

Organisation du travail autour de R

Eric Marcon

5 juillet 2018

Organisation
du travail
d'Eric M.

Eric Marcon

L'histoire

Organiser les
données

Analyser les
données

Ecrire

Organisation

Limites

Avantages

L'histoire

Contexte

Laboratoire de recherche.

Equipes de chercheurs - étudiants - techniciens.

Production de données, méthodes et documents.



Eric Marcon

L'histoire

Organiser les données

Analyser les données

Ecrire

Organisation

Limites

Avantages

Organisation spontanée

Organisation
du travail
dans le P

Eric Marcon

L'histoire

Organiser les
données

Analyser les
données

Ecrire

Organisation

Limites

Avantages

On pilote difficilement une équipe de chercheurs.

Haut niveau technique.

Tendance à diverger.

Objectifs

Organisation
du travail
d'un projet de P

Eric Marcon

L'histoire

Organiser les
données

Analyser les
données

Ecrire

Organisation

Limites

Avantages

Etre plus efficace.

Echanger plus facilement.

Recherche reproductible.

Méthode

Organisation
du travail
Amont de P

Eric Marcon

L'histoire

Organiser les
données

Analyser les
données

Ecrire

Organisation

Limites

Avantages

Les outils ne font pas l'organisation.

Les objectifs sans outils non plus.

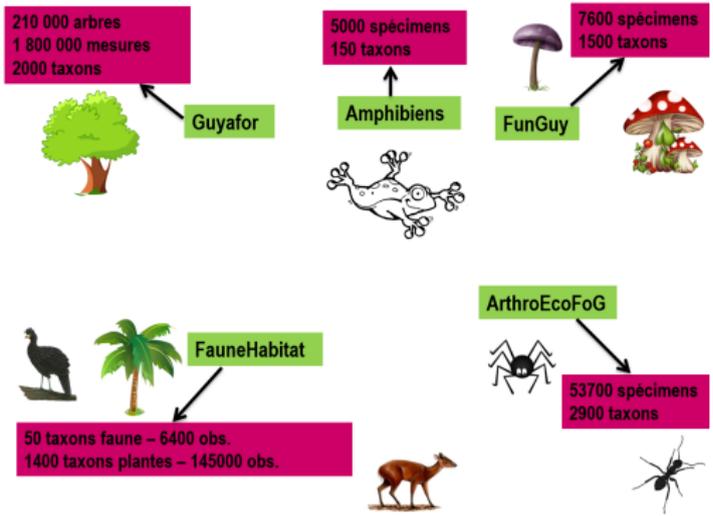
Itérations besoins ↔ outils

Organiser les données

Données précieuses

Relativement peu de données en écologie.

Eric Marcon



Prix unitaire élevé.

Choix

Présentation
Le travail
du projet de P

Eric Marcon

L'histoire

**Organiser les
données**

Analyser les
données

Ecrire

Organisation

Limites

Avantages

Données standardisées : SGBDR

Données ponctuelles :

- tableaux, format CSV
- Accessibles en ligne (partages de fichiers, HTTP).

Une ingénieure de recherche dédiée dans l'unité.

Analyser les données

Script pour :

- la reproductibilité
- la versatilité
- l'explicitation

Communauté, gratuité. . .

Code R avec commentaires

```
# Addition  
2 + 2
```

```
## [1] 4
```

Document RMarkdown avec chunks

Addition

Utiliser l'opérateur + :

```
2+2
```

```
## [1] 4
```

Organisation
du travail
amont de P

Eric Marcon

L'histoire

Organiser les
données

Analyser les
données

Ecrire

Organisation

Limites

Avantages

Ecrire

Enjeux

Organisation
du travail
dans le P

Eric Marcon

L'histoire

Organiser les
données

Analyser les
données

Ecrire

Organisation

Limites

Avantages

Beaucoup de temps passé à produire des documents.

Processus collaboratif nécessaire.

Réutilisation.

Au début

Documents Word:

- Structuration possible, mais rare,
- Limites dans le rendu final.

