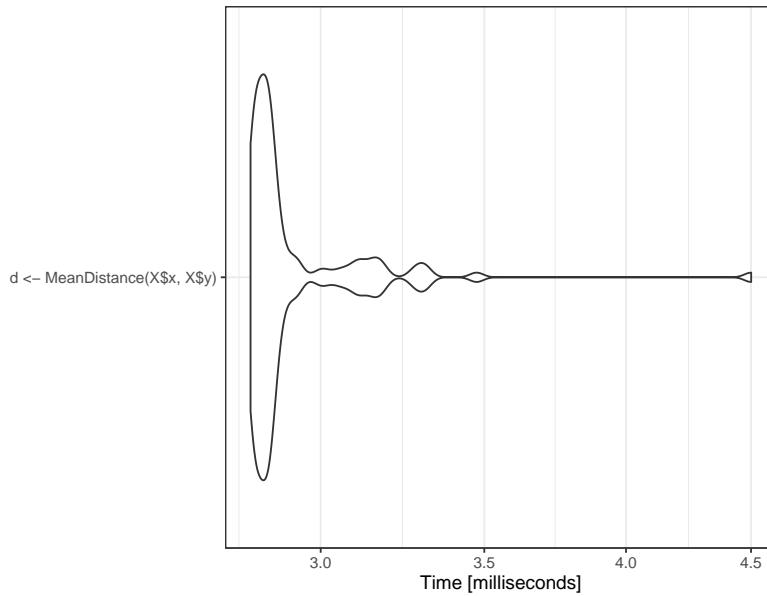
La fonction C++ permettant de calculer les distances est la suivante.

```
#include <Rcpp.h>
using namespace Rcpp;

// [[Rcpp::export]]
double MeanDistance(NumericVector x, NumericVector y) {
  double distance=0;
  double dx, dy;
  for (int i=0; i < (x.length()-1); i++) {
    for (int j=i+1; j < x.length(); j++) {
      // Calculate distance
      dx = x[i]-x[j];
      dy = y[i]-y[j];
      distance += sqrt(dx*dx + dy*dy);
    }
  }
  return distance/(double)(x.length()/2*(x.length()-1));
}
```

Elle est appelée dans R très simplement. Le temps d'exécution est très court.

```
mb <- microbenchmark(d <- MeanDistance(X$x, X$y))
# suppress messages pour éliminer les messages superflus
suppressMessages(autoplot(mb))
```



```
d
## [1] 0.5154062
```

2.7.7 RcppParallel

RcppParallel permet d’interfacer du code C++ parallélisé, au prix d’une syntaxe plus complexe qu’avec **RCpp**. Une documentation est disponible¹¹.

La fonction C++ exportée dans R ne réalise pas les calculs mais organise seulement l’exécution en parallèle d’une autre fonction, non exportée, de type Worker.

Deux fonctions (C++) de parallélisation sont disponibles pour deux types de tâches :

- `parallelReduce` pour l’accumulation d’une valeur, utilisée ici pour additionner les distances.
- `parallelFor` pour remplir une matrice de résultats.

La syntaxe du Worker est un peu laborieuse mais assez simple à adapter : les constructeurs initialisent les variables C à partir des valeurs transmises par R et déclarent la parallélisation.

```
// [[Rcpp::depends(RcppParallel)]]
#include <Rcpp.h>
#include <RcppParallel.h>
using namespace Rcpp;
using namespace RcppParallel;

// Fonction de travail, non exportée
```

¹¹<http://rcppcore.github.io/RcppParallel/>

2. UTILISER R

```

struct TotalDistanceWrkr : public Worker
{
    // source vectors
    const RVector<double> Rx;
    const RVector<double> Ry;

    // accumulated value
    double distance;

    // constructors
    TotalDistanceWrkr(const NumericVector x, const NumericVector y) :
        Rx(x), Ry(y), distance(0) {}
    TotalDistanceWrkr(const TotalDistanceWrkr& totalDistanceWrkr, Split) :
        Rx(totalDistanceWrkr.Rx), Ry(totalDistanceWrkr.Ry), distance(0) {}

    // count neighbors
    void operator()(std::size_t begin, std::size_t end) {
        double dx, dy;
        unsigned int Npoints = Rx.length();

        for (unsigned int i = begin; i < end; i++) {
            for (unsigned int j=i+1; j < Npoints; j++) {
                // Calculate squared distance
                dx = Rx[i]-Rx[j];
                dy = Ry[i]-Ry[j];
                distance += sqrt(dx*dx + dy*dy);
            }
        }
    }

    // join my value with that of another Sum
    void join(const TotalDistanceWrkr& rhs) {
        distance += rhs.distance;
    }
};

// Fonction exportée
// [[Rcpp::export]]
double TotalDistance(NumericVector x, NumericVector y) {

    // Declare TotalDistanceWrkr instance
    TotalDistanceWrkr totalDistanceWrkr(x, y);

    // call parallel_reduce to start the work
    parallelReduce(0, x.length(), totalDistanceWrkr);

    // return the result
    return totalDistanceWrkr.distance;
}

```

L'usage dans R est identique à celui des fonctions C++ interfacées par **RCpp**.

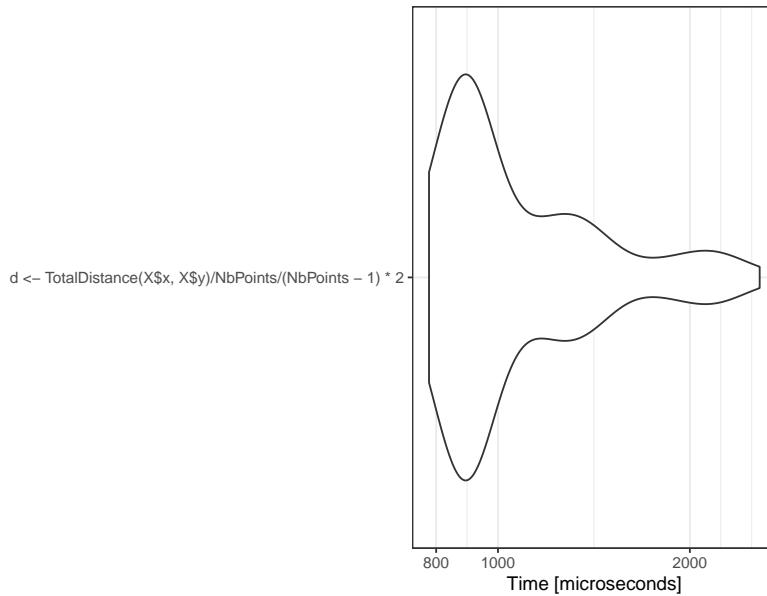
```
(mb <- microbenchmark(d <- TotalDistance(X$x, X$y)/NbPoints/(NbPoints -
 1) * 2))
```

```

## Unit: microseconds
##                                              expr
## d <- TotalDistance(X$x, X$y)/NbPoints/(NbPoints - 1) * 2
##   min     lq      mean    median      uq      max neval
## 779.633 885.3555 1127.122 915.284 1281.08 2573.438   100

```

```
# suppress messages pour éliminer les messages superflus
suppressMessages(autoplot(mb))
```



```
d
```

```
## [1] 0.5154062
```

Le temps de mise en place des tâches parallèles est bien plus long que le temps de calcul en série.

En multipliant le nombre de points par 50, le temps de calcul en série doit être multiplié par 2500 environ.

```
NbPoints <- 5000
X <- runifpoint(NbPoints)
system.time(d <- MeanDistance(X$x, X$y))
```

```
##      user    system elapsed
##  7.104   0.036   7.179
```

En parallèle, le temps augmente peu : la parallélisation devient réellement efficace. Ce temps est à comparer à celui de la boucle for de référence, multiplié par 2500, soit 5024 secondes.

```
system.time(d <- TotalDistance(X$x, X$y)/NbPoints/(NbPoints -
1) * 2)
```

```
##      user    system elapsed
##  4.957   0.052   2.224
```

2.7.8 Conclusions sur l'optimisation de la vitesse du code

De cette étude de cas, plusieurs enseignements peuvent être retirés :

- une boucle `for` est une bonne base pour des calculs répétitifs, plus rapide que `vapply()`, simple à lire et à écrire ;
- des fonctions optimisées peuvent exister dans les packages de R pour des tâches courantes (ici, la fonction `pairdist()` de **spatstat** est deux ordres de grandeur plus rapide que la boucle `for`) ;
- le recours au code C++ permet d'accélérer significativement les calculs, de trois ordres de grandeur ici ;
- la parallélisation du code C++ divise encore le temps de calcul par environ la moitié du nombre de coeurs pour de longs calculs.

Au-delà de cet exemple, l'optimisation du temps de calcul sous R peut être compliquée si elle passe par la parallélisation et l'écriture de code C++. L'effort doit donc être concentré sur les calculs réellement long alors que la lisibilité du code doit rester la priorité pour le code courant. Le code C est assez facile à intégrer grâce à **RCpp** et sa parallélisation n'est pas très coûteuse avec **Rcpp-Parallel**.

L'utilisation de boucles `for` n'est plus pénalisante depuis la version 3.5 de R. L'écriture de code vectoriel, utilisant `sapply()` se justifie toujours pour sa lisibilité.

Le choix de paralléliser le code doit être évalué selon le temps d'exécution de chaque tâche parallélisable. S'il dépasse quelques secondes, la parallélisation se justifie. `mclapply()` remplace `lapply()` sans aucun effort, mais nécessite un hack (fourni ici) sous Windows. `foreach()` ne remplace pas `for()` aussi simplement et ne se justifie que pour des tâches très lourdes en termes de mémoire et de temps de calcul, en particulier sur des clusters de calcul.

2.8 Flux de travail

Le package **targets** permet de gérer un flux de travail (*workflow*), c'est-à-dire de décomposer le code en tâches élémentaires appelées *cibles* qui s'enchaînent, dont le résultat est stocké dans une variable, elle-même enregistrée sur le disque. En cas de changement dans le code ou les données utilisées, seules les cibles concernées sont réévaluées.

Le fonctionnement du flux est proche de celui d'un cache, mais ne dépend pas de l'ordinateur sur lequel il s'exécute. **targets** permet de visualiser les tâches obsolètes, d'intégrer le flux à un projet de document (voir section 4.9), et même de faire appel à un cluster de calcul pour traiter les tâches en parallèle.

2.8.1 Principe de fonctionnement

La documentation¹² de **targets** est détaillée et fournit un exemple travaillé pour apprendre à utiliser le package¹³. Elle n'est pas reprise ici, mais les principes du fonctionnement du flux sont expliqués.

Le flux de travail est unique pour un projet donné. Il est codé dans le fichier `_targets.R` à la racine du projet. Il contient :

- des commandes globales, comme le chargement des packages ;
- une liste de cibles, qui décrivent le code à exécuter et la variable qui stocke leur résultat.

Le flux est exécuté par la fonction `tar_make()` qui met à jour les cibles qui le nécessitent. Son contenu est placé dans le dossier `_targets`. Les variables stockées sont lues par `tar_read()`.

Si le projet nécessite de longs calculs, **targets** permet de n'exécuter que ceux qui sont nécessaires. Si le projet est partagé ou placé sous contrôle de source (chapitre 3), le résultat des calculs est intégré l'est aussi. Enfin, si le projet est un document (chapitre 4), son formatage est complètement indépendant du calcul de son contenu, pour un gain de temps qui peut être considérable.

2.8.2 Exemple minimal

L'exemple suivant est encore plus simple que celui du manuel de **targets**, qui permettra d'aller plus loin. Il reprend l'étude de cas précédente : un jeu de points est généré puis la distance moyenne entre les points obtenus est calculée. Une carte des points est tracée en plus. Chacune de ces trois opérations est une cible dans le vocabulaire de **targets**.

Le fichier du flux de travail est donc le suivant :

```
# Fichier _targets.R
library("targets")
tar_option_set(packages = c("spatstat", "dbmss"))
list(
  # Tirage des points
  tar_target(X,
    runifpoint(NbPoints)
  ),
  # Paramétrage
  tar_target(NbPoints,
    1000
  ),
  # Distance moyenne
  tar_target(d,
    sum(pairdist(X)) / NbPoints / (NbPoints - 1)
  ),
  # Carte
  tar_target(map,
    autoplot(as.wmppp(X))
  )
)
```

¹²<https://books.ropensci.org/targets/>

¹³<https://books.ropensci.org/targets/walkthrough.html>

2. UTILISER R

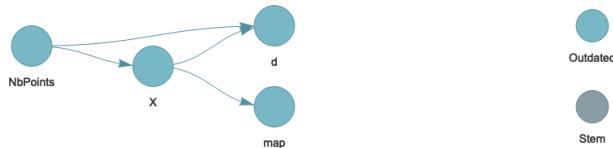
```
)
```

Les commandes globales consistent à charger le package **targets** lui-même puis lister les packages nécessaires au code. L'exécution du flux a lieu dans une nouvelle instance de R.

Les cibles sont ensuite listées. Chacune est déclarée par la fonction `tar_target()` dont le premier argument est le nom de la cible, qui sera celui de la variable qui recevra le résultat. Le deuxième argument est le code qui produit le résultat. Les cibles sont très simples ici et peuvent être écrites en une seule commande. Quand ce n'est pas le cas, chaque cible peut être écrite sous la forme d'une fonction, stockée dans un fichier de code séparé chargé par la fonction `source()` au début du fichier de flux.

La commande `tar_visnetwork` permet d'afficher l'enchaînement des cibles et leur état éventuellement obsolète.

```
library("targets")
tar_visnetwork()
```



L'ordre de déclaration des cibles dans la liste sans importance : elles sont ordonnées automatiquement.

Le flux est exécuté par `tar_make()`.

```
tar_make()
```

```
## • start target NbPoints
## • built target NbPoints
## • start target X
## • built target X
## • start target d
## • built target d
## • start target map
## • built target map
## • end pipeline: 3.877 seconds
```

Le flux est maintenant à jour et `tar_make()` ne refait aucun calcul.

```
tar_visnetwork()
```



```
tar_make()
```

```
##  skip target NbPoints
##  skip target X
##  skip target d
##  skip target map
##  skip pipeline: 0.314 seconds
```

Les résultats sont lus par `tar_read()`.

```
tar_read(d)
```

```
## [1] 0.5189867
```

```
tar_read(map)
```



2.8.3 Intérêt pratique

Dans cet exemple, **targets** complique l'écriture du code et `tar_make()` est beaucoup plus lent que la simple exécution du code qu'il traite parce qu'il doit vérifier si les cibles sont à jour. Dans un projet réel qui nécessite de longs calculs, le traitement du statut des cibles est négligeable et le gain de temps apporté par la seule évaluation des cibles nécessaires est considérable. La définition des cibles reste une contrainte, mais force à bien structurer son projet.

GIT ET GITHUB

Le contrôle de source consiste à enregistrer l'ensemble des modifications apportées sur les fichiers suivis. Les avantages sont nombreux : traçabilité et sécurité du projet, possibilité de collaborer efficacement, de revenir en arrière, de tenter de nouveaux développements sans mettre en péril la version stable...

3.1 Principes

3.1.1 Contrôle de source

L'outil standard est aujourd'hui *git*.

Les commandes de git peuvent être exécutées dans le terminal de RStudio.



A screenshot of the RStudio interface showing the Terminal tab active. The terminal window displays a command-line session:

```
Console Terminal x R Markdown x Jobs x
Terminal 1  /c/Users/Eric.Marcon/AppData/Local/Gitted/Enseignement/travailR
$ 
Eric.Marcon@Wacapou20 ~/AppData/Local/Gitted/Enseignement/travailR
$ git status
fatal: not a git repository (or any of the parent directories): .git
Eric.Marcon@Wacapou20 ~/AppData/Local/Gitted/Enseignement/travailR
$ 
```

FIG. 3.1 : Capture d'écran du terminal de RStudio. La commande `git status` supposée décrire l'état du dépôt renvoie une erreur si le projet R n'est pas sous contrôle de source.

La commande `git status` (figure 3.1) retourne l'état du dépôt (*repository*), c'est-à-dire l'ensemble des données gérées par git pour suivre le projet en cours.

RStudio intègre une interface graphique pour git suffisante pour se passer de la ligne de commande dans le cadre d'une utilisation standard, présentée ici.

3.1.2 git et GitHub

git est le logiciel installé sur le poste de travail.

GitHub est une plateforme, accessible par le web¹, qui permet de partager le contenu des dépôts git (pour travailler à plusieurs) et de partager de la documentation sous la forme d'un site web (*GitHub Pages*).

Comme GitHub permet au minimum la sauvegarde des dépôts git, les deux sont toujours utilisés ensemble. GitHub n'est pas la seule plateforme utilisable mais la principale. Les alternatives sont Bitbucket² et GitLab³ par exemple.

3.2 Crée un nouveau dépôt

3.2.1 A partir d'un projet existant

Dans un projet R existant, activer le contrôle de source dans les options du projet (figure 3.2). La commande exécutée est `git init`. Redémarrer RStudio à la demande.

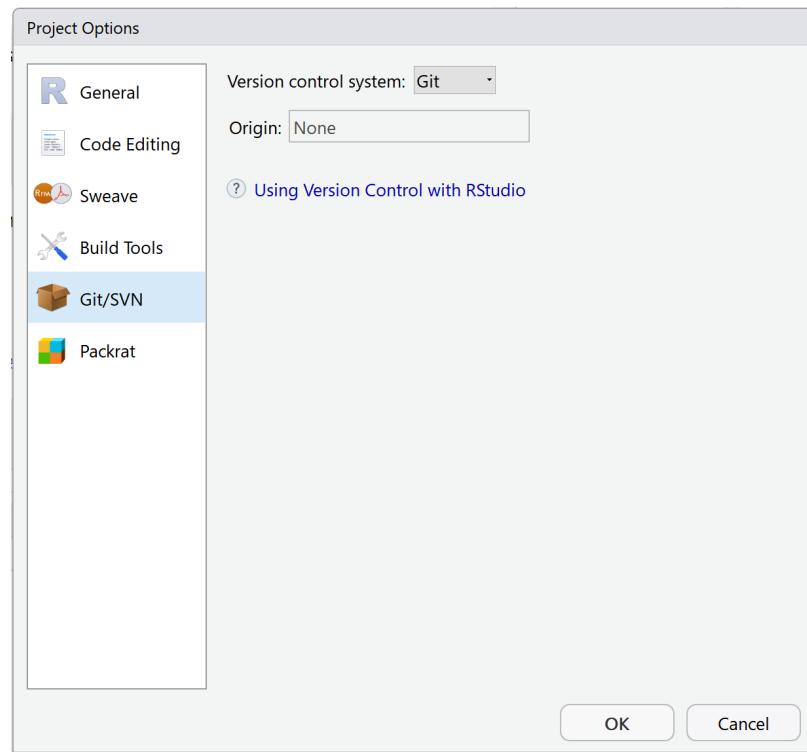


FIG. 3.2 : Activation du contrôle de source dans le menu “Tools > Project Options...”.

¹<https://github.com/>

²<https://bitbucket.org/>

³<https://about.gitlab.com/>

Une nouvelle fenêtre *Git* apparaît dans le panneau supérieur droit. Elle contient la liste des fichiers du projet (figure 3.3).

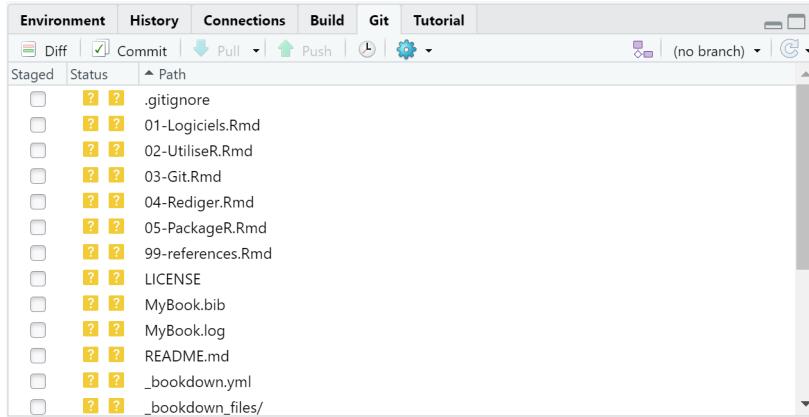


FIG. 3.3 : Fichiers du projet, pas encore pris en compte par git.

A ce stade, les fichiers ne sont pas pris en compte par git : leur statut est un double point d'interrogation jaune. Pour git, le répertoire de travail local est un *bac à sable* où toutes les modifications sont possibles sans conséquences.

Le fichier `.gitignore` contient la liste des fichiers qui n'ont jamais vocation à être pris en compte, qu'il est donc inutile d'afficher dans la liste : les fichiers intermédiaires produits automatiquement par exemple. La syntaxe des fichiers `.gitignore` est détaillée dans la documentation de git⁴. En règle générale, utiliser un fichier existant : les modèles de documents notamment incluent leur fichier `.gitignore`.

3.2.2 Prendre en compte des fichiers

Dans la fenêtre git, cocher la case *Staged* permet de prendre en compte (*Stage*) chaque fichier. La commande exécutée est `git add <NomDeFichier>`. Les fichiers pris en compte une première fois ont le statut “A” pour “Added”.

Les fichiers pris en compte font partie de l'*index* de git.

3.2.3 Valider des modifications

Les fichiers pris en compte peuvent être validés (*Commit*) en cliquant sur le bouton “Commit” dans la fenêtre *Git*. Une nouvelle fenêtre s’ouvre (figure 3.4), qui permet de visualiser toutes les modifications par fichier (ajouts en verts, suppressions en rouge). Le grain de modification traité par git est la ligne de texte, terminée par un retour à la ligne. Les fichiers binaires comme les images sont traités en bloc.

Chaque validation (*Commit*) est accompagnée d'un texte de description. La première ligne est la description courte. Une description détaillée peut être ajoutée.

⁴<https://git-scm.com/docs/gitignore>

3. GIT ET GITHUB

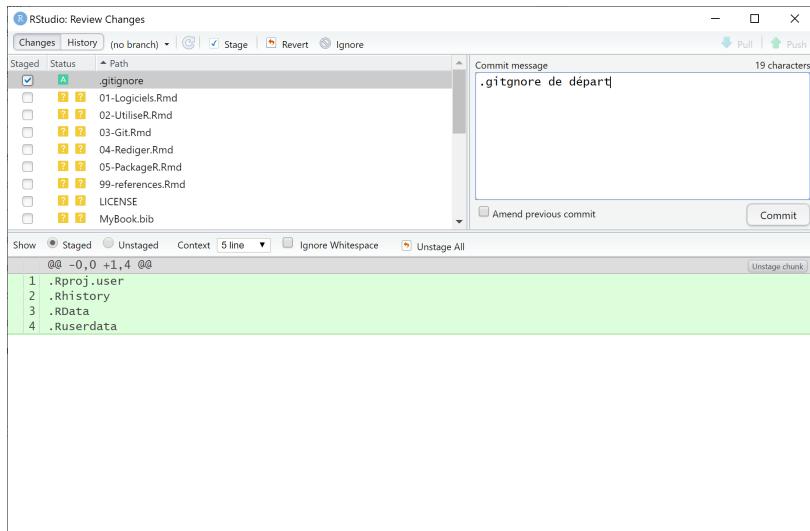


FIG. 3.4 : Fenêtre de validation des modifications prises en compte.

tée après un saut de ligne. Pour la lisibilité de l'historique du projet, chaque *commit* correspond donc à une action, correspondant à la description courte : tous les fichiers modifiés ne sont pas forcément pris en compte et validés en une fois. La commande exécutée est `git commit -m "Message de validation"`.

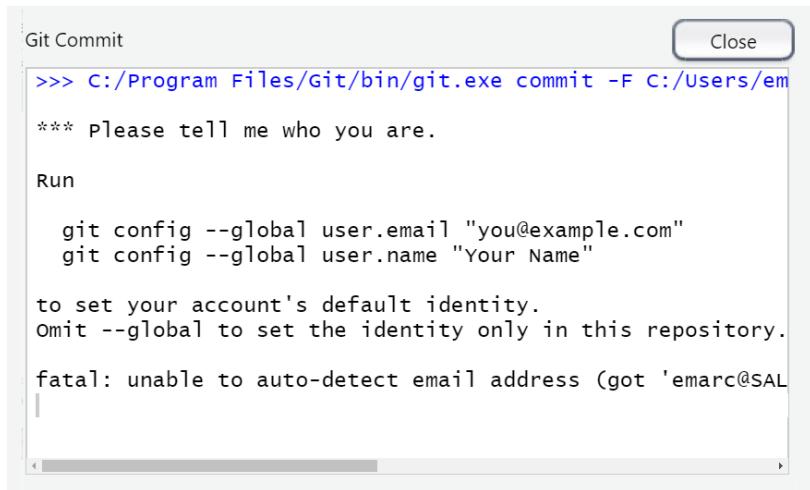


FIG. 3.5 : Fenêtre de demande d'identification.

Les validations sont liées à leur auteur, qui doit être identifié par git. En règle générale, git utilise les informations du système. S'il n'y parvient pas, une fenêtre demande à l'utilisateur de s'identifier avant d'effectuer son premier *commit* (figure 3.5). Les commandes indiquées sont à exécuter dans le terminal de RStudio. Elles peuvent aussi être utilisées pour vérifier les valeurs connues par git :

```
git config user.name  
git config user.email
```

Dès la première validation, la branche principale du dépôt, appelée “master”, est créée. Une branche est une version du dépôt, avec son propre historique et donc ses propres fichiers. Les branches permettent :

- de développer de nouvelles fonctionnalités dans un projet, sans perturber la branche principale qui peut contenir une version stable. Si le développement est retenu, sa branche pourra être fusionnée avec la branche *master* pour constituer une nouvelle version stable.
- de contenir des fichiers totalement différents de ceux de la branche principale, pour d’autres objectifs. Sur GitHub, les pages web de présentation du projet peuvent être placés dans une branche appelée “gh-pages” qui ne sera jamais fusionnée.

Le dépôt git est complètement constitué. Dans le vocabulaire de git, il comprend trois *arbres* (figure 3.6) :

- le répertoire de travail, ou bac à sable, qui contient les fichiers non pris en compte : inconnus, modifiés, supprimés ou renommés (case *Staged* décochée) ;
- l’index, qui contient les fichiers pris en compte (case *Staged* cochée) ;
- la tête, qui contient les fichiers validés.

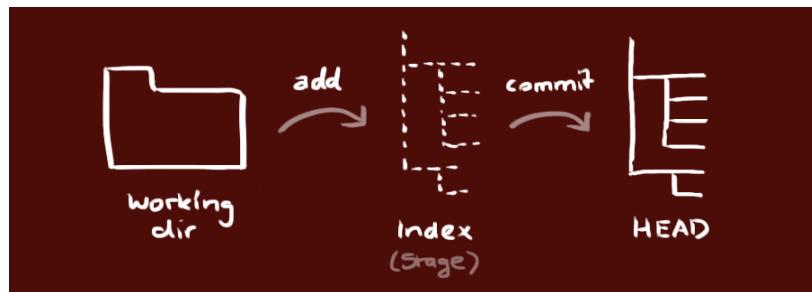


FIG. 3.6 : Les trois arbres de git. Source : <https://rogerdudler.github.io/git-guide/index.fr.html>

Le statut des fichiers est représenté par deux icônes dans la fenêtre *Git* de RStudio : deux points d’interrogation quand ils n’ont pas été pris en compte par git. Ensuite, l’icône de droite décrit la différence entre le répertoire de travail et l’index. Celle de gauche décrit la différence entre l’index et la tête. Un fichier modifié aura donc l’icône M affichée à droite avant d’être pris en compte, puis à gauche après prise en compte. Il est possible, même s’il vaut mieux l’éviter, de modifier à nouveau un fichier pris en compte avant qu’il soit validé : alors, les deux icônes seront affichées.

3.2.4 Crée un dépôt vide sur GitHub

Un dépôt vide sur GitHub doit être créé (figure 3.7) :

3. GIT ET GITHUB

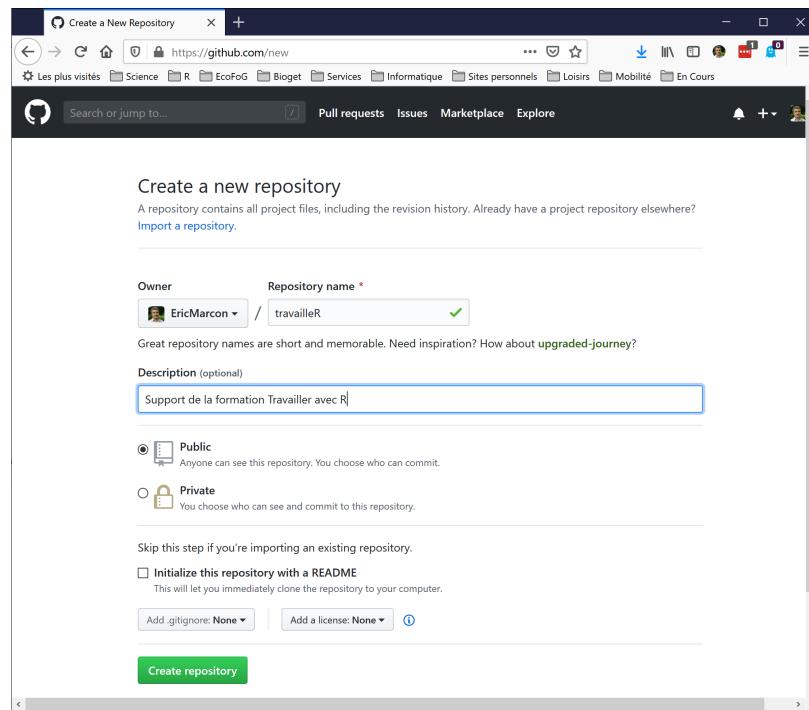


FIG. 3.7 : Crédation d'un dépôt sur GitHub.

- Sur GitHub, cliquer sur le bouton vert “New repository”.
- Saisir le nom du dépôt, identique à celui du projet R local.
- Ajouter une description, qui apparaîtra uniquement sur la page GitHub du dépôt.
- Choisir le statut du dépôt :
 - Public : visible par tout le monde
 - Privé : visible seulement par les collaborateurs du projet, ce qui exclut de compléter par des pages web de présentation.
- Ne pas ajouter de README, .gitignore ou licence : le projet doit être vide.
- Cliquer sur “create Repository”.
- Copier l'adresse du dépôt (<https://github.com/>... ou git@github.com :...)

Le choix de l'adresse est lié à la méthode d'authentification. L'authentification SSH (voir section 1.4.3) est à privilégier.

3.2.5 Lier git et GitHub

Dans RStudio, un premier *commit* doit au moins avoir eu lieu pour que la branche principale du projet, nommée “master”, existe. En haut à droite de la fenêtre *Git* (figure 3.3), il est affiché “(no branch)” avant cela. Ensuite, il est affiché “master”, le nom par défaut de la branche principale du projet. Le projet peut alors être lié au dépôt GitHub.

Méthode graphique

Cliquer sur le bouton violet à côté de “master” : une fenêtre apparaît (habituellement utilisée pour la création d'une nouvelle branche, voir section 3.4). Saisir le nom de la branche “master”, cliquer sur “Add Remotes” et compléter :

- Remote Name : origin ;
- Remote URL : coller l'adresse du dépôt GitHub ;
- Cliquer sur “Add”.

Cocher la case “Sync with Remote”.

Au message indiquant qu'une branche *master* existe déjà, cliquer sur “Overwrite”.

En ligne de commande

Plutôt que la manipulation précédente, le lien entre Git et GitHub peut être mis en place par quelques commandes de git exécutées dans le terminal de RStudio. Elles sont affichées sur la page d'accueil de tout dépôt vide nouvellement créé sur GitHub et peuvent donc être copiées et collées directement vers le terminal.

```
git remote add origin git@github.com:GitHubID/NomDuDepot.git  
git branch -M master  
git push -u origin master
```

La première commande déclare le dépôt GitHub comme dépôt distant. Le nom *origin* est une convention de git. Il peut être modifié mais l'organisation du projet sera plus lisible en respectant la convention. L'adresse du dépôt est <https://github.com/GitHubID/NomDuDepot.git> si l'authentification HTTPS est choisie.

Les commandes suivantes activent la branche principale du projet et poussent son contenu vers GitHub.

Attention au nom de la branche principale (voir section 3.4) : par défaut, elle s'appelle “master” dans un projet créé dans RStudio mais “main” sur GitHub. Les lignes de commande ci-dessus fournies par GitHub remplacent donc `master` par `main` et doivent être corrigées pour correspondre au nom de la branche créée par RStudio.

