Diversité des pixels

Eric Marcon

8 juillet 2022

1 Thérorie de la diversité neutre

1.1 Entropie de Shannon

La théorie de la diversité fondée sur la théorie de l'information (Shannon, Shannon) définit la diversité comme la difficulté à prévoir la nature d'un individu choisi au hasard (l'espèce d'un arbre tiré dans une communauté).

On définit la *probabilité* d'une espèce comme la probabilité qu'un arbre choisi au hasard lui appartienne. On définit la *rareté* d'une espèce comme l'inverse de sa probabilité.

L'*information* apportée par un individu est une fonction de sa rareté. Tirer un individu d'une espèce rare apporte beaucoup d'information (l'information peut être comprise comme la surprise). L'information est nulle pour une rareté égale à 1 (il n'y a qu'une seule espèce, on est donc sûr de la tirer) et doit croître avec la rareté. La fonction d'information fondamentale est celle de Shannon : le logarithme de la rareté.

L'*entropie* d'une communauté est l'information moyenne apportée par un individu. Comme tous les individus de l'espèce s ont la même rareté, l'entropie s'écrit

$$H = \sum_s p_s \ln (1/p_s). \quad (1)$$

Exemple d'une communauté de 3 espèces, de probabilités 1/2, 1/4 et 1/4 :

```
library("entropart")
library("dplyr")
# Création de la communauté
c(s1 = 1/2, s2 = 1/4, s3 = 1/4) %>%
    as.ProbaVector -> C
# Entropie
Shannon(C)
```

```
##     None
## 1.039721
```

« None » signifie qu'aucune correction n'a été apportée à l'estimation. Dans la réalité, la diversité est calculée à partir d'échantillons par des estimateurs mathématiques essaient de corriger l'incomplétude des données. Par exemple, l'estimateur de Chao permet d'évaluer la richesse en évaluant le nombre d'espèces présentes dans la communautés mais pas observées.

1.2 Entropie généralisée

D'autres fonctions d'information peuvent être utilisées, à condition qu'elles croissent avec la rareté, à partir de 0 pour une rareté égale à 1.

Une façon efficace et simple de généraliser l'entropie de Shannon consiste à déformer la fonction logarithme. Le logarithme déformé d'ordre q est défini par

$$\ln_q x = \frac{x^{1-q} - 1}{1 - q}, \tag{2}$$

Le logarithme déformé converge vers le logarithme naturel quand $q \to 1$ (figure 1).

```
ln0 <- function(p) lnq(1/p, q = 0)
ln1 <- function(p) log(1/p)
ln2 <- function(p) lnq(1/p, q = 2)
lnm1 <- function(p) lnq(1/p, q = -1)
ggplot(data.frame(x = c(0, 1)), aes(x)) + stat_function(fun = ln1) +
    stat_function(fun = ln0, lty = 2, col = "red") +
    stat_function(fun = ln2, lty = 3, col = "blue") +
    stat_function(fun = lnm1, lty = 4, col = "green") +
    coord_cartesian(ylim = c(0, 10)) + labs(x = "p",
    y = expression(ln[q](1/p)))
```

L'entropie généralisée, dite HCDT ou de (Tsallis, Tsallis) est encore la moyenne du logarithme (déformé) de la rareté.

$$^qH = \sum_s p_s \ln_q (1/p_s). \tag{3}$$

L'intérêt de ce formalisme est que tous les « indices de diversité » classiques sont des cas particuliers d'entropie HCDT pour des valeurs particulières de q :

- ^0H est la richesse (le nombre d'espèces) moins 1 ;
- ^1H est l'entropie de Shannon ;
- ^2H est l'entropie de Simpson (la probabilité que deux arbres soient d'espèces différentes).

Plus q est petit, plus les espèces rares influent sur l'entropie :

- Pour $q = 0$, une espèce a la même importance quelle que soit son abondance.
- Pour $q = 2$, les espèces rares ont un effet négligeable.
- Pour $q = +\infty$, seule l'espèce la plus abondante est prise en compte.