Eric Marcon

L'histoire

Organiser les données

Analyser les données

Ecrire

Organisation

Limites

Avantages

Statistiques spatiales en écologie forestière

Extension aux domaines rectangulaires

La démarche précédente a permis de montrer qu'un test global de la fonction de Ripley était possible, y compris pour des petits semis de points. Sa limite est que la démonstration a été faite pour un domaine carré, dans le but de montrer les propriétés asymptotiques du test en augmentant la taille du carré. Pour une application pratique, les calculs doivent être repris pour un domaine rectangulaire, noté A_{r_1, r_2} , dans lequel les tailles des côtés seront notées l_1 et l_2 . Comme l'objectif est l'application empirique, seul l'estimateur $\hat{K}_{r_1, r_2}(r)$ présente un intérêt : l'intensité du processus n'est jamais connue. La notation sera désormais le $\hat{K}(r)$.

Les étapes du raisonnement sont exactement les mêmes que pour le carré. Il est estimé par $n(A_{r_1, r_2})/(l_1 l_2)$.

La valeur du biais est :

$$E(\hat{K}(r) - K(r)) = \frac{4r^3(l_1 + l_2)}{3l_1 l_2} + \frac{r^4}{2l_1^2 l_2^2} \quad (24)$$

La variance vaut :

$$\begin{aligned} \text{Var}(\hat{K}(r)) = & 2l_1^2 l_2^2 E\left(\frac{\mathbf{1}(N(A_{r_1, r_2}) > 1)}{N(A_{r_1, r_2})(N(A_{r_1, r_2}) - 1)}(e_{r_1, r_2} - e_{r_1, l_2}^2)\right) \\ & + 4l_1^2 l_2^2 E\left(\frac{\mathbf{1}(N(A_{r_1, r_2}) > 1)(N(A_{r_1, r_2}) - 2)}{N(A_{r_1, r_2})(N(A_{r_1, r_2}) - 1)}\right) E(\hat{K}(U, r)) \\ & + l_1^2 l_2^2 e^{-\lambda l_1 l_2} (1 + \lambda l_1 l_2) (1 - e^{-\lambda l_1 l_2} - \lambda l_1 l_2 e^{-\lambda l_1 l_2})^2 r_1^2 l_2^2 \end{aligned} \quad (25)$$

Où :

Le problème traité est la non-linéarité de l'indice de Shannon par rapport aux probabilités qui entraîne un biais d'estimation. La fonction logarithmique fournit un exemple simple : l'espérance de $\ln(p_n)$ n'est pas le logarithme de l'espérance de p_n parce que la fonction ln est concave. Chaque estimateur \hat{p}_n fluctue autour de p_n mais tout p_n est mesuré. À cause de la concavité, $\ln(\hat{p}_n)$ est en moyenne inférieur à $\ln(p_n)$, cette relation est connue sous le nom d'inégalité de Jensen.¹⁰⁵ L'indice de Shannon est concave (Figure 3.10¹⁰⁵) donc son estimateur (3.20) est biaisé négativement, même sans prendre en considération les espèces non observées.

Le biais peut être évalué par simulation : 10000 tirages sont réalisés dans une loi normale d'espérance p_n choisie et d'écart-type 0.01. Le biais est la différence entre $-p_n \ln p_n$ (comm) et la moyenne des 1000 valeurs de $-p_n \ln \hat{p}_n$ (la probabilité est estimée par sa réalisation à chaque tirage). Le valeur du biais en fonction de p_n est en Figure 3.10¹⁰⁵. Le biais de l'indice de Shannon est la somme des biais pour toutes les probabilités spécifiques de la communauté étudiée, et son calcul est toujours l'objet de recherches.

Grassberger¹⁰³ a fourni la correction de référence :

$$\hat{H} = - \sum_{n=1}^{s_n} \frac{m_n}{n} \left(\ln(n) - \Psi(n_n) - \frac{(-1)^{n_n}}{n_n + 1} \right) \quad (3.42)$$

Grassberger¹⁰³ l'a perfectionné :

$$\hat{H} = - \sum_{n=1}^{s_n} \frac{m_n}{n} \left(\Psi(n) - \Psi(n_n) - (-1)^{n_n} \int_0^1 \frac{e^{-t} - 1}{1+t} dt \right) \quad (3.43)$$

FIG. 3.9 - Courbe de $-x \ln x$ entre 0 et 1.

