Détermination du pool d'espèces en vue de la prédiction des pH et C/N

Mohamed Akodad

Master Mathématiques et applications Année 2012-2013

Université de Strasbourg **UFR Mathématique-Informatique**

Année Universitaire 2012-2013

Spécialité Statistique

Détermination du pool d'espèces en vue de la prédiction des pH et C/N

Présenté par : Mohamed AKODAD

Date de la soutenance : 29 Août 2013

Maîtres de stage : M. Jean-Claude PIERRAT et M. Pierre MONTPIED
Du 1er Mars 2013 au 31 Août 2013

Laboratoire d'accueil : Laboratoire d'Etudes des Ressources Forêts-Bois (LERFoB)

Adresse : Centre AgroParisTech site de Nancy
14 rue Girardet 54042 Nancy Cedex

Directeur : Mme Meriem FOURNIER
Responsable du Master : Mme Armelle GUILLOU

…, et celui qui les poursuit est condamné
à ne jamais connaître le repos.
Henri POINCARE
La valeur de la science

Remerciements

Je tiens tout d'abord à remercier M. Jean-Claude PIERRAT et M. Pierre MONTPIED, enseignants chercheurs au sein du Laboratoire d'Etudes des Ressources Forêts-Bois (LERFoB), ainsi que M. Nicolas POULIN, ingénieur de recherche, pour m'avoir encadré durant ce stage, pour le temps qu'ils m'ont consacré tout au long de cette période et pour avoir répondu à mes questions.

Je remercie, M. Jean-Claude GEGOUT, responsable de l'équipe, pour m'avoir accueilli, pour ses conseils et recommandations, ainsi que Mme Meriem FOURNIER, directrice du LERFoB, Mme Paulina PINTO, Mme Ingrid SEYNAVE et M. Christian PIEDALLU. De plus, ce projet a bénéficié du soutien du Laboratoire d'Excellence ARBRE ((ANR-12-LABXARBRE-01)).

Je tiens à remercier également, M. Robin GENUER, maître de conférences à l'université de Bordeaux, pour le temps qu'il m'a consacré et pour avoir répondu à mes questions sur les forêts aléatoires tout au long de ce stage.

Un grand merci également à toutes les personnes rencontrées au sein du LERFoB durant ce stage pour leur accueil et leur bonne humeur, en particulier à Vincent, Corinne, Sophie, mais aussi aux doctorants, Emilien, Raphael, Nicolas, Marie aux stagiaires et contractuels Abel, Nuria, Marc, Lisa, Ping Lian, Ruben, Ian, Pierre sans oublier le jardiner Bernard l'ami de tous.

Résumé

La pédologie est la science qui s'intéresse à la formation et à l'évolution des sols, d'un site donné. Néanmoins, une autre méthode indirecte permet l'étude des sols, c'est la bioindication. Un bioindicateur est un indice construit à partir d'une ou plusieurs espèces végétales, dont la présence ou l'absence, renseigne sur certaines caractéristiques écologiques de l'environnement.

Les modèles statistiques couramment utilisés sont des modèles linéaires généralisés (ter Braak & Looman 1986), ce sont des modèles paramétriques. Néanmoins, les modèles non paramétrique permettent plus de liberté à l'utilisateur et sont moins contraignant à la différence des modèles paramétriques. Ces dernières années, avec l'essor de l'apprentissage statistique, le développement de nouveau modèle est apparu. L'un des plus répandu est le modèle des forêts aléatoires introduite par Breiman (2001). Cette méthode consiste à construire de manière aléatoire un très grand nombre d'arbres de régression (Breiman 1984), avant d'en faire la moyenne.

C'est une méthode très intéressante et performante, car elle permet de faire la sélection de variables. En effet, lorsqu'on est en présence de centaines voire de milliers de variables, il peut être intéressant de réduire ce nombre pour garder uniquement les variables qui expliquent au mieux la variable réponse.

Le jeu de données sur lequel j'ai pu travailler, provient de la base de données EcoPlant. Dans ce jeu de données, il y a des relevés phytosociologiques des espèces végétales (1 : présence et 0 : absence de l'espèce), il y a la mesure du pH et du C/N (nutrition azoté) mais aussi plein d'autres relevés comme la température moyenne, le bilan hydrique, etc… Le but de mon stage a été de sélectionner un ensemble d'espèces végétales pour prédire au mieux le pH et le C/N par la méthode des forêts aléatoires. En effet, Breiman (2001), a introduit un indice qui permet de choisir les variables les plus importantes. Cet indice, m'a permis de faire, une sélection générale, puis de faire une sélection plus développé à l'aide de la méthode de Genuer et al. (2010).

Des résultats intéressants (en termes d'erreur de prédiction et de variance expliquée) ont été obtenus, notamment en retenant moins d'une centaine d'espèces en vue de prédire le pH et le C/N. Cependant, cela reste encore trop élevé pour les spécialistes forestiers. C'est pour cela qu'un découpage, judicieux de la France, en cinq zones a permis de sélectionner moins de variables. De plus, en ajoutant du bruit aux variables (pH et C/N), on a montré qu'il n'y avait pas de phénomène de surparamétrisation.

Mots clés : bioindication, forêts aléatoires, arbres de régression, sélection de variables, apprentissage statistique.

Table des matières

1) Introduction

1.1) Présentation du laboratoire

J'ai réalisé mon stage du 1er Mars 2013 au 31 Août 2013, au Laboratoire d'Etudes des Ressources Forêts-Bois (LERFoB), situé à Nancy, dirigé par Mme Meriem Fournier, au sein de l'équipe écologie forestière (EF) dirigée par M. Jean-Claude GEGOUT.

Les équipes
- Ecologie Forestière
- Croissance, Production, Qualité des Bois
- Sylviculture et reboisement
- Mission Gestion végétation Forestière
- Installation Expérimentale Croissance
- Cellule technique GIS Coopérative
- Plateau Xylosciences

Organisation du LERFoB
Avril 2013

Fonctions transversales	
Directrice	Meriem Fournier (IC PEF)
Directeur Adjoint	François Ningre (IR2)
Gestion	Hélène Hurpeau (TREX), Nathalie Morel (TRN), Sophie Barthélémy (TRN), Virginie Friley (TR contractuelle)
Communication	Corinne Martin (SA)
Informatique	Christian Herbé (TRN)

Equipes	Ecologie forestière	Croissance, Production, Qualité des Bois	Sylviculture et Reboisement	Mission Gestion Végétation Forestière
Type	Recherche	Recherche	Enseignement et Expertise	
Acronyme	EF	CPQB	SR	MGVF
Chercheurs et Enseignants- Chercheurs	Jean-Claude Gégout (PR) Bruno Ferry (MC) Bernard Jabiol (MC) François Lebourgeois (MC) Damien Marage (MC)	Ignacio Barbeito (CR2) Francis Colin (CR1) Catherine Collet (CR1) Thiéry Constant (CR1) Jana Dlouha (CR2) Fleur Longuetaud (CR2) Frédéric Mothe (CR1) Cyrille Rathgeber (CR1) Holger Wernsdörfer (MC) Noël Le Goff (mission) Gérard Nepveu (mission) Jean-Marc Ottorini (mission)	Holger Wernsdörfer (MC)	Catherine Collet (CR1)
Ingénieurs	Jean-Daniel Bontemps (IPEF) Christian Piedallu (IR2) Jean-Claude Pierrat (IR1)	Meriem Fournier (IC PEF) Mathieu Fortin (IR) Responsable adjoint : François Ningre **(IR2)**	Eric Lacombe (IDAE) Philippe Durand (IAE) Yves Ehrhart (IDAE) Mathieu Fortin (IR)	
Assistants Ingénieurs et Techniciens	Sylvie Lehmann (AI) Vincent Perez (TRN) Fabien Spicher (TRN)	Loïc Dailly (TR) Bruno Garnier (TRN) Michel Pitsch (TREX)	Nicole Ory (SA)	Léon Wehrlen (AI)
Doctorants	Romain Bertrand Gabriella Riofrio Dillon Raphaël Trouvé Emilien Kuhn	Nicolas Bilot Vivien Bonnesoeur Julie Bossu Henry Cuny Félix Hartmann Zineb Kebbi Benkeder		
Contractuels (longue durée)	Mathilde Duverger (TR) Ian David Ondo (IR) David Thibaut (TR)	Damien Bourreau (Post-Doc) Marin Chaumet (IR Apprenti) Jean-Baptiste Morisset (Post-Doc)	Matthieu Fellmann (IE)	Xavier Auzuret **(IR)** Erwin Thirion (AI)

Équipes	Installation Expérimentale Croissance	Cellule Technique Coopérative Modélisation	Plateau Xylosciences
Acronyme	IEC	CT-Coop	
Chercheurs et Enseignants- Chercheurs			
Ingénieurs		Ingrid Seynave (IR2)	Julien Ruelle (IR) Philippe Jacquin (IE)
Assistants Ingénieurs et Techniciens	Daniel Rittié (AI) Frédéric Bordat (ATP2) Guy Maréchal (TRN) Florian Vast (TRN)	Responsable adjoint : Sébastien Daviller **(TRN)**	Responsable adjoint : Etienne Farré **(AI)** Emmanuel Cornu (ATP2) Charline Freyburger (TRN) Pierre Gelhaye (TRX) Maryline Harroué (TRN) Alain Mercanti (TRN)
Contractuels (longue durée)			

NB : en couleur = responsable de l'équipe

Figure 1 : Organisation du LERFoB

L'équipe écologie forestière est intégrée au département Sciences et Ingénierie Agronomiques, Forestières, de l'Eau et de l'Environnement (SIAFEE) de l'école d'ingénieur AgroParisTech et assurant des missions d'enseignement dans le cadre des cursus : d'ingénieurs AgroParisTech, des ingénieurs des ponts, des eaux et des forêts (IPEF) et de formation universitaires (Master Nancy, Metz, Dijon). De plus, l'équipe assure des missions de développement à l'attention des gestionnaires forestiers et des milieux naturels, des enseignants, par la réalisation d'ouvrages et d'articles spécifiques, en organisant et participant à diverses formations continues ou séminaires.

1.2) Présentation du sujet et des données

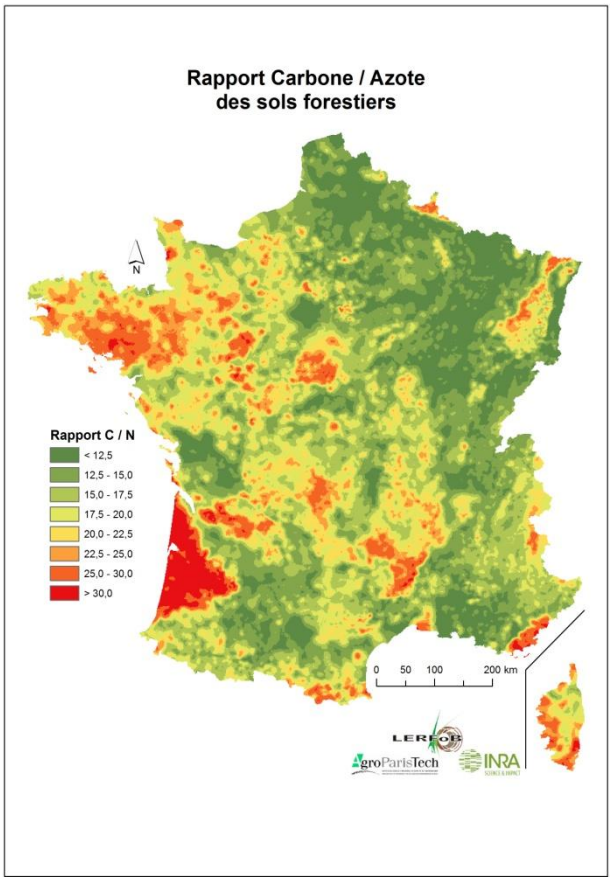
L'utilisation de la végétation de sous-bois par les forestiers pour estimer les caractéristiques stationnelles et les potentialités forestières remonte à près d'un siècle (Cajander, 1926). Les caractéristiques stationnelles représentent les propriétés d'un site étudié (géologie, topographie, climat, végétation) pouvant impacter la productivité du peuplement forestier, c'est-à-dire la population d'arbres présents sur le site. Les potentialités forestières représentent la capacité d'un sol, d'un site donné, pour la productivité d'une espèce. Ce n'est que depuis la fin du 20ème siècle que des méthodes formalisées ont été développées (ter Braak & Looman 1986) pour définir le caractère bioindicateur des espèces par rapport aux paramètres environnementaux. Un bioindicateur est un indice construit à partir d'une ou plusieurs espèces végétales, dont la présence ou l'absence, renseigne sur certaines caractéristiques écologiques de l'environnement.