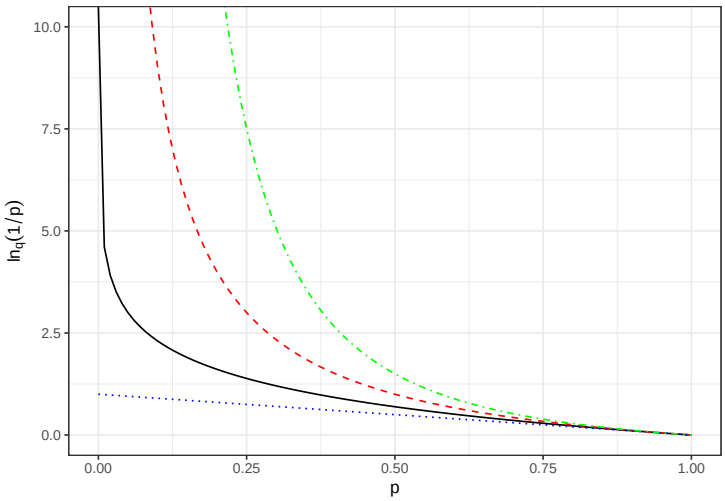


FIG. 1 : Valeur du logarithme d'ordre q de la rareté pour des probabilités entre 0 et 1 et différentes valeurs de q : $q = -1$ (pointillés alternés verts) ; $q = 0$ (pointillés longs rouges) ; $q = 1$ (trait plein) : logarithme naturel ; $q = 2$ (pointillés courts bleus). L'ordre de la diversité donne d'autant plus de poids aux espèces rares (l'information est plus grande) qu'il est petit.

Exemple : la fonction `Tsallis()` retourne l'entropie de Tsallis.

```
# Indices de diversité et entropie q=0
Richness(C) - 1
```

```
## None
##    2
```

```
Tsallis(C, q = 0)
```

```
## None
##    2
```

```
# q=1
Shannon(C)
```

```
##     None
## 1.039721
```

```
Tsallis(C, q = 1)
```

```
##     None
## 1.039721
```

```
# q=2
Simpson(C)
```

```
##  None
## 0.625
```

```
Tsallis(C, q = 2)
```

```
##  None
## 0.625
```

1.3 Nombres de Hill

Sauf dans les cas particuliers $q = 0$ et $q = 2$, l'entropie est difficile à interpréter (Hurlbert, Hurlbert). Hill (Hill) propose de la transformer en nombre effectifs d'espèces, c'est-à-dire le nombre d'espèces équiprobables qui formeraient une communauté avec la même entropie que les données.

Le nombre effectif d'espèces correspondant à l'entropie HCDT est son exponentielle déformée (la fonction réciproque du logarithme déformé) :

$$ {}^{q}D = e_{q}^{{}^{q}H}. \tag{4}$$

Le terme *diversité* doit être réservé aux nombres de Hill (Jost, Jost), l'entropie doit être appelée entropie et le terme « indice » évité.

Exemple :

```
# q=0
Richness(C)
```

```
## None
##    3
```

```
expq(Tsallis(C, q = 0), q = 0)
```

```
## None
##    3
```

```
Diversity(C, q = 0)
```

```
## None
##    3
```

```
# q=1
expq(Tsallis(C, q = 1), q = 1)
```

```
##     None
## 2.828427
```

```
Diversity(C, q = 1)
```

```
##     None
## 2.828427
```

```
# q=2
expq(Tsallis(C, q = 2), q = 2)
```

```
##     None
## 2.666667
```

```
Diversity(C, q = 2)
```

```
##     None
## 2.666667
```

En pratique, on utilise la fonction `Diversity()`. A l'ordre 2 (c'est-à-dire au sens de la diversité de Simpson), une communauté contenant 2,66 espèces équiprobable aurait la même diversité que la communauté étudiée.

1.4 Profils de diversité

Comme la diversité est un nombre d'espèces, on peut tracer sa courbe en fonction de q qui résume la diversité d'une communauté quel que soit le poids donnée à ses espèces rares.

Exemple de deux prodils de diversité d'un hectare de forêt de Paracou, parcelles 6 et 18

```
# Valeurs de q utilisées pour tracer la courbe
q.seq <- seq(0, 2, 0.1)
# Parcelle 6
P6D <- CommunityProfile(Diversity, Paracou618.MC$Nsi[,
    1], q.seq)
# Parcelle 18
P18D <- CommunityProfile(Diversity, Paracou618.MC$Nsi[,
    2], q.seq)
# Figure
autoplot(P6D, xlab = "q", ylab = "Diversité") + geom_line(aes(x,
    y), as.data.frame.list(P18D), lty = 2)
```

La figure 2 montre que la parcelle 18 de Paracou est plus diverse que la parcelle 6.