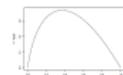


FIGURE 3.10 - Biais de $-p_n \ln p_n$.

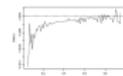


FIGURE 3.10 - Biais de $-p_n \ln p_n$.

Echange par messagerie.

Besoin individuel

Présentation
Présentation
Présentation

Eric Marcon

L'histoire

Organiser les
données

Analyser les
données

Ecrire

Organisation

Limites

Avantages

Se concentrer sur le fond :

- LaTeX plutôt que Word,
- Markdown plutôt que LaTeX.

Construire sa pensée ↔ rédiger :

- Intégrer les traitements au texte
- knitr et LaTeX puis RMarkdown.

Besoin collectif

Organisation
du travail
collectif de P

Eric Marcon

L'histoire

Organiser les
données

Analyser les
données

Ecrire

Organisation

Limites

Avantages

Ecriture en parallèle

Suivi des versions

Progrès possibles

Google Docs : collaboration.

SharePoint puis Office 365.

Overleaf:

The screenshot shows the Overleaf web editor interface. The browser address bar displays the URL: <https://www.overleaf.com/14180831kmbvtqtcq#57121282>. The user's name, Eric Marcon, is visible in the top right corner.

The interface is split into two main panes:

- Source Pane (Left):** Displays the LaTeX source code. Line 55 is `\maketitle`. Line 56 is `\begin{abstract}`. Lines 58-63 contain the abstract text:


```

58 \begin{abstract}
59 For a decade, distance-based methods have
60 been widely employed and constantly
61 improved in spatial economics.
62 These methods are a very useful tool for
63 accurately evaluating the spatial
64 distribution of economic activity. We
65 introduce a new distance-based
66 statistical measure for evaluating the
67 spatial concentration of industries.
68 The  $S_m$  function is the first relative
69 density function to be proposed in
70 economics. This tool supplements the
71 typology of distance-based methods
72 recently drawn up by \cite{Marcon2017}.
73 By considering several theoretical and
74 empirical examples, we show the
75 advantages and the limits of the  $S_m$ 
76 function for detecting spatial structures
77 in economics.
      
```
- Preview Pane (Right):** Shows the rendered output of the LaTeX code. It includes:
 - ORIGINAL RESEARCH ARTICLE**
 - Title:** Distance-Based Measures of Spatial Concentration: Introducing a Relative Density Function
 - Authors:** Gaël Lapeere¹ and Eric Marcon² and Florence Pasqui³
 - Abstract:** For a decade, distance-based methods have been widely employed and constantly improved in spatial economics. These methods are a very useful tool for accurately evaluating the spatial distribution of economic activity. We introduce a new distance-based statistical measure for measuring the spatial concentration of industries. The function is the first relative density function to be proposed in economics. This tool supplements the typology of distance-based methods drawn up by [27]. By considering several theoretical and empirical examples, we show the advantages and the limits of the S_m function for detecting spatial structures in economics.
 - KEYWORDS:** Spatial Concentration, Aggregation, Point Pattern, Agglomerative Economic Geography
 - 1. Introduction**
 - Text:** Industrial agglomerations are one of the main features of today's economic geography [20, 30]. Thus, it is not surprising that each recent research has attempted to improve the measurement of the spatial concentration of activities. Distance-based methods are the least statistical measures to be proposed in the field of spatial economics for detecting spatial structures (empirical concentration or dispersion). By introducing space as continuous, distance-based methods provide a detailed analysis without models. Consequently, many authors consider them to be very promising techniques [8, 20, 21, 31] and have thus open the way for new developments in the spatial concentration of activities [1, 12, 26, 32]. Today, the Distance and Centring K_d function [25] is the most used distance-based method in economics. It is a density function that includes absolute concentration [8, 12]. In this article for the first time we shall introduce a relative density function in spatial economics. In a similar vein to the well-known kernel estimator [20], we have developed S_m

Eric Marcon

L'histoire

Organiser les données

Analyser les données

Ecrire

Organisation

Limites

Avantages

Outils retenus

Présentation
Le travail
Amont de P

Eric Marcon

L'histoire

Organiser les
données

Analyser les
données

Ecrire

Organisation

Limites

Avantages

Environnement de travail unique : RStudio.

