

Aide-mémoire : Analyses de données & Cartographie sur R

Alexis Mérot

Modifié le : 2020-08-06

Table des matières

Introduction	5
1 R Markdown, Bookdown & Blogdown	7
1.1 Pourquoi R Markdown ?	7
Liste de ressources Internet utiles	8
2 Extraction/importation, nettoyage & manipulation des données	9
2.1 Extraction de données provenant d'un PDF	9
2.2 Récupération de données provenant du web : le <i>web scraping</i> . .	9
Liste de ressources Internet utiles	9
3 Statistique	11
3.1 Quelques notions clés	11
3.2 Statistique fréquentiste	11
Liste de ressources Internet utiles	11
3.3 Statistique bayésienne	11
Liste de ressources Internet utiles	12
4 Visualisation des données : la <i>Dataviz</i>	13
5 Les Systèmes d'Information Géographiques	15
5.1 Qu'est-ce qu'un SIG ?	15
5.2 R en tant que SIG	16
Liste de ressources Internet utiles	16

Introduction



Cet aide-mémoire n'est pour l'instant qu'un brouillon. Il est donc loin d'être complet et l'écriture est pour l'instant très succincte.

Ce projet est un ensemble de notes écrites en R Markdown (Allaire et al., 2020) et via le package **bookdown** (<https://github.com/rstudio/bookdown>). Ces notes s'accumuleront au fur et à mesure de mon apprentissage des différents outils et concepts dont j'ai besoin pour les analyses de données et la programmation. Cela me permet de les comprendre, les mémoriser, ainsi que de les partager.

L'aide-mémoire s'insérera peut-être dans un autre plus gros projet : la création d'un blog répertoriant tous mes projets et mon CV. Il commencera certainement lorsque je pourrai démarrer la lecture de la documentation de l'excellent package **blogdown** (<https://bookdown.org/yihui/blogdown/>).

Cet aide-mémoire intégrera les notions théoriques indispensables en statistique et en cartographie, ainsi que les outils proposés par R (et si besoin d'autres langages) pour la mise en pratique à travers d'exemples. Étant principalement intéressé par la Biologie de la conservation et globalement l'Écologie, les exemples se focaliseront pour la plupart sur des données en lien à ces domaines.

Toutes les sources qui m'ont été utiles pour acquérir ces connaissances seront accessibles dans les références bibliographiques ou dans les ressources Internet à la fin des chapitres.

Chapitre 1

R Markdown, Bookdown & Blogdown



Work In Progress

1.1 Pourquoi R Markdown ?

R Markdown est un format de fichier (à l'extension **.Rmd**) fournissant un cadre de création pour faire des rapports scientifiques automatisés. Ces documents peuvent ainsi être totalement reproductibles et plusieurs formats de rendu final (statiques ou dynamiques) sont supportés.

Le fichier est écrit via le langage Markdown et des sections de code R peuvent y être insérées facilement (ainsi que du code écrit via d'autres langages tels que Python ou SQL). Cela offre une syntaxe facile à lire et à écrire tout en permettant de générer un rapport structuré et élégant.

Pour que cela fonctionne, R Markdown est lié à deux packages : **knitr** et le convertisseur universel de document **pandoc** (Fig. 1.1).

Le package **knitr** permet la création, à partir du fichier **.Rmd**, d'un fichier au format **md** contenant le code et sa sortie. Ce fichier est alors converti dans le format de rendu final voulu via **pandoc** (**.html**, **.pdf**, etc).

Toutes mes notes seront donc écrites via R Markdown, et cette section intégrera toutes les astuces intéressantes que je rencontre au fur et à mesure des besoins.

Pour ne pas paraphraser tout le livre de Yihui Xie, je vous invite à lire son excellent guide gratuit : <https://bookdown.org/yihui/rmarkdown/>.



FIGURE 1.1 – Source : <https://rmarkdown.rstudio.com/lesson-2.html>

Liste de ressources Internet utiles

- R Markdown :
 - Vue d'ensemble de R Markdown
 - Cours sur la communication avec R Markdown
 - Comment utiliser R Markdown comme base pour le développement de packages bien organisés
 - Quelques trucs et astuces sur R Markdown
 - Comment donner du *peps* à mon document RMD
 - Un autre guide de R Markdown
 - Guide complet de R Markdown
 - Nouveau guide de R Markdown (*en cours d'écriture*)
 - Création d'un template R Markdown
 - Bookdown :
 - Site officiel de Bookdown
 - Guide complet de Bookdown
 - Extension à Bookdown : **bookdownplus**
 - Guide en français de Bookdown
 - Introduction en français à Bookdown
 - Blogdown :
 - Guide complet sur Blogdown
 - Court tutoriel d'introduction sur R Markdown, Bookdown et Blogdown
 - Guide pour le package **knitr**
 - Options valables pour les *chunks* et le package **knitr**
-

Chapitre 2

Extraction/importation, nettoyage & manipulation des données



Work In Progress

2.1 Extraction de données provenant d'un PDF

2.2 Récupération de données provenant du web :
le *web scraping*

Liste de ressources Internet utiles

Chapitre 3

Statistique



Work In Progress

3.1 Quelques notions clés

Concepts à comprendre :

- Théorie des probabilités
- Variable aléatoire réelle
- Fonction de répartition (ou fonction de distribution cumulative) d'une variable aléatoire
- Fonction de densité ou densité de probabilité