Authentification

Si l'authentification HTTPS est choisie, à la première connexion de RStudio à GitHub, une fenêtre permet de saisir ses identifiants GitHub (figure 3.8).

Depuis août 2021, GitHub n'accepte plus le mot de passe du compte de l'utilisateur pour cette authentification : le jeton personnel (PAT) créé en section 1.4.4 doit être saisi à sa place.

Si l'authentification SSH est choisie et a été configurée à l'installation de git (section 1.4.3), aucune action n'est nécessaire.

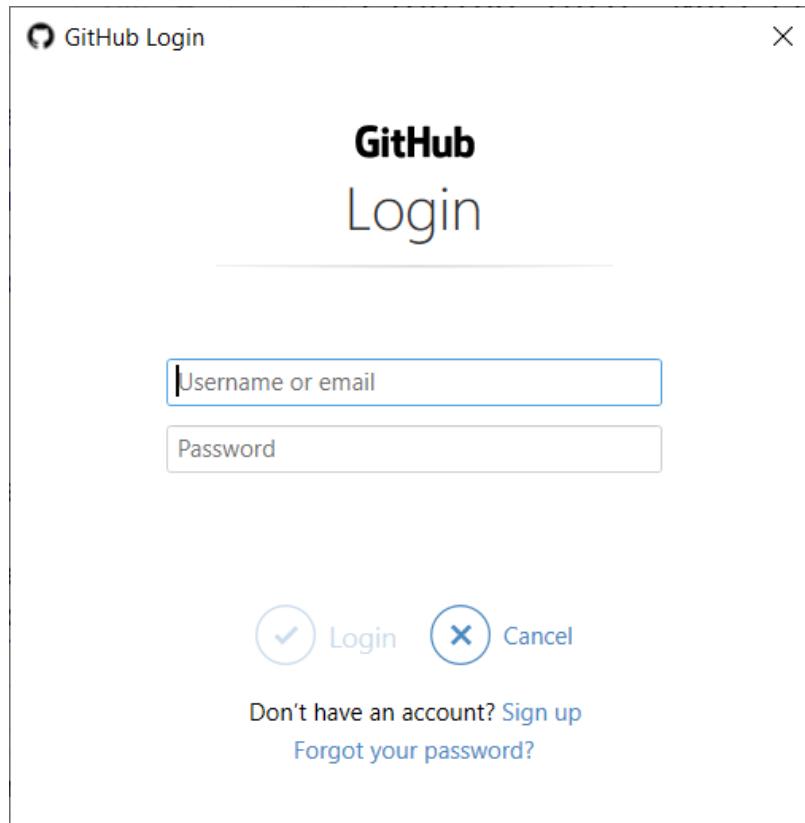


FIG. 3.8 : Identification HTTPS sur GitHub.

3.2.6 Pousser les premières modifications

La manipulation précédente a automatiquement poussé (*Push*) les modifications validées sur GitHub. Par la suite, il faudra cliquer sur le bouton “Push” de la fenêtre *Git* pour le faire.

Sur GitHub, les fichiers résultant des modifications enregistrées par git sont maintenant visibles.

Chaque *commit* réalisé localement est compté par git et un message “Your branch is ahead of ‘origin/master’ by *n* commits” affiché dans en haut de la fenêtre *Git* indique qu’il est temps de mettre à jour GitHub en poussant l’ensemble de ces *commits*. Cliquer sur le bouton “Push” pour le faire.

A ce stade, le projet doit disposer d’un fichier README.md qui présente son contenu sur GitHub. Son contenu minimal est un titre et quelques lignes de description :

```
# Nom du Projet  
Description en quelques lignes.
```

Il est conseillé d’utiliser des badges⁵, à placer juste après le titre, pour déclarer l’état de maturité du projet, par exemple :

⁵<https://github.com/orangemug/stability-badges>

```
![stability-wip](https://img.shields.io/badge/-)
stability-work_in_progress-lightgrey.svg)
```

3.2.7 Cloner un dépôt de GitHub

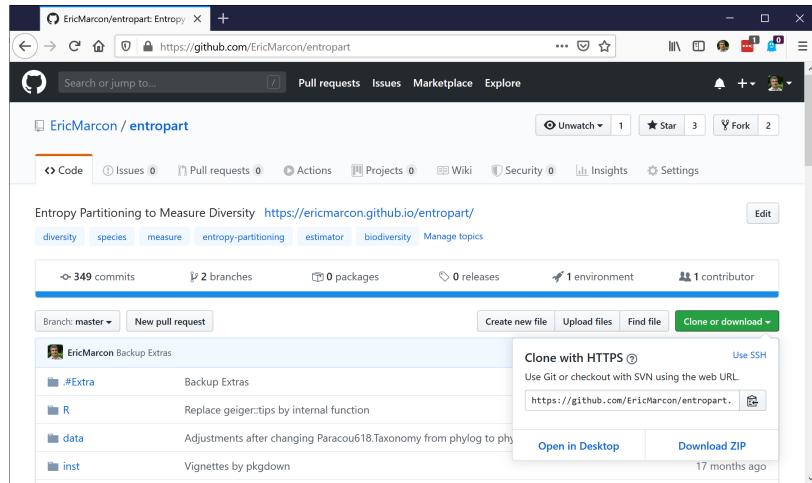


FIG. 3.9 : Clonage d'un dépôt à partir de *GitHub*.

Tout dépôt sur GitHub peut être installé (on dit *cloné*) sur le poste de travail en copiant son adresse qui apparaît en cliquant sur le bouton vert (figure 3.9).

Dans RStudio, créer un nouveau projet et, dans l'assistant, choisir “Version Control”, “Git” et coller l'adresse dans le champ “Repository URL”. Le nom répertoire à créer pour le projet est déduit automatiquement de l'adresse. Choisir le répertoire dans lequel celui du projet va être créé et cliquer sur “Create Project”. Le projet créé est lié au dépôt distant sur GitHub.

Pour travailler à plusieurs sur le même projet, le propriétaire du projet doit donner l'accès au projet à des collaborateurs (figure 3.10), c'est-à-dire d'autres utilisateurs GitHub dans les réglages du dépôt (*Settings*).

Les collaborateurs sont invités par un message envoyé par *GitHub*.

3.3 Usage courant

3.3.1 Tirer, modifier, valider, pousser

Toute séance de travail sur un projet commence en tirant (Bouton “Pull”) de la fenêtre *Git* pour intégrer au dépôt local les mises à jour effectuées sur GitHub par d'autres collaborateurs.

Les modifications apportées aux fichiers du projet sont ensuite prises en compte (cocher les cases *Staged*) et validées (*Commit*) avec un message explicatif. Une bonne pratique consiste à valider les modifications à chaque fois qu'une tâche élémentaire, qui peut être décrite dans le message explicatif, est terminée

3. GIT ET GITHUB

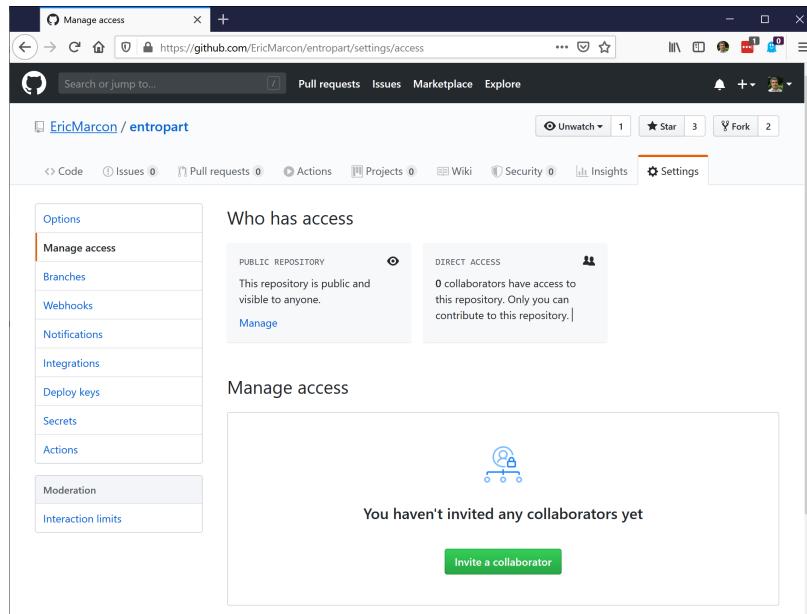


FIG. 3.10 : Attribution des droits d'accès sur GitHub.

plutôt que d'effectuer des *commits* regroupant de nombreux changements avec une description vague.

Dès que possible, pousser (*Push*) les mises à jour pour qu'elles soient visibles par les collaborateurs.

3.3.2 Régler les conflits

Il n'est pas possible de pousser les modifications validées si un collaborateur a modifié le dépôt distant sur GitHub. Il faut alors les tirer pour les intégrer au dépôt local avant de pousser les modifications fusionnées.

Un conflit a lieu si un *Pull* importe dans le fichier local une modification qui ne peut pas être fusionnée automatiquement parce qu'une modification contradictoire a eu lieu localement. Git considère chaque ligne comme un élément indivisible : la modification de la même ligne sur le dépôt distant et le dépôt local génère donc un conflit.

Git insère dans le fichier contenant un conflit les deux versions avec une présentation particulière :

```
<<<<<< HEAD # Version importée du conflit
Lignes en conflit, version importée
===== # limite entre les deux versions
Lignes en conflit, version locale
>>>>>> # Fin du conflit
```

Les lignes de formatage contenant les <<<, les ===== et les >>> doivent être supprimés et une seule version des lignes problématiques conservée, qui peut être différente des deux versions originales. La résolution du conflit doit être prise en compte et validée.

Pour limiter les conflits dans un document contenant du texte (typiquement, un document R Markdown), une bonne pratique consiste à traiter chaque phrase comme une ligne, terminée par un retour à la ligne qui ne sera pas visible dans le document mis en forme : un saut de ligne est nécessaire pour séparer les paragraphes.

3.3.3 Voir les différences

Dans la fenêtre *Git* de RStudio, le menu contextuel (affiché par un clic droit) “Diff” peut être utilisé pour afficher les modifications apportées à chaque fichier (figure 3.11).

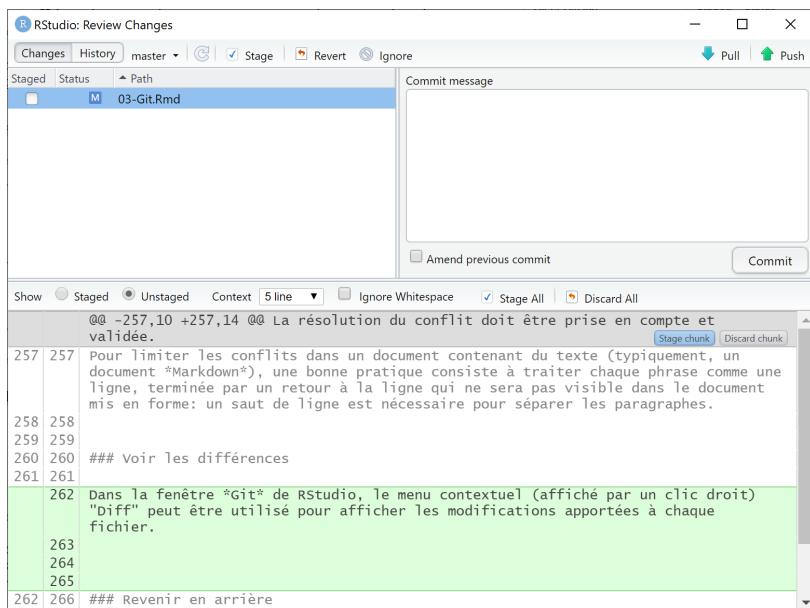


FIG. 3.11 : Différences entre le répertoire de travail et la tête.

3.3.4 Revenir en arrière

Le menu contextuel “Revert” permet d’annuler toutes les modifications apportées à un fichier (affichées par *Diff*) et de rétablir son contenu validé la dernière fois (son état dans la tête).

Il n’est pas simple de revenir en arrière au-delà de la dernière validation parce que les modifications ont pu être prises en compte par des collaborateurs : leur suppression rendrait le projet incohérent.

3.3.5 Voir l’historique

Le bouton en forme d’horloge de la fenêtre *Git* de RStudio affiche l’historique du projet (figure 3.12).

3. GIT ET GITHUB

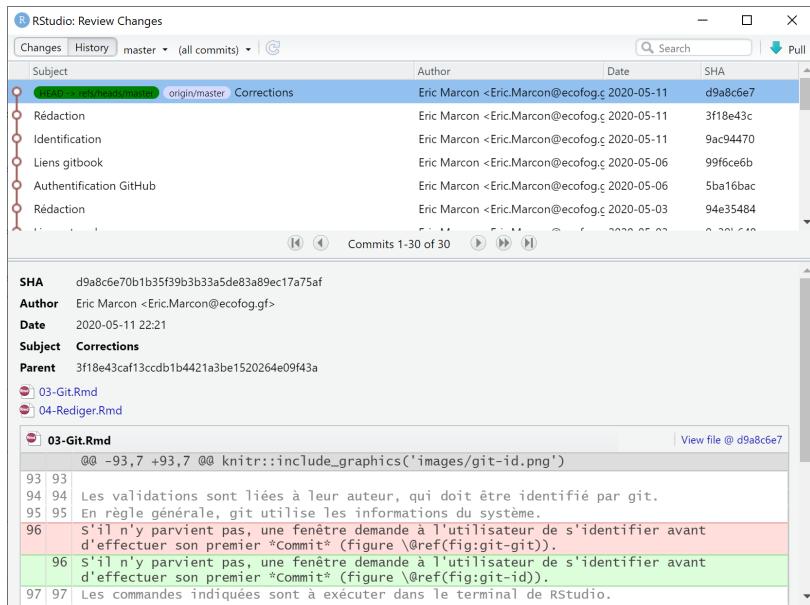


FIG. 3.12 : Historique des validations dans le dépôt.

En haut se trouve la tête, puis toutes les validations (*commits*) qui l'ont constituée. Pour chaque validation, les différences de chaque fichier peuvent être affichées en cliquant sur le nom du fichier dans la partie basse de la fenêtre.

3.4 Branches

Les branches d'un projet sont des versions différentes mais simultanées. Un usage typique est le développement d'une nouvelle fonctionnalité. Si son écriture prend du temps, le projet est perturbé par le chantier en cours : le code peut ne plus fonctionner. Si le développement s'avère impossible ou inutile, il faut pouvoir l'abandonner sans dommage. Pour l'isoler pendant sa réalisation et se permettre de le valider ou de l'abandonner à la fin, il faut le placer dans une branche.

La branche principale du projet s'appelle “master” ou “main” à partir de novembre 2020⁶. Elle doit toujours être dans un état stable : c'est elle qui est clonée à partir de GitHub par d'autres utilisateurs éventuels.

Le changement de convention pour le nom de la branche “master” fait qu'à partir de novembre 2020, les projets créés sur GitHub clonés dans RStudio ont pour branche principale “main” alors que les projets créés sur RStudio puis liés à GitHub conservent le nom “master”.

⁶<https://github.com/github/renaming>

3.4.1 Créer une nouvelle branche

Cliquer sur le bouton violet “New Branch” dans la fenêtre *git* de RStudio. Saisir son nom et cliquer sur “Create”.

La nouvelle branche est maintenant active.

Les commandes *git* peuvent aussi être exécutées dans le terminal (pour créer la branche et l’activer) :

```
git branch new_branch  
git checkout new_branch
```

3.4.2 Changer de branche

Sélectionner la branche à activer dans la liste des branches locales de la la fenêtre *git*.

Les *commits* s’appliquent à la branche active. Chaque branche se comporte comme une version différente du projet.

Attention : pour éviter la confusion, sauvegarder les modifications, prendre en compte et valider les changements avant de changer de branche.

3.4.3 Pousser la nouvelle branche

Les premières modifications de la nouvelle branche doivent être poussées en ligne de commande parce que les boutons “Push” et “Pull” de la fenêtre *Git* ne fonctionnent pas tant que la branche n’existe pas sur le dépôt distant.

Exécuter, dans le terminal :

```
git push -u origin new_branch
```

3.4.4 Comportement du système de fichier

A chaque changement de branche, *git* réécrit les fichiers du projet pour qu’ils reflètent l’état de la branche. Les changements peuvent être observés hors de RStudio, dans l’explorateur de fichier par exemple.

Les fichiers ignorés par *.gitignore* ne sont pas modifiés. Il est donc indispensable que les fichiers *.gitignore* des différentes branches soit identiques, sinon des fichiers ignorés dans une branche apparaîtront comme ajoutés dans la branche affichée après un changement.

Les branches de développement ont un contenu proche de celui de la branche principale. Ce n’est pas le cas de branches spécialisées vues plus loin, comme *gh-pages* (voir section 3.7) qui contient le site web de présentation du dépôt. Il est préférable de ne pas tenter d’afficher ces branches dans RStudio : leur contenu est produit automatiquement et ne doit pas être modifié manuellement. Si c’est indispensable, il faudra y copier le fichier *.gitignore* de la branche principale et garder à l’esprit que les fichiers ignorés appartiennent en réalité à une autre branche que celle affichée.

3.4.5 Fusionner avec merge

La fusion d'une branche de développement avec la branche principale marque l'atteinte de son objectif : son code va être intégré au projet. L'interface graphique de RStudio ne prévoit pas les fusions, il faut donc utiliser le terminal : tout d'abord, se placer dans la branche cible (possible avec l'interface graphique) :

```
git checkout master
```

Ensuite, fusionner :

```
git merge new_branch
```

Dans la majorité des situations, la fusion sera automatique (“Fast Forward”). Il est possible que des conflits apparaissent : utiliser la commande `git status` pour afficher la liste des fichiers concernés, les ouvrir, régler le conflit et effectuer un *commit*.

La branche fusionnée n'est pas supprimée : elle peut être utilisée à nouveau pour d'autres développements ou supprimée manuellement avec la commande suivante :

```
git branch -d new_branch
```

3.4.6 Fusionner avec une requête de tirage

L'autre façon de fusionner est plus formelle mais aussi plus générale : elle permet de fusionner une branche dans un dépôt d'un autre utilisateur pour y contribuer, ou de faire valider sa branche par un autre membre de l'équipe dans un projet collaboratif.

Pour contribuer au projet d'un autre utilisateur de GitHub⁷, il faut commencer par en créer un *fork*, c'est-à-dire une copie sous la forme d'un dépôt lié à l'original. Il sera possible de tirer les modifications de l'original pour rester à jour⁸ (par opposition à une simple copie instantanée possible en téléchargeant un Zip du projet) et, à la fin du développement, de fusionner le *fork* au dépôt original (par opposition à un clone qui ne permettrait pas de contribuer par la suite).

Ensuite, il faut créer une branche de développement comme précédemment, la modifier et finalement demander au propriétaire du dépôt de la fusionner. Ce processus est décrit en détail dans la documentation de git .

Dans le cadre plus simple d'une branche de son propre projet comme dans le cas d'un *fork*, la branche de développement est prête à être fusionnée. Elle doit avoir été poussée sur GitHub. Sur la page GitHub du projet, un bouton

⁷<https://git-scm.com/book/fr/v2/GitHub-Contributing-%C3%A0-un-projet>

⁸<https://ardalis.com/syncing-a-fork-of-a-github-repository-with-upstream/>

“Create Pull Request” permet de demander la fusion. Un message décrivant les modifications proposées avec leur argumentaire doit être ajouté.

Le propriétaire du projet (les membres de l’équipe dans le cadre d’un projet collaboratif, ou soi-même si l’équipe se réduit à une personne) est averti de la requête de tirage. Sur la page du projet original, il est possible de voir le message, la liste des modifications (chronologie des *commits* ou comparaison des fichiers), d’engager un discussion avec l’auteur de la requête... Si la requête n’est pas retenue, elle peut être fermée. Si elle est validée, le bouton “Merge Pull Request” permet de fusionner la branche de développement avec la branche “master” (ou une autre) du projet source.

Les requêtes de tirage sont le seul moyen de contribuer à un dépôt sur lequel on ne dispose pas de droits d’écriture. C’est aussi le moyen de fusionner une branche de développement dans son propre projet en gardant une trace explicite (dans la rubrique *Pull requests* de la page GitHub du projet). Dans le cadre d’un projet collaboratif, les propositions d’un membre (auteur de la requête) peuvent être validées par un autre (qui accepte la fusion).

3.5 Usage avancé

3.5.1 Commandes de git

Au-delà de l’usage courant permis par l’interface graphique de RStudio, des manipulations avancées des projets sont permises en utilisant git en ligne de commande. Quelques exemples utiles sont présentés ici.

Un petit guide des commandes est proposé par Roger Dudler⁹. Il résume les commandes essentielles, donc intégrées à l’interface graphique de RStudio. Des liens vers des références plus complètes sont donnés en bas de la page.

3.5.2 Taille d’un dépôt

Pour connaître l’espace disque occupé par un dépôt, utiliser la commande `git count-objects -vH`¹⁰.

Les données pour ce document au stade de la rédaction sont présentées à titre d’exemple.

```
$ git count-objects -v
count: 200
size: 2.66 MiB
in-pack: 0
packs: 0
size-pack: 0
prune-packable: 0
garbage: 0
size-garbage: 0
```

⁹<https://rogerdudler.github.io/git-guide/index.fr.html>

¹⁰<https://git-scm.com/docs/git-count-objects>

La taille totale est sur la ligne *size*. Les packs sont une méthode utilisée par git pour réduire la taille du dépôt : des fichiers similaires sont stockés sous la forme d'une partie commune et de différences. La ligne *prune-packable* donne la taille d'objets stockés à la fois sous forme individuelle et dans des packs. Si leur taille est importante, exécuter `git prune-packed` pour la ramener à zéro.

La ligne *size-garbage* donne la taille des objets qui peuvent être supprimés. `git gc` les supprime, mais pas seulement : il optimise le stockage.

```
$ git gc
Enumerating objects: 194, done.
Counting objects: 100% (194/194), done.
Delta compression using up to 8 threads
Compressing objects: 100% (188/188), done.
Writing objects: 100% (194/194), done.
Total 194 (delta 83), reused 0 (delta 0)

$ git count-objects -vH
count: 1
size: 5.72 KiB
in-pack: 194
packs: 1
size-pack: 4.00 MiB
prune-packable: 0
garbage: 0
size-garbage: 0 bytes
```

Ici, la majorité des objets du dépôt a été placée dans un pack (mais sa taille est supérieure à celle des objets individuels).

Il est généralement inutile d'effectuer la collecte des déchets manuellement : git gère bien l'organisation de ses dépôts.

GitHub limite la taille des dépôts. En mai 2020, la limite est de 100 Go. La taille de tous les dépôts d'un utilisateur authentifié peut être affichée dans les réglages de son compte (“Personal Settings”, “Repositories”)¹¹.

3.5.3 Supprimer un dossier

Toutes les modifications apportées à un dépôt sont stockées dans son historique. Il peut être utile d'en supprimer dans quelques cas particuliers :

- si un fichier contenant des informations confidentielles a été validé par mégarder. La validation de sa suppression ne le retire pas de l'historique, et les informations confidentielles restent visibles en consultant l'historique.
- si des fichiers volumineux ne sont plus nécessaires, par exemple des fichiers PDF produits par R Markdown (chapitre 4), binaires (donc inadaptables à git) et reproductibles à partir du code.

Typiquement, le dossier `docs` est utilisé pour stocker les documents produits à partir de code R Markdown. Les fichiers HTML et PDF doivent s'y trouver

¹¹<https://github.com/settings/repositories>

pour constituer les pages GitHub du projet. Chaque modification du dépôt génère une nouvelle version de ces fichiers dont le volume de l'historique devient rapidement considérable. Une solution efficace consiste à déléguer la création de ces fichiers à un système d'intégration continue (chapitre 6) et à retirer le dossier `docs` de la branche principale (*master*) du dépôt. Il faut alors supprimer tout son historique pour récupérer la place qu'il occupe, qui peut être l'essentiel de la taille du dépôt.

Les commandes de suppression complète d'un dossier d'un dépôt sont présentées ici¹². Le dépôt doit être propre, c'est-à-dire sans modifications non validées, et les versions distantes et locales synchronisées.

Les trois commandes suivantes suppriment complètement le dossier `docs` de l'historique du dépôt git :

```
git filter-branch --tree-filter "rm -rf docs" |>
    --prune-empty HEAD
git for-each-ref --format="%({refname})" refs/original/ |>
    | xargs -n 1 git update-ref -d
```

Le dossier n'est pas supprimé du répertoire de travail. Il doit donc être ajouté au fichier `.gitignore` pour ne plus être suivi. La modification de `.gitignore` doit être validée. Ces opérations peuvent être réalisées avec l'interface de RStudio ou en ligne de commande :

```
echo docs/ >> .gitignore
git add .gitignore
git commit -m 'Removing docs folder from git history'
```

Le nettoyage du dépôt est nécessaire pour supprimer physiquement les données retirées :

```
git gc
```

Enfin, le dépôt doit être poussé. L'option `--force` implique le remplacement du contenu du dépôt distant par celui du dépôt local : toutes les modifications faites par des collaborateurs sont effacées, c'est pourquoi cette opération de nettoyage implique l'arrêt complet du projet pendant qu'elle a lieu.

```
git push origin master --force
```

Ce code peut être utilisé pour supprimer totalement n'importe quel fichier ou dossier d'un dépôt en remplaçant simplement `docs` dans la commande `git filter-branch` initiale. La réduction de la taille du dépôt peut être suivie en utilisant `git count-objects -vH` avant l'opération, avant `git gc` (la taille du dépôt reste stable mais a été déplacée vers *garbage*) et à la fin (la taille du dépôt est sensiblement réduite).

¹²<https://stackoverflow.com/questions/10067848/remove-folder-and-its-contents-from-git-githubs-history>

3.5.4 Revenir en arrière

Il est possible de restaurer un dépôt dans un état précédent en plaçant sa tête (figure 3.6) au niveau d'un ancien *commit*. Toutes les modifications ultérieures sont alors détruites. Cette opération ne doit pas être réalisée sur un dépôt partagé : les autres utilisateurs ne pourraient plus pousser leurs modifications.

Afficher l'historique du dépôt et rechercher l'identifiant (SHA) du dernier *commit* à conserver. Dans le terminal de RStudio, exécuter :

```
git reset --hard <SHA>
git push -f
```

Tout l'historique du dépôt après le point de restauration choisi est perdu.

Une méthode moins radicale et utilisable sur un dépôt partagé consiste à exécuter un *commit* qui annule les modifications d'un autre mais ne détruit aucune donnée de l'historique. Cette opération n'annule qu'un seul *commit* à la fois et doit donc être répétée pour en annuler plusieurs, en commençant par le plus récent. Dans le terminal de RStudio, exécuter :

```
git revert <SHA>
```

Pour annuler le dernier *commit*, exécuter :

```
git revert HEAD
```

Utiliser HEAD évite simplement de rechercher l'identifiant correspondant.

3.6 Données confidentielles dans un dépôt public

Un dépôt public sur GitHub pose problème quand des données utilisées dans le projet ne le sont pas.

Une solution peu satisfaisante consiste à ne pas inclure les données au projet, ce qui le rend non reproductible. Une meilleure solution est de les crypter, en permettant à certains utilisateurs de les décrypter. C'est l'objet du package *secret*.

Un coffre-fort (dossier *vault*) est créé dans le projet. Il contient une liste d'utilisateurs autorisés : chacun d'entre eux doit disposer d'une paire de clés de cryptage, une clé publique incluse dans le coffre-fort et une clé privée, gardée secrète. Les données sont cryptées avec toutes les clés publiques disponibles (et donc dupliquées). Les utilisateurs utilisent ensuite chacun sa clé privée pour le décryptage.

Pour ne pas multiplier les copies des données, le propriétaire du dépôt a intérêt à créer un utilisateur générique pour le projet, dont il communiquera la clé privée hors de GitHub. Le coffre contiendra les clés du propriétaire du projet et de l'utilisateur générique seulement. En cas de compromission de la clé privée de l'utilisateur générique, il suffira de le retirer du coffre-fort et d'en créer un nouveau.

3.6.1 Génération d'une paire de clés pour le propriétaire du projet

Les clés sont générées par le logiciel *ssh*, installé avec *git* ou par défaut sous Linux.

La procédure est identique à celle de la section 1.4.3, mais la clé utilisée doit être au format RSA (pris en charge par le package **secret**, contrairement au format ed25519, plus sûr, utilisé pour l'authentification sur GitHub).

Exécuter la commande suivante dans le terminal de RStudio pour créer une clé RSA :

```
ssh-keygen -t rsa -b 4096 -C "user@email"
```

Stocker la clé publique sur GitHub dans “Settings > SSH and GPG Keys”. Repérer la position de la clé : si une clé d'authentification a déjà été enregistrée pour deux postes de travail par exemple, la clé RSA sera la troisième.

3.6.2 Génération d'une paire de clés pour le projet

Générer une clé au format RSA dans le terminal de RStudio :

```
ssh-keygen -t rsa -b 4096"
```

- Entrer le nom de la clé : NomDuProjet.rsa.
- Ne pas saisir de phrase de validation (mot de passe) pour permettre l'utilisation de la clé sans interaction.

La clé privée NomDuProjet.rsa ne doit être diffusée qu'aux ayant-droits du projet. Il faut donc ajouter la ligne *.rsa au fichier .gitignore du projet pour ne pas pousser la clé sur GitHub.

Pour permettre l'intégration continue du projet (chapitre 6), la clé privée doit être stockée comme un secret du dépôt GitHub contenant le projet. Appliquer la procédure de la section 6.2.2 pour créer un secret nommé “RSA” et coller le contenu du fichier NomDuProjet.rsa dans le champ “Value” du formulaire.

L'utilisation du secret est décrite dans la section 6.2.5.

3.6.3 Création d'un coffre-fort

Exécuter :

```
library("secret")
vault <- "vault"
create_vault(vault)
```

3.6.4 Ajout des utilisateurs

Le propriétaire du projet est ajouté à partir de sa clé publique stockée sur GitHub, qui est la troisième dans notre exemple.

```
# Identifiant GitHub du propriétaire du projet
github_user <- "EricMarcon"
# Lecture et stockage de la clé, i est le numéro de la clé
add_github_user(github_user, vault = vault, i = 3)
```

La clé de l'utilisateur générique du projet est ajoutée par :

```
library("openssl")

## Linking to: OpenSSL 1.1.1m  14 Dec 2021

project_id <- "NomDuProjet"
# Lecture de la clé
rsa_project <- read_pubkey(paste0(project_id, ".rsa.pub"))
# Ajout au coffre-fort
add_user(project_id, public_key = rsa_project, vault = vault)
```

3.6.5 Stockage des données

Les données, stockées dans des variables de R, sont stockées une à une par la fonction `add_secret()`. Dans l'exemple suivant, la variable s'appelle X et vaut 1.

```
X <- 1
add_secret(
  # Nom de la donnée
  "X",
  # Valeur
  value = X,
  # Utilisateurs autorisés: propriétaire et générique
  users = c(paste0("github-", github_user), project_id),
  # Coffre-fort
  vault = vault)
```

Le contenu du coffre-fort peut être vérifié :

```
# Liste des données du coffre
list_secrets(vault = vault)

##   secret      email
## 1      X github-E.....

# Liste des propriétaires de la donnée 'X'
list_owners("X", vault = vault)

## [1] "github-EricMarcon" "NomDuProjet"
```

Les données seront lues dans le code du projet par la commande `get_secret()`. La clé privée de l'utilisateur générique du projet, communiquée par un moyen sécurisé aux ayant-droits, doit se trouver dans le dossier du projet.

```
# Sélection de la clé privée
Sys.setenv(USER_KEY = usethis::proj_path(paste0(project_id, ".rsa")))
```

```
## v Setting active project to '/Users/runner/work/travailR/
## travailR'
```

```
# Lecture de la donnée 'X'
get_secret("X", vault = vault)
```

```
## [1] 1
```

La clé peut être vérifiée :

```
local_key()
```

```
## [4096-bit rsa private key]
## md5: e81dcb0745a755286c2dc1fc4c6ad117
## sha256: cca11ef82e17c3b77b699e7f3c23e083e8f0f79cb70be8274799f076c44b0c2d
```

3.7 Pages GitHub

Tout projet sur GitHub doit avoir contenir un fichier `README.md` pour le présenter. Ce fichier est écrit au format Markdown.