Markdown.

Git et GitHub.

Tout document est un projet R.

Tout groupe de méthodes diffusable est un package (GitHub + Travis + CodeCov).

Organisation

Un dépôt commun

Sur GitHub: EcoFoG.

Eric Marcon

L'histoire

Organiser les données

Analyser les données

Ecrire

Organisation

Limites

Avantages

UMR Ecologie des Forêts de Guyane

Kourou, French Guiana | <http://www.ecofog.gf> | webmaster@ecofog.gf

Repositories 12 | People 10 | Teams 2 | Projects 0 | Settings

Search repositories... | Type: All | Language: All | Customize pinned repositories

ForestData
 Post-inventory processing of forest plot data
 HTML GPL-3.0 Updated 7 days ago

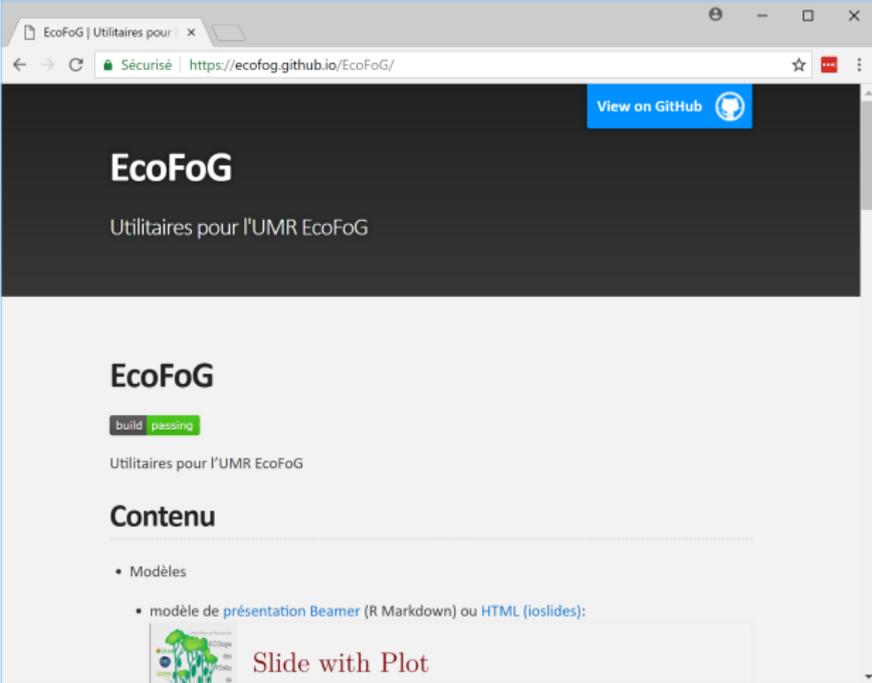
AppurementInventaire
 Script R générant des feuilles de terrains d'appurement
 R Updated 11 days ago

Top languages: TeX, R, PHP

People: [Profile pictures]

Un package commun

Package EcoFoG.



The screenshot shows a web browser window displaying the EcoFoG GitHub repository page. The browser's address bar shows the URL `https://ecofog.github.io/EcoFoG/`. The page has a dark header with the text "EcoFoG" and "Utilitaires pour l'UMR EcoFoG". A blue button labeled "View on GitHub" with the GitHub logo is in the top right. Below the header, the text "EcoFoG" is repeated, followed by a green badge that says "build passing". Underneath, it says "Utilitaires pour l'UMR EcoFoG". A section titled "Contenu" contains a list item "Modèles" with a sub-item "modèle de présentation Beamer (R Markdown) ou HTML (ioslides):". Below this is a small thumbnail image of a presentation slide titled "Slide with Plot" which features the EcoFoG logo and a plot.

