24/04/2019

# Rapport final

## Quantification de la pollution par suivi de la biodiversité du lichen

Duy Anh Alexandre, Brunilde Bachelet, Margot Besseiche, Elvire Fauchet

TUTRICE : LAURE TURCATI
COORDINATEURS : YVES MECHULAM, PIERRE-DAMIEN COUREUX

# Remerciements

Nous tenons à remercier tout d'abord notre tutrice Mme. Laure Turcati pour sa disponibilité, sa bienveillance et ses conseils tout au long de notre travail durant cette année.

Nous remercions aussi Mme. Linda Seggi pour nous avoir fourni ses précédents travaux sur le protocole *Lichens Go !*, qui nous ont été fort utiles pour notre PSC.

Enfin, nous remercions tous les chercheurs avec qui nous avons été en contact pour l'obtention de données supplémentaires ou des précisions sur les données que nous avons utilisées.

# Table des matières

# I – Introduction

## A- Contexte de travail

Au cours de l'année 2017, l'observatoire participatif et transdisciplinaire de l'environnement urbain, PartiCitaE a lancé une nouvelle version de l'enquête participative *Lichens Go !*. Cette enquête permet à des volontaires de récolter et enregistrer des données sur la présence de lichens en France ; pour le moment, seules les données de Paris et Lyon sont disponibles. L'analyse de ces données doit permettre d'étudier la pollution, les lichens étant un très bon bioindicateur de qualité de l'air.

Notre travail consiste à analyser ces données pour vérifier si elles sont exploitables pour un diagnostic de la qualité de l'air et déterminer si le protocole peut être simplifié davantage, pour une ouverture plus vaste dans le cadre d'une enquête participative.

Nous avons donc dans un premier temps recherché les propriétés des lichens et le contexte dans lequel s'inscrit l'enquête *Lichens Go !*. Nous avons ensuite détaillé les différents calculs effectués sur notre base de données : diversité fonctionnelle, corrélation entre lichens et pollution, et pertinence du protocole.

## B- Les lichens

On appelle lichen une association stable entre une algue, aussi appelée photobionte, et un champignon ou mycobionte [1]. Le lichen est, malgré cette dualité, vu comme un corps homogène, nommé thalle [2] ; on parle de symbiose. En effet, l'association est ici nécessaire à la vie : le photobionte, souvent aquatique, est protégé de l'environnement terrestre, et procure l'énergie et les nutriments à l'association ; le mycobionte assure une fonction mécanique de protection et d'ancrage au support, et produit l'acide lichénique permettant de repousser les prédateurs.

La présence de lichens est importante pour la vie. Par exemple, en colonisant une roche, le lichen y emprisonne de la poussière, de l'eau… Cela érode la roche et permet la formation de sols fertiles.

Il existe 3 formes majeures de lichens :

- Crustacé : ce type est très lié au substrat via la surface en-dessous, il ne peut être retiré sans être détruit
- Foliacé : ce type est plat, dorso ventral, et se propage horizontalement vers l'extérieur ; il en existe une grande diversité
- Fruticuleux: ce type est constitué d'un disque basal d'attache, avec un ratio volume/surface élevé : les cycles sec/humide sont plus rapides pour cette forme, sa sensibilité plus élevée

Le développement d'une espèce donnée de lichen en un lieu donné dépend entre autres du climat, la résistance aux plus extrêmes températures étant souvent due au champignon. La colonisation d'un tronc est gouvernée par l'humidité, le pH, la lumière et les nutriments présents ; la croissance du thalle dépend aussi de l'environnement [1] (pluie, température, nutriments, position…) : ainsi, les lichens sur des surfaces non verticales ont un meilleur taux de croissance que ceux présents sur des surfaces verticales, car ils retiennent mieux l'humidité par exemple.

Parmi les facteurs environnementaux conditionnant le développement des lichens, on peut citer les suivants :

Lumière : facteur écologique vital pour un champignon en formation, dépendant de l'association symbiotique formée. La quantité de lumière reçue couplée à l'hydratation détermine la croissance du lichen. Ex. Lumière élevée + haute température = inhibition de la croissance.

Humidité / sécheresse : dans les régions tempérées, où les pluies sont intermittentes, les foliacés et fruticuleux sont avantagés.

Qualité de l'air **:** sensibilité du thalle à différents gaz. Par exemple, le $SO_2$ détériore la photosynthèse en convertissant une chlorophylle en phaeophytine.

Vent : plus il y a de vent, plus les crustacés ont l'avantage par rapport aux autres formes de lichens.

L'Homme ayant une action sur la qualité de l'air, il peut donc en avoir une sur le développement plus ou moins efficace de certaines espèces de lichens.

## C- Lichens et biosurveillance

La **biosurveillance** [2] est un terme désignant l'utilisation d'organismes pour obtenir des valeurs quantitatives sur certaines caractéristiques de l'environnement. Les informations nécessaires sont obtenues à partir du changement de distribution ou de comportement d'un organisme en particulier. Si la sélection de l'organisme à surveiller est correctement effectuée, ce-dernier permet une surveillance généralisée (obtention de données sur des lieux reculés ou inaccessibles par d'autres méthodes), une facilité de traitement et d'échantillonnage, et une diminution des coûts (par exemple, si les lichens donnent une bonne indication de la qualité de l'air, cela peut permettre de s'affranchir d'une construction de station de mesure, plus coûteuse).

Les lichens sont souvent considérés comme des bioindicateurs idéaux [1]. En effet, ils :

- Absorbent nutriments (et polluants) directement dans l'atmosphère.
- Ont des sensibilités différentes aux différents environnements.
- Possèdent une capacité à coloniser une grande part des espaces.

De plus, leur croissance est lente, leur espérance de vie élevée et leur morphologie maintenue. Autre avantage : leur surface qui facilite la rétention des particules permet l'augmentation de la concentration de certains éléments chimiques ce qui rend plus simple leur détection en cas de prélèvement.

Les effets de la pollution de l'air sur les lichens ont été démontrés depuis les années 1800 [4], et l'idée d'utiliser la présence ou la fréquence d'apparition d'un organisme biologique comme indicateur de qualité de l'air découle du travail de Nylander, dans les années 1860 [5]. Les lichens sont aujourd'hui l'un des bioindicateurs les plus répandus pour ce qui est de la pollution.

Les lichens ont d'ores et déjà été utilisés pour leur sensibilité aux faibles niveaux de sulfure, dans la détection de $SO_2$ dans l'air (années 1970). Plus récemment, des études ont montrées une corrélation entre les lichens et les niveaux d'oxyde de nitrogène (NOx).

Différentes méthodes de biosurveillance sont possibles : analyse de la composition des communautés, analyses biochimiques des tissus de lichens, ou étude de transplants. Dans la suite de ce PSC, c'est une analyse de composition des communautés qui a été menée.

## D- Les enquêtes participatives

Il existe de nombreuses manières d'associer des personnes non-expertes du domaine à la production de connaissances scientifiques, que ce soit dans le cadre de dispositifs académiques, plus cadrés et centralisés, ou dans le cadre de réseaux collaboratifs en émergence [3]. Cette idée de donner plus de place et de parole aux différents acteurs sociaux (professionnels, citoyens, usagers…) émerge de la forme qu'a pris la société : l'aspiration à des modèles délibératifs plus démocratiques et à des rapprochements science-société. Cette aspiration a permis l'essor de ce qui est appelé communément « sciences participatives ». Ces dispositifs existent aujourd'hui sous plusieurs formes, et pour plusieurs approches. Dans le cadre de la recherche et du suivi de la biodiversité à grande échelle, comme cela peut être le cas pour l'observation des lichens afin de déterminer la qualité de l'air, l'idée est de faire de la science avec les participants bénévoles, sans pour autant les impliquer dans toutes les étapes de construction de la recherche. Pour autant, une telle démarche peut s'avérer transformative pour la société, via l'implication des acteurs locaux en science.

L'approche des sciences participatives possède de nombreux avantages [4,5,6] : elle permet notamment d'éduquer et de sensibiliser les citoyens au monde scientifique, comme mentionné ci-dessus, tout en obtenant une base de données plus conséquente que ce qu'un chercheur seul, ou un groupe de chercheurs, auraient pu rassembler. C'est également l'un des seuls moyens d'obtenir des données de manière suffisamment répétée pour permettre un suivi du bioindicateur.

Cependant, la participation de « non-scientifiques » demande souvent une simplification des protocoles d'obtention de données, ce qui peut les rendre moins fiables que le plus petit nombre collecté par des chercheurs [6]. Ce potentiel manque de fiabilité est compensé par un grand nombre de données. Il s'agit donc pour les chercheurs à l'origine de la recherche de trouver le juste équilibre entre un protocole trop compliqué, qui amène à une récolte de données lacunaires voire erronées, et un protocole trop simplifié, qui ne permet pas de tirer des conclusions satisfaisantes.

## E- Orientation de notre travail

Nous avons récupéré les données de Lyon et Paris récoltées selon le protocole de PartiCitaE (détaillé ci-dessous). Nos analyses et interprétations sont liées essentiellement aux deux questions suivantes, centrales de notre travail :

- **Les données récoltées vérifient-elles la corrélation entre qualité de l'air et espèce de lichen présente ? Le protocole permet-il de tirer des conclusions satisfaisantes sur la qualité de l'air ?**
- **Le protocole peut-il être simplifié ?**

Les méthodes employées et analyses réalisées sont détaillées dans la suite de ce travail.

Nous avons également fait un parallèle avec le protocole OPAL [7], débuté en décembre 2007 en Grande-Bretagne : il s'agit également d'une enquête participative s'intéressant au lien entre lichens et taux de NOx dans l'air, et ayant permis l'obtention de résultats satisfaisants et précis.

# II - Matériels et méthodes

## A- Protocole

Le protocole à l'origine de nos données est celui de *Lichens Go !*, un protocole suivant la norme Afnor NF en 16413. Détaillé pour les bénévoles et chercheurs récoltant des données, il est disponible sur le site Web de PartiCitaE [8].
La démarche à suivre est la suivante :

Il faut choisir trois arbres espacés de 2 à 10 mètres les uns des autres, et d'une circonférence supérieure à 30 centimètres. L'espacement est important, l'idée étant d'éviter l'effet de « sous-bois ». Les arbres doivent également être droits et sans branches basses, afin que l'écoulement soit homogène sur l'ensemble du tronc.

Ensuite, les bénévoles doivent estimer l'abondance des différentes espèces de lichens sur le tronc. Ils utilisent une grille de 5 carrés, disponible sur le site de l'enquête : c'est leur référence pour mesurer l'étendue de colonisation du tronc. Ils choisissent la face la plus riche en lichens sur chaque arbre, et observent cette face plus les deux faces latérales (en utilisant l'orientation Nord, Est, Sud et Ouest pour entrer les données sur le site). Ils identifient ensuite les lichens présents dans chacun de ces carrés. Certains bénévoles sont spécialistes en lichens et peuvent précisément identifier les espèces de lichen présentes, mais la plupart peuvent seulement identifier le type de lichen parmi les trois grandes catégories – crustacé, foliacé ou fruticuleux.

Enfin, l'observateur doit entrer ces données sur le site internet de l'enquête en indiquant la position géographique des trois arbres et leur circonférence.

## B- Données environnementales et calculs d'indices

Nous avons utilisé pour nos analyses différentes variables issues de l'application du protocole *Lichens Go !* à 197 arbres répartis sur 68 sites à Paris, Lyon (et un seul site à Lille).

Les paramètres qui nous ont été utiles peuvent être répartis en 3 catégories :

- **Les paramètres concernant l'arbre sur lequel est effectué le relevé :**
  - La circonférence de l'arbre
  - La longitude de l'arbre
  - La latitude de l'arbre
  - L'essence de l'arbre (très peu renseignée)

- **Les paramètres concernant le site du relevé et son environnement :**
  - Le nom de la station de relevé de NO2 la plus proche
  - La ville du relevé : Paris/Lyon/Lille
  - La moyenne des longitudes des arbres du site
  - La moyenne des latitudes des arbres du site
  - La moyenne des circonférences des arbres du site
  - La distance entre le site et le centre-ville
  - La distance entre le site et la station de mesure du NO2
  - La distance entre le site et la forêt la plus proche

- La distance entre le site et la pièce d'eau la plus proche. Ce paramètre est une manière approximative de mesurer le taux d'humidité dans l'environnement du lichen.
- (La distance à la pièce d'eau la plus proche) / (La taille de cette pièce d'eau) (DistxArea_divsumArea) *Cette donnée n'a été calculée que pour les relevés de Paris*
- Le pourcentage de territoire urbain dense dans un rayon de 50m/100m/500m autour du site

*(Toutes les données concernant l'environnement viennent de l'inventaire Corine Land Cover 2012, accessibles sur le site land.copernicus.eu. Toutes celles concernant les densités de population viennent de l'INSEE.)*

- Le taux de NO2 moyen enregistré entre 2014 et 2017 à la station de NO2 la plus proche (*Les stations de NO2 utilisées sont celles d'AirParif pour Paris, et AtmoAURA pour Lyon.)*

- **Les paramètres de diversité calculés à partir des relevés :**

Lors de l'incorporation des données issues des relevés de Lyon à celles de Paris, il nous était indispensable de recalculer certains paramètres en fonction de la présence de lichens sur les grilles d'observation.

Tous ces calculs tirent leurs données d'un tableau brut où sont répertoriées les présences de différentes espèces de lichen sur chaque grille (voir le dispositif du protocole PartiCitaE partie II-A).

- Pourcentage de foliacés/fruticuleux/crustacés parmi les lichens relevés,

défini comme le rapport du nombre de carrés contenant le thalle particulière sur le nombre de carrés contenant un lichen.

- Diversité :

Le premier indice de diversité, dénommé « Diversité », comptabilise le nombre d'espèces différentes présentes sur chaque arbre ou site. Pour rappel, un site comporte de deux à trois arbres.

- Fréquence :

Cet indice de diversité, dénommé « Fréquence », est calculé comme la somme des fréquences de chaque espèce présente sur une face/un arbre/un site.

Pour une face, la Fréquence d'une espèce est le rapport du nombre de carrés où l'espèce est présente sur le nombre total de carrés. Le dénominateur est toujours 5 dans ce cas. Comme c'est une somme, l'indice peut être plus grand que 1, alors que l'appellation Fréquence peut laisser penser le contraire.

Pour un arbre, où 3 faces sont observées (protocole uniformisé), la Fréquence correspond à la moyenne des fréquences sur les 3 faces.

Pour un site, la Fréquence est la moyenne des fréquences correspondant aux arbres présents sur ce site.

### o Indice de Shannon :

Cet indice, couramment utilisé en biologie, donne une idée du nombre d'espèces du milieu et de la répartition des individus au sein de ces espèces. Il est calculé selon la formule :

$$H = -\sum_{i=1}^{S} p_i log_2(p_i)$$

où $p_i$ est la proportion d'une espèce *i* par rapport au nombre total d'espèces (S) dans le milieu d'étude.

C'est cet indice que nous utilisons principalement pour faire notre analyse statistique inférentielle.

Les scripts du code R pour le calcul de ces indices sont en annexe 1.

### o La diversité fonctionnelle

La diversité fonctionnelle peut être définie comme la diversité de certains "traits fonctionnels", souvent reliés à la notion de résilience des écosystèmes. Les traits utilisés dans notre étude sont au nombre de cinq : le pH du lichen, la photophilie, la poléotolérance, l'aridité et l'eutrophisation. Il existe différentes façons de mesurer la diversité fonctionnelle. Dans notre cas, nous avons calculé les distances phénotypiques entre espèces, qui ne sont qu'une somme des carrés de la différence entre deux espèces, traits par traits.