Profil de diversité de la communauté étudiée :

```
CommunityProfile(Diversity, C) %>%
    autoplot
```

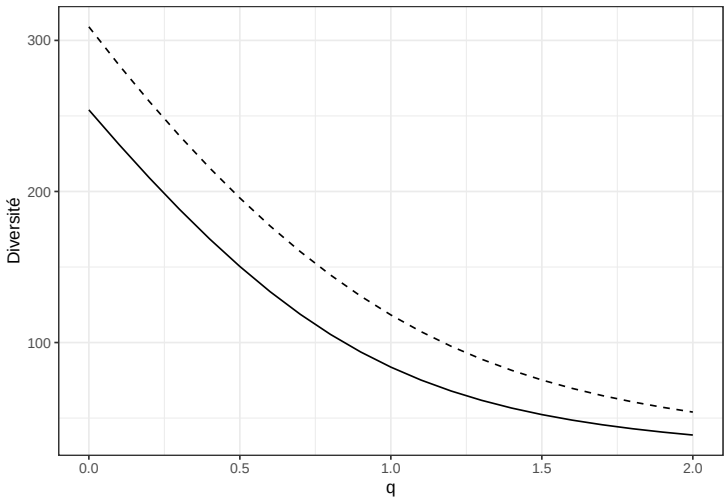
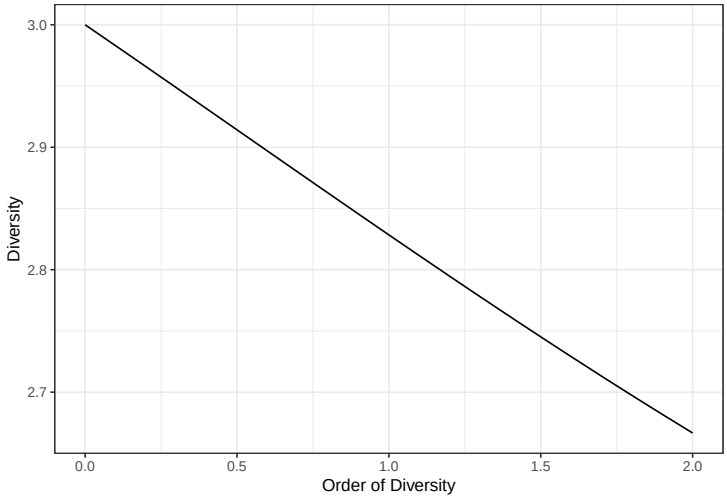


FIG. 2 : Profil de diversité calculé pour deux parcelles de Paracou (Parcelle 6 : trait plein et Parcelle 18 : trait pointillé). La correction du biais d'estimation est celle de Chao et Jost.



2 Diversité fonctionnelle

La diversité définie précédemment est dite *neutre* parce que les différences entre espèces sont les mêmes. L'identité des espèces n'a pas d'importance : deux individus sont de la même espèce ou d'espèces différentes, mais cette différence ne varie pas.

La définition de la rareté peut être généralisée pour prendre en compte la proximité plus ou moins grande des espèces (Leinster and Cobbold, Leinster and Cobbold).

On définit une distance entre toutes les paires d'individus (qui peut ne dépendre que de leurs espèces ou varier pour chaque individu). De façon équivalente, on peut écrire une matrice de distance ou placer les individus dans un espace multidimensionnel (la transformation est une Analyse en Coordonnées Principales : PcoA). On obtient ce type de données en calculant les distances entre individus en partir de leurs valeurs de traits fonctionnels (d'où le nom de diversité fonctionnelle) ou de n'importe quelle valeurs numériques multidimensionnelles (comme la distance entre pixels d'une image dans l'espace colorimétrique).

Les distances doivent être transformée en similarités. Classiquement :

- on normalise les distances entre 0 et 1 et on définit la similarité comme 1 moins la distance.
- ou (les propriétés mathématiques sont plus solides), on définit la similarité comme l'exponentielle négative de la distance. La similarité entre les espèces s et t est $z_{s,t} = e^{-ud_{s,t}}$ où u est un réel positif qui fixe la décroissance de la similarité en fonction de la distance : plus u est grand, plus les espèces sont dissimilaires pour une même distance.