Eric Marcon

L'histoire

Organiser les données

Analyser les données

Ecrire

Organisation

Limites

Avantages

Des modèles de documents

Eric Marcon

L'histoire

Organiser les données

Analyser les données

Ecrire

Organisation

Limites

Avantages

Présentation.

Article

Titre de l'article

Prénom Nom¹

Deuxième Auteur²

Résumé
 Résumé de l'article.

Mots-clés
 mots-clés, séparés par des virgules

¹USRI Ecologie, Agronomie, Climat, Environnement, INRA, Université des Antilles, Université de Guyane, Campus Agronomique, 67315 Kourou, France.
²Department of Ecology, University of Edinburgh
 Street address, Zip code, Country.
 Contact: prenom.nom@ecology.gf, http://www.ecology.gf/tpj/article47

Table des matières	
1	Introduction
2	R Markdown
2.1	Intérêt
2.2	Commentaire
3	Code
3.1	Code R
3.2	TikZ&LaTeX
3.3	Figures
3.4	Listes
3.5	Maths
3.6	Bibliographie
4	Types de document
4.1	Document HTML, PDF
4.2	Document Word
4.3	Présentation Beamer
4.4	Autres Modèles

1. Introduction

Ce modèle permet la rédaction d'articles au format Markdown. Il produit directement des articles bien formatés par auto-boutage (déjà sur R!A), par exemple et le code

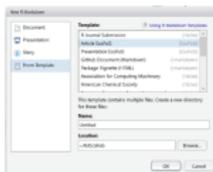


FIGURE 1. Nouveau document Markdown

2.1 Intérêt

Markdown est très simple à apprendre. Markdown permet d'écrire son code R pour un rendu reproductible.

Markdown permet de produire, sans réviser le texte, un document dans différents formats : article LaTeX ou Word par exemple.

2.2 Comment faire

Ouvrage

CHAPITRE 1

Notions de Diversité

LE TERME *biodiversité* est attribué¹ à Walter Rosen, un membre du National Research Council américain, qui a commencé à contracter les termes *biological diversity* pendant la préparation d'un colloque dont les actes seront publiés sous le titre "*Biodiversity*".² La question de la diversité biologique intéressait les écologues bien avant l'invention de la biodiversité, mais le néologisme a connu un succès fulgurant³ en même temps qu'il devenait une notion floue, dans lequel chacun peut piocher ce qu'il souhaite y trouver, au point de lui retirer son caractère scientifique.⁴ Une cause de ce glissement est que la biodiversité a été nommée pour attirer l'attention sur son érosion, en lien avec la biologie de la conservation. Cette érosion concernait potentiellement de nombreux aspects du monde vivant, la définition de la biodiversité fluctue selon les besoins : DeLong⁵ en recense 85 dans les dix premières années de littérature. Les indicateurs de la biodiversité peuvent englober bien d'autres choses que la diversité du vivant : le nombre d'espèces menacées (par exemple la liste rouge de l'UICN), la taille des populations ou la surface des écosystèmes préservés, la dégradation des habitats, la menace pesant sur des espèces emblématiques... Une mesure rigoureuse et cohérente de la diversité peut pourtant être construite pour clarifier beaucoup (mais pas tous) des concepts qui constituent la biodiversité.

Dans l'introduction du premier chapitre des actes de ce qui était devenu le « Forum sur la Biodiversité », Wilson utilise le mot dans le sens étroit de nombres d'espèces. L'élargissement de la notion aura « systèmes naturels » et à l'opposé à la diversité génétique intraspécifique est venu du monde de la conservation.⁶

La déclaration de Michel Leveson, président du comité scientifique de la conférence de Paris en 2005⁷ en donne une définition actuelle :

¹ C. Moore et al. (2006). « A century-driven discipline ». The growth of environmental biology, in : *Conservation Biology* 20, 3, p. 433-452. doi : 10.1111/j.1365-1730.2006.06440.x.