Ensuite, pour chaque arbre (et site), nous calculons sa diversité fonctionnelle selon la formule de l'entropie quadratique de Rao [18] :

$$FD = \sum_{i,j}^{n} d_{ij} p_i p_j$$

où $d_{ij}$ est la distance fonctionnelle entre l'espèce i et j, et $p_i$ est l'abondance relative de l'espèce i, ici prise égale à sa "fréquence" calculée ci-dessus puis normalisée.

Nous remarquons que le calcul de la diversité fonctionnelle nécessite l'identification des espèces précises de lichen, ce qui paraît difficile lorsque les données sont fournies par des personnes non expertes. En effet, à Lyon, cette identification était impossible, nous avons donc appliqué les analyses de corrélation seulement sur la base de données à Paris.

Pour ces calculs, nous avons pu nous inspirer des indices de diversité et des modèles testés par Mme Linda Seggi, qui a effectué un stage à l'OSU Ecce Terra avant que nous ne commencions ce PSC.

Par la suite, la variable à expliquer que nous avons principalement utilisée pour les analyses est l'indice de Shannon par arbre. L'utilisation de cet indice est presque équivalente à d'autres indices de diversité calculés, comme le montre ce tableau de corrélation de Pearson :

| | Diversité | Fréquence | Shannon |
|---|---|---|---|
| **Diversité** | 1 | 0,82 | 0,93 |
| **Fréquence** | 0,82 | 1 | 0,66 |
| **Shannon** | 0,93 | 0,66 | 1 |

## C- Les analyses

Nous avons utilisé pour l'analyse des données une approche inférentielle, en se basant sur la régression linéaire multiple.

En statistique, la **régression linéaire multiple** est une méthode de régression mathématique étendant la régression linéaire simple pour décrire les variations d'une variable à expliquer, en fonction des variations de plusieurs variables explicatives [9][10].

Dans notre cas, la variable à expliquer est la diversité des lichens, qui se manifeste à travers des paramètres que nous avons calculés (*expliqué plus haut*), par exemple l'indice de diversité de Shannon.

Les variables explicatives sont nombreuses, et leur choix fait l'objet d'une problématique entière dans le domaine des statistiques de données. Comme nous souhaitons expliciter un lien entre le degré de pollution (taux de dioxyde d'azote atmosphérique) et la diversité lichénique, le taux de NO2 doit être présent parmi ces variables.

Notre choix du modèle linéaire résulte de sa simplicité. De plus, même si le lien semble relativement faible quand on affiche l'indice de Shannon en fonction du taux de NO2 (Fig.1), dans les modèles que nous avons développés, l'analyse de la variance (ANOVA)

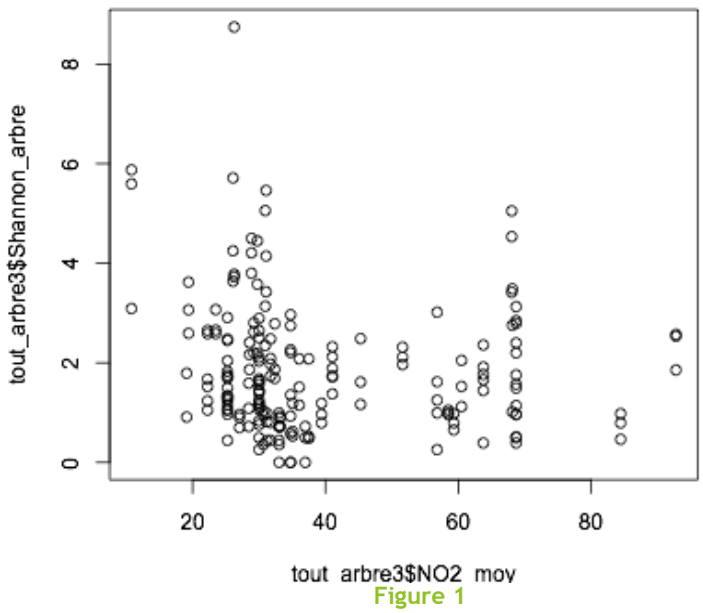


Figure 1

démontre des corrélations entre les variables.

## 1° Appréhension des données : Analyse en Composantes Principales

L'ACP est une méthode formalisée dans les années 30 par le statisticien américain Harold Hotelling.

Les K variables et leurs N réalisations sont représentées dans une matrice X (où elles correspondent donc à K vecteurs dans $R^N$)

$$X = \begin{bmatrix} X_{1,1} & \cdots & X_{1,N} \\ \vdots & \ddots & \vdots \\ X_{K,1} & \cdots & X_{K,N} \end{bmatrix}$$

Cette matrice est centrée et réduite dans le calcul effectué par la fonction prcomp de R [11], ce qui est nécessaire car les différentes variables ne sont pas du tout homogènes, en la matrice :

$$M = \begin{bmatrix} \frac{X_{1,1} - \underline{X_1}}{\sigma(X_1)} & \cdots & \frac{X_{1,N} - \underline{X_N}}{\sigma(X_N)} \\ \vdots & \ddots & \vdots \\ \frac{X_{K,1} - \underline{X_1}}{\sigma(X_1)} & \cdots & \frac{X_{K,N} - \underline{X_N}}{\sigma(X_N)} \end{bmatrix}$$

La matrice $C = \frac{1}{K} \cdot M^T \cdot M$ est donc la matrice de corrélation des K variables.

L'ACP consiste à diagonaliser C (symétrique et réelle, donc diagonalisable selon le théorème spectral) ; on appelle les vecteurs propres trouvés les Composantes Principales (« PC » dans les tableaux en annexe 2), et ces vecteurs propres sont des combinaisons linéaires de nos variables d'origine. Ils correspondent à K variables indépendantes, dont la somme des variances (i.e. de leurs valeurs propres) est maximale [12].

Les Composantes Principales ayant les valeurs propres les plus élevées permettent donc d'expliquer la plus grande proportion de la variabilité de notre échantillon, et représenter nos variables $X_i$ dans l'espace de ces Composantes Principales permet de distinguer parmi les $X_i$ lesquelles sont très corrélées, et lesquelles ont la plus grande part de responsabilité dans la variabilité de l'échantillon. On peut ainsi réduire le nombre de variables utilisées ensuite dans la recherche d'un modèle.

## 2° Un peu de théorie sur la régression linéaire

La régression linéaire simple (une variable explicative) sur un ensemble de données Y est définie ainsi :

On note Y la variable aléatoire réelle à expliquer et X la variable explicative ou effet fixe (exogène). (En pratique ce sont des échantillons d'observation, dont nous n'avons pas besoin de connaître la loi.)
Le modèle revient à supposer qu'en moyenne Y est une fonction affine de X. L'écriture du modèle suppose implicitement une notion préalable de causalité dans le sens où Y dépend de X car le modèle n'est pas symétrique. Ce qui s'écrit mathématiquement :

$Y = \beta_0 + \beta_1 X + \epsilon$ où $\epsilon$ (les résidus) suit une distribution gaussienne centrée en 0.

L'estimation des paramètres $\beta_0$ et $\beta_1$ est obtenue par minimisation de la somme des carrés des résidus i.e. écarts entre observations et modèle (moindres carrés). Le critère des moindres carrés s'écrit :

$$\min_{\beta_0,\beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 = \min_{\beta_0,\beta_1} \sum_{i=1}^{n} {e_i}^2$$

On pose :

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i,$$

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2, \qquad s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2,$$

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}), \qquad r = \frac{s_{xy}}{s_x s_y};$$

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2.$$

Total sum of squares : $SST = (n-1)s_y^2$

Regression sum of squares : $SSR = (n-1)\frac{s_{xy}^2}{s_x^2}$

Error sum of squares : $SSE = (n-2)s^2$

Et on vérifie que SST = SSR + SSE

On appelle **coefficient de détermination** la quantité :

$$R^2 = r^2 = \frac{s_{xy}^2}{s_x^2 s_y^2} = 1 - \frac{n-2}{n-1}\frac{s^2}{s_y^2} = \frac{\text{SSR}}{\text{SST}}$$

Il exprime le rapport entre la variance expliquée par le modèle et la variance totale. **Plus il est proche de 1, plus la qualité de la régression est bonne.**

Régression linéaire multiple [13] :
C'est le même modèle à plusieurs variables explicatives $X_1$, $X_2$, ….

$$Y = a_o + a_1 X_1 + a_2 X_2 + \ldots + a_n X_n$$

Les autres définitions sont des extensions naturelles de celles appliquées à la régression linéaire simple.

**Les hypothèses du modèle :**

Les hypothèses de validation du modèle linéaire (simple ou multiple) portent sur les résidus :

- Ils doivent suivre une distribution gaussienne. Ceci peut se vérifier en traçant un diagramme des quartiles (QQ plot gaussien) [14].
- Homoscédasticité des résidus (i.e. les résidus ont la même variance quel que soit le groupe considéré, ou quelle que soit la valeur de la variable explicative considérée).

Néanmoins, ces critères ne sont dans la plupart des données statistiques réelles que partiellement respectés. Aussi, lorsque le nombre d'observations n est élevé, le non-respect de ces hypothèses n'invalide pas pour autant l'utilisation justifiée d'un modèle de régression linéaire [15].

### 3° Modèle non-linéaire : General Additive Model

Un autre type de modèle est le GAM (General Additive Model), qui s'écrit :

$$g(y) = \beta_0 + f_1(x_1) + f_2(x_2) + f_3(x_3) + \varepsilon, \varepsilon \sim N(0, \sigma^2)$$

En complément d'une régression linéaire usuelle (qui nous donnera des coefficients paramétriques), une fonction de liaison est appliquée à la variable à expliquer (g) ; et les fonctions $f_1$ et $f_2$ permettent de lisser les variables $x_1$ et $x_2$, et sont déterminées à partir d'une famille de fonctions, par maximum de vraisemblance. La distribution de ε peut aussi être modifiée (loi de poisson, exponentielle…) [16].

La fonction que nous utilisons pour calculer ces modèles est la fonction gam de la library mgcv [17], où :

- g est par défaut l'identité,
- $f_1$, $f_2$ etc. sont des « splines », i.e. des fonctions polynomiales par morceaux,
- l'estimation de $f_1$ et $f_2$ peut être contrôlée pour obtenir des fonctions qui ne varient pas trop fortement et éviter un sur-ajustement des données.

En plus des coefficients paramétriques, la fonction gam renseignera le taux de déviance expliqué par le modèle, ainsi que, pour chaque variable explicative lissée :

- edf, une estimation du nombre de degrés de liberté ajoutés par le lissage spline cubique des termes. Plus il est élevé, plus il y a un risque de sur-ajustement des données.
- Une p-value pour chaque variable explicative correspondant à l'hypothèse : « La variable explicative lissée est indépendante de la variable explicative »

### 4° Critère de sélection, AIC

La sélection des variables à intégrer dans le modèle est très importante. Elle détermine fortement la qualité de l'analyse de corrélation :

- La sélection de « mauvaises » variables (décorrélation avec la variable à expliquer) fausse et baisse drastiquement la qualité du modèle.
- L'utilisation de variables redondantes i.e. fortement corrélées entre elles rajoute des effets d'interaction dans le modèle, effets qui ne sont pas toujours souhaités.
- Généralement, on souhaite limiter le nombre de variables explicatives, pour ne pas complexifier inutilement le modèle. Dans notre cas, plus il y a de paramètres, moins le rôle de la pollution au NO2 est visible dans le modèle.

Il peut d'abord être intéressant de visionner les corrélations entre couples de variables (avec un test de corrélation de Pearson, très simple à effectuer avec R, ou en affichant une variable en fonction de l'autre…) et de faire une première sélection visuelle des variables.

| | **Dist Centre** | **Dist stati** | **Dist Foret** | **Dist Water** | **altitude** | **perc_dense _500m** | **Diversité** | **Fréquence** | **Shannon** |
|---|---|---|---|---|---|---|---|---|---|
| **NO2_ moy** | -0,28 | 0,124 | 0,012 | -0,093 | -0,208 | -0,020 | -0,138 | -0,111 | -0,130 |

Figure 2 : Test de corrélation de Pearson ayant pour référence le taux de NO2

Un critère souvent utilisé pour la sélection des variables est le critère d'information d'Akaike (AIC), défini par :

$$AIC = n \times log(\frac{SSE}{n}) + 2(l+1)$$

avec n le nombre d'observations, SSE la somme des carrés des résidus du modèle estimé et l le nombre de variables explicatives.
On cherche alors à choisir les variables de façon à minimiser ce critère.

Intuitivement, si le modèle s'ajuste mieux aux données, la somme des carrés des résidus (SSE) va diminuer et donc le terme va également diminuer. Le deuxième terme de la somme traduit une pénalisation de l'ajout de variables supplémentaires.

Dans notre cas, nous avons utilisé une méthode dite descendante. L'idée est de commencer par un modèle linéaire avec beaucoup de paramètres explicatifs, puis de les enlever au fur et à mesure pour arriver à une valeur d'AIC optimale.

### 5° Analyse de la variance (ANOVA)

Cette méthode, souvent couplée à la régression linéaire, teste la significativité globale du modèle. La F-valeur est définie par :

$$F = (n-2)\frac{SSR}{SSE}$$

Elle exprime la corrélation linéaire entre la variable de réponse et la variable explicative. Plus F est grand, plus cette corrélation est marquée. Le seuil déterminant correspond au $95^e$ centile de la loi de Fisher [14]. Sous l'hypothèse $H_0$ : « Il n'existe pas de corrélation linéaire entre Y et $X_i$ », la probabilité d'obtenir une valeur d'au moins F est donnée par la dernière colonne « Pr(>F) » du tableau d'ANOVA de R. Si cette probabilité est très faible (le seuil de référence est habituellement inférieur à 5%), l'hypothèse $H_0$ est rejetée, donc il existe une corrélation entre Y et $X_i$.

# III - Résultats obtenus

## A- Liens entre un indice de diversité des lichens et le taux de $NO_2$

### 1- Recherche des paramètres influents par ACP et critère AIC

**L'ACP** effectuée sur la plupart des variables quantitatives des données de Lyon et Paris qui ont un sens physique (Annexe 2) révèle 3 Composantes Principales qui sont responsables de 47% de la variance de l'échantillon, et qui ont toutes une variance supérieure à 3 (Annexe 2).

Nous décidons donc de représenter les variables normalisées de nos données dans l'espace formé par ces 3 Composantes Principales. La fusion des variables qui sont à peu près colinéaires dans cet espace en une seule variable, et la suppression des variables qui ont une composante de très faible norme dans cet espace, nous pousse à conserver les variables :

- L'altitude de l'arbre (altitude)
- La distance à la pièce d'eau la plus proche (DistWater_)
- Le pourcentage de territoire urbain dense dans un rayon de 500m (perc_dense_500m)
- La fréquence des foliacés parmi l'ensemble des lichens de l'arbre (freq_fol_arbre)
- Le taux de NO2 moyen enregistré par la station la plus proche entre 2014 et 2017(NO2_moy)

La même analyse réalisée avec les entrées regroupées pas site donne des projections moins lisibles, mais qui semblent confirmer le choix de nos 5 variables principales (Annexe 2).