Le fichier peut être placé dans le dossier `docs` pour fournir à fois la page d'accueil du dépôt et de son site web. Le package **memoiR** fournit des commandes permettant d'automatiser ces tâches dans les projets de documents. Un dépôt contenant un article écrit en R Markdown (voir section 4.3.2) est utilisé comme exemple¹³.

Son fichier `README.md` existe aux deux emplacements : il est écrit par le développeur à la racine du projet et dupliqué par `GitHubPages.R`.

3.7.1 Activation

Pour activer les pages GitHub, il faut ouvrir les propriétés du dépôt (*Settings*) et modifier la rubrique “GitHub Pages” (dans “Options”). Sélectionner la branche du projet et le dossier contenant les pages web, ici : `master` et `/docs`. En option, le choix d'un thème personnalise l'apparence des pages.

Le site web est accessible à une adresse¹⁴ du domaine `github.io`.

Le fichier `README.md` affiché en page d'accueil a un aspect très différent mais le même contenu que celui affiché avec le code sur la page du dépôt dans GitHub.

¹³<https://github.com/EricMarcon/Krigeage>

¹⁴<https://EricMarcon.github.io/Krigeage/>

3. GIT ET GITHUB

L'intérêt des pages GitHub est de permettre un accès simple aux documents formatés quand le dépôt contient une production écrite et ou à la documentation des packages R. Ces contenus seront présentés dans le chapitre suivant.

Un site web principal est proposé avec chaque compte GitHub, à l'adresse <https://GitHubID.github.io>¹⁵. Il sera utilisé pour héberger un site web personnel produit par **blogdown**.

3.7.2 Badges

Les badges sont de petites images, éventuellement mises à jour dynamiquement, qui renseignent rapidement sur le statut d'un projet. Ils doivent être placés immédiatement après le titre du fichier README.md.

Une bonne pratique consiste à indiquer l'avancement dans le cycle de vie du projet. Les badges correspondants sont listés sur le site du Tidyverse¹⁶.

Leur code Markdown est le suivant :

```
![stability-wip]
(https://img.shields.io/badge/lifecycle-maturing-blue.svg)
```

Le package **usethis** simplifie leur création en plaçant le code nécessaire dans le presse-papier. Il suffit ensuite de le coller dans le fichier.

```
usethis::use_lifecycle_badge("maturing")
```

¹⁵Exemple : <https://EricMarcon.github.io/Krigeage/>

¹⁶<https://www.tidyverse.org/lifecycle/>

RÉDIGER

R et RStudio permettent de rédiger efficacement des documents de tous formats, du simple bloc-note à la thèse, en passant par des diaporamas. Les outils pour le faire sont l'objet de ce chapitre, complété par la production de sites web (y compris un site personnel).

4.1 Bloc-note Markdown (R Notebook)

Dans un fichier .R, le code doit toujours être commenté pour faciliter sa lecture. Quand l'explication du code nécessite plusieurs lignes de commentaire par ligne ou bloc de code, il est temps d'inverser la logique et de placer le code dans un texte.

Le concept de programmation lettrée (*literate programming*) a été développé par KNUTH (1984). Il s'agit de décrire les objectifs et les méthodes par du texte, dans lequel le code s'intègre.

L'outil le plus simple est le bloc-note Markdown (Menu “File > New File > R Notebook”). Le modèle de document contient son mode d'emploi.

Le langage qui permet de formater le texte est Markdown¹, un langage de balisage simple à utiliser :

- Les paragraphes sont séparés par des sauts de ligne ;
- Le document est structuré par des titres : leur ligne commence par un nombre de # correspondant à leur niveau ;
- Les formats de caractères sont limités à l'essentiel : italique ou gras (texte entouré par une ou deux *);
- D'autres codes simples permettent tous les formatages utiles.

¹<https://fr.wikipedia.org/wiki/Markdown>

4. RÉDIGER

Ce langage est le pivot du logiciel pandoc², dédié à la conversion de documents de formats différents.

Le package **rmarkdown** (XIE 2015) fait le lien entre R et Markdown, en s'appuyant sur l'interface de RStudio qui n'est pas indispensable mais simplifie énormément son utilisation. Le dialecte de Markdown utilisé par le package est appelé *R Markdown*. Sa syntaxe est résumée dans une antisèche³. Sa documentation complète est en ligne (XIE et al. 2018).

Les équations sont écrites au format LaTeX⁴.

L'organisation la plus simple d'un document *R Markdown* est visible dans le modèle de bloc-note. Il commence par un en-tête au format YAML⁵ :

```
---
```

```
title: "R Notebook"
```

```
output: html_notebook
```

```
--
```

La première entrée est le titre, la seconde le format de sortie : plus précisément le nom de la fonction chargée de traiter le document.

Le document contient du texte formaté en Markdown et des bouts de code (*code chunks*) entourés par trois accents graves (la syntaxe markdown d'un bloc de code) et une description du langage, ici r. Ces bouts de code sont traités par **knitr** qui transforme le résultat de l'exécution du code R en Markdown et l'intègre au texte du document.

Traiter un document R Markdown s'appelle le *tricoter (knit)*. La chaîne de production est la suivante :

- **knitr** traite les bouts de code : calculs, production de figures ;
- **rmarkdown** intègre la production des bouts de code et texte pour produire un fichier Markdown standard ;
- pandoc (installé avec RStudio) convertit ce fichier au format HTML, LaTeX ou Word ;
- LaTeX produit un fichier PDF quand ce format est demandé.

RStudio permet de lancer le tricot par des boutons plutôt que par des commandes : dans la fenêtre source (celle du haut à gauche), un bouton “Knit” accompagne les documents R Markdown. Pour les bloc-notes R Markdown, il est remplacé par un bouton “Preview” avec les mêmes fonctions. Il peut être déroulé pour choisir le format de sortie : HTML, Word, PDF (en passant par LaTeX) et, pour les bloc-notes, une commande “Preview” qui affiche le document en HTML sans exécuter les bouts de code pour gagner du temps. Dès le premier tricot au format Word ou HTML, on remarquera que le bouton “Preview” disparaît.

Au final, l'utilisation de R Markdown combine plusieurs avantages :

²<https://fr.wikipedia.org/wiki/Pandoc>

³<https://rstudio.com/wp-content/uploads/2015/02/rmarkdown-cheatsheet.pdf>

⁴https://fr.wikibooks.org/wiki/LaTeX/%C3%89crire_des_math%C3%A9matiques

⁵<https://fr.wikipedia.org/wiki/YAML>

- La simplicité de la rédaction : le texte brut est plus facile à lire et à formater qu'en LaTeX par exemple ;
- L'automatisation de la production : le formatage et la mise en page sont entièrement automatiques ;
- La reproductibilité : chaque document peut être autosuffisant accompagné de ses données. Relancer le tricotage régénère entièrement le document, y compris les calculs nécessaires et la production des figures.

Elle a aussi quelques inconvénients :

- Le formatage dépend de modèles, et développer de nouveaux modèles n'est pas simple ;
- Les erreurs de tricot sont parfois difficiles à corriger, notamment quand elles interviennent à l'étape de la compilation LaTeX ;
- La reproductibilité consomme du temps de calcul. Pour limiter ce problème, un système de cache permet de ne pas réévaluer tous les bouts de code R à chaque modification du texte. La production de gros documents peut aussi être déléguée à un système d'intégration continue (chapitre 6).

4.2 Modèles R Markdown

Des modèles de document plus élaborés que le bloc-note sont fournis par des packages, dont **rmarkdown**. Ils sont accessibles par le menu “File > New File > R Markdown...” (figure 4.1)).

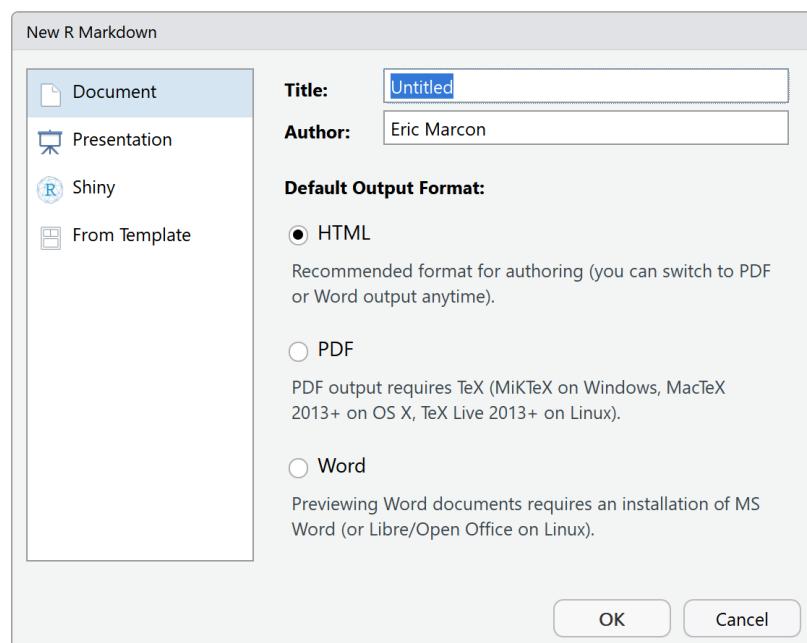


FIG. 4.1 : Nouveau document Markdown à partir d'un modèle.

4. RÉDIGER

Les modèles les plus simples sont *Document* et *Presentation*. Les informations à fournir sont le titre et le nom de l'auteur, et le format du document attendu (qui pourra être modifié plus tard). Ces modèles créent un seul fichier dont l'enregistrement ne sera obligatoire qu'au moment de tricoter.

La syntaxe est la même que celle du bloc-note. Dans l'entête, une entrée supplémentaire est utilisée pour la date, qui peut être calculée par R à chaque tricot :

```
date: "|r format(Sys.Date(), '%d/%m/%Y')|"
```

Remplacer les barres verticales | de l'exemple ci-dessus par des guillemets inversés : ce document étant écrit avec R Markdown, la date serait calculée et affichée à la place du code si les guillemets inversés étaient utilisés directement.

Le code R en ligne (par opposition aux bouts de code) peut être utilisé partout dans un document R Markdown, y compris dans l'entête pour l'affichage de la date. Il commence par un guillemet inversé suivi de r et se termine par un autre guillemet inversé.

Les documents peuvent être tricotés au format HTML, PDF (via LaTeX) ou Word. L'entête du fichier R Markdown est réécrit quand le tricot est lancé par le bouton de RStudio qui place en premier le format de sortie utilisé et l'ajoute si nécessaire.

Les présentations peuvent être tricotées dans deux formats HTML, ioslide⁶ ou Slidy⁷, au format Beamer (PDF)⁸ ou en Powerpoint⁹.

Le niveau 2 de plan (##) marque le changement de diapositive.

Du code supplémentaire, présenté dans les documentations des formats HTML, permet d'utiliser des fonctionnalités spécifiques.

Ces modèles sont simples mais assez peu utiles : le bloc-note R est plus facile à utiliser que le modèle de document pour des documents minimalistes. Des modèles plus élaborés sont disponibles.

4.3 Articles avec bookdown

R Markdown ne permet pas de rédiger un article scientifique. La bibliographie ne pose pas de problème parce qu'elle est gérée par pandoc pour les documents HTML ou Word et sous-traitée à LaTeX pour les documents PDF. Les équations, figures et tableaux sont numérotés par LaTeX mais pas en HTML. Les références croisées (les renvois à un numéro de figure par exemple) ne sont pas supportés. Enfin, les légendes de figures ou tableaux ne supportent que du texte brut, sans aucun formatage.

⁶<https://bookdown.org/yihui/rmarkdown/ioslides-presentation.html>

⁷<https://bookdown.org/yihui/rmarkdown/slidy-presentation.html>

⁸<https://bookdown.org/yihui/rmarkdown/beamer-presentation.html>

⁹<https://bookdown.org/yihui/rmarkdown/powerpoint-presentation.html>

bookdown comble ces manques. Le package a été conçu pour la rédaction d'ouvrages comportant plusieurs chapitres mais peut être utilisé pour des articles. Le package ne fournit pas directement de modèles.

Le package **memoiR** fournit les modèles présentés ici. Il doit être installé.

4.3.1 Ecrire

Les principales caractéristiques de Markdown sont résumées ici. Une formation rapide et plus complète est proposée par RStudio¹⁰.

Le texte est écrit sans aucun autre formatage que les retours à la ligne. Un simple retour à la ligne n'a aucun effet sur le document produit : il permet de séparer les phrases pour simplifier le suivi du code source par git.

Un saut de ligne marque un changement de paragraphe.

Les différents niveaux de plan sont désignés par le nombre de croisillons correspondant en début de ligne : # pour un titre de niveau 1, ## pour un titre de niveau 2, etc. Un espace sépare les croisillons et le texte du titre.

Les listes à puces sont marquées par un tiret (suivi d'un espace) en début de ligne. Un saut de ligne est nécessaire avant le début de la liste mais les éléments de la liste sont séparés par un simple retour à la ligne. Les listes indentées sont créées en insérant 4 espaces avant le tiret de début de ligne. Enfin, les listes numérotées sont créées de la même façon en remplaçant les tirets par des nombres, dont la valeur n'a pas d'importance.

Dans le texte, les parties en italique sont entourées par une étoile ou un tiret bas (*italique*), alors que deux étoiles marquent le gras.

Code R

Le code R est inclus dans des bouts de code (*code chunks*) créés facilement en cliquant sur le bouton “Insert a new code chunk” au-dessus de la fenêtre du code source dans RStudio. Ils commencent et se terminent par trois guillemets inversés sur une nouvelle ligne. Ces bouts de code peuvent contenir du code R mais aussi Python par exemple : le type de code est indiqué dans l'entête sur la première ligne, avant le nom du bout de code, puis une liste d'options séparées par des virgules, par exemple :

```
```{r cars, echo=TRUE}
```

Le nom et les options sont facultatifs : l'entête minimal est {r}.

Les options les plus utiles sont :

- echo pour afficher (=TRUE) ou cacher (=FALSE) le code ;
- message=FALSE pour cacher les messages d'ouverture de certains packages ;

---

<sup>10</sup><https://rmarkdown.rstudio.com/lesson-1.html>

#### 4. RÉDIGER

---

- `warning=FALSE` pour cacher les avertissements.

Les options par défaut sont déclarées dans le bout de code nommé “Options” au début du document Markdown, dans la fonction `opts_chunk$set()`. L’option `echo` doit être mise à `FALSE` par défaut pour un article scientifique par exemple.

### Figures

```
plot(pressure)
```

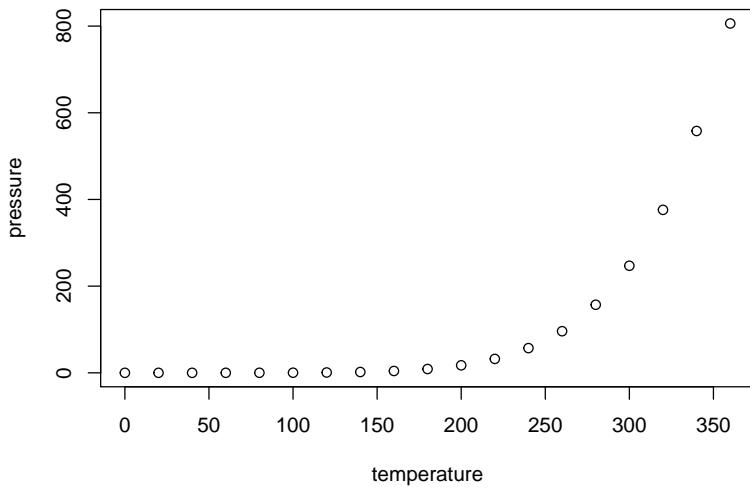


FIG. 4.2 : Titre de la figure

Les figures peuvent être créées par le code R (figure 4.2). Avec Bookdown, une étiquette est associée à chaque figure : son nom est `fig:xxx` où `xxx` est le nom du bout de code R. Les renvois se font avec la commande `\@ref(fig:xxx)`.

L’entête du bout de code de la figure 4.2 est :

```
```{r pressure, fig.cap="Titre de la figure"}
```

Il contient au minimum le nom de la figure et sa légende. Si la légende est longue, l’entête est peu lisible. De plus, la légende est limitée à du texte simple. Pour des légendes plus élaborées, il est possible de déclarer la légende dans un paragraphe séparé qui commence par le texte `(ref:NomFigure)`. La figure 4.3 bénéficie d’une légende améliorée.

Le texte de `fig.cap`, “Titre de la figure” précédemment, est remplacé par `(ref:pressure)` à l’intérieur des guillemets qui sont conservés et la légende est

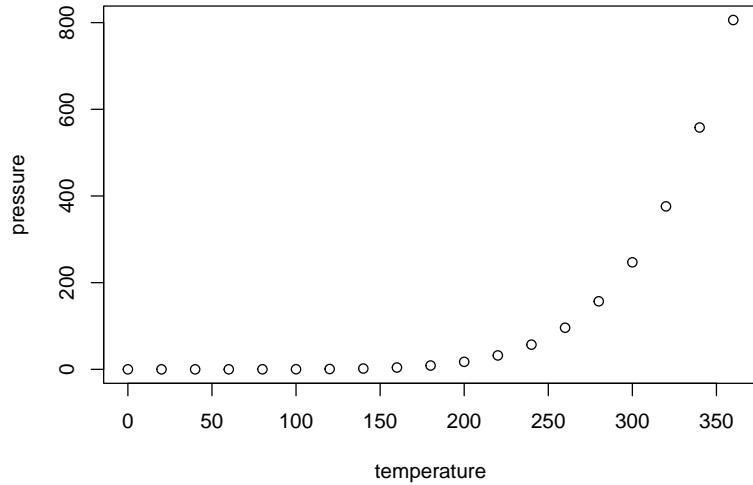


FIG. 4.3 : Titre avec *italique*, maths ($\sqrt{\pi}$) et renvoi vers la figure 4.2

saisie dans un paragraphe commençant par (ref:pressure) suivi d'un espace. Les légendes sont limitées à un paragraphe unique.

Si une table des figures est utilisée (option `lof: true` dans l'entête), une légende courte est nécessaire en plus de la légende complète. Elle est déclarée dans `fig.scap`.

Les figures qui ne sont pas créées par R mais proviennent de fichiers sont intégrées dans un bout de code par la fonction `include_graphics()` dont l'argument est le fichier contenant l'image à afficher. Placer systématiquement ces fichiers dans le dossier `images` pour une bonne organisation.

Tableaux

Les séparateurs horizontaux - et verticaux | permettent de dessiner un tableau selon la syntaxe de Markdown, mais ce n'est pas la meilleure méthode.

Les tableaux peuvent aussi être produits par du code R. Le contenu du tableau est dans un dataframe. La fonction `kable` du package `knitr` prépare le tableau pour l'affichage et passe le résultat à la fonction `kable_styling` du package `kableExtra` pour le formatage final.

```
library("tidyverse")
mes_iris <- head(iris)
names(mes_iris) <- c("Longueur sépales ($l_s$)", "Largeur",
"Longueur pétales", "Largeur", "Espèce")
knitr::kable(mes_iris, caption = "Tableau créé par kable",
booktabs = TRUE, escape = FALSE) %>%
kableExtra::kable_styling(bookstrap_options = "striped",
full_width = FALSE)
```

TAB. 4.1 : Tableau créé par kable

Longueur sépales (l_s)	Largeur	Longueur pétales	Largeur	Espèce
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5.0	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa

La légende est précisée par l'argument `caption` et le référencement est possible parce que le tableau reçoit une étiquette dont le nom est `tab`: suivi du nom du bout de code (tableau 4.1). Comme pour les figures, une légende améliorée peut être écrite dans un paragraphe séparé. Une légende courte pour une éventuelle liste des tableaux (option `lot`: `true` dans l'entête) est déclarée dans l'argument `caption.short` de `kable()`.

Utiliser systématiquement l'argument `booktabs = TRUE` pour que l'épaisseur des lignes de séparation soit optimale en LaTeX. Comme le tableau contient des mathématiques (dans le nom de la première colonne), l'option `escape = FALSE` est nécessaire.

L'option de style `bootstrap_options = "striped"` fournit des tableaux plus lisibles en HTML. Enfin, l'option `full_width = FALSE` permet d'ajuster la largeur du tableau à son contenu au lieu d'occuper toute la largeur disponible.

Le package **flextable** permet de réaliser des tableaux plus élaborés, comme dans l'exemple suivant qui affiche en couleur les longs sépales.

```
library("flextable")

## 
## Attaching package: 'flextable'

## The following objects are masked from 'package:spatstat.geom':
## 
##     border, rotate

## The following object is masked from 'package:purrr':
## 
##     compose

## The following objects are masked from 'package:kableExtra':
## 
##     as_image, footnote

# Rappel du jeu de données initial iris
iris %>%
  # Premières lignes
head() %>%
  # Création d'un objet flextable
```

```
flextable() %>%
  # Titre des colonnes
set_header_labels(Sepal.Length = "Longueur sépales", Sepal.Width = "Largeur",
  Petal.Length = "Longueur pétales", Petal.Width = "Largeur",
  Species = "Espèce") %>%
  # Sélection des longs sépales (>5) et affichage en
  # rouge
color(~Sepal.Length > 5, ~Sepal.Length, color = "red")
```

Longueur sépales	Largeur	Longueur pétales	Largeur Espèce
5.1	3.5	1.4	0.2 setosa
4.9	3.0	1.4	0.2 setosa
4.7	3.2	1.3	0.2 setosa
4.6	3.1	1.5	0.2 setosa
5.0	3.6	1.4	0.2 setosa
5.4	3.9	1.7	0.4 setosa

La documentation du package¹¹ est disponible en ligne, ainsi qu'une galerie¹².

flextable ne supporte pas la numérotation des légendes hormis dans les documents Word. Cette limite est rédhibitoire.

Maths

Les équations au format LaTeX peuvent être insérées en ligne, comme $A = \pi r^2$ (code : \$A=\pi r^2\$) ou isolées (les \$ sont doublés) comme

$$e^{i\pi} = -1.$$

Elles peuvent être numérotées : voir équation (4.1), en utilisant l'environnement \equation.

$$A = \pi r^2. \tag{4.1}$$

L'équation numérotée est créée par le code suivant :

```
\begin{equation}
A = \pi r^2.
\label{eq:disque}
\end{equation}
```

Références croisées

Les figures et tableaux ont une étiquette générée automatiquement, identique au nom du bout de code préfixé par fig: et tab::

¹¹<https://ardata-fr.github.io/flextable-book/>

¹²<https://ardata-fr.github.io/flextable-gallery/gallery/>

4. RÉDIGER

Pour les équations, l'étiquette est ajoutée manuellement par le code (\#eq:xxx) avant la fin de l'équation.

Les sections peuvent recevoir une étiquette en terminant leur titre par {#yyy}. Les sections reçoivent par défaut une étiquette implicite¹³ correspondant à leur texte, en minuscules, où les caractères spéciaux sont remplacés par des tirets. Les étiquettes implicites sont instables (elles changent avec le titre de la section) et difficiles à prévoir : c'est pourquoi il est conseillé d'ajouter une étiquette explicite à chaque section faisant l'objet d'un renvoi. C'est le cas des chapitres, pour lesquels le nom du fichier HMTL produit est identique à l'étiquette. Les étiquettes de chapitres doivent respecter les règles de nomenclature des fichiers en ne contenant pas de caractères spéciaux.

Des signets peuvent aussi être placés librement dans le texte avec la commande (ref:zzz).

Dans tous les cas, l'appel à la référence est fait par la commande \@ref(ref:zzz).

Bibliographie

Les références bibliographiques au format bibtex doivent être incluses dans le fichier .bib déclaré dans l'entête du document Markdown.

```
bibliography: references.bib
```

Ce fichier peut être créé et maintenu à jour par Zotero installé avec l'extension Better BibTeX (voir section 1.6). Il suffit pour cela de créer une collection Zotero correspondant au projet et d'y glisser les références pertinentes. Utiliser ensuite le menu contextuel “Exporter la collection...” et sélectionner :

- Format : “Better BibTeX” pour les articles et présentations ou “Better Bi-bLaTeX” pour les mémoires, selon que la bibliographie est gérée par bibtex et natbib ou biber et biblatex pour la production de documents PDF.
- Cocher la case “Garder à jour” pour que toute modification dans Zotero soit exportée automatiquement.
- Cliquer sur “OK” puis choisir le nom du fichier (references.bib) et son emplacement (le dossier du projet R).

Les références peuvent être appelées dans le texte, entre parenthèses par le code [@Reference], ou dans le texte, en supprimant les crochets.

La bibliographie est traitée par pandoc lors de la production de documents Word ou HTML. Le style bibliographique peut être précisé, en ajoutant la ligne

```
csl:nom_du_fichier.csl
```

¹³https://pandoc.org/MANUAL.html#extension-implicit_header_references

dans l’entête du document et en copiant le fichier de style *.csl* dans le dossier du projet. Plus d’un millier de styles sont disponibles¹⁴.

Pour les documents PDF, la bibliographie est gérée par LaTeX.

Pour préparer la soumission d’un manuscrit à une revue, il faudra ouvrir le fichier *.tex* intermédiaire produit par pandoc et copier le contenu de l’environnement `{document}` dans le modèle proposé par la revue, qui se chargera du formatage.

Langues

Les langues sont à déclarer dans l’entête des document produits par les modèles de **memoiR**.

La langue principale du document modifie le nom de certains éléments, comme la table des matières. Les langues supplémentaires permettent la rédaction de documents multilingues.

Les champs de l’entête sont :

```
lang: fr-FR
otherlangs: [en-US, it]
```

Le changement de langue dans le document est géré en LaTeX mais pas en HTML en insérant sur une nouvelle ligne la commande suivante :

```
\selectlanguage{english}
```

La langue en cours n’a d’effet que dans les sorties LaTeX : un espace est ajouté devant les ponctuations doubles en Français, la taille des espaces est plus grande en début de phrase en Anglais, etc. La commande `\selectlanguage` est simplement ignorée en HTML.

Les noms de langues sont différents dans l’entête (codes IETF) et dans le texte (nom de la langue). La correspondance et la liste complète des langues se trouve dans le tableau 3 de la documentation du package **polyglossia**¹⁵.

4.3.2 Modèle Simple Article

Le modèle *Simple Article* de **memoiR** produit un document HTML simple avec une table des matières flottante (voir l’exemple¹⁶). D’autres formats HTML sont disponibles : voir la galerie¹⁷ du package. Le format PDF est proche du modèle *article* de LaTeX (exemple¹⁸).

Le modèle contient sa propre documentation.

¹⁴<https://github.com/citation-style-language/styles>

¹⁵<http://mirrors.ctan.org/macros/unicodetex/latex/polyglossia/polyglossia.pdf>

¹⁶<https://EricMarcon.github.io/Krigeage/Krigeage.html>

¹⁷<https://ericmarcon.github.io/memoiR/>

¹⁸<https://EricMarcon.github.io/Krigeage/Krigeage.pdf>

Créer

Utiliser le menu “File > New File > R Markdown...” puis sélectionner “From template” (figure 4.1). La liste des modèles disponible et le package qui les propose est alors affichée.

Sélectionner le modèle *Simple Article* du package **memoiR**, choisir le nom du projet (“Name :”, qui sera le nom du dossier dans lequel il sera créé, et son dossier parent (“Location :”). Dans l’organisation proposée en section 1.2.4, le dossier parent est %LOCALAPPDATA%\ProjetsR. Le nom du projet ne doit contenir aucun caractère spécial (accent, espace...) pour assurer sa portabilité sur tous les systèmes d’exploitation (Windows, Linux, MacOS).

Les modèles élaborés créent un dossier avec de nombreux fichiers (bibliographie, styles, modèle LaTeX...), contrairement aux modèles simples qui créent seulement un fichier.

Quand un dossier est créé, par exemple par le modèle *Simple Article*, il faut en faire un projet RStudio : dans le menu des projets (en haut à droite de la fenêtre de RStudio), utiliser le menu “New Project...” puis “Existing Directory” et sélectionner le dossier qui vient d’être créé.

Ecrire

Les instructions pour utiliser le modèle sont contenues dans le texte fourni par défaut.

Tricoter

Le document peut être tricoté en plusieurs formats :

- *html_document2* est le format HTML pour lequel le modèle a été conçu : un bloc-note avec une table des matières flottante ;
- *gitbook* est un format HTML alternatif, utilisé normalement pour les ouvrages ;
- *downcute* est un format HTML proposé par le package **rmdformats** ;
- *pdf_book* produit un document PDF suivant le modèle LaTeX *article*, couramment utilisé directement en LaTeX ;
- *word_document2* crée un fichier Word.

Mettre en ligne

Le package **memoiR** simplifie la mise en ligne des documents produits.

La fonction `build_gitignore()` crée un fichier `.gitignore` pour le contrôle de source qui doit être activé (voir section 3.1.1).

La fonction `build_readme()` crée un fichier `README.md` nécessaire à GitHub. Il contient le titre du projet, son résumé et des liens vers les versions HTML et PDF des documents produits.

Le projet doit être lié à un dépôt GitHub (section 3.2).

Deux stratégies de publications sont possible. Dans la première, les documents sont tricotés localement et placés dans le dossier `docs`, qui sera le support des pages GitHub. Dans la seconde, les documents sont tricotés par GitHub Actions à chaque fois que des modifications sont poussées sur le dépôt : on parle d'intégration continue (section 6).

La stratégie de production locale est traitée ici ; l'intégration continue le sera dans la section 6.3.1.

La fonction `build_githubpages()` place tous les documents tricotés (HTML et PDF) dans le dossier `docs`, avec une copie du fichier `README.md`. De cette façon, il est possible d'activer les pages GitHub du projet (sur le dossier `docs` de la branche `master`). Le fichier `README.md` sera la page d'accueil du site web produit.

En pratique, on tricote au format HTML pendant toute la phase de rédaction, parce que la production est très rapide. Quand le document est stabilisé, il faut le tricoter au format HTML et au format PDF. Enfin, l'exécution de `build_githubpages()` place tous les fichiers produits dans `docs`. Il reste à pousser le dépôt sur GitHub et activer les pages GitHub.

4.3.3 Autres modèles

Le modèle *Stylish Article* de **memoiR** est destiné à la production d'articles PDF pour l'autoarchivage (typiquement, le dépôt sur HAL) bien formatés, au format A4 en double colonne¹⁹.

Le format HTML est le même que celui du modèle *Simple Article*.

Le package **rticles** a pour ambition de fournir des modèles pour toutes les revues scientifiques qui acceptent une soumission d'articles en LaTeX. Il propose donc des modèles Markdown qui produisent des fichiers PDF conformes aux exigences des revues et la possibilité de récupérer le fichier `.tex` intermédiaire (pandoc produit un fichier `.tex` transmis au compilateur LaTeX). Le package ne permet pas de tricot HTML parce qu'il utilise la syntaxe LaTeX dans le document R Markdown au lieu d'utiliser **bookdown** pour gérer les références bibliographique et les références croisées. Il n'est pas possible d'échanger directement du contenu R Markdown standard avec des documents écrits pour **rticles**, ce qui limite beaucoup l'intérêt du package.

4.4 Présentation Beamer

Le modèle *Beamer Presentation* de **memoiR** permet de créer des présentations au format HTML et PDF (beamer) simultanément, comme le montre l'exemple²⁰.

La démarche est identique à celle des articles du même package. Les niveaux de titre permettent de séparer les parties de la présentation (#) et les diapositives

¹⁹Exemple : <https://EricMarcon.github.io/Rochebrune2018/Entropie.pdf>

²⁰<https://EricMarcon.github.io/Chao1/>, choisir Lecture (HTML) ou Téléchargement (PDF).

(##). Deux formats sont disponibles en HTML : ioslides²¹ et Slidy²². Quelques spécificités dans le code permettent d'affiner la présentation des diapositives, pour un affichage sur deux colonnes par exemple : elles sont documentées dans le modèle.

4.5 memoir

Le modèle *Memoir* du package **memoiR** est destiné aux documents longs, qui présentent une différence importante avec les documents précédents : un document long est composé de plusieurs chapitres, chacun placé dans son fichier .Rmd.

Le format HTML est gitbook²³, le standard de la lecture en ligne de documents de ce type. Le format PDF est dérivé du modèle LaTeX *memoir*²⁴, optimisé aussi pour les documents longs.

Ce document a été écrit avec ce modèle.

4.5.1 Créer

La création d'un projet d'ouvrage est identique à celle présentée plus haut : le modèle est : *Memoir*. Le dossier créé doit être transformé en projet.