² W. C. Wilson et R. M. Peterkin. (1984). *Biodiversity*. Washington, D.C. : The National Academies Press.

³ P. Bouché (2014). « La diversité des vivant comme un aspect de la diversité : aspects historiques et géographiques ». In : *La biodiversité des écosystèmes. Écologie, philosophie, éthique et sociologie*. Sous la dir. G. Courlet et J. Dubois. Paris : Editions Météorologiques, Chap. 1, p. 31-48.

⁴ J. Dubois (2014). « La biodiversité : impasses scientifiques et non épistémologiques ». In : *La biodiversité en question. Écologie et sociologie*. Sous la dir. G. Courlet et J. Dubois. Paris : Editions Météorologiques, Chap. 2, p. 53-118. doi : 10.3929/ethz-b0001261441.01.000.

⁵ W. C. DeLong (1986). « Defining Biodiversity ». In : *Biodiversity*. Ecology Institute 24, 4, p. 739-745.

⁶ G. Speth et al. (1992). « Framework for Global Biodiversity Assessment ». In : W. W. Conner, Washington, D.C. : WWI, USNC, UNDP, 1992.

Des outils communs

Sans vocation à être publiés sur CRAN

Eric Marcon

L'histoire

Organiser les données

Analyser les données

Ecrire

Organisation

Limites

Avantages

Cruteur de carte auto

Forêt :
 Paracou

Campagne :
 2016

Parcelles :
 1

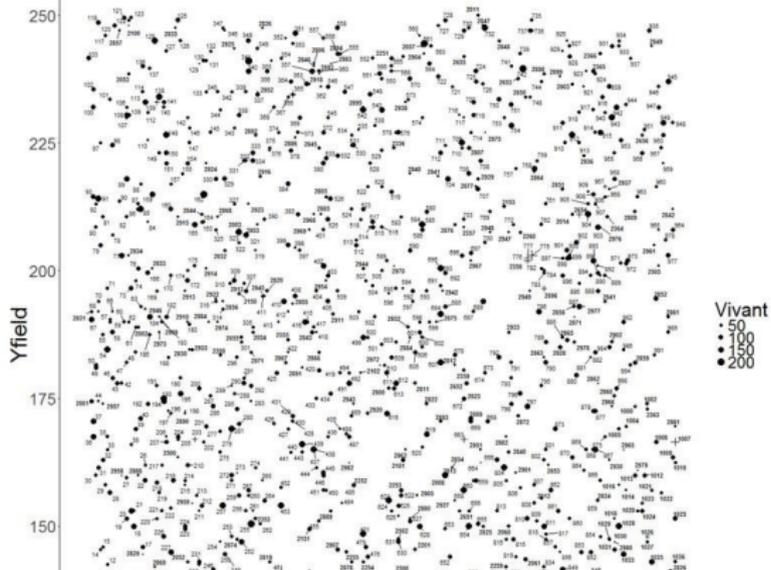
Taille du texte des libelles :
 7

Etiquetage intelligent

Extension :
 svg

Sauvegarder Aperçu

Paracou - Parcelle 1 - C 1



Packageurs communs / Packageurs personnels

Eric Marcon

L'histoire

Organiser les
données

Analyser les
données

Ecrire

Organisation

Limites

Avantages

Dans le dépôt EcoFoG : industrialisation des méthodes.

Dans les dépôts des chercheurs : recherche propre.

Le tout publié sur CRAN.