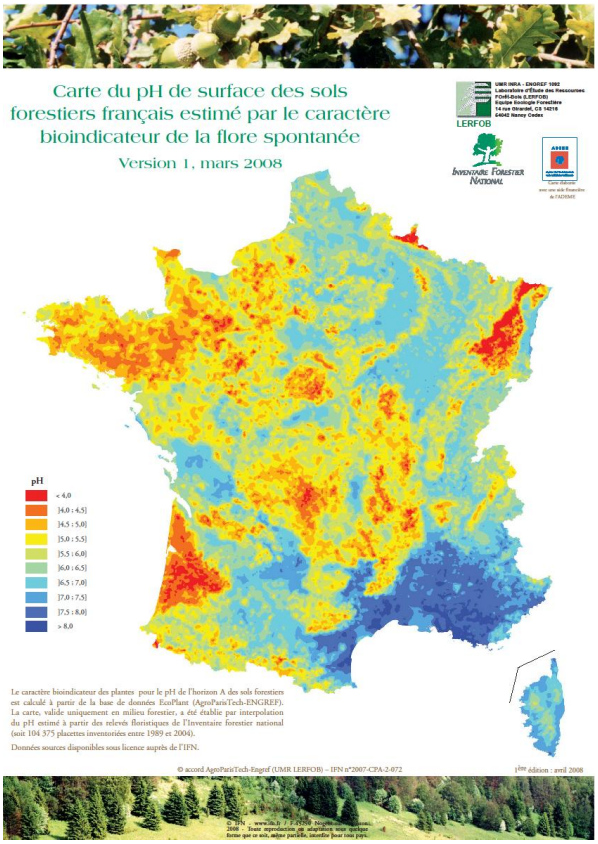
Les modèles statistiques utilisés pour la bioindication sont fréquemment des modèles linéaires généralisés (ter Braak & Looman 1986) reliant la variable du milieu aux relevés floristiques composés de variables binaires. Celles-ci indiquent les présences/absences de nombreuses espèces, les modèles sont soit paramétrique, soit non paramétrique en utilisant les arbres de régression ou de classification (Vayssières & al. 2000). De plus, les prédicteurs sont en grand nombre, souvent fortement corrélés et en interaction (Vayssières & al. 2000). De nombreux modèles peuvent être équivalents (au sens du critère utilisé : erreur quadratique moyenne, AIC, BIC, ...), ce qui rend délicate la sélection des variables à inclure pour obtenir le modèle final. D'autre part, les méthodes de bioindication actuellement applicables sur la France entière restent onéreuses car elles reposent sur des inventaires floristiques exhaustifs longs à réaliser (de 30 à 60 minutes). Par ailleurs, elles sont malaisées à mettre en œuvre par un non spécialiste, car elles présupposent la connaissance d'une part importante de la flore forestière française (plus de 500 espèces communes), et la capacité de déterminer les espèces dans une flore.

C'est pour cela que les objectifs du stage, visent à diminuer les connaissances préalables nécessaires à la bioindication, par les plantes, en diminuant le nombre d'espèces nécessaires (50 à 200 espèces). Le but étant d'optimiser la méthode de réalisation d'inventaires floristiques pour bioindiquer les qualités nutritionnelles comme le pH (potentiel hydrogène) et le C/N (nutrition

azotée) des sites par la méthode des forêts aléatoires (partie 3 & Breiman 2001). Le pH mesure la concentration d'une solution aqueuse en protons H+ et le degré d'acidité d'une solution.

Le C/N est le rapport entre le carbone organique et l'azote total d'un horizon (couche du sol). C'est un indice qui sert à caractériser globalement les matières présentes dans le sol mais également les apports et restitutions organiques.





Puis, j'étudierai le problème de stabilité en ajoutant du bruit à ces qualités nutritionnelles. Le jeu de données sur lequel j'ai travaillé est issu de la base de données EcoPlant qui comprend 3766 relevés phytosociologiques dans lesquels 680 espèces végétales sont notés en présence/ absence (1 : présence et 0 : absence de l'espèce). De plus, d'autres variables sont mesurées : pH, C/N (nutrition azoté), la température moyenne, le bilan hydrique, etc… Dans la première partie, je rappellerai les modèles utilisés au sein du LERFoB. Ensuite, j'expliquerai ce que sont les forêts aléatoires. Enfin, j'exposerai différentes façons de sélectionner les variables.

2) Modèles

Dans cette première partie, j'introduirai les modèles additifs généralisés (Wood 2006), sans rentrer dans les détails, car en début de stage j'ai pensé les utiliser pour résoudre la problématique mais cette méthode était inadaptée. Dans un premier temps, je rappellerai ce qu'est le modèle linéaire.

2.1) Modèles statistiques

L'introduction de modèles statistiques est un moyen de prendre en compte l'information à priori.

Définition :

Soit X une variable aléatoire dans (Ⅹ, Ғ), où Ғ est la tribu borélienne de Ⅹ.

On appelle **modèle statistique paramétrique** pour X tout triplet (Ⅹ, Ғ, $\{\mathbf{P}_\theta\}_{\theta\in\Theta}$), où $\{\mathbf{P}_\theta\}_{\theta\in\Theta}$ est une collection de probabilités sur (Ⅹ, Ғ) indexée par un ensemble d'indices quelconque Θ tel que $\mathbf{P}_X \in \{\mathbf{P}_\theta\}_{\theta\in\Theta}$.

De plus, un **modèle non paramétrique** est un modèle qui ne peut se mettre sous la forme paramétrique. On considèrera une collection de probabilités plus large pouvant avoir des formes différentes.

2.1.1) Modèle linéaire

On veut trouver une fonction *f* linéaire en X telle que Y = *f*(X), où Y est la variable à expliquer et X l'ensemble des p variables explicatives. Dans mon cas, Y représenterai le pH ou le C/N et X les espèces végétales.

Le modèle linéaire s'écrit alors de la manière suivante :

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i,\ 1 \leq i \leq n$$

Avec $\mathbb{E}(\varepsilon_i) = 0$ (l'espérance des ε_i), mais pas nécessairement de même variance, ni indépendants.

Sous forme matricielle, le modèle peut s'écrire :

$Y = X\beta + \varepsilon$ où $Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$, $\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$ représentant le vecteur des résidus, $X = \begin{pmatrix} 1 & \cdots & x_{p1} \\ \vdots & \ddots & \vdots \\ 1 & \cdots & x_{pn} \end{pmatrix}$, la matrice déterministes des prédicteurs, $\beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix}$, le vecteur des paramètres inconnus à la régression. Comme $\mathbb{E}(\varepsilon_i) = 0$, $\mathbb{E}(Y) = X\beta$ et $Var(\varepsilon) = V$.

Si V=$\sigma^2 I_n$ on parle de moindres carrés ordinaires et si V est quelconque symétrique semi définie positive, on parle de moindres carrés généralisés.

Pour estimer β, dans le cadre moindre carré ordinaire, on utilise le théorème de Gauss Markov (vu en cours de modèle linéaire) :

Théorème de Gauss Markov :

Si (X'X), où X' est la matrice transposé de X, est inversible, alors $\hat{\beta}$ est l'estimateur de Gauss Markov de β : $\hat{\beta} = (X'X)^{-1}X'Y$, c'est l'estimateur sans biais à variance minimale (ESBVM) parmi tous les estimateurs sans biais linéaire de β et Var($\hat{\beta}$) = $\sigma^2(X'X)^{-1}$.

La démonstration de ce théorème ne sera pas faite dans ce rapport, car le but de cette partie est uniquement d'introduire le modèle additif généralisé.

Sous R, (R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http: //www.R-project.org/.), pour créer un modèle on utilise la fonction **lm()** du package stats (R Core Team and contributors worldwide).

Cette fonction s'utilise de la manière suivante :

lm(formula,data), où **formula** est la formule voulue pour le modèle et **data** représente le jeu de données.

Ensuite, en faisant un **summary()**, de la sortie de la fonction **lm()**, on obtient l'estimation des paramètres dans la colonne *Estimate*. De plus, il est intéressant de regarder au niveau du *Multiple R-squared*, pour juger de la qualité du modèle. Il peut être intéressant de simplifier le modèle, en ne gardant que les variables les plus informatives, cela se fait avec la fonction **step(model)** où **model** représente le modèle créé par la fonction **lm()**. Enfin, on peut faire la prédiction sur un nouveau jeu de données (**newdata**), sur le modèle linéaire créé en utilisant la fonction **predict(model,newdata)**.

2.1.2) Modèle additif généralisé

Le modèle additif généralisé (GAM) est une extension du modèle linéaire généralisé, car dans le modèle GAM, les prédicteurs ne seront pas forcément linéaire. En effet, on pourra approcher la variable à expliquer Y par une combinaison linéaire de fonctions non paramétriques en les variables explicatives X_j, $1 \leq j \leq p$. Ces modèles font partie des modèles non paramétriques.

Dans les modèles GAM, il est nécessaire que la loi de distribution de Y appartienne à la famille exponentielle. Le modèle s'écrit:

$$g(\mu_i) = \alpha + \sum_{j=1}^{p} f_j x_{ij} \ , 1 \leq i \leq n$$

Où $\mu_i = \mathbb{E}(y_i)$, g est une fonction connue, appelée fonction de lien, monotone et deux fois dérivables. Cette fonction peut avoir plusieurs formes comme logistique ou identité. Les fonctions f_j sont des fonctions décomposables sur une base de fonction (polynomiale, splines,..). Chaque f_j aura pour forme:

$$f(x) = \sum_{i=1}^{q} b_i(x)\beta_i$$

Avec b_i les éléments de la base de fonction et β_i les paramètres à estimer. Ces fonctions peuvent être estimées par décomposition sur une base de fonctions splines. Celles-ci représentent des transformations polynomiales par morceaux.

Sous R, on peut utiliser les packages mgcv (Wood, S.N. (2006) Generalized Additive Models: An Introduction with R. Chapman and Hall/CRC.) ou le package gam (Trevor Hastie (2013). gam: Generalized Additive Models. R package version 1.08. http://CRAN.R project.org/package=gam)

Les modèles additifs généralisés peuvent être construits avec le package mgcv, en utilisant la fonction **gam()**, elle s'utilise de la manière suivante:

gam(formula,data), où **formula** est la formule voulue pour le modèle et **data** représente le jeu de données.

La différence avec un modèle linéaire classique est qu'on peut y inclure des termes de lissage **s()** et **te()**, correspondant aux splines. Le terme **te()** sera utilisé dans un cadre multidimensionnel pour prendre en compte des effets d'interaction entre les variables. Enfin, on peut faire la prédiction sur un nouveau jeu de données (**newdata**), sur le modèle GAM créé (**modelgam**), et en utilisant la fonction **predict(modelgam,newdata)**.

Néanmoins, cette méthode est inadaptée à notre jeu de données car il est impossible de mettre des splines sur des variables binaires, mais uniquement sur des variables quantitatives. C'est pour cela qu'avec mon tuteur, nous nous sommes intéressés aux forêts aléatoires (Breiman, 2001). Cette méthode fait partie des modèles non paramétrique et est utilisée en apprentissage statistique, introduite par Vladimir Vapnik (1995).

3) Apprentissage statistique

3.1) L'apprentissage statistique

En apprentissage statistique, nous disposons d'un échantillon d'apprentissage $\{(X_1,Y_1), \ldots, (X_n,Y_n)\}$, qui est une suite de vecteurs aléatoires indépendants et identiquement distribuées (i.i.d.), qui est de même loi que (X,Y). De plus, (X, Y) est indépendant de l'échantillon d'apprentissage et sa loi est inconnue. L'entier naturel *n* représente le nombre d'observations de l'échantillon d'apprentissage. Etant donné que la loi de (X, Y) est inconnue, on essaie d'apprendre cette loi par l'échantillon d'apprentissage. En particulier, on essaie de comprendre le lien entre la variable Y (considérée comme la variable réponse) et la variable X (considérée comme variable explicative). La méthode statistique doit être capable de prédire la variable réponse et celle-ci doit être la plus proche possible de la vraie valeur, associé à la variable explicative.

Il existe deux cadres en apprentissage statistique, la régression et la classification. Dans les limites de mon stage je m'intéresse uniquement à la partie régression. C'est pour cela que j'introduis les arbres de régression afin de mieux comprendre les forêts aléatoires.

3.2) Arbre de régression

Les arbres de décision, classification et régression, ont été introduit par Léo Breiman et al. (1984). On parle souvent de CART (Classification And Regression Trees), qui est l'abréviation du titre du livre, qui construit des prédicteurs par arbre en régression. La méthode consiste à découper l'espace des variables explicatives $X_1,\ldots, X_p$ dans le but d'expliquer Y.

Un arbre de régression est construit par une procédure itérative. Dans cette procédure, on commence par chercher une règle de division binaire $d = d(x^m,s)$, du type $x^m \leq s$ (s appartenant à $\mathbb{R}$ (ensemble des réels), si X^m est quantitative) ou x^m appartient à S (où S est est un sous-ensemble de l'ensemble des modalités de X^m, X^m est qualitative). Ceci permet de partager l'ensemble des observations initiales, noté t_0, et dit racine de l'arbre, en deux sous-ensembles, t_g et t_d, dits nœuds descendants de t_0. Parmi tous les partages possibles explorés sur toutes les variables explicatives et tous les seuils, on retient celui qui minimise le critère suivant :

$$\sum_k (y_k - \overline{y_t})^2$$

Où $\overline{y_t}$ est la moyenne des y_k des observations présentes dans le nœud t. Ce critère représente, la somme des carrés des écarts à la moyenne. Une fois, cette règle obtenue et le partage effectué, on recommence la même procédure de partage appliquée aux nœuds t_g et t_d.