Exécuter `build_git()` et `build_readme()`, activer le contrôle de source et pousser le projet sur GitHub, de la même façon que pour un article (section 4.3.2).

Chaque chapitre de l'ouvrage est un fichier Rmd, dont le nom commence normalement par son numéro (ex. : 01-intro.Rmd). Tous les fichiers Rmd présents dans le dossier du projet sont en réalité traités comme des chapitres, triés par ordre de nom de fichier, dont ceux fournis par le modèle (démarrage et syntaxe) qui doivent être supprimés à l'exception de 99-references.Rmd qui contient la bibliographie, placée à la fin. Le fichier index.Rmd est particulier : il contient l'entête du document et le premier chapitre.

4.5.2 Ecrire

Le premier chapitre est placé dans l'avant-propos de l'ouvrage imprimé : il ne doit pas être numéroté (d'où le code {-} à côté du titre) dans la version HTML. Il se termine obligatoirement par la commande LaTeX \mainmatter qui marque le début du corps de l'ouvrage.

Les niveaux de plan commencent par # pour les chapitres (un seul par fichier), ## pour les sections, etc.

²¹<https://bookdown.org/yihui/rmarkdown/ioslides-presentation.html>

²²<https://bookdown.org/yihui/rmarkdown/slidy-presentation.html>

²³<https://www.gitbook.com/>

²⁴<https://www.ctan.org/pkg/memoir>

4.5.3 Tricoter

La compilation au format PDF est faite par XeLaTeX, qui doit être installé.

Pendant la rédaction, il est fortement conseillé de ne créer que le fichier HTML, ce qui est beaucoup plus rapide qu'une compilation LaTeX. Chaque chapitre peut être visualisé très rapidement en cliquant sur le bouton "Knit" au-dessus de la fenêtre de source. Le livre entier est créé en cliquant sur le bouton "Build Book" de la fenêtre *Build* de RStudio. La liste déroulante du bouton permet de créer tous les documents ou de se limiter à un format.

Les fichiers produits sont placés directement dans le dossier `docs`, qui sera utilisé par les pages GitHub pour permettre la lecture en ligne et le téléchargement du PDF. La page d'accueil du site web est créée par bookdown à partir du fichier `index.Rmd` : le fichier `README.md` n'est pas dupliqué dans `docs`.

4.5.4 Finitions

La mise en page est assurée de façon totalement automatique par pandoc (en HTML) et LaTeX (en PDF).

Il est souvent utile d'aider LaTeX à résoudre quelques dépassements de marge dus à de trop grandes contraintes de mise en page : pour la lisibilité optimale, les colonnes sont étroites, mais le code (texte formaté entre deux apostrophes inversées) n'autorise pas la césure.

Si une ligne de texte dépasse dans la marge de droite dans le document PDF, la solution consiste à ajouter manuellement le code `\break` à l'emplacement désiré pour le retour à la ligne dans le document R Markdown. La commande n'a aucun effet sur le document HTML mais force la césure en LaTeX. Pour couper du texte formaté (entre astérisques pour l'italique ou plus fréquemment entre apostrophes inversées pour du code), il faut terminer le formatage avant `\break` et le recommencer après. Exemple, pour forcer le retour à la ligne avant `fichier.Rmd` :

```
Le fichier `/chemin/`\break`fichier.Rmd`
```

En HTML, un espace sera ajouté entre les deux portions de code.

Les bouts de code R sont formatés automatiquement par **knitr** quand l'option `tidy=TRUE` leur est appliquée. Le comportement par défaut est indiqué dans les options de **knitr**, dans un bout de code au début du fichier `index.Rmd` :

```
# knitr options
knitr:::opts_chunk$set(
  cache=TRUE, warning=FALSE, echo = TRUE,
  fig.env='SCfigure', fig.asp=.75,
  fig.align='center', out.width='80%',
  tidy=TRUE,
  tidy.opts=list(blank=FALSE, width.cutoff=55),
  size="scriptsize",
  knitr.graphics.auto_pdf = TRUE)
```

4. RÉDIGER

La largeur maximale d'une ligne de code formaté est ici de 55 caractères, optimal pour le modèle. Il arrive que le formatage automatique ne fonctionne pas parce que **knitr** ne parvient pas à trouver une coupure de ligne respectant toutes les contraintes, ce qui provoque un dépassement de marge dans le code. Dans ce cas, formater manuellement le bout de code en lui ajoutant l'option `tidy=FALSE`.

Les blocs de code littéral, délimités par trois apostrophes inversées, doivent être formatés manuellement, en évitant toute ligne de plus de 55 caractères.

4.5.5 Site gitbook

Le site web contenant le document gitbook doit être paramétré dans `_output.yml` pour que :

- Le titre du document apparaisse en haut de la table des matières ;
- Une indication de l'usage de GitHub et bookdown soit affichée en bas de la table des matières ;
- Un bouton GitHub dans la barre de titre permette d'ouvrir le dépôt du projet ;
- Un autre bouton permette de télécharger le document PDF.

Le fichier `_output.yml` de ce document est le suivant :

```
bookdown::gitbook:
  css: style.css
  config:
    sharing:
      github: yes
      facebook: false
      twitter: false
    toc:
      before: |
        <li><a href=".//">Travailler avec R</a></li>
      after: |
        <li>
          <a href="https://github.com/EricMarcon/travailleR" target="blank">
            Hébergé sur GitHub, publié par bookdown
          </a>
        </li>
    download: pdf
```

La section `sharing:` gère les boutons de la barre de titre. Par défaut, les liens vers Facebook et Twitter sont activés mais celui vers GitHub ne l'est pas. Pour qu'il fonctionne, le dépôt GitHub doit être déclaré dans l'entête du fichier `index.rmd` :

```
github-repo: EricMarcon/travailleR
```

La section `toc:` contient deux portions de code HTML dans lesquelles le titre du document et le lien vers son dépôt GitHub doivent être adaptés au projet.

Enfin, la section `download:` liste les formats de documents téléchargeables et permet d'afficher un bouton de téléchargement dans la barre de titre.

4.5.6 Intégration continue

La construction d'un ouvrage prend du temps, surtout s'il contient des calculs. Elle doit être lancée au format gitbook et au format PDF. En production, elle peut être confiée à GitHub (chapitre 6.3.2).

4.5.7 Google Analytics

Le suivi de l'audience de l'ouvrage peut être confié à Google Analytics. Pour cela, il faut créer un compte et ajouter une *propriété* Google Analytics, c'est-à-dire un site web, puis un flux de données, ici un flux web²⁵.

Google Analytics fournit un script de configuration nommé `gtag.js` à placer à la racine du dossier du projet. Enfin, déclarer le script dans l'entête des pages web en ajoutant une instruction dans `_output.yml`, dans sa première section.

```
bookdown::gitbook:
  includes:
    in_header: gtag.js
```

4.6 Site web R Markdown

Un site web constitué de pages écrites avec R Markdown (sans les fonctionnalités de **bookdown**) et un menu peut être créé très simplement, avec un résultat de bonne facture²⁶.

4.6.1 Modèle

Dans RStudio, dans le menu des projets en haut à droite, cliquer sur “New Project...” puis “New Directory” puis “Simple R Markdown website”. Saisir le nom du projet, sélectionner le dossier dans lequel le projet sera créé en cliquant sur “Browse” et enfin cliquer sur “Create Project”.

Le site par défaut contient deux pages : `index`, la page d'accueil, et `about`, la page “A propos”. Le fichier `_site.yml` contient le nom du site et le contenu de sa barre de navigation : un titre et le fichier correspondant. D'autres pages seront ajoutées en créant de nouveaux fichiers `.Rmd` et en les ajoutant au fichier `_site.yml`.

4.6.2 Améliorations

Le modèle de site peut facilement être amélioré en complétant `_site.yml` :

- en ajoutant une icône GitHub dans la barre de navigation pour renvoyer vers le code source du site :

²⁵https://support.google.com/analytics/answer/9304153?hl=fr&ref_topic=9303319

²⁶<https://rstudio.github.io/learnr/> par exemple.

4. RÉDIGER

- en choisissant la méthode de tricot des pages, pour utiliser **bookdown** au lieu de **rmarkdown** ;
- en plaçant les fichiers du site dans le dossier `docs` et ainsi séparer le code et la production.

Le fichier `_site.yml` complété est le suivant :

```
name: "my-website"
navbar:
  title: "My Website"
  left:
    - text: "Home"
      href: index.html
    - text: "About"
      href: about.html
  right:
    - icon: fa-github
      href: https://github.com/rstudio/rmarkdown
output_dir: "docs"
output:
  bookdown::html_document2:
    theme: sandstone
    highlight: tango
    toc: true
    toc_float: yes
```

L'icône de GitHub fait partie de la collection Font Awesome dont toutes les icônes gratuites²⁷ sont utilisables avec la même syntaxe : “fa-nom”.

Le lien correspondant à l'icône doit être celui du dépôt GitHub du site web.

La syntaxe de la section `output` est la même que celle des documents vus plus haut. Elle s'applique à toutes les pages (dont l'entête YAML est réduite au minimum). Les thèmes disponibles sont ceux de `rmarkdown`²⁸.

L'option `highlight` indique la façon dont le code R éventuellement affiché sera formaté. Enfin, la table des matières est flottante, ce qui signifie que sa position s'ajuste quand la fenêtre défile.

4.6.3 Contôle de source

Le projet doit être placé sous contrôle de source et poussé sur GitHub (chapitre 3). Le fichier `.gitignore` est le suivant :

```
# R
.Rbuildignore
.RData
.Rhistory
.Rprofile
.Rproj.user

# Web Site
/_site/
/*_cache/
/*_files/
```

²⁷<https://fontawesome.com/icons?d=gallery&m=free>

²⁸<https://bookdown.org/yihui/rmarkdown/html-document.html#appearance-and-style>

Activer les pages GitHub (section 3.7) sur le dossier docs pour héberger le site. Ajouter un fichier vide nommé `.nojekyll` dans `docs` pour que les pages GitHub ne tentent pas de reformater le site. On peut utiliser le terminal de RStudio pour exécuter :

```
touch docs/.nojekyll
```

4.7 Site web personnel : blogdown

Pour créer une page web personnelle, *Hugo* est un générateur de site statique capable de produire des pages HTML à partir de code Markdown. Les sites statiques ont l'avantage, en comparaison aux sites dynamiques gérés par un système de gestion de contenu (CMS, par exemple : Wordpress, Joomla, SPIP), d'être portables sur n'importe quel serveur web sans support de base de données ni de code à exécuter côté le serveur (tel que PHP) et d'être très rapides puisque les pages sont créées une seule fois et non à chaque consultation. Un site Hugo peut être hébergé par exemple sur la page personnelle de tout utilisateur de GitHub dont l'adresse est de la forme “GitHubID.github.io”.

Hugo propose de nombreux thèmes, qui sont des modèles de structure de sites, donc le thème **Academic**, destiné aux chercheurs. Dans RStudio, le package **blogdown** est prévu pour produire facilement des pages web avec Hugo. Ces pages peuvent contenir du code R : elles sont très proches d'un article, vu plus haut, dont le contenu peut être facilement copié et collé. Nous utiliserons donc cette solution, pour un site comme celui proposé en exemple²⁹.

La structure du site web est simple :

- une page d'accueil, contenant divers composants paramétrables comme la biographie de l'auteur, une sélection de publications, de billets de blogs ou d'autres éléments et un formulaire de contact;
- des pages détaillant les divers éléments (publications, billets, etc.) écrites en R Markdown.

4.7.1 Installation des outils

La première étape consiste à installer le package **blogdown** dans R.

```
install.packages("blogdown")
```

blogdown est capable d'installer Hugo sous Windows, macOS ou Linux.

```
blogdown::install_hugo()
```

La documentation complète de **blogdown** est disponible³⁰.

²⁹<https://EricMarcon.github.io/>

³⁰<https://bookdown.org/yihui/blogdown/>

Les versions récentes de Hugo utilisent *Go* (le langage de programmation) pour installer leurs modules à la volée : ici le thème Academic est chargé depuis GitHub au moment de la création du site. Go doit donc être installé³¹.

4.7.2 Créer

La façon la plus simple consiste à créer un dépôt sur GitHub à partir du modèle. Sur la page du dépôt *starter-academic*³², cliquer sur le bouton “Use this template”, s’authentifier éventuellement sur GitHub, puis saisir le nom du dépôt qui contiendra le projet, par exemple “MySite”.

Le dépôt peut être celui du site principal de son compte GitHub (voir section 3.7), à l’adresse <https://GitHubID.github.io>³³. Le nom à saisir est simplement “GitHubID.github.io” (*GitHubID* est le nom du compte GitHub).

Créer le dépôt. Copier l’adresse du dépôt en cliquant sur le bouton “Code” puis sur le bouton à droite de l’adresse (figure 4.4).

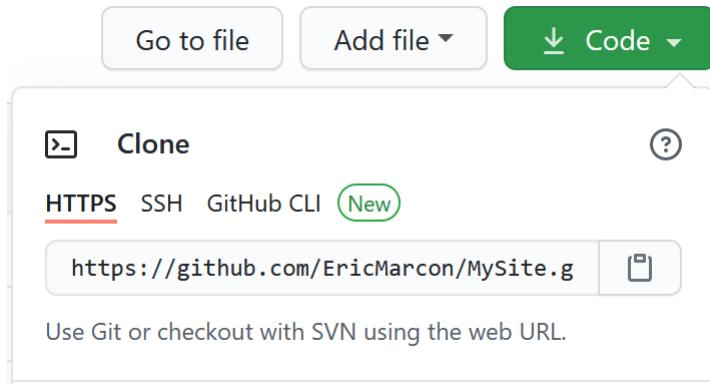


FIG. 4.4 : Copie de l’adresse d’un dépôt à cloner sur GitHub.

Dans RStudio, créer un nouveau projet à partir de GitHub : dans le menu des projets en haut à droite, cliquer sur “New Project...” puis “Version Control” puis “Git” puis coller l’adresse dans le champ “Repository URL” (figure 4.5). Sélectionner le dossier dans lequel le projet sera créé en cliquant sur “Browse” et enfin cliquer sur “Create Project”.

Le projet créé est une copie exacte du modèle, qui doit être personnalisée.

RStudio ajoute automatiquement à la fin du fichier `.gitignore` une ligne pour ignorer ses fichiers de travail (dossier `.Rproj.user`). Ajouter une ligne de commentaire pour le signaler. Le contenu de `.gitignore` doit être le suivant :

```
# R
.Rbuildignore
.RData
```

³¹<https://golang.org/doc/install>

³²<https://github.com/wowchemy/starter-academic>

³³Exemple : <https://EricMarcon.github.io/Krigeage/>

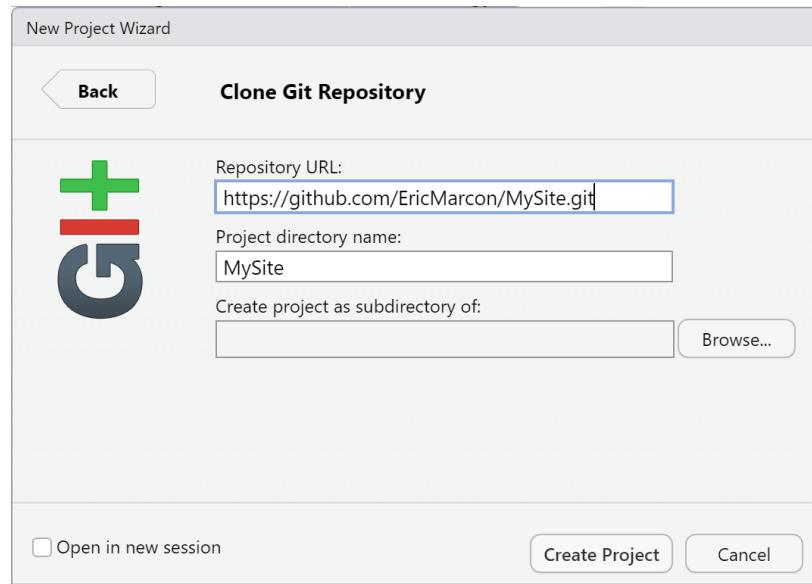


FIG. 4.5 : Collage de l'adresse du dépôt à cloner.

```
.Rhistory
.Rprofile
.Rproj.user
```

```
# Hugo
/resources/
/public/
```

```
# blogdown
/static/en/
/static/fr/
*.rmarkdown
_index.html
index.html
**/index_files/
```

Un bug de **blogdown** nécessite de déplacer le fichier `config.toml` du dossier `config/_default/` à la racine du projet.

Prendre en compte ces modifications dans git en faisant un commit.

4.7.3 Construction du site

Exécuter

```
blogdown::build_site(build_rmd = TRUE)
```

pour construire le site web, y compris ses futures pages R Markdown.

Pour afficher le site, exécuter :

```
blogdown:::serve_site()
```

Il apparaît dans la fenêtre *Viewer* de RStudio, dont le bouton d agrandissement permet l affichage dans le navigateur internet par défaut du système.

Pour modifier le contenu du site, il est préférable d arrêter le serveur web par la commande :

```
blogdown:::stop_server()
```

Le site produit par **blogdown** se trouve dans le dossier `public` qui peut être copié directement sur un serveur web qui l hébergera. Une solution simple consiste à déclarer ce dossier comme racine des pages GitHub du projet (section 3.7). La méthode optimale consiste à utiliser l intégration continue (voir section 6.3.4) pour le copier à la racine de la branche `gh-pages` qui sera déclarée comme emplacement du site sur GitHub.

4.7.4 Site multilingue

Si le site est multilingue (Français et Anglais par exemple), son contenu (dossier `content`) doit être copié dans un dossier correspondant à chaque langue. Par exemple, le fichier `content/authors/admin/_index.md` qui contient les informations sur le propriétaire du site est remplacé par `content/en/authors/admin/_index.md` et `content/fr/authors/admin/_index.md` si le site supporte l Anglais et le Français. En pratique, créer un dossier `en` et un dossier `fr` dans `content`. Copier tout le contenu de `content` (sauf les deux nouveaux dossiers) dans `en` puis déplacer ce même contenu dans `fr`.

4.7.5 Paramétriser

Les fichiers de configuration du site sont bien documentés et offrent de nombreuses options. Les principales sont passées en revue ici pour une création rapide d'un site fonctionnel.

Le fichier `config.toml` contient les paramètres généraux du site. Les lignes à mettre à jour sont celle du titre du site (le nom du propriétaire puisqu'il s'agit d'un site personnel) et son adresse publique. Pour le site exemple :

```
title = "Eric Marcon"
baseurl = "https://EricMarcon.github.io/"
```

Il contient aussi la ligne de sélection de la langue par défaut ("en" ou "fr" au choix) et celle qui permet de placer les fichiers produits par Hugo dans chaque dossier de langue ("true" obligatoirement pour un site multilingue) :

```
defaultContentLanguage = "fr"
defaultContentLanguageInSubdir = true
```

Le dossier `config/_default/` contient les autres fichiers de configuration.

`languages.toml` contient les paramètres linguistiques et les traductions de menus. Pour chaque langue, la version utilisée et le dossier de contenu sont précisés :

```
[en]
languageCode = "en-us"
contentDir = "content/en"
[fr]
languageCode = "fr-fr"
contentDir = "content/fr"
```

Pour les langues additionnelles, le titre du site, les paramètres d'affichage des dates et la traduction des menus sont ajoutés. Dans la section `[fr]` :

```
[fr]
languageCode = "fr-fr"
contentDir = "content/fr"
title = "Eric Marcon"
description = "Page personnelle d'Eric Marcon"
[fr.params]
description = ""
date_format = "02-Jan-2006"
time_format = "15:04"
[[fr.menu.main]]
name = "Accueil"
url = "#about"
weight = 20
(...)
```

Ces lignes sont commentées dans le modèle et doivent donc être décommentées en retirant les `#` en têtes de lignes.

Les menus sont décrits plus bas.

`params.toml` décrit l'aspect du site. Les options sont regroupées par sujet, par exemple “Theme” pour l'apparence générale. Dans “Basic Info”, la ligne

```
site_type = "Person"
```

sélectionne un site personnel. Il est possible d'utiliser Academic pour un site de projet scientifique ou un site d'unité, non documentés en détail ici. Les principales différences sont, pour un site collectif :

- la gestion des auteurs : dans le dossier `/contents/<langue>/authors`, un seul dossier `admin` est utilisé pour un site personnel alors qu'un dossier par personne est nécessaire pour un site collectif;
- un composant décrit plus bas, qui permet de présenter les personnes, doit être activé.

La description du site dans la langue par défaut est saisie, à destination des moteurs de recherche :

```
description = "Eric Marcon's Homepage"
```

4. RÉDIGER

Elle doit être traduite dans le fichier `languages.toml`, dans chaque langue.

Dans “Site Features”, nous sélectionnons la coloration du code R, l’activation du formatage des équations et l’avertissement légal pour l’utilisation des cookies.

```
highlight_languages = ["r"]
math = true
privacy_pack = true
```

La ligne `edit_page` doit être mise à jour : remplacer le dépôt par défaut “<https://github.com/gcushen/hugo-academic>” par celui du site.

“Contact details” contient les informations pour contacter le propriétaire du site. Elles doivent être saisies.

“Regional Settings” contient les paramètres d’affichage de date pour la langue par défaut (ceux des autres langues sont dans `languages.toml`). Ils n’ont normalement pas à être modifiés.

“Comments” permet d’activer les commentaires des visiteurs en bas de pages, avec Disqus ou Comment.io (un compte est nécessaire chez le fournisseur). “Marketing” permet d’activer le suivi de fréquentation du site en saisissant simplement son identifiant Google Analytics (à créer avec un compte Google). “Content Management System” contient la ligne `netlify_cms` dont la valeur doit être `false` si le site n’est pas hébergé par Netlify. Enfin “Icon Pack Extensions” permet d’activer les icônes Academicicons si nécessaire.

4.7.6 Ecrire

Utiliser la documentation en ligne³⁴ en complément des informations principales détaillées ici. L’exemple utilisé ici est le site personnel de l’auteur³⁵.

La méthode de travail consiste à progresser pas à pas en testant puis validant chaque étape :

- Effectuer les modifications ;
- Construire le site et vérifier le résultat : `blogdown:::serve_site()` ;
- Arrêter le site : `blogdown::stop_server()` ;
- Si le résultat n’est pas satisfaisant, recommencer ;
- Valider les modifications (*commit*).

Page d’accueil

La page d’accueil du site est constituée par une suite d’éléments (*widgets*) qui se trouvent dans `/contents/<langue>/home`. Chaque élément est décrit par un fichier markdown. Le premier est `index.md`. Il n’est normalement jamais modifié. Son contenu est le suivant :

³⁴<https://wowchemy.com/docs/page-builder/>

³⁵<https://EricMarcon.github.io/>

```
+++
# Homepage
type = "widget_page"
headless = true # Homepage is headless, other widget
pages are not.
+++
```

Le fichier ne contient qu'un entête au format TOML, encadré par une ligne de `+++`. Le type de composant (`type`) indique qu'il s'agit d'une page de composants, dans laquelle les autres composants du dossier trouveront leur place. `headless = true` signifie que la page n'a pas d'en-tête.

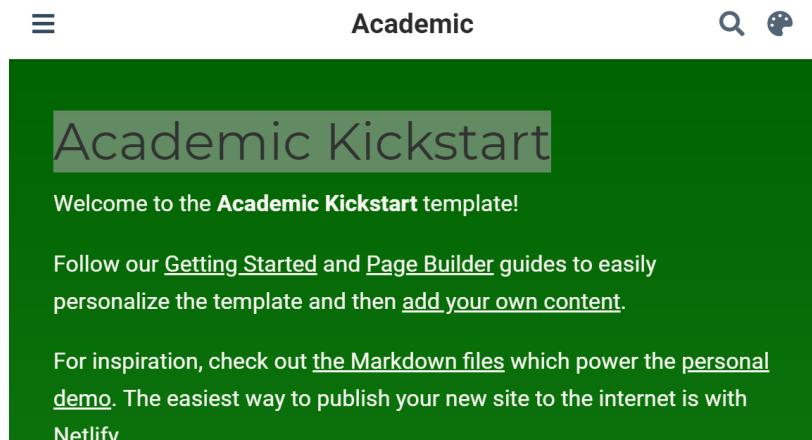


FIG. 4.6 : Composant `demo` dans Academic.

Le composant `demo.md` (figure 4.6) est un composant de type “blank”, c'est-à-dire une page de texte libre : il sert ici à présenter le modèle Academic Kickstart et doit donc être désactivé. L'entête contient ses informations de formatage (titre, nombre de colonnes, couleurs...) et le contenu de la page est écrit en markdown. Les composants apparaissent par ordre croissant de poids (weight dans l'entête) : 15 marque le premier composant dans le modèle Academic. Le composant peut être désactivé en supprimant son fichier ou en modifiant sa propriété `active` dans l'entête :

```
active = false # Activate this widget? true/false
```

Le composant suivant est `about.md` (figure 4.7). Il présente le propriétaire du site. Son titre doit être localisé. Dans le dossier `/content/fr/home`, sa valeur sera :

```
title = "Biographie"
```

L'auteur (`author`) doit correspondre à un dossier de `/contents/<langue>/authors`. `admin` convient parfaitement pour un site personnel. Academic permet de créer des sites d'équipes : dans cette configuration, un dossier par personne serait nécessaire. L'image affichée par le composant est le fichier

4. RÉDIGER

Eric Marcon
Chercheur en écologie
AgroParisTech

g p Q

Biographie

Je suis chercheur en écologie tropicale à l' [UMR Amap](#), chargé de mission à la Direction de la Recherche et de la Valorisation d' [AgroParisTech](#) et coordinateur du parcours [BioGET](#) du master Biodiversité, Ecologie et Evolution (AgroParisTech et Université de Montpellier).

Intérêts
• Ecologie des communautés
• Foresterie tropicale
• Ecologie Statistique
• Développement avec R

Formation
• Habilitation à Diriger des Recherches en écologie, 2016 Université de Guyane
• Doctorat en écologie, 2010 AgroParisTech
• Ingénieur du Génie Rural, des Eaux et des Forêts, 1999 Ecole National du Génie Rural, des Eaux et des Forêts
• DEA en économie internationale, 1999

FIG. 4.7 : Composant `about` dans Academic.

avatar.jpg placé dans ce dossier. Limiter la taille du fichier pour la performance du site (moins d'un mégaoctet est une taille raisonnable), tout en assurant une taille minimale de quelques centaines de pixels de côté pour la qualité de l'affichage.

Le contenu du composant est lu dans le fichier `_index.md` du même dossier, qui contient toutes les informations sur l'auteur. Son organisation est assez claire : modifier son contenu à partir de l'exemple fourni. Si des icônes de type ai sont utilisées, activer le pack d'icône AcademicIcons dans `config/_default/params.toml`.



FIG. 4.8 : Composant `skills` dans Academic.

Le composant talents (`skills`, figure 4.8) présente les compétences de l'auteur de façon graphique. Une collection d'icônes est disponible, et des icônes nouvelles peuvent être ajoutées.

Le composant expérience (`experience`, figure 4.9) liste les expériences professionnelles. Toutes les informations sont saisies dans son entête.

Le composant accomplishments présente les formations professionnelles et permet d'accéder à leurs certificats.

Le composant posts va chercher son contenu dans le dossier `/contents/<langue>/post` qui contient les billets de blog (voir plus bas). Le fichier

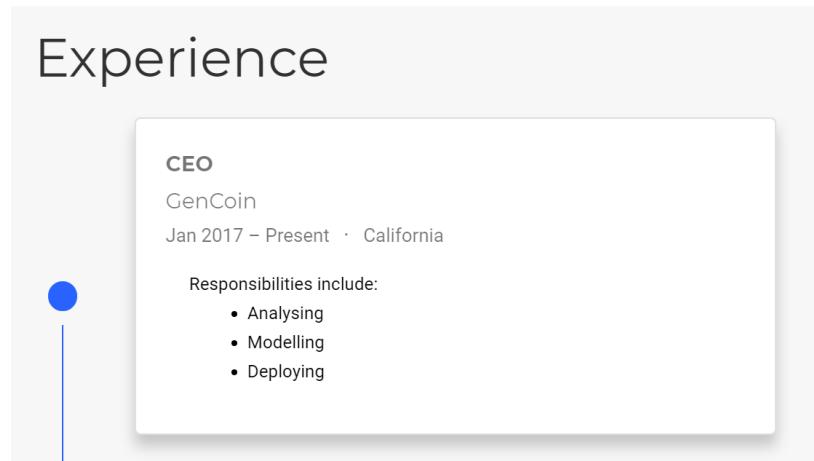


FIG. 4.9 : Composant experience dans Academic.

`posts.md` contient des options de mise en page dans son entête.

Le composant `projects` fonctionne de la même façon. La différence entre les deux composants est leur mise en forme : `posts` est du type “pages”, qui affiche les éléments les plus récents, alors que `projects` est de type “portfolio”, qui affiche les éléments sélectionnés qui contiennent la description `featured: true` dans leur propre entête. Il est possible de créer des composants de ces types librement, en spécifiant le dossier contenant les éléments dans “page-type”. Exemple : créer un composant nommé `software.md` en renommant `projects.md`, modifier sa ligne `page_type = "software"` et créer un dossier `/contents/<langue>/software` pour y placer du contenu.

Les composants `publications` et `featured` sont de type “pages” et “portfolio” respectivement et prennent leur contenu dans le dossier `publication`.

Le composant `tags` présente un nuage de mots à partir des mots-clés déclarés dans tous les fichiers de contenu (billets de blog, publications...) sous la forme suivante dans leur entête :

```
tags = ["Mot Clé 1", "Autre Mot Clé"]
```

Enfin, le composant `contact` permet d'afficher un formulaire de contact. Il utilise les informations du fichier `config/_default/params.toml` dans sa partie commençant par :

```
#####
## Contact details
##
```

Pour afficher une carte, entrer la latitude et la longitude de l'adresse dans la ligne `coordinates`. Pour afficher un formulaire de messagerie, choisir le service `formspre.io` (`email_form = 2` dans `contact.md`). Pour activer le service de messagerie, il faudra construire le site web, s'envoyer un premier message en utilisant le formulaire et suivre les instructions de Formspre.

4. RÉDIGER

Le composant `people` est utilisé dans les sites collectifs pour présenter les membres. Le composant `slider` permet d'afficher un carrousel (des éléments défilants) en haut de page. Pour comprendre son fonctionnement, le plus simple consiste à l'activer.

Menu de la page d'accueil

La page d'accueil comporte un menu qui permet de naviguer rapidement vers ses composants ou vers d'autres pages. Il est paramétré dans `config/_default/menus.toml`. Les éléments du menu ont un nom affiché, un lien (commençant par `#` pour pointer vers un composant ou un chemin relatif dans le site comme `publication/`), et un poids qui définit leur ordre d'affichage, de la même façon que celui des composants de la page d'accueil.

Un menu à deux éléments pour pointer vers l'accueil du site et les billets de blogs est donc le suivant :

```
[[main]]
name = "Home"
url = "#about"
weight = 10

[[main]]
name = "Posts"
url = "#posts"
weight = 20
```

Le menu doit être traduit dans chaque langue dans le fichier `config/_default/languages.toml` :

```
[fr]
[[fr.menu.main]]
name = "Accueil"
url = "#about"
weight = 10
[[fr.menu.main]]
name = "Articles"
url = "#posts"
weight = 20
```

Billets

Le site est alimenté par des billets de blog placés dans le dossier `/contents/<langue>/post`. Il doivent être traduits et placés dans le dossier `post` de chaque langue pour être disponibles dans la langue correspondante. L'exemple utilisé ici est un guide pour estimer correctement la densité d'une variable bornée³⁶.

Son code est sur GitHub³⁷.

Un billet est placé dans un dossier (`/content/fr/post/densite`) qui contient son code R Markdown et éventuellement des images, des données pour

³⁶<https://EricMarcon.github.io/post/densite/>

³⁷<https://github.com/EricMarcon/HomePage2020/tree/master/content/fr/post/densite>

alimenter le code et d'autres éléments appelés par le code. Hugo supporte des fichiers markdown natifs. L'apport de **blogdown** relativement à un site Hugo natif est le support de R Markdown, donc la possibilité d'exécuter tout code R comme dans un bloc-note (dont le contenu peut être réutilisé sans modification).