Nous cherchons à exprimer les variables exprimant la diversité des lichens en fonction des variables environnementales, nous réalisons donc un test ANOVA sur le modèle qui explique l'indice de Shannon et la proportion de foliacés sur l'arbre, par la moyenne des

latitudes du site, le logarithme de la distance à la pièce d'eau la plus proche, le pourcentage de territoire urbain dense aux 500m, et le taux de NO2 moyen :

lm(formula = Shannon_arbre + freq_fol_arbre ~ altitude + DistWater_ + perc_dense_500m + NO2_moy, data = arbre_tout_moi)

| | Estimates | Standard Error | Anova : F value | Anova : Pr>F |
|---|---|---|---|---|
| Intercept (ordonnée à l'origine) | 2.466e+00 | 3.480e-01 | | |
| altitude | 2.803e-04 | 1.313e-03 | 0.0389 | 0.843886 |
| DistWater_ | 2.059e-05 | 2.032e-04 | 0.0003 | 0.986162 |
| perc_dense_500m | -6.390e-03 | 2.431e-03 | **7.0969** | **0.008522** |
| NO2_moy | 4.418e-03 | 4.870e-03 | 0.8229 | 0.365723 |

| | |
|---|---|
| F-statistics | 1.99 |
| P-value | 0.09864 |
| R adjusted | 0.02386 |
| Max residuals (valeur absolue) | 3.8984 (en indice de Shannon) |

Ce modèle ne révèle que très peu de corrélations entre les différentes variables.

Pour la méthode de sélection par **AIC**, nous commençons avec un modèle exhaustif avec un maximum de variables explicatives. Comme pour l'ACP, nous avons tout de même fait une première sélection en ne prenant en compte que les variables avec un sens physique et susceptibles d'avoir un effet sur la diversité. Ces variables sont : la moyenne des latitudes du site, la distance au centre-ville, le taux de $NO_2$moyen, la distance à la station de mesure du $NO_2$, la distance à la forêt, la distance au prochain point d'eau, l'altitude et le pourcentage de couverture urbaine dense dans les 500m. De plus, nous avons appliqué un logarithme aux distances pour accentuer le poids des distances proches. Nous avons choisi d'ajouter un terme d'interaction entre le taux de $NO_2$ moyen et la distance à la station de $NO_2$, car c'est le choix qui a été fait par Mme. Linda Seggi préalablement.

En appliquant la fonction step de R, les variables retenues sont:

- Le logarithme de la distance au point d'eau le plus proche
- L'altitude du site

- Le logarithme de la distance au centre-ville
- Le taux de $NO_2$ moyen
- Le logarithme de la distance à la station de mesure du $NO_2$
- Le taux d'urbanisation dense dans les 500m

Nous avons testé un modèle linéaire avec ces variables, en ajoutant des logarithmes sur les distances, et avons obtenu des corrélations avec une significativité concentrée sur les paramètres log(DistCentre) et altitude :

lm(formula = Shannon_arbre ~ DistCentre + log(DistForet_) + altitude + perc_dense_500m + NO2_moy*log(Dist_stati), data = arbre_tout_moi)

| | Estimates | Standard Error | Anova : F value | Anova : Pr>F |
|---|---|---|---|---|
| Intercept (ordonnée à l'origine) | 0.9956 | 1.4520 | | |
| log(DistCentre) | 0.2498 | 0.1230 | **6.3641** | 0.0126 |
| NO2_moy | -0.0213 | 0.0138 | 0.4232 | 0.5163 |
| log(Dist_stati) | -0.3759 | 0.1144 | 0.7746 | 0.3801 |
| altitude | 0.0044 | 0.0015 | **8.9485** | 0.0032 |
| perc_dense_500m | -0.0051 | 0.0029 | 3.3060 | 0.0709 |
| NO2_moy :log(Dist_stati) | 0.0054 | 0.0024 | 5.0403 | 0.0261 |

| | |
|---|---|
| F-statistics | 4.143 |
| P-value | 0.0006858 |
| R adjusted | **0.1043** |
| Max residuals (valeur absolue) | 3.8663 (en indice de Shannon) |

## 2- Le modèle de Linda Seggi appliqué à la diversité fonctionnelle

En reprenant le modèle linéaire qui a été retenu par Linda Seggi (Annexe 3), mais en prenant la Diversité Fonctionnelle comme variable à expliquer, nous obtenons le modèle :

Formule :
div_fcn~NO2_moyenne*log(Dist_stati)+percentage.500m+log(DistxArea_divsumArea)

| | Estimates | Standard Error | Anova : F value | Anova : Pr>F |
|---|---|---|---|---|
| Intercept (ordonnée à l'origine) | 574.4688 | 107.2602 | | |
| NO2_moy | -6.9460 | 2.2499 | **4.7349** | **0.037299** |
| log(Dist_stati) | -92.6413 | 24.3109 | 2.4939 | 0.124442 |
| percentage.500m | -1.5183 | 0.3952 | **12.0047** | **0.001575** |
| log(DistxArea_divsumArea) | -14.0684 | 6.1860 | **3.8098** | 0.060036 |
| NO2_moy:log(Dist_stati) | 1.3704 | 0.4577 | **8.9662** | **0.005365** |

| | |
|---|---|
| F-statistics | 6.402 |
| P-value | **0.000342** |
| R adjusted | **0.4287** |
| Max residuals (valeur absolue) | 179.41 (en indice de diversité fonctionnelle) |

En comparant au modèle où la variable à expliquer est l'indice de Shannon (voir annexe 3), nous observons :

- que le **coefficient de détermination est significativement meilleur**: on passe de R=0.26 à R=0.43
- **l'effet de la distance au point d'eau le plus proche**, proxy de l'humidité de l'environnement ambiant, est devenu significatif (voir le tableau ANOVA), ce qui constitue un résultat particulièrement intéressant. Ceci pousse à l'utilisation d'un tel indice de diversité dans un cadre plus répandu. Cependant, comme nous l'avions

indiqué, le calcul nécessite la connaissance précise des espèces de lichens observées, ce qui peut s'avérer difficile en pratique.

### 3- Modèle simple : taux de $NO_2$ comme seul paramètre

Une autre idée de départ serait d'essayer d'expliquer l'indice de Shannon seulement à partir du logarithme du taux de $NO_2$ moyen entre 2014 et 2017.

Call:

lm(formula = Shannon_arbre ~ log(NO2_moy), data = tout_arbre3)

| | Estimates | Standard Error | Anova : F value | Anova : Pr>F |
|---|---|---|---|---|
| Intercept (ordonnée à l'origine) | 4.1288 | 0.8037 | | |
| log(NO2_moy) | -0.6355 | 0.2227 | **8.1409** | **0.004805** |

| | |
|---|---|
| F-statistics | 8.141 |
| P-value | 0.004805 |
| R adjusted | **0.03586** |
| Max residuals (valeur absolue) | 6.6916 (en indice de Shannon) |

Ce modèle donne des résultats bons pour l'analyse de la variance mais médiocres pour ce qui est de la part expliquée par le modèle dans la variance totale.

### 4- Approfondissement de ce modèle : ajout de l'altitude, de la distance au centre-ville, à la forêt la plus proche et à la station de mesure du $NO_2$

Le modèle précédent n'étant pas suffisant, et en s'appuyant sur les autres modèles déjà présentés, nous décidons d'ajouter parmi les paramètres explicatifs le logarithme de la distance à la forêt la plus proche et à la station de mesure du $NO_2$ la plus proche. Un terme d'interaction est ajouté entre ce dernier et le logarithme du taux de $NO_2$, car la

corrélation de Pearson entre ces deux variables est relativement significative (voir figure 2).

Formule

Shannon_arbre ~ log(NO2_moy) * log(Dist_stati) + log(DistForet)

| | Estimates | Standard Error | Anova : F value | Anova : Pr>F |
|---|---|---|---|---|
| Intercept (ordonnée à l'origine) | 15.94941 | 2.34853 | | |
| log(NO2_moy) | -3.12025 | 0.64580 | **9.8759** | 0.0019460 |
| log(Dist_stati) | -1.89025 | 0.41230 | **8.7886** | 0.0034238 |
| log(DistForet) | -0.37470 | 0.09598 | **15.5507** | 0.0001133 |
| log(NO2_moy):log(Dist_stati) | 0.50204 | 0.11408 | 19.3656 | 1.806e-05 |

| | |
|---|---|
| F-statistics | 13.4 |
| P-value | 1.266e-09 |
| R adjusted | **0.2052** |
| Max residuals (valeur absolue) | 5.3857 (en indice de Shannon) |

Ce modèle donne des résultats tout à fait satisfaisants, mais nous avons poussé plus loin pour essayer de maximiser la part de la variance expliquée par le modèle.

En ajoutant les paramètres « Altitude » et « Distance au centre-ville », nous sommes arrivés à ce modèle qui est le dernier retenu en ce qui concerne les modèles linéaires :

- Variable à expliquer : l'indice de Shannon par arbre
- Variable explicative : le taux de $NO_2$, la distance au centre-ville, la distance à la station de mesure de $NO_2$, la distance à la zone forestière la plus proche et l'altitude du site de relevés. Utiliser l'altitude est, dans notre cas, une manière d'intégrer le lieu du relevé dans le modèle, l'altitude moyenne parisienne étant très différente de l'altitude moyenne lyonnaise.

Nous comparons ce modèle à celui qui a été utilisé pour les données de Paris par Linda Seggi:

model21 <- lm(Shannon ~ NO2_moyenne*log(Dist_stati) + percentage.500m+ log(DistxArea_divsumArea), data=parametre)

ainsi qu'aux modèles obtenus avec les méthodes ACP et AIC.

Pour notre modèle, nous n'avons pas retenu :

- Le pourcentage de couverture urbaine dans les 500m, car il y avait beaucoup de valeurs manquantes (33). Nous avons aussi constaté que ce paramètre était très influent sur la diversité. Sa présence dans le modèle baissait le coefficient de détermination et éclipsait le rôle des autres variables, notamment celui du taux de NO2 atmosphérique qui nous intéresse.
- Pour DistxArea_divsumArea, la superficie du point d'eau manquait pour les données de Lyon. C'est le paramètre « distance à la forêt la plus proche» qui le remplace comme proxy du taux d'humidité, car il donnait des résultats plus significatifs que l'utilisation de la distance au point d'eau le plus proche (peut-être parce que la surface du point d'eau n'est pas prise en compte).
- Nous n'avons pas retenu non plus le paramètre « fréquence en foliacés » car il faisait baisser la significativité globale du modèle, sans augmenter la part de variance expliquée. Les foliacés ne sont pas les plus fragiles en présence de $NO_2$ (grande capacité d'eutrophisation).
- Nous avons ajouté le paramètre « Altitude », suite à l'intégration des données sur deux villes aux reliefs distincts: Paris et Lyon.

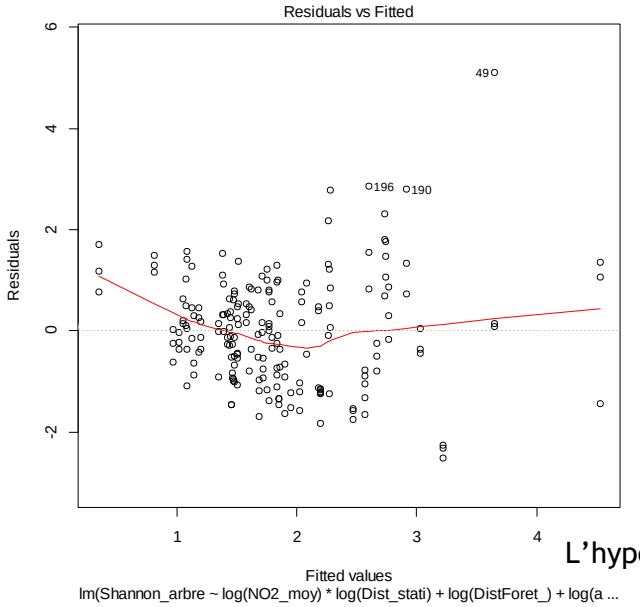
Formule : Shannon_arbre ~ log(NO2_moy) * log(Dist_stati) + log(DistForet_) + log(altitude) + log(DistCentre)

| | Estimates | Standard Error | Anova : F value | Anova : Pr>F |
|---|---|---|---|---|
| Intercept (ordonnée à l'origine) | 5.7083 | 2.924 | | |
| log(NO2_moy) | **-2.0390** | 0.6094 | **10.8891** | 0.0011594 |
| log(Dist_stati) | **-1.5319** | 0.3787 | **9.6903** | 0.0021451 |
| log(DistForet) | **-0.2520** | 0.0915 | **17.1462** | 5.244e-05 |
| log(altitude) | **0.5440** | 0.1464 | **6.0627** | 0.0147168 |
| log(DistCentre) | **0.3946** | 0.1030 | **21.9494** | 5.392e-06 |
| log(NO2_moy):log(Dist_stati) | 0.3842 | 0.1046 | 14.6290 | 0.0001785 |

| | |
|---|---|
| F-statistics | 13.39 |
| P-value | 1.335e-12 |
| R adjusted | **0.2792** |
| Max residuals (valeur absolue) | 5.0974 (en indice de Shannon) |

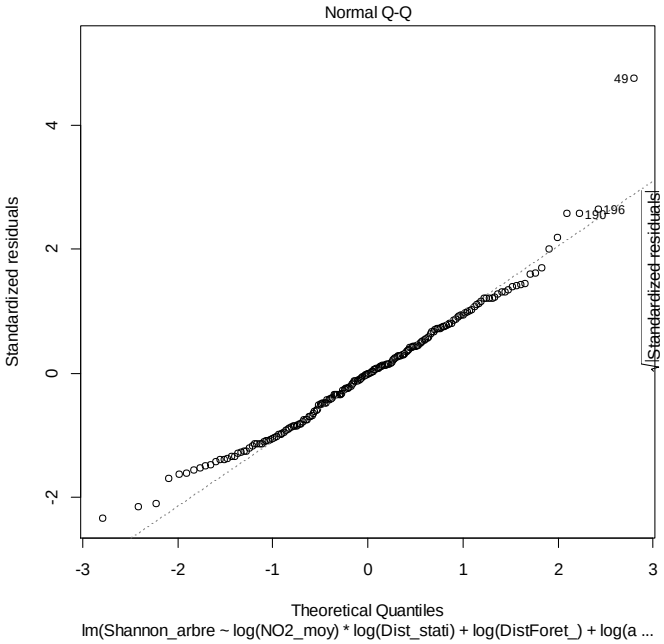
Comme pour les modèles précédents, nous avons étudié les résidus, dans le but de repérer si un relevé particulièrement anormal n'altérait pas nos résultats.