Documents communs / Documents personnels

Présentation
à l'usage
des collègues

Eric Marcon

L'histoire

Organiser les
données

Analyser les
données

Ecrire

Organisation

Limites

Avantages

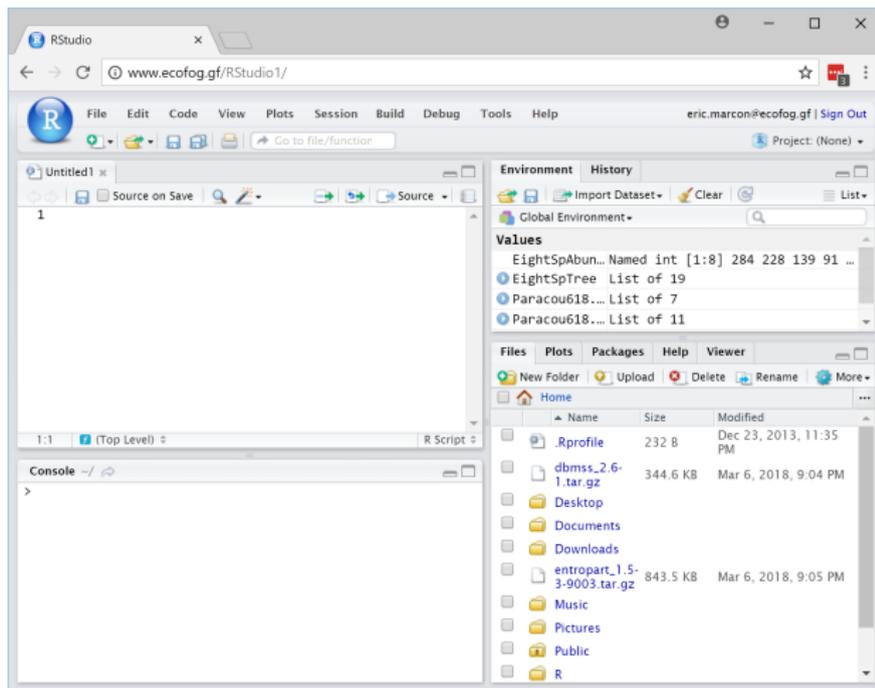
Même fonctionnement.

Utilisation systématique des pages GitHub.

Documents pas forcément publics : dépôt BitBucket.

Serveur RStudio

Pour les calculs longs ou parallélisés.



Applications Shiny.

Eric Marcon

L'histoire

Organiser les données

Analyser les données

Ecrire

Organisation

Limites

Avantages

Utilisation systématique

Organisation
du travail
Amont de P

Eric Marcon

L'histoire

Organiser les
données

Analyser les
données

Ecrire

Organisation

Limites

Avantages

Formation des étudiants.

Cours en ligne.

Support des nouveaux projets.

Limites

Adhésion ou pas.

Arguments :

- perte de contrôle,
- rigidité,
- pas Wywiwyg.

Compétences

Présentation
Présentation
Présentation de P

Eric Marcon

L'histoire

Organiser les
données

Analyser les
données

Ecrire

Organisation

Limites

Avantages

Formations nécessaires :

- à R,
- à Git,
- à Markdown... à LaTeX.

Manques :

- Correcteur d'orthographe en temps réel

Chaine complexe \leftrightarrow fragile.

Exemples :

- undefined control sequence
\@@magyar@captionfix;
- R et RTools 3.5.0 et devtools.

Avantages

Recherche reproductible

Présentation
à travers
l'exemple de R

Eric Marcon

L'histoire

Organiser les
données

Analyser les
données

Ecrire

Organisation

Limites

Avantages

Intégration complète de toute la chaîne.

Données → Traitements → Figures → Texte.

Multiples formats de sortie

Systématiquement HTML et PDF → Pages GitHub.