Le fichier principal d'un billet est `index.Rmd`. **blogdown** crée un fichier `index.html` pendant la construction du site : il peut être ignoré (dans `.gitignore`) et supprimé à tout moment. Si une image `featured.png` (optimale pour un schéma) ou `featured.jpg` (optimale pour un photo) est placée dans le dossier, elle sera utilisée comme vignette du billet.

`index.Rmd` comprend un entête au format yaml (entourée par des `---`) ou `toml` (entourée par des `+++`) qui décrit son affichage :

```
---
title: "Titre du billet"
subtitle: "Sous-titre"
summary: "Résumé"
authors: []
tags: ["Mot Clé 1", "Autre Mot Clé"]
categories: []
date: 2020-04-17
featured: false
draft: false

# Featured image
# To use, add an image named `featured.jpg/png` to
# your page's folder.
# Focal points: Smart, Center, TopLeft, Top, TopRight,
# Left, Right, BottomLeft, Bottom, BottomRight.
image:
  caption: ""
  focal_point: ""
  preview_only: false

bibliography: references.bib
---
```

Les auteurs sont utilisés dans les sites collectifs. Les tags permettent d'alimenter le composant nuage de mots s'il est activé dans la page d'accueil. Les catégories permettent de rechercher des pages au contenu similaire (recherche par mot-clé sur le site). L'option `featured: true` fait apparaître le billet dans les composants de type `featured` sur la page d'accueil. L'option `draft: true` cache le billet.

Les éléments suivants précisent l'affichage de la vignette : légende et position. L'option `preview_only: true` limite l'affichage aux miniatures (sur la page d'accueil), retirant donc l'image du billet lui-même.

Les éléments d'entête nécessaires au corps de texte R Markdown, comme le nom du fichier contenant les références bibliographiques, placé dans le même dossier, sont ajoutés.

Le corps du texte est celui d'un document R Markdown standard, avec du code R inclus. Un bout de code initial permet de fixer les options de R et charger les packages nécessaires.

4. RÉDIGER

En pratique, la façon la plus efficace de créer un nouveau billet est de copier le dossier complet d'un billet précédent, de le renommer et de modifier son contenu. La commande `blogdown::new_post()` peut aussi être utilisée mais ne gère pas les langues multiples (et crée donc le billet dans le dossier `/contents/post` à moins de préciser l'argument `subdir`).

La reconstruction du site ne met par défaut pas à jour les pages basées sur un fichier `.Rmd`. Pour le faire, il faut forcer la commande `build_site()`.

```
blogdown::build_site(build_rmd = TRUE)
blogdown::serve_site()
```

Publications

Les publications sont organisées comme les billets, mais placées dans le dossier `/contents/<langue>/publications`.

L'exemple utilisé est un article de revue³⁸ avec son code³⁹.

Un fichier `cite.bib` contenant la référence au format BibTex est placé dans le dossier. Le nom du dossier est de préférence celui de l'identifiant de la publication. L'entête du fichier `index.md` (ici au format Markdown, mais `.Rmd` est possible si du code R est nécessaire) contient les mêmes informations que le fichier BibTex, mais au format approprié (yaml), et les éléments propres à Academic (`featured`) :

```
---
title: "Evaluating the geographic concentration of |>
industries using distance-based methods"
authors: ["Eric Marcon", "Florence Puech"]
publication_types: ["2"]
abstract: "We propose (...)"
publication: "*Journal of Economic Geography*"
doi: "10.1093/jeg/lbg016"

date: 2003-10-01
featured: false
---
```

Les types de publication sont :

- 0 = Uncategorized;
- 1 = Conference paper;
- 2 = Journal article;
- 3 = Preprint / Working Paper;
- 4 = Report;
- 5 = Book;
- 6 = Book section;
- 7 = Thesis;

³⁸<https://EricMarcon.github.io/publication/marcon-2003-a/>

³⁹<https://github.com/EricMarcon/HomePage2020/tree/master/content/fr/publication/marcon-2003-a>

- 8 = Patent.

Des boutons sont affichés en haut de la page de la publication en fonction des informations trouvées :

- PDF : si la ligne `url` est présente dans l'en-tête ;
- Citation : si le fichier `cite.bib` est présent dans le dossier ;
- DOI : si la ligne `doi` est présente dans l'en-tête.

Le corps de la publication contient un lien (au format HTML) vers le site Dimension qui fournit des informations bibliométriques. Ce lien peut être réutilisé très simplement, en remplaçant simplement le DOI du document :

```
<span class="__dimensions_badge_embed__"
      data-doi="10.1093/jeg/lbg016"></span>
<script async src="https://badge.dimensions.ai/
  badge.js" charset="utf-8"></script>
```

Enfin, un fichier `/contents/<langue>/publications/_index.Rmd` permet de présenter la bibliographie complète. Il est accessible à partir du composant `publications` de la page d'accueil qui affiche un lien “Plus de Publications”.

Le fichier exemple⁴⁰ avec son code⁴¹ permet d'interroger Google Scholar pour obtenir le réseau de coauteurs, l'indice h et le nombre de citations annuelles de l'auteur. Il est réutilisable en modifiant simplement l'identifiant Google Scholar à la ligne 30.

En faisant exécuter le code régulièrement, par exemple par GitHub (voir ci-dessous), les statistiques affichées sont maintenues à jour sans intervention humaine.

Communications

Les communications sont organisées comme les publications, dans le dossier `/contents/<langue>/talk`.

L'exemple utilisé est une communication en Français, donc dans `/contents/fr/talk`⁴² avec son code⁴³.

Une image peut être utilisée plus facilement que pour une publication.

L'en-tête contient des lignes particulières adaptées aux communications :

```
---
title: "Construction de l'estimateur de biodiversité |>
Chao1"
```

⁴⁰<https://EricMarcon.github.io/publication/>

⁴¹<https://github.com/EricMarcon/HomePage2020/tree/master/content/fr/publication/marcon-2003-a>

⁴²<https://EricMarcon.github.io/talk/chao1/>

⁴³<https://github.com/EricMarcon/HomePage2020/tree/master/content/fr/talk/chao1>

4. RÉDIGER

```
event: "Semaine des mathématiques 2020"
event_url: https://eduscol.education.fr/cid59178/|>
semaine-des-mathematiques.html

location: Université de Guyane

summary: []
abstract: |
Pour estimer le nombre d'espèces (richesse
spécifique) d'une communauté à partir d'un
échantillon, l'estimateur Chao1 est l'outil
le plus utilisé.

Sa construction est expliquée et son efficacité
est testée sur des données simulées.

# Talk start and end times.
# End time can optionally be hidden by
# prefixing the line with `#`.
date: "2020-03-11T11:00:00Z"
date_end: "2020-03-11T12:00:00Z"
all_day: false

# Schedule page publish date (NOT talk date).
publishDate: "2020-04-14"

# Is this a featured talk? (true/false)
featured: false

image:
caption: 'Produit scalaire des vecteurs $v_0$ |>
et $v_2$'
focal_point: Smart

url_code: "https://github.com/EricMarcon/Chao1"
url_pdf: "https://EricMarcon.github.io/Chao1/|>
Chao1.pdf"
url_slides: "https://EricMarcon.github.io/Chao1/|>
Chao1.html"

# Enable math on this page?
math: true
---
```

Les liens (url_code par exemple) font apparaître des boutons qui permettent d'afficher respectivement le code source de la présentation, un fichier pdf et les diapositives en ligne.

Autres éléments

Il est possible d'ajouter librement des éléments supplémentaires sur le site :

- dans /contents/<langue>/, créer un dossier dont le nom est le type d'éléments (exemple : recette);
- ajouter des éléments dans ce dossier, chacun dans son propre dossier;
- le fichier obligatoire est index.md ou index.Rmd avec un en-tête contenant possiblement tous les champs rencontrés dans les éléments post, publication et talk;
- le fichier de vignette, featured.png ou featured.jpg, est facultatif;

- tous les fichiers nécessaires au tricot (images, données) peuvent être ajoutés dans le même dossier ;
- dans `/contents/<langue>/home`, ajouter un composant de la page d'accueil en copiant-collant un élément existant de type “pages” (comme `publications`) ou “portfolio” (comme `featured`) et le paramétrier pour qu'il pointe sur le bon dossier (dans l'exemple : `page-type=recette`) et ajuster son apparence (nombre d'éléments par exemple) et sa position (poids) ;
- ajouter éventuellement une entrée de menu pour pointer sur le composant, avec le même poids que le composant.

Les fichiers d'index peuvent porter l'extension `.Rmd` ou `.md`. Dans le premier cas, ils seront traités par **blogdown**, qui supporte l'intégration de code R. Dans l'autre cas, ils seront traités par Hugo, qui ne gère que le format markdown standard. Les fichiers `.md` nécessitent moins de ressource et sont donc préférés quand ils suffisent.

Finitions

L'icône du site, qui apparaît dans la barre d'adresse des navigateurs web, se trouve dans `assets/images`. Le fichier `icon.png` peut être remplacé.

4.7.7 Intégration continue

La construction du site web en production peut être confiée à GitHub (section 6.3.4), y compris sa mise à jour périodique si des pages du site traitent des données qui évoluent dans le temps.

4.7.8 Mises à jour

Le thème Academic est régulièrement mis à jour. La version utilisée est indiquée dans le fichier `go.mod`. Pour utiliser la dernière version officielle, exécuter dans la console R la commande suivante :

```
blogdown::hugo_cmd("mod get -u")
```

Les fichiers `go.mod` et `go.sum`, qui contiennent les codes de hachage des fichiers du module, sont mis à jour.

Chaque changement de version peut nécessiter des adaptations du contenu du site, référencées dans la documentation en ligne du thème⁴⁴.

Mettre Hugo à jour en même temps :

```
blogdown::update_hugo()
```

⁴⁴<https://wowchemy.com/updates/>

4.8 Exportation de figures

Quand la production de documents avec R Markdown n'est pas possible, les figures issues de R doivent être exportées sous forme de fichiers pour être intégrés dans un autre processus d'écriture. Il est préférable de créer des scripts pour créer les figures de façon reproductible et au format optimal.

4.8.1 Formats vectoriels et raster

Les figures doivent en général être produites dans un format vectoriel :

- SVG pour la publication d'affiches ou de posters ;
- EMF (Extended Meta-File) pour Word ou la suite Microsoft Office qui ne supporte pas d'autres formats ;
- EPS (Encapsulated PostScript) ou PDF (Portable Document Format) pour LaTeX.

Les figures raster (composées d'un ensemble de points, comme les photographies) sont rares dans R. La fonction `image()` utilisée pour afficher des cartes utilise par défaut des polygones plutôt que des points. La figure 4.10 montre le résultat du code suivant :

```
x <- 10 * (1:nrow(volcano))
y <- 10 * (1:ncol(volcano))
image(x, y, volcano, col = hcl.colors(100, "terrain"), axes = FALSE)
contour(x, y, volcano, levels = seq(90, 200, by = 5), add = TRUE,
        col = "brown")
axis(1, at = seq(100, 800, by = 100))
axis(2, at = seq(100, 600, by = 100))
box()
```

Elle est composée d'un ensemble de rectangles colorés : il s'agit bien d'une image vectorielle.

Si nécessaire, des images peuvent être produites aux formats BMP (bitmap, sans compression), JPEG (compressées avec perte de qualité), PNG (compressées sans perte de qualité, avec transparence possible) ou Tiff (compressées ou non).

4.8.2 Fonctions

La fonction `postscript()` produit un fichier EPS. Le code R doit appeler la fonction pour créer le fichier, produire la figure, puis fermer le fichier, par exemple :

```
# Ouverture du fichier
postscript("Fig1.eps", width = 6, height = 4, horizontal = FALSE)
# Création de la figure
plot(cars)
# Fermeture du fichier
dev.off()
```

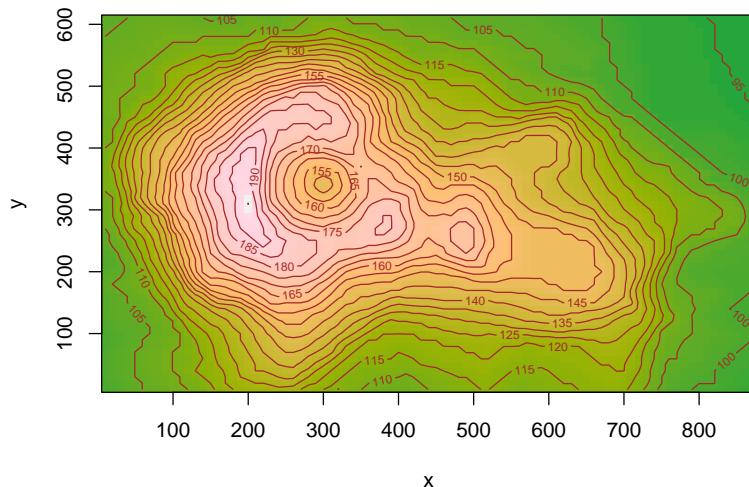


FIG. 4.10 : Courbes de niveau du volcan Maunga Whau, code fourni en exemple de l'aide de la fonction `image()`.

```
## pdf  
## 2
```

La largeur et la hauteur (en pouces) d'un fichier vectoriel n'ont pas d'importance, mais leur rapport fixe l'aspect de la figure. La taille des textes est fixe : augmenter la taille de la figure revient donc à diminuer la taille relative des textes : procéder par essais successifs, en veillant à ce que les légendes restent lisibles à la taille finale de la figure.

L'argument horizontal fixe l'orientation de la figure de façon assez imprévisible : procéder par essais.

Les fonctions `eps()`, `pdf()`, `bmp()`, `jpeg()`, `png()` et `tiff()` fonctionnent de la même manière. Se référer à l'aide des fonctions pour le choix des options (résolution, niveau de compression, etc.). La fonction `emf()` est fournie par le package **devEMF**.

Les polices de caractères ne sont pas incluses dans les fichiers EPS ou PDF. Si nécessaire, la fonction `embedFonts()` permet d'y remédier, à condition que GhostScript soit installé.

4.8.3 Package ragg

Le package **ragg**⁴⁵ améliore la qualité des fichiers PNG, JPEG et TIFF. Les fonctions optimisées sont `agg_png()`, `agg_jpeg()` et `agg_tiff()`. Leur usage est le même que celui des fonctions de **grDevices**.

⁴⁵<https://ragg.r-lib.org/>

4. RÉDIGER

Les documents R Markdown produisent des images au format PNG pour leur version HTML. **ragg** améliore leur qualité : le package doit être installé et dev = "ragg_png" doit être ajoutée aux options de **knitr**. Pour ce document, les options déclarées dans index.Rmd sont les suivantes :

```
knitr::opts_chunk$set(  
  cache=FALSE, # Cache chunk results  
  echo = TRUE, # Show/Hide R chunks  
  warning=FALSE, # Show/Hide warnings  
  # Figure alignment and size  
  fig.align='center', out.width='80%', fig.asp=.75,  
  # Graphic devices (ragg_png is better than standard png)  
  dev = c("ragg_png", "pdf"),  
  # Code chunk format  
  tidy=TRUE, tidy.opts=list(blank=FALSE, width.cutoff=60),  
  size="scriptsize", knitr.graphics.auto_pdf = TRUE  
)  
options(width=60)
```

Enfin, **ragg** peut être utilisé comme moteur de rendu graphique par défaut dans RStudio à partir de la version 1.4 (Menu “Tools > Global Options > General > Graphics > Backend”).

4.9 Flux de travail

Un flux de travail (voir section 2.8) peut être intégré dans un document R Markdown à partir de la version 0.5 du package **targets**.

```
library("targets")
```

4.9.1 Déclaration du flux

Le flux est géré par des bouts de code de type **targets**. Leur entête minimal est {targets} au lieu de {r}, et ils doivent être nommés. Ces bouts de code permettent de créer le fichier _targets.R quand ils sont exécutés en mode non interactif, notamment pendant que le document est tricoté. S’ils sont lancés en mode interactif, par exemple dans R Studio, leur code est exécuté. L’option tar_interactive = FALSE dans leur entête permet de les tester sans tricoter tout le document.

Un ancien flux éventuel doit être supprimé avant d’écrire le nouveau :

```
tar_unscript()
```

Le premier bout de code, avec l’option tar_globals=TRUE, écrit les options globales du flux. Pour créer le flux présenté en section 2.8, le code est simplement :

```
```{targets targets_global, tar_globals=TRUE}  
Packages
tar_option_set(packages = c("spatstat", "dbmss"))
```
```

Les fonctions utilisées par les cibles sont déclarées dans ce type de bout de code : elles sont ajoutées à un fichier dans le dossier de travail `_targets_r` (différent du dossier `_targets` qui contient les fichiers de calcul des cibles).

4.9.2 Déclaration des cibles

Les cibles elles-mêmes sont déclarées dans des bouts de code dont le nom est celui de la variable de destination.

```
```{targets X, tar_simple=TRUE}
runifpoint(NbPoints)
```

```

Chaque cible nécessite un bout de code construit de cette manière. La valeur de la cible est la dernière valeur renournée, à la manière d'une fonction qui n'utilisera pas `return()`.

Pendant le tricot, ce code simplifié (`tar_simple=TRUE`) est transformé automatiquement en écriture de cible :

```
tar_target(X, {
  runifpoint(NbPoints)
})

## Define target X from chunk code.
## Establish _targets.R and _targets_r/targets/X.R.
```

La lecture du document est alourdie par cette syntaxe particulière : `targets` n'est pas utile pour des documents dont le code, rapide à exécuter, doit être affiché dans le texte. En revanche, si le code est long à exécuter et n'est pas affiché, son intérêt est considérable pour limiter le temps de calcul.

Les autres bouts de code nécessaires pour compléter le flux sont les suivants :

- `NbPoints` :

```
tar_target(NbPoints, {
  1000
})

## Define target NbPoints from chunk code.
## Establish _targets.R and _targets_r/targets/NbPoints.R.
```

- `d` :

```
tar_target(d, {
  sum(pairdist(X))/NbPoints/(NbPoints-1)
})

## Define target d from chunk code.
## Establish _targets.R and _targets_r/targets/d.R.
```

4. RÉDIGER

- map :

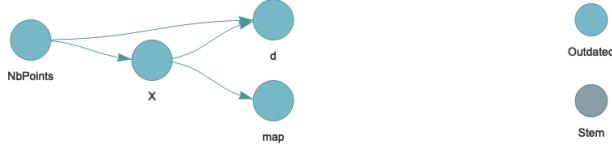
```
tar_target(map, {  
  autoplot(as.wmppp(X))  
})
```

```
## Define target map from chunk code.  
## Establish _targets.R and _targets_r/targets/map.R.
```

4.9.3 Exécution du flux

Pour lancer le calcul des cibles, un bout de code standard (`{r}`) doit appeler `tar_make()` :

```
tar_visnetwork()
```



```
tar_make()
```

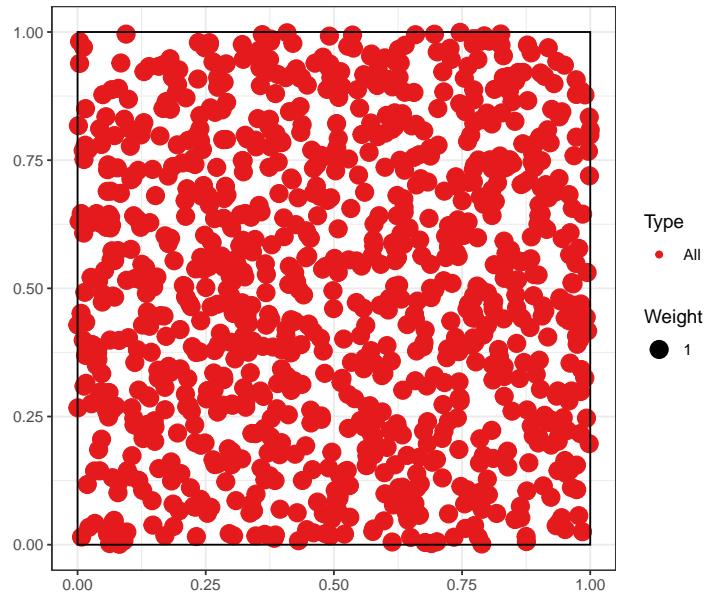
```
## • start target NbPoints  
## • built target NbPoints  
## • start target X  
## • built target X  
## • start target d  
## • built target d  
## • start target map  
## • built target map  
## • end pipeline: 3.423 seconds
```

`tar_visnetwork()` permet de vérifier que le flux est correct avant de l'exécuter. Au moment de la production finale du document, l'option `include=FALSE` peut être ajoutée à l'entête de ce bout de code pour qu'il ne produise aucun affichage.

4.9.4 Utilisation des résultats

Les bouts de code qui utilisent les valeurs des cibles doivent les lire avec `tar_read()` :

```
tar_read(map)
```



4.9.5 Contrôle de source

Les fichiers de **targets** doivent être inclus au contrôle de source. De cette façon, les calculs effectués localement ne seront pas répétés par GitHub Actions (chapitre 6) et la construction du document sera rapide.

PACKAGE

Les packages de R permettent d'étendre les fonctionnalités du logiciel par du code fourni par la communauté des développeurs. Ils sont la clé du succès de R parce qu'ils permettent de diffuser rapidement de nouvelles méthodes issues de la recherche ou d'ajouter de nouveaux outils qui peuvent devenir des standards, comme le **tidyverse**.

Il est utile de produire un package quand on a écrit des nouvelles fonctions qui forment un ensemble cohérent. Un package à usage personnel ou limité à une équipe de travail est simple à mettre en place et le temps gagné en utilisant facilement la version à jour de chaque fonction amortit très rapidement le temps consacré à la fabrication du package. Ce type de package a vocation à être hébergé sur GitHub.

Des packages à usage plus large, qui fournissent par exemple le code correspondant à une méthode publiée, sont placés dans le dépôt CRAN, d'où ils pourront être installés par la commande standard `install.packages()`. CRAN effectue des vérifications poussées du code et n'accepte que les packages passant sans aucun avertissement sa batterie de tests. Ils doivent respecter la politique¹ du dépôt.

La documentation pour la création de packages est abondante. L'ouvrage de référence est celui de WICKHAM (2015), à consulter en tant que référence.

L'approche utilisée ici consiste à créer un premier package très rapidement pour comprendre que la démarche est assez simple. Il sera ensuite enrichi des éléments nécessaires à un package diffusé à d'autres utilisateurs que son concepteur : une documentation complète et des tests de bon fonctionnement notamment.

¹<https://cran.r-project.org/web/packages/policies.html>

5.1 Premier package

Cette introduction reprend les recommandations du blog *Créer un package en quelques minutes*² de ThinkR.

5.1.1 Crédit

Les packages ont une organisation stricte dans une structure de fichiers et de répertoires figée. Il est possible de créer cette structure manuellement mais des packages spécialisés peuvent s'en charger :

- **usethis** automatise la création des dossiers ;
- **roxygen2** permet d'automatiser la documentation obligatoire des packages ;
- **devtools** est la boîte à outils du développeur, permettant notamment de construire et tester les packages ;

Les trois sont à installer en premier lieu :

```
install.packages(c("usethis", "roxygen2", "devtools"))
```

Le package à créer sera un projet RStudio. Dans le menu des projets, sélectionner “New Project > New Directory > R package using devtools...”, choisir le nom du projet et son dossier parent. Le package s'appellera **multiple**, dans le dossier %LOCALAPPDATA%\ProjetsR en suivant les recommandations de la section 1.2.4.

Le nom du package doit respecter les contraintes des noms de projets : pas de caractères spéciaux, pas d'espaces... Il doit aussi être évocateur de l'objet du package. Si le package doit être diffusé, toute sa documentation sera rédigée en Anglais, y compris son nom.

La structure minimale est créée :

- un fichier DESCRIPTION qui indique que le dossier contient un package et précise au minimum son nom ;
- un fichier NAMESPACE qui déclare comment le package intervient dans la gestion des noms des objets de R (son contenu sera mis à jour par roxygen2) ;
- un dossier R qui contient le code des fonctions offertes par le package (vide à ce stade).

Le package peut être testé tout de suite : dans la fenêtre *Build* de RStudio, cliquer sur “Install and Restart” construit le package et le charge dans R, après avoir redémarré le programme pour éviter tout conflit.

Dans la fenêtre *Packages*, **multiple** est maintenant visible. Il est chargé, mais ne contient rien.

²<https://thinkr.fr/creer-package-r-quelques-minutes/>

5.1.2 Première fonction

Fichiers

Les fonctions sont placées dans un ou plusieurs fichier .R dans le dossier R. L'organisation de ces fichiers est libre. Pour cet exemple, un fichier du nom de chaque fonction sera créé. Des fichiers regroupant les fonctions similaires ou un seul fichier contenant tout le code sont des choix possibles.

Le choix fait ici est le suivant :

- un fichier qui contiendra le code commun à tout le package : package.R ;
- un fichier commun à toutes les fonctions : fonctions.R.

Création

La première fonction, double(), est créée et enregistrée dans le fichier fonctions.R :

```
double <- function(number) {
  return(2 * number)
}
```

A ce stade, la fonction est interne au package et n'est pas accessible depuis l'environnement de travail. Pour s'en persuader, construire le package (*Install and Restart*) et vérifier le bon fonctionnement de la fonction :

```
double(2)
```

Le résultat est un vecteur composé de deux 0 parce que la fonction appelée est un homonyme du package **base** (voir sa documentation en tapant ?double) :

```
base::double(2)
```

```
## [1] 0 0
```

Pour que la fonction de notre package soit visible, elle doit être *exportée* en la déclarant dans le fichier NAMESPACE. C'est le travail de **roxygen2** qui gère en même temps la documentation de chaque fonction. Pour l'activer, placer le curseur dans la fonction et appeler le menu “Code > Insert Roxygen Skeleton”. Des commentaires sont ajoutés avant la fonction :

```
#' Title
#'
#' @param number
#'
#' @return
#' @export
#'
#' @examples
double <- function(number) {
  return(2 * number)
}
```

5. PACKAGE

Les commentaires à destination de **roxygen2** commencent par `#'` :

- la première ligne contient le titre de la fonction, c'est-à-dire un descriptif très court : son nom en général ;
- la ligne suivante (séparée par un saut de ligne) peut contenir sa description (rubrique *Description* dans l'aide) ;
- la suivante (après un autre saut de ligne) peut contenir plus d'informations (rubrique *Details* dans l'aide) ;
- les arguments de la fonction sont décrits par les lignes `@param` ;
- `@return` décrit le résultat de la fonction ;
- `@export` déclare que la fonction est exportée : elle sera donc utilisable dans l'environnement de travail ;
- des exemples peuvent être ajoutés.

La documentation doit être complétée :

```
'#' double
'#'
#' Double value of numbers.
'#'
#' Calculate the double values of numbers.
'#'
#' @param number a numeric vector.
'#'
#' @return A vector of the same length as `number` containing the
#'         transformed values.
#' @export
'#'
#' @examples
#' double(2)
#' double(1:4)
double <- function(number) {
  return(2 * number)
}
```

Ne pas hésiter à s'inspirer de l'aide de fonctions existantes pour respecter les standards de R (ici : `?log`) :

- penser que les fonctions sont normalement vectorielles : `number` est par défaut un vecteur, pas un scalaire ;
- certains éléments commencent par une majuscule et se terminent par un point parce que ce sont des paragraphes dans le fichier d'aide ;
- le titre n'a pas de point final ;
- la description des paramètres ne commence pas par une majuscule.

La prise en compte des changements dans la documentation nécessitent d'appeler la fonction `roxygenize()`. Dans la fenêtre *Build*, le menu “More > Document” permet de le faire. Ensuite, construire le package (*Install and Restart*) et vérifier le résultat en exécutant la fonction et en affichant son aide :

```
double(2)
`?` (double)
```

Il est possible d'automatiser la mise à jour de la documentation à chaque construction du package par le menu “Build > Configure Build Tools...” : cliquer sur “Configure” et cocher la case “Automatically reoxygenize when running Install and Restart”. C'est un choix efficace pour un petit package mais pénalisant quand le temps de mise à jour de la documentation s'allonge avec la complexité du package. La reconstruction du package est le plus souvent utilisée pour tester des modifications du code : sa rapidité est essentielle.

La documentation pour **roxygen2** supporte le format Markdown³.

A ce stade, le package est fonctionnel : il contient une fonction et un début de documentation. Il est temps de lancer une vérification de son code : dans la fenêtre *Build*, cliquer sur “Check” ou utiliser la commande `devtools::check()`. L'opération *réoxygène* le package (met à jour sa documentation), effectue un grand nombre de tests et renvoie la liste des erreurs, avertissements et notes détectées. L'objectif est toujours de n'avoir aucune alerte : elles doivent être traitées immédiatement. Par exemple, le retour suivant est un avertissement sur la non-conformité de la licence déclarée :

```
> checking DESCRIPTION meta-information ... WARNING
  Non-standard license specification:
    `use_gpl3_license()`
  Standardizable: FALSE

0 errors v | 1 warning x | 0 notes v
Erreur : R CMD check found WARNINGS
```

Pour la corriger, mettre à jour, exécuter la commande de mise à jour de la licence, en commençant par votre nom :

```
options(usethis.full_name = "Eric Marcon")
usethis::use_gpl3_license()
```

La liste des licences valides est fournie par R⁴.

Après correction, relancer les tests jusqu'à la disparition des alertes.

5.1.3 Contrôle de source

Il est temps de placer le code sous contrôle de source.

Activer le contrôle de source dans les options du projet (figure 3.2). Redémarrer RStudio à la demande.

Créer un dépôt sur GitHub et y pousser le dépôt local, comme expliqué dans le chapitre 3.

Créer le fichier `README.md` :

³<https://roxygen2.r-lib.org/articles/markdown.html>

⁴<https://svn.r-project.org/R/trunk/share/licenses/license.db>

5. PACKAGE

```
# multiple  
An R package to compute mutiple of numbers.
```

Le développement du package est ponctué par de nombreux commits à chaque modification et une publication (push) à chaque étape, validée par une incrémentation du numéro de version.

5.1.4 package.R

Le fichier package.R est destiné à recevoir le code R et surtout les commentaires pour **roxygen2** qui concernent l'ensemble du package.

Le premier bloc de commentaire produira l'aide du package (?multiple).

```
#' multiple-package  
#'  
#' Multiples of numbers  
#'  
#' This package allows simple computation of multiples  
#' of numbers, including fast algorithms for integers.  
#'  
#' @name multiple  
#' @docType package  
NULL
```

Son organisation est identique à celle des documentations de fonctions, avec deux déclarations particulières pour le nom du package et le type de documentation. Le code NULL après les commentaires indique à **roxygen2** qu'il n'y a pas de code R lié.

La documentation est mise à jour par la commande roxygen2::roxygenise(). Après reconstruction du package, vérifier que l'aide est apparue : ?multiple.

5.2 Organisation du package

5.2.1 Fichier DESCRIPTION

Le fichier doit être complété :

```
Package: multiple  
Title: Calculate multiples of numbers  
Version: 0.0.0.9000  
Authors@R:  
  person(given = "Eric",  
         family = "Marcon",  
         role = c("aut", "cre"),  
         email = "e.marcon@free.fr",  
         comment = c(ORCID = "0000-0002-5249-321X"))  
Description: This package allows simple computation  
            of multiples of numbers, including fast algorithms  
            for integers.  
License: GPL-3  
Encoding: UTF-8  
LazyData: true  
Roxygen: list(markdown = TRUE)  
RoxygenNote: 7.1.1
```

Le nom du package est figé et ne doit pas être modifié.

Son titre doit décrire en une ligne à quoi il sert. Le titre est affiché dans la fenêtre *Packages* à côté des noms des packages.

La version doit respecter les conventions :

- Le premier nombre est la version majeure, 0 tant que le package n'est pas stable puis 1. La version majeure ne change que si le package n'est plus compatible avec ses versions précédentes, ce qui oblige les utilisateurs à modifier leur code.
- Le deuxième est la version mineure, incrémentée quand des fonctionnalités nouvelles sont ajoutées.
- Le troisième est la version de correction : 0 à l'origine, incrémentée à chaque correction de code sans nouvelle fonctionnalité.
- Le quatrième est réservé au développement, et commence à 9000. Il est incrémenté à chaque version instable et disparaît quand une nouvelle version stable (*release*) est produite.

Exemple : une correction de bug sur la version 1.3.0 produit la version 1.3.1. Les versions de développement suivantes (instables, non destinées à l'usage en production) sont 1.3.1.9000 puis 1.3.1.9001, etc. Le numéro de version doit être mis à jour à chaque fois que le package est poussé sur GitHub. Quand le développement est stabilisé, la nouvelle version, destinée à être utilisée en production, est 1.3.2 si elle n'apporte pas de nouvelle fonctionnalité ou 1.4.0 dans le cas contraire.

