Organisation du travail autour de R

Eric Marcon

5 juillet 2018

Organisation du travail autour de R

Eric Marcon

L'histoire

Contexte

Laboratoire de recherche.

Equipes de chercheurs - étudiants - techniciens.

Production de données, méthodes et documents.

Organisation spontanée

On pilote difficilement une équipe de chercheurs.

Haut niveau technique.

Tendance à diverger.

Objectifs

Etre plus efficace.

Echanger plus facilement.

Recherche reproductible.

Méthode

Les outils ne font pas l'organisation.

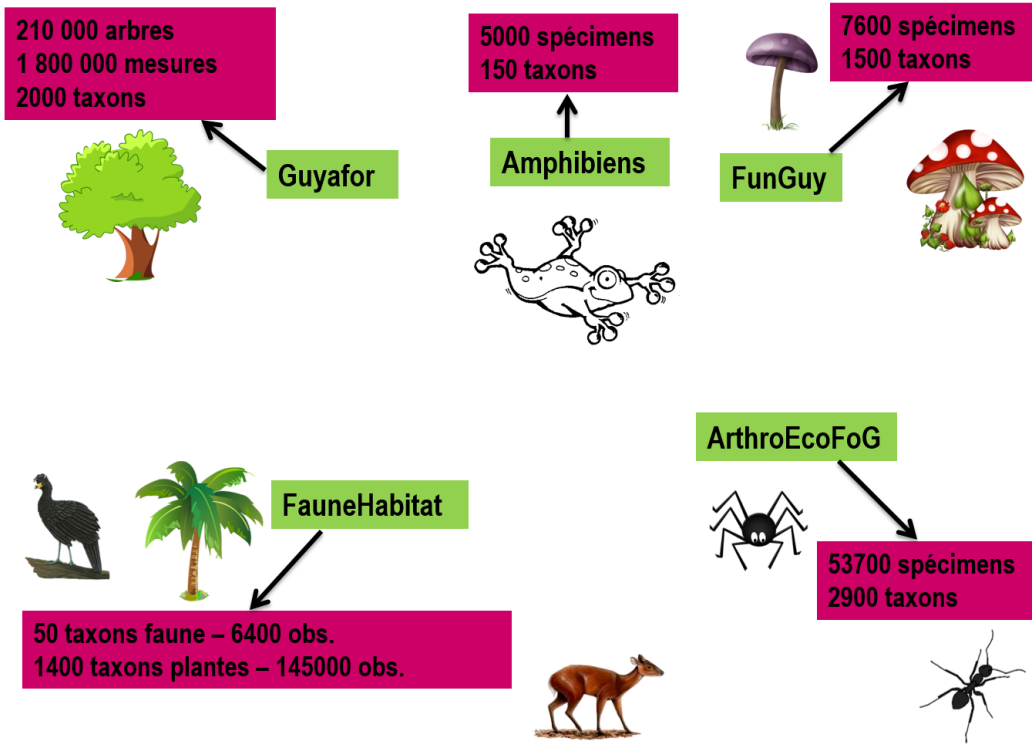
Les objectifs sans outils non plus.

Itérations besoins $\leftrightarrow$ outils

Organiser les données

Données précieuses

Relativement peu de données en écologie.



Prix unitaire élevé.

Choix

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

R

Script pour :

- la reproductibilité
- la versatilité
- l'explicitation

Communauté, gratuité...

Documentation

Code R avec commentaires

```
# Addition
2 + 2
```

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

Document RMarkdown avec chunks

Addition

Utiliser l'opérateur + :

```
2+2
```

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

Ecrire

Enjeux

Beaucoup de temps passé à produire des documents.

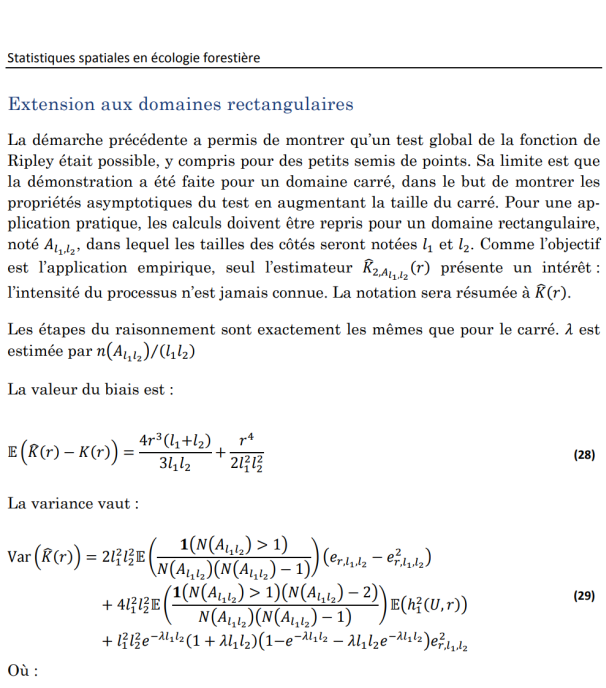
Processus collaboratif nécessaire.