Voici un exemple d'arbre :

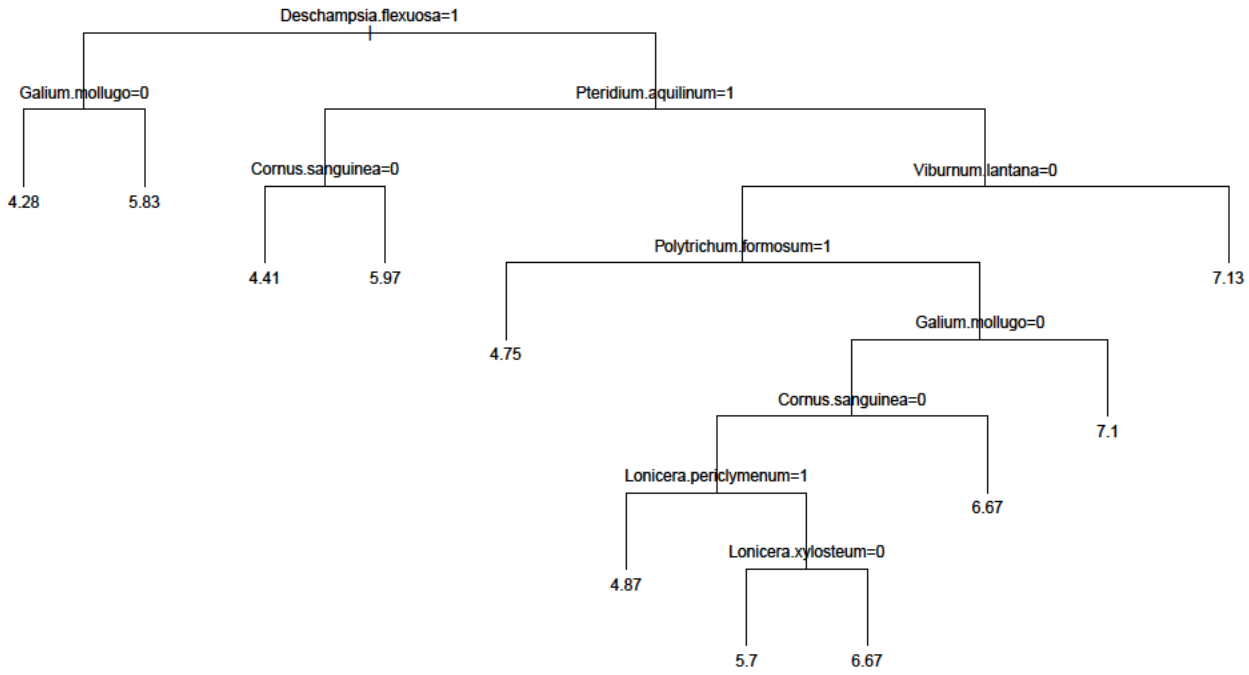


Figure 2 : Arbre de régression

Dans mon cas, les variables à expliquer sont le pH et le C/N (nutrition azotée), les variables explicatives sont les espèces végétales formant le jeu de donnée. Ces variables sont des variables binaire, 1 : présence, 0 : absence, de l'espèce. L'arbre construit ci-dessus n'est formé que par quelques espèces afin de mieux comprendre et de mieux visualiser les arbres de régression mais sinon le jeu de donnée possède 680 espèces végétales. Au bout de chaque branche qui constitue l'arbre, il y a la valeur prédite de la variable à prédire, dans l'exemple la variable est le pH. Pour construire un tel arbre sous R, j'utilise le package rpart (Terry Therneau, Beth Atkinson and Brian Ripley (2012). rpart: Recursive Partitioning. R package version 4.1-0.) et en utilisant la fonction **rpart()** de la manière suivante :

rpart(formula, data), où **formula** est la formule voulue pour le modèle et **data** représente le jeu de données.

Chacun des nœuds est divisé comme je l'ai montré précédemment, jusqu'à ce l'on décide de s'arrêter. Le critère de l'arrêt du découpage du nœud est arbitraire, mais le plus répandu est de ne pas découper le nœud si le nombre d'observations dans l'échantillon d'apprentissage contenus dans le nœud est inférieur à un certain nombre. De plus, j'évoquerai aussi, sans entrer dans les détails, la notion d'élagage, qui consiste à choisir de manière convenable les nœuds pour éviter d'avoir un arbre trop complexe ou des problèmes de surapprentissage. En effet, lorsque le modèle est suffisamment complexe pour passer très précisément par tous les points, c'est le phénomène de surapprentissage. Pour élaguer l'arbre, j'utilise la fonction **prune()** du package rpart cité ci-dessus. Cette fonction s'emploie de la manière suivante : **prune(arbre)**, où **arbre** représente l'arbre construit avec la fonction **rpart()**.

Voici un exemple d'arbre élagué :

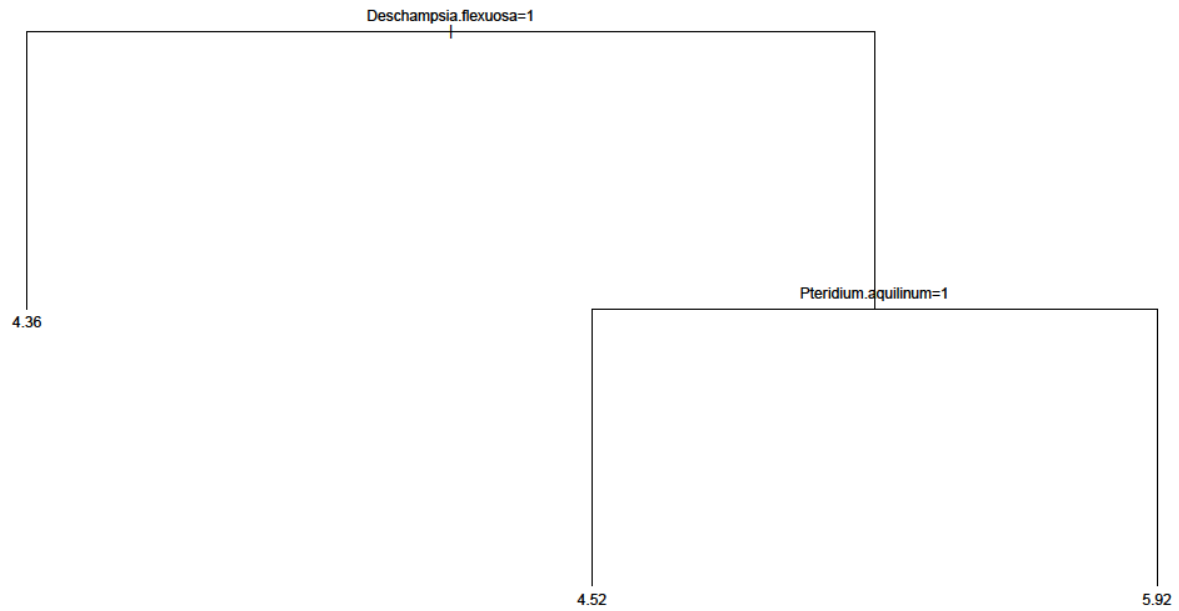


Figure 3 : Arbre ci-dessus élagué

Les avantages des arbres de régression sont multiples : c'est une méthode simple, facile à comprendre, à utiliser et à interpréter mais aussi c'est une méthode qui tient en compte les interactions entre les variables. Néanmoins, c'est une méthode instable qui présente une grande variance.

Pour remédier à ce problème, Breiman (2001), introduit les forêts aléatoires. Ceci va être le sujet de la partie suivante. Cette méthode consiste à la construction de manière aléatoire d'un très grand nombre d'arbres de régression avant d'en faire la moyenne.

3.3) Forêts aléatoires

Définition : *Soit $\{\hat{h}(., \Theta_1), \ldots, \hat{h}(., \Theta_q)\}$ une collection de prédicteurs par arbre, où $(\Theta_1, \ldots, \Theta_q)$ est une suite de variables aléatoires indépendantes et identiquement distribuées (i.i.d.), indépendante de l'échantillon d'apprentissage. Le prédicteur des forêts aléatoires est obtenu par agrégation (en faisant la moyenne des arbres) de cette collection de prédicteurs.*

Les arbres qui composent la forêt sont construits ainsi : On génère tout d'abord plusieurs échantillons bootstrap. Ensuite, sur chaque échantillon une variante de CART est appliquée. Plus exactement, on choisit aléatoirement *l* (*l*<*n*) observations parmi les *n* observations qui composent le jeu de données. Pour découper un nœud, on tire aléatoirement *m* variables parmi les *p* et on cherche parmi celles-ci la meilleure coupure (avec toujours le même critère). De plus, les arbres construits ne sont pas élagués. Sous R, pour générer les forêts aléatoires, j'utilise la fonction randomForest qui est incluse dans le package possédant le même nom (A. Liaw and M. Wiener (2002). Classification and Regression by randomForest. R News 2(3), 18--22.). Il y a deux paramètres dans cette fonction à préciser, *ntree* représentant le nombre d'arbres composant la forêt et le paramètre *mtry* représentant le nombre de variables d'entrées choisit de manière aléatoire (bootstrap) à chaque division.

Le tirage, à chaque nœud, des *m* variables, se fait, sans remise, et uniformément parmi toutes les variables (chaque variable a une probabilité 1/*p* d'être choisie). Le nombre *m* ($m \leq p$) est fixé au début de la construction de la forêt et est identique pour tous les arbres. C'est un paramètre très important de la méthode.

Enfin, pour faire la prédiction, on fait la moyenne de toutes les prédictions des arbres composant la forêt et qui n'ont pas étaient pris dans le bootsrap. Plus précisément, fixons une observation (X_i, Y_i) de l'échantillon d'apprentissage. On considère maintenant, l'ensemble des arbres construits sur les échantillons bootstrap ne contenant pas cette observation, c'est-à-dire pour lesquels cette observation est « Out-Of-Bag ». Nous agrégeons alors uniquement les prédictions de ces arbres pour fabriquer la prédiction $\widehat{y_k}$ de y_k.

Liaw et al. 2002 proposent un algorithme pour mieux comprendre

1) Construire ntree (nombres d'arbres) échantillons bootstrap à partir des données,
2) Pour chaque échantillon, construire l'arbre non élagué de la manière suivante : à chaque nœud, au niveau des prédicteurs, le choix de la variable à partagé se fera sur un échantillon aléatoire mtry (le nombre de variables sélectionnées pour chaque nœud) et on choisit parmi celles-ci la meilleure coupure.

3) Pour la prédiction sur un nouveau jeu de données, on agrège les prédictions des ntree arbres.

Hastie et al.2008 proposent un autre algorithme :

I) Pour b=1,…, B
- Construire un échantillon bootsrap Z* de taille N à partir des données
- Générer une forêt aléatoire T_b avec les données de l'échantillon bootstrap, en répétant les étapes suivantes pour chaque nœud terminal de l'arbre, jusqu'à ce que la taille minimale de nœud n_{min} soit atteinte
  i) Sélectionner de manière aléatoire *m* parmi les p disponibles,
  ii) Choisir la meilleure variable entre les *m*,
  iii) Diviser le nœud en deux nœuds fils.

II) Récupérer l'ensemble des arbres $\{T_b\}_1^B$

Pour faire une prédiction au point y : $\widehat{f_{rf}^B}(y) = \frac{1}{B}\sum_{b=1}^{B} T_b(y)$

De plus, Hastie et al. 2008, donne une explication au fait que les forêts aléatoires permettent de réduire la variance. Si les prédictions des arbres sont identiquement distribuées, de variance σ², avec un coefficient de corrélation deux à deux ρ, la variance de la moyenne de B prédictions est alors : $\rho\sigma^2 + \frac{1-\rho}{B}\sigma^2$

Quand le nombre d'arbres B de la forêt est grand, le second terme disparaît, mais si la corrélation entre les arbres est trop forte le premier terme reste. L'idée des forêts aléatoires, est de diminuer cette variance en décorrélant, les arbres, autant que possible les uns avec les autres.

Pour faire la prédiction, on fait la moyenne de toutes les prédictions de certains arbres composant la forêt (comme vu précédemment). Néanmoins, lors de cette prédiction une erreur est commise, elle est appelée l'erreur Out Of Bag (erreur OOB).

L'erreur OOB

L'erreur OOB se calcule de la manière suivante : soit une observation (X_i,Y_i), on considère maintenant les arbres construits sur les échantillons ne contenant pas cette observation, cette observation est dite out of bag, on agrège alors les prédictions de ces arbres pour former notre prédiction $\widehat{y_k}$ dey_k.

Cette erreur est l'erreur quadratique moyenne donnée par la formule suivante :

$$errOOB_s = \frac{1}{n}\sum_{k=1}^{n}(y_k - \widehat{y_k})^2$$

En utilisant le logiciel R, on utilise la fonction la fonction **randomForest()** pour générer une forêt aléatoire. Elle s'utilise de la manière suivante :

randomForest(formula, data, ntree, mtry), où **formula** est la formule voulue pour le modèle et **data** représente le jeu de données, **ntree** le nombre d'arbres formant la forêt (par défaut le nombre est fixé à 500) et **mtry**, le nombre de variables sélectionnées pour chaque nœud. Ce paramètre peut être choisi avec la fonction **tuneRF()** afin d'obtenir une erreur quadratique plus petite. Car, par défaut, et dans un cadre de régression ce paramètre vaut p/3, où p représente de nombre de variables explicatives. Cette fonction estime l'erreur OOB, pour mtry=p/3, p/6 et 2p/3, c'est ce que suggère Breiman. De plus, c'est la procédure de Liaw et al. (2002), décrite ci-dessus, qui est implémentée dans ce package.

En sortie R, après avoir utiliser la fonction **randomForest()**, on a « *mean of squared residuals* », qui représente l'erreur OOB et « *percent variance explained* », qui est calculé par la formule suivante : 1 - mean of squared residuals/var(Y), avec var(Y) la variance de la variable à expliquer, celui-ci représente un 'genre' de R^2, afin de juger de la qualité du modèle. Enfin, il est possible de prédire sur un nouveau jeu de données suivant le modèle de forêts aléatoire, en utilisant la fonction **predict()** : **predict(modelrf,newdata)**, avec **modelrf** est le modèle de forêt aléatoire et **newdata**, le nouveau jeu de données.