La description des auteurs est assez lourde mais simple à comprendre. Les identifiants Orcid des auteurs académiques peuvent être utilisés. Si le package a plusieurs auteurs, ils sont placés dans une fonction `c()` : `c(person(...), person())` pour deux auteurs. Dans ce cas, il faut préciser le rôle de chacun :

- “cre” pour le créateur du package
- “aut” pour un auteur parmi les autres
- “ctb” pour un contributeur, qui peut avoir signalé un bug ou fourni un peu de code.

La description du package en un paragraphe permet de donner plus d'informations.

La licence précise la façon dont le package peut être utilisé et modifié. GPL-3 est une bonne valeur par défaut, mais d'autres choix sont possibles⁵.

L'option `LazyData` signifie que les données d'exemples fournies avec le package peuvent être utilisées sans les appeler au préalable par la fonction `data()` : c'est le standard actuel.

Enfin, les deux dernières lignes sont gérées par **roxygen2**.

⁵<https://r-pkgs.org/description.html#description-license>

5.2.2 Fichier NEWS.md

Le fichier NEWS .md contient l'historique du package. Les nouvelles versions sont ajoutées en haut du fichier.

Créer une première version du fichier :

```
# multiple 0.0.0.9000  
## New features  
* Initial version of the package
```

Les titres de premier niveau doivent contenir le nom du package et sa version. Les titres de niveau 2 sont libres, mais contiennent en général des rubriques comme “New features” et “Bug Fixes”.

Pour ne pas multiplier les versions décrites, il est conseillé de modifier la version en cours et de compléter la documentation jusqu’au changement de version de correction (troisième nombre). Ensuite, l’entrée correspondant à cette version reste figée et une nouvelle entrée est ajoutée.

5.3 Vignette

Une vignette est indispensable pour documenter correctement le package :

```
usethis::use_vignette("multiple")
```

Le fichier `multiple.Rmd` est créé dans le dossier `vignettes`. Ajouter un sous-titre dans son entête : la description courte du package :

```
title: "multiple"  
subtitle: "Multiples of numbers"
```

Le reste de l’entête permet à R de construire la vignette à partir de code R Markdown.

Le corps de la vignette contient par défaut du code R pour déclarer les options de présentation des bouts de code et le chargement du package. Une introduction à l’utilisation du package doit être écrite dans ce document, en R Markdown.

Pendant le développement du package, la vignette peut être construite manuellement en exécutant :

```
devtools::build_vignettes("multiple")
```

Les fichiers produits sont placés dans `doc/` : ouvrir le fichier `.html` pour contrôler le résultat.

RStudio ne crée pas la vignette du package quand la commande “Install and Restart” de la fenêtre Build est appelée. Pour une installation complète, deux solutions sont possibles :

- Construire le fichier source du package (“Build > More > Build Source Package”) puis l’installer (“Packages > Install > Install from > Package Archive file”). Le fichier source se trouve à côté de celui du projet.
- Pousser le code du package sur GitHub puis exécuter :

```
remotes::install_github("multiple", build_vignettes = TRUE)
```

La vignette peut ensuite être affichée par la commande :

```
vignette("multiple")
```

5.4 pkgdown

Le package **pkgdown** permet de créer un site d’accompagnement du package⁶, qui reprend le fichier README.md comme page d’accueil, la vignette dans une rubrique “Get Started”, l’ensemble des fichiers d’aide avec leurs exemples exécutés (section “Reference”), le fichier NEWS.md pour un historique du package (section “Changelog”) et des informations du fichier DESCRIPTION.

Créer le site avec **usethis**

```
usethis::use_pkgdown()
```

Construire ensuite le site. Cette commande sera exécutée à nouveau à chaque changement de version du package :

```
pkgdown::build_site()
```

Le site est placé dans le dossier docs. Ouvrir le fichier index.htm avec un navigateur web pour le visualiser. Dès que le projet sera poussé sur GitHub, activer les pages du dépôt pour que le site soit visible en ligne (voir section 3.7).

pkgdown place le site dans le dossier docs.

Ajouter l’adresse des pages GitHub dans une nouvelle ligne du fichier DESCRIPTION :

URL: <https://GitHubID.github.io/multiple>

L’ajouter aussi dans le fichier _pkgdown.yml qui a été créé vide, ainsi que l’option suivante :

```
url: https://GitHubID.github.io/multiple
```

```
development:  
  mode: auto
```

⁶Exemple : <https://EricMarcon.github.io/entropart/>

pkgdown place le site dans le dossier `docs/dev` si le site d'une version stable (à trois nombres) du package existe dans `docs` et que la version en cours est une version de développement (à quatre nombres). De cette façon, les utilisateurs d'une version de production du package ont accès au site sans qu'il soit perturbé par les versions de développement.

Le site peut être enrichi de plusieurs façons :

- En ajoutant des articles au format R Markdown dans le dossier `vignettes/articles`. La vignette ne peut pas mobiliser d'importantes ressources de calcul pour présenter des exemples parce qu'elle est construite en même temps que le package. Les articles sont générés par **pkgdown**, indépendamment, et peuvent donc être plus ambitieux ;
- En améliorant sa présentation (regroupement des fonctions par thèmes, ajout de badges, d'un sticker⁷...) : se référer à la vignette de **pkgdown**.

Pour enrichir la documentation du package, il est possible d'utiliser un fichier `README.Rmd` au format R Markdown, à tricoter pour créer le `README.md` standard de GitHub, utilisé comme page d'accueil du site **pkgdown**, qui peut de cette façon présenter des exemples d'utilisation du code. La démarche est détaillée dans *R Packages*⁸. La complexité ajoutée est à comparer au gain obtenu : une page d'accueil simple (sans code) avec des liens vers la vignette et les articles est plus simple à mettre en œuvre.

5.5 Code spécifique aux packages

5.5.1 Importation de fonctions

Créons une nouvelle fonction dans `fonctions.R` qui ajoute un bruit aléatoire à la valeur double :

```
fuzzydouble <- function(number, sd = 1) {  
  return(2 * number + rnorm(length(number), 0, sd))  
}
```

Le bruit est tiré dans une loi normale centrée d'écart-type `sd` et ajouté à la valeur calculée.

`rnorm()` est une fonction du package **stats**. Même si le package est systématiquement chargé par R, le package d'appartenance de la fonction doit obligatoirement être déclaré : les seules exceptions sont les fonctions du package **base**.

Le package **stats** doit d'abord être déclaré dans `DESCRIPTION` qui contient une instruction `Imports`:. Tous les packages utilisés par le code de **multiple** seront listés, séparés par des virgules.

⁷L'application Shiny **hexmake** permet de créer facilement un sticker : <https://connect.thinkr.fr/hexmake/>

⁸<https://r-pkgs.org/release.html?q=readme#readme-rmd>

```
Imports: stats
```

Cette “importation” signifie simplement que le package **stats** doit être chargé, mais pas nécessairement attaché (voir section 2.2), pour que **multiple** fonctionne.

Ensuite, la fonction `rnorm()` doit être trouvable dans l'environnement du package **multiple**. Il y a plusieurs façons de remplir cette obligation. D'abord, le commentaire suivant pourrait être fourni pour **roxygen2** :

```
#' @import stats
```

Tout l'espace de nom du package **stats** serait attaché et accessible au package **multiple**. Ce n'est pas une bonne pratique parce qu'elle multiplie les risques de conflits de noms (voir section 2.2). Notons que la notion d'importation utilisée ici est différente de celle de DESCRIPTION, bien qu'elles aient le même nom.

Il est préférable d'importer uniquement la fonction `rnorm()` en la déclarant dans la documentation de la fonction :

```
#' @importFrom stats rnorm
```

Ce n'est pas une pratique idéale non plus parce que l'origine de la fonction n'apparaîtrait pas clairement dans le code du package.

La bonne pratique est de ne rien importer (au sens de **roxygen2**) et de qualifier systématiquement les fonctions d'autres packages avec la syntaxe `package::fonction()`. C'est la solution retenue ici parce que la directive `@importFrom` importerait la fonction dans tout le package **multiple**, pas seulement dans la fonction `fuzzydouble()`, au risque de créer des effets de bord (modifier le comportement d'une autre fonction du package qui n'assumerait pas l'importation de `rnorm()`). Finalement, le code de la fonction est le suivant :

```
#' fuzzydouble
#'
#' Double value of numbers with an error
#'
#' Calculate the double values of numbers
#' and add a random error to the result.
#'
#' @param number a numeric vector.
#' @param sd the standard deviation of the Gaussian error added.
#'
#' @return A vector of the same length as `number`
#' containing the transformed values.
#' @export
#'
#' @examples
#' fuzzydouble(2)
#' fuzzydouble(1:4)
fuzzydouble <- function(number, sd = 1) {
  return(2 * number + stats::rnorm(length(number), 0, sd))
}
```

5.5.2 Méthodes S3

Les méthodes S3 sont présentées en section [2.1.2](#).

Classes

Les objets appartiennent à des classes :

```
# Classe d'un nombre
class(2)

## [1] "numeric"

# Classe d'une fonction
class(sum)

## [1] "function"
```

En plus des classes de base, les développeurs peuvent en créer d'autres.

Méthodes

L'intérêt de créer de nouvelles classes est de leur adapter des méthodes existantes, le cas le plus courant étant `plot()`. Il s'agit d'une méthode générique, c'est-à-dire un modèle de fonction, sans code, à décliner selon la classe d'objet à traiter.

```
plot

## function (x, y, ...)
## UseMethod("plot")
## <bytecode: 0x7fc28a75ac30>
## <environment: namespace:base>
```

Il existe dans R de nombreuses déclinaisons de `plot` qui sont des fonctions dont le nom est de la forme `plot.class()`. **stats** fournit une fonction `plot.lm()` pour créer une figure à partir d'un modèle linéaire. De nombreux packages créent des classes adaptées à leurs objets et proposent une méthode `plot` pour chaque classe. Les fonctions peuvent être listées :

```
# Quelques fonctions plot()
head(methods(plot))

## [1] "plotANY-method"    "plotcolor-method"
## [3] "plotAccumCurve"   "plot.acf"
## [5] "plotACF"           "plot.addvar"

# Nombre total
length(methods(plot))

## [1] 150
```

Inversement, les méthodes disponibles pour une classe peuvent être affichées :

```
methods(class = "lm")

## [1] add1      alias      anova
## [4] as_flextable case.names coerce
## [7] confint    cooks.distance deviance
## [10] dfbeta    dfbetas   drop1
## [13] dummy.coef effects   extractAIC
## [16] family    formula   fortify
## [19] hatvalues influence initialize
## [22] kappa     labels   logLik
## [25] model.frame model.matrix nobs
## [28] plot      predict   print
## [31] proj      qqnorm   qr
## [34] residuals response rstandard
## [37] rstudent   show     simulate
## [40] slotsFromS3 summary  variable.names
## [43] vcov
## see '?methods' for accessing help and source code
```

La méthode `print` est utilisée pour afficher tout objet (elle est implicite quand on saisit seulement le nom d'un objet) :

```
my_lm <- lm(dist ~ speed, data = cars)
# Equivalent de '> my_lm'
print(my_lm)
```

```
##
## Call:
## lm(formula = dist ~ speed, data = cars)
##
## Coefficients:
## (Intercept)      speed
## -17.579        3.932
```

La méthode `summary` affiche un résumé lisible de l'objet :

```
summary(my_lm)

##
## Call:
## lm(formula = dist ~ speed, data = cars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -29.069 -9.525 -2.272  9.215 43.201
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -17.5791    6.7584  -2.601  0.0123 *
## speed       3.9324    0.4155   9.464 1.49e-12 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.38 on 48 degrees of freedom
## Multiple R-squared:  0.6511, Adjusted R-squared:  0.6438
## F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

5. PACKAGE

Les autres méthodes ont été créées spécifiquement pour les besoins du package **stats**.

Attribution d'un objet à une classe

Pour qu'un objet appartient à une classe, il suffit de le déclarer :

```
x <- 1
class(x) <- "MyClass"
class(x)
```

```
## [1] "MyClass"
```

Une façon plus élégante de le faire est d'ajouter la nouvelle classe à l'ensemble des classes auquel l'objet appartient déjà :

```
y <- 1
class(y) <- c("MyClass", class(y))
class(y)
```

```
## [1] "MyClass" "numeric"
```

Il n'y a aucune vérification de cohérence entre la structure réelle de l'objet et une structure de la classe qui serait déclarée ailleurs : le développeur doit s'assurer que les méthodes trouveront bien les bonnes données dans les objets qui déclarent lui appartenir. Dans le cas contraire, des erreurs se produisent :

```
class(y) <- "lm"
tryCatch(print(y), error = function(e) print(e))
```

```
## <simpleError: $ operator is invalid for atomic vectors>
```

5.5.3 En pratique

Création d'une méthode générique

De nouvelles méthodes génériques peuvent être créées et déclinées selon les classes.

A titre d'exemple, créons une méthode générique `triple` qui calculera le triple des valeurs dans le package **multiple**, déclinée en deux fonctions distinctes : une pour les entiers et une pour les réels. Les calculs sur les nombres entiers plus rapides que ceux sur les réels, ce qui justifie l'effort d'écrire deux versions du code.

```
# Méthode générique
triple <- function(x, ...) {
  UseMethod("triple")
}
```

La méthode générique ne contient pas de code au-delà de sa déclaration. Sa signature (c'est-à-dire l'ensemble de ses arguments) est importante parce que les fonctions dérivées de cette méthode devront obligatoirement avoir les mêmes arguments dans le même ordre et pourront seulement ajouter des arguments supplémentaires avant . . . (qui est obligatoire). Comme la nature du premier argument dépendra de la classe de chaque objet, l'usage est de l'appeler `x`.

La méthode est déclinée en deux fonctions :

```
triple.integer<- function (x, ...){
  return(x * 3L)
}
triple.numeric<- function (x, ...){
  return(x * 3.0)
}
```

Dans sa version entière, `x` est multiplié par `3L`, le suffixe `L` signifiant que `3` doit être compris comme un entier. Dans sa version réelle, `3` peut être noté `3.0` pour montrer clairement qu'il s'agit d'un réel. Sous R, `3` sans autre précision est compris comme un réel.

Le choix de la fonction dépend de la classe de l'objet passé en argument.

```
# Argument entier
class(2L)

## [1] "integer"

# Résultat entier par la fonction triple.integer
class(triple(2L))

## [1] "integer"

# Argument réel
class(2)

## [1] "numeric"

# Résultat réel par la fonction triple.numeric
class(triple(2))

## [1] "numeric"

# Performance
microbenchmark::microbenchmark(triple.integer(2L), triple.numeric(2),
  triple(2L))

## Unit: nanoseconds
##                                expr  min    lq     mean   median    uq
##  triple.integer(2L)  333  346 23799.26  357.5  380.0
##  triple.numeric(2)  336  346 21341.21  366.0  393.5
##      triple(2L) 1353 1375 1506.29 1384.5 1435.0
##                                max  neval
##  2342942    100
##  2095442    100
##      6063    100
```

La mesure des performances par le package **microbenchmark** ne montre pas de différence entre les fonctions `triple.integer()` et `triple.numeric` comme attendu parce que le temps consacré au calcul lui-même est négligeable en comparaison du temps d'appel de la fonction. La méthode générique consomme beaucoup plus de temps que les calculs très simples ici. R teste en effet l'existence de fonctions correspondant à la classe de l'objet passé en argument aux méthodes génériques. Comme un objet peut appartenir à plusieurs classes, il recherche une fonction adaptée à la première classe, puis aux classes suivantes successivement. Cette recherche prend beaucoup de temps et justifie de réserver l'usage de méthodes génériques à la lisibilité du code plutôt qu'à une recherche de performance : l'intérêt des méthodes génériques est de fournir à l'utilisateur du code une seule fonction pour un objectif donné (plot pour réaliser une figure) quelles que soient les données à traiter.

Création d'une classe

Dans un package, on créera des classes si les résultats des fonctions le justifient : structure de liste et identification de la classe à un objet (“lm” est la classe des modèles linéaires). Pour toute classe créée, les méthodes `print`, `summary` et `plot` (si une représentation graphique est possible) doivent être écrites.

Ecrivons une fonction `multiple()` dont le résultat sera un objet d'une nouvelle classe, “multiple”, qui sera une liste mémorisant les valeurs à multiplier, le multiplicateur et le résultat.

```
multiple <- function(number, times = 1) {  
  # Calculate the multiples  
  y <- number * times  
  # Save in a list  
  result <- list(x = number, y = y, times = times)  
  # Set the class  
  class(result) <- c("multiple", class(result))  
  return(result)  
}  
# Classe du résultat  
my_multiple <- multiple(1:3, 2)  
class(my_multiple)  
  
## [1] "multiple" "list"
```

L'appel à la fonction `multiple()` renvoie un objet de classe “multiple”, qui est aussi de classe “list”. En absence de fonction `print.multiple()`, R cherche la fonction `print.list()` qui n'existe pas et se rabat sur la fonction `print.default()` :

```
my_multiple  
  
## $x  
## [1] 1 2 3  
##  
## $y  
## [1] 2 4 6
```

```
##  
## $times  
## [1] 2  
##  
## attr(,"class")  
## [1] "multiple" "list"
```

La fonction `print.multiple` doit donc être écrite pour un affichage lisible, limité au résultat :

```
print.multiple <- function(x, ...) {  
  print.default(x$y)  
}  
# Nouvel affichage  
my_multiple
```

```
## [1] 2 4 6
```

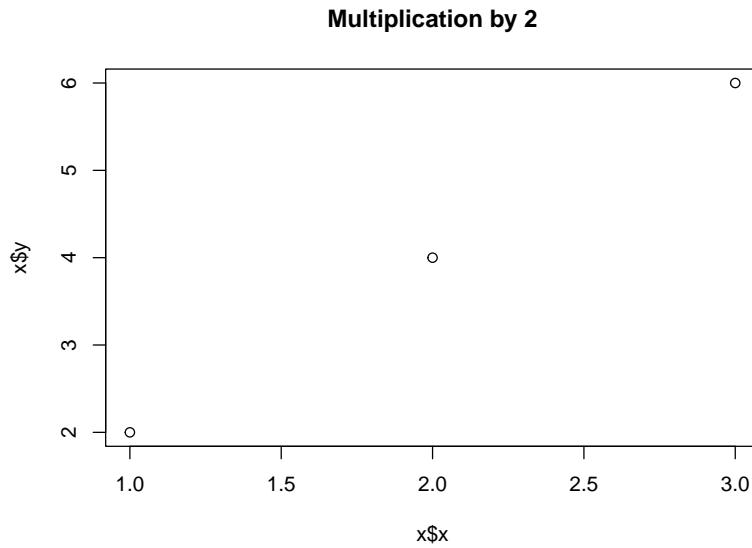
Les détails peuvent être présentés dans la fonction `summary` :

```
summary.multiple <- function(object, ...) {  
  print.default(object$x)  
  cat("multiplied by", object$times, "is:\n")  
  print.default(object$y)  
}  
# Nouvel affichage  
summary(my_multiple)
```

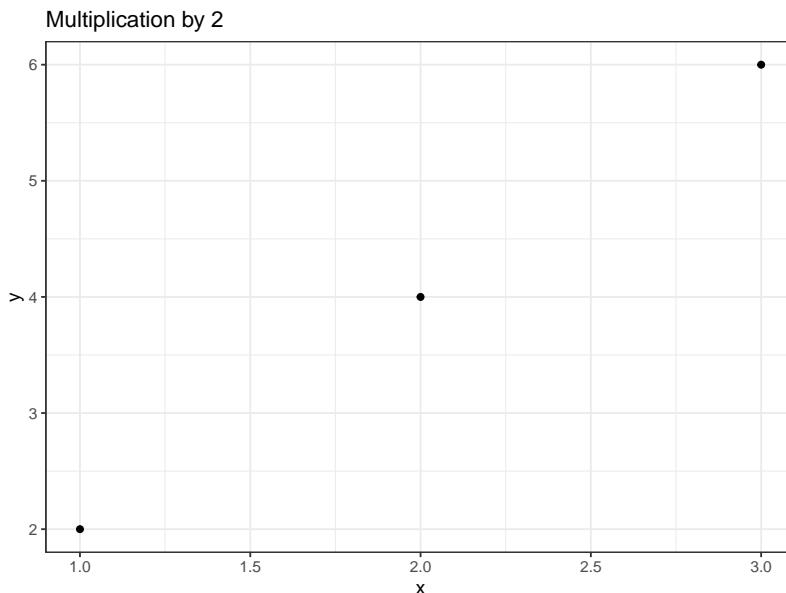
```
## [1] 1 2 3  
## multiplied by 2 is:  
## [1] 2 4 6
```

Enfin, une fonction `plot` et une fonction `autoplot` complètent l'ensemble :

```
plot.multiple <- function(x, y, ...) {  
  plot.default(y=x$y, x=x$x, type = "p",  
    main = paste("Multiplication by", x$times), ...)  
}  
  
autoplot.multiple <- function(object, ...) {  
  data.frame(x = object$x, y = object$y) %>%  
    ggplot2::ggplot() +  
    ggplot2::geom_point(ggplot2::aes(x = .data$x, y = .data$y)) +  
    ggplot2::labs(title = paste("Multiplication by",  
      object$times))  
}  
  
plot(my_multiple)
```



```
autoplot(my_multiple)
```



Pour des raisons techniques liées à l'évaluation non conventionnelle dans le tidyverse, les noms de variables utilisées par `aes()` doivent être préfixées par `.data$` dans les packages et `rlang::data` doit être importé. Dans le cas contraire, la vérification du package renvoie une note indiquant que les variables `x` et `y`, utilisées par les arguments de `aes()` n'ont pas été déclarées et n'existent peut-être pas dans l'environnement local (voir section 2.2).

Documentation

Les méthodes génériques et les fonctions qui les déclinent doivent être documentées comme n'importe quelle autre fonction.

La gestion de l'espace des noms est un peu plus complexe :

- Les méthodes génériques doivent être exportées :

```
#' @export
```

- Les fonctions dérivées de méthodes génériques ne doivent pas être exportées mais être déclarées comme méthodes, avec le nom de la méthode générique et la classe traitée. **roxygen2** demande qu'une directive d'exportation soit ajoutée mais ne l'applique pas (comme il se doit) dans le fichier **NAMESPACE** qui est utilisé par R :

```
#' @method plot multiple
#' @export
```

- Depuis la version 3 de **roxygen2**, la déclaration `@method` est inutile tant que le nom de la fonction est décomposable sans ambiguïté, comme `plot.multiple` : `@export` suffit. Si le nom de la fonction dérivée comporte plusieurs points, **roxygen2** peut ne pas détecter automatiquement le générique et l'objet et `@method` doit être maintenu.
- Les fonctions dérivées de méthodes génériques venant d'un autre package nécessitent d'importer la méthode générique, sauf si elle est fournie par **base** (`print` est fourni par **base** et n'est donc pas concerné) :

```
#' @importFrom graphics plot
#' @importFrom ggplot2 autoplot
```

- Les génériques importés de cette manière doivent être réexportés par une directive à placer par exemple juste après le code de la fonction dérivée :

```
#' @export
graphics::plot
#' @export
ggplot2::autoplot
```

- **roxygen2** crée automatiquement un fichier d'aide `reexports.Rd` dans lequel se trouve un lien vers la documentation originale des génériques ré-exportés.

Dans `DESCRIPTION`, le package d'origine de chaque générique doit être listé dans la directive `Imports` :

```
Imports: ggplot2, graphics
```

Enfin, l'importation de fonctions du tidyverse nécessite aussi quelques précautions :

5. PACKAGE

- le package **tidyverse** est réservé à l’usage interactif de R : il n’est pas question de l’importer dans DESCRIPTION parce que ses dépendances peuvent changer et aboutir à des résultats imprévisibles. Le package **magrittr** fournit les tuyaux, principalement `%>%`. Le package **rlang** fournit l’objet `.data` présenté ci-dessous. Il doivent être importés dans DESCRIPTION.

```
Imports: magrittr, rlang, stats
```

- Comme il n’est pas possible de préfixer les `%>%` par le nom du package, il faut importer la fonction en utilisant les délimiteurs prévus pour les fonctions dont le nom contient des caractères spéciaux :

```
#' @importFrom magrittr `)%>%`
```

- Les fonctions du tidyverse qui utilisent des noms de colonnes de tibbles ou dataframes génèrent des avertissements au moment de la vérification du package parce que ces noms sont confondus avec des noms de variables non définies. Pour éviter cette confusions, l’objet `.data` du package **rlang** est utilisé (par exemple dans `aes()` vu plus haut). Il doit être importé :

```
#' @importFrom rlang .data
```

Finalement, le code complet est le suivant :

```
## Multiplication of a numeric vector
#'
#' @param number a numeric vector
#' @param times a number to multiply
#'
#' @return an object of class `multiple`
#' @export
#'
#' @examples
#' multiple(1:2, 3)
multiple <- function(number, times = 1) {
  # Calculate the multiples
  y <- number * times
  # Save in a list
  result <- list(x = number, y = y, times = times)
  # Set the class
  class(result) <- c("multiple", class(result))
  return(result)
}

## Print objects of class multiple
#'
#' @param x an object of class `multiple`.
#' @param ... further arguments passed to the generic method.
#'
#' @export
#'
#' @examples
#' print(multiple(2,3))
print.multiple <- function(x, ...) {
  print.default(x$y)
}
```

```

#' Summarize objects of class multiple
#'
#' @param object an object of class `multiple`.
#' @param ... further arguments passed to the generic method.
#'
#' @export
#'
#' @examples
#' summary(multiple(2,3))
summary.multiple <- function(object, ...) {
  print.default(object$x)
  cat("multiplied by", object$times, "is:\n")
  print.default(object$y)
}

#' Plot objects of class multiple
#'
#' @param x a vector of numbers
#' @param y a vector of multiplied numbers
#' @param ... further arguments passed to the generic method.
#'
#' @importFrom graphics plot
#' @export
#'
#' @examples
#' plot(multiple(2,3))
plot.multiple <- function(x, y, ...) {
  plot.default(y=x$y, x=x$x, type = "p",
               main = paste("Multiplication by", x$times), ...)
}

#' @export
graphics::plot

## function (x, y, ...)
## UseMethod("plot")
## <bytecode: 0x7fc28a75ac30>
## <environment: namespace:base>

#' autoplot
#'
#' ggplot of the `multiple` objects.
#'
#' @param object an object of class `multiple`.
#' @param ... ignored.
#'
#' @return a `ggplot` object
#' @importFrom ggplot2 autoplot
#' @importFrom magrittr `%>%
#' @importFrom rlang .data
#' @export
#'
#' @examples
#' autoplot(multiple(2,3))
autoplot.multiple <- function(object, ...) {
  data.frame(x = object$x, y = object$y) %>%
    ggplot2::ggplot() +
    ggplot2::geom_point(ggplot2::aes(x = .data$x, y = .data$y)) +
    ggplot2::labs(title = paste("Multiplication by",
                                object$times)))
}

#' @export
ggplot2::autoplot

## function (object, ...)

```

```
## {
##     UseMethod("autoplot")
## }
## <bytecode: 0x7fc289b9cd90>
## <environment: namespace:ggplot2>
```

5.5.4 Code C++

L'utilisation de code C++ a été vue en section 2.5. Pour intégrer ces fonctions dans un packages, il faut respecter les règles suivantes :

- les fichiers .cpp contenant le code sont placés dans le dossier /src du projet ;
- le code est commenté pour **roxygen2** de la même façon que les fonctions R, mais avec le marqueur de commentaire du langage C :

```
#include <Rcpp.h>
using namespace Rcpp;

/**' timesTwo
/**'
/**' Calculates the double of a value.
/**'
/**' @param x A numeric vector.
/**' @export
// [[Rcpp::export]]
NumericVector timesTwo(NumericVector x) {
  return x * 2;
}
```

- dans DESCRIPTION, importer les packages. **Rcpp**, et **RcppParallel** si du code parallélisé est utilisé (supprimer ses références sinon), doivent être déclarés dans LinkingTo :

```
Imports: Rcpp, RcppParallel
LinkingTo: Rcpp, RcppParallel
```

- les commentaires pour **roxygen2** doivent être ajoutés à package.R (“multiple” est le nom du package) :

```
' @importFrom Rcpp sourceCpp
#' @importFrom RcppParallel RcppParallelLibs
#' @useDynLib multiple, .registration = TRUE
```

- les fichiers de travail de C++ sont exclus du contrôle de source dans .gitignore :

```
# C binaries
src/*.o
src/*.so
src/*.dll
```

Ces modifications sont en partie effectuées automatiquement, pour **Rcpp** seulement, par **usethis**, mais l'insertion manuelle du code est plus rapide et fiable : ne pas utiliser cette commande.