Réutilisation.

Au début

Documents Word:

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



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_{l_1,l_2}, dans lequel les tailles des côtés seront notées l_1 et l_2. Comme l'objectif est l'application empirique, seul l'estimateur $\hat{K}_{2,A_{l_1,l_2}}(r)$ présente un intérêt : l'intensité du processus n'est jamais connue. La notation sera résumée à $\hat{K}(r)$.

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

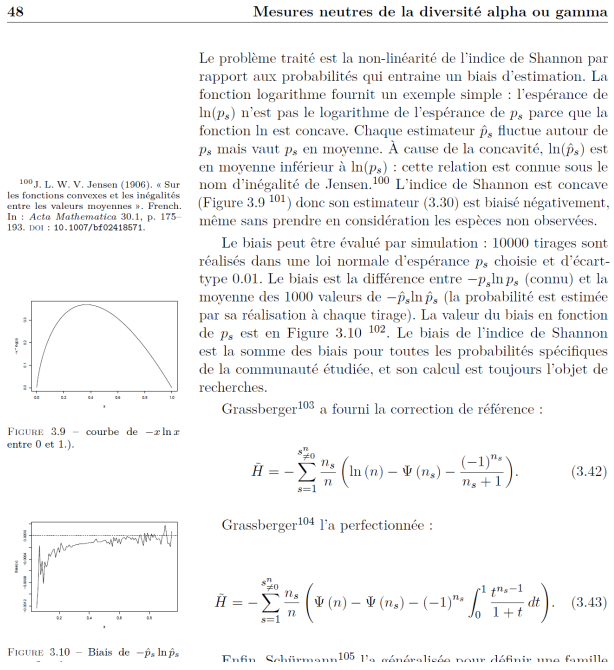
La valeur du biais est :

$$\mathbb{E}\left(\hat{K}(r) - K(r)\right) = \frac{4r^3(l_1+l_2)}{3l_1l_2} + \frac{r^4}{2l_1^2l_2^2} \quad (28)$$

La variance vaut :

$$\begin{aligned}\mathrm{Var}\left(\hat{K}(r)\right) = {} & 2l_1^2l_2^2\mathbb{E}\left(\frac{\mathbf{1}(N(A_{l_1l_2})>1)}{N(A_{l_1l_2})(N(A_{l_1l_2})-1)}\right)\left(e_{r,l_1,l_2} - e^2_{r,l_1,l_2}\right) \\ & + 4l_1^2l_2^2\mathbb{E}\left(\frac{\mathbf{1}(N(A_{l_1l_2})>1)(N(A_{l_1l_2})-2)}{N(A_{l_1l_2})(N(A_{l_1l_2})-1)}\right)\mathbb{E}\left(h_1^2(U,r)\right) \\ & + l_1^2l_2^2e^{-\lambda l_1l_2}(1+\lambda l_1l_2)(1-e^{-\lambda l_1l_2} - \lambda l_1l_2e^{-\lambda l_1l_2})e^2_{r,l_1,l_2}\end{aligned} \quad (29)$$

Où :



48 Mesures neutres de la diversité alpha ou gamma

Le problème traité est la non-linéarité de l'indice de Shannon par rapport aux probabilités qui entraine un biais d'estimation. La fonction logarithme fournit un exemple simple : l'espérance de $\ln(p_s)$ n'est pas le logarithme de l'espérance de p_s parce que la fonction ln est concave. Chaque estimateur $\hat{p}_s$ fluctue autour de p_s mais vaut p_s en moyenne. À cause de la concavité, $\ln(\hat{p}_s)$ est en moyenne inférieur à $\ln(p_s)$: cette relation est connue sous le nom d'inégalité de Jensen.[100] L'indice de Shannon est concave (Figure 3.9 [101]) donc son estimateur (3.30) 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_s choisie et d'écart-type 0.01. Le biais est la différence entre $-p_s \ln p_s$ (connu) et la moyenne des 1000 valeurs de $-\hat{p}_s \ln \hat{p}_s$ (la probabilité est estimée par sa réalisation à chaque tirage). La valeur du biais en fonction de p_s est en Figure 3.10 [102]. 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.

Grassberger[103] a fourni la correction de référence :

$$\check{H} = -\sum_{s=1}^{s^n_{\neq 0}} \frac{n_s}{n}\left(\ln(n) - \Psi(n_s) - \frac{(-1)^{n_s}}{n_s+1}\right). \quad (3.42)$$

Grassberger[104] l'a perfectionnée :

$$\check{H} = -\sum_{s=1}^{s^n_{\neq 0}} \frac{n_s}{n}\left(\Psi(n) - \Psi(n_s) - (-1)^{n_s}\int_0^1 \frac{t^{n_s-1}}{1+t}\,dt\right). \quad (3.43)$$

Enfin, Schürmann[105] l'a généralisée pour définir une famille

[100] J. L. W. V. Jensen (1906). « Sur les fonctions convexes et les inégalités entre les valeurs moyennes ». French. In : *Acta Mathematica* 30.1, p. 175–193. DOI : 10.1007/bf02418571.