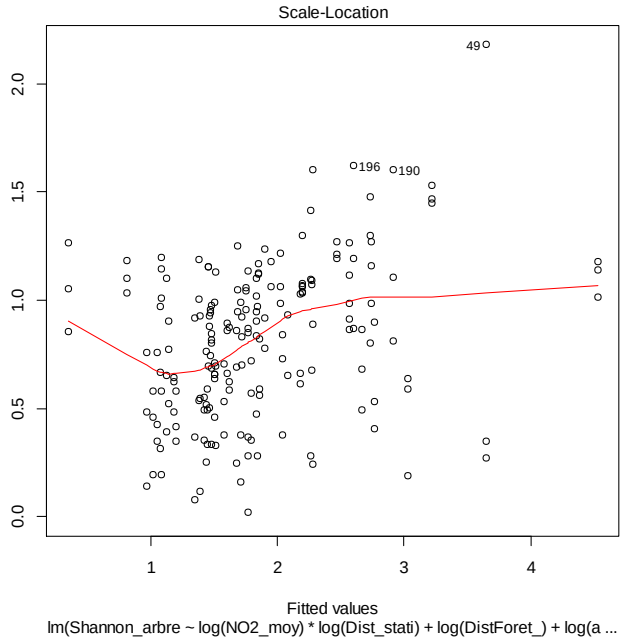
<u>Etude des résidus avec les plots</u>

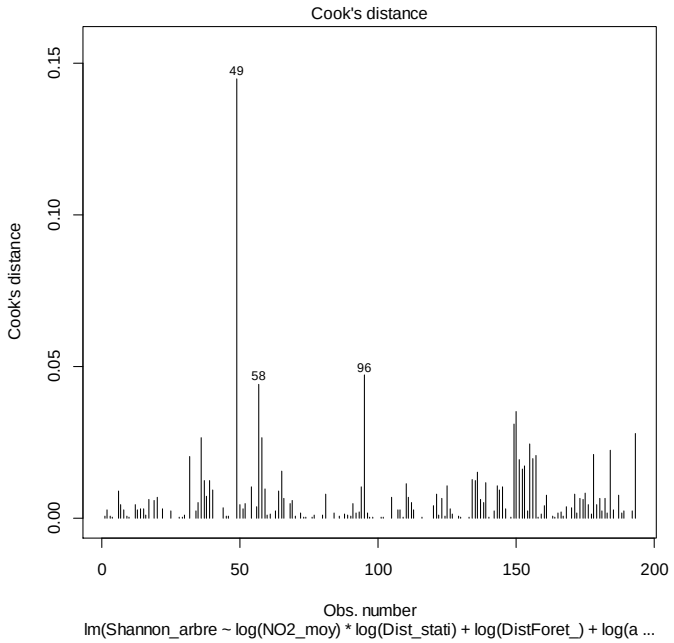



L'affichage met en évidence un point excentré (le point 49), où le résidu est important.

L'hypothèse de normalité des résidus semble respectée (voir le Q-Q plot gaussien) ainsi que l'homoscédasticité des résidus (relativement).






La distance de Cook mesure l'effet de la suppression d'une donnée. Les données avec d'importants résidus (données aberrantes) ayant un fort effet de levier peuvent fausser le résultat et la précision d'une régression. Les points ayant une distance de Cook importante sont considérés comme méritant un examen plus approfondi dans l'analyse.

Ainsi, tous les graphes mettent en évidence un point "aberrant" : le point 49, qui correspond à l'arbre 117_A3, où la diversité est particulièrement élevée (19 espèces relevées).
En testant le modèle sur la base de donnée privée de ce point, nous arrivons à un modèle où le maximum des résidus est réduit de 5.0974 à 2.9962**, mais le coefficient de détermination est moins bon : 0.2555 au lieu de 0.2792.**

## 5- Application d'un modèle GAM inspiré du modèle linéaire précédent

Pour essayer une autre approche, nous avons appliqué un lissage sur la variable log(NO2_moy) du modèle précédent avec un modèle GAM. La variable log(DistCentre) est alors peu significative et n'améliore pas le coefficient de détermination, nous l'avons donc enlevée du modèle :

Shannon_arbre ~ s(log(NO2_moy))+log(NO2_moy) * log(Dist_stati) + log(DistForet_) + log(altitude)

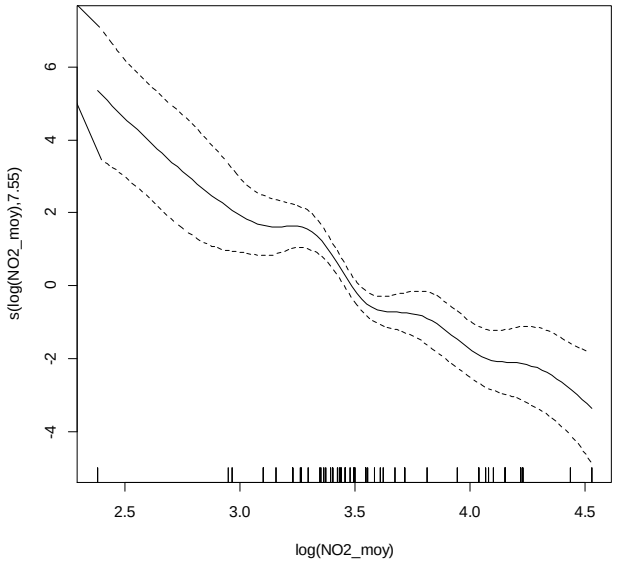
Les tests statistiques nous donnent :

| | Estimates | Standard Error | Anova : F value | Anova : Pr>F | Edf : Estimated Degrees of Liberty |
|---|---|---|---|---|---|
| Intercept (ordonnée à l'origine) | 0.4826 | 0.1231 | | | |
| log(NO2_moy) | 0.7783 | 0.1980 | **15.451** | **0.000121** | |
| log(Dist_stati) | -1.6165 | 0.4016 | **16.200** | **8.39e.05** | |
| log(DistForet_) | -0.2678 | 0.1049 | **6.515** | **0.011532** | |
| log(altitude) | 0.2138 | 0.1442 | 2.200 | 0.139766 | |

| | | | | | |
|---|---|---|---|---|---|
| log(NO2_moy):log(Dist_stati) | 0.4141 | 0.1110 | **13.921** | **0.000256** | |
| s(log(NO2_moy)) | | | **6.513** | **1.11e-07** | **7.541** |

| | |
|---|---|
| R adjusted | **0.263** |
| Max residuals (valeur absolue) | 4.977513 |

Le lissage du paramètre log(NO2_moy) calculé par la fonction gam est le suivant :



## B- Comment simplifier le protocole ?

### 1- N'observer qu'un arbre par site ?

Pour essayer de déterminer si le protocole pourrait être simplifié en n'observant qu'un arbre par site, nous avons simulé cette simplification en tirant aléatoirement un arbre sur chaque site, et en recréant une base de données à partir de ces tirages (Annexe 4).

En refaisant les mêmes tests que précédemment avec le dernier modèle linéaire retenu, nous moyennons les résultats des tests ANOVA sur ces tirages.

**Moyennes des coefficients de nos tests statistiques (10 000 tirages) :**

| | Estimates | Standard Error | Anova : F value | Anova : Pr>F |
|---|---|---|---|---|
| Intercept (ordonnée à l'origine) | 3.6840832 | 5.2273864 | | |
| log(NO2_moy) | -2.2126849 | 1.1472284 | **3.481317** | **0.130084242** |
| log(Dist_stati) | -1.6804498 | 0.7060114 | 3.359022 | 0.114515890 |
| log(altitude) | 0.5942650 | 0.2692516 | 1.797913 | 0.258077874 |
| log(DistCentre) | 0.5161894 | 0.1840072 | **10.887440** | **0.005021483** |
| log(NO2_moy) :log(Dist_stati) | 0.4113194 | 0.1956681 | 4.615350 | 0.061315191 |

| | |
|---|---|
| F-statistics | 4.828208 |
| P-value | 0.002914124 |
| R adjusted | 0.2344062 |
| Max residuals (valeur absolue) | 3.586775 (en indice de Shannon) |

**Variance des coefficients (10 000 tirages) :**

| | Estimates | Standard Error | Anova : F value | Anova : Pr>F |
|---|---|---|---|---|
| Intercept (ordonnée à l'origine) | 7.13585431 | 0.2773485634 | | |

| | | | | |
|---|---|---|---|---|
| log(NO2_moy) | 0.41919734 | 0.0134512406 | 4.132571 | 2.130119e-02 |
| log(Dist_stati) | 0.14465003 | 0.0051666635 | 2.903931 | 1.121349e-02 |
| log(altitude) | 0.01820256 | 0.0007341640 | 1.515724 | 3.243605e-02 |
| log(DistCentre) | 0.01168562 | 0.0003423654 | 11.373610 | 9.428898e-05 |
| log(NO2_moy) :log(Dist_stati) | 0.01056572 | 0.0004001800 | 3.852878 | 4.443938e-03 |

| | |
|---|---|
| F-statistics | 0.987645 |
| P-value | 3.291148e-05 |
| R adjusted | 0.05218037 |

## 2- S'inspirer du Protocole OPAL ?

Dans un second temps, nous avons essayé de simuler la base de données que nous aurions obtenue si le protocole OPAL [7] avait été utilisé pour les récolter.

Les particularités du protocole OPAL par rapport au protocole *Lichens Go !* sont les suivantes :

- Les participants doivent donner l'abondance de lichens sur le tronc, et uniquement leur présence ou non sur les branches (ce que nous ne pouvons pas faire à partir des données de PartiCitaE). Il n'y a pas de notion de site, les paramètres MEAN_lon, SUM_circo et tous ceux qui sous-entendent un regroupement par site sont donc à supprimer.
- Seuls 9 lichens doivent être analysés, classés en 3 catégories :
  - Sensibles au nitrogène : Usnea(fruticuleux), Evernia(fruticuleux), Hypogymnia (foliacé)
  - Intermédiaires : Melanelixia (foliacé), Flavoparmelia(foliacé), Parmelia (foliacé)
  - Non sensibles : Xanthoria parietina (foliacé), Xanthoria polycarpa (foliacé), Physcia(foliacé)

  Ces espèces étaient en effet parmi les plus répandues dans les relevés de PartiCitaE, et comme 21 sites sur 29 renseignent les espèces de lichen à Lyon, et 31 sur 38 à Paris, nous avons gardé ce critère.

- Les participants doivent se concentrer sur trois espèces d'arbres répandues : Fraxinus excelsior, Quercus et Acer pseudoplatanus. Ces trois espèces sont dans trois catégories de pH différentes au regard de leur écorce. Cependant, sur les données de PartiCitaE, seul un arbre est identifié à Lyon, et seulement 2/3 des arbres le sont à Paris. Nous abandonnons donc ce paramètre (essence de l'arbre) pour la simulation du protocole OPAL.
- Les données de mesure de circonférence, d'emplacement, de pollution correspondent à la façon dont la plupart des paramètres concernant l'environnement du site d'observation sont relevés, nous les gardons donc tels quels dans la base de données.

- Pour l'abondance des lichens sur le tronc entre 50cm et 1.5m du sol, les participants choisissent la face avec le plus de lichens (face 1 systématiquement pour le protocole PartiCitaE), puis une valeur entre 0 et 3 basée sur la comparaison avec une feuille A4 :
  - 0 = aucun (i.e. 0 carrés pour le protocole PartiCitaE)
  - 1 = moins d'un quart de la feuille (i.e. 1 carré pour le protocole PartiCitaE)
  - 2 = entre un quart et la feuille entière (i.e. entre 2 et 4 carrés pour le protocole PartiCitaE)
  - 3 = plus qu'une feuille (i.e. 5 carrés pour le protocole PartiCitaE)

Le détail des modifications de la base de données pour simuler le protocole OPAL est en Annexe 5. Notons que les indices de Diversité et de Shannon calculés sont légèrement différents, car la notation de l'abondance sur une échelle de 0 à 3 telle que demandée dans le protocole OPAL n'est pas réellement proportionnelle à la fréquence des lichens au $m^2$ sur le tronc.

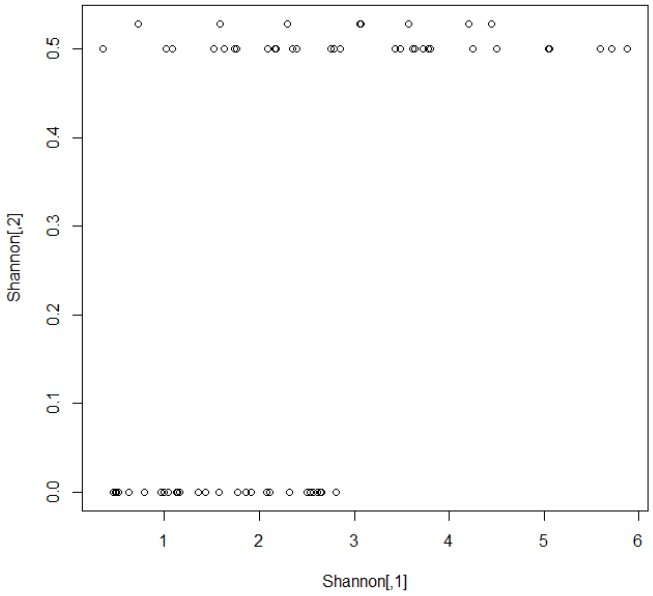
Les paramètres « altitude » et « DistForet » n'ont pas de corrélations très significatives au taux de $NO_2$ quand on l'applique à la base de données modifiée selon le protocole OPAL, et ne font pas augmenter beaucoup le coefficient de détermination. Nous les avons donc retirés du modèle pour tester la régression linéaire suivante :

Shannon_arbre ~ s(log(NO2_moy))+log(NO2_moy) * log(Dist_stati) + log(DistCentre)

| | Estimates | Standard Error | Anova : F value | Anova : Pr>F |
|---|---|---|---|---|
| Intercept (ordonnée à l'origine) | 1.42173 | 0.74641 | | |
| log(NO2_moy) | **-0.53630** | 0.15789 | **15.1154** | 0.0002522 |
| log(Dist_stati) | -0.36977 | 0.11170 | **11.1646** | 0.0014273 |
| Log(DistCentre) | 0.10572 | 0.03372 | **15.4385** | 0.0002200 |
| log(NO2_moy):log(Dist_stati) | 0.09250 | 0.02991 | **9.59619** | 0.0029943 |

| | |
|---|---|
| F-statistic | 12.82 |
| P-value | **1.237e-07** |
| R adjusted | **0.4211** |
| Max residuals (valeur absolue) | 0.47477 (en indice de Shannon) |

Nous avons aussi comparé l'indice de Shannon obtenu avec le protocole OPAL (Shannon[,2]) , et celui obtenu avec le protocole de PartiCitaE (Shannon[,1]) :



# IV. Interprétation des résultats

## A- Liens entre un indice de diversité des lichens et le taux de $NO_2$

### 1- Recherche des paramètres influents par ACP et AIC

D'un point de vue biologique, un modèle où l'indice de Shannon d'un arbre et la fréquence en foliacés sur son tronc dépend de la ville (dont « altitude » est un proxy), du taux d'humidité (approximé par DistWater_), de la densité de population et du taux de $NO_2$, serait assez intuitif. L'analyse grâce à l'ACP sélectionne bien ces variables.

Cependant, la régression linéaire sur ces paramètres ne fournit pas de résultats concluants, sauf sur la densité de population. Notamment, le taux de $NO_2$ est corrélé positivement à l'indice de Shannon (alors que nous attendions naturellement une corrélation négative), avec un intervalle de confiance à 95% très important. Cela ne nous étonne pas pour autant car il s'agit d'un premier modèle très peu sophistiqué (pas de termes d'interaction entre les variables explicatives, et nous verrons par la suite qu'appliquer des échelles logarithmiques peut être pertinent).

Le critère de sélection AIC révèle quant à lui un paramètre important qui n'est pas intuitif : la distance entre le point de relevé et la station de mesure du $NO_2$.

Dans ces deux modèles, la corrélation entre l'indice de Shannon et le taux de $NO_2$ reste par contre de mauvaise qualité, nous avons donc cherché d'autre modèle nous permettant de le mettre en avant, en s'appuyant sur les paramètres mis en avant par ces deux méthodes de sélection.