```
# usethis::use_rcpp()
```

La construction du package entraînera la compilation du code : les Rtools sont donc indispensables.

5.5.5 Package bien rangé

Tout package moderne doit être compatible avec le tidyverse, ce qui nécessite peu d'efforts :

- pour permettre l'utilisation de pipelines, l'argument principal des fonctions doit être le premier ;
- les fonctions qui transforment des données doivent accepter un dataframe ou un tibble comme premier argument et retourner un objet du même format ;
- les méthodes `plot()` doivent être doublées de méthodes `autoplot()` avec les mêmes arguments qui produisent le même graphique avec **ggplot2**.

5.6 Bibliographie

La documentation d'un package fait appel à des références bibliographiques. Elles peuvent être gérées automatiquement avec **Rdpack** et **roxygen2**. Les références utilisées dans les fichiers R Markdown (vignette, site produit par **pkgdown**) ne sont pas concernées.

5.6.1 Préparation

Les références bibliographiques doivent être placées dans un fichier `bibtex REFERENCES.bib` placé dans le dossier `inst`. Ce dossier contient des fichiers qui seront placés à la racine du dossier du package quand il sera installé.

Ajouter la ligne suivante à `DESCRIPTION` :

```
RdMacros: Rdpack
```

Ajouter aussi le package **Rdpack** à la liste des packages importés :

```
Imports: magrittr, stats, Rcpp, Rdpack
```

Enfin, importer la fonction `reprompt()` de **Rdpack** en ajoutant les lignes suivantes à la documentation pour **roxygen2** dans `package.R` :

```
#' @importFrom Rdpack reprompt
```

5.6.2 Citations

Les références sont citées par la commande `\insertCite{key}{package}` dans la documentation destinée à **roxygen2**. package est le nom du package dans lequel le fichier REFERENCES.bib doit être cherché : ce sera normalement le package en cours, mais les références d'autres packages sont accessibles, à la seule condition qu'ils utilisent **Rdpack**.

key est l'identifiant de la référence dans le fichier. Exemple⁹ : documentation du package **SpatDiv** hébergé sur GitHub, fichier .R du package :

```
#' SpatDiv
#'
#' Spatially Explicit Measures of Diversity
#'
#' This package extends the **entropart** package
#' \insertCite{Marcon2014c}{SpatDiv}.
#' It provides spatially explicit measures of
#' diversity such as the mixing index.
```

La référence citée se trouve dans inst/REFERENCES.bib :

```
@Article{Marcon2014c,
  author = {Marcon, Eric and Herault, Bruno},
  title = {entropart, an R Package to Partition
           Diversity},
  journal = {Journal of Statistical Software},
  year = {2015},
  volume = {67},
  number = {8},
  pages = {1--26},
```

Les citations sont entre parenthèses. Pour placer le nom de l'auteur hors de la parenthèse, ajouter la déclaration ;textual :

```
\insertCite[Marcon2014c;textual]{SpatDiv}
```

Pour citer plusieurs références (forcément du même package), les séparer par des virgules.

A la fin de la documentation d'un objet utilisant des citations, ajouter systématiquement une liste des références :

```
#' @references
#' \insertAllCited{}
```

⁹Package **SpatDiv** sur GitHub : <https://github.com/EricMarcon/SpatDiv/blob/master/R/package.R>

5.7 Données

Des données peuvent être intégrées à un package, notamment pour la clarté des exemples.

La méthode la plus simple consiste à utiliser **use_this**. Créer des variables contenant les données à sauvegarder puis les sauvegarder :

```
seq1_10 <- 1:10
seq1_100 <- 1:100
usethis::use_data(seq1_10, seq1_100)
```

Un fichier .rda est créé dans le dossier data pour chaque variable créée. Avec l'option LazyData activée dans DESCRIPTION, les variables seront disponibles dès le chargement du package, mais ne seront effectivement chargées en mémoire qu'après leur première utilisation.

Chaque variable doit être documentée dans le fichier package.R :

```
#' seq1_10
#'
#' A sequence of numbers from 1 to 10
#'
#' @format A numeric vector.
#' @source Values computed by the R software,
#'   \url{https://www.r-project.org/}
"seq1_10"
```

Le nom de la variable est donné entre guillemets après le bloc de commentaires (à la place du code R d'une fonction). @format décrit le format des données et @source permet d'indiquer leur source.

5.8 Tests unitaires

Dans l'idéal, tout le code inclus dans un package devrait être testé de multiples façons :

- contre les erreurs de syntaxe : les procédures de vérification de R s'en chargent assez bien ;
- pour vérifier la conformité des résultats de calculs aux valeurs attendues ;
- contre la survenue d'erreurs si les utilisateurs n'utilisent pas le code comme le développeur l'a prévu (arguments incorrects passés aux fonctions, données inadéquates...).

Les tests unitaires sont utilisés dans les deux derniers objectifs. Ils s'appuient sur **testthat** à intégrer au package :

```
usethis::use_testthat()

## 
## Attaching package: 'testthat'
```

5. PACKAGE

```
## The following object is masked from 'package:targets':
##
##     matches

## The following object is masked from 'package:dplyr':
##
##     matches

## The following object is masked from 'package:purrr':
##
##     is_null

## The following objects are masked from 'package:readr':
##
##     edition_get, local_edition

## The following object is masked from 'package:tidyverse':
##
##     matches
```

Les tests doivent être ajoutés sous la forme de fichiers .R dont le nom commence obligatoirement par `test` dans le dossier `tests/testthat`.

Chaque test (donc le contenu de chaque fichier) commence par son contexte, c'est-à-dire ce un ensemble de tests. Exemple, dans un fichier `test_double.R` :

```
context("function double")
```

Les tests sont contenus dans des fichiers qui les regroupent par thème, par exemple `test_double.R`. Le nom de chaque test est passé comme argument de la fonction `test_that()` :

```
test_that("Double values are correct", {
  skip_on_cran()
  x <- 1:2
  # 2 x 2 should be 4
  expect_equal(double(x), c(2, 4))
  # The result should be a number (type = 'double')
  expect_type(double(x), "double")
  # Error management
  expect_error(double("a"))
})
```

```
}
```

```
## Test passed
```

Toutes les fonctions commençant par `expect` permettent de comparer leur premier argument à un résultat : dans l'exemple ci-dessus, le résultat de `double(1:2)` doit être 2 4 et le type de ce vecteur doit être réel à double précision. Le dernier test vérifie qu'une chaîne de caractère passée comme argument génère une erreur, ce qui n'est pas optimal : si le package traitait l'erreur, le message retourné pourrait être testé.

La commande `skip_on_cran()`, à utiliser systématiquement, évite de lancer les tests sur CRAN quand le package y sera déposé : CRAN dispose de ressources limitées et restreint strictement le temps de vérification des packages sur

sa plateforme. Les tests devront donc être réalisés sur GitHub, grâce à l'intégration continue, voir section 5.10.

Les tests peuvent être lancés par le menu “More > Test package” de la fenêtre *Build* ou par la commande `devtools::test()`.

Il est conseillé d'écrire les tests dès qu'une fonction du package est stabilisée.

5.9 Fichier .gitignore

Le fichier `.gitignore` obtenu à ce stade est incomplet. Il peut être remplacé par celui-ci :

```
# History files
.Rhistory
.Rapp.history
# Session Data files
.RData
# Example code in package build process
*-Ex.R
# Output files from R CMD build
/*.tar.gz
# Output files from R CMD check
/*.Rcheck/
# RStudio files
.Rproj.user/
.Rprofile
# knitr and R markdown default cache directories
*_cache/
/cache/
# Temporary files created by R markdown
*.utf8.md
*.knit.md
# C binaries
src/*.o
src/*.so
src/*.dll
/src-i386/
/src-x64/
# uncomment if pkgdown is run by CI
# docs/
```

La dernière ligne concerne le dossier `docs/`, qui reçoit le site web produit par **pkgdown**. Elle est commentée tant que la production du site est réalisée localement, mais décommentée si elle est confiée à GitHub Actions (voir section suivante).

5.10 Intégration continue

La vérification (*Check*) du package doit être effectuée à chaque étape du développement, ce qui consomme un temps considérable. Elle peut être automatisée très simplement avec le service GitHub Actions, déclenché à chaque modification du dépôt sur GitHub. L'analyse de la couverture du code par les tests (quelles parties du codes sont testées ou non) sera ajoutée.

GitHub est également capable de reconstruire la documentation du package avec **pkgdown**, autre opération consommatrice de ressources, après la réussite des tests.

La section [6.3.5](#) détaille le moyen de le faire.

5.11 CRAN

Les packages dont l'audience dépasse l'entourage de l'auteur peuvent être déposés sur CRAN. Les règles à respecter sur CRAN sont nombreuses¹⁰. Elles sont vérifiées par la commande de vérification R CMD check avec l'option --as.cran. La vérification ne doit renvoyer aucune erreur, aucun avertissement, ni aucune note avant de soumettre le package.

5.11.1 Test du package

La vérification du package par GitHub dans le cadre de l'intégration continue n'est pas suffisante. Le package doit être testé sur la version de développement de R. Le site *R-hub builder*¹¹ permet de le faire simplement.

Le package, dont la version ne doit pas être de développement (limitée à trois nombres, voir section [5.2.1](#)), doit être construit au format source : dans la fenêtre *Build* de RStudio, cliquer sur “More > Build Source Package”. Sur le site *R-hub builder*, cliquer sur “Advanced”, sélectionner le fichier source du package et la plateforme de test : *Debian Linux, R-devel, GCC*.

Le package **rhub** permet d'utiliser la même plateforme de vérification que le site *R-hub builder* depuis RStudio. La première étape consiste à valider son adresse de messagerie avec la commande validate_email(). Ensuite, il suffit d'appeler la fonction check_for_cran() pour lancer une vérification complète.

5.11.2 Soumission

Quand le package est au point, la soumission à CRAN se fait par le site web dédié¹².

En cas de rejet, traiter les demandes et soumettre à nouveau en incrémentant le numéro de version.

5.11.3 Maintenance

Des demandes de corrections sont envoyées par CRAN de temps à autre, notamment lors des changements de version de R. L'adresse de messagerie du responsable du package (*maintainer*) doit rester valide et les demandes doivent être traitées rapidement. Dans le cas contraire, le package est archivé.

¹⁰<https://cran.r-project.org/web/packages/policies.html>

¹¹<https://builder.r-hub.io/>

¹²<https://xmpalantir.wu.ac.at/cransubmit/>

Les nouvelles versions du package sont soumises de la même façon que la première.

INTÉGRATION CONTINUE

L'intégration continue consiste à confier à un service externe la tâche de vérifier un package, produire des documents Markdown pour les pages web d'un dépôt GitHub ou tricoter entièrement un site web à partir du code.

Toutes ces tâches peuvent être accomplies localement sur le poste de travail mais prennent du temps et risquent de ne pas être répétées à chaque mise à jour. Dans le cadre de l'intégration continue, elles le sont systématiquement, de façon transparente pour l'utilisateur. En cas d'échec, un message d'alerte est envoyé.

La mise en place de l'intégration continue se justifie pour des projets lourds, avec des mises à jour régulières. plutôt que pour des projets contenant un simple document Markdown rarement modifié.

6.1 Outils

6.1.1 GitHub Actions

L'outil utilisé le plus fréquemment pour des projets R déposés sur GitHub était *Travis CI*¹ mais le service est devenu payant en 2021.

Les Actions GitHub remplacent avantageusement Travis. Ce service est intégré à GitHub.

6.1.2 codecov

Pour évaluer le taux de couverture du code des packages R, c'est-à-dire la proportion du code testé d'une façon ou d'une autre (exemples, tests unitaires, vignette), le service *Codecov*² s'intègre parfaitement à GitHub.

¹<https://travis-ci.org/>

²<https://codecov.io/>

Il faut ouvrir un compte, de préférence en s’authentifiant par GitHub.

6.1.3 GitHub Pages

Les pages web de GitHub peuvent être hébergées dans le répertoire `docs` de la branche master du projet : c’est la solution retenue quand elle sont produites sur le poste de travail.

Si elles sont produites par intégration continue, elle le seront obligatoirement dans une branche dédiée appelée `gh-pages`.

6.2 Principes

Un projet de document est traité en exemple. L’objectif est de faire tricoter par GitHub un projet Markdown. Cette pratique est appropriée pour les projets d’ouvrages, qui nécessitent beaucoup de ressources pour leur construction. Dans ce type de projet, le code est tricoté par knitr pour produire plusieurs documents, typiquement aux formats HTML et PDF, accessibles sur les pages GitHub. Quand les documents sont produits localement, ils sont placés dans le dossier `docs` et poussés sur GitHub.

Pour que GitHub s’en charge, quelques réglages sont nécessaires.

6.2.1 Obtention d’un jeton d’accès personnel

Pour écrire sur GitHub, le service d’intégration continue devra s’authentifier au moyen d’un jeton d’accès personnel (*Personal Access Token* : PAT) dont la création est décrite en section [1.4.4](#).

Générer un nouveau jeton, le décrire en tant que “GitHub Actions” et lui donner l’autorisation “repo”, c’est-à-dire modifier *tous* les dépôts (il n’est pas possible de limiter l’accès à un dépôt particulier).

6.2.2 Secrets du projet

Sur GitHub, afficher les paramètres du projet et sélectionner “Secrets”. Le bouton “New Repository Secret” permet de stocker des variables utilisées dans les scripts des Actions GitHub (visibles publiquement) sans en diffuser la valeur. Le jeton d’accès personnel est indispensable pour que les Actions GitHub puissent écrire leur production dans le projet. Créer un secret nommé “GH_PAT” et saisir la valeur du jeton sauvegardée précédemment. Après avoir cliqué sur “Add Secret”, le jeton ne pourra plus être lu.

Pour permettre l’envoi de messages de succès ou d’échec sans diffuser son adresse de messagerie, créer un secret nommé “EMAIL” qui la contient.

6.2.3 Activation du dépôt sur CodeCov

L’analyse de la couverture du code des packages est utile pour détecter les portions de code non testées. En revanche, l’analyse de la couverture des projets de document n’a pas d’intérêt.

Pour activer un dépôt, il faut d’authentifier sur le site de CodeCov avec son compte GitHub. La liste des dépôts est affichée et peut être actualisée. Si les dépôts à traiter sont hébergés par une organisation, par exemple les dépôts d’une salle de classe GitHub, il faut actualiser la liste des organisations en suivant les instructions (un lien permet de modifier rapidement les options de GitHub pour autoriser la lecture d’une organisation par Codecov) et à nouveau mettre à jour la liste des dépôts. Enfin, quand le dépôt recherché est visible, il faut l’activer. Il est inutile d’utiliser le système de jetons de Codecov.

6.2.4 Scripter les actions de GitHub

Un flux de travail (*workflow*) de GitHub est une succession de tâches (*jobs*) comprenant des étapes (*steps*). Un flux de travail est déclenché par un évènement, généralement chaque *push* du projet, mais aussi à intervalles réguliers (*cron*).

Typiquement, les flux créés ici contiennent deux tâches : la première installe R et les composants nécessaires et exécute des scripts R (ce qui constitue ses étapes successives); la seconde publie des fichiers obtenus dans les pages GitHub.

Les flux de travail sont configurés dans un fichier au format YAML placé dans le dossier `.github/workflows/` du projet. Les différentes parties du script sont présentées ci-dessous. Le script complet est celui de ce document, accessible sur GitHub³.

Déclenchement

L’action est déclenchée à chaque fois que des mises à jour sont poussées sur GitHub :

```
on:
  push:
    branches:
      - master
```

La branche prise en compte est *master* (à remplacer par *main* le cas échéant).

Pour déclencher l’action périodiquement, il faut utiliser la syntaxe de *cron* (le système de planification des tâches sous Unix) :

```
on:
  schedule:
    - cron: '0 22 * * 0'  # every sunday at 22:00
```

³<https://github.com/EricMarcon/travailleR/blob/master/.github/workflows/bookdown.yml>

6. INTÉGRATION CONTINUE

Les valeurs successives sont celles des minutes, des heures, du jour (quatrième du mois), du mois et du jour de la semaine (0 pour dimanche à 6 pour samedi). Les * permettent d'ignorer une valeur.

Les entrées push et schedule peuvent être utilisées ensemble :

```
on:  
  push:  
    branches:  
      - master  
  schedule:  
    - cron: '0 22 * * 0'
```

Actuellement, la planification n'est prise en compte que dans la branche *master*.

Nom du flux de travail

Le nom du flux est libre. Il sera affiché par le badge qui sera ajouté dans le fichier README.md du projet (voir section 6.4).

```
name: bookdown
```

Première tâche

Les tâches sont décrites dans la rubrique jobs. renderbook est le nom de la première tâche : il est libre. Ici, l'action principale consistera à produire un ouvrage bookdown avec la fonction `render_book()`, d'où son nom.

```
jobs:  
  renderbook:  
    runs-on: macOS-latest
```

La déclaration `runs-on` décrit le système d'exploitation sur lequel la tâche doit s'exécuter. Les choix possibles sont Windows, Ubuntu ou MacOS⁴. L'intégration continue de R sur GitHub utilise habituellement MacOS qui a l'avantage d'utiliser des packages R compilés donc beaucoup plus simples (certains packages nécessitent des librairies extérieures à R pour leur compilation) et rapides à installer, tout en permettant l'usage de scripts.

Premières étapes

Les étapes sont décrites dans la rubrique steps.

```
steps:  
  - name: Checkout repo  
    uses: actions/checkout@v2  
  - name: Setup R  
    uses: r-lib/actions/setup-r@v1  
  - name: Install pandoc  
    run: |  
      brew install pandoc
```

⁴https://docs.github.com/en/free-pro-team@latest/actions/reference/workflow-syntax-for-github-actions#jobsjob_idruns-on

Chaque étape est décrite par son nom (libre) et ce qu'elle réalise.

La force de GitHub Actions est de permettre l'utilisation d'*actions* écrites par d'autres et stockées dans un projet public GitHub. Une action est un script accompagné de métadonnées qui décrivent son usage. Son développement est accompagné par des numéros de version successifs. On appelle une action par l'instruction `uses` : le projet GitHub qui la contient et sa version.

Dans leur projet GitHub respectif, les actions existent dans leur version de développement (@master) et dans des versions d'étape (*release*) accessibles par leur numéro (@v1). Ces versions d'étape sont préférables parce qu'elles sont stables.

Les actions généralistes sont mises à disposition par GitHub dans l'organisation GitHub Actions⁵. L'action "actions/checkout" permet de se placer dans la branche principale du projet traité par le flux de travail : c'est en général la première étape de tous les flux.

L'action suivante est l'installation de R, mise à disposition par l'organisation *R infrastructure*⁶.

L'installation de pandoc (logiciel extérieur à R mais nécessaire à R Markdown) peut être réalisée par une commande exécutée par MacOS. Elle est appelée par `run` : et peut contenir plusieurs lignes (d'où le `|`). Ce script dépend du système d'exploitation : `brew` est le gestionnaire de paquets de MacOS. Pour éviter les spécificités d'un système, il est préférable d'utiliser une action :

```
- name: Install pandoc
  uses: r-lib/actions/setup-pandoc@v1
```

Caches

L'installation des packages de R prend du temps, beaucoup s'ils sont installés à partir des sources (la procédure standard sous Ubuntu, mais pas sous MacOS et Windows où les packages binaires sont utilisés par défaut). Le calcul des bouts de code est en général l'étape la plus longue du flux de travail. L'action `cache` permet de mettre en cache les résultats des deux opérations.

```
- name: Cache Renv packages
  uses: actions/cache@v2
  with:
    path: $HOME/.local/share/renv
    key: r-${{ hashFiles('renv.lock') }}
    restore-keys: r-
- name: Cache bookdown results
  uses: actions/cache@v2
  with:
    path: _bookdown_files
    key: bookdown-${{ hashFiles('**/*Rmd') }}
    restore-keys: bookdown-
```

Le cache est mis à jour en cas de modification d'un package ou d'un bout de code, ce qui nécessite un moyen rapide de vérifier les modifications : une valeur

⁵<https://github.com/actions/>

⁶<https://github.com/r-lib/>

6. INTÉGRATION CONTINUE

de contrôle (*hashtag*) est calculée par la fonction `hashFiles()` à partir du fichier `renv.lock` (voir ci-dessous) pour les packages et l'ensemble des fichiers `.Rmd` pour les bouts de code. Tout changement entraîne la réinstallation des packages ou le recalcul de l'ensemble du code : la gestion du cache est moins fine que celle de R sur un poste de travail, qui ne recalcule que les bouts de code modifiés.

Packages

L'installation des packages est gérée par la fonction `install.packages()`. Plutôt que d'énumérer les packages à installer dans les arguments de la fonction, source d'erreur, il est préférable d'utiliser le package `renv` pour enregistrer tous les packages utilisés par le projet et les installer en une fois pour l'intégration continue. `renv` installera les packages dans la version enregistrée, ce qui permet d'éviter des effets imprévus dus à des versions différentes entre le poste de travail et GitHub Actions.

```
- name: Install packages
  run: |
    R -e 'install.packages("renv")'
    R -e 'renv::restore()'
```

Il est nécessaire d'installer `renv` sur le poste de travail utilisé pour le développement du projet.

Il faut utiliser un fichier `DESCRIPTION` pour lister les packages de tout projet, comme si c'était un package R. Pour ce document :

```
Package: travailleR
Title: Travailler avec R
Version: 1.1.0
Authors@R: c(
  person("Eric", "Marcon", , "e.marcon@free.fr", c("aut", "cre"))
)
URL: https://github.com/EricMarcon/travailleR
Imports:
  bookdown,
  (...)
```

Avant de déclencher le flux de travail, il est nécessaire de créer la liste des packages dans leur version en cours sur le poste de travail :

```
renv::snapshot(type = "explicit")
```

A sa première utilisation, le package `renv` informe de quelques adaptations de l'environnement de travail, qu'il faut accepter.

Cette commande crée le fichier `renv.lock` qui est utilisé par GitHub Actions pour installer les packages pendant l'intégration continue. Il pourra être mis à jour à tout moment pour prendre en compte leur mise à jour.

Alternativement, les packages nécessaires peuvent être installés sans l'aide de `Renv` :

```
- name: Install packages
  run: |
    options(pkgType = "binary")
    options(install.packages.check.source = "no")
    install.packages(c("remotes", "bookdown", "formatR", "tinytex"))
    tinytex::install_tinytex()
    remotes::install_deps(dependencies = TRUE)
  shell: Rscript {0}
```

Cette étape utilise Rscript comme environnement de commande, ce qui lui permet d'exécuter directement des commandes R (à comparer à l'utilisation de R -e dans les exemples précédents).

Les packages servant à produire le document sont listés :

- **remotes** pour sa fonction `install_deps()`;
- **bookdown** pour tricoter;
- **formatR** pour la mise en forme des bouts de code (`tidy=TRUE`);
- **tinytex** pour disposer d'une distribution LaTeX.

Les autres packages, ceux utilisés par le projet, sont lus dans le fichier DESCRIPTION par la fonction `install_deps()`.

Sous MacOS, les packages sont installés par défaut en version binaire, mais à partir de leur code source s'il est plus récent. La création des packages binaires prend quelques jours à CRAN : cette situation n'est donc pas rare. Les packages ne contenant que du code R ou du code C++ sans référence à des librairies externes s'installent en revanche sans problème. En revanche, si le package nécessite des librairies externes à R ou une compilation de code Fortran, l'installation échoue. Il serait donc nécessaire d'installer préalablement les librairies nécessaires (et éventuellement un compilateur Fortran) à l'ensemble des packages dont le projet dépend : cette solution n'est pas réaliste parce qu'elle implique l'inventaire de l'ensemble des dépendances, qui peuvent changer, et un nombre important d'installations chronophages et inutiles la plupart du temps, quand les packages binaires sont à jour. Une meilleure solution est de forcer l'installation des packages binaires même si le code source est plus récent : c'est l'objet des deux options de R définies avant l'appel à `install.packages()`.

Dans cette approche, les packages ne sont pas mis en cache.

Tricot

La production de l'ouvrage est lancée par une commande R.

```
- name: Render pdf book
  run: |
    bookdown::render_book("index.Rmd", "bookdown::pdf_book")
  shell: Rscript {0}
- name: Render gitbook
  run: |
    bookdown::render_book("index.Rmd", "bookdown::gitbook")
  shell: Rscript {0}
```

6. INTÉGRATION CONTINUE

Les formats paramétrés dans `_output.yml` sont ignorés.

Le fichier PDF doit être produit avant le format GitBook pour que son lien de téléchargement soit ajouté à la barre de menu du site GitBook. D'autre part, R doit être fermé et rouvert entre les deux rendus faute de quoi les tableaux ne sont pas créés correctement dans le GitBook⁷. Les deux étapes ne doivent pas être regroupées en une seule.

Sauvegarde

Le résultat du tricot, placé dans le dossier `docs` de la machine virtuelle en charge de l'intégration continue, doit être préservé pour que la tâche suivante puisse l'utiliser.

La dernière étape de la tâche de production utilise l'action `upload-artifact` pour cela.

```
- name: Upload artifact
  uses: actions/upload-artifact@v1
  with:
    name: _book
    path: docs/
```

Le contenu de `docs` est sauvegardé en tant qu'*artefact* nommé “`_book`”. Les artefacts sont visibles publiquement sur la page des Actions du projet GitHub.

Après sa dernière étape, la machine virtuelle utilisée pour cette étape est détruite.

Publication

La publication de l'artefact dans la branche `gh-pages` du projet nécessite une autre tâche.

```
deploy:
  runs-on: ubuntu-latest
  needs: renderbook
  steps:
    - name: Download artifact
      uses: actions/download-artifact@v1
      with:
        # Artifact name
        name: _book
        # Destination path
        path: docs
    - name: Deploy to GitHub Pages
      uses: Cecilapp/GitHub-Pages-deploy@v3
      env:
        GITHUB_TOKEN: ${{ secrets.GH_PAT }}
      with:
        email: ${{ secrets.EMAIL }}
        build_dir: docs
        jekyll: no
```

⁷<https://stackoverflow.com/questions/46080853/why-does-rendering-a-pdf-from-markdown-require-closing-rstudio-between-renders/46083308#46083308>

La tâche est nommée “deploy” (le nom est libre). Elle s’exécute sur une machine virtuelle sous Ubuntu. Elle ne peut se lancer que si la tâche “renderbook” a réussi. Ses étapes sont les suivantes :

- *Download artifact* : Restauration du dossier docs ;
- *Deploy to GitHub Pages* : copie du dossier docs dans la branche gh-pages.

Cette dernière étape utiliser l’action GitHub-Pages-deploy mise à disposition par l’organisation *Cecilapp*. Elle utilise une variable d’environnement, GITHUB_TOKEN, pour s’authentifier et des paramètres :

- *email* : l’adresse de messagerie destinataire du rapport d’exécution. Pour ne pas exposer l’adresse publiquement, elle a été stockée dans un secret du projet ;
- *buid_dir* : le répertoire à publier,
- *jekyll* :*no* pour créer un fichier vide nommé .nojekyll qui indique aux pages GitHub de ne pas essayer de traiter leur contenu comme un site web Jekyll.

6.2.5 Données confidentielles dans un dépôt public

Si le projet contient des données confidentielles (section 3.6), GitHub Actions doit utiliser la clé privée du projet pour les extraire de leur coffre-fort.

La clé privée doit être stockée dans un secret du projet, nommé “RSA”. L’étape suivante, à insérer avant l’étape de tricot, écrit la clé dans un fichier pour que le code du projet y ait accès.

```
- name: Private key
  run: |
    cat("${{ secrets.rsa }}", file="NomDuProjet.rsa"
    shell: Rscript {0}
```

6.3 Modèles de scripts

Des modèles de scripts pour tous les types de projets sont présentés ici. Tous nécessitent même préparation :

- les secrets GH_PAT et EMAIL doivent être enregistrés dans le projet GitHub (section 6.2.2) ;
- un fichier DESCRIPTION doit être utilisé pour lister les packages nécessaires (section 5.2.1), quel que soit le type de projet ;
- un instantané des packages installés (`renv.lock`) doit être réalisé si `renv` (section 6.2.4) est utilisé.

La branche `gh-pages` est créée automatiquement par les scripts. Vérifier après la première exécution que les pages GitHub sont bien activées sur cette branche (section 3.7). Supprimer ensuite le dossier `docs` s'il existait, pousser la modification sur GitHub et enfin ajouter la ligne suivante au fichier `.gitignore` pour pouvoir tricoter localement les projets sans perturber GitHub :

```
docs/
```

6.3.1 memoiR

La fonction `build_ghworkflow()` du package **memoiR** crée automatiquement les scripts nécessaires à la production des modèles du package. Le script est toujours nommé `memoir.yml`.

Ces scripts n'utilisent ni `renv` ni `cache`. Ils n'ont pas besoin d'un fichier `DESCRIPTION` pour l'installation des dépendances mais chaque document doit contenir son le bout de code de paramétrage (`Options`) la liste de tous les packages nécessaires à son tricot (stockés dans la variable `Packages`).

6.3.2 Projet d'ouvrage

Le flux de travail s'appelle `rmarkdown`; sa tâche de production `render`.

```
on:
  push:
    branches:
      - master

name: rmarkdown

jobs:
  render:
    runs-on: macOS-latest
    steps:
      - name: Checkout repo
        uses: actions/checkout@v2
      - name: Setup R
        uses: r-lib/actions/setup-r@v1
      - name: Install pandoc
        uses: r-lib/actions/setup-pandoc@v1
      - name: Install dependencies
        run: |
          options(pkgType = "binary")
          options(install.packages.check.source = "no")
          install.packages(c("memoiR", "rmdformats", "tinytex"))
          tinytex::install_tinytex()
        shell: Rscript {0}
      - name: Render pdf book
        run: |
          bookdown::render_book("index.Rmd", "bookdown::pdf_book")
        shell: Rscript {0}
      - name: Render gitbook
        run: |
          bookdown::render_book("index.Rmd", "bookdown::gitbook")
        shell: Rscript {0}
      - name: Upload artifact
        uses: actions/upload-artifact@v1
        with:
```

```

        name: ghpages
        path: docs
deploy:
  runs-on: ubuntu-latest
  needs: render
  steps:
    - name: Download artifact
      uses: actions/download-artifact@v1
      with:
        name: ghpages
        path: docs
    - name: Deploy to GitHub Pages
      uses: Cecilapp/GitHub-Pages-deploy@v3
      env:
        GITHUB_TOKEN: ${{ secrets.GH_PAT }}
      with:
        email: ${{ secrets.EMAIL }}
        build_dir: docs
        jekyll: no

```

6.3.3 Articles et présentations

Le flux de travail s'appelle `rmarkdown`; sa tâche de production `render`.

```

on:
  push:
    branches:
      - master

name: rmarkdown

jobs:
  render:
    runs-on: macOS-latest
    steps:
      - name: Checkout repo
        uses: actions/checkout@v2
      - name: Setup R
        uses: r-lib/actions/setup-r@v1
      - name: Install pandoc
        uses: r-lib/actions/setup-pandoc@v1
      - name: Install dependencies
        run: |
          options(pkgType = "binary")
          options(install.packages.check.source = "no")
          install.packages(c("memoir", "rmdformats", "tinytex"))
          tinytex::install_tinytex()
        shell: Rscript {0}
      - name: Render Rmarkdown files
        run: |
          RMD_PATH=$(ls | grep "[.]Rmd$")
          Rscript -e 'for (file in commandArgs(TRUE)) |>
            rmarkdown::render(file, "all")' ${RMD_PATH[*]}
          Rscript -e 'memoir::build_githubpages()'
      - name: Upload artifact
        uses: actions/upload-artifact@v1
        with:
          name: ghpages
          path: docs
deploy:
  runs-on: ubuntu-latest
  needs: render
  steps:
    - name: Download artifact

```

6. INTÉGRATION CONTINUE

```
uses: actions/download-artifact@v1
with:
  name: ghpages
  path: docs
- name: Deploy to GitHub Pages
  uses: Cecilapp/GitHub-Pages-deploy@v3
  env:
    GITHUB_TOKEN: ${{ secrets.GH_PAT }}
  with:
    email: ${{ secrets.EMAIL }}
    build_dir: docs
    jekyll: yes
```

L'étape chargée du tricot utilise un script pour lister tous les fichiers .Rmd, les traiter (tous les formats de sortie listés dans leur entête yaml sont produits). La fonction `build_githubpages()` (voir section 4.3.2) place les résultats dans `docs`.

La tâche de déploiement indique aux pages GitHub d'utiliser Jekyll, c'est-à-dire d'utiliser le fichier `README.md` comme page d'accueil.

Si l'étape de tricot nécessite de modifier la langue utilisée par R, par exemple pour afficher correctement la date de production des documents, elle peut être modifiée comme ceci :

```
- name: Render Rmarkdown files
  run: |
    Sys.setlocale("LC_TIME", "fr_FR")
    lapply(list.files(pattern="*.Rmd"), function(file) rmarkdown::render(file, "all"))
    memoirR::build_githubpages()
    shell: Rscript {0}
```

La sélection des fichiers est ici réalisée par un script R, qui inclut une commande de localisation, ici en Français.

Cette étape peut être complétée par la sélection d'un thème GitHub Pages pour que la page d'accueil contienne un lien vers le code :

```
run: |
  echo 'theme: jekyll-theme-slate' > docs/_config.yml
```

Le thème est ici “Slate”, un des choix proposés par les pages GitHub.

6.3.4 Site web blogdown

Le fichier appelé `blogdown.yml` est très similaire. Le nom du flux de travail est `blogdown` et celui de la tâche de production est `buildsite`.