FIGURE 3.9 – courbe de $-x \ln x$ entre 0 et 1.).

FIGURE 3.10 – Biais de $-\hat{p}_s \ln \hat{p}_s$

Echange par messagerie.

Besoin individuel

Se concentrer sur le fond :

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

Construire sa pensée $\leftrightarrow$ rédiger :

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

Besoin collectif

Ecriture en parallèle

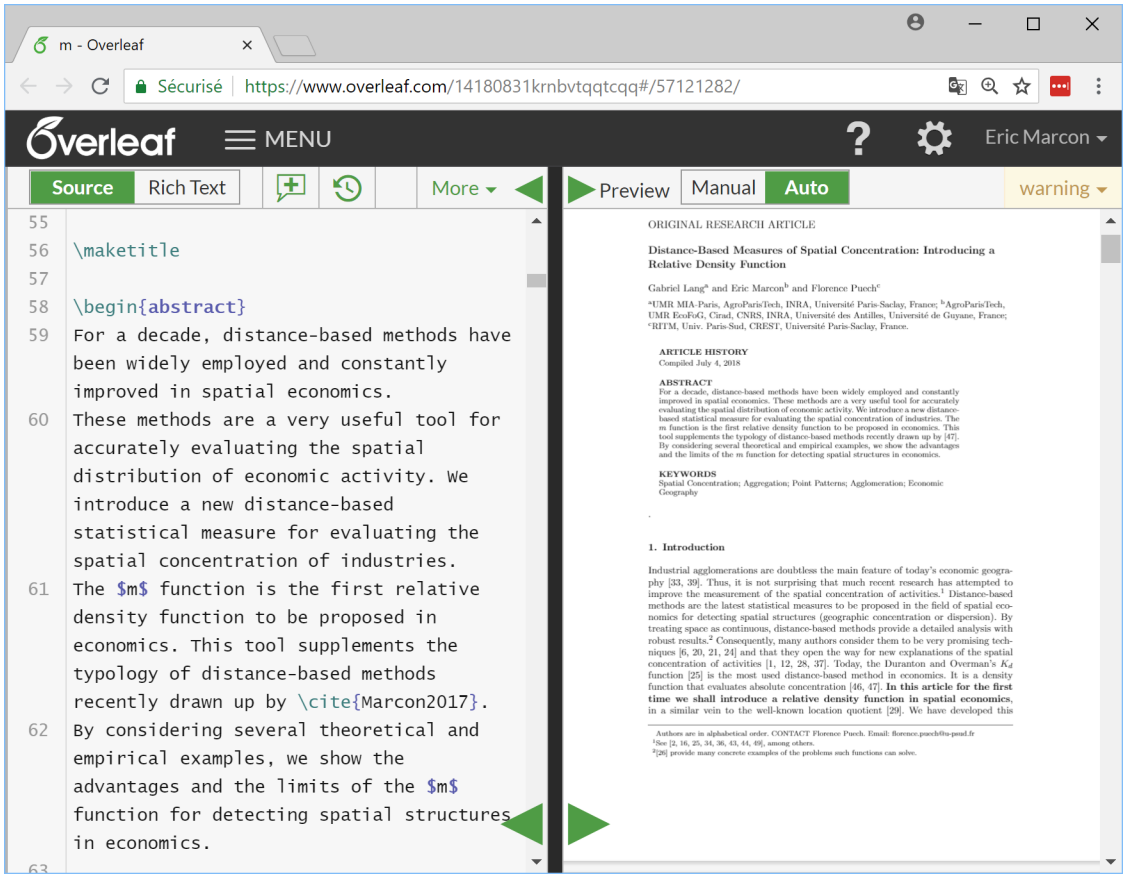
Suivi des versions

Progrès possibles

Google Docs : collaboration.

SharePoint puis Office 365.

Overleaf:



Outils retenus

Environnement de travail unique : RStudio.

Markdown.

Git et GitHub.

