Modèles de prédiction de biomasse à partir de données de scan laser aéroporté (LiDAR)

Cheikh Ahmadou Bamba Diop

UFR Mathématique-Informatique

Pôle R&D de Nancy

Rapport de stage

Master 2 mention Mathématiques et Applications : Spécialité Statistique.

Modèles de prédiction de biomasse à partir de données de scan laser aéroporté (LiDAR)

Stage réalisé à l'Office National des Forêts, Pôle de Nancy, sous la direction de Jérôme BOCK et de Jean-Pierre RENAUD

Février-Août 2011

Présenté par DIOP Cheikh Ahmadou Bamba

Remerciements

Je tiens à remercier tous ce qui m'ont aidé de prêt ou de loin à la réalisation de ce travail.

Je remercie en particulier mes maîtres de stage Jérôme BOCK et Jean-Pierre RENAUD pour leur soutien et leur disponibilité jusqu'à la fin.

Je remercie aussi tous les membres de l'équipe qui ont tous été des tuteurs disponibles pour moi.

Je n'oublie pas mes collègues stagiaires avec qui l'ambiance a été très bonne.

Je remercie également Laurent Saint ANDRE du pôle INRA Nancy pour sa collaboration et ses conseils.

Enfin, je remercie mes amis de Nancy qui ont toujours été là pour moi.

RESUME

Ces dernières années, la technologie LiDAR a nettement progressé en foresterie. En Scandinavie par exemple, elle sert actuellement de manière opérationnelle à estimer la ressource forestière. Afin de juger de la précision de son utilisation dans un contexte forestier français, cette étude a été mise sur pied. Elle avait pour objectif d'évaluer la robustesse de l'estimation de 2 paramètres dendrométriques clés (i.e. la hauteur dominante (Ho) et la surface terrière (G)), à partir des nuages de points LiDAR.

Au total, des données de 3 forêts feuillues et 3 forêts résineuses de montage situées dans l'Est de la France, ont été utilisées. Pour les forêts feuillues, l'acquisition LiDAR a été effectuée en hiver (hors feuilles). Pour chacun des sites, 20 à 120 placettes de terrain ont servi à mettre en relation les mesures dendrométriques avec les différents «métriques» issus du LiDAR. Pour l'estimation de Ho, la robustesse de 3 modèles préexistants a été comparée, alors que pour G, de nouveaux modèles ont été établis. Deux procédures de sélection de variables ont été mise en œuvre pour la construction de ces modèles: l'une paramétrique utilisant la PLS, et l'autre, non-paramétrique utilisant le Random Forest.

Les résultats obtenus montrent une précision acceptable des modèles pour l'estimation de Ho, qui est du même ordre de grandeur que l'erreur attendue de mesure. Le modèle de Ho qui intègre un indice spatiale semble le plus robuste, indépendamment de la forêt étudiée (avec une erreur relative inférieur de 7%, sans recalibration des paramètres). Cependant l'estimation de G est plus problématique. Les modèles trouvés dans la littérature sont imprécis et très peu robustes pour le type de forêts que nous avons étudié. Certains de nos modèles peuvent être localement précis lors de la phase de calibration (<7% d'erreur pour un site), mais se révèlent rapidement très bruités voire aberrants (de l'ordre de 20% ou plus) lorsqu'ils sont appliqués à d'autres forêts. Ils nécessitent alors une recalibration. Les erreurs relatives que nous avons obtenu pour les modèles de G varient entre 6.37% et 17.8 % lors des calibrations.

Cette étude montre qu' après quelques adaptations simples, l'utilisation du LiDAR permet d'estimer Ho de façon opérationnelle. Par contre, il s'avère que l'estimation de G est encore trop imprécise. Afin d'améliorer cette situation nous proposons d'explorer l'utilisation d'autres types de « métriques » qui tiendraient compte de l'information spatiale contenue dans le nuage de points LiDAR. Des indices de textures, ou l'utilisation de voxels (cubes) pourraient s'avérer complémentaire aux « métriques » utilisés jusqu'à présent, qui sont basés uniquement sur la distribution en hauteur des nuages lidar.

MOTS CLES : LiDAR, inventaire forestier, Random Forest, PLS, validation de modèles.

Sommaire

1. Introduction :

En 2007, le Grenelle de l'environnement s'est donné comme objectif d'augmenter la part des énergies renouvelables dans notre consommation afin de diminuer nos émissions de CO_2 (http://www.fne.asso.fr/fr/themes/question.html?View=entry&EntryID=249). « Produire plus de bois en optimisant ressource et récolte » passe par une meilleure connaissance de la disponibilité de la ressource. En effet, une gestion durable de la forêt nécessite une précision importante dans l 'évaluation de l'état de la ressource (type de peuplement, quantité de matière, surface enjeux, localisation) pour ajuster au mieux les prélèvements et répondre aux attentes du public en matière de bois énergie.

Traditionnellement, le forestier procède par inventaires en plein ou inventaires statistiques par points de sondages pour évaluer la ressource. Mais cette méthode est longue, fastidieuse, en particulier pour les mesures de hauteur. D'autre part, elle est souvent entachée d'erreurs en raison d'un fort effet observateur, ou par des biais de mesure.

Ces inventaires peuvent être améliorés par l'utilisation d'outils de télédétection, qui laissaient entrevoir depuis longtemps, des perspectives intéressantes pour obtenir une connaissance exhaustive de la ressource à grande échelle. Cependant, cette échelle de description, souvent en 2 dimensions (*e.g.* images satellites), n'est pas assez fine pour le forestier qui a besoin d'une précision locale forte.

Depuis une quinzaine d'années, une nouvelle technologie de télédétection qui utilise le laser s'est développée : le LiDAR (Light Detection and Ranging). Elle apporte une troisième dimension qui renseigne sur la spatialisation de la végétation et sa structure sur l'ensemble du territoire. Elle consiste en l'envoi d'impulsions laser à haute fréquence (150 000 impulsions à la seconde) à partir d'un avion. Ces rayons laser sont en partie réfléchis par la végétation ou transmis jusqu'au sol. Les retours forment un nuage de points 3D caractéristiques des structures rencontrées (*e.g.* forêt, champs).

Ces données ont servi à deux études (Dez 2008 [1], Martins 2009 [2]) visant à mettre au point une méthode opérationnelle d'estimation de la hauteur dominante (Ho) en forêt. Elles se basent sur des méthodes de régression linéaire utilisant une sélection des estimateurs LiDAR par procédure STEPWISE. Une première analyse de sensibilité de ces modèles, envers des paramètres externes, tels que la densité du nuage LiDAR, la précision du positionnement GPS des placettes de calibration, ou la structure des peuplements a été effectuée [2] mais se limite à la forêt de Haye uniquement. Cependant, depuis 2007 de nombreuses autres forêts ont été survolées au LiDAR, ce qui offre l'opportunité d'évaluer la robustesse des modèles sous différents contextes forestiers, ainsi que pour des caractéristiques variées de vol LiDAR (*e.g.* densité de points).

Certes la hauteur n'est pas la seule variable importante pour caractériser les forêts. Le forestier a également besoin de connaître la surface terrière (G), le volume (V) en matière ligneuse ou la densité des tiges (N) de ses peuplements. Jusqu'à présent, aucun modèle ne lui fourni de telles données de façon opérationnelle dans des contextes de forêts feuillues.

Notre étude a comme objectif d'évaluer, sur la base des données LiDAR à notre disposition, la robustesse de différents modèles d'estimation de la hauteur dominante dans des contextes forestiers variés, puis d'en décrire la précision et les limites d'utilisation. De plus, pour l'estimation de la surface terrière G, notre objectif est d'établir un modèle robuste de prédiction de ce paramètre en sélectionnant parmi les meilleurs variables LiDAR. Comme un grand nombre de variables peut être généré à partir des nuages de points LiDAR, deux méthodes de sélections ont été comparés : la première, basée sur une approche non-paramétrique, utilise une méthode de forêts aléatoires [3] pour évaluer l'importance relative des variables LiDAR, alors que la seconde utilise la méthode des moindres carrés partiels (PLS). Une description de la précision des modèles ainsi obtenus est donnée, de même qu'une discussion sur les limites d'utilisation de ces derniers. La robustesse des modèles est évaluée par validation croisée, ainsi que par validation indépendante, en comparant les mesures prises sur une forêt indépendante aux prédictions d'un modèle calibré sur une autre forêt.

2. Contexte de l'étude :

Cette étude s'insère dans un projet ANR Foresee qui regroupe les principaux partenaires français intéressés par l'évaluation des ressources forestières. Ce projet vise à fournir des méthodes et des outils pour estimer et cartographier la ressource forestière sur pied, ainsi que ses conditions d'exploitation (*e.g*. zones non bûcheronnables en montagne, accessibilité par rapport à la desserte), à l'échelle de grands massifs forestiers ou de bassins d'approvisionnement.

L'objectif de notre étude est globalement d'établir des modèles robustes de prédiction de données dendrométriques (hauteur dominante (Ho) et surface terrière (G)) à partir d'estimateurs LiDAR. Des procédures de sélections de variables ont été comparées et la robustesse des modèles a été évaluée par validation croisée (pour un même site) ou par validations indépendantes (sur d'autres sites).

2.1. Fonctionnement de la technique du LIDAR multi-échos :

Le LiDAR aéroporté est un système actif de télédétection basé sur l'émission et la réception d'un faisceau laser . Il comprend [1] :

- ♦ un vecteur aérien : un avion dans notre cas ;
- ♦ un laser, transmettant les ondes lumineuses ;
- ♦ un récepteur de la lumière rétrodiffusée ;
- ♦ un système de géo- référencement : un GPS et une centrale inertielle (INS pour Internal Navigation System).

Le laser produit une impulsion sous forme de signal optique qui est ensuite réfléchi par une cible et retourné au récepteur. Connaissant la vitesse du signal qui est la même que celle de la lumière, le temps de parcours est converti en distance entre l'avion et la cible. En fonction de la position de l'avion, donnée par le GPS, et de la direction des impulsions, connue grâce à l'INS, les mesures de distances peuvent être converties en coordonnées cartographiques et fournir la position de la cible. Chaque impulsion reçue est donc associée à trois coordonnées dans l'espace, constituées de X, Y et Z.

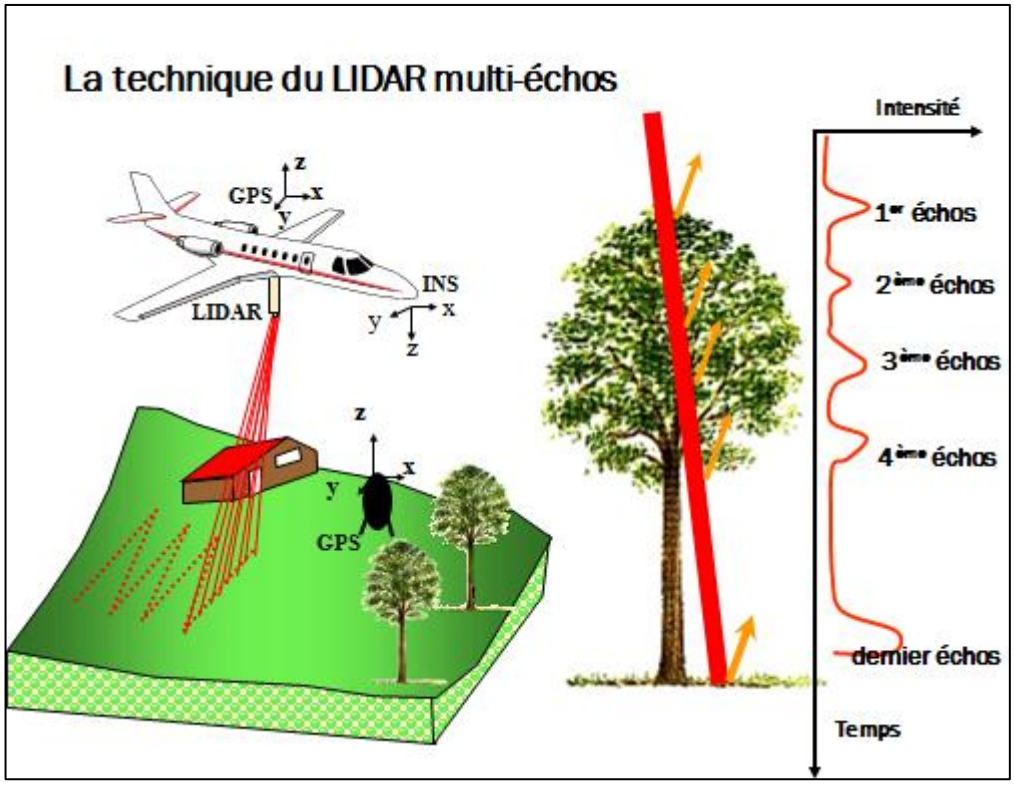

Figure 1:la technique lidar multi-échos (Bock et al. 2011)

Cette technologie LiDAR est capable d'enregistrer plusieurs retours, appelés échos, pour chaque impulsion émise. En fonction du nombre d'échos souhaités, ne sont gardés, en plus du premier et du dernier échos, qu'un certains nombre d'échos dits intermédiaires. Des algorithmes permettent de détecter les points renvoyés par le sol. Ainsi il est possible de classer les points enregistrés en deux groupes : les points sol et les points végétation.

Les données se présentent sous la forme d'un nuage de points géoréférencés en trois dimensions(longitude, latitude, altitude). La quantité d'énergie du signal reçu ou intensité, peut également être enregistrée. Elle renseigne sur la nature du matériau revoyant l'impulsion.

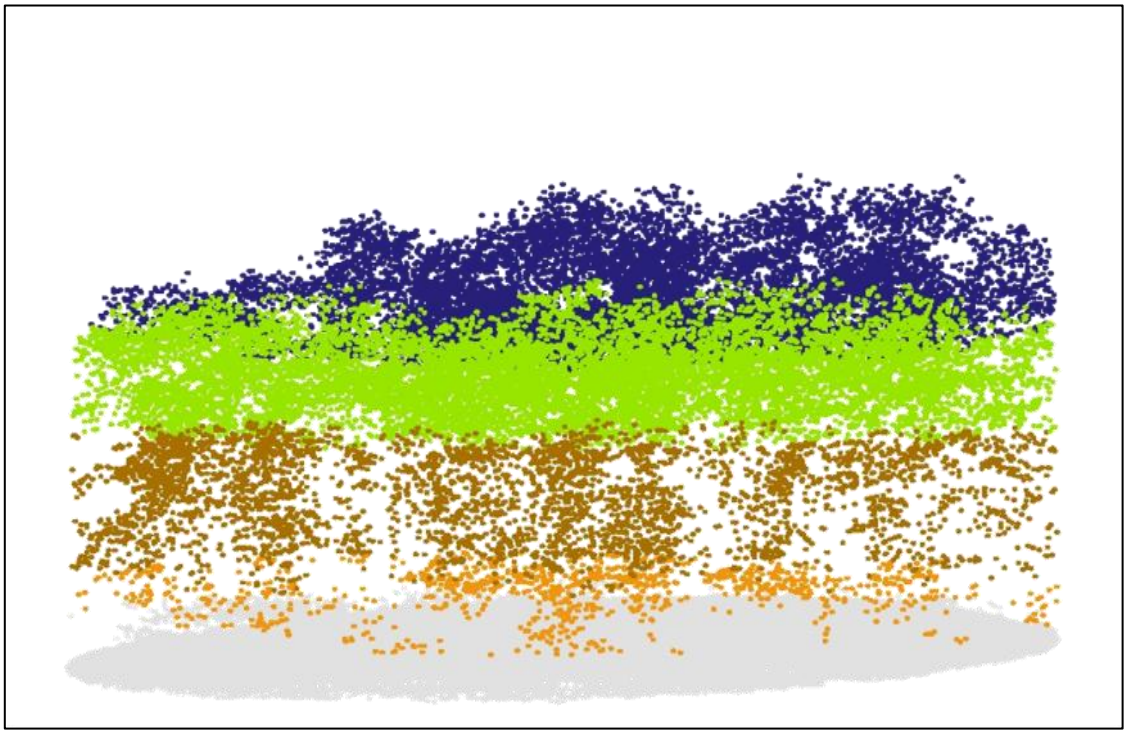

Figure 2:Exemple de nuage LiDAR sur une placette colorié en fonction des niveaux de hauteur(Bock et al. 2009)

2.2. Intérêt pour la gestion forestière

Le système LiDAR permet l'obtention rapide d'information, avec une numérisation qui se fait en même temps que l'acquisition. En outre, il fournit une information très dense avec un rendement surfacique élevé d'environ 40 km²/heure. En plus, il est même possible de survoler des zones inaccessibles (montagnes, par exemple) de jour comme de nuit.

Toutefois le volume important des données, certains problèmes d'interprétation, ainsi que le coût important de l'opération constituent les principales difficultés associées à l'utilisation du LiDAR.

En foresterie, le LiDAR présente un intérêt pour le gestionnaire, qui doit par exemple, procéder à certaines opérations sylvicoles sur la base de la hauteur des jeunes peuplements, ou procéder à des récoltes en fonction du capital disponible. Il lui est donc nécessaire de connaître avec précision la ressource dont il dispose, sa spatialisation, ainsi que les conditions liées à sa mobilisation. Un autre type d'intérêt du LiDAR concerne son utilisation pour évaluer l'adéquation entre la ressource disponible d'un territoire et l'implantation d'usines de transformation. Il est important de bien connaître la ressource forestière présente sur un bassin d'approvisionnement avant d'implanter des usines ayant un fort besoin en biomasse forestière.

2.3. Etat de l'art sur l'utilisation du LIDAR

L'utilisation du LiDAR en foresterie occupe actuellement plusieurs chercheurs en raison des espoirs qu'il suscite en termes d'efficacité pour la caractérisation de la ressource. Il doit son essor aux récents progrès réalisés dans les domaines des lasers, du GPS et de la navigation inertielle. Il

permet l'acquisition de données à haute résolution, conduisant à la réalisation de modèles numériques de terrain (MNT) d'une précision inférieure au mètre, ou de modèles numériques de surface (MNS) représentant l'altitude du « sursol ».

Depuis le début des années 1990, les scanners laser aéroportés LiDAR ont démontré une grande précision dans la mesure de la hauteur du couvert forestier ([4], [5]). Une des étapes de traitement consiste à séparer par filtrage les échos issus d'une réflexion avec un élément du couvert de ceux issus du sol. Un modèle numérique de canopée (MNC) est alors calculé par différence entre le MNS et le MNT donnant ainsi la hauteur, et non plus l'altitude, du couvert végétal. De plus, [5]ont démontré que la distribution verticale des points LiDAR est intimement reliée à celle du matériel végétal, donnant un sens biophysique aux échos LiDAR reçus. La distribution verticale des différents échos sur une même placette, représentée sous forme de quantiles, ou de composantes canoniques, a été ainsi utilisée pour estimer diverses variables forestières ([6] , [7], [8], [9]). Les précisions obtenues à cette échelle sont de l'ordre de 3 à 12 % pour la hauteur moyenne, 8 à 19 % pour le volume à l'hectare et 6 à 30% pour le diamètre moyen. Mais pour le moment, les équations utilisées sont partiellement dépendantes des paramètres d'acquisition des données LiDAR et de la structure des peuplements, ce qui réduit leur généricité.

En Scandinavie, l'utilisation du LiDAR pour des applications forestières est opérationnelle depuis 2002 [8]. Son utilisation pour estimer la hauteur des peuplements, leurs surfaces terrières, ou leur volume sur pied y est arrivée à maturité. Cette technologie émerge également dans d'autres pays, tels l'Autriche [10], le Canada [11], l'Allemagne [12], la Russie [13]ou les Etats-Unis [14]. Cependant, bien que la précision des valeurs estimées obtenus par LiDAR en forêt boréale soit souvent équivalente ou supérieure à celles obtenues de façon conventionnelle([4] et [13]) l'adaptation de ces méthodologies à d'autres types de peuplements, tels que ceux présents en France (feuillus à structures forestières complexes ou hétérogènes, taillis sous futaies) nécessite encore un important travail de recherche-développement. Pour généraliser l'utilisation des données LiDAR il faut donc envisager le développement d'estimateurs statistiques plus génériques permettant un transfert facilité à différentes conditions d'acquisition ou contextes forestiers.

Pour estimer des valeurs dendrométriques tels que Ho, G, ou le volume à l'aide du LiDAR, différentes variables numériques (ou métriques) calculées à partir des points LiDAR ont été utilisées. Les hauteurs moyennes ou médianes des points LiDAR, les percentiles de hauteur des premiers ou des derniers échos sont des exemples de métriques qui résument l'information contenue dans le nuage de points d'une placette en une seule dimension, (i.e. distribution en hauteur). D'autres métriques exploitent l'information « spatiale », contenue dans plus d'une dimension. Ainsi, l'identification de maxima locaux, ou de sommets, la quantification d'indices de texture (Haralick [15]), de distances moyennes entre maxima, ou l'utilisation de voxels [16]sont des exemples de métriques qui peuvent être extraits des nuages LiDAR. Dans cette étude nous nous sommes limité aux métriques qui résument l'information contenue dans le nuage de points en information de distribution en hauteur et aux métriques développés par Martins [2], utilisant une forme de maxima locaux.

2.3.1. Définitions des variables dendrométriques:

Dans cette étude, l'unité d'échantillonnage est représenté par une placette circulaire de rayon r prise dans une forêt donnée. Soit n (en are) la surface d'une placette de rayon r mètres (1 are=100 m²). On a alors $n=\pi.r^2/100$ en are.