Eric Marcon

L'histoire

Organiser les données

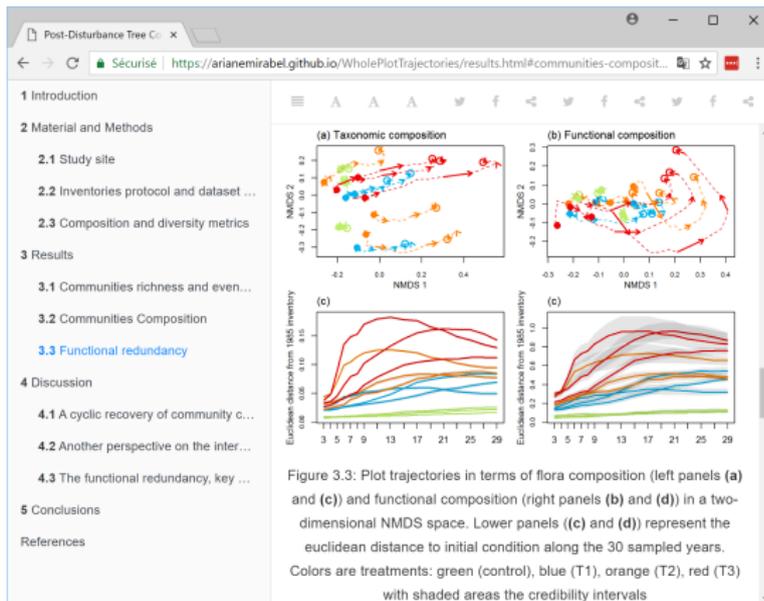
Analyser les données

Ecrire

Organisation

Limites

Avantages



Reformatage facile, même vers Word.

Qualité des documents

Eric Marcon

L'histoire

Organiser les données

Analyser les données

Ecrire

Organisation

Limites

Avantages

Seulement possible avec LaTeX:

- Respect des règles typographiques ;
- Usage des marges.

Mais tout LaTeX n'est pas disponible avec RMarkdown.

CHAPITRE 8

Entropie phylogénétique

L'essentiel
 L'entropie phylogénétique est la moyenne de l'entropie HCDT le long d'un arbre phylogénétique. Son minimum est simplement celle de l'entropie HCDT à chaque période de l'arbre. Elle va de pair avec la diversité phylogénétique qui est son nombre effectif d'espèces, c'est-à-dire le nombre d'espèces équiprobables, dans un arbre où toutes les espèces descendent d'un ancêtre unique, dont l'entropie serait la même que celle de la communauté réelle. Dans un tel arbre, la diversité phylogénétique se réduit à la diversité taxon.

L'entropie HCDT peut être utilisée pour définir une mesure de diversité prenant en compte l'histoire évolutive des espèces.

8.1 Généralisation de l'entropie HCDT

Provine et Bonnell¹ découpent l'arbre phylogénétique en périodes. À partir de la racine de l'arbre, une nouvelle période est définie à chaque ramification d'une branche quelconque. Les débuts et fins de périodes sont notés t_0 , la racine de l'arbre est fixée à $t_0 = 0$. L'arbre est ultramétrique.

Nous suivrons plutôt les notations de Chao et al.² en numérotant les périodes à partir du présent et en notant T_i leur durée. Figure 8.7, la première période se termine quand les branches des espèces 3 à 5 se rejoignent. L'arbre comprend $K = 3$ périodes.

L'entropie HCDT (9) de l'équation (4.6) est calculée à chaque période. Figure 8.1, à la deuxième période (T_2), l'arbre a trois feuilles, avec des probabilités égales à celle des espèces 1 et 2 et la somme de celles des espèces 3 à 5. H_2 peut être calculée avec ces valeurs de probabilités. De même cette valeur d'entropie H_3 ou 8 est le quantile de la période.

¹Provine et Bonnell (2006), « Biological diversity: Limited classification can lead to the overestimation of biodiversity entropy », *et. vol. 63, p. 117*

²Chao et al. (2014), « Phylogenetic diversity measure based on HCD method », *et. vol. 67, p. 65*

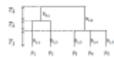


FIGURE 8.1. Arbre phylogénétique en branchement logarithmique. Majorité de la figure 7.7. L'entropie peut être calculée à $t = 0$, avec une probabilité égale à 1/3. Les autres des branches sont égales.

Rigueur

Présentation
Le travail
Amont de P

Eric Marcon

L'histoire

Organiser les
données

Analyser les
données

Ecrire

Organisation

Limites

Avantages

Documentation au même niveau que la réflexion.

Possibilité de revenir en arrière, historique.

Réduction des zones d'ombre.

Capitalisation.

GitHub:

- <https://github.com/EcoFoG/>

UMR EcoFoG :

- <https://www.ecofog.gf/>