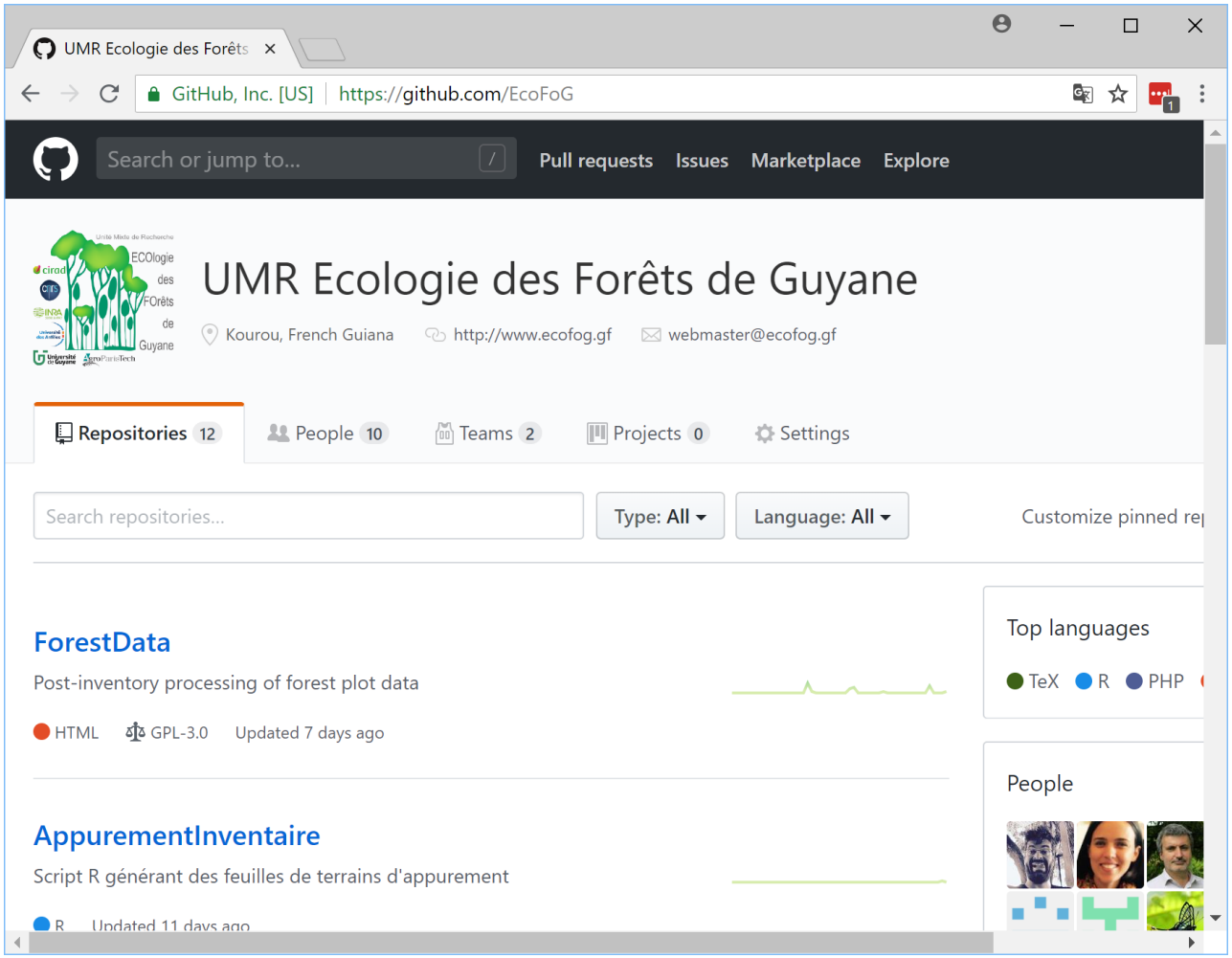
Tout document est un projet R.