La hauteur dominante (Ho) est définie en France comme étant, pour une futaie régulière, la moyenne des hauteurs des 100 plus gros arbres à l'hectare. Comme l'unité d'échantillonnage est généralement inférieur à l'hectare, ceci conduit à un biais d'estimation de Ho, qui est lié à des effets de sélection et d'auto-corrélation spatiale [17]. D'un point de vue pratique, Pardé et Bouchon [18] ont proposé une « règle du pouce » pour corriger ce biais. Elle consiste, sur des placettes de N ares ($N<100$), a faire la moyenne des hauteurs des N-1 plus gros arbres de chaque placette pour obtenir Ho. Ceci corrige en pratique le biais d'estimation.

La surface terrière (G) concerne tous les arbres recensables d'une placette (i.e. dont le diamètre est supérieur ou égal à un seuil de précomptage de 7.5 cm ou 17.5 cm selon les utilisateurs ; 7.5 cm dans notre cas). Pour chaque arbre A, le diamètre du tronc est mesuré à hauteur de poitrine (1.3 m). La surface correspondante est calculée et rapportée à l' hectare après pondération par la surface de la placette échantillonnée. Ce résultat, exprimé en m²/ha, défini la surface terrière de l'arbre A. La somme des surfaces de l'ensemble des arbres de la placette constitue la surface terrière (en m²/ha) de l'unité d'échantillonnage.

2.3.2. Calcul de métriques LIDAR.

Nos modèles utilisent des métriques LiDAR qui sont des valeurs numériques résumant la distribution des points LiDAR dans l'espace. Ces métriques sont établis en se basant entre autre sur la densité des points LiDAR, leur hauteur par rapport au sol, leur nature (premier ou dernier écho), ou leur nombre (total et/ou par tranche horizontale)…Il constituent les variables explicatives que l'on peut classer en trois grandes catégories :(voir annexe pour la définition de tous les métriques).

⇒ Les indicateurs de distribution de hauteur : moyenne, percentile, coefficient de variation, écart-type, Kurtosis ... ;

⇒ Les indicateurs de densité de points : par tranche de hauteur fixe (entre 2-6 m, 6-12 m, 12-24 m , ...), ou relative par dixième de la hauteur maximale ou de la hauteur du 95ième percentile (Hp95)), ou cumulée, en sommant la densité des différentes tranches jusqu'à une hauteur donnée ;

⇒ Des indicateurs spatiaux : qui se base sur des maxima locaux. Par exemple l'indice Hmv5, défini par Martins [2], représente la moyenne des hauteurs des 5 points les plus hauts du nuage LiDAR en excluant autour de chaque maxima une zone similaire à la couronne des arbres.

Depuis une dizaine d'années, différents métriques LiDAR ont été utilisés dans les modèles d'estimation de la biomasse forestière. Par exemple, Naesset [8] arrive de cette façon à estimer la biomasse de 3 forêts scandinaves d'âges variés caractérisées par différents types de sols. Dans notre étude, nous avons retenu un de ses modèles utilisé pour des forêts matures sur sols fertiles, considérant que cette situation était plus près de notre objet d'étude.

Pionnier dans la modélisation de paramètres forestiers en utilisant le LiDAR, Naesset (Cité 203 d'après Scopus) utilise un modèle de forme multiplicative :

$$Y = \beta_0 {h_{0f}}^{\beta_1} {h_{10f}}^{\beta_2} \dots {h_{90f}}^{\beta_{10}} {h_{0l}}^{\beta_{11}} {h_{10l}}^{\beta_{12}} \dots {h_{90f}}^{\beta_{20}} \times {h_{meanf}}^{\beta_{21}} {h_{meanl}}^{\beta_{22}} {h_{cvf}}^{\beta_{23}} {h_{cvl}}^{\beta_{24}} {d_{0f}}^{\beta_{25}} {d_{1f}}^{\beta_{26}} \dots {d_{9f}}^{\beta_{34}} {d_{0l}}^{\beta_{35}} {d_{1l}}^{\beta_{36}} \dots {d_{9l}}^{\beta_{44}}$$

Où Y est la variable dendrométrique à prédire (Hdom, G, N, V ...) voir l' annexe pour la signification des variables du modèle.

Cette formulation a servi de base à de nombreuses autres études publiées dans les revues scientifiques forestières. Elle nécessite une transformation logarithmique pour obtenir un modèle linéaire qui estime log(Y). Ceci introduit, par conséquent, un biais dans l'estimation finale de la variable initiale Y [19]. Le modèle de départ prend ainsi la forme suivante :

$$\begin{aligned} lnY = {} & \beta_{0+}\beta_1 lnh_{0f} + \beta_2 lnh_{10f} + \dots + \beta_{10} lnh_{90f} + \beta_{11} lnh_{10l} + \dots + \beta_{20} lnh_{90l} \\ & + \beta_{21} lnh_{meanf} + \beta_{22} lnh_{meanl} + \beta_{23} lnh_{cvf} + \beta_{24} lnh_{cvl} + \beta_{25} lnd_{0f} \\ & + \beta_{26} lnd_{1f} + \dots + \beta_{34} lnd_{9f} + \beta_{35} lnd_{0l} + \beta_{36} lnd_{1l} \dots + \beta_{44} lnd_{9l} \end{aligned}$$

Afin de réduire le nombre de paramètres du modèle, Naesset procède à une sélection de type stepwise et ne conserve dans le modèle final que la meilleure combinaison de métriques LiDAR triée parmi un groupe de variables souvent très corrélées.

Il travaille sur des forêts résineuses scandinaves, ce qui rend ses modèles probablement peu adaptés aux contextes français de peuplements feuillus mélangés. De plus, la détransformation de log(Y), pour revenir aux unités initiaux du Y risque d'introduire de l'instabilité dans les résultats. Dans des contextes plus proches de ceux que nous étudions, Heurich [12] a produit des modèles permettant de prédire différents paramètres dendrométriques pour des hêtraies mélangées de Bavière. Ses modèles (additifs) ont été retenus en raison de la similitude des caractéristiques

forestières bavaroises, en plus de ne pas nécessiter de détransformation. Le modèle de Heurich prend la forme suivante

$$\boldsymbol{Y = \beta_{0+}\beta_1 h_{0f} + \beta_2 h_{10f} + \cdots + \beta_{10} h_{90f} + \beta_{11} h_{10l} + \cdots + \beta_{20} h_{90l} + \beta_{21} h_{meanf} + \beta_{22} h_{meanl} + \beta_{23} h_{cvf} + \beta_{24} h_{cvl} + \beta_{25} d_{0f} + \beta_{26} d_{1f} + \cdots + \beta_{34} d_{9f} + \beta_{35} lnd_{0l} + \beta_{36} d_{1l} \ldots \beta_{44} d_{9l} + \beta_{45} h_{0vf} + \beta_{46} h_{10vf} + \cdots + \beta_{54} h_{90vf} + \beta_{55} h_{10vl} + \cdots + \beta_{69} h_{90vl} + \beta_{70} h_{meanvf} + \beta_{71} h_{meanvl} + \beta_{72} h_{cvvf} + \beta_{73} h_{cvl} + \beta_{74} h_{med\%vf} + \beta_{75} h_{med\%vl} + \beta_{76} h_{dvmvf} + \beta_{77} h_{dvmvl} + \beta_{78} PR_{totvf} + \beta_{79} PR_{ulvf} + \beta_{80} PR_{ilvf} + \beta_{81} PR_{llvf} + \beta_{83} PR_{totvl} + \beta_{84} PR_{ulvl} + \beta_{85} PR_{ilvl} + \beta_{86} PR_{llvl}}$$

Tout comme Naesset, Heurich réduit la dimension de ses modèles en sélectionnant les métriques LiDAR les plus adaptés à l'aide d'une procédure stepwise. Il propose des modèles pour l'estimation de G et Ho basés uniquement sur des métriques résumant le nuage de points LiDAR en une seule dimension (i.e. sa distribution en hauteur). Très peu d'études tentent de faire ressortir l'information contenue dans la dimension spatiale (X,Y) des placettes échantillonnées. Pourtant, en analyses d'images, des indices de textures sont disponibles depuis de nombreuses années (Harallick [15]). Popescous [16] est un des seuls, à notre connaissance, à avoir tenté d'extraire ce genre d'information ; mais à l'échelle d'arbres individuels, Martins [2] en voulant imité l'indicateur de Ho a également construit un métrique qui extrait les maxima locaux sur les placettes échantillonnées. Elle a constaté que ce paramètre lui permettant d'atteindre une plus grande robustesse dans l'estimation de Ho pour une forêt où le mélange d'espèces (chêne/hêtre) était présent [20]. Le modèle de Martin a donc été évalué dans notre étude et ses indicateurs de maxima locaux ont également été retenus.

Enfin, certains auteurs tentent de prédire des paramètres dendrométriques tel que le volume de bois sur pieds, à partir de l'identification des sommets des arbres [21]. Cette approche, bien que séduisante, car on imagine facilement compter les arbres dont on obtient la hauteur, reste encore expérimentale. De plus, en forêt feuillus, elle génère des biais importants car une large partie des arbres ne sont pas détectés correctement (*e.g.* les petits, ou les arbres jumelés possèdent un taux de détection très faible). Comme nous cherchions une méthode robuste, les métriques LiDAR associés à la détection d'arbres individuels n'ont pas été retenus pour cette étude.

3. Matériels et méthodes

3.1. Données :

Les données à notre disposition sont de deux natures : les variables dendrométriques mesurées sur le terrain aussi précisément que possible, qui constituent les variables à prédire; et les différents métriques qui sont calculés à partir du nuage de points LiDAR, qui constituent les variables explicatives. Par simplification, nous considérons que le géo-référencement des ces 2 sources de variables est sans erreur, c'est-à-dire qu'elles se superposent parfaitement sur le plan spatial. Ceci n'est pas vraiment le cas car sur le terrain l'erreur de positionnement du GPS est de l'ordre de 2 à 5 m dans notre contexte forestier . Cette erreur a été estimée par rapport à des repères fixes ou à des cibles ([2], Bock com. pers.). Nous considérons que, pour les placettes étudiées (d'une superficie moyenne de 600 m2), cette erreur est négligeable comme il a été déjà été démontré par Martins [2] pour Ho.

3.1.1. Le domaine d'étude

Pour cette étude nous disposons de données LiDAR sur 4 sites différents : les forêts domaniales de Haye (54) et de Languimberg (57), l'Observatoire Pérenne de l'Environnement (OPE) à Montiers (55) ainsi que des forêts de montagne : forêts de Chamonix (74), du Chablais (74) et de Vaujany (38). Pour chaque site, des placettes circulaires dont le rayon varie d'un site à l'autre entre 13.82 et 20 m ont servi à calibrer et/ou valider les modèles estimant les paramètres dendrométriques à partir des données LiDAR. Chaque placette, constitue une unité d'observation, à l'intérieure de laquelle Ho et G sont mesurés. Les informations concernant la calibration des modèles sont résumées au Tableau 1.

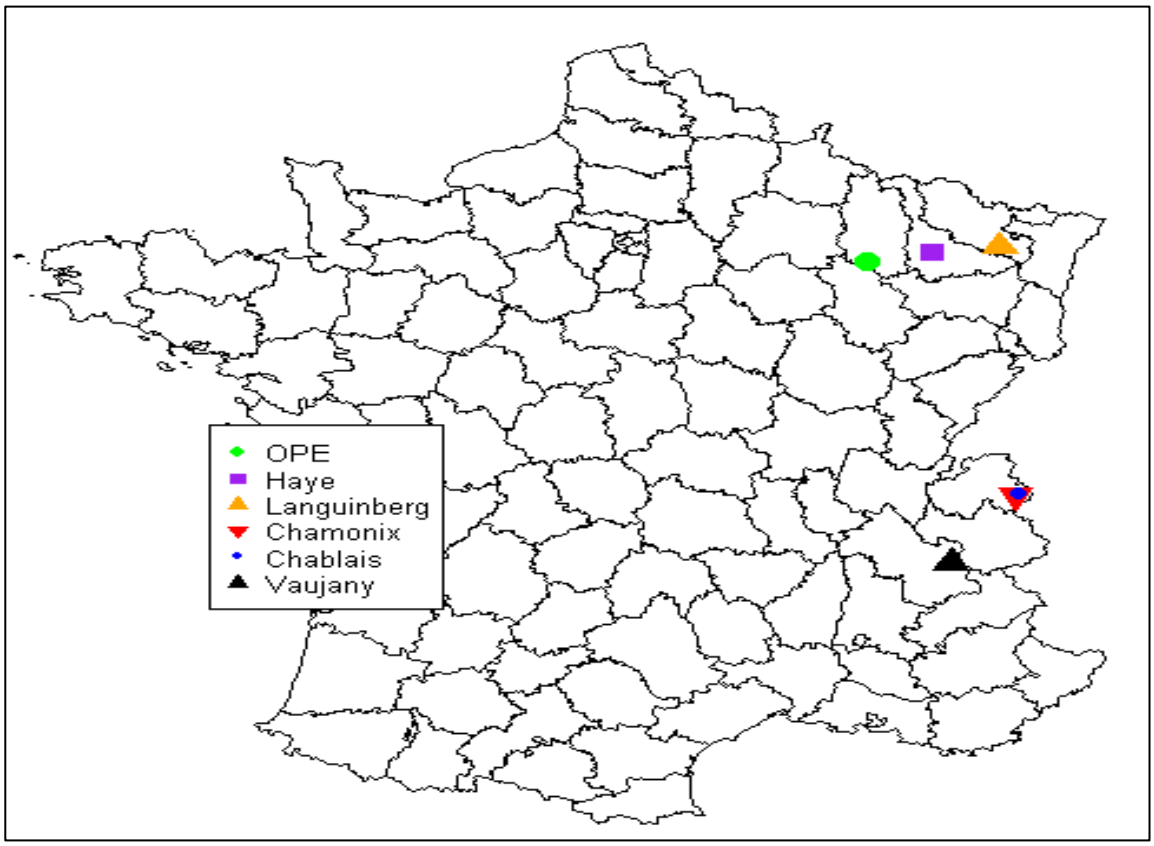

Figure 3:Localisation des sites d'étude (J.P. Renaud ONF 2011)

Site/ Information	Haye	Montiers (OPE)	Languimberg	Montagne : Chamonix Chablais, Vaujany
Nombre de placettes pour G	25	32	46	64
Nombre de placettes pour Ho	120	27	51	64
Densité (nombres de points/m²)	5pts/m²	15pts/m²	0.5 pts/m²	5pts/m²
Etat de végétation/ Type	Hors Feuille/ feuillus	Hors Feuille/feuillus	En feuille / feuillus	En feuille / résineux
Rayon des placettes	13 .82 m	15 m	18 ou 20 m	13 .82 m

Tableau 1:résumé des informations sur les données placettes

3.1.2. Données mesurées sur le terrain

♦ **Données mesurées :**

La localisation du centre de chaque placettes a été effectuée grâce à l'acquisition de plus de 80 points GPS obtenus à l'aide d'un Trimble GéoXT. Une correction différentielle a été réalisée pour chaque placette à l'aide de l'antenne fixe la plus proche du site de mesure. Dans certaines configurations (env. 20% des cas) les caractéristiques du milieu (*e.g.* trous, arbres tombés, intersections de chemins) ou l'installation de cibles visibles au LiDAR ont permis de localiser plus précisément les placettes (env. 1-2 m près, voire moins de 1 m pour les cibles). L'erreur associée au géo-référencement est supposé négligeable dans cette étude.

Pour chaque placette, la circonférence de tous les arbres recensables vivants (>7.5cm de diamètre) a été mesurée au ruban gradué (ou au compas forestier).

Sur les placettes de plaine (OPE, FD Haye, Languimberg) les n-1 plus gros arbres ont été mesurés en hauteur (où « n » est équivalent à la surface de la placette en are). Pour limiter l'effet observateur, un seul mesureur par site a réalisé les relevés de hauteur. De plus, afin de limiter l'effet de parallaxe, deux mesures à 180° de la hauteur du bourgeon terminal ont également été effectuées au vertex pour chaque arbre retenu([1], Dupouey com. Pers). De plus, sur l'OPE et Languimberg, la séquence de mesure était répétée jusqu'à avoir une différence inférieure à 1m entre les 2 mesures. La hauteur attribuée à l'arbre représente la moyenne arithmétique de ces 2 mesures.

Pour les placettes de montagne, tous les arbres ont été mesurés en hauteur à raison d'une seule mesure par arbre. L'observateur placé en amont de l'arbre (pratiquement au même niveau que l'apex), identifie avec certitude le bourgeon terminal sur les résineux et obtient ainsi une précision acceptable pour la mesure de hauteur (cette dernière n'a toutefois pas été quantifiée, CEMAGREF com. pers).

♦ **Données calculées par placette :**

Ho : La hauteur dominante (en m) sur les placettes de « n » ares représente la moyenne arithmétique des hauteurs des n-1 arbres les plus gros de la placette. Cette façon de calculer Ho réduit le biais associé à la différence de taille entre la surface échantillonnée et celle considérée dans la définition de Ho (qui est l'hectare)

G : La surface terrière à l'hectare représente la somme des surfaces de toutes les sections de troncs des arbres recensables mesurées à 1.3m. Cette surface est ramenée à l'hectare conventionnellement après pondération par la surface de la placette.

3.1.3. Calcul des métriques à partir des points lidar

De manière globale, l'objectif des métriques est de renseigner sur la distribution dans l'espace des points LiDAR. Il s'agit d'étudier leur densité, leur hauteur par rapport au sol, leur nature (premiers ou derniers échos), leur nombre (total et par tranche). Nous avons utilisé une terminologie où un « f » à la fin d'un nom de métrique signifie qu'elle est calculée uniquement à partir des premiers échos (first pulses) ; dans le cas des derniers échos (last pulses) le symbole utilisé est un « l ». Cela est valable pour tous les métriques calculées.

$hp_{0f}, hp_{10f}, hp_{20f}, \dots, hp_{95f} hp_{99f}$ représentent respectivement la **hauteur du percentile** 0%, 10%, 20%, …. 95%,99 des premiers échos.

$d_{0f}, d_{1f}, \dots d_{9f}$: représentent les densités de points par tranche horizontale. On divise de façon horizontale, l'espace compris entre une hauteur supérieure à 1m [8] ou 2m [12] et la hauteur du percentile $hp_{95\%}$ en 10 parties égales (de 0 à 9). Ainsi, on obtient la densité d_x en faisant le rapport entre le nombre de points sur cette tranche x par le nombre total de points dans l'ensemble des 10 tranches.

♦ **Exemple1 : le calcul du hmv5 :**

⇒ Sur une placette de surface n ares, on localise le point lidar le plus haut (Hi) et on suppose qu'il est le sommet de l'arbre le plus haut ;

⇒ On fixe à partir de ce point, un rayon d'exclusion $\boldsymbol{r = e^{(b+a*\log(H))}}$ avec a= 0.800431424 et b= -1.266961153 [20] ;

⇒ On supprime du nuage l'ensemble des points LIDAR inclus dans la zone d'exclusion de rayon r autour de Hi ;

⇒ Puis on répète n-1 fois la même opération en identifiant toujours les maxima locaux du nuage de points résiduel. On obtient alors la hauteur de n-1 maxima locaux dont on fait la moyenne;

⇒ Ainsi pour une placette de 6 ares, on obtient un Hmv5 qui est le résultats de la moyenne des hauteurs de 5 maxima locaux obtenus en suivant cet algorithme.

♦ **Exemple2 : calcul des taux de pénétration :**

Soit N= nombre total de first et de last.

$n_{<x}$ = nombre total de first et de last pulses se trouvant à une hauteur de moins de x mètres.

$$Taux\ de\ pénétration\ total\ PR_{tot} = (n_{<1})/N$$
$$Taux\ de\ pénétration\ de\ la\ couche\ supérieure\ PR_{ul} = (n_{<0.8\times h_{100}})/N$$
$$Taux\ de\ pénétration\ de\ la\ couche\ intermédiare\ PR_{il} = (n_{<0.5\times h_{100}})/N_{(<0.8\times h_{100})}$$
$$Taux\ de\ pénétration\ de\ la\ couche\ inférieure\ PR_{ll} = (n_{<1})/N_{(<0.5\times h_{100})}$$

3.2. Méthodes statistiques

3.2.1. Rappels sur la régression linéaire multiple.

On considère le modèle de régression linéaire suivant :

$y_i = \beta_0 + \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + \cdots + \beta_p x_i^{(p)}\ avec\ i = 1, \dots, n\ obsevations\ et$
$p = nombre\ de\ variables\ explicatives\ et\ \beta_0, \beta_1, \beta_2 \dots \beta_p\ les\ coefficients\ associés.$

Écriture matricielle du modèle : $Y = XB + \mathrm{E}$ avec $Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$, $X = \begin{bmatrix} 1 & x_1^{(1)} & \cdots & x_1^{(p)} \\ 1 & x_2^{(1)} & \cdots & x_2^{(p)} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & x_n^{(1)} & \cdots & x_n^{(p)} \end{bmatrix}$ et,

les paramètres à estimer $\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}$ et l'erreur $\mathrm{E} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$

Avec comme hypothèses $\varepsilon \overset{iid}{\rightarrow} N(0, \sigma^2 I_n)$, X1,…,Xp indépendantes et non aléatoires.