L'usage des forêts aléatoires est utilisé en bio-informatique, dans un cadre de régression ou dans un cadre de classification (Diaz Uriarte et al. 2006) pour sélectionner les variables. Car dans ces cas là, le nombre de variables explicatives est bien supérieur aux nombre d'observation. C'est ce que je développerai dans la partie suivante.

4) Sélection de variables

Il y a deux objectifs distincts en sélection de variables, l'un est l'interprétation et l'autre la prédiction. L'objectif d'interprétation cherche à sélectionner toutes les variables X^r fortement reliées à la variable réponse Y (même si les variables X^r sont corrélées entre elles). L'objectif de prédiction cherche à sélectionner un petit sous ensemble de variables suffisant, pour bien prédire la variable réponse. Je présenterai une méthode générale et une méthode un peu plus complète, Genuer et al. (2010), pour faire ces sélections.

4.1) Sélection générale

L'objectif est de trouver un sous-ensemble de variables importantes suffisant pour la prédiction. Pour calculer l'importance d'une variable X^r, noté VI(X^r), Breiman (2001) introduit un indice, dont j'ai pu trouver une explication assez simple dans la thèse de Genuer (2010) : si on permute aléatoirement la r-ième variable, on obtient un échantillon perturbé, noté $\widetilde{OOB}_s^r$, plus les permutations causent une forte augmentation de l'erreur, plus elle est importante. De plus, cet indice permet de distinguer les variables importantes et pertinentes des variables qui ne le sont pas. Cet indice est calculé par la formule suivante :

$$\mathrm{VI}(\mathrm{X^r}) = \frac{1}{ntree} \sum_{s=1}^{ntree} (err\widetilde{OOB}_s^r - errOOB_s)$$

Avec *ntree* représentant le nombre d'arbres qui composent la forêt.

Une première procédure de sélection a consisté à retenir les q<p variables les plus importantes, puis de générer de nouveau une forêt avec ces variables. Puis, on continue ainsi, jusqu'à obtenir un nombre suffisant de variables pour lesquels l'erreur OOB n'est pas trop élevé ni trop faible. Car si l'erreur OOB est trop élevé le modèle n'est pas bon, à l'inverse s'il est trop faible le modèle est trop bon, c'est alors le surapprentissage. Voici les résultats pour cette sélection de variables. Avec le logiciel R, il y a un graphique intéressant dans le package randomForest qui s'appelle le varImPlot(), en voici un exemple, celui se trace avec la fonction **varImPlot()**. Ce graphique permet de voir le classement des variables des plus importantes au moins importantes. Dans l'exemple, il n'y a que quelques espèces parmi les 680 composants le jeu de données.

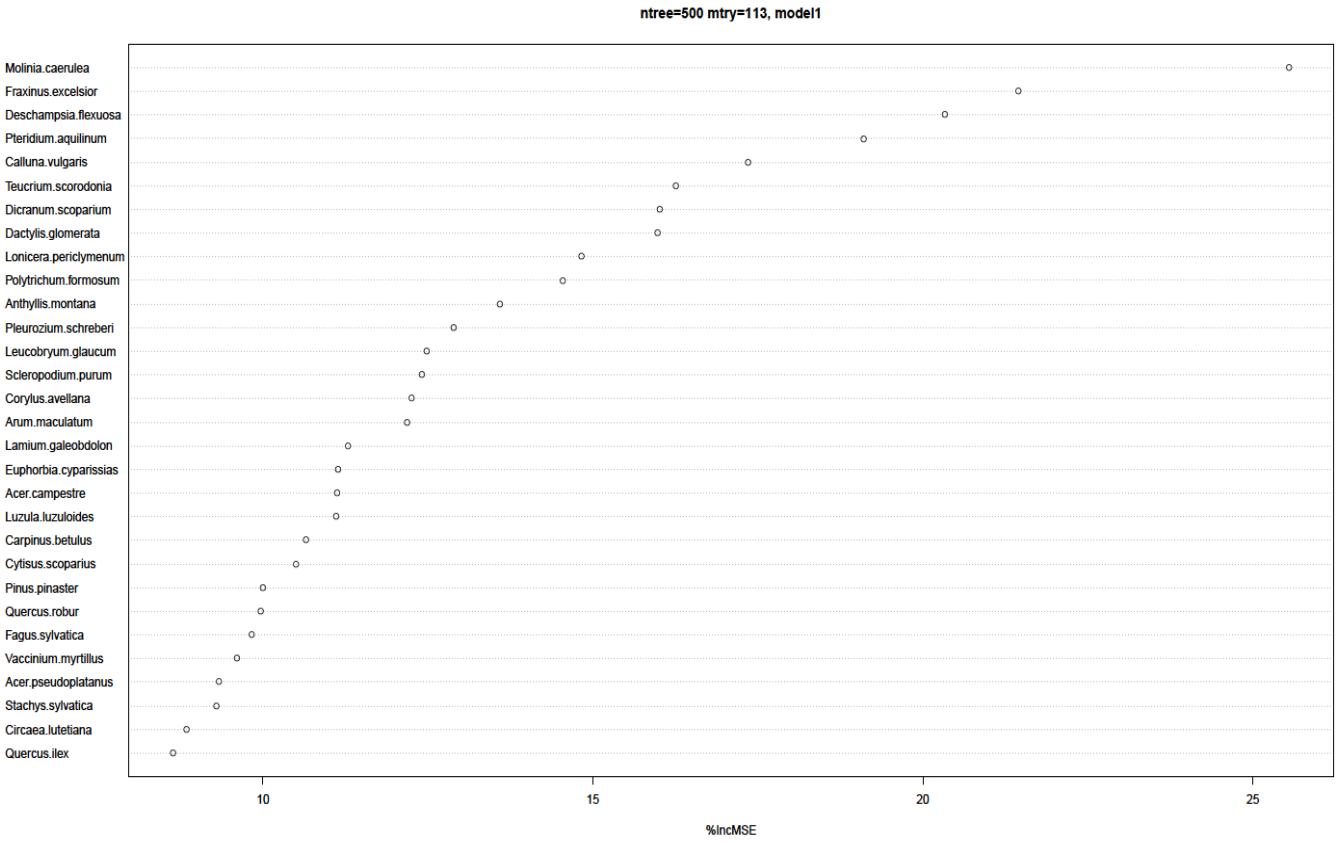


Figure 4 : les 30 espèces les plus importantes pour modéliser le C/N.

Ensuite, je traçais l'importance de chaque variable de manière croissante, afin de supprimer les variables qui possédaient une importance négative, voici le graphique :

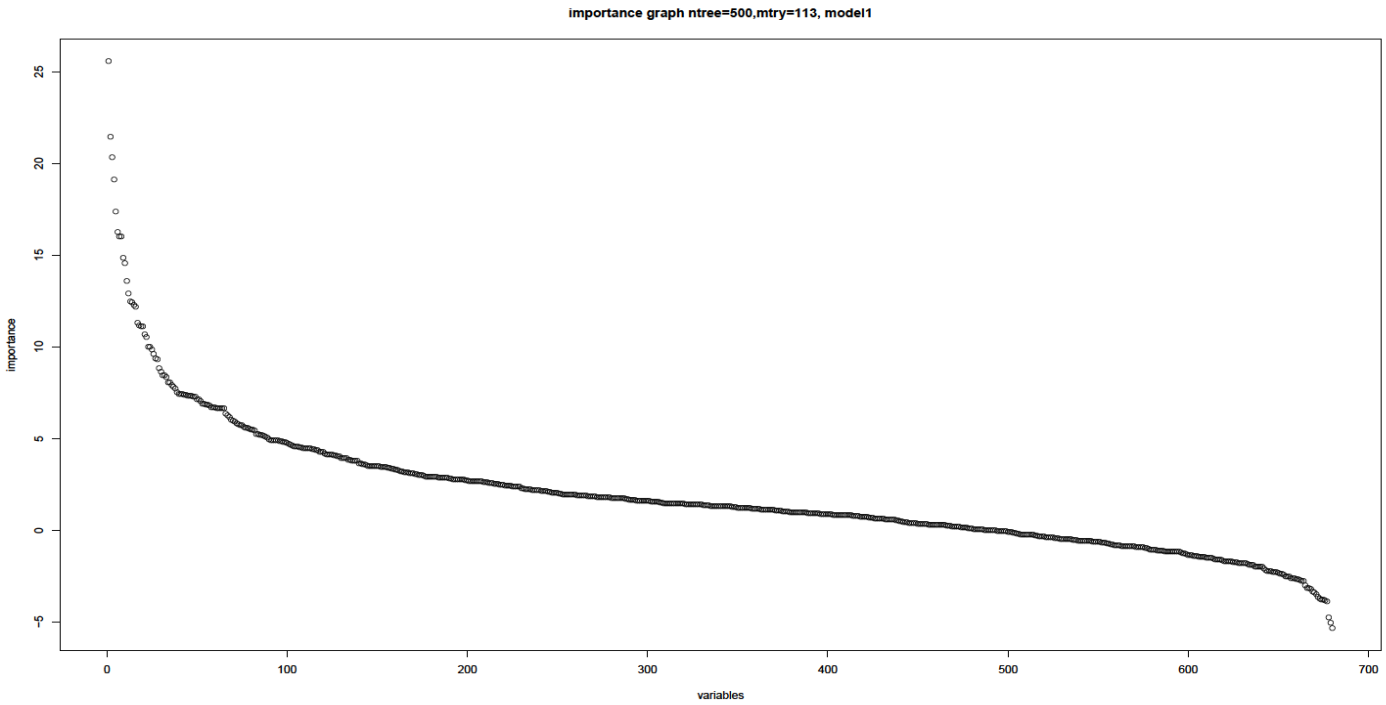


Figure 5 : Importance de toutes les variables explicatives pour modéliser le C/N.

4.1.1) Résultats

Tableau 1 : Résultats pour le pH

Modèles	ntree	mtry	% var explained	OOB	OOB test
model1	500	113	80.71	0.3701461	0.408727
model2	500	95	80.54	0.3792106	0.3629095
model3	500	83	80.7	0.3701666	0.3744479
model4	500	75	80.66	0.3697253	0.3606537
model5	500	133	79.57	0.3854059	0.3913594
model6	500	100	80.1	0.376245	0.3602282
model7	500	66	80.19	0.3788355	0.3719771
model8	500	17	78.65	0.4035619	0.3499556
model9	500	8	78.81	0.4154642	0.4494755
model10	500	6	69.7	0.580142	0.6045078
model11	500	6	62.13	0.7279489	0.7031214

Dans ce tableau (et les suivants), ntree représente le nombre d'arbres constituant la forêt, mtry le nombre de variables d'entrées choisit de manière aléatoire à chaque division, % var explained le pourcentage de variance expliquée, OOB l'erreur OOB donnée ci-dessus et OOB test l'erreur OOB sur un échantillon test.

Je constate que lorsque je diminue le nombre de variables, l'erreur OOB à tendance a stagné entre 0.37 et 0.385. Puis, lorsque je diminue fortement le nombre de variables, il augmente jusqu'à atteindre 0.72. De plus, l'erreur dans l'échantillon test est du même ordre de grandeur que l'erreur OOB. On voit que le modèle à 200 variables est intéressant, mais cela reste encore trop élevé pour les spécialistes forestiers. Le modèle a seulement 50 variables reste possible.

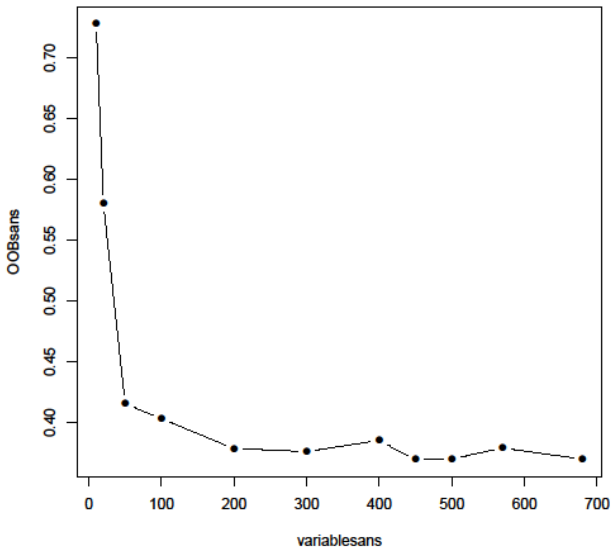


Figure 6 : erreur OOB en fonction des variables pour le pH

Je fais de même pour le C/N, et les résultats sont résumés dans le tableau suivant :

Tableau 2 : Résultats pour le C/N

Modeles	Ntree	Mtry	OOB	% var explained	OOB test
model1	500	113	13.91272	48.97	14.91156
model2	500	82	14.29021	48.62	13.54001
model3	500	65	14.17396	47.23	14.02107
model4	500	110	14.86044	47.04	12.65513
model5	500	200	14.30345	47.42	14.6458
model6	500	33	14.02941	49.46	13.79493
model7	500	33	13.36386	50.41	15.9965
model8	500	8	14.98252	47.27	13.07953
model9	500	3	16.02821	41.9	16.5562
model10	500	3	17.20526	37.13	18.24582

Je constate que lorsque je diminue le nombre de variables, l'erreur OOB à tendance a stagné entre 13.36 et 14.98. Puis, lorsque je diminue fortement le nombre de variables, il augmente jusqu'à atteindre 17.2. De plus, l'erreur dans l'échantillon test n'est pas vraiment du même ordre de grandeur que l'erreur OOB. Le modèle à 100 variables est intéressant.