### 2- Le modèle de Linda Seggi appliqué à la diversité fonctionnelle

Cet indice améliore notre modèle car il prend en compte l'interaction des différentes espèces de lichen avec leur environnement. En effet, nos analyses confirment le constat que la diversité fonctionnelle explique mieux le fonctionnement des écosystèmes que les mesures classiques de diversité.

Cependant, nous ne pouvons utiliser cet indice que sur les données de Paris, ce qui limite son intérêt dans le cadre de l'étude de la pertinence du protocole PartiCitaE. Il serait peut-être possible toutefois d'utiliser un indice de diversité fonctionnelle simplifié. Par exemple, une première approche serait de moyenner les valeurs des cinq traits (poléotolérance, pH, photophilie, aridité, eutrophisation) pour chaque type de thalle et de calculer la diversité fonctionnelle à partir des thalles relevés et de ces caractéristiques moyennes.

### 3- Modèle simple : taux de $NO_2$ comme seul paramètre

Nous observons bien une corrélation négative entre l'indice de Shannon et le taux de $NO_2$. Cependant, la part de variance expliquée par le modèle est très faible (R=3,6%).

La F-valeur élevée, égale à 8.141 met bien en avant une corrélation entre indice de Shannon et taux de $NO_2$ moyen. Mais le taux de $NO_2$ à lui seul ne permet pas d'expliquer toute la variabilité de la diversité lichénique dans l'échantillon étudié, ce qui était prévisible car la diversité des lichens dépend d'une multitude de facteurs, comme vu dans la partie I.

Pour expliquer une plus grande part de la variance de l'indice de Shannon, nous avons essayé d'intégrer d'autres paramètres dans les modèles, en s'inspirant des parties précédentes.

### 4- Approfondissement de ce modèle : ajout de l'altitude, de la distance au centre-ville, à la forêt et à la station de mesure du $NO_2$ la plus proche

Le dernier modèle linéaire retenu apporte un coefficient de détermination assez important : **R=27.92%**. En effet, en ajoutant d'autres paramètres du modèle, on explique mieux les observations faites sur la diversité des lichens.

Ici, nous observons bien une corrélation négative et significative entre la diversité lichénique et le taux de $NO_2$, ainsi que le logarithme des distances à la station de mesure de $NO_2$ et de la distance à la forêt. Le lien physique entre la distance à la station de mesure est peu clair. Tout de même, nous supposons que ces stations étant situées dans des lieux relativement peu pollués, la diversité des lichens est plus importante à mesure qu'on s'approche de ces lieux.

On observe aussi, conformément à notre intuition physique, que la diversité est corrélée positivement à l'altitude (la pollution est sans aucun doute moins importante à Lyon qu'à Paris) et à la distance au centre (la pollution est plus présente en centre-ville qu'en banlieue).

L'intégration du terme d'interaction entre NO2_moy et Dist_stati est largement inspirée du modèle réalisé par Linda Seggi. Elle est justifiée au vu de la corrélation de Spearman entre ces deux variables.

Ainsi, la corrélation entre le taux de pollution et l'indice de diversité de Shannon, fondée sur la diversité lichénique, est mise en évidence. Le protocole de PartiCitaE qui invite à une participation des citoyens dans les relevés de lichens semblerait fonctionnel, dans la mesure où les corrélations ne sont pas faussées après l'ajout des données de relevés de Lyon (réalisés par des non-experts) aux relevés de Paris (réalisés par des experts en biologie).

### 5- Application d'un modèle GAM inspiré du modèle linéaire précédent

Même si l'analyse de la variance semble plus concluante que pour les modèles linéaires simples (la F-value de la variable log(NO2) passe de 10,8891 à 15,451 pour le terme linéaire et 6,513 pour le terme lissé), et si la façon dont la variable NO2_moy est lissée n'est pas aberrante (pas de changement de la monotonie de la fonction), le nombre

de degrés de libertés rajoutés (7,541) est très élevé. De plus, la version lissée du logarithme du taux de $NO_2$ moyen reste quasiment proportionnelle au logarithme de NO2_moy, sauf aux alentours de 3,5, qui est la moyenne et la médiane des taux de $NO_2$ de l'ensemble des relevés. Ces deux éléments laissent supposer qu'il y a eu un sur-ajustement des données par le lissage du taux de NO2 autour de log(NO2)=3,5.

Ainsi, les approximations données par un modèle GAM laissent supposer que la dépendance de Shannon_arbre en fonction de NO2_moy n'est pas tout à fait logarithmique, et qu'en l'ajustant on peut limiter l'influence de la distance au centre-ville, mais aussi de l'altitude (et donc de la ville choisie), qui a une très faible F-value dans ce modèle. Cependant, nous n'avons pas réussi à atteindre un modèle plus performant par rapport aux modèles linéaires, grâce à la méthode GAM.

## B- Comment simplifier le protocole ?

### 1- N'observer qu'un arbre par site ?

Nous observons une baisse globale de la qualité des corrélations, avec des valeurs de F value et Pr(>F) assez peu significatives pour l'ANOVA, comparé au modèle 5 sur l'ensemble des données. Ce n'est pas surprenant, car nous n'avons que 58 relevés à chaque simulation (58 sites ont été observés au total seulement).

De plus, les énormes variances obtenues sur l'ensemble des tirages confirment que les arbres d'un même site sont loin de fournir des relevés homogènes, et qu'il y a donc une variance au sein d'un site non négligeable qui justifie de faire les relevés sur des sites de 2 ou 3 arbres.

La seule corrélation maintenue est celle entre l'indice de Shannon et le logarithme de la distance au centre-ville, ce qui vient du fait que c'était déjà de loin la corrélation la plus forte quand la base de données utilisée était la table complète.

### 2- S'inspirer du protocole OPAL ?

Les coefficients obtenus avec ce modèle sont comparables à ceux obtenus avec la table de données complète et le modèle 5, malgré le fait que nous n'ayons plus les données que de 87 arbres, et d'une seule face par arbre.

Ainsi, le protocole OPAL semblerait être un bon moyen d'obtenir des données qui apportent une information suffisante sur la qualité de l'air, et des résultats plus significatifs, même à partir de données récoltées par des bénévoles, qui comportent beaucoup d'erreurs. Il semble logique en effet que le taux d'erreur soit plus faible lorsque les participants se concentrent sur quelques espèces bien reconnaissables, sans prendre en compte des espèces rares qu'ils ne pourraient identifier avec certitude.

Cependant, la perte de corrélation pour les paramètres « altitude » et « DistForet » montre que l'indice de Shannon issu de ce protocole reflète moins bien l'ensemble des

facteurs réels (humidité par exemple) qui ont une influence sur l'abondance des différentes espèces de lichens. Cela provient à la fois de la simplicité du protocole et du choix des espèces à repérer. On le remarque notamment quand on affiche l'indice de Shannon issu du protocole OPAL, en fonction de celui issu du protocole PartiCitaE : on aperçoit un effet de seuil, l'indice de Shannon issu d'OPAL prenant principalement les valeurs 0.0 ou 0.5, ce qui réduit énormément l'information qu'il transporte.

Enfin, la base de données OPAL comporte 22 espèces relevées entre 2007 et 2015, et il pourrait être intéressant d'adapter la liste des espèces de lichens à observer en fonction des espèces le plus souvent rencontrées à Paris et Lyon, qui sont peut-être différentes des 9 espèces utilisées ici.

# V. Conclusion

Notre objectif est d'abord de vérifier la cohérence des données de lichens issues du protocole *Lichens Go !* avec leur capacité à prédire la qualité de l'air, qui se manifeste par le taux de $NO_2$. Ainsi, d'après nos modèles linéaires et non linéaire, une corrélation entre l'indice de Shannon et le taux moyen de $NO_2$ atmosphérique au niveau d'un arbre est mise en évidence, en faisant intervenir de nombreux autres paramètres. Si la plupart de ces paramètres ont une influence qui peut s'expliquer d'un point de vue biologique, l'importance de la distance à la station de mesure du $NO_2$ reste difficile à expliquer.

Le protocole *Lichens Go !* semblerait opérant et il serait donc possible d'utiliser ces modèles en prédicteurs de la pollution atmosphérique à partir de relevés de lichens issus d'enquêtes participatives.

Ce protocole a l'avantage d'être assez simple pour être réalisable par des personnes non expertes du domaine. Un certain compromis est à trouver entre la simplicité du protocole pour permettre sa réalisabilité, et la perte d'information qui peut en résulter. Par exemple, nous remarquons que la diversité fonctionnelle qui semble produire des résultats très significatifs (jusqu'à 44% de la variance expliquée par le modèle), ne peut pas être calculée telle quelle avec le protocole *Lichens Go !*, par manque d'informations. De même, on pourrait penser (sans avoir une quelconque certitude) que les modèles gagneraient en précision si les paramètres qui servent de proxy au taux d'humidité étaient remplacés par la mesure de ce taux lui-même.

Et si le protocole pouvait être davantage simplifié, sans altérer la corrélation existante entre le taux de $NO_2$ et la diversité lichénique ? La participation citoyenne pourrait alors s'avérer particulièrement intéressante et efficace. En effet, plus le protocole est simple, plus il est facile d'impliquer du monde au service de la science.

La comparaison avec le protocole OPAL génère des résultats intéressants sur nos données : les corrélations sont toujours significatives. Un protocole qui s'inspire des principes du protocole OPAL demanderait moins de connaissances techniques de la part des bénévoles. Le risque d'erreurs liées à la méconnaissance des bénévoles devient aussi plus réduit, car il y a moins d'espèces à observer. Là encore, il faut trouver un compromis, car un protocole trop simple donne moins d'informations sur l'environnement des lichens. (On rappelle que l'importance est tout de même de garder des informations sur la pollution atmosphérique).

Enfin, ne faire les relevés que sur un arbre par site ne semblerait pas être une meilleure idée, car la diversité des arbres au sein d'un site peut s'avérer importante, et étudier plusieurs arbres proches n'apporte pas de redondance. Il faut tout de même relativiser cette affirmation, qui n'est fondée que sur l'analyse d'un jeu de données de 38 arbres.

# BIBLIOGRAPHIE


[1] V. Shukla et al. in *Lichens to biomonitor the environment*, pp. 1-60 (Springer, 2014)

[2] B. A. Markert et al. in *Trace Metals and other Contaminants in the Environment* 6, pp. 396-398 (Elsevier, 2003)

[3] F. Chlous et al., Introduction. Foisonnement participatif : des questionnements communs ?, *Natures Sciences Sociétés* 25, 4, 327-335 (2017)

[4] L. Seed et al., Modelling relationships between lichen bioindicators, air quality and climate on a national scale: Results from the UK OPAL air survey, *Environmental Pollution xxx*, 1-11 (Elsevier, 2013)

[5] L. Gosling et al., Citizen science identifies the effects of nitrogen dioxide and other environmental drivers on tar spot of sycamore, *Environmental Pollution* 214, 549-555 (Elsevier, 2016)

[6] D. J. Tregidgo et al., Can citizen science produce good science ? Testing the OPAL Air Survey methodology, using lichens as indicators of nitrogenous pollution, *Environmental Pollution* 182, 448-451 (Elsevier, 2013)

[7] OPAL Air Survey, at <https://www.opalexplorenature.org/AirSurvey>

[8] PartiCitaE, Participez! Lichens Go!, at <http://www.particitae.upmc.fr/fr/participez/suivez-les-lichens.html>

[9] E. Gallic, Régression linéaire avec R : Sélection de modèle, at <http://egallic.fr/l3-eco-gestion-regression-lineaire-avec-r-selection-de-modele/>

[10] C. Prieur, Professeur, Régression linéaire multiple en R, Exemple détaillé, at <http://ljk.imag.fr/membres/Clementine.Prieur/M1SSD/02exemple.pdf>

[11] R Documentation, Principal Components Analysis, at <https://stat.ethz.ch/R-manual/R-patched/library/stats/html/prcomp.html>

[12] Wikipédia, Analyse en Composantes Principales, at <https://fr.wikipedia.org/wiki/Analyse_en_composantes_principales>

[13] M. Genin, Régression linéaire multiple, at <http://cerim.univ-lille2.fr/fileadmin/user_upload/statistiques/michael_genin/Cours/Modelisation/Regression-lineaire-multiple_DU.pdf>

[14] C. Blisle, Table de la loi de Fisher, at <https://archimede.mat.ulaval.ca/stt1920/STT-1920-Loi-de-Fisher.pdf>

[15] R-atique, Non-respect des hypothèses du modèle linéaire (ANOVA, régression) : c'est grave docteur ??, at <http://perso.ens-lyon.fr/lise.vaudor/non-respect-des-hypotheses-du-modele-lineaire-anova-regression-cest-grave-docteur/>

[16] STATISTICA, Concepts Fondamentaux en Statistique, Data Mining : Modèles Additifs Généralisés, at <http://www.statsoft.fr/concepts-statistiques/modeles-additifs-generalises/modeles-additifs-generalises.php#.XLRxfOgzaM8>

[17]R Documentation, Summary for a GAM fit, at<https://stat.ethz.ch/R-manual/R-devel/library/mgcv/html/summary.gam.html>

[18] Botta-Dukát et al., Rao's quadratic entropy as a measure of functional diversity based on multiple traits, Journal of Vegetation Science 16, 533-540 (Opulus Press Uppsala, 2005)

# ANNEXE 1

### Calculs de fréquences, diversités, Shannon

```
#transfert de données et regroupement
obs_tout <- read.csv("observation_tout.csv",sep=";")
par_site <- obs_tout %>% group_by(pk_reponse)
par_arbre <- obs_tout %>% group_by(pk_reponse,pk_arbre)
par_arbre_face <- obs_tout %>% group_by(pk_reponse,pk_arbre,num_face)


#calcul de la Diversité (nombre d'espèces de lichens différents)
Diversite_site <- par_site %>% summarise(n_distinct(sp))
write.csv(Diversite_site,"Diversite_site.csv")
Diversite_arbre <- par_arbre %>% summarise(n_distinct(sp))
write.csv(Diversite_arbre,"Diversite_arbre.csv")


#calcul de "fréquences"
freq_face <- par_arbre_face %>%
 summarise((sum(Q1)+sum(Q2)+sum(Q3)+sum(Q4)+sum(Q5))/5)

colnames(freq_face)[colnames(freq_face)=="(sum(Q1) + sum(Q2) + sum(Q3) + sum(Q4) +
 sum(Q5))/5"] <- "diversite"
write.csv(freq_face,"diversite_face.csv")

freq_arbre <- freq_face %>% group_by(pk_reponse,pk_arbre) %>%
 summarise(mean(diversite))

colnames(freq_arbre)[colnames(freq_arbre)=="mean(diversite)"] <- "diversite"
write.csv(freq_arbre,"diversite_arbre.csv")

freq_site <- freq_arbre %>% group_by(pk_reponse) %>% summarise(mean(diversite))

colnames(freq_site)[colnames(freq_site)=="mean(diversite)"] <- "diversite"
write.csv(freq_site,"div_site.csv")


#calcul de Shannon
espece_par_arbre <- z %>% group_by(pk_reponse,pk_arbre,sp)
freq_esp_par_arbre <- espece_par_arbre %>%
 summarise((sum(Q1)+sum(Q2)+sum(Q3)+sum(Q4)+sum(Q5))/5/min(cb_face))

colnames(freq_esp_par_arbre)[4] <- "freq"
shannon_par_arbre <- freq_esp_par_arbre %>% summarise(-sum(freq*log(freq,2)))

write.csv(shannon_par_arbre,"shannon_arbre.csv")

espece_par_site <- z %>% group_by(pk_reponse,sp)
```

```
freq_esp_par_site <- espece_par_site %>%
 summarise((sum(Q1)+sum(Q2)+sum(Q3)+sum(Q4)+sum(Q5))/15/min(cb_face))

colnames(freq_esp_par_site)[3] <- "freq"
shannon_par_site <- freq_esp_par_site %>% summarise(-sum(freq*log(freq,2)))

write.csv(shannon_par_site,"shannon_site.csv")

#calcul de fréquences de thalle

par_arbre_thalle <- obs_tout %>% group_by(pk_reponse,pk_arbre,thalle)
par_arbre_nb_thalle <- par_arbre_thalle %>%
 summarise(sum(Q1)+sum(Q2)+sum(Q3)+sum(Q4)+sum(Q5))
colnames(par_arbre_nb_thalle)[4]<-"nb"

par_arbre_thalle <- full_join(par_arbre_nb_thalle,par_arbre_nb_thalle %>%
 summarise(sum(nb)),by="pk_arbre")

c <- par_arbre_thalle$nb/par_arbre_thalle[6]
write.csv(par_arbre_thalle,"par_arbre_thalle.csv")
write.csv(c,"freq.csv")

par_site_thalle <- obs_tout %>% group_by(pk_reponse,thalle)
par_site_nb_thalle <- par_site_thalle %>%
 summarise(sum(Q1)+sum(Q2)+sum(Q3)+sum(Q4)+sum(Q5))
colnames(par_site_nb_thalle)[3]<-"nb"

par_site_thalle <- full_join(par_site_nb_thalle,par_site_nb_thalle %>%
 summarise(sum(nb)),by="pk_reponse")

c <- par_site_thalle$nb/par_site_thalle[4]
write.csv(par_site_thalle,"par_site_thalle.csv")
write.csv(c,"freq_site.csv")
```