Pour la suite, nous définissons les quantités suivantes :
Variable endogène estimée : $\hat{y}_i = \beta_0 + \hat{\beta}_1 x_i^{(1)} + \hat{\beta}_2 x_i^{(2)} + \cdots + \hat{\beta}_p x_i^{(p)}$
Residual Sum of Squares ou Somme des carrés résiduels $RSS(\beta) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
Total Sum of Squares ou Somme des carrés totaux $TSS(\beta) = \sum_{i=1}^{n}(y_i - \bar{y})^2$

Dans notre démarche d'analyse et d'interprétation de nos modèles de régression, nous avons retenu quelques valeurs classiques nous permettant d'avoir un aperçu sur la qualité et les limites de nos modèle.

Un modèle de régression peut être en très bonne adéquation avec les données observées (R^2 élevé) sans pour autant garder cette qualité si ces données changent (par ex. pour une autre forêt) ou si l'on prédit au hasard l'une des observations. C'est pourquoi nous allons utiliser plusieurs méthodes de validation de modèle après avoir vérifié les hypothèses liées à la régression linéaire (analyse de résidus et des covariables). Les méthodes utilisées sont : la validation croisée leave-one-out et leave-n-out ; la variation moyenne de la RMSE après n validations croisées ou sa variation lorsqu'on passe d'un site à un autre ; l'importance du biais…

3.2.1.1. R^2 et R^2 ajusté ou coefficient de détermination multiple :

La différence entre TSS et RSS correspond à l'amélioration de la prédiction du modèle de régression, par rapport au modèle moyen [22]. Nous avons :$R^2 = \frac{(TSS-RSS)}{TSS}$. Il constitue l'amélioration proportionnelle de la prédiction du modèle de régression, par rapport au modèle moyen. Il indique la qualité de l'ajustement du modèle sur les données.
R^2 varie de zéro à un. Zéro indique que le modèle proposé n'améliore pas la prédiction par rapport au modèle moyen et 1 indique une prévision parfaite.
Un problème avec R^2, c'est qu'il croît mécaniquement tant que des prédicteurs sont ajoutés au modèle de régression. Cette augmentation artificielle se produit même si les prédicteurs ajoutés n'améliorent pas réellement l'ajustement.

Pour y remédier, on calcule le R^2 ajusté, qui intègre le nombre de degré de liberté du modèle. En effet, R^2 ajusté diminue si l'augmentation de la qualité de l'ajustement induite par de nouvelles prédicteurs ne compense pas la perte de degré de liberté $\overline{R^2} = 1 - \frac{RSS(\beta)/(n-(p+1))}{TSS(\beta)/(n-1)}$ [23]. De même, il augmente si les prédicteurs ajoutés apportent plus de précision au modèle. R^2 ajusté doit toujours être utilisé avec les modèles de plus d'une variable explicative. Il est interprété comme la proportion de la variance totale qui est expliquée par le modèle après pondération par le nombre de ses paramètres.

Enfin, dans notre étude, une valeur élevée du R^2 est importante car elle nous apporte de l'information sur la qualité de prédiction.

3.2.1.2. Le test global ou partiel de Fisher : le F-test

Le F-test évalue l'hypothèse nulle selon laquelle les coefficients de régression sont égaux à zéro par rapport à l'hypothèse alternative qui stipule qu'au moins un parmi eux est différent de zéro [24]. Une hypothèse nulle équivalente est que R^2 est égal à zéro. Un F-test significatif indique que le R^2 observé est fiable, et n'est pas dû au hasard dans l'ensemble des données. Ainsi, le F-test détermine si la relation proposée entre la variable réponse et l'ensemble des prédicteurs est statistiquement fiable, et peut être utile lorsque l'objectif de recherche est soit la prédiction ou l'explication.

3.2.1.3. RMSE (Root Mean Square Error ou racine carrée de l'erreur quadratique moyenne)

La RMSE est la racine carrée de la variance des résidus [24]. Elle indique l'adéquation absolue du modèle aux données. Elle s'obtient de la façon suivante :

Contrairement au R^2 est une mesure relative de l'ajustement, la RMSE en est une mesure absolue. Comme la racine carrée de la variance, RMSE peut être interprété comme l'écart-type de la variance inexpliquée, et a le mérite d'être de même unité que la variable de réponse. Des valeurs faibles de la RMSE indiquent un meilleur ajustement. La RMSE est une bonne mesure de la précision avec laquelle le modèle prédit la réponse. Dans le cas de la prédiction, elle reste le critère le plus

important pour juger de la qualité du modèle. Elle favorise les modèles qui se trompent souvent ,mais avec une faible erreur, aux modèles qui se trompent rarement mais avec de gros écarts.

3.2.2. Robustesse, généricité, fiabilité des modèles.

Un objectif important de notre étude est d'évaluer la généricité et la robustesse des modèles de prédiction de paramètres dendrométriques d'une forêt à l'autre. Un modèle développé pour une forêt est il applicable directement sur une autre, ou nécessite-t-il une recalibration, voir un reparamétrage complet ? Une façon de répondre a ces interrogations passe par des validations croisées ou des validations indépendantes.

3.2.2.1. R^2 VC, PRESS, Validation croisée leave-n-out et répétition de validation croisée

Une validation croisée simple(leave-one-out). a été effectuée en retirant une fois chaque observation de l'échantillon de calibration du modèle Ensuite le modèle est ajusté sur les n-1 observations restantes et l'écart entre la valeur de l'observation retirée et sa valeur prédite par le modèle est calculé. On en déduit alors le R^2 VC (coefficient de détermination multiple de validation croisée) et le PRESS (Predicted Residual Sum of Squares ou Somme des carrés des erreurs de prédiction) qui sont obtenus après n validations croisées leave-one-out. Le PRESS est la somme des carrés de ces écarts calculés : $PRESS = \sum_i \left(y_i - \hat{y}_{i,-i}\right)^2$. Le Root Mean PRESS est égal à la racine carré du PRESS déjà divisé par le nombre d'observations. Le R^2 VC est le coefficient de détermination multiple calculé en se basant sur ces n-1 erreurs de ces validations leave-one-out.

La validation croisée leave-n-out consiste à retirer une partie des observations de l'échantillon (10% par exemple [25]), d'estimer le modèle avec les d'observations restantes(90%) avant de calculer les erreurs de prédiction sur les observations retirées. Il est possible de se contenter de cette erreur de validation croisée mais par la suite nous répétons de cette procédure en calculant à chaque fois la RMSE de validation croisée. L'objectif est d'évaluer la moyenne et l'écart-type de cette liste d' erreurs(cinq(5) par exemples). La moyenne nous permet de juger de l'importance de l'erreur et l'écart type nous renseigne sur la variabilité de cette erreur d'une validation croisée à une autre.

3.2.2.2. Validation indépendante

La validation indépendante permet d'apprécier la robustesse d'un modèle d'un peuplement forestier à l'autre. Ainsi, les coefficients d'un modèle estimé sur les données d'un site sont utilisés pour prédire les variables observées sur un autre site. Il est alors possible de calculer l'erreur de prédiction et le biais d'un modèle. En réalité cette approche est plus robuste puisque les données utilisées sont totalement indépendantes. En effet dans la validation croisée (leave-one-out ou leave-n-out), même si l'échantillon de validation est choisi aléatoirement , les données récoltées sur un même site ont forcément des caractéristiques semblables qui font que la prédiction restera plus corrélée en passant de l'échantillon d'estimation à celui de validation. La validation indépendante nous semble obligatoire lorsque l'on désire évaluer la robustesse d'une méthode, avant d'atteindre une phase « opérationnelle ». Elle permet de juger de la généricité des outils que l'on teste. Après cette validation indépendante, un test de Chow peut être mis en œuvre pour comparer les coefficients obtenus pour un modèle donné sur deux sites différents.

3.3. Techniques de sélection de variables

Dans le contexte de notre étude, nous souhaitons connaître la contribution de chaque variable explicative à la capacité de prédiction du modèle. Or le nombre de variables explicatives (plus de 300 métriques) est très élevé par rapport au nombre de placettes de calibration (entre 25 ou 57 pour la surface terrière G selon les sites et 120 pour la hauteur dominante Ho).Ceci pose un problème d'estimation des coefficients dans la recherche de modèle. En plus, nous savons que, par construction, certaines métriques sont fortement corrélées.

Par ailleurs, il ne faut pas perdre de vue qu'une variable explicative peut être importante dans la prédiction de la variable endogène sans qu'il y ait une relation linéaire entre elles. En effet, si la relation entre la variable expliquée et l'une des variables explicatives n'est pas linéaire, une

transformation simple peut augmenter de manière considérable sa significativité dans un modèle de régression linéaire.

Ainsi, notre problème revient, avant tout, à mettre en place une stratégie de sélection de variables capable de tenir compte de ces préoccupations de colinéarité et de linéarité tout en restant parcimonieux dans l'établissement du modèle de prédiction.

Il existe deux objectifs dans la sélection de variable [3] : soit trouver des variables importantes fortement liées à la réponse à des fins d'interprétation, soit trouver une petite combinaison optimale de variables pour construire un modèle parcimonieux de prédiction de cette même réponse. Nous allons par la suite exposer deux méthodes statistiques qui permettent de sélectionner le « meilleur modèle » parmi une grande quantité de combinaisons possibles de variables. Il s'agit des forêts aléatoires et des Moindres Carrés Partiels(PLS ou Partial Least Squares). La première utilise l'algorithme développé par Breiman, celui des forêts aléatoires (Random Forest, voir [26] chapitre 15); et la deuxième méthode est paramétrique et utilise la régression PLS (Partial Least Squares). L'utilisation d'une STEPWISE tend à sélectionner des variables très corrélées alors que les deux méthodes de sélection précédentes permettent de diversifier cette sélection. Notre démarche a été de sélectionner un sous ensemble de métriques, identifiées comme important par ces méthodes, afin de réduire le nombre de variables à considérer, puis de procéder à une Stepwise pour obtenir un modèle final parcimonieux.

3.3.1. Random Forest : méthode non-paramétrique

Dans son article de 2001 [27] , Leo Breiman expose deux philosophies de modélisation au sein de la communauté des statisticiens. Il s'agit pour lui de la culture de modélisation des données et celle de modélisation algorithmique. L'usage du Random Forest rentre dans la deuxième catégorie. En effet, il nous permet de classer les variables explicatives en fonction de leur poids dans la prédiction .

Le Random Forest, comme son nom l'indique, est une technique de ré-échantillonnage qui utilise des arbres aléatoires [26]. L'idée principale dans le ré-échantillonnage est de faire la moyenne d'un grand nombre de modèles peu précis dans le but d'obtenir un modèle sans biais et de variance minimale. En effet B arbres sont construits à partir des variables explicatives. Ainsi, puisque les erreurs sont identiquement distribuées alors l'espérance de **B** arbres est équivalente à l'espérance de chacun de ces arbres. Ce qui fait que le biais dans l'ensemble des arbres est le même que celui dans chaque arbre et que l'amélioration de la précision ne passe que par réduction de la variance. La moyenne de B variables aléatoires identiquement distribuées, de variance σ^2 , admet une variance de $\frac{1}{B}\sigma^2$. Mais si ces variables aléatoires sont seulement identiquement distribuées (et pas forcément indépendantes, ce qui est notre cas) alors cette variance est évaluée à $\rho\sigma^2 + \frac{1-\rho}{B}\sigma^2$ [26] ,avec ρ la corrélation positive entre les arbres(donc entre les variables).

On remarque que plus **B** (nombre d'échantillons bootstrap) augmente plus le second terme décroit. Mais cette diminution est compensée en partie par l'augmentation de la corrélation entre les arbres générés. Au final ce mécanisme limite la réduction de la variance et par conséquent celle du bruit.

L'idée du Random Forest est d'améliorer la réduction de la variance en réduisant la corrélation entre les arbres tout en évitant de trop augmenter la variance. En effet, lors de la construction d'un échantillon bootstrap, avant de générer chaque nœud, un nombre $m \leq p$ de variables explicatives est choisi au hasard parmi p candidats. Intuitivement, réduire m permet de réduire la corrélation entre les arbres et par conséquent de réduire la variance à travers la formule $\rho\sigma^2 + \frac{1-\rho}{B}\sigma^2$, avec ρ la corrélation entre les arbres aléatoires.

L'importance de la variable X^j liée à la forêt aléatoire peut se définir de la façon suivante. Pour chaque arbre a de la forêt, on considère l'échantillon OOB_a associé. En effet, avant de construire l'échantillon bootstrap, une partie des observations est retirée de l'échantillon servant à construire l'arbre en question (voir annexe A). Ces observations qui constituent l'échantillon out of bag OOB_a (en dehors du sac) serviront à calculer l'erreur de prédiction commises par l'arbre concerné. On note $errOOB_a$ l'erreur quadratique moyenne(MSE) de l'arbre a sur l'échantillon de validation

OOB_a. Ensuite, les valeurs de X^j dans OOB_a sont permutées aléatoirement pour obtenir un échantillon perturbé noté $\widetilde{OOB_a^j}$. $err\widetilde{OOB_a^j}$, l'erreur de prédiction de l'arbre **a** sur l'échantillon perturbé, est calculée. L'importance de la variable X^j est donnée par :

$VI(X^j) = \frac{1}{ntree}\sum_a(err\widetilde{OOB_a^j} - errOOB_a)$, sachant que la somme se fait sur tous les arbres et que $ntree$ est le nombre d'arbres dans la forêt aléatoire (Random Forest)

3.3.2. La régression par moindres carrés partiels ou PLS (Partial Least Squares)

La régression PLS est une technique récente, mais bien répandue, qui généralise et combine les caractéristiques de l'analyse sur composantes principales et de la régression multiple. Elle est particulièrement utile quand on a besoin de prédire un ensemble de variables dépendantes à partir d'un ensemble très grand de variables explicatives(prédicteurs) qui peuvent être très fortement corrélées entre elles. PLS est donc une méthode pour construire des modèles de prédiction quand les facteurs sont nombreux et très colinéaires. Dans le cas où il y a qu'une seule variable à prédire, comme dans notre cas, on parle de la PLS1.

Considérant que la PLS est un sujet assez répandu (ouvrage de référence en français de M. Tenenhaus [28]), les détails ne seront pas exposés dans ans cette étude. Toutefois le calcul des quelques grandeurs, relatives à cette méthode et dont nous avons fait usage, sera exposé.

En résumé, la PLS est une méthode factorielle et linéaire qui, à partir des variables de départ, cherche *h(h=1,…,A)* combinaisons linéaires successives notées $t_h = X_{h-1}b_h$, qui sont liées avec la variables expliquée *y*. ***A*** représente la dimension du modèle. Ensuite, la variable *y* est modélisée linéairement à partir ces combinaisons de variables, aussi appelées variables latentes. Enfin, cette régression est exprimée en fonction de X sous la forme : $y = b_1x_1 + b_2x_2 + \cdots + b_mx_m + résidu$

♦ **poids B :**

C'est la valeur estimée, pour chaque variable, des coefficients *b* dans la formule de régression ci-dessus.

♦ **Le PRESS : Predicted Residual Sum of Squares**

Les observations sont partagées en N groupes, et on réalise N fois $t_h = X_{h-1}b_h$ en enlevant à chaque fois un groupe. On obtient [28]:

$$PRESS_h = \sum_i\left(y_{(h-1),i} - \hat{y}_{(h-1),-i}\right)^2$$

Où $\hat{y}_{(h-1),-i}$ est calculé dans l'analyse réalisée sans le groupe contenant l'observation (i).

♦ **Variable importance in the Prediction : VIP :**

Composant PLS : $t_h = X_{h-1}b_h$, avec $\|b_h\| = 1$

L'importance de la variable $x_j(j = 1, \dots, p)$ pour la prédiction de *y* dans un modèle à *m* composantes :

$$VIP_{mj} = \sqrt{\frac{p}{\sum_{h=1}^{m} cor^2(y,t_h)}\sum_{h=1}^{m} cor^2(y,t_h)\, b_{hj}^2}$$

La somme des carrés des VIP sur les variable est égale à 1. Une variable est jugée importante dans la prédiction si son VIP est supérieur à 0.8 [28].

3.4. Mise en œuvre des méthodes

Nous allons, dans un premier temps, analyser, tester et discuter quelques modèles linéaires proposés dans la littérature. En effet, une prédiction de Ho et de G sera réalisée en gardant les coefficients des modèles originaux. Ensuite nous réalisons un ajustement de ces mêmes modèles sur nos propres données (forêt de Haye principalement). Enfin, nous comparons numériquement et graphiquement ces résultats.

Dans un second temps, à l'aide des deux méthodes de sélection citées plus haut, nous nous attacherons à définir nos propres modèles, pour G seulement, avant de procéder à différents tests de validation.

Les métriques sont classées en fonction des types de points LiDAR qui ont servi à leur calcul et de leur sens biologique. D'une part, il peut s'agir d'un filtre sur la nature des points(type d'échos :

premier, dernier, intermédiaire ou unique), sans filtre au cas où tous les points sont pris en comptes(Sans filtre), avec filtres tels que les points dessus de 2 mètres ou de 1 mètre, pour éliminer le « bruit » associé à la strate herbacée. D'autre part, cette classification sera combiner avec l'indice d'importance des variables (increasing MSE pour random Forest et VIP pour PLS) pour faire le tri et choisir parmi les variables candidates.

3.4.1. Procédure de sélection avec Random Forest

Pour sélectionner des variables à l'aide du Random Forest, nous procédons aux étapes suivantes :

Etape1 :Lancer la procédure randomForest du package du même dans le logiciel R avec toutes les variables (plus de 300) ;
Etape2 :Voir de manière globale quelles sont les groupes selon notre classification de variables qui sortent le mieux au sens du increasing MSE de chaque variable(voir dictionnaire Annexe B) ;
Etape3 : Sélectionner les variables/groupes de variables les plus importantes au sens du increasing MSE (70% environ)
Etape4 :Sélectionner la variable la plus pertinente parmi les groupes de variables corrélées.

Reprendre ces étapes avec les variables sélectionnées jusqu'à obtenir un minimum de variables(environ 40 ou 45 variables). A partir de ce moment, la PROC REG de SAS est utilisée dans le but de sélectionner le meilleur modèle à quatre ou cinq variables, avec les options selecte=stepwise, best=5 et stop=5, en choisissant une significativité pour rentrer dans le modèle de 15% pour significativité plus petite d'y rester de 5%.

3.4.2. Procédure de sélection de variables avec PLS.

L'utilisation de la PLS se justifie, dans notre cas, par le besoin de réduire le rang du modèle de régression recherché en regroupant les variables corrélées et d'en choisir un nombre restreint. Comme dans le cas du Random Forest, la démarche suivie pour sélectionner les variables est la suivante :
Etape1 :Faire une PLS avec toutes les variables explicatives disponibles ;

Etape2 : Eliminer les variables/classes avec une importance faible(un VIP, variable importance factor inférieur à 0.80 [29])et dont le coefficient associé dans la régression est inférieur en valeur absolu à 0.002 (arbitraire) ;

Etape3 : Refaire la même procédure jusqu'à ce que plus de 90% des variables explicatives soit exprimées dans la PLS ;

Etape4 : A partir de ce moment faire une sélection manuelle : si deux variables sont corrélées, choisir celles avec un plus grand VIP et/ou une meilleure interprétation biologique ;

Etape5 :Choisir un modèle final à l'aide des variables restantes (environ 40) à l'aide d'une stepwise en choisissant le seuil d'entré et de sortie des variables candidates comme dans la cas du Random Forest.

3.4.3. Vérification des hypothèses :

Pour chaque modèle établi et analysé, une vérification des hypothèses de validité a été mise en œuvre pour que les résultats soient fiables. Il n'est pas jugé nécessaire, dans cette étude, de présenter ces vérification mais seulement les techniques utilisées seront décrites.
Pour chaque site d'étude, l'échantillon a été choisi de sorte qu'il soit représentatif de la forêt étudiée. Les métriques, qui constituent les variables explicatives, sont calculés de façon indépendante les uns des autres et dépendent tous du nuage de points lidar.

Dans chaque modèle, pour vérifier la **normalité des résidus**, le QQplot est tracé et la normalité est supposée si le nuage de points est aligné sur un droite. De plus, le test de normalité de Shapiro et Wilk est aussi mise en œuvre pour plus de précision au seuil de 5%.

Par ailleurs, nous avons tracé le graphique qui met, en abscisse, les résidus studentisés de la régression, et en ordonnée, les valeurs observées de la variable expliquée. Ainsi, **l'homoscédasticité** est respectée si ces résidus sont répartis de façon homogène de part et d'autre de l'axe des abscisses. En plus, le test de White est données en sorti pour vérifier cette tendance graphique.

Dans tous les cas, si une des hypothèses n'est pas vérifiée, à cause souvent d'observations aberrantes et/ou influentes , alors en se basant sur la distance de Cook ou sur les caractéristiques graphiques, ces points sont écartés de l'échantillon jusqu'à nouvelle ordre. Toutefois, ces observations mal prédites sont, dans la plupart des cas, des valeurs tout à fait raisonnables et réellement existantes.

3.5. Logiciels utilisés

Deux logiciels sont essentiellement utilisés dans nos traitements statistiques et dans la gestion de données. Il s'agit du logiciel libre R (R Development Core Team (2011). 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/.) et de SAS(SAS Institute).

R est utilisé tout d'abord dans un souci de conformité et de continuité avec l'environnement de travail ONF. En effet, le service de Recherche&Développement de l'ONF à Velaine a fait le choix de faire de R son principal outil de traitement statistique. Ce qui permettra certainement de valoriser ce travail dans l'avenir. De plus, c'est le logiciel dont la pratique nous est beaucoup plus familière puisqu'il a servi très amplement à notre formation de statisticien. En dehors de ces considérations, il faut noter que R est capable de réaliser toutes les manipulations statistiques que nous avons prévues de mettre en œuvre et présente l'avantage d'être acquis gratuitement.