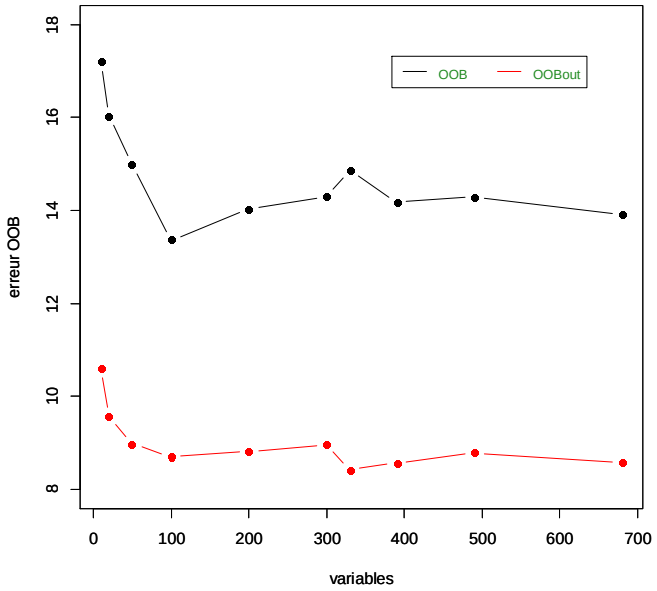


Figure 7 : erreur OOB en fonction des variables pour le C/N

La courbe en rouge représente les résultats sans les valeurs extrêmes. En effet, pour le C/N, il y avait des valeurs très grandes et qui sortaient de la boîte à moustache. Voici, la représentation de la boîte pour le pH et le C/N.

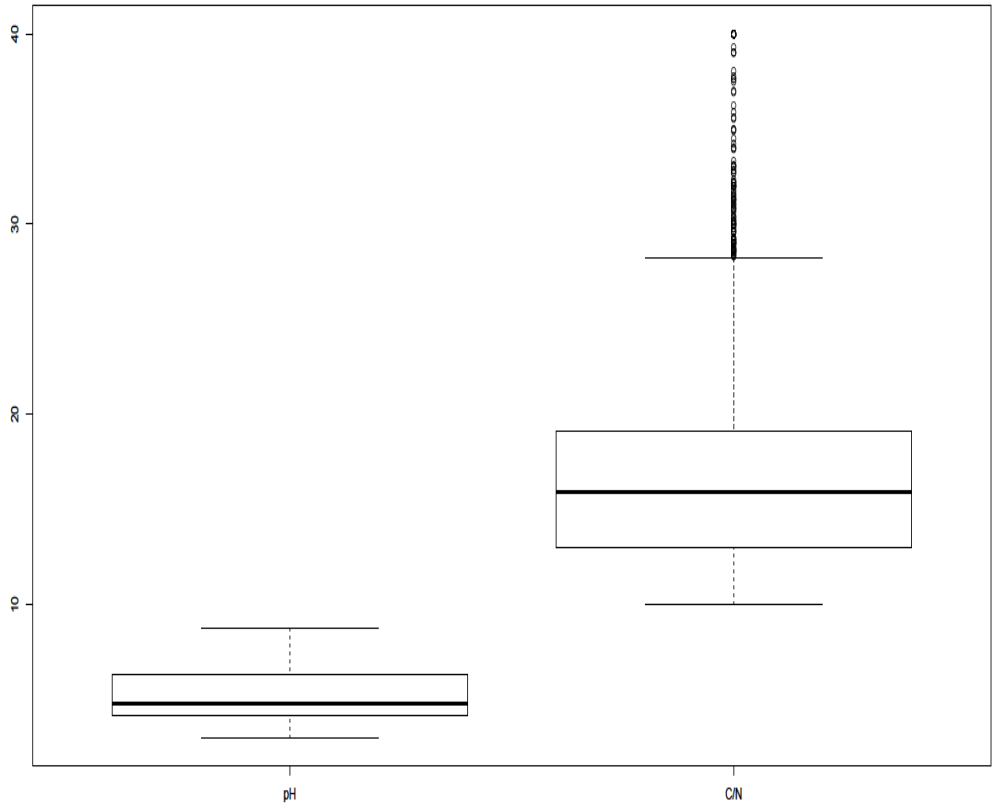


Figure 8 : boîte à moustache du pH (à droite) et C/N

Voici les résultats du C/N sans les valeurs extrêmes :

Tableau 3 : Résultats pour le C/N sans valeurs extrêmes

modèles	ntree	mtry	OOB	% var explained	OOB test
model1out	500	113	8.587089	48.73	8.576001
model2out	500	82	8.7909	47.15	8.286449
model3out	500	65	8.574686	49.25	8.682527
model4out	500	55	8.420992	49.54	9.016289
model5out	500	50	8.960605	45.99	7.681703
model6out	500	33	8.817038	47.72	8.213279
model7out	500	17	8.710304	47.88	8.439058
model8out	500	8	8.973449	44.11	8.960491
model9out	500	6	9.565722	42.41	10.62441
model10out	500	6	10.61029	36.74	10.6733

Je remarque que, sans les valeurs extrêmes, l'erreur OOB diminue significativement. Je constate que lorsque je diminue le nombre de variables, l'erreur OOB à tendance a stagné

entre 8.42 et 8.97. Puis, lorsque je diminue fortement le nombre de variables, il augmente jusqu'à atteindre 10.61. De plus, l'erreur dans l'échantillon test n'est pas vraiment du même ordre de grandeur que l'erreur OOB. Le modèle à 100 variables est intéressant.

De plus, le pourcentage de variance expliquée pour le pH est assez élevé, allant parfois même atteindre les 80%. Néanmoins, pour le C/N, avec ou sans valeurs extrêmes, le pourcentage est assez faible. Ceci peut se voir en traçant les graphiques des valeurs prédictes en fonction des valeurs observées.

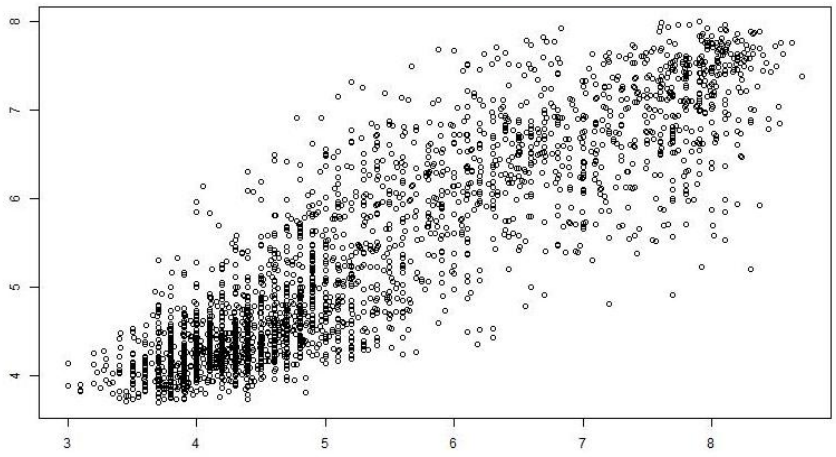


Figure 9 : Graphique représentant les valeurs prédictes en fonction de celles observées pour le pH

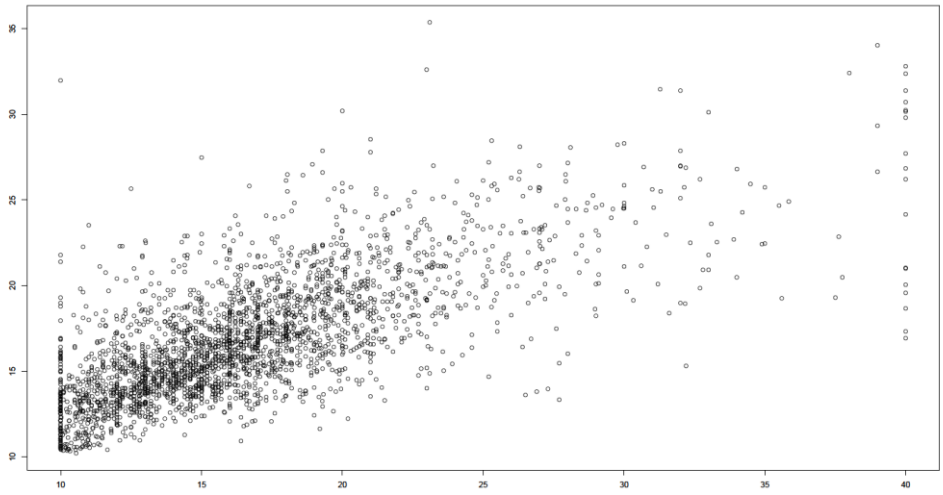


Figure 10 : Graphique représentant les valeurs prédictes en fonction de celles observées pour le C/N (avec les valeurs extrêmes)

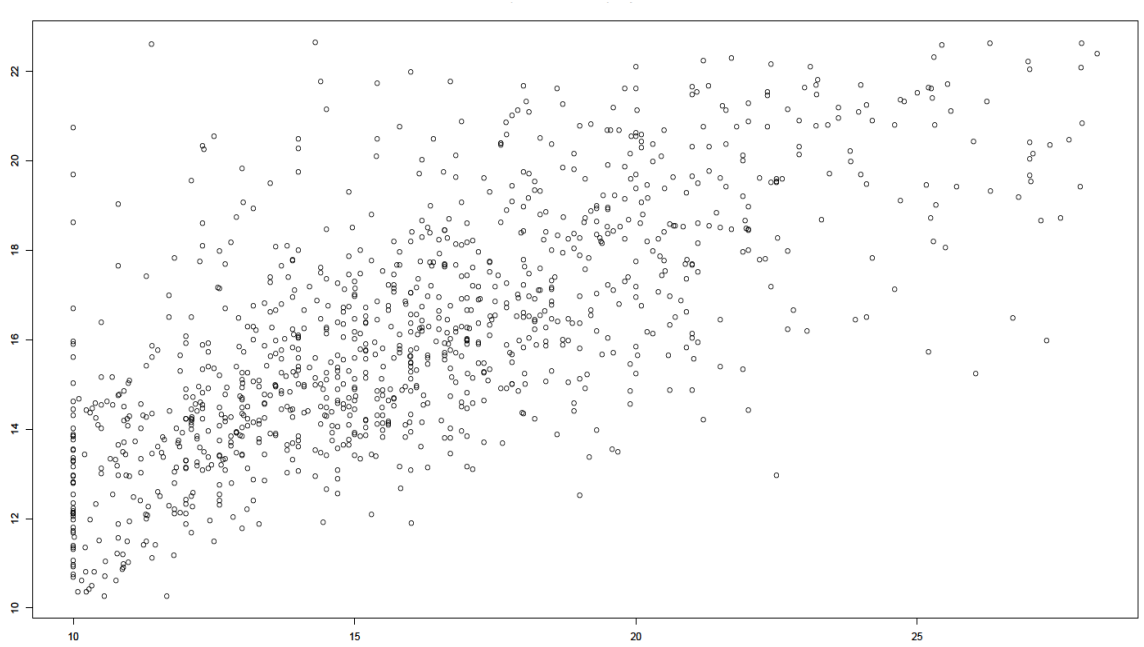


Figure 11 : Graphique représentant les valeurs prédictes en fonction de celles observées pour le C/N (sans les valeurs extrêmes)

Néanmoins, pour le C/N, les bornes proviennent d'une convention empirique (en dehors les valeurs ne paraissent pas plausibles). En effet, pour les valeurs de C/N trop élevées (40 unités de C/N) ou basses (10 unités de C/N), on peut remarquer qu'il y a une bande qui se forme sur la figure 10. Cette bande disparait (figure 11) pour les grandes valeurs de C/N, les valeurs extrêmes, par contre cette bande reste pour les faibles valeurs de C/N.

Enfin, avec cette sélection de variables, les variables (espèces végétales) les plus importantes sont les suivantes :

Pour le pH : *Pteridium aquilinum, Lonicera periclymenum, Deschampsia flexuosa, Polytrichum formosum,Viburnum lantana, Acer campestre, Teucrium chamaedrys, Cornus sanguinea, Erica arborea,Calluna vulgaris, Lonicera xylosteum, Fraxinus excelsior, Thymus vulgaris, Euphorbia cyparissias,Viola reichenbachiana, Fagus sylvatica, Quercus robur, Brachypodium sylvaticum, Fragaria vesca,Atrichum undulatum, Athyrium filix femina, Ligustrum vulgare, Dryopteris filix mas, Crataegus laevigata, Carpinus betulus, Rubus fruticosus, Galium odoratum, Galium mollugo,Dryopteris dilatata, Dryopteris carthusiana, Betula pendula, Hieracium murorum, Teucrium scorodonia, Ribes alpinum, Vaccinium myrtillus, Sesleria caerulea, Abies alba, Sorbus aria, Brachypodium pinnatum, Epipactis helleborine, Phillyrea angustifolia, Rosa arvensis,Bromus erectus, Ulmus minor, Sorbus aucuparia, Dactylis glomerata, Juniperus communis,Cardamine pentaphyllos, Lonicera alpigena, Galium aparine.*

Pour le C/N : *Pinus pinaster, Pteridium aquilinum, Dicranum scoparium, Deschampsia flexuosa, Polytrichum formosum,Dactylis glomerata, Fraxinus excelsior, Molinia caerulea, Lonicera periclymenum, Teucrium scorodonia,Euphorbia cyparissias, Corylus avellana, Abies alba, Calluna vulgaris, Pleurozium schreberi, Luzula luzuloides,Atrichum undulatum, Quercus suber, Fagus sylvatica, Anthyllis montana, Picea abies, Acer pseudoplatanus,Carpinus betulus, Cytisus scoparius, Lamium galeobdolon, Acer campestre, Rubia peregrina, Prenanthes purpurea, Crataegus monogyna, Plagiomnium undulatum, Euphorbia characias, Scleropodium purum, Dryopteris dilatata, Quercus ilex, Arum maculatum, Leucobryum glaucum,Vaccinium myrtillus, Circaea lutetiana, Quercus petraea, Brachypodium sylvaticum,Linum suffruticosum subsp. appressum, Galeopsis tetrahit, Thymus serpyllum, Glechoma hederacea,Eryngium campestre, Viola reichenbachiana, Castanea sativa, Rubus idaeus, Galium mollugo, Prunus spinosa Oxalis acetosella, Athyrium filix femina, Rubus fruticosus, Quercus robur, Hedera helix, Polygonatum verticillatum, Carex pilulifera, Galium odoratum, Festuca altissima, Holcus mollis, Agrostis capillaris, Sorbus aucuparia, Orthilia secunda, Frangula alnus, Carex sylvatica, Dryopteris carthusiana,Dryopteris filix mas, Cornus sanguinea, Eurhynchium striatum, Alnus glutinosa, Bromus erectus, Sorbus aria,Urtica dioica, Sorbus torminalis, Lavandula angustifolia, Quercus humilis, Polygonatum multiflorum Galium corrudifolium, Rosa arvensis, Hieracium pilosella, Anemone nemorosa, Genista pilosa, Ajuga reptans,Quercus rubra, Vincetoxicum hirundinaria, Carex brizoides, Digitalis purpurea, Plantago media, Prunus avium,Hieracium murorum, Trifolium pratense, Carex flacca, Crataegus laevigata, Melampyrum sylvaticum, Phyteuma orbiculare, Deschampsia cespitosa, Genista anglica, Cistus salviifolius, Evonymus europaeus, Valeriana montana*.