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

Organisation

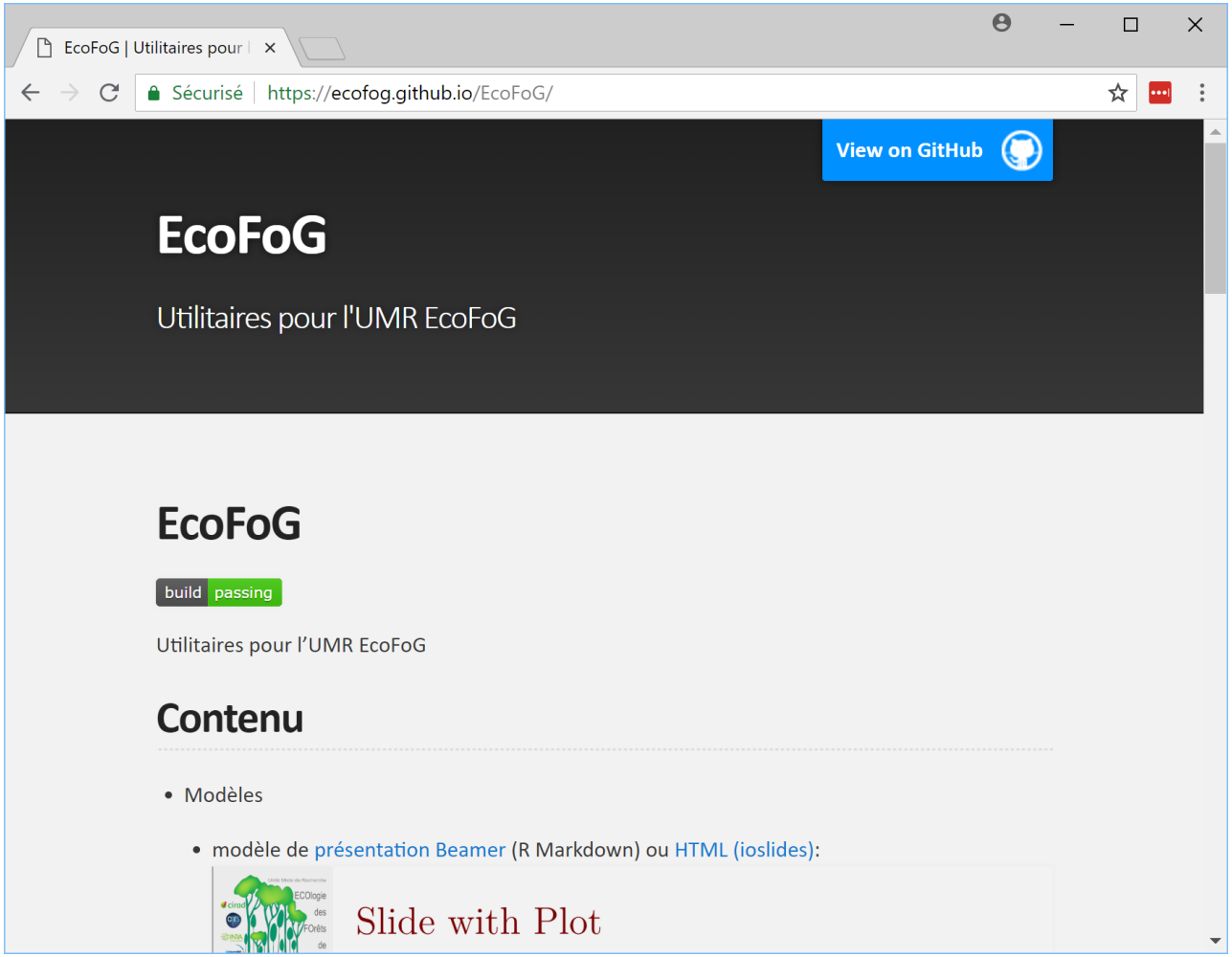
Un dépôt commun

Sur GitHub: EcoFoG.



Un package commun

Package EcoFoG.



Des modèles de documents

Présentation.

Article

Titre de l'article

Prénom Nom[1*]
Deuxième Auteur[2]

Résumé
Résumé de l'article.

Mots-clés
mot-clés, séparés par des virgules

[1]UMR EcoFoG, AgroParistech, CNRS, Cirad, INRA, Université des Antilles, Université de Guyane.
Campus Agronomique, 97310 Kourou, France.
[2]Department of Ecology, University of Edimburgh
Street address, Zip code, Country.
*Contact: prenom.nom@ecofog.gf, http://www.ecofog.gf/spip.php?article47

Table des matières

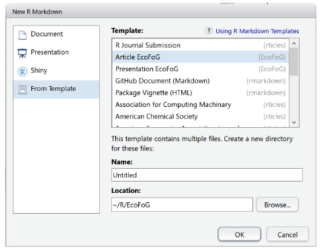


FIGURE 1. Nouveau document Markdown

1. Introduction

Ce modèle permet la rédaction d'articles au format Markdown. Il produit directement des articles bien formatés pour l'auto-harchivage (dépôt sur HAL par exemple) et le code

2.1 Intérêt

Markdown est très simple à apprendre.

Markdown permet d'intégrer son code R pour un résultat *reproductible*.

Markdown permet de produire, sans réécrire 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 *biodiversity* est attribué[1] 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 fulgurant[3] en même temps qu'il devenait une notion floue, dans lequel chacun peut placer 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 concernant potentiellement de nombreux aspects du monde vivant, la définition de la biodiversité fluctue selon les besoins : DeLong[5] 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'IUCN), 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 aux « systèmes naturels » et à l'opposé à la diversité génétique intraspécifique est venu du monde de la conservation.[6] La déclaration de Michel Loreau, président du du comité scientifique de la conférence de Paris en 2005[7] en donne une définition aboutie :

[1] C. Meine et al. (2006). « "A mission-driven discipline" : The growth of conservation biology ». In : *Conservation Biology* 20.3, p. 631–651. DOI : 10.1111/j.1523-1739.2006.00449.x.

[2] E. O. Wilson et F. M. Peter, éds. (1988). *Biodiversity*. Washington, D.C. : The National Academies Press.

[3] P. Blandin (2014). « La diversité du vivant avant (et après) la biodiversité : repères historiques et épistémologiques ». In : *La biodiversité en question. Enjeux philosophiques, éthiques et scientifiques*. Sous la dir. d'É. Casetta et J. Delord. Paris : Editions Matériologiques. Chap. 1, p. 31–68.