Néanmoins, le logiciel SAS est utilisé dans la réalisation de la PLS. Cette emploi est essentiellement dû au fait que nous avons travaillé avec nos partenaires de l'INRA(Institut Nationale de la Recherche Agronomique) qui l'utilise. La procédure SAS PLS est utilisée pour produire les sorties nécessaires au calcul du VIP et des poids de chaque variable dans la régression PLS. La procédure REG a servi dans la sélection par stepwise de variables après en avoir isolé un nombre restreint à l'aide des méthodes de sélection choisies dans ce travail. Ce choix se fonde sur le fait que l'option « selection » de cette procédure nous donne la possibilité de fixer la significativité nécessaire à une variable pour renter et/ou rester dans le modèle. Ce mécanisme nous permet d'éviter de saturer ce dernier.

4. Résultats

4.1. La hauteur dominante

Au vu des résultats des modèles du Tableau 2, il apparaît que l'estimation de la hauteur dominante est assez précise. En effet les deux modèles de régression (Martins et Heurich) présentent chacun un R^2 de 0.99 ; ce qui signifie que les informations apportées par le LiDAR sont capables d'expliquer jusqu'à de 99% la hauteur dominante. En outre la RMSE est d'environ 76cm et 75 cm soit une erreur respective de 3.2% et de 2.7% en moyenne. Par contre le modèle de Naesset qui nécessite une détransformation (voir 0) donne un R^2 qui ne dépasse pas 0.89. Pire, l'erreur réelle recalculé est beaucoup plus importante par rapport aux deux premiers modèles et atteint les 1.14m soit une erreur moyenne 4.8% .

La validation croisée leave-one-out fournit un Root Mean PRESS de 0.69m et 0.78m pour les modèles de Martins et de Heurich respectivement. Cette valeur, qui est d'autant plus importante si certaines observations sont mal reproduites, montre que ces deux modèles résistent bien à cette épreuve de validation croisée. La RMSE moyenne, obtenue après cinq validations croisées leave-n-out, confirme cette situation. Celle-ci est très proche de celle obtenue dans l'estimation du modèle dans le cas de Martins mais passe néanmoins à 1.11m pour dans le cas de Heurich.

Modèle	**Naesset**[1]	**Heurich**	**Martins**
Formule	$log(Ho) = 0.736 - 0.896\,log(hp80l1m) + 1.702log(hp90l1m) + 0.03log(d9l)$	$Ho = 17.6 + 0.4h_{maxvf} - 4PR_{lll} + 0.6hp90\text{vf} + 11.3d9l - 0.3\,h_{medvf} + 0.3hp60\text{vf} - 0.1hp20\text{af} - 0.5hcval$	$Ho = 0.14 + 0.98 * Hmv5 + 0.039 * d9$
Estimation des modèles			
R2	0.889	0.99	0.99
R2 ajusté	0.879	0.989	0.99
RMSE(m)	1.14 soit 4.8% en moyenne	0.647 soit 2.7% en moyenne	0.761 soit 3.2% en moyenne
Validation croisée			
R2 CV	0.934	0.989	0.99
Root Mean PRESS(m)	1.15	0.69	0.78
Moyenne de 5 RMSE(m)		1.114	0.73
SD de 5 RMSE(m)		0.01	0.08

Tableau 2: Estimation et validation croisée des modèles de hauteur dominante (Ho)

Lors du passage d'un site à l'autre dans le cadre de la validation indépendante(voir Tableau 3), il apparaît que le modèle de Matins reste le plus stable avec une erreur de prédiction de 1.71m sur le site de l'OPE si on conserve les coefficients du modèle obtenus sur les placettes de forêt de Haye. Même en réajustant le modèle aux données de OPE, la RMSE, qui est l'erreur théorique

[1] Il faut noter que le modèle de Naesset est sous forme log. Donc une détransformation est effectuée pour revenir en mètre(voir les détails dans le paragraphe 0 sur la partie Naesset).

d'ajustement, est de 1.59m soit une amélioration de seulement 12cm (Figure 4). Or cette RMSE est la plus petite erreur possible du modèle sur les données concernées; ce qui pourrait montrer qu'il n'est pas nécessaire de réévaluer les coefficients de ce modèle obtenu sur le site de Haye. Les résultats sont similaires si la validation se fait sur les forêts de Languimberg ou de Montagne avec des erreurs de prédiction respectives de 1.08 et 2.63m. De même que sur OPE, après un réajustement on ne gagne qu'une précision de 12cm sur Languimberg et de 7cm sur Montagne. Néanmoins, il faut noter que l'erreur du modèle de Martins, réajusté sur les données des autres sites, est plus importante que celle obtenue en forêt de Haye. Pour le modèle de Heurich, il reste précis s'il est recalibré mais par contre, il prédit des valeurs largement différentes de la réalité de terrain si on conserve les coefficients obtenus sur la forêt de Haye(Figure 5).

Modèle	**Naesset**	**Heurich**	**Martins**
Validation indépendante sur OPE			
Placette de calibration	Modèle original : avec les coefficients de l'article	Modèle original	Haye : 120 placettes
Placette de validation	Haye : 120 placettes	Haye : 120 placettes	OPE : 27 placettes
biais	0	Trop grand : 487.69	0.03
Ecart type (en m)	2.15	Trop grand : 76.66	1.71
RMSE théo[2] (m)		1.67	1.59
Validation indépendante sur Languimberg			
Placette de calibration		Haye	Haye : 120 placettes
Placette de validation		Languimberg : 51 placettes	Languimberg : 51 placettes
Biais(en m)		29.88	0.12
Ecart type (en m)		97.92	1.08
RMSE théo(m)		0.83	0.96
Validation indépendante en Montagne			
Placette de calibration	Modèle original	Modèle original	Haye
Placette de validation	Montagne : 64 placettes	Montagne : 64 placettes	Montagne : 64 placettes
Biais(en m)		120.14	1.33
Ecart type (en m)		57.33	2.63
RMSE théo(m)		2.23	2.56

Tableau 3: Validation indépendante des modèles de hauteur dominante (Ho)

[2] C'est la RMSE obtenue en réajustant le modèle sur les données de validation, s'exprime mètres.

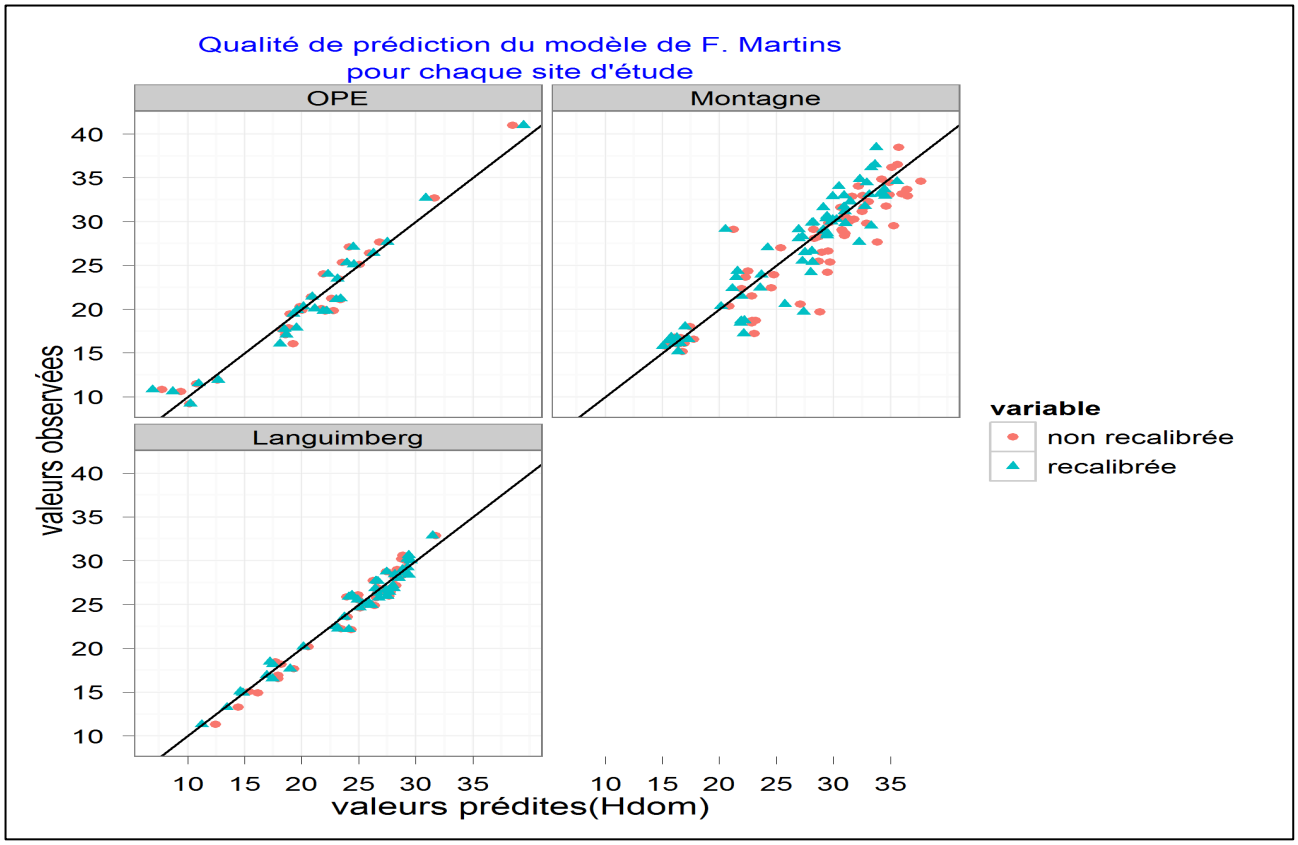

Figure 4: Robustesse du modèle Martins d'un site à l'autre

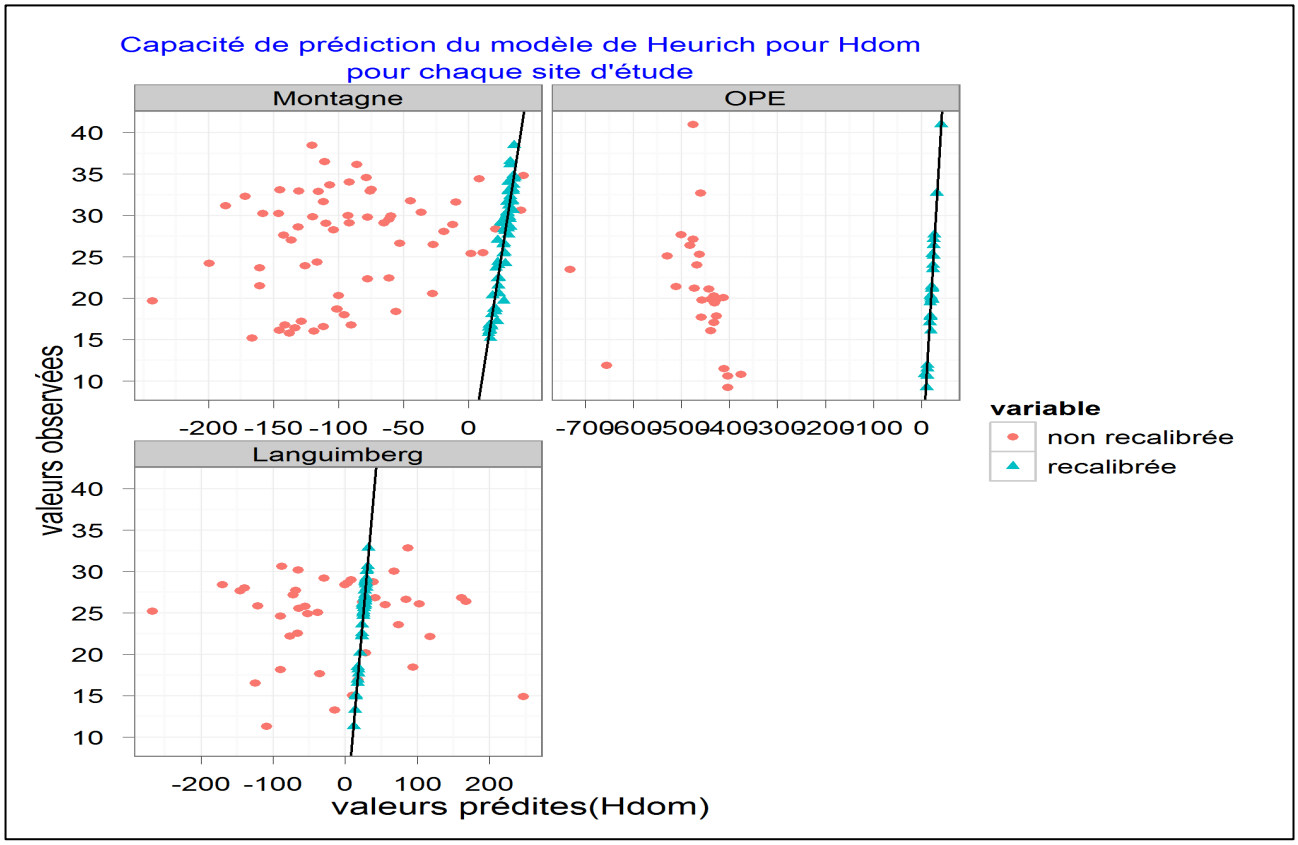

Figure 5: illustration de l'inefficacité du modèle de Heurich pour Hdom d'un site à l'autre.

4.2. La surface terrière(G)

Pour chaque méthode de sélection, nous avons retenu deux modèles en choisissant le premier en fonction des données issues de tous les sites, et le second en n'utilisant uniquement que les données de la forêt de Haye. Cela veut dire que les modèles 1 et 3 tiennent compte des sensibilités de tous les sites et nous fait espérer une plus grande robustesse dans la validation indépendante. Les Tableau 4 et Tableau 5 résument l'ensemble des résultats des modèles de prédiction de surface terrière obtenus.

D'abord les modèles de Naesset et de Heurich, proposés dans la littérature, présentent respectivement des R^2 ajusté de 0.31 et 0.46 lorsqu'ils sont estimés sur les 25 placettes de la forêt de Haye. Ce qui correspond à des RMSE respectives de 1.14m²/ha et 4.30m²/ha. En passant à une validation croisée leave-one-out on obtient des Root Mean PRESS plus élevés respectivement de 1.25m²/ha et de 6.86m²/ha. Le modèle de Heurich résiste moins bien à cette épreuve à cause certainement du nombre important de variables explicatives utilisées (10) et des coefficients associés à ces variables qui sont trop fluctuants pour être stables(Tableau 4).

Ensuite, les modèles 1 et 2(Figure 6), obtenus grâce à la procédure de sélection de variable basée sur l'utilisation de Random Forest, possèdent des caractéristiques différentes l'un de l'autre. En effet le modèle 1, avec un pouvoir prédictif plus faible (R^2=0.58), présente l'avantage d'être plus robuste que le modèle 2 (R^2=0.93) en passant d'un site à l'autre sans réévaluation des coefficients. Même si cette erreur de prédiction est plus importante dans le premier cas, celle-ci varie dans des proportions nettement moindre que celle du modèle 2 qui explose en validation indépendante. Pour ce derniers, elle passe de 1.92 m²/ha en validation croisée à 23,23.7 et 27 m²/ha lors des validations indépendantes respectives sur les sites de l'OPE, de Languimberg et de Montagne alors que l'erreur du modèle 1 passe de 5.88 à 7.19, 9.63 et 14.92 m²/ha, soit respectivement 0.470 ,3.486 et 5.340m²/ha de plus par rapport à l'erreur minimale obtenue en réajustant(Tableau 5). Toutefois le modèle 2 reste plus précis si une ré-estimation est opérée sur chaque site d'étude.

Enfin, les modèles 3 et 4 sont à l'image des modèles 1 et 2(Figure 7). Ils présentent respectivement des R^2 de 0.492 et de 0.852 pour une RMSE de 5.354m²/ha et de 2.896 m2/ha . Comme dans le cas précédent le modèle 4 résiste moins bien à la validation croisée et présente des erreurs très importantes d'un site à l'autre avec 40.72 m2/ha , 15.26 et 18.73 m2/ha pour respectivement les sites de l'OPE, Languimberg et Montagne. Ces valeurs sont de 14.81, 8.13 et 26.08 m2/ha pour le modèle 3. On constate qu'exceptionnellement, l'erreur sur les sites de Montagne est beaucoup plus important pour le modèle 3 par rapport au modèle 4 contrairement à la logique des autres sites.

	Modèles existants		Sélection random Forest		Sélection PLS	
Modèle	**Naesset**	**Heurich**	**Modèle 1**	**Modèle 2**	**Modèle 3**	**Modèle 4**
Formule	$log(G) = 0.883 + 1.081\,log(hp70l1m) + 0.468log(d4l)$	$G = 59.9 + d9f - 48.8habs_{2m} + 17.4hcv_{f2m} - 11.2hp30f2m - 0.5hp40f2m + 24.5\,PRil\_f - 231.2d1l + 207.6\,d2l$	$G = 24.48 - 3.12d2f + 0.73hp95\text{vf} - 0.25\,PRl\text{l\_f} + 0.109\text{PRtot\_l}$	$G = 192.56 - 79.09\text{D\_2\_12} - 1.79\text{D9}_{\text{f}} - 15.89hp302m + 22.59hp402m - 8.42hp752m$	$G = 12.33 - 72.88\text{D}_{\text{dbh}} + 0.53\text{d4}_{\text{a}} + 0.313hmean\_2m + 0.66hhmv5 - 0.10PRll\_f$	$G = 26.89 - 92.12\text{D\_2\_6} - 20.71\text{hmean}_{2\text{m}} + 3.43hp202m + 10.19hp402m + 5.33hp902m$
Estimation des modèles(données forêt de Haye)						
R2	0.374	0.67	0.581	0.935	0.492	0.852
R2 ajusté	0.317	0.46	0.498	0.918	0.359	0.812
RMSE(m)	1.14 (4.16%)	4.30 (14.3%)	4.863 (16%)	1.916 (6.37%)	5.354 (17.8 %)	2.896(9.6 %)
Validation croisée(données forêt de Haye)						
R2 CV	0.191	0.311	0.408	0.86	0.214	0.74
Root Mean PRESS(m)	1.25	6.84	5.88	2.84	7.04	3.88
Moyenne RMSE	1.22	6.26	4.04	1.92	4.45	2.9
SD RMSE	1.07	3.25	1.13	0.55	1.38	1.11

Tableau 4:Estimation et validation croisée des modèles de surface terrière(G)

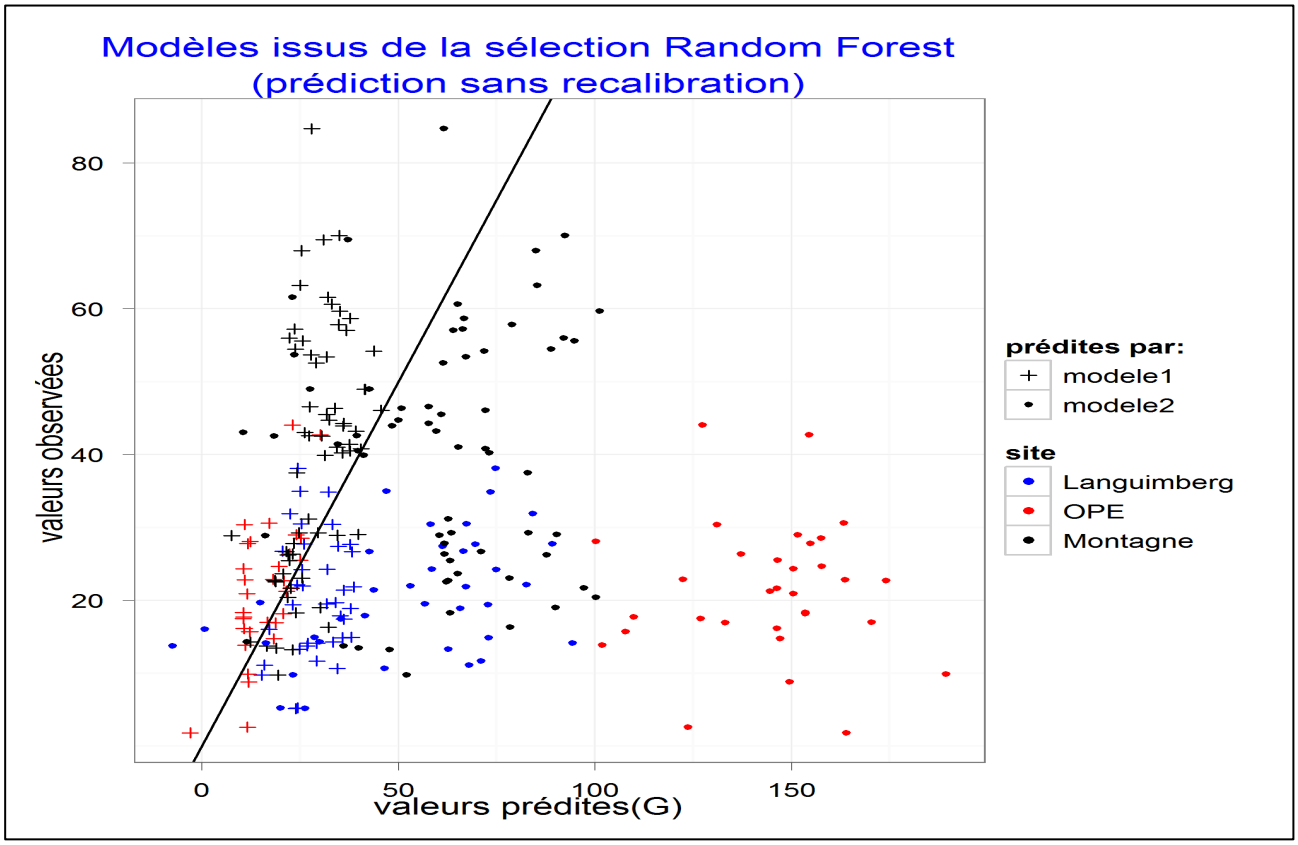

Figure 6:Evaluation de la qualité de prédiction de G des modèles issus de la sélection par Random Forest