Pour le C/N (sans valeurs extrêmes) : *Molinia caerulea, Fraxinus excelsior, Deschampsia flexuosa, Dicranum scoparium, Arum maculatum, Dactylis glomerata, Fagus sylvatica, Anthyllis montana, Euphorbia cyparissias,,Abies alba, Pteridium aquilinum, Scleropodium purum, Vaccinium myrtillus, Polytrichum formosum, Pinus pinaster, Quercus ilex, Lamium galeobdolon, Prunus spinosa, Calluna vulgaris, Atrichum undulatum, Carpinus betulus, Corylus avellana, Rubia peregrina, Pleurozium schreberi, Rumex acetosa, Galeopsis tetrahit, Crataegus monogyna, Quercus petraea, Fragaria vesca, Leucobryum glaucum, Brachypodium sylvaticum, Phillyrea angustifolia, Galium mollugo, Lonicera periclymenum, Teucrium scorodonia, Acer pseudoplatanus, Orthilia secunda, Geranium robertianum, Viola.reichenbachiana, Plagiomnium.undulatum, Sorbus.aucuparia, Hippocrepis.comosa, Evonymus europaeus, Acer campestre, Laserpitium gallicum, Carex flacca, Picea abies, Thuidium tamariscinum, Prenanthes purpurea, Linum suffruticosum.subsp. appressum, Rubus idaeus, Rosa arvensis, Dryopteris filix.mas, Oxalis acetosella, Lavandula angustifolia, Ulmus minor, Quercus robur, Circaea lutetiana, Castanea sativa, Melampyrum pratense, Anemone nemorosa, Vicia sepium, Lonicera xylosteum, Rubus fruticosus, Polygonatum verticillatum, Ajuga reptans, Galium odoratum, Bromus erectus, Festuca altissima, Ligustrum vulgare, Carex sylvatica, Cornus sanguinea, Bupleurum falcatum, Athyrium filix femina, Hedera helix, Quercus suber, Allium ursinum, Globularia cordifolia, Paris quadrifolia,Thymus serpyllum Eurhynchium striatum, Polygonatum multiflorum, Lotus delortii, Glechoma*

hederacea, Agrostis capillaris, Hordelymus europaeus, Sorbus aria, Carex pilulifera, Urtica dioica, Holcus mollis, Teucrium chamaedrys, Crataegus laevigata, Brachypodium pinnatum, Prunus avium, Dryopteris carthusiana, Hypnum cupressiforme, Conopodium majus, Plagiomnium affine, Geum urbanum, Pistacia terebinthus.

Je remarque qu'entre la sélection de variables, pour le C/N, avec et sans valeurs extrêmes, il y a 75 variables similaires, ce qui me paraît assez élevé. En fait, même si j'enlève les valeurs extrêmes, le nombre de variables sélectionnées restent inchangées mais les espèces choisies sont quand même assez différentes.

4.2) Sélection selon Robin GENUER et al.

La procédure de sélection de variable se base selon la méthode de Genuer et al. (2010). La première étape consiste à classer les variables par ordre décroissant d'importance, puis à retirer les variables de faible importance. Ensuite, avec les variables sélectionnées, on compare l'erreur OOB de modèles de forêts aléatoires emboités (à chaque pas on rajoute une variable dans le modèle) et on sélectionne le modèle réalisant la plus faible erreur. Enfin, la dernière étape, qui a pour but de trouver un petit sous-ensemble de variables, suffisant pour bien prédire la variable réponse. Cette étape consiste d'ajouter (à chaque pas), une variable dans le modèle que si elle fait suffisamment diminuer l'erreur OOB. Cette procédure est appliquée aux jeux de données, afin de bien prédire le pH et le C/N. En voici les résultats.

4.2.1) Résultats

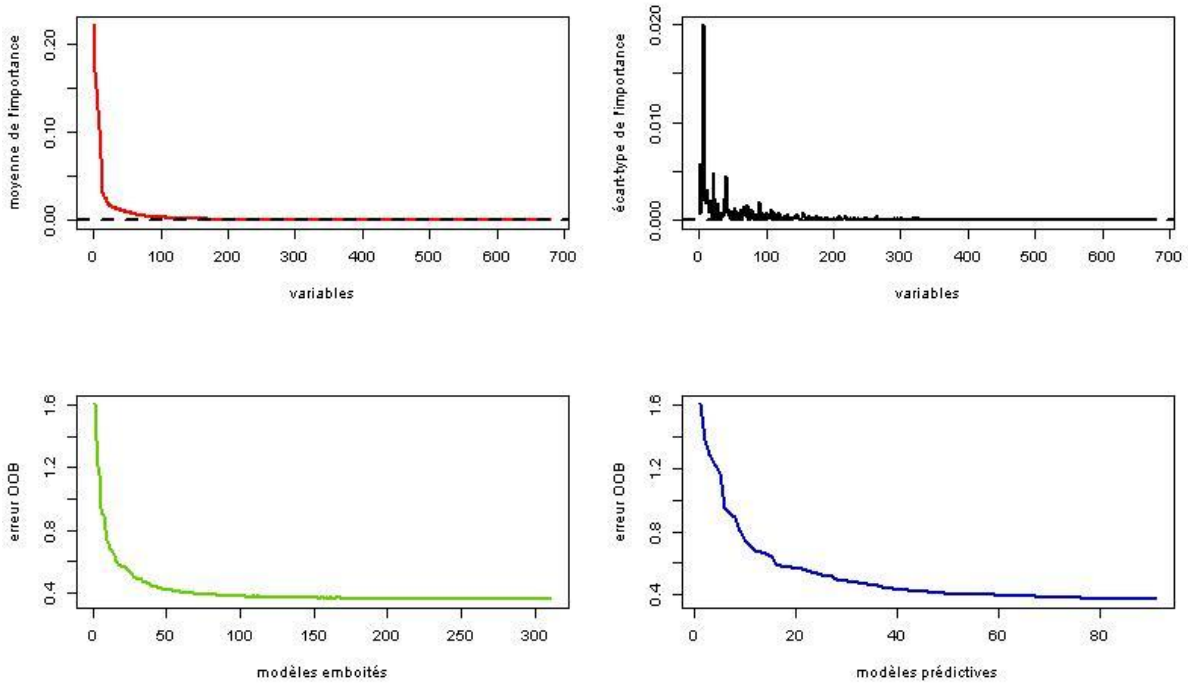


Figure 12 : sélection de variables selon Genuer et al. (2010) pour le pH

Je constate que lorsque le nombre de variables augmente, l'erreur OOB varie entre 1.61 et 0.35 pour les modèles emboités et conserve 311 variables. Mais cela reste encore trop élevé, c'est pour cela que, la dernière étape des modèles prédictives, vise à diminuer le nombre de variables dans le but de bien prédire le pH en ne conservant que 91 variables, l'erreur OOB est comprise entre 1.61 et 0.37. C'est le modèle qui sera retenu. De plus, dans la sélection générale, j'avais choisi le modèle à 50 variables, en comparant avec les espèces choisies lors de cette sélection, il y a 41 espèces qui sont identiques, ce qui est élevé.

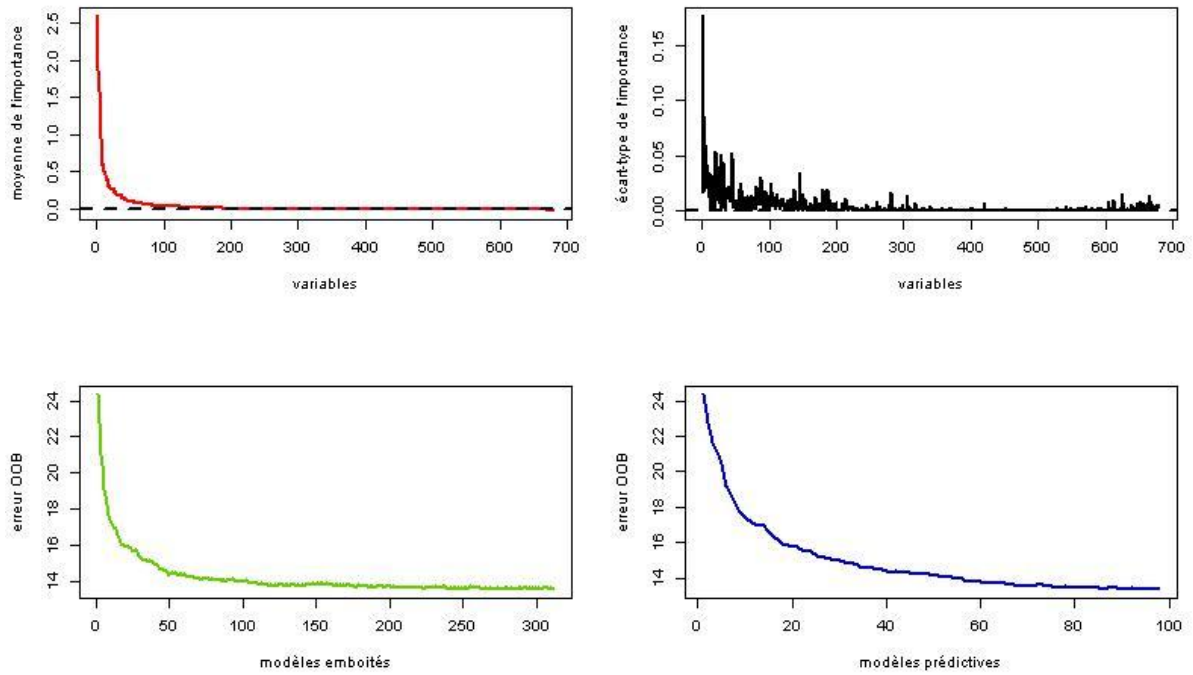


Figure 13 : sélection de variables selon Genuer et al. (2010) pour le C/N

Je constate que lorsque le nombre de variables augmente, l'erreur OOB varie entre 13.56 et 24.4 pour les modèles emboités et conserve 312 variables. Mais cela reste encore trop élevé, c'est pour cela que, la dernière étape des modèles prédictives, vise à diminuer le nombre de variables dans le but de bien prédire le C/N en ne conservant que 98 variables, l'erreur OOB est comprise entre 13.37 et 24.41. C'est le modèle qui sera retenu. De plus, dans la sélection générale, j'avais choisi le modèle à 100 variables, en comparant avec les espèces choisies lors de cette sélection, il y a 56 espèces qui sont identiques, ce qui représente un peu plus de la moitié, ce qui est normal car la méthode n'est pas la même.