# ANNEXE 2

**Sélection des variables quantitatives qui ont un sens physique :**

table <- arbre_tout_moi[c("MEAN_lon", "MEAN_lat","MEAN_circo","SUM_circon", "DistCentre", "Dist_stati", "DistForet_", "DistWater_","altitude", "AltiMIN_in", "Diff_alti", "Pente_degr", "DiffAlti_x", "Exposition", "CLC2012", "AREA_ha_CL", "perc_dense_50m", "perc_dense_100m", "perc_dense_500m", "Div_arbre", "Shannon_arbre", "freq_crus_arbre", "freq_fol_arbre", "Div_site", "Shannon_site", "freq_crus_site", "freq_fol_site", "freq_fruti_site", "NO2_moy")]

**ACP sur les paramètres par arbres :**

**Importance of components:**

| | PC1 | PC2 | PC3 | PC4 | PC5 | PC6 | PC7 |
|---|---|---|---|---|---|---|---|
| Standard deviation | 2.4096 | 2.1430 | 1.8272 | 1.65383 | 1.51899 | 1.40887 | 1.21593 |
| Proportion of Variance | 0.2002 | 0.1584 | 0.1151 | 0.09432 | 0.07956 | 0.06844 | 0.05098 |
| Cumulative Proportion | 0.2002 | 0.3586 | 0.4737 | 0.56802 | 0.64758 | 0.71603 | 0.76701 |

| | PC8 | PC9 | PC10 | PC11 | PC12 | PC13 | PC14 |
|---|---|---|---|---|---|---|---|
| Standard deviation | 1.16642 | 1.03270 | 0.95339 | 0.80791 | 0.77493 | 0.65254 | 0.62194 |
| Proportion of Variance | 0.04692 | 0.03678 | 0.03134 | 0.02251 | 0.02071 | 0.01468 | 0.01334 |
| Cumulative Proportion | 0.81392 | 0.85070 | 0.88204 | 0.90455 | 0.92526 | 0.93994 | 0.95328 |

| | PC15 | PC16 | PC17 | PC18 | PC19 | PC20 | PC21 |
|---|---|---|---|---|---|---|---|
| Standard deviation | 0.56533 | 0.52867 | 0.43986 | 0.41262 | 0.36993 | 0.27921 | 0.2411 |
| Proportion of Variance | 0.01102 | 0.00964 | 0.00667 | 0.00587 | 0.00472 | 0.00269 | 0.0020 |
| Cumulative Proportion | 0.96430 | 0.97394 | 0.98061 | 0.98648 | 0.99120 | 0.99389 | 0.9959 |

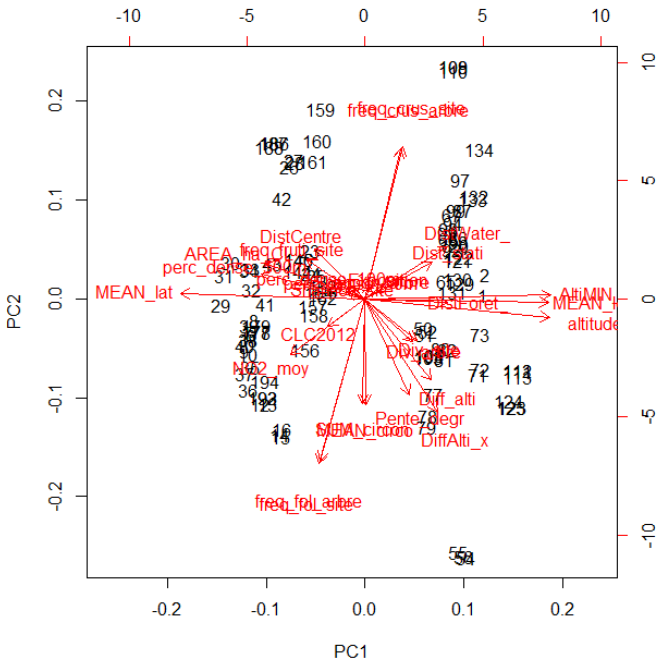
| | PC22 | PC23 | PC24 | PC25 | PC26 | PC27 | PC28 |
|---|---|---|---|---|---|---|---|
| Standard deviation | 0.22809 | 0.19278 | 0.13693 | 0.08096 | 0.05571 | 0.03543 | 0.01903 |
| Proportion of Variance | 0.00179 | 0.00128 | 0.00065 | 0.00023 | 0.00011 | 0.00004 | 0.00001 |
| Cumulative Proportion | 0.99768 | 0.99896 | 0.99961 | 0.99984 | 0.99994 | 0.99999 | 1.00000 |

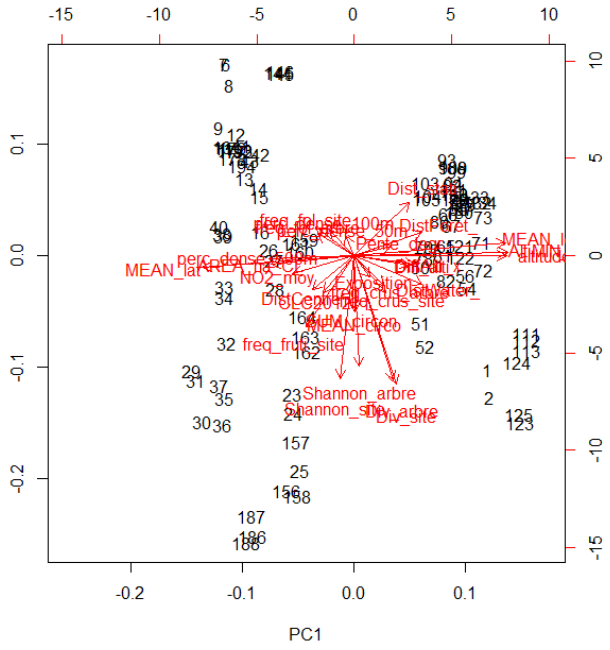
| | PC29 |
|---|---|
| Standard deviation | 4.805e-16 |
| Proportion of Variance | 0.000e+00 |
| Cumulative Proportion | 1.000e+00 |

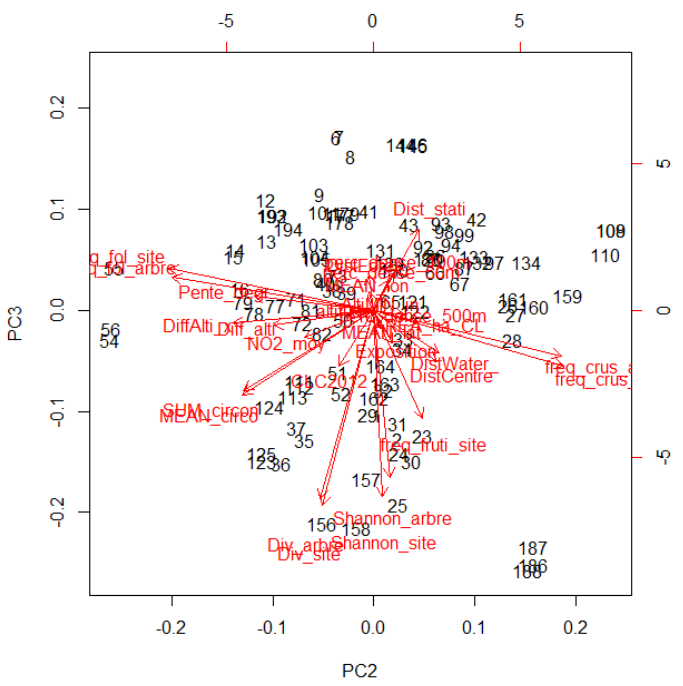
**Standard deviation^2 = Variance**

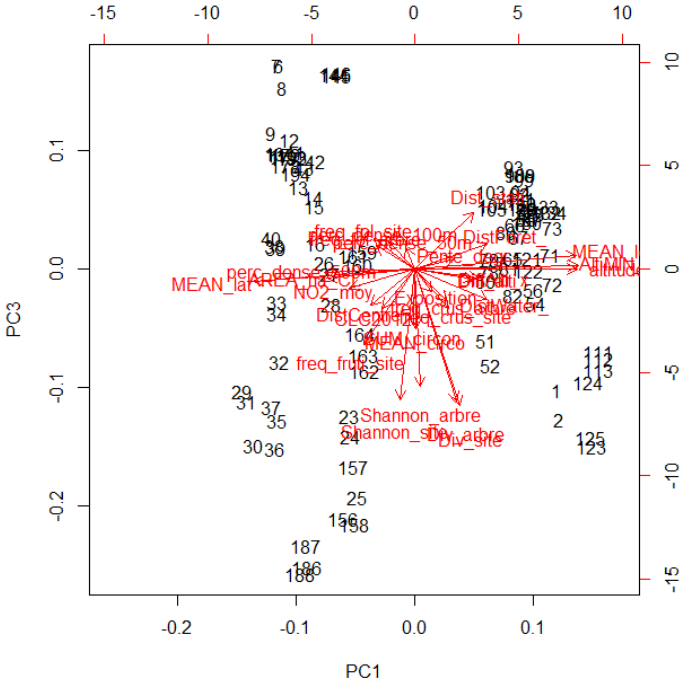
[1] 5.806128e+00 4.592460e+00 3.338765e+00 2.735159e+00 2.307332e+00
[6] 1.984902e+00 1.478483e+00 1.360541e+00 1.066477e+00 9.089579e-01
[11] 6.527161e-01 6.005120e-01 4.258056e-01 3.868074e-01 3.195938e-01
[16] 2.794887e-01 1.934752e-01 1.702541e-01 1.368484e-01 7.795884e-02
[21] 5.812266e-02 5.202404e-02 3.716382e-02 1.874895e-02 6.554217e-03
[26] 3.103690e-03 1.255090e-03 3.622465e-04 2.308940e-3

**Projection des paramètres dans l'espace des 3 premières PC :**






**ACP sur les paramètres par sites :**

**Importance of components:**

| | PC1 | PC2 | PC3 | PC4 | PC5 | PC6 | PC7 |
|---|---|---|---|---|---|---|---|
| Standard deviation | 2.3903 | 1.8681 | 1.6585 | 1.51596 | 1.38612 | 1.32804 | 1.15576 |
| Proportion of Variance | 0.2285 | 0.1396 | 0.1100 | 0.09193 | 0.07685 | 0.07055 | 0.05343 |
| Cumulative Proportion | 0.2285 | 0.3681 | 0.4782 | 0.57009 | 0.64694 | 0.71749 | 0.77092 |

| | PC8 | PC9 | PC10 | PC11 | PC12 | PC13 | PC14 |
|---|---|---|---|---|---|---|---|
| Standard deviation | 1.11779 | 0.98406 | 0.93518 | 0.7954 | 0.67382 | 0.64127 | 0.53690 |
| Proportion of Variance | 0.04998 | 0.03874 | 0.03498 | 0.0253 | 0.01816 | 0.01645 | 0.01153 |
| Cumulative Proportion | 0.82090 | 0.85963 | 0.89462 | 0.9199 | 0.93808 | 0.95453 | 0.96606 |

| | PC15 | PC16 | PC17 | PC18 | PC19 | PC20 | PC21 |
|---|---|---|---|---|---|---|---|
| Standard deviation | 0.5316 | 0.50236 | 0.37929 | 0.24734 | 0.21061 | 0.19533 | 0.13346 |
| Proportion of Variance | 0.0113 | 0.01009 | 0.00575 | 0.00245 | 0.00177 | 0.00153 | 0.00071 |
| Cumulative Proportion | 0.9774 | 0.98746 | 0.99321 | 0.99566 | 0.99744 | 0.99896 | 0.99967 |

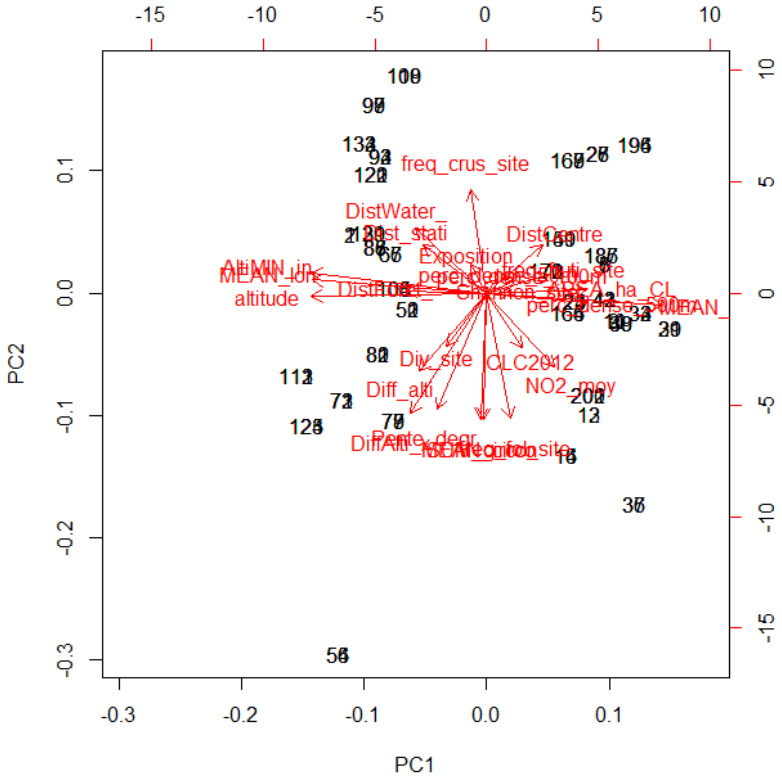
| | PC22 | PC23 | PC24 | PC25 |
|---|---|---|---|---|
| Standard deviation | 0.08051 | 0.03601 | 0.01912 | 2.988e-16 |
| Proportion of Variance | 0.00026 | 0.00005 | 0.00001 | 0.000e+00 |
| Cumulative Proportion | 0.99993 | 0.99999 | 1.00000 | 1.000e+00 |