3.2 Statistique fréquentiste

Liste de ressources Internet utiles

3.3 Statistique bayésienne

Liste de ressources Internet utiles

Chapitre 4

Visualisation des données : la *Dataviz*



Work In Progress

Chapitre 5

Les Systèmes d'Information Géographiques



Work In Progress

5.1 Qu'est-ce qu'un SIG ?

Un Système d'Information Géographique est, comme tout Système d'Information¹, un ensemble organisé de ressources ayant pour fonction de collecter, stocker, traiter et diffuser des informations². Ici, ces informations (généralement informatisées) sont des données géospatiales stockées sous forme de couches d'informations superposées et reliées les unes aux autres par un référentiel cartographique³. Les SIG sont devenus des outils essentiels dans de nombreux domaines tels que l'écologie, la médecine ou la sociologie.

Pour aider les utilisateurs au traitement des données géospatiales, il existe de performants et très utilisés logiciels gratuits ou payants tels que QGis ou ArcGIS. Ces logiciels offrent une approche graphique à la lecture, l'écriture, la manipulation et la visualisation des données. Ceci peut néanmoins limiter la reproductibilité et l'automatisation des projets. Pour remédier à cela, de nombreux langages de programmation peuvent être utilisés pour écrire et partager des scripts. Parmi les plus utilisés, il y a Python (qui est notamment utilisé pour la conception de plugins dans les logiciels de SIG) et R (dont les scripts sont maintenant exécutables dans QGis). En plus de cela, l'approche en lignes de

1. Cf. le cours sur Openclassroom.

2. Cf. la page Wikipédia sur le Système d'Information Géographique.

3. Fond de carte représentant un territoire géographique sur lequel peuvent s'appuyer de nouvelles données cartographiques.

commande permet de se libérer de certaines contraintes imposées par ces logiciels ainsi que d'avoir plus de contrôle sur ce que l'ont fait (même si ces logiciels sont de plus en plus performants).

5.2 R en tant que SIG

Afin d'avoir un bon aperçu et une bonne base sur l'utilisation de R en tant que SIG, je vous invite à lire le livre *Geocomputation with R* de Lovelace et al. (2019). Ce livre est mis à jour régulièrement et est mis à disposition gratuitement à cette adresse : <https://geocompr.robinlovelace.net/>.

Ayant commencé à apprendre les analyses statistiques avec R, c'est naturellement que je me suis tourné vers ce langage pour la cartographie et l'analyse des données géospatiales. En effet, la communauté de R a créé de performants et magnifiques packages de cartographie et de géocalcul libres, gratuits et bien documentés. Je m'intéresserai donc de plus près à ce qu'offre par exemple Python lorsque j'aurai maîtrisé suffisamment R. Un autre langage élégant et très récent qui est à regarder de très près est Julia, qui offrira certainement des packages rapides et performants au fur et à mesure de sa maturité. Par ailleurs, même si des programmes manqueraient à R ou si d'autres langages possèdent des programmes plus adaptés pour certaines tâches, des packages R offrent la possibilité d'en faciliter l'accès. Par exemple, les packages tels que **Rcpp** et **Reticulate** permettent l'utilisation de programmes écrits respectivement en C++ et Python.

D'autres caractéristiques intéressantes de R sont sa flexibilité et sa constante évolution. Par exemple, il offre maintenant la possibilité de faire facilement des applications web et des cartes interactives notamment via les packages **Shiny** et **Leaflet**. Il offre par la même occasion divers outils d'analyses avancées, de modélisation et de visualisation qui sont mis à jour et améliorés régulièrement.

Pour plus d'informations concernant les atouts de R en tant que SIG ainsi qu'un bref aperçu de l'utilité des autres langages tels que Python, Java et C++, je vous invite à lire le chapitre Why use R for geocomputation du livre de Lovelace et al. (2019).

Liste de ressources Internet utiles

- Guide sur les analyses de données géographiques, leur visualisation et leur modélisation sur R
- Introduction à l'utilisation des packages de cartographie sur R
- Introduction au package **sf**
- Édition interactive de cartes avec **mapedit**
- introduction à l'utilisation de R comme un SIG

- Introduction pour créer des cartes avec R
- Introduction en français pour créer des cartes avec R
- Introduction en français sur le package **rgeoapi**
- Zoomer sur une carte avec R
- Tracer des cartes avec **ggplot2** via des fichiers *shapefiles*
- Tutoriel pour dessiner des cartes avec R, **sf** et **ggplot2**
- Cartes interactives avec **mapview**
- Cartes interactives avec **leaflet**
- Guide pour faire des cartes en 3D à partir d'une imagerie satellite
- Utilisation du package **rayshader** pour la création de cartes en 2D et 3D
- Manipulation et visualisation de données LiDAR pour la foresterie avec **lidr**
- Blog français contenant divers tutoriels sur la SIG et QGis

Bibliographie

Allaire, J., Xie, Y., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., Wickham, H., Cheng, J., Chang, W., and Iannone, R. (2020). *rmarkdown : Dynamic Documents for R*. R package version 2.3.3.

Lovelace, R., Nowosad, J., and Muenchow, J. (2019). *Geocomputation with R*. Chapman & Hall/CRC The R Series. CRC Press.