```
on:
  push:
    branches:
      - master
  schedule:
    - cron: '0 22 * * 0'

name: blogdown
```

```

jobs:
  buildsite:
    runs-on: macOS-latest
    steps:
      - name: Checkout repo
        uses: actions/checkout@v2
      - name: Setup R
        uses: r-lib/actions/setup-r@v1
      - name: Install pandoc
        uses: r-lib/actions/setup-pandoc@v1
      - name: Install packages
        run: |
          options(pkgType = "binary")
          options(install.packages.check.source = "no")
          install.packages(c("remotes", "blogdown", "formatR"))
          remotes::install_deps(dependencies = TRUE)
        shell: Rscript {0}
      - name: Build website
        run: |
          blogdown::install_hugo(force = TRUE)
          blogdown::build_site(local = TRUE, build_rmd = TRUE)
        shell: Rscript {0}
      - name: Upload artifact
        uses: actions/upload-artifact@v2
        with:
          name: _website
          path: public/
  deploy:
    runs-on: ubuntu-latest
    needs: buildsite
    steps:
      - name: Download artifact
        uses: actions/download-artifact@v1
        with:
          # Artifact name
          name: _website
          # Destination path
          path: public
      - name: Deploy to GitHub Pages
        uses: Cecilapp/GitHub-Pages-deploy@v3
        env:
          GITHUB_TOKEN: ${{ secrets.GH_PAT }}
        with:
          build_dir: public
          email: ${{ secrets.EMAIL }}
          jekyll: no

```

L'action `checkout` se place dans la branche source avec sa variable `ref`.

La tâche `Build website` utilise le package **blogdown** pour installer Hugo (le générateur de sites web) et ensuite construire le site.

Si le site web utilise des données en ligne qui justifient de le mettre à jour périodiquement, GitHub Actions peut être lancé tous les jours, toutes les semaines ou tous les mois en plus des reconstructions déclenchées par une modification du dépôt (voir section 6.2.4). Ici, le site est reconstruit tous les dimanches à 22h.

Exemple : la page qui affiche la bibliométrie du site web⁸ de l'auteur interroge Google Scholar pour afficher les citations des publications. Le site est mis à jour toutes les semaines pour que les statistiques soient à jour.

⁸<https://EricMarcon.github.io/fr/publication/>

6.3.5 Packages R

Un script optimal pour la vérification d'un package est le suivant :

```
on:
  push:
    branches:
      - master

  name: R-CMD-check

  jobs:
    R-CMD-check:
      runs-on: macOS-latest
      env:
        GITHUB_PAT: ${{ secrets.GH_PAT }}
      steps:
        - uses: actions/checkout@v2
        - uses: r-lib/actions/setup-r@v1
        - name: Install pandoc
          uses: r-lib/actions/setup-pandoc@v1
        - name: Install dependencies
          run: |
            options(pkgType = "binary")
            options(install.packages.check.source = "no")
            install.packages(c("remotes", "rcmdcheck", "covr", "pkgdown"))
            remotes::install_deps(dependencies = TRUE)
            shell: Rscript {0}
        - name: Check
          run: rcmdcheck::rcmdcheck(args = "--no-manual", error_on = "warning")
          shell: Rscript {0}
        - name: Test coverage
          run: covr::codecov(type="all")
          shell: Rscript {0}
        - name: Install package
          run: R CMD INSTALL .
        - name: Pkgdown
          run: |
            git config --local user.email "actions@github.com"
            git config --local user.name "GitHub Actions"
            Rscript -e 'pkgdown::deploy_to_branch(new_process = FALSE)'
```

Le fichier est nommé `check.yml`. Il ne contient qu'une seule tâche, nommée `R-CMD-check` comme le flux.

Le script n'utilise pas **Renv** pour gérer les packages parce que la vérification d'un package doit fonctionner avec les versions en cours sur CRAN. **remotes** installe les packages nécessaires à partir du fichier `DESCRIPTION`.

L'étape `Check` vérifie le package. Les avertissements sont traités comme des erreurs.

L'étape `Test coverage` utilise le package **covr** pour mesurer le taux de couverture et téléverse les résultats sur le site [Codecov](#).

Enfin, les deux dernières étapes installent le package puis utilisent **pkgdown** pour créer le site de documentation du package et le pousser dans la branche `gh-pages` du projet.

Ce script ne contient qu'une tâche : le déploiement du site de documentation est directement exécuté par **pkgdown**. Son succès est affiché par un badge à afficher dans le fichier `README.md` (voir section 6.4)

Des scripts plus complexes sont proposés par R-lib⁹, notamment pour exécuter les tests sur plusieurs systèmes d'exploitation et plusieurs versions de R. Ces tests poussés sont à effectuer avant de soumettre à CRAN (section 5.11) mais consomment trop de ressource pour un usage systématique.

6.4 Ajouter des badges

Le succès des Actions GitHub est visible en ajoutant un badge dans le fichier README.md, juste après le titre du fichier. Sur la page du projet, choisir “Actions” puis sélectionner l’action (dans “Workflows”). Cliquer sur le bouton “...” puis sur “Create Status Badge”. Coller le code Markdown :

```
# Nom du projet
! [bookdown] (https://github.com/<GitHubID>/<Depot>/workflows/<NomDuFlux>/badge.svg)
```

Le nom du flux a été déclaré dans l’entrée name : du fichier de configuration des actions GitHub.

Le taux de couverture mesuré par Codecov peut aussi être affiché par un badge :

```
[! [codecov] (https://codecov.io/github/<GitHubID>/<Depot>/branch/master/graphs/badge.svg)
(https://codecov.io/github/<GitHubID>/<Depot>)]
```

⁹<https://github.com/r-lib/actions/tree/master/examples#standard-ci-workflow>

SHINY

Shiny permet de publier sous la forme d'un site web une application interactive utilisant du code R. Le site peut fonctionner localement, sur le poste de travail d'un utilisateur qui le lance à partir de RStudio, ou en ligne, sur un serveur dédié exécutant Shiny Server¹.

De façon basique, un formulaire permet de saisir les arguments d'un fonction et une fenêtre de visualisation d'afficher les résultats du calcul.

L'utilisation d'une application Shiny rend très simple l'exécution du code, y compris pour des utilisateurs étrangers à R, mais limite évidemment les possibilités.

7.1 Première application

Dans RStudio, créer une application avec le menu “File > New File > Shiny Web App...”, saisir le nom de l'application “MonAppShiny” et sélectionner le dossier où la placer.

Le nom de l'application a servi à créer un dossier qu'il faut maintenant transformer en projet (menu des projets en haut à droite de RStudio, “New Project > Existing Directory”, sélectionner le dossier de l'application).

Le fichier de l'application nommé `app.R` contient deux fonctions : `ui()` qui définit l'interface graphique et `server()` qui contient le code R à exécuter. L'application peut être lancée en cliquant sur “Run App” dans la fenêtre du code.

La correspondance entre la fenêtre affichée (figure 7.1) et le code de la fonction `ui()` est simple à voir :

- le titre de l'application est affiché par la fonction `titlePanel()` ;

¹<https://rstudio.com/products/shiny/download-server/>

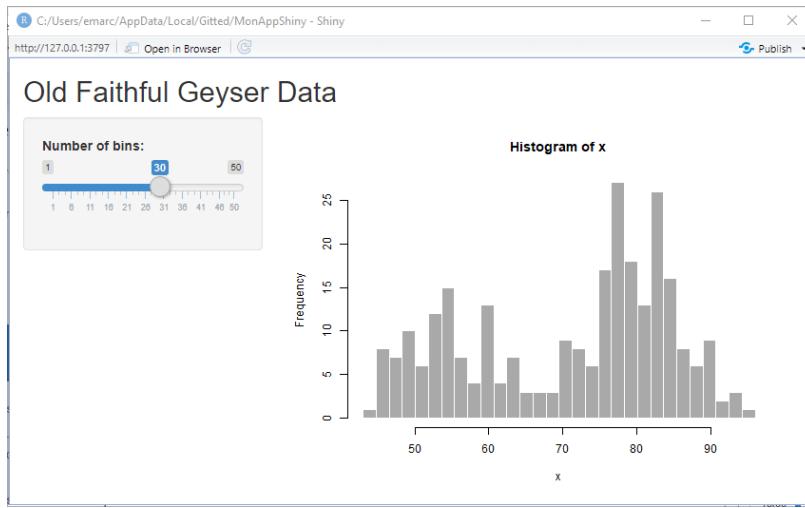


FIG. 7.1 : Application Shiny *Old Faithful Geyser Data*.

- le curseur qui fixe le nombre de barres de l’histogramme est créé par `sliderInput()` ;
- la fonction `sidebarLayout()` fixe la disposition des éléments de la page, `sidebarPanel` pour les contrôles de saisie et `mainPanel()` pour l’affichage du résultat.

Le résultat est affiché par la fonction `plotOutput()` dont l’argument est le nom d’un élément de `output`, la variable remplie par la fonction `server()`.

Toute modification d’un élément de l’interface, précisément d’un élément affiché par une fonction dont le nom se termine par `Input()` (il en existe pour tous les types d’entrées, par exemple `textInput()`) de **Shiny** provoque l’exécution de `server()` et la mise à jour des éléments de `output`.

7.2 Application plus élaborée

7.2.1 Méthode de travail

Une application est créée en choisissant :

- une disposition de la fenêtre (*layout*) ;
- les contrôles de saisie des paramètres (*input*) ;
- les contrôles d’affichage des résultats (*output*) .

Le code pour traiter les entrées et produire les sorties est ensuite écrit dans `server()`.

Le tutoriel de RStudio² est très détaillé et doit être utilisé pour aller plus loin.

²<https://shiny.rstudio.com/tutorial/>

7.2.2 Exemple

Cette application simple utilise le package **scholar** pour interroger Google Scholar et obtenir les données bibliométriques d'un auteur à partir de son identifiant.

Le fichier app.R contient tout le code et est construit progressivement ici. L'application complète, avec des sorties graphiques en plus de sa version simplifiée présentée ici est disponible sur GitHub³.

Le début du code consiste à préparer l'exécution de l'application en chargeant les packages nécessaires :

```
# Prepare the application #####
# Load packages
library("shiny")
library("tidyverse")
```

Le code de l'application complète intègre une fonction pour installer les packages manquants, à n'exécuter que quand l'application est exécutée sur un poste de travail (sur un serveur, la gestion des packages n'est pas du ressort de l'application).

L'interface utilisateur est la suivante :

```
# UI #####
ui <- fluidPage(
  # Application title
  titlePanel("Bibliometrics"),

  sidebarLayout(
    sidebarPanel(
      helpText("Enter the Google Scholar ID of an author."),
     textInput("AuthorID", "Google Scholar ID", "4iLBmbUAAAJ"),
      # End of input
      br(),
      # Display author's name and h
      uiOutput("name"),
      uiOutput("h")
    ),
    # Show plots in the main panel
    mainPanel(
      plotOutput("network"),
      plotOutput("citations")
    )
  )
)
```

La fenêtre de l'application est fluide, c'est-à-dire qu'elle se réorganise seule quand sa taille varie, et est composée d'un panneau latéral (pour la saisie et l'affichage de texte) et d'un panneau principal, pour l'affichage de graphiques.

Les éléments du panneau latéral sont :

- un texte d'aide : helpText();

³<https://github.com/EricMarcon/bibliometrics>

- un champ de texte à saisir, `textInput()`, dont les arguments sont le nom, le texte affiché, et la valeur par défaut (l'identifiant d'un auteur) ;
- un saut de ligne : `br()` ;
- des contrôles de sortie au format HTML : `uiOutput()`, dont l'argument unique est le nom.

Le panneau principal contient deux contrôles de sortie graphiques, `plotOutput()` dont l'argument est aussi le nom.

Le code à exécuter pour traiter les entrées et produire les sorties est dans la fonction `server()`.

```
# Server logic #####
server <- function(input, output) {
  # Run the get_profile function only once #####
  # Store the author profile
  AuthorProfile <- reactiveVal()
  # Update it when input$AuthorID is changed
  observeEvent(input$AuthorID,
    AuthorProfile(get_profile(input$AuthorID)))

  # Output #####
  output$name <- renderUI({
    h2(AuthorProfile()$name)
  })

  output$h <-
    renderUI({
      a(href = paste0(
        "https://scholar.google.com/citations?user=",
        input$AuthorID),
        paste("h index:", AuthorProfile()$h_index),
        target = "_blank"
      )
    })

  output$citations <- renderPlot({
    get_citation_history(input$AuthorID) %>%
      ggplot(aes(year, cites)) +
      geom_segment(aes(xend = year, yend = 0),
                   size = 1,
                   color =
                     'darkgrey') +
      geom_point(size = 3, color = 'firebrick') +
      labs(title = "Citations per year",
           caption = "Source: Google Scholar")
  })

  output$network <- renderPlot({
    ggplot() + geom_blank()
  })
}
```

Les informations nécessaires aux champs de sortie `$name` et `$h` (nom de l'auteur et indice h) sont obtenus par la fonction `get_profile()` du package **scholar**. Cette fonction interroge la page web Google Scholar de l'auteur et extrait les valeurs du résultat : c'est une traitement lourd, qu'il vaut mieux n'exécuter qu'une seule fois plutôt que deux, dans les fonctions `renderUI()` chargées de calculer les valeurs de `output$h` et `output$name`.

Le code le plus simple pour le faire serait le suivant :

```
# Run the get_profile function only once ##### Store the
# author profile
AuthorProfile <- get_profile(input$AuthorID)
```

La difficulté de la programmation d'une application Shiny est que tout calcul se référant à un élément de l'interface d'entrée doit être *réactif*. Si ce dernier code était exécuté, le message d'erreur suivant apparaît : “Operation not allowed without an active reactive context. (You tried to do something that can only be done from inside a reactive expression or observer.)”.

En pratique, l'exécution du code est lancée par la modification d'un contrôle d'entrée (ici : `input$AuthorID`). Le code faisant référence à un de ces contrôles doit être en permanence en attente d'une modification : il doit donc placé dans des fonctions particulières comme `renderPlot` dans l'application *Old Faithful Geyser Data* ou `renderUI()` ici. Le code suivant s'exécuterait sans erreur :

```
# Output #####
output$name <- renderUI({
  AuthorProfile <- get_profile(input$AuthorID)
  h2(AuthorProfile$name)
})
```

L'appel à la valeur du contrôle `input$AuthorID` a bien lieu dans une fonction réactive (mais `get_profile()` devrait être utilisé une deuxième fois dans le calcul de `output$h`, ce que nous voulons éviter). La fonction `h2(AuthorProfile$name)` produit du code HTML, un paragraphe de titre de niveau 2 dont la valeur est passée en argument.

Toutes les fonctions dont le nom commence par `render` dans le package **shiny** sont réactives, et chacune est destinée à produire un type de sortie différent, par exemple du texte (`renderText()`) ou du code HTML (`renderUI()`).

Si du code est nécessaire pour calculer des variables communes à plusieurs contrôles de sortie (`output$name` et `output$h`), il doit lui-même être réactif. Deux fonctions sont très utiles :

- `observeEvent()` surveille les changements d'un contrôle d'entrée et exécute du code quand ils se produisent ;
- `reactiveVal()` permet de définir une variable réactive, qui sera modifiée par le code de `observeEvent()` et entraînera à son tour l'exécution d'autres fonctions réactives qui utilisent sa valeur.

Le code optimal crée donc une variable réactive pour y stocker le résultat de l'interrogation de Google Scholar :

```
# Store the author profile
AuthorProfile <- reactiveVal()
```

La variable réactive est vide à ce stade. Son utilisation est ensuite celle d'une fonction : `AuthorProfile(x)` lui attribue la valeur `x` et `AuthorProfile()`, sans argument, renvoie sa valeur. La fonction `observeEvent()` est déclenchée quand `input$AuthorID` est modifié et exécute le code passé en deuxième argument, ici la mise à jour de `AuthorProfile`.

```
# Update it when input$AuthorID is changed
observeEvent(input$AuthorID, AuthorProfile(get_profile(input$AuthorID)))
```

Enfin, les fonctions `renderUI()` qui fournissent les valeurs des contrôles de sortie utilisent la valeur de `AuthorProfile` :

```
# Output ####
output$name <- renderUI({
  h2(AuthorProfile()$name)
})
```

Remarquer les parenthèses de `AuthorProfile()`, variable réactive, par opposition à la syntaxe `AuthorProfile$name` pour une variable classique.

La valeur de `output$h` est un lien internet, `<a href=...` en HTML, écrit par la fonction `a()` du package **htmltools** utilisé par `renderUI()`.

```
output$h <- renderUI({
  a(href = paste0("https://scholar.google.com/citations?user=",
    input$AuthorID), paste("h index:", AuthorProfile()$h_index),
  target = "_blank")
})
```

Le lien est vers la page Google Scholar de l'auteur. La valeur affichée est son indice `h`. L'argument `target = "_blank"` indique que le lien doit être ouvert dans une nouvelle fenêtre du navigateur.

Le graphique `output$citations` est créé par la fonction réactive `renderPlot()`. Les données fournies par la fonction `get_citation_history()` de **scholar** (qui interroge l'API de Google Scholar) sont traitées par `ggplot()`.

Enfin, le graphique `output$network` est un graphique vide dans cette version simplifiée de l'application.

L'application complète reprend ce code en y ajoutant le traitement des erreurs dans le cas où le code de l'auteur n'existe pas sur Google Scholar et le graphique du réseau des co-auteurs.

7.3 Hébergement

Une application Shiny n'est pas forcément hébergée par un serveur web : elle peut être exécutée sur les postes de travail des utilisateurs s'ils disposent de R.

Pour un usage plus large, un serveur dédié est nécessaire. [Shinyapps.io⁴](https://www.shinyapps.io/) est un service de RStudio qui permet d'héberger gratuitement 5 applications Shiny avec un temps de fonctionnement maximal de 5 heures par mois.

⁴<https://www.shinyapps.io/>

Il faut tout d'abord ouvrir un compte sur le site, de préférence avec ses identifiants GitHub. Pour permettre la gestion des applications en ligne directement depuis RStudio, il faut installer ensuite le package **rsconnect** et le paramétrier :

```
rsconnect::setAccountInfo(name = "prenom.nom", token = "xxx",
                           secret = "<SECRET>")
```

Le code exact, avec le nom d'utilisateur et le jeton à utiliser, sont affichés sur la page d'accueil de Shinyapps.io : cliquer sur “Show Secret”, copier le code et le coller dans la console de RStudio pour l'exécuter. Un bouton “Publish” est disponible juste à droite du bouton “Run App”. Cliquer dessus et valider la publication (figure 7.2).

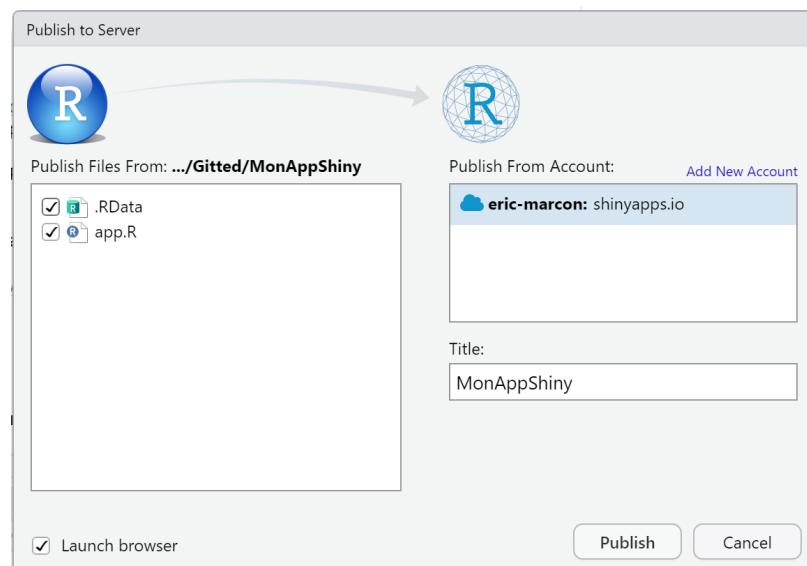


FIG. 7.2 : Publication de l'application Shiny sur Shinyapps.io.

L'application est maintenant accessible à l'adresse <https://prenom-nom.shinyapps.io/MonAppShiny/>

L'application “Bibliometrics” ne fonctionne pas sur Shinyapps.io parce que la façon dont le package **Scholar** interroge Google Scholar n'est pas supportée. La plupart des applications Shiny fonctionnent sans difficulté, tant qu'elles ne nécessitent pas de fonctionnalités réseau complexes.

ENSEIGNER AVEC R

R, RStudio et GitHub fournissent des outils pour enseigner.

Le package **learnr** permet de réaliser des tutoriels interactifs.

On verra aussi comment utiliser les salles de classe GitHub (*GitHub Classrooms*) qui permettent de diffuser à une classe (une liste d'étudiants disposant d'un compte GitHub) un modèle de dépôt (un début de projet R) que chaque étudiant devra développer et publier. Les outils de la salle de classe permettent d'évaluer le travail fourni assez simplement.

8.1 learnr

learnr permet de rendre interactifs les bouts de code de n'importe quel document produit par R Markdown en HTML, en les transformant en applications Shiny. La documentation sur le site de RStudio¹ est très claire et ne sera pas reprise ici : nous verrons seulement comment commencer et comment diffuser les tutoriels.

8.1.1 Premier tutoriel

Utiliser comme pour tous les documents le menu “File > New File > RMarkdown...” et créer un nouveau document à partir d'un modèle “Interactive Tutorial”. L'assistant crée un dossier du nom choisi, à transformer en projet R et passer sous contrôle de source, comme pour tous les documents vus précédemment (voir section 4.3.2).

Pour exécuter le tutoriel, cliquer sur le bouton “Run Document” qui se trouve à place habituelle du bouton “Tricoter”.

¹<https://rstudio.github.io/learnr/>

Les tutoriels peuvent inclure des exercices, qui sont des bouts de code avec l'option `exercise=TRUE`. Ces exercices sont affichés sous la forme d'une fenêtre de code modifiable et exécutable par l'utilisateur. Des indices peuvent être donnés², un bouton ajouté pour afficher la solution, une limite de temps peut être fixée³, et le code comme son résultat peuvent être comparés à une valeur attendue⁴.

Des quizz⁵ peuvent être ajoutés, sous la forme de questionnaires à choix multiples ou uniques.

La progression de l'utilisateur dans le tutoriel (code saisi, réponses aux questions...) est sauvegardée par `learnr` sur le poste de travail. Un tutoriel peut être arrêté puis repris sans perte de données. En revanche, il n'y a pas de moyen simple de récupérer ces données pour une évaluation par le formateur par exemple.

8.1.2 Diffusion

Les tutoriels peuvent être diffusés en copiant les fichiers ou en indiquant aux utilisateurs de cloner les projets GitHub qui les contiennent.

Ils peuvent aussi être hébergés sur Shinyapps.io (voir section 7.3).

Enfin, ils peuvent être inclus dans un package⁶.

8.2 GitHub Classrooms

GitHub Classrooms permet de diffuser à un public étudiant des dépôts GitHub à modifier et de contrôler le résultat. Les applications sont aussi bien l'apprentissage de R que la production de documents, pour un travail personnel ou un examen par exemple.

8.2.1 Inscription

Pour commencer à utiliser l'outil, il faut ouvrir un compte. Sur le site de GitHub Classrooms⁷, cliquer sur "Sign in" et utiliser son compte GitHub pour s'authentifier.

8.2.2 Organisations

L'étape suivant consiste à créer une organisation GitHub. Une organisation GitHub contient essentiellement des membres (titulaires d'un compte GitHub) et des dépôts accessibles à l'adresse <https://github.com/Organisation/Depot>.

²https://rstudio.github.io/learnr/exercises.html#Hints_and_Solutions

³https://rstudio.github.io/learnr/exercises.html#Time_Limits

⁴https://rstudio.github.io/learnr/exercises.html#Exercise_Checking

⁵<https://rstudio.github.io/learnr/questions.html>

⁶https://rstudio.github.io/learnr/publishing.html#R_Package

⁷<https://classroom.github.com/>

La façon la plus simple de travailler consiste à créer une organisation par cours mais d'autres approches sont possibles dans des structures utilisant intensivement l'outil. L'organisation créée pour l'exemple est ici "Cours-R"⁸.

Une adresse de messagerie est nécessaire (utiliser la même que celle de son compte GitHub) et l'organisation doit être déclarée comme appartenant à son compte personnel.

Si l'organisation n'est pas visible sur la page de GitHub Classrooms, cliquer sur "Grant us access".

8.2.3 Nouvelle salle de classe

Une salle de classe (*classroom*) est peuplée d'étudiants qui recevront des tâches (*assignments*) à exécuter.

Cliquer sur *New Classroom*. Sélectionner l'organisation en charge de l'administration de la salle de classe.

Saisir le nom de la salle de classe : une bonne pratique est de la préfixer par le nom du cours et d'ajouter le nom de la session, par exemple "Cours-R-2020-EdGuyane".

Ne pas ajouter de collaborateurs (ce sera possible plus tard), et saisir éventuellement la liste des étudiants (un nom par ligne, possible plus tard aussi). La classe est créée.

Toutes les salles de classe sont visibles depuis la page d'accueil de GitHub Classrooms⁹. Cliquer sur un nom pour en ouvrir une. Le bouton "Settings" permet de changer son nom ou de la supprimer. Le bouton "TAs and Admins" permet d'ajouter des collaborateurs, c'est-à-dire d'autres utilisateurs GitHub qui pourront administrer la salle de classe.

Le bouton "Students" permet d'ajouter des étudiants. La liste de nom est libre, sans format obligatoire. Cliquer sur "Create Roster" pour l'activer. Les noms doivent ensuite être liés à des comptes GitHub : ce travail peut être fait par l'administrateur ou par les étudiants eux-mêmes quand ils recevront la première tâche à effectuer. Chaque étudiant doit avoir un compte sur GitHub.

8.2.4 Préparer un modèle de dépôt

Une tâche est un dépôt GitHub à modifier. Par exemple¹⁰, créer un dépôt contenant un projet R avec un fichier Markdown décrivant le travail à faire et éventuellement une partie du code nécessaire pour y parvenir, les autres fichiers du modèle R Markdown utilisé et un fichier de données.

Ouvrir les propriétés du dépôt sur GitHub et cocher la case *Template Repository* pour en faire un modèle.

⁸<https://github.com/Cours-R>

⁹<https://classroom.github.com/classrooms>

¹⁰<https://github.com/EricMarcon/Cours-R-Memo/settings>

Assigner une tâche

Ouvrir une salle de classe et cliquer sur “New Assignment”.

Saisir un titre explicite pour les étudiants, une date limite optionnelle et choisir “Individual Assignment”.

Par défaut, le nom de la tâche sert de préfixe pour le nom des dépôts des étudiants mais il peut être remplacé par un préfixe choisi. Quand les étudiants rendront leur travail, tous les dépôts de toutes les tâches seront stockés dans l’organisation.

Le dépôt crée sur le compte de chaque étudiant peut être privé ou public, selon que l’on souhaite que les étudiants puissent voir le travail des autres ou non. Donner le droit d’administration et rendre le site public si les étudiants doivent pouvoir activer les pages GitHub pour présenter le résultat de leur travail. Cliquer sur “Continue”.

Sélectionner le dépôt modèle (*starter code*) puis cliquer sur “Continue” puis “Create Assignment”.

La nouvelle tâche est créée. Elle est associée à un lien d’invitation qu’il faut copier et envoyer aux étudiants. Quand ils cliqueront sur le lien, ils atteindront une page GitHub qui leur permettra d’associer leur compte à un nom de la liste (aucun contrôle n’est possible : le premier connecté peut s’associer à n’importe quel nom). Ils pourront ensuite créer un nouveau projet RStudio à partir du dépôt GitHub créé automatiquement par GitHub Classrooms, modifier ce projet selon les consignes de travail et le pousser sur GitHub. Le dépôt se trouve sur le compte de l’organisation à laquelle la classe est reliée, et est suffixé par l’identifiant GitHub de l’étudiant.

Contrôler le travail des étudiants

Il est possible d’afficher chaque dépôt créé par les étudiants à partir de la page de la tâche sur GitHub Classrooms. Si le travail à produire est un document rédigé, demander aux étudiants de le placer dans les pages GitHub du dépôt pour le lire directement en ligne.

L’assistant GitHub Classrooms¹¹ permet de télécharger en une fois tous les dépôts des étudiants pour les corriger sur son poste de travail.

¹¹<https://classroom.github.com/assistant>

CONCLUSION

L'environnement de travail de R et RStudio permet de produire tous types de documents avec un langage unique.

L'objectif de reproductibilité des résultats est atteint en intégrant les traitements statistiques et la rédaction. Le travail collaboratif est permis par l'utilisation systématique du contrôle de source et de GitHub. La présentation des résultats est assurée par les pages GitHub et des modèles de documents couvrant la majorité des besoins.

Pour les pauses, R fournit même quelques jeux dans le package **fun**, dont le célèbre démineur :

```
# Installation du package
install.packages("fun")
# Ouverture d'une fenêtre X et exécution
if (interactive()) {
  if (.Platform$OS.type == "windows")
    x11() else x11(type = "Xlib")
  fun::mine_sweeper()
}
```

Ce document n'a pas pour objectif d'être exhaustif sur les possibilités de R mais plutôt de présenter une méthode de travail et des moyens simples de l'appliquer rapidement. On se reportera aux ouvrages plus détaillés cités dans le texte pour approfondir tel ou tel point.

Ce document est mis à jour régulièrement en fonction de l'évolution des outils disponibles.

BIBLIOGRAPHIE

- BARNIER, J. (2020). *Introduction à R et Au Tidyverse*. URL : <https://juba.github.io/tidyverse/> (cf. p. 15).
- GANDRUD, C. (2015). *Reproducible Research with R and RStudio*. 2^e éd. Chapman and Hall/CRC (cf. p. ix).
- GILLESPIE, C. et R. LOVELACE (2016). *Efficient R Programming*. O'Reilly Media. URL : <https://csgillespie.github.io/efficientR/> (cf. p. 15).
- KNUTH, D. E. (1^{er} jan. 1984). «Literate Programming». In : *The Computer Journal* 27.2, p. 97-111. DOI : [10.1093/comjnl/27.2.97](https://doi.org/10.1093/comjnl/27.2.97). URL : <https://doi.org/10.1093/comjnl/27.2.97> (visité le 11/06/2021) (cf. p. 77).
- WICKHAM, H. (2010). «A Layered Grammar of Graphics». In : *Journal of Computational and Graphical Statistics* 19.1, p. 3-28. URL : <http://vita.had.co.nz/papers/layered-grammar.pdf> (cf. p. 21).
- (2014). *Advanced R*. Chapman and Hall/CRC. URL : <http://adv-r.had.co.nz/> (cf. p. 15).
- (2015). *R Packages*. 1st. O'Reilly Media, Inc. (cf. p. 117).
- (2017). *Ggplot2 : Elegant Graphics for Data Analysis*. 2^e éd. Springer. DOI : [10.1007/978-3-319-24277-4](https://doi.org/10.1007/978-3-319-24277-4). URL : <http://had.co.nz/ggplot2/book> (cf. p. 22).
- WICKHAM, H. et G. GROLEMUND (2016). *R for Data Science*. O'Reilly Media. URL : <http://r4ds.had.co.nz/> (cf. p. 15, 22).
- XIE, Y. (2015). *Dynamic Documents with R and Knitr*. 2^e éd. Boca Raton, Florida : Chapman and Hall/CRC. URL : <https://yihui.name/knitr/> (cf. p. 78).
- XIE, Y., J. ALLAIRE et G. GROLEMUND (2018). *R Markdown : The Definitive Guide*. Boca Raton, Florida : Chapman and Hall/CRC. URL : <https://bookdown.org/yihui/rmarkdown> (cf. p. 78).

TABLE DES FIGURES

| | | |
|------|---|-----|
| 1.1 | Activation du droit de modifier la bibliothèque système sous Windows. | 4 |
| 1.2 | Dossier pour les projets sous contrôle de source, sous Windows. | 6 |
| 1.3 | Activation du droit de modifier la bibliothèque système sous Windows. | 6 |
| 2.1 | Prix des diamants en fonction de leur poids. Démonstration du code de ggplot2 combiné au traitement de données du tidyverse. | 23 |
| 2.2 | Temps d'exécution en parallèle | 39 |
| 3.1 | Capture d'écran du terminal de RStudio. La commande <code>git status</code> supposée décrire l'état du dépôt renvoie une erreur si le projet R n'est pas sous contrôle de source. | 55 |
| 3.2 | Activation du contrôle de source dans le menu “Tools > Project Options...” | 56 |
| 3.3 | Fichiers du projet, pas encore pris en compte par <code>git</code> | 57 |
| 3.4 | Fenêtre de validation des modifications prises en compte. | 58 |
| 3.5 | Fenêtre de demande d'identification. | 58 |
| 3.6 | Les trois arbres de <code>git</code> . Source : https://rogerdudler.github.io/git-guide/index.fr.html | 59 |
| 3.7 | Création d'un dépôt sur GitHub. | 60 |
| 3.8 | Identification HTTPS sur GitHub. | 62 |
| 3.9 | Clonage d'un dépôt à partir de <i>GitHub</i> | 63 |
| 3.10 | Attribution des droits d'accès sur GitHub. | 64 |
| 3.11 | Différences entre le répertoire de travail et la tête. | 65 |
| 3.12 | Historique des validations dans le dépôt. | 66 |
| 4.1 | Nouveau document Markdown à partir d'un modèle. | 79 |
| 4.2 | Titre de la figure | 82 |
| 4.3 | Titre avec <i>italique</i> , maths ($\sqrt{\pi}$) et renvoi vers la figure 4.2 | 83 |
| 4.4 | Copie de l'adresse d'un dépôt à cloner sur GitHub. | 96 |
| 4.5 | Collage de l'adresse du dépôt à cloner. | 97 |
| 4.6 | Composant <code>demo</code> dans Academic. | 101 |
| 4.7 | Composant <code>about</code> dans Academic. | 102 |
| 4.8 | Composant <code>skills</code> dans Academic. | 102 |
| 4.9 | Composant <code>experience</code> dans Academic. | 103 |
| 4.10 | Courbes de niveau du volcan Maunga Whau, code fourni en exemple de l'aide de la fonction <code>image()</code> | 111 |

TABLE DES FIGURES

| | |
|--|-----|
| 7.1 Application Shiny <i>Old Faithful Geyser Data</i> | 164 |
| 7.2 Publication de l'application Shiny sur Shinyapps.io. | 169 |

Résumé Cet ouvrage propose une organisation du travail autour de R et RStudio pour, au-delà des statistiques, rédiger des documents efficacement avec R Markdown, aux formats variés (mémos, articles scientifiques, mémoires d'étudiants, livres, diaporamas), créer son site web et des applications R en ligne (Shiny), produire des packages et utiliser R pour l'enseignement.