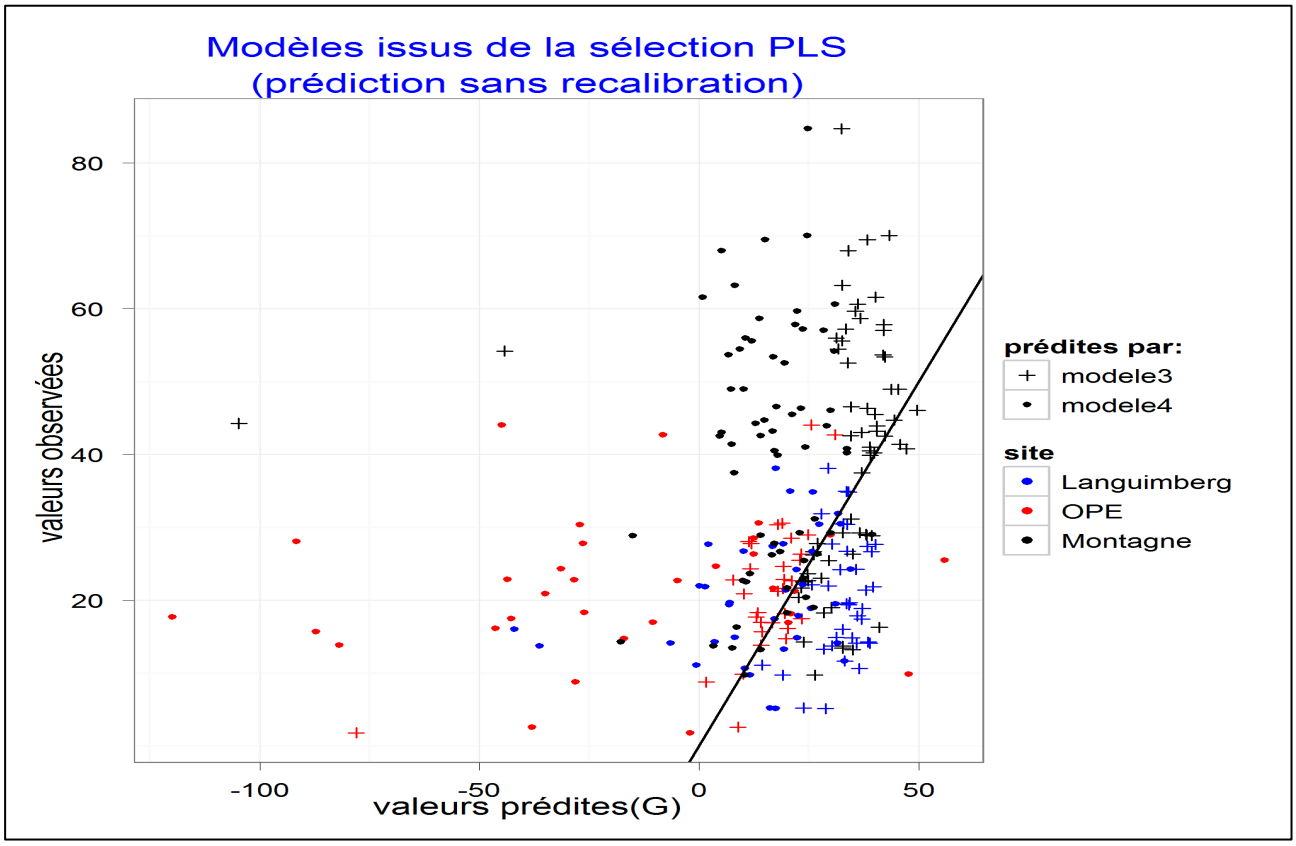

Figure 7: Evaluation de la qualité de prédiction de G des modèles issus de la sélection par PLS calibré sur Haye.

	Modèles existants		Sélection Random Forest		Sélection PLS	
Modèle	**Naesset**	**Heurich**	**Modèle 1**	**Modèle 2**	**Modèle 3**	**Modèle 4**
Validation indépendante sur OPE						
Placette de calibration	Modèle original	Modèle original	Haye :	Haye :	Haye	Haye
Placettes de validation	OPE	OPE	OPE	OPE	OPE	OPE
Biais(en m)		1959.15 !!!	5.44	-122	7.43	39.64
Ecart type		258.62 !!!	7.19	23.19	14.81	40.72
RMSE théo		258.62 !!!	6.72	6.45	4.94	5.65
Validation sur Languimberg						
Placette de calibration	Modèle original	Modèle original	Haye	Haye	Haye	Haye
Placettes de validation	Languimberg	Languimberg	Languimberg	Languimberg	Languimberg	Languimberg
Biais(en m)		315.87 !!!	-8 .52	-32.16	-12.23	6.82
Ecart type		740.18 !!!	9.63	23.73	8.13	15.26
RMSE théo		5.3	6.144	4.75	5.786	6.02
Validation en Montagne						
Placette de calibration	Modèle original	Modèle original	Haye	Haye	Haye	Haye
Placettes de validation	Montagne	Montagne	Montagne	Montagne	Montagne	Montagne
Biais(en m)		861.05 !!!	11.34	-21.29	8.26	23.5
Ecart type		582.65 !!!	14.92	27.15	26.08	18.73
RMSE théo		7.07	9.58	7.70	7.89	8.09

Tableau 5: Validation indépendante des modèles de surface terrière(G)

5. Discussions

5.1. Hauteur dominante

Dans le cadre de l'estimation de la hauteur, le modèle de Martins est le plus stable parmi les trois modèles étudiés. Celui de Naesset, en plus d'avoir un pouvoir prédictif plus faible que les autres, présente l'inconvénient d'utiliser la transformation logarithmique. Le biais introduit par la détransformation, nécessaire car log(Ho) n'admet pas d'interprétation biologique connue, dépend en partie de la variance des hauteurs qui est associée elles aux peuplements étudiés. Le modèle de Heurich, bien que présentant un pouvoir prédictif « local » élevé, est pénalisé en matière de robustesse, d'une part par le nombre très important de variables explicatives et, d'autre part, par les intervalles de confiance très larges des coefficients associés à ces variables.

Ainsi, le modèle de Martins reste le meilleur candidat pour estimer Ho. En effet, il ne comporte que deux variables explicatives que sont le ***hmv5*** et la densité des échos lidar de la couche supérieure de la canopée ***d9***. Sa stabilité en validation indépendante nourrit l'espoir d'en faire un modèle opérationnel capable d'estimer Ho, à l'échelle d'un massif forestier pour lequel des données LiDAR seraient disponibles. Toutefois, il est noté que l'erreur minimale commise même avec une recalibration est de 1.59m et 2.56m pour respectivement OPE et Montagne contre 0.76m et 0.96m pour Haye et Languimberg. Donc il est important de comprendre pourquoi ce modèle se comporte relativement mal sur les deux sites précités ?

Un élément important présent dans le modèle retenu est la variable ***hmv5***. Cette variable vise à identifier des maximum locaux, tout en excluant des zones circulaires assimilables aux houppiers des arbres(voir 3.1.3 exemple 1). Comme cette variable dépend de la forme des couronnes (rayon d'exclusion) et qu'elle a été calibrée sur des feuillus et non sur des résineux, son application hors de son domaine de calibration est donc hasardeuse. Par exemple, les conifères ont en moyenne des houppiers de plus faibles ampleur que les feuillus, bien que la présence de feuilles chez les résineux et la forme coniques des arbres conduisent à une probabilité plus forte de bien identifier l'apex de l'arbre. Ceci nous laisse penser que le rayon d'exclusion du ***hmv5*** conduit probablement à éliminer des zones importantes du nuage laser et tend donc a sous-estimer Ho sur le site de Montagne (voir Figure 8). C'est pourquoi une autre formule de détermination du rayon d'exclusion a été proposée [30]. Cette méthode propose de tenir compte, d'une part, de l'altitude qui jouerait un rôle sur la forme des houppiers, du type de peuplement d'autre part(résineux/feuillu). En plus, la formule proposée admet des coefficients différents selon que le peuplement soit composé de conifères ou de feuillus. Ceci pose cependant une condition supplémentaire d'un point de vue opérationnelle, celle de pouvoir a priori identifier l'essence de l'arbre en automatique, ce qui est difficile à partir du seul nuage LiDAR. Enfin, la pente joue également sur l'estimation de la hauteur. En effet comme les apex sont rarement à l'aplomb direct du pied de l'arbre, souvent leur projection verticale tombe en aval de la souche [31]. Ainsi la hauteur lidar se trouve surestimée en situation de pente. Rappelons que par convention, la mesure de terrain, se fait par rapport à un point de référence situé du côté amont . Il peut alors y avoir en fonction du diamètre des arbres et de l'amplitude de la pente, de 0.1 à 0.5 m d'écart dans de telles situations.

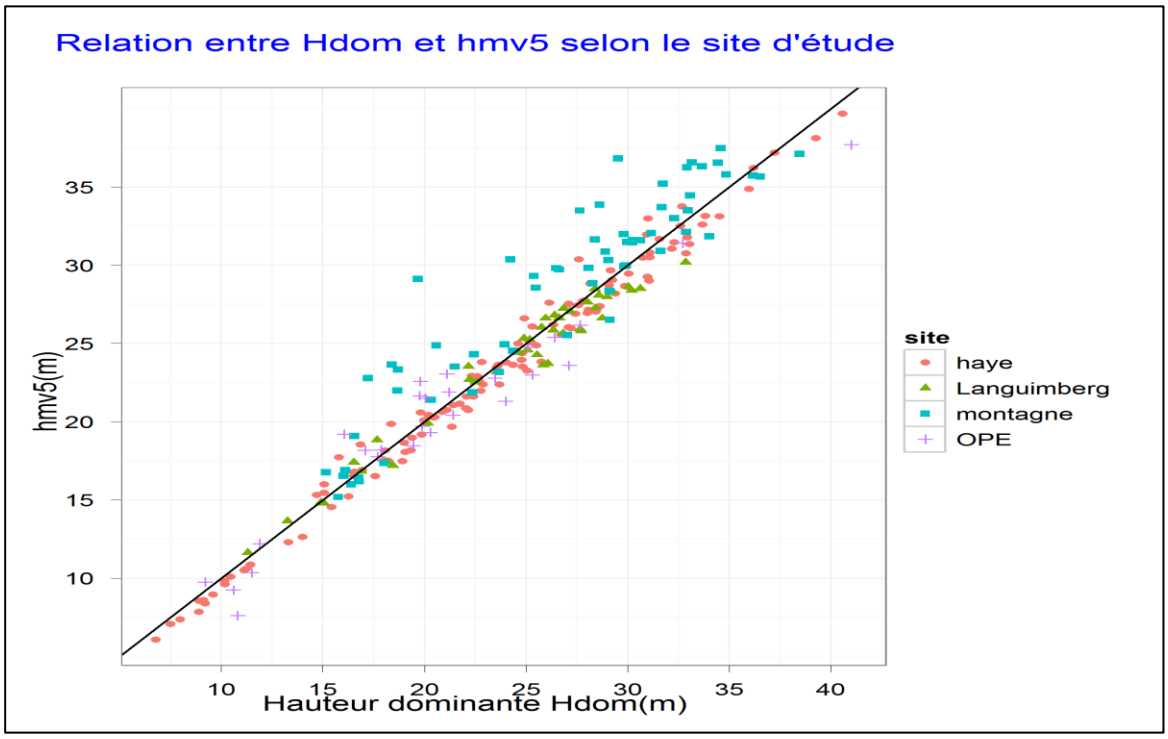

Figure 8: illustration du mauvais mode de calcul de *hmv5* sur Montagne

Dans le modèle de Martins retenu, une deuxième variable d'importance est d9. Il est intéressant de noter que cette variable est liée à l'état de végétation (en feuille ou hors feuille) au moment du vol LiDAR. Ainsi, en pleine saison de végétation (pour les feuillus du moins), la présence de feuilles freine la pénétration du signal lidar et la hauteur du 90ième percentile se trouve par conséquent plus élevée qu'en condition de porosité plus grande. La densité des échos se trouve également à être plus élevée dans cette couche supérieure en présence de feuilles. Il apparaît ainsi, au vu de la Figure 9 , que les sites de Haye et de l'OPE, pour lesquels les données lidar ont été acquises hors feuille, sont pris en sandwich entre les deux autres sites (Languimberg et Montagne), acquis alors qu'ils étaient en feuille. En effet, d9, qui représente la densité relative dans la couche supérieure, augmente si la quantité de feuille sur cette tranche augmente. Ce constat se confirme si une analyse de la variance est mise en œuvre pour déterminer l'effet de l'état de végétation. Cette analyse se fait en introduisant un facteur « végétation » dans le modèle de Martins. La p-value obtenue pour ce facteur est très largement inférieure au seuil de risque de première espèce fixé de 5%. De même, l'effet type de peuplement (résineux/feuillu) est aussi fortement significatif. En clair, la faible densité d9 dans les placettes montagne s'explique par la forme de la canopée des peuplements de conifères qui présente des vides importantes à hauteur de la première couche. Dans le cas du site de Languimberg, l'importance de d9 est, en partie, due à l'interceptions des échos par les feuilles nombreuses de la couche supérieure.

Enfin, la concordance entre le vol lidar et les mesures de terrain doit également être prise en compte. Un décalage d'une saison de végétation s'est écoulé en forêt de haye entre le vol lidar et les mesures de terrain. Ainsi les mesures de ce site ont eu lieu en fin 2007 alors que le vol lidar avait préalablement eu lieu dès le début de cette même année. Cette saison de végétation écoulée, conduit à une croissance en hauteur estimée entre 20cm et 50cm entre la mesure de terrain et le vol lidar. Dans ces conditions, la calibration lidar a tendance à sous estimer la hauteur mesurée de 20cm à 50cm. En tenant compte de cette erreur, il est possible de réduire le biais calculé sur les autres sites de cette même quantité.

Pour résumer les résultats portant sur Ho, cette étude montre que le modèle de Martins est capable de prédire la hauteur avec une précision de plus de 95% sans recalibration, mais à condition, d'une

part, de réaliser les vols lidar dans le même état de végétation (hors feuille de préférence), et d'autre part, d'adapter le calcul du hmv5 dans le cas des conifères et en tenant compte de l'altitude comme l'ont proposé Hollaus et ses collaborateurs [10] . Chose intéressante également à noter, c'est que l'ajout au modèle de variables tel que d9 permet possiblement d'atteindre une plus grande robustesse, en raison d'un rôle d'ajustement que pourrait jouer cette variable vis à vis des conditions de vol et tout particulièrement de la densité des émission LiDAR, ou de l'état de végétation (en feuille ou hors feuille).

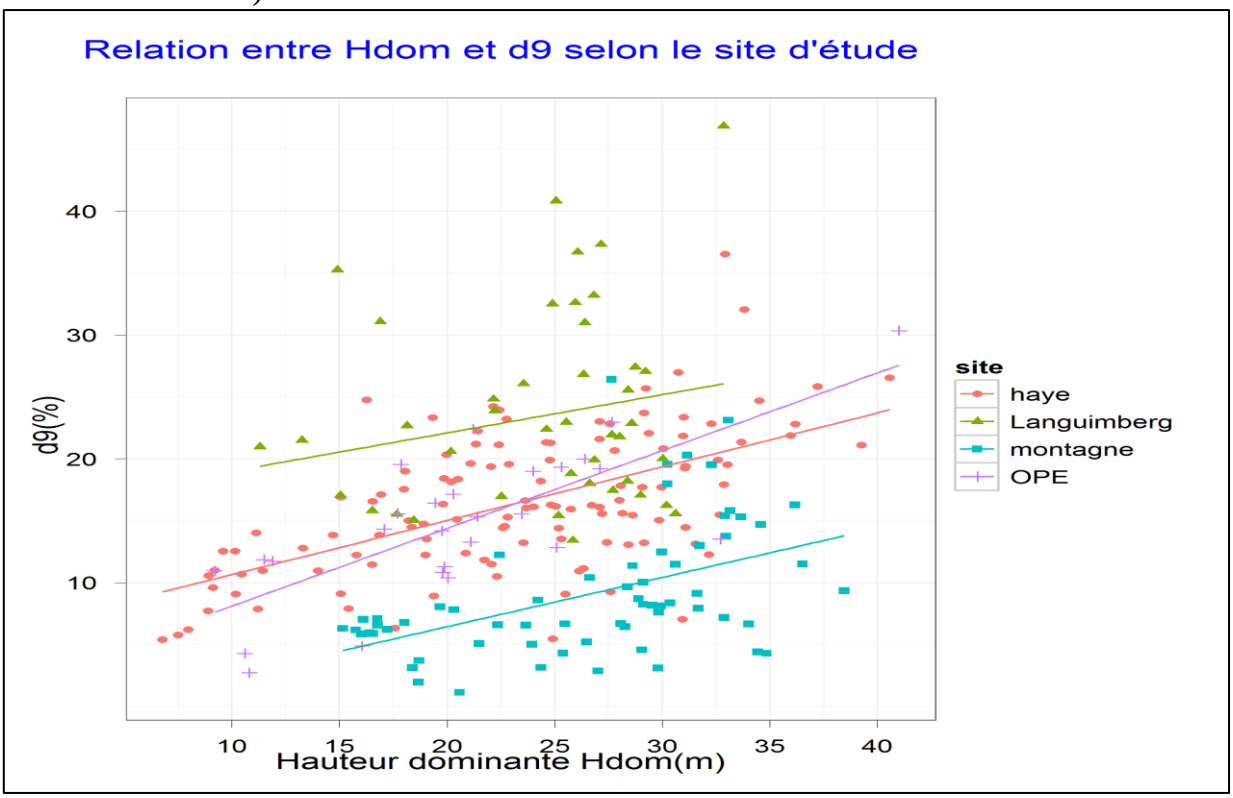

Figure 9: illustration de la dépendance de *d9* à la saison de végétation(feuille/hors feuille) et du type(feuillus/résineux)

5.2. Surface terrière :

Pour l'évaluation du G, d'un point de vue opérationnel, une précision de l'ordre de 10% est souhaitée par le forestier. Par exemple, une erreur inférieure à 3 m²/ha semble être une limite acceptable pour un peuplement de 30 m²/ha (les peuplements étudiés faisaient en moyenne entre 20m²/ha et 80m²/ha). Malheureusement, nos modèles n'ont pas conduit à une telle précision en prédiction indépendante. En effet, même si certains parmi eux admettent une RMSE de moins de 10% sur(modèle 2 et modèle 4 du Tableau 4) sur les données de calibration, cette erreur explose dans la plupart des cas en validation indépendante. Si un test de Chow est mis en œuvre sur les sites deux à deux, l'hypothèse nulle selon laquelle les coefficients des modèles sont pareillement estimés sur les observations des deux sites, est rejetée au seuil de 5% pour tous les modèles et pour toutes les combinaisons de sites possibles. Cela montre qu'il y a des différences importantes entre les caractéristiques respectives des sites d'étude. Cela n'est pas surprenant car la densité des points lidar à hauteur de 1. 30 m, supposé liée à la surface terrière, dépend du type de peuplement (TSF, Taillis, Futaie régulière…), de sa sylviculture (densité, sous étage, placettes ouvertes ou en régénération…). Or le LiDAR ne fait pas de distinction entre les tiges recensables ou pas (diamètre supérieur à 7.5cm). Ce qui fait qu'il est difficile d'attribuer une surface terrière à une densité de

points lidar par exemple, d'autant plus que cette densité dépend ,par ailleurs, de la présence ou non de feuilles. Il faut donc imaginer d'autres métriques permettant de mieux traduire la diversité des conditions sylvicoles étudiées. Par exemple, on sait que le nombre de tiges à l'hectare influence G. Comme cette variable possède un caractère spatial marqué, il serait donc judicieux de chercher à tirer profit d'indices spatiaux.

Une avenue à considérer pour améliorer l'estimation du G serait d'inclure des indices de textures issue du LiDAR (ou voire même de photos aériennes). Par exemple Couteron [32] utilise des photographies aérienne de la forêt tropicale pour dégager des indices de texture du couvert végétal en utilisant des transformés de fourrier. Ce genre d'approche permettrait d'ajouter des variables différentes de celles basées uniquement sur la distribution en hauteur des points lidar. L'utilisation d'indices d'agrégation spatiale tels que ceux proposés par Riplay [33], qui explorent la dimension horizontale des peuplements, ou l'utilisation de voxels (Popescus [16]), qui pourraient permettre de calculer des textures en 3 dimensions, représentent certes des pistes à explorer. Les approches développées pour identifier les arbres à partir des nuages lidar (Monnet)permettraient également d'ajouter une information spatiale aux métriques utilisés. Évidement la réalisation de tels indices et l'évaluation de leur contribution à l'amélioration des modèles d'estimation de G dépassent largement le cadre de ce travail, mais reste des avenues de recherche chargées d'espoir pour le développement robuste et opérationnel de l'outil LiDAR en foresterie.