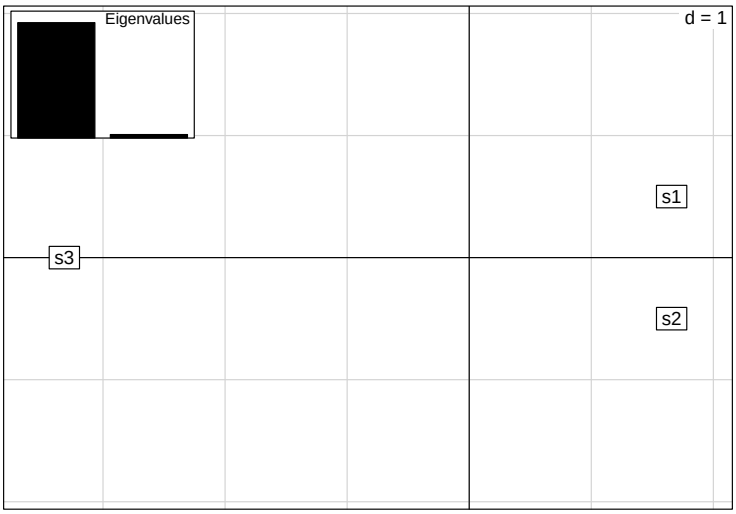
Un individu est complètement similaire à lui-même (distance nulle, similarité égale à 1).

> Exemple : les distances entre les 3 espèces sont 1 (entre s1 et s2) et 5 entre s3 et les autres.

```
D <- matrix(
  c(
    0, 1, 5,
    1, 0, 5,
    5, 5, 0
    ),
  nrow = 3)
colnames(D) <- rownames(D) <- names(C)
```

> L'espace engendré par cette matrice de distance est :

```
library("ade4")
D %>%
    as.dist %>%
    dudi.pco(scannf = FALSE, nf = 2) %>%
    scatter
```



La matrice de similarité est obtenue en prenant l'exponentielle négative de la similarité, avec $u = 1$ pour commencer.

```
(Z <- exp(-D))
```

```
##             s1          s2          s3
## s1 1.000000000 0.367879441 0.006737947
## s2 0.367879441 1.000000000 0.006737947
## s3 0.006737947 0.006737947 1.000000000
```

La *banalité* d'un individu est sa similarité moyenne avec tous les autres, y compris lui-même.

La banalité des individus est la même à l'intérieur de chacune des trois espèces : on parle de banalité des espèces. Elle vaut :

```
(B <- Z %*% C)
```

```
##         [,1]
## s1 0.5936543
## s2 0.4356242
## s3 0.2550535
```

Dans le cadre de la diversité neutre vue précédemment, tous les arbres d'une même espèce ont une similarité de 1, et tous les arbres d'espèces différentes ont une similarité nulle (la distance entre individus d'espèces différentes est infinie). Dans ce cas particulier, la banalité des arbres d'une espèce est la probabilité de l'espèce.

Dans le cas plus général d'individus regroupés en espèces placées dans un espace multidimensionnel, leur banalité est supérieure à leur probabilité : un individu est d'autant plus banal qu'il existe des individus proches de lui. Dans une approche d'écologie fonctionnelle, la diversité est celle de l'occupation des

niches. La niche occupée principalement par un individu l'est aussi un peu par les individus proches fonctionnellement.

La *rareté* est définie comme l'inverse de la banalité, l'entropie est la moyenne du logarithme de la rareté et la diversité est son exponentielle : la théorie est identique à celle de la diversité neutre.

La rareté des espèces est

```
1/B
```

```
##        [,1]
## s1 1.684482
## s2 2.295557
## s3 3.920747
```