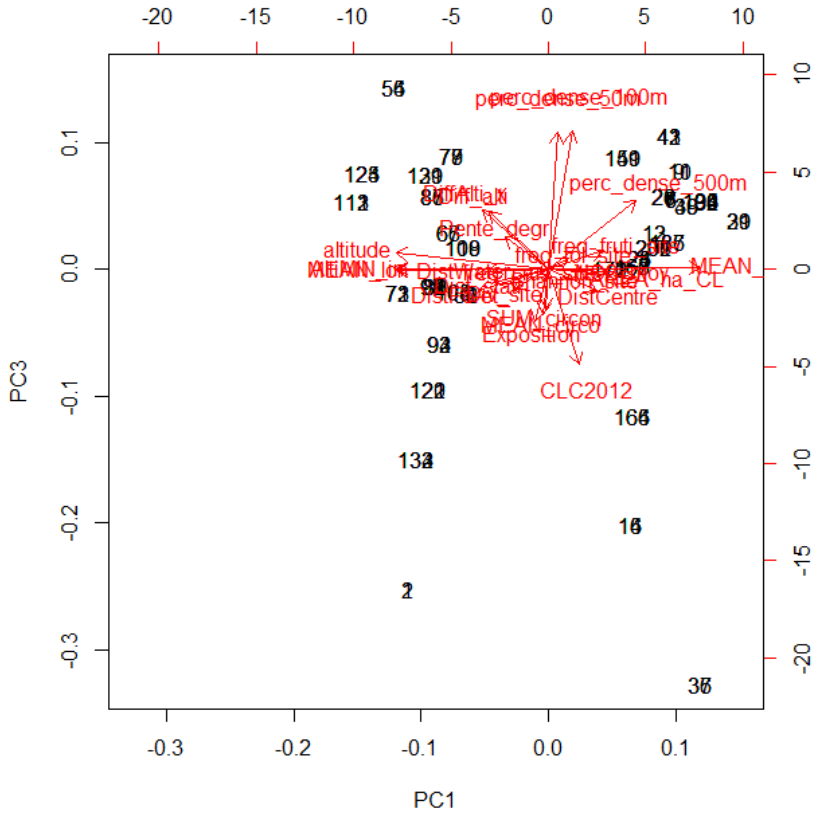
**Standard deviation^2 = Variance**

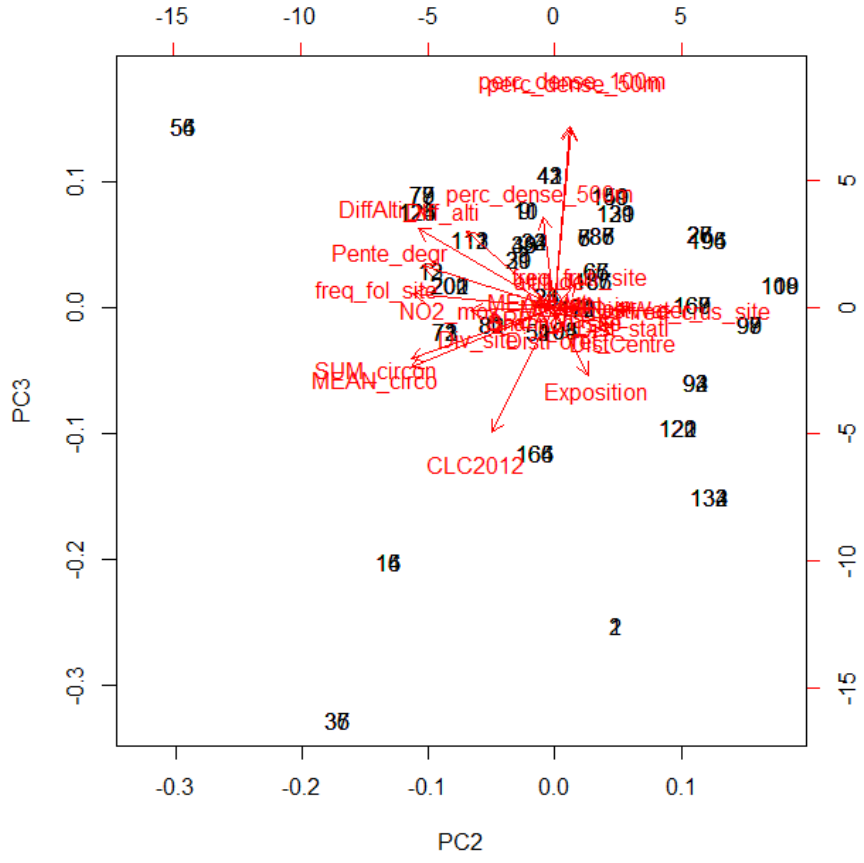
[1] 5.713522e+00 3.489975e+00 2.750601e+00 2.298146e+00 1.921329e+00
[6] 1.763691e+00 1.335778e+00 1.249444e+00 9.683803e-01 8.745622e-01
[11] 6.325980e-01 4.540389e-01 4.112320e-01 2.882639e-01 2.825752e-01
[16] 2.523672e-01 1.438586e-01 6.117579e-02 4.435525e-02 3.815245e-02

[21] 1.781142e-02 6.481187e-03 1.296400e-03 3.657313e-04 8.927955e-32

**Projection des paramètres dans l'espace des 3 premières PC :**




PC3
PC1
perc_dense_500m
Pente_degr
altitude
MEAN_
SUM_circon
Exposition
CLC2012


PC3
PC2
perc_dense_500m
Pente_degr
freq_fol_site
SUM_circon
MEAN_circo
Exposition
CLC2012

# ANNEXE 3

## L'Analyse de Linda Seggi :

```
model21 <- lm( Shannon ~ NO2_moyenne*log(Dist_stati) + percentage.500m+
log(DistxArea_divsumArea), data = data.arb)
anova(model21)

Analysis of Variance Table

Response: Shannon
                          Df Sum Sq Mean Sq F value    Pr(>F)
NO2_moyenne                1  3.895  3.8951  6.3644   0.01320 *
log(Dist_stati)            1  3.481  3.4812  5.6882   0.01894 *
percentage.500m            1 16.163 16.1629 26.4097 1.352e-06 ***
log(DistxArea_divsumArea)  1  1.227  1.2266  2.0043   0.15993
NO2_moyenne:log(Dist_stati)  1  1.084  1.0838  1.7709   0.18627
Residuals                101 61.813  0.6120
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


summary(model21)

Call:
lm(formula = Shannon ~ NO2_moyenne * log(Dist_stati) + percentage.500m +
    log(DistxArea_divsumArea), data = data.arb)

Residuals:
    Min      1Q  Median      3Q     Max
-1.7451 -0.4602  0.1330  0.5579  1.5126

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               4.056029   0.592666   6.844 6.08e-10 ***
NO2_moyenne              -0.017457   0.012382  -1.410    0.162
log(Dist_stati)          -0.358734   0.136733  -2.624    0.010 *
percentage.500m          -0.011142   0.002121  -5.252 8.38e-07 ***
log(DistxArea_divsumArea)  -0.050596   0.033038  -1.531    0.129
NO2_moyenne:log(Dist_stati)  0.003427   0.002575   1.331    0.186
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7823 on 101 degrees of freedom
  (7 observations deleted due to missingness)
Multiple R-squared:  0.2949, Adjusted R-squared:   0.26 (part de la variance expliquée par
le modèle)
F-statistic: 8.447 on 5 and 101 DF,  p-value: 1.045e-06
```

# ANNEXE 4

### Tirage aléatoire d'un arbre par site :

```
TirageAnova<- function (ParamPA, tirage){
	if(tirage){
		j <- sample(1:3, 1)
	}
	else {
	j<-1
	}
	h<-0
	ParamN <- ParamPA[j,]
	while(h<length(ParamPA[[1]])) {
	if(tirage){
		n <-ParamPA[["nbarbre"]][h+1]
	}
	else{
		n<-1
	}
	if(n>0){
	if(tirage){
		j <- sample(1:n,1)
	}
	seqN<-c()
	seqR<-c()
	for (k in 1:length(ParamPA)){
		if(tirage){
			seqN<- c(seqN, as.numeric(ParamPA[[j+h,k]]))
		}
		else{seqN<- c(seqN, as.numeric(ParamPA[[h+1,k]]))}
	}
	}
	h<- h+max(1,n)
	ParamN <- rbind(ParamN, seqN)

}

model1N <- lm(Shannon_arbre~log(NO2_moy)*log(Dist_stati) +
log(altitude)+log(DistCentre), data = ParamN)
model1N}
```

### Moyenne et variance des résultats sur plusieurs tirages :

```
# Appelle TirageAnova() N fois, et calcule la moyenne et la variance
# des coefficients du test Anova effectué sur les modèles renvoyés par TirageAnova()

Tiragesnew<-function(ParamPA, N){
# mat contiendra tous les résultats des tests Anova à la suite
	model<-TirageAnova(ParamPA, T)
	ano<-anova(model)
```

```
        sum<-summary(model)
        estim <-sum[["coefficients"]][,1:2]
        ftrucs <- ano[c("F value", "Pr(>F)")]
        fstats <- sum[["fstatistic"]][1]
        pvalue<- pf(sum$fstatistic[1], sum$fstatistic[2], sum$fstatistic[3], lower.tail=FALSE)
        adjr <- sum[["adj.r.squared"]]
        maxres<-max(abs(sum[["residuals"]]))
        matestim <-estim
        matftrucs <- ftrucs
        matfstats <- fstats
        matpvalue<- pvalue
        matadjr <- adjr
        matmaxres<-maxres
        moyestim <-estim
        moyftrucs <- ftrucs
        moyfstats <- fstats
        moypvalue<- pvalue
        moyadjr <- adjr
        print(adjr)
        moymaxres<-maxres
        varestim <-matrix(nrow=nrow(estim), ncol=2, data=0)
        varftrucs <-matrix(nrow=nrow(ftrucs), ncol=2, data=0)
        varfstats <- 0
        varpvalue<- 0
        varadjr <- 0
        varmaxres<-0
        for (i in 2:N){
                model<-TirageAnova(ParamPA, T)
                ano<-anova(model)
                sum<-summary(model)
                estim <-sum[["coefficients"]][,1:2]
                ftrucs <- ano[c("F value", "Pr(>F)")]
                fstats <- sum[["fstatistic"]][1]
                pvalue<- pf(sum$fstatistic[1], sum$fstatistic[2], sum$fstatistic[3],
lower.tail=FALSE)
                adjr <- sum[["adj.r.squared"]]
                maxres<-max(abs(sum[["residuals"]]))
                matestim <-cbind(matestim, estim)
                matftrucs <- cbind(matftrucs, ftrucs)
                matfstats <- cbind(matfstats, fstats)
                matpvalue<- cbind(matpvalue, pvalue)
                matadjr <- cbind(matadjr, adjr)
                matmaxres<cbind(matmaxres, maxres)
                moyestim <-moyestim +estim
                moyftrucs <- moyftrucs +ftrucs
                moyfstats <- moyfstats +fstats
                moypvalue<- moypvalue +pvalue
                moyadjrr<- moyadjr +adjr
                moymaxres<-moymaxres + maxres
        }
        moyestim <-moyestim/N
        moyftrucs <- moyftrucs/N
        moyfstats <- moyfstats/N
        moypvalue<- moypvalue/N
```

```
        moyadjr <- moyadjr/N
        moymaxres<-moymaxres/N
        for (i in 1:N){
                varestim <-varestim +(matestim[,((i-1)*2+1):((i-1)*2+2)]-moyestim)^2
                varftrucs <- varftrucs +(matftrucs[,((i-1)*2+1):((i-1)*2+2)]-moyftrucs)^2
                varfstats <- varfstats +(matfstats[i]-moyfstats)^2
                varpvalue<- varpvalue +(matpvalue[i]-moypvalue)^2
                varadjr <- varadjr +(matadjr[i]-moyadjr)^2
                varmaxres<-varmaxres +(matmaxres[i]-moymaxres)^2
                #On récupère les résultats du test Anova du i° Tirage, pour le rajouter
                # au calcul de la Variance
        }
varestim <-varestim/N
varftrucs <- varftrucs/N
varfstats <- varfstats/N
varpvalue<- varpvalue/N
varadjr <- varadjr/N
varmaxres<-varmaxres/N
print(moyestim)
print(moyftrucs)
print(moyfstats)
print(moypvalue)
print(moyadjr)
print(moymaxres)
print(varestim)
print(varftrucs)
print(varfstats)
print(varpvalue)
print(varadjr)
print(varmaxres)
result<- moyestim
model}
```