6. Conclusion :

Cette étude a permis de démontrer que l'estimation opérationnelle de la hauteur dominante (Ho) à l'échelle d'un massif forestier est possible dès maintenant, en adaptant un modèle développé par Martins aux conditions des forêts résineuses. Cet ajustement devrait permettre d'améliorer la précision des estimations, qui sont de toutes façons déjà acceptables. En fonction du type de forêt (résineuse/feuillue) une équation devra être adaptée, rendant ainsi la méthode d'estimation de Ho plus précise et largement accessible.

Par contre, pour la surface terrière (G), il n y a pas encore de modèle capable d'estimer ce paramètre avec une précision qui soit acceptable pour le forestier. Il faut reconnaître que cette tache est loin d'être facile et devra faire l'objet d'un travail de recherche et de développement approfondi qui devrait explorer d'autres pistes telles que l'utilisation des textures(2D) ou des voxels(3D) ou en les combinant avec l'approche des métriques liés à la hauteur.

7. Annexes

ANNEXE A. Le Random Forest : méthode non-paramétrique

A.1: Définition de Random Forest :

L'idée principale dans le ré-échantillonnage est de faire la moyenne de beaucoup d'erreur dans le but d'obtenir un modèle sans biais et de minimiser la variance. Les arbres constituent un candidat idéal dans le cadre du ré-échantillonnage dans la mesure où ils prennent en compte les interactions complexes existantes dans la structure des données et sont capables de fournir un biais faible s'ils sont efficacement utilisés. [26]. Puisque les erreurs générées sont très bruitées, les arbres sont avantagés par le calcul de la moyenne de ces erreurs. En plus, puisque dans le ré-échantillonnage les erreurs sont identiquement distribuées donc la l'espérance de **B** arbres est équivalente à l'espérance de chacun de ces arbres. Cela veux dire que le biais dans l'ensemble des arbres est le même que celui dans chaque arbre et qu'on ne puisse espérer une amélioration de la précision que par réduction de la variance. La moyenne de B variables aléatoires identiquement distribuées, de variance σ^2 , admet une variance de $\frac{1}{B}\sigma^2$. Mais si ces variables aléatoire sont seulement identiquement distribuées (et pas forcément indépendant, ce qui est notre cas) alors cette variance est évalué à $\rho\sigma^2 + \frac{1-\rho}{B}\sigma^2$ [26], avec ρ la corrélation positive entre les arbres (donc entre les variables).

On remarque que plus **B** (nombre d'échantillons bootstrap) augmente plus le second terme décroit. Mais cette diminution est compensée en partie par l'augmentation de la corrélation entre les arbres générés. Au final ce mécanisme limite le bénéfice de la variance et par conséquence celle du bruit. L'idée du Random Forest est d'améliorer la réduction de la variance en réduisant la corrélation entre les arbres tout en évitant de trop augmenter la variance. Ceci est possible grâce au processus de sélection aléatoire des variables explicatives (voir ***Algorithme 1***).

En effet, lors de la construction d'un échantillon bootstrap : avant de générer chaque nœud, un nombre $m \leq p$ de variables explicatives est choisi au hasard parmi les p candidats. Dans la pratique, m est choisi égal à √p ou égal au moins à 1.

Algorithme 1 : Random Forest pour la régression [26]

1.Pour b =1 jusqu'à B

(a) Construire un échantillon bootstrap Z^* de taille N à partir des données

(b) Générer une forêt aléatoire T_b avec les données de l'échantillon bootstrap en répétant les étapes suivantes pour chaque nœud terminal jusqu'à atteindre le

i. Sélectionner au hasard m variables parmi les p disponibles

ii. Prendre la meilleure variable

iii. Diviser le nœud obtenu en deux nœuds fils

2. Récupérer l'ensemble des arbres $\{T_b\}_1^B$.

Pour effectuer une prédiction à un nouveau point x par régression :

Valeur prédite de x : $\hat{f}_{rf}^B = \frac{1}{B}\sum_{b=1}^{B} T_b\,(x)$

Intuitivement, réduire m permet de réduire la corrélation entre les arbres et par conséquent de réduire la variance à travers la formule $\rho\sigma^2 + \frac{1-\rho}{B}\sigma^2$, avec ρ la corrélation entre les arbres aléatoires.

A.2: L'importance d'une variable

Evaluer l'importance d'une variable est un exercice difficile non seulement pour le classement des variables avant une stepwise mais aussi pour l'interprétation et le sens qu'on puisse donner à ces variables par rapport aux données étudiées.

L'importance de la variable X^j liée à la forêt aléatoire peut se définir de la façon suivante. Pour chaque arbre a de la forêt, on considère l'échantillon OOB_a associé. En effet, dans l'algorithme 1 avant de construire l'échantillon bootstrap, une partie des observations est retirée de l'échantillon servant à construire l'arbre en question. Ces observations qui constituent l'échantillon out of bag OOB_a (en dehors du sac) serviront à calculer l'erreur de prédiction commises par l'arbre concerné. On note $errOOB_a$ l'erreur quadratique moyenne(MSE) de l'arbre a sur l'échantillon de validationOOB_a. Ensuite, les valeurs de X^j dans OOB_a sont permutées aléatoirement pour obtenir un échantillon perturbé noté $\widetilde{OOB_a^j}$ et calculer$\widetilde{errOOB_a^j}$, l'erreur de prédiction de l'arbre **a** sur l'échantillon perturbé. L'importance de la variable X^j est donnée par :

$VI(X^j) = \frac{1}{ntree}\sum_a(\widetilde{errOOB_a^j} - errOOB_a)$, sachant que la somme se fait sur tous les arbres et que $ntree$ est le nombre d'arbres dans la forêt aléatoire (Random Forest).

ANNEXE B. Dictionnaires des variables

Nom de la variable	Description	classification1	classification2	classification3	classification4
hmv5	Moyenne des hauteur des 5 maxima locaux	hauteur	espace	pasdeseuil	all
hmvPredom	Moyenne des hauteur des maxima locaux impairs 1,3,5				
hp5a	hauteur du percentille 5% sans aucun filtre				
hp5v	hauteur du percentille 5% des points végétation				
hp51m	hauteur du percentille 5% si on considère les points de hauteur sup à 1m				

hp52m	hauteur du percentille 5% si on considère tous les points de hauteur sup à 2m				
hp5af	hauteur du percentille 5% de tous les firsts				
hp5al	hauteur du percentille 5% de tous les lasts				
hp5afl	hauteur du percentille 5% des first et last réunis				
hp5vf	hauteur du percentille 5% des firsts de type végétation				
hp5vl	hauteur du percentille 5% des lasts de type végétation				
hp5f1m	hauteur du percentille 5% si on considère tous les first de hauteur sup à 1m				
hp5l1m	hauteur du percentille 5% si on considère tous les lasts de hauteur sup à 1m				
hp5f2m	hauteur du percentille 5% si on considère tous les first de hauteur sup à 2m				
hp5l2m	hauteur du percentille 5% si on considère tous les lasts de hauteur sup à 2m	hauteur	espace	pasdeseuil	all
hp10a	de manière analogue à hp5 %	hauteur	espace	pasdeseuil	all
hp10v	de manière analogue à hp5 %	hauteur	percentil	pasdeseuil	all
hp101m	de manière analogue à hp5 %	hauteur	percentil	seuilveget	all
hp102m	de manière analogue à hp5 %	hauteur	percentil	seuil1m	all
hp10af	de manière analogue à hp5 %	hauteur	percentil	seuil2m	all
hp10al	de manière analogue à hp5 %	hauteur	percentil	pasdeseuil	first
hp10afl	de manière analogue à hp5 %	hauteur	percentil	pasdeseuil	last
hp10vf	de manière analogue à hp5 %	hauteur	percentil	pasdeseuil	first_last
hp10vl	de manière analogue à hp5 %	hauteur	percentil	seuilveget	first
hp10f1m	de manière analogue à hp5 %	hauteur	percentil	seuilveget	last
hp10l1m	de manière analogue à hp5 %	hauteur	percentil	seuil1m	first
hp10f2m	de manière analogue à hp5 %	hauteur	percentil	seuil1m	last
hp10l2m	de manière analogue à hp5 %	hauteur	percentil	seuil2m	first
hp20a	de manière analogue à hp5 %	hauteur	percentil	seuil2m	last
hp20v	de manière analogue à hp5 %	hauteur	percentil	pasdeseuil	all
hp201m	de manière analogue à hp5 %	hauteur	percentil	seuilveget	all
hp202m	de manière analogue à hp5 %	hauteur	percentil	seuil1m	all
hp20af	de manière analogue à hp5 %	hauteur	percentil	seuil2m	all
hp20al	de manière analogue à hp5 %	hauteur	percentil	pasdeseuil	first
hp20afl	de manière analogue à hp5 %	hauteur	percentil	pasdeseuil	last
hp20vf	de manière analogue à hp5 %	hauteur	percentil	pasdeseuil	first_last
hp20vl	de manière analogue à hp5 %	hauteur	percentil	seuilveget	first
hp20f1m	de manière analogue à hp5 %	hauteur	percentil	seuilveget	last
hp20l1m	de manière analogue à hp5 %	hauteur	percentil	seuil1m	first
hp20f2m	de manière analogue à hp5 %	hauteur	percentil	seuil1m	last
hp20l2m	de manière analogue à hp5 %	hauteur	percentil	seuil2m	first
hp25a	de manière analogue à hp5 %	hauteur	percentil	seuil2m	last
hp25v	de manière analogue à hp5 %	hauteur	percentil	pasdeseuil	all
hp251m	de manière analogue à hp5 %	hauteur	percentil	seuilveget	all
hp252m	de manière analogue à hp5 %	hauteur	percentil	seuil1m	all
hp25af	de manière analogue à hp5 %	hauteur	percentil	seuil2m	all
hp25al	de manière analogue à hp5 %	hauteur	percentil	pasdeseuil	first
hp25afl	de manière analogue à hp5 %	hauteur	percentil	pasdeseuil	last
hp25vf	de manière analogue à hp5 %	hauteur	percentil	pasdeseuil	first_last
hp25vl	de manière analogue à hp5 %	hauteur	percentil	seuilveget	first
hp25f1m	de manière analogue à hp5 %	hauteur	percentil	seuilveget	last
hp25l1m	de manière analogue à hp5 %	hauteur	percentil	seuil1m	first
hp25f2m	de manière analogue à hp5 %	hauteur	percentil	seuil1m	last
hp25l2m	de manière analogue à hp5 %	hauteur	percentil	seuil2m	first
hp30a	de manière analogue à hp5 %	hauteur	percentil	seuil2m	last
hp30v	de manière analogue à hp5 %	hauteur	percentil	pasdeseuil	all
hp301m	de manière analogue à hp5 %	hauteur	percentil	seuilveget	all

hp302m	de manière analogue à hp5 %	hauteur	percentil	seuil1m	all
hp30af	de manière analogue à hp5 %	hauteur	percentil	seuil2m	all
hp30al	de manière analogue à hp5 %	hauteur	percentil	pasdeseuil	first
hp30afl	de manière analogue à hp5 %	hauteur	percentil	pasdeseuil	last
hp30vf	de manière analogue à hp5 %	hauteur	percentil	pasdeseuil	first_last
hp30vl	de manière analogue à hp5 %	hauteur	percentil	seuilveget	first
hp30f1m	de manière analogue à hp5 %	hauteur	percentil	seuilveget	last
hp30l1m	de manière analogue à hp5 %	hauteur	percentil	seuil1m	first
hp30f2m	de manière analogue à hp5 %	hauteur	percentil	seuil1m	last
hp30l2m	de manière analogue à hp5 %	hauteur	percentil	seuil2m	first
hp40a	de manière analogue à hp5 %	hauteur	percentil	seuil2m	last
hp40v	de manière analogue à hp5 %	hauteur	percentil	pasdeseuil	all
hp401m	de manière analogue à hp5 %	hauteur	percentil	seuilveget	all
hp402m	de manière analogue à hp5 %	hauteur	percentil	seuil1m	all
hp40af	de manière analogue à hp5 %	hauteur	percentil	seuil2m	all
hp40al	de manière analogue à hp5 %	hauteur	percentil	pasdeseuil	first
hp40afl	de manière analogue à hp5 %	hauteur	percentil	pasdeseuil	last
hp40vf	de manière analogue à hp5 %	hauteur	percentil	pasdeseuil	first_last
hp40vl	de manière analogue à hp5 %	hauteur	percentil	seuilveget	first
hp40f1m	de manière analogue à hp5 %	hauteur	percentil	seuilveget	last
hp40l1m	de manière analogue à hp5 %	hauteur	percentil	seuil1m	first
hp40f2m	de manière analogue à hp5 %	hauteur	percentil	seuil1m	last
hp40l2m	de manière analogue à hp5 %	hauteur	percentil	seuil2m	first
hp50a	de manière analogue à hp5 %	hauteur	percentil	seuil2m	last
hp50v	de manière analogue à hp5 %	hauteur	percentil	pasdeseuil	all
hp501m	de manière analogue à hp5 %	hauteur	percentil	seuilveget	all
hp502m	de manière analogue à hp5 %	hauteur	percentil	seuil1m	all
hp50af	de manière analogue à hp5 %	hauteur	percentil	seuil2m	all
hp50al	de manière analogue à hp5 %	hauteur	percentil	pasdeseuil	first
hp50afl	de manière analogue à hp5 %	hauteur	percentil	pasdeseuil	last
hp50vf	de manière analogue à hp5 %	hauteur	percentil	pasdeseuil	first_last
hp50vl	de manière analogue à hp5 %	hauteur	percentil	seuilveget	first
hp50f1m	de manière analogue à hp5 %	hauteur	percentil	seuilveget	last
hp50l1m	de manière analogue à hp5 %	hauteur	percentil	seuil1m	first
hp50f2m	de manière analogue à hp5 %	hauteur	percentil	seuil1m	last
hp50l2m	de manière analogue à hp5 %	hauteur	percentil	seuil2m	first
hp60a	de manière analogue à hp5 %	hauteur	percentil	seuil2m	last
hp60v	de manière analogue à hp5 %	hauteur	percentil	pasdeseuil	all
hp601m	de manière analogue à hp5 %	hauteur	percentil	seuilveget	all
hp602m	de manière analogue à hp5 %	hauteur	percentil	seuil1m	all
hp60af	de manière analogue à hp5 %	hauteur	percentil	seuil2m	all
hp60al	de manière analogue à hp5 %	hauteur	percentil	pasdeseuil	first
hp60afl	de manière analogue à hp5 %	hauteur	percentil	pasdeseuil	last
hp60vf	de manière analogue à hp5 %	hauteur	percentil	pasdeseuil	first_last
hp60vl	de manière analogue à hp5 %	hauteur	percentil	seuilveget	first
hp60f1m	de manière analogue à hp5 %	hauteur	percentil	seuilveget	last
hp60l1m	de manière analogue à hp5 %	hauteur	percentil	seuil1m	first
hp60f2m	de manière analogue à hp5 %	hauteur	percentil	seuil1m	last
hp60l2m	de manière analogue à hp5 %	hauteur	percentil	seuil2m	first
hp70a	de manière analogue à hp5 %	hauteur	percentil	seuil2m	last
hp70v	de manière analogue à hp5 %	hauteur	percentil	pasdeseuil	all
hp701m	de manière analogue à hp5 %	hauteur	percentil	seuilveget	all
hp702m	de manière analogue à hp5 %	hauteur	percentil	seuil1m	all
hp70af	de manière analogue à hp5 %	hauteur	percentil	seuil2m	all
hp70al	de manière analogue à hp5 %	hauteur	percentil	pasdeseuil	first

hp70afl	de manière analogue à hp5 %	hauteur	percentil	pasdeseuil	last
hp70vf	de manière analogue à hp5 %	hauteur	percentil	pasdeseuil	first_last
hp70vl	de manière analogue à hp5 %	hauteur	percentil	seuilveget	first
hp70f1m	de manière analogue à hp5 %	hauteur	percentil	seuilveget	last
hp70l1m	de manière analogue à hp5 %	hauteur	percentil	seuil1m	first
hp70f2m	de manière analogue à hp5 %	hauteur	percentil	seuil1m	last
hp70l2m	de manière analogue à hp5 %	hauteur	percentil	seuil2m	first
hp75a	de manière analogue à hp5 %	hauteur	percentil	seuil2m	last
hp75v	de manière analogue à hp5 %	hauteur	percentil	pasdeseuil	all
hp751m	de manière analogue à hp5 %	hauteur	percentil	seuilveget	all
hp752m	de manière analogue à hp5 %	hauteur	percentil	seuil1m	all
hp75af	de manière analogue à hp5 %	hauteur	percentil	seuil2m	all
hp75al	de manière analogue à hp5 %	hauteur	percentil	pasdeseuil	first
hp75afl	de manière analogue à hp5 %	hauteur	percentil	pasdeseuil	last
hp75vf	de manière analogue à hp5 %	hauteur	percentil	pasdeseuil	first_last
hp75vl	de manière analogue à hp5 %	hauteur	percentil	seuilveget	first
hp75f1m	de manière analogue à hp5 %	hauteur	percentil	seuilveget	last
hp75l1m	de manière analogue à hp5 %	hauteur	percentil	seuil1m	first
hp75f2m	de manière analogue à hp5 %	hauteur	percentil	seuil1m	last
hp75l2m	de manière analogue à hp5 %	hauteur	percentil	seuil2m	first
hp80a	de manière analogue à hp5 %	hauteur	percentil	seuil2m	last
hp80v	de manière analogue à hp5 %	hauteur	percentil	pasdeseuil	all
hp801m	de manière analogue à hp5 %	hauteur	percentil	seuilveget	all
hp802m	de manière analogue à hp5 %	hauteur	percentil	seuil1m	all
hp80af	de manière analogue à hp5 %	hauteur	percentil	seuil2m	all
hp80al	de manière analogue à hp5 %	hauteur	percentil	pasdeseuil	first
hp80afl	de manière analogue à hp5 %	hauteur	percentil	pasdeseuil	last
hp80vf	de manière analogue à hp5 %	hauteur	percentil	pasdeseuil	first_last
hp80vl	de manière analogue à hp5 %	hauteur	percentil	seuilveget	first
hp80f1m	de manière analogue à hp5 %	hauteur	percentil	seuilveget	last
hp80l1m	de manière analogue à hp5 %	hauteur	percentil	seuil1m	first
hp80f2m	de manière analogue à hp5 %	hauteur	percentil	seuil1m	last
hp80l2m	de manière analogue à hp5 %	hauteur	percentil	seuil2m	first
hp90a	de manière analogue à hp5 %	hauteur	percentil	seuil2m	last
hp90v	de manière analogue à hp5 %	hauteur	percentil	pasdeseuil	all
hp901m	de manière analogue à hp5 %	hauteur	percentil	seuilveget	all
hp902m	de manière analogue à hp5 %	hauteur	percentil	seuil1m	all
hp90af	de manière analogue à hp5 %	hauteur	percentil	seuil2m	all
hp90al	de manière analogue à hp5 %	hauteur	percentil	pasdeseuil	first
hp90afl	de manière analogue à hp5 %	hauteur	percentil	pasdeseuil	last
hp90vf	de manière analogue à hp5 %	hauteur	percentil	pasdeseuil	first_last
hp90vl	de manière analogue à hp5 %	hauteur	percentil	seuilveget	first
hp90f1m	de manière analogue à hp5 %	hauteur	percentil	seuilveget	last
hp90l1m	de manière analogue à hp5 %	hauteur	percentil	seuil1m	first
hp90f2m	de manière analogue à hp5 %	hauteur	percentil	seuil1m	last
hp90l2m	de manière analogue à hp5 %	hauteur	percentil	seuil2m	first
hp95a	de manière analogue à hp5 %	hauteur	percentil	seuil2m	last
hp95v	de manière analogue à hp5 %	hauteur	percentil	pasdeseuil	all
hp951m	de manière analogue à hp5 %	hauteur	percentil	seuilveget	all
hp952m	de manière analogue à hp5 %	hauteur	percentil	seuil1m	all
hp95af	de manière analogue à hp5 %	hauteur	percentil	seuil2m	all
hp95al	de manière analogue à hp5 %	hauteur	percentil	pasdeseuil	first
hp95afl	de manière analogue à hp5 %	hauteur	percentil	pasdeseuil	last
hp95vf	de manière analogue à hp5 %	hauteur	percentil	pasdeseuil	first_last
hp95vl	de manière analogue à hp5 %	hauteur	percentil	seuilveget	first