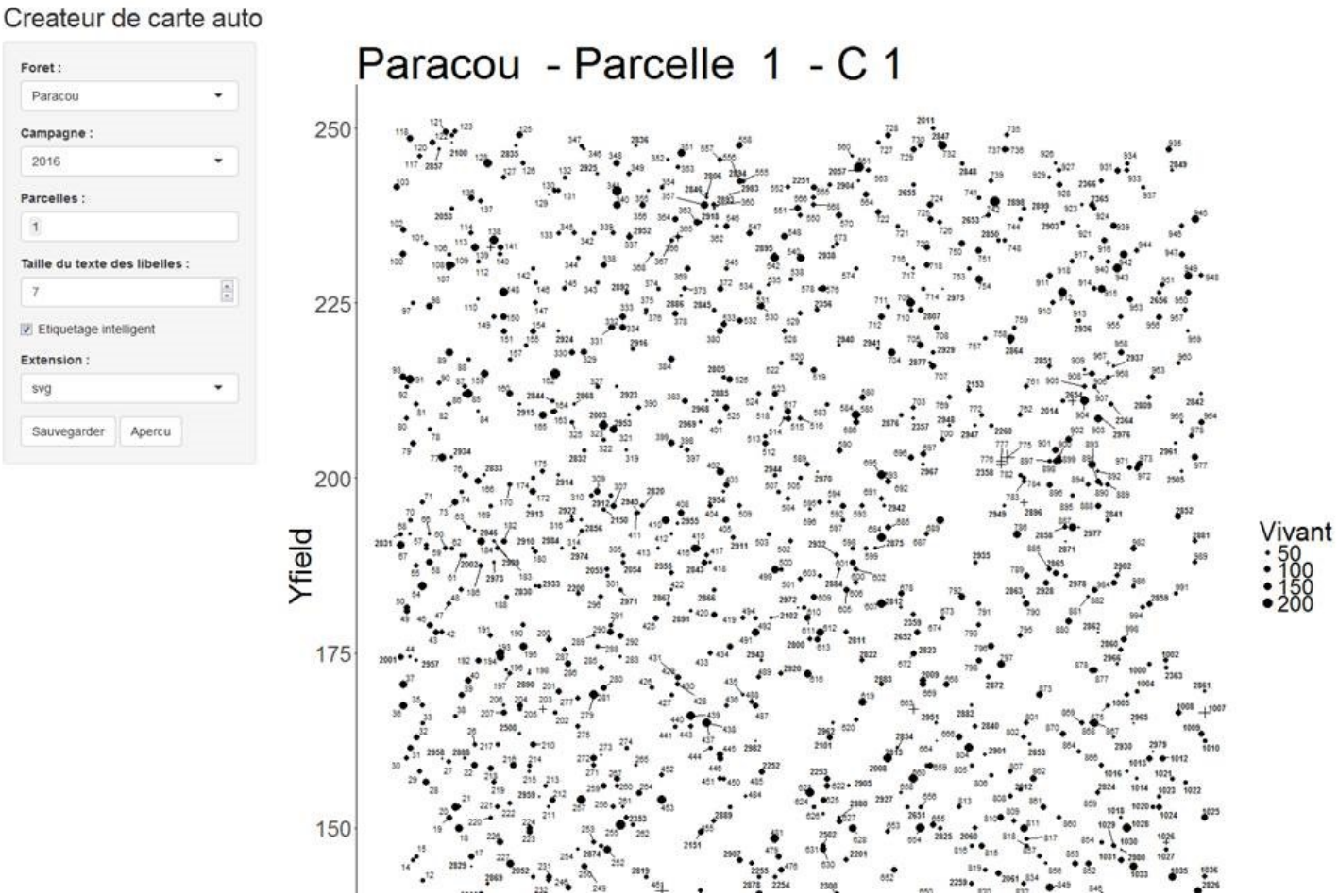
[4] J. Delord (2014). « La biodiversité : imposture scientifique ou ruse épistémologique ? » In : *La biodiversité en question. Enjeux philosophiques, éthiques et scientifiques*. Sous la dir. d'É. Casetta et J. Delord. Paris : Editions Matériologiques. Chap. 3, p. 83–118. DOI : 10.3917/ednat.delor.2014.01.0083.

[5] D. C. J. DeLong (1996). « Defining Biodiversity ». In : *Wildlife Society Bulletin* 24.4, p. 738–749.

[6] J. G. Speth et al. (1992). « Foreword ». In : *Global Biodiversity Strategy*. Sous la dir. de K. Courrier. Washington, D.C. : WRI, IUCN, UNEP,

Des outils communs

Sans vocation à être publiés sur CRAN



Packages communs / Packages personnels

Dans le dépôt EcoFoG : industrialisation des méthodes.

Dans les dépôts des chercheurs : recherche propre.

Le tout publié sur CRAN.