à comparer avec l'inverse des probabilités

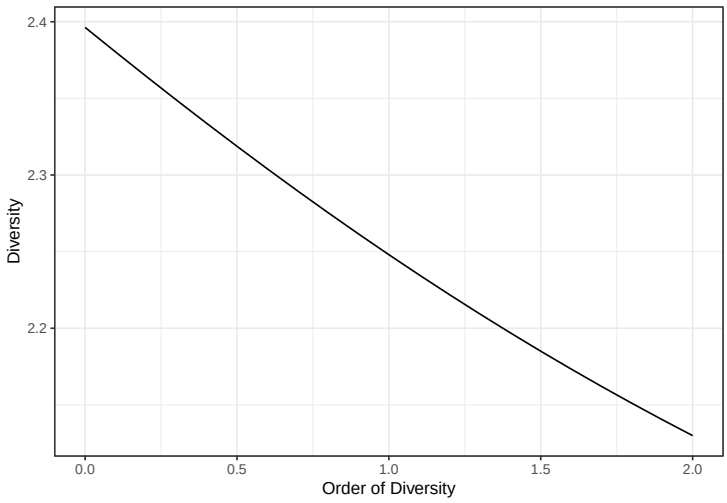
```
1/C %>%
    as.matrix
```

```
##    [,1]
## s1    2
## s2    4
## s3    4
```

Les espèces sont moins rares parce que chacune est « complétée » par les effectifs des espèces proches.

Le nouveau profil des diversité est le suivant :

```
CommunityProfile(Dqz, NorP = C, Z = Z) %>%
    autoplot
```



3 Application à une image

On peut définir une distance entre pixels de l'image dans l'espace défini par les différents canaux de couleur : par exemple, la distance euclidienne.

3.1 Création d'une image

L'image est composée de 4 points : un noir, un rouge, un vert, un bleu.

```
points_df <- data.frame(id = 1:4, x = rep(1:2, 2),
    y = rep(1:2, each = 2), R = c(0, 255, 0, 0), V = c(0,
        0, 255, 0), B = c(0, 0, 0, 255))
points_df
```

```
##   id x y   R   V   B
## 1  1 1 1   0   0   0
## 2  2 2 1 255   0   0
## 3  3 1 2   0 255   0
## 4  4 2 2   0   0 255
```

3.2 Distance entre points

Dans l'espace des couleurs, la distance entre points est la distance euclidienne.

3.2.1 Code R

```
# Nombre de points
points_n <- nrow(points_df)
# Matrice de distances dans l'espace des couleurs
points_d <- matrix(0, nrow = points_n, ncol = points_n)
# Calcul des distances
for (i in 1:(points_n - 1)) {
    for (j in (i + 1):points_n) {
        points_d[i, j] <- sqrt(sum((points_df[i, c("R",
            "V", "B")] - points_df[j, c("R", "V", "B")])^2))
    }
}
# Symétrisation
for (i in 1:(points_n - 1)) {
    for (j in (i + 1):points_n) {
        points_d[j, i] <- points_d[i, j]
    }
}
points_d
```

```
##      [,1]     [,2]     [,3]     [,4]
## [1,]    0 255.0000 255.0000 255.0000
## [2,]  255   0.0000 360.6245 360.6245
## [3,]  255 360.6245   0.0000 360.6245
## [4,]  255 360.6245 360.6245   0.0000
```

3.2.2 Code C

Une fonction vide (« void », c'est-à-dire qui ne renvoie pas de résultat) a pour arguments la matrice (`NumericMatrix`) des réflectances (entrée) et la matrice

de distance entre les points (sortie). Il est très difficile de définir une fonction qui renvoie une matrice en C++ parce que la taille de l'objet n'est pas connue : il est plus efficace de passer la matrice de sortie comme argument.

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

// [[Rcpp::export]]
void distance_c(NumericMatrix reflectances, NumericMatrix distances) {
  for (int i=0; i < (reflectances.nrow()-1); i++) {
    for (int j=i+1; j < reflectances.nrow(); j++) {
      // Calcul de la distance
      distances(i,j) = sqrt(sum(pow(reflectances(i, _)-reflectances(j, _), 2)));
      // Symétrisation
      distances(j,i) = distances(i,j);
    }
  }
}
```

L'appel de la fonction ne renvoie rien mais modifie la matrice de distance passée en argument : C++ traite les arguments par référence (la variable originale est modifiée) au contraire de R qui les traite par valeur (une copie de la variable originale est utilisée).

```
# Appel de la fonction
distance_c(
  # Le premier argument est une matrice qui contient une ligne par pixel et tous les canaux en colonne
  as.matrix(points_df[, c("R", "V", "B")]),
  # Le deuxième argument est la matrice de résultat, qui doit être crée avant et avoir la bonne taille
  # Si la taille n'est pas bonne, R crashe.
  points_d
)
points_d
```

```
##      [,1]     [,2]     [,3]     [,4]
## [1,]    0 255.0000 255.0000 255.0000
## [2,]  255   0.0000 360.6245 360.6245
## [3,]  255 360.6245   0.0000 360.6245
## [4,]  255 360.6245 360.6245   0.0000
```

3.3 Similarité

La similarité entre pixels est calculée à partir des distances. Ici, la similarité est le complément à 1 de la distance normalisée.