hp95f1m	de manière analogue à hp5 %	hauteur	percentil	seuilveget	last
hp95l1m	de manière analogue à hp5 %	hauteur	percentil	seuil1m	first
hp95f2m	de manière analogue à hp5 %	hauteur	percentil	seuil1m	last
hp95l2m	de manière analogue à hp5 %	hauteur	percentil	seuil2m	first
hp99a	de manière analogue à hp5 %	hauteur	percentil	seuil2m	last
hp99v	de manière analogue à hp5 %	hauteur	percentil	pasdeseuil	all
hp991m	de manière analogue à hp5 %	hauteur	percentil	seuilveget	all
hp992m	de manière analogue à hp5 %	hauteur	percentil	seuil1m	all
hp99af	de manière analogue à hp5 %	hauteur	percentil	seuil2m	all
hp99al	de manière analogue à hp5 %	hauteur	percentil	pasdeseuil	first
hp99afl	de manière analogue à hp5 %	hauteur	percentil	pasdeseuil	last
hp99vf	de manière analogue à hp5 %	hauteur	percentil	pasdeseuil	first_last
hp99vl	de manière analogue à hp5 %	hauteur	percentil	seuilveget	first
hp99f1m	de manière analogue à hp5 %	hauteur	percentil	seuilveget	last
hp99l1m	de manière analogue à hp5 %	hauteur	percentil	seuil1m	first
hp99f2m	de manière analogue à hp5 %	hauteur	percentil	seuil1m	last
hp99l2m	de manière analogue à hp5 %	hauteur	percentil	seuil2m	first
hmean_a	moyenne des hauteurs de tous les points sans filtre	hauteur	percentil	seuil2m	last
hmean_v	moyenne des hauteurs de tous les points végétation	hauteur	percentil	pasdeseuil	all
hmean_1m	moyenne des hauteurs de tous les points à plus de 1m	hauteur	percentil	seuilveget	all
hmean_2m	moyenne des hauteurs de tous les points à plus de 2m	hauteur	percentil	seuil1m	all
hmean_af	moyenne des hauteurs de tous les first	hauteur	percentil	seuil2m	all
hmean_al	moyenne des hauteurs de tous les last	hauteur	percentil	pasdeseuil	first
hmean_afl	moyenne des hauteurs de des first et last réunis	hauteur	percentil	pasdeseuil	last
hmean_vf	moyenne des hauteurs de tous les first de type végétation	hauteur	percentil	pasdeseuil	first_last
hmean_vl	moyenne des hauteurs de tous les last de type végétation	hauteur	percentil	seuilveget	first
hmean_f1m	moyenne des hauteurs de tous les first à plus de 1m	hauteur	percentil	seuilveget	last
hmean_l1m	moyenne des hauteurs de tous les last à plus de 1m	hauteur	percentil	seuil1m	first
hmean_f2m	moyenne des hauteurs de tous les first à plus de 2m	hauteur	percentil	seuil1m	last
hmean_l2m	moyenne des hauteurs de tous les last à plus de 2m	hauteur	percentil	seuil2m	first
hcv_a	coéf de variation= 100*(ecart-type des hauteur/moyenne des hauteurs) pour tous les points sans filtre	hauteur	percentil	seuil2m	last
hcv_v	coéf de variation= 100*(ecart-type des hauteur/moyenne des hauteurs) pour tous les	hauteur	dispertion	pasdeseuil	all
hcv_1m	coéf de variation= 100*(ecart-type des hauteur/moyenne des hauteurs) pour tous les	hauteur	dispertion	seuilveget	all
hcv_2m	coéf de variation= 100*(ecart-type des hauteur/moyenne des hauteurs) pour tous les	hauteur	dispertion	seuil1m	all
hcv_af	coéf de variation= 100*(ecart-type des hauteur/moyenne des hauteurs) pour tous les	hauteur	dispertion	seuil2m	all
hcv_al	coéf de variation= 100*(ecart-type des hauteur/moyenne des hauteurs) pour tous les	hauteur	dispertion	pasdeseuil	first
hcv_afl	coéf de variation= 100*(ecart-type des hauteur/moyenne des hauteurs) pour tous les	hauteur	dispertion	pasdeseuil	last
hcv_vf	coéf de variation= 100*(ecart-type des hauteur/moyenne des hauteurs) pour tous les	hauteur	dispertion	pasdeseuil	first_last
hcv_vl	coéf de variation= 100*(ecart-type des hauteur/moyenne des hauteurs) pour tous les	hauteur	dispertion	seuilveget	first
hcv_f1m	coéf de variation= 100*(ecart-type des hauteur/moyenne des hauteurs) pour tous les	hauteur	dispertion	seuilveget	last
hcv_l1m	coéf de variation= 100*(ecart-type des hauteur/moyenne des hauteurs) pour tous les	hauteur	dispertion	seuil1m	first
hcv_f2m	coéf de variation= 100*(ecart-type des hauteur/moyenne des hauteurs) pour tous les	hauteur	dispertion	seuil1m	last
hcv_l2m	coéf de variation= 100*(ecart-type des hauteur/moyenne des hauteurs) pour tous les	hauteur	dispertion	seuil2m	first
hmax_a	hauteur max des points (est le même pour tous les filtres)	hauteur	dispertion	seuil2m	last
hmax_v	hauteur max des points	hauteur	dispertion	pasdeseuil	all
hmax_1m	hauteur max des points	hauteur	dispertion	seuilveget	all
hmax_2m	hauteur max des points	hauteur	dispertion	seuil1m	all
hmax_af	hauteur max des points	hauteur	dispertion	seuil2m	all
hmax_al	hauteur max des points	hauteur	dispertion	pasdeseuil	first

hmax_afl	hauteur max des points	hauteur	dispertion	pasdeseuil	last
hmax_vf	hauteur max des points	hauteur	dispertion	pasdeseuil	first_last
hmax_vl	hauteur max des points	hauteur	dispertion	seuilveget	first
hmax_f1m	hauteur max des points	hauteur	dispertion	seuilveget	last
hmax_l1m	hauteur max des points	hauteur	dispertion	seuil1m	first
hmax_f2m	hauteur max des points	hauteur	dispertion	seuil1m	last
hmax_l2m		hauteur	dispertion	seuil2m	first
hmed_a	hauteur médiane de tous les points sans filtre	hauteur	dispertion	seuil2m	last
hmed_v	hauteur médiane	hauteur	dispertion	pasdeseuil	all
hmed_1m	hauteur médiane	hauteur	dispertion	seuilveget	all
hmed_2m	hauteur médiane	hauteur	dispertion	seuil1m	all
hmed_af	hauteur médiane	hauteur	dispertion	seuil2m	all
hmed_al	hauteur médiane	hauteur	dispertion	pasdeseuil	first
hmed_afl	hauteur médiane	hauteur	dispertion	pasdeseuil	last
hmed_vf	hauteur médiane	hauteur	dispertion	pasdeseuil	first_last
hmed_vl	hauteur médiane	hauteur	dispertion	seuilveget	first
hmed_f1m	hauteur médiane	hauteur	dispertion	seuilveget	last
hmed_l1m	hauteur médiane	hauteur	dispertion	seuil1m	first
hmed_f2m	hauteur médiane	hauteur	dispertion	seuil1m	last
hmed_l2m	hauteur médiane	hauteur	dispertion	seuil2m	first
habs_a	absolue deviation=moyenne |hauteur-hmean|	hauteur	dispertion	seuil2m	last
habs_v	absolue deviation=moyenne |hauteur-hmean|	hauteur	dispertion	pasdeseuil	all
habs_1m	absolue deviation=moyenne |hauteur-hmean|	hauteur	dispertion	seuilveget	all
habs_2m	absolue deviation=moyenne |hauteur-hmean|	hauteur	dispertion	seuil1m	all
habs_af	absolue deviation=moyenne |hauteur-hmean|	hauteur	dispertion	seuil2m	all
habs_al	absolue deviation=moyenne |hauteur-hmean|	hauteur	dispertion	pasdeseuil	first
habs_afl	absolue deviation=moyenne |hauteur-hmean|	hauteur	dispertion	pasdeseuil	last
habs_vf	absolue deviation=moyenne |hauteur-hmean|	hauteur	dispertion	pasdeseuil	first_last
habs_vl	absolue deviation=moyenne |hauteur-hmean|	hauteur	dispertion	seuilveget	first
habs_f1m	absolue deviation=moyenne |hauteur-hmean|	hauteur	dispertion	seuilveget	last
habs_l1m	absolue deviation=moyenne |hauteur-hmean|	hauteur	dispertion	seuil1m	first
habs_f2m	absolue deviation=moyenne |hauteur-hmean|	hauteur	dispertion	seuil1m	last
habs_l2m	absolue deviation=moyenne |hauteur-hmean|	hauteur	dispertion	seuil2m	first
PRtot	taux de pénétration total =nbre de points de h<1m /N avec N=nbre de points total	hauteur	dispertion	seuil2m	last
PRul	taux de pénétration à travers la couche supérieur =nbre de points de h<hp80 /N avec N=nbre de points total	hauteur	dispertion	pasdeseuil	all
PRil	taux de pénétration à travers la couche intermédiaire =nbre de points de h<hp50 /nbre de points de h<hp80	hauteur	dispertion	seuilveget	all
PRll	taux de pénétration à travers la couche inferieur =nbre de points de h<1m /nbre de points de h<hp50	hauteur	dispertion	seuil1m	all
PRtot_f	taux de pénétration total =nbre de points de h<1m /N avec N=nbre de points total	hauteur	dispertion	seuil2m	all
PRul_f	taux de pénétration supérieur =nbre de points de h<hp80 /N avec N=nbre de points total	hauteur	dispertion	pasdeseuil	first
PRil_f	taux de pénétration intermédiaire =nbre de points de h<hp50 /nbre de points de h<hp80	hauteur	dispertion	pasdeseuil	last
PRll_f	taux de pénétration total =nbre de points de h<1m /nbre de points de h<hp50	hauteur	dispertion	pasdeseuil	first_last
PRtot_l	taux de pénétration total =nbre de points de h<1m /N avec N=nbre de points total	hauteur	dispertion	seuilveget	first
PRul_l	taux de pénétration supérieur =nbre de points de h<hp80 /N avec N=nbre de points total	hauteur	dispertion	seuilveget	last
PRil_l	taux de pénétration intermédiaire =nbre de points de h<hp50 /nbre de points de h<hp80	hauteur	dispertion	seuil1m	first
PRll_l	taux de pénétration total =nbre de points de h<1m /nbre de points de h<hp50	hauteur	dispertion	seuil1m	last
PRtot_on	taux de pénétration total =nbre de points de h<1m /N avec N=nbre de points total	hauteur	dispertion	seuil2m	first
PRul_on	taux de pénétration supérieur =nbre de points de h<hp80 /N avec N=nbre de points total	hauteur	dispertion	seuil2m	last
PRil_on	taux de pénétration intermédiaire =nbre de points de h<hp50 /nbre de points de h<hp80	penetration	sol	tranchevariable	all

PRll_on	taux de pénétration total =nbre de points de h<1m /nbre de points de h<hp50	penetration	coucheSup	tranchevariable	all
densite	densité totale sur la placette= nombre de points N de la placette/ surface de la placette	penetration	coucheInt	tranchevariable	all
d0a	densité de la tranche n° 0 (de 2m à 2m+(Hp95-2m)/10) avec Htranche0=2m+(Hp95-2m)/10	penetration	coucheBas	tranchevariable	all
d1a	densité de la tranche n° 1 (de Htranche0 à Htranche0+(Hp95-2m)/10	penetration	sol	tranchevariable	first
d2a	densité de la tranche n° 1 (de Htranche0 à Htranche0+(Hp95-2m)/10	penetration	coucheSup	tranchevariable	first
d3a	densité de la tranche n° 1 (de Htranche0 à Htranche0+(Hp95-2m)/10	penetration	coucheInt	tranchevariable	first
d4a	densité de la tranche n° 1 (de Htranche0 à Htranche0+(Hp95-2m)/10	penetration	coucheBas	tranchevariable	first
d5a	densité de la tranche n° 1 (de Htranche0 à Htranche0+(Hp95-2m)/10	penetration	sol	tranchevariable	last
d6a	densité de la tranche n° 1 (de Htranche0 à Htranche0+(Hp95-2m)/10	penetration	coucheSup	tranchevariable	last
d7a	densité de la tranche n° 1 (de Htranche0 à Htranche0+(Hp95-2m)/10	penetration	coucheInt	tranchevariable	last
d8a	densité de la tranche n° 1 (de Htranche0 à Htranche0+(Hp95-2m)/10	penetration	coucheBas	tranchevariable	last
d9a	densité de la tranche n° 1 (de Htranche0 à Htranche0+(Hp95-2m)/10	penetration	sol	tranchevariable	only
d0f	densité de la tranche n° 1 (de Htranche0 à Htranche0+(Hp95-2m)/10	penetration	coucheSup	tranchevariable	only
d1f	densité de la tranche n° 1 (de Htranche0 à Htranche0+(Hp95-2m)/10	penetration	coucheInt	tranchevariable	only
d2f	densité de la tranche n° 1 (de Htranche0 à Htranche0+(Hp95-2m)/10	penetration	coucheBas	tranchevariable	only
d3f	densité de la tranche n° 1 (de Htranche0 à Htranche0+(Hp95-2m)/10	densite	nonCumule	tranchefixe	all
d4f	densité de la tranche n° 1 (de Htranche0 à Htranche0+(Hp95-2m)/10	densite	nonCumule	tranchevariable	all
d5f	densité de la tranche n° 1 (de Htranche0 à Htranche0+(Hp95-2m)/10	densite	nonCumule	tranchevariable	all
d6f	densité de la tranche n° 1 (de Htranche0 à Htranche0+(Hp95-2m)/10	densite	nonCumule	tranchevariable	all
d7f	densité de la tranche n° 1 (de Htranche0 à Htranche0+(Hp95-2m)/10	densite	nonCumule	tranchevariable	all
d8f	densité de la tranche n° 1 (de Htranche0 à Htranche0+(Hp95-2m)/10	densite	nonCumule	tranchevariable	all
d9f	densité de la tranche n° 1 (de Htranche0 à Htranche0+(Hp95-2m)/10	densite	nonCumule	tranchevariable	all
d0l	densité de la tranche n° 1 (de Htranche0 à Htranche0+(Hp95-2m)/10	densite	nonCumule	tranchevariable	all
d1l	densité de la tranche n° 1 (de Htranche0 à Htranche0+(Hp95-2m)/10	densite	nonCumule	tranchevariable	all
d2l	densité de la tranche n° 1 (de Htranche0 à Htranche0+(Hp95-2m)/10	densite	nonCumule	tranchevariable	all
d3l	densité de la tranche n° 1 (de Htranche0 à Htranche0+(Hp95-2m)/10	densite	nonCumule	tranchevariable	all
d4l	densité de la tranche n° 1 (de Htranche0 à Htranche0+(Hp95-2m)/10	densite	nonCumule	tranchevariable	first
d5l	densité de la tranche n° 1 (de Htranche0 à Htranche0+(Hp95-2m)/10	densite	nonCumule	tranchevariable	first
d6l	densité de la tranche n° 1 (de Htranche0 à Htranche0+(Hp95-2m)/10	densite	nonCumule	tranchevariable	first
d7l	densité de la tranche n° 1 (de Htranche0 à Htranche0+(Hp95-2m)/10	densite	nonCumule	tranchevariable	first
d8l	densité de la tranche n° 1 (de Htranche0 à Htranche0+(Hp95-2m)/10	densite	nonCumule	tranchevariable	first
d9l	densité de la tranche n° 1 (de Htranche0 à Htranche0+(Hp95-2m)/10	densite	nonCumule	tranchevariable	first
D0	densité cumulée des tranches: D0=d0	densite	nonCumule	tranchevariable	first
D1	densité cumulée des tranches: D1=d0+d1=d1+D0	densite	nonCumule	tranchevariable	first
D2	densité cumulée des tranches: D2	densite	nonCumule	tranchevariable	first
D3	densité cumulée des tranches: D3	densite	nonCumule	tranchevariable	first
D4	densité cumulée des tranches: D4	densite	nonCumule	tranchevariable	last
D5	densité cumulée des tranches: D5	densite	nonCumule	tranchevariable	last
D6	densité cumulée des tranches: D6	densite	nonCumule	tranchevariable	last
D7	densité cumulée des tranches: D7	densite	nonCumule	tranchevariable	last

D8	densité cumulée des tranches: D8	densite	nonCumule	tranchevariable	last
D9	densité cumulée des tranches: D9=d9+D8	densite	nonCumule	tranchevariable	last
D0f	densité cumulée des tranches pour les first	densite	nonCumule	tranchevariable	last
D1f	densité cumulée des tranches pour les first	densite	nonCumule	tranchevariable	last
D2f	densité cumulée des tranches pour les first	densite	nonCumule	tranchevariable	last
D3f	densité cumulée des tranches pour les first	densite	nonCumule	tranchevariable	last
D4f	densité cumulée des tranches pour les first	densite	cumul	tranchevariable	all
D5f	densité cumulée des tranches pour les first	densite	cumul	tranchevariable	all
D6f	densité cumulée des tranches pour les first	densite	cumul	tranchevariable	all
D7f	densité cumulée des tranches pour les first	densite	cumul	tranchevariable	all
D8f	densité cumulée des tranches pour les first	densite	cumul	tranchevariable	all
D9f	densité cumulée des tranches pour les first	densite	cumul	tranchevariable	all
D0l	densité cumulée des tranches pour les last	densite	cumul	tranchevariable	all
D1l	densité cumulée des tranches pour les last	densite	cumul	tranchevariable	all
D2l	densité cumulée des tranches pour les last	densite	cumul	tranchevariable	all
D3l	densité cumulée des tranches pour les last	densite	cumul	tranchevariable	all
D4l	densité cumulée des tranches pour les last	densite	cumul	tranchevariable	first
D5l	densité cumulée des tranches pour les last	densite	cumul	tranchevariable	first
D6l	densité cumulée des tranches pour les last	densite	cumul	tranchevariable	first
D7l	densité cumulée des tranches pour les last	densite	cumul	tranchevariable	first
D8l	densité cumulée des tranches pour les last	densite	cumul	tranchevariable	first
D9l	densité cumulée des tranches pour les last	densite	cumul	tranchevariable	first
D_dbh	densité de points dans la tranche 0 2m =nbre de points entre 0 et 2m/ nbre de points total	densite	cumul	tranchevariable	first
D_2_6	densité de points dans la tranche 2-6m	densite	cumul	tranchevariable	first
D_2_12	densité de points dans la tranche 2-12m	densite	cumul	tranchevariable	first
D_2_24	densité de points dans la tranche 2-24m	densite	cumul	tranchevariable	first
D_6_12	densité de points dans la tranche 6-12m	densite	cumul	tranchevariable	last
D_6_24	densité de points dans la tranche 6-24m	densite	cumul	tranchevariable	last
D_12_24	densité de points dans la tranche 12-24m	densite	cumul	tranchevariable	last
D_24p	densité de points dans la tranche plus de 24m	densite	cumul	tranchevariable	last