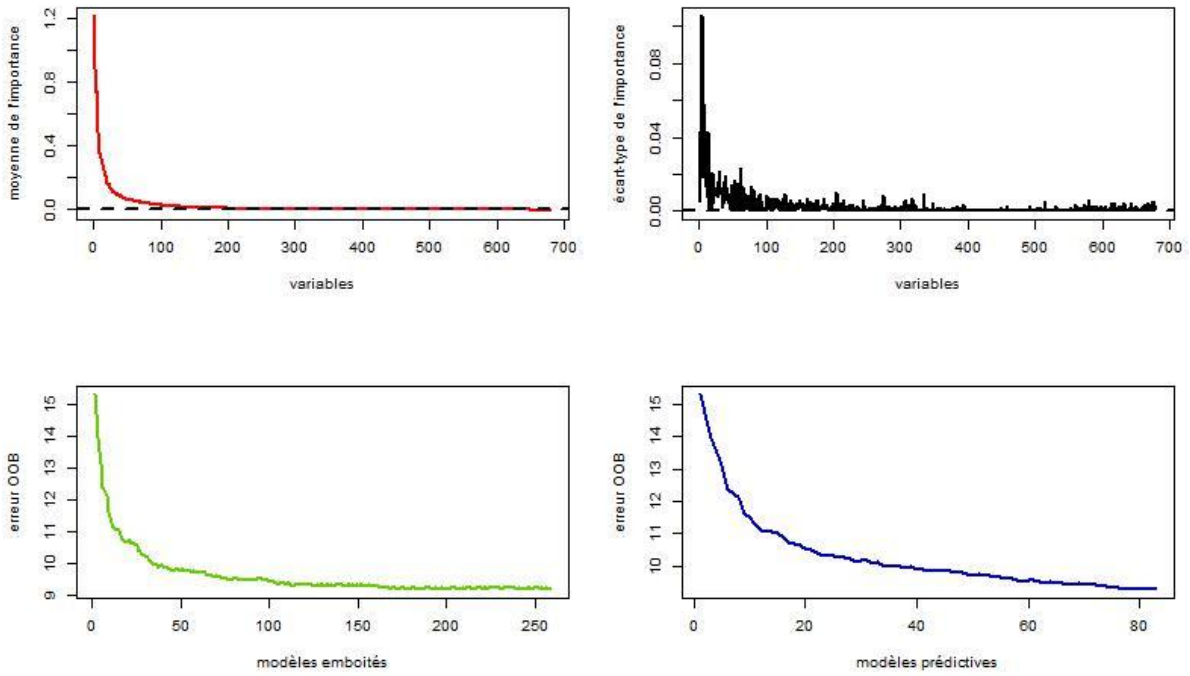


Figure 14 : sélection de variables selon Genuer et al. (2010) pour le C/N sans les valeurs extêmes.

Je constate que lorsque le nombre de variables augmente, l'erreur OOB varie entre 9.17 et 15.34 pour les modèles emboités et conserve 259 variables. Mais cela reste encore trop élevé, c'est pour cela que, la dernière étape des modèles prédictives, vise à diminuer le nombre de variables dans le but de bien prédire le C/N (sans les valeurs extrêmes) en ne conservant que 83 variables, l'erreur OOB est comprise entre 9.26 et 15.34. C'est le modèle qui sera retenu. Comme précédemment, pour la sélection générale, pour le C/N (sans les valeurs extrêmes), j'ai réussi à diminuer l'erreur OOB. De même, dans la sélection générale, j'avais choisi le modèle à 100 variables, en comparant avec les espèces choisies lors de cette sélection, il y a 51 espèces qui sont identiques, ce qui représente un peu plus de la moitié. Avec cette sélection, je remarque qu'entre la sélection de variables, pour le C/N, avec et sans valeurs extrêmes, il y a 47 variables similaires, ce qui est moins élevé que pour la sélection générale.

Cette méthode de sélection de variables est implémentée sous R dans le package VSURF (Robin Genuer, Jean-Michel Poggi and Christine Tuleau-Malot (2013). VSURF: Variable Selection Using Random Forests. R package version 0.5. http://CRAN.R-project.org/package=VSURF), en utilisant la fonction **VSURF()** de la manière suivante : **VSURF(X, Y, ntree, mtry)**, où **X** est la matrice des variables explicatives, **Y** la variable réponse, **ntree** le nombre d'arbre formant la forêt et **mtry**, le nombre de variables sélectionnées pour chaque nœud.

Enfin, avec cette sélection de variables, les variables (espèces végétales) les plus importantes sont les suivantes :

Pour le pH : *Pteridium aquilinum,Polytrichum formosum, Acer campestre, Viburnum lantana, Deschampsia flexuosa,Teucrium chamaedrys,Cornus sanguinea, Ligustrum vulgare, Lonicera periclymenum,Euphorbia cyparissias, Lonicera xylosteum, Galium mollugo, Helleborus foetidus, Fraxinus excelsior,Sorbus aria,Rubia peregrina, Sesleria caerulea,Ribes alpinum, Brachypodium pinnatum, Bromus erectus, Quercus robur, Arum maculatum, Athyrium filix femina, Clematis vitalba,Carpinus betulus,Dryopteris dilatata, Dryopteris carthusiana, Thymus vulgaris, Paris quadrifolia, Rubus fruticosus, Phillyrea angustifolia, Brachypodium sylvaticum, Calluna vulgaris,Galium aparine,Galium odoratum, Quercus ilex, Amelanchier ovalis, Quercus petraea, Vaccinium myrtillus, Atrichum undulatum, Cardamine pentaphyllos, Tilia platyphyllos, Crataegus laevigata, Viola reichenbachiana, Epipactis helleborine, Dryopteris filix mas, Erica arborea,Prunus avium,Carex digitata,Mercurialis perennis, Cytisus scoparius, Geranium robertianum, Neottia nidus avis, Eurhynchium striatum, Carex humilis, Loniceraalpigena, Castanea sativa, Hedera helix, Primula.elatior, Dactylis glomerata, Galium corrudifolium, Lathraea squamaria,Cardamine pratensis, Stellaria holostea, Anthoxanthum odoratum, Ilex aquifolium,Picea abies, Rosmarinus officinalis,Plagiomnium undulatum, Agrostis capillaris, Anemone nemorosa, Hypericum pulchrum, Poa nemoralis, Potentilla sterilis, Festuca heterophylla, Lysimachia vulgaris, Lathyrus linifolius subsp. montanus, Geum urbanum, Rubus saxatilis, Tamus communis, Populus tremula, Laserpitium gallicum, Ranunculus flammula, Frangula alnus, Daphne laureola, Rhytidiadelphus triquetrus, Listera.ovata, Orthilia secunda,Genista pilosa,Luzula sylvatica,Polystichum aculeatum.*

Pour le C/N : *Calluna vulgaris, Pteridium aquilinum, Dicranum scoparium, Molinia caerulea, Fraxinus excelsior, Deschampsia flexuosa, Pleurozium schreberi, Pinus pinaster, Leucobryum glaucum, Corylus avellana, Lonicera periclymenum, Polytrichum formosum, Scleropodium purum, Stachys sylvatica, Anthyllis montana, Lamium galeobdolon, Teucrium scorodonia, Dactylis glomerata, Arum maculatum, Euphorbia cyparissias, Abies alba, Cytisus scoparius, Carpinus.betulus, Acer pseudoplatanus, Luzula luzuloides, Quercus petraea, Crataegus monogyna, Sorbus aria, Prunus spinosa, Picea abies, Quercus robur, Quercus humilis, Linum suffruticosum subsp. Appressum, Galeopsis tetrahit, Hypnum cupressiforme, Ligustrum vulgare, Arbutus unedo, Plagiomnium undulatum, Brachypodium sylvaticum, Quercus ilex, Rubus idaeus, Polygonatum multiflorum, Evonymus europaeus, Dryopteris carthusiana, Melampyrum pratense, Convallaria majalis, Polygonatum verticillatum, Erica scoparia, Deschampsia cespitosa, Festuca altissima, Prenanthes purpurea, Dryopteris filix mas, Quercus rubra, Carexbrizoides, Cistus.salviifolius, Alnus glutinosa, Orthilia secunda, Urtica dioica, Geranium robertianum, Valeriana montana, Agrostis capillaris, Listera ovata, Rosa arvensis, Hylocomium brevirostre, Paris quadrifolia, Pseudarrhenatherum longifolium, Ulmus minor, Lavandula angustifolia, Hordelymus europaeus, Lonicera nigra, Valeriana tripteris,*

Hypericum perforatum, Salix caprea, Melica uniflora, Vaccinium vitis idaea, Moehringia muscosa, Sambucus racemosa, Carex umbrosa, Staehelina dubia, Quercus x calvescens, Rumex acetosa, Ribes rubrum, Conopodium majus, Blechnum spicant, Prunus padus, Hypericum androsaemum, Poa nemoralis, Malus sylvestris, Rhamnus alaternus, Luzula campestris, Genista anglica, Cephalanthera rubra, Pinus halepensis, Phalaris arundinacea, Viola sp. , Impatiens noli tangere, Onobrychis viciifolia, Rumex acetosella.

Pour le C/N (sans valeurs extrêmes) : *Deschampsia flexuosa, Dicranum scoparium, Fraxinus excelsior, Pteridium aquilinum, Molinia caerulea, Calluna vulgaris, Leucobryum glaucum, Polytrichum formosum, Euphorbia cyparissias, Pleurozium schreberi, Arum maculatum, Lamium galeobdolon, Vaccinium myrtillus, Scleropodium purum, Abies alba, Corylus avellana, Dactylis glomerata, Anthyllis montana, Lonicera periclymenum, Pinus pinaster, Quercus petraea, Carpinus betulus, Prunus spinosa, Plagiomnium undulatum, Thymus serpyllum, Luzula luzuloides, Galeopsis tetrahit, Glechoma hederacea, Crataegus monogyna, Anemone nemorosa, Picea abies, Dryopteris carthusiana, Prunus avium, Quercus robur, Frangula alnus, Eurhynchium striatum, Luzula sylvatica, Carex brizoides, Melampyrum pratense, Hypnum cupressiforme, Dryopteris filix mas, Hedera helix, Rubus idaeus, Ligustrum vulgare, Laserpitium gallicum, Viburnum opulus, Rosa arvensis, Circaea lutetiana, Rubia peregrina, Lavandula angustifolia, Urtica dioica, Quercus humilis, Clematis vitalba, Lonicera xylosteum, Erica scoparia, Galium odoratum, Phillyrea angustifolia, Ulmus minor, Lathyrus linifolius subsp. Montanus, Galium aparine, Eryngium campestre, Berberis vulgaris, Crataegus laevigata, Carex umbrosa, Dicranella heteromalla, Populus tremula, Sambucus racemosa, Festuca heterophylla, Poa nemoralis, Carex digitata, Mycelis muralis, Rubus caesius, Juniperus communis, Euphorbia characias, Rumex acetosa, Ononis minutissima, Galium saxatile, Malus sylvestris, Agrostis capillaris, Heracleum sphondylium, Arrhenatherum elatius, Tilia platyphyllos, Tortella tortuosa.*

Enfin, pour juger de la qualité des trois modèles j'ai fait tourné les forêts aléatoires pour constater quelle était le pourcentage expliquée par ces modèles, ceci est résumé dans le tableau suivant :

Tableau 4 : Résultats de forêts aléatoires pour les 3 modèles ci dessus

modèles	ntree	mtry	OOB	% var explained	OOB test
modele.genuer.CN	500	16	13.80092	49.12	13.68881
modele.genuer.CN (sans valeurs extrêmes)	500	14	8.828234	47.68	8.534824
modele.genuer.pH	500	15	0.3885236	80.17	0.3764868

Je conclue que, comme pour la sélection générale, la sélection de Genuer et al. (2010) permet d'avoir un pourcentage de variance expliquée élevé pour le pH, tandis que celui-ci est assez faible pour le C/N. De plus, dans la bibliographie, il y a une autre façon de sélectionner les variables grâce aux forêts aléatoires (Diaz Uriarte et al. 2006). Néanmoins, je n'ai pas trouvé d'articles affirmant que telles méthodes étaient meilleures qu'une autre. Cependant, je pense que l'introduction pas à pas de Genuer et al. (2010), me parait plus pertinente que la sélection générale, qui elle me paraît un peu trop global.

5) Discussion

La bioindication est une méthode indirecte de l'étude des sols. J'ai montré que grâce aux modèles des forêts aléatoires, on pouvait faire de la sélection de variables. L'avantage de cette méthode est qu'elle prend en compte les interactions et les corrélations entre les variables explicatives. De plus, elle est performante aussi bien pour des problèmes classiques (n>>p, n est le nombre d'observations et p le nombre de variables explicatives) que pour des problèmes de grande dimension (n<<p). Néanmoins, l'un des inconvénients de cette méthode, est sa difficulté d'interprétation. En effet, pour un arbre de régression, il suffit uniquement de lire au niveau des nœuds les variables les plus importantes et de suivre les branches pour avoir la prédiction en fonction des variables. Nous perdons tous ces informations dans le cadre des forêts aléatoires et il est difficile de répondre à la question, pourquoi ces variables ont-elles étaient sélectionnées et pas les autres ? C'est pour cela, qu'il ne suffira pas uniquement de dire que telle espèce est très présente dans le jeu de données (comme par exemple le *Pteridium aquilinium* 26% de présence, ce qui est élevé pour une espèce), pour dire que c'est pour cela qu'elle a été sélectionnée. Car, il y a par exemple, l'espèce *Teucrium chamaedrys*, qui a seulement 5% de présence, mais qui a été aussi sélectionnée. De plus, ce n'est pas possible de dire que telle espèce a été choisi car cette espèce pour dans un sol acide ou un sol basique. En effet, l'espèce *Acer campestre* pousse sur des sols basiques tandis que l'espèce *Deschampsia flexuosa* pousse sur des sols acides. Pourtant, toutes les deux espèces ont été sélectionnées comme étant importante pour le pH. La même réflexion pourrait être faite pour le C/N.

De plus, j'ai voulu savoir s'il y avait des espèces qui étaient à la fois importante pour le pH et le C/N, en traçant l'importance des variables du pH en fonction de l'importance des variables du C/N.