```
# Normalisation des distances
points_d_norm <- points_d/max(points_d)
# Similarité
(Z <- 1 - points_d_norm)
```

```
##           [,1]      [,2]      [,3]      [,4]
## [1,] 1.0000000 0.2928932 0.2928932 0.2928932
## [2,] 0.2928932 1.0000000 0.0000000 0.0000000
## [3,] 0.2928932 0.0000000 1.0000000 0.0000000
## [4,] 0.2928932 0.0000000 0.0000000 1.0000000
```

La banalité de chaque pixel est calculée comme la similarité moyenne avec les autres.

```
(points_ordinariness <- apply(Z, MARGIN = 1, mean))
```

```
## [1] 0.4696699 0.3232233 0.3232233 0.3232233
```

Sa rareté est l'inverse de sa banalité.

```
(points_rarity <- 1/points_ordinariness)
```

```
## [1] 2.129155 3.093836 3.093836 3.093836
```

3.4 Entropie et diversité

L'entropie de l'image est la moyenne du log de la rareté des pixels.

```
(points_entropy <- mean(log(points_rarity)))
```

```
## [1] 1.03599
```

Enfin, la diversité de l'image est l'exponentielle de son entropie.

```
(points_diversity <- exp(points_entropy))
```

```
## [1] 2.817895
```

Plus simplement, avec entropart :

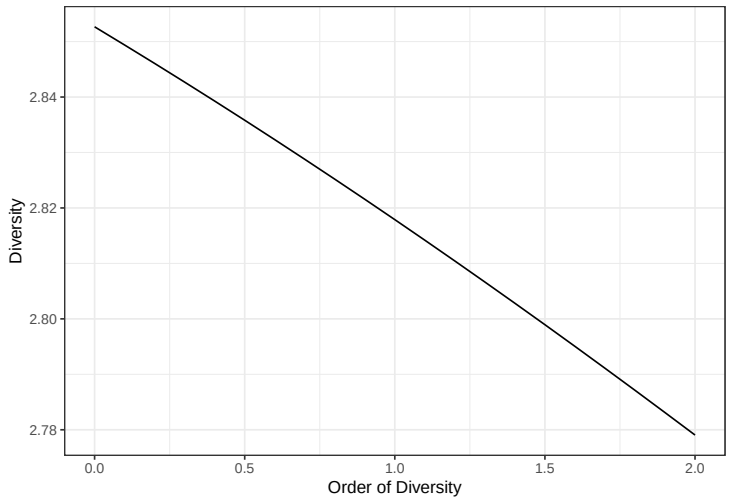
```
library("entropart")
Dqz(rep(1/points_n, points_n), q = 1, Z = Z)
```

```
##     None
## 2.817895
```

Le nombre effectif de pixels est légèrement inférieur à 3. Elle est comprise entre 1 si tous les pixels étaient de la même couleur et 4 s'ils étaient de 4 couleurs équidistantes.

On peut tracer un profil de diversité en fonction de q, qui donne un poids plus ou moins importants aux pixels rares.

```
autoplot(CommunityProfile(FUN = Dqz, NorP = rep(1/points_n,
    points_n), q.seq = seq(from = 0, to = 2, by = 0.1),
    Z = Z))
```



Cette diversité peut être comparée à celle des arbres obtenue par ailleurs.

4 Cours complet

A lire en ligne ou à télécharger : https://ericmarcon.github.io/MesuresBioDiv2/.

Références

Hill, M. O. Diversity and evenness : A unifying notation and its consequences. *54*(2), 427–432.

Hurlbert, S. H. The nonconcept of species diversity : A critique and alternative parameters. *52*(4), 577–586.

Jost, L. Entropy and diversity. *113*(2), 363–375.

Leinster, T. and C. Cobbold. Measuring diversity : The importance of species similarity. *93*(3), 477–489.

Shannon, C. E. A mathematical theory of communication. *27*(3), 379–423, 623–656.

Tsallis, C. What are the numbers that experiments provide ? *17*(6), 468–471.