# Annexe 5

```
Opal <- function (ParamPRarbre){
        tablefiltree<-filter(ParamPRarbre,
(ParamPRarbre[["num_face"]]==1)&(ParamPRarbre[["circonference"]]>=40))
        #On ne garde que la face 1, i.e. celle avec le plus de lichens. On ne
connaitra pas son orientation
        qtot                                                                    <-
tablefiltree[["Q1"]]+tablefiltree[["Q2"]]+tablefiltree[["Q3"]]+tablefiltree[["Q4"]]+tabl
efiltree[["Q5"]]
        tablefiltree<-cbind(tablefiltree, qtot)
        abondanceusnea<-c()
        abondanceevernia<-c()
        abondancehypogymnia<-c()
        abondancemelanelixia<-c()
        abondanceflavoparmelia <- c()
        abondanceparmelia <- c()
        abondancexanthoriapar<-c()
        abondancexanthoriapoly<-c()
```

```
abondancephyscia<-c()
abondancesensible <- c()
abondanceintermediaire <-c()
abondancenonsensible <- c()
abondancefruti <- c()
abondancefolia <- c()
n <- length(tablefiltree[["pk_reponse.x"]])
i<-1
while (i<=n)
{
	abondance<-NA
	if(!is.na(tablefiltree[["nom"]][i])){
	if(tablefiltree[["nom"]][i]=="USNEA"){
		if(tablefiltree[["qtot"]]==0){
			abondance <- 0}
		if (tablefiltree[["qtot"]]==1){
			abondance <-1}
		if ((tablefiltree[["qtot"]]>=2)&(tablefiltree[["qtot"]]<=4)){
			abondance <-2}
		if (tablefiltree[["qtot"]]==5){
			abondance <-3}
		abondanceusnea<-rbind(abondanceusnea, abondance)
		abondanceevernia<-rbind(abondanceevernia, 0)
		abondancehypogymnia<-rbind(abondancehypogymnia, 0)
		abondancemelanelixia<-rbind(abondancemelanelixia, 0)
		abondanceflavoparmelia <- rbind(abondanceflavoparmelia, 0)
		abondanceparmelia <- rbind(abondanceparmelia, 0)
		abondancexanthoriapar<-rbind(abondancexanthoriapar, 0)
		abondancexanthoriapoly<-rbind(abondancexanthoriapoly, 0)
		abondancephyscia<-rbind(abondancephyscia, -1)

	}
	if(tablefiltree[["nom"]][i]=="EVERNIA PRUNASTRI"){
		if(tablefiltree[["qtot"]]==0){
			abondance <- 0}
		if (tablefiltree[["qtot"]]==1){
			abondance <-1}
		if ((tablefiltree[["qtot"]]>=2)&(tablefiltree[["qtot"]]<=4)){
			abondance <-2}
		if (tablefiltree[["qtot"]]==5){
			abondance <-3}
		abondanceusnea<-rbind(abondanceusnea, 0)
		abondanceevernia<-rbind(abondanceevernia, abondance)
		abondancehypogymnia<-rbind(abondancehypogymnia, 0)
		abondancemelanelixia<-rbind(abondancemelanelixia, 0)
		abondanceflavoparmelia <- rbind(abondanceflavoparmelia, 0)
		abondanceparmelia <- rbind(abondanceparmelia, 0)
		abondancexanthoriapar<-rbind(abondancexanthoriapar, 0)
		abondancexanthoriapoly<-rbind(abondancexanthoriapoly, 0)
		abondancephyscia<-rbind(abondancephyscia, 0)
```

```
		}
		if(tablefiltree[["nom"]][i]=="HYPOGYMNIA PHYSODES"){
			if(tablefiltree[["qtot"]]==0){
				abondance <- 0}
			if (tablefiltree[["qtot"]]==1){
				abondance <-1}
			if ((tablefiltree[["qtot"]]>=2)&(tablefiltree[["qtot"]]<=4)){
				abondance <-2}
			if (tablefiltree[["qtot"]]==5){
				abondance <-3}
			abondanceusnea<-rbind(abondanceusnea, 0)
			abondanceevernia<-rbind(abondanceevernia, 0)
			abondancehypogymnia<-rbind(abondancehypogymnia,
abondance)
			abondancemelanelixia<-rbind(abondancemelanelixia,0)
			abondanceflavoparmelia <- rbind(abondanceflavoparmelia, 0)
			abondanceparmelia <- rbind(abondanceparmelia, 0)
			abondancexanthoriapar<-rbind(abondancexanthoriapar, 0)
			abondancexanthoriapoly<-rbind(abondancexanthoriapoly, 0)
			abondancephyscia<-rbind(abondancephyscia, 0)

		}
		if(tablefiltree[["nom"]][i]=="MELANELIXIA SUBARGENTIFERA"){
			if(tablefiltree[["qtot"]]==0){
				abondance <- 0}
			if (tablefiltree[["qtot"]]==1){
				abondance <-1}
			if ((tablefiltree[["qtot"]]>=2)&(tablefiltree[["qtot"]]<=4)){
				abondance <-2}
			if (tablefiltree[["qtot"]]==5){
				abondance <-3}
			abondanceusnea<-rbind(abondanceusnea, 0)
			abondanceevernia<-rbind(abondanceevernia, 0)
			abondancehypogymnia<-rbind(abondancehypogymnia, 0)
			abondancemelanelixia<-rbind(abondancemelanelixia,
abondance)
			abondanceflavoparmelia <- rbind(abondanceflavoparmelia, 0)
			abondanceparmelia <- rbind(abondanceparmelia, 0)
			abondancexanthoriapar<-rbind(abondancexanthoriapar, 0)
			abondancexanthoriapoly<-rbind(abondancexanthoriapoly, 0)
			abondancephyscia<-rbind(abondancephyscia, 0)
		}

	if((tablefiltree[["nom"]][i]=="FLAVOPARMELIA")|(tablefiltree[["nom"]][i]=="FL
AVOPARMELIA CAPERATA")){
			if(tablefiltree[["qtot"]]==0){
				abondance <- 0}
			if (tablefiltree[["qtot"]]==1){
				abondance <-1}
```

```
                    if ((tablefiltree[["qtot"]]>=2)&(tablefiltree[["qtot"]]<=4)){
                        abondance <-2}
                    if (tablefiltree[["qtot"]]==5){
                        abondance <-3}
                    abondanceusnea<-rbind(abondanceusnea, 0)
                    abondanceevernia<-rbind(abondanceevernia, 0)
                    abondancehypogymnia<-rbind(abondancehypogymnia, 0)
                    abondancemelanelixia<-rbind(abondancemelanelixia, 0)
                    abondanceflavoparmelia <- rbind(abondanceflavoparmelia,
abondance)
                    abondanceparmelia <- rbind(abondanceparmelia, 0)
                    abondancexanthoriapar<-rbind(abondancexanthoriapar, 0)
                    abondancexanthoriapoly<-rbind(abondancexanthoriapoly, 0)
                    abondancephyscia<-rbind(abondancephyscia, 0)
                }

        if((tablefiltree[["nom"]][i]=="PARMELIA")|(tablefiltree[["nom"]][i]=="PARMELIA
SULCATA")){
                    if(tablefiltree[["qtot"]]==0){
                        abondance <- 0}
                    if (tablefiltree[["qtot"]]==1){
                        abondance <-1}
                    if ((tablefiltree[["qtot"]]>=2)&(tablefiltree[["qtot"]]<=4)){
                        abondance <-2}
                    if (tablefiltree[["qtot"]]==5){
                        abondance <-3}
                    abondanceusnea<-rbind(abondanceusnea, 0)
                    abondanceevernia<-rbind(abondanceevernia, 0)
                    abondancehypogymnia<-rbind(abondancehypogymnia, 0)
                    abondancemelanelixia<-rbind(abondancemelanelixia, 0)
                    abondanceflavoparmelia <- rbind(abondanceflavoparmelia, 0)
                    abondanceparmelia <- rbind(abondanceparmelia, abondance)
                    abondancexanthoriapar<-rbind(abondancexanthoriapar, 0)
                    abondancexanthoriapoly<-rbind(abondancexanthoriapoly, 0)
                    abondancephyscia<-rbind(abondancephyscia, 0)
                }

        if((tablefiltree[["nom"]][i]=="XANTHORIA")|(tablefiltree[["nom"]][i]=="XANTH
ORIA PARIETINA")){
                    if(tablefiltree[["qtot"]]==0){
                        abondance <- 0}
                    if (tablefiltree[["qtot"]]==1){
                        abondance <-1}
                    if ((tablefiltree[["qtot"]]>=2)&(tablefiltree[["qtot"]]<=4)){
                        abondance <-2}
                    if (tablefiltree[["qtot"]]==5){
                        abondance <-3}
                    abondanceusnea<-rbind(abondanceusnea, 0)
                    abondanceevernia<-rbind(abondanceevernia, 0)
                    abondancehypogymnia<-rbind(abondancehypogymnia, 0)
```

```
                abondancemelanelixia<-rbind(abondancemelanelixia, 0)
                abondanceflavoparmelia <- rbind(abondanceflavoparmelia, 0)
                abondanceparmelia <- rbind(abondanceparmelia, 0)
                abondancexanthoriapar<-rbind(abondancexanthoriapar,
abondance)
                abondancexanthoriapoly<-rbind(abondancexanthoriapoly, 0)
                abondancephyscia<-rbind(abondancephyscia, 0)
        }
        if(tablefiltree[["nom"]][i]=="XANTHORIA POLYCARPA"){
                if(tablefiltree[["qtot"]]==0){
                        abondance <- 0}
                if (tablefiltree[["qtot"]]==1){
                        abondance <-1}
                if ((tablefiltree[["qtot"]]>=2)&(tablefiltree[["qtot"]]<=4)){
                        abondance <-2}
                if (tablefiltree[["qtot"]]==5){
                        abondance <-3}
                abondanceusnea<-rbind(abondanceusnea, 0)
                abondanceevernia<-rbind(abondanceevernia, 0)
                abondancehypogymnia<-rbind(abondancehypogymnia, 0)
                abondancemelanelixia<-rbind(abondancemelanelixia, 0)
                abondanceflavoparmelia <- rbind(abondanceflavoparmelia, 0)
                abondanceparmelia <- rbind(abondanceparmelia, 0)
                abondancexanthoriapar<-rbind(abondancexanthoriapar, 0)
                abondancexanthoriapoly<-rbind(abondancexanthoriapoly,
abondance)
                abondancephyscia<-rbind(abondancephyscia, 0)
        }

      if((tablefiltree[["nom"]][i]=="PHYSCIA")|(tablefiltree[["nom"]][i]=="PHYSCIA
ADSCENDENS")|(tablefiltree[["nom"]][i]=="PHYSCIA TENELLA")){
                if(tablefiltree[["qtot"]]==0){
                        abondance <- 0}
                if (tablefiltree[["qtot"]]==1){
                        abondance <-1}
                if ((tablefiltree[["qtot"]]>=2)&(tablefiltree[["qtot"]]<=4)){
                        abondance <-2}
                if (tablefiltree[["qtot"]]==5){
                        abondance <-3}
                abondanceusnea<-rbind(abondanceusnea, 0)
                abondanceevernia<-rbind(abondanceevernia, 0)
                abondancehypogymnia<-rbind(abondancehypogymnia, 0)
                abondancemelanelixia<-rbind(abondancemelanelixia, 0)
                abondanceflavoparmelia <- rbind(abondanceflavoparmelia,0)
                abondanceparmelia <- rbind(abondanceparmelia, 0)
                abondancexanthoriapar<-rbind(abondancexanthoriapar, 0)
                abondancexanthoriapoly<-rbind(abondancexanthoriapoly, 0)
                abondancephyscia<-rbind(abondancephyscia, abondance)
        }
        }
```

```
        if (is.na(abondance)){
                tablefiltree<-tablefiltree[-i,]
                n<-n-1
        }
        else {
                i<-i+1
        }
}
tablefiltree<-cbind(tablefiltree, abondanceusnea, abondanceevernia,abondancehypogymnia,abondancemelanelixia,abondanceflavoparmelia,abondanceparmelia,abondancexanthoriapar, abondancexanthoriapoly, abondancephyscia)
tablefiltree<- summarise(group_by(tablefiltree, pk_arbre), pk_reponse=pk_reponse.x[1],Nom_Statio=Nom_Statio[1],date=date[1],CentreVill=CentreVill[1],Code_Stati = Code_Stati[1],DistCentre=DistCentre[1], Dist_stati=Dist_stati[1], DistForet_=DistForet_[1], DistWater_=DistWater_[1],altitude=altitude[1], Pente_degr=Pente_degr[1], Exposition=Exposition[1], CLC2012=CLC2012[1], AREA_ha_CL=AREA_ha_CL[1], perc_dense_50m=perc_dense_50m[1], perc_dense_100m=perc_dense_100m[1], perc_dense_500m=perc_dense_500m[1], Div_arbre=n_distinct(sp), usnea = max(abondanceusnea), evernia = max(abondanceevernia),hypogymnia= max(abondancehypogymnia),melanelixia = max(abondancemelanelixia),flavoparmelia = max(abondanceflavoparmelia),parmelia = max(abondanceparmelia),xanthoriapar = max(abondancexanthoriapar), xanthoriapoly=max(abondancexanthoriapoly), physcia = max(abondancephyscia), NO2_moy=NO2_moy[1], circonference=circonference[1], type_coord=type_coord[1], lon=lon[1], lat=lat[1], adresse=adresse[1], essence=essence[1], genre=genre[1])
Abondance_arbre<-tablefiltree[["usnea"]] +tablefiltree[["evernia"]] + tablefiltree[["hypogymnia"]] +tablefiltree[["melanelixia"]]+ tablefiltree[["flavoparmelia"]]+ tablefiltree[["parmelia"]]+ tablefiltree[["xanthoriapar"]]+ tablefiltree[["xanthoriapoly"]]+ tablefiltree[["physcia"]]
Absensible_arbre<-tablefiltree[["usnea"]] +tablefiltree[["evernia"]] + tablefiltree[["hypogymnia"]]
Abintermediaire_arbre<- tablefiltree[["melanelixia"]]+ tablefiltree[["flavoparmelia"]]+ tablefiltree[["parmelia"]]
Abnonsensible_arbre<-tablefiltree[["xanthoriapar"]]+ tablefiltree[["xanthoriapoly"]]+ tablefiltree[["physcia"]]
Abfruti_arbre<-tablefiltree[["usnea"]] +tablefiltree[["evernia"]]
Abfolia_arbre<-tablefiltree[["hypogymnia"]] +tablefiltree[["melanelixia"]]+ tablefiltree[["flavoparmelia"]]+ tablefiltree[["parmelia"]]+ tablefiltree[["xanthoriapar"]]+ tablefiltree[["xanthoriapoly"]]+ tablefiltree[["physcia"]]
Perc_sensible_arbre <-Absensible_arbre/Abondance_arbre
Perc_intermediaire_arbre <-Abintermediaire_arbre/Abondance_arbre
Perc_nonsensible_arbre <-Abnonsensible_arbre/Abondance_arbre
Perc_fruti_arbre <-Abfruti_arbre/Abondance_arbre
Perc_folia_arbre <-Abfolia_arbre/Abondance_arbre
Perc_usnea <- tablefiltree[["usnea"]]/Abondance_arbre
```

```
	Perc_evernia <- tablefiltree[["evernia"]]/Abondance_arbre
	Perc_hypogymnia <- tablefiltree[["hypogymnia"]]/Abondance_arbre
	Perc_melanelixia <- tablefiltree[["melanelixia"]]/Abondance_arbre
	Perc_flavoparmelia <- tablefiltree[["flavoparmelia"]]/Abondance_arbre
	Perc_parmelia <- tablefiltree[["parmelia"]]/Abondance_arbre
	Perc_xanthoriapar <- tablefiltree[["xanthoriapar"]]/Abondance_arbre
	Perc_xanthoriapoly <- tablefiltree[["xanthoriapoly"]]/Abondance_arbre
	Perc_physcia <- tablefiltree[["physcia"]]/Abondance_arbre
	Shannon_arbre<-Perc_physcia*0
	if(Perc_usnea>0){
		Shannon_arbre<-Shannon_arbre - (Perc_usnea*log(Perc_usnea, 2))
	}
	if(Perc_evernia>0){
		Shannon_arbre<-Shannon_arbre - (Perc_evernia*log(Perc_evernia, 2))
	}
	if(Perc_hypogymnia>0){
		Shannon_arbre<-Shannon_arbre -
(Perc_hypogymnia*log(Perc_hypogymnia, 2))
	}
	if(Perc_melanelixia>0){
		Shannon_arbre<-Shannon_arbre -
(Perc_melanelixia*log(Perc_melanelixia, 2))
	}
	if(Perc_flavoparmelia>0){
		Shannon_arbre<-Shannon_arbre -
(Perc_flavoparmelia*log(Perc_flavoparmelia, 2))
	}
	if(Perc_parmelia>0){
		Shannon_arbre<-Shannon_arbre - (Perc_parmelia*log(Perc_parmelia,
2))
	}
	if(Perc_xanthoriapar>0){
		Shannon_arbre<-Shannon_arbre -
(Perc_xanthoriapar*log(Perc_xanthoriapar, 2))
	}
	if(Perc_xanthoriapoly>0){
		Shannon_arbre<-Shannon_arbre -
(Perc_xanthoriapoly*log(Perc_xanthoriapoly, 2))
	}
	if(Perc_physcia>0){
		Shannon_arbre<-Shannon_arbre - (Perc_physcia*log(Perc_physcia, 2))
	}
	length(Shannon_arbre)
	tablefiltree<-cbind(tablefiltree, Abondance_arbre, Absensible_arbre,
Abnonsensible_arbre, Abfruti_arbre, Abfolia_arbre, Perc_sensible_arbre,
Perc_intermediaire_arbre, Perc_nonsensible_arbre, Perc_fruti_arbre,
Perc_folia_arbre, Shannon_arbre)
	tablefiltree
}
```