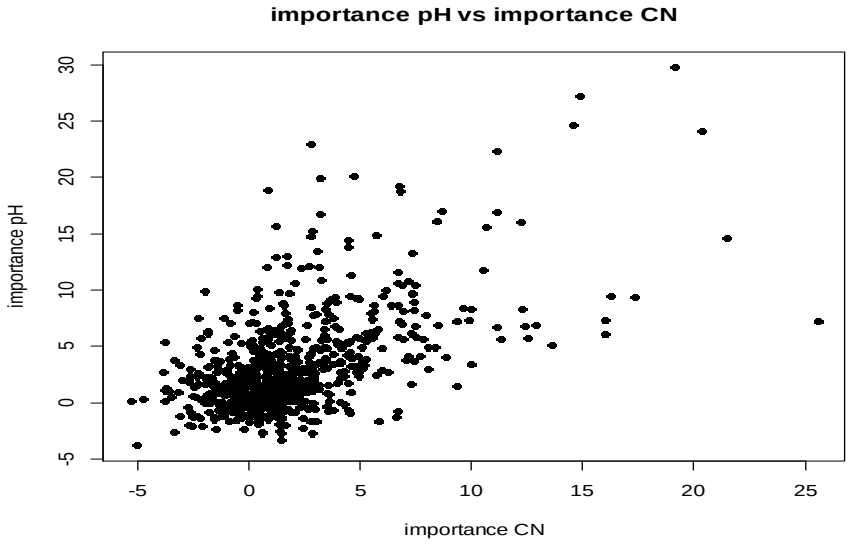


Figure 15 : l'importance des variables du pH en fonction de l'importance des variables du C/N.

Mais il n'y a que les espèces suivantes qui sont à la fois importantes au pH et au C/N : *Pteridium aquilinium*, *lonicera periclymenum*, *Fraxinus excelsior*, *Deschampsia flexuosa*, *polytrichum formosum*, *Euphorbia cyparissias*. Toutes les autres sont moins importantes.

En discutant, avec les spécialistes forestiers, ils pensent que pour les deux sélections de variables, cela représente beaucoup trop de variables, c'est pour cela que dans un rapport de Gégout et al. (2008), ils préconisent un découpage en zone biogéographique de la France. Voici le découpage en 5 zones :

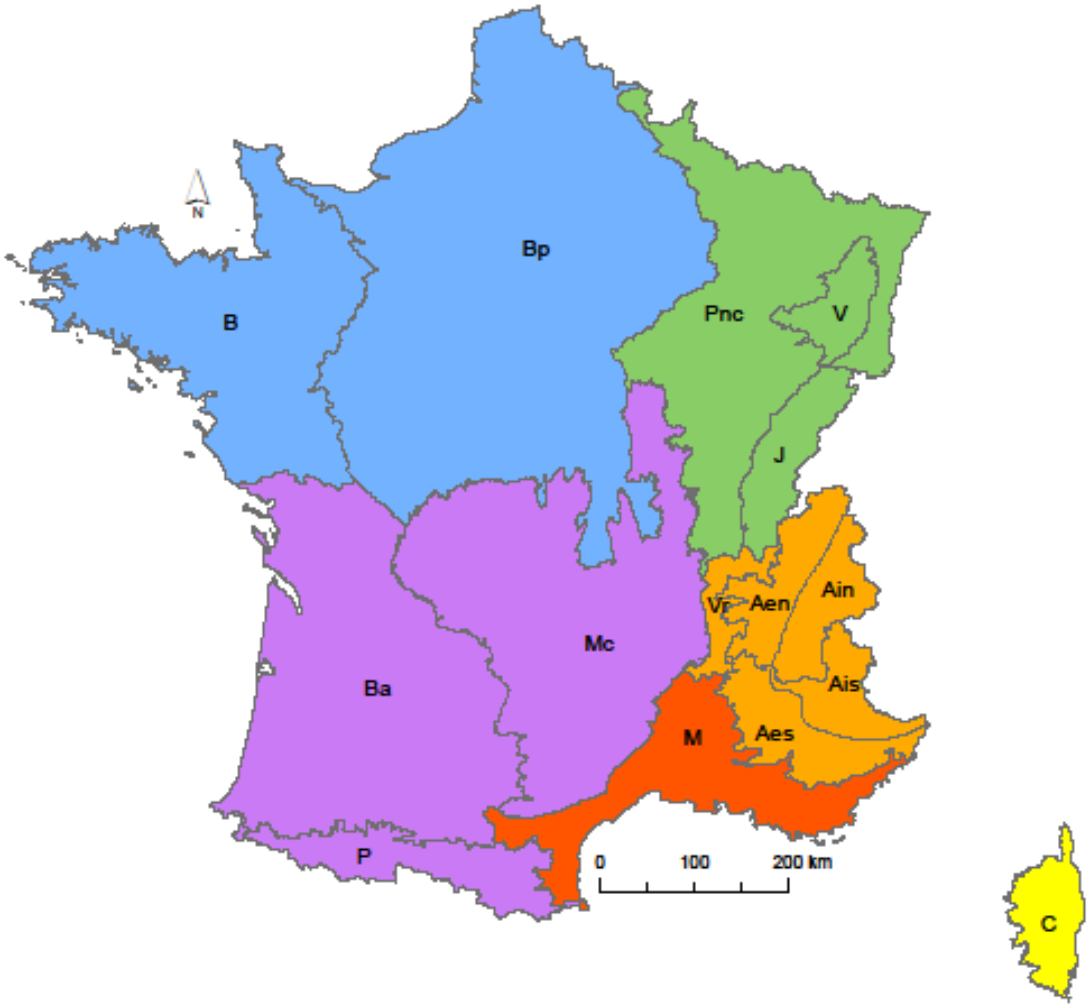


Figure 16 : découpage en 5 zones biogéographique de la France (la Corse faisant partie de la méditerranée (M))

Voici les résultats résumé dans le tableau suivant :

Zone 1 : B-Bp (Bretagne-Bassin parisien), Zone 2 : Pnc-V-J (Plaines et collines, Vosges, Jura) Zone 3 : Mc-P-Ba (Massif central, Pyrénées et Bassin aquitain), Zone 4 : Ain-Ais-Aes-Aen-Vr (Alpes) et Zone 5 : M-C (Méditerranée et Corse).

Tableau 5 : Résultats découpage en 5 zones

	min OOB (pour le pH)	min OOB (pour le C/N)	Nombre de variables selectionnées (pH)	Nombre de variables selectionnées (C/N)
Zone 1	0.3	16.29	36	46
Zone 2	0.36	12.5	40	32
Zone 3	0.35	14.26	13	12
Zone 4	0.43	9.26	44	38
Zone 5	0.34	9.52	67	56

La sélection de variables que j'ai utilisées est celle de Genuer et al. .

Dans ce tableau, on remarque que pour le pH, l'erreur OOB pour le pH est du même ordre de grandeur que pour la sélection au niveau national, mais l'avantage est qu'on diminue le nombre de variables sélectionnées. De même que pour la sélection pour le C/N, on réussit même à diminuer l'erreur OOB dans certaines zones.

Enfin, ce qui est intéressant de voir c'est d'étudier si lorsque je perturbe les variables à expliquer (pH et C/N), l'erreur OOB reste-t-elle inchangée ?les variables sélectionnées sont-elles les même ? Pour cela, j'ai ajouté, une loi normale centrée mais d'écart-type 0.05, 0.1, 0.2, 0.3, 0.4, 0.5 pour le pH, et de 0.5, 0.6, 0.7, 0.8, 0.9, 1 pour le C/N respectivement.

Les résultats sont résumés dans les tableaux suivants :

Tableau 6 : Résultats après perturbation du pH

	min OOB	Nombre de variables sélectionnées	Nombre de variables similaires au pH sans bruit
Simul 1	0.371	100	60
Simul 2	0.376	94	55
Simul 3	0.41	98	56
Simul 4	0.473	96	56
Simul 5	0.537	78	56
Simul 6	0.62	100	59

Je rappelle que pour le pH, l'erreur OOB minimum était de 0.35 et le nombre de variables sélectionnées étaient de 91. Comme j'ajoute une variable indépendante au pH, la variance de prédiction (erreur OOB) devrait être la somme entre l'erreur OOB du pH non perturbé et la variance de la loi normale (Luo et al. 2006). Or, dans mon cas, l'erreur est supérieure pour chaque simulation. D'après l'article (Luo et al. 2006), si l'erreur était égale à la somme entre l'erreur OOB du pH non perturbé et la variance de la loi normale, on aurait eu une

surparamétrisation, c'est-à-dire que le modèle ajusterait le bruit. Comme ce n'est pas le cas, le modèle n'est pas surparamétré et ceci peut être expliqué par un biais d'estimation lors de la sélection. De plus, on s'aperçoit aussi que le nombre de variables sélectionnées est du même ordre de grandeur, que pour le pH sans perturbation, mais qu'il y a un peu plus de la moitié qui sont similaires.

Tableau 7 : Résultats après perturbation du C/N

	min OOB	Nombre de variables sélectionnées	Nombre de variables similaires au C/N sans bruit
Simul 1	14.114	72	48
Simul 2	14.168	81	43
Simul 3	14.579	68	42
Simul 4	14.349	93	53
Simul 5	14.638	82	47
Simul 6	15.182	53	38

Je rappelle que pour le C/N, l'erreur OOB minimum était de 13.56 et le nombre de variables sélectionnées étaient de 98. De même que pour le pH, j'ajoute une variable indépendante au C/N, la variance de prédiction (erreur OOB) devrait être la somme entre l'erreur OOB du C/N non perturbé avec la variance de la loi normale. Or, dans mon cas, l'erreur est supérieure pour chaque simulation. La même explication, que pour le pH, peut être faite. Le modèle n'est pas surparamétré. De plus, on s'aperçoit aussi que le nombre de variables sélectionnées est moins élevé, que pour le C/N sans perturbation, mais qu'il y a un peu plus de la moitié qui sont similaires.

6) Conclusion

Durant ce stage, j'ai pu m'initier à l'apprentissage statistique et à la recherche bibliographique. Bien que l'apprentissage soit une matière jeune, ces méthodes sont des plus performantes, que ce soit dans le cadre de la régression ou la classification. La méthode que j'ai étudiée tout au long de mon stage, les forêts aléatoires, semble être des plus efficace. Dans mon cas, je les ai utilisées pour faire de la sélection de variables.

En effet, les forêts aléatoires, m'ont permis de sélectionner moins d'une centaine de variables pour bien prédire le pH avec une erreur quadratique moyenne faible et un pourcentage de variance expliquée élevé. De même, pour le C/N, malgré que l'erreur soit élevée et un faible pourcentage de variance expliquée, j'ai réussi à réduire le nombre de variables explicatives.

De plus, de l'aveu même des spécialistes forestiers, le C/N est une variable assez compliquée à expliquer. Cependant, cette méthode possède un défaut majeur, l'interprétation des résultats. Le temps de calcul pour les modèles à plusieurs centaines de variables était relativement long. Pour gagner du temps, par exemple dans la sélection de Genuer, j'ai dû uniquement faire de la sélection avec deux forêts alors que dans son article (Genuer et al. 2010), ils en utilisent cinquante. En outre, la sélection faite dans les cinq zones va aider les techniciens sur le terrain car il s'agira de relever que les espèces sélectionnées dans ces zones-là. Cela sera, d'une part, plus facile pour eux, car ils n'auront qu'à connaître ces espèces, et d'autre part, ils prendront moins de temps à faire leurs relevés floristiques.

D'un point de vue personnel, ce stage dans un autre laboratoire de recherche, m'a permis de rencontrer des chercheurs et doctorants formidables, mais aussi des stagiaires venant des quatre coins monde.

7) Bibliographie

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., (1984). *Classification And Regression Trees*. Chapman & Hall, New York.

Breiman, L., (1996). Bagging predictors. *Machine Learning*, **26(2)**, 123--140.

Breiman, L., (2001). Random Forests, *Machine Learning*, **45**, 5--32.

Cajander, A. K., (1926). The theory of forest types, *Acta Forestalia Fennica*, **29**, 1--108.

Díaz-Uriarte, R., Alvarez de Andrés, S., (2006). Gene Selection and classification of microarray data using random forest, *BMC Bioinformatics*, **7,** 3.

Gégout J.-C., Rameau J.-C., Renaux B., Jabiol B., Bar M., Marage D., (2008). Les habitats forestiers de la France tempérée; typologie et caractérisation phytoécologique. AgroParisTech-ENGREF, Nancy. 720 pages, 6 annexes. Document financé par l'Office National des Forêts et l'ADEME.

Genuer, R., Poggi, J.-M., Tuleau, C., (2008). Random Forests : some methodological insights, *Rapport de recherche 6729, Inria*.

Genuer, R., Poggi, J.-M., Tuleau, C., (2010a). Variable selection using Random Forests. *Pattern Recognition Letters*, **31(14)** ,2225--2236, 2010a.

Genuer, R., (2010). *Forêts aléatoires : aspects théoriques, sélection de variables et applications.* Thèse, Université Paris-Sud, Orsay.

Ghattas B., (1999). Importance des variables dans les méthodes CART, *Modulad*, **24**, 29--39.

Guillou A., (201-2012). Master 1 : cours de Modèles linéaires.

Hastie, T., Tibshirani, R., Friedman, J., (2008). *The Elements of Statistical Learning*, Second edition. Springer.

Liaw A., Wiener M., (2002). Classification and Regression by randomForest, *R News 2(3)*, 18--22.

Luo, X., Stefanski L., A., Boos, D., D., (2006). Tuning variable selection procedures by ading noise, *American Statistical Association,* **48**, 165--175.

Vapnik V., (1995). *The nature of statistical learning theory*. Springer.

Vayssières, M. P., Plant R. E. Allen-Diaz B. H., (2000). Classification trees: An alternative non-parametric approach for predicting species distributions, *Journal of Vegetation Science*, **11**, 679--694.

Ter Braak C.J.F., Looman C.W.N., (1986). Weighted averaging, logistic regression and the gaussian response model, *Vegetatio*, **78**, 57--72.

Wood, S. N., (2006). *Generalized Additive Models An Introduction with R*. Chapman & Hall, New York.