Documents communs / Documents personnels

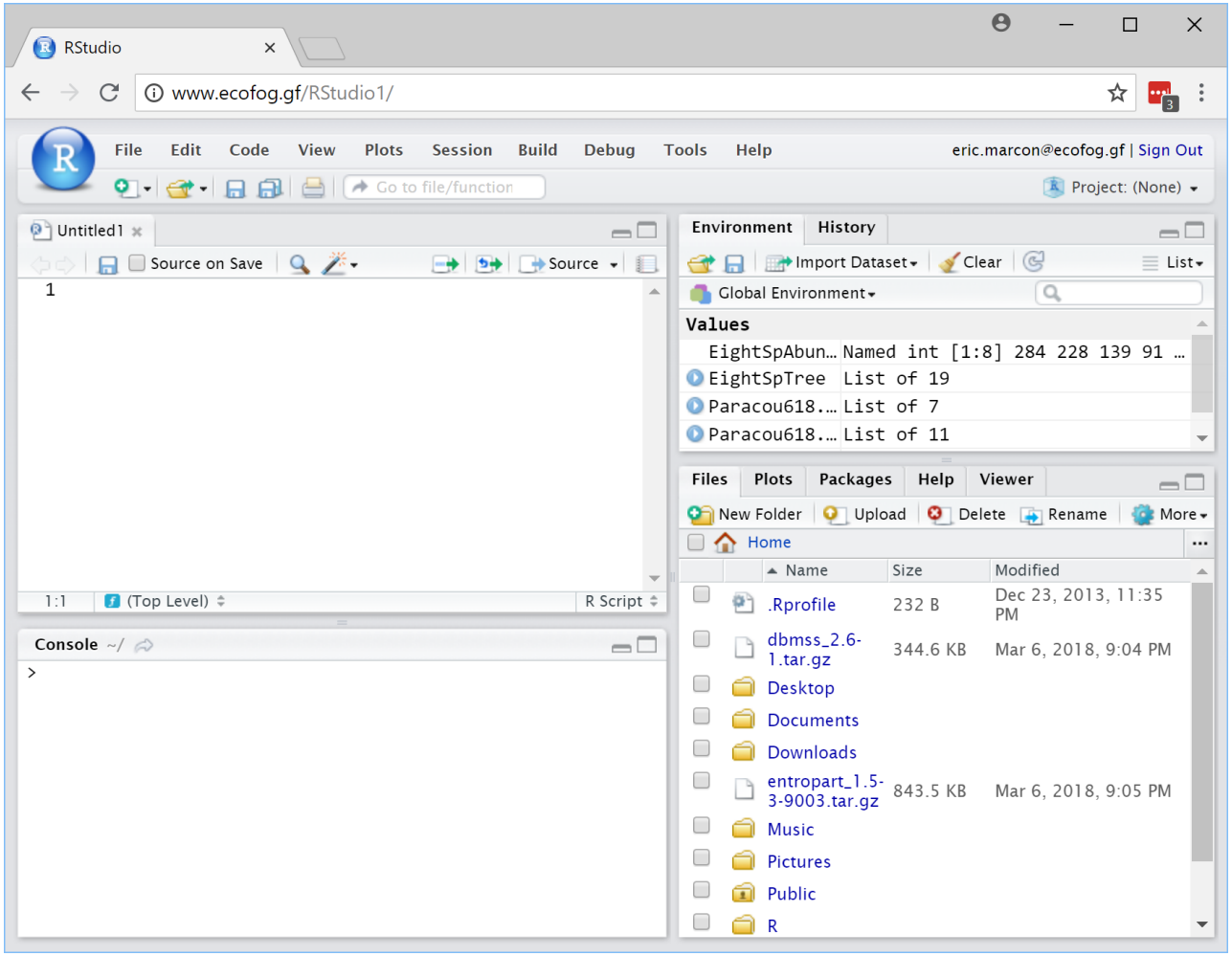
Même fonctionnement.

Utilisation systématique des pages GitHub.

Documents pas forcément publics : dépôt BitBucket.

Serveur RStudio

Pour les calculs longs ou parallélisés.



Applications Shiny.

Utilisation systématique

Formation des étudiants.

Cours en ligne.

Support des nouveaux projets.

Limites

Envie

Adhésion ou pas.

Arguments :

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

Compétences

Formations nécessaires :

- à R,
- à Git,
- à MarkDown... à LaTeX.

Technique

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

Intégration complète de toute la chaîne.

Données → Traitements → Figures → Texte.

Multiples formats de sortie

Systématiquement HTML et PDF → Pages GitHub.