ANNEXE C. Scripts R et SAS

C.1: Exemple de programme R qui a parmi le traitement statistique des modèles de régression

```
############################## Fonctions qui concernent les analyses sur les modèles ################
library(lmtest)
PRESS<-function(x)sum((resid(x)^2/(1-hatvalues(x))^2))
## Cross validation estimate =MSPE= mean of square predicticted error
CVE<-function(x)sum((resid(x)^2/(1-hatvalues(x))^2))/length(fitted(x))
## calcul de R2 et de R2CV= R2 cross validation
R2<-function(x){ f=fitted(x)
                                1-(sum((resid(x)^2))/sum(((x$y-mean(f))^2)))
}
                                ## R2=1-(Som des carrés residuel/som des carrées tataux)
R2CV<-function(x){ yj<-(fitted(x)-hatvalues(x)*x$y)/(1-hatvalues(x))
                                    cor(x$y,yj)^2
}
RMSE=function(x)sqrt(sum(resid(x)^2/length(fitted(x)))) #
biais<-function(x){return(mean(predict(x)-x$y))}
graph_predites_vs_obs=function(x){plot(predict(x),x$y,col="red",pch=19,xlab="valeurs prédites",ylab=" Valeurs observées",main=names(x$model))
                        abline(a=0,b=1, col="blue",pch=19)}
#### somme des carrés résiduels( il se trouve aussi sur la table anova)
RSS<-function(x)sum(resid(x)^2)
analyse_lm=function(x,donnees){
  resume<-summary.lm(x,correlation=TRUE,signif.stars=TRUE)
  print("********** RESUME DE LA REGRESSION **********",quote=F)
  show(resume$call)
  print("========== Coéfficients de la régression:==========",quote=F)
  show( resume$coefficients)
  print("==========Anova :==========",quote=F)
  show(avov<-anova(x))
  print("==========Qualité de la régression :==========",quote print(paste("R2=",round(R2(x),3)," et R2 cross Validation(leave-one-out) R2CV=",round(R2CV(x),3)),quote=FALSE)
      print(paste("R2 ajusté= ",round(resume$adj.r.squared,3)," RMSE=",round(RMSE(x),3), "soit ",round(100*RMSE(x)/mean(x$y),3),"% d'erreur "),quote=FALSE)
      print(paste("PRESS=",round(PRESS(x),2),"RSS=",round(RSS(x),2)," si PRESS>RSS alors ya des observations mal prédites"),quote=FALSE)
      print(paste("RootMeanPRESS=",round(sqrt(PRESS(x)/length(x$y)),2)," est l'erreur de prédicion à comparer à la RMSE qui est l'erreur de l'ajustement"),quote=F)

}
diagnostic_variables=function(x){
  print("********** ANALYSE SUR LES CO-VARIABLES **********",quote=F)
  resume<-summary.lm(x,correlation=TRUE,signif.stars=TRUE)
  print("==========Matrice de corrélation ==========",quote=F)
  show(resume$correlation)
  print(" Les valeurs proppres de la matrice: si une des valeurs est inférieure à 0.001 , il y a multicolinéarité",quote=F)
  show(eigen(resume$correlation)$values )
  print(" Test de colinéarité des co-variables :", quote=F)
  show(resume$aliased)
  print(" Variance Inflation Factors :VIF = 1 / Tolérance =1/(1-R²). Des valeurs élevées de VIF indiquent donc la présence de multicolinéarité.", quote=F)
  show(vif(x))
```

```
  }

  cross_val=function(x,donnees,M){
        ## c'est l'erreur moyen de validation croisée= SS/nbre
        ##de paquets dans cross validation c'est a comparer avec l'erreur de prediction
        library(DAAG)
        print("Validation croisée")
        print("On calcule la RMSE moyenne d'une suite de cross validation et son ecart type ")
        rmse=vector()
        i=1
        for(seed in c(10,20,29,100,150,500,1000))
        {rmse[i]=round(sqrt(cv.lm(donnees,x,m=M,seed=seed,plotit=FALSE,
printit=FALSE)["ss"]/length(fitted(x))),2)
        i=i+1
        }
        print(paste(" moyenne des RMSE =", round(mean(rmse),3), " écart type=  " ,round(sd(rmse),3)),quote=F)
  }
  print(paste("********************* Validation coisée répétée ",nbrepet, "fois
*****************************"),quote=F)
  ##print(Tot,quote=F)
  print(paste("RMSE moyenne cross validation: ",round(mean(Tot["RMSE"]),2), " Ecart-type"
,round(sd(Tot["RMSE"]),2)),quote=F)
  print(paste(" R2 CV moyen:" ,round(mean(Tot["Rcv"]),2)," Ecart-type" ,round(sd(Tot["Rcv"]),2)),quote=F)


  }

  ## destection des observations influtentes ou aberantes: prend en entrées le nom de la regression , les données qui ont
permis de l'avoir
  diagnostic_observations=function(x,donnees,alpha){
                                    library(car)
                                    library(DAAG)
                                    n=dim(donnees)[1]
                                    p=ncol(x$x)        #ncol(x$x)=p c'est le nombre de paramètre à estimer
                                    PLAC=donnees$PLAC
                                    #-----------------------------------------------------résidus studentisés externes
                                    r_i_hat=rstudent(x)  ## résidus studentisés externes
                        table_r_i_hat=data.frame(PLAC,r_i_hat)
                                                ## t_student est la statistique de student(n-1-(p-1),1-alpha/2)
                                    table_r_i_hat$t_student=qt(1-alpha/2,(n-1-(p-1)))
                                    #-----------------------------------------------------Cook
                                    Cook_dsit=cooks.distance(x)
        ## on considère qu'une observation est influente si cook_dsit>4/(n-(p-1)) sert a la detection des influences
                                    Cook_table=data.frame(PLAC,Cook_dsit)
                                    Cook_table$seuil=4/(n-(p+1))
                                    #-----------------------------------------------------dfbetas
                                    Dfbetas_i_j=dfbetas(x)    #on considere que l'obs i pèse indument sur l'estimation de
Bj lorsque Dfbetas_i_j>2/sqrt(n). sert à l'identification de la covariable surlaquelle l'influence s'exerce
                                    Dfbeta_table=data.frame(PLAC,Dfbetas_i_j) ## Dfbetas_i_j comporte p colonnes
coresponsdant aux paramètre à estimer
                                    Dfbeta_table$seuil=2/sqrt(n)
                                    #-----------------------------------------------------Covratio
                                    Covratio_i=covratio(x)
                                    Covratio_table=data.frame(PLAC,Covratio_i) ## le covratio est le rapport entre la
variance mesurée sans l'obs i sur la variance totale (avec l'obs i incluse)

                        ## interprétation choisie: donc si covratio>1 cela veut dire que l'obs apporte de la précision et pas du
tout au cas contraire.
                                    print("****************************** DIAGNOSTIC SUR LES
OBSERVATIONS ******************************",quote=F)
```

```
                print("Observations mal prédites selon les résidus studentisés externes",quote=F)  #(
voir les annexes pour mieux le comprendre)

        print(subset(table_r_i_hat,table_r_i_hat$r_i_hat>table_r_i_hat$t_student),quote=F,digits=2)
                print("Observations mal prédites selon la distance de Cook ",quote=F)
                print("on considère qu'une observation est influente si cook_dsit>4/(n-(p-1)) sert a la
détection des influences",quote=F)

        print(subset(Cook_table,Cook_table$Cook_dsit>Cook_table$seuil),quote=F,digits=2)
                print("Localisation des covariables et des observations à l'origine des ecarts:
",quote=F)
                print("cas du dfbetas :on considere que l'obs i pèse indument sur l'estimation de Bj
lorsque Dfbetas_i_j>2/sqrt(n). sert à l'identification de la covariable surlaquelle l'influence s'exerce ",quote=F,digits=2)
                print(subset(Dfbeta_table,(Dfbeta_table[,2]>Dfbeta_table$seuil
|Dfbeta_table[,3]>Dfbeta_table$seuil|Dfbeta_table[,4]>Dfbeta_table$seuil)),quote=F,digits=2)
                print("cas du Covration :si covratio>1 cela veut dire que l'obs apporte de la précision
et pas du tout au cas contraire ",quote=F)
                print(subset(Covratio_table,Covratio_table$Covratio_i<=1),quote=F,digits=2)

                ##print(outlierTest(x,cutoff=0.05),quote=FALSE)
  }
  ##### analyse graphique des résidus : ici on choisit soit les résidus standardisé(normalisés) soit ceux studentisés.
  ## il faut noté que les graphes sont directement enregistrés dans le directory courant de R
  diagnostic_graphique=function(x,nom){
      library(car) #pour utiliser scatterplotMatrix
                png(file=paste(paste(nom,"_residus"),".png"),width = 700, height = 700)
                plot(x,main="",which=1,caption = list(" Graphe des résidus "),sub.caption="")
                dev.off()
                png(file=paste(paste(nom,"_qqplot"),".png"),width = 700, height = 700)
                plot(x,main="",which=2,caption = list(""),sub.caption="")
                SW=shapiro.test(rstandard(x))
                text(x=0,labels=paste("p-value shapiro Wilk =" ,round(SW$p.value,3) ))
                dev.off()
                png(file=paste(nom,"_scatterplot.png"),width = 700, height = 700)
                scatterplotMatrix(data.frame(residus=resid(x),x$model),diagonal
="boxplot",smooth=FALSE,main="")
                dev.off()
                png(file=paste("ajustement_",nom,".png"),width = 700, height = 700)
                plot(x$y,predict(x),xlab="valeurs observées",ylab="valeurs
prédites",xlim=c(0,max(x$y)+1),ylim=c(0,max(predict(x))+1),col="blue",pch=19, main="")
      abline(a=0,b=1,col="red",pch=19)
      abline(coef(lm(predict(x)~x$y)),col="green")
      leg.txt <- c(" droite y=x ", "droite de regression: pred~obs")
      couleur=c("red","green")
      ligne_type=c(4,4)
      pch_type=c(19,19)
      legend(list(x=0,y=max(predict(x))+1), legend = leg.txt, col=couleur, pch=pch_type,lty=ligne_type)
      dev.off()
  }
  #### Test sur les résidus
  test_residus=function(x,alpha){
      print("*********************** TEST SUR LES RESIDUS
**************************************",quote=F)
      print(paste("Test de normalité des résidus pour un risque de première espèce alpha=",alpha),quote=F)
                print(shapiro.test(rstandard(x)),quote=FALSE,digits=2)      # test de normalité des
résidues
                show(bptest(x,data=x$model))
  }
  # M = nombre de out of bag => généralement 10 % du Nb échant.
  traitement_global_model=function(x,donnees,nom,M,alpha){
       sortie_txt(x,donnees,nom,M,alpha)
       analyse_lm(x,donnees)
```

```
        test_residus(x,alpha)
        cross_val_ val (x,donnees,M)
        diagnostic_observations(x,donnees,alpha)
         diagnostic_variables(x)
        diagnostic_graphique(x,nom)
}
```

C.2: Programme SAS pour la PLS

```
/* Loading files */
%macro table_base;
"D:\Stage\sas\St_Andre_version2\analyse_G_PLS_version3_0908.xls"
%mend table_base;
proc import out=dico
DATAFILE=  %table_base
        DBMS=EXCEL REPLACE;
                range="dico$";
   GETNAMES=YES;
run;


PROC IMPORT OUT= tableSAS1
        DATAFILE=  %table_base
        DBMS=EXCEL REPLACE;
                range="tableSAS1$";
   GETNAMES=YES;
run;
/* Loading files */
PROC IMPORT OUT= tableSAS2
        DATAFILE=  %table_base
        DBMS=EXCEL REPLACE;
                range="tableSAS2$";
   GETNAMES=YES;
run;

Data tableSAS;
Merge tableSAS1 tableSAS2;
by IdPlacette;
run;

/***************************** DataSelection *************************/
data GPLS;
set tableSAS;
*if site='haye';
if G2_ha>0;
run;

/***************************** PLS ***********************************/
ods graphics on;

proc pls data=GPLS cv=split cvtest(seed=27 STAT=PRESS);
model G2_ha = V16-V349 /solution; /*ATTENTION REMPLACER VARIABLE*/
/*ATTENTION REMPLACER VARIABLE*/ /*ATTENTION REMPLACER VARIABLE*/
```

```
  output out=outpls predicted = yhat1
                     yresidual = yres1
                     xresidual = xres1-xres15;
  run;
  proc gplot data = outpls;
  plot G2_ha * yhat1;
  run;
  proc reg data=outpls;
  model G2_ha = yhat1; /*Mesures = Simulation*/
  h1: test Intercept=0;
  h2: test yhat1=1;
  h3: test Intercept=0, yhat1 = 1;
  run;
  quit;

  /* REAJUSTEMENT en prennant uniquement le nombre de facteurs optimums, ANALYSE DES
PARAMETRES */
  ods listing close;
  ods output    XLoadings=xloadings
                      PercentVariation  = pctvar
             XWeights         = xweights
             CenScaleParms     = solution
                          ParameterEstimates = solbis;

  proc pls data=GPLS nfac=4 details method=pls; /*ATTENTION REMPLACER NOMBRE*/

  model    G2_ha=    V16-V349    /solution;    /*ATTENTION    REMPLACER    VARIABLE*/
/*ATTENTION REMPLACER VARIABLE */

  output out=outpls predicted = yhat1
                yresidual = yres1;
  run;
  ods listing;
  /* ANALYSE DES PARAMETRES 1, Contribution de chaque frequence à chaque facteur */
  proc transpose data=xloadings(drop=NumberOfFactors)
                out =xloadings;
  run;
  data xloadings; set xloadings;
    n = _n_;
    rename col1=Factor1 col2=Factor2 col3=Factor3 col4=Factor4;
  run;
  border
  axis1 Order=(-0.6 to 0.6 by 0.1) label=("Loading"  ) major=(number=5) minor=none ;
  axis2 label=("Frequency")                minor=none;
  symbol1 v=none i=join c=red    l=1;
  symbol2 v=none i=join c=green  l=1 /*l= 3*/;
  symbol3 v=none i=join c=blue   l=1 /*l=34*/;
  symbol4 v=none i=join c=yellow l=1 /*l=46*/;
  symbol5 v=none i=join c=black l=1 /*l=46*/;
  legend1 label=none cborder=black;
  proc gplot data=xloadings;
  plot (Factor1 Factor2 Factor3 Factor4)*n
```

```
        / overlay legend=legend1 vaxis=axis1
          haxis=axis2 vref=0 lvref=2 frame cframe=ligr;
  run; quit;
  /* ANALYSE DES PARAMETRES 2, Poids de chaque frequence, lesquelles sont importantes */
  data solution; set solution;
    if (RowName = 'Intercept') then delete;
    rename RowName = Predictor G2_ha = B; /*ATTENTION REMPLACER VARIABLE*/
/*ATTENTION REMPLACER VARIABLE*/
  run;
    /*  Transpose weights and R**2's.
    /----------------------------------------------------*/
    data xweights; set xweights; _name_='W'||trim(left(_n_));
    data pctvar  ; set pctvar  ; _name_='R'||trim(left(_n_));
    proc transpose data=xweights(drop=NumberOfFactors InnerRegCoef)
                out =xweights;
  run;
    proc transpose data=pctvar(keep=_name_ CurrentYVariation)
                out =pctvar;
    run;
    /*
    /  Sum the normalized squared weights times the
    /  normalized R**2's.  The VIP is defined as the square
    /  root of this weighted average times the number of
    /  predictors.
    /----------------------------------------------------*/
    proc sql;
      create table vip as
        select *,
          w1 /sqrt(uss(w1)) as wnorm1, /*ATTENTION METTRE A JOUR VARIABLE*/
/*ATTENTION METTRE A JOUR VARIABLE*/
            w2 /sqrt(uss(w2)) as wnorm2,
                        w3 /sqrt(uss(w3)) as wnorm3,
                        w4 /sqrt(uss(w4)) as wnorm4
          from xweights left join pctvar(drop=_name_) on 1;
    data vip; set vip; keep _name_ vip;
      array wnorm{4}; /*ATTENTION REMPLACER NOMBRE*/ /*ATTENTION
REMPLACER NOMBRE*/
      array r{4}; /*ATTENTION REMPLACER NOMBRE*/ /*ATTENTION REMPLACER
NOMBRE*/
      VIP = 0;
      do i = 1 to 4; /*ATTENTION REMPLACER NOMBRE*/ /*ATTENTION REMPLACER
NOMBRE*/
        VIP = VIP + r{i}*(wnorm{i}**2)/sum(of r1-r4); /*ATTENTION REMPLACER
NOMBRE*/ /*ATTENTION REMPLACER NOMBRE*/
        end;
      VIP = sqrt(VIP * 334);
    data vipbpls; merge solution vip(drop=_name_);
    proc print data=vipbpls;
    run;
```

8. Bibliographie

[1] G. DEZ, "Utilisation du LIDAR aéroporté pour l'estimation de la hauteur dominante des peuplements forestiers: Application en forêt de Haye," Office National des Forêts Mémoire de stage de fin d'études, 2008.

[2] F. Martine, "Estimation de la hauteur dominante," Office National des Forêts, INRA nancy rapport de stage, 2009.

[3] R. Genuer, "Forêts aléatoires: aspects théoriques, sélection de variables et applications," 2010.

[4] E. Naesset, "Determination of mean tree height of forest stands using airborne laser scanner data," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 52, pp. 49-56, 1997, cited By (since 1996) 167.

[5] B. P. Magnussen,, "Reconciling multivariate calibrations of volume estimates in two-phase forest inventories," *Forest Science*, vol. 44, pp. 266-271, 1998, cited By (since 1996) 0.

[6] N. M. Olsson, "Simulating the effects of lidar scanning angle for estimation of mean tree height and canopy closure," *Canadian Journal of Remote Sensing*, vol. 29, pp. 623-632, 2003, cited By (since 1996) 40.

[7] A. Y. b. Okano, "Estimation of fIPAR in deciduous forest stands in summer and winter using airborne MSS," in , vol. 7, 2004, pp. 4584-4585.

[8] E. Naesset, "{Practical large-scale forest stand inventory using a small-footprint airborne scanning laser}," *Scandinavian Journal of Forest Research*, vol. 19, pp. 164-179, 2004.

[9] M. Woods, K. Lim, and P. Treitz, "{Predicting forest stand variables from LiDAR data in the Great Lakes--St. Lawrence forest of Ontario}," *The Forestry Chronicle*, vol. 88, pp. 827-838, 2008.

[10] M. Hollaus, W. Wagner, B. Maier, and K. Schadauer, "{Airborne laser scanning of forest stem volume in a mountainous environment}," *Sensors*, vol. 7, pp. 1559-1577, 2007.

[11] Boudreau, "Estimating Quebec provincial forest resources using ICESat/GLAS," *Canadian Journal of Forest Research*, vol. 39, pp. 862-881, 2009, cited By (since 1996) 7.

[12] M. Heurich and F. Thoma, "{Estimation of forestry stand parameters using laser scanning data in temperate, structurally rich natural European beech (Fagus sylvatica) and Norway spruce (Picea abies) forests}," *Forestry*, vol. 81, p. 645, 2008.

[13] P. S. a. Nelson, "Lidar remote sensing of forest biomass: A scale-invariant estimation approach using airborne lasers," *Remote Sensing of Environment*, vol. 113, pp. 182-196, 2009, cited By (since 1996) 18.

[14] C. Y. a. Nelson, "Wildfire smoke injection heights: Two perspectives from space," *Geophysical Research Letters*, vol. 35, 2008, cited By (since 1996) 45.

[15] R. M. Haralick, "Statistical and Structural Approaches to Texture," *Procedings of the IEEE*, vol. 67, pp. 786-804, 1979.

[16] S. C. Popescu, et al., "{A voxel-based lidar method for estimating crown base height for deciduous and pine trees}," *Remote Sensing of Environment*, vol. 112, pp. 767-781, 2008.

[17] O. García, "Estimating top height with variable plot sizes," *Canadian Journal of Forest Research*, vol. 28, pp. 1509-1517, 1998.

[18] J. Pardé and J. Bouchon, *Dendrométrie*, 2èth ed. Ecole Nationale du génie rural, des Eaux et Forêts, 1988.

[19] G. Baskerville, "{Use of logarithmic regression in the estimation of plant biomass}," *Canadian Journal of Forest Research*, vol. 2, pp. 49-53, 1972.

[20] J. Bock, et al., "Towards site index mapping in deciduous stands using multi-echo LIDAR data".

[21] H. M. a. b, "An application-oriented automated approach for co-registration of forest inventory and airborne laser scanning data," *International Journal of Remote Sensing*, vol. 31, pp. 1133-1153, 2010, cited By (since 1996) 3.

[22] Miscellaneous: TO DO.

[23] M. H. Kutner, C. Nachtsheim, and J. Neter, *{Applied linear regression models}*. McGraw-Hill New York, NY, USA, 2004.

[24] T. P. Ryan, "Modern regression methods," 1996.

[25] R. H. Myers, *{Classical and modern regression with applications}*.

[26] T. Hastie, R. Tibshirani, and J. H. Friedman, *{The elements of statistical learning: data mining, inference, and prediction}*. Springer Verlag, 2009.

[27] L. Breiman, "Statistical modeling: The two cultures," *Statistical Science*, vol. 16, pp. 199-215, 2001.

[28] M. Tenenhaus, *La régression PLS: théorie et pratique*. Editions Technip, 1998.

[29] D. DESBOIS, "INTRODUCTION A LA REGRESSION DES MOINDRES CARRES PARTIELS AVEC LA PROCEDURE PLS DE SAS".

[30] L. EYSN, et al., "Adapting alpha shapes for forest delineation using ALS Data".

[31] D. S. Véga,, "Extraction de parametres d'arbre a partir de modeles numeriques de canopee lidar," *Revue Francaise de Photogrammetrie et de Teledetection*, pp. 62-71, 2010, cited By (since 1996) 0.

[32] O. S. c. Couteron, "A generalized, variogram-based framework for multi-scale ordination," *Ecology*, vol. 86, pp. 828-834, 2005, cited By (since 1996) 11.

[33] B. D. Ripley, *Spatial statistics*. {Wiley-IEEE}, 2004.

UNIVERSITÉ DE STRASBOURG

UFR Mathématique-Informatique
Strasbourg

Master 2 mention Mathématiques et Applications : Spécialité Statistique.

RESUME

Ces dernières années, la technologie LiDAR a nettement progressé en foresterie. En Scandinavie par exemple, elle sert actuellement de manière opérationnelle à estimer la ressource forestière. Afin de juger de la précision de son utilisation dans un contexte forestier français, cette étude a été mise sur pied. Elle avait pour objectif d'évaluer la robustesse de l'estimation de 2 paramètres dendrométriques clés (i.e. la hauteur dominante (Ho) et la surface terrière (G)), à partir des nuages de points LiDAR.

Au total, des données de 3 forêts feuillues et 3 forêts résineuses de montage situées dans l'Est de la France, ont été utilisées. Pour les forêts feuillues, l'acquisition LiDAR a été effectuée en hiver (hors feuilles). Pour chacun des sites, 20 à 120 placettes de terrain ont servi à mettre en relation les mesures dendrométriques avec les différents «métriques» issus du LiDAR. Pour l'estimation de Ho, la robustesse de 3 modèles préexistants a été comparée, alors que pour G, de nouveaux modèles ont été établis. Deux procédures de sélection de variables ont été mise en œuvre pour la construction de ces modèles: l'une paramétrique utilisant la PLS, et l'autre, non-paramétrique utilisant le Random Forest.

Les résultats obtenus montrent une précision acceptable des modèles pour l'estimation de Ho, qui est du même ordre de grandeur que l'erreur attendue de mesure. Le modèle de Ho qui intègre un indice spatiale semble le plus robuste, indépendamment de la forêt étudiée (avec une erreur relative inférieur de 7%, sans recalibration des paramètres). Cependant l'estimation de G est plus problématique. Les modèles trouvés dans la littérature sont imprécis et très peu robustes pour le type de forêts que nous avons étudié. Certains de nos modèles peuvent être localement précis lors de la phase de calibration (<7% d'erreur pour un site), mais se révèlent rapidement très bruités voire aberrants (de l'ordre de 20% ou plus) lorsqu'ils sont appliqués à d'autres forêts. Ils nécessitent alors une recalibration. Les erreurs relatives que nous avons obtenu pour les modèles de G varient entre 6.37% et 17.8 % lors des calibrations.

Cette étude montre qu' après quelques adaptations simples, l'utilisation du LiDAR permet d'estimer Ho de façon opérationnelle. Par contre, il s'avère que l'estimation de G est encore trop imprécise. Afin d'améliorer cette situation nous proposons d'explorer l'utilisation d'autres types de « métriques » qui tiendraient compte de l'information spatiale contenue dans le nuage de points LiDAR. Des indices de textures, ou l'utilisation de voxels (cubes) pourraient s'avérer complémentaire aux « métriques » utilisés jusqu'à présent, qui sont basés uniquement sur la distribution en hauteur des nuages lidar.

MOTS CLES : LiDAR, inventaire forestier, Random Forest, PLS, validation de modèles

AGENCE NATIONALE DE LA RECHERCHE ANR

Office National des Forêts
Département R&D

FORESEE
Forest Resource Estimation for Energy