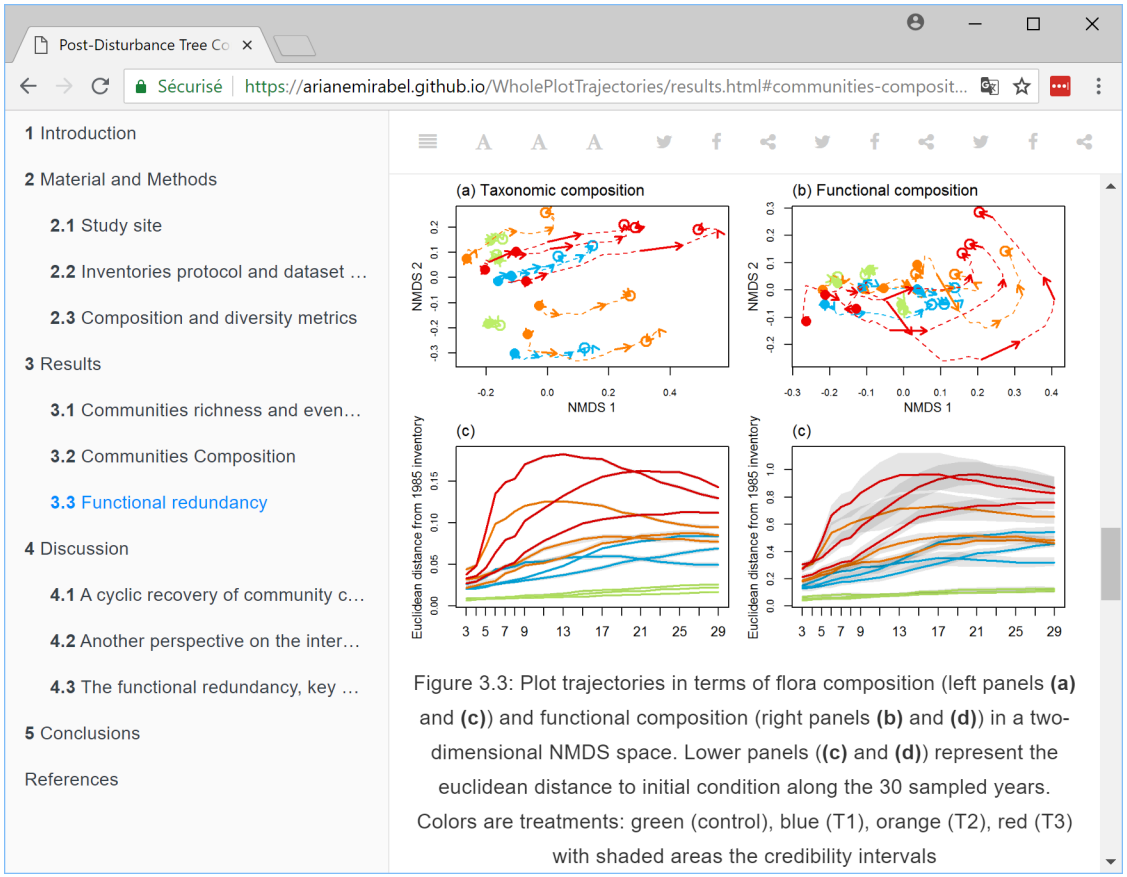


Figure 3.3: Plot trajectories in terms of flora composition (left panels **(a)** and **(c)**) and functional composition (right panels **(b)** and **(d)**) in a two-dimensional NMDS space. Lower panels (**(c)** and **(d)**) represent the euclidean distance to initial condition along the 30 sampled years. Colors are treatments: green (control), blue (T1), orange (T2), red (T3) with shaded areas the credibility intervals

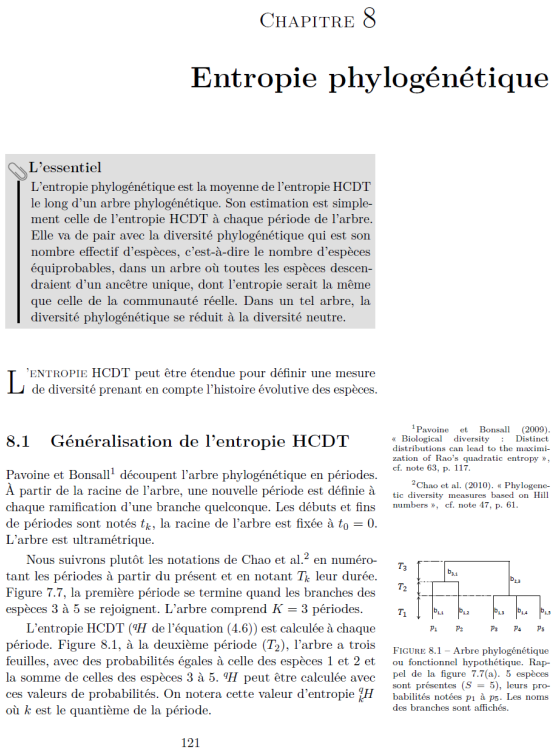
Reformatage facile, même vers Word.

Qualité des documents

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 estimation 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 descendraient 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 neutre.

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

8.1 Généralisation de l'entropie HCDT

Pavoine et Bonsall[1] 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_k, 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_k leur durée. Figure 7.7, la première période se termine quand les branches des espèces 3 à 5 se rejoignent. L'arbre comprend $K = 3$ périodes.

L'entropie HCDT (${}^qH$ 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. ${}^qH$ peut être calculée avec ces valeurs de probabilités. On notera cette valeur d'entropie ${}^q_kH$ où k est le quantième de la période.

[1] Pavoine et Bonsall (2009). « Biological diversity : Distinct distributions can lead to the maximization of Rao's quadratic entropy », cf. note 63, p. 117.

[2] Chao et al. (2010). « Phylogenetic diversity measures based on Hill numbers », cf. note 47, p. 61.

FIGURE 8.1 – Arbre phylogénétique ou fonctionnel hypothétique. Rappel de la figure 7.7(a). 5 espèces sont présentes ($S = 5$), leurs probabilités notées p_1 à p_5. Les noms des branches sont affichés.

121

Rigueur

Documentation au même niveau que la réflexion.

Possibilité de revenir en arrière, historique.

Réduction des zones d'ombre.

Capitalisation.

Références

GitHub:

- https://github.com/EcoFoG/

UMR EcoFoG :

- https://www.ecofog.gf/